Hindawi
Complexity
Volume 2021, Article ID 3721661, 14 pages
https://doi.org/10.1155/2021/3721661

Research Article

On the Investigation of Monthly River Flow Generation Complexity Using the Applicability of Machine Learning Models

Ma Shaofu[1] **Anas Mahmood Al-Juboori**[2] **Asmaa Hussein Alwan**[3] **and Abdel-Salam G. Abdel-Salam**[4]

[1]*College of Physical Education, Lanzhou City University, Lanzhou 730070, China*
[2]*Dams and Water Resources Research Center, University of Mosul, Mosul, Iraq*
[3]*College of Education for Human Science-Ibn Rushd, University of Baghdad, Baghdad, Iraq*
[4]*Department of Mathematics, Statistics and Physics, College of Arts and Sciences, Qatar University, Doha, Qatar*

Correspondence should be addressed to Asmaa Hussein Alwan; asmaahussienalwan@gmail.com and Abdel-Salam G. Abdel-Salam; abdo@qu.edu.qa

Received 4 June 2021; Revised 20 June 2021; Accepted 24 June 2021; Published 7 July 2021

Academic Editor: Haitham Afan

Streamflow is associated with several sources on nonstationaries and hence developing machine learning (ML) models is always the motive to provide a reliable methodology to understand the actual mechanism of streamflow. The current research was devoted to generating monthly streamflows from annual streamflow. In this study, three different ML models were applied for this purpose, including Multiple Additive Regression Trees (MART), Group Methods of Data Handling (GMDH), and Gene Expression Programming (GEP). The models were developed based on annual streamflow and monthly time index of three rivers (i.e., Upper Zab, Lower Zab, and Diyala) located in the north region of Iraq. The modeling results indicated an optimistic simulation for generating the monthly streamflow time series from annual streamflow time series. The potential of the MART model was superior to the GMDH and GEP models for Upper Zab River (R^2 0.84, 0.64, and 0.47), Lower Zab River (R^2 0.75, 0.46, and 0.40), and Diyala River (R^2 0.78, 0.42, and 0.5). The results of RMSE were 113, 169, and 208 for Upper Zab River, 95, 149, and 0.5 for Lower Zab River, and 73, 118, and 109 for Diyala River. The results have proved the possibility of changing the timescale in generating streamflow data.

1. Introduction

The hydrological processes are associated with several elements such as evaporation, evapotranspiration, precipitation, runoff, river flow, infiltration, and groundwater. In nature, the hydrological cycle is featured by high stochasticity, nonstationarity, and nonlinearity [1], and thus studying the hydrological process is one of the significant topics in the field of water resources engineering. Over the past literature, several models have been introduced for modeling hydrology cycle processes and evidently proofed their capacity [2–4]. Between several components of the hydrology cycle, streamflow is a very important process and has received major interest by the hydrologists and computer scientists [1]. The establishment of accurate and reliable models "forecasting, prediction, or optimization" for the long scale, such as yearly, seasonally, or monthly, is very magnificent for reliable water resources management and planning [5]. In addition, for short scale like day, hour, and minutes, streamflow recording is very essential for flooding warning and monitoring in order to lessen and mitigate their effects on various structure and human well-being [6].

The data-driven streamflow models are regression-based where the relationships between model inputs and output are directly defined [7, 8]. With the advances of computer aided models, ML models such as fuzzy logic, neural network, nature-based algorithm, support vector machine, decision tree, and optimizers have been successfully implemented for modeling streamflow patterns. These models can help in detecting the nonlinear, dispensable, and dynamic pattern of the time series [9–12]. However, a number of problems are associated with most of the ML-based techniques due to their inherent limitations [13]. The ML-based models need previous information of the stochastic behavior of the addressed research issue (i.e., hydrology or climatology processes or water quality data) [14, 15]. Hence, it is essential to configure reliably in terms of learning process to obtain the important information from the chronological data of streamflow. In addition, it is required to optimize a number of model internal parameters [16, 17]. Over the time, many hybrid models have been also implemented such as fractionally autoregressive integrated moving average (FARIMA) and self-exciting threshold autoregressive (SETAR) with GEP, MARS, and MLR [18]. Similarly, another authors used autoregressive conditional heteroscedasticity (ARCH) to hybridized GEP and MARS models [19]. In particular, the conventional ML-based models need numerous trial-and-error processes to determine the optimum architecture design. For example, hydrological models using neural network require optimization of the number of hidden layers, the type of the transfer function, and the number of neurons in a hidden layer's choices [20]. Hybrid models are one of the updated models that have started to be used extensively in hydrology science [21]. Correspondingly, fuzzy models are one of the traditional models that lack handling complex problems and too many rules [22] and same goes for MLR model when dealing with multiple output and complexity [23]. The highlighted limitations of the existing ML-based streamflow forecasting models have necessitated the search for more sophisticated ML-based modeling techniques.

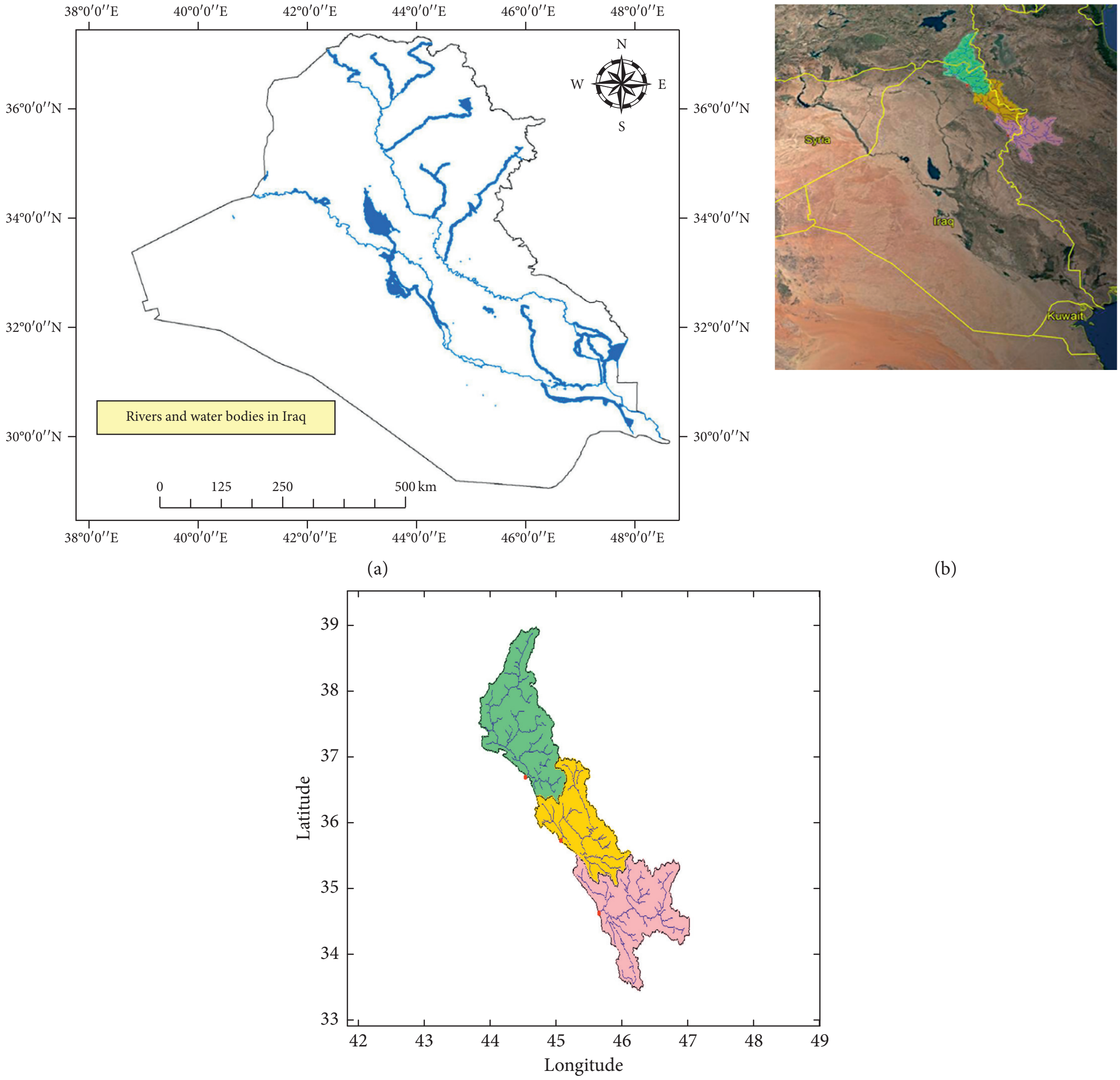
Streamflow forecasting plays an essential role for the researchers and engineers to better understand the river pattern which in turn helps to design more sustainable and efficient infrastructure and management project. Streamflow data is important yet presents itself with various issues such as missing data, noncontinuous data, nonlinearity, and extreme events [24, 25]. Researchers have devised various techniques and tools to overcome them, yet to grasp the full scope of such data in terms of seasonality, point source pollution, and sudden changes due to event of heavy rain or other calamities, and more work is needed to be done. Disaggregating streamflow can sever an essential procedure for reservoir operation and river basin management in general [26, 27]. This topic has received an extensive capacity by several hydrology scholars. Stedinger and Vogel [28] developed a simple class of a disaggregation model that can reproduce a covariance matrix of streamflow and reasonable approximation to the lead times that should be imposed for the disaggregation approach. Of recent advanced computer models, the disaggregation procedure was investigated by several scholars. A stochastic model was proposed to disaggregate streamflow at multiple sites preserving their temporal and spatial dependencies [29]. An integrated nonparametric model with genetic algorithm was to simulate seasonal streamflow disaggregating [30]. Monthly streamflow scale was disaggregated into daily scale using simple stochastic, as conducted in [31]. Various other research studies were conducted on the streamflow disaggregation [32–35]. All the reported research over the literature evidenced the capacity of studying the streamflow disaggregation. However, the implementation of the ML models for the streamflow disaggregation is limited and needs to be investigated. ML models such as Multiple Additive Regression Trees (MART), Group Methods of Data Handling (GMDH), and Gene Expression Programming (GEP) are yet to be explored for the generating monthly streamflow time series from annual streamflow time series. There was no established research over the literature using those models yet to be tested.

