APPLICATION OF MULTILAYER PERCEPTRON BASED ARTIFICIAL NEURAL NETWORK FOR MODELING OF RAINFALL RUNOFF IN A HIMALAYAN WATERSHED

Sanjarambam Nirupama Chanu¹, Pravendra Kumar²

¹,² Department of Soil & Water Conservation Engineering, College of Technology, G. B. Pant University of Agriculture and Technology, Pantnagar - 263145 (U. S. Nagar) Uttarakhand, (India)

ABSTRACT

In the present study, multilayer perceptron (MLP) based neural network, which is one of the efficient artificial neural network (ANN) was applied for modeling daily rainfall-runoff in a Himalayan watershed called Bino watershed in Almora and Pauri Garhwal districts of Uttarakhand, India using the time series monsoon data of ten years (2000-2009) of rainfall and runoff. Gamma test (GT) technique was applied for selection of the best input combinations. Performance of model developed was evaluated qualitatively as well as quantitatively using indices viz. correlation coefficient (r), root mean square error (RMSE) and coefficient of efficiency (CE). The results of the study showed that model (MLP7) with 5 inputs and one hidden layer with 8 neurons was found to be the best followed by model (MLP 19) having 10 hidden neurons for first hidden layer and 11 for second hidden layer. The r, RMSE and CE values for MLP 7 during testing were determined to be 0.9207, 0.9644 (mm) and 0.7974, respectively. The result of study revealed that ANN can be successfully applied for rainfall-runoff modeling in the study area with good accuracy.

Keywords: Artificial Neural Network, Rainfall, Runoff, Perceptron, Watershed

I. INTRODUCTION

Modeling of rainfall-runoff is considered one of the prerequisite of hydrological processes for various applications involving conservation and management of water resources. Due to its non-linear, multi-dimensional and inter-relationships nature of underlying climatic and physiographic factors, it is become extremely complex to model such phenomena and it exhibits both temporal and spatial variability. So far many hydrological models have been developed for simulating such a hydrological process viz. conceptual, physically based distributed models and black box models. Artificial neural network (ANN) is such a black box artificial intelligence technique that have frequently applied in last few decades in a variety of areas because to its great flexibility with highly adaptive nature in modeling the non-linear processes overcoming the limitations posed by conceptual and physical based models. So far number of studies have been conducted by many researchers around the world for modeling hydrological phenomena using ANN [1-9]. Above studies explained the
capability of ANN superior than the conventional models without requiring an explicit description of the complex nature of the underlying process in a mathematical form. Sudheer et al. stated that this is one of the main advantages of the ANN approach over traditional methods [10]. Tokar and Markus applied ANN for rainfall-runoff modeling and demonstrated the impact of the training data selection on the accuracy of runoff prediction[11]. Sarangi et al. developed ANN and regression models using watershed-scale geomorphologic parameters to predict surface runoff and sediment losses of the St. Esprit watershed, Quebec, Canada [12]. Yadav et al. investigated the applicability of ANN for inflow forecasting of a salty lake known as Flat bay, situated on the northern side of Port Blair in Andaman and Nicobar Islands, India [13]. Zounemat-Kermani et al. investigated the effects of upstream stations’ flow records on the performance of two different artificial neural network (ANN) models, multi layer feed-forward neural network using Levenberg–Marquardt learning algorithm (LMFF) and radial basis function (RBF) models for predicting daily watershed runoff [14]. Patil and Valunjkar applied multi-layer perceptron (MLP) for Gunjwani water in lower Bhima sub-basin (Maharashtra, India) to forecast next-day runoff and compared results with MLR [15]. Nemati et al. compared the performance of MLP with two kinds of statistical neural networks, radial based function (RBF) and general regression neural network (GRNN), in rainfall-runoff simulation [16]. Phukoetphim et al. compared the performance of a symbolic regression combination method based on gene expression programming (GEP) with two different neural network combination methods, the multilayer perceptron neural network (MLPNN) and the radial basis function neural network (RBFNN), for development of multimodel systems [17].

