HYDROLOGICAL PROCESSES
Hydrol. Process. **27**, 1399–1410 (2013)
Published online 17 April 2012 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.9300

Dimensionality reduction in drought modelling

João Filipe Santos,[1]* Maria Manuela Portela[2] and Inmaculada Pulido-Calvo[3]

[1] *Departamento Engenharia, ESTIG, Instituto Politécnico de Beja, Rua Afonso III, 7800-050 Beja, Portugal*
[2] *Departamento Engenharia Civil, SHRH, Instituto Superior Técnico (Lisboa), Portugal, Avda. RoviscoPais, 1049-001 Lisboa, Portugal*
[3] *Departamento Ciencias Agroforestales, Escuela Técnica Superior de Ingeniería, Campus La Rábida, Universidad de Huelva, 21819 Palos de la Frontera, Huelva, Spain*

Abstract:

For monitoring hydrological events characterized by high spatial and temporal variability, the number and location of recording stations must be carefully selected to ensure that the necessary information is collected. Depending on the characteristics of each natural process, certain stations may be spurious or redundant, whereas others may provide most of the relevant data. With the objective of reducing the costs of the monitoring system and, at the same time, improving its operational effectiveness, three procedures were applied to identify the minimum network of rain gauge stations able to capture the characteristics of droughts in mainland Portugal. Drought severity is characterized by the standardized precipitation index applied to the timescales of 1, 3, 6 and 12 consecutive months. The three techniques used to reduce the dimensionality of the network of rain gauges were as follows: (i) artificial neural networks with sensitivity analysis, (ii) application of the mutual information criterion and (iii) *K*-means cluster analysis using Euclidean distances. The results demonstrated that the best dimensionality reduction method was case dependent in the three regions of Portugal (northern, central and southern) previously identified by cluster analysis. All the reduction techniques lead to the selection of a subset of rain gauges capable of reproducing the original temporal patterns of drought. For specific severe drought events in Portugal in the past, the comparison between drought spatial patterns obtained with the original stations and the selected subset indicated that the subset produced statistically satisfactory results (correlation coefficients higher than 0.6 and efficiency coefficients higher than 0.5). Copyright © 2012 John Wiley & Sons, Ltd.

KEY WORDS rain gauge network; drought monitoring; standardized precipitation index; mutual information; sensitivity analysis; artificial neural network; Portugal

Received 27 October 2011; Accepted 2 March 2012

INTRODUCTION

Drought is a recurrent natural phenomenon that can lead to a severe reduction in the availability of fresh water for a certain period and can affect wide areas. Although complex, attempts are made to characterize this phenomenon, and the process generally involves the calculation of drought indices, derived from meteorological and hydrological records. These provide information about historical droughts and therefore can also be used to monitor current conditions. According to Tsakiris (2008), such indices are useful for planning and management purposes as they provide standardized measures of the deviation of the water availability from normal conditions. Among these drought indices, the most widely used is the standardized precipitation index (SPI), which is based on precipitation data, as implied by the name (McKee *et al*. 1993; Santos *et al*. 2010).

Various authors, such as Bonaccorso *et al*. (2003), have emphasized that the monitoring of meteorological and hydrological events characterized by high spatial and temporal variability, such as droughts, requires careful selection of the optimal number of gauge stations able to describe the phenomenon within the area under study. One of the most widely used criteria for network design is based on geostatistical techniques (Bastin *et al*. 1984; Bogardi and Bardossy 1985; Pardo-Igúzquiza 1998, and more recently Chen *et al*. 2008) the aim of being able to improve the operational performance of a monitoring network with fewer gauges by selecting just the most important stations. In the case of drought monitoring, to achieve effective modelling of the phenomenon, most common methods still depend on a relatively large network of meteorological stations with long time series.

For forecasting purposes, the size and the quality of the input network of a model are crucial, as reported by many authors including Murphy (1991), Zheng and Billings (1996), Maier and Dandy (2000), Back and Trappenberg (2001) and, more recently, Guest and Smith-Genut (2010). If relevant inputs (independent variables) are omitted, the model cannot fully capture the input–output pattern (i.e. the model is underspecified). On the other hand, if the model includes redundant or unnecessary inputs (i.e. the model is overspecified), one or more of the following may occur: (i) the size, computational complexity and memory requirements of the model increase; (ii) the calibration of the model becomes more difficult due to an increase in the size of the search space and the greater number of local optima; (iii) the interpretation of the physical meaning of results from calibrated models becomes more difficult; and (iv) more data are needed to efficiently estimate the optimal values of the model parameters.

*Correspondence to: João Filipe Santos, Departamento Engenharia, ESTIG, Instituto Politécnico de Beja, Rua Afonso III, 7800-050 Beja, Portugal.
E-mail: joaof.santos@estig.ipbeja.pt

The aim of this study was to identify the minimum number of rain gauges that need to be considered in a given drought monitoring network to improve the monitoring and, consequently, the predictability of drought-related phenomena on the basis of one of the main drought indicators usually considered, namely, the SPI (McKee *et al.* 1993). The objectives here presented are not to eliminate rain gauges of the meteorological network of a given region but instead to select those rain gauges that should be considered for an optimal drought monitoring and thus to facilitate the planning of drought events to the responsible agencies. Recently, this purpose is supported by Chebbi *et al.* (2011), who concluded that it is possible to obtain different optimal rain gauge locations if rainfall intensity or erosion is evaluated to obtain an optimal extension of a monitoring network.

Three different reduction techniques were applied: (i) a method that relies on extracting knowledge contained within trained artificial neural networks (ANNs) by incorporating sensitivity analyses (ANN-SA), the method most commonly used with calibrated neural network models (Maier and Dandy 1996, 1997; Schleiter *et al.* 1999); (ii) a model-free approach, namely, the mutual information (MI) criterion (Shannon, 1948), which is a measure of the strength of dependence between two variables; and (iii) a *K*-means clustering technique using the Euclidean distance measure (K_{means}-ED), which selects the most important input rain gauges from each of the existing clusters (Holliday *et al.* 2004).

The performance of these dimensionality reduction techniques was tested for mainland Portugal by comparing the temporal and spatial drought patterns provided by the SPI index derived from an initial large network of rain gauges with those from the reduced network that resulted from the application of each of the techniques. This large network has previously been used to study drought spatial and temporal variability (Santos *et al.* 2010) and to analyse drought frequency (Santos *et al.* 2011). To describe the spatial pattern of the droughts, we tested two interpolation techniques, namely, the inverse distance weighting (IDW) and the kriging technique. The results achieved were assessed in terms of error measures and, additionally, for the spatial patterns by mapping interpolation errors.

MATERIALS AND METHODS

ANNs: sensitivity analysis

The first method used to identify the minimum number of rain gauges that ensures a monitoring network able to describe the temporal and spatial patterns of the droughts in mainland Portugal was based on ANNs, which are mathematical models inspired by the neural architecture of biological nervous systems.

The most widely studied and used ANN models involve multilayer feed-forward networks or multilayer perceptrons (Rumelhart *et al.* 1986; Senthil-Kumar *et al.* 2005). These models 'learn' in an iterative way, from data being input several times to the neural network until a predetermined error level (calculated as the sum of the squared errors) is reached; each iteration is called an 'epoch'. These supervised ANNs allow the analysis of complex data sets and the assessment of nonlinear relationships between dependent and independent variables. Detailed descriptions of the performance of multilayer perceptron ANNs have been given by Hsu *et al.* (1995), ASCE (2000a,b), Shrestha *et al.* (2005), Gutiérrez-Estrada *et al.* (2007), and Pulido-Calvo and Portela (2007). In particular, there are many methods for calibrating and training the ANNs; in this study, the standard back-propagation algorithm was used.

