Hindawi
Advances in Meteorology
Volume 2022, Article ID 1433835, 13 pages
https://doi.org/10.1155/2022/1433835

Research Article

Evaporation Rate Prediction Using Advanced Machine Learning Models: A Comparative Study

Zainab Abdulelah Al Sudani[1] **and Golam Saleh Ahmed Salem**[2]

[1]*Water Resources Department, College of Engineering, University of Baghdad, Baghdad, Iraq*
[2]*Department of Electrical and Electronic Engineering, Trust University, Nobogram Road, Barishal-8200, Bangladesh*

Correspondence should be addressed to Golam Saleh Ahmed Salem; dr.salem@trustuniversity.edu.bd

Received 14 November 2021; Revised 17 January 2022; Accepted 20 January 2022; Published 21 February 2022

Academic Editor: Upaka Rathnayake

Accurately estimating the amount of evaporation loss is necessary for scheduling and calculating irrigation water requirements. In this study, four machine learning (ML) modeling approaches, extreme learning machine (ELM), gradient boosting machine (GBM), quantile random forest (QRF), and Gaussian process regression (GPR), have been developed to estimate the monthly evaporation loss over two stations located in Iraq. Monthly climatical parameters have been used as an input variable for simulating the evaporation rate. Several statistical measures (e.g., mean absolute error (MAE), correlation coefficient (R), mean absolute percentage error (MAPE), and modified index of agreement (Md)), as well as graphical inspection, were used to compare the performances of the applied models. The results showed that the GBM model has much better performance in predicting monthly evaporation over two stations compared to other applied models. For the first case study which was in Diyala, the results showed a prediction enhancement in terms of MAE and RMSE by 7.17%, 21.01%; 16.51%, 15.74%; and 23.14%, 26.64%; using GBM compared to ELM, GPR, and QRF, respectively. However, for the second case study (in Erbil), the prediction enhancement was improved in terms of reduction of MAE and RMSE by 10.88%, 9.24%; 15.24%, 5%; and 16.06%, 15.76%; respectively, compared to ELM, GPR, and QRF models. The results of the proposed GMBM model can therefore assist local stakeholders in the management of water resources.

1. Introduction

In the hydrological cycle, evaporation plays a major role; therefore, monitoring evaporation is important for managing water resources, optimizing irrigation schedules, and modeling agricultural production [1,2]. Besides, evaporation rate has significant importance in studying climate change and global warming because this parameter dissipates a good proportion of the global precipitation [3–5]. The evaporation loss is influenced primarily by the vapor pressure gradient and the available heat energy, which are determined by the weather data like air temperature, relative humidity, wind speed, and solar radiation [6–8]. These variables are strongly associated with other aspects like the current season, time of day, geographical location, and sort of climate [9,10]. The evaporation process is therefore extremely nonlinear and complex.

For computing and evaluating evaporation, there are two procedures, direct and indirect [11]. Pan evaporation E_{pan} is considered as a well-known direct method used extensively for the estimation of evaporation rate. In particular, evaporimeters cannot be placed everywhere, especially in inaccessible regions where precise instrumentation is not possible [12]. Furthermore, the process of installing and maintaining this evaporation equipment in several regions is expensive [13]. However, the indirect method includes empirical equations used for measuring the evaporation rate [14]. These empirical equations can be established utilizing meteorological and hydrological parameters such as temperature, sunshine hour, wind speed, humidity, and rainfall [15,16]. Precise measurement of some of these meteorological factors requires advanced tools and skilled labor [17]. Often, instrument malfunctions, improper maintenance, and harsh weather conditions make it difficult to gauge these

data minus any errors, which is essential for the prediction of evaporation via empirical equations [18]. Thus, it would be problematic to project evaporation by gauging these factors incorrectly [19].

Thus, indirect systems of estimating evaporation by applying empirical equations are dependent on data and are also influenced by different assumptions. In other words, these approaches are considered as data-sensitive procedures and the accuracy of prediction would mainly depend on the data validity [20]. Additionally, such climatic data are generally scarce or hard to find at a particular hydrological station, and they tend to be discontinuous in certain places [21]. Evaporation is difficult to model through empirical techniques due to its extremely complex physical and nonlinear nature. In addition, an empirical model designed for a specific scenario might not perform well in another scenario, requiring recalibrations of the coefficients before execution. Several empirical models have been created by many researchers in literature to model evaporation loss [22]. The selection of the predictors is one of the main challenges for the nonlinear regression process. Therefore, creating a robust predictive model using empirical procedures is very difficult.

