Identification of dominant sources of sea level pressure for precipitation forecasting over Wales

Azadeh Ahmadi and Dawei Han

ABSTRACT

Downscaling methods are utilized to assess the effects of large scale atmospheric circulation on local hydrological variables such as precipitation and runoff. In this paper, a methodology of statistical downscaling using a support vector machine (SVM) approach is presented to simulate and predict the precipitation using general circulation model (GCM) data. Due to the complexity and issues related to finding a relationship between the large scale climatic parameters and local precipitation, the climate variables (predictors) affecting monthly precipitation variations over Wales are identified using a combination of the methods including the principal component analysis (PCA), fuzzy clustering, backward selection, forward selection, and Gamma test (GT). The effectiveness of those tools is illustrated through their implementations in the case study. It has been found that although the GT itself fails to identify the best input variable combination, it provides useful and narrowed-down options for further exploration. The best input variable combination is achieved by the GT and forward selection method. This approach can be a useful way for assessing the impacts of climate variables on precipitation forecasting.

Key words | backward selection, cross validation, forward selection, gamma test (GT), precipitation prediction, support vector machine

Azadeh Ahmadi (corresponding author)
Department of Civil Engineering,
Isfahan University of Technology,
Isfahan,
Iran
E-mail: *aahmadi@cc.iut.ac.ir*

Dawei Han
Water and Environmental Management Research Centre,
Department of Civil Engineering,
University of Bristol,
Bristol,
UK

INTRODUCTION

A general circulation model (GCM) is a numerical model based on the conservation laws of momentum, energy, and water vapor utilized for analysis of atmosphere in all three spatial dimensions. GCMs are the most reliable and powerful tool for estimating the climate variables at large spatial scale. GCMs provide outputs at nodes of grid boxes, which are tens of thousands of square kilometers in size. However, most hydrological models used in climate studies need to simulate sub-grid scale (higher spatial resolution). Therefore, there is a need to downscale the GCM outputs to higher resolutions.

Statistical downscaling methods establish a statistical relationship between one or several large-scale meteorological variables (including air temperature, relative humidity, specific humidity, geo-potential height, zonal, vertical, and meridional wind velocities at various pressure levels and sea level pressure (SLP)), and local-scale variables such as precipitation, temperature, and stream flow. This is done by translating anomalies of the large scale climate data (predictors) into anomalies of some local scale variable (predictand). The most popular approach of downscaling is the regression method that relies on the direct quantitative relationship between the local scale climate variable (predictand) and the variables containing the large scale climate information (predictors) through some form of regression.

Many previous studies about the selection of predictors have been based on regression analysis on the fundamental assumption that regional climate is conditioned by large scale atmospheric state and antecedent predictands variables. In order to find the most effective nodes of grid boxes of GCM outputs, the regression analysis is utilized as well. At the scale of the UK, several studies have investigated the connections between surface temperature and pressure as predictors and rainfall as predictand (Barrow &

doi: 10.2166/hydro.2012.110

Semenov 1995; Colman 1997; Wilby 1998, 2001, 2005; Colman & Davey 1999; Haylock *et al.* 2006). Statistical downscaling based on empirical relationships between the large scale climate and regional climate variables is also used to carefully select large scale predictor variables (Wilby *et al.* 2004; Schmidli *et al.* 2007).

Individual downscaling schemes differ according to the choice of mathematical transfer function, predictor variables, or statistical fitting procedure. Commonly applied methods include multiple regression (Murphy 1999), canonical correlation analysis (CCA) (von Storch *et al.* 1993), artificial neural networks (Crane & Hewitson 1998) and support vector machine (SVM) (Tripathi *et al.* 2006; Ghosh & Mujumdar 2008; Najafi *et al.* 2011), gene expression programming (GEP) (Hashmi *et al.* 2011) which is similar to nonlinear regression for prediction.

Recently, owing to the advancement of modern computing technology, a new algorithm from the computing science community called the Gamma test (GT) (Agalbjörn *et al.* 1997; Končar 1997) has been proposed to identify the model input selection. GT is a data-driven analysis tool with a potential to select the input variables without the excessive detailed model development. The GT is a technique for estimating the noise level present in a data set, which is directly estimated from the data without assuming anything regarding the parametric form that governs the system (Končar 1997; Stefánsson *et al.* 1997). The GT has been explored in some studies related to hydrological modeling (Remesan *et al.* 2008; Ahmadi *et al.* 2009; Moghaddamnia *et al.* 2009a, 2009b; Piri *et al.* 2009; Wan Jaafar *et al.* 2011). Its usefulness is derived from the fact that the low noise levels will only be encountered when all of the principal causative factors that determine output have been included in the input. Variables selection process could be performed in the GT for estimating noise levels for every possible subset of the input variables.

Alternatively, principal component analysis (PCA) is one of the multivariate statistical methods which can be used to reduce the input variable complexity when we have a huge volume of information (Camdevyren *et al.* 2005). PCA changes the input variables into principal components (PCs) that are independent and linear compounds of input variables (Lu *et al.* 2003). Instead of the direct use of input variables, they are transformed into PCs and then used as input variables. In this method, the information of input variables will be presented with minimum losses in PCs (Helena *et al.* 2000).

In this study, the relationship between rainfall of Wales in the United Kingdom and SLP from GCMs is examined for precipitation forecasting. Although other weather variables such as temperature may also affect rainfall, only pressure is considered in this study to illustrate and compare the alternative methods. First, the historical monthly rainfall and grid point SLP data whose latitude ranges from 25°N to 75°N and longitude ranges from 75°W to 15°E at a spatial resolution of 3.75° by National Center for Environmental Prediction (NCEP) are extracted. The number of SLP gridded data clusters is determined based on the PCA. Then, the SLP gridded data are clustered in similar groups using the fuzzy clustering method. In order to determine the most important cluster centers in downscaling modeling, the regression method, GT, backward selection, and forward selection are used to select the most effective input variables for SVM modeling. Finally, the results of different approaches for input selection are compared. The paper is organized as follows: the methods are briefly described in the next section, including the PCA, fuzzy clustering, the GT, backward/forward selection and leave-one-out cross validation (LOOCV) and SVM. This is followed by a case study for implementing the proposed methodology. Next, the obtained results are presented, followed by a summary and conclusion.

MATERIALS AND METHODS

The PCA technique

The PCA was invented in 1901 by Karl Pearson (Pearson 1901). PCA is a statistical procedure to transfer correlated variables into a set of uncorrelated variables by identifying the patterns of multidimensional variables. In the PCA method, the eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix is calculated. Through the PCA procedure, the internal structure of data in an unbiased way is revealed for selecting the model predictors. PCA identifies patterns in data and presents the data in a way to highlight their similarities and differences.

PCA is based on the statistical representation of random variables. Suppose a random vector population X, where

$$X = (x_1, \ldots x_n)^T \tag{1}$$

and the mean of that population is denoted by:

$$\mu_X = E\{X\} \tag{2}$$

For performing PCA, first the covariance matrix of the normalized variables is computed:

$$C_x = E\left\{(X - \mu_X)(X - \mu_X)^T\right\} \tag{3}$$

where the components of C_x, denoted by c_{ij}, represent the covariance between the random variable components x_i and x_j. The component c_{ii} is the variance of the component x_i.

Using the symmetric covariance matrix, an orthogonal basis is obtained by finding its eigenvalues and eigenvectors. The eigenvectors e_i and the corresponding eigenvalues λ_i are the solution of the following equation:

$$C_x e_i = \lambda_i e_i \quad i = 1, \ldots, n \tag{4}$$

These values can be determined by finding the solution of the characteristic equation:

$$|C_x - \lambda I| = 0 \tag{5}$$

where the I is the identity matrix having the same order as C_x and |.| denotes the determinant of the matrix. The characteristic equation has an order n. Then the eigenvectors are ordered by descending eigenvalues. The percentage of information transferred with different variables can be presented by calculating variance proportion as follows:

$$w_i = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{6}$$

where w_i is the percentage of the total variance explained by the ith PCs. The PCA is based on the linear systems and may not be very effective when the system is very nonlinear.

The Fuzzy clustering method

Cluster analysis is used to divide data into groups (clusters) in such a way that similar data objects are assigned to the same cluster and dissimilar data objects to different clusters. The data clustering improves data understanding and reveals its internal structure. In fuzzy logic, we speak about degrees of belief between '0 and 1' or 'true and false' where the degree of membership is expressed by a membership function, ranging from '0 to 1.' In fuzzy clustering method, the membership function is used to assign the data to the clusters. The most prominent fuzzy clustering algorithm is the fuzzy c-means, a fuzzification of k-means (Bezdek 1981).

The fuzzy c-means algorithm is utilized to partition a collection of elements $X = (x_1, \ldots, x_n)$ into a collection of c fuzzy clusters with respect to some given criteria. The algorithm provides a list of c cluster centers $X = (c_1, \ldots, c_c)$ and a partition matrix $U = u_{ij} \quad i = 1, \ldots, n, \; j = 1, \ldots, c,$ including the degree to which element x_i belongs to cluster c_j.

The LOOCV method

Cross validation is one of the most commonly used model selection criteria in data mining (Allen 1974; Geisser 1975). This method can be used to estimate the generalization error which is the modeling error on the unseen data of the model. Cross validation can be used for model selection considering the smaller generalization error. The idea behind cross validation is that any assumed model fitted to the data should be verified to make sure this model can be generalized to the future data (Hawkins *et al.* 2003). In a cross-validation method, the historical data are split into the training data for calibrating the model and the testing data for measuring the generalization model performance. There are several types of cross-validation methods and among them, the LOOCV is used in this study as follows:

1. The N observed data are split into two categories, the N–1 data for training (i.e., calibrating) the model and the one left out is for testing.

