Computers & Geosciences 106 (2017) 109–117

Contents lists available at ScienceDirect

Computers & Geosciences

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

Research paper

Optimal estimation of areal values of near-land-surface temperatures for testing global and local spatio-temporal trends

Hong Wang[a,b], Eulogio Pardo-Igúzquiza[b], Peter A. Dowd[c], Yongguo Yang[a,*]

[a] School of Resources and Geosciences, China University of Mining and Technology, XuZhou, JiangSu Province 221116, China
[b] Geological Survey of Spain, Ríos Rosas 23, 28003 Madrid, Spain
[c] University of Adelaide, Australia

ARTICLE INFO

Keywords:
Constrained spatial clustering
Temperature-altitude correlation
Regression kriging
Time series
Temperature trend detection
Global warming

ABSTRACT

This paper provides a solution to the problem of estimating the mean value of near-land-surface temperature over a relatively large area (here, by way of example, applied to mainland Spain covering an area of around half a million square kilometres) from a limited number of weather stations covering a non-representative (biased) range of altitudes. As evidence mounts for altitude-dependent global warming, this bias is a significant problem when temperatures at high altitudes are under-represented. We correct this bias by using altitude as a secondary variable and using a novel clustering method for identifying geographical regions (clusters) that maximize the correlation between altitude and mean temperature. In addition, the paper provides an improved regression kriging estimator, which is optimally determined by the cluster analysis. The optimal areal values of near-land-surface temperature are used to generate time series of areal temperature averages in order to assess regional changes in temperature trends. The methodology is applied to records of annual mean temperatures over the period 1950–2011 across mainland Spain. The robust non-parametric Theil-Sen method is used to test for temperature trends in the regional temperature time series. Our analysis shows that, over the 62-year period of the study, 78% of mainland Spain has had a statistically significant increase in annual mean temperature.

1. Introduction

Changes in near-land-surface temperatures are perhaps the most common and reliable indicator of global warming (Robeson, 1994). Near-land-surface temperature is usually measured at a finite number of irregularly spaced sampling locations comprising networks of weather stations. Although temperature measurements are affected by many factors, including longitude, latitude, altitude, slope orientation, atmospheric circulation and proximity to the sea, altitude is the most significant variable and explains most of the spatially dependent variance in temperature (Hudson and Wackernagel, 1994). In mountainous areas, altitude is the simplest direct measurement that is most highly correlated with temperature (Dodson and Marks, 1997; Benavides et al., 2007). The correlation is usually linear and negative so that temperature decreases as altitude increases with, in general, a mean gradient of 0.6 °C per 100 m of altitude (Viers, 1975). However, for large areas (several degrees of latitude), the many other factors listed above may affect the temperature in such a way that the linear relationship between altitude and temperature is much weaker because, for example, different climate factors are merged within the large area. For example, for mainland Spain the Mediterranean marine influence is different to the Atlantic marine influence. In addition, temperature measurements are biased because weather stations tend to be located at low altitudes (Rolland, 2002) and areas at high altitudes (for example, mountainous areas) are poorly represented (Robeson, 1994). This under-representation is particularly important as evidence mounts for altitude-dependent global warming (see, for example, Pepin and Lundquist, 2008 and Mountain Research Initiative EDW Working Group, 2015). Fig. 1 shows a histogram of altitudes obtained from a digital elevation model (DEM) of mainland Spain together with a histogram of the altitudes of weather stations for the year 1994. This figure shows that 25% of the surface of mainland Spain has altitudes less than 400 m and 20% of the surface has altitudes greater than 1000 m; whereas, 42% of the temperature monitoring stations (i.e., the data collection points) are located at altitudes less than 400 m and only 10% of the stations are at altitudes greater than 1000 m. This problem can be solved by using the DEM altitude as a secondary variable together with the linear relationship between altitude and temperature. However, the correlation of altitude and temperature over large areas is relatively small because of the influence of other factors such as

* Corresponding author.
E-mail addresses: wanghongcumt@hotmail.com (H. Wang), e.pardo@igme.es (E. Pardo-Igúzquiza), peter.dowd@adelaide.edu.au (P.A. Dowd), ygyang88@hotmail.com (Y. Yang).

http://dx.doi.org/10.1016/j.cageo.2017.06.002
Received 6 February 2017; Received in revised form 25 May 2017; Accepted 5 June 2017
Available online 06 June 2017
0098-3004/ © 2017 Elsevier Ltd. All rights reserved.

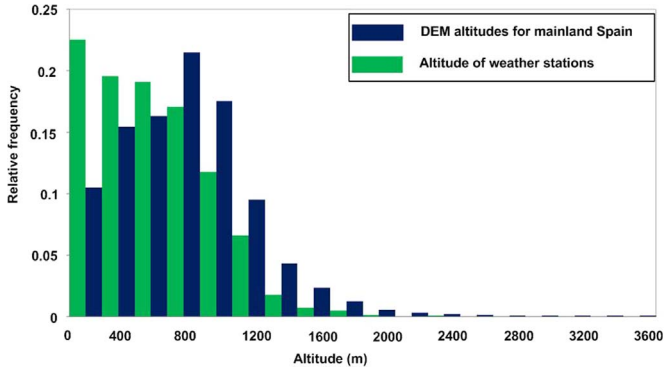

Fig. 1. Histogram of digital elevation model (DEM) altitudes for mainland Spain (blue) and altitudes of weather stations for the year 1994 (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

latitude, longitude, proximity to the sea, pressure and wind patterns. Thus, it is useful to identify zones in which the correlation between altitude and temperature is as strong as possible. Cluster analysis is highly suited to this purpose.

Clustering algorithms (Stooksbury and Michaels, 1991; Fovell and Fovell, 1993; Gerstengarbe et al., 1999; DeGaetano, 2001; Unal et al., 2003; Huth et al., 2008; Mahlstein and Knutti, 2010; Cannon, 2012; Tang et al., 2012; Zscheischler et al., 2012) have been used for similar, but not identical, problems to the one dealt with here. In this work we propose a new cluster method that has two novel aspects. The first is the recognition that the problem is a particular form of a constrained cluster analysis problem. The second is accounting for the spatial correlation of the data when testing the spatial correlation of the residuals of the regression of temperature on altitude for the clusters. There is no requirement for the obtained clusters to coincide with climatic regions because the definition of the latter differs from that of the obtained clusters. For example, the Spanish state meteorological agency (la Agencia Estatal de Meteorología) defines climate regions on the basis of the Köppen-Geiger Climate Classification (AEMET, 2011), which is a classification system based on the assumption that native vegetation is the best expression of climate. The purpose of our clustering approach is not to identify climate regions but to obtain regions with a high correlation between temperature and altitude. The regions resulting from the cluster analysis are not interpreted climatologically, they are used solely to obtain optimal estimates of mean areal temperatures. In addition, the regional clusters implicitly take account of secondary variables such as latitude, longitude and proximity to the sea. A detailed explanation of the methodology employed in this study is given in the following section.