Before the calibration of any ANN, the data set was divided in two subsets: the first one (calibration subset [CS] + select subset [SS], CSS) was composed of 75% gauging stations of each region identified in mainland Portugal by Santos *et al.* (2010) and the second one (test subset, TS) was composed of the 25% remaining gauging stations (Gutiérrez-Estrada and Bilton 2010). In the first subset (CSS), 75% gauging stations (also randomly selected) composed the select subset (SS), which were used to avoid overtraining or overcalibration of the ANN. The best method to ensure that overtraining does not occur is to monitor periodically (at the end of each epoch) the sum square error for both the CS subset and the SS subset (internal validation). It is normal that the sum square error for the CS subset decreases continuously with training. However, this may be forcing the neural network to fit the noise in the CS subset. To avoid this problem, training is stopped at the end of each epoch, and the sum square error of the SS subset is calculated. When the sum square error of the SS subset begins to increase, training must be stopped, and the weights of the epoch that provided a minimum error for the SS subset should be tested with the TS subset. This last phase is also known as the *generalization phase* or *external validation*. Iyer and Rhinehart (1999) recommend repeating this process at least 30 times for each model, and this recommendation was followed here.

The selection of the input variables from among those available is a fundamental issue in ANN modelling (Bowden *et al.* 2005; Shrestha *et al.* 2005; May *et al.* 2008; Fernando *et al.* 2009). For calibration and validation phases, the objective is to select input variables that explain most of the variance with the least error. The selection of inappropriate input variables may cause undesirable effects (e.g. outputs provided by the model for each period are systematically very close to values observed in the previous periods), the presence of more local optima and extreme difficulties in extracting physical meaning from the calibrated models. To overcome these problems, many neural approaches have adopted methods, including autocorrelation and partial autocorrelation function analysis, linear cross-correlation analysis and spectral analysis (Jain *et al.* 2004; Shrestha *et al.* 2005; Elgaali and García 2007; Pulido-Calvo and Portela 2007; Gutiérrez-Estrada *et al.* 2007). However, such methods tend to be designed to capture linear

dependence between model inputs and outputs, which can mask the strength of nonlinear relationships and increase the risk of omitting relevant inputs (i.e. the model will be underspecified).

In the present study, the candidate inputs of the model were standardized by monthly precipitation in the 144 gauging stations of the three regions identified in mainland Portugal by Santos *et al*. (2010), and the outputs were weighted SPI values at different timescales also obtained by the same authors. ANNs with one and two hidden layers were assessed; in each case, 5−50 neurones (5, 6, 7,..., 50) were tested. The sigmoid activation function was used in hidden neurones and the linear activation function was used in output neurones.

The selection of the inputs of the model was made using an alternative form of sensitivity analysis applied to each ANN model, based on the approach of the missing value problem (Hunter *et al*. 2000). This method has been applied previously in ecological modelling and studies of fisheries resource management (Gutiérrez-Estrada *et al*. 2009; Gutiérrez-Estrada and Bilton 2010; Yáñez *et al*. 2010). The analysis was carried out by replacing each input variable with missing values and assessing the effect of this on the output error. The newly calculated error was compared with the original error to obtain a ratio (error of the model with an input variable with missing values/error of the model with all selected input variables). Thus, for any input variable x, a ratio equal or very close to 1 indicates that this variable had a very low weight in the general structure of the model (Hunter *et al*. 2000).

Mutual information algorithm

The second approach tested involved estimation of the MI. In information theory (Shannon 1948), the MI of two random variables is a measure of the mutual dependence of those variables. Although first developed for binary variables, it has been extended to continuous variables and so can be used to quantify nonlinear as well as linear dependence between any two variables (Bowden *et al*. 2005). This makes MI an attractive alternative to the use of the correlation coefficient, which can only describe patterns of linear dependence. In particular, MI maybe especially well suited to hydrological problems, as the dependence between hydrological variables and related indices is often nonlinear in nature.

The MI algorithm used to identify the relevant rain gauges was the one modified and extended by Moddemeijer (1989). This method can be compared with cross-correlation analysis, as both search for the maximum correspondence between two samples. The MI function between variables x and y is

$$\mathrm{MI}(x,y) = \int\int f_{x,y}(x,y)\log_e\frac{f_{x,y}(x,y)}{f_x(x)f_y(y)}\mathrm{d}x\ \mathrm{d}y \quad (1)$$

where $f_x(x)$ and $f_y(y)$ are the marginal probability density functions of x and y, respectively, and $f_{x,y}(x,y)$ is the joint probability density function of x and y. The joint probability density function of the two variables is equal to the product of the marginal densities if there is no statistical dependence between the variables (Sharma 2000). Hence, if x and y are independent, the joint probability density $f_{x,y}(x,y)$ is $f_x(x)f_y(y)$, and the result of MI is 0; conversely, a strong dependence between the two variables is indicated by a high value of MI.

For the case of a bivariate sample, the MI score of the discrete approach in Equation (1) can be approximated by

$$\mathrm{MI}(\mathrm{x},\mathrm{y}) = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\log_{\mathrm{e}}\left[\frac{\mathrm{f}_{\mathrm{x,y}}(\mathrm{x_i},\mathrm{y_i})}{\mathrm{f_x}(\mathrm{x_i})\mathrm{f_y}(\mathrm{y_i})}\right] \quad (2)$$

where $f_x(x_i)$, $f_y(y_i)$ and $f_{x,y}(x_i,y_i)$ are the marginal and joint density functions. The key to successful computation of the MI score using Equation (2) is the accurate estimation of such functions from a finite set of continuous random samples.

To calculate the marginal and joint probabilities, we used a two-dimensional histogram to approximate the joint probability density function between the variables x (inputs) and y (outputs), as in Moddemeijer (1989). The most widely used nonparametric approach for estimating MI is based on histograms, one of the main advantages of which is computational efficiency (Fernando *et al*. 2009), and this is especially important when dealing with long time series, such as those we have analysed in this study.

K-means clustering: Euclidean distances

The third approach to identifying the relevant rain gauges for describing the regional and spatial drought patterns of the SPI was the *K*-means clustering method (Hartigan and Wong 1979). According to this method, the rain gauges are grouped homogeneously so that similar precipitation variations on different timescales are assigned to the same group (or cluster), whereas different variations are grouped separately. Santos *et al*. (2010) found that nonhierarchical methods such as the *K*-means algorithm outperformed hierarchical methods tested with precipitation data.

In general, the *K*-means method will produce exactly K different clusters of the greatest possible distinction, with the goal of (i) minimizing variability within clusters and (ii) maximizing variability between clusters. Accordingly, the *K*-means method is an iterative approach whereby clusters are built around k central points called *centroids*. The process starts with k random centroids and, based on a distance measure, each rain gauge is assigned to its nearest centroid. For continuous variables, the centroid value is simply the mean of the values of all members assigned to the cluster. Then, iteratively, on the basis of the content of each cluster, the respective centroids are updated and the rain gauges are reassigned to their new closest centroids. In our analysis, the algorithm stopped after a certain number of iterations.

The distance measure used to assign each rain gauge to the nearest centroid or to measure the distance between two cluster centres was the Euclidean distance, D, expressed as

$$D_{i,j} = |x-y| = \sqrt{\sum_{i=1}^{N}|x_i-y_i|^2} \quad (3)$$

where x and y are the elements of the time series to be compared and N is the sample dimension. Although there are many other distance metrics, the Euclidean distance is the dissimilarity measure most commonly used in K-means cluster analysis (Kahya *et al.* 2008).

