

Development of Flood Forecasting System Using Statistical and ANN Techniques in the Downstream Catchment of Mahanadi Basin, India

Anil Kumar Kar¹, Anil Kumar Lohani², Narendra Kumar Goel¹, Gopal Prasad Roy³

¹*Department of Hydrology, Indian Institute of Technology, Roorkee, India*

²*National Institute of Hydrology, Roorkee, India*

³*Department of Water Resources, Government of Orissa, Bhubaneswar, India*

E-mail: aklnih@gmail.com, lohani@nih.ernet.in

Received May 6, 2010; revised July 27, 2010; accepted August 12, 2010

Abstract

The floods in river Mahanadi delta are due to either dam release of Hirakud or due to contribution of intercepted catchment between Hirakud dam and delta. It is seen from post-Hirakud periods (1958) that out of 19 floods 14 are due to intercepted catchment contribution. The existing flood forecasting systems are mostly for upstream catchment, forecasting the inflow to reservoir, whereas the downstream catchment is devoid of a sound flood forecasting system. Therefore, in this study an attempt has been made to develop a workable forecasting system for downstream catchment. Instead of taking the flow time series concurrent flood peaks of 12 years of base and forecasting stations with its corresponding travel time are considered for analysis. Both statistical method and ANN based approach are considered for finding the peak to reach at delta head with its corresponding travel time. The travel time has been finalized adopting clustering techniques, there by differentiating high, medium and low peaks. The method is simple and it does not take into consideration the rainfall and other factors in the intercepted catchment. A comparison between both methods are tested and it is found that the ANN methods are better beyond the calibration range over statistical method and the efficiency of either methods reduces as the prediction reach is extended. However, it is able to give the peak discharge at delta head before 24 hour to 37 hour for high to low peaks.

Keywords: Flood Forecasting, Mahanadi Basin, Hirakud Dam, Statistical Method, ANN Architecture, Clustering

1. Introduction

Flood is a regular phenomenon in the delta of Mahanadi basin. The discharge of the intercepted catchment of below Hirakud dam has contributed largely to the flood at delta. The delta is highly fertile and thickly populated with population density 400-450 persons per sq.km. The drainage capacity of deltaic rivers is 25500 m³/s only. Any discharge above this can create a flood like situation.

Establishment of a workable flood forecasting method for the downstream catchment of Mahanadi basin is always under scanner. A lot of hydro-meteorological information is needed for establishment of a model. The dam at upstream is operated by a rule curve. The intercepted catchment is very big around 40000 sq.km. In

order to establish a Rainfall-Runoff model, rainfall data of all station is highly essential. For further strengthening of model evaporation, soil moisture, temperature, dam release etc. are needed. Sometimes all information is not available on real time basis making it difficult for establishment of a physical based model. In that case statistical methods are more useful.

The statistical methods are generally graphical or in the form of mathematical relationship, developed with the help of historical data, using statistical analysis. These include simple gauge to gauge relationships, gauge to gauge relationship with additional parameters and rainfall-peak stage relationship. These relationships can be easily developed and are most commonly used in India as well as other countries of the world [1]. Spok-

In our study the station at Khairmal is taken as the base station (B.S) which is 115 km away and with an average lead time of 12-16 hour for Hirakud release to reach here. Barmul is the intermediate station (I.S) and Mundali is the forecasting station (F.S) just upstream of Cuttack city.

3. Data Availability

Adequate data is needed for the formulation of the forecasting services. The development of the river forecasting procedure requires historical hydrological data and for the preparation of operational forecasts sufficient current information is required. Discharge data of (3 hourly) is available at base, intermediate and forecasting station for a period of 1996-2007. Discharge of 2830 cumecs and above at forecasting stations are considered as peaks and its corresponding peaks at intermediate and base station are collected with their related travel times after drawing the corresponding time series. Out of this 80 peaks are considered for our analysis, taking 60 peaks for calibration and 20 peaks for testing of models. Generally, it can be said that it is necessary to have a minimum of 10 years of basic hydraulic data, available to develop adequate river forecasting procedures, the primary requirement being that the period of record should contain a representative range of peak flows [20].

