M5 Model Tree for Regional Mean Annual Flood Estimation

K. K. SINGH
Department of Civil Engineering
National Institute of Technology, Kurukshetra
India

Abstract: Reliable and precise estimation of floods in a river is critical for efficient flood management and surface water planning. Hydrologists use catchment and hydrological data to establish regional relationships between mean annual flood and various catchment and rainfall characteristics. The relationship is used for predicting floods of different return periods. For Indian catchments Swamee et al. [5] developed a regional relationship using dimensional analysis from 93 catchments spread over the entire country. Swamee et al. [5] model predicts considerable amount of data much beyond the scatter line of ± 50 % and predicted mean annual flood value is sometimes as high as 8 times the actual value. The correlation coefficient (CC) of model is as low as 0.09 and the root mean square error (RMSE) is very high. Thus, a retrospect of the model is warranted. In the present paper, models of the mean annual flood relationship are developed using an M5 model tree. The data is analyzed by using a cross-validation method. The model predictions with an M5 model tree fall well within a scatter line of ± 15 % with correlation coefficient (CC) as high as 0.9936 and very low RMSE. The predictive value of the M5 model tree is 1.25 times the actual value. It is concluded that the M5 model tree-based modelling approaches is superior in accuracy to the Swamee et al. [5] model.

Key words: Flood, dimensional analysis, India, M5 model tree, correlation coefficient, root mean square error

1 Introduction

Reliable and precise estimation of floods in a river is key for efficient flood management and surface water planning. Hydrologists use catchment and hydrological data to establish regional relationships for mean annual flood estimation. Once a relationship for a region is established, it can be used for predicting floods of different return periods. Using dimensional analysis, Swamee et al. [5] derived a relationship for mean annual flood from 93 catchments in India spread over the entire country having a return period of 2.33 years. The proposed relation is as follows:

$$Q = 1.724 \frac{A^{0.8925} p^{0.92} (S_0 + 0.012)^{0.55}}{D^{0.24} T^{0.17} (C_f + 0.049)^{0.55}} \tag{1}$$

where Q, A, p, D, T, S_0 and C_f are, respectively, mean annual flood in m^3/s, drainage area in sq. km., average annual rainfall in cm, average annual rainfall duration in minutes, return period in years, catchment slope in %, and the fraction of forest area as ratio. This relationship is considered more general, dimensionally correct and it satisfies the boundary conditions. However, the average error associated with Eq. (1) is significant and coefficient of determination is less when used with different data in Indian catchments including the data used in the study of Swamee et al. [5]. The predicted mean annual flood values for most of the data set using Eq. (1) fall much beyond the scatter line of ± 50 % and predicted values are sometimes as high as 8 times the actual value. The correlation coefficient (CC) of Eq. (1) is low and the root mean square error (RMSE) is very high. Thus, there is a need of using some alternate approaches in modeling mean annual floods.

Successful applications of machine learning in water management by Solomatine and Dulal [3], and Bhattacharya and Solomatine [1] have inspired the exploration of its applicability to modeling the complex flood relationships. Among machine learning techniques, artificial neural network (ANN) is the one of the widely used approaches in various areas of water-related research. M5 model tree (Quinlan, [2]) based modeling is not as popular as ANN but has been proved to be very efficient and robust in water resource applications (Solomatine and Xue, [4]; Bhattacharya and Solomatine, [1]). The main objective of this study is to use M5 model tree for Indian catchments for predicting flood discharges based on the known hydrological system data. The predictive accuracy of this model is compared with the method of Swamee et al. [5] using the same data.

2 M5 Model Tree

In machine learning approaches a non-linear parametric function approximator is used. In the function approximator the coefficients of the function decomposition are obtained from the input–output data pairs, some chosen model structure and systematic learning rules. Once trained, the machine learning model becomes a parametric description of the function. Learning a general principle from a set of specific training examples is achieved by trying out different model structures and the related parameters. Out of several possible methods, ANN is the most widely used method in the water sector, whereas M5 model trees which is almost unknown to the water sector.

In this technique, the parameter space is split into areas (subspaces) and it builds in each of them a linear regression model. The resulting model can be seen as a modular model, or a committee machine, with the linear models being specialized on the particular subsets of the input space. Combination of specialized models ("local" models) is used quite often for modelling. M5 model tree approach is based on the principle of information theory that makes it possible to split the multi-dimensional parameter space and generate the models automatically according to the overall quality criterion. It allows for variation in the number of models created.

The splitting in the M5 modal tree approach follows the idea of a decision tree, but instead of the class labels, it has linear regression functions at the leaves, which can predict continuous numerical attributes. Model trees generalize the concepts of regression trees, which have constant values at their leaves (Witten & Frank, [6]). Therefore, they are analogous to piece-wise linear functions (and hence nonlinear). Computational requirements for model trees grow rapidly with increase in the dimensionality of the data set. Model trees learn efficiently and can tackle tasks with very high dimensionality. The major advantage of model trees over regression trees is that model trees are much smaller than regression trees and regression functions do not normally involve many variables. The working of M5 algorithm used in the present study for inducing a model tree is described in what follows:

The splitting criterion for the M5 model tree algorithm is based on treating the standard deviation of the class values that reach a node as a measure of the error at that node, and calculating the expected reduction in this error as a result of testing each attribute at that node.

