sustainability

MDPI

Article

An Integrated Statistical-Machine Learning Approach for Runoff Prediction

Abhinav Kumar Singh [1], Pankaj Kumar [1], Rawshan Ali [2], Nadhir Al-Ansari [3,*], Dinesh Kumar Vishwakarma [4,*], Kuldeep Singh Kushwaha [5], Kanhu Charan Panda [6], Atish Sagar [7], Ehsan Mirzania [8], Ahmed Elbeltagi [9], Alban Kuriqi [10,11] and Salim Heddam [12]

[1] Department of Soil and Water Conservation Engineering, G. B. Pant University of Agriculture and Technology, Pantnagar 263145, India; avi4913@gmail.com (A.K.S.); pankaj591@gmail.com (P.K.)

[2] Department of Petroleum, Koya Technical Institute, Erbil Polytechnic University, Erbil 44001, Iraq; rawshan.ali@epu.edu.iq

[3] Department of Civil, Environmental and Natural Resources Engineering, Lulea University of Technology, 97187 Lulea, Sweden

[4] Department of Irrigation and Drainage Engineering, G. B. Pant University of Agriculture and Technology, Pantnagar 263145, India

[5] Centre for Water Engineering and Management, Central University of Jharkhand, Ranchi 835205, India; kuldeepkushwaha@hotmail.com

[6] Department of Agricultural Engineering, Institute of Agricultural Sciences, Banaras Hindu University, Varanasi 221005, India; kanhucharan.bm@gmail.com

[7] Division of Agricultural Engineering, ICAR—Indian Agriculture Research Institute, New Delhi 110012, India; atishmicky.sagar@gmail.com

[8] Department of Water Engineering, Faculty of Agriculture, University of Tabriz, Tabriz 5166616471, Iran; e.mirzania99@ms.tabrizu.ac.ir

[9] Agricultural Engineering Department, Faculty of Agriculture, Mansoura University, Mansoura 35516, Egypt; ahmedelbeltagy81@mans.edu.eg

[10] CERIS, Instituto Superior Técnico, University of Lisbon, 1649-004 Lisbon, Portugal; alban.kuriqi@tecnico.ulisboa.pt

[11] Civil Engineering Department, University for Business and Technology, 10000 Pristina, Kosovo

[12] Laboratory of Research in Biodiversity 17 Interaction Ecosystem and Biotechnology, Agronomy Department, Hydraulics Division, Faculty of Science, University 20 Août 1955, Route El Hadaik, Skikda 21000, Algeria; heddamsalim@yahoo.fr

* Correspondence: nadhir.alansari@ltu.se (N.A.-A.); dinesh.vishwakarma4820@gmail.com (D.K.V.); Tel.: +91-967-079-1406 (D.K.V.)

Citation: Singh, A.K.; Kumar, P.; Ali, R.; Al-Ansari, N.; Vishwakarma, D.K.; Kushwaha, K.S.; Panda, K.C.; Sagar, A.; Mirzania, E.; Elbeltagi, A.; et al. An Integrated Statistical-Machine Learning Approach for Runoff Prediction. *Sustainability* **2022**, *14*, 8209. https://doi.org/10.3390/su14138209

Academic Editor: Zhenliang Yin

Received: 2 June 2022
Accepted: 30 June 2022
Published: 5 July 2022

Abstract: Nowadays, great attention has been attributed to the study of runoff and its fluctuation over space and time. There is a crucial need for a good soil and water management system to overcome the challenges of water scarcity and other natural adverse events like floods and landslides, among others. Rainfall–runoff (R-R) modeling is an appropriate approach for runoff prediction, making it possible to take preventive measures to avoid damage caused by natural hazards such as floods. In the present study, several data-driven models, namely, multiple linear regression (MLR), multiple adaptive regression splines (MARS), support vector machine (SVM), and random forest (RF), were used for rainfall–runoff prediction of the Gola watershed, located in the south-eastern part of the Uttarakhand. The rainfall–runoff model analysis was conducted using daily rainfall and runoff data for 12 years (2009 to 2020) of the Gola watershed. The first 80% of the complete data was used to train the model, and the remaining 20% was used for the testing period. The performance of the models was evaluated based on the coefficient of determination (R^2), root mean square error (RMSE), Nash–Sutcliffe efficiency (NSE), and percent bias (PBAIS) indices. In addition to the numerical comparison, the models were evaluated. Their performances were evaluated based on graphical plotting, i.e., time-series line diagram, scatter plot, violin plot, relative error plot, and Taylor diagram (TD). The comparison results revealed that the four heuristic methods gave higher accuracy than the MLR model. Among the machine learning models, the RF (RMSE (m^3/s), R^2, NSE, and PBIAS (%) = 6.31, 0.96, 0.94, and −0.20 during the training period, respectively, and 5.53, 0.95, 0.92, and −0.20 during the testing period, respectively) surpassed the MARS, SVM, and the MLR models in forecasting daily runoff for all cases studied. The RF model outperformed in all four models' training and testing periods. It can be summarized that the RF model is best-in-class and delivers a strong potential for the runoff prediction of the Gola watershed.

Keywords: MARS; SVM; RF; rainfall; runoff; rainfall–runoff modeling

1. Introduction

Forecasting heavy precipitation is an important function in estimating the runoff and flooding in the short to medium term [1–4], flood warning [5], real-time flood forecasting [6], and flood mitigation [7,8]. Nonetheless, rainfall directly affects runoff generation in streams, rivers, and even floods, making it one of the most specific hydrological phenomena [2]. The socioeconomic impacts of rainfall are significant, from physical damage in floods to disruptions in transport networks [3,9]. Simultaneously, India is challenged with increasing population and climate change, which have threatened the present freshwater need for irrigation and drinking [10–13]. To overcome the challenges of water scarcity and the deterioration of cultivable land, modeling rainfall–runoff plays an important role. Many aspects of our daily lives depend on the rain we receive [14,15]. Rainfall remains one of the most influential meteorological variables [16]. The rainfall–runoff modeling in water resource management attracts many researchers and practitioners worldwide. Planning and managing water properly is the only way to prevent water stress and to balance supply and demand [17–19]. In addition to natural disasters such as floods caused by runoff from precipitation and river flow and droughts caused by short rainfall for a long duration, we can also determine the occurrence of these natural disasters by assessing the rainfall–runoff relationship [20].

The major role of several nonlinear and nonstationary variables in converting rainfall into runoff is difficult to comprehend [21]. The response to the catchment precipitation becomes more complex due to the spatiotemporal variability in rainfall intensity and uniformity [22]. However, the direct contribution of rainfall in runoff generation and runoff in streams, rivers, and even floods are one of the most focused upon hydrological phenomena. To understand the accurate relationship between the two hydrological variables, the concept of rainfall–runoff (*R-R*) plays a critical role in the area of hydrological science [23]. However, despite the remaining inconsistency in the rainfall–runoff relation, the application of machine learning is promising. These computational techniques either reduce the requirement of modeling parameters, improve modeling accuracy, or are even applicable for both purposes [24]. The main aim of this modeling is to improve our understanding of the major hydrological phenomenon, which influences all watershed systems. It also helps to develop a simulation tool to help decision makers optimize and plan the operational rules of the water resource system [25].

In rainfall–runoff modeling [2,26] and rainfall forecasting [27,28], the use of artificial intelligence (AI) and machine learning (ML) established modeling in water resources in a new direction. Several studies attempt the application of AI and ML, whether for *R-R* or for rainfall forecasting [4,9,29,30], streamflow [31–36], suspended sediment-load prediction [36–42], flood forecasting [5,6,43], stage–discharge modeling [44–48], soil temperature estimation [49–56], pan evaporation [57–68], reference evapotranspiration [69–78], soil parameter estimation such as infiltration, permeability, and saturated hydraulic conductivity [79–88], groundwater quality index [89–92], drought and stress tolerance in maize crops [93], water footprint [94,95], rice yield estimation [96], and crop coefficients [97]. Artificial neural networks (ANN) gained immense popularity in rainfall–runoff modeling [22,23,28,98] as well as rainfall forecasting [99–101], although there is no requirement for deep knowledge of hydrological processes in AI-based rainfall–runoff modeling [102]. The MLR linear mode is the most common statistical tool to predict the output–input variables. It develops a linear relationship between multiple variables [12,103,104]. A quantitative relationship exists between the dependent and independent variables in MLR [105]. The values of the independent variables in MLR are affiliated with the values of the dependent variables [106]. The dependent, independent, and intercept variables are local behavior calculated by the least square rule or other regression rules [45].

For *R-R* modeling, AI and ML have been extensively used for decades [28]. These models have been compared to traditional statistical methods and conceptual models. The nonlinear MARS is a nonlinear and nonparametric regression [107,108]. MARS built several MLR models in the dataset range [109]. It is done by creating knots based on the splitting strategy and running a suite of the linear model for each subset; the nonlinear responses between the input and output of a dataset are divided into piecewise linear segments (splines) of different gradients [110]. The SVM is a generalized nonlinear model for both classification and regression analysis, and it was introduced by Vapnik [111]. The basic concept of SVM is to minimize the structural risk. The algorithm converts the patterns that are not linearly separable to higher-dimensional feature space using kernel functions. It attempts to reduce the upper limit of the generalization error. For its advantages over other general algorithm such as ANN, it is a better method in the hydrological field for simulation and forecasting hydrological events. The RF is a supervised ML algorithm based on bagging or bootstrap aggregation, a part of ensemble learning [112].

Al-Sudani et al. [113] hybridized the MARS model using the differential evolution algorithm (DE). They compared its performances, i.e., MARS-DE, with those of the single MARS and the least square support vector machine model (LSSVM). They reported the superiority of the hybrid MARS-DE. Adnan et al. [114] compared the performance of four ML models, i.e., the optimal pruning extreme learning machine (OPELM), MARS, M5Tree, and MARS-K_{means}. It was found that MARS-K_{means} surpassed all other models for multi-step forecasting, i.e., one, six, and twelve hours in advance. In another study, Li et al. [115] evaluated the performances of extreme learning machines (ELM), RF, and SVM for forecasting daily, low, and peak streamflow. They reported the superiority of the ELM model.

Therefore, the present paper aims to compare the MLR, SVM, MARS, and RF models for forecasting daily runoff at the Gola watershed, located in the south-eastern part of Uttarakhand. The study was conducted with the major objectives of selecting the most relevant input variables for R-R forecasting and comparing the models' performance across the studied stations. The rest of the paper is organized as follows: Section 2 presents brief information about the study cases and data collation, Sections 3 and 4 describe the methods, and Section 5 presents the main findings and discusses their relevance in light of the literature. Finally, the main conclusions drawn from this study are given in Section 6.

2. Materials and Methods

2.1. General Description of Study Area

The Gola watershed in the south-eastern part of Uttarakhand state is shown in Figure 1. The Gola River originates in the Bhirapani valley near the village of Paharpani, Uttarakhand state, in the lesser Himalayas. The river's major tributaries are Kanchi, Kharkai, and Karkari. The watershed lies between 29°16′18″ to 29°27′33″ N latitudes and 79°46′5″ to 79°32′51″ E longitudes in northern India. The total catchment area of the Gola watershed is about 611 km^2. The climatic condition of the Gola watershed is mild and generally warm. The minimum and maximum elevations of the watershed are 252 m and 2302 m, respectively, above mean sea level. The Gola watershed comes under a subtropical climate with predominant seasonal rainfall. The average annual rainfall is 1699 mm, heavily influenced by monsoon rainfall. As the watershed lies on the eastern edge of the Himalayan ranges, it is subjected to heavy rainfall. It is mainly a spring-fed river; this river is a source of water for Haldwani and Kathgodam. The monsoon season extends from July to September and produces 90% of the annual rainfall. The watershed receives heavy rainfall in the months of July and August. Due to this, the mainstream of rainfed rivers like the Gola River subsequently has high discharge in these months of the year. The barrage is a landmark for the residents, and provides irrigation water for the bhabar fields. For this reason, it is very important to know the daily forecasting of the river flow to avoid any risk/distress/fatality.

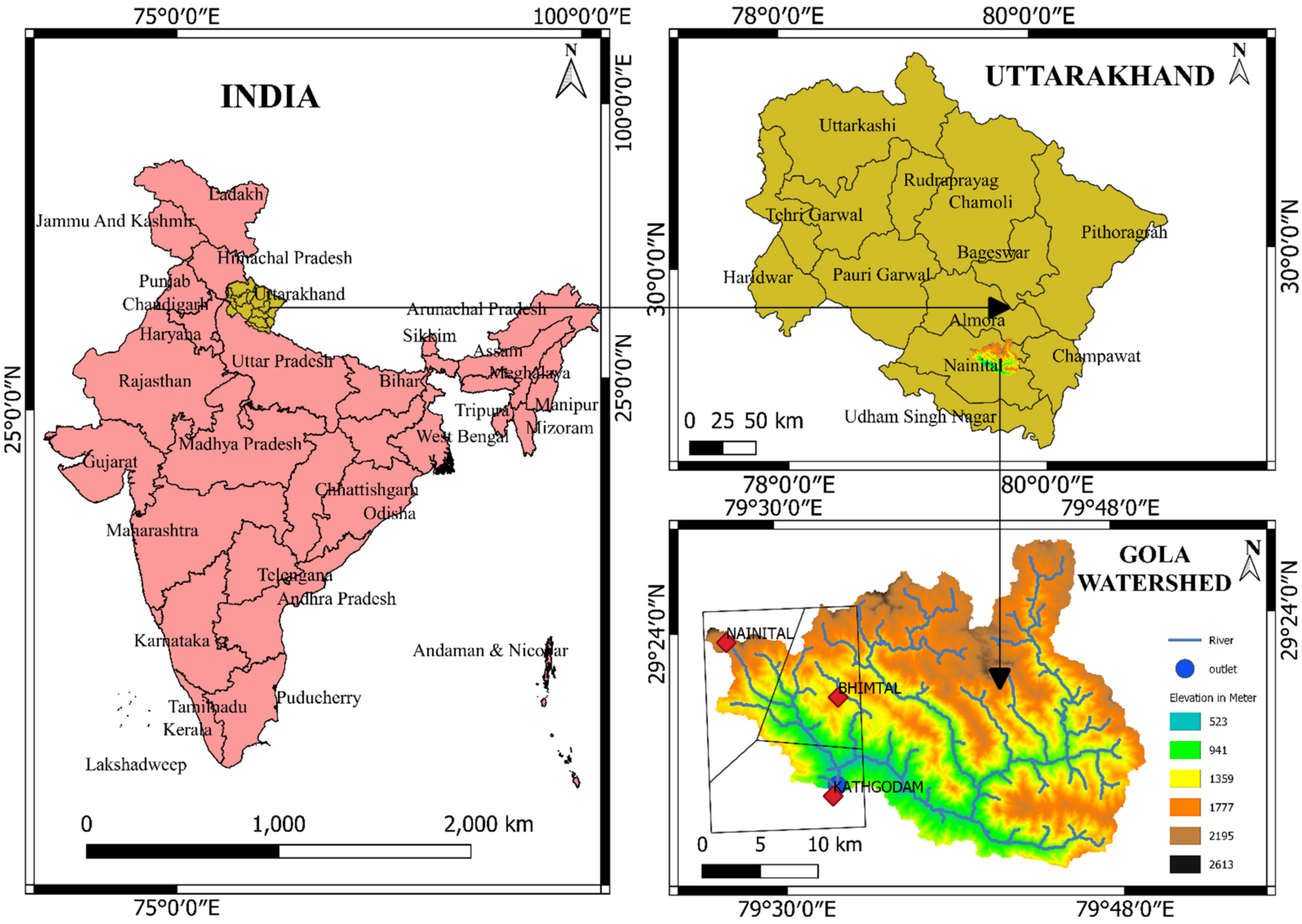
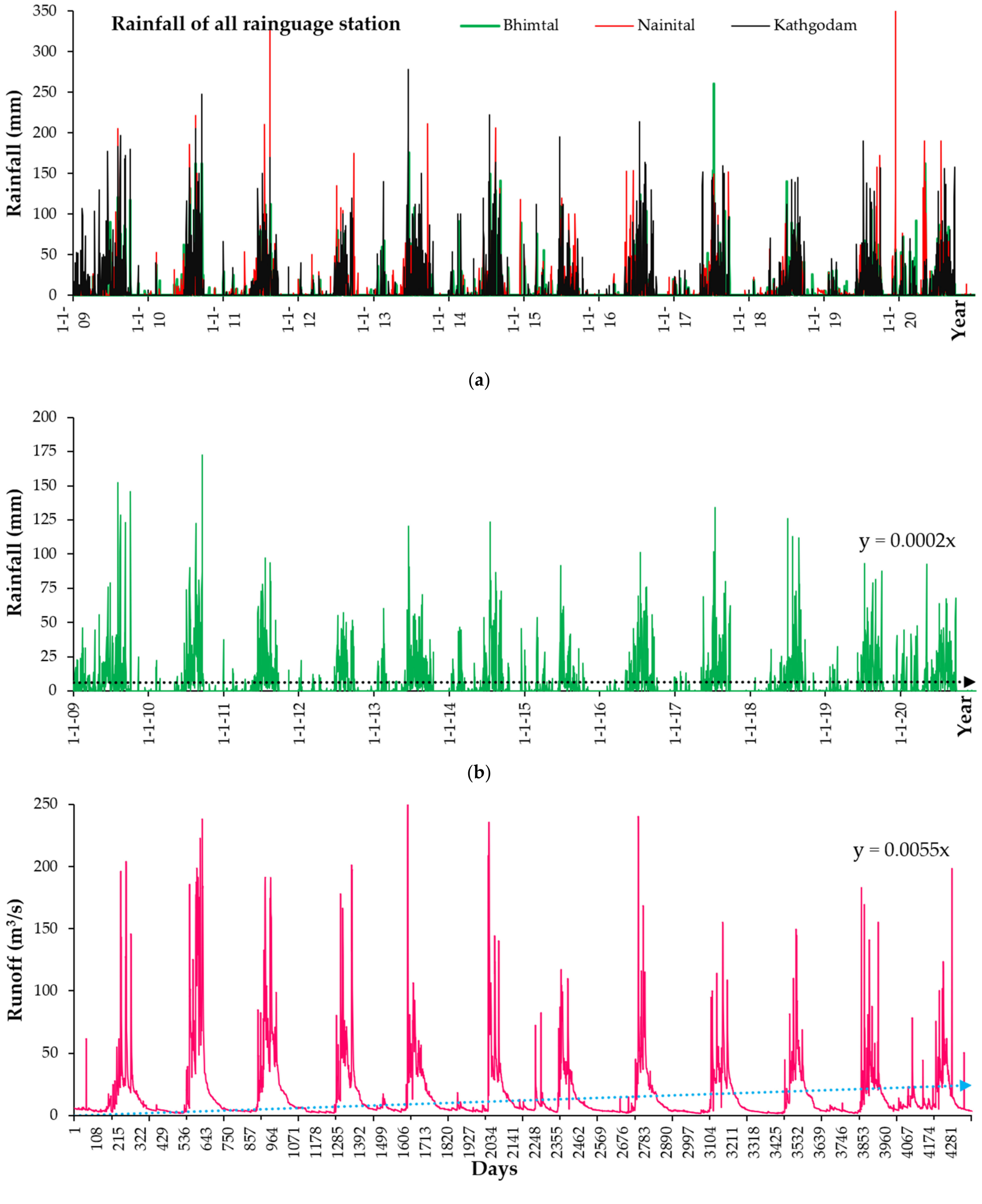


Figure 1. Location map of the study area.