2. Methodology

Geostatistical methods are widely used for mapping temperature (Hudson and Wackernagel, 1994) and estimating areal values of temperature (Ishida and Kawashima, 1993). The mean areal value of temperature over a particular area is defined by:

$$\overline{T} = \frac{1}{\chi}\int_{\chi} T(u)du \tag{1}$$

where $\chi \subset \mathcal{R}^2$ is the zone of interest of finite area and $T(u)$ is the temperature at the spatial point location $u \in \chi$.

The integral in Eq. (1) is approximated by summing the temperatures of a discrete pixel or small cell representation of the zone of interest:

$$\overline{\overline{T}} = \frac{1}{k}\sum_{i=1}^{k} T(u_i) \tag{2}$$

where k is the number of discrete cells comprising the zone $\chi \subset \mathcal{R}^2$ and $T(u_i)$ is the temperature at the i^{th} cell.

The value $T(u_i)$ is usually unknown and must be estimated from a finite set of data values. To avoid the bias introduced by the data (because of over-representation of low altitudes) and to account for the correlation between altitude and temperature, the altitude of each cell is determined from a DEM of the zone of interest. For very large zones, such as mainland Spain with a surface area of 492,072 km^2, the relationship between altitude and temperature would be obscured if data from all temperature stations were considered together. This is because the topography of the Iberian Peninsula is complex and there are many specific effects that change with latitude and longitude; for example, the different Atlantic and Mediterranean marine influences, the different frequencies of easterly winds in the Mediterranean area and westerly winds in the Atlantic area, the heating and cooling of hillsides depending on their orientation and perturbation effects such as the incursion of relatively cold air masses from the Atlantic. For these reasons we divide the zone of interest into smaller areas in which the relationship between altitude and temperature is stronger (higher negative correlation between altitude and temperature). These areas, which maximize the correlation between altitude and temperature, are identified by a new cluster analysis procedure.

Classical cluster analysis identifies groups of objects that are similar. It does so by maximising the similarity of objects (in our case, temperature measurements from weather stations) within a group and maximising the dissimilarity of different groups of objects (Gordon, 1996). There are two broad types of clustering methods: hierarchical clustering and non-hierarchical clustering. Among the non-hierarchical clustering algorithms the most widely used is the k-means algorithm. The similarity of objects is usually defined in terms of a distance (e.g., Euclidean, Mahalanobis) according to the measured characteristics of the objects.

For the problem addressed in this paper, the first difference with respect to classical clustering is that, instead of defining the similarity measure as a distance between the objects of a group, it is an objective function to be maximised or minimised. The second difference is that the problem addressed in this paper is a case of constrained clustering in which a contiguity constraint restricts the sets of allowable solutions (Gordon, 1996), i.e., the objects in each group must comprise a spatially contiguous set. Thus, given a number of groups, an object can change its membership from group A to group B if two requirements are met: (i) groups A and B are contiguous and (ii) the value of the objective function is improved. Clustering temperatures into regions with high linear correlation between altitude and temperature can thus be seen as a contiguity-constrained optimisation problem.

The first issue is the definition of clusters and contiguity. The locations of the weather stations are used as the seeds of a Voronoi tessellation of the geographic space covered by the stations. A cluster, or group, of weather stations (or of the corresponding temperature measurements) is a union of contiguous Voronoi cells and the boundary of the cluster is the outermost sequence of its constituent cell boundaries. Two clusters are contiguous if they share a boundary. A member, or object, belonging to cluster A is contiguous with cluster B if its Voronoi cell shares a boundary with the Voronoi cell of any member of cluster B. These definitions are used in the application of the contiguity constraint.

In the proposed algorithm for contiguity-constrained classification of a set of N objects (weather stations) the algorithm starts with an exhaustive classification into M groups. The manner in which this starting classification is obtained is described below. The classification is exhaustive in the sense that the N objects have been classified and each belongs to one of the M groups.

For any given configuration of groups $(G_1, \ldots, G_M)$ the objective function, $OF(G_1, \ldots, G_M)$, of the configuration is defined by:

$$OF(G_1, \ldots, G_M) = \sum_{i=1}^{M} n_i r_i^* \tag{3}$$

where n_i is the number of objects that belong to the i^{th} group and r_i^* is the value of the modified Pearson product-moment correlation coefficient:

$$r_i^* = r_i + 1.96\sigma(r_i) \tag{4}$$

where r_i is the estimated Pearson correlation coefficient of the i^{th} group and $\sigma(r_i)$ is the associated standard error. The correlation between altitude and temperature is negative and thus, from Eq. (4), $r_i < r_i^*$. The modified coefficient, r_i^*, can be used instead of r_i as an experimental measure of correlation between altitude and temperature that accounts for the size of the group, i.e., the uncertainty of the estimated value of the correlation coefficient as quantified by the standard error in Eq. (4). The value $1.96\sigma(r_i)$ is the lower bound of the 95% confidence interval and has been chosen as a conservative value for including cluster size in the comparison of the correlation coefficients of two different clusters.

The estimated Pearson product-moment correlation coefficient, r_i, for the i^{th} group is:

$$r_i = \sum_{j=1}^{n_i} (T_{ij} - \overline{T_i})(H_{ij} - \overline{H_i}) \Bigg/ \left(\sqrt{\sum_{j=1}^{n_i} (T_{ij} - \overline{T_i})^2} \sqrt{\sum_{j=1}^{n_i} (H_{ij} - \overline{H_i})^2} \right) \tag{5}$$

where T_{ij} is the temperature at the j^{th} station of the i^{th} group and H_{ij} is the altitude of the j^{th} station of the i^{th} group. Weather stations are thus defined by their pair (T, H) of temperature and altitude. Stations are also defined by their geographical co-ordinates (X, Y), which are implicitly included in the proposed methodology by the contiguity constraint.

It is obvious that

$$OF(G_1, \ldots, G_M) \geq -N \tag{6}$$

with the minimum value $-N$ being unattainable in practice because it would imply the unlikely case of a perfect (negative) correlation between altitude and temperature for a given partition $(G_1, \ldots, G_M)$, with correlation coefficient of -1 for each group of the partition.

The standard error, $\sigma(r_i)$, of the estimated Pearson correlation coefficient can be calculated by a parametric method such as Student's t-distribution or by using a non-parametric method such as the bootstrap. The advantage of the latter is that it works when the sampling distribution of the correlation coefficient is asymmetrical and the data are (spatially) correlated, as is the case in the application described here.

The clustering process is applied to each set of annual temperatures. In the clustering process there are two permitted operations: coalescence of two groups and moving an element from one group to another. Both operations use the definitions of contiguity given above.

