IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

RAINFALL-RUNOFF MODEL IN RIVER BASIN USING DATA DRIVEN TECHNIQUES

[1]Shilpa .S. kasulla, [2]Sanjana Londhe, [3]Dr. Seema jagtap
[1]Lecturer, [2]Lecturer, [3]Head of Department
Department of Civil,
Government polytechnic, Thane, India

Abstract: The focus of work will be on rainfall-runoff modeling i.e. how the transformation of rainfall into runoff can be simulated with different data driven tools describing runoff generation processes. Based on personal and professional experiences, the use of hydrological models, their advantages and challenges are described. Rainfall to runoff transformation is one of the imperative non linear hydrological processes.

1.INTRODUCTION

This report explores rainfall-runoff models, their generation methods, and the categories under which they fall. Runoff plays an important role in the hydrological cycle by returning excess precipitation to the oceans and controlling how much water flows into stream systems. Modeling runoff can help to understand, control, and monitor the quality and quantity of water resources. A few categories of rainfall-runoff models are described by the model structure and spatial processes within the model. Both control the way models calculate runoff. Model structure is based on the governing equations a model uses to determine runoff; categories can be generalized into empirical, conceptual, and physical structures. Spatial processes within a model are the interpretation of the catchment characteristics to be modeled. This category separates models into lumped, semi-distributed, and distributed models, which is a generalization because many models overlap and contain elements from each of the categories. A discussion about comparing different runoff models and observed runoff values is presented as well. This report aims to inform modelers about various rainfall-runoff models and their strengths and weaknesses.

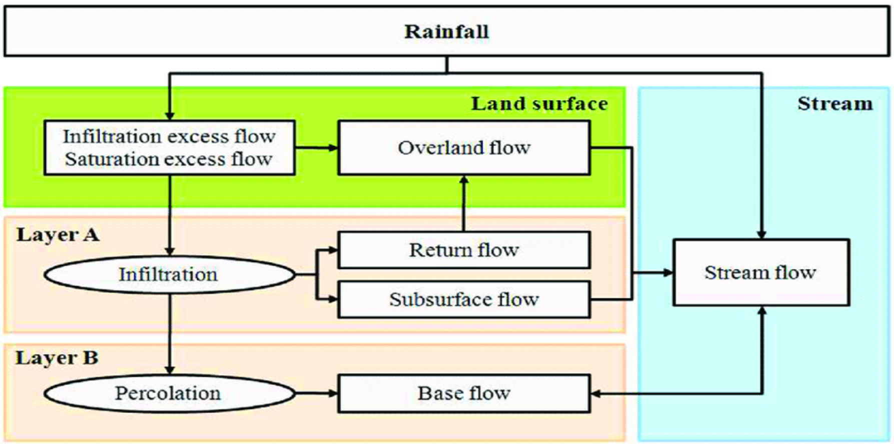

Figure1: Schematic Diagram of Rainfall Runoff Modelling

In general we can say that the field of hydrology was one of the earlier areas of science and engineering to use statistical concepts in an effort to analyze and forecast the natural phenomena. To understand the variations in hydrologic variables and successfully predict the future behavior of the same, a lot of research is taking place and consequently many new methodologies have been introduced.

2. STUDY AREA AND METHODOLOGY

The present work aims at forecasting of stream flow one day ahead at four stations Paud, Rakshewadi, Khamgaon and Nighoje in Maharashtra, India, using previous values of stream flow and rainfall and the techniques of M5 model trees.

Krishna River is India’s fourth-largest river basin, which covers 258,948km2 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 outfall into the bay of Bengal traversing tree states Karnataka (113,271 km2), Andhra Pradesh (76,252 km2), Maharashtra (69,425 km2). The basin is divided into twelve sub basins for hydrological studies namely Upper Krishna, Middle Krishna, Ghataprabha, Malaprabha, Upper Bhima, Lower Bhima, Lower Krishna, Tungabhadra, Venavati, Musi, Pallaru and Munnaru. The data was collected by surface and Ground water hydrology department (M.E.R.I.) through Hydro-project Nashik.