The main objective of the current research is to investigate the feasibility of MART, GMDH, and GEP models for generating monthly streamflow time series from annual streamflow time series. The proposed models represent three different types of ML models. The MART model is one of the most popular decision tree models that strengthen the weak learning, which results in strong learning process and better generalization [36], while the GMDH model is chosen to represent self-learning models. The GEP model was applied as revolutionary model. The proposed models were evaluated statistically among each other and analyzed based on their predictability capacity. The study aims to demonstrate the possibility of changing the timescale in generating streamflow. This is the first application of using the GMDH, GEP, and MART models to generate monthly streamflow data from annual monthly streamflow data without using method of fragments which is usually used to disaggregate the annual streamflow to monthly streamflow.

2. Materials and Models

2.1. Study Area and Data. Upper Zab, Lower Zab, and Diyala Rivers are the major tributaries of Tigris River in Iraq, which were selected for the case study in this research. The Tigris River is one of the largest rivers in the Middle East. The river is about 1718 km long that goes through Turkey then Syria then Iraq. However, the major percentage (253,000 km) of about 85% of the river travels through Iraq region. The Tigris River along with the Euphrates River contributes to the Iraqi region as the main natural resources of fresh water that is required for diverse necessity of water usages. The Upper Zab River headwaters are located in Turkey's territory, while the headwaters of the Lower Zab and Diyala rivers are located in Iran's territory [37]. Figure 1 shows the location of Upper Zab, Lower Zab, and Diyala Rivers in Iran. Table 1 summarizes the morphological and flow data characteristics for the Upper Zab, Lower Zab, and Diyala upstream Bekhme, Dokan, and Derbindi-Khan flow gauging stations, respectively. The climate of the basin is predominantly semiarid. The temperature in the basin varies from maximum 45°C during summer to minimum 10°C in winter. The mean monthly discharge and the standard deviation of Tigris River flow at Baghdad station are 411.35 m^3/s and 234.52 m^3/s, respectively. Monthly flow data for the period 1932–2004 were selected. This period was selected because there was no missing data during this period. The first 70% of data was selected for training the models, while the second 30% of data was selected to validate the models.

2.2. Introduction to the Gene Expression Programming (GEP) Model. GEP was invented by Ferreira as an extension of traditional genetic programming. The program is developed



(a)

(b)

(c)

Figure 1: Upper Zab, Lower Zab, and Diyala Rivers' location.

as linear strings of fixed chromosome's length and then encoded as a nonlinear form with different dimensions [38]. In GEP, expressions are generated automatically by encoding the expression in the form of a tree consisting of nodes representing functions and leaves (terminal) representing constants and variables. The generated candidates were evaluated by a fitness function. The genes included two parts: tail that includes variables and head that includes variables and constants [39]. Five steps are used to develop the GEP model: (i) selecting a set of predictor variables, which can be used in discrete programs; (ii) selecting the specific functions and arithmetic operations; (iii) choosing the fitness measure; (iv) selecting the appropriate head length, quantity of genes, and the linking function; and (v) selecting the genetic operators which include inversion rate and mutation rate [40]. More details for GEP are found in [41]. Figure 2 shows the flowchart of gene expression programming algorithm.

2.3. Introduction to the Multiple Additive Regression Trees (MART) Model. MART was developed by Derrig and Francis [42] to increase the accuracy of the traditional decision tree model result. The researchers found that the models developed using MART are more accurate models in comparison with any known modeling methodologies. The model can handle categorical and continuous inputs and target variables. The model is more stable due to the use of

Table 1: Morphological and flow data characteristics.

	Upper Zab	Lower Zab	Diyala
Basin area (km^2)	16863	11706	15765
Basin slope (m/m)	0.350	0.265	0.252
Average over land flow (m)	5.878	6.05	6.92
Perimeter (km)	1141.2	1053	1368.16
Basin length (km)	189.9	125	165.76
Mean basin elevation (m)	1870	1381.60	1551.15
Flow data record	1932–2004	1932–2004	1932–2004
Maximum flow (m^3/s)	1681	1135	947
Minimum flow (m^3/s)	32	9	3
Standard deviation (m^3/s)	277	180	148
Mean (m^3/s)	369	193	150

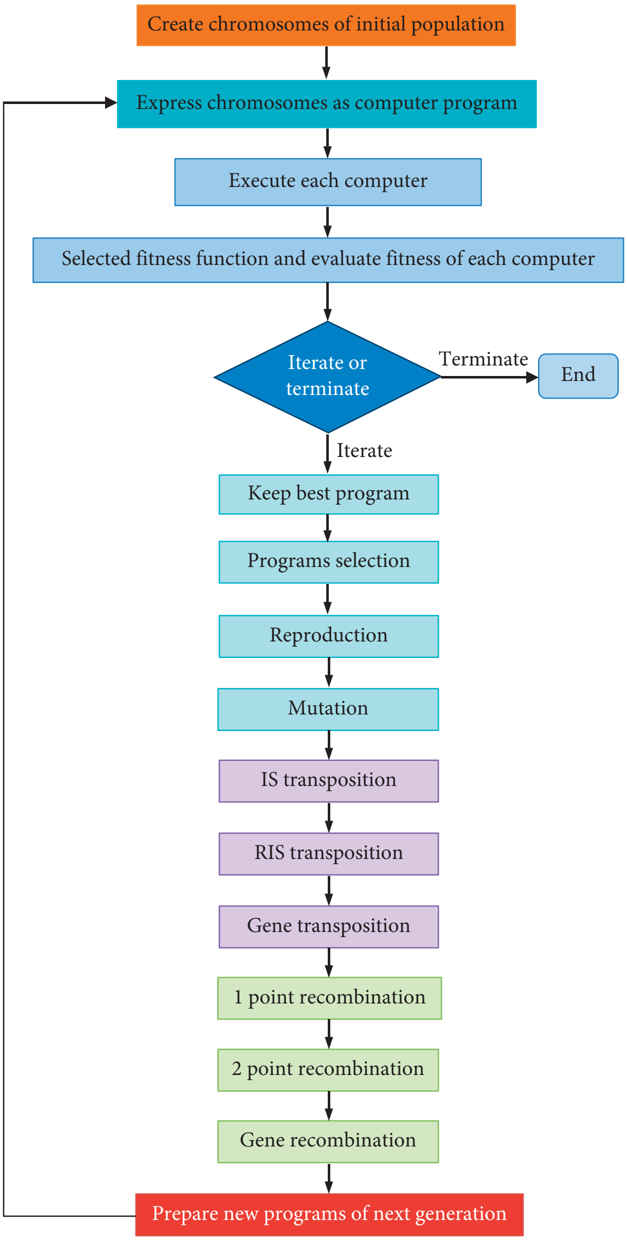


Figure 2: Flowchart of gene expression programming algorithm (Ferreira, 2001).

