American Journal of Water Science and Engineering
2016; 2(1): 1-5
Published online March 1, 2016 (http://www.sciencepublishinggroup.com/j/ajwse)
doi: 10.11648/j.ajwse.20160201.11

Multiple Linear Regression Model for Stream Flow Estimation of Wainganga River

Sharad Patel[1], M. K. Hardaha[2], Mukesh K. Seetpal[3], K. K. Madankar[4]

[1]Water Resources Division, Department of Civil Engineering, IIT Bombay, Mumbai, Maharastra, India
[2]Department of Soil and Water Engineering, College of Agricultural Engineering, Jawaharlal Nehru Krishi Vishwa Vidyalaya, Jabalpur, India
[3]Department of Civil Engineering, AKS University, Satna, Madhya Pradesh, India
[4]NM Sadguru Water and Development Foundation, Chosala, Gujarat, India

Email address:

Sharadp56@gmail.com (S. Patel), mkhardaha@rediffmail.com (M. K. Hardaha), mukeshseetpal@gmail.com (M. K. Seetpal), kapil_seonicity@rediffmail.com (K. K. Madankar)

To cite this article:

Sharad Patel, M. K. Hardaha, Mukesh K. Seetpal, K. K. Madankar. Multiple Linear Regression Model for Stream Flow Estimation of Wainganga River. *American Journal of Water Science and Engineering.* Vol. 2, No. 1, 2016, pp. 1-5. doi: 10.11648/j.ajwse.20160201.11

Abstract: This paper is based on the study of multiple linear regression based technique in determining the rainfall-runoff relations. Without considering the temperature, topography or other parameters of the study area, simply using the data of rainfall and runoff to predict the future runoff is the key characteristic of this technique. Ignoring the parameters of temperature, topography may lead to inaccuracy in forecasting future runoff, but the use of Multiple linear regression based technique provide a simple and fast way to determine the runoff. In the present investigation, 14 years data on rainfall and runoff at the Balaghat district of Madhya Pradesh have been analyzed to develop regression models for stream flow estimation with rainfall as input and different regression model tested with varying input length of data record.

Keywords: Multiple Linear Regression Model, Correlation Coefficient, Standard Error

1. Introduction

Rainfall and Runoff data measurement always require wide and complex field work due to a complex hydrological phenomenon of nature and have a greater role in the design of hydraulic structures. The complexities associated with the measurement of these parameters increase with spatial and temporal variability. Rainfall runoff model (RRMs) is a unique tool for hydrological parameter estimation in engineering and environmental science. They are applied to extend stream flow time-series in space and time to evaluate management strategies and/or catchment response to climate and/or land use variability for real-time flood forecasting. Since the rainfall- runoff process is a highly complex, nonlinear, and dynamic hence it is extremely difficult to model.

Regression based statistical model is one of the effective and simplest statistical tool for stream - flow prediction derived from concurrent limited data of rainfall and runoff of the drainage basin if it is gauged (Pilgrim, Chapman, & Doran, 1988). Often, records are not available at the site, and the relation must be derived from a gauged basin of similar size in the region and transferred to the site. Usually monthly or annual yield is required for the design of a water supply for domestic use or irrigation, either with or without storage. In most cases in the arid zone, the drainage basin is of relatively small area. Major water supplies and large storages involving large drainage basins are rarely required in arid and semiarid regions. Similarly, if the basin is large enough to contain a few hundred meters of channel, account should be taken of these losses. Possible approaches using only surface flow data are regression relations between reach inflow and outflow (Lane, Diskin, & Renard, 1971).

Jarboe & Haan, (1974) used a multiple linear regression model to relate four parameters (maximum possible infiltration rate, the maximum possible seepage rate, the maximum capacity of that part of the soil's moisture- holding capacity, which is less readily available for evapotranspiration, and the constant defining the fraction of seepage that becomes runoff.) of Haan's (1972) model and measurable catchment characteristics.

Magette, Shanholtz, & Carr, (1976) used Jone's, (1976) procedure to fit a subset of six parameters of the Kentucky watershed model, and were able to obtain acceptable multiple regression equation using indices of 15 watershed characteristics. A study assessing commonly used floodfrequency methods compared deterministic and regression models for determining peak flood flow frequencies for rural ungauged watersheds (Newton & Herrin, 1983). Similarly a study comparing results of deterministic and regression models of storm- runoff loads and volumes in Denver, Colorado, indicated that neither type of model had consistently been more accurate than the other when applied to a particular basin (Lindner - Lunsford & Ellis, 1987).