2.1 DESCRIPTION OF STUDY AREA

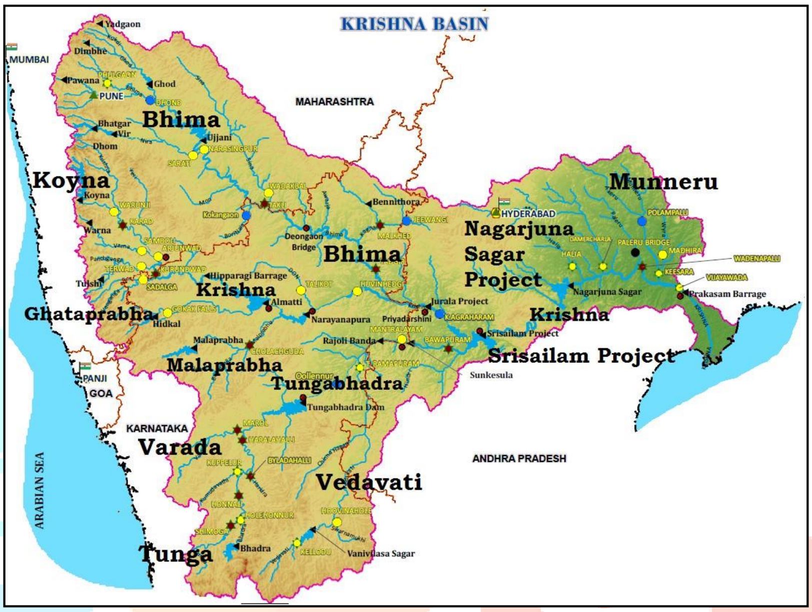

Figure2.1: Krishna River Basin Map

2.1.1 KRISHNA RIVER BASIN DETAILS

- The Krishna is the second largest eastward draining interstate river in Peninsular India.
- It rises in the Mahadev range of the Western Ghats at an altitude of 1,337 m near Mahabaleshwar in Maharashtra State, about 64 km from the Arabian Sea.
- It flows for a distance of 305 km in Maharashtra, 483 km in Karnataka and 612 km in Andhra Pradesh before finally out falling into the Bay of Bengal.
- The length of the river is about 1,400 km
- Krishna basin lies between latitudes 13° 07’ N and 19° 20’ N and longitudes 73° 22’ E and 81° 10’ E.
- On the north, the basin is bounded by the range separating it from the Godavari basin, on the south and east by the Eastern Ghats and on the west by the Western Ghats.
- Drainage area of the basin is 258,948 km^2

Drainage area of the Basin (in Km^2)

Name of State	Drainage area
Maharashtra	69,425
Karnataka	113,271
Andhra Pradesh	76,252
Total Drainage Area of Krishna Basin (in Km^2) =258,948	

2.1.2. GEOLOGY OF THE BASIN

The geology of the Krishna basin is dominated in the northwest by the Deccan Traps, in the central part by unclassified crystallines, and in the east by the Cuddapah Group. The Dharwars (southwest central) and the Vindhian (east central) form a significant part of the outcrops within the unclassified crystallines. Krishna delta is predominantly formed by Pleistocene to recent material.

2.1.3. WATER POTENTIAL OF THE BASIN

Surface Water potential	78.1 km^3
Ground Water potential	26.41 km^3

2.1.4. HYDROPOWER POTENTIAL

Hydropower Station	(Installed capacity 50MW and above)
Nagarjunsagar	815
Srisailam	770
Nagarjunsagar RC	91
Nagarjunsagar CH	61
Srisailam LB	900

2.1.5. TRIBUTARIES WITH DRAINAGE AREA IN SQ. KM.

Name of the Tributary	Drainage area (Km^2)
Koyna	4,890
Panchganga	2,575
Dudhganga	2,350
Ghataprabha	8,829
Malaprabha	11,549
Bhima	70,614
Tungabhadra	71,417
Dindl	3,490
Peddavagu	2,343
Halia	3,780
Musi	11,212

2.1.6 MAJOR PROJECTS

Upper Krishna Project Stage – 1,Upper Krishna Project Stage – 2, Srisailam Dam, Pulichintala Project, Nagarjunasagar Project, Ghatprabha Dam, Tungabhadra Project, Vanivilas Sagar Project, Bennihora Project, Bhadha Reservoir Project, Bhima Irrigation Project, Hipparagi Barrage, Malprabha Project, Upper Tunga Project, Koyna dam, Markendaya Project, Singatalur Lift Irrigation, Krishna Irrigation Project, Osmansagar Reservoir and Prakasam barrage

2.1.7 WATER QUALITY OF THE BASIN

Based on the systematic sampling of river water at many locations in the basin, its suitability for various purposes is determined by CPCB and as per the results, the quality is not as per the desired class and BOD remains the most critical parameter. At some places, DO and total coliform are also causing problem.

