A Review on Artificial Neural Networks for Streamflow Prediction

Mohd Saleem[1], Sanat Nalini Sahoo[2]
[1]*Civil engineering Department, NIT Rourkela, Rourkela, 769008, India, Email: masaleem029@gmail.com*
[2] *Civil engineering Department, NIT Rourkela, Rourkela, 769008, India, Email:sahoosanat@nitrkl.ac.in*

ABSTRACT

Stream flow prediction provides the information of various problems related to the design and effective operation of river balancing system. So it is an essentially important aspect of any watershed modelling. The black box models(ANN) have proven to be an efficient alternative to physical (traditional) methods for simulating streamflow and sediment yield of the catchments. This present study focusses on development of models using ANN for predicting the stream flow for Subarnarekha river basin. By reviewing the earlier research works, it is observed that the procedure addresses the selection of input variables, the definition of model architecture and the strategy of the learning process. The input vectors used for the models were daily rainfall, mean daily evaporation, mean daily temperature and lag streamflow. Further, it is observed that the model performance was evaluated by statistical parameters like root-mean square error measures (RMSE), Nash-Sutcliffe efficiency (N-S) and squared correlation coefficient (R^2) and found that ANN model performance improved with increasing input vectors.

Keywords: Streamflow, RMSE, N-S efficiency, R^2.

1.INTRODUCTION

Stream flow or discharge is the volume of water that moves through a specific point in a stream during the given period of time. It is one part of the spill over of water from the land to waterbodies, the other segment being surface overflow. Water streaming in channels originates from surface overflow from nearby hillslopes, from ground water stream out of the ground, and from water discharged from pipes. Discharge is typically measured in units of cubic feet per second (cfs). To determine discharge, a cross-sectional area of the stream or river is measured. The discharge of water streaming in a channel is measured utilizing stream gauges or can be assessed by the Manning condition. The record of stream after some time is known as a hydrograph. Flooding occurs when the volume of water exceeds the capacity of the channel.

Modelling of streamflow process is important especially for planning, operating, and management of water resources. The exact measure of streamflow from precipitation possesses a critical place in the hydrological cycle. The measure of streamflow from precipitation is important to anticipate for keeping away from hazard and assessment of flood. Therefore, numerous hydrological models have been developed in order to simulate this complex process. Due to advances in computing systems, use of computerized reasoning (AI) strategies for cross-station or single station every day or month to month streamflow expectation has been examined and successful results have been reported.

Streamflow predictions on a river system can be done in two ways:
1)Short term prediction. 2)Long term prediction.

Short-term prediction forecasts, with lead times of hours or days, are necessary for flood warning systems and real-time reservoir operation and as a long-term prediction or transition from short- to long-term prediction forecasts, Monthly period prediction, is useful for many water resource applications such as environmental protection, drought management and optimal reservoir operation. Both short-term and long-term streamflow predictions are required to plan, operate and optimize the activities associated with water resource system. Each of them has their own benefits and applications in operational hydrology. Based upon the aim of forecasting issue, different kinds of daily, monthly, and annual streamflow prediction models were developed.

Artificial Neural Networks have a structure where nonlinear capacities are available and the parameter identification process is based on techniques which search for global maximums in the space of feasible parameter values, and hence can represent the nonlinear effects present in the rainfall-runoff processes. ANN were created as a data stockpiling models and their parameters are ascertained in a way that looks like normal procedures (McCulloch and Pitts, 1943). Details of their properties and the computational process have been presented by Hopfield (1982) and the learning procedure of ANN is depicted by Rumelhart and McClelland (1986). The utilization of ANN systems in water assets and streamflow forecast is moderately new and has been accounted for by French et al (1992), Zurada (1992), Hall and Minns (1993), Zealand et al (1999) etc., In many previous studies, ANN type such as multi-layer feed-foreword back-propagation neural network was commonly adopted and it proved to be the most powerful tool to 80% of practical application in all fields of hydrologic engineering and sciences. An important advantage of ANN compared to classical stochastic models is that they do not require variables to be stationary and normally distributed. Nonstationary effects present in global phenomena, in morphological changes in rivers and others can be captured by the inner structure of ANN. Furthermore, ANN are relatively stable with respect to noise in the data and have a good generalization potential to represent input-output relationships.

