

Hydrologic Modelling and Analysis Using A Self-Organizing Linear Output Network

K. Hsu, S. Sorooshian, H.V. Gupta, X. Gao, and B. Imam

*The University of Arizona, Tucson, AZ 85715, USA
(hsu@hwr.arizona.edu)*

Abstract: Artificial neural networks (ANNs) have been broadly applied to many hydrological applications for which their underlying processes are complicated nonlinear. Although many networks, such as multi-layer feedforward neural networks (MFNs), provide excellent capability in function fittings, very often, they are referred to as black-box models. In this study, a multivariate ANN procedure, entitled SOLO (Self-Organizing Linear Output mapping network) is introduced. This model architecture has been designed for rapid estimation of network structure/parameters and system outputs. Furthermore, the SOLO provides features that facilitate insight to the input-output processes, thereby extending its usefulness as a tool for investigations into the underlying processes through the data classification processes. A case study using SOLO model in a hydrologic rainfall-runoff forecasting is demonstrated. Uncertainty of model estimates is also evaluated.

Keywords: Self-organizing feature map; Principal component regression; rainfall-runoff process

1. INTRODUCTION

ANN models are popular from their flexible structures that learn the system behavior from data without priori information, and hence provide a cost-effective means in the complex system development and simulation. Many hydrological applications of ANNs were published in many conference proceedings and journal articles during 1990s. [Hsu et al., 1995, 1997a, 1997b, 1999; Mason, 1996; Minns and Hall, 1996; Smith and Eli, 1995; Tokar and Johnson, 1999; Sorooshian et al., 2000] For those ANN applications to the hydrologic and water resources systems, summarized articles are available from Maier and Dandy (2000) and published article from the ASCE task committee on application of Artificial Neural Networks in hydrology (2000).

In this study, a hybrid ANN model, named Self-Organizing Linear Output network (SOLO), is introduced. This network architecture includes a data clustering procedure (Kohonen 1989) and a group of linear mapping functions connecting in between the input and output variables of classified data. Some features that provided by the SOLO model include: 1) efficient and effective in model calibration, 2) explanation of the input/output relationship through the data classification map, and 3) uncertainty assessment of model estimates.

In the case study, the SOLO network was applied to the daily streamflow prediction. The performance of the model was evaluated using testing data consisted of 36 years (10/1/1948~9/30/1983) of daily rainfall and streamflow data for the Leaf River basin (1949 km²) near Collins, Mississippi.

2. THE SOLO NETWORKS

The SOLO network consists of three layers (see Figure 1). The input layer includes n_0 units connecting to the input variables. The classification and mapping layers consist of $n_1 \times n_1$ matrixes. The input data are classified into $n_1 \times n_1$ clusters using a Self-Organizing Feature Map (SOFM) (Kohonen, 1989). In the mapping layer, model outputs are generated from multivariate linear regression. Let w_{ji} represent the connection strength (weight, parameter) linking the i^{th} input variable ($i = 1, \dots, n_0$) to the j^{th} SOFM unit ($j \in n_1 \times n_1$), and let v_{ji} represent the connection strength linking the same (i^{th}) input variable to the j^{th} regression unit. Compute the Euclidian "distance" d_j between the input vector, $x = \{x_i, i = 1, \dots, n_0\}$, and the j^{th} SOFM unit as follows:

$$d_j = \left[\sum_{i=1}^{n_0} (x_i - w_{ji})^2 \right]^{0.5} \quad (1)$$

The classification unit, c , on the SOFM classification layer is selected according to the distance measurement, where $d_c = \min(d_j)$, for all

j . The output, z , is then determined by using a linear regression function of input vector (x) associated with the selected classification unit, as shown below:

$$z = \sum_{i=1}^{n_0} v_{ji}x_i + v_{j0} \quad \text{if } j = c \quad (2)$$

$$= \phi \quad \text{otherwise}$$

Finding the network connection weights w_{ji} is based on a non-supervised (self-organizing) data clustering procedure. The weight matrix, w_{ji} , are initialized randomly and are then sequentially adjusted as shown below:

$$w_{ji}(k) = w_{ji}(k-1) + \eta(k)[x_j - w_{ji}(k-1)], j \in A_c(k)$$

$$w_{ji}(k) = w_{ji}(k-1) \quad \text{otherwise} \quad (3)$$

Where, k is the training iteration, $A_c(k)$ is the size of a neighborhood around the winner unit c , and $\eta(k)$ is the learning rate at the iteration k . The sizes of both $A_c(k)$ and $\eta(k)$ are progressively reduced during the training iteration. Training is stopped when the training weights, w_{ji} , are stabilized.

