4th Intercontinental Geoinformation Days

igd.mersin.edu.tr

Predicting the monthly flow of the Kaleybar Chay River based on M5 model tree

Kambiz Falsafian *[1]

[1]*University of Tabriz, Marand Technical College, Tabriz, Iran*

Keywords
River flow
Statistical indicators
M5 model tree

Abstract

Due to its importance in the designing of water projects, flood investigation and estimation has always been one of the key issues in the field of hydrology and researchers have tried to estimate the river flow more accurately by using different methods. In this regard and in the present study, the monthly discharge values of Kaleybar Chay River were predicted using machine learning method of M5 model tree. Based on this, time series data of discharge and precipitation in monthly delays are used as input parameters of the models and the results are evaluated using the statistical indicators of correlation coefficient, root mean square error and mean absolute error. Finally, the values obtained from the M5 models showed that among the studied models, M5(6) model with the root mean square error of 0.968 and the mean absolute error of 0.625, had the highest correlation with the observed values and recorded the most accurate results in this study. It was also found that most of the M5 models had a successful performance in estimating the monthly flow in the study area.

1. Introduction

Control and use of surface water have great important especially in areas with water shortage problems and seasonal rivers. Accurate estimation of river flow is necessary for planning and managing water resources and organizing the river. It is also important to predict the flow rates of the river from various aspects such as the optimal operation of dam reservoirs. In addition, river flow is one of the influential factors in the phenomena of drought, floods, sources of drinking water supply and in general issues related to water systems. In recent years, researchers have used various methods to estimate river discharge values, among which, machine learning models have shown better performance due to high accuracy and the need for less cost and time. Among the studies conducted in this field, the following can be mentioned:

Cannas et al. (2005) estimated the monthly flow of the Tirso Basin River in Sardinia, Italy, using a combined neural network model and wavelet analysis, and investigated the effect of data preprocessing on the neural network using discrete and continuous wavelet transforms. The results showed that the integrated model is more accurate than the neural network model. Huang et al. (2014) predicted monthly discharge values in Wei River Basin using a specific hybrid model of empirical mode decomposition-support vector machine (EMD-SVM). Comparison of the results showed that in all studied stations, the model performed better than the ANN and SVM methods. In another study, Nouri and Kalin (2016) simulated daily river flow rates in Atlanta, USA. They first used the SWAT model to simulate daily flow and used the results obtained from this model as input to the artificial neural network method. Finally, it was shown that the combination of semi-distributed models with artificial neural network method improves the accuracy of river flow prediction in the study area.

The M5 model tree is also one of the machine learning methods that has been used in recent years to predict many hydrological phenomena. One of them is the study of Sattari et al. (2013). They evaluated the ability of the M5 model tree to predict the daily discharge of the Sohu Stream in Ankara, Turkey. Sattari et al. (2013) compared the results using statistical indicators with the support vector machine method and showed that in general the M5 model had a better performance. Also, using M5 model tree, Zahiri and Azamathulla (2014) estimated river flow, Singh et al. (2010) estimated the average annual flood, Shaghaghi et al. (2019) predicted the

* **Corresponding Author**

*(kfalsafian@tabrizu.ac.ir) ORCID ID 0000–0002–1567–7976

Cite this study

Falsafian, K. (2022). Predicting the monthly flow of the Kaleybar Chay River based on M5 model tree. 4th Intercontinental Geoinformation Days (IGD), 74-77, Tabriz, Iran

dimensions of the river regime and Unes et al. (2020) modeled river flow.

According to the studies conducted in the field of using machine learning methods in modeling hydrological phenomena, the high importance of estimating river discharge values by applying the mentioned techniques can be inferred. Therefore, the purpose of this study is to predict the monthly discharge values of Kaleybar Chay River in Kaleybar station applying M5 model tree and using discharge (in two-time delays of one and two months) and precipitation (without time delay and time delay of one Month) data.

