International Journal of Engineering Research and Development
e-ISSN : 2278-067X, p-ISSN : 2278-800X
Volume 2, Issue 5 (July 2012), PP. 01-12

Forecasting Stream Flow using Support Vector Regression and M5 Model Trees

Shreenivas N. Londhe, Pradnya R. Dixit
Department of Civil Engineering, Vishwakarma Institute of Information Technology, Survey No. 2/3/4, Kondhwa (Bk), Pune, MH 411048. India.

Abstract*—The paper presents use of two recent data driven techniques namely Model Tree and Support Vector Regression to forecast stream flow at two different locations one in Narmada river basin and the other location in Krishna river basin of India. The stream flow models are developed using the previous values of measured stream flow and rainfall to forecast stream flow one day in advance. All the models (total 63) show reasonable accuracy as evident by high values of correlation coefficient, coefficient of efficiency and low value of root mean square error. Additionally scatter plots and hydrographs were also drawn to asses the model performance. The models developed using Support Vector Regressions performed better compared to MT models. The effect of rainfall as an input for forecasting stream flow was also investigated.

***Keywords*—Data driven modelling, model trees, stream flow forecasting, support vector regression**

I. INTRODUCTION

Data-driven modeling can be considered as an approach that focuses on using the Machine Learning methods in building models that would complement or replace the "knowledge-driven" models describing physical behavior [1. Examples of the most common methods used in data-driven modeling of river basin systems are: statistical method which is perhaps the oldest followed by the techniques which try to imitate human brain and perception like Artificial Neural Networks, fuzzy rule-based systems, Genetic Programming, Support Vector Machines (or regression) and Model Trees [2]. Due to availability of data these methods are gaining popularity since last two decades or so. ANN is now an established technique in the field of Hydrology as it is being used since last 2 decades or so. However it has been shown by [3] and [4] that a technique of Model Tree can also yield comparable results for stream flow forecasting. The recent technique of Support Vector Machines (SVM) has shown promising results in forecasting of waves by [5] and multi-time scale stream flow predictions by [6]. The present work aims at forecasting of stream flow one day in advance at two stations in India using previous values of stream flow and rainfall and the techniques of Support Vector Regression (SVR) and M5 Model Tress (MT). The first station, Rajghat is in the Narmada river basin of India and the second one at Paud, in the Krsihna river basin of India. Results of both approaches will be compared for accuracy of prediction. It is a first work of its kind where in MT and SVR are compared for stream flow forecasting to our knowledge. The succeeding section will present brief information about SVR and MT techniques along with their applications particularly for stream flow forecasting followed by section on study area and data. The model formulation will be discussed later followed by results and discussion. The concluding remarks will be presented in the last.

II. SUPPORT VECTOR REGRESSION (SVR)

The support vector machine (SVM), introduced by [7] is a technique used in pattern recognition and is one of the most attractive forecasting tools in recent years, but applications are rare in the field of civil engineering. Support vector machines are methods of supervised learning, which are commonly used for classification and regression purposes. Their formulation embodies the Structural Risk Minimization (SRM) Principle, which has been shown to be superior to traditional Empirical Risk Minimization (ERM) Principle, employed by many of the other modeling techniques like ANN.

SRM minimizes the error on the training data. It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning. SVMs were first developed to solve the classification problem, but recently they have been extended to the domain of regression problems. A SVM constructs a separating hyperplane between the classes in the n-dimensional space of the inputs. This hyperplane maximizes the margin between the two data sets of the two input classes. The margin is defined as the distance between the two parallel hyperplanes, on each side of the separating one, pushed against each of the two datasets. Simply, the larger the margin, the better the generalization error of the classifier would be. For the case of regression, the only difference is that SVM attempts to fit a curve, with respect to the kernel used in the SVM, on the data points such that the points lie between the two marginal hyperplanes as much as possible, the aim is to minimize the regression error. In the present work a least squares version of SVM's for the function estimation problem (LS-SVR) of prediction of stream flow is used. While in classical SVM's many support values are zero (nonzero values correspond to support vectors), in least squares SVM's the support values are proportional to the errors.

Radial Basis Kernel is used for calibrating the SVR models. Readers are referred to [8] for details of SVM and SVR. The details are not provided in this paper to avoid repetition. [9] used support vector machines for Lake water level prediction.[10] used SVM for long term discharge prediction. [11] employed SVR for real time fold stage forecasting. [6]

used SVR for multi-time scale stream flow predictions. Recurrent support vector machines were used by [12] for rainfall forecasting. [13] used distributed Support Vector Regression for river stage prediction.

III. M5 MODEL TREE (MT)

The M5 Model Tree is a data driven method based on the idea of decision tree that follows the principle of recursive partitioning of input space using entropy-based measures, and finally assigning class labels to resulting subsets. M5 algorithm splits the parameter space into areas (subspaces) and builds in each of them a local specialized linear regression model. The splitting in MT follows the idea used in building a decision tree, but instead of the class labels it has linear regression functions at the leaves, which can predict continuous numeric attributes. Model trees generalize the concepts of regression trees which have constant values at their leaves. So, they are analogous to piecewise linear functions (and hence non-linear). Model trees learn efficiently and can tackle tasks with very high dimensionality — up to hundreds of attributes. The major advantage of model trees over regression trees is that model trees are much smaller than regression trees, the decision strength is clear, and the regression functions do not normally involve many variables [14]. The M5 algorithm is used for inducing a model tree. Readers are referred to [15] for details of the procedure. In Hydrology the application of M5 Model Trees is relatively new and more research in this field is called for. The works published so far are by [3], [16], [17], [14] and [4] which are related to rainfall-runoff modeling, flood forecasting, flow predictions, water level discharge relationship and forecasting of runoff. As discussed in 'Introduction' the present work aims at comparing results of both the approaches for forecasting of stream flow one day in advance at two locations distinctly apart in the country of India.