Many studies have been conducted to solve different water-resource problems employing different artificial intelligence (AI) approaches such as random forest (RF), support vector machine (SVM), extreme learning machine (ELM), feed-forward neural network (FFNN), extra-tree, Gaussian process regression (GPR), gradient boosting model (GBM), and quantile regression forest (QRF) [23–29]. Goyal et al. [30]. presented a study to estimate the daily evaporation loss over subtropical areas using different AI modeling approaches. The study used six meteorological parameters to establish the applied models. The findings of the study illustrated that the Adaptive Neurofuzzy Inference System (ANFIS) and least square support vector regression (LS-SVR) provide the best accuracy compared to the other used models. Another study was performed in [31] to estimate the evaporation loss of the Beysehir lake located in the southern part of Turkey. This study employed several machine learning approaches coupled with cross-validation technique to predict the monthly evaporation over that case study which is characterized as an arid and semiarid area. The study found that both ANN and SVR had a good prediction accuracy. Qasem et al. [32] developed a complicated model based on the incorporation of the ML models such as SVR and ANN with wavelet transforms (WT) for modeling the monthly rate of evaporation in arid and humid climates. The obtained results showed that the WT did not significantly enhance the prediction accuracy in some cases. Besides, the standard model (ANN) showed satisfactory accuracy in terms of predicting the evaporation rates. As ANN showed higher performance in prediction evaporation loss, it is significant to compare ANN with other machine learning methods such as RF and ELM. A study introduced by [33] provided a good comparison between the performances of ANN and random forest in the prediction of evaporation. The study's result proved that the RF has better performance than ANN as well as providing very accurate estimates. Furthermore, Althoff et al. [34] presented a study using different ML approaches to estimate the small dams' evaporation loss in Brazil. The findings of the study illustrated that the performance of RF was very satisfactory in the prediction of evaporation loss over small dams. Several other research evidenced the contribution of the AI models in simulating the catchment evaporation processes [35–37]. Recently, kernel-based models, fuzzy algorithms, and their hybrids with other algorithms have been successfully used for predicting evaporation [38]. However, developed gradient boosting models were rarely applied in modeling reference evapotranspiration worldwide. According to our knowledge, no study has focused on evaluating and comparing the capability of newly developed gradient boosting models for evaporation estimation in arid to semiarid climate zones of Iraq. Therefore, it is interesting to evaluate the performance of GBM and compare it with reliable AI models such as extreme learning machine (ELM), quantile regression forest (QRF), and Gaussian process regression (GPR) for estimating evaporation rate (E_p) in arid to semiarid climate zones of Iraq.

The contribution of this study is to determine the efficiency of the gradient boosting model (GBM) in estimating the evaporation rate (E_p) using data collected from two meteorological stations located in Iraq. The performance of GBM was compared with those of reliable AI models such as extreme learning machine (ELM), quantile regression forest (QRF), and Gaussian process regression (GPR). Furthermore, it is the first time to use GBM model for predicting the monthly evaporation loss related to several stations located in Iraq.

2. Data and Case Study

Iraq is geographically located in the Middle East and has almost two major climate zones, semiarid in the south and semihumid in the north [39]. The Iraqi region lacks sufficient water resources and suffers from droughts [40,41]. As temperatures rise in Iraq, surface water availability decreases, and groundwater levels in aquifers decrease. Iraq's hydrological cycle has been affected severely by evaporation, which currently depletes about 61% of its total precipitation [16,42]. Thus, it is very important to accurately predict the evaporation loss in Iraq. In this study, two case studies are selected to estimate the evaporation rate. The first case study is in Diyala state, while the second station is in Erbil state (see Figure 1). Diyala is located in the central part of the region, while Erbil is located in the northern region. The evaporation rate was predicted as function of six metrological parameters such as sunshine hours, minimum and maximum temperature, wind speed, rainfall, and relative humidity.

