FORECASTING OF DAILY RUNOFF USING ARTIFICIAL NEURAL NETWORKS

Santosh Patil
Shreemati Kashibai Navale College of Engineering, Vadgoan (Bk), Sinhagad Road,
Pune University, Pune, INDIA

Shriniwas Valunjkar
Government College of Engineering, Karad, Shivaji University, Kolhapur,
Maharashtra, INDIA

ABSTRACT

The modeling of hydrological processes is important in view of the many uses of water resources. The use of artificial neural networks (ANN) is becoming increasingly common in the analysis of hydrology and water resources problems. In the present work, ANN is applied for Gunjwani watershed in lower Bhima sub-basin (Maharashtra, India) to forecast next-day runoff. The ANN architecture, namely multi-layer perceptron (MLP) was adopted. Daily rainfall, runoff, evaporation, humidity, temperature data (full year as well as seasonal) for 7 year 10 months were used for model development. Combinations of the different input data series were studied using correlation between rainfall and runoff. The number of statistical parameters i.e. mean square error (MSE), mean absolute error (MAE) and coefficient of correlation (R) were used for performance evaluation of the developed models. Further, the results of the developed ANN models were compared with multiple linear regression models (MLR). It was found that the MLP model showed a better forecasting of runoff as compared to MLR models.

Key words: Multiple Linear Regression; Artificial Neural Networks; Rainfall-Runoff; Lower Bhima, India.

1. INTRODUCTION

Assessment of correct water resources is a pre-requisite for the successful planning, execution and operation of project. Water is one of the essential commodities, which is available cheaply as a natural resource. This resource is random in nature, rare and become costly sometimes.
It is necessary that, availability of water resources for the schemes be reviewed from time to time, as more and more historical hydrological data becomes available. The appraisal of hydrological data and precisely arriving at the availability of water resources is a science, at the same time; its utilization for proper planning is an art. Generally, the hydrological data available is of short duration. Using more advanced methodologies it is possible to design the hydro-electric and irrigation schemes successfully. Reappraisal of the project will definitely indicate the change required in working system of the scheme. Accordingly, the monitoring operation of reservoirs could be safely and suitably modified. Most of the research on rainfall-runoff models is involved in developing mathematical models, which not only quantify the watershed-scale hydrologic process but also understand the hidden knowledge from that quantification. According to explanation of governing process, these models are classified into two types-physical based models and system based theoretical models.

The physical based model consisted of detailed explanation of different physical processes where these processes are used to control hydrologic behavior of system. On the other hand system based theoretical models do not consider the parameters used in physical models. ANN is used to solve all types of complex non-linear problems. ANN achieves maximum reliabilities in describing rainfall-runoff processes.

ANN is considered as suitable tool in order to establish the relationships between rainfalls and runoffs. Another advantage of ANN is that it always gives better performance than distributed parameter models when the less data is required and long term forecasting is needed.

Hapuarachchi, H.A.P et.al. [1] modeled the rainfall-runoff process using ANN and found it for the problems which are difficult to describe the processes based on physical equations. They described the procedure of linear least square and simplex optimization for identifying the structure and parameters of three-layer feed forward ANN models. Smith and Eli [2] applied a back propagation (BP) neural network model to predict peak discharge and time to peak over a hypothetical watershed. Data sets for training and validation were generated by either a linear or non-linear reservoir model. By representing the watershed as a grid of cells, it was possible to incorporate the spatial and temporal distribution information of rainfall in to ANN model. Minns and Hall [3] pointed out the importance of standardization based maximum and minimum values of inputs and outputs. Kumar et.al [4] and, Cigizoglu [5] applied the radial based function (RBF) type ANN using orthogonal least squares (OLS) algorithm to model the rainfall runoff process and compared with the multi layered perceptron (MLP) ANN using BP algorithm. Thandaveswara, and Sajikumar [6] compared the temporal back propagation neural network model for evaluating and performance of calibration of small data length. Elshobagy et.al. [7] compared ANN technique with linear and non-linear regression techniques. The feed forward back propagation algorithm was considered for the network for the evaluation of combined effect of mean square error and mean regression error called pooled mean square error. Zhang and Govindraju [8] used the modular neural network (MNN) for prediction of catchment runoff, and utilized Bayesian concepts in deriving the training algorithm. The performance of the MNN showed improved results compared with the standard ANN.