4. Methodology

The selection of an appropriate flood forecasting model depends on the availability of the data, output desired etc. On the basis of the analytical approach for the development of flood forecasting method can be classified as:

- 1) Methods based on statistical approach.
- 2) Methods based on mechanism of formation and propagation of floods.

4.1. Statistical Approach

Finding correlation between stage and discharges between upstream and downstream gauging stations is one of the simplest methods. This gives better result when there is less influence of tributaries joining the main stream in between or the intercepted catchment is not influenced by heavy rainfall.

4.2. ANN Approach

There are many ANN architectures and algorithms developed. Out of them most common are Multi layer feed forward, Hopfield networks, Radial basis function network, Recurrent network, Self organization feature maps, Counter propagation networks. Selection of a particular

network is application oriented. However, the multi layer feed forward networks are most commonly used for hydrological applications [7].

Generally four distinct steps applied to any ANN – based solution including flood forecasting problems. First of all data transformation is the initial step. The input always varies in a larger range. So in order to lessen the range the data is put in logarithmic scale. Then normalization and scaling is done. Data sets with many variations make training difficult. In order to prevent it the data are usually scaled using statistical, min-max, sigmoidal or principal component transformations [21]. Also the absolute input values are scaled to avoid asymptotic issues [22].

The second step is the fixing the network architecture in which for a particular problem the number of hidden layers, neurons in each layer and the connectivity between neurons are set. Many experimental results say that one hidden layer may be enough for most forecasting problems [6,17]. The studies of Cybenko [23], Hornik *et al.* [24] revealed that a single hidden layer is sufficient for ANNs to approximate any non-linear transfer function. The number of neurons in each layer depends upon the problem being studied. Less number of neurons in hidden layer will make the network with less degree of freedom for learning and more number of neurons will lead towards more time and over fitting [25]. Validation set error is often used to determine the optimal number of hidden neurons for a given study.

In the third step is the finalization of a learning algorithm for training the network. The parameters are finalized for the training data set to be applicable for any kind of testing data. ANN architecture is considered to be trained when the difference between ANN output and observed output is very small.

Finally the validation step is applied to get the performance of the network. The optimal number of hidden neurons and training iterations can also be determined through validation. The selection of acceptable model is finalized on the basis of RMSE, R^2 and efficiency.

4.3. Clustering Approach

As the data set varies in a wide range, it requires some distinction and in particular when different peaks are available for the same travel time. In order to make proper justice for different peaks with respective travel time, it requires clustering. Both K -mean and Fuzzy C -mean clustering methods are attempted.

4.3.1. K-Mean

This method was developed by MacQueen [26]. It is best described as a partitioning method. It partitions the data into K mutually exclusive clusters and returns a vector of

indices indicating to which of the K -clusters it has assigned each observation. The algorithm to clusters N objects based on attributes into K partitions where $K < N$. The optimization function

$$V = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2 \quad (1)$$

It tries to achieve minimum intra cluster variance or the squared error function. Where there are K clusters $S_i = 1, 2, \dots, K$, and V_i is the centroid or mean point of all the points $X_j \in S_i$. [27].

4.3.2. Fuzzy C-Mean (FC)

In this method the affinity of a site to undergo either two or more clusters are visualized. Earlier developed by Dunn [28] and improved by Bezdek [29] is basically used for pattern recognition. Here the data are bound to each cluster by means of a membership function which represents the Fuzzy behavior of this algorithm. It shows how to group data points that populate some multidimensional space into a specific number of different clusters. The objective function

$$Jm = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2, \quad 1 \leq m \leq \infty \quad (2)$$

$m =$ any real number, u_{ij} = degree of membership of x_i in cluster j , $x_i =$ i th of d -dimensional measured data, $c_j = d$ -dimension center of cluster.