2.1.8 MYTHOLOGY

Krishna is a mighty east flowing river of peninsular India. It is the same river as Krsnavena in the Puranas or Krsnaveni in the Yoginitantra. It is also known as Kanhapenna in Jatakas and Kanhapena in the Hathigumpha inscription of Kharavela. The word Krishna also indicates dark color.

1. Data of rainfall & runoff segregation for model Training & Testing, so that mode give most accurate results
2. Make well trained rainfall runoff and runoff runoff model
3. Use the model for predicting one day ahead runoff
4. Find root mean square error and coefficient of correlation to give statistical validation and verification of results of model.

2.2 MODEL TRRE (M5)

A decision tree is a logical model represented as a binary (two-way split) tree that shows how the values of a target (dependent) variable can be predicted by using the values of a set of predictor (independent) variables. These are basically two types of decision trees:

1. Classification trees are the most common and are used to predict a symbolic attribute (class).
2. Regression trees which are be used to predict the value of a numeric attribute (Witten and Frank, 1999).

If each leaf in the tree contains a linear regression model, that is used to predict the target variable at that leaf, is called a model tree. The M5 model tree algorithm was originally developed by Quinlan (1992). Detail description of this technique is beyond of this paper and can be found in Witten and Frank (1999). A bit description of this technique follows. The M5 algorithm constructs a regression tress by recursively splitting the instance space using tests on a single attributes that maximally reduce variance in the target variable. Figure 1 illustrates this concept. The formula to compute the standard deviation reduction (SDR) is: (Quinlan, 1994)

$$\text{SDR}=\text{sd}(\text{T})\text{-}\sum\frac{t}{t}\text{sd}(\text{T})$$

Where T represents a set of example that reaches the node; T represents the subset of examples that have the I th outcome of the potential set; and sd represents the standard deviation.

2.2.1 DEFINITION

Model tree is a machine learning technique based on an idea of spitting the input data into sub areas and building "local" linear regression models in each of them. MT is the linear regression method based on an assumption of linear dependencies between input and output. The 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. In MT a step towards non-linearity is made- since it builds a model that locally linear but overall is non linear MT approach based on the principal of information theory makes it possible to spit the multi dimensional parameter space and to generate model automatically according to the overall quality criteria; it also allows for varying the number of models. MT learns efficiently and can tackle task with very high dimensionality- up to hundred of attribute. A model tree splits the input progressively. The set *T* of examples is either associated with a leaf, or some test is chosen that splits *T* into subsets corresponding to the test outcomes and the same process is applied recursively to the subsets. Splits are based on minimizing the intra-subset variation in the output values down each branch. In each node, the standard deviation of the output values for the examples reaching a node is taken as a measure of the error of this node and calculating the expected reduction in error as a result of testing each attribute and all possible split values. The attribute that maximizes the expected error reduction is chosen. A model tree (MT) is an extension of the concept of classification tree and regression tree in which the computational process is represented by a tree structure consisting of a root node (decision box) branching out to numerous other nodes and leaves. In a model tree the entire input (or parameter) domain is divided into sub-domains and a (multivariate) linear regression model is developed for each sub-domain. It therefore considers a piecewise linear model to approximate a given nonlinear relationship between a dependent variable and corresponding independent variables. In case of MT the output as a whole does not depend on input but it involves a set of functional relationships within the input domain. Depending upon considerations like a domain splitting criterion there are alternative algorithms to build a MT.

2.2.2 PRUNING THE MT

By applying pruning that is making tree smaller by combining sub trees in one node. If a generated tree has too many leaves, it may be "too accurate" and hence over fit and be a poor generalize. It is possible to make a tree more robust by simplifying it, i.e., by pruning that is by merging some of the lower sub trees into one node.