Xu, Seibert, & Halldin, (1996) developed relationships between parameters of a monthly water balance model and the land-use data for Swedish catchments. Regression equations were used to calculate model parameters from catchment characteristics. Simulation of annual runoff based on these parameter values showed an average absolute error of prediction for nine catchments to be less than 1% with a maximum error of 20% for one catchment. Asati & Rathore, (2012) developed an autoregressive model, multiple linear regression model and ANN for a complex non-linear relationship between rainfall as input data and output as runoff, without considering the nature of process and compared the performance of each model.

This paper presents an approach that combines rainfall-runoff data for generation of multiple linear regression rainfall- runoff models for stream flow estimation. A particular advantage of this statistical model is limited hydrological data (except rainfall and runoff) required without considering the physics of hydrological process of the catchment.

2. Study Area

The study has been carried for Wainganga River in Godavari basin. The stream flow model has been developed based on the stream flow data recorded at Kumhari gauging station near Balaghat of Madhya Pradesh. The location of the Wainganga sub-basin is shown in Figure 1Kumhari is located at Latitude of N 21º52'58"and longitude of E 80º10'41" on Kumhari road. The total catchment area at Kumhari outlet is 8070 sq. km. The Wainganga is a river in Madhya Pradesh, which originates about 12 km from Mundara village of the Seoni district in the southern slopes of the Satpura Range of Madhya Pradesh, and flows south through Madhya Pradesh and Maharashtra in a very winding course of approximately 360 miles. After joining the Wardha, the united stream, known as the Pranahita, ultimately falls into the River Godavari.

In the present study Stream flow data of Kumhari gauging station (code AGH40R6) obtained from the Central Water Commission (CWC), Nagpur is used. Daily stream flow data for the period 1st June 1995 to 31st May 2008 have been used for the analysis and development of the model. The daily rainfall data for the period of 14 years of the Balaghat has been obtained from the Office of the Sub Divional Officer, Left Bank Dhooti Canal, Subdivision no.6, Balaghat (M. P.).

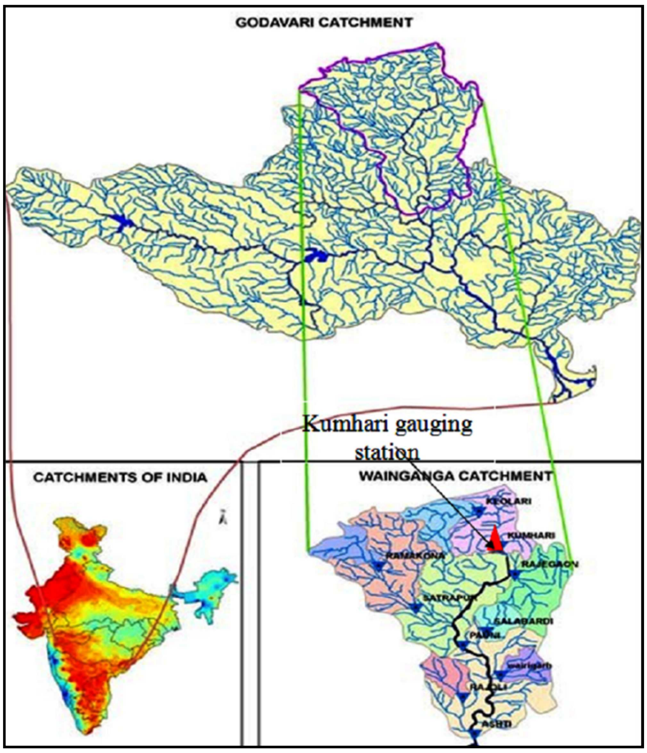


Figure 1. *Catchment area of Waingnga River with Kumhari gauging station.*

2.1. Multiple- Linear Regression (MLR) Model

Simple Linear Regression establishes the relation between response variable or dependent variable and explanatory variable or independent variable, in order to minimize the square residual. Multiple linear Regression (MLR) is simply extended form of Simple regression in which two or more variables are independent variables are used and can be expressed as:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots\cdots\cdots \beta_p X_p \qquad (1)$$

Where: Y = Dependent variable; α = Constant or intercept; β_1 = Slope (Beta coefficient) for X_1; X_1 =First independent variable that is explaining the variance in Y; β_2 = Slope (Beta coefficient) for X_2; X_2 = Second independent variable that is explaining the variance in Y; p= Number of independent variables; β_p= Slope coefficient for X_p; X_p= p^{th} independent variable explaining the variance in Y.