Similarly, different ANN architectures are attempted over the discharge and corresponding travel times. The calibration and validation data sets remain same for both the methodology. The same methodology is also adopted

for finding the travel time. But in order to further distinguish K -mean and Fuzzy C mean clustering methods are adopted to find a better answer for travel time of different peaks.

5. Results and Discussion

First of all statistical approach is applied into calibration dataset (60 nos.). The relationships between the discharges between Khairmal-Barmul, Barmul-Mundali, Khairmal-Mundali and Khairmal-Barmul-Mundali are developed. These relationships are put on the testing data sets (20 nos.) and the results are recorded in **Table 1**.

Where, $Q_M =$ discharge at Mundali (Forecasting Station), $Q_B =$ discharge at Barmul (Intermediate Station) and $Q_K =$ discharge at Khairmal (Base Station).

The same dataset is again put into ANN architecture using MATLAB codes. The trial has been taken with a 3-layer feed forward network. Different combinations of feed forward network with changing transfer function, number of neurons and epochs varying at an increment of 50 are trailed. The combinations which are mostly as per performance criteria fixed are noted. The Cascade feed forward network has been most successful for Khairmal-Barmul and Khairmal-Barmul-Mundali and other two cases are with Feed forward network. Numbers of neurons in input and output layers are mentioned in column (2) of **Table 2**. In all cases ‘tansig’ neurons are used in first layer, ‘purelin’ in second layer and ‘trainbr’ remains the training function. Both ‘tansig’ and ‘purelin’ are better compatible with feed forward networks. Again the two layer sigmoid/linear network can represent any

Table 1. Relationship between discharges using simple statistical method (peak to peak).

Between stations	Equation	RMSE (m ³ /s)		R ²		Efficiency	
		Training	Testing	Training	Testing	Training	Testing
Khairmal-Barmul	$Q_B = 1.122 * Q_K - 289.4$	2382.62	1287.4	0.94	0.974	0.952	0.9678
Barmul-Mundali	$Q_M = 1.009 * Q_B + 678.17$	1740.26	1275.16	0.968	0.989	0.948	0.9735
Khairmal-Mundali	$Q_M = 1.144 * Q_K + 225.04$	2623.76	1769	0.929	0.9826	0.929	0.9490
Khairmal-Barmul-Mundali	$Q_M = 0.84197 * Q_B + 0.1996 * Q_K + 468.72$	1705.85	1272.12	0.970	0.990	0.934	0.9736

Table 2. Relationship between discharges using ANN architecture.

Between stations	ANN architecture	Epochs	RMSE (m ³ /s)		R ²		Efficiency	
			Training	Testing	Training	Testing	Training	Testing
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Khairmal-Barmul	CFF 1,7,1	3000	2249.1	1132.2	0.9720	0.9891	0.9446	0.9752
Barmul-Mundali	FF 1,7,1	3000	1681.5	1251.9	0.9852	0.9962	0.9705	0.9745
Khairmal-Mundali	FF 1,5,1	2000	2441.5	1741.7	0.9684	0.9789	0.9375	0.9506
Khairmal-Barmul-Mundali	CFF 2,15,1	4000	1697	1324.1	0.9850	0.9922	0.9700	0.9714

functional relationship between inputs and outputs if the sigmoid layer has enough neurons [30]. The ‘trainbr’ function updates the weight and bias values as per Levenberg-Marquardt optimization. It minimizes the combination of squared errors and weights and then determines the correct combination so as to produce a network that generalizes well as per Bayesian regularization process. It reduces the difficulty of determining the optimum network architecture also over fitting of training dataset is prevented, whereas in other networks generalization and early stopping is required for reducing over fitting. The performance functions are set as RMSE, R^2 and efficiency. The details of the network applied for different combinations are shown in **Table 2**.