2.2. Data Acquisition and Input Data Preparation

The daily data of rainfall and runoff for 12 years (2009 to 2020) of the study location (Gola watershed) were used to analyze rainfall–runoff modeling. The runoff data of the Gola River were taken from an observation station at Kathgodam barrage. The rainfall data of three rain-gauge stations—Nainital, Bhimtal, and Kathgodam—were taken from the respective irrigation departments (Figure 2a). The Thiessen polygon method was used to calculate the mean areal rainfall of the Gola watershed. Plots of the rainfall and runoff time-series data are shown in Figure 2b,c, respectively. The daily rainfall and runoff data for 12 years were used to develop and validate the models.

Statistical parameters were used to analyze the time-series dataset for rainfall–runoff modeling of the Gola watershed and are presented in Table 1. The complete dataset was divided into training and testing datasets. The first 80% of the complete data was used in training, and the remaining 20% was used for the testing period. During the division of the datasets into training and testing subsets, cross-validation of the dataset is necessary to obtain the same statistical population. The skewness value of the dataset showed that the distribution was highly skewed. Figure 3 shows the flowchart of the proposed methodology.



(**a**)

(**b**)

(**c**)

Figure 2. (**a**) Rainfall data from three raingauge stations, namely, Bhimtal, Nainital and Kathgodam; (**b**) mean areal rainfall time-series data for Gola watershed and (**c**) runoff time-series data for Gola watershed.

Table 1. Basic statistics of training, testing, and total rainfall and runoff datasets at study stations.

Statistical Parameters	Mean	Median	Minimum	Maximum	Standard Deviation	CV (%)	Skewness
			Total dataset				
Rainfall (mm)	6.45	0	0	172.38	15.60	24.18	3.87
Runoff (m^3/s)	17.22	6.06	1.88	250.03	27.51	15.97	3.80
			Training data				
Rainfall (mm)	6.45	0	0	172.38	15.87	24.60	4.02
Runoff (m^3/s)	17.35	5.66	1.38	250.03	28.69	16.54	3.76
			Testing data				
Rainfall (mm)	6.89	0	0	111.69	14.47	21.00	3.07
Runoff (m^3/s)	16.83	7.5	1.61	197.08	22.16	13.16	3.59

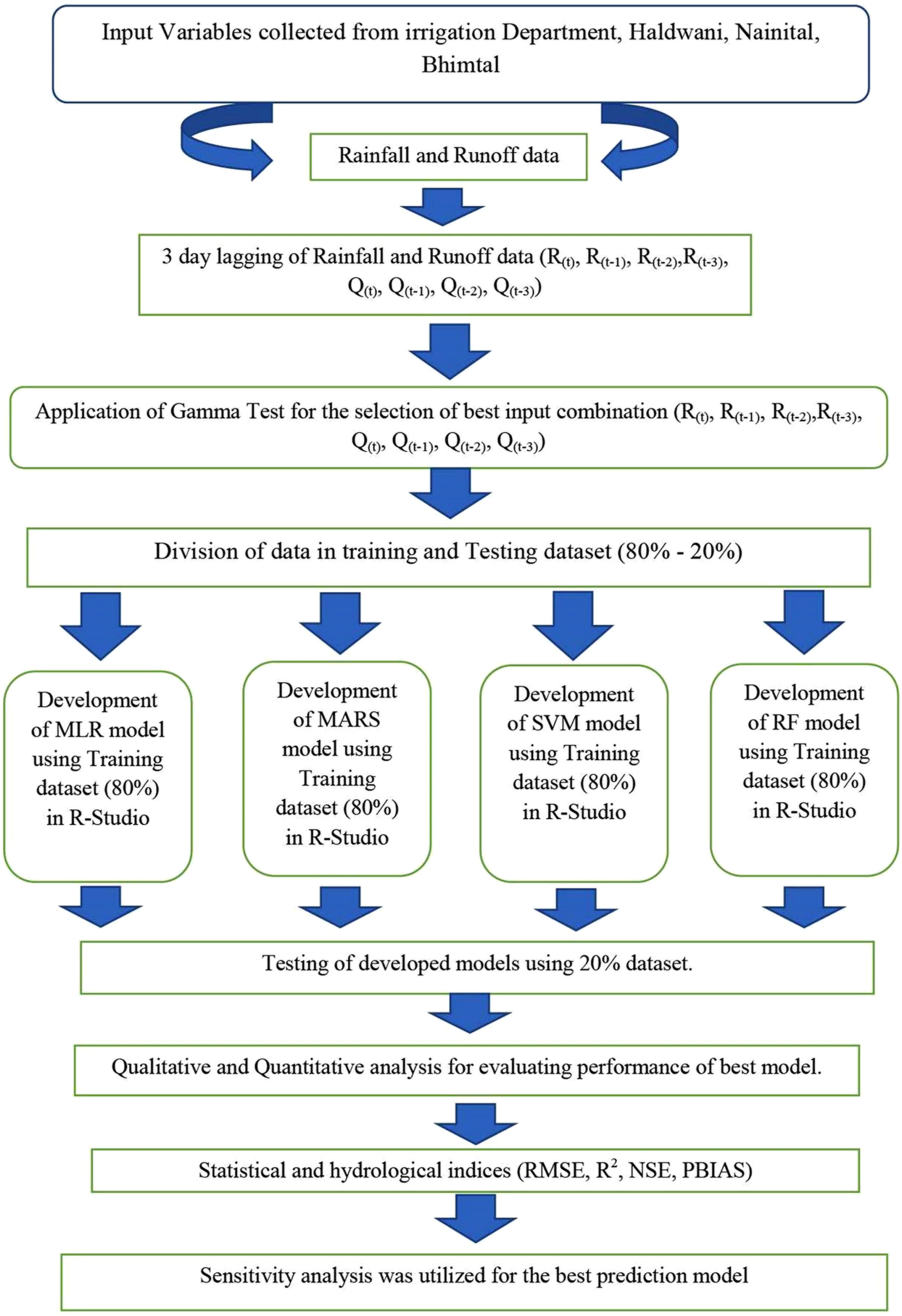

Input Variables collected from irrigation Department, Haldwani, Nainital, Bhimtal
Rainfall and Runoff data
3 day lagging of Rainfall and Runoff data ($R_{(t)}$, $R_{(t-1)}$, $R_{(t-2)}$,$R_{(t-3)}$, $Q_{(t)}$, $Q_{(t-1)}$, $Q_{(t-2)}$, $Q_{(t-3)}$)
Application of Gamma Test for the selection of best input combination ($R_{(t)}$, $R_{(t-1)}$, $R_{(t-2)}$,$R_{(t-3)}$, $Q_{(t)}$, $Q_{(t-1)}$, $Q_{(t-2)}$, $Q_{(t-3)}$)
Division of data in training and Testing dataset (80% - 20%)
Development of MLR model using Training dataset (80%) in R-Studio
Development of MARS model using Training dataset (80%) in R-Studio
Development of SVM model using Training dataset (80%) in R-Studio
Development of RF model using Training dataset (80%) in R-Studio
Testing of developed models using 20% dataset.
Qualitative and Quantitative analysis for evaluating performance of best model.
Statistical and hydrological indices (RMSE, R^2, NSE, PBIAS)
Sensitivity analysis was utilized for the best prediction model

Figure 3. Flowchart of methodology.

2.3. Gamma Test

The Gamma test selects the best input variables in modeling a dataset [47,116–119]. It is a flexible and unbiased tool for evaluating the potential of each input parameter. Traditionally, trial and error methods were used to select the best input variable, making it a time-consuming and tedious job that includes training and testing every possible input combination to select the best suitable input vector. It also fails to provide information about the number of data points necessary for calibration to achieve the accuracy of the optimum model. The Gamma test plays a significant role by guiding the selection of various input parameters to develop reliable and smooth models. Nonlinear correlation between random variables is evaluated by Gamma tests, like the nonlinear correlation between the input and output pairs. The idea of the Gamma test was first discovered by Stefansson et al. [120] for simulation; it was then another researcher who further adapted it for research activity [38,47,121–123]. It was used for estimating the minimum standard error for nonlinear models for any input variables [121].

The mathematical Gamma test is represented as:

$$\{X_1(i), \ldots \ldots \ldots X_n(i), Y_i\} = \{(X_i, Y_i)|1 \le i \le N\} \tag{1}$$

where X is the $X_1, \ldots, X_n$ corresponds to the predictor's variables, i.e., m variables for a total of N data points, scalar Y_i is the output variable, and Gamma (Γ) is calculated by building up a linear regression between input (X) and output (Y), as:

$$Y = f(X_i, \ldots \ldots \ldots X_n) + \Gamma \tag{2}$$

where f is a smooth function and Γ corresponds to the noise. The overall model complexity is evaluated according to Equation (2). More suitable input variables were reflected by a low value of Γ, i.e., close to zero. In addition, a complicated model is obtained based on the obtained gradient value, and based on the standard error (SE) of Γ, a more reliable Gamma value is obtained. In addition, $V_{ratio,}$ given by Equation (3), indicates the predictability of the output variables. A model's complexity can be determined from the output Y of Equation (2). A value of Γ close to zero indicates a suitable input variable. We have a complicated model when the gradient is high; we have a simple model if the gradient is low. The Gamma value is more reliable if the standard error (SE) of Γ is smaller. $V_{ratio,}$ given in Equation (3), measures the predictability of a variable.

$$V_{ratio} = \frac{\Gamma}{\sigma^2(Y)} \tag{3}$$

where $\sigma^2(y)$ is the output variance of Y, and Γ is the Gamma function. When Vratio is near 0, the predictability is higher. We can build a more qualitative mathematical model with smaller values for Gamma (Γ), gradient, SE, and V_{ratio}.

3. Software Application Used in the Study

The machine learning techniques, MLR, MARS, SVM, and RF, were used in the rainfall–runoff modeling of the Gola watershed. The description of these models is as follows.

3.1. Multiple Linear Regression

Multiple linear regression is the simplest statistical technique to predict the output from several input variables. The linear relationship is developed between multiple variables. In this regression, a quantitative relationship is made by independent and dependent variables. The values of independent variables are related to dependent variables [106].

MLR predicts runoff from input rainfall by taking the dataset into training and testing data periods. The expression for MLR is defined as follows:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 \ldots \ldots \ldots + \alpha_n X_n \tag{4}$$

where Y = the output or the modeled variable; X_1, X_2 . . . . X_n = the inputs variables; α_0 = intercept, and α_1, α_2 . . . . . . α_n = regression coefficients.

3.2. Multivariate Adaptive Regression Splines (MARS)

Multivariate adaptive regression splines (MARS) is a nonlinear and nonparametric regression method that built several MLR models across the range of predictor values. It is done by splitting the data and running a linear model on each different partition. The nonlinear responses between the input and output of a dataset are divided into piecewise linear segments (splines) of different gradients [124]. The extensible regression models proposed by MARS have the solution space for each model divided into intervals, and splines fit every interval space [125]. There is the creation of a bias function and finding a potential knot location to improve the model's performance and over-fit. The backstage of the MARS model is done by pruning the ineffective term [126]. The comparison of different model subsets is done by the less expansive technique of generalized cross-validation [127]. It is expressed as follows:

$$\mathrm{GCV} = \frac{MSE}{\left[1 - \frac{(f+1)+pf}{n}\right]^2} \tag{5}$$

where *MSE* = mean square error; *f* = number of bias functions; *p* = bias function penalty; and *n* = number of observations.

The MARS model performs under two types of forward and backward functions [126]. In the forward function stage, the model develops a huge quantity of bias functions introduced by the MARS model. The generalized form of the MARS model is given below [128]:

$$\mathrm{Y} = \beta_o + \sum_{i=1}^{m} \beta_i \mathrm{H}_{ei}(\mathrm{X}_v(e,\ \mathrm{I})) \tag{6}$$

where Y = output parameter; β_o = constant value; I = number of bias functions; $\mathrm{H}_{ei}(\mathrm{X}_v(e, \mathrm{I}))$ = ith bias function; and β_i = the corresponding coefficient of $\mathrm{H}_{ei}(\mathrm{X}_v(e,\ \mathrm{I}))$.

The model has a collection of bias functions. In the second stage, it can estimate the least square model. The MARS model is defined as follows [129]:

$$\mathrm{Y} = \partial_o + \sum_{i=1}^{n} \partial_i h_i(\mathrm{X}) \tag{7}$$

where $h_i(X)$ = the splines function; ∂ = the coefficient of the spline function; and i = the total number of functions in the model.

3.3. Support Vector Machine (SVM)

The support vector machine (SVM) is a supervised ML model that uses a nonlinear generalization algorithm to classify two groups and regression problems. The foundation of the SVM was made by Vapnik [111,130], and Bray and Han [131] introduced it. SVMs are generalized linear classifiers and supervised learning methods for regression and data classification. The kernel allows SVMs to form nonlinear boundaries. The different kernel functions include linear kernel, polynomial kernel, radial basis, and sigmoid kernel. The algorithm's expression by the dataset's inner products is called the dual problem. Support vector regression (SVR) was developed by Vladimir Vapnik [111]. It is characterized by using kernels, sparse solution, and Vapnik–Chervonenkis (VC) controls of margins and many support vectors.

SVR is a powerful tool in real-value function estimation. It estimates continuous-value multivariate functions. It uses a supervised learning approach: SVR trains by taking symmetrical loss functions and reducing high and low misestimation [132]. Support vector regression (SVR) attempts to minimize the upper limit of the generalization error instead of fixing the empirical error. The first formulation of SVR is a hard-margin solution that contributes to overfitting. The soft margin appears to generalize in the presence of outliers and noise. It prevents overfitting, which makes it favorable for forecasting research work. It has high generalization capability and great prediction accuracy. SVMs formulate binary classification problems to convex optimization problems [111]. The ε-intense region around the function allows SVM generalization to SVR, also called ε-tube. It helps reformulate the optimization problem in continuous-value functions, which helps to balance model complexity and prediction error. Considering the training dataset, T, represented as:

$$\mathrm{T} = \{(\mathrm{x}_1,\ \mathrm{y}_1),\ (\mathrm{x}_2,\ \mathrm{y}_2)\ldots\ \ldots(\mathrm{x}_\mathrm{m},\ \mathrm{y}_\mathrm{m})\} \tag{8}$$

where x ϵ X R^{n} are the training inputs and y ϵ Y $\subset$ R^{n} are the training outputs. Assume a nonlinear function y is given by:

$$f(x) = \mathbf{w}^{\mathbf{T}}\mathbf{\Phi}(\mathbf{x_i}) + b \tag{9}$$

where *f(x)* = nonlinear function; T = training data; w = weight vector, *b* = bias; and **Φ** (x_i) = higher dimensional feature space by the linear mapping function of input space x. The main objective is to fit the dataset T with the help of function *f(x)*, having the highest deviation ε from the training dataset T. The equation is now transformed into a constrained complex problem, as follows:

$$\text{min.}\ \left(\frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w}\right)\text{subject to :}\ \begin{cases}\mathrm{yi} - \left(\mathbf{w}^{\mathbf{T}}\mathbf{\Phi}(\mathbf{x_i}) + b\right) \leq\ \varepsilon \\ \mathrm{yi} - \left(\mathbf{w}^{\mathbf{T}}\mathbf{\Phi}(\mathbf{x_i}) + b\right) \geq \varepsilon\end{cases} \tag{10}$$

where ε ($\geq$0) is the maximum acceptable deviation. Equation (7) can be written as:

$$\text{min.}\ \left(\frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w}\right)\text{subject to :}\ \begin{cases}\mathrm{yi}\ \mathbf{-w}^{\mathbf{T}}\mathbf{\Phi}(\mathbf{x_i}\ ) - b\ \leq\ \varepsilon \\ \mathbf{w}^{\mathbf{T}}\mathbf{\Phi}(\mathbf{x_i}\ ) + b - \mathrm{y}_\mathrm{i} \leq \varepsilon\end{cases} \tag{11}$$

Further, the final expression for SVM becomes:

$$f(x) = \sum_{i=1}^{m}\left(\ \alpha_i^+ - \alpha_i^-\right)K\left(x_i,\ x_j\right) + b \tag{12}$$

where α_i^+ and α_i^- are the Langrangian multipliers, and $K\left(x_i,\ x_j\right)$ is the kernel function. [133].