Two groups G_i and G_j will coalesce to form a new group, G_k, if

1) The two groups, G_i and G_j, are contiguous.
2) The value of the objective function improves.

An object, o_k, that belongs to group G_i can move to group G_j if:

1) The object o_k is contiguous with the group G_j.
2) The value of the objective function improves.

Finally, the constrained clustering algorithm comprises the following steps:

(i) Start with an initial partition of M clusters, where M is greater than the expected optimal number of clusters. For example, $M = 100$ is used in the case study. Select at random M stations from the total N stations ($N > M$). Call these locations the seed stations. Next, each of the N stations is assigned to the nearest seed station and then M groups that satisfy the contiguity constraint are formed. This is the starting random partition with M groups:

$$\{G_1, G_2, \ldots, G_M\} \tag{7}$$

for which the initial value of the objective function is given by Eq. (3).

(ii) Each group, $\{G_i\}$, is taken in turn and the closest contiguous group (in terms of geographical distance) $\{G_j\}$ is found. The closest group is the one that contains an object that has the shortest geographical distance (calculated using geographical co-ordinates (X, Y)) of a member of the group $\{G_i\}$. Obviously the two objects and the two groups are contiguous. Three operations are then tried. Operation 1 ($O1$): if either of the two groups has less than n_{min} objects, the groups are merged; Operation 2 ($O2$): try to merge both groups; Operation 3 ($O3$): try to move an object from one group to the other. The purpose of $O1$ is that if either of the two groups has a small number of elements, defined by the threshold value n_{min} (for example, $n_{min} = 20$ is used in the case study), there is no point in calculating a very unreliable correlation coefficient and thus the groups are merged to create a larger group. The meaning of $O2$ is that two groups are merged if the resulting group is better than the worse of the two groups. In other words, operation 2 consists in merging group $\{G_i\}$ and group $\{G_j\}$ into a single group if the following criterion is satisfied:

$$r_{ij}^* \leq max(r_i^*, r_j^*) \tag{8}$$

where r_{ij}^* is the modified Pearson correlation coefficient, defined in Eq. (4), for the merged group $\{G_{ij}\} = \{G_i\} \cup \{G_j\}$. Note that a maximum operator is used in Eq. (8) because the correlations are negative and groups with the largest possible negative correlation are required.

(iii) If the operation of merging the two groups fails (because Eq. (8) is not satisfied) then the operation of moving an element to the closest group is tried. If the pair of elements $\{o_{ik}, o_{jl}\}$ are the two closest elements between groups $\{G_i\}$ and $\{G_j\}$, such that $\{o_{ik}\} \in \{G_i\}$ and $\{o_{jl}\} \in \{G_j\}$, there are two possibilities to try: (1) station $\{o_{ik}\}$ leaves group $\{G_i\}$ and joins group $\{G_j\}$ and (2) station $\{o_{jl}\}$ leaves group $\{G_j\}$ and joins group $\{G_i\}$. Note that the possibility of both stations swapping groups is not allowed because it violates the contiguity constraint. Thus, for possibility (1), let r_i^* and r_j^* be the correlation coefficients of $\{G_i\}$ and $\{G_j\}$ respectively and let r_{i-}^* and r_{j+}^* be the correlation coefficients of $\{G_{i-}\}$ and $\{G_{j+}\}$, where $\{G_{i-}\} = \{G_i\} - \{o_k\}$, i.e., group $\{G_i\}$ without object $\{o_k\}$, and $\{G_{j+}\} = \{G_j\} + \{o_k\}$ is group $\{G_j\}$ with object $\{o_k\}$ added. The proposal to move an object from group i to group j is accepted if

$$\max(r_{i-}^*, r_{j+}^*) < max(r_i^*, r_j^*) \tag{9}$$

where max(A, B) is the operator that selects the maximum of A and B.

Similarly, for possibility (2), the proposal to move an object from group j to group i is accepted if

$$\max(r_{i+}^*, r_{j-}^*) < max(r_i^*, r_j^*) \tag{10}$$

Combining (9) and (10), and noting that both conditions cannot hold simultaneously, gives:

$$\min\{\max(r_{i-}^*, r_{j+}^*), \max(r_{i+}^*, r_{j-}^*)\} < max(r_i^*, r_j^*) \tag{11}$$

(iv) Go to (ii) to operate on the next group until all M groups have been visited in turn.

Obviously this process could be repeated many times with a

different starting number of groups or with the same starting number of groups but with different group configurations (i.e., different station numbers chosen as seeds) or with a different collection of random numbers used in the merging process of forming groups. In our work we repeat it for each year of the time span considered. Next, the solution for each year is used as the starting configuration for each year and the cluster that is the best for all the years is selected. The unique final cluster is chosen such that it is better than the solution obtained by considering one cluster for every year. However, the main reason for having only one cluster classification is to avoid temporal discontinuities in the time series at the pixel level (i.e., time series of the pixel temperatures). As each pixel belongs to the same cluster for all years there are no temporal discontinuities. To avoid spatial discontinuities between clusters, a local moving window is used to smooth the estimated slope and the intercept for each cluster. This smoothing only affects borders between clusters because inside a cluster the slope and intercept values are constant. Spatial discontinuities are minimal and appear only at the contacts between the clusters; they are of no significance because our purpose is to provide optimal estimates of areal temperatures at the local (pixel) level and/or at the Spanish mainland level (aggregation of all pixels).

The final configuration of groups provides a means of estimating regional temperatures for all years as follows:

(i) Use the DEM of the region of interest to provide a discrete cell representation of the region.

(ii) For any cell of any group use a linear regression of temperature on altitude to estimate temperature from altitude:

$$T^*(u_0) = a + bH(u_0) \tag{12}$$

where $T^*(u_0)$ is the estimated temperature over a given cell located at the coordinates of its centre $u_0 = \{x_0, y_0\}$; $H(u_0)$ is the altitude of the cell and a and b are the least squares estimates of intercept and slope of the regression line. The estimation variance of the temperature estimate in Eq. (12) is given by :

$$V(T^*(u_0)) = s^2\left(1/n + (H(u_0) - \overline{H})^2 / \sum_{i=1}^{n} (H(u_i) - \overline{H})^2\right) \tag{13}$$

where n is the number of data in the cluster and $\overline{H}$ is the mean of the altitudes of the n data in the cluster:

$$\overline{H} = 1/n \sum_{i=1}^{n} H(u_i) \tag{14}$$

s^2 is the estimated variance of the regression residuals $\{R(u_i) = T^*(u_i) - T(u_i); i = 1, \ldots, n\}$:

$$s^2 = 1/(n-2) \sum_{i=1}^{n} (R(u_i))^2 \tag{15}$$

The final estimate of the temperature of a cell at location u_0 is given by the temperature estimated by the regression line plus the estimate of the residual, which accounts for the variability not explained by the regression:

$$T^{**}(u_0) = T^*(u_0) + R^*(u_0) \tag{16}$$

where $T^*(u_0)$ is given by Eq. (12) and $R^*(u_0)$ is the ordinary kriging (Olea, 1999) estimate of the residual:

$$R^*(u_0) = \sum_{j=1}^{m} \lambda_j R(u_j) \tag{17}$$

where m is the number of neighbours of, or locations close to, u_0, the residuals of which are used in the estimation process; λ_j is the weight assigned to the j^{th} residual in the estimation of the residual at u_0. These weights are obtained by solving the kriging system, which can be found in any standard textbook on geostatistics (e.g., Olea, 1999). In geostatistics, the estimator given in Eq. (16) is known as regression kriging (Hengl et al., 2007). The variance of the estimator given in Eq. (17) is the kriging estimation variance:

$$\sigma^2(u_0) = \mathbf{B}^{\mathrm{T}}\boldsymbol{\lambda} \tag{18}$$

where $\mathbf{B}$ and $\boldsymbol{\lambda}$ are described in Olea (1999). Because the regression estimates are independent of the kriged estimates, the uncertainty of the estimator given in Eq. (16) is the sum of the uncertainties:

$$V(T^{**}(u_0)) = V(T^*(u_0)) + \sigma^2(u_0) \tag{19}$$

(iii) Any regional average obtained as the union of a given number of K cells, can be calculated as the mean of the estimates of the K cells:

$$\chi = \bigcup_{i=1}^{K} u_i \tag{20}$$

$$T^{**}(\chi) = 1/K \sum_{i=1}^{K} T^{**}(u_i) \tag{21}$$

A measure of the uncertainty of the estimator in Eq. (16) is given by the variance of the mean of the estimates of the K cells:

$$V(T^{**}(\chi)) = 1/K^2 \sum_{i=1}^{K} \sum_{j=1}^{K} \mathrm{Cov}(T^{**}(u_i), T^{**}(u_j)) \tag{22}$$

where $\mathrm{Cov}(T^{**}(u_i), T^{**}(u_j))$ is the covariance of the estimates for cells i and j. The off-diagonal terms ($i \neq j$) in Eq. (22) are:

$$\begin{aligned} &\mathrm{Cov}(T^{**}(u_i), T^{**}(u_j)) = \mathrm{Cov}(T^*(u_i), T^*(u_j)) + \mathrm{Cov}(R^*(u_i), R^*(u_j)) \\ &\mathrm{Cov}(T^*(u_i), T^*(u_j)) = \boldsymbol{X}_i^T \boldsymbol{C}_{Reg} \boldsymbol{X}_j \\ &\mathrm{Cov}(R^*(u_i), R^*(u_j)) = \boldsymbol{\lambda}_i^T \boldsymbol{C}_R \boldsymbol{\lambda}_j \end{aligned} \tag{23}$$

where:

$$\boldsymbol{X}_i^T = [1 H(u_i)]$$

$$\boldsymbol{\lambda}_i^T = [\lambda_1 \lambda_2 \ldots \lambda_m]$$

m : number of neighbours used in the estimation of $R^*(u_i)$
$\boldsymbol{C}_{Reg}$: variance-covariance matrix of the two coefficients of the linear regression
$\boldsymbol{C}_R$: covariance of the residual.

We use a two-step approximation of Eq. (22). First, we follow the widely used convention in global estimation of ignoring the off-diagonal terms on the assumption that they are generally small with respect to the diagonal terms (see, for example, Matheron, 1962; Journel and Huijbregts, 1978, 2003; David, 1977, 1982). In our application, the error incurred in this approximation decreases as the size of the cell increases; in addition, as groups do not have common data, the approximation error reduces even further. Second, we approximate the diagonal terms in Eq. (22) by

$$\mathrm{Cov}(T^{**}(u_i), T^{**}(u_i)) = V(T^{**}(u_i)) \approx V(T^*(u_i)) + \sigma^2_{Nugget}(u_i) \tag{24}$$

$\sigma^2_{Nugget}(u_i)$ is the kriging variance as in Eq. (18) but assuming a pure nugget variance for the variogram model (i.e., no spatial correlation). This is a conservative approximation because:

$$\sigma^2_{Nugget}(u_i) \geq \sigma^2(u_i) \tag{25}$$

Eqs. (21) and (22) give the areal estimates of annual mean temperature and the uncertainty of the estimate for any given year and any region or sub-region. Thus, repeating the estimations for a number of years gives a climatic time series for a given region or sub-region.

The time series, starting in year 1 and ending in year N, is obtained as:

$$\{T_1^{**}(\chi),\ T_2^{**}(\chi),\ldots,\ T_N^{**}(\chi)\} \tag{26}$$

and the general linear regression equation fitted to it can be written as:

$$\vec{T} = \beta_0 + \beta_1 T^{**} \tag{27}$$

where β_0 and β_1 are, respectively, the intercept at the origin and the slope.

3. Case study

3.1. Study area and research materials

The study area of this work is all of mainland Spain. It is bordered to the northeast by France; to the south and the east by the Mediterranean Sea; and to the west and northwest by Portugal and the Atlantic Ocean. A DEM of mainland Spain with a resolution (i.e., cell size) of 0.8 km × 0.8 km is used in this study to estimate the regional temperature so as to include the effect of elevation on average temperature. We used the Universal Transverse Mercator (UTM) projection and selected UTM-30N as the reference system. Temperatures are recorded daily at the official mainland weather stations of Spanish Meteorological Agency (Agencia Estatal de Meteorologia). We used annual mean temperature (temperatures averaged over a whole year) for 62 years (1950–2011) and for a given number of stations. Over the 62-year period, the number of stations at which annual mean temperatures were recorded ranges from a minimum of 306 in 1950 to a maximum of 1361 in 1994. The primary reason for this variation in the number of stations is the failure to record temeperatures due to severe weather conditions or lack of maintenance of meteorological equipment. For illustrative purposes,

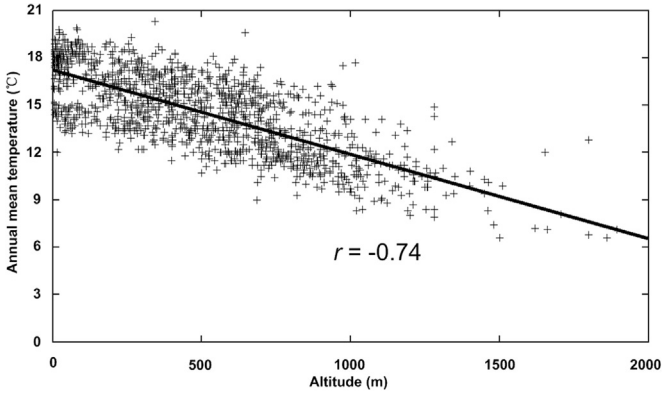

Fig. 3. Linear relationship between altitude and annual mean temperature for mainland Spain, as recorded at the 1303 stations for the year 2000; r is the correlation coefficient and the black line is the linear regression line.