The mean of the Euclidean distances was used as a threshold to retain the key members of each cluster, that is, to reduce the number of rain gauges in each region while maintaining the main regional and spatial information content. Specifically, the cluster members selected were those that were closer than the mean distance (K_{means}-ED). First, the optimum number of clusters was obtained by evaluating the appropriateness of the classification using the aforementioned distance measure between clusters, as in Webster and Oliver (1990) and Hair *et al.* (2005). Then, the distances from each of the cluster members to their respective cluster centre (mean) were used to identify potential 'bad' cluster members, that is, rain gauges that were very distant from the cluster centre and therefore could be omitted from the set. It was also assumed that if a given cluster member did not belong to any other cluster and was distant from its respective cluster centre, then it was even further away from the centres of alternative clusters.

Testing of the proposed dimensionality reduction techniques: Portuguese case study

Portugal is located in the western part of the Iberian Peninsula, and its climate has both Atlantic and Mediterranean influences. The mean annual rainfall varies from more than 3000 mm in the northwest to less than 400 mm in the south and follows a complex spatial pattern (N–S/E–W), which is closely related to the topography of the country. Indeed, topography is one of the major factors affecting the distribution of precipitation. As water availability decreases, the hydrological regime within and between years becomes more irregular (Portela and Quintela 2006). This makes the monitoring and prediction of droughts an important issue for the management of water resources, particularly in the semiarid central and southern regions of Portugal.

Santos *et al.* (2010) identified the general spatial pattern of droughts in Portugal using the SPI. On the basis of 144 rain gauges (sited at the locations shown in Figure 1), they identified three well-defined subregions with different drought patterns (Figure 1): the northern region (cluster 1, CL1), the central region (cluster 2, CL 2) and the southern region (cluster 3, CL 3). The number of rain gauges in each of these regions was 56, 53 and 35, respectively. Drought information was obtained and represented by the SPI1, SPI3, SPI6 and SPI12 time series (data for 1, 3, 6 and 12 months periods, respectively).

Using precipitation data as input, two main requisites were considered to achieve dimensionality reduction: first, there must be a set of stations in each of the three regions of Figure 1 able to reproduce the respective temporal drought patterns, that is, the SPI-weighted time series obtained by the Thiessen interpolation method; and second, these sets should also reproduce a spatial drought pattern as close as possible to the pattern obtained using all the original rain gauges, without a significant loss of spatial information.

With the ANN-SA approach, all the precipitation data were used as the input to the ANN trained with the objective of reproducing the weighted drought temporal patterns of the regions for SPI1, SPI3, SPI6 and SPI12. To avoid errors in the training of the ANN with respect to multiscale input/output relationships and because the target SPI values are already normalized values, we also normalized the input variables (precipitation). For SPI1, the precipitation measurements used as input were normalized by the corresponding monthly values, whereas, similarly, data for SPI3, SPI6 and SPI12 were normalized on the basis of the precipitation in the 3-, 6- and 12-month periods, respectively, according to the expression

$$P_{\text{norm}} = \frac{x_i - \bar{x}_i}{s_{x_i}} \tag{4}$$

where x_i is the precipitation in the 1-, 3-, 6- or 12-month periods and $\bar{x}_i$ and s_{x_i} are the respective mean and standard deviation.

In the sensitivity analysis developed for the best ANN-SA models, the variables (rain gauges) that exhibited a ratio (error of the model with an input variable with missing values/error of the model with all selected input variables) greater than the mean model ratio were selected because of their relative importance in the model.

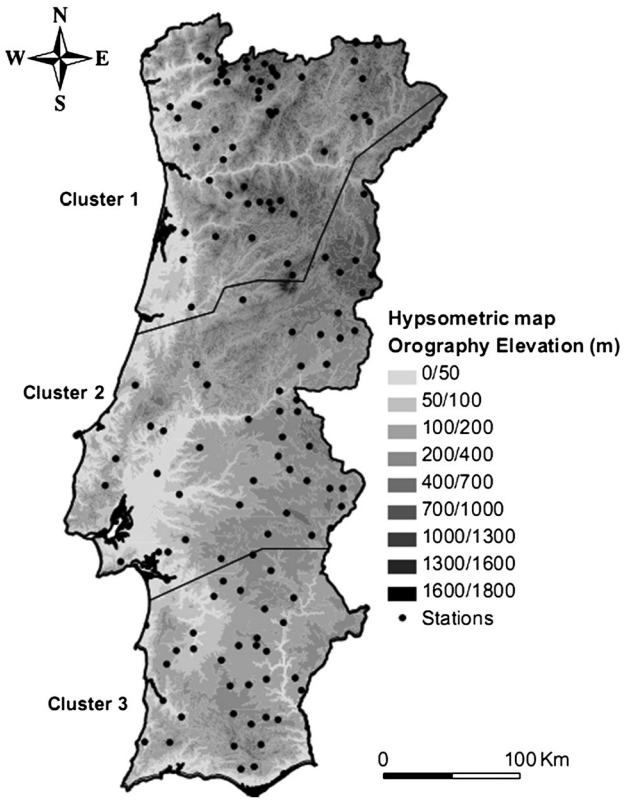

Figure 1. Spatial drought patterns in mainland Portugal using SPI. A map showing the three regions previously identified by cluster analysis (Santos *et al.*, 2010). The dots represent the location of the 144 rain gauges used for the drought analysis

For each timescale and region, the MI approach was used to assess the dependence between the standardized precipitation (inputs) and the corresponding weighted SPI (outputs). As previously stated, the purpose of using the MI criterion is to select the most relevant rain gauges among the regional set. It means that the data from a specific rain gauge must contain as much information as possible with respect to the prediction SPI values. As the information content is precisely what is measured by the MI, the first rain gauges to be selected are, naturally, the ones that maximize the MI. In each region for each SPI timescale, the threshold criterion we adopted was the average MI value, and the inputs above that threshold were selected (i.e. further ongoing research will consider the use of other thresholds). Other authors, such as Sharma (2000), have suggested instead using the 95th percentile confidence limit for the sample MI to decide whether the MI of a candidate input is significantly different from zero and should therefore be added to the set of already selected inputs.

The K_{means}-ED technique with standardized precipitation data was used for selecting the optimum number of clusters in each region (northern, central and southern), and the Euclidean distance measure of all the rain gauges in each cluster was analysed. The mean Euclidean distance of the rain gauges to their cluster centre was used to exclude the least relevant rain gauges from each cluster. All the rain gauges further away than this threshold distance were removed from the set. We made the assumption that the rain gauges selected in this way, for each region and precipitation timescale, were those that should be considered in the SPI calculation.

Measures of accuracy (validation phase)

Four accuracy measures were used to evaluate each reduction technique: the root mean square error (RMSE); the correlation coefficient (R), to compare between simulated and observed values; the efficiency coefficients (E_1 and E_2); and the persistence index (PI) (Nash and Sutcliffe 1970; Kitanidis and Bras 1980; Legates and McCabe, 1999; Pulido-Calvo and Portela 2007). It is advisable to use the RMSE to quantify the error in the same units as the variables (in this case dimensionless SPI values). The RMSE is given by:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(\text{SPI}_t - \text{S}\hat{P}\text{I}_t\right)^2}{N}} \tag{5}$$

where SPI_t is the original temporal drought pattern at the timescale t, $\text{S}\hat{P}\text{I}_t$ is the equivalent pattern based on selected rain gauges and N is the total number of observations of the validation set. The coefficient of efficiency E_j is used to assess how well the selected rain gauges describe the original drought pattern. E_j is given by

$$E_j = 1.0 - \frac{\sum_{i=1}^{N} \left|\text{SPI}_t - \text{S}\hat{P}\text{I}_t\right|^j}{\sum_{i=1}^{N} \left|\text{SPI}_t - \overline{\text{SPI}_t}\right|^j} \tag{6}$$

The sensitivity to outliers due to the squaring of the difference terms is associated with E_2. The measure E_1 (here called the modified coefficient of efficiency) reduces the effect of the squared terms. A value of 0 for E_2 indicates that the original average $\overline{\text{SPI}}$ is as good a predictor as the selected drought temporal pattern, whereas negative values indicate that the original average is a better predictor than the selected pattern (Legates and McCabe, 1999). For a good match, the values of R^2 and of E_j should be close to one.