2.ARTIFICIAL NEURAL NETWORKS(ANN)

An ANN is a very interconnected system of numerous straightforward preparing units called neurons, which are comparable to the natural neurons in the human mind. ANNs have been produced as a speculation of scientific models of human insight or neural science.

Their development is based on the following rules:
1.Data processing happens at many single components called nodes additionally called as units, cells, or neurons.
2. Signals are passed between nodes through connection links.
3. Each connection link has an associated weight that represents its connection strength.
4. Each node typically applies a nonlinear transformation called an activation function to its net input to determine its output signal.

Neurons having similar characteristics in an ANN are organised in groups called layers. The neurons in one layer are associated with those in the neighbouring layers, however not to those in the same layer. The strength of connection between the two neurons in adjacent layers is

represented as ‘connection strength’ or ‘weight’. An ANN normally consists of three layers: an input layer, a hidden layer and an output layer. The input neurons receive and process the input signals and send an output signal to other neurons in the network. Each neuron can be connected to the other neurons and has an activation function and a threshold function, which can be continuous, linear or nonlinear functions. The signal passing through a neuron is transformed by weights which modify the functions and thus the output signal that reaches the following neuron. Modifying the weights for all neurons in the network, changes the output. Once the architecture of the network is defined, weights are calculated so as to represent the desired output through a learning process where the ANN is trained to obtain the expected results.

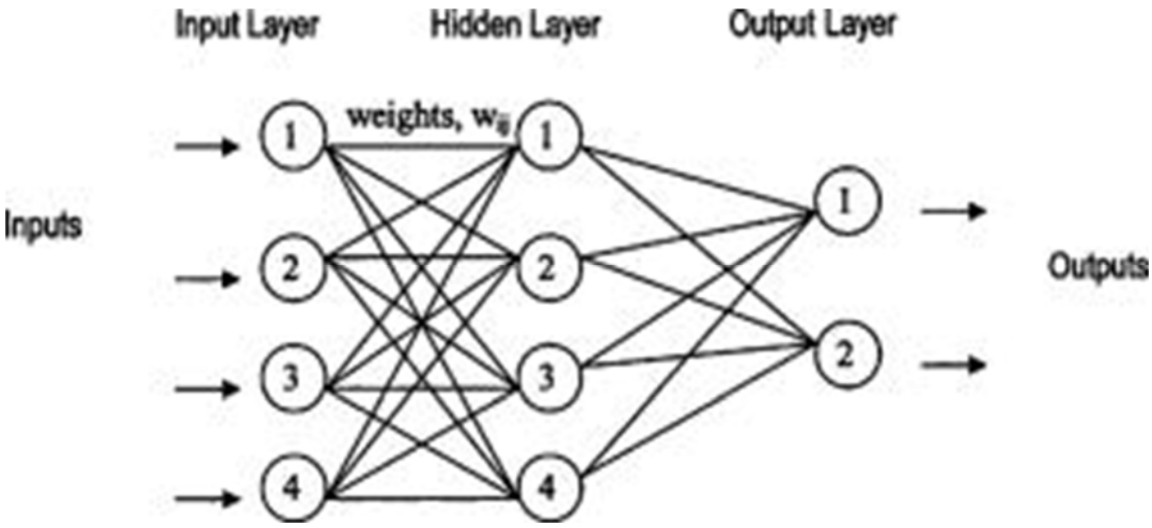

Figure 1: Multilayer feedforward Artificial Neural Network

One method for grouping neural networks is by the quantity of layers:
1)single layer (Hopfield nets); 2) bilayer (Carpenter/Grossberg versatile reverberation systems); and 3) multilayer (generally backpropagation systems). A linear model can be represented adequately by a single layer network, while a nonlinear model is generally associated with a multiple layer network. ANNs can likewise be sorted in light of the direction of data stream and processing.
1) Feed forward system. 2)Recurrent systems.

In a feedforward network, the nodes are generally arranged in layers, starting from a first input layer and ending at the final output layer. There can be several hidden layers, with each layer having one or more nodes. Information passes from the input to the output side. The nodes in one layer are connected to those in the next, but not to those in the same layer. Thus, the output of a node in a layer is only a dependent on the inputs it receives from previous layers and the corresponding weights. On the other hand, in a recurrent ANN, information flows through the nodes in both directions, from the input to the output side and vice versa. This is generally achieved by recycling previous network outputs as current inputs, thus allowing for feedback.