IV. STUDY AREA AND DATA

The present work deals with forecasting of stream flow at 2 stations, namely Rajghat and Paud in Narmada river basin and Krishna river basin of India respectively. Narmada, the largest west flowing and seventh largest river in India, covers a large area of Madhya Pradesh state besides some area of Maharashtra state & Gujarat state before entering into the Gulf of Camby, Arabian Sea. Narmada Basin lies between East Longitudes 72^0 32' to 82^0 45' and North Latitudes 21^0 20' to 23^0 45'. The total catchment area covered is 98796 Sq.Km. The observations of daily average stream flow values and rainfall pertained to Rajghat on Narmada River were available from the records of the Central Water Commission, Bhopal, India for the years of 1987 to 1997. India is peculiar by its monsoon season in which it receives rainfall almost for 4 months all over the country. The Narmada catchment receives rainfall starting from late June continuing till early October.

Krishna Basin is India's forth-largest river basin, which covers 258,948 Km^2 of southern India. Krishna river originates in the Western Ghats at an elevation of about 1337 m just north of Mahabaleshwar in Maharashtra, India about 64 km from the Arabian Sea and flows for about 1400 km and outfalls into the Bay of Bengal traversing three states Karnataka (113,271 Km^2), Andhra Pradesh (76,252 Km^2), Maharashtra (69,425 Km^2). The selected rain gauge and discharge station Paud is on Mula river in the Pune district of Maharashtra State of India. The data was collected by Surface and Ground Water Hydrology department, through Hydro-Project, Nasik [18] . Total fourteen years data for daily rainfall and discharge (stream flow) from the year 1994 to 2007 was available for developing the stream flow models. The location under consideration receives rainfall in the monsoon months starting from early June continuing till early September.

V. MODEL FORMULATION

After examining the data it was found that the average discharge values for the monsoon months of July to October were differing considerably. Table 1 shows statistical parameters of the observed flow at Rajghat and Paud for the months of July to October (July to September for Paud) which indicates a large variation as a result of which separate monthly models were decided to be developed for the months of July, August, September and October (July, August and September for Paud). The next task was to determine the number of antecedent discharges as well as rainfalls to be used for predicting discharge one day in advance. Input data selection can be done in a variety of ways such as by noticing the significant lag effect through evaluation of the serial correlations, saliency analysis, and cross-correlation statistics and also by trials. The method of trials was used in this work as it is simpler and does not assume linearity underlying the definition of the correlation coefficients. For a monthly model it was started with 2 previous values of discharges (the current day and one previous day) as inputs to predict discharge of the next day. To these two values of discharges then 2 previous values of rainfall (the current day and one previous day) were added one by one while discharge on the next day was maintained as output. In the next step 3 and 4 previous values of discharges (the current day and two and three previous days respectively) were used one by one. Finally these discharges were also supplemented by 2 previous rainfall values as above though the out put was discharge on the next day in every case. It was found that addition of any further stream flow or rainfall value (previous) did not further improve the accuracy of the developed models in prediction of stream flow for both SVR and MT. Thus in all 9 models were developed for one month. In functional form it can be stated as

$Q_{t+1} = f(Q_t, Q_{t-1})$
$Q_{t+1} = f(Q_t, Q_{t-1}, R_t)$
$Q_{t+1} = f(Q_t, Q_{t-1}, R_t, R_{t-1})$
$Q_{t+1} = f(Q_t, Q_{t-1}, Q_{t-2})$
$Q_{t+1} = f(Q_t, Q_{t-1}, Q_{t-2}, R_t)$
$Q_{t+1} = f(Q_t, Q_{t-1}, Q_{t-2}, R_t, R_{t-1})$
$Q_{t+1} = f(Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3})$
$Q_{t+1} = f(Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t)$
$Q_{t+1} = f(Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1})$

Where Q = discharge and R = Rainfall

For four months of July, August, September and October total 36 models were developed at Rajghat. For Paud in Krishna river basin data for the month of October was not available and hence models for July, August and September months were developed (total 27). Each model was calibrated using approximately 70% of data and the remaining 30% of data was used to test the model. Details of all the models viz., inputs, training and testing data for Rajghat and Paud are shown in table 2 and 3 respectively. The Least Square – Support Vector Machines toolbox based on MATLAB is used in the present work. The readers are directed to [19] for details of LS-SVM and [20] for the LS-SVM tool box. The models were developed using M5 Model Trees following the same data division to compare the results of both the approaches. The software WEKA developed by University of Waikato, New Zealand was used to develop M5 Model Trees models.

VI. MODEL ASSESSMENT

Results of the developed models in testing were assessed by plotting the scatter plot between the observed and predicted flow and drawing the hydrograph of observed and predicted stream flow by both the approaches. The coefficient of correlation between the observed and predicted stream flow was also calculated to judge the accuracy of model prediction quantitatively. Need for more than one model assessment technique has been emphasized by [21] Dawson and Wilby (2001). Accordingly, two conventional evaluation criteria, RMSE (root mean square error) and E (coefficient of efficiency), were used in the present study to measure the performances of models in testing.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_{\hat{i}}\right)^2}{n}} \quad (1)$$

$$\mathbf{CE = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} \quad (2)$$

Where, Y = observed stream flow, $\hat{Y}$ = Predicted Stream flow, $\overline{Y}$ = mean of the observed stream flow, n = total number of target and output pairs.

RMSE provides a quantitative indication of the model absolute error in terms of the units of the variable, with the characteristic that larger errors receive greater attention than smaller ones. This characteristic can help eliminate approaches with significant errors [13]. Lower the value of RMSE, greater is the accuracy. CE does not depend on data scale and hence is more suited when different scales are involved [21]. CE indicates prediction capabilities of values different from the mean and varies from -∞ to +1. CE of 0.9 and above is very satisfactory and that of below 0.8 is unsatisfactory.