In addition, the PI, used to evaluate the performance of the models, is given by (Kitanidis and Bras 1980)

$$\text{PI} = 1 - \frac{\sum_{i=1}^{N} \left(\text{SPI}_t - \text{S}\hat{P}\text{I}_t\right)^2}{\left(\text{SPI}_t - \text{SPI}_{t\text{-}L}\right)^2} \tag{7}$$

where SPI_{t-L} is the observed SPI value at the time step $t - L$ and L is the lead time. In our calculations, L was set equal to one (only 1 month ahead). A PI value of 1 reflects a perfect adjustment between selected and original temporal patterns, and PI a value of 0 is equivalent to saying that the model is no better than a 'naïve' model.

Temporal and spatial droughts patterns (validation phase)

To test the validity of the selected subsets of rain gauges, we compared the temporal patterns of the droughts resulting from such sets and from the original rain gauges on the basis of the weighted time series of SPI given by the Thiessen interpolation. The accuracy measures presented in the Measures of Accuracy section and the scatter plots were also used to analyse the results.

To obtain more information about the performance of the best reduction techniques, we also calculated and compared some drought spatial patterns for the entire country. In particular, the mapping of specific extreme drought events was carried out using the subset of rain gauges obtained by the reduction models and the original set of rain gauges. Following the suggestion of Willmott (1982), such maps were compared using the accuracy measures between results with the original and results with the selected rain gauges. Using Arcgis 9.3 software, two common interpolation methods were tested: a geostatistical method (or Kriging) and a local interpolation procedure, IDW. Kriging was conceived by Matheron (1963) and is based on a statistical model of a phenomenon rather than an interpolating function. The method assumes that the spatial variation of a continuous climate-related variable, such as SPI, is too irregular to be modelled by a mathematical function, and its spatial variation would be better predicted by a probabilistic surface. Such a continuous variable, called a *regionalized variable* by Vicente-Serrano *et al.* (2003), is assumed to be based on a drift component and a random but spatially correlated component. At any point, $z(x)$ is given by (Borrough and McDonnell 1998)

$$z(x) = m(x) + \varepsilon'(x) + \varepsilon''(x) \tag{8}$$

where $m(x)$ is the drift component that indicates the structural variation of the drought variable, $\varepsilon'(x)$ is the residual or the difference between the drift component and the sampling data values, which are spatially dependent and $\varepsilon''(x)$ represents the spatially independent residuals.

The second method, IDW, is based on the assumption that the SPI value at an unsampled point $z(x)$ is a distance-weighted average of the SPI values at the sampling points $z(x_1)$, $z(x_2)$, ..., $z(x_n)$, within a radius around $z(x)$. The inverse of the distance, d_i, between $z(x_i)$ and $z(x)$ is the weighting factor because, generally, climate-related variables (such as the SPI) are more alike between the nearer points than between distant points. This supposes that predictions are obtained from the nearest sampling points:

$$z(x) = \frac{\sum_{i=1}^{N} z(x_i) d_{ij}^{-r}}{\sum_{i=1}^{N} d_{ij}^{-r}} \quad (9)$$

where $z(x)$ is the predicted value, $z(x_i)$ is the SPI of a neighbouring weather station, d_{ij} is the distance between $z(x)$ and $z(x_i)$ and r is a positive exponent. To test the quality of the results obtained with this method, we created maps with $r = 1$, 2 and 3.

RESULTS AND DISCUSSION

Validation of temporal drought patterns simulated with the reduction techniques

Of all the ANN architectures applied to the original set of rain gauges, to characterize the SPI1, SPI3, SPI6 and SPI12 time series, the configuration of one hidden layer with 10 neurones (e.g. 56-10-1) always produced the best results in all of the three cluster regions shown in Figure 1. The accuracy measures for such ANN models are presented in Table I.

To test the effectiveness of the ANN-SA, MI and K_{means}-ED methods, we applied them to data sets known to be interdependent, namely, precipitation and SPI time series. The number of rain gauges that resulted from each reduction method is reported in Table II. As mentioned before, the initial numbers of stations were 56, 53 and 35, respectively, in CL1, CL2 and CL3.

In global terms, for all the regions and for all the SPI timescales, the greatest reduction in the number of stations was achieved using the ANN-SA method and the least reduction with the K_{means}-ED method, with one exception in this latter case: SPI6 in the south (CL3), for which the MI method was found to be even less powerful.

The performance of the three different reduction methods was assessed with respect to their ability to reproduce the original temporal pattern of the droughts in each region on the basis of data from the subset of rain gauges resulting from each of the methods, which we refer to as the selected rain gauges. For that purpose, the Thiessen polygon method was first applied to all the original rain gauges and then only to the selected ones. The drought patterns thus obtained were compared using the accuracy measures described in the Measures of Accuracy section.

In general terms, these comparisons showed that all the reduction techniques led to a set of rain gauges able to reproduce the original drought temporal patterns. For brevity, we focus on the best techniques identified, and the corresponding accuracy measures are summarized in Table III.

The calculations showed that the accuracy measures were not as good for CL3 as for the other regions. Specifically, although for all SPI timescales we obtained $R > 0.94$ and E_1, E_2 and PI close to 1, the RMSE was relatively high and the E_1 and PI were relatively low as compared with the values for other regions (CL2 and CL1). These higher RMSE values and lower E_1 and PI values might be related to the loss of information content for the CL3 region due to the reducing network. Even the small reduction of the number of rain gauges in the least effective of the techniques applied, the MI caused a noticeable deterioration in the reproduction of the original drought pattern.

Table I. Original set of rain gauges. Accuracy measures for the best ANN models applied to simulate the SPI1, SPI3, SPI6 and SPI12 time series

Original stations	No. input rain gauges	RMSE	R	E_1	E_2	PI
CL1	ANN architecture			56: 56-10-1: 1		
SPI1	56	0.502	0.794	0.417	0.630	0.788
SPI3		0.611	0.720	0.310	0.517	0.294
SPI6		0.488	0.829	0.448	0.683	0.194
SPI12		0.043	0.999	0.952	0.997	0.982
CL2	ANN architecture			53: 53-10-1: 1		
SPI1	53	0.583	0.714	0.368	0.501	0.675
SPI3		0.561	0.783	0.413	0.611	0.438
SPI6		0.408	0.886	0.548	0.785	0.443
SPI12		0.058	0.998	0.940	0.995	0.972
CL3	ANN architecture			35: 35-10-1: 1		
SPI1	35	0.660	0.697	0.356	0.485	0.682
SPI3		0.646	0.757	0.386	0.555	0.420
SPI6		0.609	0.789	0.383	0.623	−0.197
SPI12		0.081	0.997	0.946	0.995	0.961