3.4. Random Forest (RF)

Random forest is a supervised ML that uses ensemble learning techniques for classification and regression problems. It is a technique that predicts from different machine learning algorithms or the same algorithm several times for more accurate predictions. RF uses the bagging technique or bootstrap aggregation, part of ensemble learning. It generates several subsets of data from training samples chosen randomly with replacement. Each subset of data is used to train its decision tree.

The bootstrap aggregation technique reduces the variance of an estimated prediction function. Bagging works excellently for high variance and low bias, such as decision trees. Random forest constructs multiple decision trees during training. It combines the prediction results from each decision tree to give the final output. Decision trees are computationally expansive; these are very sensitive to data on which they are trained and may experience deviation in predictions if the underlying data are changed. Several decision trees are constructed by the algorithm that operates the model. RF is the aggression of tree predictors. The trees are estimated by the values of a random vector computed from the same distribution for each forest tree [134]. In the RF model, every tree is grown with a random subset of variables [135]. It ensures that the bagged trees are in the way that a single tree reduces the correlation and variance between trees of the model. Each decision tree picks a random sample from the dataset, while generating its split adds a further element of randomness to minimize the problem of overfitting. The random forest chooses nodes from a subset of available features that breaks variables at each node to reduce the association between the trees. The mean square error (*MSE*) can be calculated as [127]:

$$MSE = \frac{1}{N}\sum_{i=1}^{m}(Z_i - i)^2 \tag{13}$$

where Z_i = the measured variable value and i = the mean of all out-of-bag (OOB) predictions.

The RF model comes under the classification of regression tree (CART) tools and is used for classification and regression problems. Many RF trees are key parameters; the model performance can be evaluated by out-of-bag (OOB). Random forest can help over-fitting the model for the training dataset, which can be evaded by selecting input data during the training cycle and establishing variation in weak learners [136]. The RF model makes multiple decision trees, and the output of the models can be estimated by taking the mean output of every tree. The predicted values are calculated as:

$$Y = \frac{1}{N}\sum_{i=1}^{N} R(x) \tag{14}$$

where Y = the predicted output by the RF model, N = the number of trees (n-tree) utilized in the RF model, and *R(x)* = the results of every random tree.

4. Performance Evaluation of Models

The performances of the MLR, MARS, SVM, and RF models were evaluated based on the coefficient of determination (R^2), root mean square error (*RMSE*), Nash–Sutcliffe efficiency (*NSE*), and percent bias (*PBIAS*), and visual interpretation using a line diagram, scatter diagram, violin plot, and relative and Taylor diagrams. The R^2, *RMSE* [12,47,117,137,138], *NSE* [139], and *BIAS* [119,140,141] are described as:

$$R^2 = \left[\frac{\left[\sum Q_o Q_p\right] - \left[\frac{\sum Q_o \sum\ Q_p}{N}\right]}{\sqrt{\left[\sum Q_p{}^2 - \frac{(\sum Q_p)^2}{N}\right]\left[\sum Q_o{}^2 - \frac{(\sum Q_o)^2}{N}\right]}}\right]^2 \tag{15}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_o - Q_p)^2} \tag{16}$$

$$NSE = 1 - \left[\frac{\sum_{i=1}^{N}(Q_o\ - Q_p)^2}{\sum_{i=1}^{N}(Q_o\ - \overline{Q_o}\,)^2}\right] \tag{17}$$

$$PBIAS = \left[\frac{\sum_{i=1}^{N}(Q_o\ - Q_p)}{\sum_{i=1}^{N}(Q_o)}\right] \times 100 \tag{18}$$

where Q_o = the observed runoff value; Q_p = the predicted runoff value; N = the total number of values of the variable in the dataset; and $\overline{Q_o}$ = the mean of the observed discharge data.

The coefficient of determination describes the statical relationship (collinearity) between the variables and helps to show the nature of association among the predicted and observed data. R^2 is the ratio of explained variation compared to the total variation [142]. The coefficient of multiple determination measures the percentage of various independent variables, which can be explained by variations in the independent variables when taken together [143]. It ranges from 0 to 1; its higher value indicates less error variance, and generally, a value greater than 0.5 is considered acceptable [144,145]. It is famously used in model evaluation. This statistical tool is highly sensitive to outliers and insensitive to additive and proportional differences between observed and predicted data [146]. The square root of the average square of all of the errors is called root mean square error (*RMSE*) [104]. It is an excellent general-purpose error matrix commonly used for the numerical prediction model. *RMSE* has a good measure of accuracy, but it can only compare the prediction error of models or configure a particular variable and not between two different variables, making it scalar-dependent.

RMSE lies between 0 to ∞ [47]. NSE is a normalized statistical tool determining the relative magnitude of residual variance or noise. NSE lies between $-\infty$ and 1 and is less sensitive to high extreme values [139]. Percent bias measures the relationship between the observed data and its predicted data; it measures the average tendency of observed data to be larger or smaller than the predicted data [147]. Percent bias describes whether the simulated model is overestimated or underestimated. A low PBIAS value or a value that tends to zero indicates the optimal model. A negative value indicates the overestimation of the model.

In contrast, a positive PBIAS value indicates an underestimation of the model [119,140,141,147]. When the data are evaluated, PBIAS reveals any deviation of the data as a percentage [148]. A model with higher R^2 and NSE values, and lower RMSE and PBAIS values, decrees a relatively better model for the simulation of Q_t.

5. Results and Discussion

This section deals with the development and results of runoff prediction models using ML techniques for the Gola watershed. This study categorizes runoff modeling into two approaches based on the machine learning model context. One approach considers that a forecasting model can be based on the river flow data by including the correlated lag times as an attribute to forecast one step. On the other hand, the second approach to modeling river flow entails an appropriate exogenous hydrological variable apart from the river flow data. Multiple linear regression (MLR), multivariate adaptive regression splines (MARS), support vector machine (SVM), and random forest (RF) models were applied to develop the rainfall–runoff models. The qualitative performance evaluation of the models was achieved by visual observations such as time series, scatterplot, violin plot, relative error, and Taylor plot, and quantitative evaluations were carried out using different statistical and hydrological performance indices, namely, root mean square error (RMSE), coefficient of determination (R^2), the Nash–Sutcliffe coefficient of efficiency (NSE), and percent bias (PBIAS).

5.1. Selection of Best Input Combination

Selecting the most appropriate input variables is vital to model development [79]. In the present investigation, the GT algorithm was used for selecting the relevant input variable combination for runoff prediction. In this study, various combinations of present-day runoff ($Q_{(t)}$), previous day runoff ($Q_{(t-1)}$), previous two days' runoff ($Q_{(t-2)}$), previous three days' runoff ($Q_{(t-3)}$), present-day rainfall ($R_{(t)}$), previous day rainfall ($R_{(t-1)}$), previous two days' rainfall ($R_{(t-2)}$) and previous three days' rainfall ($R_{(t-3)}$) were used for Gamma testing (Table 2). The models with low Gamma (Γ) and V_{ratio} values were considered the most appropriate for developing the models [149]. It was noticed that the Gamma value and V_{ratio} decreased with an increase in the number of predictors. However, after a certain point, the Gamma value began to increase again. This might be due to the following two reasons: (i) the inclusion of a high number of input variables may be the cause of overfitting, and (ii) the inclusion of a smaller number of input variables results in the incapacity of the model to correctly explain the total variance of the forecasted subset. The minimum Gamma (Γ) and V_{ratio} values were 0.407 and 0.191, respectively, for the M19 predictor set. Hence, the M19 predictor combination was employed for further analysis. It could be stated that using rainfall with a two-day lag and the discharge from one to three days' lag as a predictor would produce an optimum rainfall–runoff model. It was also noticed that the Gamma value and V_{ratio} increased when the rainfall of the three-day lag was included in the predictors. This might be due to the low correlation of the predictor variable with the predictand.

Table 2. Gamma statistics for different input combinations.

Model No.	Model Input Combination	Mask	Gamma	V-Ratio
M1	$Q_{(t-1)}$	0000001	0.082	0.329
M2	$R_{(t)}, Q_{(t-1)}$	1000001	0.061	0.244
M3	$R_{(t)}, R_{(t-1)}, Q_{(t-1)}$	1100001	0.064	0.256
M4	$R_{(t)}, R_{(t-1)}, R_{(t-2)}, Q_{(t-1)}$	1110001	0.056	0.225
M5	$R_{(t)}, R_{(t-1)}, R_{(t-2)}, R_{(t-3)}, Q_{(t-1)}$	1111001	0.051	0.207
M6	$R_{(t)}, R_{(t-1)}, R_{(t-2)}, R_{(t-3)}, Q_{(t-3)}, Q_{(t-1)}$	1111101	0.054	0.217
M7	$Q_{(t-1)}, Q_{(t-2)}$	0000011	0.081	0.320
M8	$R_{(t)}, Q_{(t-1)}, Q_{(t-2)}$	1000011	0.063	0.254
M9	$R_{(t)}, R_{(t-1)}, Q_{(t-1)}, Q_{(t-2)}$	1100011	0.051	0.212
M10	$Q_{(t-1)}, Q_{(t-2)}, Q_{(t-3)}$	0000111	0.076	0.307
M11	$R_{(t)}, R_{(t-1)}, Q_{(t-1)}, Q_{(t-2)}, Q_{(t-3)}$	1000111	0.053	0.213
M12	$R_{(t)}, R_{(t-2)}, Q_{(t-1)}, Q_{(t-2)}, Q_{(t-3)}$	1100111	0.048	0.194
M13	$R_{(t)}, R_{(t-1)}, R_{(t-2)}, R_{(t-3)}$	1111000	0.107	0.430
M14	$R_{(t)}, R_{(t-1)}, R_{(t-2)}, R_{(t-3)}, Q_{(t-1)}$	1111001	0.061	0.207
M15	$R_{(t)}, R_{(t-1)}, R_{(t-2)}, Q_{(t-1)}, Q_{(t-2)}, Q_{(t-3)}$	1110111	0.050	0.200
M16	$R_{(t)}, R_{(t-1)}, R_{(t-2)}$	1110000	0.114	0.459
M17	$R_{(t)}, R_{(t-1)}, R_{(t-2)}, Q_{(t-1)}$	1110001	0.056	0.225
M18	$R_{(t)}, R_{(t-1)}, R_{(t-2)}, Q_{(t-1)}, Q_{(t-2)}$	1110011	0.052	0.209
M19	$\mathbf{R_{(t)}, R_{(t-1)}, R_{(t-2)}, Q_{(t-1)}, Q_{(t-2)}, Q_{(t-3)}}$	1110111	0.047	0.191
M20	$R_{(t)}, R_{(t-1)}$	1100000	0.124	0.498
M21	$R_{(t)}, R_{(t-1)}, Q_{(t-1)}$	1100001	0.064	0.256
M22	$R_{(t)}, R_{(t-1)}, Q_{(t-1)}, Q_{(t-2)}$	1100011	0.053	0.212
M23	$R_{(t)}, R_{(t-1)}, Q_{(t-1)}, Q_{(t-2)}, Q_{(t-3)}$	1100111	0.073	0.194
M24	$R_{(t)}, R_{(t-1)}, R_{(t-3)}\ Q_{(t-1)}, Q_{(t-2)}, Q_{(t-3)}$	1101111	0.050	0.203
M25	$R_{(t)}, Q_{(t-1)}$	1000001	0.061	0.244
M26	$R_{(t)}, Q_{(t-1)}, Q_{(t-2)}$	1000011	0.063	0.254
M27	$R_{(t)}, Q_{(t-1)}, Q_{(t-2)}, Q_{(t-3)}$	1000111	0.053	0.213
M28	$R_{(t)}, R_{(t-1)}, Q_{(t-1)}, Q_{(t-2)}, Q_{(t-3)}$	1100111	0.048	0.194
M29	$R_{(t)}, R_{(t-1)}, R_{(t-2)}, R_{(t-3)}$	1111000	0.107	0.430
M30	$R_{(t)}, R_{(t-1)}, R_{(t-2)}, R_{(t-3)}, Q_{(t-1)}$	1111001	0.051	0.207
M31	$R_{(t)}, R_{(t-1)}, R_{(t-2)}, Q_{(t-1)}, Q_{(t-2)}$	1110011	0.052	0.209
M32	$R_{(t)}\ R_{(t-1)}\ R_{(t-2)}\ R_{(t-3)}\ Q_{(t-1)}\ Q_{(t-2)}$	1111011	0.050	0.200
M33	$R_{(t)}, R_{(t-1)}, R_{(t-2)}\ R_{(t-3)}, Q_{(t-1)}, Q_{(t-2)}, Q_{(t-3)}$	1111111	0.048	0.194

5.2. Application of Machine Learning Techniques for Rainfall–Runoff Modeling

As per the GT result, the best input combination for the development of the MLR, MARS, SVM, and RF models was made based on the following equation:

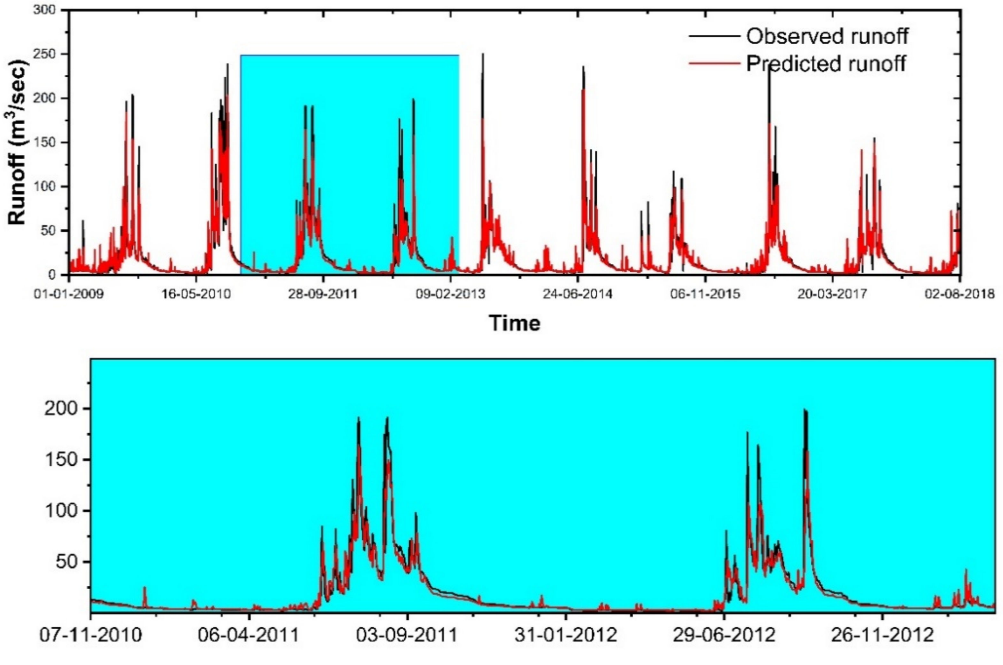
$$Q_{(t)} = f\left(R_{(t)}, R_{(t-1)}, R_{(t-2)}, Q_{(t-1)}, Q_{(t-2)}, Q_{(t-3)}\right) \tag{19}$$

5.2.1. MLR Model for Runoff Prediction

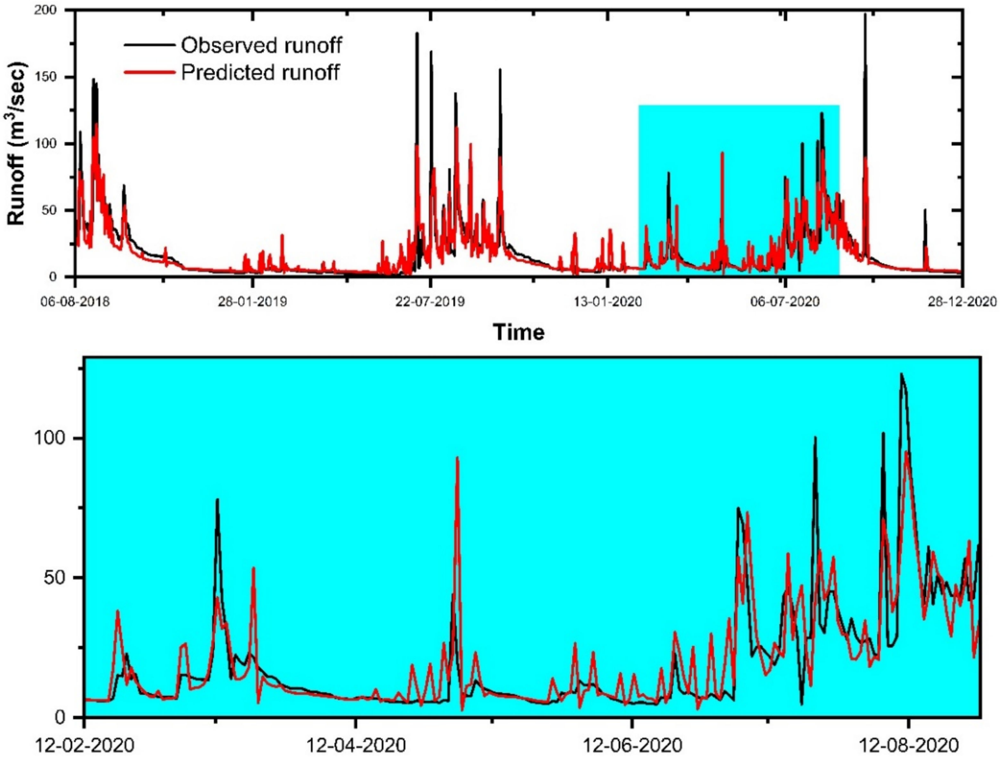
The MLR technique was used to predict the runoff of the Gola watershed using the best input combination based on the Gamma test results in the R-Studio environment. The developed MLR model with its training dataset can be formulated as follows:

$$Q_{(t)} = 1.11 + 0.55R_{(t)} + 0.01R_{(t-1)} - 0.03R_{(t-2)} + 0.48Q_{(t-1)} + 0.21Q_{(t-2)} + 0.03Q_{(t-3)} \tag{20}$$

The qualitative performance assessment of the MLR model for predicting the runoff of the Gola watershed was done using a graphical comparison between the ordinates of the observed and predicted runoff values (Figures 4 and 5). The visual observation revealed that there was a large variation in the observed and predicted peak values of the runoff. It was also observed that the MLR model underpredicts higher flow/runoff and overpredicts lower flow/runoff in both the training and testing periods (Figure 4a,b). The MLR, being a linear model, could not capture the nonlinearity in the predictor–predictand relationship. Hence, the MLR model explained the medium range of variance in the predictand variable better than the extreme values. In other words, these models simulated the average runoff values more effectively than extreme events. The average points are well simulated in most cases.