The plots of the statistical method and ANN network over the testing value of 20 numbers of observation as obtained from MATLAB is also produced in **Figures 2(a)** to **(d)** respectively.

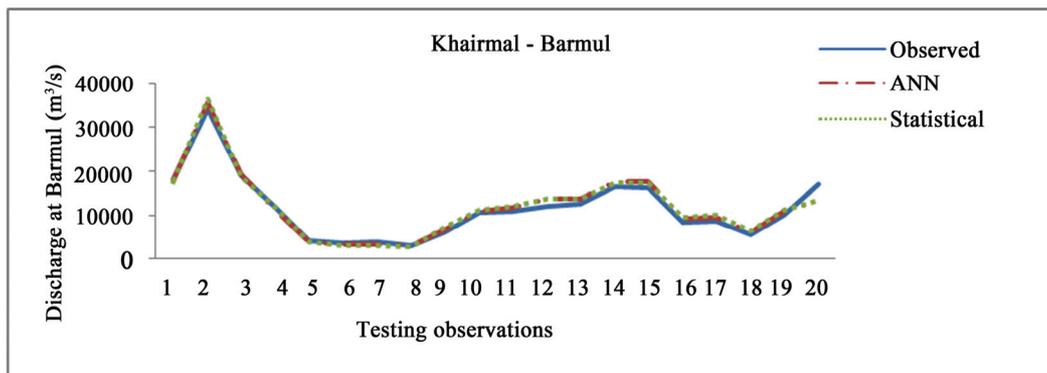
For travel time the corresponding discharge and travel times between the stations are trailed with the same calibration and testing data range applying both Statistical (**Table 3**) and ANN architectures (**Table 4**). The ANN is also applied in the same way with many numbers of trials.

It has been revealed that none of the case has a good efficiency, RMSE and R^2 . All the relations are ended

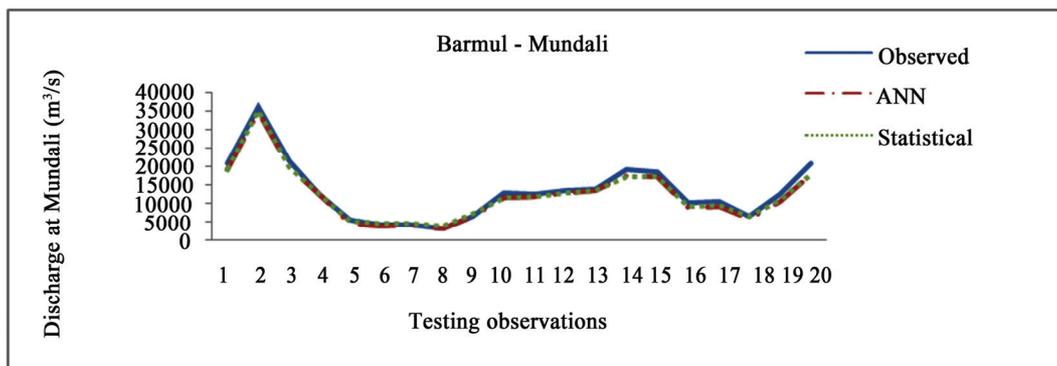
with a poor correlation. The lack of correlation does not imply lack of association. Because ‘r’ measures only linear association, a strict curvilinear relationship would not necessarily be reflected in a high ‘r’ value. Conversely, correlation between two variables does not guarantee that they are causatively connected [31].

However, to have a better answer for efficiency and RMSE values clustering methods are adopted for the discharge value at base station and its corresponding travel time for the calibration dataset. At least 4 clusters are attempted applying K-mean and Fuzzy c-mean method using the same MATLAB code. The datasets are divided into 4 clusters and their varying ranges are recorded. For each range the average travel time has been calculated. In both methods low peak (2660-8256 m^3/s) has large travel time (37.77 hours) to reach at delta head where as high peaks has taken nearly 25 hours to reach at delta head. As each peak is categorized the respective travel time for different discharge values are fixed. However when the result is put over the test dataset K-mean clustering has given 80% efficiency with RMSE value of 2.5 hour (**Table 5**). One value of testing dataset has been removed because it seems that it has some recording errors. So travel time test series has been carried out with 19 numbers of data.