Table II. Results from the dimensional reduction methods

	No. stations			Sample reduction (%)		
	ANN-SA	MI	K_{means}-ED	ANN-SA	MI	K_{means}-ED
CL1						
Original stations		56		–	–	–
SPI1	20	29	34	64	48	39
SPI3	19	28	31	66	50	45
SPI6	17	29	32	70	48	43
SPI12	18	22	30	68	61	46
CL2						
Original stations		53		–	–	–
SPI1	15	28	30	72	47	43
SPI3	19	24	33	64	55	38
SPI6	15	27	29	72	49	45
SPI12	23	27	31	57	49	42
CL3						
Original stations		35		–	–	–
SPI1	12	17	17	66	51	51
SPI3	12	17	18	66	51	49
SPI6	11	18	17	69	49	51
SPI12	7	14	18	80	60	49

Table III. Drought temporal patterns for the best reduction methods

Technique	SPI1	SPI3	SPI6	SPI12
CL1	K_{means}-ED	MI	K_{means}-ED	ANN-SA
RMSE	0.066	0.068	0.072	0.061
R	0.997	0.997	0.997	0.997
E_1	0.925	0.924	0.921	0.930
E_2	0.993	0.994	0.993	0.995
PI	0.996	0.991	0.981	0.964
CL2	K_{means}-ED	K_{means}-ED	MI	ANN-SA
RMSE	0.065	0.087	0.088	0.067
R	0.997	0.995	0.995	0.997
E_1	0.921	0.900	0.896	0.920
E_2	0.993	0.989	0.989	0.993
PI	0.996	0.985	0.972	0.956
CL3	MI	ANN-SA	MI	MI
RMSE	0.205	0.239	0.257	0.301
R	0.970	0.963	0.960	0.945
E_1	0.776	0.746	0.735	0.675
E_2	0.940	0.928	0.920	0.888
PI	0.965	0.905	0.809	0.294

Accuracy measures between results based on the original and the selected rain gauges.

The best overall results were achieved by MI, especially for SPI1, SPI6 and SPI12 in CL3, but also for SPI3 in CL1 and SPI6 in CL2. The best accuracy measures were achieved in CL1, with R values higher than 0.99 and the efficiency measures and PI were also close to one. In CL1, these good results could be related to the higher density of rain gauges in some areas of the region, which increases the cross correlation and means that the reduction process results in relatively little loss of information content.

To analyse the relationship between patterns produced by the original set and by the reduced set of rain gauges, we generated scatter plots. Such plots are shown in Figure 2 for the northern and southern regions only, as the pattern in the central region (CL2) was almost the same as that for the north. The scatter plots illustrate that the relationship between drought temporal patterns in CL3, although acceptable, is not as good as in CL1 or CL2 (not shown, but similar to CL1).

Comparison of the yearly and monthly drought temporal patterns

Yearly and monthly analyses were also conducted to better understand the relationships between SPI based on the original set and on the selected subset of rain gauges. For all the SPI timescales in the south, CL3, Table IV lists the years in which correlation coefficients between these sets were less than or equal to 0.85. In the other regions, the correlation coefficients were always higher than 0.85.

In the MI technique, the yearly R values were always higher than 0.61 for SPI1, SPI3 and SPI6 and were higher than 0.56 for SPI12. Further, R had a value of 0.57 in 1944, a year in which there was an extreme drought. In the ANN-SA technique, the R values for SPI1, SPI3 and SPI6 were slightly better than those with MI: higher than 0.75. For SPI12, the R values were higher than 0.61, with the minimum value also occurring in the extremely dry year of 1944. The results from K_{means}-ED are intermediate between those of the other two techniques: R values are higher than 0.69 for SPI1, SPI3 and SPI6 and higher than 0.46 for SPI12. The same extremely dry year of 1944 had an R value of 0.47. In general terms, the three techniques performed well on an annual timescale.

An equivalent analysis at the monthly level also found very good agreement. In fact, even for the least good results, obtained for the south (CL3), the correlation coefficients were always higher than 0.9, as shown in Figure 3.

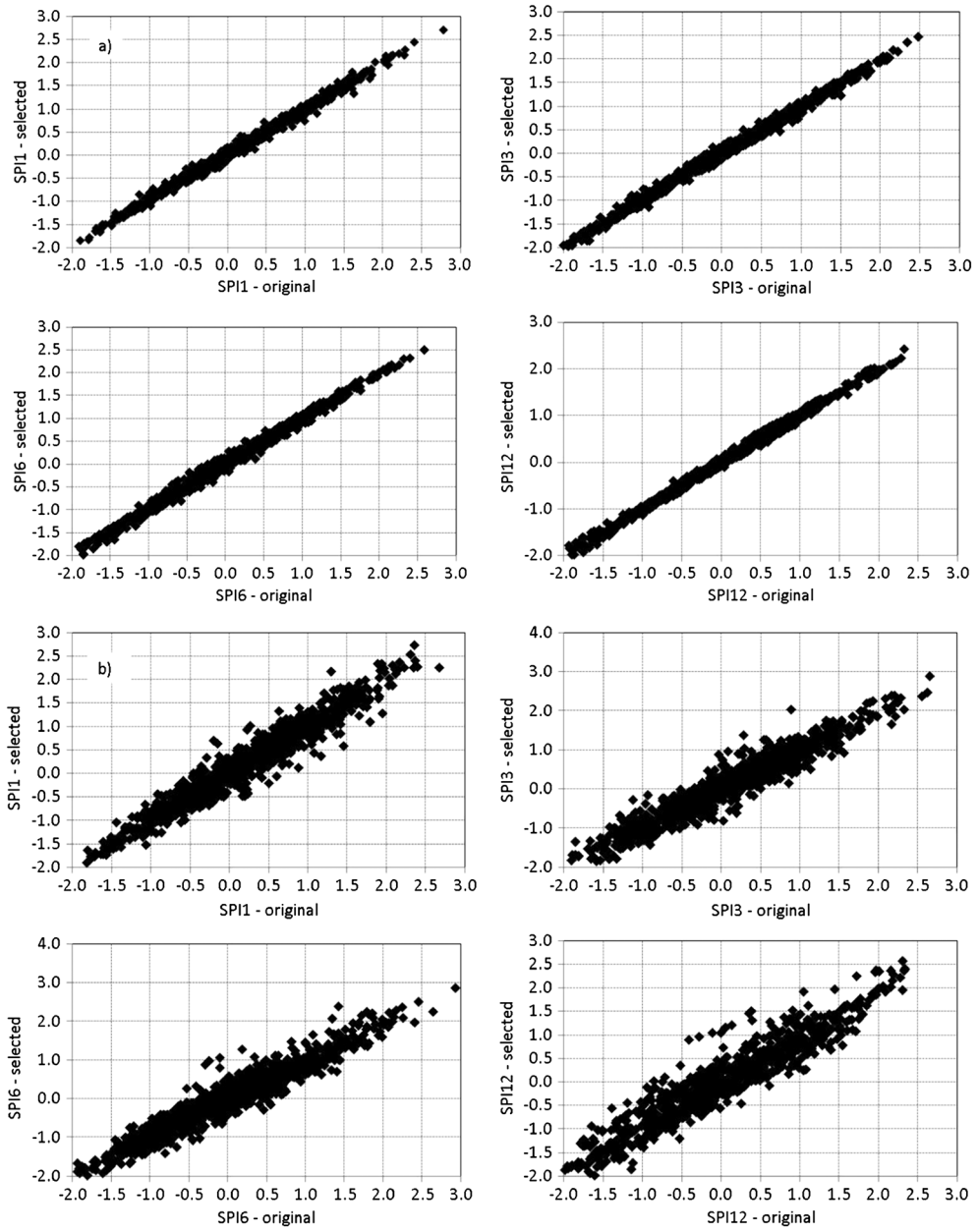

Figure 2. Scatter plots for SPI1, SPI3, SPI6 and SPI12 between the original and the selected set of rain gauges: a) Northern region; b) Southern region.