2. Method

2.1. Study area

Kaleybar Chay Basin with an area of 144960 hectares is located in northwestern Iran and north of East Azerbaijan province. This basin is considered as a subset of Aras basin and is located between the geographical coordinates of 46° 40′ to 47° 13′ east longitude and 38° 39′ to 39° 09′ north latitude. Kaleybar Chay River is the most important river of Kaleybar Chay Basin and one of the permanent rivers of East Azerbaijan province. This river originates from the heights of Qara Dagh and the main branch of this river passes through Kaleybar and finally flows into Aras River. In the present study, the flow and precipitation data of Kaleybar Chay River in Kaleybar station from 2002 to 2015 have been used on a monthly scale, so that the flow data with two-time delays of one and two months and the precipitation data has been applied with a delay of one month and without a time delay.

In this study, using different combinations of time series of discharge and precipitation data as input of M5 model tree, river discharge values were estimated. Table 1 shows the different combinations of input parameters of the models.

Table 1. Different combinations of input parameters of the studied models

Combination Number	Input Parameters	Output Parameter
1	P_t	Q_t
2	Q_{t-1}	Q_t
3	P_t, Q_{t-1}	Q_t
4	P_{t-1}, Q_{t-1}	Q_t
5	P_{t-1}, P_t, Q_{t-1}	Q_t
6	P_{t-1}, P_t, Q_{t-1}, Q_{t-2}	Q_t

2.1. M5 model tree

The M5 model tree (Quinlan 1992) is a subset of machine learning and data mining methods. Data mining refers to the process of searching for and discovering various models, summarizing, and obtaining values from a set of known values. Data mining methods are designed for large data sets with many variables, so they are different from older statistical methods designed for small data sets with small variables. Decision tree-based methods as one of the most well-known data mining techniques, predict or classify the objective property as output in the form of a model with a tree structure using input data. M5 model is a tree model for predicting continuous numerical traits in which linear regression functions are displayed on the leaves of this tree (Sattari et al. 2013), which in recent years has made a significant change in classification and predictions issues. Decision trees are a useful solution to many classification problems that use complex databases and complex or erroneous information. Decision trees, which have predictive and descriptive properties, are the most widely used classification models because of their easy installation, interpretation, and integration into database systems, and better reliability. The division criterion is based on the standard deviation of the subset values. The mathematical formula for calculating the standard deviation reduction (SDR) is as follows:

$$SDR = SD(T) - \sum \frac{T_i}{T} \times SD(T_i) \quad (1)$$

In Equation (1), T represents a group of samples that are bound, Ti represents a subset of samples that is the product of a potential group, and SD represents a standard deviation. After examining all possible structures, a structure is selected that has the maximum expected error reduction. This division process often produces an excellent tree-like structure that leads to an over-appropriate structure (Unes et al. 2020).

2.2. Criteria for evaluating the accuracy of models

The error values between the applied models and the observational data were evaluated by correlation coefficient (R), root mean square error (RMSE) and mean absolute error (MAE) using the Equations 2 to 4.

$$R = \frac{(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)}{(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2)(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2)} \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \quad (4)$$

In Equations 2 to 4, x_i and y_i are the observed and predicted monthly flow rates, respectively, and n is the number of observations.

3. Results

In this study, using time series of discharge and precipitation data and using M5 model tree, the monthly discharge values of Kaleybar Chay River in Kaleybar station were estimated. Then the results of the mentioned methods were compared by statistical indices

of correlation coefficient, root mean square error and mean absolute error and the most appropriate and best model for predicting river discharge in the study area was determined and introduced. Table 2 shows the values of statistical indicators for different models with different combinations of input parameters.

