Wavelet-based feature extraction of rainfall-runoff process via Self-Organizing Map

VAHID NOURANI[1], MASOUMEH PARHIZKAR[2], TOHID REZAPOUR KHANGHAH, AIDA HOSSEINI BAGHANAM, ELNAZ SHARGHI

Department of Water Resources Engineering, Faculty of Civil Engineering

University of Tabriz

29 Bahman Ave., Tabriz, Iran.

IRAN

[1]E-mail: vnourani@yahoo.com [2]E-mail: m.parhizkar66@gmail.com

Abstract: - One of the essential steps in any modeling is the determination of dominant input variables, else it may lead to poor model accuracy. Rainfall and runoff time series due to various seasonalities involved in the processes may have numerous possible subsets. Therefore, looking for an efficient mathematical-based tool in order to capture dominant subsets seems inevitable. The wavelet transform is a drastic mathematical tool to capture the multi-scale features of a signal where on the other hand the Self-Organizing Map (SOM) can be a powerful technique capable of ordering multivariate data by similarity as preserving the topological structure of the data. The current paper presents a procedure to extract dominant features of rainfall and runoff process obtained from the Delaney Creek Sub-basin located in Tampa Bay Watershed at Florida, USA. Firstly, the wavelet technique was used to decompose the main rainfall and runoff time series into several sub-series. Subsequently, several independent sub-series were chosen via SOM. Results revealed the efficiency of the applied methodology in choosing dominant subsets of the input data.

Key-Words: - Rainfall-runoff modeling; features extraction; Wavelet transform; Clustering; Self-Organizing Map; Delaney Creek Sub-basin

1 Introduction

Accurate modeling of hydrological processes such as rainfall-runoff modeling, which can be helpful in city planning, land uses, flood and water resource management, is of prime importance for hydrologists and water resource engineers. One of the essential steps when using any mathematical tool is to determine dominant input variables of the process. Since some of the inputs may be correlated, noisy or have no significant relationship with output variables, they are not equally informative [1]. More difficult learning, divergence, obscurity and poor model accuracy are some shortcomings which come along with application of a mathematical model without proper data pre-processing.

Therefore, a challenge in using a mathematical tool is to extract input subsets that are independent, informative and efficiently cover the proposed input domain. Numerous possible subsets of input variables, especially in the rainfall-runoff modeling where appropriate lags must also be chosen, is a cause to look for an efficient tool.

Potency of the wavelet in decomposing non-stationary time series into different scales which can explain simultaneously both spectral and temporal information of a signal creates an effective implement in facing with high non-stationary signal fluctuations and seasonalities features involved in the process. Remesan et al. used the wavelet transform in Runoff prediction [2]. Tiwari and Chatterjee developed a hybrid wavelet–bootstrap–Artificial Neural Network (WBANN) model to explore the potential of wavelet and bootstrapping techniques for developing an accurate and reliable ANN model for hourly flood forecasting [3]. Nourani et al. introduced two hybrid artificial

Intelligence approaches, including Wavelet-ANFIS model for developing a rainfall–runoff model [4].

Selection of the most relevant and appropriate wavelet- based features is an important step in modeling of the rainfall-runoff process when various data sources are available over the watershed. The Self-Organizing Map (SOM) due to its capability in visualization of data relationships is a useful mathematical-based tool to determine the prominent sub-series among diverse sets. SOM is a kind of ANN method which has the authority to classify, cluster, estimate, predict and data mining [5]. It is an effective tool to convert complex, nonlinear, statistical relationship between high-dimensional data items into simple, geometric relationship on a low-dimensional display [6]. This method is applied increasingly in hydrology and water resources. Bowden et al. divided ANN data into training, testing and validation subsets using SOM [7]. Lin and Chen have used SOM in identification of homogeneous regions for regional frequency analysis [8]. Lin and Wu presented a SOM-based approach to estimate design hyetographs of ungauged sites [9]. Kalteh et al. have reviewed the published applications of the SOM in different fields of water resources problems [5].

In this paper the wavelet transform is used to decompose the main rainfall-runoff time series into several sub-series. These sub-series are then clustered via SOM to choose some independent sub-series which can efficiently be used in modeling rainfall-runoff process.

2 The wavelet transform

The wavelet transform has increased in usage and popularity in recent years. A comprehensive literature survey of wavelet in geosciences can be found in Foufoula-Georgiou and Kumar [10], and the most recent contributions are cited by Labat [11].