Table IV. Southern region (CL3; years with correlation coefficients, *R*, between original and selected SPI variables less than or equal to 0.85)

SPI	No.	*R*	Years
MI			
SPI1	1	0.83	1952
SPI3	4	0.62/0.74/0.84/0.79	1920/1929/1984/1988
SPI6	5	0.74/0.80/0.79/0.83/0.70	1919/1923/1955/1963/1987
SPI12	8	0.79/0.84/0.80/0.57/0.75/0.80/0.67/0.67	1919/1934/1940/1944/1950/1953/1957/1975
ANN-SA			
SPI1	1	0.79	1951
SPI3	2	0.76/0.82	1920/1988
SPI6	8	0.84/0.80/0.76/0.83/0.82/0.84/0.83/0.84	1914/1920/1924/1933/1956/1962/1988/1990
SPI12	8	0.74/0.62/0.67/0.78/0.69/0.83/0.71/0.81	1934/1944/1950/1953/1957/1974/1975/1986
K_{means}-ED			
SPI1	2	0.80/0.81	1951/1952
SPI3	3	0.77/0.81/0.84	1920/1984/1988
SPI6	5	0.70/0.82/0.80/0.72/0.70	1924/1933/1943/1956/1988
SPI12	9	0.73/0.75/0.47/0.76/0.76/0.61/0.83/0.78/0.83	1919/1940/1944/1953/1957/1975/1976/1986/1997

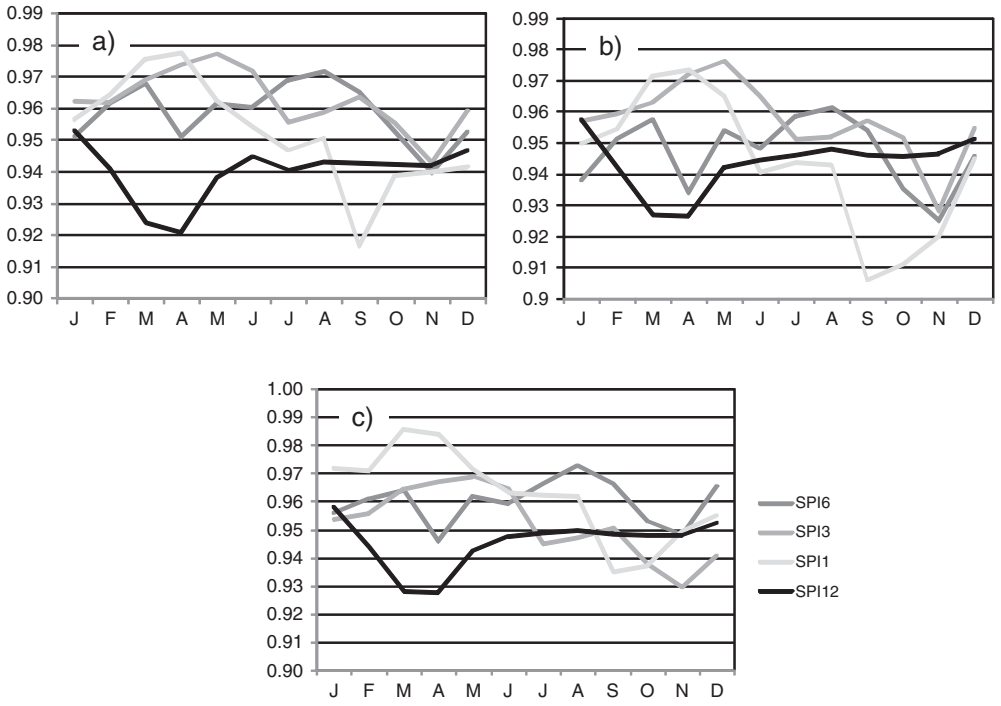

Figure 3. Southern region (CL3). Monthly correlation coefficients for (a) ANN-SA, (b) K_{means}-ED and (c) MI

Validation of spatial drought patterns simulated with the reduction techniques

Following the example of Akhtari *et al.* (2009), we evaluated the performance of the two interpolation methods applied (IDW and kriging), comparing, in particular, how they performed for certain severe historical drought events, namely, in November 1981 for SPI1, in November 1971 for SPI3, in December 1917 for SPI6 and in December 1980 for SPI12.

The best interpolation method varies as a function of the cluster region and of the scale of the mapping. Areas of great topographic complexity and regions with contrasting atmospheric or oceanic influences present more problems than flatter areas or regions with constant atmospheric patterns (Vicente-Serrano *et al.*, 2003). In the present case study, the north of Portugal (CL1) represents an area in which there are significant geographical differences and spatial climatic diversity; despite this complexity, we carried out the comparison of interpolation techniques to select the best possible model.

The accuracy measures were calculated for the errors between predictions based on the selected subset of stations and on the original set of rain gauges. Table V shows the results that give a preliminary idea of the model's validity. In general, with the exception of the SPI on a 1-month timescale, the IDW interpolation model with a parameter $r = 1$ gave the best performance.

For each SPI timescale, for the best reduction technique and for the interpolation method with best accuracy measures, we mapped a dimensionless error on the basis of the comparison of the maps of drought events using the original stations and the selected ones (Figure 4). This dimensionless error (D_{error}) was calculated using the following expression:

Table V. Accuracy measures for different interpolation procedures

	IDW $r=1$	IDW $r=2$	IDW $r=3$	Kriging
SPI1				
RMSE	0.281	0.277	0.276	0.282
R	0.580	0.598	0.603	0.581
E_1	0.513	0.520	0.521	0.484
E_2	0.335	0.356	0.362	0.332
PI	0.653	0.664	0.667	0.651
SPI3				
RMSE	0.412	0.412	0.416	0.423
R	0.778	0.780	0.779	0.766
E_1	0.532	0.544	0.539	0.505
E_2	0.602	0.602	0.595	0.582
PI	0.755	0.755	0.751	0.742
SPI6				
RMSE	0.551	0.551	0.568	0.564
R	0.799	0.798	0.785	0.789
E_1	0.517	0.514	0.497	0.502
E_2	0.602	0.602	0.577	0.583
PI	0.757	0.756	0.741	0.745
SPI12				
RMSE	0.550	0.562	0.586	0.555
R	0.833	0.820	0.803	0.827
E_1	0.570	0.571	0.561	0.567
E_2	0.680	0.666	0.637	0.674
PI	0.643	0.627	0.595	0.637

$$D_{\text{error}} = \frac{\text{error} - \text{min}}{\text{max} - \text{min}} \tag{10}$$

where error is the difference between the predicted and observed SPI and min and max are the minimum and maximum error values for each timescale and interpolation method, respectively. The mapping of these