VII. RESULTS AND DISCUSSION

The calibrated models were tested with unseen values of inputs and results were obtained. For Rajghat models it was found that the influence of previous values of stream flow on the prediction accuracy was fluctuating for SVR models. Table 4 shows results of all the models developed for Rajghat (Model 1 to 36) with both SVR and MT approaches. For July models (Models 1, 4, 7) the prediction accuracy decreased with increase in previous values of stream flow as inputs (r = 0.9, 0.86 and 0.84 respectively). For August (Model 10, 13, 16) similar performance was observed (r = 0.73, 0.7 and 0.68 respectively). However for September (Model 19, 22, 25) the r value increased for one additional stream flow as input (0.74 to 0.83) but then decreased for further addition of stream flow as input (0.83 to 0.79). For October the model accuracy first decreased slightly (0.91 to 0.9) and then increased (0.9 to 0.92) with increase in the stream flow values as inputs. It can also be observed that by adding 1 or 2 values previous values of rainfall as inputs did not necessarily improve the stream flow forecasting accuracy almost for all the models. On the other hand in some models (for eg. Model No. 25, 26, 27) accuracy of prediction decreased as indicated by decreasing values of the correlation coefficient from 0.79 to 0.75 to 0.69.

For models developed using M5 Model tree algorithm a similar fluctuating trend was observed for models with previous stream flow used as input in that the accuracy of prediction neither increased nor decreased consistently with increase in number of previous stream flow values as inputs. Similarly addition of rainfall values as inputs did not improve the accuracy of prediction significantly though it did not decrease as in case of SVR models. The reason may be duplication of information by adding rainfall as input along with stream flow values. The cause of stream flow is rainfall particularly in the monsoon months and the effect (stream flow) is being modeled by using both cause and effect (rainfall and runoff) simultaneously making the rainfall information ineffective in predicting the stream flow. One more reason may be absence of rainfall values in many data sets particularly in the month of October which shows constant results after addition of rainfall values. It may also be noted that rainfall does not occur on all days of monsoon months.

When results of both the approaches were compared it was found that SVR is superior to MT in majority of models as evident by higher 'r' and 'CE values and less 'RMSE' values. When results of each monthly model were compared for SVR approach, model 1(ip2july) with 2 previous values of stream flow as inputs seems to be the best with highest 'r' and CE values and lowest RMSE value. However in MT approach model 8 (ip4rtjuly) with 4 previous values of stream flow and one previous rainfall as input seems to be the best though it was less accurate as compared to Model 1

(ip2july). For the month of August, model 10 (ip2august) with 2 previous values of stream flow as inputs wins the race for SVR model where as model 13 (ip3august) was the best amongst MT models. In August models MT results were at par with SVR results except RMSE which slightly more for MT model. For the month of September model 24 (ip3rtrt-1september) was the best in SVR models while model 22 (ip3sept) wins the race in MT models. In this case also results of SVR model (24) are better than MT model (22). For the month of October results of 34 (ip4oct) were the best for both SVR and MT approaches with SVR results slightly better. It may be noted that results are not mentioned in brackets to avoid repetition of information provided in table 4. Thus for July model 1, for August model 10, for September model 24 and for October model 34 all with SVR approach are the best though results of MT are not far behind. Out of all the models (both SVR and MT) model 34 (ip4oct) developed using SVR approach seems to be the best with highest 'r' (0.92), Highest 'CE' (0.84) and lowest 'RMSE' (105.68 m^3/s). Same model developed using MT shows the best results amongst all MT models with r = 0.91, CE = 0.81, and RMSE = 113.38 m^3/s. It may be noted that for both the approaches the same model was the best (model 34) when the other error measures of CE and RMSE were compared. Results of August models are the worst amongst all Rajghat SVR models where as results of July models were found to be the least in MT models. By observing table 1 it seems the average stream flow and standard deviation values play a role in accuracy of the developed models. The month of October which has the least average and the least standard deviation exhibits the best performance by both the approaches.

Figure 1 shows scatter plot for Rajghat July model (model1, ip2 july) which indicates that though both the approaches under predict the extreme events SVR approach is better. Figure 2 shows hydrograph of Rajghat October model (model 34, ip4oct) where in SVR predicts the peak (1582 m^3/s) better (1520 m^3/s) than MT approach (1319 m^3/s).

For Paud models it was found that for the month of July there was an increase in the value of correlation coefficient from 0.79 to 0.82 for SVR models and 0.79 to 0.85 for MT models when the number of previous discharges was increased (Model 37, 40, 43). With addition of two previous values of rainfall one by one as inputs there was again a slight increase in correlation coefficient in each case (Model 37 to 45). Model no. 44 with 4 previous values of discharges and 1 previous rainfall as inputs, (ip4rtjuly) with MT approach seems to be best amongst all July models with correlation coefficient equal to 0.84, coefficient of efficiency equal to 0.64 and root mean squared error equal to 37.61 m^3/s. The SVR model results though not far behind with r = 0.83, CE = 0.64 and RMSE = 37.51 m^3/s. The peak discharge of 328 m^3/s was predicted as 199 m^3/s bye the SVR model and 190 m^3/s by the MT model. Thus SVR is superior in predicting the peak values.

For August however such a trend was not obtained and the 'r' value was increased from 0.55 to 0.83 with addition of 2^{nd} day stream flow values (Model 46 and 49) but then decreased with subsequent addition of one more stream flow value from 0.83 to 0.67 (Model 49 and 52) for SVR approach. However for MT approach values consistently increased from 0.45 to 0.58 for the same models. Figure 4 shows hydrograph of August model in testing. Similarly there was not much change rather a decrease in 'r' values after addition of rainfall values was also observed in many cases of August models (Models 46 to 54). Model No. 49 with 3 previous values discharges as inputs, (ip3august) using SVR approach seems to be best amongst all August models with correlation coefficient equal to 0.83, coefficient of efficiency equal to 0.68 and root mean squared error equal to 51.56 m^3/s. For MT approach model 50 (ip3rtaugust) was found to be the best with r = 0.71, CE = 0.35 and RMSE = 75.46 m^3/s. Maximum discharge of 575 m^3/s was predicted as 500 m^3/s by SVR approach and 229 m^3/s by the MT approach.