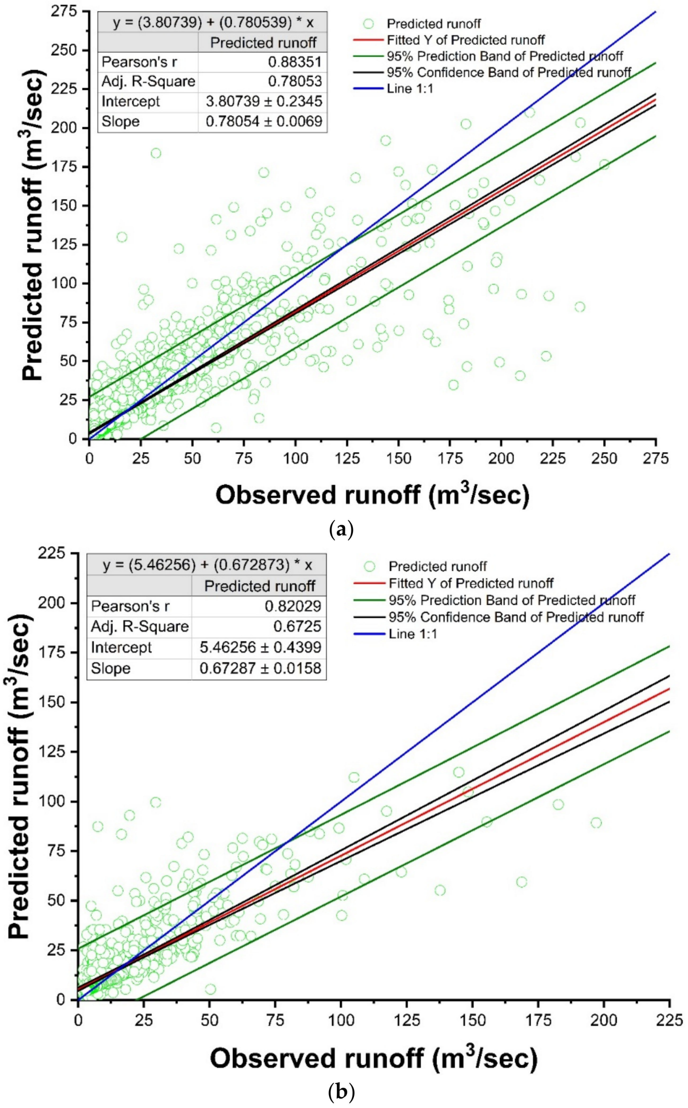


(**a**)



(**b**)

Figure 4. Results of simulation of the runoff (m^3/s) of Gola watershed using MLR model from 2009 to 2020 (**a**) training and (**b**) testing period.



(**a**)

(**b**)

Figure 5. Scatter plots, evaluation statistics, and confidence intervals of the observed and predicted daily runoff Q(t) using MLR model. (**a**) Training dataset and (**b**) testing dataset.

The values of RMSE, R^2, NSE, and PBIAS were 13.44 m^3/s, 0.78, 0.72, and 0.00%, respectively, during the training and 12.67 m^3/s, 0.67, 0.51, and 0.80%, respectively, during the testing period for the MLR model (Table 3). The model revealed low training bias and underestimated runoff values in the testing period. It was seen that the MLR model lacked the satisfactory mapping of the Gola watershed's runoff. According to Figure 5a,b, using 95% confidence intervals, the results showed that most of the points of simulated runoff values (m^3/s) are outside of the confidence range, which indicates overestimation and underestimation of the target points in both periods. However, according to the presented results, the model's performance is not acceptable.

Table 3. Comparison of different machine learning models for daily runoff prediction.

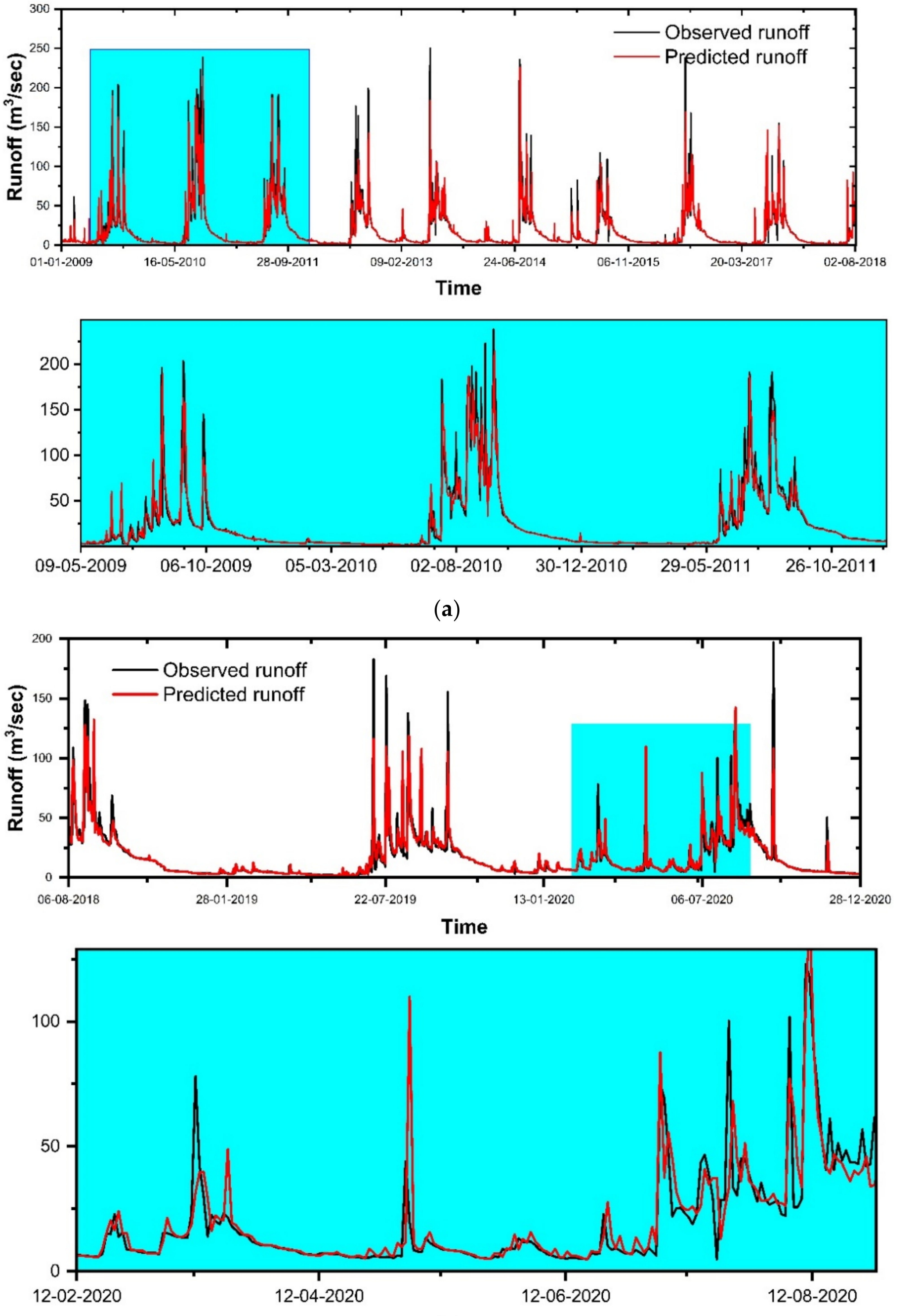
Model	Training				Testing			
	RMSE (m^3/s)	R^2	NSE	PBIAS (%)	RMSE (m^3/s)	R^2	NSE	PBIAS (%)
MLR	13.44	0.78	0.72	0	12.67	0.67	0.51	0.80
MARS	12.55	0.81	0.76	0	10.07	0.79	0.74	0.20
SVM	12.62	0.83	0.81	4.10	14.02	0.60	0.60	−0.40
RF	6.31	0.96	0.94	−0.20	5.53	0.95	0.92	−0.20

5.2.2. MARS Model for Runoff Prediction

In the case of MARS modeling, the backward pass was used to prune the model by deleting unnecessary bias functions at every stage until the supermodel was found and good generalization ability was achieved. The values of the GCV parameter for best models were 159.80 and 107.05 for the training and testing sets, respectively. RMSE, R^2, NSE, and PBIAS were 12.55 m^3/s, 0.81, 0.76, and 0.00%, respectively, during training, and 10.07 m^3/s, 0.79, 0.74, and 0.20%, respectively, during the testing period for the MARS model (Table 4). The temporal variations and scatter plots of the observed and predicted runoff during the training and testing period are displayed in Figures 6 and 7, respectively. As can be seen from Figure 6a,b, there is good agreement between the observed runoff values and the corresponding values simulated by the MARS model in the training and testing periods, respectively. The trend of predicting runoff was satisfactory for the observed runoff of the Gola watershed. The peak values of the runoff were not predicted with great accuracy. Low values of PBIAS were found in the MARS model during the training period, which indicates an accurate model simulation. The PBIAS value of 0.00% and the positive value of 0.20% indicated slight underestimation during the testing period. According to Figure 7a,b, using 95% confidence intervals, the results showed that some of the points of the simulated runoff values (m^3/s) are outside of the confidence range, which indicates the overestimation and underestimation of the target points in both periods. However, the model's performance is acceptable according to the presented results.

Table 4. Results of different performance indicators for RF models during training and testing sets.

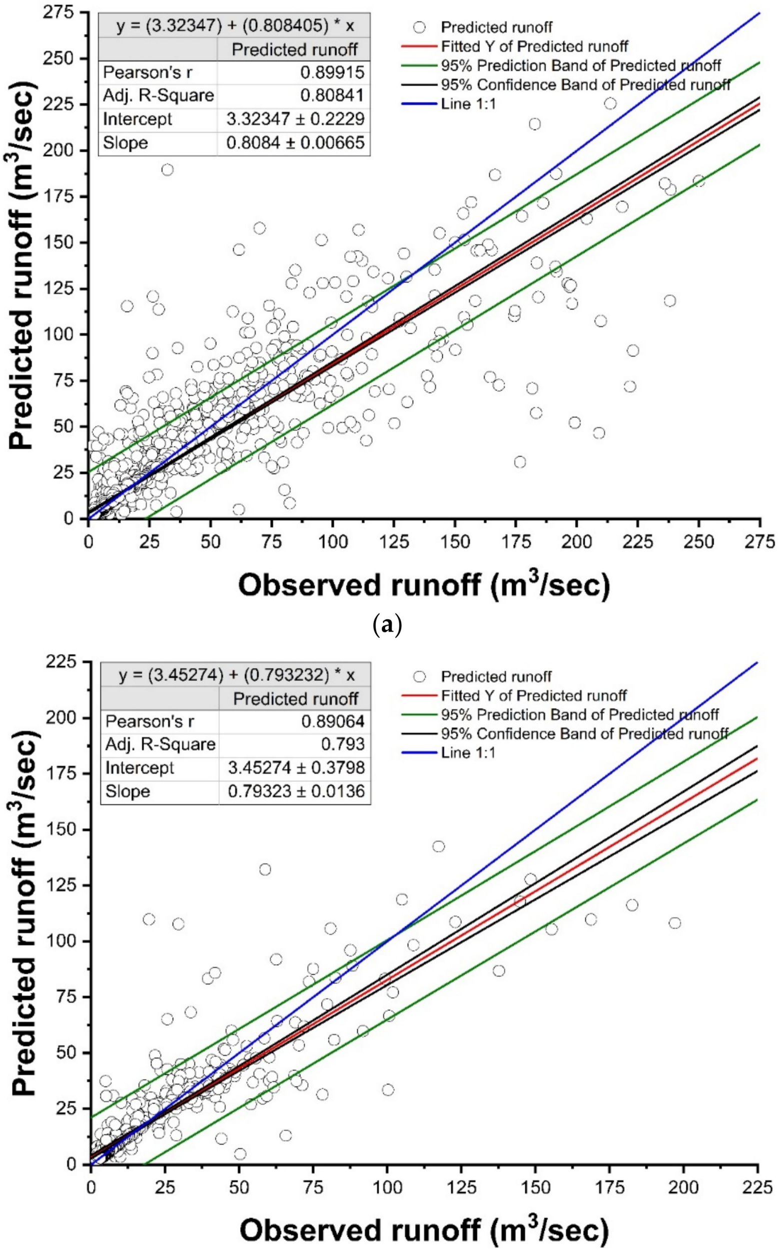
Models	Training		Testing	
	RMSE	R^2	RMSE	R^2
RF-1	6.443	0.95	5.553	0.95
RF-2	6.388	0.95	5.576	0.95
RF-3	6.351	0.96	5.572	0.94
RF-4	6.423	0.95	5.550	0.95
RF-5	6.442	0.95	5.621	0.94
RF-6	6.480	0.95	5.459	0.95
RF-7	6.415	0.95	5.609	0.95
RF-8	6.371	0.95	5.677	0.94
RF-9	6.404	0.95	5.598	0.95
RF-10	6.378	0.95	5.521	0.95
RF-11	6.403	0.95	5.430	0.95
RF-12	6.445	0.95	5.481	0.95
RF-13	6.373	0.95	5.563	0.95
RF-14	6.449	0.95	5.500	0.95
RF-15	6.447	0.95	5.468	0.95
RF-16	6.435	0.96	5.571	0.94
RF-17	6.386	0.95	5.580	0.95
RF-18	6.395	0.95	5.539	0.95
RF-19	6.481	0.95	5.451	0.95
RF-20	6.408	0.95	5.575	0.95
RF-21	6.375	0.95	5.626	0.94
RF-22	6.389	0.95	5.529	0.95
RF-23	6.451	0.95	5.467	0.95
RF-24	6.446	0.95	5.613	0.94
RF-25	6.369	0.95	5.453	0.95
RF-26	6.427	0.95	5.536	0.95
RF-27	6.322	0.95	5.547	0.95
RF-28	**6.318**	**0.96**	**5.565**	**0.95**
RF-29	6.375	0.95	5.480	0.95



(**a**)

(**b**)

Figure 6. Results of simulation of the runoff (m^3/s) of Gola watershed using MARS model from 2009 to 2020 (**a**) training and (**b**) testing period.



