ISEIS

Journal of Environmental Informatics 31(2) 137-152 (2018)

Journal of Environmental Informatics

www.iseis.org/jei

Application of PCA and Clustering Methods in Input Selection of Hybrid Runoff Models

R. Remesan[1,*], M. Bray[2], and J. Mathew[3]

[1]*School of Water Resources, Indian Institute of Technology Kharagpur 721302, India*
[2]*Cardiff School of Engineering, Cardiff University, Cardiff CF24 3AA, UK*
[3]*Department Computer Science and Engineering, Indian Institute of Technology Patna 801103, India*

Received 15 October 2013; revised 7 November 2014; accepted 15 March 2015; published online 15 May 2017

ABSTRACT. This study has proposed and investigated a novel input variable selection method for nonlinear modelling based on principle component analysis (PCA) and cluster analysis. The proposed approach was applied to daily rainfall-runoff modelling of the Brue catchment of the United Kingdom using wavelet based hybrid forms of two nonlinear models, Artificial Neural Networks (ANNs) and Local Linear Regression (LLR), to identify meaningful wavelet decomposed sub-series. The homogenous group formation capability of cluster analysis and redundancy assessment capability of PCA were applied effectively in this study to solve input selection uncertainties associated with wavelet based hybrid models. Though this concept has been represented in the selection of effective wavelet decomposed subseries in runoff modelling, the application has gotten wider implications in time series modelling with highly redundant and large input space. The study revealed the weakness of conventional forms of cross-correlation analysis and also suggested that input selection could be improved by making sufficient natural clusters (equal to the desired number of input data series) of input space and restricting the search within each cluster according to silhouette or correlation value. The study also highlighted the higher modelling capability of ANN over traditional LLR models in rainfall-runoff modelling of the Brue catchment.

Keywords: input identification, wavelet subseries, ANN, LLR

1. Introduction

Artificial Intelligence (AI) based techniques are part of many popular data-driven models that have been used extensively in the past couple of decades in stream flow forecasting and rainfall-runoff modelling. A comprehensive review by the ASCE Task Committee on Application of ANN in Hydrology (2000) shows the acceptance of ANN techniques among hydrologists. The major criticism against AI techniques in hydrology is their limited ability to account for any physics of the hydrological processes in a catchment. That concern was partially ruled out by Jain et al. (2004a) who proved that the distributed structure of the ANN is able to capture certain physical properties. The physics involved in the ANNs ability to express physical processes in a watershed through proper training have been investigated by several recent studies (e.g., Sudheer et al., 2002; Jain et al., 2004; Sudheer and Jain, 2004). Wavelet analysis is a well-defined concept with increasing applications to the quantitative explanation of time-series and is a useful tool for analysing both rainfall and runoff time-series (Lane, 2007). By combining these AI techniques with one another or with other dynamic predicttive models, the individual strengths of each approach can be exploited in a synergistic manner for the construction of powerful intelligent systems (Nayak et al., 2004; Sudheer, 2005; Nourani et al., 2009; Krishna et al., 2012; Krishna, 2014; Budu, 2014).

The added advantage of wavelet transforms (in comparison to the classical Fourier analysis) in the analysis of time series signals to detect detailed temporal patterns has attracted many researchers to apply this technique in various aspects of hydrology including rainfall runoff modelling (Kisi, 2008; Remesan et al., 2009; Nourani et al., 2011), precipitation forecasting, water level forecasting (Kisi, 2009) and sediment estimation (Partal and Cigizoglu, 2009). Recently, Sang et al. (2012) performed trend in the hydrological time series using wavelets. See Maheswaran and Khosa (2012) for a review of application of wavelets in hydrology. Nourani et al. (2014) has provided an extensive review of wavelet based hybrid wavelet modelling existing in hydrology. Despite the application of many wavelet based hybrid models in conjunction with AI techniques, the determination of effective wavelet components still remains a dilemma. The cross-correlation method is the conventional approach which is used to identify effective and useful 'detail sub-series' for modelling to overcome issues such as redundancy and overtraining.

In the data-based artificial intelligent model development, the selection of an appropriate subset of variables from the available set of potential input variable space plays a vital role. Real-world modelling of hydrological processes ideally requi-

* Corresponding author. Tel.: +91 3222 281888; fax: +91 3222 282212.
E-mail address: renji.remesan@swr.iitkgp.ernet.in (R. Remesan).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.201700378

res a complex input structure and very lengthy training data to represent inherently complex dynamic systems. In hybrid environmental modelling, data sets of redundant variables can be discarded without the loss of much information. Modelling with a large number of the available inputs can lead technical issues such as computational complexity and lack of memory. The chances of such issues arising are elevated in rainfall runoff modelling using antecedent information as such models possess high nonlinearity and may have a large number of parameters. Therefore, there is a need to identify techniques which adequately reduce the number of inputs in nonlinear models. Despite successful applications of AI techniques in hydrology, there are still many unsolved issues, particularly in the selection of training data length and data structure (Bowden et al., 2005). Maier and Dandy (2000) reviewed more than 43 journal papers in hydrology and pointed out that, in most cases, the inputs were chosen arbitrarily without scientific reasoning and some studies used trial and error approach or validation data. The Gamma Test (GT) by Stefánsson et al. (1 997) can successfully overcome these issues and has been demonstrated in water level and flow modelling of the River Thames (Durrant, 2001), daily solar radiation prediction (Remesan et al., 2008) and monthly stream flow prediction (Noori et al., 2011). GT helps to identify the best embedded structure and data length for training any smooth model prior to modelling. Research by Ahmadi et al. (2009) explored data selection capabilities of different approaches including the Gamma test, entropy theory, AIC (Akaike's information criterion) and BIC (Bayesian information criterion) in the context of solar radiation modelling and highlighted the merits and demerits associated with those approaches.

Studies (e.g., Ssegane et al., 2012; Ganti and Jain, 2011) have also shown the capabilities of principal components analysis (PCA) to discard redundant data variables and data. Procrustes analysis and a measure of similarity are also used to identify the redundancy in the data sets used in modelling. King and Jackson (1999) used both procrustes analysis and principal components to identify redundant variables from a dataset consisting of 36 meteorological variables spanning 37 years. Their approach used a measure of similarity and bivariate plots to assess the success of the alternative variable selection methods. Back and Trappenberg (1999) proposed an algorithm for model free input variable selection, called independent component analysis, which allows a straightforward statistical test to identify the redundant data from available input pool. Cluster Analysis is another excellent statistical tool for multivariate segmentation and could be used as a useful tool to drive the model development process through effective use of variable section. Zhi-hang (2009) identified and discussed its competency to act as an aid in modelling in the context of its application in analytical modelling in service Industries. Noori et al. (2011) have applied PCA, Gamma test (GT), and forward selection (FS) techniques to reduce the number of input variables in the monthly stream flow prediction utilizing ANNs and Support Vector Machines (SVMs). Levi and Rasmussen (2014) have applied an iterative principal component analysis for predicting physical soil properties in in a semiarid ecosystem. Caraway et al. (2014) have applied cluster analysis and k-nearest neighbour time series resampling for multisite, stochastic weather generator for hydrologic simulation

Advent of conjunctive application of wavelets in data based models created a general trend among researchers to decompose each and every original data series to an acceptable number of sub-series and then to use all these subseries as model inputs of databased methodical models. Inclusion of large number of these extra input variables (sub series) is growing as one of the main concerns in wavelet based hybrid modelling process. Situation further complicates in typical case studies like rainfall runoff modelling such hydrological system considers several meteorological variable time series at various antecedent time steps as original input series. There is high possibility that wavelet procedure complicates the model structure further without adding to the predictive accuracy if we consider all detail subseries of all antecedent information indiscriminately. Cross-correlation approach is the most commonly adopted approach to identify effective sub-series but often it is meaningless to check correlation of 'high frequency and low scale' information in the actual series with predictand data. Considering these facts, the aim of this paper is to introduce an alternative approach in input data series selection combining use of PCA and cluster analysis that could be adopted in hydrological time series modelling. We present this approach along with two alternative data-driven hybrid models for rainfall-runoff modelling based on wavelet transform methodology in conjunction with ANNs and LLR. This article explores the capabilities of cluster analysis and PCA to identify the effective decomposed subseries when used in conjunctt-ion. The Silhouette values based approach and traditional cross-correlation analysis was used to check the authenticity of this procedure. The performance of the proposed hybrid wavelet models were compared with that of traditional Local Linear Regression (LLR) and ANNs to assess the improvements in prediction.

2. Materials and Methods

2.1. Data Sets and Study Area

The River Brue catchment, located in Somerset, south west England, UK, was selected for the analysis. The Brue catchment is the one of the best representative catchments to express hydrological responses in the England due to its long series of high data quality data. This catchment has been extensively used for research weather radar, quantitative precipitation and flood forecasting and rainfall-runoff modelling (Bell and Moore, 2000). The catchment has a well maintained dense rain gauge network and is covered by three weather radars. The River Brue catchment was the site of the Natural Environment Research Council (NERC) funded HYREX (Hydrological Radar Experiment) project from 1993 to 1997. The catchment has a drainage area of 135 square kilometres and an elevation range between 35 metres to 190 metres above mean sea level. The catchment is located at 51.075º North and 2.58º West; the

location map of the catchment is shown in Figure 1. The river gauging point at the catchment is located at Lovington. An automatic weather station (AWS) and automatic soil water station (ASWS) recorded global solar radiation, net radiation and other weather parameters including wind speed, wet and dry bulb temperatures and atmospheric pressure at hourly intervals. Six years of daily rainfall-runoff data from the Brue catchment, spanning from 1993 to 2000, was used in this study. Three step antecedent runoff values ($Q(t\text{-}1)$, $Q(t\text{-}2)$, $Q(t\text{-}3)$), one step antecedent rainfall ($P(t\text{-}1)$) and the current rainfall information ($P(t)$) were used for hybrid modelling as recommended by Remesan et al. (2009) for this study area. Remesan et al. (2009) identified that 1056 data points are sufficient to make a reliable rainfall-runoff data model for the current study area on this particular data set. We have used 1056 data points for training and remaining data out of total 2240 data points were used as the validation data set.