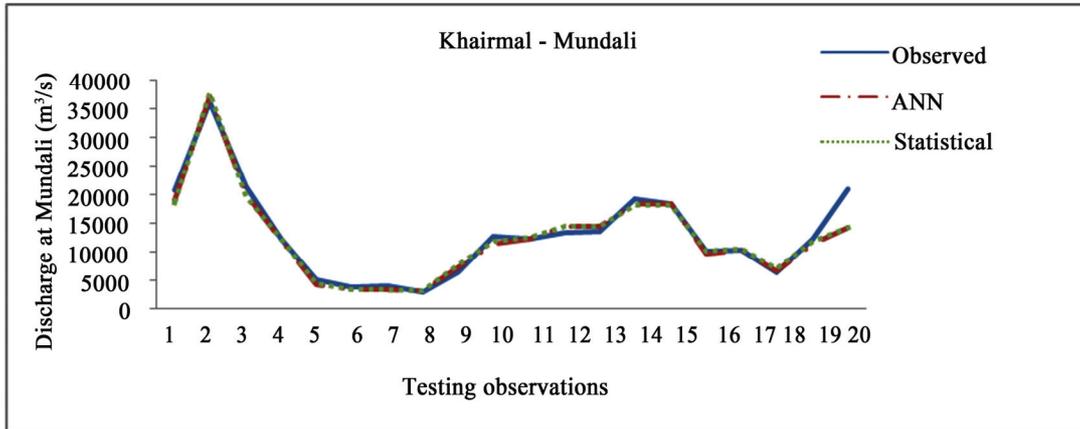
A plot (**Figure 3**) has been drawn with observed test-



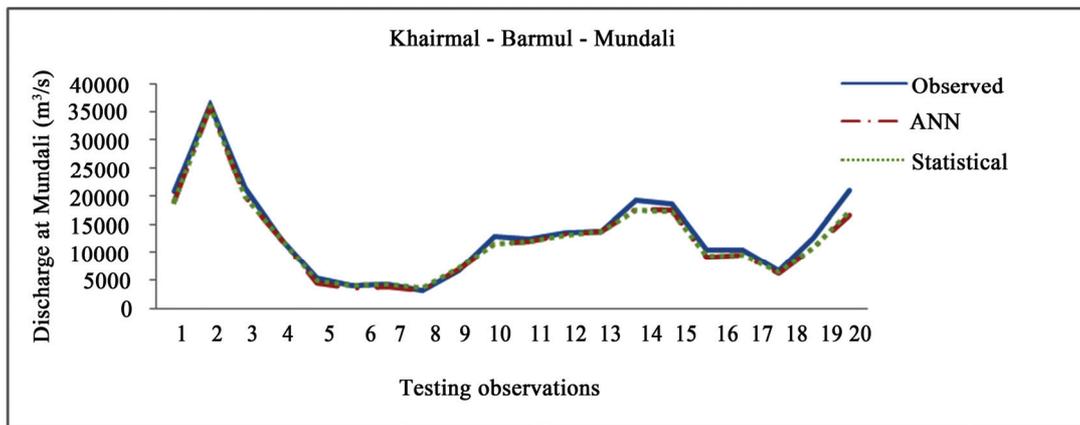
(a)



(b)



(c)



(d)

Figure 2. (a) Comparison between observed and computed discharge (Khairmal-Barmul); (b) Comparison between observed and computed discharge (Barmul-Mundali); (c) Comparison between observed and computed discharge (Khairmal-Mundali); (d) Comparison between observed and computed discharge (Khairmal-Barmul-Mundali).