Table 1. *Multiple Regression Models showing R- square values.*

Sr. no.	Model	Multiple Regression equations	R^2
1	M-I	$Q_t = 65.0135 + 7.8686P_t$	0.0846
2	M-II	$Q_t = 46.4340 + 6.4757P_t + 5.0612P_{t-1}$	0.1170
3	M-III	$Q_t = 27.7460 + 5.5538P_t + 3.6597P_{t-1} + 6.0122P_{t-2}$	0.1615
4	M-IV	$Q_t = 18.4306 + 5.1739P_t + 3.2303P_{t-1} + 5.0833P_{t-2} + 3.3770P_{t-3}$	0.1754

Sr. no.	Model	Multiple Regression equations	R^2
5	M-V	$Q_t = 15.3336 + 5.0417P_t + 3.1176P_{t-1} + 5.1405P_{t-2} + 3.1210P_{t-3} + 1.2550P_{t-4}$	0.1773
6	M-VI	$Q_t = 12.6012 + 5.0067P_t + 3.0043P_{t-1} + 5.0419P_{t-2} + 2.994P_{t-3} + 1.0258P_{t-4} + 1.1421P_{t-5}$	0.1789
7	M-VII	$Q_t = 12.0431 + 4.9819P_t + 3.0014P_{t-1} + 5.019P_{t-2} + 2.9741P_{t-3} + 0.9996P_{t-4} + 1.0911P_{t-5} + 0.2582P_{t-6}$	0.1790
8	M-VIII	$Q_t = 6.2859 + 4.4669P_t + 0.3742P_{t-1} + 3.4340P_{t-2} + 0.2662P_{t-3} - 0.5868P_{t-4} + 0.5378P_{t-5} - 0.3147P_{t-6} + 0.5440Q_{t-1}$	0.4226
9	M-IX	$Q_t = 7.5562 + 4.2603P_t + 0.1807P_{t-1} + 3.8968P_{t-2} + 0.3162P_{t-3} - 0.0957P_{t-4} + 0.8805P_{t-5} - 0.2119P_{t-6} + 0.6208Q_{t-1} - 0.1385Q_{t-2}$	0.4336
10	M-X	$Q_t = 6.0268 + 4.2722P_t + 0.2837P_{t-1} + 4.0705P_{t-2} - 0.1899P_{t-3} - 0.15517P_{t-4} + 0.4197P_{t-5} - 0.5631P_{t-6} + 0.6388Q_{t-1} - 0.2198Q_{t-2} + 0.1304Q_{t-3}$	0.4432
11	Stepwise	$Q_t = .4677 + 4.2490P_t + 4.0503P_{t-2} + 0.6387Q_{t-1} - 0.2193Q_{t-2} + 0.1276Q_{t-3}$	0.4426

In present paper rainfall and runoff at various time step recorded at the Balaghat (i. e. P_t, P_{t-1}, P_{t-2}, P_{t-3}, P_{t-4}, P_{t-5}, P_{t-6}, Q_{t-1}, Q_{t-2}, Q_{t-3}) shall be considered as independent variable for estimation of the stream flow of the day (Q_t) being dependent variable.

To get different regression models out of 4750 days length of the record, 60% of it (i. e. 2850 days) has been used to calibrate the models and 11 regression models are developed with input variables showing R-square as a performance indicator for each model shown in Table-1. Among various regression models M-X model is based upon 10 input variables (as maximum input) with naturally provides higher R- square value, but handling that much amount of data is a rigorous task. Hence, to get the most influential variables used in model step wise regression (As it removes the variables having lower influence on correlation based on critical p value i. e. 0.05.) analysis is carried out. It is suggested five input variables viz. P_t, P_{t-2}, Q_{t-1}, Q_{t-2}, Q_{t-3} as momentous for better performance of the model and keeping the R- square values nearly the same.