3. Methodology

3.1. Gaussian Process Regression. Rasmussen and Williams were the first to introduce the Gaussian process regression (GPR) [43]. This approach is a well-known and nonparametric method used for solving classification and regression

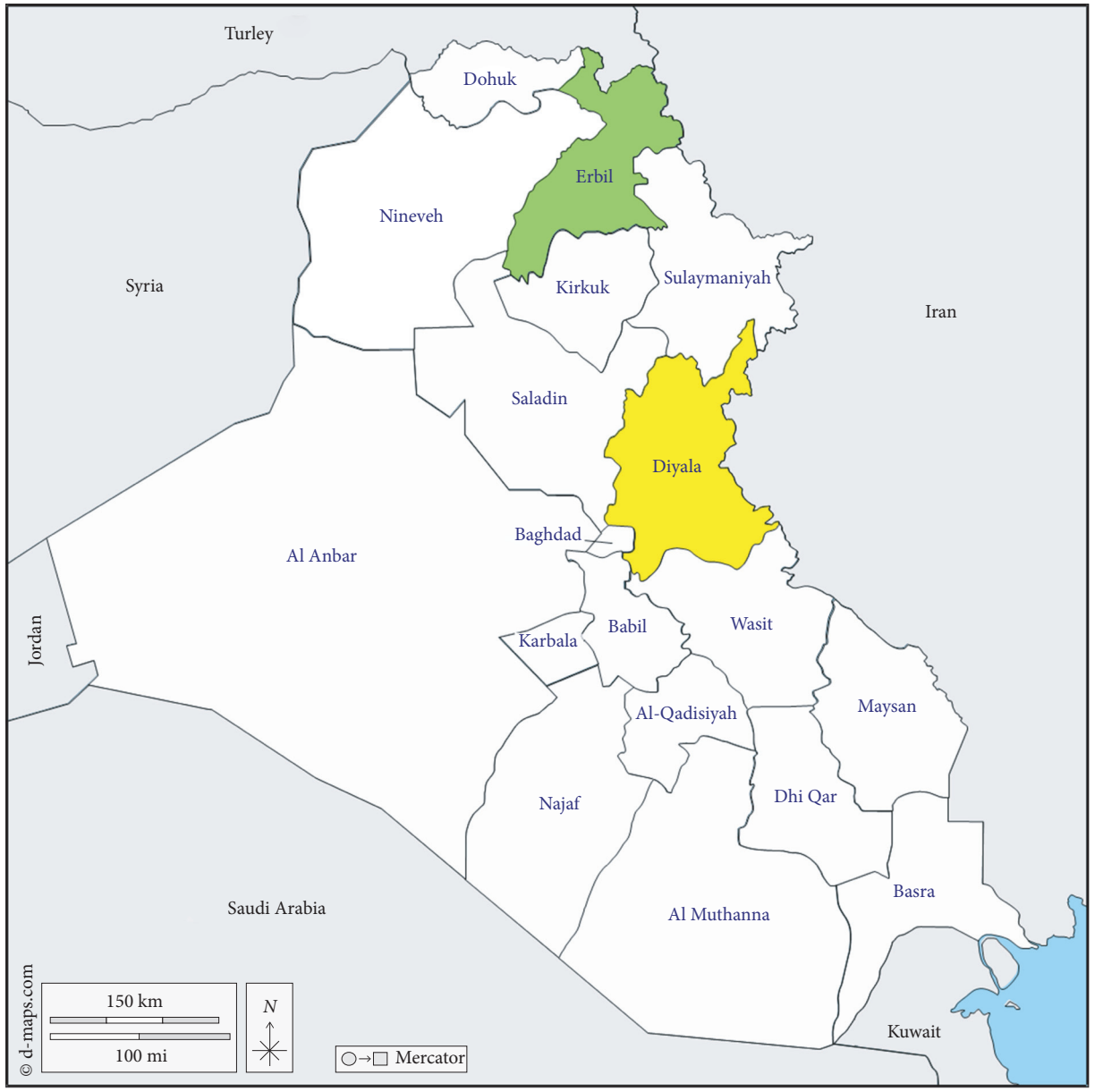

Figure 1: The locations of the selected region and the states.

problems. Furthermore, GPR model has been commonly employed to address several water resources concerts [44–47]. GPR combines Bayesian learning and kernel machines to form a principled and probabilistic approach to create a regression model. A model prediction's uncertainty can be directly outputted alongside the projected value [48].