Table 3. Relationship between discharges and travel time using simple statistical method.

Between stations	Equation	RMSE(hours)		R ²		Efficiency	
		Training	Testing	Training	Testing	Training	Testing
Khairmal-Barmul	$T_B = 270.7*(Q_K)^{-0.32}$	3.42	2.93	0.409	0.261	0.465	0.2973
Barmul-Mundali	$T_M = 64.96*(Q_B)^{-0.15}$	4.34	2.14	0.134	0.607	0.1366	0.507
Khairmal-Mundali	$T_M = 239.1*(Q_K)^{-0.22}$	5.67	3.95	0.394	0.568	0.4191	0.508

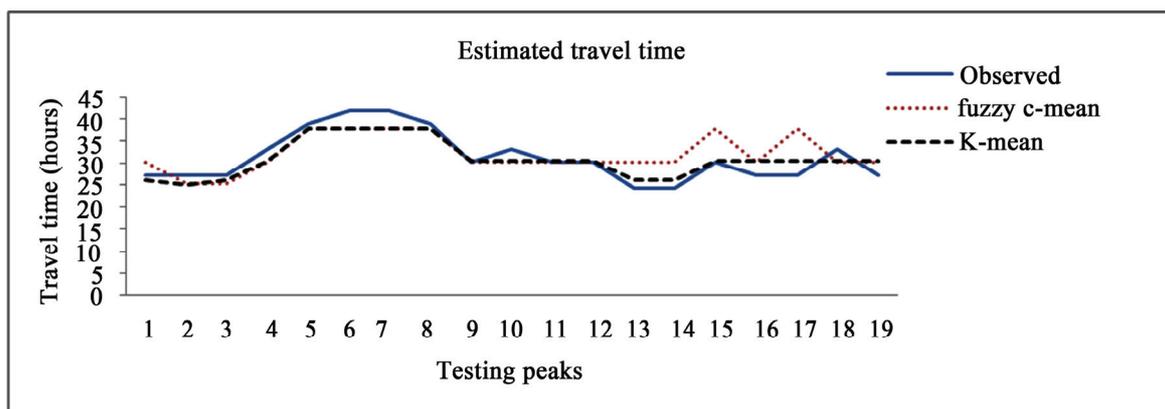
where, T_M = time (hour) to peak at Mundali, T_B = time (hour) to peak at Barmul.

Table 4. Relationship between discharges and travel time using ANN.

Between stations	Equation	RMSE (hours)		R ²		Efficiency	
		Training	Testing	Training	Testing	Training	Testing
Khairmal-Barmul	CFE,15,1 (2000)	3.449	2.998	0.6826	0.6259	0.4546	0.2634
Barmul-Mundali	CFE,10,1 (3000)	4.350	2.173	0.3933	0.7961	0.1352	0.4906
Khairmal-Mundali	FF,15,1 (2000)	5.683	3.043	0.6506	0.8306	0.4170	0.6965

Table 5. Peak values with corresponding travel time obtained by using different clustering methods

K-mean				Fuzzy-C-Mean			
Discharge range (m ³ /s)	Travel time (hours)	Testing RMSE (hours)	Testing Efficiency	Discharge range (m ³ /s)	Travel time (hours)	Testing RMSE (hours)	Testing Efficiency
2660-8256	37.77			2660-8256	37.77		
10329-15508	30.53			10329-16300	30		
16102-22810	26.06	2.52	0.8011	16499-25470	27.07	4.27	0.4003
25470-36300	25			28102-36300	25.2		

**Figure 3. Plot between testing travel time and estimated by different clustering methods.**

ing values and validation result of different clustering values. The plot shows the result of K-mean clustering is almost close with that of observed, whereas result of Fuzzy C mean for some low peak values. For high peaks both K-mean and Fuzzy have same results. As K-mean has able to model both high and low peaks with an efficiency of 80.1%, which is higher than Fuzzy C-mean methods, the discharge range values and its corresponding travel time fixed by this method may be taken as the forecast travel time between base station and forecasted station. The travel time has not calculated for other intermediate stations as our concentration is with the peak at base station and travel time of at least 24 hour.