The time-scale wavelet transform of a continuous-time signal, x (t), is defined as [12]:

$$T(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} g^* \left(\frac{t-b}{a} \right) x(t)\, dt \qquad (1)$$

Where * corresponds to the complex conjugate and g(t) is called wavelet function or mother wavelet. The parameter *a* acts as a dilation factor, while *b* corresponds to a temporal translation of the function g (t), which allows the study of the signal around b. The main property of wavelet transform is to provide a time-scale localization of process, which derives from the compact support of its basic function. This is opposed to the classical trigonometric function of Fourier analysis. The wavelet transform searches for correlations between the signal and wavelet function.

For practical applications, the hydrologists do not have at their disposal a continuous-time signal process but rather a discrete-time signal. A discrete mother wavelet has the form [12]:

$$g_{m,n}(t) = \frac{1}{\sqrt{a_0{}^m}}\, g\left(\frac{t - n b_0 a_0{}^m}{a_0{}^m}\right) \qquad (2)$$

Where m and n are integers that control the wavelet dilation and translation respectively; a_0 is a specified fined dilation step greater than 1; and b_0 is the location parameter and must be greater than zero. The most common and simplest choice for parameters are $a_0 = 2$ and $b_0 = 1$.

This power-of-two logarithmic scaling of the translation and dilation is known as the dyadic grid arrangement. The dyadic wavelet can be written in more compact notation as [12]:

$$g_{m,n} = 2^{-m/2}\, g(2^{-m} i - n) \qquad (3)$$

For a discrete time series, x_i, the dyadic wavelet transform becomes [13]:

$$T_{m,n} = 2^{-m/2} \sum_{i=0}^{N-1} g(2^{-m} i - n) x_i \qquad (4)$$

Where $T_{m,n}$ is wavelet coefficient for the discrete wavelet of scale a = 2^m and location b = 2^m n. Equation 4 considers a finite time series, x_i, i = 0, 1, 2, . . ., N − 1; and N is an integer power of 2 so that N = 2^M. This gives the ranges of m and n as, respectively, 0 < n < 2^{M-m} − 1 and 1 < m < M.

The inverse discrete transform is given by [13]:

$$x_i = \overline{T} + \sum_{m=1}^{M} \sum_{n=0}^{2^{M-m}-1} T_{m,n} 2^{-m/2} g(2^{-m} i - n) \qquad (5)$$

Or in a simple format as [13]:

$$x_i = \overline{T} + \sum_{m=1}^{M} W_m(t) \qquad (6)$$

Which $\overline{T}$ *(t)* is called approximation sub-series at level *M* and W_m *(t)* are details sub-series at levels *m* = 1, 2, . . ., *M*.

The wavelet coefficients, W_m *(t)* (*m* = 1, 2, . . .,*M*), provide the detail signals, which can capture small features of interpretational value in the data; the residual term, *T (t)*, represents the background information of data.

3 Self-Organizing Map

The self-organizing map (SOM) is an effective software tool for the visualization of high-dimensional data. It implements an orderly mapping of a high-dimensional distribution onto a regular low-dimensional grid. Thereby, it is able to convert complex, nonlinear statistical relationships between high-dimensional data items into simple geometric relationships on a low-dimensional display while preserving the topology structure of the data [6]. The way SOMs go about reducing dimensions is by producing a map of usually 1 or 2 dimensions which plot the similarities of the data by grouping similar data items together. Thus, SOMs accomplish two things; they reduce dimensions and display similarities. The basic SOM network consists of two layers, an input layer and a Kohonen layer. The input layer is fully connected to the Kohonen layer, which in most common applications is two-dimensional. A two-level SOM neural network is a better approach to catch a preliminary overview on intricate data set. It augments the conventional SOM network with an additional one-dimensional Kohonen layer in which each neuron is connected to neurons in the previous Kohonen layer. The schematic view of the two-level SOM network is shown in Fig. 1.