3.CLASSIFICATION OF ALGORITHMS

3.1 Backpropagation

Back-propagation is perhaps the most popular algorithm for training ANNs. It is essentially a gradient descent technique that minimizes the network error function. Each input vector of the training data set is passed through the network from the input layer to the output layer. The network output is compared with the target output, and an error is computed based on that. This error is propagated backward through the network to each node, and correspondingly the connection weights are adjusted. They are typically composed of three parts: (a) input layer including a number of input nodes, (b) one or more hidden layers, and (c) a number of output layer nodes

3.2 Radial basis function

A radial basis function (RBF) network can be considered as a three-layer network in which the hidden layer performs a fixed nonlinear transformation with no adjustable parameters. This layer consists of a number of nodes and a parameter vector called a ‘‘center,’’ which can be considered the weight vector of the hidden layer. The standard Euclidean distance is used to measure how far an input vector is from the centre. For each node, the Euclidean distance between the centre and the input vector of the network input is computed and transformed by a nonlinear function that determines the output of the nodes in the hidden layer. A typical RBF network has a feedforward structure consists of an input layer, a hidden layer with a radial basis activation function, and a linear output layer.

3.3 Cascade correlation algorithm

It differs from other approaches in that it starts with a minimal network without any hidden nodes and grows during the training by adding new hidden units one by one, maximizing the impact of the new node on the network error, creating a multilayer structure. Once a new hidden node has been added to the network, its input-side weights are frozen. The hidden nodes are trained in order to maximize the correlation between output of the nodes and output error. A training cycle is divided into two phases. First, the output nodes are trained to minimize the total output error. Then a new node is inserted and connected to every output note and all previous hidden nodes. The new node is trained to correlate with the output error. The addition of new hidden nodes is continued until maximum correlation between the hidden nodes and error is attained.

4. REVIEW OF RESEARCHERS

Mehr *et al.* (2014) conducted experiments on applicability of successive-station prediction models, as a practical alternative to streamflow prediction in poor rain gauge catchments. They have investigated using monthly streamflow records of two successive stations on Coruh River, Turkey. To achieve this goal, at the first stage, they considered eight different successive-

station prediction scenarios, they applied feed-forward back-propagation (FFBP) neural network algorithm as a brute search tool to find out the best scenario for the river. Then, they used two other artificial neural network (ANN) techniques, namely generalized regression neural network (GRNN) and radial basis function (RBF) algorithms to generate two new ANN models for the selected scenario. Finally, they compared the different algorithms using Nash–Sutcliffe efficiency, squared correlation coefficient, and root-mean square error measures. Performance analysis showed that only 1-month-lagged record of both stations was satisfactory to achieve accurate models with high-efficiency value. They found that the RBF network resulted in higher performance than FFBP and GRNN.

Cascade-correlation-training algorithm is the efficient method to find the optimal architecture but alternatively they used trial and error procedure for selecting the number of hidden layers to decide the optimal architecture which is time consuming process.

Dolling and Varas (2012**)** predicted the streamflow on San Juan River basin, Argentina which represents a mountainous watershed, subject to rainfall and snowmelt in conditions of scarce hydrologic information during spring and summer seasons. Input variables used were ENSO index for Zone 3 of the Pacific, monthly temperature, precipitation and snow course information. They used backpropagation momentum method along with SNNS software which proved to be a valuable and easy to use tool, for model identification and validation. Calculated flows showed that monthly spring and summer streamflow represented by neural network models have a better performance than alternative procedures.

It is observed that in input variables, the usage of ENSO index is because ENSO (El Niño-Southern Oscillation) is a naturally occurring phenomenon that involves fluctuating ocean temperatures arises from a complex interaction of a variety of climate systems. ENSO index is a method used to characterize the intensity of an **El Niño Southern Oscillation**(**ENSO**) event which is helpful in predicting monthly streamflow with reasonable accuracy.

Asati and Rathore (2012) conducted experiments on river Wainganga which was subjected to water level rise during 2004-2005. They made an attempt to use the conventional method such as Autoregressive model(AR), more deterministic approach through multi-Linear Regression model(MLR) and Artificial Neural Network(ANN) which are capable of identifying complex non-linear relationship between input and output data from 1 to 5 hours was performed individually, it was observed that AR Model gave satisfactory results compared to MR and ANN.ANN model was found to be better in simulation and prediction the flow characteristics under consideration compared to MLR and AR models for one hour ahead prediction.