2. The model parameters are obtained through the training process on the training data followed by calculating the error on the testing data.
3. This procedure is then repeated for different data until all the N data are utilized once as the testing data. Then the averaged N testing error is computed on the basis of all the testing errors.

In this study, there are 720 observed monthly data including SLP gridded data clusters as the predictors and the monthly Welsh precipitation as the predictant. In each iteration (720 iterations), 719 observed data are used for training and one left out for testing. In each iteration, the precipitation prediction error is estimated for one-monthly data which were left out for testing. The root mean squared error (RMSE) and mean averaged error (MAE) are estimated for a single data point. Therefore, 720 values for the precipitation prediction errors from iterations are obtained. Finally, the average of the precipitation prediction error over 720 iterations is calculated.

The LOOCV method is used to generate a model with good generalization for future prediction. When the total number of data is very high or the model calibration/testing is time consuming, this approach would increase the computation burden.

Backward and forward selection procedure

In the backward selection method, the variables are iteratively incorporated into larger and larger sets of variables.

1. Assume the total number of variables is P. At the beginning, all the variables are used to create a model with P input variables.
2. One variable is removed from the input variables. Thus, P states of input variables combination are generated. For each combination, one model type is developed, so there are P model types. For each model type, N LOOCV tasks are performed. The number of models explored is NP. The best model of P–1 input variables is selected based on the minimum LOOCV testing errors.
3. The best model of the previous step is used as a base to explore a second input variable that should be removed. As in Step 2, for each P–1 input variable combination, a model is constructed with N LOOCV tasks and an average testing error is estimated for the input variable combination. The number of models explored is $N(P–1)$. The best model is chosen based on the minimum errors.
4. This procedure is iterated until enough input variables are included in the model.

In the forward selection procedure which is similar to the backward selection procedure, at the beginning one variable is used and variables are progressively added one by one.

In this study, at first all SLP gridded data clusters are used as the predictors (variables) for Welsh precipitation prediction. In order to evaluate the model performance, the general prediction error is obtained using the LOOCV method. Then one variable, a SLP cluster, is removed from the predictors. The remaining clusters are used to develop the precipitation prediction model and the model's performance is assessed by the LOOCV technique. This procedure of removing the input variables is iterated until a stopping criterion is reached (in this study, it is when the error starts to grow again). One of the aims of this paper is assessment of different methods including correlation analysis, GT, and multicollinearity methods in selecting the removable variables.

The Gamma test (GT)

The GT is used to examine the relationship between inputs and outputs in numerical data sets without a need to construct a prediction model. The GT is used prior to modeling and estimating the variance of the output, even though the model is unknown. This error variance estimate presents a target mean squared error that any smooth nonlinear function should attain on the unseen data. Suppose we have a set of observed data represented by:

$$((x1, \ldots, x_M), y) = (x, y) \tag{7}$$

where the vector $X = (x_1, \ldots, x_M)$ is the *input*, confined to a closed bounded set $C \in R^M$ and the scalar y is the corresponding *output*, without loss of generality. The only assumption made is that the relationship of the system is

in the following form:

$$y = f(x_1, \ldots, x_M) + r \tag{8}$$

where f represents a smooth function and r denotes an indeterminable part, which may be due to real noise or lack of functional determination in the assumed input/output relationship. The GT is used for a data-derived estimation for *Var(r)* without knowing the underlying function f, just directly from the data. The estimate of the model's output variance called the Gamma statistic and represented by Γ cannot be accounted for by a smooth data model. The GT is derived from the Delta function of the input vectors:

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} \left|x_{N[i,k]} - x_i\right|^2 \tag{9}$$

where *N[i,k]* shows the index of the kth $((1 \le k \le k_{\mathrm{Max}}))$ nearest neighbor to x_i, and $|.|$ denotes Euclidean distance. Thus $\delta_M(k)$ is the mean square distance to the kth nearest neighbor. The corresponding Gamma function of the output values is:

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M} \left(y_{N[i,k]} - y_i\right)^2 \tag{10}$$

The GT computes the mean-squared kth nearest neighbor distances $\delta(k)$, $(1 \le k \le k_{\mathrm{Max}})$ and the corresponding $\gamma(p)^2$ where k_{Max} is the maximum number of nearest neighbors. In order to compute Γ the best line is constructed for the p points $(\delta_M(k), \gamma_M(k))$, and the vertical intercept, Γ is the GT value. The regression line slope shows the complexity of the model f. The V_{ratio} is the standardized results considering $\Gamma/\mathrm{Var(y)}$. It estimates a scale invariant noise which normally lies between zero and one.

Support vector machines (SVM)

The concept of SVM has been developed by Vapnik (1995, 1998). SVM is based on the principle of structural risk minimization from statistical learning theory. The application of SVM has received attention in the field of hydrological engineering and water resources management due to its many interesting features and promising empirical performance (Choy & Chan 2003; Bray & Han 2004; Yu *et al*. 2004; Sivapragasam & Liong 2005; Karamouz *et al*. 2009).

The SVM model is produced by support vectors included in the training data and presents the means of small subsets of training points. The cost function for building the model ignores any training data that are within a threshold ε to the model prediction. In the SVM method, the generalization bounds are relied on defining the loss function that ignores errors. In SVM, the problem is to find a linear function that best interpolates a set of training points for the following equation:

$$y = Wx + b \tag{11}$$

The parameters (W,b) should be determined to minimize the sum of the squared deviations of the data utilizing the least squares approach:

$$\sum_{i=1}^{l} (y_i - Wx_i - b)^2 \tag{12}$$

Some deviation ε between the eventual targets y_i and the function y is allowed by defining the following constraint:

$$(y_i - Wx_i \pm b) \le \varepsilon \tag{13}$$

A band or a tube around the hypothesis function y can be visualized with points outside the tube regarded as training errors, otherwise called slack variables ξ_i. For points inside the tube, the slack variables are zero and increase gradually for points outside the tube. This approach to regression is called ε-SV regression (Vapnik 1998). It can be shown that this regression problem can be expressed as the following convex optimization problem:

$$\mathrm{Min}\,\frac{1}{2}||w|| + C\sum_{i=1}^{l} \left(\xi_i + \xi_i^*\right) \tag{14}$$

Subject to:

$$\begin{aligned} &y_i - (W.x_i + b) \le \varepsilon + \xi_i \\ &(W.x_i + b) - y_i \le \varepsilon + \xi_i^* \\ &\xi_i, \xi_i^* \ge 0, i = 1, 2, \ldots, l \end{aligned} \tag{15}$$

where C is a pre-specified and positive constant that determines the degree of penalized loss when a training error occurs, ξ_i and ξ_i^* are slack variables that represent the upper and the lower training errors subject to an error tolerance ε. Then the Lagrange function is constructed from both the objective function and the corresponding constraints to solve the optimization problem. SVMs are characterized by the usage of kernel function used to change the representation of the data in the input space to a linear representation in a higher dimensional space called a feature space.

Genetic algorithm (GA)

GA is an evolutionary algorithm which utilizes a random search technique to imitate the biological and genetic process to evolve the potential solution. Each chromosome represents a potential solution to a problem. The potential solution is evaluated based on the fitness function which should progressively be improved towards an optimum solution. In each iteration, the new chromosomes of the new population are generated through genetic operators including Selection, Crossover, and Mutation. The selection operator chooses the parents for generating the next population. It can significantly improve the divergence process and decrease the run-time of the model. Several methods have been used to select the parent chromosomes, such as Roulette Wheel and Tournament methods. Through the crossover operator, certain parts of two chromosomes are randomly exchanged and new chromosomes are built based on a specific probability (Pc). The mutation operator is used to avoid being trapped in the local optimal solution. The mutation operator changes the gene value randomly considering the range of variation of each gene regarding a probability of occurrence (Pm). More details of genetic algorithms can be obtained in the works of Michalewicz (1992) and Gen & Cheng (2000).

APPLICATION OF THE METHODOLOGY TO A CASE STUDY

Rainfall in Wales varies widely, with the highest average annual totals recorded in the mountainous areas of Snowdonia and the Brecon Beacons, where the yearly fall is comparable with that in the English Lake District or the western Highlands of Scotland. In the east, close to the border with England, annual totals are similar to those over much of the English Midlands. Snowdonia is the wettest part of Wales with average annual totals exceeding 3,000 mm, but coastal areas and the east receive less than 1,000 mm a year (UK Meteorological Office). The spatial distribution of rainfall in Wales is presented in Figure 1.