For September models it was observed that the prediction accuracy increases with increase in number of previous stream flow values as inputs (Models 55, 58 and 61) with r = 0.72, 0.75 and 0.81 respectively for SVR models. Same trend was noticed for MT models as well with r = 0.73, 0.76 and 0.82 respectively. When the rainfall values were used as additional inputs the accuracy of prediction increased for one previous rainfall value in case of Model 56 for both SVR and MT models. The accuracy was then decreased with addition of one more rainfall value (model No. 57) for both SVR and MT approach. The trend however was not continued for further models (Models 59, 60 and 62, 63), which actually showed decrease in performance with addition of rainfall values. Out of all September models model 56 with 2 previous of discharges and 1 previous rainfall as inputs, (ip2rtsept) was found to be the best with 'r' value of 0.91, CE of 0.81 and very low RMSE of 3.8 m^3/s with SVR approach. The MT model for the same inputs exhibit a similar performance with r = 0.9, CE = 0.81 and RMSE = 3.88 m^3/s. Figure 3 shows hydrograph of model 56 (ip2rtsept) in testing. The maximum discharge of 74m^3/s was predicted as 79 m^3/s by SVR and 61 m^3/s by MT model in this case.

Thus for July model 44 (MT), for August model 49 (SVR) and for September model 56 (SVR) were the best by virtue of their highest 'r' and 'CE' values and lowest 'RMSE' values. The best model out of all Paud models was model 56 (ip2sept) with highest 'r' and 'CE' and lowest 'RMSE'. It can be seen from table 1 that the month of September has the lowest average rainfall as well as the lowest standard deviation. The better performance of the model can perhaps be attributed to these two statistical parameters. Consolidated results of all the Paud models are presented in table 5. Figure 4 shows typical model tree developed for model 34 (ip4oct) and figure 5 shows linear models developed at the leaves (of figure 6).

Based on the above discussion it can be said that the over all performance of SVR models is better as compared to MT models. As mentioned in the 'Introduction' the underlying principal SRM (Structural Risk Minimization) seems to make SVR perform better compared to MT.

VIII. CONCLUDING REMARKS

The paper presented comparison of stream flow models at 2 stations Rajghat in Narmada basin and Paud in Krishna river of India developed using two data driven techniques namely SVR and M5 Model Trees. The models were developed to forecast stream flow one day in advance. All the models performed reasonably well in testing with a few exceptions. The SVR models perform better compared to MT models though marginally as evident by better correlation coefficient, Coefficient of efficiency and Root mean squared error of SVR models. It was found that addition of rainfall

values as inputs with the stream flow values did not improve the model accuracy significantly. The results of models seem to be influenced by average and standard deviation value of stream flow. For prediction of peaks SVR worked better as compared to MT. It can be said that the data driven techniques like Support Vector Regression (SVR) and M5 Model Trees (MT) are worth exploring further at least in the field of Hydrology

REFERENCES

[1]. D.P. Solomatine,, A. Ostfeld, "Data-driven modelling: some past experiences and new approaches", Journal of Hydroinformatics. **10.1**, 3-22, 2008

[2]. D.P. Solomatine, "Data Driven Modelling:paradigm, methods, experiences"' Proc., 5th International Conference on Hydroinformatics, Cardiff, UK, 2002

[3]. D.P. Solomatine, and K. Dulal, "Model tree as an alternative to neural network in rainfall-runoff modeling", Hydrological Sciences. 48(3), 399–41, 2003

[4]. S.N.Londhe, and S.B. Charhate, "Comparison of data driven modeling techniques for river flow forecasting", Hydrological Sciences. 55(7), 1163-1174, 2010

[5]. J. Mahjoobi, and E.A. Mosabbeb, "Prediction of significant wave height using regressive support vector machines", Ocean Engineering. 36(5), 339-347, 2009

[6]. T. Asefa, M. Kemblowski, M. McKee, and A. Khalil, "Multi time scale stream flow predictions: the Support Vector Machines Approach",. Journal of Hydrology. 318, 7-16, 2006

[7]. V.N. Vapnik, "An overview of statistical learning theory", IEEE Transactions on Neural Networks. 10 (5), 988–999, 1999

[8]. Y.B. Dibike, S. Velickov, D. Solomatine, D. and M.B. Abbott, "Model induction with support vector machines: introduction and applications", Journal of Computing in Civil Engineering. 15 (3), 208–216, 2001

[9]. M.S. Khan, and P. Coulibaly, "Application of Support Vector Machine in Lake Water Level Prediction", ASCE Journal of Hydrologic Engineering. 11(3), 199-206, 2006

[10]. J. Lin, C. Cheng, C. and K. Chow, "Using Support Vector Machines for long term Discharge Prediction",. Hydrological Sciences. 51(4), 599-612, 2006

[11]. Yu, Pao-Shan, Chen, Shien-Tsung, Chang, and I-Fan "Support vector regression for real-time flood stage forecasting", Journal of Hydrology. 328, 704– 716, 2006

[12]. P.F. Pai, and W.C. Hong, "A recurrent support vector regression model in rainfall forecasting", Hydrological Processes. 21, 819-827, 2007

[13]. C.L. Wu, K.W. Chau, Y.S. Li, "River stage prediction based on a distributed support vector regression", Journal of Hydrology. 358, 96-111, 2008

[14]. B. Bhattacharya, and D.P. Solomatine, "Neural networks and M5 model trees in modelling water level–discharge relationship", Neurocomputing. 63, 381–396, 2005