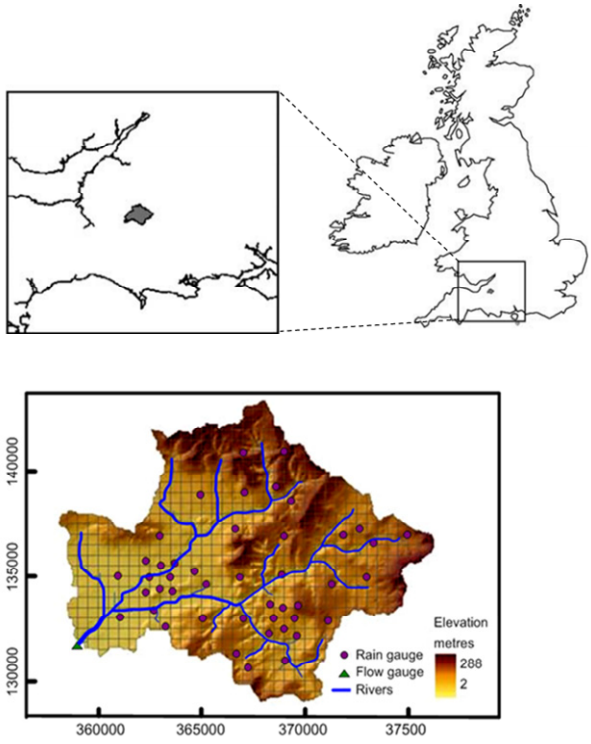


Figure 1. The River Brue Catchment, southwest England, United Kingdom (modified from FLOOD site project).

2.2. Approaches and Models

The aim of this study is to predict the 1-day-ahead runoff in Brue catchment employing essential sub-series components obtained using discrete wavelet transforms (DWT) on the original data. Selection of appropriate and effective subseries from the decomposed details (D) and approximations (A) is a challenging issue in DWT modelling in hydrology and related fields. The existing methodology is based on correlation coefficients to choose effective detail (D) subseries and use it along with the final approximation subseries. There is not much evidence in the literature of studies which test whether the adopted resolution level is suitable for modelling or of whether it adds more redundancy through extra decomposition. However, in this study we have adopted the equation INT (log n) as a thumb rule for quick estimation of resolution. In the above equation INT stands of integer part, n stands for data length of the series, and log is common logarithm (Wang and Ding, 2003). The study used a combined application of cluster analysis and PCA for identification of effective wavelet decomposed subseries for rainfall-runoff modelling. The overall procedure adopted in this study is shown in the Figure 2.

The methodology adopted in this study can be summarized as follows:

1. Using DWT, the three step antecedent runoff values ($Q(t\text{-}1)$, $Q(t\text{-}2)$, $Q(t\text{-}3)$), one step antecedent rainfall ($P(t\text{-}1)$) and current rainfall information ($P(t)$) were decomposed [3 detail series and 3 approximation series from each antecedent data series (a total of 30 wavelet decomposed subseries)].

2. Then we have used the capabilities of principal component analysis (PCA) in redundancy analysis on the 30 decomposed sub series to identify the minimum number of subseries required for building a model of the desired accuracy.

3. Cluster analysis was used to identify required clusters (2-12 constructed) and effective subseries in each cluster were identified using silhouette values and cross-correlation values.

4. These sub-series selected by both methods were modelled separately (for all 2-12 clusters) using neuro-wavelet (NW) and wavelet local linear regression (W-LLR) hybrid models.

5. Their performance was compared using statistics obtained in both training and validation, and the authenticity of combined use of PCA and cluster analysis checked. Whether the number of subseries suggested by the PCA could produce reasonable modelling results in comparison to other input combinations was evaluated.

6. The results obtained using the subseries suggested by [PCA + cluster analysis] were compared with traditional ANN and LLR to assess improvements in the modelled results.

7. The hybrid modelling results obtained using the subseries recommended by the [PCA + cluster analysis] were also compared with neuro-wavelet (NW), and wavelet local linear regression (W-LLR) considering reconstructed series constituted by adding effective sub series (Approximation and Detail series). Pre-processing of raw data is performed in this modelling phase before choosing them as input space for NW and W-LLR model. The reconstructed series was constituted without the ineffective components. Correlation coefficients between decomposed sub series and the 'time series to be modelled' were used as an indicator to filter out ineffective components.

Items 1 to 5 shown above are denoted as Method 3 in the methodology and Figure 2, whilst items 6 and 7 correspond to Method 1 and Method 2, respectively, in Figure 2.

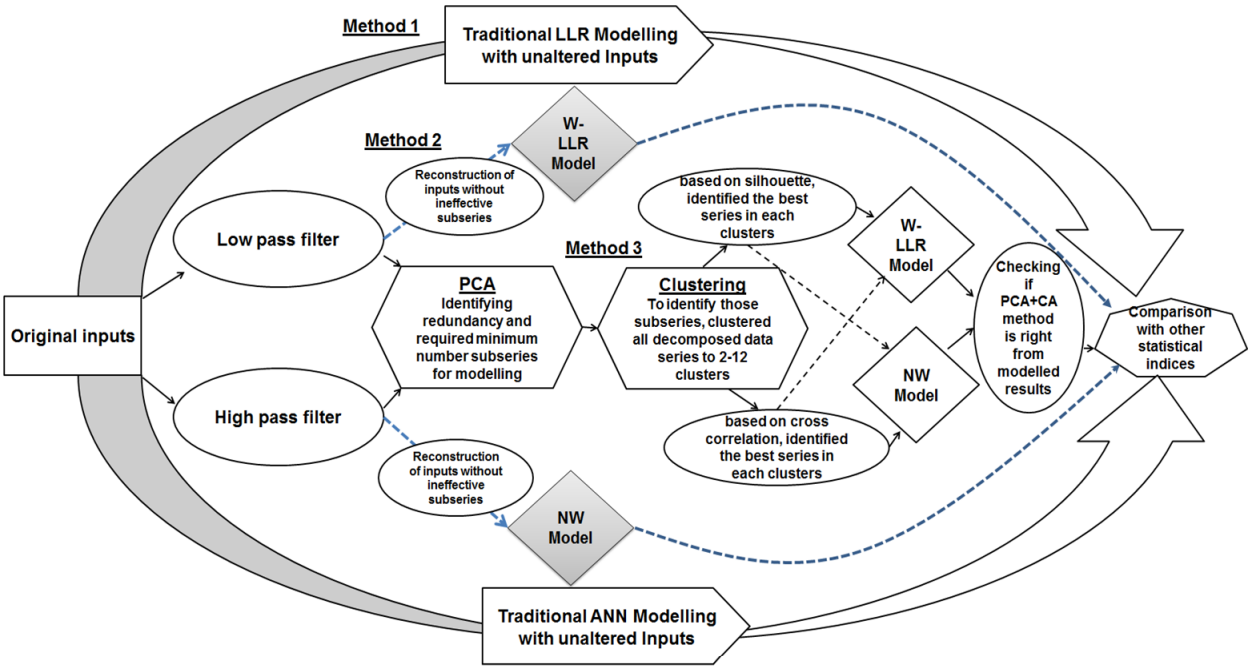


Figure 2. The methodology adopted in this study.

2.2.1. Implementation of Principal Component Analysis

The basic background of PCA can be explained as follows: assume an event for which p variables (attributes) X_i are being measured sequentially through time for n instances. The corresponding dataset $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}]$ consists of p vectors $\mathbf{x_i}$, where $\mathbf{x_i} = [\mathbf{x_{1i}}, \mathbf{x_{2i}}, \ldots, \mathbf{x_{ni}}]$ is a column vector which contains the n measurements of the variable X_i (where $\mathbf{x_i}$ includes a univariate time series). Each row of $\mathbf{X}$ corresponds to the measurements of all variables at a specific time instance. Therefore, each row of $\mathbf{X}$ can be considered as a point in p-dimensional space. PCA derives a new set of orthogonal and uncorrelated composite variates $Y_{(j)}$, which are called principal components:

$$Y_{(j)} = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p,\ j = 1,\ 2,\ \ldots,\ p \qquad (1)$$

where a_j are a related eigenvectors and X_i are input variables. The PCA information can be achieved by solving Equation 2:

$$|\boldsymbol{R} - \boldsymbol{I}\lambda| = 0 \qquad (2)$$

In the above equation, $\mathbf{R}$ is the variance-covariance matrix and λ are the eigenvalues. The $\mathbf{I}$ matrix is a unit matrix (Davis, 1986; Manly, 1986).

We have used MATLAB Statistics Toolbox for PCA application and plotted fraction of cumulative Variance associated with each principal component. This study uses the variable reduction and redundancy assignment capability of principal component analysis in nonlinear modelling. PCA is useful to assess the redundancy due to the possible correlation of one input to another in a modelling data set with large number of input data series. PCA could be used to reduce the data input series into a smaller number of principal components (artificial variables) that will account for most of the variance in the actual data. The first principal component is a combination of original data series used in the study which explains the greatest amount of existing variation. The second principal component defines the next largest amount of variation and is independent of the first principal component. These sets of uncorrelated variables (principal components) can be ordered by reducing variability and the last few items of these variables can be removed with minimum loss of real data. In this study, PCA was carried out for both the correlation matrix and the covariance matrix to see the possible differences in both approaches.

2.2.2. Implementation of Cluster Analysis

Cluster analysis is an investigative data analysis tool widely used for solving classification problems in different scientific fields. We used the k-means algorithm with different distances and similarities measures for the comparative study. There are many commonly used algorithms in hierarchical clustering which differ in the way that similarity or distance between an element and a group of elements is defined and which consequently produce different results using the same data. Hierarchical methods require input such as how similar or different objects are in order to identify different clusters. One could use Matlab to calculate a measure of (dis)similarity by estimating the distance between elements. Elements with smaller distances between one another are more similar, whereas objects with larger distances are more dissimilar. For example,

in Matlab D = pdist(X) computes the Euclidean distance between pairs of objects in m-by-n data matrix **X**. Rows of **X** correspond to observations, and columns correspond to variables. **D** is a row vector of length $m(m–1)/2$, corresponding to pairs of observations in X. The distances are arranged in the order (2, 1), (3, 1), ..., (m, 1), (3, 2), ..., (m, 2), ..., (m, m–1)). **D** is commonly used as a dissimilarity matrix in clustering or multidimensional scaling.

The k-means clustering follows an entirely different concept to the hierarchical methods. It is not based on distance measures but uses the within-cluster variation as a measure to segmenting the data in such a way that the within-cluster variation is minimized. The clustering process starts by randomly assigning elements to a number of clusters. The elements are then successively reassigned to other clusters to minimize the within-cluster variation, which is basically the (squared) distance from each observation to the center of the associated cluster. If the reallocation of an object to another cluster decreases the within-cluster variation, this element is reassigned to that cluster. Therefore, the approach is non-hierarchical. We have used MATLAB for clustering purpose. *kmeans* function in MATLAB uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm separate objects between clusters until the sum cannot decrease further. To get an idea of how well the clusters or each element in the clusters are separated, we can use silhouette plot. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters. The silhouette value $S(i)$ was defined as the indicator of dissimilarity between clusters. Assume any object i in the dataset belong to cluster A. If cluster A contains objects apart from i, we calculate $a(i)$ as the average dissimilarity of i to all other objects of A.