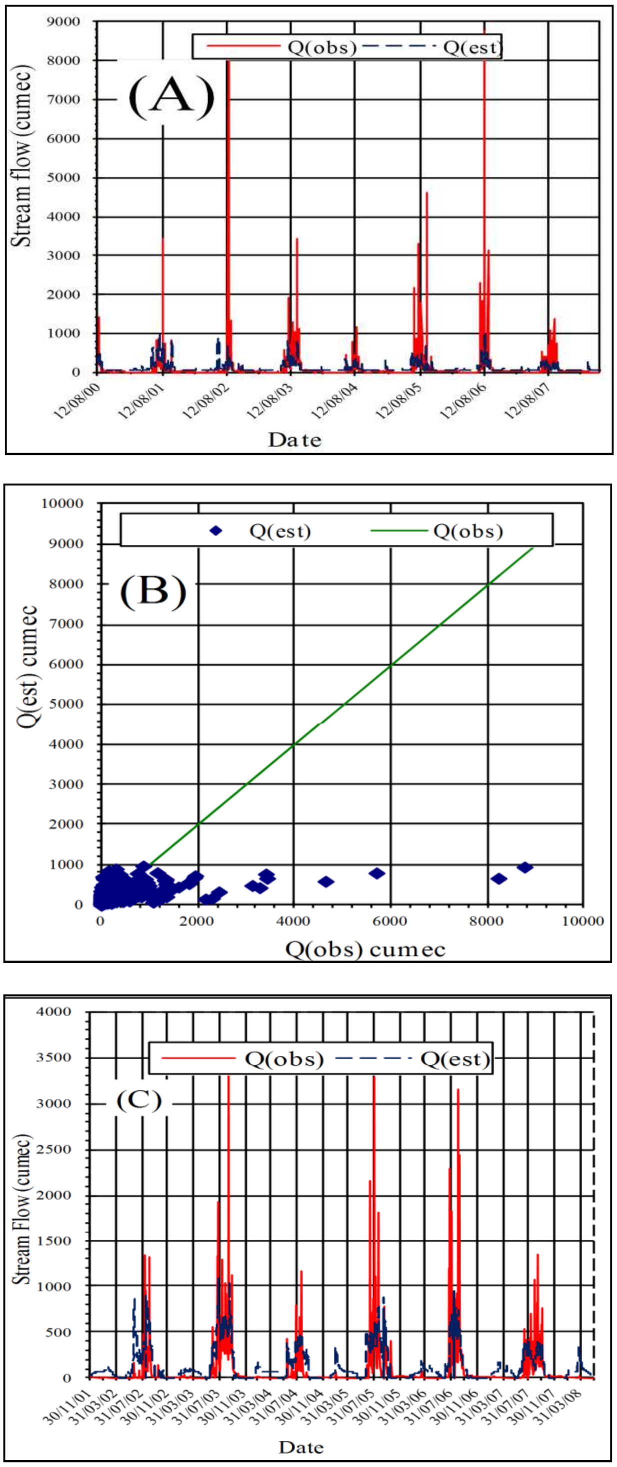
2.2. The Effect of the Length of the Data Set on Calibration and Validation of Models over the Stepwise Model

Since the stepwise model is performed well with the independent dataset. Now the same model is calibrated and validated with varying length of independent data set, Hence five different kinds of possible scenario of calibration and validation data records are analyzed which is shown in Table 2.

***Table 2.** Models calibrated and validated with varying length of the records.*

Sr. No.	Model	Calibration (% of the total data length)	Validated (% of the total data length)
1	M40-60	40	60
2	M50-50	50	50
3	M60-40	60	40
4	M70-30	70	30
5	M80-20	80	20

Observed and predicted stream flow hydrograph and deviation of estimated runoff values with observed values are depicted in Figure 2 representing five different scenarios.