2.2.3 SMOOTHING THE MT

This process is used to compensate for the sharp discontinuities that will inevitably occur between adjacent linear models at the leaves of the pruned trees. This is a particular problem for models constructed from a small number of training samples. Smoothing can be accomplished by producing linear models for each internal node, as well as for the leaves at the time the tree is built. Experiments show that smoothing substantially increases the accuracy of prediction. Details of Model tree can be seen in Quinlan (1992).

The M5 algorithm is used for inducing a model tree (Quinlan, 1992), which works as follows (Figure 2.1).

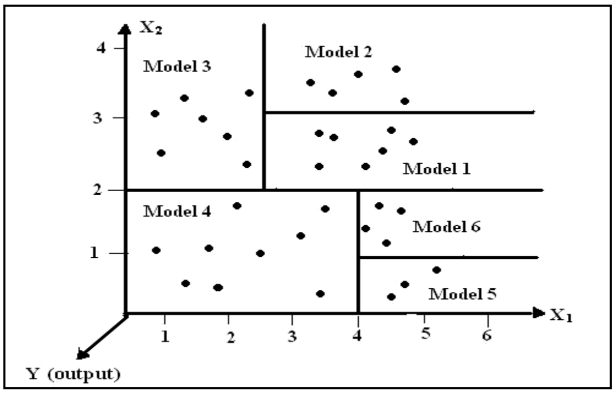

Figure 2.2: A Model Tree

2.3 GENETIC PROGRAMMING

Like MT, another latest soft computing tool that has caught attention of researchers in water resources is genetic programming (GP). The GP can iteratively generate new values till they reach a certain level of acceptance as per the selected criterion and thus look attractive in the current problem. Genetic programming is modeled out of the process of evolution occurring in nature, where the species survive following the principle of 'survival of the fittest'. Unlike the more widely known genetic algorithms (GA), its solution is a computer program or an equation as against a set of numbers in the GA. Koza (1992) explains various concepts related to GP. In GP a random population of individuals (equations or computer programs) is created, the fitness of individuals is evaluated and then the 'parents' are selected out of these individuals. The parents are then made to yield 'offspring's' by following the process of reproduction, mutation and crossover. The creation of offspring's continues (in an iterative manner) till a specified number of offspring's in a generation are produced and further till another specified number of generations is created. The resulting offspring at the end of all this process (an equation or a computer program) is the solution of the problem. The GP transforms one population of individuals into another one in an iterative manner by following the natural genetic operations like reproduction, mutation and crossover. Genetic programming (GP) is an automated method for creating a working computer program from a high-level problem statement of a problem. Genetic programming starts from a high-level statement of "what needs to be done" and automatically creates a computer program to solve the problem. The main difference between GA and GP is that individuals (models) in GAs are represented by fixed-length strings (usually binary) whereas individuals in GP, which constitute a population, are symbol string codes for mathematical models.

2.3.1 LINEAR GENETIC PROGRAMMING

The software used in this study (discipulus) uses AIMGP. In brief summary, the linear genetic programming algorithm in Discipulus is surprisingly simple. It starts with a population of randomly generated computer programs. These programs are the "primordial soup" on which computerized evolution operates. Then, GP conducts a "tournament" by selecting four programs from the population—also at random—and measures how well each of the four programs performs the task designated by the GP developer. The two programs that perform the task best "win" the tournament.