Table 2. Statistical indicators of different flow estimation models

Model	R	RMSE	MAE
M5(1)	0.437	1.38	0.815
M5(2)	0.678	1.1	0.693
M5(3)	0.708	1.06	0.675
M5(4)	0.678	1.1	0.693
M5(5)	0.708	1.06	0.675
M5(6)	0.77	0.968	0.625

Figure 1 shows the bar graph of statistical indicators of all studied models. Figure 2 demonstrates the temporal changes of river flow using the best studied models. Also figure 3 shows the distribution diagram of discharge values calculated by the superior models compared to the observed discharge.

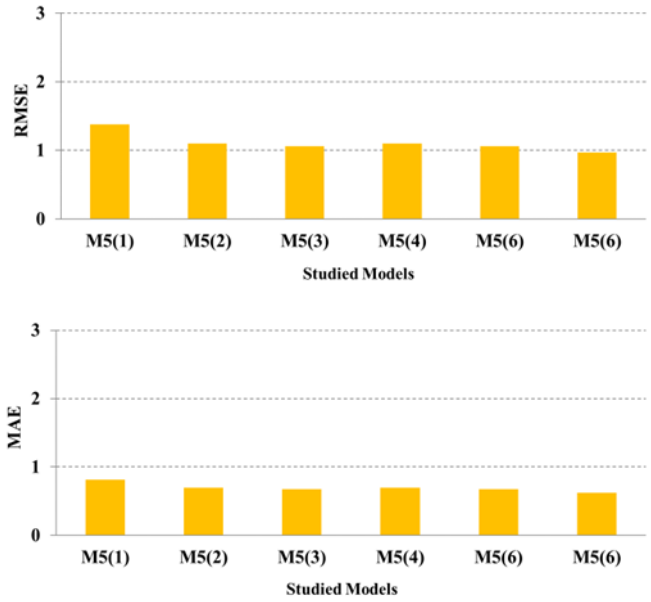

Figure 1. Bar graph of statistical indicators for all studied models

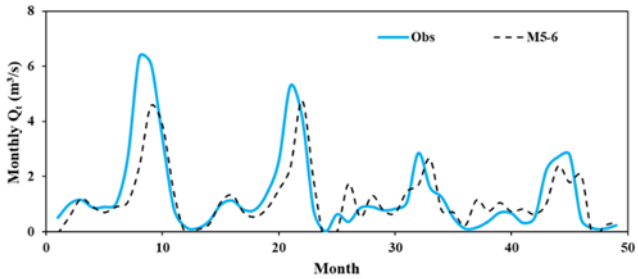

Figure 2. Diagram of temporal changes in river flow using the best studied models

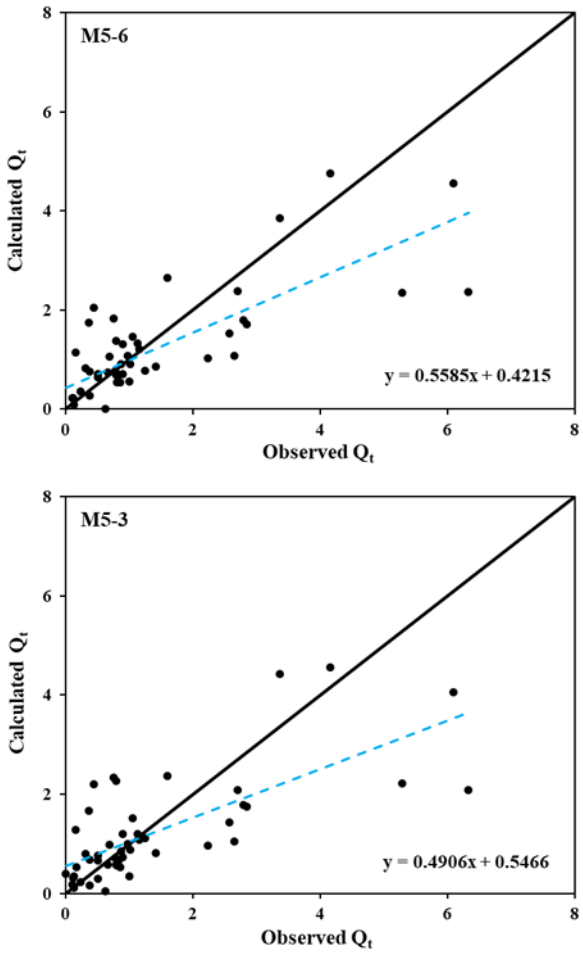

Figure 3. Distribution diagram of discharge values calculated by the superior models in comparison with the observed discharge