Moreover, one of the major phases in modeling using artificial intelligence techniques is identifying the best input combination of the network [18]. There are different methods for reducing the number of input variables such as principal component analysis (PCA), Gamma test (GT), forward selection (FS) and other techniques. The GT was firstly reported by Stefansson et al. [19], Koncar [20] and Agalbjörn et al [21]. Noori et al. explored the role of pre-processing of input parameters using Principal Component Analysis (PCA) techniques, GT and Forward Selection (FS) techniques to assess the performance of the support vector machine (SVM) model for monthly stream flow prediction and authors recommended to use the PCA and GT techniques for increasing the SVM model performance especially in cases where lack of knowledge about the input variables exists [22]. Noori et al. stated that the use of the GT in input variable pre-processing is new and there are only a few studies involving the application of this method to water resources management [22]. Maier and Dandy mentioned that determining of adequate model inputs and development of suitable network architecture are key aspects requiring further attention [23]. In development of nonlinear simulation models, proper selection of input variables is a challenging task because a false combination of input variables could prevent the model from achieving the optimal solution.

Keeping above points in view, the present study was taken up for modeling of daily runoff applying multilayer perceptron neural network (MLPNN) using GT as the best input combination selection technique in Bino watershed, India.
II. MATERIALS AND METHODS

2.1 Study area and data

The Bino watershed with a drainage area of 296.366 Km$^2$ is situated in North-Eastern part of Ramganga catchment in middle and outer ranges of Himalayas between 79° 6’ 14.4” E to 79° 17’ 16.8” E longitude and 29° 47’ 6” N to 30° 02’9.6” N latitude in Almora and Pauri Garhwal districts of Uttarakhand, India (Fig. 1). The watershed has very undulating topography with mean length of 28.46 Km and 17.27 Km and irregular slopes varying from moderate to steep in valley areas on either sides of the Bino River. The climate of the watershed varies from Himalayan sub-tropical to sub-temperate with mean annual maximum and minimum air temperature of 30°C to 18°C, respectively. The daily mean temperature remains higher during the months of May and June and minimum in December and January. Based on the rainfall data for the years 2000 to 2009, the mean annual rainfall in the area is 687.53 (mm). Daily rainfall, runoff data of 10 years (2000-2009) were collected from Divisional Forest and Soil Conservation Office, Ranikhet, Uttarakhand, India.

![Fig.1 Location map of Bino watershed](image-url)
2.2 Gamma test (GT)

Gamma test is one of the non-linear modeling and analysis tools that can investigate an underlying input-output relationship in a numerical data set as well as establishing a smooth model. GT estimates the minimum mean square error (MSE) that can be achieved when modeling the unseen data using any continuous non-linear models [24]. Suppose there exists a set of data observations as \( \{(x_i, y_i), 1 \leq i \leq M\} \) where the input vectors \( x_i \in \mathbb{R}^m \) are \( m \) dimensional vectors (with a record length of \( M \)) confined to some closed bounded set \( C \in \mathbb{R}^m \) and \( y_i \in \mathbb{R} \) is corresponding outputs scalar. If the underlying relationship between input-output can be expressed as:

\[
y = f(x_1, \ldots, x_m) + r
\]

(1)

where \( f \) is a smooth unknown function and \( r \) is a random variable representing noise. GT allows the variance of the noise variable \( r \) (\( \text{Var}(r) \)) to be estimated, despite the fact that \( f \) is unknown. GT calculates model output variance that cannot be accounted by a smooth data model called Gamma statistic \( (\varrho) \). GT is based on the \( k \)th \((1 \leq k \leq p)\) nearest neighbors \( x_{N[k]} \) for each vector \( x_i \) \((1 \leq i \leq M)\) and \( p \) is the number of near neighbors, typically \( p = 10 \) [25]. It can be derived from Delta function of the input vectors which calculates the mean squared distance of the \( k \)th neighbor:

\[
\delta_{k}(k) = \frac{1}{2M} \sum_{i=1}^{M} \| x_{N[k]} - x_i \|^2 ; \quad (1 \leq k \leq p)
\]

(2)

Where \( \| \ldots \| \) denotes Euclidean distance, and corresponding Gamma function output is given as:

\[
\gamma_{k}(k) = \frac{1}{2M} \sum_{i=1}^{M} \| y_{N[k]} - y_i \|^2 ; \quad (1 \leq k \leq p)
\]

(3)

where \( y_{N[k]} \) is the corresponding \( y \)-value for the \( k \)th nearest neighbor of \( x_i \) in Eq. (2). To compute \( \varrho \), a least squares regression line which is fitted for \( p \) points \((\delta_{M}(k), \gamma_{M}(k))\) as:

\[
\gamma = \delta + \varrho
\]

(4)

The intercept on the vertical this axis \((\delta=0)\) is the \( \varrho \) value as \( \gamma_{M}(k) \rightarrow \text{Var}(r) \) in probability as \( \gamma_{M}(k) \rightarrow 0 \). Selecting the most important and influencing parameters of a nonlinear and unknown function is one of the most difficult steps in model development. If \( n \) number of the input variables exists, the combination of \( 2^n - 1 \) would be among them and analysing all these combinations consumes lots of time. Therefore, GT was used in the present study for selecting the best combination of the input variables and it was achieved through winGamma™ software implementation [26].