In general, the mean and kernel function can be used to calculate a GPR [49]. According to this definition, GPR is an assemblage of random variables representing the value of function $f(t)$ at the given location (t). It can be expressed as follows:

$$\begin{aligned} f(t) &\sim GPR(m(t), k(t, t')), \\ m(t) &= E[f(t)], \\ k(t, t') &= E[(f(t) - m(t))(f(t') - m(t'))]. \end{aligned} \tag{1}$$

$f(t)$ is the prior distribution of the regression function, and $k(t)$, *an d* $m(t)$ are the kernel and function, respectively. By considering that the training set T includes input finite numbers in a matrix form $t_1, t_2, \ldots t_n$, the joint distribution of GPR is defined as follows:

$$p(f|TR) = N(f|M, K), \tag{2}$$

where $M(T)$ is the mean function which can be calculated by the mean function $m(t)$ as follows:

$$M(T) = \begin{bmatrix} m(t_1) \\ m(t_2) \\ \cdots \\ m(t_n) \end{bmatrix}. \tag{3}$$

Moreover, the kernel function $K(T,T)$ of the applied model can be determined by mean function $k(t,t')$ as follows:

$$K(T,T) = \begin{bmatrix} k(t_1, t_1') \dots (t_1, t_N') \\ \dots\dots \\ k(t_N, t_1') \dots (t_N, t_N') \end{bmatrix}. \quad (4)$$

In this study, the mean function is set to zero for simplicity to produce a widely used GPR prior. Besides, this technique has been widely used in previous studies [43,50]. Finally, (1) will be rewritten as follows:

$$f(t) \sim GPR(m(t) = 0, k(t,t')). \quad (5)$$

3.2. Extreme Learning Machine. Extreme learning machine (ELM) has the advantages of being a single hidden layer feedforward neural network (FFNN) with good global search ability, simple structure, fast learning speed, and excellent generalization abilities [51]. There are two types of weights in the ELM: the input weights related to the hidden layer which are assigned randomly and the output weights which are attained by analysis and calculation [52]. In other words, unlike traditional neural networks, the ELM does not require iterative learning [53]. The outputs weights of the ELM can be easily computed by determining the generalized inverse of the output matrix of the hidden output weight values. The structure of the ELM is greatly simplified by this process. The training process of ELM is summarized by few steps as follows:.

(i) Input the training dataset, and select the ELM's structure (hidden nodes) and the activation function of the hidden layer (see Figure 2).

(ii) Calculate the H matrix (output of hidden layer) as follows:

$$H = (a_1, \dots, a_{l;}\, b_1, \dots, b_{l;}\, x_1, \dots, x_n) = \begin{bmatrix} g(a_1, b_1, x_1) g(a_l, b_l, x_n) \\ \dots \\ g(a_1, b_1, x_n)\, g(a_l, b_l, x_n) \end{bmatrix}_{l*n}. \quad (6)$$

(a_i, b_i), $i = 1, 2, \dots, l$, are hidden nodes parameters which are randomly assigned.

(iii) Determine the output weight matrix (β):

$$\beta = H^{+}T, \quad (7)$$

where T is the actual label vector of the training dataset and H^{+} is Moore-Penrose generalized inverse matrix (H). $H^{+} = (H^{T}H)^{-1}H^{T}$.

$$\beta = \begin{bmatrix} \beta_1^{\,T} \\ \cdot \\ \cdot \\ \beta_l^{\,T} \end{bmatrix}_{l*m}, \qquad T = \begin{bmatrix} T_1^{\,T} \\ \cdot \\ \cdot \\ T_n^{\,T} \end{bmatrix}_{n*m}. \quad (8)$$

3.3. Quantile Random Forest (QRF). Random forest (RF) is an ensemble and supervised learning algorithm invented by Breiman [54]. The core concept of this approach is to integrate multiple trees through ensemble learning procedures. Furthermore, RF is a modified version of the Bagging algorithm with the basic idea that, for the original dataset, S_n are selected as a new data and S_n would be trained by using put back sampling method separately. The CART decision tree in RF is employed as a weak learner; however, for each tree is generated, the required number of features will be selected randomly from the original dataset labels. Thus, in a regression problem, the results of weak learners (T) are averaged to obtain the final model output. Averaging approach of RF has a significant importance in reducing the bias, as well as variance and correlation between trees [23].