[15]. J.R. Quinlan, "Learning with continuous classes", Proc., 5thAustralian Joint Conf. on Artificial Intelligence, Adams & Sterling, eds., World Scientific, Singapore, 343–348, 1992

[16]. D.P. Solomatine, 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. 9(6), 491-501, 2004

[17]. D.P. Solomatine and M.B. Siek, "Flexible and optimal M5 model trees with applications to flow predictions", Proceedings of the Sixth international Conference on Hydroinformatics, June 2004, World Scientific, Singapore, 2004

[18]. www.mahahp.org

[19]. J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B.De. Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. World Scientific, Singapore

[20]. http://www.esat.kuleuven.be/sista/lssvmlab/

[21]. C.W. Dawson, R. L. Wilby, "Hydrological Modeling using Artificial Neural Networks." Progress in Physical Geography, 25(1), 80-108, 2001

Table 1: Statistical Parameters of the Daily Observed Flow Data

Month	**Rajghat**				**Paud**			
	Average (m^3/s)	**Std. deviation**	**skewness**	**kurtosis**	**Average (m^3/s)**	**Std. deviation**	**skewness**	**kurtosis**
July	2306.04	4265.93	3.86	17.81	38.6	62.59	3.45	16.62
August	4179.81	4653.33	3.23	14.72	45.4	74.83	4.25	21.53
September	3066.75	3876.58	7.64	78.81	15.4	21.03	4.61	23.46
October	743.76	704.01	5.63	50.20	-	-	-	-

Table2: Details of Rajghat Models

Sr. No.	Month/Model	Inputs	Training data	Testing data
1.	ip2july	Q_t, Q_{t-1}	223	96
2.	ip2rtjuly	Q_t, Q_{t-1}, R_t	223	96
3.	ip2rtrt-1july	Q_t, Q_{t-1}, R_t, R_{t-1}	223	96
4.	Ip3july	Q_t, Q_{t-1},Q_{t-2}	215	93
5.	Ip3rtjuly	Q_t, Q_{t-1}, Q_{t-2},R_t	215	93
6.	Ip3rtrt-1july	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	215	93
7.	Ip4july	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	207	90
8.	Ip4rtjuly	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	207	90
9.	Ip4rtrt-1july	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	207	90
10.	ip2august	Q_t, Q_{t-1}	223	96
11.	ip2rtaugust	Q_t, Q_{t-1}, R_t	223	96
12.	ip2rtrt-1august	Q_t, Q_{t-1}, R_t, R_{t-1}	223	96
13.	Ip3august	Q_t, Q_{t-1},Q_{t-2}	215	93
14.	Ip3rtaugust	Q_t, Q_{t-1}, Q_{t-2},R_t	215	93
15.	Ip3rtrt-1august	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	215	93
16.	Ip4august	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	207	90
17.	Ip4rtaugust	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	207	90
18.	Ip4rtrt-1august	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	207	90
19.	ip2sept	Q_t, Q_{t-1}	215	93
20.	ip2rtsept	Q_t, Q_{t-1}, R_t	215	93
21.	ip2rtrt-1sept	Q_t, Q_{t-1}, R_t, R_{t-1}	215	93
22.	Ip3sept	Q_t, Q_{t-1},Q_{t-2}	207	90
23.	Ip3rtsept	Q_t, Q_{t-1}, Q_{t-2},R_t	207	90
24.	Ip3rtrt-1sept	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	207	90
25.	Ip4sept	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	200	86
26.	Ip4rtsept	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	200	86
27.	Ip4rtrt-1sept	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	200	86
28.	ip2oct	Q_t, Q_{t-1}	203	87
29.	ip2rtoct	Q_t, Q_{t-1}, R_t	203	87
30.	ip2rtrt-1oct	Q_t, Q_{t-1}, R_t, R_{t-1}	203	87
31.	Ip3oct	Q_t, Q_{t-1},Q_{t-2}	196	84
32.	Ip3rtoct	Q_t, Q_{t-1}, Q_{t-2},R_t	196	84
33.	Ip3rtrt-1oct	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	196	84
34.	Ip4oct	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	189	81
35.	Ip4rtoct	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	189	81
36.	Ip4rtrt-1oct	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	189	81

Table 3 Details Of Paud Models

Sr. No.	Month/Model	Inputs	Training data	Testing data
37.	ip2july	Q_t, Q_{t-1}	247	104
38.	ip2rtjuly	Q_t, Q_{t-1}, R_t	247	104
39.	ip2rtrt-1july	Q_t, Q_{t-1}, R_t, R_{t-1}	247	104
40.	Ip3july	Q_t, Q_{t-1},Q_{t-2}	225	97
41.	Ip3rtjuly	Q_t, Q_{t-1}, Q_{t-2},R_t	225	97
42.	Ip3rtrt-1july	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	225	97
43.	Ip4july	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	210	97
44.	Ip4rtjuly	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	210	97
45.	Ip4rtrt-1july	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	210	97
46.	ip2august	Q_t, Q_{t-1}	250	107
47.	ip2rtaugust	Q_t, Q_{t-1}, R_t	250	107
48.	ip2rtrt-1august	Q_t, Q_{t-1}, R_t, R_{t-1}	250	107
49.	Ip3august	Q_t, Q_{t-1},Q_{t-2}	220	94
50.	Ip3rtaugust	Q_t, Q_{t-1}, Q_{t-2},R_t	220	94
51.	Ip3rtrt-1august	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	220	94
52.	Ip4august	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	228	98
53.	Ip4rtaugust	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	228	98
54.	Ip4rtrt-1august	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	228	98
55.	ip2sept	Q_t, Q_{t-1}	173	75
56.	ip2rtsept	Q_t, Q_{t-1}, R_t	173	75
57.	ip2rtrt-1sept	Q_t, Q_{t-1}, R_t, R_{t-1}	173	75
58.	Ip3sept	Q_t, Q_{t-1},Q_{t-2}	147	63
59.	Ip3rtsept	Q_t, Q_{t-1}, Q_{t-2},R_t	147	63
60.	Ip3rtrt-1sept	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	147	63
61.	Ip4sept	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	142	80
62.	Ip4rtsept	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	142	80
63.	Ip4rtrt-1sept	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	142	80