Now assume a cluster C, and we can calculate $d(i, C)$, the average dissimilarity of i to all objects in the cluster C. After calculating, $d(i,C)$ for all clusters $C \neq A$. Now consider an another term $b(i)$; which can be defined as:

$$b(i) = \min_{C \neq A} d(i, C) \tag{3}$$

Using $a(i)$ and $b(i)$, the silhouette value $S(i)$ can be defined as follows:

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{4}$$

The silhouette value ranges between −1 and +1. If the value is +1 the element is distinct from other clusters, and 0 means the element is not distinctly in one cluster or another. If silhouette value −1 indicates that element is most probably in wrong cluster.

2.3. Implementation of Wavelet Hybrid Models

The study used two wavelet based hybrid models, neuro-wavelet (NW) and wavelet-Local Linear Regression combining discrete wavelet transforms (DWT) with ANNs and LLRs respectively. Wavelet transform theory and its application to multi-resolution signal decomposition has been thoroughly developed and well documented (e.g., Daubechies, 1988; Daubechies, 1992). Daubechies (1988) introduced the concept of orthogonal wavelet; generally referred to as Daybecies wavelet. Stream flow and precipitation time series are generally discrete in nature and, for analysing such series, discrete wavelet transforms (DWT) are more suitable. Thus for a discrete stream flow or precipitation time series $\mathbf{x_i}$, the DWT is defined as follows:

$$W_{a,b} = 2^{-a/2} \sum_{i=0}^{N-1} x_i \psi(2^{-a} i - b) \tag{5}$$

In the above equation, $W_{a,b}$ is DWT coefficient for scale a and time shift b, in which a and b are positive integers. N is the data length of the time series.

Mallat (1989) proposed an efficient way of decomposing the time series into 'approximations' (As) and 'details' (Ds) wavelet components/subtime series. The approximation series shows slowly changing information in the actual time series whereas the detail time series is the rapidly changing information in the time series. In other words, 'Approximation series' is 'low frequency and high scale' information and 'Detail series' is 'high frequency and low scale' information in the actual series.

The study has used 4 major antecedent information of runoff and rainfall series [three steps antecedent runoff values (Q(t-1), Q(t-2), Q(t-3)), one step antecedent rainfall, P(t-1)] and the rainfall information, P(t) for modelling. For wavelet hybrid modelling the antecedent and current input series were decomposed into three resolution levels and these decomposed subseries of the antecedent runoff and rainfall information were used to estimate the present value of runoff. A total of 30 subseries were obtained after decomposition, including the approximation subseries in these three resolution levels. We have used MATLAB for the wavelet application. The above [(Q(t-1), Q(t-2), Q(t-3)), P(t-1) and P(t)] are decomposed (i.e. the runoff and rainfall time series of 2-day mode ($D_q^{j-3}1$, $D_q^{j-2}1$, $D_q^{j-1}1$, $D_p^{i-1}1$, D_p^i1), 4-day mode ($D_q^{j-3}2$, $D_q^{j-2}2$, $D_q^{j-1}2$, $D_p^{i-1}2$, D_p^i2), 8-day mode ($D_q^{j-3}3$, $D_q^{j-2}3$, $D_q^{j-1}3$, $D_p^{i-1}3$, D_p^i3), and the approximate mode, 2-day mode ($A_q^{j-3}1$, $A_q^{j-2}1$, $A_q^{j-1}1$, $A_p^{i-1}1$, A_p^i1), 4-day mode ($A_q^{j-3}2$, $A_q^{j-2}2$, $A_q^{j-1}2$, $A_p^{i-1}2$, A_p^i2), 8-day mode ($A_q^{j-3}3$, $A_q^{j-2}3$, $A_q^{j-1}3$, $A_p^{i-1}3$, A_p^i3) , where q denotes runoff, p denotes rainfall and i and j denote the number of antecedent data sets of rainfall and runoff respectively. The wavelets decomposed the input data into three wavelet decomposition levels (2-4-8) is shown in Figures 3 and 4. These figures show detailed coefficient series and the first approximate series of the original runoff and rainfall data (Remesan et al., 2009).

2.3.1. Hybrid Neuro-Wavelet (NW) Model

In this study, a multi-layer feed-forward network type of

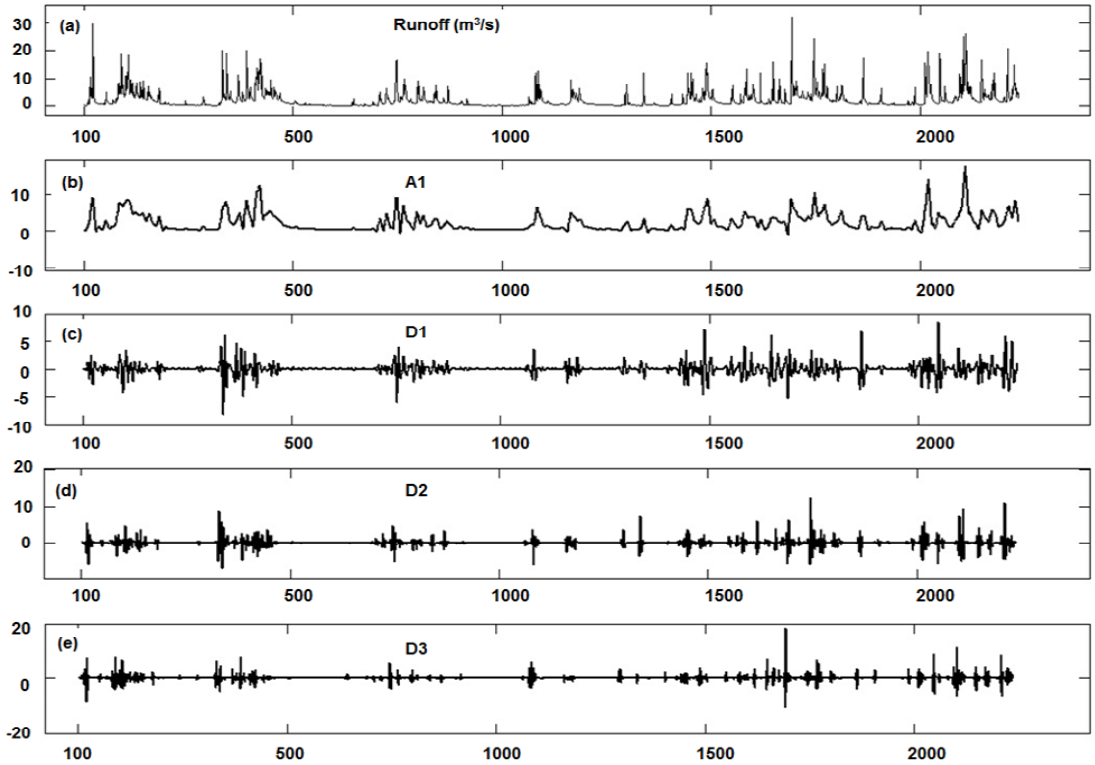


Figure 3. Three level decomposition sub-series of runoff data in Brue catchment (modified from Remesan et al. 2009).

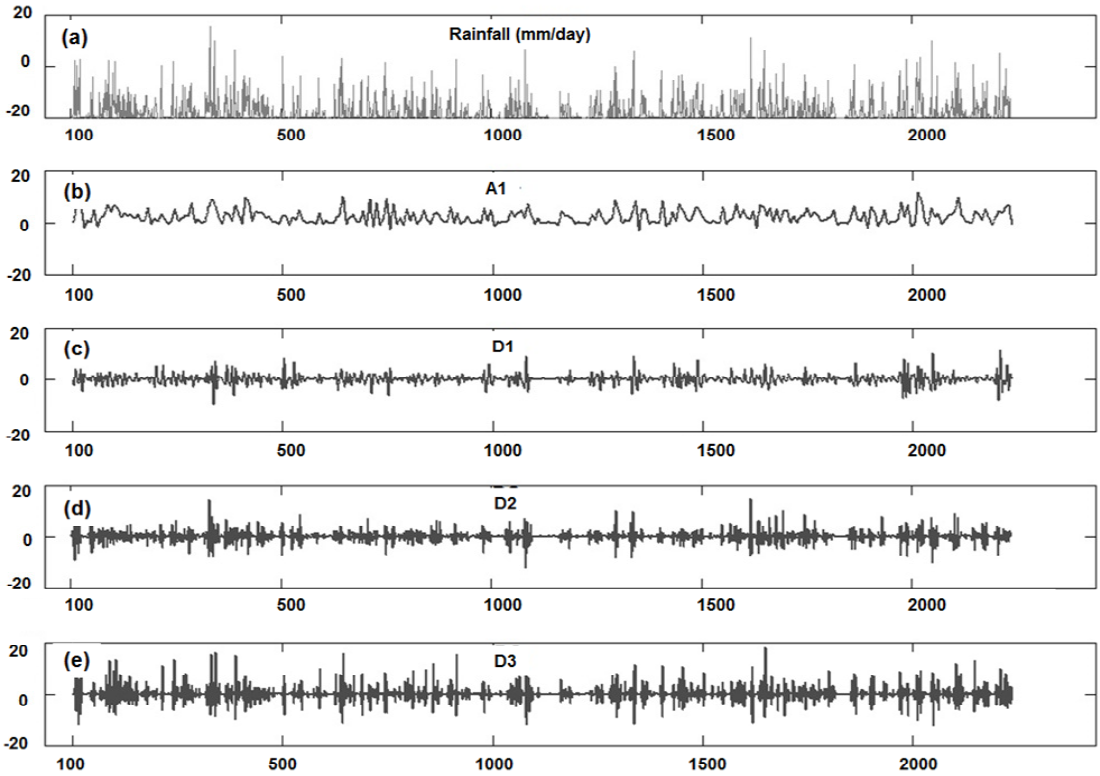


Figure 4. Three-level decomposition sub-series of precipitation data in Brue catchment (modified from Remesan et al., 2009).