the Humber M-regression loss function in its algorithm. MART algorithm is started by fitting the inputs to first tree and then the biases from the first tree are inserted to the next tree to minimize the error [43]. This procedure is repeated through a series of following trees. The final results are adjusted by adding contribution weight of each tree. The MART algorithm can be expressed as [36]

$$\text{Target} = S + C_1 \times T_1(N) + C_2 \times T_2(N) + \cdots + C_n \times T_n(N), \tag{1}$$

where S is the mean value of the target variable; N is a pseudoresidual as set value's vector, $T_1(N), T_2(N), \ldots T_n(N)$ is tree fixed to the pseudoresiduals, and $C_1, C_2, \ldots, C_n$ are the tree node predicted coefficients. Figure 3 shows a simple MART structure and Figure 4 shows the flowchart of random trees algorithm.

2.4. Introduction to the Group Method of Data Handling (GMDH). GMDH was developed to solve the problems of predication, complex system, and optimization by using a nonlinear regression algorithm. GMDH structure is classified as a self-organizing polynomial neural network's method [44]. GMDH is a specific type of supervised artificial neural network. The algorithm of GMDH uses the concept of natural selection to control the network size, complexity, and accuracy [45]. The GMDH model starts by selecting a set of functions that showed highest prediction accuracy at previously unseen data. In GMDH model, layers of neurons are created using one or more inputs. The connections between neurons in the network are self-selected during training phase. The determination of number of layers and neurons in the network is automatic. The GMDH solutions are subsets of functions called partial models [46]. The best model is reached by gradually increasing the number of partial models. The GMDH algorithm uses the two variables' quadratic equation to develop the model.

$$\begin{aligned} y &= a_1 + a_2 x_1 + a_3 x_2^1 + +a_4 x_2 + a_5 x_2^2 + a_6 x_1 x_2, \\ A &= \left(X^T X\right)^{-1} X^T Y, \end{aligned} \tag{2}$$

where $Y = [y_1, y_2, \ldots, y_n]^T$ and $A = [a_1, a_2, a_3, a_4, a_5]$.

$$X = \begin{bmatrix} 1 & x_{1I} & x_{1J} & x_{1I}x_{1J} & x_{1I}^2 & x_{1J}^2 \\ 1 & x_{2I} & x_{2J} & x_{2I}x_{2J} & x_{2I}^2 & x_{2J}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{pI} & x_{pJ} & x_{pI}x_{pJ} & x_{PI}^2 & x_{PJ}^2 \end{bmatrix}, \tag{3}$$

where m presents the number of variables, (x_1, x_2, x_6) are vectors of input variables, and $(a_1, a_2, \ldots, a_6)$ are vectors of parameters. Figure 5 shows the structure of GMHD and Figure 6 shows the flowchart of GMDH algorithm. More details of GMHD are found in [44].

2.5. Performance Evaluation. In this research, two different performance metrics were selected to evaluate the proposed models: coefficient of determination (R^2) and root mean square error (RMSE) [47].

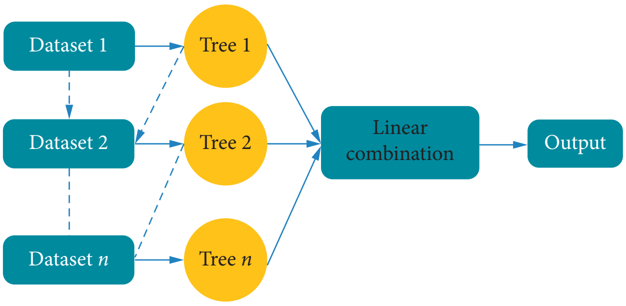


Figure 3: The structure of the MART model.

$$\begin{aligned} R^2 &= \left[\frac{\sum_{i=1}^{n} \left(Q^{\circ} - \overline{Q}_o\right)\left(Q_p - \overline{Q}_p\right)_i}{\sqrt{\sum_{i=1}^{n} \left(Q^{\circ} - \overline{Q}_o\right)^2 \left(Q_p - \overline{Q}_p\right)_i^2}} \right]^2, \\ \text{RMSE} &= \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(Q_o - Q_P\right)^2}, \end{aligned} \tag{4}$$

where Q_{io} and Q_{ip} are the observed and generated streamflow values, respectively, $\overline{Q}_o$ is the observed streamflow mean value, and n is the data record number. The best models are those which showed low RMSE and are close to 1 value for R^2.

3. Modeling Results and Discussion

In this study, the three ML models were applied to develop the best models to generate monthly streamflow from annual streamflow. The models were developed using monthly streamflow as a target variable while the annual streamflow and monthly time index as predictor variables. The time index is an index that represents the monthly sequence within a year, and its values range from 1 (January) to 12 (December). Selecting the best model for predicting the monthly streamflow of the three proposed rivers requires choosing the best model settings for the MART, GMDH, and GEP models. The best MART model requires selecting the best settings for the three parameters in the model that includes the amount of trees in series, depth of discrete trees, and number of splits (least size). These values for the three rivers are 600, 5, and 10 for the Upper Zab River, 800, 5, and 10 for the Lower Zab River, and 300, 5, and 10 for Diyala River. The best GMDH model requires selecting the best settings for the four parameters in the model that include maximum network layers, maximum polynomial order, number of neurons per layer, and network layer connections type. The optimum parameter's settings of the GMDH model for the three rivers in this study are 20 for maximum network layers and 16 for maximum polynomial order, same number of neurons as inputs' option for the number of neurons per layer, and previous layer and original input variables for the network layer connections' type. There are five major steps for GEP modeling in this study: (1) selecting the set of functions to be used: 5 basic mathematical functions were used: +; −; ×; ÷; and power; (2) selecting the

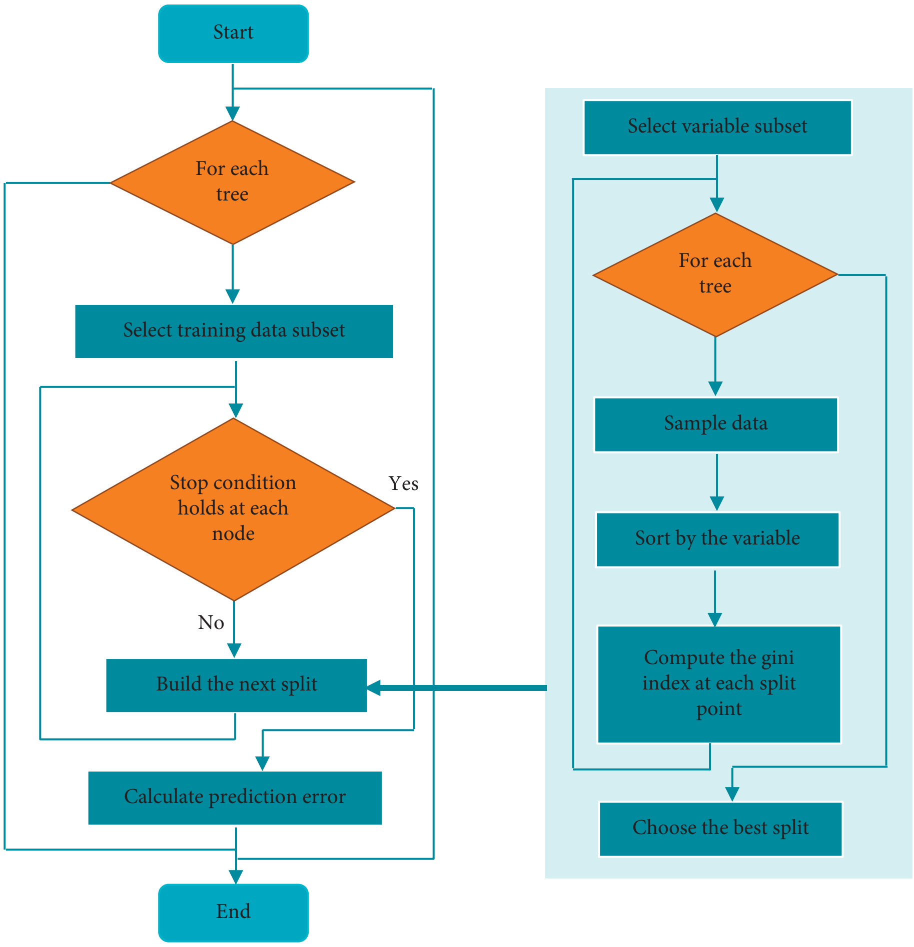


Figure 4: Flowchart of random trees algorithm.