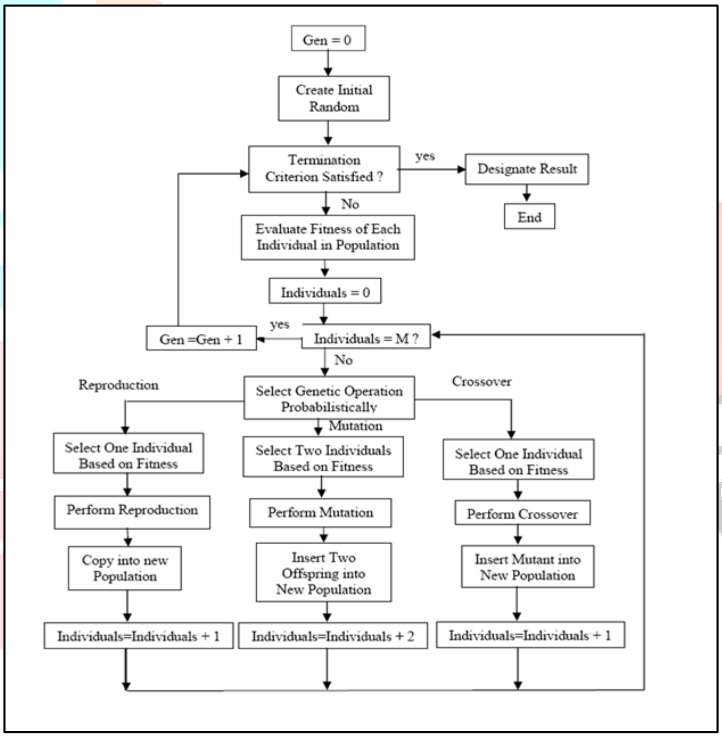
Gen = 0
Create Initial Random
Termination Criterion Satisfied ?
yes
Designate Result
End
No
Evaluate Fitness of Each Individual in Population
Individuals = 0
yes
Gen =Gen + 1
Individuals = M ?
No
Reproduction
Select Genetic Operation Probabilistically
Crossover
Mutation
Select One Individual Based on Fitness
Select Two Individuals Based on Fitness
Select One Individual Based on Fitness
Perform Reproduction
Perform Mutation
Perform Crossover
Copy into new Population
Insert Two Offspring into New Population
Insert Mutant into New Population
Individuals=Individuals + 1
Individuals=Individuals + 2
Individuals=Individuals + 1

Figure 2.3 FLOWCHART OF THE GENETIC PROGGRAMMING PARADIGM

2.4 CREATING TRAINING, VALIDATION AND APPLIED DATA FILES

A text file to be in a precise format before it may be imported as a training file, a validation file, or an applied data file. Following are the rules for creating them:

- Data files must be ACSII text files, which may be created using Word Pad, notepad or Microsoft Excel.
- The training, validation, and applied data files should be identical in structure, with the same number of inputs and outputs, i.e. the same number of columns. These files may, however, have a different number of examples – that is, a different number of rows in the file.
- The data must be arranged in columns in the training, validation, and applied data files.
- Each column represents an input or an output (the output being held in the farthest right column).
- Each row in the training, validation, and applied data files must have a separate "example" that contains both inputs and one projected output.
- A tab or a space on each row must separate the columns of data in the training, validation, and applied data files.
- The output data that we want to have GP learn must be the right hand column.
- The training, validation, and applied data files must have the same number of columns of data in each row and must have two or more rows and two or more columns of data.
- Every value in the training, validation, and applied data files must be an integer or a real number.
- Non-printing characters should not be put at the end of a line of the end of the file.

2.5 EVALUATION OF MODEL PERFORMANCE

The developed models were evaluated for their accuracy by employing statistical parameters like correlation coefficient (r), root mean squared error (rmse) along with the hydrographs and scatter plots between the model predicted and observed discharge values. The error measures are explained below (Dawson 2001).

2.5.1 CORRELATION COEFFICIENT

A correlation coefficient is a number between -1.0 and +1.0, which measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, correlation coefficient is equal to 1.The measure however is very sensitive to deviations at larger observations.

2.5.2 ROOT MEAN SQUARE ERROR

This measure gives an overall agreement between the observed and modeled datasets. It has no upper bound with zero as the value for a perfect model. This measure is good for iteratively arrived at predictions and gives only an overall picture of errors.

3. RESULTS & DISCUSSION

3.1 GENERAL

The above-developed nine different models for each monsoon month that is June, July, August, September, and October for each station were tested for their performance using statistical parameters and plotting hydrographs and scatter plots (total 180 models for both GP and MT approach).