Table 4: Results of Rajghat Models

Sr. No.	Month/Model	Inputs	r_{SVM}	CE	RMSE	r_{MT}	CE	RMSE
1.	ip2july	Q_t, Q_{t-1}	**0.9**	**0.81**	**2328.11**	0.64	0.36	4228.93
2.	ip2rtjuly	Q_t, Q_{t-1}, R_t	0.87	0.75	2642.31	0.67	0.41	4085.3
3.	ip2rtrt-1july	Q_t, Q_{t-1}, R_t, R_{t-1}	0.86	0.73	2757.95	0.65	0.36	4254.21
4.	ip3july	Q_t, Q_{t-1},Q_{t-2}	0.86	0.73	2767	0.65	0.35	4335.26
5.	iIp3rtjuly	Q_t, Q_{t-1}, Q_{t-2},R_t	0.86	0.74	2773.71	0.65	0.36	4328.28
6.	ip3rtrt-1july	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	0.8	0.62	3321.62	0.65	0.36	4328.28
7.	ip4july	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	0.84	0.71	2944.71	0.65	0.35	4393.48
8.	ip4rtjuly	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	0.8	0.62	3386.27	**0.79**	**0.63**	**3318.96**
9.	ip4rtrt-1july	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	0.81	0.61	3399.5	0.65	0.35	4393.48
10.	ip2august	Q_t, Q_{t-1}	**0.73**	**0.52**	**2701.3**	0.7	0.48	2817.34
11.	ip2rtaugust	Q_t, Q_{t-1}, R_t	0.72	0.48	2817.35	0.7	0.48	2817.34
12.	ip2rtrt-1august	Q_t, Q_{t-1}, R_t, R_{t-1}	0.72	0.5	2749.88	0.7	0.48	2817.34
13.	ip3august	Q_t, Q_{t-1},Q_{t-2}	0.7	0.47	2825.17	**0.73**	**0.51**	**2743.46**
14.	ip3rtaugust	Q_t, Q_{t-1}, Q_{t-2},R_t	0.68	0.44	2950.23	0.7	0.49	2803.16
15.	ip3rtrt-1august	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	0.71	0.49	2797.88	**0.73**	**0.51**	**2743.46**
16.	ip4august	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	0.68	0.43	3000.79	0.73	0.51	2789.48
17.	ip4rtaugust	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	0.69	0.46	2927.77	0.73	0.51	2789.48
18.	ip4rtrt-1august	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	0.7	0.43	3011.8	0.73	0.51	2789.48
19.	ip2sept	Q_t, Q_{t-1}	0.74	0.44	1144.67	0.73	0.49	1111.08
20.	ip2rtsept	Q_t, Q_{t-1}, R_t	0.74	0.52	1081.2	0.73	0.49	1111.08
21.	ip2rtrt-1sept	Q_t, Q_{t-1}, R_t, R_{t-1}	0.74	0.49	1113.35	0.72	0.38	1230.26
22.	ip3sept	Q_t, Q_{t-1},Q_{t-2}	0.83	0.52	927.7	**0.77**	**0.53**	**913.35**
23.	ip3rtsept	Q_t, Q_{t-1}, Q_{t-2},R_t	0.84	0.52	924.33	**0.77**	**0.53**	**913.35**
24.	ip3rtrt-1sept	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	**0.84**	**0.61**	**827.47**	0.76	0.27	1137.14
25.	ip4sept	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	0.79	0.53	818.71	0.67	0.37	949.05
26.	ip4rtsept	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	0.75	0.46	877.32	0.68	0.42	911.69
27.	ip4rtrt-1sept	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	0.69	0.43	904.48	0.67	0.37	949.05
28.	ip2oct	Q_t, Q_{t-1}	0.91	0.77	123.92	0.89	0.74	131.43
29.	ip2rtoct	Q_t, Q_{t-1}, R_t	0.91	0.73	133.65	0.89	0.74	131.43
30.	ip2rtrt-1oct	Q_t, Q_{t-1}, R_t, R_{t-1}	0.89	0.73	134.46	0.89	0.74	131.43
31.	ip3oct	Q_t, Q_{t-1},Q_{t-2}	0.9	0.8	117.03	0.87	0.72	137.04
32.	ip3rtoct	Q_t, Q_{t-1}, Q_{t-2},R_t	0.85	0.71	138.93	0.87	0.72	137.04
33.	ip3rtrt-1oct	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	0.85	0.7	141.79	0.87	0.72	137.04
34.	ip4oct	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	**0.92**	**0.84**	**105.68**	**0.91**	**0.81**	**113.38**
35.	ip4rtoct	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	0.9	0.8	119.02	0.91	0.81	113.38
36.	ip4rtrt-1oct	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	0.9	0.79	121.07	0.91	0.81	113.38