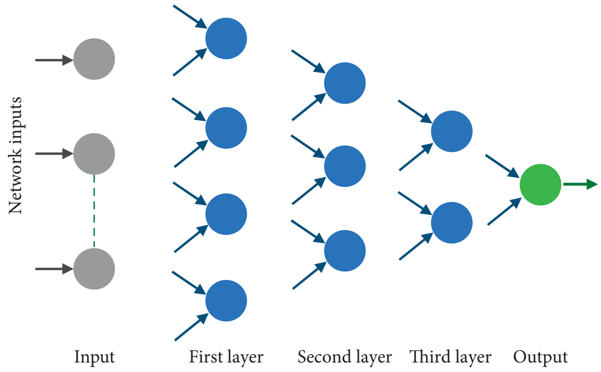


Figure 5: The structure of the GMDH model.

fitness function: the root relative square error (RRSE) was selected; (3) choosing the optimum general parameters: here, population size, genes per chromosomes, and gene head length were chosen; (4) choosing linking function: addition was chosen; and (5) we selected genetic operators, mutation rate of 0.044, and inversion rate of 0.1. Table 2 summarizes the best GEP model setting for the three rivers.

The results of the optimum symbolic fit regression functions from GEP model are explained in the following generated expressions for the Upper Zab, Lower Zab, and Diyala Rivers, respectively:

$$\begin{aligned} Q_m &= \frac{-90151.84}{T} - 0.0012708T^2 + 80992.678 \\ &\quad + Q_a + \frac{89130.39}{T}, \\ Q_m &= 167.44 - T + Q_a - 16.4182T - T, \\ Q_m &= 262.276 - 22.8876T + Q_a \\ &\quad + \sqrt{Q_a} + 383.9346 + \frac{64.230}{T - 2.6546}, \end{aligned} \tag{5}$$

where Q_m is a monthly flow, Q_a is an annual flow, and T is a time index (1, 2, 3, . . ., 12). The values of monthly streamflow change with the change in the value of the time index in the previous functions. The performance of the proposed models was evaluated utilizing the couple of statistical

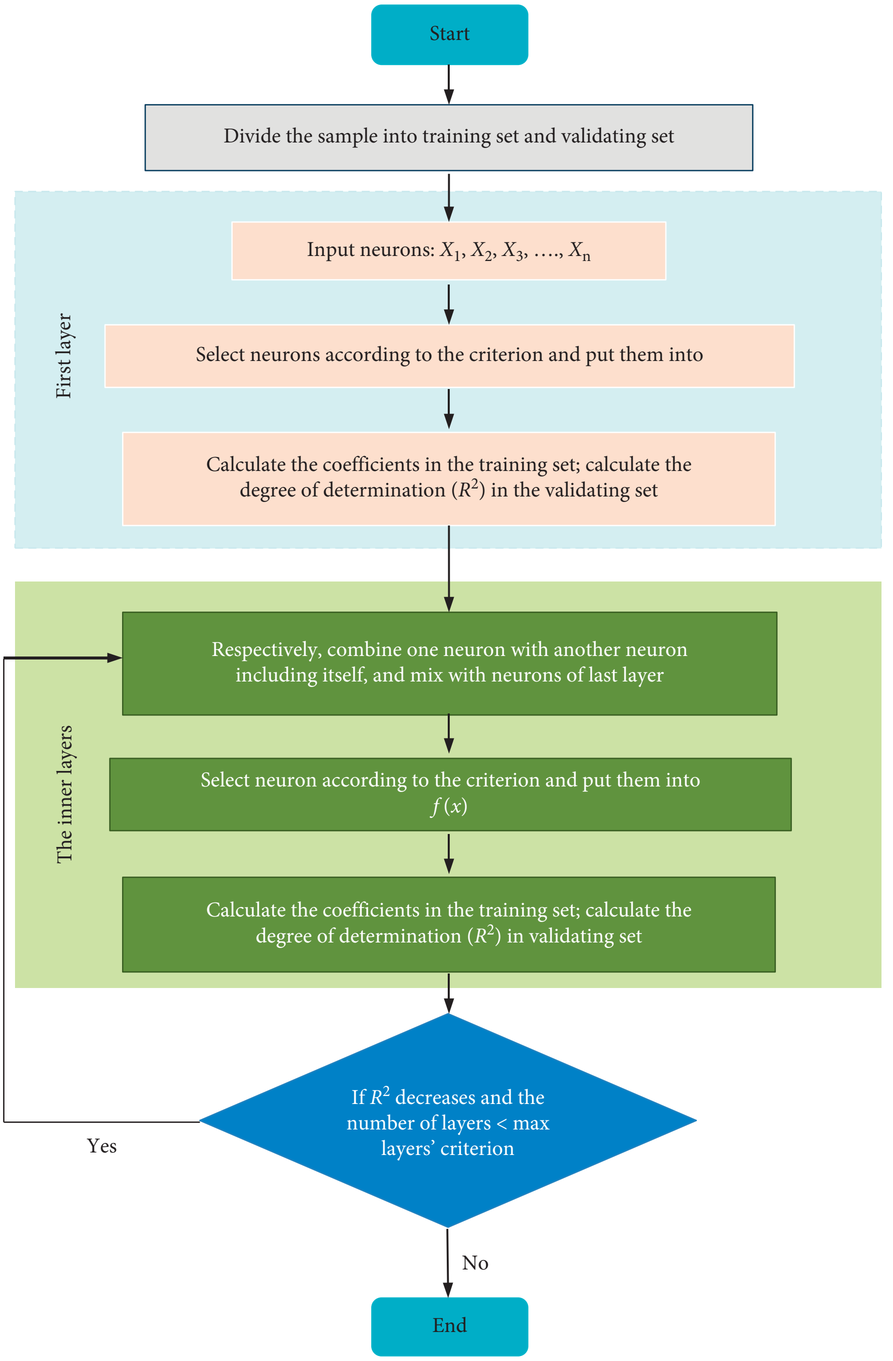

Start
Divide the sample into training set and validating set
First layer
Input neurons: X_1, X_2, X_3, …., X_n
Select neurons according to the criterion and put them into
Calculate the coefficients in the training set; calculate the degree of determination (R^2) in the validating set
The inner layers
Respectively, combine one neuron with another neuron including itself, and mix with neurons of last layer
Select neuron according to the criterion and put them into $f(x)$
Calculate the coefficients in the training set; calculate the degree of determination (R^2) in validating set
If R^2 decreases and the number of layers < max layers' criterion
Yes
No
End

Figure 6: Flowchart of GMDH algorithm.

metrics and graphical visualization. Based on the reported statistical result in Table 3, the Upper Zab River simulation has shown that the performance of the MART model is superior over the performance of GMDH and GEP models. The R^2 values are 0.93, 0.81, and 0.53 for the training phase and 0.84, 0.64, and 0.47 in the validating phase for MART, GMDH, and GEP models, respectively. On the other hand, and using the absolute error measures, the results of RMSE have proved the accuracy of MART model in comparison with the other models, which can be due to the model of learning used by MART model where weak learning is boosted by regression learning leading to higher accuracy. The RMSE values are 85, 141, and 222 m^3/s in training phase and 113, 169, and 208 m^3/s in the validating phase for the MART, GMDH, and GEP models, respectively. For the Lower Zab River, R^2 values are 0.90, 0.47, and 0.41 in training phase and 0.75, 0.46, and with lowest of 0.40 through the validating phase for the MART, GMDH, and GEP models, respectively. The RMSE results have also proved the performance of MART model in comparison with the GMDH and GEP models. The RMSE values during the validating phase are 62, 144, and 151 m^3/s, and during the training

Table 2: Model setting.