The most commonly used ANN methods for modeling rainfall-runoff processes is the multi layer perceptron (MLP).

2. MULTIPLE LINEAR REGRESSION AND ARTIFICIAL NEURAL NETWORK METHODS

Artificial neural networks are flexible mathematical structures that are capable of identifying complex non-linear relationships between input and output data sets. A neural network consists of a large number of simple processing elements that are called neurons. Each neuron is connected to other neurons by means of direct communication links, each with an associated weight that
represents information being used by the net to solve a problem. The network usually has two or more layers of processing units where each processing unit in each layer is connected to all processing units in the adjacent layers.

2.1 Multiple Linear Regression Model

A polynomial curve for two variables (i.e. rainfall and runoff) is fitted by correlation and regression method. Polynomial Regression [9] an algorithm is developed to derive the first and higher order equations using least-square criterion. A strategy for fitting a best curve of \( m \)\(^{th} \) degree through the data is to minimize the sum of the square of residual errors.

\[
y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \ldots \ldots a_m x^m + e
\]

where \( e \) = residual or error between the model and observation.

For this case; the sum of the square of the residuals (Sr) is:

\[
S_{r} = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2 - a_3 x_i^3 \ldots \ldots - a_m x_i^m)^2
\]

Taking the derivative of Equation (3) with respect to each unknown coefficient of the polynomial and equating these set of derivatives to zero, the set of normal equations is generalized as:

\[
a_0 \sum x_i^m + a_1 \sum x_i^{m+1} + a_2 \sum x_i^{m+2} + a_3 \sum x_i^{m+3} + \ldots \ldots + a_m \sum x_i^{2m} = \sum x_i^m y_i
\]  

where, all summations are from \( i = 1 \) to \( n \). These generalized set of equations are linear and are in the order of \( m, m+1, m+2, m+3 \ldots 2m \) with corresponding unknowns \( a_0, a_1, a_2, a_3, \ldots , a_n \). The unknown coefficients of first degree equation are generated from the observed data.

In the multiple linear regression models (MLR), the dependant variables is assumed to be a linear function of one or more independent variables plus an error introduced to account for all other factors. The multiple linear regression models are represented in equation as:

\[
y = m_1 x_1 + m_2 x_2 + m_3 x_3 + m_4 x_4 \ldots \ldots m_n x_n
\]

Where \( m_i \) and \( c \) are constants, \( y \) is the dependant variable and \( x_i \) is the \( i^{th} \) independent variables. The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations that are provided.

2.2 Multi-layer Perceptron Network

Multi-layer perceptron (MLP) are feed-forward networks with one or more hidden layers. Given a training set of input-output data, the most common learning rule for MLP is the back propagation algorithms. Considering networks with one hidden layer shown in Figure 1, the processing of a single neuron is broken into two steps, i.e. the weighted sum of the inputs followed by the activation function. For example, consider a neuron in the hidden layer that receives inputs from neurons in the input layer. The net input, \( y \), to the hidden neuron is the sum of the weighted signals from the input neurons (that is: \( y' = w_1 x_1 + w_2 x_2 + w_3 x_3 \ldots \ldots w_n x_n \)).
The activation $y$ of this hidden neuron is then given by some function of its net input, $y = f(y_1)$. The most common activation function and the one implemented in this study is a sigmoid function and is described as follows:

$$f(x) = \frac{1}{1 + \exp(-x)}$$

This procedure is repeated for each input vector and at the completion of a pass through the entire data set, all the nodes change their weights based on the accumulated derivatives of the error with respect to each weight and these changes move the weights in the direction in which the error declines most quickly.

The measured data employed for ANN model development are again divided into training and testing. First the ANN model is trained to characterize the associations and procedures inside the measured training data set. Once the model is satisfactorily trained, it is able to generalize appropriate output for the set of input data. This output is afterward compared with the measured testing data set. The model is considered to behave satisfactorily if and only if its performance throughout the testing period is exactly similar to that during the training period.