4. Discussion

According to the obtained results (Table 2), M5(6) model with correlation coefficient of 0.77, root mean square error of 0.968 and mean of absolute error of 0.625 had the best performance. M5(3) and M5(5) models were in the next position with the same performance and correlation coefficient of 0.708, root mean square error of 1.06 and mean absolute error of 0.675 with input parameters of P_t, Q_{t-1} and P_{t-1}, P_t, Q_{t-1}, respectively. In general, all implemented models provided acceptable and good performance. However, by comparing the results, it was found that except for the first model, other models can be used to estimate the flow rate with the desired accuracy in the Kaleybar Chay River.

According to Figure 1, the mentioned trend about the high accuracy of M5 models can also be concluded from this figure. Figure 2 also shows the high agreement of the superior models with the observational data. Similarly, Figure 3 displays the lower distribution of points of the superior models around the axis of the half-instrument.

5. Conclusion

Estimating the flow rate of rivers in each region is one of the most important and fundamental issues in planning and managing water resources. Therefore, in this study, the discharge values of Kaleybar Chay River

were modeled on a monthly scale. Accordingly, the M5 model tree technique was used and the time series of precipitation and discharge data were utilized as input data of this model in different combinations. The results showed that most of the studied models had acceptable and good performance so that M5(6) model with the root mean square error of 0.968 and the mean of absolute error of 0.625 estimated the most accurate values. In general, it can be concluded that using the superior models of this research, the flow rate of Kaleybar Chay River can be estimated with good accuracy.

References

Cannas, B. & Fanni, A. & Sias. G. & Tronci, S. and Zedda, M.K. (2005), River flow forecasting using neural networks and wavelet analysis. Geophysical Research Abstracts, 7, 08651, 1-11.

Huang, S., Chang, J., Huang, Q., & Chen, Y. (2014). Monthly streamflow prediction using modified EMD-based support vector machine. Journal of Hydrology, 511, 764-775.

Noori, N., & Kalin, L. (2016). Coupling SWAT and ANN models for enhanced daily streamflow prediction. Journal of Hydrology, 533, 141-151.

Quinlan, J. R. (1992). Learning with continuous classes. Fifth Australian Joint Conference on Artificial Intelligence, 343-348, World Scientific. Singapore.

Sattari, M. T., Pal, M., Apaydin, H., & Ozturk, F. (2013). M5 model tree application in daily river flow forecasting in Sohu Stream, Turkey. Water Resources, 40, 233-242.

Shaghaghi, S., Bonakdari, H., Gholami, A., Kisi, O., Binns, A., & Gharabaghi, B. (2019). Predicting the geometry of regime rivers using M5 model tree, multivariate adaptive regression splines and least square support vector regression methods. International Journal of River Basin Management, 17, 333-352.

Singh, K. K., Pal, M., & Singh, V. P. (2010). Estimation of mean annual flood in Indian catchments using backpropagation neural network and M5 model tree. Water Resources Management, 24, 2007-2019.

Unes, F., Demirci, M., Zelenakova, M., Calisici, M., Tasar, B., Vranay, F., & Ziya Kaya, Y. (2020). River Flow Estimation Using Artificial Intelligence and Fuzzy Techniques. Water, 12(2427), 1-21.

Zahiri, A., & Azamathulla, H. M. (2014). Comparison between linear genetic programming and M5 tree models to predict flow discharge in compound channels. Neural Computing and Applications, 24, 413-420.