(**a**)

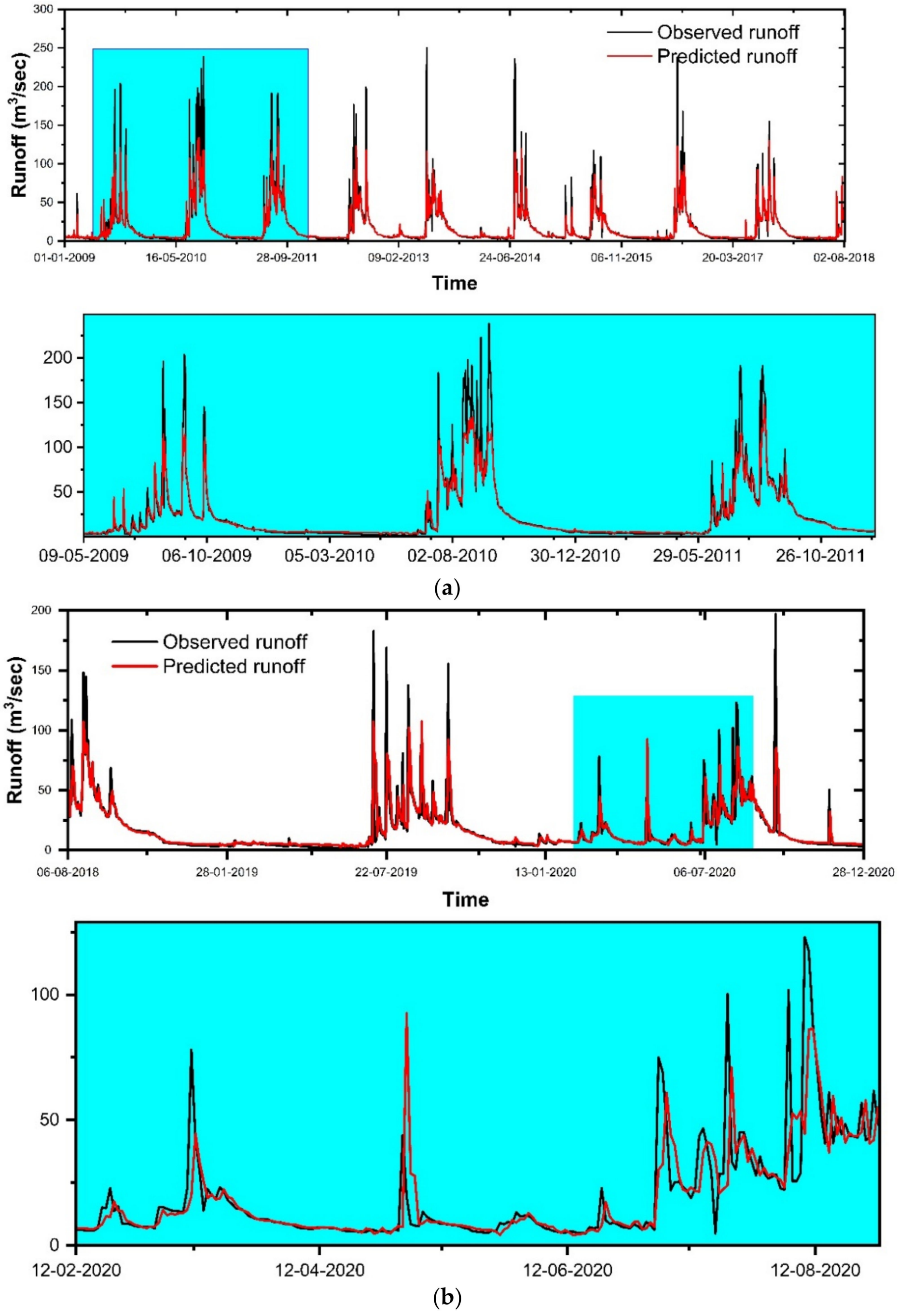
(**b**)

Figure 7. Scatter plots, evaluation statistics, and confidence intervals of the observed and predicted daily runoff Q(t) using MARS model. (**a**) Training dataset and (**b**) testing dataset.

5.2.3. SVM Model for Runoff Prediction

The present study uses the radial basis and linear kernel functions for SVM-based runoff modeling. The radial basis function performed better than the linear function, which is the reason it was selected for the current study. Because the trial-and-error method cannot achieve the best performance, optimization algorithms were used in the SVM [150].

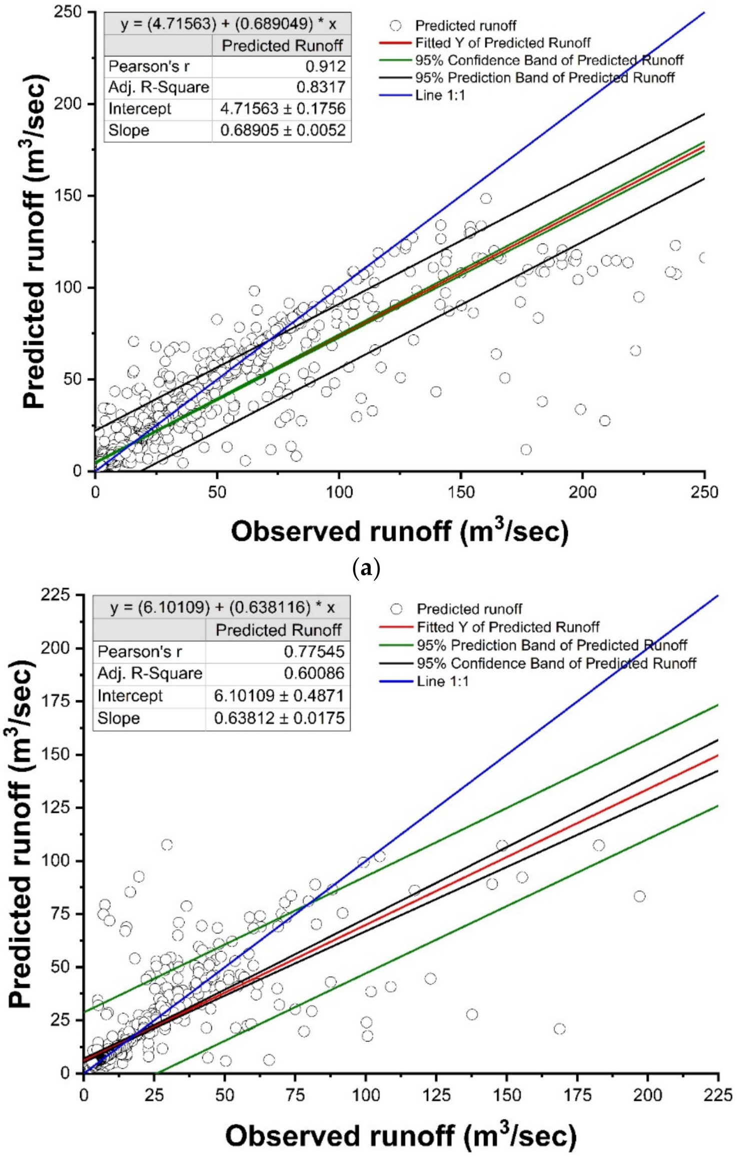
The temporal variations and scatter plots of the observed and predicted runoff during the training and testing period are displayed in Figures 8 and 9, respectively. The time-series plot revealed the fact that the model was underpredicting large values in the training period as well as in the testing period (Figure 8a,b). The values of RMSE, R^2, NSE, and PBIAS for the SVM model were 12.614 m3/s, 0.83, 0.81, and −3.90% for the training period and 14.02 m^3/s, 0.60, 0.60, and 0.40% for the testing period, respectively (Table 3). The R^2 value (0.83) shows a strong linear relationship between the observed and predicted variables in the training period. It was found satisfactory (0.60) during the testing period. The NSE value (0.81) revealed good model predictive skills during training. The 0.60 value in the testing period shows satisfactory predictive skills during the testing period. The PBIAS value was found to be −3.90% during the training period, which shows the model was overpredicting the runoff values during the training period, and the testing period (0.40%) reveals that the model was underpredicting the runoff values. According to Figure 9a,b, using 95% confidence intervals, the results showed that some of the points of the simulated runoff values (m^3/s) are outside of the confidence range, which indicates underestimation in the training dataset and overestimation and underestimation of the target points in the testing dataset. However, the model's performance is acceptable according to the presented results.



(**a**)

(**b**)

Figure 8. Results of simulation of the runoff (m^3/s) of Gola watershed using SVM model from 2009 to 2020 (**a**) training and (**b**) testing period.



(**a**)

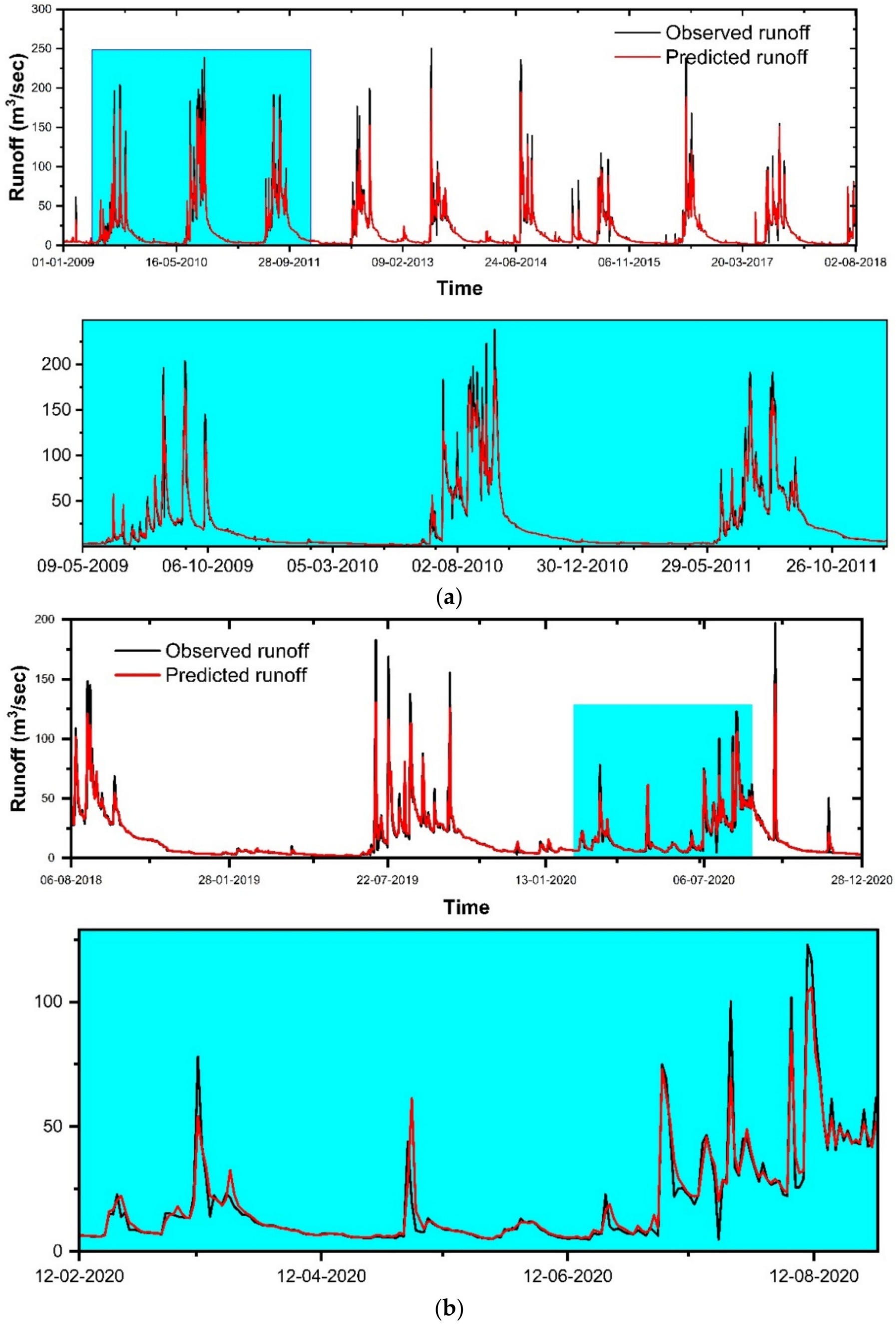
(**b**)

Figure 9. Using SVM model, scatter plots, evaluation statistics, and confidence intervals of the observed and predicted daily runoff Q(t). (**a**) Training dataset and (**b**) testing dataset.

5.2.4. Random Forest Model for Runoff Prediction

Two parameters were used in tuning the random forest model *ntree* (number of trees) and *mtry* (number of variables) [150]. In the present study, a trial-and-error technique was used in which the *n-tree* values varied from 200 to 600, and the *m-try* values varied from 2 to 6 to find the best performing random forest model. It was found that 400 decision trees (*n-tree*) and seven variables (*m-try*) were optimal for the best model fit. It was apparent from Table 4 that the RMSE values were in the range of 6.318 m^3/s to 6.480 m^3/s and R^2 values were in the range of 0.95 to 0.96 during the training period. The values of RMSE lay in the range of 5.430 m^3/s to 5.677 m^3/s, and R^2 values lay in the range of 0.94 to 95 during the testing period. From the evaluation of all of the results, it was observed that the RF-28 model was superior to the other RF models.

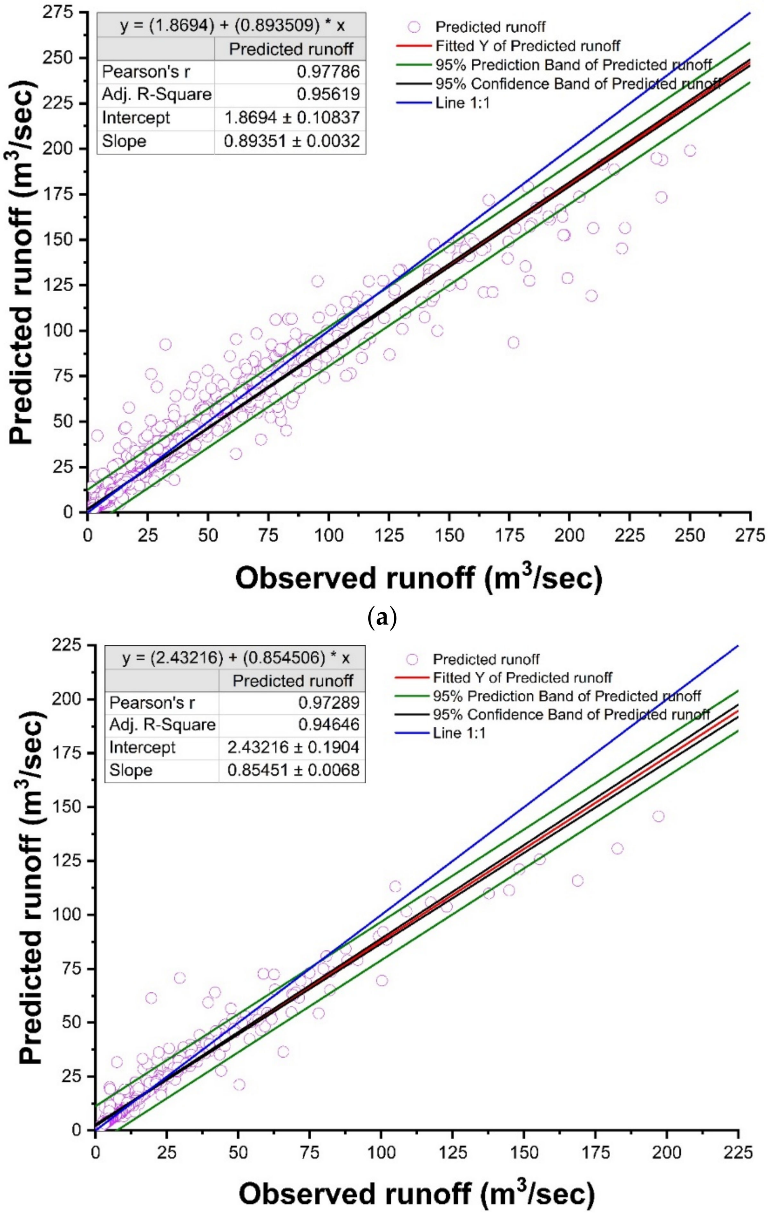
The RMSE, R^2, NSE, and PBIAS values of the RF-28 model were 6.318 m^3/s, 0.96, 0.94, and −0.20% for the training period and 5.565 m^3/s, 0.95, 0.92, and −0.10% for the testing period (Table 3). The low RMSE values show a concentration of data around the best fit line. The R^2 values (0.96 and 0.95) during the training and testing period revealed a strong linear relationship between the observed and predicted runoff values. The NSE values were found to be 0.94 and 0.95 during the training and testing periods, respectively, which shows the good predictive ability of the model. The PBIAS values revealed that the model slightly overpredicted the runoff values during training and testing. The temporal variations and scatter plots of the observed and predicted runoff during the training and testing periods are displayed in Figures 10 and 11, respectively. The time-series plot revealed the fact that the model slightly overpredicted in the training period as well as in the testing period (Figure 10a,b). The simulation results of the RF model are also shown in Figure 11a,b: except for a few overestimated and underestimated cases in the testing period, all of the simulated data are in the 95% confidence intervals. The model's accuracy is also confirmed. According to the statistics presented in Figure 11a,b, it can be concluded that the RF model has a high ability to simulate the runoff value.



(**a**)

(**b**)

Figure 10. Results of simulation of the runoff (m^3/s) of Gola watershed using RF model from 2009 to 2020 (**a**) training and (**b**) testing period.



(**a**)

(**b**)

Figure 11. Using RF model, scatter plots, evaluation statistics, and confidence intervals of the observed and predicted daily runoff Q(t). (**a**) Training dataset and (**b**) testing dataset.

5.3. Model Comparison

It was noticed that the MLR model showed the least accuracy among all models (Table 3). As reported by Panda et al. [151] (2022), the poor performance of the MLR model might be attributed to the following two reasons: (1) the inability of the MLR to address predictor–predictand nonlinearity, and (2) the reduced efficiency of the model due to the presence of outliers and serial correlation. The RF-developed hydrological model exhibited significantly improved performance over the rest of the models. The superior performance of the RF model could be due to the following reasons: (1) the superior ability of the model to address nonlinearity compared to the rest of the models [152], (2) its capacity to handle noisy data efficiently [153], and (3) its ability to reduce the overfitting problem [154]. It was also found that the MARS model outperformed the SVM model during the testing period. This indicated that the MARS could handle the predictor–predictand nonlinearity better than the SVM model.

