International Journal of Civil & Structural Engineering – IJCSE
Volume 1 : Issue 1 *[ISSN 2372 – 3971]*
Publication Date : 09 January 2014

The Learning of K-Nearest Neighbours Model and Multivariate Adaptive Regression Splines Model in Rainfall-Runoff Processes at Pahang River, Malaysia.

Dewi Anneka binti Halid
Faculty of Civil Engineering,
Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia.
dewianneka@yahoo.com

Ismail bin Atan
Faculty of Civil Engineering
Universiti Teknologi MARA
Shah Alam, Selangor, Malaysia.
ismail585@yahoo.com

Abstract— **Ongoing needs to achieve the best accuracy of flood forecasting has stimulate this study to investigates the potential of two data driven model, where their application are relatively new in hydrology problems. The approaches studied here are K-Nearest Neighbours (KNN) and Multivariate Adaptive Regression Splines (MARS). To analyze and compares the performance of these two approaches in flood prediction, Pahang River in Malaysia has been selected as area of study. 30-years historical data set of daily rainfall and runoff at upstream tributaries of Pahang River were used to develop and validate the capability of both approaches in one-year-ahead prediction of flood discharge. The effect of different length of record data to the performance of models was also examined. Simulation results showed that longer period data can provide significant improvement to the performance of both approaches. However, satisfactory result of flood prediction only appeared superior for MARS model.**

Keywords—Flood Prediction, K-Nearest Neighbours, Multivariate Adaptive Regression Splines, Pahang River and Performance.

I. Introduction

The role of rainfall-runoff model is very essential in discharge prediction application. Generally, these models can be divided into conceptual rainfall-runoff model (CRR) and stochastic model (Yaacob, Jamaluddin, and Harun, 2005). The basic concept of these models is developed by using the mathematical formulation, but CRR required the physical law in their calculation while another model is vice versa. CRR have been developed long time ago and has labelled as complex process and consuming-time because it composed a large number of parameters that related to the physical system. To find alternative for CRR, recently most of hydrologist stressed their hydrologic forecasting study on system theoretic model compared to the conceptual model.

The successful of stochastic model in rainfall-runoff modelling have been revealed in many previous studies. It can be considered as technique that posses a very fast computation, less data requirements, and not related to the physical behaviour of the system. For better knowledge, the other basic terms for stochastic models, namely numerical model, linear and non-linear regression model, and data driven modelling. Among a variety of data driven modeling approaches that widely applied in hydrologic forecast are such as Auto-Regressive Moving Average (ARMA), Artificial Neural Networks (ANN), K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Model Tree and other statistical analysis. Based on careful review of the system theoretic model literature in Malaysia, it can be concluded that Neural Network is the most popular non-linear regression for flood forecasting. Various studies and papers of Neural Network in flood forecasting can be found, and most of it successful.

Although most of present approaches in flood prediction give the satisfactory results, they still have their own weaknesses and limitation such as time consuming, complex procedure, large data requirement and etc. The new models still need to be continuing study to improve the present flood forecasting system and to make it adequate with the current world situations. Therefore, objective of the study presented here was to seek new alternatives for the shortcomings of the present approach in flood prediction. The approaches focused here are K-Nearest Neighbours (KNN) and Multivariate Adaptive Regression Splines (MARS).

A. *K-Nearest Neighbours (KNN)*

The K-Nearest Neighbors (KNN) technique is a one of the simple non-parametic regression in rainfall runoff modeling that very intuitive and not implying any structured interaction, but nevertheless possesses powerful statistical (Toth, Brath, and Montanari, 2000). It calculates a prediction for unknown observations by exploiting the closeness between the most recent observation and K similar sets of observations chosen in any adequately large training simple (Toth et al, 2000), and then some function of their response values will be use to make the prediction, such as an average (Myatt, 2007). This method was originally developed by Farmer and Sidorowich in 1987 but has been introduced to the hydrological research world by Karlsson and Yakowitz (1987a, b).