Result by Model Tree (M5) for Paud station for June month is-

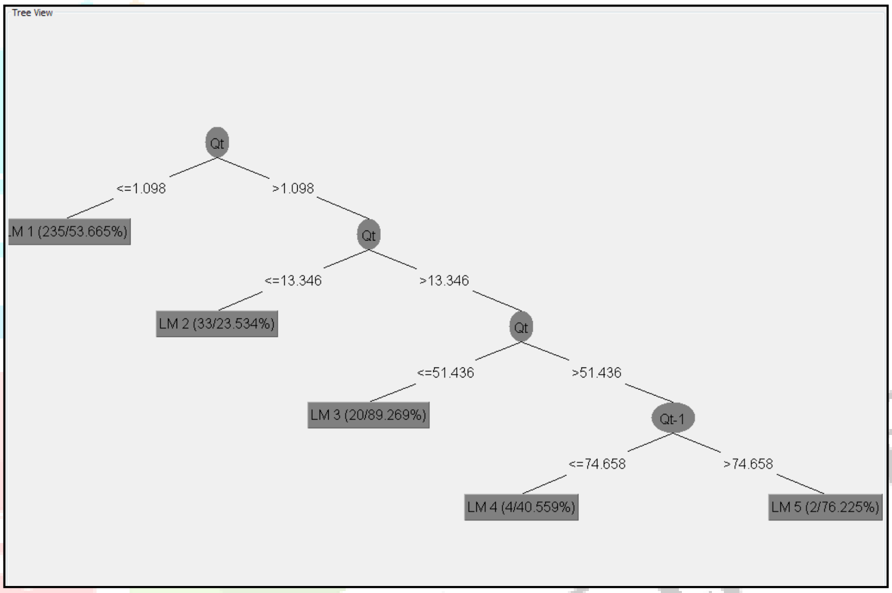

Figure 3.1 Model Tree for PJUNE1 Model

```
M5 pruned model tree:
(using smoothed linear models)
Qt <= 1.098 : LM1 (235/53.665%)
Qt >  1.098 :
|   Qt <= 13.346 : LM2 (33/23.534%)
|   Qt >  13.346 :
|   |   Qt <= 51.436 : LM3 (20/89.269%)
|   |   Qt >  51.436 :
|   |   |   Qt-1 <= 74.658 : LM4 (4/40.559%)
|   |   |   Qt-1 >  74.658 : LM5 (2/76.225%)
LM num: 1
Qt+1 =
        -0.0081 * Qt-1
        + 0.0427 * Qt
        + 1.139
LM num: 2
Qt+1 =
        -0.0994 * Qt-1
        + 0.7115 * Qt
        + 2.5174
```

LM num: 3
Qt+1 =
-0.1117 * Qt-1
+ 0.4603 * Qt
+ 9.6179

LM num: 4
Qt+1 =
-0.1117 * Qt-1
+ 0.5736 * Qt
+ 13.1764

LM num: 5
Qt+1 = -0.1117 * Qt-1
+ 0.528 * Qt
+ 14.7751

3.2 COMPARISON OF M5 MT AND DISCIPULUS (GP)

Best models selected above for both MT and GP are same. Plotting their combined hydrographs like observed, MT and GP on one plot compared best models. Scatter plots was also used to compare MT and GP. It was observed that inclusion of rainfall has improved the results.

The hydrograph dawned to compare Model Tree (M5) and Genetic Programming results for all models. The result for same is shown in figure 3.2 below for station Paud, june month.

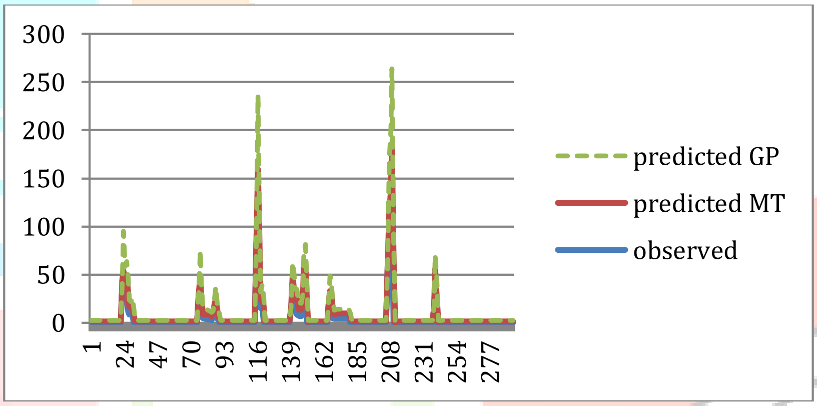

Figure 3.2 Hydrograph for PJUNE1 Model to compare MT and GP

Scatter plot was obtained as graph of predicted stream flow verses observed stream flow to compare Model Tree (M5) and Genetic Programming results. Scatter plot obtained is shown in figure 3.3 for Paud station, july month, model number 2.