Fig. 2 shows the locations of the 1303 stations that recorded temperatures for the year 2000.

3.2. Results of constrained spatial clustering

Scatterplots of altitude and annual mean temperature for each year (for example Fig. 3 shows the scatterplot for the year 2000), show a clear linear relationship but with significant dispersion within a broad band. In Fig. 3 the dispersion of temperatures is approximately 7 °C, which is most evident for altitudes less than around 1000 m comprising almost 90% of the weather stations. The total correlation coefficient is −0.74 with a 95% confidence interval of [−0.76, −0.72]. Student's t-statistic was used to calculate this interval; for such a large sample, the non-parametric bootstrap evaluation gives virtually the same value.

When applying the proposed clustering approach, the uncertainty introduced by the possibility of choosing among different configurations has been evaluated and shown to be very small. Hence, empirically we set the initial number of groups M = 100 and the threshold number of elements for merging (operation 1) n_{min} = 20 for

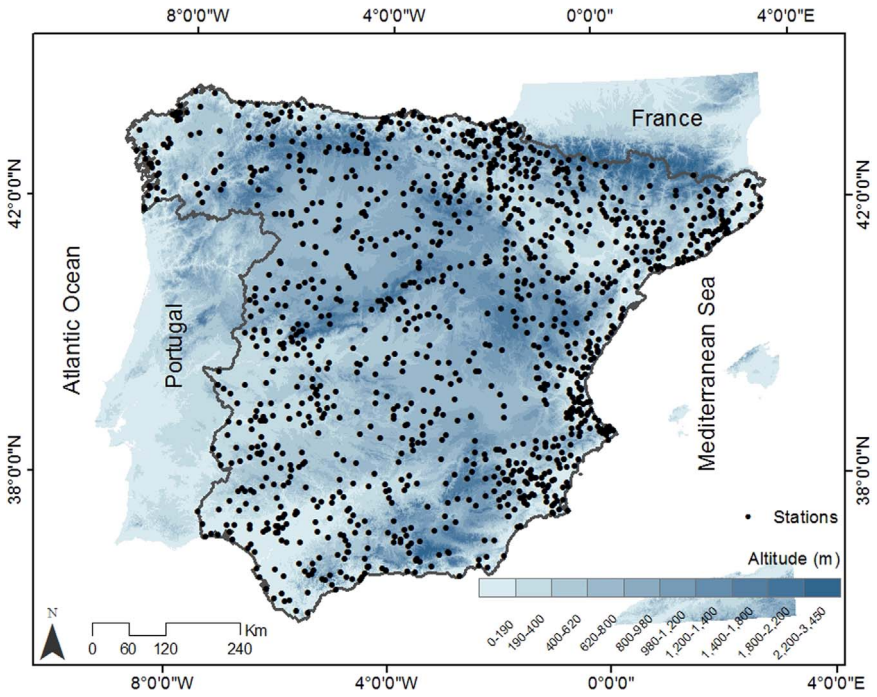

Fig. 2. Locations on the mainland Spain DEM of the 1303 stations at which temperature was recorded for year 2000.

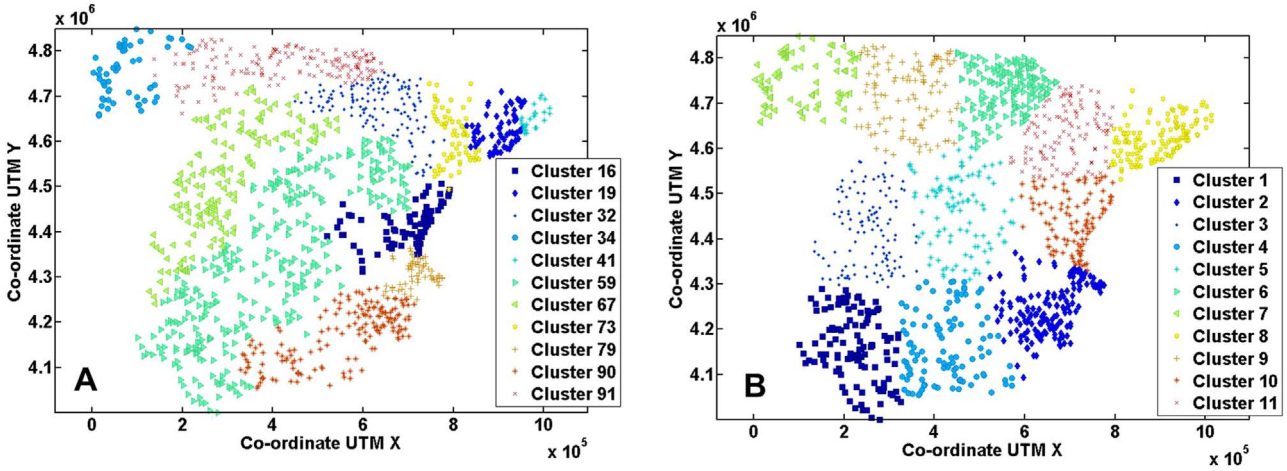

Fig. 4. Clusters obtained by A: constrained spatial clustering algorithm; B: *k*-means algorithm.

all years. Clusters retained by the spatial clustering algorithm are shown in Fig. 4A for year 2000. When applied to all years, the minimum value of the objective function (Eq. (3)) is achieved for a configuration of 11 groups for year 2000; this configuration was then used for all years.

For comparison, we use the classical *k*-means clustering method, implemented using coordinate X, coordinate Y, altitude and temperature in standardized units. For comparative purposes, the two methods should have the same number of groups. Thus, for all years we have specified 11 groups, the number that gave the minimum objective function value using our proposed clustering method. The initial number of groups was 11 with centriods chosen at random for all years. The *k*-means algorithm was applied to the stations for each of the 62 years and the results for year 2000 are shown in Fig. 4B.

The scatterplots of altitude and annual mean temperature for each spatially constrained cluster are shown in Fig. 5 in which stations with different symbols and colours represent the different clusters. The 11 clusters are clearly distinguishable in these plots and the stations in each cluster display a strong (negative) correlation between altitude and annual mean temperature. In particular, the weakest correlation coefficient among all spatial clusters is −0.86, which is significantly higher than the correlation coefficient (−0.74) for all temperatures taken as a single group and the weakest correlation coefficient (−0.74) for the *k*-means clusters. The strongest correlation coefficient for the spatial clusters is −0.95, which is marginally higher than the strongest value (−0.94) of the *k*-means clusters. The objective function value (Eq. (3)) for the 11 retained clusters is −1113.30, which is less than the value of −918.30 when all stations are considered as one group and the value of −1025.70 for the *k*-means algorithm. These results demonstrate the ability of the constrained spatial clustering approach to identify geographical regions with a strong linear relationship between altitude and annual mean temperature. In addition, the use of the modified correlation coefficient in the spatial clustering algorithm is a simple way of accounting for the effects of the uncertainty caused by different cluster sizes. Similar results were obtained for other years.