In locally, the application of this method in hydrological problem is less established and still new among local

researcher. However, there have ample studies have been conducted internationally. Toth et al (2000) have study on comparative of K-Nearest Neighbours method with Auto-Regressive Moving-Average models and Artificial Neural Networks method for real time flood forecasting. The results of study indicates that performance of the rainfall forecasts appeared superior for the KNN method due to non-linear and threshold effects characterizing the rainfall–runoff transformation modeling. KNN model also can provide a good fit for low-intensity precipitation and flood forecasting accuracy can be further increased with respect to the use of intuitive. Other than that, the successful application of KNN method in flood forecasting also has revealed in study of Galeati (1990), Shamseldin and O'connor (1996), Moore and Bell (2001), Eskandarinia, Nazarpour, Teimouri, and Ahmadi (2010), and Azmi, Araghinejad, and Kholghi (2010).

B. *Multivariate Adaptive Regression Splines (MARS)*

Multivariate Adaptive Regression Splines (MARS) is a relatively new statistical technique in regression modeling where is well suited for high dimension problem and used in nonlinear model estimation when the exact nonlinear model is unknown (Hasti, Tibshirani, and Friedman, 2001). The mathematical basis for MARS was developed by the american statistician Jerome Friedman in 1990–1991. The power of MARS lies in its ability to estimate the contributions of the basic functions so that both the additive and the interactive effects of the predictors are allowed to determine the response variable (Friedman, 1991).

Unlike KNN model, the study of MARS in hydrological field still less established either in national or internationally. Nevertheless, Dwinnell (2000) has remark that MARS approach has proven effective at variety learning problem and competitive with Neural Network. The application of MARS in hydrology has been studied earlier by Lewis and Ray (1993), and Lall and Keppene (1996) in forecasting, then by Ames (1998) and, Abraham and Steinberg (2001) in rainfall prediction. All these study found that forecast technique using MARS model can produce more accurate forecasting. Besides that, the successful of MARS model in estimating the runoff from hilly watersheds also has been demonstrated in study of Sharda, Patel, Prasher, Ojasvi, and Prakash (2006) and Sharda et al (2010).

II. Methodology

A. *Site Description and Data*

The area of study is located in the largest state of Peninsular Malaysia,which is a Pahang River in Pahang State. The river also known as the longest river with the length of 459km and its upstream is located at the Main Range of Titiwangsa. The catchment area of Pahang River spans seven districts in Pahang which are Maran, Jerantut, Bentong, Lipis, Temerloh, Bera and Cameron Highlands and one sub district in Kuantan, eleven sub districts in Pekan and also two districts in Negeri Sembilan State which are Jelebu and Kuala Pilah. Usually, Pahang River experience flood every year due to the northeast monsoon and the floods of 1971, 1982, 1993, 1994, 1995, 1999, 2000, 2007 and 2010 were particularly high (DID, 2003).

Pahang River runs from Kuala Tembeling at the confluence of two equally large and long rivers which are the Jelai and the Tembeling as shown in Figure 1 below. Jelai River originates from the Central Mountain Range while Tembeling River has its origin at the Besar Mountain Range. The daily rainfall and water discharge which are composed 30 years of historical data from 1973 to 2003 at 12 stations along upstream of Pahang river were used to developed and validate the capability of KNN and MARS model in predict flood at upstream (Sungai Yap) , middle stream (Sungai Temerloh) , and downstream (Sungai Lubok Paku) of Pahang River. To determine the performance of model in short term flood prediction and long term flood prediction, data has been arranged into 5-years data set, 10 years data set, 20 years data set, and 30 years data set. All these data sets used 1 year data of 2003 for validation process while the others were applied for calibration.