	Upper Zab	Lower Zab	Diyala
Function set			
Addition	+	+	+
Subtraction	−	−	−
Multiplication	×	×	×
Division	÷	÷	÷
Power	**	**	**
General parameters			
Population size	50	100	100
Genes per chromosomes	4	4	4
Gene head length	8	8	8
Fitness function	RRSE	RRSE	RRSE
Linking function	Addition	Addition	Addition
Genetic operators			
Mutation rate	0.044	0.044	0.044
Inversion rate	0.1	0.1	0.1

Table 3: The performance metrics of the applied ML predictive models through the training and validating phases when modeling the three investigated rivers.

	Upper Zab				Lower Zab				Diyala			
Model	Training		Validating		Training		Validating		Training		Validating	
	R^2	RMSE (m^3/s)	R^2	RMSE (m^3/s)	R^2	RMSE (m^3/s)	R^2	RMSE (m^3/s)	R^2	RMSE (m^3/s)	R^2	RMSE (m^3/s)
MART	0.93	85	0.84	113	0.90	62	0.75	95	0.85	66	0.78	73
GMDH	0.81	141	0.64	169	0.47	144	0.46	149	0.51	120	0.42	118
GEP	0.53	222	0.47	208	0.41	151	0.40	157	0.50	121	0.5	109

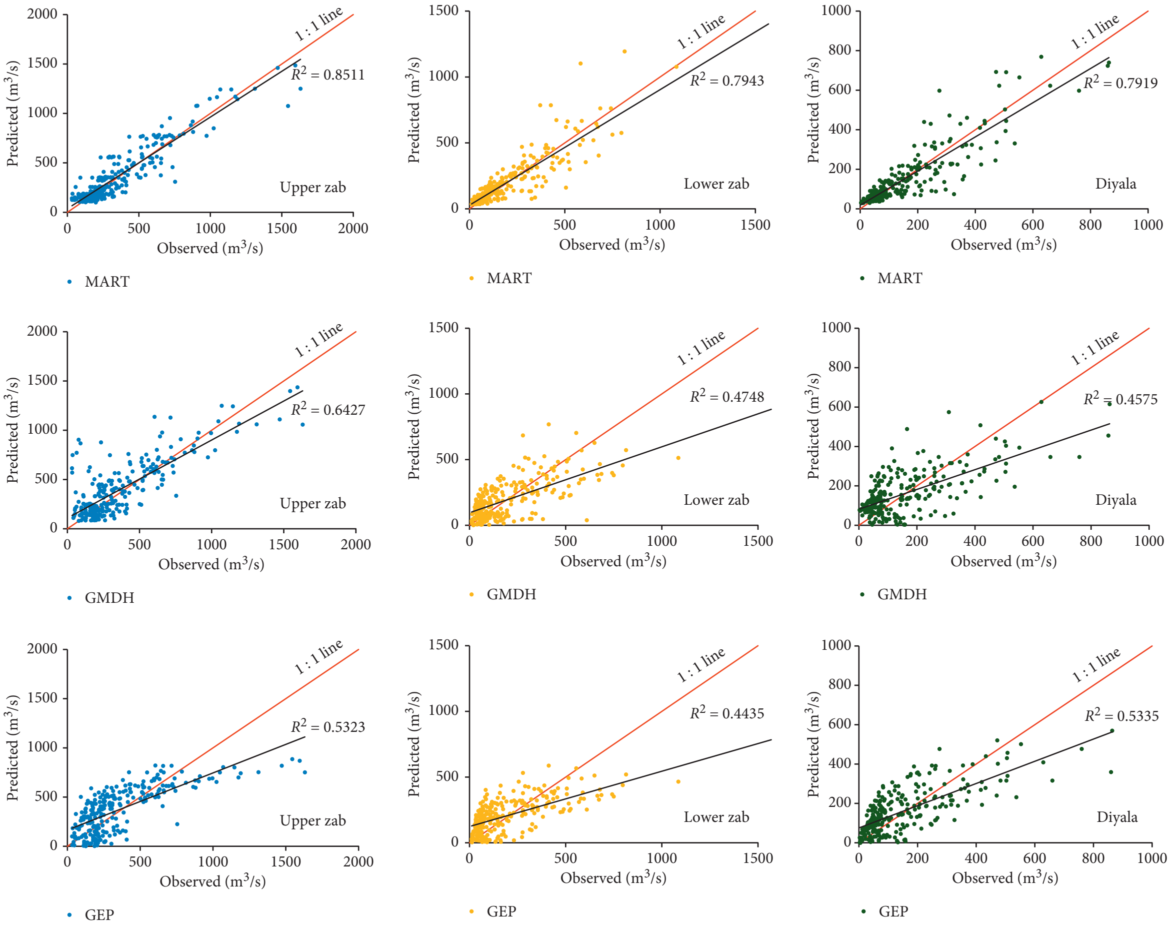
phase the errors produced are 95, 149, and 157 m^3/s for the MART, GMDH, and GEP models, respectively. The R^2 values are 0.85, 0.51, and 0.50 in training phase and 0.78, 0.42, and 0.50 in the validating phase.

The RMSE values in the validating period are 66, 120, and 121 m^3/s in the training phase and 73, 118, and 109 m^3/s for the MART, GMDH, and GEP models, respectively, for the Diyala River. Both the R^2 and RMSE metrics results have proved the accuracy of the MART model to disaggregate annual flows to monthly streamflow in comparison with the GMDH and GEP models. The quality of the proposed models was measured by equating between the three statistical time series parameters which are maximum flow, standard deviation, and mean. Table 4 exhibits the results of these parameters. In accordance with the reported results in Table 4, it is apparent that the performance capacity of the MART model was superior to the performance of GMDH and GEP models. The observed data of the maximum monthly streamflow with the results of the applied models shows that the maximum monthly streamflow values in the validating phase were 1631, 1486, 1435, and 885 m^3/s for the Upper Zab River, 1569, 1215, 769, and 588 m^3/s for the Lower Zab River, and 864, 769, 626, and 570 m^3/s for the Diyala River of the observed, MART, GMDH, and GEP models, respectively. Comparing the results of the statistical parameters, standard deviation, and the mean of the observed monthly streamflow with the results of the applied models as in Table 4 shows improved competence of the MART model compared to the GMDH and GEP models.

The predicted monthly streamflow over the validating period was assessed using the scatter plots variation as illustrated in Figures 7(a)–7(c). The plots in Figure 7 demonstrated a good relationship between the observed value and the generated monthly streamflow using the potential of the MART model in comparing to the other models. Also, the efficiency of the applied models was evaluated by comparing monthly statistical parameters for each month. The models' capability to handle streamflow data decreases with increased stochasticity of the data; however, the results depict that MART is more capable of predicting such data. The results of the maximum monthly flow, mean flow, and standard deviation for each month were compared with the results of observed monthly streamflow in the validating phase. The comparisons were made by plotting the maximum monthly flow, mean flow, and standard deviation against the months; see Figure 8. The results of Figure 8 also demonstrated that MART model is accurate and thus superior with respect to the GMDH and GEP models' performance to generate monthly streamflow from annual streamflow. As per the results of the statistical parameters, it is apparent that the MART model performance is accurate when compared to the observed data for the three studied rivers (Figure 8). It is apparent that rivers have diverging hydrological characteristics and model behavior and performance can change greatly according to that. As per the figure, it can be observed that each river presents different seasonality and deviation over the time period because of which models generate more error during modeling.