Another preliminary aspect to evaluate is the magnitude and spatial variability structure of the regression. Fig. 6 shows the omni-directional semi-variogram of the regression residuals for all 11 constrained clusters for the year 2000 taken together. The fitted model is a nugget variance of 0.45 $(°C)^2$ and a spherical model with a structural variance of 0.34 $(°C)^2$ and a range of approximately 130 km. The variance of the temperatures for the year 2000 is 9.5$(°C)^2$ and thus the structural variance of the residual represents only 3.5% of the total temperature variance. The regression thus explains most of the temperature variability. Similar results were obtained for other years.

3.3. Trend detection in optimal temperature time series

The time series shown in Fig. 7 was obtained by using the improved regression kriging estimator to calculate the annual mean temperature over mainland Spain for each year of the period 1950–2011. In Fig. 7, this time series shows a significant variability in estimated annual mean temperature from year to year of up to 2.4 °C between 1956 and 2011. This raises the question of whether there is any statistically significant trend in the time series in Fig. 7. This question can be answered by applying statistical methods of trend detection together with the principle of parsimony. The latter suggests a simple trend model, such as the linear trend shown in Fig. 7, which would indicate (if confirmed as statistically significant at a given confidence level such as 95%) that there has been an increase in mean temperature over the entire time interval [1950,2011]. The slope of the linear trend of this time series is 0.014 with a 95% confidence interval of [0.0071, 0.0211]. The slope is statistically significant and supports the evidence for warming over mainland Spain over the 62-year time interval.

Results for the non-parametric Theil-Sen estimator (Theil, 1950; Sen, 1968) are shown in Fig. 8. Fig. 8A shows the slope of the local (pixel) time series for the time interval [1950,2011]; Fig. 8B and C show the lower and upper limits respectively of a 95% confidence interval for the values plotted in Fig. 8A. Because of the equivalence between confidence intervals and hypothesis testing, if the slope is positive in Fig. 8A and the lower and upper limits are also positive, then the positive slope is statistically significant with a significance level of 0.05, or with a confidence level of 95%. Alternatively, if the slope is negative in Fig. 8A and the lower and upper limits are both negative, then the negative slope is statistically significant with a significance level of 0.05, or with a confidence level of 95%. Note that there are areas with negative slope indicating that the mean temperature is decreasing with time. However, these areas of negative slope represent only 1% of the total surface area of mainland Spain (Fig. 9A). As can be seen in Fig. 9A, 99% of the total area has a positive slope and 78% of the total area has a statistically significant positive slope (pink area in Fig. 9B). Thus it is clear that the annual mean temperature has increased over most of mainland Spain from 1950 to 2011.

4. Conclusions

We have introduced a constrained clustering algorithm for identifying geographical regions that maximize the correlation between altitude and mean temperature. This algorithm provides a means of estimating mean areal temperatures by using a digital elevation model that accounts for temperature information in areas where there are no

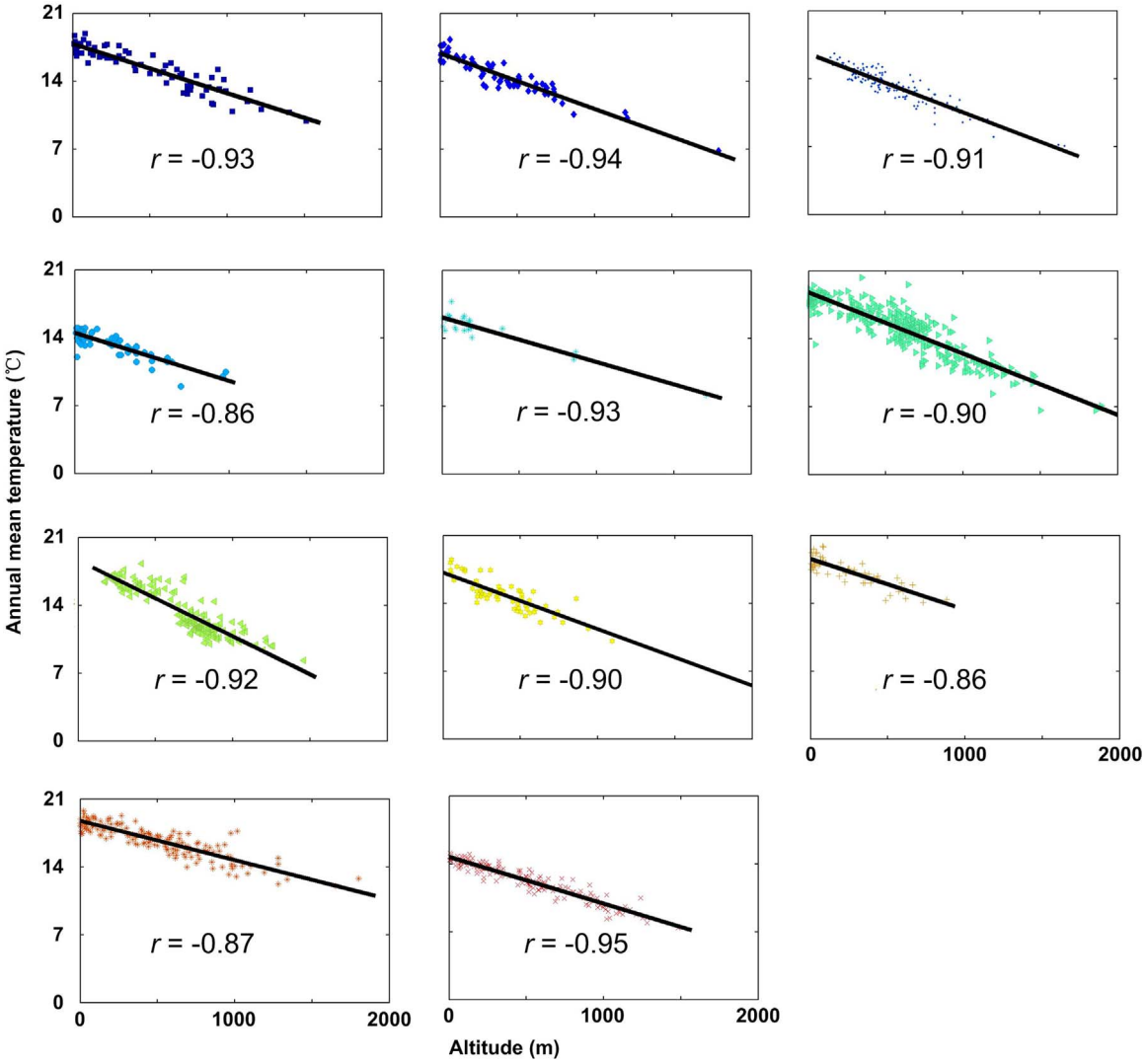

Fig. 5. Scatterplots of altitude and annual mean temperature based on constrained spatial clusters for year 2000.