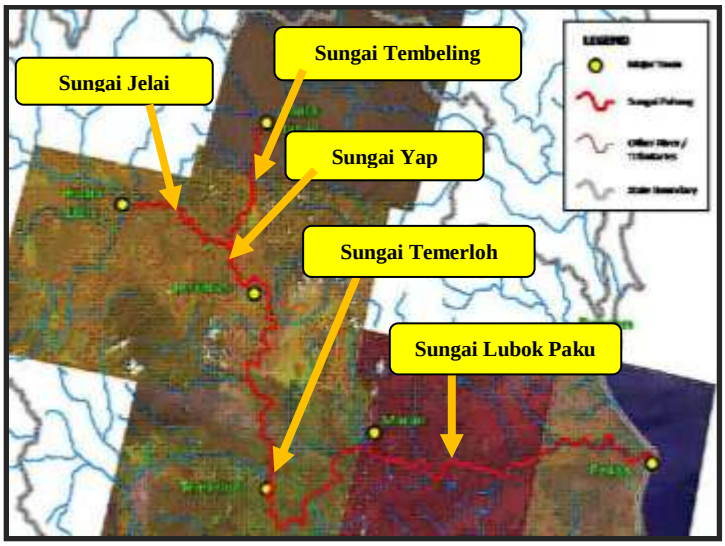

Figure 1: Pahang River and Tributaries (Ghani et al, 2012)

B. *KNN Approach*

The prediction concept of K-Nearest Neighbours is based on local approximation, where it making use only the nearby observations. For better understand (Knight and Shamseldin, 2006), the learning set $D_i = \{v,k\}$ is taken as collection of known cases $D_t = [v,k]$ and search is made for a given pattern v to be recognized for the best among the precedent v_j. The best class label K of the nearest neigbour $v_{nearest}$ will be determine using the weighted Euclidean distance (Karlsson and Yakowitz,1987) where the value will forwarded as a result

of the predcition. The formula of weighted Euclidean distance is shown in Equation 1 below (Toth et al, 2000).

$$r_{it} = \sqrt{\sum_{j=1}^{d} w_j (v_{ij} - v_{tj})^2} \quad (1)$$

Where v_{tj} is the jth component of [v,k] and w_j are scaling weights such as the standard deviation or range of v_j. Then the forecast is obtained by averaging the temporal evolution of the nearest neighbours as described in Equation 2, where it assumes to be similar to the evolution of the current situation. For the higher lead time L, the forecast is obtained with straight-forward as shown in Equation 3 (Toth et al, 2000).

$$v_{ij+1} = \frac{1}{K} \sum_{j=1}^{K} v_{tj+1} \quad (2)$$

$$v_{ij+L} = \frac{1}{K} \sum_{j=1}^{K} v_{tj+L} \quad (3)$$

The percentage of data sampling and number of nearest neighbours K are main parameter in this model. To get the optimum model of KNN in this study, 10% of data were used as sampling data and remained data were used as verification. As suggested by Lall and Sharma (1996), the generalized cross validation (GCV) function was applied to select the best value of K. 10-fold cross validation was used in this analysis, where all data are divided into 10 subsets and the process of cross-validation are repeated into 10 times.

C. *MARS Approach*

A key concept apply in MARS is the notion of knots, which are the points that mark the end of regions in data space where a distinct linear regression is run or the behavior of modeled function changes ((Briand,Freimut, and Vollei, 2004). As example, Figure 2 below show two knots X_1 and X_2 delimit three intervals where different linear relationships are identified.

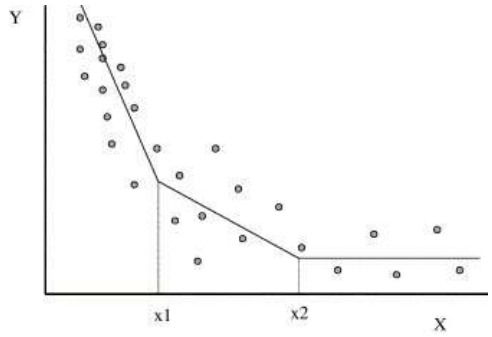

Figure 2: Example of Knots in MARS

MARS find the notion of knots through two processes which is a forward stepwise selection process then followed by a backward "pruning" process (Zhou and Leung, 2007). At the first stage, model will be built using forward stepwise by repeatedly adding basis function until the model reaches some predetermined maximum number of basis functions. The basis function here is an input-output relationship between the knots that have been determined. The basis functions will be defined by either of the two equations given below (Sharda et al, 2006).