6. Conclusions

Both statistical and ANN methods are applied for downstream Hirakud catchment between stations Khairmal, Barmul and Mundali. The forecasting obtained by both the methods is encouraging. Although the ANN method has a better performance, the existence of statistical method cannot be denied as far as discharge measurement is considered. However, the adoption of clustering method in order to find travel time with respect to different peaks is another finding. The best result for travel time is for

Khairmal-Mundali reach (efficiency = 80.11% and RMSE = 2.55 hour). As the data recording is of 3 hour interval this kind of result is still encouraging. The travel time between base and forecasting stations varies between 24-37 hour. By applying clustering method justice has been made for different magnitude of peaks. The results are handy to operate. The moment one gets the discharge at Khairmal (Base Station), he can immediately calculate the discharge to be obtained at Mundali (Forecasting Station) with its corresponding travel time. The result at Mundali (Forecasting Station) with respect to Khairmal (Base Station) basing on the ANN network is best for taking the flood forecasting effectively because in taking Khairmal-Barmul and Barmul-Mundali efficiency is above 97% and RMSE value is 500 m³/s less than Khairmal-Mundali but lead time of only 12 hours remains after Barmul. So considering the factor of higher lead time for more relief measures we are making a tradeoff with efficiency. Thus calculating peak discharge and travel time at Mundali on the basis of Khairmal discharge and time of occurrence will be a best workable method for practicing hydrologists and field engineers. Still the results are running with the limitation of data insufficiency, large recording intervals, non-consideration of influence of other climatic variation.