In this paper, selection of model architecture, number of hidden layers and number of nodes in each hidden layer are not defined clearly which is why AR model performed better than ANN. A rigorous exercise on selection of transfer function which best suits the data, optimal architecture, number of epochs could have led to much better prediction of ANN.

Kothari and Gharde (2010**)** used ANN and FL algorithms for predicting the stream flow for catchment of savitri river basin. They used input variables like daily rainfall, mean daily

evaporation, mean daily temperature and lag streamflow. They considered the hydrological data of 20 years (1992–2011), of which 13 years (1992–2004) was for training and rest 7 years (2005–2011) for validation of the models. The mode performance was evaluated by RMSE, EV, CE, and MAD statistical parameters. It is observed that ANN model performance increases significantly with increase in the number of inputs whereas FL model performs better with single input as rainfall. Comparatively, ANN model performance was found to be superior as compared to FL model in forecasting streamflow for Savitri Basin.

5.METHODOLOGY

5.1 Selection of input & output variables

In hydrology, the values of input vectors can be casual variables such as rainfall, temperature, water levels, spatial locations, evaporation, basin area, elevation, slopes, pump operating status, contaminant loads, meteorological data, and so on. The values of output vectors can be hydrological responses such as runoff, streamflow, ordinates of a hydrograph, optimal pumping patterns, rain fields, hydraulic conductivities, contaminant concentrations, and others. The selection of an appropriate input vector that will allow an ANN to successfully map to the desired output vector is not an easy task. For example, a deep physical study can lead to better choice of input variables for proper mapping. This will help in avoiding loss of information that may result in omission of key input variables, and also prevent inclusion of false inputs that tend to confuse the training process. A sensitivity analysis can be done to determine the importance of a variable when sufficient data is available. The input variables that do not have a significant effect on the performance of an ANN can be removed from the input vector which results in a more compact network.

5.2 Collection and pre-processing data:

Mostly hydrological data is obtained either from gauges that are placed on site or through remote sensing instruments. It is observed that there is no fixed method for determining the number of input-output data pairs. To ensure a good approximation, Carpenter and Barthelemy (1994) stated that the number of data pairs used for training should be equal to or greater than the number of parameters (weights) in the network. Inclusion of unnecessary patterns could slow network learning. In contrast, an insufficient data set could lead to poor learning. This makes it useful to analyse and pre-process the data before it is used for an ANN application. Routine procedures such as plotting and examining the statistics are sometimes effective in judging the reliability of the data and possibly to remove outliers. In many cases, the data needs to be encoded, normalized, or transformed before being applied to an ANN.

5.3 Design of ANN:

Designing ANN involves the determination of the ANN architecture and selection of a training algorithm. An optimal architecture might be viewed as the one yielding the best execution as far as error minimization while retaining a simple and compact structure. There is no particular theory exists for determination of such an optimal ANN architecture. The flexibility lies in selecting the number of hidden layers and in assigning the number of nodes to each of these layers. A trial-and-error procedure is generally applied to decide on the optimal architecture. But the cascade-correlation-training algorithm is an efficient method to find the optimal architecture.

5.4 Training and Cross training

The available data set is generally partitioned into three parts for training, cross training, and validation. The purpose of training is to determine the set of connection weights and nodal thresholds that cause the ANN to estimate outputs that are sufficiently close to target values. The dataset reserved for training is used to achieve this goal. This fraction of the complete data to be employed for training should contain sufficient patterns so that the network can mimic the underlying relationship between input and output variables adequately. During training, the adjustment of errors is continued till weight space is found which results in the smallest overall prediction error. However, there is the danger of overtraining a network in this process, also called as overfitting. This happens when the network parameters are too fine-tuned to the training data set. To prevent this kind of overfitting, a cross training procedure is usually recommended. The goal of this procedure is to stop training when the network begins to overtrain. The second portion of the data is reserved for this purpose. After the adjustment of network parameters with each epoch, the network is used to find the error for this data set. Initially, errors for both the training and cross training data sets go down. After an optimal amount of training has been achieved, the errors for the training set continue to decrease, but those associated with the cross training data set begin to rise. This is an indication that further training will likely result in the network overfitting the training data. The process of training is stopped at this time, and the current set of weights and thresholds are assumed to be the optimal values. The network is ready for use as a predictive tool.