$$Y = f_i(X) = \max(0, X_f - C_f) \quad (4)$$

$$Y = f_i(X) = \max(C_f - X_f, 0) \quad (5)$$

Where in the sub-group i, Y is the output of a basis function f, and C is the constant, call the knot, for input variable of X. According to Eq. 4, output Y= X-C when X is larger than C; otherwise Y=0. From Eq.5, Y= C-X when X is smaller than C; otherwise Y=0. These two basis functions are mirror images for each other. Then, in second stage which is backward "pruning" process, the contribution of each basis function will be evaluated using the "lack of fit" criterion (Zhou et al,2007). The main purpose of this process is to eliminate the basis functions which those contributes less to the overall goodness of fit. Therefore, the least effective terms will be removed one by one until the best model found. Here, the "lack of fit" will be computed by using cross-validation criterion (GCV) as shown in Equation 6, and the best predictive fits is selected.

$$GCV(M) = \frac{1}{N} \sum_{i=1}^{n} [y_i - f(x_i)]^2 \Big/ \left[1 - \frac{C(M)}{N}\right]^2 \quad (6)$$

The performance of MARS model is much relies on the total number of basis functions. By adding more number of basis functions to this model, the performance of MARS model can be improved. Therefore, the limit number of basis functions should be large enough to ensure that all best effective basis functions can be captured from the all data. Unlike KNN model, this model does not require a data sampling, it learns from the all previous data to develop the best predictive model.

D. *Performance Evaluation*

The predictive accuracy of the KNN and MARS techniques were evaluated using the mean relative squared error (MRSE), the mean relative absolute error (MRAE), and the percentage of correlation efficiency (CE). MRSE and MRAE is the lack-of-fit indicator, where if this value is 0, the forecast is perfect. Correlation efficiency here is used to assess the predictive power of hydrological discharge model. An efficiency of 100% corresponds to a perfect match between model and observations. For best model, the value of these three statistics should be: (MRSE=0.0; MRAE=0.0; CE=100%). The authors believe that these three error statistics along with the visual comparison between observed and

predicted values of three points of Pahang River are sufficient to reveal any significant differences with regard to their performance. The improvement of each model here is measured in part by the change in MRSE,MRAE and CE.

III. Result and Discussion

This section presented the results and analysis of KNN model and MARS model. All model development in this study using 30 years of data sets, where one year data of 2003 was used for validation process while the others were applied for calibration. Data set of upstream of Pahang River was used as training data sets to obtain the optimum model for each approach. The optimum model that obtained from river upstream analysis were used as model to predict further flood at middle stream and downstream of Pahang river. The improvement of performance of each models in variable length input data were also analyzed to determine the effect of longer period data to the performance of each model.

A. *Training sets*

Based on 10% of data sampling in KNN model, cross validation method was implemented for a selection Number of nearest neighbours, K, ranging from 1to 35 to get the optimum number of K . The model performance of KNN in term of MRSE, MRAE and CE for one-year-ahead flood prediction 2003 was summarized in table 1.

Table 1: Model Performance Statistic of KNN (Upstream-Sungai Yap)

Data Set	$K_{Optimum}$	MRSE	MRAE	CE
5 Years	5	0.83	0.47	7.37
10 Years	3	0.28	0.35	29.01
20 Years	5	0.21	0.33	38.84
30 Years	**5**	**0.18**	**0.33**	**41.43**

As seen from table 1, MRSE, MRAE and CE of model for each data set were presented. There was improvement due to the increasing of period data sets for this model but is minor. The increasing number of K also does not give any advantages to the performance of the model and the optimum value of K here only in range from 3 to 5. KNN produced an unsatisfactory CE-value range from 0.07 to 0.41, MRSE value range from 0.83 to 0.18, and MRAE from 0.47 to 0.33. The highlight row in table 1 was selected as the optimum model of KNN.

With regard to MARS model, the number of basic function was setted in ranging from 1-100 basic function. The value of threshold also has been reduced to increase the development of number of basic function. The performance of MARS model for one-year-ahead flood prediction 2003 in term of MRSE, MRAE and CE for Sungai Yap at Pahang River was summarized in table 2 below. The significant improvement of MARS model due to the longer period of data sets can be seen clearly from table 2.