ANN and DWT models were combined together to obtain a neuro-wavelet (NW) model. The discrete wavelet transfer model is functioned through two set of filters viz. high-pass and low-pass filters which decompose the signal into two sets of series namely detailed coefficients (D) and approximation (A) sub-series, respectively. In the proposed NW model, these decomposed sub series obtained from DWT on the original data directly are used as inputs to the ANN model. The study established a three-layer feed forward neural network (one input layer, one hidden layer, and one output layer), because of its proven ability in modelling many real-world functional problems. In this study, the Hecht-Nielsen (1990) approach has been adopted for the selection of hidden neurons in ANN modelling (i.e. number of hidden nodes set to be twice the input layer dimension plus one). To crosscheck the authenticity of this approach, we have performed some trial and error analysis before proceeding to actual modelling. The Levenberg-Marquardt training algorithm was used to adjust the weights of the feed forward neural network. The sigmoid and linear activation functions were employed for the hidden and output layers, respectively.

2.3.2. Implementation of Wavelet Local Linear Regression (W-LLR)

The nonparametric model, LLR, has been considered in this study along with traditional ANN and hybrid NW model for the estimation and comparison of the time series responses in hybrid rainfall runoff modelling. The LLR model has gained great acceptance among time series modellers because of its positive modelling abilities in low-dimensional forecasting problems. The LLR technique does not require a long time series for the development of a predictive model, in comparison to various statistically and analytical methods including neural network modelling. In fact, the LLR technique can make a prediction once three representative data points are available. Deciding the size of $p_{\max}$ (the number of near neighbours to be included for the local linear modelling) is the delicate phase in LLR based time series modelling.

Given a neighbourhood of $p_{\max}$ points, we must solve a linear matrix equation:

$$\boldsymbol{Xm} = \boldsymbol{y} \tag{6}$$

where $\mathbf{X}$ is a $p_{\max} \times d$ matrix of the $p_{\max}$ input points in d-dimensions, $\mathbf{x}_i (1 \le i \le p_{\max})$ are the nearest neighbour points, $\mathbf{y}$ is a column vector of length $p_{\max}$ of the corresponding outputs, and $\mathbf{m}$ is a column vector of parameters that must be determined to provide the optimal mapping from $\mathbf{X}$ to $\mathbf{y}$, such that:

$$\begin{Bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1d} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{x_{p\max}1} & x_{x_{p\max}2} & x_{x_{p\max}3} & \cdots & x_{x_{p\max}4} \end{Bmatrix} \begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ \vdots \\ m_d \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{p\max} \end{pmatrix} \tag{7}$$

The rank $\boldsymbol{r}$ of the matrix $\mathbf{X}$ is the number of linearly independent rows, which will affect the existence or uniqueness of the solution for $\mathbf{m}$.

If the matrix $\mathbf{X}$ is square and non-singular then a unique solution to Equation (8) is $\mathbf{m} = \mathbf{X}^{-1}\mathbf{y}$. If $\mathbf{X}$ is not square or singular, we should find a vector $\mathbf{m}$ which minimises:

$$| \boldsymbol{Xm} - \boldsymbol{y} |^2 \tag{8}$$

where the unique solution to this problem is provided by $\mathbf{m} = \mathbf{X}^{\#}\mathbf{y}$ where $\mathbf{X}^{\#}$ is a pseudo-inverse matrix.

In this study, the wavelet decomposed subseries were used as inputs to the LLR model. To differentiate this model from the traditional LLR with usual inputs, we called it wavelet-LLR (W-LLR)

2.4. Statistical Indices for Comparison

The study has employed the correlation coefficient (CORR), root mean squared error (RMSE) and mean bias error (MBE) for comparison of the models in both training and validation phases. These statistical terms can be defined as follows:

Correlation coefficient,

$$CORR = \frac{\sum_{i=1}^{N}(Q_{\mathrm{m}} - \bar{Q}_m)(Q_0 - \bar{Q}_0)}{\sqrt{\sum_{i=1}^{N}(Q_{\mathrm{m}} - \bar{Q}_m)^2 (Q_0 - \bar{Q}_0)^2}} \tag{9}$$

Coefficient of determination,

$$(R2) = \left| \frac{N(\sum Q_0 Q_m) - (\sum Q_m)(\sum Q_0)}{\sqrt{N\sum Q_0^2 - (\sum Q_0)^2}[N\sum Q_m^2 - (\sum Q_m)^2]} \right|^2 \tag{10}$$

Root mean squared error,

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_m - Q_0)^2} \tag{11}$$

Mean bias error,

$$MBE = \frac{\sum_{i=1}^{n}(Q_0 - Q_m)}{N} \tag{12}$$

where Q_m is the modelled or the estimated runoff by a model, Q_0 is the observation runoff, $\bar{Q}_m$ is the average of the estimated runoff, $\bar{Q}_0$ is the average of the observed runoff, and N is the number of observations.

3. Result and Discussion

3.1. Input Selection Using PCA and Cluster Analysis

This study has adopted a new approach (combining PCA and cluster analysis) for the selection of effective inputs (wavelet decomposed subseries in this case study). The first sub-

section shows how PCA was used to identify the relevant proportion of data sets in the total decomposed subseries and thus permitted the removal of redundant series. In the subsequent section, cluster analysis was used to spot effective subseries which are essential to make a meaningful hybrid model. The representative subseries in each cluster were identified based on silhouette value and later this subseries selection was compared with the cross-correlation method before being used in the final modelling with NW and W-LLR.

3.1.1. Application of PCA on the Decomposed Subseries

The main purpose of applying PCA to the whole data was to establish the predominant variations among different data series and to discover whether these variations were linked in some way to select better inputs for the modelling. Principal component analysis was performed using the covariance matrix and then again with the correlation matrix for each of the available 30 data series. High redundancy due to correlation of some decomposed data series was anticipated, due to the presence of approximation subseries in all three resolution levels in the whole data pool. The covariance matrix and correlation matrix based PCA analysis results on the decomposed daily rainfall runoff data at the Brue catchment are shown in Table 1. The comparison of the results offered no insight for establishing the significance of the correlation over the covariance matrix (or vice versa), since none of the scenarios analysed displayed any noteworthy difference. Analysis was conducted using both normalised and non-normalised data with no significant difference in the result.

From the analysis results, in an ideal scenario, 15 components out of 30 are required to express almost 100% of the information contained in the whole data sets used in the study. The remaining components have percentage variances of either zero or very close to zero. Comparing results obtained using the covariance matrix (in non-normalized data) with those of the correlation matrix, one can see that the variance contained in the first principal component produced by the covariance matrix was nearly 48%, which means a suitably selected single data at out of 30 can explain 48% of the whole information in the data series. It is also evident from Table 1 that six principal components can explain nearly 90% of the information in the whole data and seven principal components can explain 91% of the information. This indicates that around only six or seven of the components contained the significant information which in turn implies a large redundancy if we use all available 30 inputs. The analysis with PCA has found that out of 30 data series, six or seven data series can effectively make a model without any redundancy transferring more than 90% of the information inherent in the whole series. So the next step is the identifycation of these six subseries inputs precisely.

3.1.2. Clustering of Decomposed Subseries to Select Effective Inputs

Cluster analysis was used to identify which six inputs out of the 30 available inputs would be the most suitable. For this purpose, the study has applied both hierarchical clustering and k-means clustering to identify the natural clusters in the available 30 wavelet decomposed data series. However, the hierarchical clustering was not as effective as k-mean clustering since it always made separate clusters for natural groups details (D) and approximation (A) (those results are not included in this paper). Because of this limitation, the study concentrated on the use of k-means clustering and the representative elements in each cluster were identified using the silhouette values (S). The k-means clustering technique was applied to the decomposed subseries of daily rainfall runoff information from the Brue catchment, to get clusters from two to six. The clustering details are shown in Figure 5.

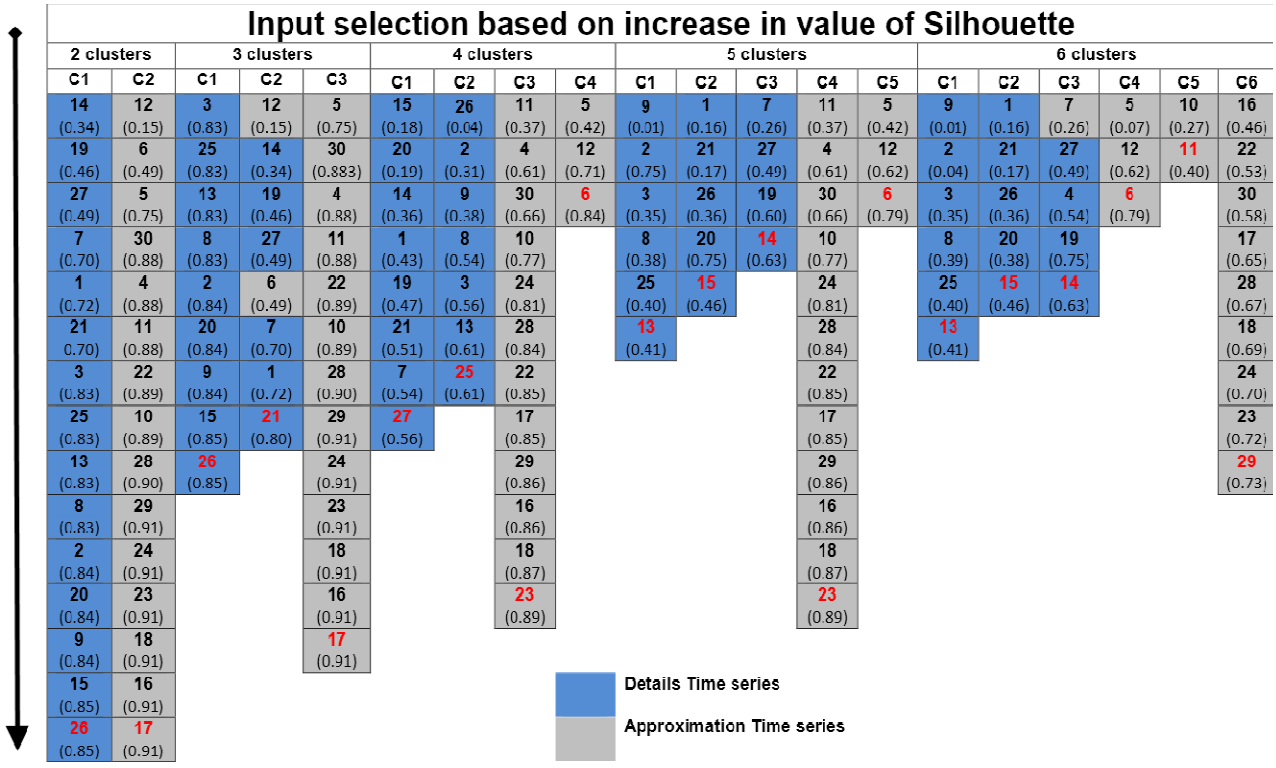
In Figure 5, the data series 1-6 is a decomposed sub series of the present value of rainfall information, $P(t)$ (the first three are details and the rest are approximation series). The data series 7-12 is t-1 antecedent rainfall information, $P(t\text{-}1)$ (the first three are details and the rest are approximation series). The data series 13-18 is t-1 antecedent runoff information, $Q(t\text{-}1)$ (the first three are details and the rest are approximation series). The data series 19-24 is t-2 antecedent runoff information, $Q(t\text{-}2)$ (the first three are details and the rest are approximation series). The data series 25-30 is t-3 antecedent runoff information, $Q(t\text{-}3)$ (the first three are details and the rest are approximation series). The dark colour corresponds to the decomposed approximation subseries and light colour is for the decomposed detailed subseries. In the 2-clustering, the approach made the two natural clusters with all details in one cluster and all approximation subseries in the other cluster. The S value indicates that the series number 26 (2^{nd} detail sub series of the $Q(t\text{-}3)$) and 17 (2^{nd} approximation of $Q(t\text{-}1)$) are the best series for modelling. The 3-clustering identifies the data series number 26 (2^{nd} detail sub series of the $Q(t\text{-}3)$), 17 (2^{nd} approximation of $Q(t\text{-}1)$) and 21(3^{rd} details sub series of $Q(t\text{-}3)$) based on the silhouette values. In the case of 3-clustering, the middle cluster in Figure 5 was a composite type containing data series from both details and approximations.