The monthly precipitation data were taken for the period 1948–2007 (720 monthly data) from the UK Meteorological Office. The monthly and annual time series of the Welsh precipitation are presented in Figures 2 and 3. The predictor used in this study was the monthly SLP of grid points whose latitudes range from 25°N to 75°N and longitudes range from 75°W to 15°E. The CGCM2 grid is uniform along the longitude with the grid box size of 5° and nearly uniform along the latitude (approximately 5°) as shown in Figure 4. The total number of grid points is $11 \times 19 = 209$. The data are extracted from the NCEP's reanalysis data bank from 1948 to 2007 (available for free

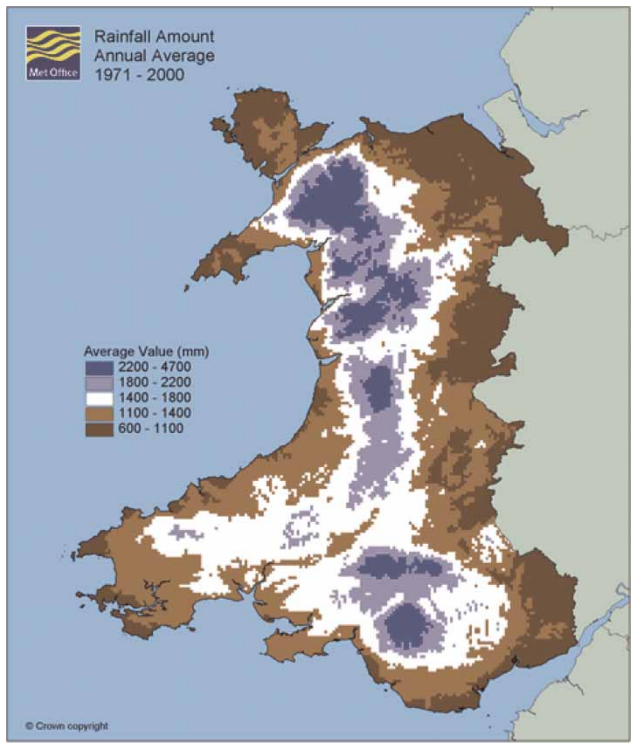

Figure 1 | The spatial distribution of annual precipitation in Wales.

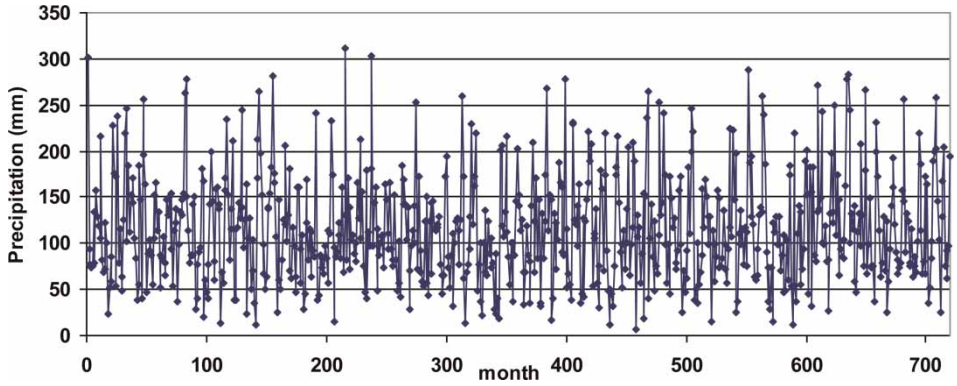

Figure 2 | The monthly variation of average Welsh precipitation.

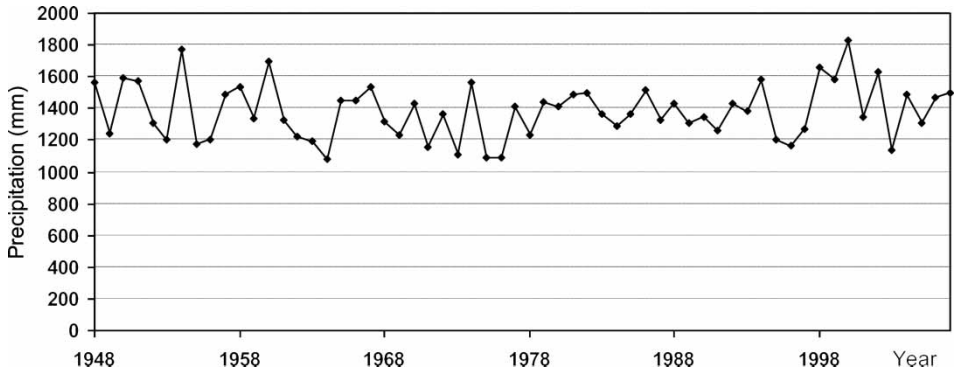

Figure 3 | The annual variation of Welsh precipitation.

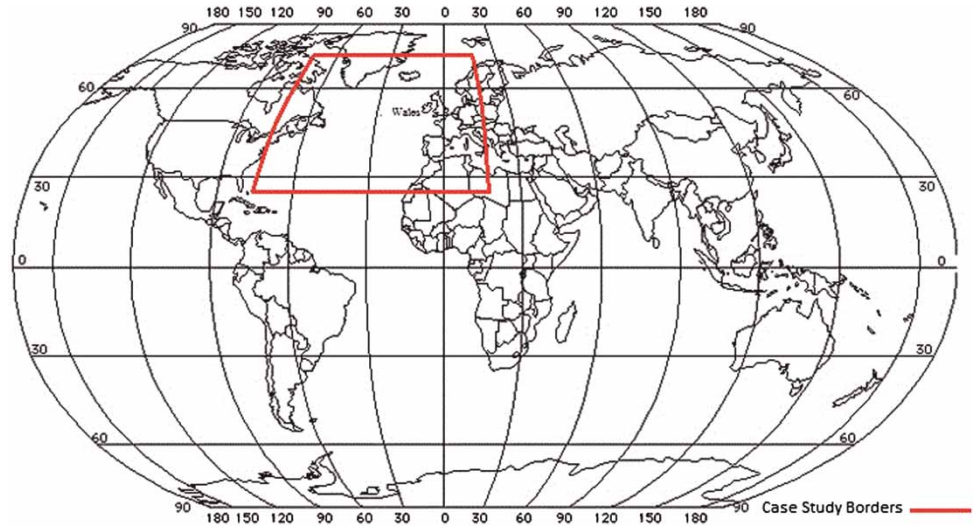

Figure 4 | The location of the case study and the geographical extent of SLP predictor.

download at the NCEP/NCAR internet site http://dss.ucar.edu/pub/reanalysis/).

RESULTS

Application of PCA and fuzzy clustering

In order to find the pattern in the 209 grid points and reduce the number of dimensions, the PCA has been applied. The eigenvalue decomposition of a data covariance matrix is calculated. Figure 5 shows the percentage of information which is transferred with different grid points. In Table 1, eigenvalues, variance proportion, and cumulative variance proportion for the top 20 components are shown.

Clearly, the first three components account for approximately 68.34, 12.79 and 6.57% of the total variance in the data sets, respectively. The first 20 components together account for about 99.47% of the total variance and the rest only accounted for about 0.53%. Also, 98.06 and 93.96% of the total variance are taken with 10 and 5 grid points, respectively. Therefore, in this stage, the number of grid points without significant information redundancy is considered as 10. In the next step, 10 SLP gridded data are selected using the GT and correlation coefficient methods.

In order to consider the similarity between the data and classify similar grid points into distinctive groups, a fuzzy *c*-means clustering has been done. The number of classifications is considered as 10 which is the number of PCs. The cluster centers are determined based on the membership function of the grid points of each cluster. Figure 6 shows the results of classifying the grid points into 10 clusters. The average monthly SLP for 10 clusters are presented in Figure 7. Figure 6 shows the distribution of the clusters in the study area. It can be seen that there is a low SLP region in the north including Clusters 8, 6, 10, and 1 and high SLP in the south including Clusters 5, 4, and 3. The SLP difference between these two regions might be a major driving force of rainfall in Wales. In the next stage, the most effective input variables on the Welsh rainfall prediction are assessed.

Identification of lag times

A lag correlation analysis between the standardized precipitation and cluster centers of SLP is carried out to find the lag time of the SLP influence on the Welsh precipitation. A lag correlation analysis of the SLP shows that the monthly precipitation has higher correlation with the SLP variation in the previous months. The results of correlation values between the precipitation and the cluster centers of SLP variables with lags from zero to six are presented in Table 2. The results show the precipitation has the most correlation with the zero lag-SLP. However, if we want to forecast the precipitation a few months ahead, the correlation of clusters with 5 month lags is relatively high. Therefore, the forecasting lead time is considered as 5 months.

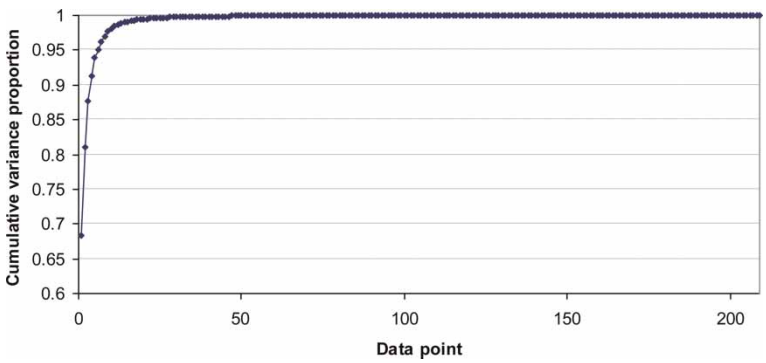

Figure 5 | Variation of cumulative variance proportion for different grid points.