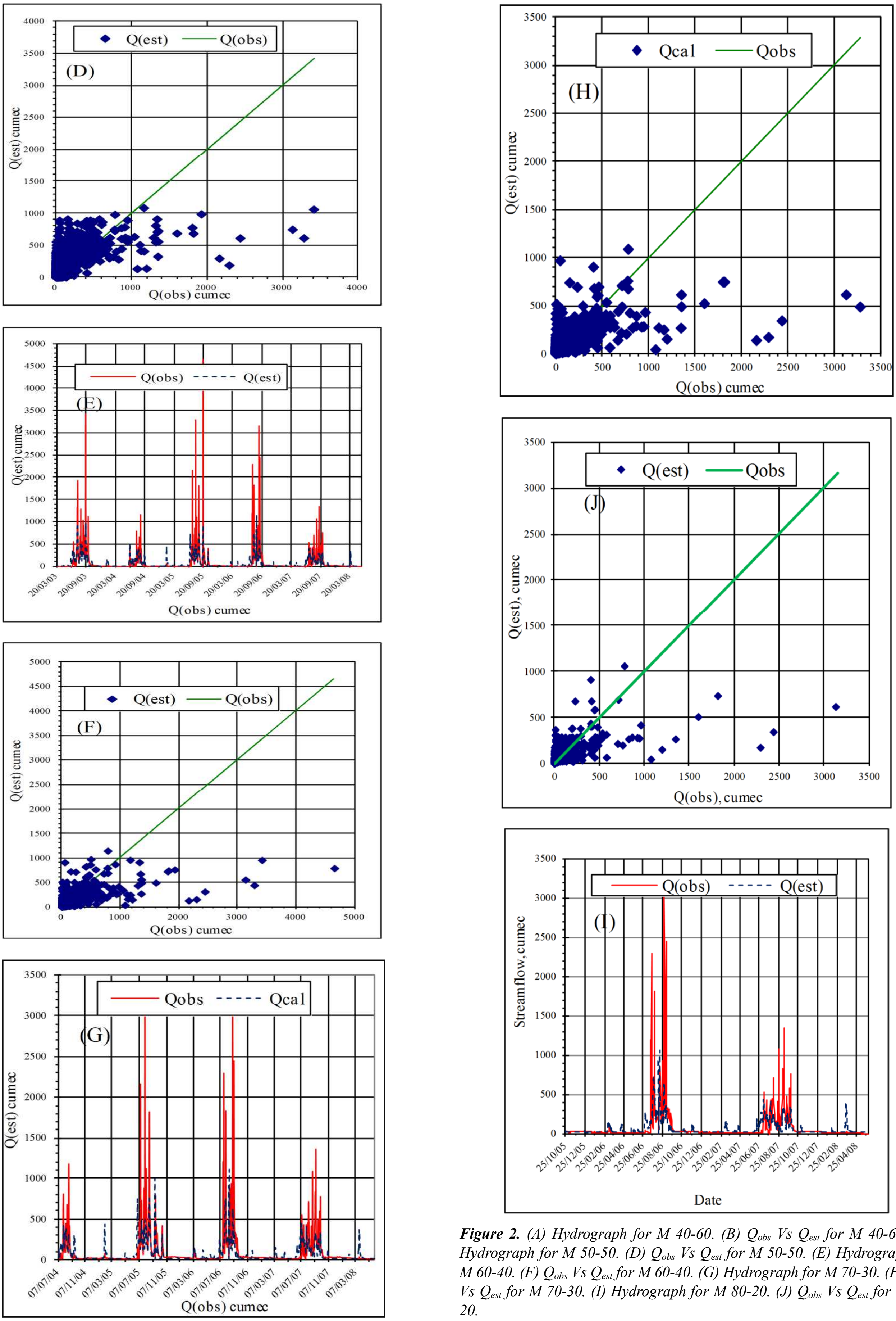


Figure 2. *(A) Hydrograph for M 40-60. (B) Q_{obs} Vs Q_{est} for M 40-60. (C) Hydrograph for M 50-50. (D) Q_{obs} Vs Q_{est} for M 50-50. (E) Hydrograph for M 60-40. (F) Q_{obs} Vs Q_{est} for M 60-40. (G) Hydrograph for M 70-30. (H) Q_{obs} Vs Q_{est} for M 70-30. (I) Hydrograph for M 80-20. (J) Q_{obs} Vs Q_{est} for M 80-20.*

3. Results and Discussion

Different models to correlate rainfall and runoff have been established through various graphs (Figure 2) and in order to compare these models different performance indicator like R (Correlation coefficient), SE (Standard error) and RMSE (Root mean square error) are used which are depicted as Table 3. However considering the various indicators Model 40-60 provides better prediction of stream flow values.

Table 3. *Different situations for calibration and validation of data with their performance indicator.*

S. No.	Model	Calibration		Validation	
		R	SE	R	RMSE
1	M40-60	0.8423	207.37	0.5505	427.85
2	M50-50	0.8067	198.26	0.4409	449.27
3	M60-40	0.6652	348.24	0.6433	257.04
4	M70-30	0.6686	329.46	0.6184	317.47
5	M80-20	0.6516	327.15	0.6537	336.43

4. Summary and Conclusions

The present study is dedicated to the statistical analysis based multiple regression model for stream flow estimation without considering the physical parameter associated with the Wainganga river basin using past 14 years (1994 to 2008) daily rainfall and runoff data. Different numbers of input data as rainfall with a past record of runoff values are used to establish a mathematical relation for estimation of the present day stream flow resulted as 11 numbers of equations (M-I to stepwise model). Since Stepwise model is performing well with available data, hence it is selected for the five different scenarios with different data range of calibration and validation and later reliabilities of different scenarios are evaluated based on the R, SE and RMSE.