In the case of 4-clustering and 5-clustering each cluster was made from the natural groups without any composite cluster. The silhouette values obtained for the 4-cluster identified series 27, 25, 23 and 6; this corresponds to 3^{rd} detail of $Q(t\text{-}3)$, 1^{st} detail of $Q(t\text{-}3)$, 2^{nd} approximation of $Q(t\text{-}2)$ and 3^{rd} approximation of $P(t)$ respectively. In the case of 5-clustering the elements 13, 14, 15, 23 and 6 had the higher values for silhouette parameter in each clusters; these elements correspond to 1^{st} detail of $Q(t\text{-}1)$, 2^{nd} detail of $Q(t\text{-}1)$, 3^{rd} detail of $Q(t\text{-}1)$, 2^{nd} approximation of $Q(t\text{-}3)$ and 3^{rd} approximation of $P(t)$ respectively. The PCA results showed 6 components can convey 90% of the information contained in the whole data where as and 15 or 16 sub-series components can convey 100% of the information. We are also aware that there are at least 10 redundant approximation subseries as the study has used all three approbation series for modelling. So the study is more interested in the finding from the 6-clustering. The silhouette values of each 6-clusters suggest that the sub-series 13, 14, 15, 6, 11, and 29 could convey 90% of the formation contained in

Table 1. The Variation of Variance for Different PC Values and Input Types

	Using covariance matrix		Using correlation matrix		Using covariance and matrix normalized inputs		Using correlation matrix and normalized inputs	
P.C.	VE	CV (%)	VE	CV (%)	VE	CV (%)	VE	CV (%)
1st	47.67	48	34.25	34	40.62	41	34.25	34
2nd	17.63	65	13.02	47	14.62	55	13.02	47
3rd	7.28	73	9.71	57	10.06	65	9.71	57
4th	6.47	79	8.54	66	6.81	72	8.54	66
5th	4.99	84	6.04	72	5.76	78	6.04	72
6th	4.17	89.6	5.88	77	5.72	84	5.88	77
7th	2.82	91	5.17	83	4.24	88	5.17	83
8th	2.35	93	4.84	87	2.85	91	4.84	87
9th	1.58	95	3.41	91	2.43	93	3.41	91
10th	1.31	96	2.70	94	1.87	95	2.70	94
11th	1.14	97	1.69	95	1.26	96	1.69	95
12th	0.74	98	1.38	97	1.19	97	1.38	97
13th	0.56	99	1.08	98	0.80	98	1.08	98
14th	0.52	99	0.79	99	0.73	99	0.79	99
15th	0.21	100	0.60	99	0.29	99	0.60	99
16th	0.21	100	0.44	100	0.29	100	0.44	100

* Footnote: VE= Variance explained, CV= cumulative variance %



Sub-series 1-6 denotes 1st 2nd and 3rd details series and 1st 2nd 3rd approximation series of P(t), Sub-series 7-12 denotes 1st 2nd and 3rd details series and 1st 2nd 3rd approximation series of P(t-1), Sub-series 13-18 denotes 1st 2nd and 3rd details series and 1st 2nd 3rd approximation series of Q(t-1), Sub-series 19-24 denotes 1st 2nd and 3rd details series and 1st 2nd 3rd approximation series of Q(t-2), Sub-series 25-30 denotes 1st 2nd and 3rd details series and 1st 2nd 3rd approximation series of Q(t-3)

Figure 5. The *k*-means clustering details of wavelet decomposed daily rainfall runoff sub-series at the Brue catchment.

all 30 element data series. These sub-series correspond to 1st detail of $Q(t\text{-}1)$, 2nd detail of $Q(t\text{-}1)$, 3rd detail of $Q(t\text{-}1)$, 2nd approximation of $Q(t\text{-}1)$ and 2nd approximation of $Q(t\text{-}3)$. This shows that in wavelet hybrid modelling, different input series requires decomposition in different resolution levels. So the proposed methodology could be used as a guideline for making decisions regarding the extent of decomposition of inputs using wavelets. However, it is always appropriate to check the credibility of this finding through a full comparison with any conventional approaches and controlled modelling experiments before making a final conclusion. Abrahart and See (2007) have argued the need to have consistent measures of merit and trust in hydrological modelling. The study has identified that proper controlled experiments are inevitable for more authenticity and before acceptance of any model or approach. Thus, the study has performed a cross-check applying a cross-correlation analysis to the same cluster element.

3.1.3. Cross Correlation Analysis on Decomposed Subseries to Select Effective Inputs

The study has performed cross correlation analysis of the entire above mentioned cluster elements with the desired current runoff information, $Q(t)$. The analysis results are shown in Figure 6. The numbering details of each subseries used in this analysis are the same as that in Figure 5. The elements in Figure 6 are arranged in terms of increasing cross correlation value. It has been found that there is a considerable difference in the findings of both approaches. The cross correlation approach identified elements 15 (3rd detail of the $Q(t\text{-}1)$) and 17 (2nd approximation of $Q(t\text{-}1)$) as the best sub-series; whereas the k-mean clustering based on silhouette values identified 26 (2nd detail sub-series of the $Q(t\text{-}3)$) and 17 (2nd approximation of $Q(t\text{-}1)$). In the case of 3-clusters, the cross correlation values were higher for elements 15 (3rd detail of the $Q(t\text{-}1)$), 12 (3rd approximation of the $Q(t\text{-}1)$) and 17 (2nd approximation of $Q(t\text{-}1)$). At the same time the elements 26, 11 and 17 had the higher silhouette values.

The cross-correlation analysis found the elements 15, 9, 17, and 5 in the case of 4-clusters; these elements correspond to 3rd detail of $Q(t\text{-}1)$, 3rd detail of $P(t\text{-}1)$, 2nd approximation of $Q(t\text{-}1)$ and 2nd approximation of $P(t)$. In the case of 5 clusters the cross correlation identified the elements 3, 14, 15, 17 and 6; whereas the corresponding elements as per silhouette values were 13, 14, 15, 17 and 6. In the case of 6-clusters the cross correlation identified elements like 3 (3rd detail of $P(t)$), 4 (1st approximation of $P(t)$), 6 (3rd approximation of $P(t)$), 15 (3rd detail of $Q(t\text{-}1)$), 11(2nd detail of $Q(t\text{-}1)$) and 17 (2nd approximation of $Q(t\text{-}1)$) where as the silhouette values identified 13, 14, 15, 6, 11 and 29. These subseries 13, 14, 15, 6, 11 and 29 correspond to 1st detail of $Q(t\text{-}1)$, 2nd detail of $Q(t\text{-}1)$, 3rd detail of $Q(t\text{-}1)$, 2nd approximation of $P(t\text{-}1)$ and 2nd approximation of $Q(t\text{-}3)$ respectively. Both approaches have their own scientific base. So it is essential to use each set of suggested inputs in turn, during the modelling phase, so as to assess the credibility of these two approaches.

3.2. Modelling for Comparison of Silhouette and Cross-Correlation Approaches

To find the reliability of the decomposed sub-series inputs selected by the cross-correlation method and silhouette values, modelling using hybrid forms of both LLR and ANN models with wavelet decomposes subseries as inputs (W-LLR and NW models) was performed. Table 2 shows the modelled results up to cluster 6 and indicates that any combination lesser than PCA suggested combination producing better results.