Table 1 | Eigenvalues obtained through PCA application

Grid point	Eigenvalue	Variance proportion	Cumulative variance proportion
1	11,430	0.683414	0.6834
2	2,139.5	0.127923	0.8113
3	1,098.1	0.065657	0.8770
4	585.88	0.035031	0.9120
5	460.52	0.027535	0.9396
6	199.83	0.011948	0.9515
7	183.34	0.010962	0.9625
8	132.15	0.007901	0.9704
9	111.45	0.006664	0.9770
10	59.299	0.003546	0.9806
11	56.245	0.003363	0.9839
12	53	0.003169	0.9871
13	28.093	0.00168	0.9888
14	23.255	0.00139	0.9902
15	16.828	0.001006	0.9912
16	14.961	0.000895	0.9921
17	14.163	0.000847	0.9929
18	11.487	0.000687	0.9936
19	10.116	0.000605	0.9942
20	8.6642	0.000518	0.9947

Data selection methods

Input selection using the correlation coefficient

In this method, the predictor variables are selected based on high correlation with the precipitation. The most correlated SLP cluster data with the precipitation are Clusters 8, 6, 10, 3, 1, 5, 2, 9, 7, and 4, respectively. A backward approach is adopted so that the first model starts with all the variables which are gradually removed one by one by assessing their correlation coefficients. Therefore, nine models with 10 to one number of highly correlated variables are generated as shown in Table 3. The masks of the precipitation forecasting models with 5 month lead time are presented in this table.

These results are then further investigated by running the sampling method of the LOOCV. This method holds one data out for testing and the remaining data for training the prediction model (i.e., the SVM model in this study). The process is repeated until all the variables are utilized as testing data and finally, the model performance can be computed by averaging the results of all the testing data. The model's performance has been evaluated using the LOOCV method in order to obtain the generalization error of the models to the testing data. The RMSE and MAE are used to evaluate the models:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{P} - P_i\right)^2} \tag{16}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{P} - P_i\right| \tag{17}$$

where $\hat{P}_i$ and P_i are the estimated and observed precipitation and N is the number of historical precipitation

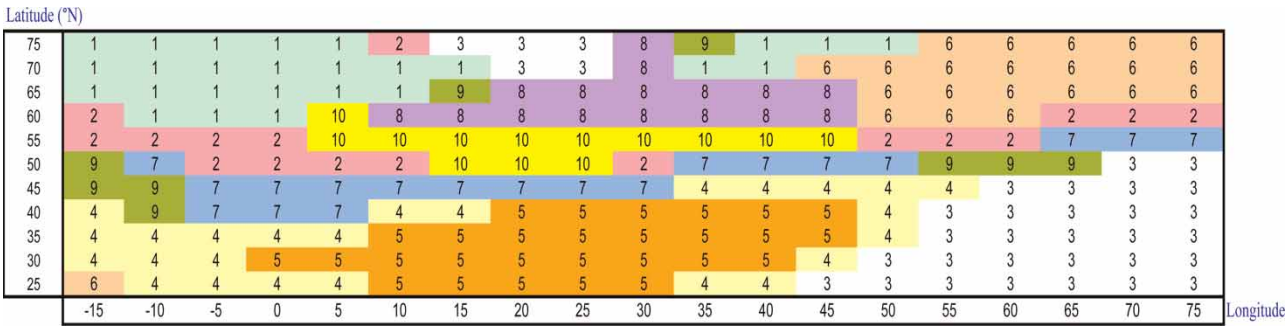

Figure 6 | Classified grid points of SLP in the region.

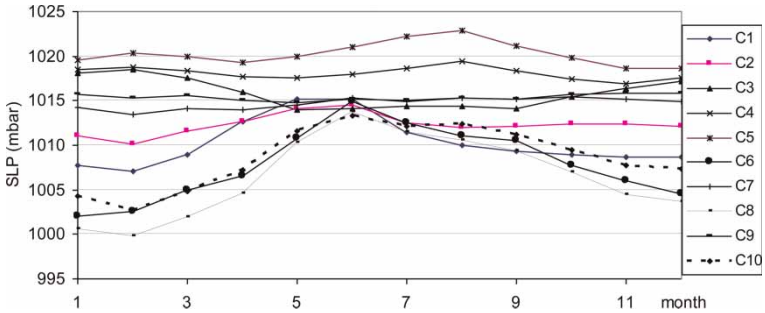

Figure 7 | The variation of average monthly SLP of different clusters.

Table 2 | The correlation coefficients between the precipitation and the mean values of cluster centers in different time lags

	Time lags						
Cluster	**0**	**1**	**2**	**3**	**4**	**5**	**6**
All	0.523	0.14	0.157	0.155	0.152	0.19	0.146

data. Finally, the model with the lowest RMSE and MAE on the testing data is selected as this is the most optimal model among those tested. Table 3 indicates the performance errors for the training and testing data sets from combinations of 10 to one variables as identified by the correlation coefficient analysis. The RMSE value is the same as MAE value, because these errors in testing process are computed for one data set. The results show that the model with 10 input variables has the lowest RMSE and MAE on the testing data and therefore, this model is the best one to chose. The SVM model's parameters are determined based on the minimum prediction error. The best model is selected as the epsilon-SVR with radial basis kernel function.

The forward method is also used for input selection. The selection procedure starts with one variable which is highly correlated with the precipitation and then adds one by one. The input selection using the forward method gives similar results to the backward method. The results presented in Table 3 show that the model with 10 variables is the best one in the testing step.

Table 3 | Model selection by the correlation analysis through the backward selection

			Training data set		Testing data set	
Number of variables	**Mask**	**Combination of variables**	**RMSE (mm)**	**MAE (mm)**	**RMSE (mm)**	**MAE (mm)**
10	1111111111	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	23.161	12.142	45.712	45.712
9	1110111111	$x_1, x_2, x_3, x_5, x_6, x_7, x_8, x_9, x_{10}$	26.807	14.341	45.857	45.857
8	1110110111	$x_1, x_2, x_3, x_5, x_6, x_8, x_9, x_{10}$	28.029	15.261	45.897	45.897
7	1110110101	$x_1, x_2, x_3, x_5, x_6, x_8, x_{10}$	31.537	17.622	46.762	46.762
6	1010110101	$x_1, x_3, x_5, x_6, x_8, x_{10}$	37.434	22.334	47.717	47.717
5	1010010101	$x_1, x_3, x_6, x_8, x_{10}$	43.283	28.648	47.961	47.961
4	0010010101	x_3, x_6, x_8, x_{10}	44.041	34.27	48.61	48.61
3	0000010101	x_6, x_8, x_{10}	44.162	38.769	52.29	52.29
2	0000010100	x_8, x_{10}	44.638	41.73	54.381	54.381
1	0000010000	x_8	46.448	43.375	55.548	55.548

Input selection using the correlation coefficient considering multicollinearity

Multicollinearity occurs when two or more predictor variables in a multiple regression model are highly correlated. In this section, the multicollinearity among predictors in the models is explored. For this purpose, the possibility of multicollinearity is assessed by carrying out the correlation matrix for all the variables. The correlation value varies between −1 and +1. The correlation coefficient values are presented in Table 4 and four pairs of variables have coefficient values more than 0.8, which can be classified as highly correlated.

In this method, the predictor variables are selected based on high correlation with the precipitation and multicollinearity analysis. A backward approach is adopted so that the first model starts with all the variables which are gradually removed one by one by assessing their correlation coefficients with the precipitation and other variables. The second model is generated with nine variables by removing Cluster 4 which has the least correlation with the precipitation. The cluster centers are ranked based on the correlation coefficients between the precipitation and the mean values of cluster centers as shown in Table 5.

Then the third model is generated by removing Cluster 7 with eight input variables. In generating the model with seven input variables, since Cluster 9 has a high correlation with Cluster 7 at 0.896, Cluster 2 is removed. Then, at the stage of input selection for the model with six variables, Cluster 5 should be removed. But, because it has a high correlation with Cluster 4 at 0.801, Cluster 5 remains and Cluster 1 is removed. The model with five variables is constructed by removing Cluster 3 without any significant correlated coefficient with other variables. Finally, Clusters 10, 6, 9, and 5 as the highly correlated variables with other variables are removed.

Finally, nine models with 10 to one variables are generated as shown in Table 6. These results are then further investigated by developing the SVM model with the LOOCV method. The results show that the model with 10 input variables has the lowest RMSE and MAE on the

Table 5 | Ranking the clusters based on the correlation coefficients between the precipitation and the mean values of cluster centers

Rank	Cluster	Correlation coefficients	Absolute (correlation coefficients)
1	4	0.031	0.031
2	7	0.057	0.057
3	9	−0.094	0.094
4	2	0.133	0.133
5	5	0.167	0.167
6	1	0.185	0.185
7	3	−0.313	0.313
8	10	0.318	0.318
9	6	0.321	0.321
10	8	0.332	0.332

Table 4 | The correlation coefficient between the SLP variables[a]

Covariance	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1	0.576	−0.469	−0.403	−0.326	0.587	−0.135	0.748	−0.310	0.552
x_2		1	−0.158	−0.023	−0.103	0.387	0.638	0.501	0.416	0.755
x_3			1	0.444	0.086	−0.694	0.287	−0.629	0.662	−0.487
x_4				1	**0.801**	−0.313	0.554	−0.342	0.635	−0.100
x_5					1	−0.123	0.365	−0.222	0.308	−0.015
x_6						1	−0.089	**0.809**	−0.358	0.574
x_7							1	0.007	**0.896**	0.472
x_8								1	−0.271	**0.804**
x_9									1	0.133
x_{10}										1

[a]Bold values indicate highly correlated.