The connection weights of the linear regression matrix, v_{ji} , are determined from a least square error solution of the linear function. The linear regression function with respect to the classification unit j is determined below:

$$Z = X\theta + \varepsilon \quad (4)$$

where Z is a $mx1$ vector with m output data; X is a mxn matrix with m sets (rows) of input vectors (x_i)^T, $i=1, \dots, n$; θ is a $nx1$ vector of regression parameters for unit j , $\theta = [v_{j0}, v_{j1}, v_{j2}, \dots, v_{jn}]$ ^T; and ε is a $mx1$ vector of estimation errors with zero mean and variance σ_e^2 . Regression parameters, θ , are determined by minimizing the root mean square error of the output residuals:

$$\hat{\theta} = (X^T X)^{-1} X^T Z \quad (5)$$

From hydrologic time series, the selected variables, such as the time delay sequences of precipitation, surface runoff, base flow, soil moisture, air and land surface temperature, latent/sensible heat fluxes, and longwave/shortwave radiation fluxes, in the regression function can be mutually correlated. Under the extreme situation, if part of the input variables is collinear to each other, a singular matrix of $(X^T X)^{-1}$ is obtained, which would generate high uncertainty of the model parameters and estimates. To avoid from obtaining highly correlated variables in the multivariate input variables, an orthogonal transformation using principal component analysis (PCA) is applied to obtain a matrix Y having independent (orthogonal) column vectors:

$$Y = XC \quad (6)$$

Where, Y is the mxn matrix of principal components, and C is the mxn transformation matrix with eigenvectors derived from the covariance matrix of X . From Equation (4) and (6), we have:

$$Z = YC^T\theta + \varepsilon = Y\beta + \varepsilon \quad (7)$$

where $\beta = C^T\theta$ and can be determined from $\hat{\beta} = (Y^T Y)^{-1} Y^T Z$. The number (p) of the largest principal components is determined from the ratio $V = \sum_{i=1}^p \lambda_i / \sum_{j=1}^n \lambda_j \cdot 100\% > 95\%$.

The expected value of the estimate \hat{z} is $y\hat{\beta}$, and the variance of \hat{z} is $\sigma^2 y^T (Y^T Y)^{-1} y$. If the estimated error, ε , comes from a normal distribution, the upper and lower bounds (U_α, L_α) of the model output predictions corresponding to a $100(1-\alpha)$ confidence range are given below:

$$U_\alpha = y\hat{\beta} + t_{1-\alpha/2, m-p} \sigma \sqrt{\gamma} \quad (8)$$

$$L_\alpha = y\hat{\beta} - t_{1-\alpha/2, m-p} \sigma \sqrt{\gamma}$$

where $t_{1-\alpha/2, m-p}$ is a t distribution with $m-p$ degrees of freedom; m is the size of data, and p is the rank of Y .

3. RAINFALL-RUNOFF MODELING

In the case study, the SOLO network was applied to the daily streamflow prediction of a watershed. The test data consisted of 36 years (10/1/1948~9/30/1983) of daily rainfall and streamflow data for the Leaf River basin (1949 km²) near Collins, Mississippi. The first 11 years of data was used for model development and calibration, and the remaining 25 years were used for performance evaluation. The input variables were selected from three-day time delayed area averaged rainfall and streamflow as shown below:

$$x = [r(t), r(t-\Delta t), r(t-2\Delta t), q(t), q(t-\Delta t), q(t-2\Delta t)]^T$$

$$= [x_1, x_2, x_3, x_4, x_5, x_6]^T$$

The model estimates is $q(t+\Delta t)$, and Δt is 1 day. The SOLO model uses data classification and piece-wise linear regression functions to predict the streamflow. The classification layer of the SOLO is the SOFM network, which consists of 15 x 15 nodes. A linear regression function is fit to the data included in each classified data nodes. The selection of network architecture is not described in details here. Further discussions of model selection are described in Hsu et al. (2002). Performance of the model is evaluated based on the ability of the model to provide accurate one-day predictions of streamflow.