Table 5: Results of Paud Models

Sr. No.	Month/Model	Inputs	r_{SVM}	CE	RMSE	r_{MT}	CE	RMSE
37.	ip2july	Q_t, Q_{t-1}	0.79	0.63	37.66	0.79	0.62	38.07
38.	ip2rtjuly	Q_t, Q_{t-1}, R_t	0.8	0.61	36.62	0.82	0.61	38.42
39.	ip2rtrt-1july	Q_t, Q_{t-1}, R_t, R_{t-1}	0.82	0.59	39.15	0.82	0.61	38.42
40.	ip3july	Q_t, Q_{t-1},Q_{t-2}	0.8	0.63	37.94	0.8	0.63	38.18
41.	ip3rtjuly	Q_t, Q_{t-1}, Q_{t-2},R_t	0.81	0.62	38.72	0.83	0.65	37.36
42.	ip3rtrt-1july	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	0.81	0.6	39.64	0.81	0.62	38.61
43.	ip4july	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	0.82	0.66	36.57	0.85	0.7	34.21
44.	ip4rtjuly	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	**0.83**	**0.64**	**37.61**	**0.84**	**0.64**	**37.51**
45.	ip4rtrt-1july	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	0.83	0.62	38.53	0.82	0.66	36.75
46.	ip2august	Q_t, Q_{t-1}	0.55	0.3	79.24	0.45	0.089	90.28
47.	ip2rtaugust	Q_t, Q_{t-1}, R_t	0.53	0.28	80.25	0.45	0.089	90.28
48.	ip2rtrt-1august	Q_t, Q_{t-1}, R_t, R_{t-1}	0.5	0.25	82	0.42	0.03	93.14
49.	ip3august	Q_t, Q_{t-1},Q_{t-2}	**0.83**	**0.68**	**51.56**	0.55	0.28	77.25
50.	ip3rtaugust	Q_t, Q_{t-1}, Q_{t-2},R_t	0.67	0.42	70.19	**0.71**	**0.31**	**75.46**
51.	ip3rtrt-1august	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	0.69	0.41	69.72	**0.71**	0.27	78.09
52.	ip4august	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	0.67	0.44	66.75	0.58	0.31	73.77
53.	ip4rtaugust	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	0.64	0.41	68.25	0.61	**0.35**	**71.76**
54.	ip4rtrt-1august	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	0.65	0.41	68.15	0.62	0.34	71.97
55.	ip2sept	Q_t, Q_{t-1}	0.72	0.52	8.35	0.73	0.51	8.48
56.	ip2rtsept	Q_t, Q_{t-1}, R_t	**0.91**	**0.81**	**3.8**	**0.9**	**0.81**	**3.88**
57.	ip2rtrt-1sept	Q_t, Q_{t-1}, R_t, R_{t-1}	0.7	0.49	8.64	0.73	0.51	8.47
58.	ip3sept	Q_t, Q_{t-1},Q_{t-2}	0.75	0.49	9.42	0.76	0.54	8.97
59.	ip3rtsept	Q_t, Q_{t-1}, Q_{t-2},R_t	0.73	0.51	9.24	0.66	0.43	10
60.	ip3rtrt-1sept	Q_t, Q_{t-1}, Q_{t-2},R_t, R_{t-1}	0.7	0.49	9.43	0.42	0.08	13.82
61.	ip4sept	Q_t, Q_{t-1},Q_{t-2},Q_{t-3}	0.81	0.42	10.04	0.82	0.59	8.47
62.	ip4rtsept	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t	0.72	0.52	9.14	0.6	0.35	10.58
63.	ip4rtrt-1sept	Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}	0.69	0.47	9.53	0.59	0.35	10.58

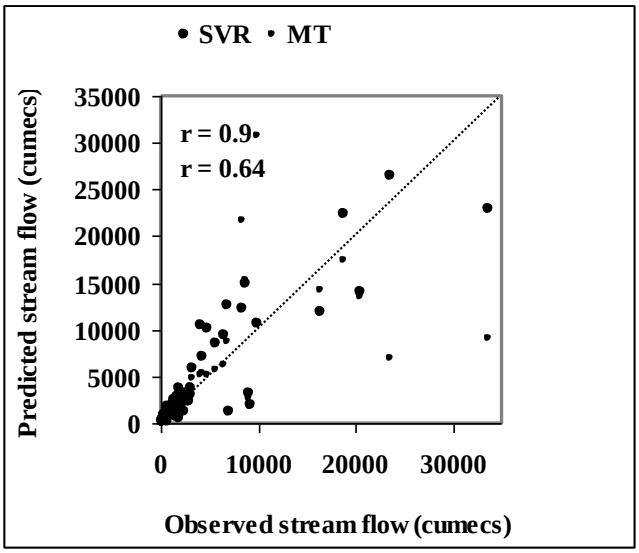

Figure 1 Scatter plot for Rajghat July model (Model 1 ip2july)

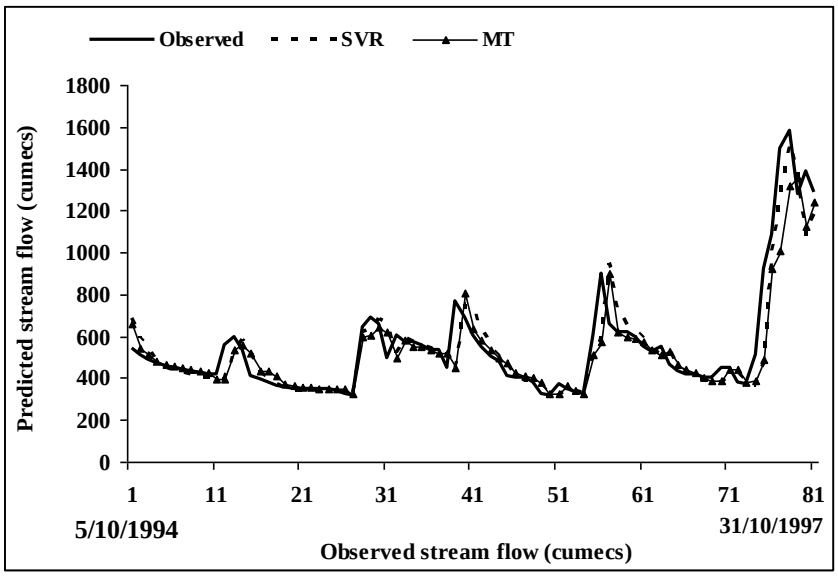

Figure 2 Hydrograph for Rajghat October model (Model 34 ip4oct)