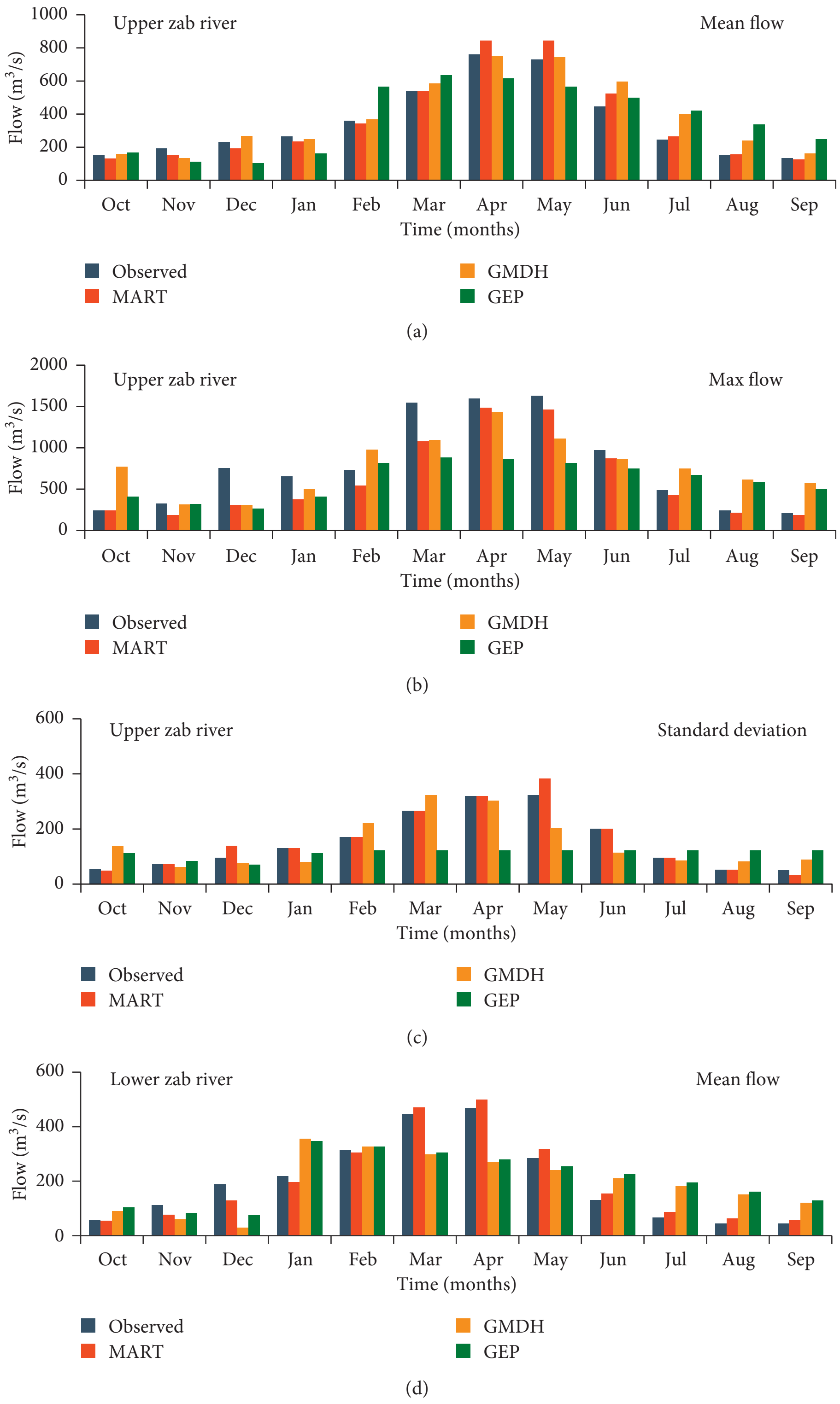
Table 4: Statistical analysis of observed and modeling results of the monthly streamflow values.

River name		Observed	MART	GMDH	GEP
			Training		
Upper Zab	Maximum flow (m^3/s)	1681	1558	1897	993
	Standard deviation (m^3/s)	326	303	293	229
	Mean (m^3/s)	397	393	393	402
Lower Zab	Maximum flow (m^3/s)	1406	1215	827	628
	Standard deviation (m^3/s)	198	179	131	125
	Mean (m^3/s)	198	196	199	203
Diyala	Maximum flow (m^3/s)	1451	850	908	617
	Standard deviation (m^3/s)	171	134	119	113
	Mean (m^3/s)	156	150	154	160
			Validating		
Upper Zab	Maximum flow (m^3/s)	1631	1486	1435	885
	Standard deviation (m^3/s)	286	287	282	228
	Mean (m^3/s)	353	364	388	370
Lower Zab	Maximum flow (m^3/s)	1569	1215	769	588
	Standard deviation (m^3/s)	205	201	149	129
	Mean (m^3/s)	198	201	195	207
Diyala	Maximum flow (m^3/s)	864	769	626	570
	Standard deviation (m^3/s)	150	145	111	116
	Mean (m^3/s)	150	148	157	160



(a) (b) (c)

Figure 7: The scatter plots between the observed and predicted streamflow values for all applied predictive models. (a) Upper Zab River, (b) Lower Zab River, and (c) Diyala River.



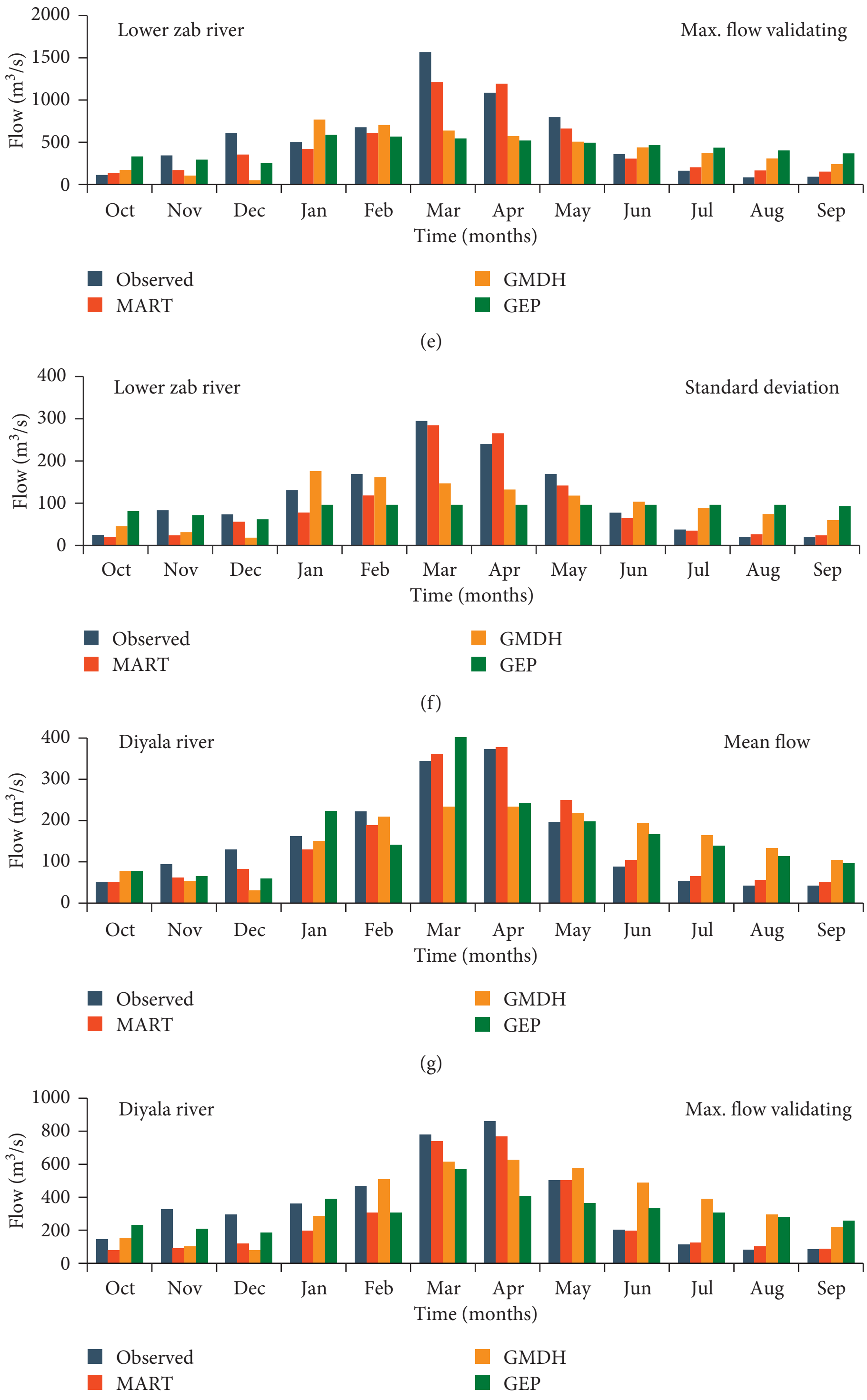
(a)

(b)

(c)

(d)

Figure 8: Continued.