5.5 Validation

The performance of a trained ANN can be evaluated by subjecting it to new patterns that it has not seen during training. The performance of the network can be determined by computing the percentage error between predicted and desired values. In addition, plotting the model output versus desired response can also be used to assess ANN performance. Since finding optimal network parameters is essentially a minimization process, it is advisable to repeat the training and validation processes several times to ensure that satisfactory results have been obtained.

5.6 Selection of the optimal model

The results obtained with the validation set for each of the selected model architectures are analysed in order to choose the best model for the required streamflow prediction. To judge which model has the best performance, graphical and analytical comparisons can be used. One can compare time series graphs of observed and predicted monthly streamflow and dispersion diagrams of observed and calculated values. Errors or residues should be analysed to test them for normality, independence, autocorrelation and cross correlation. Both numerical and graphical results should be considered with respect to predetermined criteria to select the best model.

6.STRENGTHS AND LIMITATION OF ANN

These are some reasons why ANN is more attractive computational tool.
1.They are able to recognize the relation between the input and output variables without explicit physical consideration.
2. They work well even when the training sets contain noise and measurement errors.
3. They are able to adapt to solutions over time to compensate for changing circumstances.
4. They possess other inherent information-processing characteristics and once trained are easy to use.

1.The success of an ANN application depends both on the quality and the quantity of data available. Therefore, data sets recorded over a system that should be relatively stable and unaffected by human activities.
2.It is not immediately clear how far back one must go in the past to include temporal effects. This makes the resulting ANN structure more complicated.
3.The fact that there is no standardized way of selecting network architecture also receives criticism. The choice of network architecture, training algorithm, and definition of error are usually determined by the user's past experience and preference, rather than the physical aspects of the problem.

7.CONCLUSION:

The detailed reviews presented in this paper highlights the application of Artificial Neural Networks in predicting the streamflow of a river basin. The present study put emphasis on following points:

1) Artificial neural networks (ANN) are used to establish the relationship between the input and output of specified hydro-meteorological data combination.

2) It is important to determine the dominant model inputs, as this reduces the size of the network and consequently reduces the training times and increases the generalization ability of the network for a given data set.
3) The selection of data transfer function is very important in artificial neural network modelling, which transfers the signal from input to hidden and hidden layer to output layered with appropriate weightages.
4) The Back propagation(BP), radial basis function(RBF), generalised neural networks(GRNN) are supervised learning algorithm which precisely trains the outputs of model with reference to input by calculating errors in actual data and computed data.

Acknowledgement

The authors would like to acknowledge Department of civil engineering, National institute of technology, Rourkela, India for providing support for this research

REFERENCES:

1. By the ASCE Task Committee on Application of Artificial Neural Networks in Hydrology.
2. Mahesh Kothari, k d Gharde (2010) "Application of ANN and FL algorithms for streamflow modelling on savitri catchment" Indian academy of sciences, No. 5, July 2015, pp. 933–943.
3. S R Asati and S S Rathore (2012)" comparative study of stream flow Prediction models" Vol. 1, No. 2, April 2012© 2012 IJLBPR.
4. Oscar r. Dolling, Eduardo a. Varas (2002)" Artificial neural networks for streamflow prediction", journal of hydraulic research, vol. 40, 2002, no. 5.
5. A. Danandeh Mehr, E. Kahya, A. Sahin, M. J. Nazemosadat (2013)" Successive-station monthly streamflow prediction using different artificial neural network algorithms, 20 August 2013 / Revised: 10 January 2014 / Accepted: 5 May 2014 / Published online: 21 May 2014_ Islamic Azad University (IAU) 2014.
6. Raphael M. Wambua, Benedict M. Mutua, James M. Raude (2016)" Prediction of Missing Hydro-Meteorological Data Series Using Artificial Neural Networks (ANN) for Upper Tana River Basin, Kenya, American journal of water resources, vol.4, No.2 ,2016, pp 35-43.
7. Jorge O. Pierini, Eduardo A. Gómez, Luciano Telesca (2012)" Prediction of water flows in Colorado River, Argentina. Lat. Am. J. Aquat. Res. vol.40 no.4 Valparaíso nov. 2012.
8. Karim Solaimani (2009)," Rainfall-runoff Prediction Based on Artificial Neural Network", American-Eurasian J. Agric. & Environ. Sci., 5 (6): 856-865, 2009