7. References

- [1] Central Water Commission of India, "Manual on Flood Forecasting," 1980.
- [2] E. Spokkereff, "Flood Forecasting for the River Rhine in the Netherlands," *Proceeding of the Symposium the Extreme of Extremes: Extraordinary Floods*, Reykjavic, Ireland, July 2000, pp. 347-352.
- [3] M. Bruen and J. Yang, "Functional Networks in Real-Time Flood Forecasting - A Novel Application," *Journal of Advances in Water Resources*, Vol. 28, No. 9, 2005, pp. 899-909.
- [4] M. Campolo, P. Andreeussi and A. Soldati, "River Flood Forecasting with a Neural Network Model," *Water Resources Research*, Vol. 35, No. 4, 1999, pp. 1191-1197.
- [5] M. Campolo, A. Soldati and P. Andreeussi, "Artificial Neural Network Approach to Flood Forecasting in the River Arno," *Hydrological Sciences Journal*, Vol. 48, No. 4, 2003, pp. 381-398.
- [6] P. Coulibaly, F. Anctil and B. Bobee, "Daily Reservoir Forecasting Using Artificial Neural Network with Stopped Training Approach," *Journal of Hydrology*, Vol. 230, No. 3-4, 2000, pp. 244-257.
- [7] C. W. Dawson and R. L. Wilby, "Hydrological Modeling Using ANN," *Progress in Physical Geography*, Vol. 25, No. 1, 2001, pp. 80-108.
- [8] C. E. Imrie, S. Durucan and A. Korre, "River Flow Prediction Using Artificial Neural Network: Generalization beyond the Calibration Range," *Journal of Hydrology*, Vol. 233, No. 1-4, 2000, pp. 138-153.
- [9] D. F. Lekkas, C. Onof, M. J. Lee and E. A. Baltas, "Application of Neural Networks for Flood Forecasting," *Global Nest: The International Journal*, Vol. 6, No. 3, 2004, pp. 205-211.
- [10] A. K. Lohani, "ANN and Fuzzy Logic in Hydrological Modeling and Flow Forecasting," Ph.D Dissertation, Indian Institute of Technology, Roorkee, 2007.
- [11] A. W. Minns and M. J. Hall, "Artificial Neural Networks as Rainfall-Runoff Models," *Hydrological Sciences Journal*, Vol. 41, No. 3, 1996, pp. 399-417.
- [12] J. R. Muhamad and J. N. Hassan, "Khabur River Flow Modeling Using Artificial Neural Network," *Al-Rafidain Engineering*, Vol. 13, No. 2, 2005, pp. 33-42.
- [13] A. Mukerjee, C. Chatterji, N. Singh and N. S. Raghuvanshi, "Flood Forecasting Using ANN, Neuro-Fuzzy and Neuro-GA Models," *Journal of Hydrologic Engineering*, Vol. 14, No. 6, 2009, pp. 647-652.
- [14] D. P. Solomatine and Y. Xue, "M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of The Huai River in China," *Journal of Hydrologic Engineering*, Vol. 9, No. 6, 2004, pp. 491-501.
- [15] D. P. Solomatine and R. K. Price, "Innovative Approaches to Flood Forecasting Using Data Driven and Hybrid Modeling," *6th International Conference on Hydroinformatics - Liong*, World Scientific Publishing Company, New Jersey, 2004.
- [16] C. M. Zealand, D. H. Burn and S. P. Simonovic, "Short Term Streamflow Forecasting Using Artificial Neural Network," *Journal of Hydrology*, Vol. 21, No. 4, 1999, pp. 32-48.
- [17] J. Zhang and M. J. Hall, "Regional Flood Frequency Analysis for the Gan-Ming River Basin in China," *Journal of Hydrology*, Vol. 296, No. 1-4, 2004, pp. 98-117.
- [18] Y. Xiong, X. Jun and Z. Xiang, "Flood Forecasting by Coupling Clustering Methods and Artificial Neural Networks," *Proceeding of Symposium HS 3006*, IUGG, Perugia, 2007.
- [19] Water Resources Department, Government of Orissa, "Time of Flow in Mahanadi River between Hirkud & Mundali," 2009. <http://www.dowrorissa.gov.in/Flood/DailyFloodBulletin.html>
- [20] S. N. Ghosh, "Flood Control and Drainage Engineering," A. A. Balkema Publishers, Rotterdam, 1997.
- [21] K. L. Priddy and P. E. Keller, "Artificial Neural Networks, an Introduction," SPIE Press, Bellingham, 2005.
- [22] S. Haykin, "Neural Networks," Macmillan College Publishing Company, Macmillan, 1994.
- [23] G. Cybenko, "Approximation by Superposition of a Sigmoidal Function," *Journal of Mathematics Control Signal System*, Vol. 2, No. 4, 1989, pp. 303-314.
- [24] K. Hornik, M. Stinchcombe and H. White, "Multilayer Feed forward Networks are Universal Approximators," *Neural Networks*, Vol. 2, No. 5, 1989, pp. 359-366.
- [25] N. Karunanithi, W. J. Grann, D. Whitely and K. Bovee, "Neural Networks for River Flow Prediction," *Journal of Computing in Civil Engineering*, Vol. 8, No. 2, 1994, pp. 201-220.
- [26] J. B. MacQueen, "Some Methods for Classification of Multivariate Observation," *Proceeding of 5th Berkely Symposium on Probability and Statistics*, University of California Press, Berkely, 1967, pp. 281-297.
- [27] K-Means Algorithm, 2009. <http://biocomp.bioen.uiuc.edu/oscar/tools/kmeans.html>
- [28] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, Vol. 3, No. 3, 1973, pp. 32-57.
- [29] J. C. Bezdeck, "Pattern Recognition with Fuzzy Objective Function Algorithm," Plenum Press, New York, 1967.
- [30] Neural Network Toolbox, Matlab, 2004.
- [31] World Meteorological Organization (WMO), "Guide for Hydrological Practices," Geneva, 1994.