From the study carried out, the following conclusions can be drawn:

1. M40-60 model gives the best result as predicted stream flow, due to its higher value of R and lower value of SE in the calibration section which are 0.8423 and 207.37. It also resulted as the higher value of R and lower value of RMSE as 0.5505 and 427.85 for validation section.
2. The best stream flow prediction can be drawn by using all ten parameters with M40-60 scenario, therefore it gives the highest correlation value among all ten models i. e. 0.6657 with lowest standard error value i. e. 348.34456.
3. Stepwise Regression Analysis shows that by taking the five parameters i. e. P_t, P_{t-3}, Q_{t-1}, Q_{t-2}, and Q_{t-3} using M40-60 scenario, the model can predict the Stream Flow better than other models as it gives the value of R and SE as 0.6652 and 348.2448.

References

[1] Asati, S., & Rathore, S. (2012). Comparative study of stream flow prediction model. *International Journal of sciences Biotechnology and Pharma Research, 1*, 139-151.

[2] Gosain, A., Mani, A., & Dwivedi, C. (2009). *Hydrological Modelling-Literature Review.* Climawater.

[3] Haan, C. (1995). *Statistical methods in Hydrology.* New Delhi: East-West Press Pvt. Ltd.

[4] Haan, C. T. (1972). A water yield model for small watersheds. *Water Resources Research, 8(1)*, 28-69.

[5] Holtan, H., Stiltner, G., Hensen, W., & Lopez, N. (1975). *Revised model of Watershed Hydrology.* Washington, D. C.: USDA-ARS Tech. Bulletin No. 1518.

[6] Jarboe, J., & Haan, C. (1974). Calibration of Water yield model for small ungauged watersheds. *Water Resources Research, 2*, 256-262.

[7] Jones, J. (1976). Physical data for catchment models. *Nordic Hydrol.*, 245-164.

[8] Kothyari, U. (1995). Estimation of monthly runoff from small catchments in India. *Hydrological Sciences – Journal - des Sciences Hydrologigues, 40*, 533-541.

[9] Lane, L., Diskin, M., & Renard, K. (1971). Input- Output relationship for an ephemeral stream channel system. *Journal of Hydrology, 13*, 22-40.

[10] Lindner- Lunsford, J., & Ellis, S. (1987). *Comparison of conceptually based and regression rainfall- runoff models, Denver Metropolitan Area, Colorado, and application in urban areas.* U. S. Geological Survey, Denver, CO.

[11] Magette, W., Shanholtz, V., & Carr, J. (1976). Estimation selected parameters for the Kentucky Watershed model from watershed characterstics. *Water Resources Research, 12(3)*, 462-476.

[12] Nawaz, N., & Adeloye, A. (1999). Evaluation of monthly runoff estimated by a rainfall- runoff model for reservoir yield assessment. *Hydrological Sciences—Journal—des Sciences Hydrologiques, 1*, 44.

[13] Newton, D., & Herrin, J. (1983). Estimating flood frequencies at ungauged loaction. *Hydraulic Engineering* (pp. 528-533). New York: American Society of Civil Engineers.

[14] Pilgrim, D., Chapman, T., & Doran, D. (1988). Problems of rainfall-runoff modelling in arid and semiarid regions. *Hydrological Sciences-Journal-des Sciences Hydrologiques, 4*, 33.

[15] Rutter, A., Kershaw, K., Morton, A., & Robins, P. (1971). A Predictive model of rainfall interception in forests. 1 Derivation of the model from observations in a plantation of Corsican pine. *Agricultural Meteorology*, 367-384.

[16] Wharton, G., & Tomlinson, J. (1999). Flood discharge estimation from river channel dimensions: results of applications in Java, Burundi, Ghana and Tanzania. *Hydrological Sciences—Journal—des Sciences Hydrologiques, 1*, 44.

[17] Xu, C., Seibert, J., & Halldin, S. (1996). Regional water balance modelling in the NOPEX area test on parameter estimation using catchment characteristics. *Journal of Hydrology*.

[18] Xu, C., & V. P., Singh. (1998). A review on monthly water balance models for water resources investigations. *Water Resources management, 12*, 31-50.