The SOM is trained iteratively: Initially the weights are randomly assigned. When the n-dimensional input vector x is sent through the network, the distance between the weight w neurons of SOM and the inputs is computed. The most common criterion to compute the distance is Euclidean distance [6].

$$\|x - w\| = \sqrt{\sum_{i=1}^{n} (x_i - w_i)^2} \qquad (7)$$

The weight with the closest match to the presented input pattern is called winner neuron or Best Matching Unit (BMU). The BMU and its neighbouring neurons are allowed to learn by changing the weights at each training iteration t, in a manner to further reduce the distance between the weights and the input vector [6]:

$$w(t+1) = w(t) + \alpha(t) h_{lm} (x - w(t)) \qquad (8)$$

Where α is the learning rate, ranging in [0 1], l and m are the positions of the winning neuron and its neighbouring output nodes and h_{lm} is the neighbourhood function. The most commonly used neighbourhood function is the Gaussian [6]:

$$h_{lm} = \exp\left(-\frac{\|l-m\|^2}{2\sigma(t)^2}\right) \qquad (9)$$

Where h_{lm} is the neighbourhood function of the best matching neuron l at iteration t; and l-m is the distance between neurons l and m on the map grid; and σ is the width of the topological neighbourhood. The training steps are repeated until convergence. After the SOM network is constructed, the homogeneous regions i.e., clusters, is defined on the map.

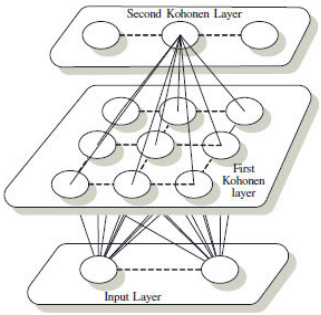


Fig.1 Architecture of the two-level SOM neural network [14]

To evaluate the performance of clustering results produced by the SOM neural network, the Silhouette Coefficients are used as the measure of

cluster validity [14]. The Silhouette Coefficient of a member can indicate the degree of similarity of a station within the cluster, which is defined as [14]:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{10}$$

where S(i) is the silhouette of station I; a(i), measured as a Euclidean distance, is the average dissimilarity of cluster i to all other stations in cluster A; and b(i) is the least average dissimilarity of station i to the stations within a cluster different from cluster A. Thus, a smaller S(i) value indicates a better similarity among stations within the same cluster. The overall quality of a clustering distribution can then be measured using the average silhouette width for the entire data set, which is defined as [14]:

$$SC = \frac{1}{n}\sum_{i=1}^{n} S(i) \tag{11}$$

where n is the total number of stations in the data set. A higher value of SC indicates better discrimination among clusters of a mining result.

4 Study Area

The Delaney Creek Sub-basin in Tampa Bay Watershed at Florida State located between 27°52□ and 27°56□ North latitude and 82°22□ and 82°24□ West longitude and its drainage area is about 42 square kilo-meters of open water which drains to Tampa Bay on the Gulf of Mexico (Fig .2).

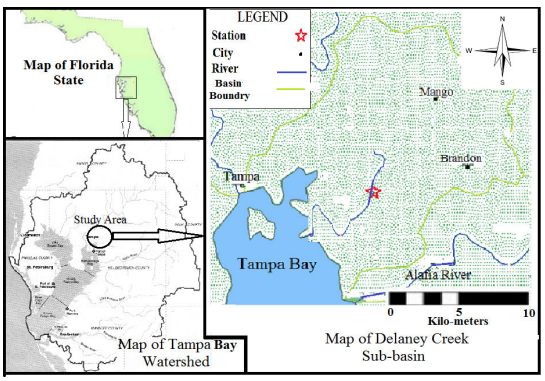


Fig. 2 Study area

This watershed is fairly flat and its elevation varies between 10 meters above and below sea level. The climate of this region is subtropical, exhibiting a transitional pattern from continental to tropical Caribbean. Long, warm and humid summers are typical as well as mild, dry winters. The annual average temperature and the total yearly rainfall are about 23°C and 1350 mm, respectively. The observed daily stream flow and rainfall time series of Delaney Creek Station, where is the outlet of Delaney Creek Watershed, were used in this research. Time series are included 6403 days data observed from August 1993 up to December 2011.

5 Results and Discussion

Data pre-processing, such as extracting dominant features of a signal, plays a crucial role in upgrading the model accuracy. Therefor in this paper a combination of wavelet transform and SOM were used to determine dominant sub-series of a signal.

Firstly, the wavelet transform was used to decompose the rainfall and runoff time series at level 3 into four sub-series (one approximation and three details). The following formula which relates the appropriate decomposition level to the number of time series data was the cause to choose level 3 as the suitable level of decomposition in this study [15]:

$$L = \text{int}[\log(N)] \tag{12}$$

Where L indicates decomposition level and N refers to the number of time series data, which is 6403 in this case study.

