### (Max. 15 Pages including figures, tables and any other supplementary data)

### Data-driven modelling in forecasting daily streamflows for Purna River, India

{leave one line blank}

**Author A**1**, Author B**2 **and Author C**3

*1PhD Research Scholar, Department of Civil Engineering, Punjab Engineering College Chandigarh, India – 160012;* Email: *abc@gmail.com*

*2Professor, Department of Civil Engineering, Punjab Engineering College Chandigarh, India – 160012;* Email: *xyz@pec.edu.in*

*3Professor, Department of Civil Engineering, Indian Institute of Technology, Bombay, Mumbai – 400076;* E-mail*:* *def@iitb.ac.in*

{leave one line blank}

**Abstract**

{leave one line blank}

The applicability and performance of the model tree machine learning technique are investigated in daily river flow forecasting of Purna River up to Yerli stream gauging station in the Upper Tapi basin, India. The model tree technique divides the function into linear patches and provides a representation that is reproducible and coherent by the practitioners. [Abstract must be between 150 – 300 words]

{leave one line blank}

***Keywords: [Include 4-5 keywords separated by a comma]***

{leave one line blank}

# Introduction

{leave one line blank}

The forecasting of hydroclimatic variables such as rainfall, river discharge, river stage, etc. is a challenging task for hydrologists due to the complexity and non-linearity of hydrological phenomena. The prediction of inflows into reservoirs is useful in diverse applications such as efficient irrigation planning, flood control and mitigation, drought management, and hydropower generation (Yeh, 1985). The data-driven or black-box models treat the hydrological system as a black box and try to explore the relationship between the historical inputs to the system (rainfall, temperature) and corresponding outputs (runoff) with the help of statistical, artificial intelligence, soft computing, machine learning and data mining techniques (Solomantine and Dulal, 2003). The data-driven techniques such as Artificial Neural Network (ANN), Genetic Programming (GP), Fuzzy Logic, and Support Vector Machine (SVM) found wide applications in hydrological forecasting (Dawson and Wilby, 1998; Zealand et al., 1999; Londhe and Charhate, 2010; Jothiprakash and Magar, 2012).

{leave one line blank}

Model Tree (MT) is an emerging and promising data mining technique (Quinlan, 1992) and found limited applications in hydrology till date. Bhattacharya and Solomantine (2005) developed a water level-discharge relationship at Swarupganj station othe n Bhagirathi River, by employing ANN and MT methods. The present study demonstrates the applicability of the MT machine learning technique in forecasting daily flows for the Purna river in the Tapi basin, India.

{leave one line blank}

# Materials and Methods

{leave one line blank}

## Model Tree technique

{leave one line blank}

The model tree (MT) is a machine learning technique that uses the idea of splitting the parameter space into sub-spaces and formulating the linear regression model for each of them (Witten and Frank, 2005).

{leave one line blank}

## Study Area and Data Source

{leave one line blank}

#### Purna River basin

{leave one line blank}

The study area is the Purna river basin (area ≈ 18,430 km2) which originates in Betul district in Gawaligarh steep mountains of the Satpura range at a latitude of 21° 38′ N and longitude of 77° 36′ E in Maharashtra, India. The index map of the Purna River basin is shown in Figure 1.

{leave one line blank}

#### Data collection

{leave one line blank}

The daily rainfall data for seventeen rain gauge stations were collected from India Meteorological Department (IMD), Pune and the daily discharge data for three stream gauging sites were obtained from Central Water Commission (CWC), Surat, India.

{leave one line blank}



##### **Figure 1** Index map of study area

{leave one line blank}

## Selection of Input Parameters

{leave one line blank}

The input parameters were selected based on the cross-correlation analysis carried out at each stream gauging site. For the present study, *Q*(*t*) and *P*(*t*) indicate the stream discharge (m3/s) and areal average rainfall [mm] in a catchment on the *t*-th day respectively. Further, (*t* – 1) and (*t* – 2) indicate values corresponding to the previous day and the previous two days. The nomenclature *L, G* and *Y* represent Lakhpuri, Gopalkheda and Yerli sub-catchments respectively. The correlation values for hydrologic variables at different time lags are given in Table 1. From cross-correlation analyses, for the Lakhpuri sub-catchment, the following input variables were selected, and represented through functional relationship by Equation (1).

$$QL\_{t}=f\left(QL\_{t-1},PL\_{t},PL\_{t-1}\right) (1)$$

{leave one line blank}

In the same manner, the input variables were selected for Gopalkheda as well as Yerli sub-catchments, and functional relationships can be expressed in a similar way.

{leave one line blank}

##### **Table 1** Cross-correlation analysis for variables in the Lakhpuri sub-catchment

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | *QLt* | *QLt-1* | *QLt-2* | *PLt* | *PLt-1* | *PLt-2* |
| *QLt* | 1 |  |  |  |  |  |
| *QLt-1* | **0.47** | 1 |  |  |  |  |
| *QLt-2* | 0.23 | 0.47 | 1 |  |  |  |
| *PLt* | **0.52** | 0.17 | 0.04 | 1 |  |  |
| *PLt-1* | **0.44** | 0.52 | 0.17 | 0.38 | 1 |  |
| *PLt-2* | 0.21 | 0.44 | 0.52 | 0.13 | 0.38 | 1 |

**Bold values** *indicate the variables selected as input to develop the functional relationship*

{leave one line blank}

# Results and Discussions

{leave one-line blank}

The detailed analysis has been carried out for Lakpuri, Gopalkheda and Yerli sub-catchments using WEKA software, and the results are described in the following paragraphs.

{leave one line blank}

For the Lakhpuri sub-catchment, the time series and scatter plot between observed and predicted discharge is shown in Figure 2.

{leave one line blank}



#####  **Figure 2** Time series plot of observed and predicted discharge at Lakhpuri station

{leave one line blank}

# Conclusions

{leave one line blank}

The following conclusions are derived from the foregoing study:

1. The input variables for M5 model trees were selected from the cross-correlation analysis between rainfall and runoff at different time lags.
2. It has been observed that the correlation between observed and predicted flows was relatively low for Lakhpuri (*R* = 0.69), moderate for Gopalkheda (*R* = 0.76) and high for Yerli (*R* = 0.92) sub-catchment.
3. It has also been observed that, for all the three sub-catchments, the model is able to predict low and medium flows satisfactorily, but the high flows are not predicted accurately.

{leave one line blank}

**Acknowledgements** [Acknowledgment of the funding/data disseminating agencies, if any]

{leave one line blank}

The authors acknowledge the financial support received from the Department of Science and Technology (DST), Ministry of Science and Technology, Government of India to carry out the present work. The authors are also thankful to India Meteorological Department (IMD) and Central Water Commission (CWC) for providing the necessary data to conduct the present study.

{leave one line blank}

**References** [References should be in APA format]

{leave one line blank}

Arunkumar, R., and Jothiprakash, V. (2012). Reservoir evaporation prediction using data-driven techniques. *Journal of Hydrologic Engineering*, 18(1), 40-49.

Bhattacharya, B., and Solomatine, D. P. (2005). Neural networks and M5 model trees in modelling water level–discharge relationship. *Neurocomputing*, 63, 381-396.

Dawson, C. W., and Wilby, R. (1998). An artificial neural network approach to rainfall-runoff modelling. *Hydrological Sciences Journal*, 43(1), 47-66.

Witten, I. H., and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, San Francisco, USA.

Yeh, W. W. G. (1985). Reservoir management and operations models: A state‐of‐the‐art review. *Water Resources Research*, 21(12), 1797-1818.