The formula to compute the standard deviation reduction (SDR) is:

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \qquad (2)$$

where T represents a set of examples that reaches the node; T_i represents the subset of examples that have the i^{th} outcome of the potential set; and sd represents the standard deviation. After examining all the possible splits, M5 chooses the one that maximizes the expected error reduction. Splitting in M5 ceases when the class values of all the instances that reach a node vary just slightly, or only a few instances remain. This division often produces a large tree like structure that must be pruned back, for instance by replacing a subtree with a leaf. In the final stage, a smoothing process is performed to compensate for the sharp discontinuities that will inevitably occur between adjacent linear models at the leaves of the pruned tree, particularly for some models constructed from a smaller number of training examples. In smoothing, the adjacent linear equations are updated in such a way that the predicted outputs for the neighboring input vectors corresponding to the different equations are becoming close in value.

3 Data

The hydrological data collected for this study include Q, A, p, D, T, S_0 and C_f. The data pertaining to mean annual flood discharge (having no flow due to snow melt and free from upstream storage effects) and catchment properties for 93 river basin catchments as shown in Fig. 1 were taken from Swamee et al. [5].

The data used in the study are same as by Swamee et al. [5], i.e., mean annual flood peaks varying from 37.52 m^3/s to 56100 m^3/s, average rainfall of durations varying from 0.75 hr to 12 hr, recurrence intervals varying from

2 to 25 years, drainage areas ranging in size from 14.5 km^2 to 935, 000 km^2, catchment slope varying from 0.004 % to 0.69 % and fraction of forest area varying from 0.01 to 0.91.

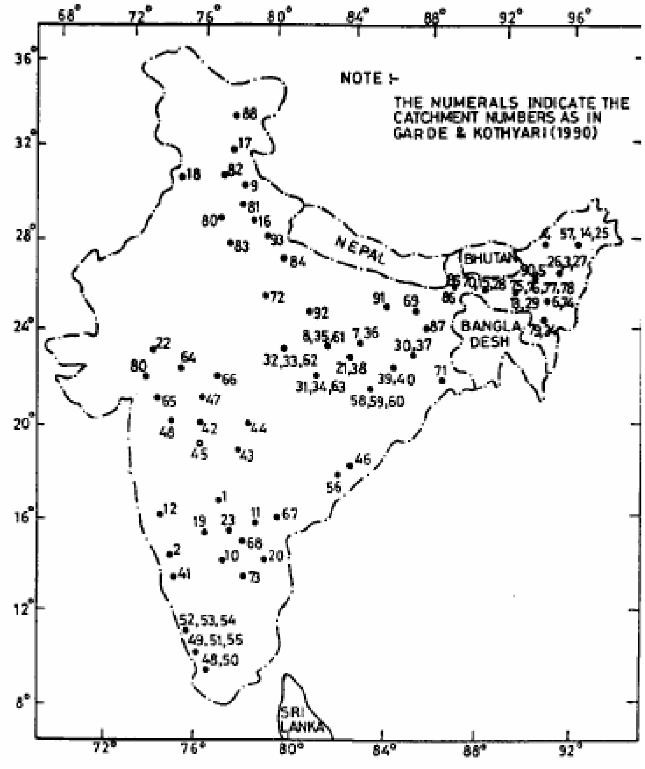

Fig. 1. Geographical Location of River Basin catchments in India

4 Experiments and Results

M5 model technique was applied for calculating the correlation coefficient (CC) and root mean square error (RMSE) by using cross-validation method to generate the model on the input data set comprising different parameters (drainage area, catchment slope, forest cover, average annual rainfall, average annual rainfall duration, return period, and mean annual flood). Cross-validation is a method of estimating the accuracy of a classification or regression model. The input data set was divided into several parts (a number defined by the user), with each part in turn used to test a model fitted to the remaining parts. For this study, a ten-fold cross-validation was carried out. For this study, M5 model tree as implemented by Witten and Frank [6] were used.