Table 2: Model Performance Statistic of MARS (Upstream-Sungai Yap)

Data Sets	$BF_{Optimum}$	MRSE	MRAE	CE
5 Years	34	0.19	0.27	80.53
10 Years	48	0.07	0.2	90.5
20 Years	66	0.05	0.15	95.94
30 Years	**87**	**0.03**	**0.11**	**98.19**

The basic factor that affects the MARS model performance is a number of basis functions. Performance statistic of MARS model improved due to the increasing of number of basis functions, while number of basis functions increased due to the longer period of data set. Process of adding basis function to this model always helps in improving the CE, MRSE and MRAE. In this study, MARS model gave satisfactory result of a CE value range from 0.81 to 0.98, MRSE value range from 0.19-0.05, and MRAE value range from 0.27-0.15. The highlight row in table 2 is the optimum model of MARS.

B. *Testing Sets*

As mentioned earlier, the optimum model of KNN and MARS at upstream were used to predict flood at midlle stream and downstream of Pahang River. The testing result and comparison performance between these models were presented in table 3 and illustrated in Figure 4. When comparing the KNN and MARS models, the best percentage error for KNN was not as good as the best for MARS model. This is also supported by the correlation efficiency values and the level improvement that increased due to the longer period of data sets by MARS model. For better understand the improvement of KNN model and MARS model due to the longer period of data sets are shown in Figure 3.

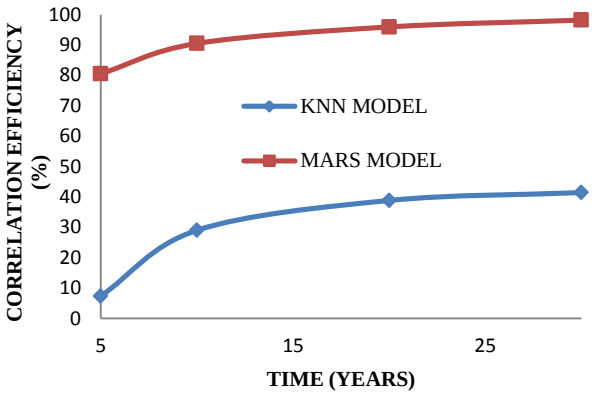

Figure 3: Relationships between CE and Length of Data Sets

Table 3: Comparison of performance statistic between KNN Model and MARS Model

Optimum Model	K-Nearest Neighbour (KNN)			Multivariate Adaptive Regression Splines (MARS)		
Performance Statistic	MRSE	MRAE	CE	MRSE	MRAE	CE
Training Set- Sungai Yap	0.18	0.33	41.43	0.03	0.11	98.19
Testing Set- Sungai Temerloh	0.5	0.5	34.63	0.02	0.10	98.45
Testing Set- Sungai Lubok Paku	0.40	0.40	22.89	0.02	0.10	98.07