### 3. STUDY AREA AND DATA ANALYSIS

The present study was conducted with the help of data obtained at Gunjvani river watershed up to Velhe on river Kanand that is a tributary to Gunjvani in western ghats of Maharashtra, India and located South-West of Pune city. The watershed is covered between latitudes 18°16'N and 18°20'N, and longitudes 73°30'E and 73°40'E. Gunjvani watershed has an area of 62.95 sq.km. upto the Velhe river gauge site. Length of the river is about 14.7 km with highest peak elevation of 1403 m above mean sea level (MSL) at Torna Fort situated south of village Velhe Budruk and the headwater of the watershed is at an elevation of 1186 m above MSL.
There are four stations with standard rain gauges in the watershed. The data collected for the studies include hydrological data and maps. The hydrological data obtained from the Hydrological Project, Nashik Maharashtra, India includes daily runoff, evaporation, humidity, rainfall, and temperature for 7 years 10 months, from March 2000 to December 2007.

3.1. Model Development

To approximate the relationship between a set of inputs and outputs, it is necessary to compare the predictive capabilities of the new model with the existing approaches. Therefore multiple linear regression models are developed as an initial step. From the literature review it is observed that in most of the studies only rainfall data is used to predict the runoff. In this study, combined (meteorological data) is used to describe the physical phenomena of the rainfall-runoff process in order to forecast runoff.

In the basin there are four rain gauge stations measuring the rainfall data and one of them also measure meteorological data. The models with different structures and different inputs are presented in the Table 1.

<table>
<thead>
<tr>
<th>Sr. No</th>
<th>Models</th>
<th>Inputs</th>
<th>No. of variables</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Model 1</td>
<td>$P_t$, $H_t$</td>
<td>2</td>
<td>$Q_{t+1}$</td>
</tr>
<tr>
<td>2</td>
<td>Model 2</td>
<td>$P_t$, $E_t$</td>
<td>2</td>
<td>$Q_{t+1}$</td>
</tr>
<tr>
<td>3</td>
<td>Model 3</td>
<td>$P_t$, $T_t$</td>
<td>2</td>
<td>$Q_{t+1}$</td>
</tr>
<tr>
<td>4</td>
<td>Model 4</td>
<td>$P_t$, $T_t$, $H_t$</td>
<td>3</td>
<td>$Q_{t+1}$</td>
</tr>
<tr>
<td>5</td>
<td>Model 5</td>
<td>$P_t$, $T_t$, $E_t$</td>
<td>3</td>
<td>$Q_{t+1}$</td>
</tr>
<tr>
<td>6</td>
<td>Model 6</td>
<td>$P_t$, $H_t$, $E_t$</td>
<td>3</td>
<td>$Q_{t+1}$</td>
</tr>
<tr>
<td>7</td>
<td>Model 7</td>
<td>$P_t$, $H_t$, $E_t$, $T_t$</td>
<td>4</td>
<td>$Q_{t+1}$</td>
</tr>
<tr>
<td>8</td>
<td>Model 8</td>
<td>$Q_t$, $P_t$, $H_t$, $E_t$, $T_t$</td>
<td>5</td>
<td>$Q_{t+1}$</td>
</tr>
</tbody>
</table>

3.2 Selection of input variables

Determination of significant input variables is one of the most important steps in the model development process. All the potential input variables will not be equally useful since some may not be correlated, noisy or have no significant relationship with the output variables being modeled by Maier and Dandy [10]. Generally, some degree of priorities knowledge is used to specify the initial set of candidate inputs [11] when the relationship to be modeled is not well understood, then an analytical technique, such as cross correlation, is often employed. Sudheer [12], [13] employed a trial and error approach to identify the appropriate input vector.