M5 model approach generates 94 linear models at leaf nodes. The pruned model tree for this data set is given below:

```
Area <= 5896.5 :
|  Area <= 471.5 : LM1 (612/2.451%)
|  Area >  471.5 :
|  |  slope <= 0.313 : LM2 (238/6.24%)
|  |  slope >  0.313 :
|  |  |  Area <= 3459.5 : LM3 (102/6.694%)
|  |  |  Area >  3459.5 : LM4 (51/3.393%)
Area >  5896.5 :
|  Area <= 84105.5 :
|  |  forest_cover <= 0.09 :
|  |  |  Area <= 12072.5 :
|  |  |  |  Area <= 8086.5 : LM5 (17/0%)
|  |  |  |  Area >  8086.5 :
|  |  |  |  |  slope <= 0.02 : LM6 (17/0%)
|  |  |  |  |  slope >  0.02 : LM7 (34/0%)
|  |  |  Area >  12072.5 :
|  |  |  |  Area <= 42101 : LM8 (34/0%)
|  |  |  |  Area >  42101 : LM9 (17/0%)
|  |  forest_cover >  0.09 :
|  |  |  slope <= 0.014 : LM10 (68/5.717%)
|  |  |  slope >  0.014 :
|  |  |  |  forest_cover <= 0.24 :
|  |  |  |  |  forest_cover <= 0.135 : LM11
(34/5.991%)
|  |  |  |  |  forest_cover >  0.135 :
|  |  |  |  |  |  forest_cover <= 0.17 : LM12
(68/5.181%)
|  |  |  |  |  |  forest_cover >  0.17 : LM13
(51/4.655%)
|  |  |  |  forest_cover >  0.24 :
|  |  |  |  |  Area <= 53357 :
|  |  |  |  |  |  Area <= 14958.5 : LM14 (17/0%)
|  |  |  |  |  |  Area >  14958.5 : LM15 (51/7.678%)
|  |  |  |  |  Area >  53357 : LM16 (34/0%)
```

Fig. 2. Pruned model tree obtained by using M5 modelling approach.

The Fig. 2 shows twenty one linear models as obtained by M5 model. This figure also suggests the importance of ‘Area’, ‘forest cover’ and ‘slope’ of the catchment to estimate the mean annual flood. This figure also suggests that the rainfall, rainfall duration and recurrence interval are not important parameters in modeling the mean annual flood as reported in several physically-based models.

Fig. 3 shows a graph between observed and predicted mean annual floods using an M5 model tree as well as and those computed by using Eq. (1). The results from Fig. 2 suggest that most of the values predicted by M5 model lie within a scatter of ±15 % error from the line of perfect agreement, whereas most of the predicted values with Eq. (1) lie outside the 30 % line, suggesting a better performance by M5 model in comparison to Eq. (1). Further, comparison of results from Table 1 and Fig. 2 as well as Fig. 3 suggest an improved performance by the M5 model tree approach in comparison to both neural network and Eq. (1). Thus, both modeling approaches outperformed the Swamee et al. [5] method and a substantial improvement in the predictive accuracy is obtained by using M5 model tree in comparison to Swamee et al. [5].

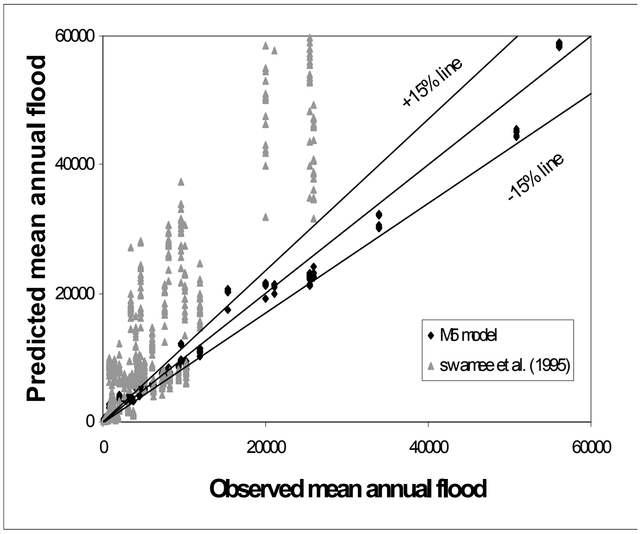

Fig. 3. Predicted mean annual flood using M 5 model & comparison with eq. (1)

Table 1: Results obtained by using M5 model tree

Algorithm used	Time taken to build model (s)	CC	RMSE
M5 model tree	0.49	0.994	1152.26
Swamee et al., [5]	-	0.897	37034.23

As the computational cost is an important parameter while using a machine learning approach, M5 model tree is also carried out in the present study. Results from Table 1 suggest that computational time taken by M5 model is quite small indicating easier handling of large data by M5 model.

5 Conclusions

This study concludes that M5 model tree-based model is superior in accuracy to the Swamee et al. [5] model. It can also be concluded that M5 model trees, provides insight into the created models, hence may be acceptable to decision makers. M5 model tree used a very small computational time and always converged. The accuracy of M5 tree is much better to Swamee et al. [5] for this data set, thus suggesting it can be an alternative to Swamee et al. [5] for regional mean annual flood estimation.

References

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

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

[3] Solomatine, D.P. and Dulal, K.N. "Model tree as an alternative to neural network in rainfall-runoff modeling", Hydrological Sci. J. 48 (3), 2003, 399–412.

[4] Solomatine, D. P. and Xue, Y. "M5 model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China," Journal of Hydrologic Engineering, Vol. 9, No. 6, 2004, 1-10.

[5] Swamee, P.K., Ojha, C.S.P. and Abbas, A. "Mean annual flood estimation" Journal of Water resources planning and management, Vol. 121, 1995, 403-407.

[6] *Witten, I. H. and Frank, E. "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.*