The violin plot distribution of the observed and simulated runoff during the training and testing periods is depicted in Figure 12. The MARS model captured the extreme values better during the training than the other models. However, the RF model demonstrated a greater ability to capture the high runoff values during the testing period. This indicated that the RF model could learn the hidden processes better than the other models. The performance of the MARS model was similar to that of the RF. Although the MLR performed better in capturing extreme events during the calibration period, it could not perform similarly during the validation period. The SVM model showed the least efficacy in simulating high values during the calibration and validation periods.

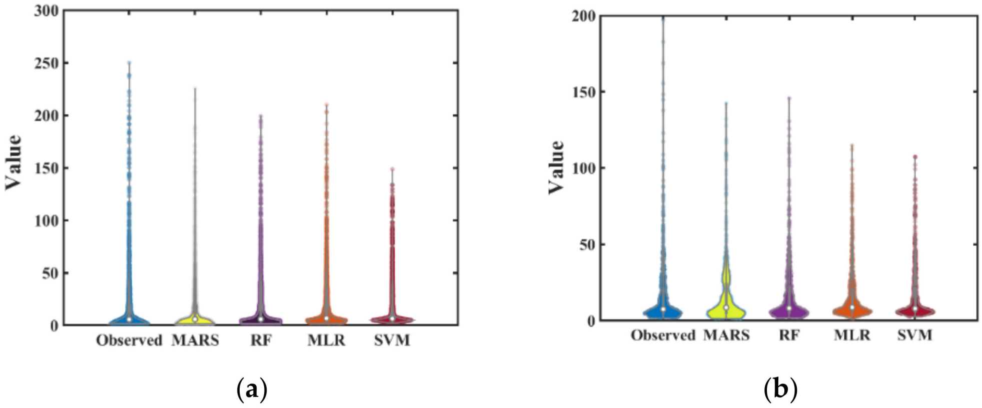


Figure 12. Violin plot displays the observed and predicted runoff distribution for the four models during the (**a**) training and (**b**) testing phase at the Gola watershed.

The relative error plot further validated the above results (Figure 13). Finally, the model efficiencies were compared using a Taylor diagram (Figure 14). It was concluded that the RF model showed the highest accuracy, followed by the MARS model. In contrast, the SVM model showed the lowest efficacy, followed by the MLR model.

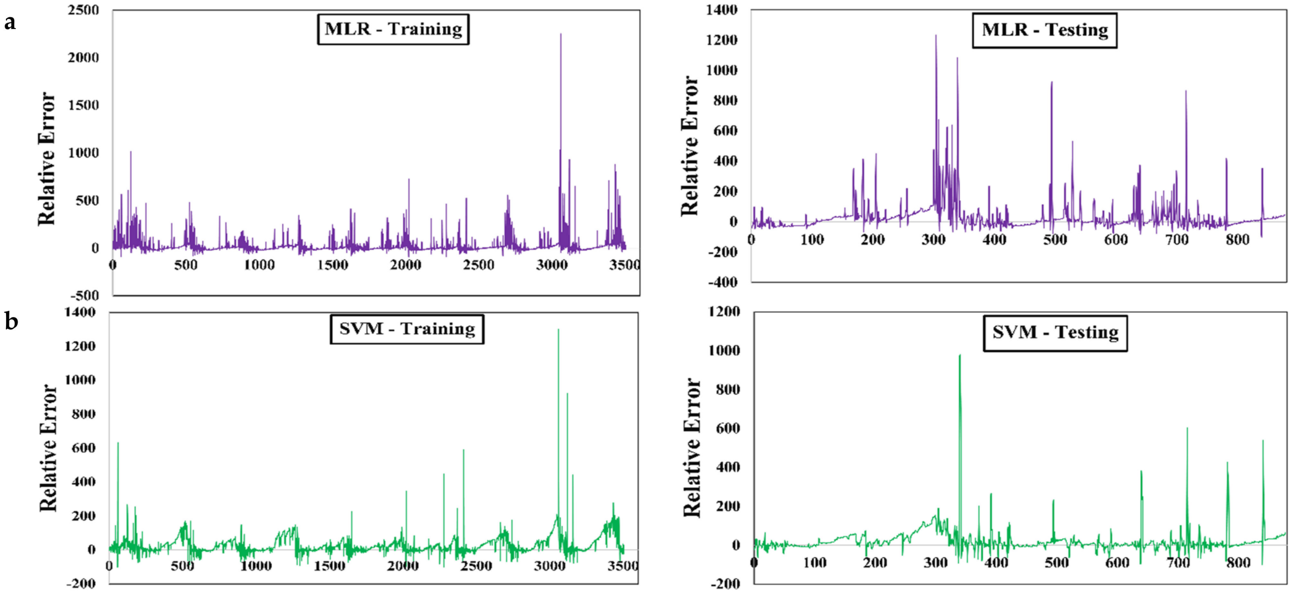


Figure 13. *Cont.*

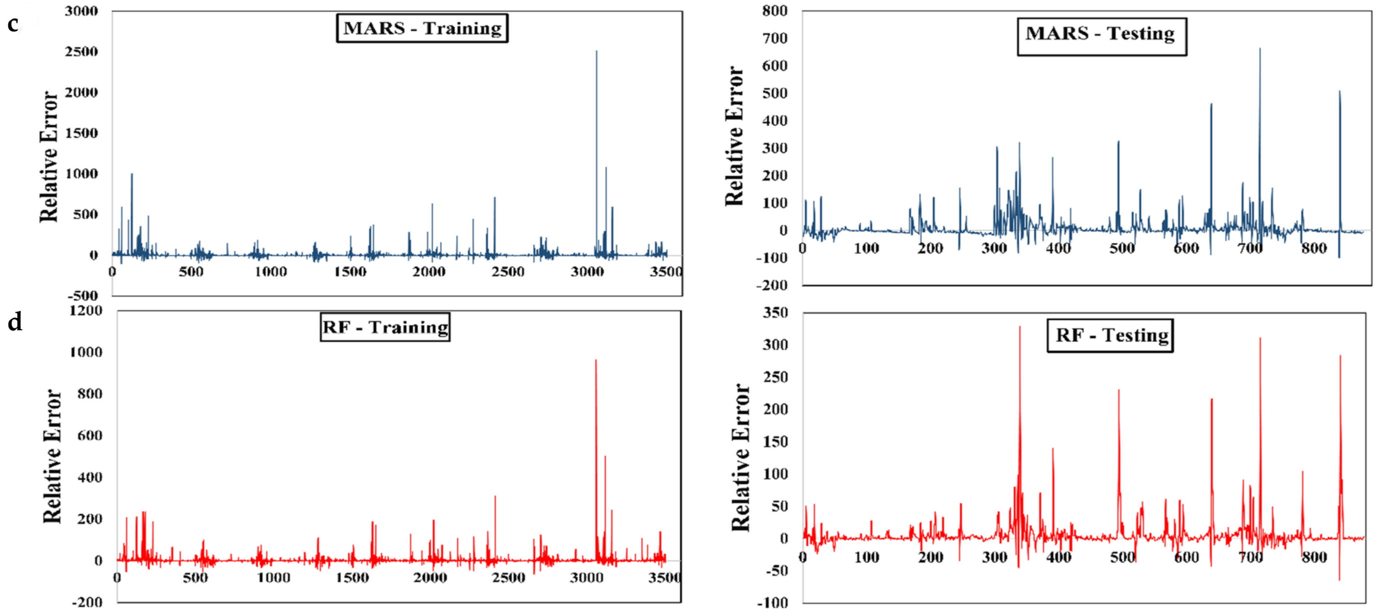
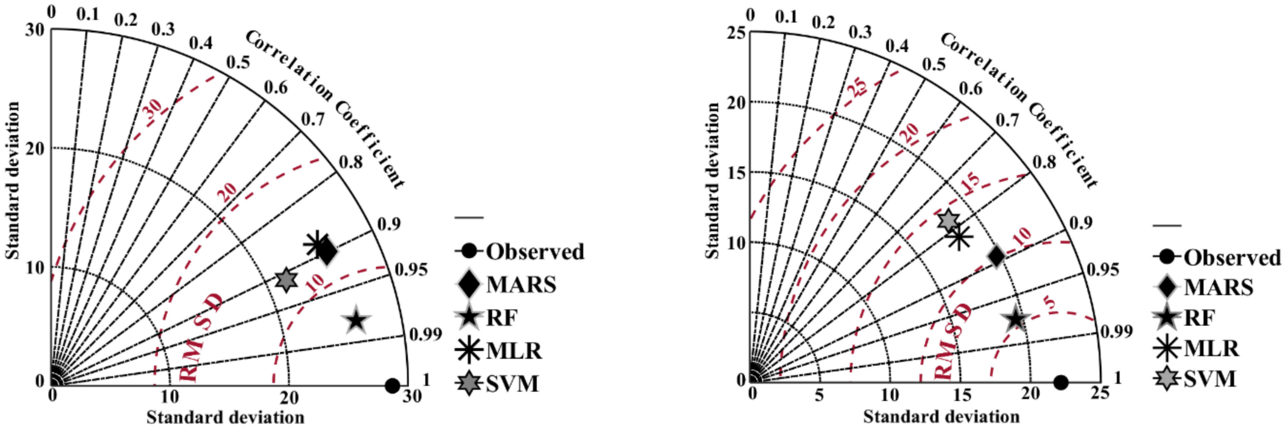


Figure 13. Relative error distribution over the training and testing phase for the daily timescale river flow for the Gola watershed, (**a**) MLR, (**b**) SVM and (**c**) MARS, and (**d**) random forest.



(**a**) (**b**)

Figure 14. Taylor diagram of SVM, random forest, MARS, and MLR models during the (**a**) training and (**b**) testing period at the Gola watershed.

6. Conclusions

Rainfall is an essential hydrological phenomenon to maintain the balance of freshwater availability for the survival and growth of life. This study was conducted to evaluate the runoff pattern of the Gola watershed. The comparative results of the training and testing datasets between the MLR, MARS, SVM, and RF models' potential to predict the runoff of the Gola watershed were investigated. Among the developed models, in terms of root mean square error (RMSE), the ranking of the models was RF, MARS, SVM, and MLR for the training period and RF, MARS, MLR, and SVM for the testing period, respectively. Based on the coefficient of determination (R^2) statistics, the models were ranked as RF, SVM, MARS, and MLR for the training dataset and RF, MARS, MLR, and SVM for the testing dataset. The models were ranked as RF, SVM, MARS, and MLR for training and RF, MARS, SVM, and MLR for testing in the case of NSE statistics, respectively. Based on the quantitative analysis and indices, the ranking of the models was RF, MARS, SVM, and MLR for the training period, and RF, MARS, MLR, and SVM for the testing period. Perhaps these results were due to data division, input uncertainties, and model parameter optimization. In order to determine the consistency of the models, these should be tested using varying data lengths and training–testing splits. The obtained results suggested that the accuracy of the MLR, MARS, SVM, and RF techniques were adequate using rainfall and runoff parameters for modeling. It was found that there was variation in the results of different machine learning models. The evaluation of the model performance revealed that the RF model outperformed the other regression models for predicting the runoff of the Gola watershed.

Author Contributions: Conceptualization, A.K.S. and P.K.; methodology, A.K.S.; software, A.K.S.; validation, P.K. and D.K.V.; formal analysis, A.K.S. and K.S.K.; investigation, P.K. and D.K.V.; resources, P.K.; data curation, A.K.S.; writing—original draft preparation, A.K.S., D.K.V., K.C.P., A.S. and K.S.K.; writing—review and editing, D.K.V., R.A., A.E., A.K. and S.H.; visualization, D.K.V. and E.M.; supervision, P.K. and D.K.V.; project administration, A.K.S.; funding acquisition, N.A.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The authors would like to collaborate, and data can be available upon request.

Acknowledgments: The authors are grateful to the Department of Soil and Water Conservation Engineering, G.B. Pant University of Agriculture and Technology, Pantnagar, Uttarakhand, India, and to Gola Barrage gauge station Haldwani–Kathgodam, Uttarakhand, India, for providing data for this research. Alban Kuriqi acknowledges the Portuguese Foundation for Science and Technology (FCT) for their support through PTDC/CTA-OHR/30561/2017 (WinTherface).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alizadeh, Z.; Yazdi, J.; Najafi, M.S. Improving the outputs of regional heavy rainfall forecasting models using an adaptive real-time approach. *Hydrol. Sci. J.* **2022**, *67*, 550–563. [CrossRef]
2. Khan, M.T.; Shoaib, M.; Hammad, M.; Salahudin, H.; Ahmad, F.; Ahmad, S. Application of machine learning techniques in rainfall–runoff modelling of the soan river basin, Pakistan. *Water* **2021**, *13*, 3528. [CrossRef]
3. Barrera-Animas, A.Y.; Oyedele, L.O.; Bilal, M.; Akinosho, T.D.; Delgado, J.M.D.; Akanbi, L.A. Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. *Mach. Learn. Appl.* **2022**, *7*, 100204. [CrossRef]
4. Basha, C.Z.; Bhavana, N.; Bhavya, P.; Sowmya, V. Rainfall prediction using machine learning & deep learning techniques. In Proceedings of the 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2–4 July 2020; pp. 92–97.
5. Yang, T.-H.; Yang, S.-C.; Ho, J.-Y.; Lin, G.-F.; Hwang, G.-D.; Lee, C.-S. Flash flood warnings using the ensemble precipitation forecasting technique: A case study on forecasting floods in Taiwan caused by typhoons. *J. Hydrol.* **2015**, *520*, 367–378. [CrossRef]
6. Liu, J.; Wang, J.; Pan, S.; Tang, K.; Li, C.; Han, D. A real-time flood forecasting system with dual updating of the NWP rainfall and the river flow. *Nat. Hazards* **2015**, *77*, 1161–1182. [CrossRef]
7. Mosavi, A.; Ozturk, P.; Chau, K. Flood Prediction Using Machine Learning Models: Literature Review. *Water* **2018**, *10*, 1536. [CrossRef]
8. You, G.J.-Y.; Thum, B.-H.; Lin, F.-H. The examination of reproducibility in hydro-ecological characteristics by daily synthetic flow models. *J. Hydrol.* **2014**, *511*, 904–919. [CrossRef]
9. Le, T.-T.; Pham, B.T.; Ly, H.-B.; Shirzadi, A.; Le, L.M. Development of 48-hour precipitation forecasting model using nonlinear autoregressive neural network. In *Innovation for Sustainable Infrastructure*; Ha-Minh, C., van Dao, D., Benboudjema, F., Derrible, S., Huynh, D.V.K., Tang, A.M., Eds.; Lecture Notes in Civil Engineering; Springer: Singapore, 2020; Volume 54, pp. 1191–1196, ISBN 978-981-150-802-8.
10. Amin, I.; Kumar, R.; Jhajharia, D.; Sherring, A. Estimation and validation of runoff and sediment models for Dachigam watershed of Kashmir Valley. *Indian J. Soil Conserv.* **2015**, *43*, 9–14.
11. Kumar, R.; Manzoor, S.; Vishwakarma, D.K.; Al-Ansari, N.; Kushwaha, N.L.; Elbeltagi, A.; Sushanth, K.; Prasad, V.; Kuriqi, A. Assessment of Climate Change Impact on Snowmelt Runoff in Himalayan Region. *Sustainability* **2022**, *14*, 1150. [CrossRef]
12. Vishwakarma, D.K.; Kumar, R.; Pandey, K.; Singh, V.; Kushwaha, K.S. Modeling of Rainfall and Ground Water Fluctuation of Gonda District Uttar Pradesh, India. *Int. J. Curr. Microbiol. Appl. Sci.* **2018**, *7*, 2613–2618. [CrossRef]