Table 6 | Model selection by multicollinearity analysis and backward selection

			Training data set		Testing data set	
Number of variables	**Mask**	**Combination of variables**	**RMSE (mm)**	**MAE (mm)**	**RMSE (mm)**	**MAE (mm)**
10	1111111111	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	23.161	12.142	45.712	45.712
9	1101111111	$x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	26.047	13.879	46.398	46.398
8	1001111111	$x_1, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	29.379	16.153	47.052	47.052
7	1001011111	$x_1, x_4, x_6, x_7, x_8, x_9, x_{10}$	34.177	19.882	48.143	48.143
6	1001011011	$x_1, x_4, x_6, x_7, x_9, x_{10}$	37.697	22.738	47.592	47.592
5	1001011010	x_1, x_4, x_6, x_7, x_9	41.908	26.906	45.932	45.932
4	0001011010	x_4, x_6, x_7, x_9	43.635	33.231	47.437	47.437
3	0001010010	x_4, x_6, x_9	44.707	37.462	51.13	51.13
2	0000010010	x_6, x_9	45.448	41.189	53.609	53.609
1	0000010000	x_6	45.511	42.853	55.453	55.453

testing data and, therefore, this model is the best one to be chosen.

The forward method is also used for input selection. The selection procedure starts with one variable and then other variables are added by considering the correlation with the precipitation and multicollinearity. The most correlated SLP cluster with the precipitation is Cluster 8. Then Cluster 6 has high correlation with the precipitation while it has high correlation with Cluster 8 as 0.809. The next highly correlated cluster with the precipitation is Cluster 10 which is also correlated with Cluster 8. Therefore, in the second model, Cluster 8 and Cluster 3 are considered as the inputs. In the same manner, gradually more variables are added one by one. Finally Clusters 4, 10, 6, and 7 with the correlation coefficients at 0.801, 0.804, 0.809, and 0.896, respectively, with other variables are added.

Therefore, 10 models with one to 10 variables are generated as shown in Table 7. The results show that the model with 10 input variables has the lowest RMSE and MAE on the testing data and therefore, this model is the best one to be chosen. The comparison of the results presented in Tables 6 and 7 demonstrates that the developed

Table 7 | Model selection by multicollinearity analysis and forward selection

			Training data set		Testing data set	
Number of variables	**Mask**	**Combination of variables**	**RMSE (mm)**	**MAE (mm)**	**RMSE (mm)**	**MAE (mm)**
1	0000010000	x_6	45.904	42.853	55.453	55.453
2	0000010001	x_6, x_{10}	44.911	40.896	53.673	53.673
3	0010010001	x_3, x_6, x_{10}	43.847	37.804	51.233	51.233
4	0010110001	x_3, x_5, x_6, x_{10}	43.635	31.312	47.065	47.065
5	0010111001	$x_3, x_5, x_6, x_7, x_{10}$	41.2	25.88	46.396	46.396
6	0010111101	$x_3, x_5, x_6, x_7, x_8, x_{10}$	37.914	22.803	47.622	47.622
7	1010111101	$x_1, x_3, x_5, x_6, x_7, x_8, x_{10}$	32.532	18.322	46.94	46.94
8	1110111101	$x_1, x_2, x_3, x_5, x_6, x_7, x_8, x_{10}$	28.275	15.297	45.753	45.753
9	1110111111	$x_1, x_2, x_3, x_5, x_6, x_7, x_8, x_9, x_{10}$	26.807	14.341	45.857	45.857
10	1111111111	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	23.161	12.142	45.712	45.712

model based on the forward selection has a lower testing error than the backward selection. Maybe it is because the forward selection picked up the promising variables in the initial steps of input selection procedure.

Input selection using GT

With the GT, the input variable selection can be carried out without estimating the model parameters. A *backward* approach is used in calculating the GT. Thus, the modeling procedure starts with all the potential input variables, and progressively one variable by turn is removed. After removing one variable, the gamma statistics are computed and this process is repeated until one variable is left. The variable with the higher GT value would be the less promising variable because of the error variance in the input–output is higher. The least promising variable is picked up by the GT value. The best combination of variables can be identified by observing its GT value. Therefore, a combination that gives the lowest GT value indicates the best input combination.

The GT values for input variable combinations are shown in Table 8, and the result is plotted in Figure 8. The masks of the best models with nine to one input variables are 1101111111, 1001111111, 1001011111, 1001011011, 1001011010, 0001011010, 0001010010, 0000010010, and 0000010000, respectively. It can be noted from Figure 8 that there is a local minimum with the GT value of 6.

In order to evaluate the GT for input selection, the SVM models are constructed using the selected combinations of these variables with the LOOCV principle for calculating the training and testing errors. Table 9 shows the RMSE and MAE for training and testing with the combination of 10 to one input variables as identified by the GT. The results show the model with 10 input variables has the lowest RMSE on the testing data and therefore, this model is the best one to be chosen. Figure 9 illustrates the structure of the model complexity (represented by the number of input variables) in relation to the training and testing RMSE values.

The forward method is also used for input selection with the GT and the results are shown in Table 10 and Figure 10.

Table 8 | Combination of variables with Gamma statistics and backward selection[a]

Number of variables	Mask	Gamma	Gradient	Standard error	V_{ratio}
10	1111111111	0.8596	−0.0225	0.0370	0.8596
9	0111111111	0.8348	−0.0116	0.0476	0.8348
	1011111111	0.8225	−0.0006	0.0344	0.8225
	1101111111	0.7874	0.0201	0.0397	0.7874
	1110111111	0.8237	−0.0071	0.0420	0.8237
	1111011111	0.8716	−0.0249	0.0324	0.8716
	1111101111	0.8449	−0.0196	0.0425	0.8449
	1111110111	0.8367	−0.0091	0.0424	0.8367
	1111111011	0.8765	−0.0328	0.0191	0.8765
	1111111101	0.8379	−0.0114	0.0273	0.8379
	1111111110	0.8045	0.0046	0.0315	0.8045
8	0101111111	0.8445	−0.0198	0.0405	0.8445
	1001111111	0.7884	0.0310	0.0261	0.7884
	1100111111	0.8056	0.0028	0.0366	0.0366
	1101011111	0.8216	0.0080	0.0280	0.8216
	1101101111	0.8320	−0.0138	0.0547	0.8320
	1101110111	0.8525	−0.0193	0.0314	0.8525
	1101111011	0.8057	0.0140	0.0285	0.8057
	1101111101	0.8521	−0.0250	0.0315	0.8521
	1101111110	0.8035	0.0077	0.0350	0.8035
7	0001111111	0.8335	−0.0141	0.0344	0.8335
	1000111111	0.8539	−0.0387	0.0365	0.8539
	1001011111	0.7690	0.0566	0.0423	0.7690
	1001101111	0.8799	−0.0523	0.0298	0.8799
	1001110111	0.8611	−0.0210	0.0317	0.8611
	1001111011	0.8008	0.0224	0.0285	0.8008
	1001111101	0.8514	−0.0241	0.0307	0.8514
	1001111110	0.7972	0.0219	0.0421	0.7972
6	0001011111	0.8446	−0.0132	0.0232	0.8446
	1000011111	0.9172	−0.0821	0.0272	0.9172
	1001001111	0.8269	0.0066	0.0397	0.8269
	1001010111	0.8749	−0.0488	0.0264	0.8749
	1001011011	0.7549	0.0788	0.0398	0.7549
	1001011101	0.7893	0.0373	0.0346	0.7893
	1001011110	0.8039	0.0346	0.0429	0.8039
5	0001011011	0.8223	0.0092	0.0473	0.8223
	1000011011	0.8721	−0.0349	0.0550	0.8721
	1001001011	0.8655	−0.0473	0.0275	0.8655
	1001010011	0.8263	−0.0108	0.0328	0.8263
	1001011001	0.8006	0.0397	0.0349	0.8006
	1001011010	0.7741	0.0788	0.0510	0.7741
4	0001011010	0.7858	0.0999	0.0438	0.7858
	1000011010	0.8767	−0.0902	0.0456	0.8767
	1001001010	0.9213	−0.1502	0.0338	0.9213
	1001010010	0.8594	−0.0383	0.0355	0.8594
	1001011000	0.7901	0.1070	0.0390	0.7901
3	0000011010	0.8399	0.0498	0.0323	0.8399
	0001001010	0.8116	0.3360	0.0357	0.8116
	0001010010	0.8034	0.0875	0.0466	0.8034
	0001011000	0.8513	−0.0570	0.0353	0.8513

(*continued*)

Table 8 | continued

Number of variables	Mask	Gamma	Gradient	Standard error	V_{ratio}
	0000010010	0.8488	0.0296	0.0437	0.8488
2	0001000010	0.9170	0.3228	0.0322	0.9170
	0001010000	0.8728	−0.1058	0.0272	0.8728
1	0000000010	0.9471	0.0291	0.0292	0.9471
	0000010000	0.8701	−0.1709	0.0253	0.8701

[a]The highlighted masks indicate the best mask in each category.

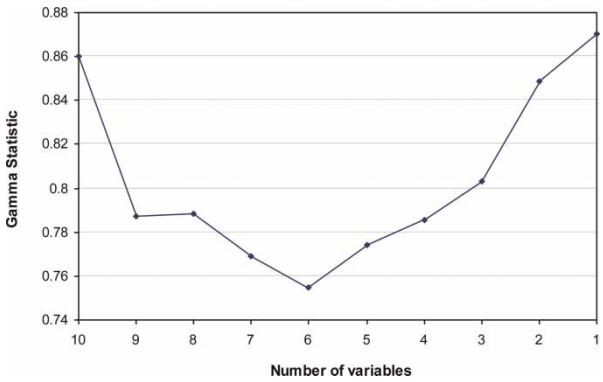

Figure 8 | GT values and number of input variables using backward selection.