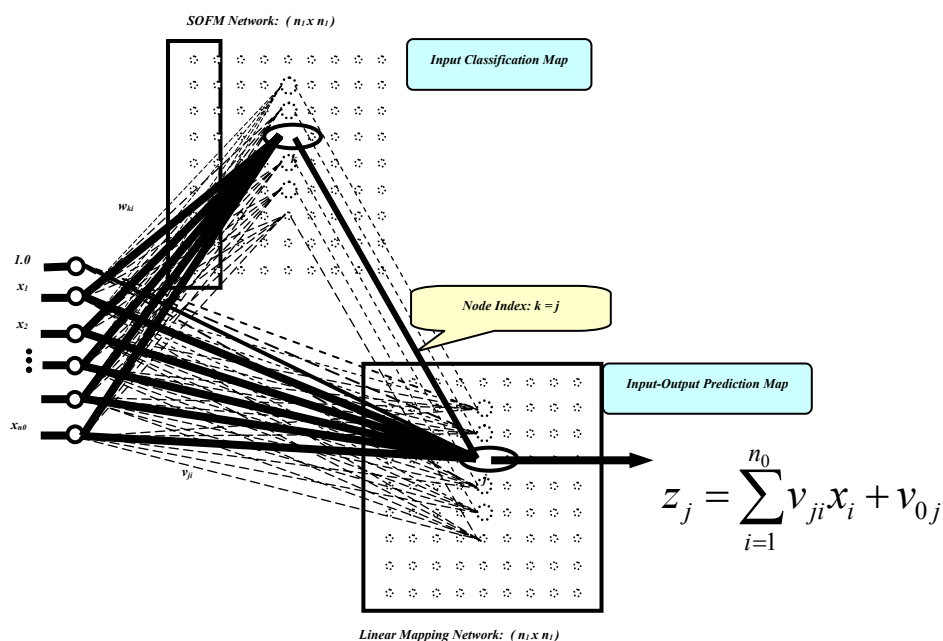


Figure 1. The architecture of a SOLO model.

Figure 2 shows the model performance based on the daily root mean square error (RMSE) (cms) plotted against volume of annual streamflow (cms). The solid squares represent calibration years, while the solid circles represent the evaluation years. It shows that the error variance increases with “wetness” of the year (i.e., the annual RMSE increases with annual streamflow). Daily rainfall time series and one-day-ahead model predictions comparing with streamflow observations for five consecutive water years (1978-1982) of the evaluation period are shown in Figure 3 and Figure 4. These Figures show that water year 1979 and 1980 are wetter than the other water years. As shown in Figure 4, the SOLO model daily flow estimates match the observed flows on all portions of the hydrograph. For these five years period, the RMSE is 18.87 cms, correlation coefficient (CORR) is 0.96, and bias (BIAS) estimate is -0.4 cms.

Figure 5 and Figure 6 show the observed stream flow and uncertainty of predicted streamflow hydrographs covering the 95% upper and lower confidence bounds for an evaluation year (water year 1980). This particular year is the wettest water year in the evaluation period having annual flow more than 65 cms. The SOLO model uses a total of 225 classification groups and linear regression functions classify and mapping various input and output hydrologic behavior. It shows that high flow regions contain the model

prediction with higher uncertainty, whereas the uncertainty bounds over the low and medium flow regions are tighter and smaller.

Figure 7 shows the SOFM network connection weights with respect to the six input variables. This SOFM classification map reveals the underlying properties of the input-output process. The vertical bars are normalized connection weights of rainfall inputs $\{r(t-2), r(t-1), r(t)\}$ and three line-connected \square symbols (representing the three streamflow inputs $\{q(t-2), q(t-1), q(t)\}$). The distribution of rainfall-runoff modes which can be identified as several classification regions described below:

- (1) *Baseflow Region (Region I)*: The behavior is characterized by no-rain and low-level, almost unchanging streamflows during a 3-day period. The corresponding streamflow prediction associated with a region is very small.
- (2) *Increasing Rainfall Region (Region II)*: Rainfall is steadily increasing during the 3-day period, but streamflows have only just begun to respond. The model predicts high streamflow levels during the next period; this region is identified as the initial stages of a storm event and is associated rising limb of the hydrograph.
- (3) *Peaking Hydrograph Region (Region III)*: The rainfall has peaked, but that the