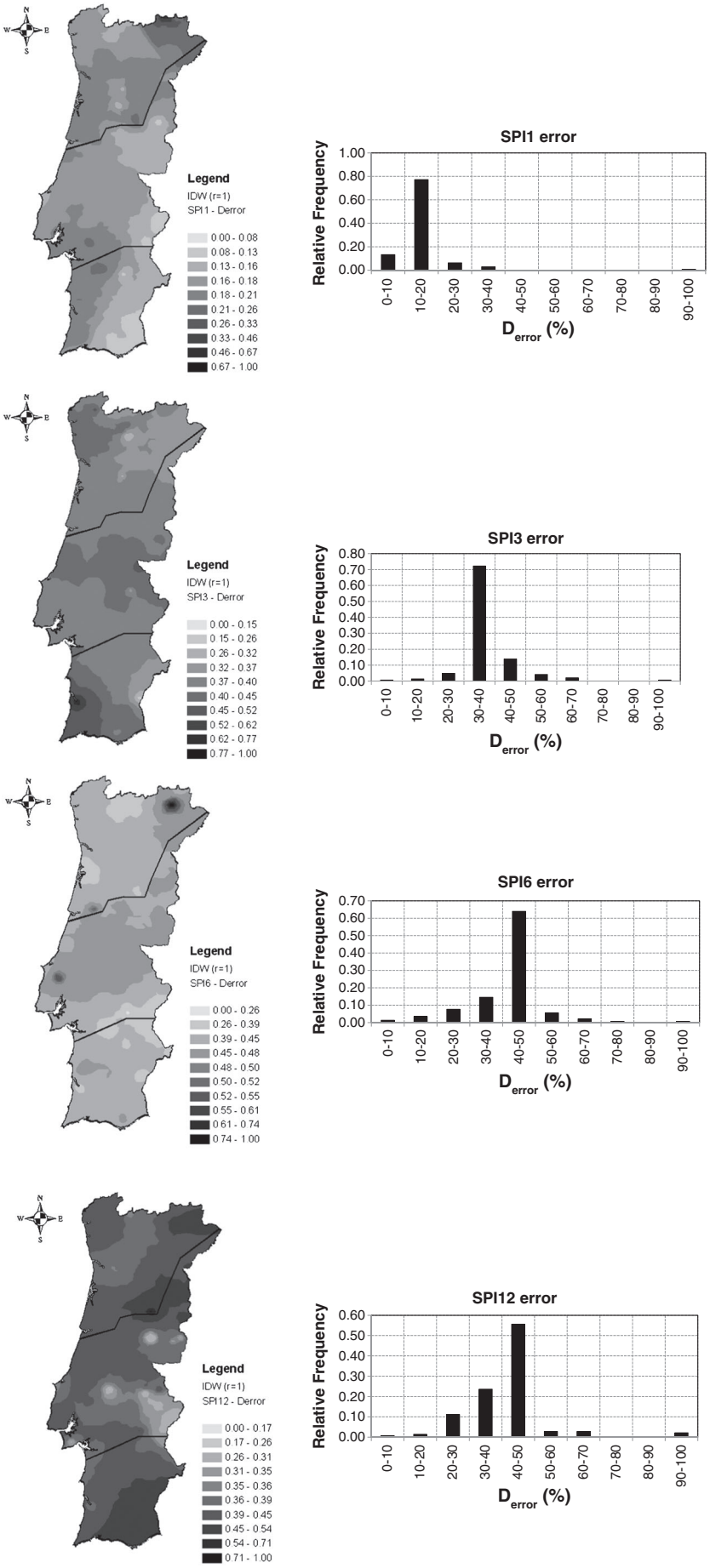

Figure 4. Dimensionless error maps for SPI timescales of 1, 3, 6 and 12 months. Histograms for the error distribution

values was made for the entire country to avoid potential discontinuities near the limits of the three cluster regions.

Figure 4 clearly shows that the margin of error grows as the SPI timescale increases: the error is approximately 10% to 20% for SPI1 and 30% to 40% for SPI3, whereas it is between 40% and 50% for SPI6 and SPI12.

The spatial distribution of the errors allows the points with poorer predictions to be identified: for SPI1, a point in the north eastern of the country; for SPI3, a point in the south west; and for SPI6, a point also in the north east. In the SPI12, the error distribution was fairly homogeneous.

CONCLUSIONS

To identify the minimum network of rain gauges capable of characterizing the regional and temporal patterns of the droughts in mainland Portugal, we tested three techniques for the reduction of the dimensionality of a preset network of rain gauges for drought monitoring using the SPI index at the 1-, 3-, 6- and 12-month timescales. The results demonstrated that although several relatively satisfactory subsets were identified, the performance of the dimensionality reduction methods was case dependent.

The first technique used the capacity of ANN trained models to capture the relationships between variables and to identify the most relevant set of input variables from the available candidates. The strength of the relationships between potential model inputs and outputs was examined using sensitivity analysis.

The second technique applied was the MI algorithm. This is called a model-free approach, as it does not require the construction of a model to assess the relationship between variables to capture nonlinear dependence between two variables and to provide useful insights concerning the relative ranking of inputs.

The third technique, *K*-means clustering (K_{means}-ED), is a commonly used data clustering method for the analysis of unsupervised learning tasks. It has the disadvantage, however, of not guaranteeing that only the nonredundant stations will be selected because stations with highly correlated data maybe selected if they are among the closest to the cluster centre.

When comparing drought temporal patterns, the values of the accurate measures obtained with the best methods, evaluated for the southern region of Portugal, were worse than those obtained for the other two regions (northern and central regions). Nevertheless, we have to point out that these values obtained for the southern region are statistically quite good ($R > 0.94$; $E_1 > 0.68$; $E_2 > 0.89$; PI > 0.81 except for SPI12). These results could be due to the absence of some rain gauges in some southern locations in the original stations network. It may be noted that this hypothesis needs to be reinforced through the research of the optimal location of rain gauges of the drought monitoring network which could be developed in future works.

Likewise, additional research needs to be carried regarding the physical mechanisms that explain the results achieved in connection with the threshold values used in the three reduction techniques to keep or reduce rain gauges from original stations network. In this sense, sensitivity analysis could be applied with the testing of variable threshold values that were considered on the three techniques used for the reduction of the dimensionality of a preset network of rain gauges for drought monitoring.

Our findings suggest that the network of rain gauges used to characterize droughts can be reduced – without compromising the accuracy of such characterization – so a more sustainable and equally efficient monitoring system could be implemented with lower costs. Indeed, with the network of automatic range gauge stations expanding, it is necessary to review the design criteria for the traditional networks to reduce installation and management costs of the new gauges and to improve monitoring in real time of extreme hydro-meteorological events, such as floods and droughts. These issues underline the current relevance of our work.

REFERENCES

Akhtari R, Morid S, Mahdianc MH, Smakhtind V. 2009. Assessment of areal interpolation methods for spatial analysis of SPI and EDI drought indices. *International Journal of Climatology* **29**: 135–145.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology. 2000a. Artificial neural networks in hydrology. I. Preliminary concepts. *Journal of Hydrology Engineering* **5**(2): 115–123.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology. 2000b. Artificial neural networks in hydrology. II. Hydrologic applications. *Journal of Hydrology Engineering* **5**(2): 124–137.

Back AD, Trappenberg TP. 2001. Selecting inputs for modeling using normalized higher order statistics and independent component analysis. *IEEE Transactions on Neural Networks* **12**(3): 612–617.

Bastin G, Lorent B, Dunqué C, Gevers M. 1984. Optimal estimation of the average areal rainfall and optimal selection of rain gauge locations. *Water Resources Research* **20**(4): 463–470

Bogardi I, Bardossy A. 1985. Multicriterion network design using geostatistics. *Water Resources Research* **21**(2): 199–208.

Bonaccorso B, Cancelliere A, Rossi G. 2003. Network design for drought monitoring by geostatistical techniques. *European Water Publications* **3/4**: 9–15.

Borrough PA, McDonnell RA. 1998. *Principles of geographical information systems*. Oxford University Press: Oxford, UK.

Bowden GJ, Dandy GC, Maier HR. 2005. Input determination for neural network models in water resources applications. Part 1—background and methodology. *Journal of Hydrology* **301**: 75–92.

Chebbi A, Bargaouir ZK, Cunha MC. 2011. Optimal extension of rain gauge monitoring network for rainfall intensity and erosivity index interpolation. *Journal of Hydrologic Engineering* **16**: 353–365.