The best model with four variables is the best model selected by the forward method.

In order to evaluate the GT for input selection, the SVM models are developed to simulate the model performance in training and testing steps with the LOOCV principle. Table 11 shows the RMSE and MAE for training and testing with the combination of 10 to one input variables as identified by the GT. The results show the model with seven input variables has the lowest RMSE on the testing data and, therefore, this model is the best one to be chosen. Figure 11 illustrates the structure of the model complexity (represented by the number of input variables) in relation to the training and testing RMSE values.

Model identification using GA

In the search for good irregular embeddings in a high dimensional input space, an alternative to the frequency analysis of a Gamma histogram is to use a genetic algorithm in which a mask's fitness is inversely proportional to its GT value.

The selection of individual solutions is performed in a probabilistic manner. The better solutions are the masks that represent the solutions with the lowest Gamma statistics and they have a greater chance of being selected for the next generation. Figure 12 is produced in real time to provide continuous feedback during the experiment. This feedback can be used to determine when to stop the algorithm, for example when there is convergence between the best individual fitness and the overall population fitness. Figure 12 shows the best and average fitness functions of the optimization model for different iterations. Figure 13 shows the GT values for different combination of inputs.

Table 9 | Model selection by Gamma statistics and backward selection

			Training data set		Testing data set	
Number of variables	Mask	Combination of variables	RMSE (mm)	MAE (mm)	RMSE (mm)	MAE (mm)
10	1111111111	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	23.161	12.142	45.712	45.712
9	1110111111	$x_1, x_2, x_3, x_5, x_6, x_7, x_8, x_9, x_{10}$	26.807	14.341	45.857	45.857
8	1110110111	$x_1, x_2, x_3, x_5, x_6, x_8, x_9, x_{10}$	28.029	15.261	45.897	45.897
7	1010110111	$x_1, x_3, x_5, x_6, x_8, x_9, x_{10}$	30.654	17.517	46.622	46.622
6	0010110111	$x_3, x_5, x_6, x_8, x_9, x_{10}$	38.265	22.988	47.372	47.372
5	0000110111	$x_5, x_6, x_8, x_9, x_{10}$	42.152	26.98	47.263	47.263
4	0000110110	x_5, x_6, x_8, x_9	43.041	31.827	46.352	46.352
3	0000100110	x_5, x_8, x_9	44.162	36.073	48.893	48.893
2	00001000100	x_5, x_8	44.71	40.042	52.272	52.272
1	00000000100	x_8	47.241	43.375	55.548	55.548

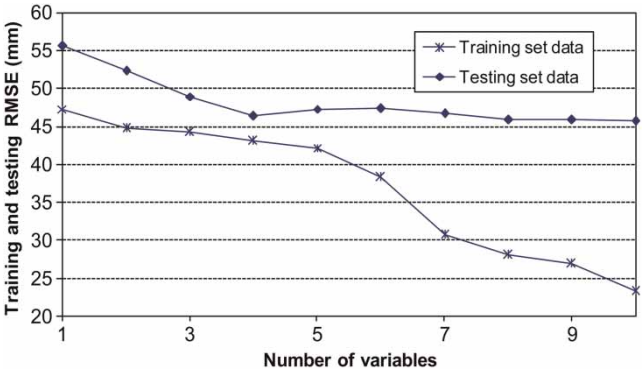

Figure 9 | Model complexity Gamma statistics and backward selection.

Table 10 | Combination of variables with Gamma statistics and forward selection[a]

Number of variables	Mask	Gamma	Gradient	Standard error	V_{ratio}
1	1000000000	0.9624	−0.0932	0.0299	0.9624
	0100000000	0.9952	−0.2475	0.0251	0.9952
	0010000000	0.8766	0.3411	0.0234	0.8766
	0001000000	0.9864	0.0115	0.0054	0.9864
	0000100000	0.9332	0.3470	0.0215	0.9332
	0000010000	0.8701	−0.1709	0.0253	0.8701
	0000001000	0.9734	−0.0744	0.0107	0.9734
	0000000100	0.8744	1.0152	0.0299	0.8744
	0000000010	0.9471	0.0291	0.0292	0.9471
	0000000001	0.8726	1.1290	0.0231	0.8726
2	1000010000	0.8808	−0.3944	0.0507	0.8808
	0100010000	0.8687	0.1389	0.0255	0.8687
	0010010000	0.9216	−0.7411	0.0425	0.9216
	0001010000	0.8728	−0.1058	0.0272	0.8728
	0000110000	0.8188	0.0154	0.0479	0.8188
	0000011000	0.8507	0.0317	0.0370	0.8507
	0000010100	0.8982	−0.5492	0.0598	0.8982
	0000010010	0.8488	0.0296	0.0437	0.8488
	0000010001	0.8082	0.7456	0.0432	0.8082
3	1000010001	0.8026	0.1559	0.0374	0.8026
	0100010001	0.8549	0.0022	0.0383	0.8549
	0010010001	0.7704	0.3532	0.0192	0.7704
	0001010001	0.8881	−0.2273	0.0426	0.8881
	0000110001	0.7954	0.0390	0.0264	0.7954
	0000011001	0.8218	0.0610	0.0544	0.8218
	0000010101	0.8368	0.2453	0.0304	0.8368
	0000010011	0.9028	−0.2101	0.0582	0.9028
4	1010010001	0.8278	0.0298	0.0487	0.8278
	0110010001	0.8540	−0.0320	0.0385	0.8540
	0011010001	0.7866	0.1062	0.0618	0.7866
	0010110001	0.7568	0.1212	0.0400	0.7568
	0010011001	0.8606	−0.0167	0.0252	0.8606
	0010010101	0.7975	0.1799	0.0294	0.7975
	0010010011	0.8719	−0.0088	0.0241	0.8719
5	1010110001	0.8323	−0.0198	0.0446	0.8323
	0110110001	0.8059	0.0266	0.0193	0.8059
	0011110001	0.7726	0.0804	0.0500	0.7726
	0010111001	0.7702	0.0688	0.0351	0.7702
	0010110101	0.8353	−0.0069	0.0315	0.8353
	0010110011	0.8189	0.0012	0.0308	0.8189
6	1010111001	0.8253	−0.0064	0.0399	0.8253
	0110111001	0.8699	−0.0499	0.0322	0.8699
	0011111001	0.8238	−0.0116	0.0323	0.8238
	0010111101	0.8229	−0.0061	0.0189	0.8229
	0010111011	0.8496	−0.0366	0.0343	0.8496
7	1010111101	0.7906	0.0232	0.0426	0.7906
	0110111101	0.8554	−0.0322	0.0282	0.8554
	0011111101	0.8524	−0.0355	0.0530	0.8524
	0010111111	0.8792	−0.0561	0.0369	0.8792
8	1110111101	0.8289	−0.0106	0.0415	0.8289
	1011111101	0.8501	−0.0167	0.0373	0.8501
	1010111111	0.8366	−0.0155	0.0391	0.8366
9	1111111101	0.8379	−0.0114	0.0273	0.8379
	1110111111	0.8237	−0.0071	0.0420	0.8237

(*continued*)

Table 10 | continued

[a]The highlighted masks indicate the best mask in each category.

The interval of 2–11 of the *x*-axis in this figure is referred to as the models with nine input variables. Intervals 12–56, 57–167, 168–386, and 387–603 are referred to the models with eight, seven, six, and five input variables, respectively. The results show that the mask '1011011010' with six input variables has the least GT value at 0.7149.

Table 12 presents the evaluation on the models selected by the GT based on the GA optimization model. The errors are evaluated using the SVM simulation model with the LOOCV. The result indicates that the model with eight variables has the lowest RMSE and MAE value on the testing data and therefore this model is selected. The variation of the simulated and recorded precipitation over Wales is presented in Figure 14.

Model comparison

This study explored the variable selections by correlation coefficient, multicollinearity analysis, the GT with backward selection, the GT with forward selection, and the GT with GA search methods. Performances for those

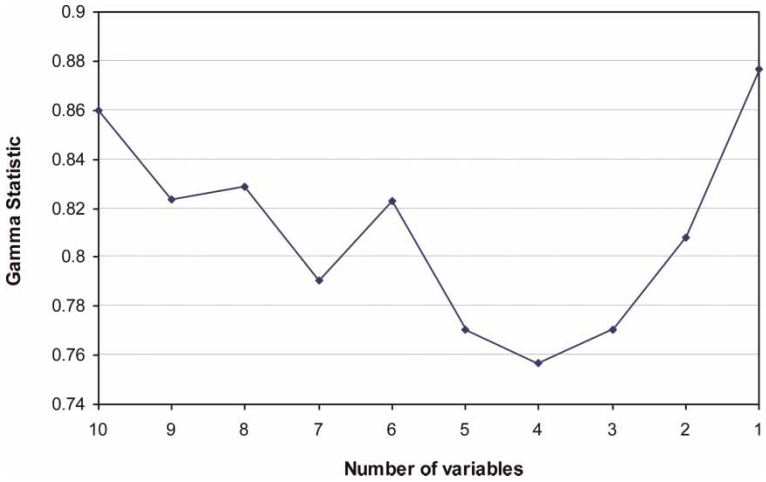

Figure 10 | GT values and number of input variables using forward selection.