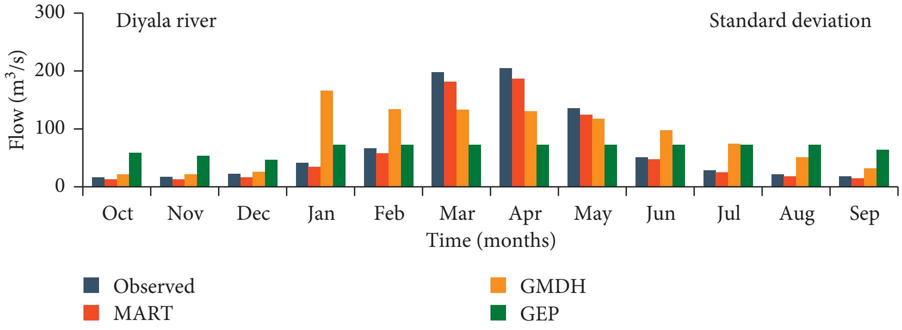
(e)

(f)

(g)

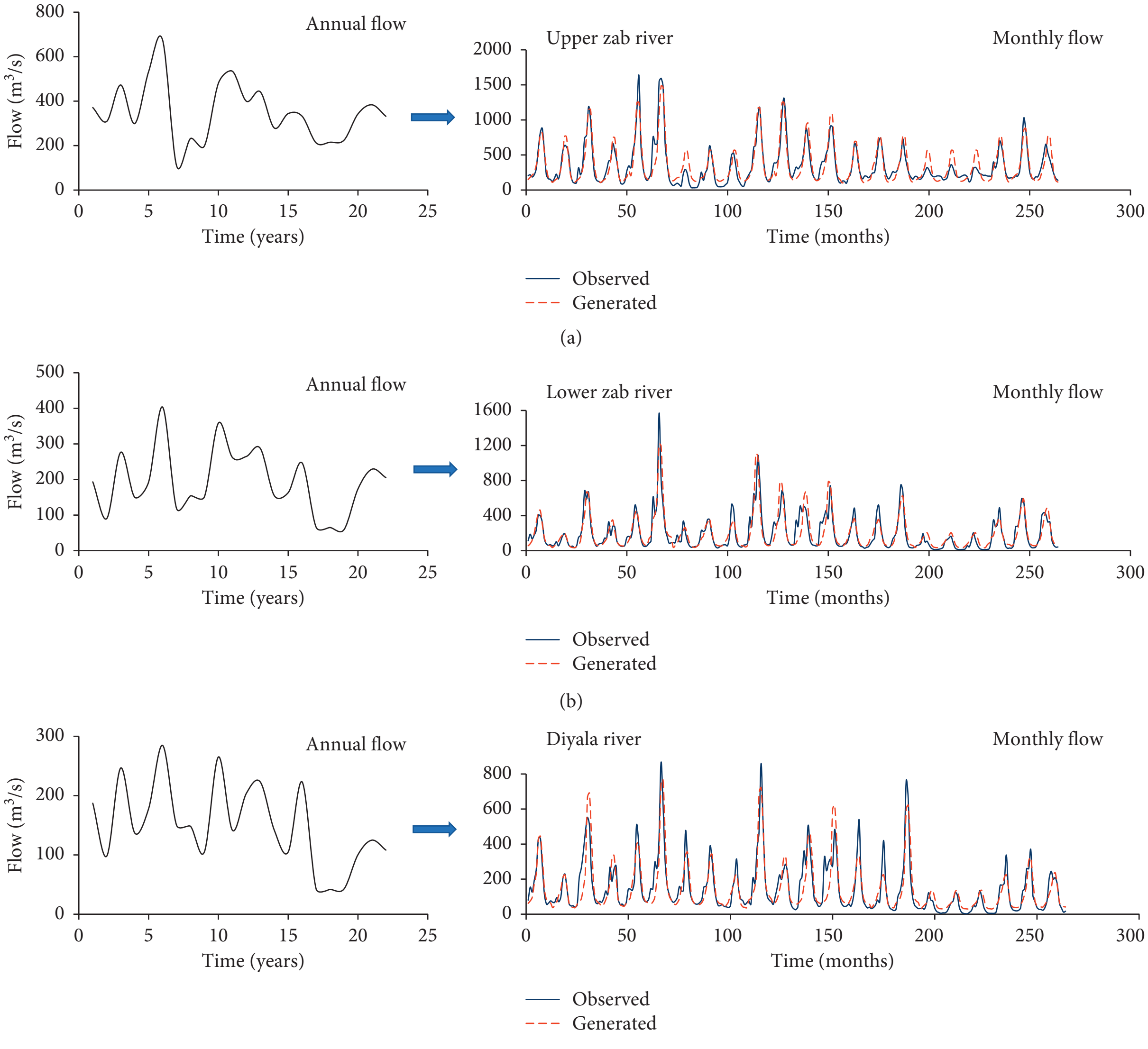
(h)

Figure 8: Continued.



(i)

Figure 8: Results of the statistical parameters against the months.



(a)

(b)

(c)

Figure 9: The graphs show the yearly time series flow of the three rivers and observed and generated monthly streamflow after application of MART model during the validating phase.

Figure 9 shows a comparison between the observed and generated monthly streamflow generated using MART model during the validation phase for the three rivers in this study and it also shows how the monthly flows were generated from the annual flows data. The results in Figure 9 show the proximity of the observed value and generated monthly streamflow which also evidenced that the MART model performance is able to produce monthly streamflow time series from annual monthly streamflow time series data.

The results indicated the efficiency of MART model in generating monthly flow data from annual flow data and this is due to MART model's structure, which enables the building of robust models with a limited number of inputs (the inputs included only the annual flow and time index). The results also showed the weakness of revolutionary and self-learning models in creating robust models with a limited number of inputs.

The results indicated the efficiency of using MART model to generate monthly streamflow from annual streamflow. This is the first application of using MART model to generate the monthly streamflow from annual streamflow. The results showed the importance of using time index to improve the accuracy of generating monthly streamflow from annual streamflow. The results of this paper are encouraging to develop new models for generating monthly streamflow data instead of the data of fragment method which is usually used for this purpose.

4. Conclusions

In this study, three different ML models were used to generate monthly streamflow time series from annual streamflow time series. The models included MART, GMDH, and GEP. The models input only included the annual streamflows and monthly time index. The results showed that the MART model is superior to the GMDH and GEP models in producing monthly streamflow time series by applying annual monthly streamflow time series data. The results indicated that the structure of MART model is better than the structure of polynomial neural networks or revolutionary models in generating modeling. The efficiency of MART model was better than the results of GMDH and GEP models for Upper Zab (R^2 0.84, 0.64, and 0.47), Lower Zab (R^2 0.75, 0.46, and 0.40), and Diyala (R^2 0.78, 0.42, and 0.5). The MART model accuracy is relating to its specific architecture, which may include number of trees growing in equivalence in addition to the use of boosting technique which helped to improve the prediction function. The results demonstrated the possibility of changing the timescale in generating streamflow. The application of MART model is easier than the method of data of fragment that is usually used to disaggregate the annual streamflow to monthly streamflow.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors have no conflicts of interest to any party.

References

[1] M. Zounemat-Kermani, M. Elena, C. Andrea, X. Xia, Q. Liang, and R. Hinkelmann, "Neurocomputing in surface water hydrology and hydraulics: a review of two decades retrospective, current status and future prospects," *Journal of Hydrology*, vol. 588, 2020.

[2] N. Sujay Raghavendra and P. C. Deka, "Support vector machine applications in the field of hydrology: a review," *Applied Soft Computing*, vol. 19, pp. 372–386, 2014.

[3] A. Mosavi, P. Ozturk, and K. W. Chau, "Flood prediction using machine learning models: literature review," *Water*, vol. 10, no. 11, p. 1536, 2018.

[4] H. Lange and S. Sippel, "Machine learning applications in hydrology," in *Forest-Water Interactions*, pp. 233–257, Springer, Berlin, Germany, 2020.

[5] L. Diop, A. Bodian, K. Djaman et al., "The influence of climatic inputs on stream-flow pattern forecasting: case study of Upper Senegal river," *Environmental Earth Sciences*, vol. 77, no. 5, p. 182, 2018.

[6] P. A. Kagoda, J. Ndiritu, C. Ntuli, and B. Mwaka, "Application of radial basis function neural networks to short-term streamflow forecasting," *Physics and Chemistry of Earth*, vol. 35, no. 13-14, pp. 571–581, 2010.