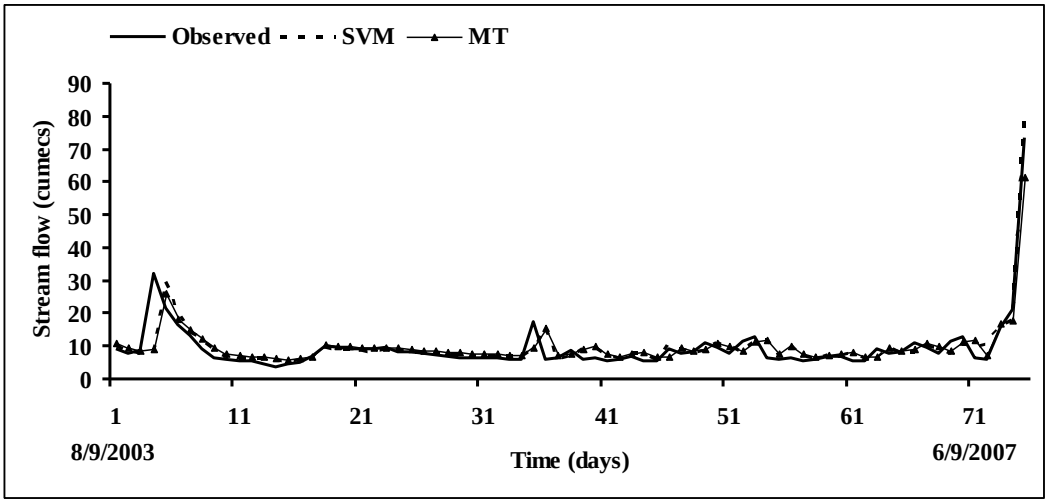

Figure 3 Hydrograph for Paud September model (Model 56 ip3tsept)

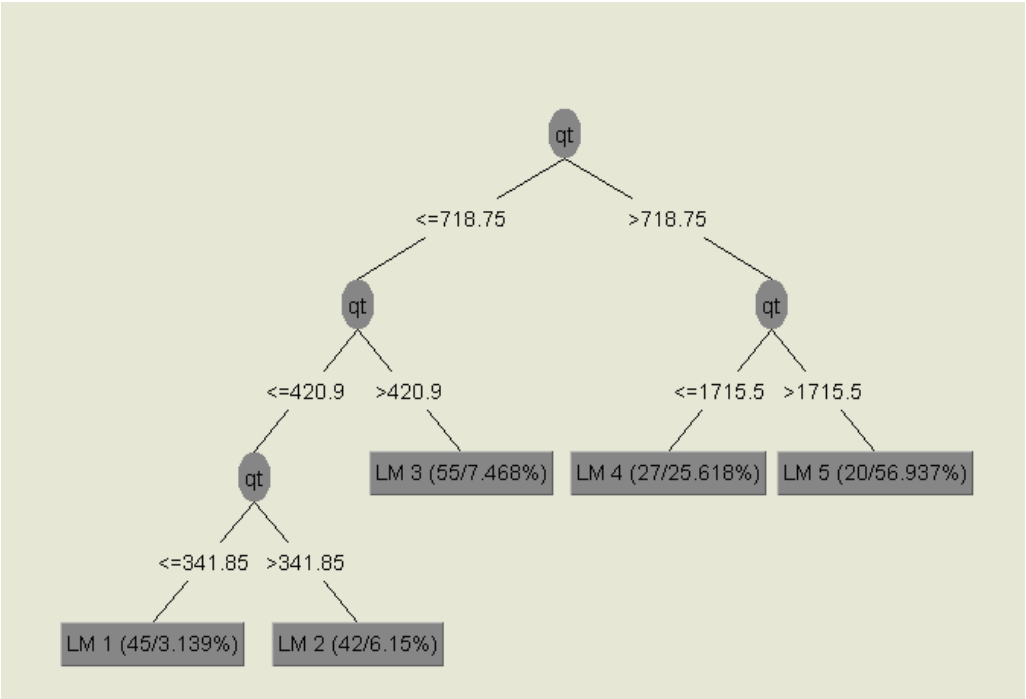

Figure 4 Typical Model Tree for Rajghat October Model (Model 34 ip4oct)

```
qt <= 718.75 :
|   qt <= 420.9 :
|   |   qt <= 341.85 : LM1 (45/3.139%)
|   |   qt >  341.85 : LM2 (42/6.15%)
|   qt >  420.9 : LM3 (55/7.468%)
qt >  718.75 :
|   qt <= 1715.5 : LM4 (27/25.618%)
|   qt >  1715.5 : LM5 (20/56.937%)

LM num: 1
qt+1 = 0.0711 * qt-3  +0.0239 * qt-2 +0.0144 * qt-1 +0.8394 * qt +12.6941

LM num: 2
qt+1 = 0.2877 * qt-3 +0.0239 * qt-2 +0.0144 * qt-1 +0.276 * qt +138.6852

LM num: 3
qt+1 = 0.0077 * qt-3 +0.0331 * qt-2 +0.0144 * qt-1+ 0.7752 * qt +74.261

LM num: 4
qt+1 = 0.0195 * qt-3 +0.0097 * qt-2 - 0.0814 * qt-1 +0.7611 * qt +246.9562

LM num: 5
qt+1 = 0.0195 * qt-3 - 0.0253 * qt-2  +0.0823 * qt-1 +0.3331 * qt +970.4434

Number of Rules : 5
```

Figure 5 MT Models developed for Rajghat October Model (Model 34 ip4oct)

(Note: The first number is the number of samples in the subset sorted to this leave and the second is root mean squared error (RMSE) of the corresponding linear model divided by the standard deviation of the samples subset for which it is built (expressed in percent)).