Table 11 | Model selection by Gamma statistics and forward selection

			Training data set		Testing data set	
Number of variables	**Mask**	**Combination of variables**	**RMSE (mm)**	**MAE (mm)**	**RMSE (mm)**	**MAE (mm)**
1	0000000100	x_8	45.664	43.375	55.548	55.548
2	0010000100	x_3, x_8	44.19	41.095	53.84	53.84
3	1010000100	x_1, x_3, x_8	44.162	37.58	50.777	50.777
4	1010100100	x_1, x_3, x_5, x_8	43.5403	31.299	45.46	45.46
5	1110100100	x_1, x_2, x_3, x_5, x_8	38.594	23.989	46.314	46.314
6	1110100110	$x_1, x_2, x_3, x_5, x_8, x_9$	34.912	20.558	44.655	44.655
7	1111100110	$x_1, x_2, x_3, x_4, x_5, x_8, x_9$	30.883	17.074	44.23	44.23
8	1111100111	$x_1, x_2, x_3, x_4, x_5, x_8, x_9, x_{10}$	28.175	15.151	45.407	45.407
9	1111110111	$x_1, x_2, x_3, x_4, x_5, x_6, x_8, x_9, x_{10}$	24.411	12.933	45.965	45.965
10	1111111111	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	23.161	12.142	45.712	45.712

models in the training and testing data are shown in Table 13 indicated by the RMSE and MAE. The test results reveal that the input selection with the GT and forward selection method has the lowest testing error followed by the GT with GA search method for input variable selection.

In order to evaluate the SVM model as a precipitation prediction model, the SVM model's performance is compared with the multiple linear regression model as a benchmark model. In both models, the input variables are selected using the GT and forward selection method. The errors are evaluated using the SVM and multiple linear regression simulation models with the LOOCV. The result presented in Table 14 indicates that the SVM model has the lowest RMSE and MAE value on the training and testing data. Therefore the SVM model is selected for precipitation prediction with the 5 month lead time.

SUMMARY AND CONCLUSION

The selection of appropriate predictor, or characteristics from the large scale atmospheric circulation, is one of the

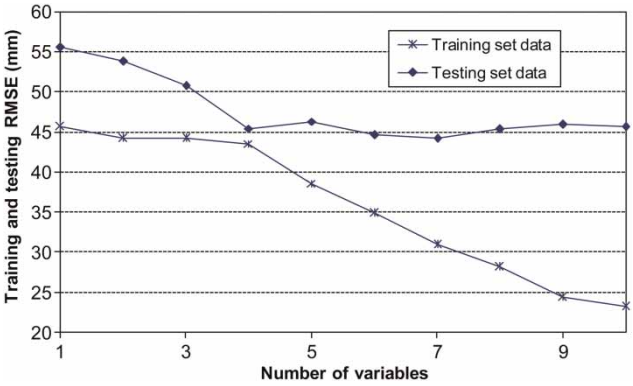

Figure 11 | Model complexity with Gamma statistics and forward selection.

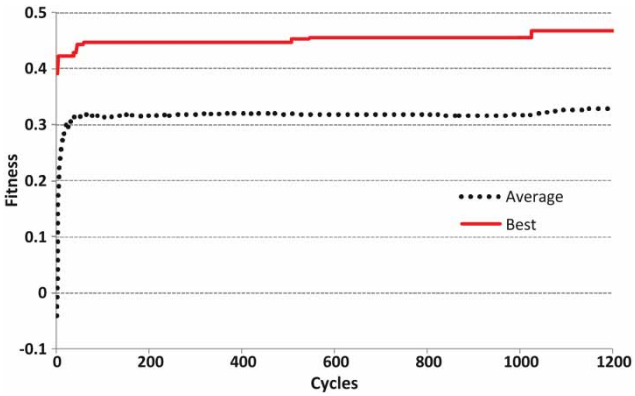

Figure 12 | The best and average fitness functions of GA model.

most important steps in a downscaling exercise. In this paper, the details of selecting the number of enough clusters, clustering the data and choosing inputs for statistical downscaling are described. The SVM model has been used for downscaling and predicting precipitation at monthly time scale from the GCM.

This study explores the application of regression, multicollinearity, the GT methods with both backward and forward selections, and GA optimization model. The LOOCV serves two purposes: selecting model input variables and assessing model performances. LOOCV is suitable as a tool for input variable selection when the data set is small because it makes full use of the data available. The results show that the multicollinearity approach selects a better model than the correlation analysis while the number of input variables is the same. The results show the combination of the forward method with correlation analysis and GT yields a better model than the backward method. Maybe it is because the forward selection picks up the promising variables in initial steps of the input selection procedure. In this study, the local minimum for the backward selection and GT is six. A further investigation reveals that a 10 variable model is the optimal. For the backward selection, the local minimum is found to be four but a further evaluation shows that a model with seven variables is the optimal. However, since both procedures are not an

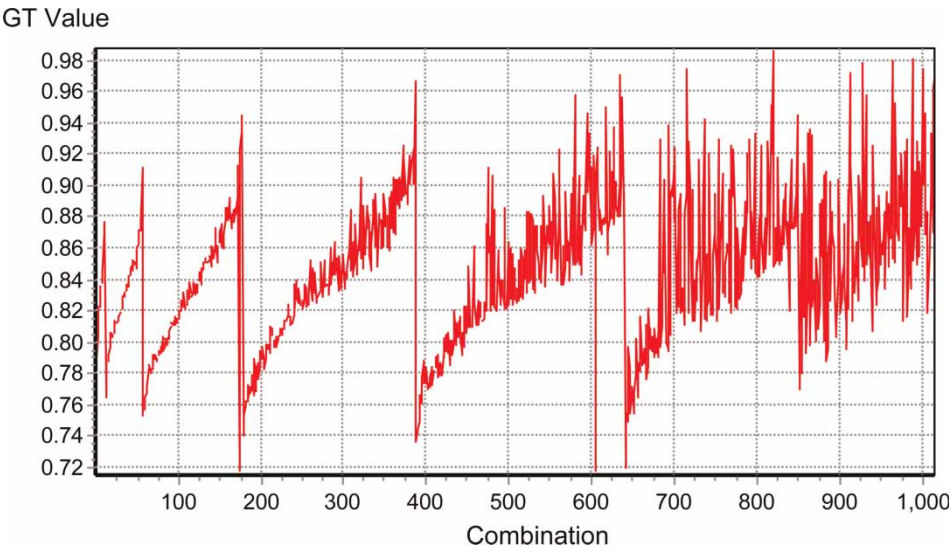

Figure 13 | The GT values for different combination of inputs.

Table 12 | Model selection by GT and genetic algorithm

			Training data set		Testing data set	
Number of variables	**Mask**	**Combination of variables**	**RMSE (mm)**	**MAE (mm)**	**RMSE (mm)**	**MAE (mm)**
10	1111111111	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	23.161	12.142	45.712	45.712
9	1101111111	$x_1, x_2, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	26.047	13.879	46.398	46.398
8	1011111110	$x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9$	28.684	15.593	45.436	45.436
7	0111011011	$x_2, x_3, x_4, x_6, x_7, x_9, x_{10}$	35.225	20.291	46.761	46.761
6	1011011010	$x_1, x_3, x_4, x_6, x_7, x_9$	37.668	22.77	46.042	46.042
5	0111010100	x_2, x_3, x_4, x_6, x_8	40.554	25.246	45.808	45.808
4	0001110010	x_4, x_5, x_6, x_9	46.678	32.737	47.198	47.198
3	0010010001	x_3, x_6, x_{10}	44.911	37.804	51.233	51.233
2	0010000001	x_3, x_{10}	44.028	40.738	53.248	53.248
1	0000010000	x_6	43.635	42.853	55.453	55.453

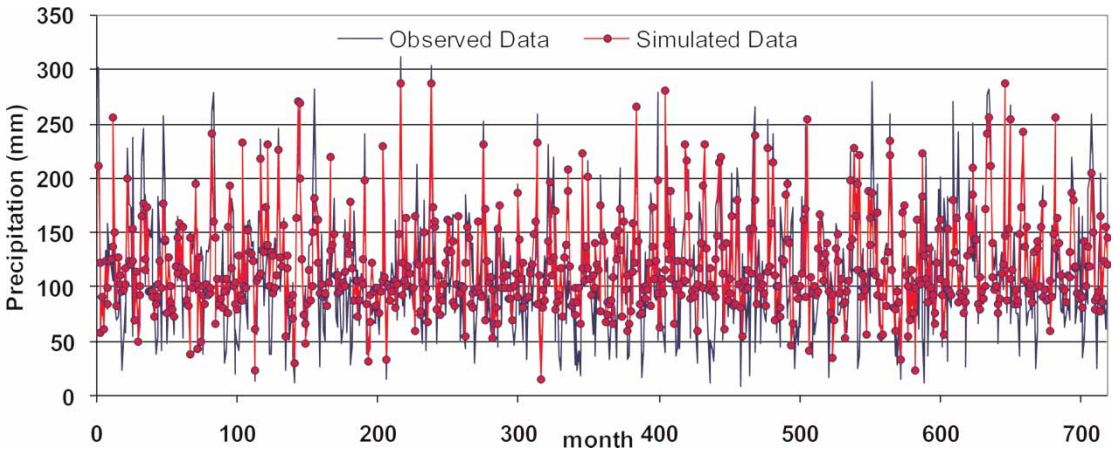

Figure 14 | The observed and simulated precipitation over Wales.