[7] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Times Series Analysis—Forecasting and Control*, Holden-Day Inc., San Francisco, CA, USA, 1976.

[8] J. D. Salas and J. T. B. Obeysekera, "ARMA model identification of hydrologic time series," *Water Resources Research*, vol. 18, no. 4, pp. 1011–1021, 1982.

[9] P. P. Balestrassi, E. Popova, A. P. Paiva, and J. W. Marangon Lima, "Design of experiments on neural network's training for nonlinear time series forecasting," *Neurocomputing*, vol. 72, no. 4–6, pp. 1160–1178, 2009.

[10] Q. Ju, Z. Yu, Z. Hao, G. Ou, J. Zhao, and D. Liu, "Division-based rainfall-runoff simulations with BP neural networks and Xinanjiang model," *Neurocomputing*, vol. 72, no. 13-15, pp. 2873–2883, 2009.

[11] S. Asadi, J. Shahrabi, P. Abbaszadeh, and S. Tabanmehr, "A new hybrid artificial neural networks for rainfall-runoff process modelling," *Neurocomputing*, vol. 121, pp. 470–480, 2013.

[12] H.-G. Han and J.-F. Qiao, "A structure optimisation algorithm for feedforward neural network construction," *Neurocomputing*, vol. 99, pp. 347–357, 2013.

[13] M. Sit, B. Z. Demiray, Z. Xiang, G. J. Ewing, Y. Sermet, and I. Demir, "A comprehensive review of deep learning applications in hydrology and water resources," 2020, https://arxiv.org/abs/2007.12269.

[14] P. A. Whigham and P. F. Crapper, "Modelling rainfall-runoff using genetic programming," *Mathematical and Computer Modelling*, vol. 33, no. 6-7, pp. 707–721, 2001.

[15] J. Guo, J. Zhou, H. Qin, Q. Zou, and Q. Li, "Monthly streamflow forecasting based on improved support vector machine model," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13073–13081, 2011.

[16] D. Labat, "Recent advances in wavelet analyses: part 1. a review of concepts," *Journal of Hydrology*, vol. 314, no. 1–4, pp. 275–288, 2005.

[17] B. Keshtegar, M. F. Allawi, H. A. Afan, and A. El-Shafie, "Optimized river stream-flow forecasting model utilizing high-order response surface method," *Water Resources Management*, vol. 30, no. 11, pp. 3899–3914, 2016.

[18] S. Mehdizadeh, F. Fathian, and J. F. Adamowski, "Hybrid artificial intelligence-time series models for monthly streamflow modeling," *Applied Soft Computing*, vol. 80, pp. 873–887, 2019.

[19] S. Mehdizadeh, F. Fathian, M. J. S. Safari, and J. F. Adamowski, "Comparative assessment of time series and artificial intelligence models to estimate monthly streamflow: a local and external data analysis approach," *Journal of Hydrology*, vol. 579, Article ID 124225, 2019.

[20] E. S. Salami and M. Ehteshami, "Simulation, evaluation and prediction modeling of river water quality properties (case study: Ireland rivers)," *International Journal of Environmental Science and Technology*, vol. 12, no. 10, pp. 3235–3242, 2015.

[21] I. Ahmadianfar, M. Jamei, and X. Chu, "A novel hybrid wavelet-locally weighted linear regression (W-LWLR) model for electrical conductivity (EC) prediction in surface water," *Journal of Contaminant Hydrology*, vol. 232, Article ID 103641, 2020.

[22] A. P. Jacquin and A. Y. Shamseldin, "Review of the application of fuzzy inference systems in river flow forecasting," *Journal of Hydroinformatics*, vol. 11, no. 3, 2009.

[23] M. Fahmi, M. F. B. Mohd Nasir, M. S. Samsudin et al., "River water quality modeling using combined principle component analysis (PCA) and multiple linear regressions (MLR): a case study at klang river," *World Applied Sciences Journal*, vol. 14, no. 2002, pp. 73–82, 2011.

[24] A. J. Jakeman, I. G. Littlewood, and P. G. Whitehead, "An assessment of the dynamic response characteristics of streamflow in the Balquhidder catchments," *Journal of Hydrology*, vol. 145, no. 3-4, pp. 337–355, 1993.

[25] M. D. Dettinger and H. F. Diaz, "Global characteristics of stream flow seasonality and variability," *Journal of Hydrometeorology*, vol. 1, no. 4, pp. 289–310, 2000.

[26] R. DarioValencia and J. C. Schakke, "Disaggregation processes in stochastic hydrology," *Water Resources Research*, vol. 9, no. 3, pp. 580–585, 1973.

[27] D. N. Kumar, U. Lall, and M. R. Petersen, "Multisite disaggregation of monthly to daily streamflow," *Water Resources Research*, vol. 36, pp. 1823–1833, 2000.

[28] J. R. Stedinger and R. M. Vogel, "Disaggregation procedures for generating serially correlated flow vectors," *Water Resources Research*, vol. 20, no. 1, 1984.

[29] J. Prairie, B. Rajagopalan, U. Lall, and T. Fulp, "A stochastic nonparametric technique for space-time disaggregation of streamflows," *Water Resources Research*, vol. 43, no. 3, 2007.

[30] T. Lee, J. D. Salas, and J. Prairie, "An enhanced nonparametric streamflow disaggregation model with genetic algorithm," *Water Resources Research*, vol. 46, no. 8, 2010.

[31] A. Acharya and J. H. Ryu, "Simple method for streamflow disaggregation," *Journal of Hydrologic Engineering*, vol. 19, no. 3, pp. 509–519, 2014.

[32] A. T. Silva and M. M. Portela, "Disaggregation modelling of monthly streamflows using a new approach of the method of fragments," *Hydrological Sciences Journal*, vol. 57, no. 5, pp. 942–955, 2012.

[33] D. A. Hughes and A. Slaughter, "Daily disaggregation of simulated monthly flows using different rainfall datasets in southern Africa," *Journal of Hydrology: Regional Studies*, vol. 4, pp. 153–171, 2015.

[34] M. M. Portela and A. T. Silva, "Disaggregation modelling of annual flows into daily streamflows using a new approach of the method of fragments," *Water Resources Management*, vol. 30, no. 15, 2016.

[35] S. H. D. Al-Zakar, N. Şarlak, and O. M. A. M. Agha, "Disaggregation of annual to monthly streamflow: a case study of Kizilirmak Basin (Turkey)," *Advances in Meteorology*, vol. 2017, Article ID 3582826, 16 pages, 2017.

[36] T. A. Awchi, "River sischarges forecasting in Northern Iraq using different ANN techniques," *Water Resources Management*, vol. 28, no. 3, pp. 801–814, 2014.

[37] C. Ferreira, "Gene expression programming: a new adaptive algorithm for solving problems," *Complex System*, vol. 13, no. 2, pp. 87–129, 2001.

[38] C. Ferreira, *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*, Springer, Berlin, Germany, 2006.

[39] T. Hong, K. Jeong, and C. Koo, "An optimized gene expression programming model for forecasting the national CO_2 emissions in 2030 using the metaheuristic algorithms," *Applied Energy*, vol. 228, pp. 808–820, 2018.

[40] C. Ferreira, "Gene expression programming in problem solving," in *Soft Computing and Industry*Springer, Berlin, Germany, 2011.

[41] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[42] R. Derrig and L. Francis, "Distinguishing the forest from the TREES: a comparison of tree based data mining methods," *Casualty Actuarial Society Forum*, 2006.

[43] A. G. Ivakhnenko, "The group method of data of handling; a rival of the method of stochastic approximation," *Soviet Automatic Control*, vol. 13, pp. 43–55, 1968.

[44] A. Sepahvand, B. Singh, P. Sihag, and A. N. Samani, "Assessment of the various soft computing techniques to predict sodium absorption ratio (SAR)," *Journal of Hydraulic Engineering*, vol. 5010, 2019.

[45] A. Ivakhnenko and G. Ivakhnenko, "The review of problems solvable by algorithms of the group method of data handling (GMDH)," *Pattern Recognition and Image Analysis*, vol. 5, no. 4, pp. 527–535, 1995.

[46] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.

[47] N. J. D. Nagelkerke, "A note on a general definition of the coefficient of determination," *Biometrika*, vol. 78, no. 3, pp. 691-692, 1991.