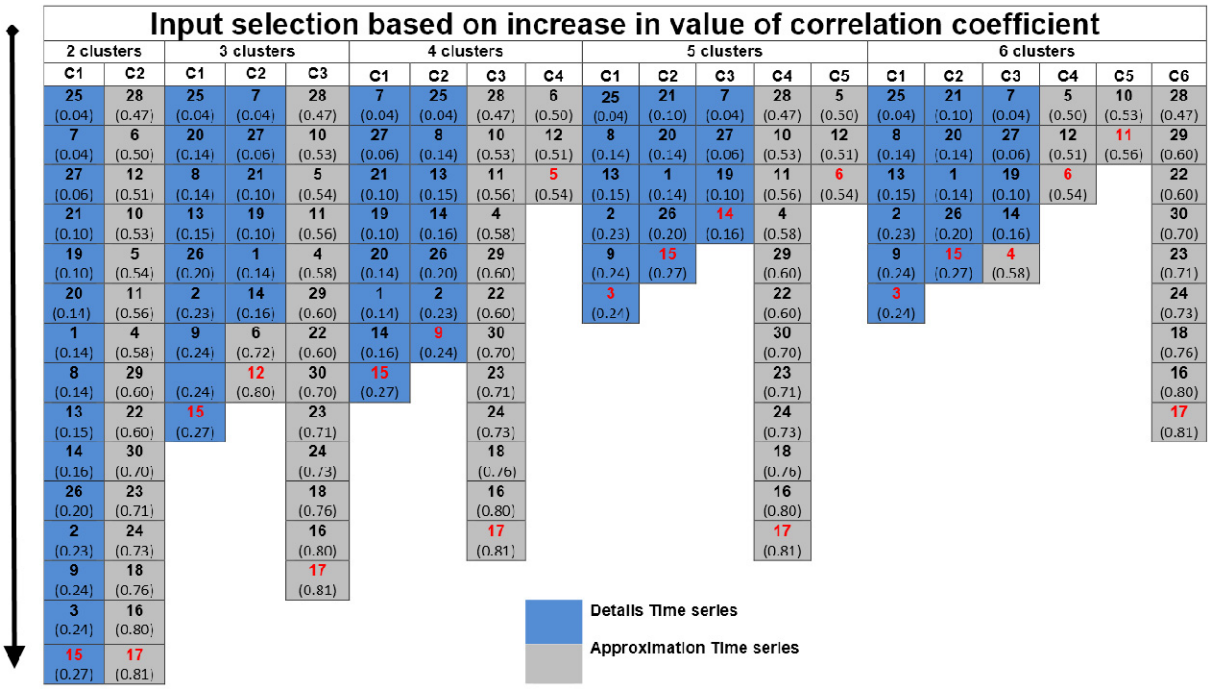
In the case of 2-clusters the inputs suggested by cross-correlation performed better than those based on the Silhouette value. In most cases the ANN models performed better or equally well than the LLR model. One should note from Table 2 that in the case of 4-clusters the models performance was quite low (even lower than 3-cluster and 2-clusters) in both the models because both the Silhouette value approach and the correlation approach failed to select proper and influential inputs. There was no detailed information from the runoff data in any of the input combinations suggested by the two approaches for 4-clusters. The 6-cluster with six inputs showed better performance than all other cases. The Silhouette value based approach succeeded to select major detail inputs like 13 (1st detail of $Q(t\text{-}1)$), 14 (2nd detail $Q(t\text{-}1)$), and 15 (3rd detail of $Q(t\text{-}1)$) in both 5-cluster and 6-cluster analysis. But the cross-correlation method failed to choose the inputs which account for sudden variations in the data spectrum. The comparison of cross correlation and Silhouette value based selection shows that in most cases the model inputs selected by Silhouette value performed better. However, in the case of 2-cluster and 4-cluster the performance of inputs selected by cross correlation outperformed that selected by Silhouette value. Even though the approach explained is in the context of selection of wavelet decomposed sub series, it could be used effectively in modelling cases where numbers of data series are available as inputs. The variation of modelled outputs (in terms of CORR) in training and validation phase using inputs obtained from different inputs (up to 12 clusters) are shown in Figure 7(a) and Figure 7(b). Figure 7(a) shows the modelled variations obtained from NW model based on inputs suggested by silhouette values and cross correlation analysis; whereas corresponding variations during training and validation phase of W-LLR model is given in Figure 7(b). The statistical indices were declining with further increase in input space in both training and validation phase. In the case of W-LLR model (using inputs obtained from silhouette values), there was sudden deterioration in statistical indices and CORR value in the validation phase when the input space increase above 7. A similar, though smaller, trend was observed in the case of W-LLR model when using inputs obtained from the cross correlation analysis.

3.3. Model Comparison with Traditional LLR and ANNs

The study has made a comparison of the hybrid LLR and ANN with wavelet decomposed inputs with traditional LLR and ANN with the unaltered antecedent information [i.e. three

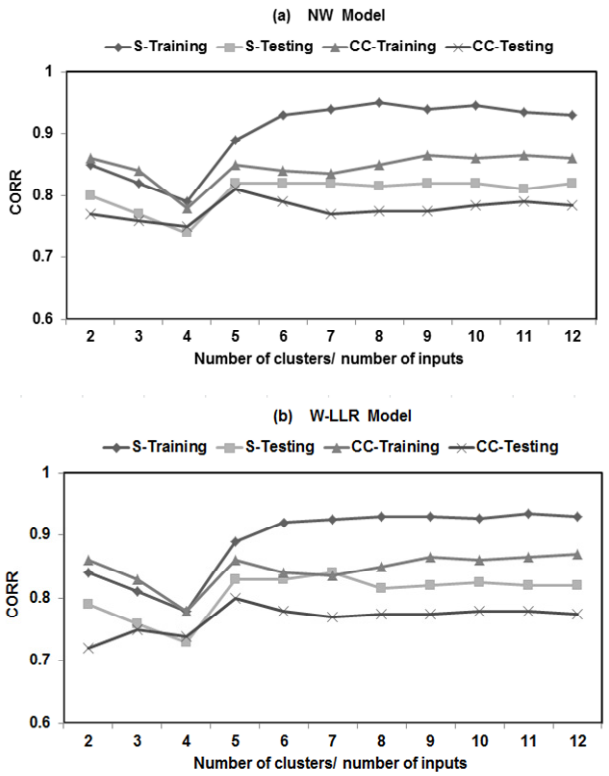
Table 2. Comparison of Silhouette Value and Cross-Correlation Based Data Selection Using Hybrid Forms of LLR and ANN Models

Clusters	Methods	Model	Training (1056 points)			Validation		
			MBE (m^3/s)	RMSE (m^3/s)	CORR	MBE (m^3/s)	RMSE (m^3/s)	CORR
2-cluster	Silhouette values	W-LLR	-0.182	0.584	0.84	-0.278	0.802	0.79
		NW	-0.179	0.523	0.85	-0.228	0.712	0.80
	Cross correlation	W-LLR	-0.180	0.534	0.86	-0.259	0.795	0.72
		NW	-0.183	0.540	0.86	-0.240	0.783	0.77
3-cluster	Silhouette values	W-LLR	-0.201	0.590	0.81	-0.301	0.810	0.76
		NW	-0.239	0.579	0.82	-0.297	0.802	0.77
	Cross correlation	W-LLR	-0.221	0.596	0.83	-0.389	0.860	0.75
		NW	-0.213	0.581	0.84	-0.368	0.819	0.76
4-cluster	Silhouette values	W-LLR	-0.356	0.625	0.78	-0.401	0.870	0.73
		NW	-0.310	0.590	0.79	-0.389	0.823	0.74
	Cross correlation	W-LLR	-0.351	0.621	0.78	-0.405	0.861	0.74
		NW	-0.305	0.619	0.78	-0.382	0.821	0.75
5-cluster	Silhouette values	W-LLR	-0.163	0.444	0.89	-0.182	0.692	0.83
		NW	-0.159	0.493	0.89	-0.178	0.681	0.82
	Cross correlation	W-LLR	-0.171	0.541	0.86	-0.188	0.689	0.80
		NW	-0.181	0.532	0.85	-0.219	0.671	0.81
6-cluster	Silhouette values	W-LLR	-0.158	0.421	0.92	-0.175	0.692	0.83
		NW	-0.155	0.415	0.93	-0.169	0.681	0.82
	Cross correlation	W-LLR	-0.232	0.578	0.84	-0.222	0.859	0.78
		NW	-0.223	0.571	0.84	-0.258	0.711	0.79



Sub-series 1-6 denotes 1st 2nd and 3rd details series and 1st 2nd 3rd approximation series of P(t), Sub-series 7-12 denotes 1st 2nd and 3rd details series and 1st 2nd 3rd approximation series of P(t-1), Sub-series 13-18 denotes 1st 2nd and 3rd details series and 1st 2nd 3rd approximation series of Q(t-1), Sub-series 19-24 denotes 1st 2nd and 3rd details series and 1st 2nd 3rd approximation series of Q(t-2), Sub-series 25-30 denotes 1st 2nd and 3rd details series and 1st 2nd 3rd approximation series of Q(t-3)

Figure 6. The cross-correlation analysis results on the clusters of wavelet decomposed daily rainfall runoff sub-series at the Brue catchment.



The lines corresponding to 'S' denotes inputs obtained from Silhouette values approach and lines corresponding to 'CC' denotes inputs obtained from cross correlation method, CORR denotes coefficient of correlation

Figure 7. The Line plots of hybrid modeling outputs: (a) NW-Model; (b) W-LLR model.

steps antecedent runoff values ($Q(t\text{-}1)$, $Q(t\text{-}2)$, and Q $(t\text{-}3)$), one step antecedent rainfall ($P(t\text{-}1)$) and current rainfall information ($P(t)$)]. The analysis results are shown in Table 3 in terms of different statistical indices.

The performances of these models are presented in Table 3, which shows that the NW model performs very well in both validation and training data. The NW model has an RMSE of 0.415 m^3/s (20.7%) during the training phase, and a validation RMSE of 0.681 m^3/s (28.38%) (30.32% improvement in comparison to traditional ANN). The correlation coefficient between the NW computed and observed were found to be 0.93 during training and 0.82 during validation; the corresponding values for traditional ANN were 0.84 and 0.76, respectively. The scatter plots of the performance of Silhouette based [PCA + cluster] hybrid NW models during training and validation phases are shown in Figures 8(a) and (b) for training and validation phase respectively. The observed and estimated runoff values of the traditional ANN model (with LM algorithm) for both training and validation data are given Figures 9 (a) and (b) in the form of a scatter plots. It is interesting to note that the LLR model has faced some difficulties in training with wavelet decomposed subseries in comparison to that with normal inputs. The RMSE for the W-LLR model was higher (0.421 m^3/s (21.05%)) compared with that of the traditional LLR model (0.414 m^3/s (20.7%)) during training phase. But the W-LLR model gave more consistency and better result during the validation phase within an RMSE of 0.692 m^3/s (28. 83%) in comparison to the corresponding result of the LLR model during validation. However, this discrepancy of wavelet based LLR model during training highlights the need to establish the hybrid models more carefully to avoid flawed modelling outcomes. The scatter plot of the observed and predicted runoff values using the traditional LLR model in the validation phase is shown in Figures10 (a) and (b). From the MBE value one can deduce that incorporation of wavelets has added more bias to the prediction during both training and validation phase.