Table 13 | Comparison of input selection methods

				Training data set		Testing data set	
Method	**Number of variables**	**Mask**	**Combination of variables**	**RMSE (mm)**	**MAE (mm)**	**RMSE (mm)**	**MAE (mm)**
Regression	10	1111111111	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	23.161	12.142	45.712	45.712
Multicollinearity – Backward	10	1111111111	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	23.161	12.142	45.712	45.712
Multicollinearity – Forward	10	1111111111	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	23.161	12.142	45.712	45.712
GT – Backward	10	1111111111	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$	23.161	12.142	45.712	45.712
GT – Forward	7	1111100110	$x_1, x_2, x_3, x_4, x_5, x_8, x_9$	30.883	17.074	44.23	44.23
GT with GA	8	1011111110	$x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9$	28.684	15.593	45.436	45.436

Table 14 | Comparison of precipitation simulation methods

Model	Number of variables	Mask	Combination of variables	Training data set		Testing data set	
				RMSE (mm)	MAE (mm)	RMSE (mm)	MAE (mm)
SVM	7	1111100110	$x_1, x_2, x_3, x_4, x_5, x_8, x_9$	30.883	17.074	44.23	44.23
Multiple Linear Regression	8	1011111110	$x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_9$	31.106	21.156	46.013	46.013

exhaustive search, the best result from them may not be the overall 'best' result and may miss the 'true' best. The GA search shows the local minimum is six and the further evaluation shows that a model with eight variables is the optimal. However, the overall best result is the model selected with the GT and forward selection. The input selection using the GTs with the forward and backward methods and GA search is more reliable than the models using the regression and multicollinearity analysis.

This methodology is simple to apply and adaptable to the precipitation prediction with other climate variables such as sea surface temperature (SST) or the antecedent precipitation in the case study. This study encourages further research in the applications of new tools of soft computing techniques with different climate variables. Clearly, more explorations into this technique are needed in order to gain valuable experience in it.

REFERENCES

Agalbjörn, S., Končar, N. & Jones, A. J. 1997 A note on the Gamma test. *Neural Comput. Appl.* **5** (3), 131–133.

Ahmadi, A., Han, D., Karamouz, M. & Remesan, R. 2009 Input data selection for solar radiation estimation. *Hydrol. Process.* **23** (19), 2754–2764.

Allen, D. M. 1974 The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16** (1), 125–127.

Barrow, E. M. & Semenov, M. A. 1995 Climate change scenarios with high spatial and temporal resolution for agricultural applications. *Forestry* **68** (4), 349–360.

Bezdek, J. C. 1981 *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Bray, M. & Han, D. 2004 Identification of support vector machines for runoff modelling. *J. Hydroinformat.* **6** (4), 265–280.

Camdevyren, H., Demyr, N., Kanik, A. & Keskyn, S. 2005 Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-*a* in reservoirs. *Ecol. Model.* **181**, 581–589.

Choy, K. Y. & Chan, C. W. 2003 Modelling of river discharges and rainfall using radial basis function networks based on support vector regression. *Int. J. Syst. Sci.* **34** (14–15), 763–773.

Colman, A. 1997 Prediction of summer central England temperature from preceding North Atlantic winter sea surface temperature. *Int. J. Climatol.* **17** (12), 1285–1300.

Colman, A. & Davey, M. 1999 Prediction of summer temperature, rainfall and pressure in Europe from preceding winter North Atlantic Ocean temperature. *Int. J. Climatol.* **19**, 513–536.

Crane, R. G. & Hewitson, B. C. 1998 Doubled CO_2 precipitation changes for the Susquehanna Basin: down-scaling from the genesis general circulation model. *Int. J. Climatol.* **18** (1), 65–76.

Geisser, S. 1975 The predictive sample reuse method with applications. *J. Am. Stat. Assoc.* **70** (350), 320–328.

Gen, M.R. & Cheng, L. 2000 *Genetic Algorithm and Engineering Optimization*, John Wiley & Sons, Inc., New York, USA.

Ghosh, S. & Mujumdar, P. P. 2008 Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Adv. Water Res.* **31**, 132–146.

Hashmi, M. Z., Shamseldin, A. Y. & Melville, B. W. 2011 Statistical downscaling of watershed precipitation using Gene Expression Programming (GEP). *Environ. Modell. Softw.* **26**, 1639–1646.

Hawkins, D. M., Basak, S. C. & Mills, D. 2003 Assessing model fit by cross validation. *J. Chem. Inf. Comput. Sci.* **43** (2), 579–586.

Haylock, M. R., Cawley, G. C., Harpham, C., Wilby, R. L. & Goodess, C. 2006 Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. *Int. J. Climatol.* **26**, 1397–1415.

Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J. M. & Fernandez, L. 2000 Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis. *Water Res.* **34**, 807–816.

Karamouz, M., Ahmadi, A. & Moridi, A. 2009 Probabilistic reservoir operation using Bayesian stochastic model and support vector machine. *Adv. Water Res.* **32**, 1588–1600.

Končar, N. 1997 Optimisation Methodologies for Direct Inverse Neurocontrol. Ph.D. Thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London, UK.

Lu, W. Z., Wang, W. J., Wang, X. K., Xu, Z. B. & Leung, A. Y. T. 2003 Using improved neural network to analyze RSP, NOx and NO2 levels in urban air in Mong Kok, Hong Kong. *Environ. Monit. Assess.* **87**, 235–254.

Meteorological Office, United Kingdom (http://www.metoffice.gov.uk/).

Michalewicz, Z. 1992 *Genetic Algorithms Data Structures Evolutionary Programs*. Springer, New York.

Moghaddamnia, A., Remesan, R., Kashani, M. H., Mohammadi, M., Han, D. & Piri, J. 2009a Comparison of LLR, MLP, Elman, NNARX and ANFIS models – with a case study in solar radiation estimation. *J. Atm. Solar-Terrest. Phys.* **71** (8–9), 975–982.

Moghaddamnia, A., Gousheh, M. G., Piri, J., Amin, S. & Han, D. 2009b Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. *Adv. Water Res.* **32** (1), 88–97.

Murphy, J. M. 1999 An evaluation of statistical and dynamical techniques for downscaling local climate. *J. Clim.* **12**, 2256–2284.

Najafi, M. R., Moradkhani, H. & Wherry, S. A. 2011 Statistical downscaling of precipitation using machine learning with optimal predictor selection. *J. Hydrol. Eng.* **16** (8), 650–664.

Pearson, K. 1901 On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **2** (6), 559–572.

Piri, J., Amin, S., Moghaddamnia, A., Keshavarz, A., Han, D. & Remesan, R. 2009 Daily pan evaporation modelling in a hot and dry climate. *J. Hydrol. Eng.* **14** (8), 803–811.

Remesan, R., Shamim, M. A. & Han, D. 2008 Model data selection using gamma test for daily solar radiation estimation. *Hydrol. Process.* **22** (21), 4301–4309.

Schmidli, J., Goodess, C. M., Frei, C., Haylock, M. R., Hundecha, Y., Ribalaygua, J. & Schmith, T. 2007 Statistical and dynamical downscaling of precipitation: an evaluation and comparison of scenarios for the European Alps. *J. Geophys. Res.* **112**, D04105.

Sivapragasam, C. & Liong, S. Y. 2005 Flow categorization model for improving forecasting. *Nord. Hydrol.* **36** (1), 37–48.

Stefánsson, A., Koncar, N. & Jones, A. J. 1997 A note on the Gamma test. *Neural Comput. Appl.* **5** (3), 131–133.

Tripathi, S., Srinivas, V. V. & Nanjundiah, R. S. 2006 Downscaling of precipitation for climate change scenarios: a support vector machine approach. *J. Hydrol.* **330** (3–4), 621–640.

Vapnik, V. N. 1995 *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Vapnik, V. N. 1998 *Statistical Learning Theory*. Wiley, New York.

von Storch, H., Zorita, E. & Cubasch, U. 1993 Downscaling of global climate change estimates to regional scale: an application to Iberian rainfall in wintertime. *J. Clim.* **6**, 1161–1171.

Wan Jaafar, W. Z., Liu, J. & Han, D. 2011 Input variable selection for median flood regionalisation. *Water Resources Research* **47**, W07503.

Wilby, R. L. 1998 Statistical downscaling of daily precipitation using daily airflow and seasonal teleconnection indices. *Clim. Res.* **10**, 163–178.

Wilby, R. L. 2001 Downscaling summer rainfall in the UK from North Atlantic ocean temperatures. *Hydrol. Earth Syst. Sci.* **5** (2), 245–257.

Wilby, R. L. 2005 Uncertainty in water resource model parameters used for climate change impact assessment. *Hydrol. Process.* **19** (16), 3201–3219.

Wilby, R. L., Wedgbrow, C. S. & Fox, H. R. 2004 Seasonal predictability of the summer hydrometeorology of the River Thames, UK. *J. Hydrol.* **295**, 1–16.

Yu, X., Liong, S. Y. & Babovic, V. 2004 EC-SVM approach for real-time hydrologic forecasting. *J Hydroinformat.* **6**, 209–223.

First received 19 June 2012; accepted in revised form 26 October 2012. Available online 13 December 2012