Chen Y-C, Wei C, Yeh H-C. 2008. Rainfall network design using kriging and entropy. *Hydrological Processes* **22**: 340–346. DOI: 10.1002/hyp.6292.

Elgaali E, García LA. 2007. Using neural networks to model the impacts of climate change on water supplies. *Journal of Water Resources Planning and Management* **133**(3): 230–243.

Fernando MKG, Maier HR, Dandy GC. 2009. Selection of input variables for data driven models: An average shifted histogram partial mutual information estimator approach. *Journal of Hydrology* **367**: 165–176.

Guest JK, Smith-Genut LC. 2010.Reducing dimensionality in topology optimization using adaptive design variable fields. *International Journal for Numerical Methods in Engineering* **81**: 1019–1045.

Gutiérrez-Estrada JC, Bilton DT. 2010. A heuristic approach to predicting water beetle diversity in temporary and fluctuating waters. *Ecological Modelling* **221**: 1451–1462.

Gutiérrez-Estrada JC, Silva C, Yáñez E, Rodríguez N, Pulido-Calvo I. 2007. Monthly catch forecasting of anchovy *Engraulisringens* in the

north area of Chile: Non-linear univariate approach. *Fisheries Research* **86**: 188–200.

Gutiérrez-Estrada JC, Yáñez E, Pulido-Calvo I, Silva C, Plaza F, Bórquez C. 2009. Pacific sardine (*Sardinopssagax*, Jenyns 1842) landings prediction.A neural network ecosystemic approach. *Fisheries Research* **100**: 116–125.

Hair JF, Anderson RE, Tatham RL, Babin B, Black B (eds). 2005. *Multivariate Data Analysis*, 6th edn. Prentice-Hall: London, UK; 928.

Hartigan JA, Wong MA. 1979. Algorithm AS 136.A *K*-means clustering algorithm. *Applied Statistics* **28**(1): 100–108.

Holliday JD, Rodgers SL, Willett P. 2004. Clustering files of chemical structures using the fuzzy *K*-means clustering method. *Journal of Chemical Information and Computer Sciences* **44**: 894–902.

Hsu K, Gupta HV, Sorooshian S. 1995. Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research* **31**(10): 2517–2530.

Hunter A, Kennedy L, Henry J, Ferguson I. 2000. Application of neural networks and sensitivity analysis to improved prediction of trauma survival. *Computer Methods and Programs in Biomedicine* **62**: 11–19.

Iyer MS, Rhinehart RR. 1999. A method to determine the required number of neural-network training repetitions. *IEEE Transactions on Neural Networks* **10**(2): 427–432

Jain A, Sudheer KP, Srinivasulu S. 2004. Identification of physical processes inherent in artificial neural network rainfall runoff models. *Hydrological Processes* **18**: 571–581.

Kahya E, Demirel MC, Beg OA. 2008. Hydrologic homogeneous regions using monthly streamflow in Turkey. *Earth Science Resources Journal* **12**(2): 181–193.

Kitanidis PK, Bras RL. 1980. Real time forecasting with a conceptual hydrological model. 2. Applications and results. *Water Resources Research* **16**(6): 1034–1044.

Legates DR, McCabe Jr. GJ. 1999. Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research* **35**(1): 233–241.

Maier HR, Dandy GC. 1996. Neural network models for forecasting univariate time series. *Neural Network World* **5**: 747–772.

Maier HR, Dandy GC. 1997. Methods for determining the inputs to artificial neural networks for forecasting water quality in rivers. *Microcomputers in Civil Engineering* **12**: 353–368.

Maier HR, Dandy GC. 2000. Neural networks for the prediction and forecasting of water resources variables: are view of modeling issues and applications. *Environmental Modelling and Software* **15**: 101–124.

Matheron G. 1963. Principles of geostatistics. *Economic Geology* **58**: 1246–1266.

May RJ, Maier HR, Dandy GC, Fernando TMKG. 2008. Non-linear variable selection for artificial neural networks using partial mutual information. *Environmental Modelling and Software* **23**(10–11): 1312–1326.

McKee TB, Doesken NJ, Kleist J. 1993.The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*. American Meteorological Society: Boston; 179–184.

Moddemeijer R. 1989. On estimation of entropy and mutual information of consinuous distributions. *Signal Processing* **16**(3): 233–246.

Murphy AH. 1991. Forecast verification: its complexity and dimensionality. *Monthly Weather Review* **119**: 1590–1601.

Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models. I. A discussion of principles. *Journal of Hydrology* **10**: 282–290.

Pardo-Igúzquiza E. 1998. Optimal selection of number and location of rainfall gauges for areal rainfall estimation using geostatistics and simulated annealing. *Journal of Hydrology* **210**: 206–220.

Portela MM, Quintela AC. 2006. Estimação em Portugal Continental de escoamento e de capacidadesuteis de albufeiras de regularização na ausência de informação. *Recursos Hidrícos* **27**(2): 7–18 (in Portuguese).

Pulido-Calvo I, Portela MM. 2007.Application of neural approaches to one-step daily flow forecasting in Portuguese watersheds.*Journal of Hydrology* **332**: 1–15.

Rumelhart DE, Hinton GE, Williams RJ. 1986. 'Learning' representations by backpropagation errors. *Nature* **323**: 533–536.

Santos JF, Portela MM, Pulido-Calvo I. 2011. Regional frequency analysis of droughts in Portugal. *Water Resources Management* **25**(14): 3537–3558.

Santos JF, Pulido-Calvo I, Portela MM. 2010. Spatial and temporal variability of droughts in Portugal. *Water Resources Research* **46**: W03503. DOI: 10.1029/2009WR008071.

Schleiter IM, Borchardt D, Wagner D, Dapper T, Schmidt KD, Schmidt HH, Werner H. 1999. Modelling water quality, bioindication and population dynamics in lotic ecosystems using neural networks. *Ecological Modelling* **120**: 271–286.

Senthil-Kumar AR, Sudheer KP, Jain SK, Agarwal PK. 2005. Rainfall-runoff modelling using artificial neural networks: comparison of network types. *Hydrological Processes* **19**: 1277–1291. DOI: 10.1002/hyp.5581

Shannon CE. 1948. A mathematical theory of communications. *The Bell System Technical Journal* **27**: 379–423 and 623–656.

Sharma A. 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part1— A strategy for system predictor identification. *Journal of Hydrology* **239**: 232–239.

Shrestha RR, Theobald S, Nestmann F. 2005. Simulation of flood flow in a river system using artificial neural networks. *Hydrology and Earth System Sciences* **9**(4): 313–321.

Tsakiris G. 2008. Uni-dimensional analysis of droughts for management decisions. *European Water* **23/24**: 3–11.

Vicente-Serrano SM, Saz-Sánchez MA, Cuadrat JM. 2003. Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): application to annual precipitation and temperature. *Climate Research Journal* **24**: 161–180.

Webster R, Oliver MA. 1990. Numerical classification: Non hierarchical methods. In *Statistical Methods in Soil and Land Resource Survey*, chap. 11. Oxford Univ. Press: Oxford, U.K; 191–212.

Willmott CJ. 1982. Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society* **63**: 1309–1313.

Yáñez E, Plaza F, Gutiérrez-Estrada JC, Rodríguez N, Barbieri MA, Pulido-Calvo I, Bórquez C. 2010. Anchovy (*Engraulis ringens*) and sardine (*Sardinops sagax*) abundance forecast off northern Chile: A multivariate ecosystemic neural network approach. *Progress in Oceanography* **87**: 242–250.

Zheng GL, Billings SA. 1996. Radial Basis Function Network Configuration using Mutual Information and the Orthogonal Least Squares Algorithm. *Neural Networks* **9**: 1619–1637.