2.3 Multilayer perceptron neural network (MLPNN)

MLP is one of the most popular ANN architecture used for hydrological modelling. Rumelhart et al. is considered to be first to introduce MLP with back propagation training algorithm for training of neural networks which considerably brought about significant growth in application of ANN in different fields [27]. MLP neural networks are multilayered feed forward networks typically trained with static back propagation and are made up of multiple layers of neurons. In this architecture, besides the input and the output layer, there is one or more than one intermediate layer(s) called hidden layer(s). Each layer is fully connected to the preceding layer by interconnection strengths or weights. A typical three layer MLP structure is shown in Fig. 2.
2.4 Development of model

In the present study, daily rainfall and runoff data of monsoon period (1\textsuperscript{st} June to 30\textsuperscript{th} September) for the period 2000-2009 were used for training and testing of MLP NN models. Out of this, 70 % of data (2000 to 2006) were used for training or calibration and remaining 30 % of data (2007 to 2009) were used for validation or testing of developed models. Best input combination was selected using GT technique and these inputs were used to train MLP NN for simulating current day runoff. Here, the MLP with both single and double hidden layers were trained using Levenberg–Marquardt as learning rule (which is an improved second order method for gradient) and hyperbolic tangent as transfer function using software NueroSolutions 5.0 designed and written by Curt Lefebvre & Jose Principe. The network training was stopped as soon as the maximum number of epochs, which was predetermined at 1000, and training threshold of 0.001 were reached. Different combinations of hidden neurons were tried and a network that yields the minimum root mean square error (RMSE), maximum correlation coefficient (r) and coefficient of efficiency (CE) was selected.

2.5 Performance evaluation

Three criteria, the root mean square error (RMSE), the correlation coefficient (r) and Coefficient of efficiency (CE) or Nash-Sutcliffe efficiency have been used to assess the goodness of fit performance of the models:

\[
RMSE = \sqrt{\frac{\sum_{j=1}^{n} (o_j - P_j)^2}{n}}
\]  \hspace{1cm} (5)

\[
r = \frac{\sum_{j=1}^{n} [(o_j - \bar{o})(P_j - \bar{P})]}{\sqrt{\sum_{j=1}^{n} (o_j - \bar{o})^2 \sum_{j=1}^{n} (P_j - \bar{P})^2}}
\]  \hspace{1cm} (6)
Where, $j$ is an integer varying from 1 to $n$, $O_j$, $P_j$, $\bar{O}$, $\bar{P}$ and $n$ are observed value, predicted value, mean of observed value, mean of predicted value and the number of observations respectively. The RMSE was used to measure prediction accuracy which produces a positive value by squaring the errors. $r$ is used as an indicator of degree of closeness between observed and predicted values. The coefficient of efficiency can be used to compare the relative performance of two approaches effectively and is commonly used to assess the predictive power of hydrological models [28]. Theoretically it varies from $-\infty$ and 1, with 1 being corresponding to perfect model.

III. RESULTS AND DISCUSSION

3.1 Gamma test

In this study, the current day rainfall ($R_t$) and previous days rainfall ($R_{t-1}$, $R_{t-2}$ .... $R_{t-n}$) as well as previous days runoff ($Q_{t-1}$, $Q_{t-2}$ .... $Q_{t-n}$), were used to simulate current day runoff ($Q_t$), where $n$ is number of lags. The results GT is shown in Table 1. According to the principals of the GT, the combination with the minimum gamma value would be the best combination for modeling and showed that the data with the provided combination has the possibility to achieve a better result in modeling [29]. As observed from the Table 1, the minimum gamma value was found to be for model no. 8 with $r$ value of 0.0642. Therefore, the combination $R_t$, $R_{t-1}$,$R_{t-2}$,$Q_{t-1}$,$Q_{t-2}$ was selected as the best input combination and optimum variables for developing MLPNN models for simulating daily runoff in Bino watershed.