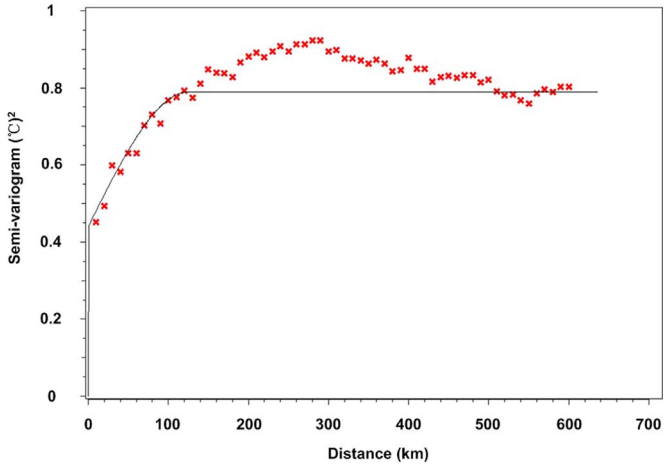

Fig. 6. Omni-directional experimental semi-variogram (red crosses) of the regression residual for the year 2000 and the fitted model (black line): a spherical model with a nugget variance of 0.45(℃)^2, a structural variance of 0.34(℃)^2 and range of approximately 130 km. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

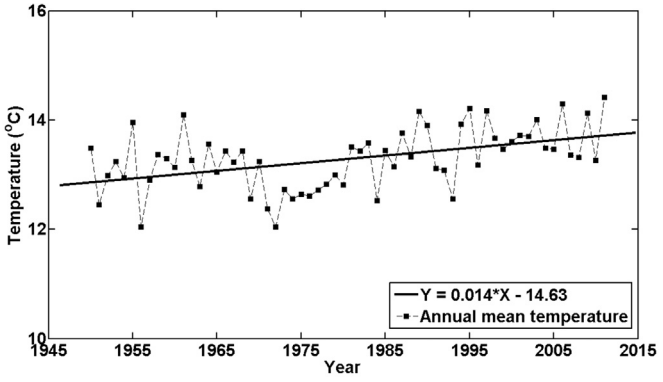

Fig. 7. Estimated mean yearly temperature for the period 1950–2011 (black dots) and fitted regression line (black line).

temperature measurements. The algorithm was used to calculate the time series of annual mean temperatures for the whole of mainland Spain. These results were then used to test an hypothesis of linear trend in mean temperature for the period of observation (from 1950 to 2011). We have shown that 99% of the total area of mainland Spain has

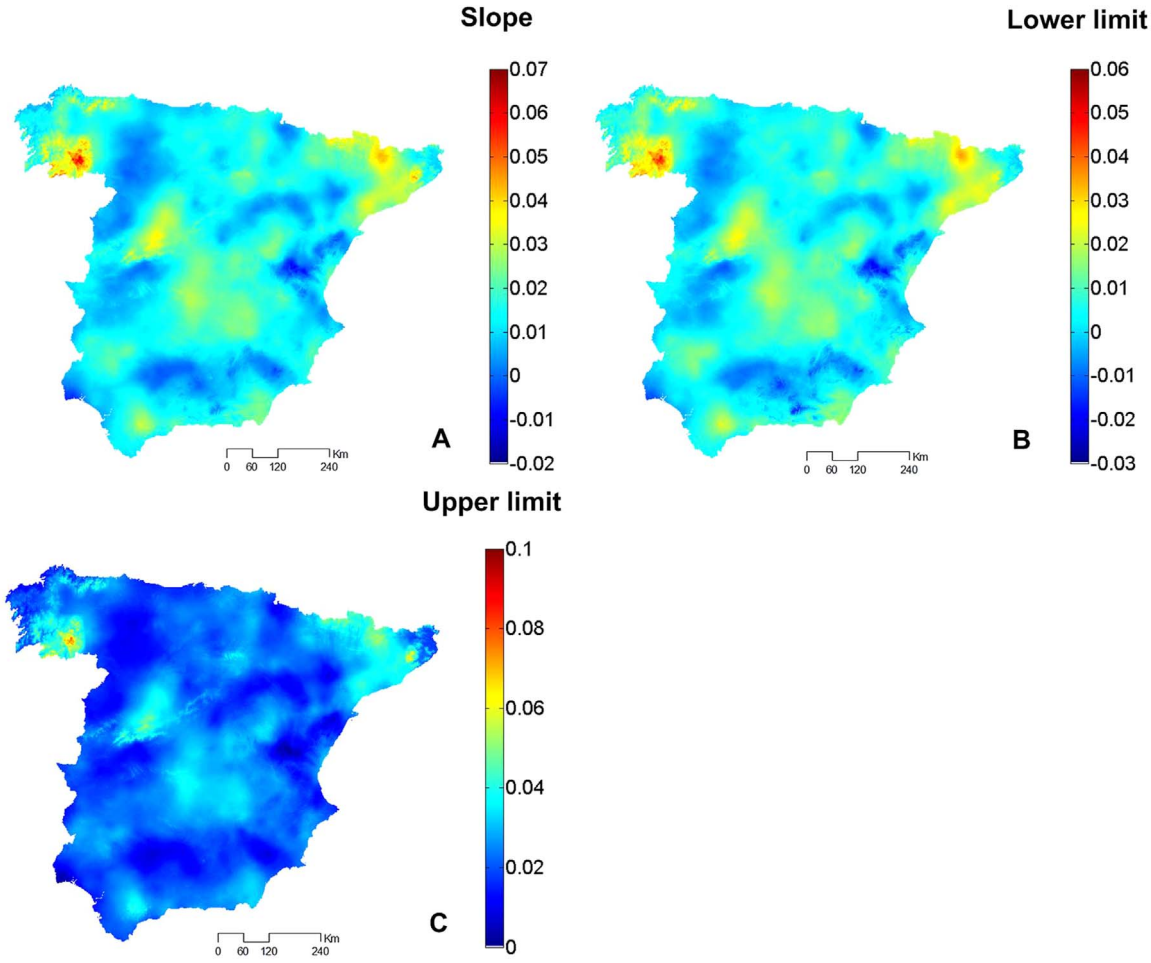

Fig. 8. Statistics of the linear regression analysis of the annual temperature data for the period 1950–2011. A: Estimated slope; B: 95% lower limit; C: 95% upper limit.