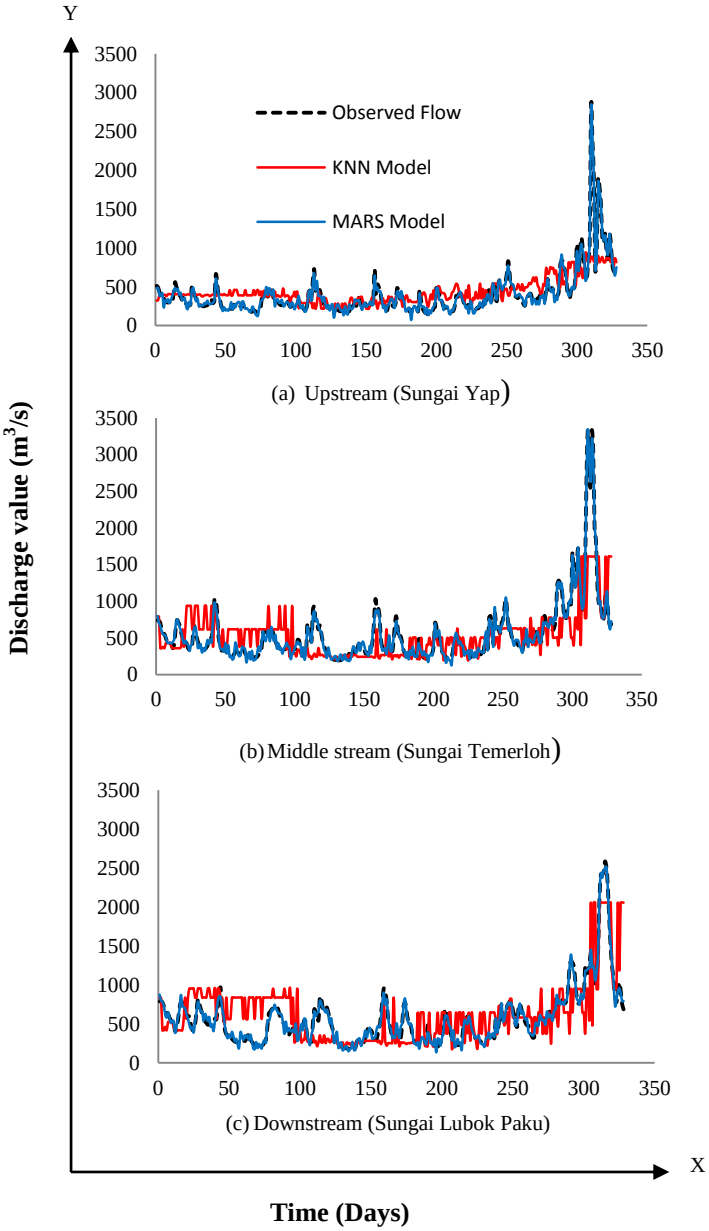

Figure 4: Comparison of observed and simulated discharged at Pahang River catchment.

IV. Conclusion

The study of of potential of KNN model and MARS model in predicting flood at Pahang River has succesfully demonstrated. The effects of different length of record data to the performance of these two models were also studied. It was revealed that both of performance of these two models can be improved by using longer period of data sets. However, based on the comparison study between KNN model and MARS model, all the best value error and efficiency percentage for KNN was not as good as the best for MARS model.This clearly indicated that MARS model shows a better performance than KNN model. Therefore, it can be concluded that MARS model can promising a best flood prediction result for this study area.

V. Acknowledgement

The authors would like to thanks the Ministry of Higher Education Malaysia and Research Management Institute (RMI), Univeriti Teknologi MARA Malaysia for technical and financial support. We also thank the Department of Irrigation and Drainage (DID) Malaysia for supplying hydrologic data.

VI. References

[1] Abraham, A., Steinberg, D., 2001a, MARS: still an alient planet in soft computing? Int. conf. Comput.Sci.2, 235-244.

[2] Abraham, A., Steinberg D., 2001b. Is neural network a reliable forecaster on earth? A MARS query!IWANN 2, 679-686.

[3] Ames, D.P., "Seasonal to Interannual Streamflow Forecasts Using Nonlinear Time Series Mehods and Climate Information", M.S thesis, Utah State University, Logan, Utah, 1998.

[4] Azmi,M., Araghinejad, S., and Kholghi, M.,(2010), **Multi Model Data Fusion for Hydrological Forecasting Using K-Nearest Neighbour Method**, *Iranian Journal of Science & Technology*,34: 81-92

[5] Briand, L.C.,Freimut, B., and Vollei, F., (2004), **Using Multiple Adaptive Regression Splines to Support Decision Making in Code Inspections**, *The Journal of Systems and Software*, 73: 205-217

[6] DID (Department of Irrigtion and Drainage Malaysia), (2003), Volume 2: Updating of Condition of Flooding in Malaysia , Annual Report 2003, Malaysia.

[7] Dwinnell, W., (2000), *Exploring MARS: an Alternative to Neural Networks*, www.salford-systems.com.