3.3 Model Performance criteria

Eight models developed in the present study were evaluated using statistical performance evaluation measures. Mean square error (MSE) is selected as a measure for indicating goodness-of-fit at high output values. Correlation coefficient, $R$ is a popular global error statistics for measuring goodness-of-fit of the models and tends to give higher weight age to the high magnitude runoff due to square of the difference between observed and predicted inflows. Mean absolute error is also used as evaluation measure.
4. RESULTS AND DISCUSSION

The developed MLR and ANN models have been applied to Gunjvani watershed data to forecast the daily runoff. The recent 94 months daily data was used to develop the models since they reflect the current land use conditions in the watershed. Out of 2860 data sets 2002 data sets (70 percent) are used for training and remaining 858 data sets 30 percent are used for testing. The comparison of statistics of MLR and ANN models are shown in Table 2.

The general MLR model used in the present study can be expressed as below. Various models have been generated by varying the value of m and n in the equation. The models which resulted in better performance are only presented here.

\[ Q(t) = f \{ P(t), Q(t), E(t), H(t), T(t) \} \]

Where, \( Q(t) = \) Runoff at time step (t), \( P(t) = \) average rainfall at time step (t), \( E(t) = \) average evaporation at time step (t), \( H(t) = \) average humidity at time step (t), \( T(t) = \) average temperature at time step (t).

<table>
<thead>
<tr>
<th>Model Input</th>
<th>MLR model</th>
<th>ANN model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>MAE</td>
</tr>
<tr>
<td>( P(t), H(t) )</td>
<td>0.724</td>
<td>6.81</td>
</tr>
<tr>
<td>( P(t), E(t) )</td>
<td>0.721</td>
<td>7.03</td>
</tr>
<tr>
<td>( P(t), T(t) )</td>
<td>0.725</td>
<td>6.79</td>
</tr>
<tr>
<td>( P(t), T(t), H(t) )</td>
<td>0.653</td>
<td>6.68</td>
</tr>
<tr>
<td>( P(t), T(t), E(t) )</td>
<td>0.725</td>
<td>6.74</td>
</tr>
<tr>
<td>( P(t), H(t), E(t) )</td>
<td>0.725</td>
<td>6.72</td>
</tr>
<tr>
<td>( P(t), H(t), E(t), T(t) )</td>
<td>0.728</td>
<td>6.63</td>
</tr>
<tr>
<td>( Q(t), P(t), H(t), E(t), T(t) )</td>
<td>0.865</td>
<td>2.68</td>
</tr>
</tbody>
</table>

The model performance of daily rainfall-runoff models with different parameters varies significantly with the parameters. The performance of model with five inputs is found to be very good out of eight models with R, MAE and MSE equal to 0.879, 2.57 and 46. It is concluded from the model results that effect of evaporation, humidity, maximum temperature, rainfall and runoff at time step (t) largely affect the prediction of runoff for next time step.
The time series and scatter plot of daily rainfall-runoff model with different parameters is shown in Fig.2 for the best model. Time series plot shows that the forecasted runoff values follow the trend of observed runoff, with average performance in estimating the peak runoff values. The zero value runoff is well predicted by the MLP (5-8-1) model. From the scatter plot, it can be seen that the runoff values are well distributed about the ideal line except for the high runoff values. Thus it may be concluded that the MLR models predicted the average runoff values and normal peak flow values in a better scenario for an intermittent river.

5. CONCLUSION

This paper presents the MLP neural network type ANN model for forecasting of one step ahead daily runoff for Gunjwani watershed in Maharashtra. Popular performance measures viz. MSE, MAE, R were explored to check the accuracy of predictions. The number of neurons in hidden layer was varied for different combinations of input. Momentum learning rule and sigmoid activation function was kept same for all MLP models. The results suggest that the choice of the number of inputs with different combinations has an impact on the model prediction efficiency.

In general, ANN models, applied to the rainfall-runoff transformation problem shows encouraging results. The fact that ANNs exhibit a comparable or even better performance than a regression model suggests that this approach could provide a useful tool in solving similar type of problems in water resources studies and management. Moreover, applying ANNs to phenomena for which no adequate physically-based models can be built allows these techniques to be used in constructing hybrid models, with optimal combination of inputs and models of various types.

REFERENCES