Table 1 Results of GT for determining the best combination out of the input variables for runoff modeling

<table>
<thead>
<tr>
<th>Model no.</th>
<th>Model input</th>
<th>Gamma value ($r$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$R_t$</td>
<td>0.1604</td>
</tr>
<tr>
<td>2</td>
<td>$R_t$, $Q_{t-1}$</td>
<td>0.2265</td>
</tr>
<tr>
<td>3</td>
<td>$R_t$, $Q_{t-1}$, $Q_{t-2}$</td>
<td>0.0972</td>
</tr>
<tr>
<td>4</td>
<td>$R_t$, $R_{t-1}$, $Q_{t-1}$</td>
<td>0.2311</td>
</tr>
<tr>
<td>5</td>
<td>$R_t$, $R_{t-1}$, $Q_{t-1}$, $Q_{t-2}$</td>
<td>0.0742</td>
</tr>
<tr>
<td>6</td>
<td>$R_t$, $Q_{t-1}$, $Q_{t-2}$, $Q_{t-3}$</td>
<td>0.1213</td>
</tr>
<tr>
<td>7</td>
<td>$R_t$, $R_{t-1}$, $R_{t-2}$, $Q_{t-1}$</td>
<td>0.2010</td>
</tr>
<tr>
<td>8</td>
<td>$R_t$, $R_{t-1}$, $R_{t-2}$, $Q_{t-1}$, $Q_{t-2}$</td>
<td><strong>0.0642</strong></td>
</tr>
<tr>
<td>9</td>
<td>$R_t$, $R_{t-1}$, $Q_{t-1}$, $Q_{t-2}$, $Q_{t-3}$</td>
<td>0.1282</td>
</tr>
</tbody>
</table>
3.2 MLPNN runoff models

Different models with the varying hidden neurons of both single and double hidden layers have been trained and tested with MLPNN to select the optimal architecture of the network. All together 20 models i.e. MLP1 to MLP 20 has been developed and out of these, 10 are single hidden layer neural networks i.e. MLP1 to MLP10 and rest are double hidden layer neural networks as shown in Table 2.