Due to proportional relationship between amount of rainfall and runoff, they are supposed to have the same seasonality levels. Therefore, both time series were decomposed at same level. Since the Haar wavelet [16] is the one with pulsed shape, it can properly capture the signal features of rainfall time series and yields comparatively high efficiency where the Daubechies-4 wavelet (db4) [16] according to its high-frequency shape is the appropriate mother wavelet for decomposing the runoff time series.

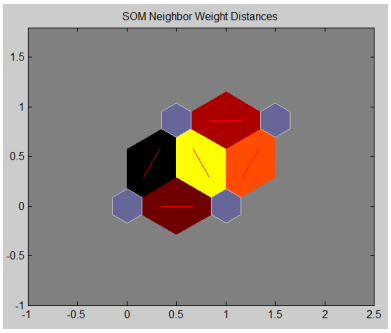
As SOM compresses information while preserving the most important topological and metric relationships of the primary data items on the display, it is an effective tool to extract dominant

features of a series. For this purpose, a 2-step SOM clustering method was employed to select the effective sub-series and reduce the dimensionality of the input space. At first, a 2-dimensional SOM was applied to have an overview on signal patterns and approximate number of clusters with regard to the SOM topology. Subsequently, in order to be ensured of the highlighted clusters, a 1-dimensional SOM was applied to classify the signals with specific numbers determined at the first step. Afterward, the Euclidean distance criterion [1] was utilized to select the centroid signal of each cluster which is the best representative of the data pattern within the cluster.

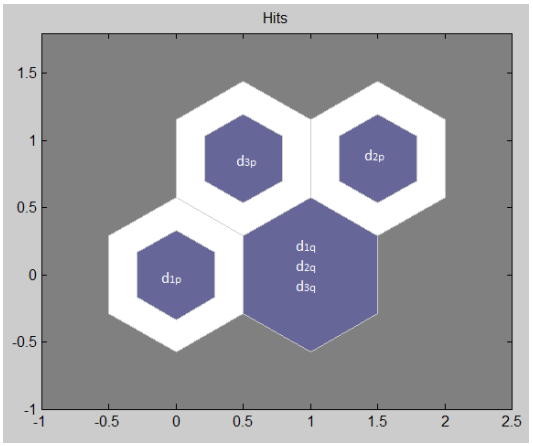
Since approximation sub-signals carry on important features of the main signals they are already supposed to be dominant inputs in the process modeling. Detailed sub-signals were imposed to SOM in order to extract dominant details which can have significant role in attaining accurate model results. Whereas the level 3 was used to decompose main time series, output layers of size 3×3, 2×2, 1×3, 1×2 were tried to catch the best pattern of clustering. Fig. 3a represents the resulted neighbor weight distances of the size 2×2. The neighbour weight distances' plan presents output neurons and their direct neighbor relationship. The regular hexagonal display the SOM output neurons while the stretched hexagonal indicate the distances between neurons. The darker colors demonstrate larger distances, and lighter colors refer to smaller distances. Fig. 3b represents the resulted hits plans of the size 2×2. The hits plan is an illustration of a SOM output layer, with each neuron showing the number of input vectors that it classifies. The relative number of vectors for each neuron is shown via the size of a colored patch. In order to easier interpretation of the result, in this paper the title of input vectors is placed instead of hits number in Fig3b.

Table 1 summarizes results obtained from clustering via SOM. Rainfall sub-series according to high stochastic features that they include do not have a specific relationship and are classified in different clusters. Since runoff subseries, in compared with rainfall sub-series, follow of an autoregressive pattern they were arranged in a same cluster.

Eventually, considering the Average Silhouette width criterion, division of the sub-series into three clusters was identified to be the best clustering pattern for this set of data. The centroid of each cluster due to Euclidean distance criterion was assigned as the representative of each cluster. Consequently, two approximation sub-signals, rainfall detail sub-signals at first and second level and runoff detail sub-signal at first level can efficiently constitute dominant model inputs in this case of study.