13. Kumar, M.; Kumar, R.; Rajput, T.B.S.; Patel, N. Efficient Design of Drip Irrigation System using Water and Fertilizer Application Uniformity at Different Operating Pressures in a Semi-Arid Region of India. *Irrig. Drain.* **2017**, *66*, 316–326. [CrossRef]
14. Thomas, D.S.G.; Twyman, C.; Osbahr, H.; Hewitson, B. Adaptation to climate change and variability: Farmer responses to intra-seasonal precipitation trends in South Africa. *Clim. Chang.* **2007**, *83*, 301–322. [CrossRef]
15. Kramer, K.L.; Hackman, J. Scaling climate change to human behavior predicting good and bad years for Maya farmers. *Am. J. Hum. Biol.* **2021**, *33*, e23524. [CrossRef]
16. Zhao, Q.; Ma, X.; Liang, L.; Yao, W. Spatial–Temporal Variation Characteristics of Multiple Meteorological Variables and Vegetation over the Loess Plateau Region. *Appl. Sci.* **2020**, *10*, 1000. [CrossRef]
17. Turgut, M.S.; Turgut, O.E.; Afan, H.A.; El-Shafie, A. A novel Master–Slave optimization algorithm for generating an optimal release policy in case of reservoir operation. *J. Hydrol.* **2019**, *577*, 123959. [CrossRef]
18. Tikhamarine, Y.; Souag-Gamane, D.; Ahmed, A.N.; Sammen, S.S.; Kisi, O.; Huang, Y.F.; El-Shafie, A. Rainfall-runoff modelling using improved machine learning methods: Harris hawks optimizer vs. particle swarm optimization. *J. Hydrol.* **2020**, *589*, 125133. [CrossRef]
19. Banadkooki, F.B.; Ehteram, M.; Ahmed, A.N.; Fai, C.M.; Afan, H.A.; Ridwam, W.M.; Sefelnasr, A.; El-Shafie, A. Precipitation Forecasting Using Multilayer Neural Network and Support Vector Machine Optimization Based on Flow Regime Algorithm Taking into Account Uncertainties of Soft Computing Models. *Sustainability* **2019**, *11*, 6681. [CrossRef]
20. Tikhamarine, Y.; Malik, A.; Kumar, A.; Souag-Gamane, D.; Kisi, O. Estimation of monthly reference evapotranspiration using novel hybrid machine learning approaches. *Hydrol. Sci. J.* **2019**, *64*, 1824–1842. [CrossRef]
21. Chang, T.K.; Talei, A.; Alaghmand, S.; Ooi, M.P.-L. Choice of rainfall inputs for event-based rainfall-runoff modeling in a catchment with multiple rainfall stations using data-driven techniques. *J. Hydrol.* **2017**, *545*, 100–108. [CrossRef]
22. Tokar, A.S.; Johnson, P.A. Rainfall-runoff modeling using artificial neural networks. *J. Hydrol. Eng.* **1999**, *4*, 232–239. [CrossRef]
23. Al Sawaf, M.B.; Kawanisi, K.; Jlilati, M.N.; Xiao, C.; Bahreinimotlagh, M. Extent of detection of hidden relationships among different hydrological variables during floods using data-driven models. *Environ. Monit. Assess.* **2021**, *193*, 692. [CrossRef]
24. Peel, M.C.; McMahon, T.A. Historical development of rainfall-runoff modeling. *WIREs Water* **2020**, *7*, e1471. [CrossRef]
25. Moradkhani, H.; Sorooshian, S. General review of rainfall-runoff modeling: Model calibration, data assimilation, and uncertainty analysis. In *Hydrological Modelling and the Water Cycle*; Sorooshian, S., Hsu, K.-L., Coppola, E., Tomassetti, B., Verdecchia, M., Visconti, G., Eds.; Water Science and Technology Library; Springer: Berlin/Heidelberg, Germany, 2008; Volume 63, pp. 1–24, ISBN 978-354-077-843-1.
26. Daniell, T.M. Neural networks. Applications in hydrology and water resources engineering. In Proceedings of the National Conference Publication—Institute of Engineers, Perth, Australia; 1991.
27. French, M.N.; Krajewski, W.F.; Cuykendall, R.R. Rainfall forecasting in space and time using a neural network. *J. Hydrol.* **1992**, *137*, 1–31. [CrossRef]
28. Asadi, H.; Shahedi, K.; Jarihani, B.; Sidle, R.C. Rainfall-Runoff Modelling Using Hydrological Connectivity Index and Artificial Neural Network Approach. *Water* **2019**, *11*, 212. [CrossRef]
29. Dash, Y.; Mishra, S.K.; Panigrahi, B.K. Rainfall prediction for the Kerala state of India using artificial intelligence approaches. *Comput. Electr. Eng.* **2018**, *70*, 66–73. [CrossRef]
30. Chau, K.W.; Wu, C.L. A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. *J. Hydroinform.* **2010**, *12*, 458–473. [CrossRef]
31. Zounemat-kermani, M.; Kisi, O.; Rajaee, T. Performance of radial basis and LM-feed forward artificial neural networks for predicting daily watershed runoff. *Appl. Soft Comput.* **2013**, *13*, 4633–4644. [CrossRef]
32. Harshburger, B.J.; Walden, V.P.; Humes, K.S.; Moore, B.C.; Blandford, T.R.; Rango, A. Generation of Ensemble Streamflow Forecasts Using an Enhanced Version of the Snowmelt Runoff Model1. *JAWRA J. Am. Water Resour. Assoc.* **2012**, *48*, 643–655. [CrossRef]
33. Humphrey, G.B.; Gibbs, M.S.; Dandy, G.C.; Maier, H.R. A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network. *J. Hydrol.* **2016**, *540*, 623–640. [CrossRef]
34. Kisi, O.; Cimen, M. A wavelet-support vector machine conjunction model for monthly streamflow forecasting. *J. Hydrol.* **2011**, *399*, 132–140. [CrossRef]
35. Thapa, S.; Zhao, Z.; Li, B.; Lu, L.; Fu, D.; Shi, X.; Tang, B.; Qi, H. Snowmelt-Driven Streamflow Prediction Using Machine Learning Techniques (LSTM, NARX, GPR, and SVR). *Water* **2020**, *12*, 1734. [CrossRef]
36. Nourani, V.; Andalib, G. Wavelet based artificial intelligence approaches for prediction of hydrological time series. In *Artificial Life and Computational Intelligence. ACALCI 2015*; Chalup, S.K., Blair, A.D., Randall, M., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 8955. [CrossRef]
37. Idrees, M.B.; Jehanzaib, M.; Kim, D.; Kim, T.-W. Comprehensive evaluation of machine learning models for suspended sediment load inflow prediction in a reservoir. *Stoch. Environ. Res. Risk Assess.* **2021**, *35*, 1805–1823. [CrossRef]
38. Kakaei Lafdani, E.; Moghaddam Nia, A.; Ahmadi, A. Daily suspended sediment load prediction using artificial neural networks and support vector machines. *J. Hydrol.* **2013**, *478*, 50–62. [CrossRef]
39. Rajaee, T.; Mirbagheri, S.A.; Zounemat-Kermani, M.; Nourani, V. Daily suspended sediment concentration simulation using ANN and neuro-fuzzy models. *Sci. Total Environ.* **2009**, *407*, 4916–4927. [CrossRef] [PubMed]

40. Melesse, A.M.; Ahmad, S.; McClain, M.E.; Wang, X.; Lim, Y.H. Suspended sediment load prediction of river systems: An artificial neural network approach. *Agric. Water Manag.* **2011**, *98*, 855–866. [CrossRef]
41. Gupta, D.; Hazarika, B.B.; Berlin, M.; Sharma, U.M.; Mishra, K. Artificial intelligence for suspended sediment load prediction: A review. *Environ. Earth Sci.* **2021**, *80*, 346. [CrossRef]
42. Azamathulla, H.M.; Cuan, Y.C.; Ghani, A.A.; Chang, C.K. Suspended sediment load prediction of river systems: GEP approach. *Arab. J. Geosci.* **2013**, *6*, 3469–3480. [CrossRef]
43. Nguyen, D.T.; Chen, S.-T. Real-Time Probabilistic Flood Forecasting Using Multiple Machine Learning Methods. *Water* **2020**, *12*, 787. [CrossRef]
44. Al-Abadi, A.M. Modeling of stage–discharge relationship for Gharraf River, southern Iraq using backpropagation artificial neural networks, M5 decision trees, and Takagi–Sugeno inference system technique: A comparative study. *Appl. Water Sci.* **2016**, *6*, 407–420. [CrossRef]
45. Kisi, Ö.; Çobaner, M. Modeling River Stage-Discharge Relationships Using Different Neural Network Computing Techniques. *Clean Soil Air Water* **2009**, *37*, 160–169. [CrossRef]
46. Lohani, A.K.; Goel, N.K.; Bhatia, K.K.S. Takagi–Sugeno fuzzy inference system for modeling stage–discharge relationship. *J. Hydrol.* **2006**, *331*, 146–160. [CrossRef]
47. Shukla, R.; Kumar, P.; Vishwakarma, D.K.; Ali, R.; Kumar, R.; Kuriqi, A. Modeling of stage-discharge using back propagation ANN-, ANFIS-, and WANN-based computing techniques. *Theor. Appl. Climatol.* **2022**, *147*, 867–889. [CrossRef]
48. Ajmera, T.K.; Goyal, M.K. Development of stage–discharge rating curve using model tree and neural networks: An application to Peachtree Creek in Atlanta. *Expert Syst. Appl.* **2012**, *39*, 5702–5710. [CrossRef]
49. Araghi, A.; Mousavi-Baygi, M.; Adamowski, J.; Martinez, C.; van der Ploeg, M. Forecasting soil temperature based on surface air temperature using a wavelet artificial neural network. *Meteorol. Appl.* **2017**, *24*, 603–611. [CrossRef]
50. Feng, Y.; Cui, N.; Hao, W.; Gao, L.; Gong, D. Estimation of soil temperature from meteorological data using different machine learning models. *Geoderma* **2019**, *338*, 67–77. [CrossRef]
51. Bilgili, M. Prediction of soil temperature using regression and artificial neural network models. *Meteorol. Atmos. Phys.* **2010**, *110*, 59–70. [CrossRef]
52. Hariharan, G.; Kannan, K.; Sharma, K.R. Haar wavelet in estimating depth profile of soil temperature. *Appl. Math. Comput.* **2009**, *210*, 119–125. [CrossRef]
53. Singh, V.K.; Singh, B.P.; Kisi, O.; Kushwaha, D.P. Spatial and multi-depth temporal soil temperature assessment by assimilating satellite imagery, artificial intelligence and regression based models in arid area. *Comput. Electron. Agric.* **2018**, *150*, 205–219. [CrossRef]
54. Mehdizadeh, S.; Fathian, F.; Safari, M.J.S.; Khosravi, A. Developing novel hybrid models for estimation of daily soil temperature at various depths. *Soil Tillage Res.* **2020**, *197*, 104513. [CrossRef]
55. Wu, W.; Tang, X.-P.; Guo, N.-J.; Yang, C.; Liu, H.-B.; Shang, Y.-F. Spatiotemporal modeling of monthly soil temperature using artificial neural networks. *Theor. Appl. Climatol.* **2013**, *113*, 481–494. [CrossRef]
56. Seifi, A.; Ehteram, M.; Nayebloei, F.; Soroush, F.; Gharabaghi, B.; Torabi Haghighi, A. GLUE uncertainty analysis of hybrid models for predicting hourly soil temperature and application wavelet coherence analysis for correlation with meteorological variables. *Soft Comput.* **2021**, *25*, 10723–10748. [CrossRef]
57. Ali Ghorbani, M.; Kazempour, R.; Chau, K.-W.; Shamshirband, S.; Taherei Ghazvinei, P. Forecasting pan evaporation with an integrated artificial neural network quantum-behaved particle swarm optimization model: A case study in Talesh, Northern Iran. *Eng. Appl. Comput. Fluid Mech.* **2018**, *12*, 724–737. [CrossRef]
58. Terzi, Ö. Daily pan evaporation estimation using gene expression programming and adaptive neural-based fuzzy inference system. *Neural Comput. Appl.* **2013**, *23*, 1035–1044. [CrossRef]
59. Guven, A.; Kişi, Ö. Daily pan evaporation modeling using linear genetic programming technique. *Irrig. Sci.* **2011**, *29*, 135–145. [CrossRef]
60. Kushwaha, N.L.; Rajput, J.; Elbeltagi, A.; Elnaggar, A.Y.; Sena, D.R.; Vishwakarma, D.K.; Mani, I.; Hussein, E.E. Data Intelligence Model and Meta-Heuristic Algorithms-Based Pan Evaporation Modelling in Two Different Agro-Climatic Zones: A Case Study from Northern India. *Atmosphere* **2021**, *12*, 1654. [CrossRef]
61. Piri, J.; Amin, S.; Moghaddamnia, A.; Keshavarz, A.; Han, D.; Remesan, R. Daily Pan Evaporation Modeling in a Hot and Dry Climate. *J. Hydrol. Eng.* **2009**, *14*, 803–811. [CrossRef]
62. Shabani, S.; Samadianfard, S.; Sattari, M.T.; Shamshirband, S.; Mosavi, A.; Kmet, T.; Várkonyi-Kóczy, A.R. Modeling daily pan evaporation in humid climates using Gaussian Process Regression. *arXiv* **2019**, arXiv:1908.04267. [CrossRef]
63. Kim, S.; Shiri, J.; Singh, V.P.; Kisi, O.; Landeras, G. Predicting daily pan evaporation by soft computing models with limited climatic data. *Hydrol. Sci. J.* **2015**, *60*, 1120–1136. [CrossRef]
64. Keshtegar, B.; Piri, J.; Kisi, O. A nonlinear mathematical modeling of daily pan evaporation based on conjugate gradient method. *Comput. Electron. Agric.* **2016**, *127*, 120–130. [CrossRef]
65. Kumar, M.; Kumari, A.; Kumar, D.; Al-Ansari, N.; Ali, R.; Kumar, R.; Kumar, A.; Elbeltagi, A.; Kuriqi, A. The superiority of data-driven techniques for estimation of daily pan evaporation. *Atmosphere* **2021**, *12*, 701. [CrossRef]

66. Malik, A.; Tikhamarine, Y.; Al-Ansari, N.; Shahid, S.; Sekhon, H.S.; Pal, R.K.; Rai, P.; Pandey, K.; Singh, P.; Elbeltagi, A.; et al. Daily pan-evaporation estimation in different agro-climatic zones using novel hybrid support vector regression optimized by Salp swarm algorithm in conjunction with gamma test. *Eng. Appl. Comput. Fluid Mech.* **2021**, *15*, 1075–1094. [CrossRef]
67. Tabari, H.; Marofi, S.; Sabziparvar, A.-A. Estimation of daily pan evaporation using artificial neural network and multivariate non-linear regression. *Irrig. Sci.* **2010**, *28*, 399–406. [CrossRef]
68. Bhagwat, S.; Kashyap, P.S.; Singh, B.P.; Singh, V.K. Daily pan evaporation modeling in hilly region of Uttarakhand using artificial neural network. *Indian J. Ecol.* **2017**, *44*, 467–473.
69. Huang, G.; Wu, L.; Ma, X.; Zhang, W.; Fan, J.; Yu, X.; Zeng, W.; Zhou, H. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *J. Hydrol.* **2019**, *574*, 1029–1041. [CrossRef]
70. Nourani, V.; Elkiran, G.; Abdullahi, J. Multi-step ahead modeling of reference evapotranspiration using a multi-model approach. *J. Hydrol.* **2020**, *581*, 124434. [CrossRef]
71. Wen, X.; Si, J.; He, Z.; Wu, J.; Shao, H.; Yu, H. Support-Vector-Machine-Based Models for Modeling Daily Reference Evapotranspiration with Limited Climatic Data in Extreme Arid Regions. *Water Resour. Manag.* **2015**, *29*, 3195–3209. [CrossRef]
72. Mor, N.; Jhajharia, D. Time series modelling of monthly reference evapotranspiration for Bikaner, Rajasthan (India). *Indian J. Soil Conserv.* **2018**, *46*, 42–51.
73. Feng, Y.; Peng, Y.; Cui, N.; Gong, D.; Zhang, K. Modeling reference evapotranspiration using extreme learning machine and generalized regression neural network only with temperature data. *Comput. Electron. Agric.* **2017**, *136*, 71–78. [CrossRef]
74. Dong, L.; Zeng, W.; Wu, L.; Lei, G.; Chen, H.; Srivastava, A.K.; Gaiser, T. Estimating the Pan Evaporation in Northwest China by Coupling CatBoost with Bat Algorithm. *Water* **2021**, *13*, 256. [CrossRef]
75. Fan, J.; Yue, W.; Wu, L.; Zhang, F.; Cai, H.; Wang, X.; Lu, X.; Xiang, Y. Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. *Agric. For. Meteorol.* **2018**, *263*, 225–241. [CrossRef]
76. Kim, S.; Kim, H.S. Neural networks and genetic algorithm approach for nonlinear evaporation and evapotranspiration modeling. *J. Hydrol.* **2008**, *351*, 299–317. [CrossRef]
77. Elbeltagi, A.; Raza, A.; Hu, Y.; Al-Ansari, N.; Kushwaha, N.L.; Srivastava, A.; Kumar Vishwakarma, D.; Zubair, M. Data intelligence and hybrid metaheuristic algorithms-based estimation of reference evapotranspiration. *Appl. Water Sci.* **2022**, *12*, 152. [CrossRef]
78. Elbeltagi, A.; Kushwaha, N.L.; Rajput, J.; Vishwakarma, D.K.; Kulimushi, L.C.; Kumar, M.; Zhang, J.; Pande, C.B.; Choudhari, P.; Meshram, S.G.; et al. Modelling daily reference evapotranspiration based on stacking hybridization of ANN with meta-heuristic algorithms under diverse agro-climatic conditions. *Stoch. Environ. Res. Risk Assess.* **2022**. [CrossRef]
79. Singh, V.K.; Panda, K.C.; Sagar, A.; Al-Ansari, N.; Duan, H.-F.; Paramaguru, P.K.; Vishwakarma, D.K.; Kumar, A.; Kumar, D.; Kashyap, P.S.; et al. Novel Genetic Algorithm (GA) based hybrid machine learning-pedotransfer Function (ML-PTF) for prediction of spatial pattern of saturated hydraulic conductivity. *Eng. Appl. Comput. Fluid Mech.* **2022**, *16*, 1082–1099. [CrossRef]
80. Sihag, P.; Tiwari, N.K.; Ranjan, S. Modelling of infiltration of sandy soil using gaussian process regression. *Model. Earth Syst. Environ.* **2017**, *3*, 1091–1100. [CrossRef]
81. Sihag, P.; Tiwari, N.K.; Ranjan, S. Prediction of unsaturated hydraulic conductivity using adaptive neuro-fuzzy inference system (ANFIS). *ISH J. Hydraul. Eng.* **2019**, *25*, 132–142. [CrossRef]
82. Sihag, P.; Tiwari, N.K.; Ranjan, S. Estimation and inter-comparison of infiltration models. *Water Sci.* **2017**, *31*, 34–43. [CrossRef]
83. Singh, B.; Sihag, P.; Parsaie, A.; Angelaki, A. Comparative analysis of artificial intelligence techniques for the prediction of infiltration process. *Geol. Ecol. Landsc.* **2021**, *5*, 109–118. [CrossRef]
84. Sihag, P.; Singh, B.; Sepah Vand, A.; Mehdipour, V. Modeling the infiltration process with soft computing techniques. *ISH J. Hydraul. Eng.* **2020**, *26*, 138–152. [CrossRef]
85. Singh, B.; Sihag, P.; Pandhiani, S.M.; Debnath, S.; Gautam, S. Estimation of permeability of soil using easy measured soil parameters: Assessing the artificial intelligence-based models. *ISH J. Hydraul. Eng.* **2021**, *27*, 38–48. [CrossRef]
86. Sihag, P.; Kumar, M.; Singh, B. Assessment of infiltration models developed using soft computing techniques. *Geol. Ecol. Landsc.* **2021**, *5*, 241–251. [CrossRef]
87. Sihag, P.; Singh, V.P.; Angelaki, A.; Kumar, V.; Sepahvand, A.; Golia, E. Modelling of infiltration using artificial intelligence techniques in semi-arid Iran. *Hydrol. Sci. J.* **2019**, *64*, 1647–1658. [CrossRef]
88. Singh, V.K.; Kumar, D.; Kashyap, P.S.; Singh, P.K.; Kumar, A.; Singh, S.K. Modelling of soil permeability using different data driven algorithms based on physical properties of soil. *J. Hydrol.* **2020**, *580*, 124223. [CrossRef]
89. Elbeltagi, A.; Pande, C.B.; Kouadri, S.; Islam, A.R.M.T. Applications of various data-driven models for the prediction of groundwater quality index in the Akot basin, Maharashtra, India. *Environ. Sci. Pollut. Res.* **2022**, *29*, 17591–17605. [CrossRef]
90. Gholami, V.; Khaleghi, M.R.; Pirasteh, S.; Booij, M.J. Comparison of Self-Organizing Map, Artificial Neural Network, and Co-Active Neuro-Fuzzy Inference System Methods in Simulating Groundwater Quality: Geospatial Artificial Intelligence. *Water Resour. Manag.* **2022**, *36*, 451–469. [CrossRef]
91. El Bilali, A.; Taleb, A.; Brouziyne, Y. Groundwater quality forecasting using machine learning algorithms for irrigation purposes. *Agric. Water Manag.* **2021**, *245*, 106625. [CrossRef]
92. Singha, S.; Pasupuleti, S.; Singha, S.S.; Singh, R.; Kumar, S. Prediction of groundwater quality using efficient machine learning technique. *Chemosphere* **2021**, *276*, 130265. [CrossRef]