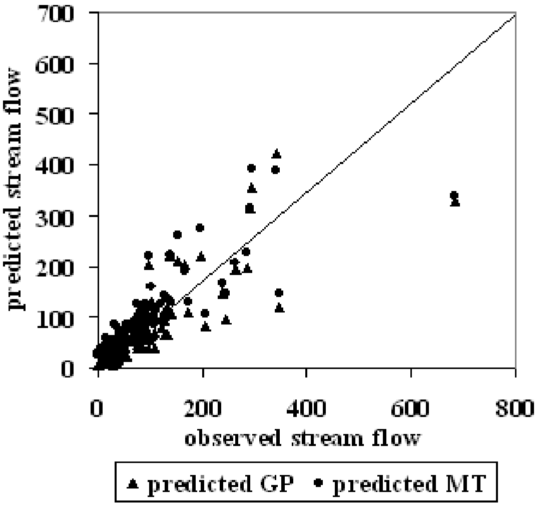

Figure 3.3 Scatter plot for PJULY2 Model to compare MT and GP

4. CONCLUSION

On comparing the statistical properties of these nine models, the model, which considered runoff as well as rainfall, is performed better with r-value of more than 0.86. The other performances of this model are also better. After various trials it was found that the testing accuracy has not increased with more than five inputs, both in precipitation and runoff. Results obtained from the best MT and GP models are compared.

The result obtained based on CC, rmse, scatter plot, hydrograph indicate that GP models can be used satisfactorily in forecasting of runoff compared to MT models.

6. REFERENCES

1. www.aimlearning.com.
2. http://www.wikipedia.com
3. http://www.google.com
4. http://www.hydrologydatainfo.com
5. http://www.cs.waikato.ac.nz/ml/weka/
6. Engineering Hydrology by K. Subramanya 2nd Edition, Publisher: Tata McGraw-Hill.
7. B. Bhattacharya, D.P. Solomatine, Neural networks and M5 model trees in modelling water leveldischarge relationship for an Indian river, in: M. Verleysen (Ed.), Proceedings of the 11th European Symposium on Artificial Neural Network, Bruges, Belgium, d-side, Evere Belgium, 2003, pp. 407–412.
8. Bishwajit Roy, Maheshwari Prasad Singh, Mosbeh R. Kaloop, Deepak Kumar, Jong-Wan Hu, Radhikesh Kumar and Won-Sup Hwang (2021), “Data-Driven Approach for Rainfall-Runoff Modelling Using Equilibrium Optimizer Coupled Extreme Learning Machine and Deep Neural Network”, Appl. Sci. 2021, 11, 6238, 1-36
9. Drounpob, A., Chang, N.B., Beaman, M. (2005), Stream low rate prediction using genetic programming model in semi-arid coastal water-shed. Proc. World water ongress, Alaska, 2005
10. K.N.V. Rama Devi, R. Venkata Ramana, Y. R. Satyaji Rao and Sanjeet Kumar, “Development of Data Driven Rainfall-Runoff Model for the Sarada River Basin.”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6C2, April 2019, 508-512
11. Koza JR (1992) Genetic Programming on the Programming of Computers by Means of Natural Selection. A Bradford Book, MIT Press, 1992.
12. Sherman, L.K., (1932). Streamflow from rainfall by the unit graph method. Eng. News Rec., 108: 501 5O5.
13. Solomatine, D.P., “Data-driven modelling: paradigm, methods, experiences”, Proc. 5th International Conference on Hydroinformatics, Cardiff, UK, (2002).
14. Solomatine D P and Rodriguez J., “Quantifying uncertainty in flood forecasting models using soft computing”, Hydrol. Sci. J. (submitted).
15. Solomatine, D.P., “Data-driven modelling: paradigm, methods, experiences”, Proc. 5th International Conference on Hydroinformatics, Cardiff, UK, (2002).
16. Solomatine, D.P. and Dulal, K. N., “Model tree as an alternative to neural network in rainfall-runoff modeling”, Hydrological Sc. J., Vol.48(3), (2003), pp 399-411.
17. Solomatine D.P. and Siek M.B., “Flexible and optimal M5 model trees with applications to flow predictions”, Proc. 6th Int. Conf. on Hydroinformatics, Singapore, World Scientific, June (2004).