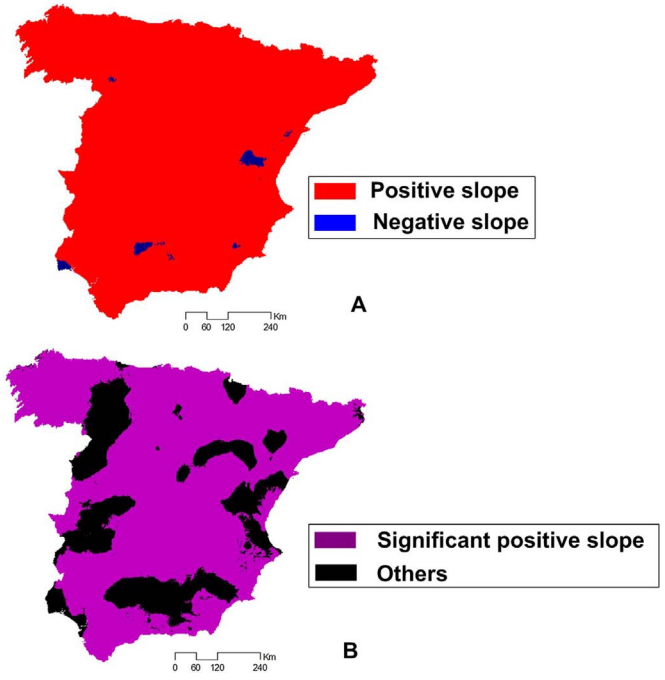

Fig. 9. A: Areas with positive slope (red) and negative slope (blue) according to Fig. 8A. B: The purple areas have statistically significant positive slope. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a positive slope and 78% of the area has a statistically significant linear trend that indicates an increase in the annual mean temperature over the period 1950–2011.

Acknowledgements

The first author and the fourth author are supported by the NSFC (41672324, 41430317), the Priority Academic Program Development of Jiangsu Higher Education Institutions and the National Science and Technology Major Projects (2016ZX05044-002). The first author is also supported by the China Scholarship Council (CSC). The work of the second author was supported by the project KARSTCLIMA, CGL2015-71510-R (Ministerio de Economía, Industria y Competitividad). The work of the third author was supported by Australian Research Council Discovery grant DP110104766. We thank the anonymous reviewers for providing constructive criticism that has helped to improve the final version of the manuscript.

References

AEMET, 2011. ⟨http://www.aemet.es/es/portada⟩.

Benavides, R., Montes, F., Agustin, R., Osoro, K., 2007. Geostatistical modelling of air temperature in a mountainous region of Northern Spain. Agr. For. Meteorol. 146, 173–188.

Cannon, A.J., 2012. Regression-guided clustering: a semi-supervised method for circulation-to-environment synoptic classification. J. Appl. Meteor. Climatol. 51, 185–190.

David, M., 1977, 1982. Geostatistical Ore Reserve Estimation (Series: Developments in Geomathematics 2). Elsevier, Netherlands, 364.

DeGaetano, A.T., 2001. Spatial grouping of United States climate stations using a hybrid clustering approach. Int. J. Climatol. 21, 791–807.

Dodson, R., Marks, D., 1997. Daily air temperature interpolated at high spatial resolution over a large mountainous region. Clim. Res. 8, 1–20.
Fovell, R.G., Fovell, M.Y.C., 1993. Climate zones of the conterminous United States defined using cluster analysis. J. Clim. 6, 2103–2135.
Gerstengarbe, F.-W., Werner, P.C., Fraedrich, K., 1999. Applying non-hierarchical cluster analysis algorithms to climate classification: some problems and their solution. Theor. Appl. Climatol. 64, 143–150.
Gordon, A.D., 1996. A survey of constrained classification. Comput. Stat. Data Anal. 21, 17–29. http://dx.doi.org/10.1016/0167-9473(95)00005-4.
Hengl, T., Heuvelink, G.B.M., Rossiter, D.G., 2007. About regression-kriging: from equations to case studies. Comput. Geosci. 33, 1301–1315.
Hudson, G., Wackernagel, H., 1994. Mapping temperature using kriging with external drift: theory and example from Scotland. Int. J. Climatol. 14, 77–91.
Huth, R., Beck, C., Philipp, A., Demuzere, M., Ustrnul, Z., Cahynová, M., Kyselý, J., Tveito, O.E., 2008. Classifications of atmospheric circulation patterns: recent advances and applications. Ann. N. Y. Acad. Sci. 1146, 105–152.
Ishida, T., Kawashima, S., 1993. Use of cokriging to estimate surface air temperature from elevation. Theor. Appl. Climatol. 47, 147–157.
Journel, A.G., Huijbregts, C.J. (1978, 2003) Mining Geostatistics. Academic Press, London (1978); Blackburn Press, New York (2003); pp. 600.
Mahlstein, I., Knutti, R., 2010. Regional climate change patterns identified by cluster analysis. Clim. Dyn. 35, 587–600.
Matheron, G., 1962. Traité de Géostatistque Appliquée; Tome 1. Mémoires du Bureau de Recherches Géologiques et Minières. Editions Technip, Paris, 333.
Mountain Research Initiative EDW Working Group, 2015. Elevation-dependent warming in mountain regions of the world. Nat. Clim. Change 5, 424–430.
Olea, R.A., 1999. Geostatistics for Engineers and Earth Scientists. Springer Science, 303.
Pepin, N.C., Lundquist, J.D., 2008. Temperature trends at high elevations: Patterns across the globe. Geophys. Res. Lett. 35, 6. http://dx.doi.org/10.1029/2008GL034026.
Robeson, S.M., 1994. Influence of spatial sampling and interpolation on estimates of air temperature change. Clim. Res. 4, 119–126.
Rolland, C., 2002. Spatial and seasonal variations of air temperature lapse rate in Alpine region. J. Clim. 16, 1032–1046.
Sen, P.K., 1968. Estimates of the regression coefficient based on Kendall's tau. J. Am. Stat. Assoc. 63, 1379–1389.
Stooksbury, D.E., Michaels, P.J., 1991. Cluster analysis of Southeastern U.S. climate stations. Theor. Appl. Climatol. 44, 143–150.
Tang, L., Su, X., Shao, G., Zhang, H., Zhao, J., 2012. Clustering-assisted regression (CAR) approach for developing spatial climate data sets in China. Environ. Modell. Softw. 38, 122–128.
Theil, H., 1950. A rank-invariant method of linear and polynomial regression analysis, I,II,III.Nederl. Akad. Wetensch. Proc. 53, pp. 386–392, 512–525,1397–1412.
Unal, Y., Kindap, T., Karaca, M., 2003. Redefining the climate zones of Turkey using cluster analysis. Int. J. Climatol. 23, 1045–1055.
Viers, G., 1975. Éléments de climatologie. Fernand Nathan, 309pp.
Zscheischler, J., Mahecha, M.D., Harmeling, S., 2012. Climate classification: the value of unsupervised clustering. Procedia Comput. Sci. 9, 897–906.