(a)



(b)

Fig. 3 (a): neighbor weight distances (b): hits plan

Table1 SOM results

Number of clusters	Details in each cluster			Center of each cluster			Average Silhouette width
	Cluster 1	Cluster 2	Cluster3	Cluster 1	Cluster 2	Cluster 3	
3	d_{3p},d_{1q},d_{2q},d_{3q}	d_{2p}	d_{1p}	d_{1q}	d_{2p}	d_{1p}	0.74
2	d_{2p},d_{3p},d_{1q},d_{2q},d_{3q}	d_{1p}	-	d_{1q}	d_{1p}	-	0.71

ISBN: 978-1-61804-092-3

6 Conclusion

In this study, the wavelet transform, which can capture the multi-scale of a signal, was used to decompose the Delaney Creek rainfall and runoff time series into different features. Rainfall time series was decomposed via Haar mother wavelet where runoff time series was decomposed via db4 mother wavelet, according to similarity of the main series shape to mother wavelets form, each into four sub-series. Then detail sub-series were imposed to SOM to be clustered into different categories and centroid of each cluster, based on the Euclidean distance criterion, was selected as the representative of each cluster. Obtained results of the SOM represent that third level detail of rainfall time series can be classified in the same cluster with runoff time series. Eventually two approximation sub-signals, rainfall detail sub-signals at first and second level and runoff detail sub-signal at first level were introduced as the dominant features of these time series.

In order to complete the current study, it is suggested to examine the efficiency of the presented methodology by developing a rainfall-runoff modeling via ANN.

Acknowledgments

The first author of the paper has been invited as plenary speaker of the conference. The kind invitation of the WSEAS is appreciated.

References:

[1] G.J. Bowden, G.C. Dandy, H.R. Maier, Input determination for neural network models in water resources applications, *Journal of Hydrology*, Vol.301, 2005, pp. 75-92.

[2] R. Remesan, M.A. Shamim, D. Han, J. Mathew, Runoff prediction using an integrated hybrid modelling scheme, *Journal of Hydrology*, Vol.372, No. 1-4, 2009, pp. 48-60.

[3] M.K. Tiwari, Ch. Chatterjee, Development of an accurate and reliable hourly flood forecasting model using wavelet-bootstrap-ANN (WBANN) hybrid approach, *Journal of Hydrology*, Vol.394, No. 3-4, 2010, pp. 458-470.

[4] V. Nourani, O. Kisi, M. Komasi, Two hybrid Artificial Intelligence approachs for modeling rainfall-runoff process, *Journal of Hydrology*, Vol.402, No.1-2, 2011, pp.41-59.

[5] A.M. Kalteh, P. Hjorth, R. Berndtsson, Review of the self-organizing map (SOM) approach in water resources: Analysis, modeling and application, *Environmental Modeling & software*, Vol.23, 2008, pp. 835-845.

[6] T. Kohonen, The self-organizing map, *Neurocomputing*, Vol.21, 1998, pp.1-6.

[7] G.J. Bowden, H.R. Maier, G.C. Dandy, Optimal division of data for neural network models in water resources applications, *Water Resources Research*, Vol.38, 2002, pp.1010-1011.

[8] G.F. Lin, L.H. Chen, Identification of homogeneous regions for regional frequency analysis using the self-organizing map, *Journal of Hydrology*, Vol.324, 2006, pp. 1-9.

[9] G.F. Lin, M.C. Wu, A SOM-based approach to estimate design hyetographs of ungauged sites, *Journal of Hydrology*, Vol.339, 2007, pp. 216-226.

[10] E. Foufoula-Georgiou, P. Kumar, *Wavelet in geophysics*, Academic Press, New York, 1995.

[11] D. Labat, R. Ababou, A. Mangin, Rainfall-runoff relation for karstic spring. Part 2: continuous wavelet and discrete orthogonal multi resolution analysis, *Journal of Hydrology*, Vol.238, pp. 149-178.

[12] P.S. Addison, K.B. Murrary, J.N. Watson, Wavelet transfom analysis of open channel wake flows, *Journal of Engineering Mechanics*, Vol.127, No. 1, 2001, pp. 58–70.

[13] V. Nourani, M.T. Alami, M.H. Aminfar, A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation, *Engineering Applications of Artificial Intelligence*, Vol.22, 2009b, pp.466-472.

[14] K.C. Hsu, S.T. Li, Clustering spatial-temporal precipitation data using wavelet transform and self-organizing map neural network, *Advances in Water Resources*, Vol.33, 2010, pp. 190-200.

[15] V. Nourani, M. Komasi, A. Mano, A multivariate ANN-Wavelet approach for rainfall- runoff modelling, *Water Resource Management*, Vol.23, 2009a, pp.2877-2894.

[16] S.G. Mallat, *A wavelet tour of signal processing*. Academic Press, San Diego, 1998.