**Table 2 Performance indices of MLP based ANN models during training and testing for runoff prediction**

<table>
<thead>
<tr>
<th>Model</th>
<th>Network</th>
<th>Training</th>
<th></th>
<th></th>
<th></th>
<th>Testing</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>r</td>
<td>RMSE</td>
<td>CE</td>
<td>r</td>
<td>RMSE</td>
<td>CE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP1</td>
<td>5-2-1</td>
<td>0.8198</td>
<td>3.3121</td>
<td>0.1892</td>
<td>0.6971</td>
<td>1.7257</td>
<td>0.3514</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 2</td>
<td>5-3-1</td>
<td>0.8249</td>
<td>3.2352</td>
<td>0.2265</td>
<td>0.7870</td>
<td>1.9638</td>
<td>0.1601</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 3</td>
<td>5-4-1</td>
<td>0.9192</td>
<td>1.7891</td>
<td>0.7634</td>
<td>0.8103</td>
<td>1.8184</td>
<td>0.2798</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 4</td>
<td>5-5-1</td>
<td>0.9327</td>
<td>1.4966</td>
<td>0.8345</td>
<td>0.8995</td>
<td>1.2797</td>
<td>0.6450</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 5</td>
<td>5-6-1</td>
<td>0.8193</td>
<td>2.8916</td>
<td>0.3820</td>
<td>0.8372</td>
<td>1.3340</td>
<td>0.6124</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 6</td>
<td>5-7-1</td>
<td>0.9284</td>
<td>1.7897</td>
<td>0.7633</td>
<td>0.8466</td>
<td>1.2796</td>
<td>0.6434</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 7</td>
<td>5-8-1</td>
<td>0.9455</td>
<td>1.2741</td>
<td>0.8800</td>
<td>0.9207</td>
<td>0.9644</td>
<td>0.7974</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 8</td>
<td>5-9-1</td>
<td>0.9025</td>
<td>1.7867</td>
<td>0.7641</td>
<td>0.8349</td>
<td>1.2564</td>
<td>0.6562</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 9</td>
<td>5-10-1</td>
<td>0.9275</td>
<td>2.1100</td>
<td>0.6710</td>
<td>0.8466</td>
<td>1.2796</td>
<td>0.6434</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 10</td>
<td>5-11-1</td>
<td>0.9160</td>
<td>1.7865</td>
<td>0.7641</td>
<td>0.7411</td>
<td>1.5139</td>
<td>0.5008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 11</td>
<td>5-6-3-1</td>
<td>0.8132</td>
<td>2.6271</td>
<td>0.4899</td>
<td>0.6212</td>
<td>1.7928</td>
<td>0.2999</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 12</td>
<td>5-6-5-1</td>
<td>0.8629</td>
<td>2.0327</td>
<td>0.6946</td>
<td>0.8284</td>
<td>1.4451</td>
<td>0.5452</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 13</td>
<td>5-7-6-1</td>
<td>0.8974</td>
<td>1.7440</td>
<td>0.7752</td>
<td>0.7342</td>
<td>1.5927</td>
<td>0.4475</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 14</td>
<td>5-8-4-1</td>
<td>0.9036</td>
<td>1.6880</td>
<td>0.7894</td>
<td>0.8780</td>
<td>1.1392</td>
<td>0.7173</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLP 15</td>
<td>5-8-7-1</td>
<td>0.9118</td>
<td>1.6459</td>
<td>0.7998</td>
<td>0.8027</td>
<td>1.4596</td>
<td>0.5360</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The results of performance evaluation indices values from the rainfall-runoff modelling of MLP based ANN models during training and testing are shown in Table 2. As observed from Table 2, the r values of the developed MLP based ANN models vary from 0.8132 to 0.9455 during training and 0.6212 to 0.9207 during testing. The RMSE values vary between 1.2741 to 3.2352 (mm) and 0.9644 to 1.9638 (mm) during training and testing, respectively. While the values of CE during training and testing varied from 0.1892 to 0.8800 and 0.1601 to 0.7974, respectively. It is also observed from the Table 2 that out of the 20 models developed, MLP7 (5-8-1) was found to the best as compared to other networks based on the performance criteria. Among the double hidden neuron networks developed, MLP19 (5-10-11-1) was found to perform good. The r, RMSE and CE for MLP19 model during training period are found to be 0.9311, 1.4796 (mm) and 0.8382, respectively and 0.9007, 1.1644 (mm), 0.7047 are their respective values during testing. Qualitative performance of developed MLP based ANN models was evaluated by comparing observed and predicted values of daily runoff graphically in the form of time series and scatter plot during training and testing for the selected networks as shown in Figs. 3 through 6. It can be observed from these figures that the observed and predicted runoffs are in close agreement during training although there are under and over predictions in some data points during testing. During testing period, model MLP7 could predict better as compared to MLP19 model. This is clear from scatter plots with $R^2$ values 0.8939 and 0.8476 for MLP7 while for MLP19, $R^2$ values were found to be 0.8670 and 0.8113 during training and testing periods, respectively.
Fig. 3 Comparison of observed and predicted runoff and their corresponding scatter plot during training period for MLP7 (5-8-1) model

\[ y = 0.826x + 0.741 \]
\[ R^2 = 0.893 \]
Fig. 4 Comparison of observed and predicted runoff and their corresponding scatter plot during testing period for MLP7 (5-8-1) model

\[ y = 0.729x + 0.976 \]

\[ R^2 = 0.847 \]
Fig. 5 Comparison of observed and predicted runoff and their corresponding scatter plot during training period for MLP19 (5-10-11-1) model
IV. CONCLUSIONS

In this study, GT technique was applied for best input selection for smooth daily rainfall-runoff modeling using MLP based ANN techniques. The results of the study showed that GT technique can be effectively used prior to actual hydrological modeling thereby saving huge time while selecting the best inputs to fed in models. Out the twenty developed models by MLP based models ANN, MLP7 was found to be the best followed by MLP 19. The results showed that MLP7 has better performance than MLP19. Therefore, MLP7 model could be successfully applied for daily rainfall-runoff modeling in Bino watershed, Uttarakhand.

V. ACKNOWLEDGEMENTS

The authors are thankful to Divisional Forest and Soil Conservation Office, Ranikhet, Uttarakhand, India for providing required data for the study and the first author is thankful to the Department of Science and Technology (DST), Ministry of Science and Technology, New Delhi for the financial support in the form of INSPIRE Fellowship rendered during her doctoral degree.

REFERENCES