In Table 3, we have also presented the results from another modelling scheme, in which model inputs are reconstructed series constituted from sub-series excluding ineffective components. We identified sub-series with cross correlations of less than 0.1 as ineffective subseries. It can be seen from the table that both NW and W-LLR with reconstructed inputs performed better than that of traditional ANN and LLR with non-decomposed inputs. The inter-comparison of LLR and ANN in this modelling phase shows that the LLR is outperforming ANN in training phase with closer performance in validation. The W-LLR with the reconstructed input series shows better [RMSE = 0.398 m^3/s (19.9%)] performance than Silhouette value suggested W-LLR model in training phase but poor performance in validation [RMSE = 0.725 m^3/s (30.21%)]. The NW with reconstructed inputs gave better performance in both training and validation period in comparison to Silhouette values based NW model (Table 3). It should be also mentioned that, although the performance of both hybrid models is satisfactory with reconstructed input series, better statistical indices are associated with Silhouette [i.e. PCA + cluster] based modelling scheme. The results indicated that the selected subseries by [PCA + cluster] method improved the modelling performance in comparison to traditional models and hybrid models with reconstructed inputs. However, Remesan et al. (2009) showed better performance by the NW model than the LLR model when using all available decomposed sub-series. But in this study, when the number of sub-series is lesser and representative, both LLR and its hybrid form have shown better performance than ANN and NW for all three modelling schemes shown in Table 3. The modelling and training time were less in this study when we used six effective subseries. This comparative study has once again confirmed that the wavelet decomposing would help model performance with less complexity if one could select effective subseries intelligently using suitable approaches.

A novel 'PCA conjunctive clustering analysis input variable pre-processing method' is explored and, subsequently, we have experimented its effect on the wavelet based hybrid LLR and the neural network models. The modelling results have shown that the proposed method provides more accurate performance on daily simulation in comparison to the input selection by cross correlation. Even though the present study is to identify better sub-series for modelling, this approach has wider implycations in selecting suitable input time series in en-

Table 3. Modelling Performance Comparison of Hybrid Forms of LLR and ANN Models with Their Traditional Forms with Unaltered Inputs

Models	Inputs	Training data (1056 points)			Validation data		
		MBE (m^3/s)	RMSE (m^3/s and %*)	R^2	MBE (m^3/s)	RMSE (m^3/s and %**)	R^2
W-LLR	6 wavelet subseries (S value)	-0.158	0.421 (21.05)	0.92	-0.175	0.692 (28.83)	0.83
NW	6 wavelet subseries (S value)	-0.155	0.415 (20.7)	0.93	-0.169	0.681 (28.38)	0.82
W-LLR	Reconstructed inputs from effective sub-series	-0.138	0.398 (19.9)	0.92	-0.212	0.725 (30.21)	0.82
NW	Reconstructed inputs from effective sub-series	-0.210	0.508 (25.4)	0.90	-0.189	0.712 (29.67)	0.81
LLR	Original antecedent data	-0.028	0.414 (20.7)	0.92	-0.171	0.922 (37.7)	0.70
ANN	Original antecedent data	-0.144	1.18 (60.3)	0.84	-0.042	1.43 (58.7)	0.76

*, ** The percentage value is percentage to mean runoff value on both training and validation phase respectively.

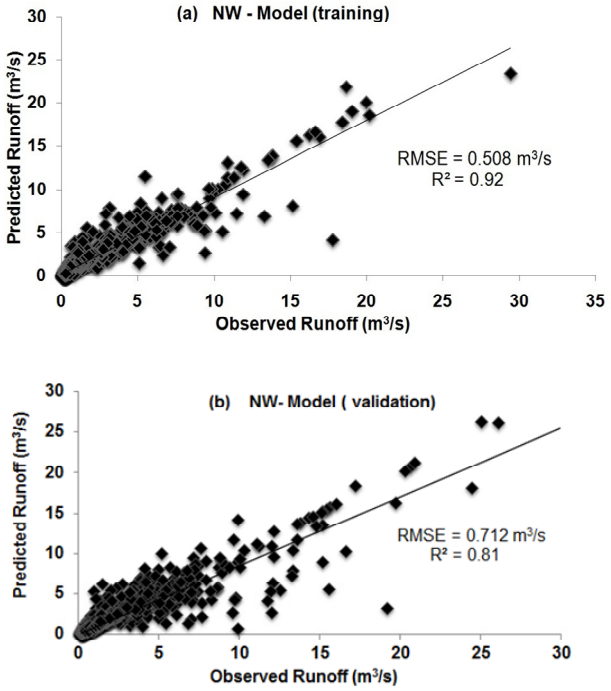


Figure 8. Scatter plots of PCA + cluster analysis silhouette based NW model: (a) training; (b) validation.

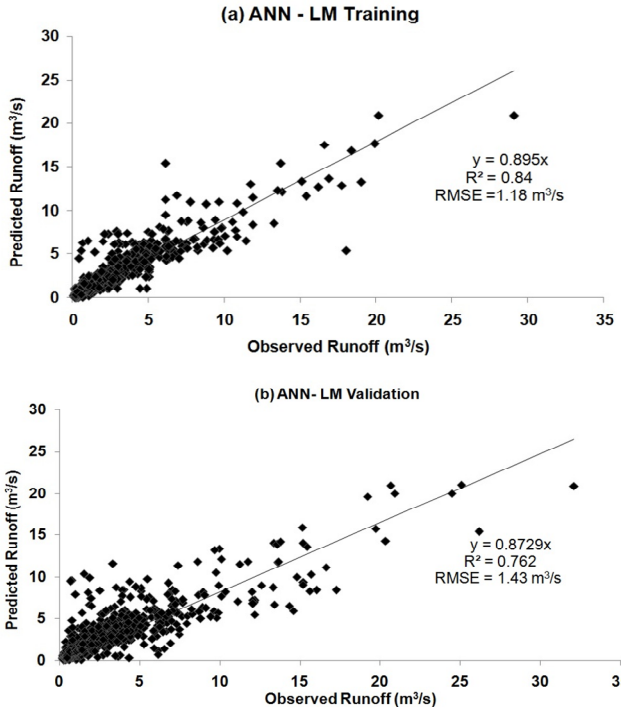


Figure 9. Scatter plots of traditional ANN-LM model: (a) training; (b) validation.

vironmental modelling scenarios with numerous input-spaces. In such cases PCA method reduces the number of input variables without fully eliminating them, then *K*-means clustering approach can identify the suitable input series for modelling with help of Silhouette plot. Albeit, in this particular case study, the input variables for pre-processing/selection by means of PCA and k-means clustering techniques gave better simulation for two data based approaches, the performance may be different depending on number of inputs, training data length (Han et al., 2007). The cross-correlation method is a linear approach and is unlikely to be very useful for nonlinear systems. In this study, we have used the first 1056 data points in the training data but the result could be different for a different set of data for training (e.g., last or somewhere in the middle).

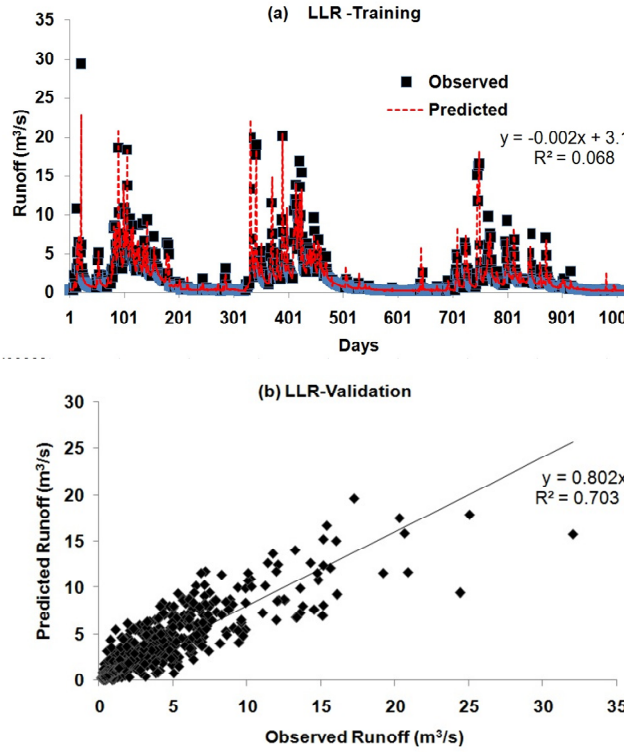


Figure 10. The observed versus the LLR predicted daily runoff at the Brue catchment: (a) time series plot of the training data set; (b) scatter plot of the validation data set.

PCA is a versatile linear transformation tool as it is completely reversible to generate the actual data from principal components. Just a word of caution even though the PCA clustering combination is suitable for better inputs selection, the modeller needs to be aware that the PCA will not properly deal with data lying on nonlinear manifolds. One of the major disadvantages of this input selection approach is that, in some cases, the PCA could not capture even simplest invariance unless the training data explicitly provides this information. This may cause unpredictable results. Brosse et al. (2001) noted that the PCA provided irrelevant ecological input information for some of their cases because of underlined linear principles in the PCA methods. The relationships between environmental variables are highly nonlinear so modellers should be cautious while using the PCA in environmental modelling. We reduced the input dimension from 30 to 6, achieving a 90% cumulative variance. Due to negative impacts of PCA method, there is a possible chance of compromise between the computation and accuracy in a given modelling technique.

4. Conclusions

Due to availability of better computational facilities, it is a general practice in neural network based modelling to assume that having more information or datasets is always better than having less. This study aimed to highlight the problem of selecting the redundant input variables for a prediction model. The study proposed a new approach which combined the PCA and cluster analysis to select parsimonious sets of inputs for prediction models. This approach was established to choose effective wavelet sub-series that could be used for hybrid rainfall runoff modelling. The main advantage of the proposed approach is that it gives an idea about how many, or the proportion of, redundant inputs that exist in the whole available input pool with the help of PCA. The study has identified that 6 of 60 sub-series can provide more than 90% of the information inherent in the system. The *K*-means clustering with the help of silhouette value identified the best and useful subseries from the input space. This approach has provided guidelines to modellers on the number of decomposition resolution levels to be adopted in each input series in wavelet hybrid modelling. The comparison with traditional cross-correlation approach has shown that the inputs selected by both the proposed approaches have only 40% similarity. The extensive modelling with selected subseries with both approaches has confirmed the advantage of the silhouette value in identifying the effective input in a natural cluster. What is an acceptable level of variance explained by the model? This question remains as a pivotal problem in PCA based data reduction procedures which decide the number of principal components we consider for case studies. We acknowledge that limiting the total variance value to 90% could be considered as a drawback of principal component analysis in this study. Nevertheless, this method is observed useful to identify effective inputs if the modelling problem constrained by large number of linear and nonlinear inputs with hidden redundancy.

Even though the study deals with the subseries selection in hybrid modelling, the precise input identification capability of this approach makes it viable for input selection in large-scale hydrological time series modelling problems. The authors urge further research on this approach including controlled experiments with different data sets (e.g., MOPEX dataset) and comparisons to alternate input selection methodologies such as entropy and mutual information.