93. Kumar, A.; Singh, V.K.; Saran, B.; Al-Ansari, N.; Singh, V.P.; Adhikari, S.; Joshi, A.; Singh, N.K.; Vishwakarma, D.K. Development of Novel Hybrid Models for Prediction of Drought- and Stress-Tolerance Indices in Teosinte Introgressed Maize Lines Using Artificial Intelligence Techniques. *Sustainability* **2022**, *14*, 2287. [CrossRef]
94. Elbeltagi, A.; Azad, N.; Arshad, A.; Mohammed, S.; Mokhtar, A.; Pande, C.; Ramezani, H.; Ahmad, S.; Reza, A.; Islam, T.; et al. Applications of Gaussian process regression for predicting blue water footprint: Case study in Ad Daqahliyah, Egypt. *Agric. Water Manag.* **2021**, *255*, 107052. [CrossRef]
95. Elbeltagi, A.; Deng, J.; Wang, K.; Hong, Y. Crop Water footprint estimation and modeling using an artificial neural network approach in the Nile Delta, Egypt. *Agric. Water Manag.* **2020**, *235*, 106080. [CrossRef]
96. Babaee, M.; Maroufpoor, S.; Jalali, M.; Zarei, M.; Elbeltagi, A. Artificial intelligence approach to estimating rice yield*. *Irrig. Drain.* **2021**, *70*, 732–742. [CrossRef]
97. Elbeltagi, A.; Zhang, L.; Deng, J.; Juma, A.; Wang, K. Modeling monthly crop coefficients of maize based on limited meteorological data: A case study in Nile Delta, Egypt. *Comput. Electron. Agric.* **2020**, *173*, 105368. [CrossRef]
98. Kumar, S.; Roshni, T.; Himayoun, D. A Comparison of Emotional Neural Network (ENN) and Artificial Neural Network (ANN) Approach for Rainfall-Runoff Modelling. *Civ. Eng. J.* **2019**, *5*, 2120–2130. [CrossRef]
99. Abbot, J.; Marohasy, J. Application of artificial neural networks to rainfall forecasting in Queensland, Australia. *Adv. Atmos. Sci.* **2012**, *29*, 717–730. [CrossRef]
100. Shoaib, M.; Shamseldin, A.Y.; Melville, B.W.; Khan, M.M. A comparison between wavelet based static and dynamic neural network approaches for runoff prediction. *J. Hydrol.* **2016**, *535*, 211–225. [CrossRef]
101. Ghumman, A.R.; Ghazaw, Y.M.; Sohail, A.R.; Watanabe, K. Runoff forecasting by artificial neural network and conventional model. *Alex. Eng. J.* **2011**, *50*, 345–350. [CrossRef]
102. Nayak, P.C.; Sudheer, K.P.; Jain, S.K. Rainfall-runoff modeling through hybrid intelligent system. *Water Resour. Res.* **2007**, *43*, W07415. [CrossRef]
103. Sinharay, S. An Overview of statistics in education. In *International Encyclopedia of Education*, 3rd ed.; Peterson, P., Baker, E., McGaw, B., Eds.; Elsevier: Amsterdam, The Netherlands, 2010; pp. 1–11, ISBN 978-0-08-044894-7.
104. Vishwakarma, D.K.; Pandey, K.; Kaur, A.; Kushwaha, N.L.; Kumar, R.; Ali, R.; Elbeltagi, A.; Kuriqi, A. Methods to estimate evapotranspiration in humid and subtropical climate conditions. *Agric. Water Manag.* **2022**, *261*, 107378. [CrossRef]
105. Chenini, I.; Khemiri, S. Evaluation of ground water quality using multiple linear regression and structural equation modeling. *Int. J. Environ. Sci. Technol.* **2009**, *6*, 509–519. [CrossRef]
106. Snedecor, G.W.; Cochran, W.G.; Fuller, J.A.R. *Métodos Estadísticos*; Continental: México City, Mexico, 1971.
107. Sekhar Roy, S.; Roy, R.; Balas, V.E. Estimating heating load in buildings using multivariate adaptive regression splines, extreme learning machine, a hybrid model of MARS and ELM. *Renew. Sustain. Energy Rev.* **2018**, *82*, 4256–4268. [CrossRef]
108. Akin, M.; Eyduran, S.P.; Eyduran, E.; Reed, B.M. Analysis of macro nutrient related growth responses using multivariate adaptive regression splines. *Plant Cell Tissue Organ Cult.* **2020**, *140*, 661–670. [CrossRef]
109. Mirabbasi, R.; Kisi, O.; Sanikhani, H.; Gajbhiye Meshram, S. Monthly long-term rainfall estimation in Central India using M5Tree, MARS, LSSVR, ANN and GEP models. *Neural Comput. Appl.* **2019**, *31*, 6843–6862. [CrossRef]
110. Zhang, J.; Zhang, H.; Xiao, H.; Fang, H.; Han, Y.; Yu, L. Effects of rainfall and runoff-yield conditions on runoff. *Ain Shams Eng. J.* **2021**, *12*, 2111–2116. [CrossRef]
111. Vapnik, V. *Statistical Learning Theory*; John Wiley & Sons, Inc.: Oxford, UK, 1998; Volume 1.
112. Li, M.; Zhang, Y.; Wallace, J.; Campbell, E. Estimating annual runoff in response to forest change: A statistical method based on random forest. *J. Hydrol.* **2020**, *589*, 125168. [CrossRef]
113. Abdulelah Al-Sudani, Z.; Salih, S.Q.; Sharafati, A.; Yaseen, Z.M. Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation. *J. Hydrol.* **2019**, *573*, 1–12. [CrossRef]
114. Adnan, R.M.; Petroselli, A.; Heddam, S.; Santos, C.A.G.; Kisi, O. Comparison of different methodologies for rainfall–runoff modeling: Machine learning vs conceptual approach. *Nat. Hazards* **2021**, *105*, 2987–3011. [CrossRef]
115. Li, X.; Sha, J.; Wang, Z.-L. Comparison of daily streamflow forecasts using extreme learning machines and the random forest method. *Hydrol. Sci. J.* **2019**, *64*, 1857–1866. [CrossRef]
116. Goyal, M.K.; Bharti, B.; Quilty, J.; Adamowski, J.; Pandey, A. Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS. *Expert Syst. Appl.* **2014**, *41*, 5267–5276. [CrossRef]
117. Malik, A.; Kumar, A.; Piri, J. Daily suspended sediment concentration simulation using hydrological data of Pranhita River Basin, India. *Comput. Electron. Agric.* **2017**, *138*, 20–28. [CrossRef]
118. Malik, A.; Kumar, A.; Kim, S.; Kashani, M.H.; Karimi, V.; Sharafati, A.; Ghorbani, M.A.; Al-Ansari, N.; Salih, S.Q.; Yaseen, Z.M.; et al. Modeling monthly pan evaporation process over the Indian central Himalayas: Application of multiple learning artificial intelligence model. *Eng. Appl. Comput. Fluid Mech.* **2020**, *14*, 323–338. [CrossRef]
119. Singh, A.; Malik, A.; Kumar, A.; Kisi, O. Rainfall-runoff modeling in hilly watershed using heuristic approaches with gamma test. *Arab. J. Geosci.* **2018**, *11*, 261. [CrossRef]
120. Stefánsson, A.; Končar, N.; Jones, A.J. A note on the Gamma test. *Neural Comput. Appl.* **1997**, *5*, 131–133. [CrossRef]
121. Noori, R.; Karbassi, A.R.; Moghaddamnia, A.; Han, D.; Zokaei-Ashtiani, M.H.; Farokhnia, A.; Gousheh, M.G. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. *J. Hydrol.* **2011**, *401*, 177–189. [CrossRef]

122. Singh, V.K.; Kumar, D.; Kashyap, P.S.; Kisi, O. Simulation of suspended sediment based on gamma test, heuristic, and regression-based techniques. *Environ. Earth Sci.* **2018**, *77*, 708. [CrossRef]
123. Singh, V.K.; Kumar, D.; Kashyap, P.S.; Singh, P.K. Predicting unsaturated hydraulic conductivity of soil based on machine learning algorithms. In Proceedings of the International Conference on Opportunities and Challenges in Engineering, Management and Science (OCEMS—2019), Bareilly, India, 15–16 February 2019.
124. Zhang, W.; Goh, A.T.C.; Zhang, Y.; Chen, Y.; Xiao, Y. Assessment of soil liquefaction based on capacity energy concept and multivariate adaptive regression splines. *Eng. Geol.* **2015**, *188*, 29–37. [CrossRef]
125. Friedman, J.H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1–67. [CrossRef]
126. Kisi, O.; Parmar, K.S. Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. *J. Hydrol.* **2016**, *534*, 104–112. [CrossRef]
127. Zhang, X. *Matrix Analysis and Applications*; Cambridge University Press: Cambridge, UK, 2017; ISBN 110-841-741-8.
128. Rezaie-Balf, M.; Zahmatkesh, Z.; Kim, S. Soft Computing Techniques for Rainfall-Runoff Simulation: Local Non–Parametric Paradigm vs. Model Classification Methods. *Water Resour. Manag.* **2017**, *31*, 3843–3865. [CrossRef]
129. Adnan, R.M.; Liang, Z.; Trajkovic, S.; Zounemat-Kermani, M.; Li, B.; Kisi, O. Daily streamflow prediction using optimally pruned extreme learning machine. *J. Hydrol.* **2019**, *577*, 123981. [CrossRef]
130. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1999; ISBN 038-798-780-0.
131. Bray, M.; Han, D. Identification of support vector machines for runoff modelling. *J. Hydroinform.* **2004**, *6*, 265–280. [CrossRef]
132. Awad, M.; Khanna, R. Support vector regression. In *Efficient Learning Machines*; Awad, M., Khanna, R., Eds.; Apress: Berkeley, CA, USA, 2015; pp. 67–80, ISBN 978-143-025-990-9.
133. Kumar, M.; Kumari, A.; Kushwaha, D.P.; Kumar, P.; Malik, A.; Ali, R.; Kuriqi, A. Estimation of Daily Stage–Discharge Relationship by Using Data-Driven Techniques of a Perennial River, India. *Sustainability* **2020**, *12*, 7877. [CrossRef]
134. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
135. Prasad, A.M.; Iverson, L.R.; Liaw, A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* **2006**, *9*, 181–199. [CrossRef]
136. Sadler, J.M.; Goodall, J.L.; Morsy, M.M.; Spencer, K. Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest. *J. Hydrol.* **2018**, *559*, 43–55. [CrossRef]
137. Malik, A.; Kumar, A. Pan Evaporation Simulation Based on Daily Meteorological Data Using Soft Computing Techniques and Multiple Linear Regression. *Water Resour. Manag.* **2015**, *29*, 1859–1872. [CrossRef]
138. Kumar, R.; Kumar, A.; Shankhwar, A.K.; Vishkarma, D.K.; Sachan, A.; Singh, P.V.; Jahangeer, J.; Verma, A.; Kumar, V. Modelling of meteorological drought in the foothills of Central Himalayas: A case study in Uttarakhand State, India. *Ain Shams Eng. J.* **2022**, *13*, 101595. [CrossRef]
139. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. [CrossRef]
140. Malik, A.; Kumar, A.; Kisi, O. Daily pan evaporation estimation using heuristic methods with gamma test. *J. Irrig. Drain. Eng.* **2018**, *144*, 4018023. [CrossRef]
141. Nury, A.H.; Hasan, K.; Alam, M.J. Bin Comparative study of wavelet-ARIMA and wavelet-ANN models for temperature time series data in northeastern Bangladesh. *J. King Saud Univ. Sci.* **2017**, *29*, 47–61. [CrossRef]
142. Wooldridge, M. *An Introduction to Multiagent Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2009; ISBN 047-051-946-0.
143. Schroeder, R.; van de Ven, A.; Scudder, G.; Polley, D. Managing innovation and change processes: Findings from the Minnesota innovation research program. *Agribusiness* **1986**, *2*, 501–523. [CrossRef]
144. Santhi, C.; Arnold, J.G.; Williams, J.R.; Dugas, W.A.; Srinivasan, R.; Hauck, L.M. Validation of the swat model on a large rwer basin with point and nonpoint sources. *JAWRA J. Am. Water Resour. Assoc.* **2001**, *37*, 1169–1188. [CrossRef]
145. Van Liew, M.W.; Arnold, J.G.; Garbrecht, J.D. Hydrologic simulation on agricultural watersheds: Choosing between two models. *Trans. ASAE* **2003**, *46*, 1539. [CrossRef]
146. Legates, D.R.; McCabe, G.J., Jr. Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **1999**, *35*, 233–241. [CrossRef]
147. Gupta, H.V.; Soroosh, S.; Yapo, P.O. Status of Automatic Calibration for Hydrologic Models: Comparison with Multilevel Expert Calibration. *J. Hydrol. Eng.* **1999**, *4*, 135–143. [CrossRef]
148. Moriasi, D.N.; Arnold, J.G.; van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **2007**, *50*, 885–900. [CrossRef]
149. Kheirfam, H.; Mokarram-Kashtiban, S. A regional suspended load yield estimation model for ungauged watersheds. *Water Sci. Eng.* **2018**, *11*, 328–337. [CrossRef]
150. Naghibi, S.A.; Ahmadi, K.; Daneshi, A. Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping. *Water Resour. Manag.* **2017**, *31*, 2761–2775. [CrossRef]
151. Panda, K.C.; Singh, R.M.; Thakural, L.N.; Sahoo, D.P. Representative grid location-multivariate adaptive regression spline (RGL-MARS) algorithm for downscaling dry and wet season rainfall. *J. Hydrol.* **2022**, *605*, 127381. [CrossRef]
152. Zhang, J.; Ma, G.; Huang, Y.; Sun, J.; Aslani, F.; Nener, B. Modelling uniaxial compressive strength of lightweight self-compacting concrete using random forest regression. *Constr. Build. Mater.* **2019**, *210*, 713–719. [CrossRef]

153. Reis, I.; Baron, D.; Shahaf, S. Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets. *Astron. J.* **2018**, *157*, 16. [CrossRef]
154. Kim, S.; Jeong, M.; Ko, B.C. Lightweight surrogate random forest support for model simplification and feature relevance. *Appl. Intell.* **2022**, *52*, 471–481. [CrossRef]