[8] Eskandarinia, A., Nazarpour, H., Teimouri, M., and Ahmadi, M.Z., (2010), **Comparison of Neural Network and K-Nearest Neighbor Methods in Daily Flow Forecasting**, *Journal of Applied Science*, 10(11): 1006-1010.

[9] Friedman, J.H., (1991) **, Multivariate adaptive regression splines**, *The Annals of Statistics*, 19(1): 1–141.

[10] Galeati, G., (1990), **A Comparison of Parametric and Non-parametric Methods for Runoff Forecasting**, *Hydrological Sciences – Journal*, 35(1): 79-94.

[11] Ghani, A.A., Chang, C.K., Leow, C.S., and Zakaria, N.A.,(2012), **Sungai Pahang Digital Flood Mapping: 2007 Flood**, *International Journal of River Basin Management*, 10:2,139-148

[12] Hastie,T., Tibshirani, R.,and Friedman, J.,(2001**), The Element of Statistical and Learning**, **Second Edition**, *Springer*, California.

[13] Karlsson, M., Yakowitz, S., (1987a), **Nearest Neighbor method for Nonparametic Rainfall- Runoff forecasting**, *Water Resour. Re,* 23 (7), pp 1300-1308.

[14] Karlsson, M., Yakowitz, S., (1987b), **Rainfall-Runoff Forecasting method, old, and new,** *Stochastic Hydrol. Hydraul,*, 1, pp 303-308.

[15] Knight, D.W., and Shamseldin, A.Y., (2006), **River Basin Modelling for Flood Risk Mitigation**, *Taylor & Francis/Balkema*, London, UK.

[16] Lall, U., and A. Sharma (1996), A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series, Water Resour. Res., 32(3), 679–693, doi:10.1029/95WR02966.

[17] Lall, U., and Keppenel, C., (1996), Complex singular spectrum analysis and multivariate adaptive regression splines applied to forecasting the Southern Oscillation, Retrieved Mac 05, 2011 from http://www.cpc.ncep.noaa.gov/products/predictions/experimental/bulletin/Mar96/article13.html

[18] Lewis, P.A.W., and Ray, B.K., (1993), "Nonlinear Modeling of Multivariate and Categorical Time Series using Multivariate Adaptive Regression Splines" in *Dimension Estimation and Model*, Ed. H.Tong, Singapore: World Scientific.

[19] Moore, R.J., and Bell, V.A., "Comparison of Rainfall-Runoff Model for Flood Forecasting", Environment Agency, Owen Wegwood, North West Region,Tech. Rep, 2001.

[20] Myatt, G.J.,(2007), Making Sense of Data; A Practical Guide to Exploratory Data Analysis and Data Mining, *John Wiley &Sons,Inc*, Canada.

[21] Shamseldin, A.Y.,and O'connor, K.M., (1996), **A Nearest Neighbour Linear Perturbation Model for River Flow Forecasting**, *Journal of Hydrology*, 179: 353-365.

[22] Sharda, V.N., Patel, R.M., Prasher, S.O., Ojasvi, P.R., and Prakash, C., (2006), **Modelling runoff from middle Himalayan watersheds employing artificial intelligence techniques**, *Journal of Agricultural Water Management*, 83: 233-242.

[23] Sharda, V.N., Prasher, S.O., Patel, R.M., Ojasvi P.R., and Prakash, C., (2010), **Performance of Multivariate Adaptive Regression Splines (MARS) in predicting runoff in Mid-Himalayan micro-watersheds with limited data**, *Hydrological Sciences Journal*,53:6, 1165-1175.

[24] Toth, E., Brath, A., and Montanari, A., (2000), **Comparison of short term rainfall prediction models for real-time flood forecasting**, *Journal of Hydrology*, 239:132-147.

[25] Yaacob, M.S., Jamaluddin, H., and Harun, S., 2006, **Daily Streamflow Forecasting Using Simplified Rule-Based Fuzzy Logic System**, *Journal - The Institution of Engineers Malaysia*,66:4.

[26] Zhou, Y., and Leung, H., (2007),Predicting Object-Oriented Software Maintainability Using Multivariate Adaptive Regression Splines, *Journal of Ssytems and Software*, 80: 1349-1361.