Acknowledgments. This study has used data sets from HYREX (Hydrological Radar Experiment) funded by NERC (Natural Environment Research Council). Authors also would like to thank Prof Dawei Han for his support throughout preparation of this manuscript.

References

Abrahart, R.J. and See L.M. (2007). Neural network modelling of non-linear hydrological relationships. *Hydrology and Earth System Sciences*., 11(5), 1563-1579. http://dx.doi.org/10.5194/hess-11-15-63-2007

Ahmadi, A., Han, D., Karamouz, M., and Remesan, R. (2009). Input data selection for solar radiation estimation. *Hydrol. Process.,* 23 (19), 2754-2764. http://dx.doi.org/10.1002/hyp.7372

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000). Artificial neural networks in hydrology -- I: Preliminary concepts. *J. Hydrol. Eng.,* 5(2), 115-123. http://dx.doi.org/10.1061/(ASCE)1084-0699(2000)5:2(115)

Back, A.D. and Trappenberg, T.P. (1999). Input variable selection us-

ing independent component analysis, *Proc. of International Joint Conference on Neural Networks*, 2, 989-992, 1999. http://dx.doi.org/10.1109/ijcnn.1999.831089

Bell, V. A. and Moore, R. J. (2000). Short period forecasting of catchment-scale precipitation. Part II: A water-balance storm model for short-term rainfall and flood forecasting. *Hydrol. Earth Syst. Sc.*, 4(4), 635-651. http://dx.doi.org/10.5194/hess-4- 635-2000

Bowden, G.J., Dandy, G.C., and Maier, H.R. (2005). Input determination for neural network models in water resources applications. Part 1 -- Background and methodology. *J. Hydrol.,* 301(1-4), 75-92. http://dx.doi.org/10.1016/j.jhydrol.2004.06.021

Brosse, S., Giraudel, J.L., and Lek, S. (2001). Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. *Ecol. Model.,* 146(1-3), 159-166. http://dx.doi.org/10.1016/S0304-3800(01)00303-9

Budu, K. (2014). Comparison of wavelet based ANN and regression models for reservoir inflow forecasting. *J. Hydrol. Eng.,* 19(7), 13-85-140. http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0000892

Caraway, N.M., McCreight, J.L., and Rajagopalan, B. (2014). Multisite stochastic weather generation using cluster analysis and k-nearest neighbor time series resampling. *J. Hydrol.,* 508(16), 197-213. http://dx.doi.org/10.1016/j.jhydrol.2013.10.054

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Commun. Pure Appl. Math.,* 41(7), 909-996. http://dx.doi.org/10.1002/cpa.3160410705

Daubechies, I. (1992). *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia, PA, 3-5. http://dx.doi.org/10.1137/1.9781611970104

Davis, J.C. (1986). *Statistical and Data Analysis in Geology*, John Wiley & Sons, New York.

Durrant, P.J. (2001). *winGamma: A Non-linear Data Analysis and Modelling Tool with Applications to Flood Prediction*, Ph.D. Thesis, Department of Computer Science, Cardiff University, Wales, UK.

Ganti, R. and Jain, A. (2011). Input variable selection for hydrologic modeling using ANNs. *Am. Geophys. Union,* Fall Meeting 2011, abstract #H23H-1380.

Han, D., Kwong, T., and Li, S. (2007). Uncertainties in real-time flood forecasting with neural networks, *Hydrol. Process.,* 21(2), 223-228. http://dx.doi.org/10.1002/hyp.6184

Hecht-Nielsen, R. (1990). *Neurocomputing*, Addison-Wesely Publishing Company, Reading, MA.

Jain, A., Sudheer, K.P., and Srinivasulu, S. (2004a). Identification of physical processes inherent in artificial neural network rainfall runoff models. *Hydrol. Process.,* 18(3), 571-581. http://dx.doi.org/10.1002/hyp.5502

Jain, S.K., Singh, V.P., and Van Genuchten, M.T. (2004b). Analysis of soil water retention data using artificial neural networks. *J. Hydrol. Eng.,* 9(5), 415-420.

King, J.R. and Jackson, D.A. (1999). Variable selection in large environmental data sets using principal component analysis. *Environmetrics,* 10(1), 67-77. http://dx.doi.org/10.1002/(SICI)1099-095X(199901/02)10:1<67::AID-ENV336>3.0.CO;2-0

Kisi, O. (2008). Stream flow forecasting using neuro-wavelet technique. *Hydrol. Process.,* 22(20), 4142-4152. http://dx.doi.org/10.1002/hyp.7014

Kisi, O. (2009). Neural network and wavelet conjunction model for modelling monthly level fluctuations in Turkey. *Hydrol. Process.,* 23(14), 2081-2092. http://dx.doi.org/10.1002/hyp.7340

Krishna, B., Satyaji Rao, Y.R., and Nayak, P.C. (2012). Wavelet neural network model for river flow time series. *Proc. ICE Water Manage.*, 165(8), 425-439. http://dx.doi.org/10.1680/wama.10.00092

Krzanowski, W.J. (1987). Selection of variables to preserve multivariate data structure using principal components. *J. Roy. Stat. Soc. Ser. C. (Appl. Stat.),* 36(1), 22-33. http://dx.doi.org/10.2307/ 2347-842

Lane, S.N. (2007). Assessment of rainfall-runoff models based upon wavelet analysis. *Hydrol. Process.,* 21(5), 586-607. http://dx.doi.org/10.1002/hyp.6249

Levi, M.R. and Rasmussen, C. (2014). Covariate selection with iterative principal component analysis for predicting physical soil properties. *Geoderma,* (219-220), 46-57. http://dx.doi.org/10.1016/j.geoderma.2013.12.013

Maheswaran, R. and Khosa, R. (2012). Comparative study of different wavelets for hydrologic forecasting. *Comput. Geosci.,* 46, 284-295. http://dx.doi.org/10.1016/j.cageo.2011.12.015

Maier, H.R. and Dandy, G.C. (2000). Neural networks for the prediction and forecasting of water resources variables: A review of modeling issues and applications. *Environ. Model. Software,* 15(1), 101-124. http://dx.doi.org/10.1016/S1364-8152(99)00007-9

Mallat, S.G. (1989). A theory for multi resolution signal decomposetion: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.,* 11(7), 674-693. http://dx.doi.org/10.1109/34.192463

Manly, B.F.J. (1986). *Multivariate Statistical Methods: A Primer*, 2nd Edition, Chapman & Hall, London.

Nayak, P.C., Sudheer, K.P., Rangan, D.M., and Ramasastri, K.S. (2004). A neuro-fuzzy computing technique for modelling hydrological time series. *J. Hydrol.,* 291(1-2), 52-66. http://dx.doi.org/10.1016/j.jhydrol.2003.12.010

Noori, R., Karbassi, A.R., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M.H., Farokhnia, A., and Gousheh., M.G. (2011). Assessment of input variables determination on the SVM model performance using PCA, Gammatest, and forward selection techniques for monthly stream flow prediction. *J. Hydrol.,* 401(3-4), 177-189. http://dx.doi.org/10.1016/j.jhydrol.2011.02.021

Nourani, V., Komasi, M., and Mano, A. (2009). A multivariate ANN-wavelet approach for rainfall-runoff modeling. *Water Resour. Manage.,* 23(14), 2877-2894. http://dx.doi.org/10.1007/s11269-009-9414-5

Nourani, V., Kisi, O., and Komasi, M. (2011) Two hybrid artificial intelligence approaches for modeling rainfall-runoff process. *J. Hydrol.,* 402(1-2), 41-59. http://dx.doi.org/10.1016/j.jhydrol.2011.03.002

Nourani, V., Baghanam, A.H., Adamowski, J., and Kisi, O. (2014). Applications of hybrid wavelet-Artificial Intelligence models in hydrology: A review. *J. Hydrol.,* 514(6), 358-377. http://dx.doi.org/10.1016/j.jhydrol.2014.03.057

Partal, T. and Cigizoglu, H.K. (2009). Prediction of daily precipitation using wavelet-neural networks. *Hydrol. Sci. J.,* 54(2), 234-246. http://dx.doi.org/10.1623/hysj.54.2.234

Remesan, R., Shamim, M.A., and Han, D. (2008). Model data selection using gamma test for daily solar radiation estimation. *Hydrol. Process.,* 22(21), 4301-4309. http://dx.doi.org/10.1002/hyp.7044

Remesan, R., Shamim, M.A., Han, D., and Jimson, M. (2009). Runoff prediction using an integrated hybrid modelling scheme. *J. Hydrol.,* 372(1-4), 48-60. http://dx.doi.org/10.1016/j.jhydrol.2009.03.034

Sang, Y.F., Wang, Z., and Liu, C. (2012). Discrete wavelet-based trend identification in hydrologic time series. *Hydrol. Process.,* 27 (14), 2021-2031. http://dx.doi.org/10.1002/hyp.9356

Sseganea, H., Tollnera, E.W., Mohamoudb, Y.M., Rasmussenc, T.C., and Dowdd, J.F. (2012). Advances in variable selection methods I: Causal selection methods versus stepwise regression and principal component analysis on data of known and unknown functional relationships. *J. Hydrol.,* (438-439), 16-25. http://dx.doi.org/10.1016/j.jhydrol.2012.01.008

Stefánsson, A., Končar, N., and Jones, A.J. (1997). A note on the Ga-

mma test. *Neural Comput. Appl.,* 5(3), 131-133. http://dx.doi.org/10.1007/BF01413858

Sudheer, K.P., Gosain, A.K., Rangan, D.M., and Saheb, S.M. (2002). Modelling evaporation using an artificial neural network algorithm. *Hydrol. Process.,* 16(16), 3189-3202. http://dx.doi.org/10.1002/hyp.1096

Sudheer, K.P. (2005). Knowledge extraction from trained neural network river flow models. *J. Hydrol. Eng.,* 10(4), 264-269. http://dx.doi.org/10.1061/(ASCE)1084-0699(2005)10:4(264)

Sudheer, K.P. and Jain, A. (2004). Explaining the internal behaviour of artificial neural network river flow models. *Hydrol. Process.,* 18(4), 833-844. http://dx.doi.org/10.1002/hyp.5517

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*, Springer-Verlag New York, USA. http://dx.doi.org/10.1007/978-1-4757-2440-0

Zhi-hang, T. (2009). Investigation and application of cluster analysis in service industries, *International Engineering an Electronic Commerce, International Symposium on IEEE*, 827-831. http://dx.doi.org/10.1109/ieec.2009.179