The quantile random forest (QRF) is considered as an improved version of RF, applying quantile regression (QR) instead of averaging approach in calculating the final form of a target [55]. Furthermore, the QRF is considered a nonparametric approach enhanced by a solid theoretical foundation [56]. The conditional distribution of the QRF can be mathematically expressed as follows:

$$F(y|X = x) = P(Y \le y|X = x)\, E\left(l_{\{y_{i\le y}\}|}\, X = x\right). \quad (9)$$

$E(l_{\{y_{i\le y}\}|}\, X = x)$ in (8) is derived by taking mean value of the observations. With regard to QRF, $E(l_{\{y_{i\le y}\}|}\, X = x)$ is representing the weighted average value of all observations $l_{\{y_{i\le y}\}|}$.

$$\widehat{F}(y \mid X = x) = \sum_{i=1}^{n} W_i(x) l_{\{y_{i\le y}\}|}. \quad (10)$$

The steps below illustrate the QRF algorithm:

(i) The M decision tree $T(\theta_t)$, $t = 1, \dots k$, is created in random forests (RF) as well as taking into account the observations of each node related to a decision tree.

(ii) For $X = x$, o it will be repeated for all decision trees and then determine all observations of each decision tree. Finally, the weight $w_i(x, \theta_t)$ of each observation $i \in \{1, \dots, n\}$ is calculated by averaging the weights of tree decisions.

(iii) For all $y \in R$, calculate the estimate of the distribution function with (9) by using the weights obtained in step (2).

Figure 3 presents the flowchart of the QRF model.

3.4. Gradient Boosting Machine. Gradient boosting machine (GBM) model is one of the most famous supervised algorithms introduced as a robust technique to solve problems related to classification and regression [57]. Decision tree is a faster algorithm but it still suffers instability, so GBM is introduced to solve this serious problem [58–60]. Furthermore, GBM has combined the decision tress and boosting algorithms' advantages [61]. The GBM works mainly on the formulation of the gradient descent of

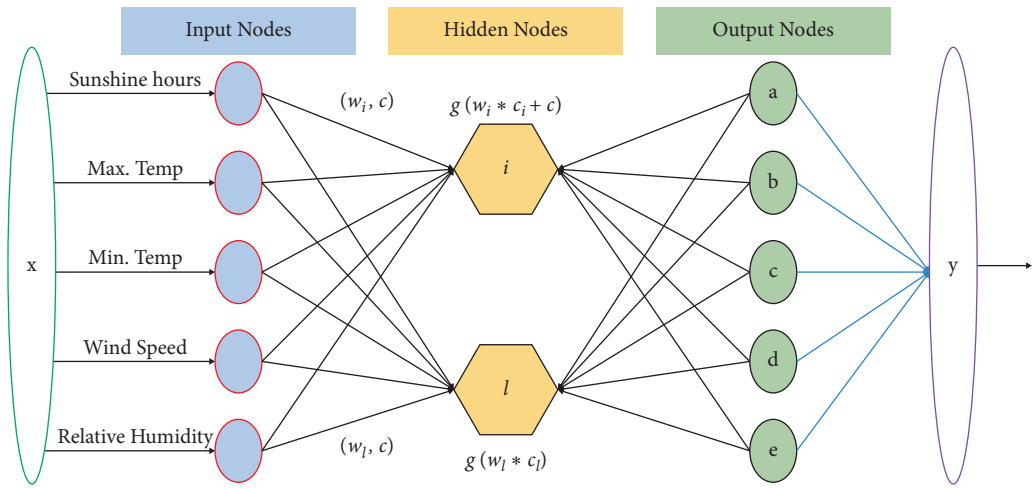

Figure 2: The basic structure of ELM model.

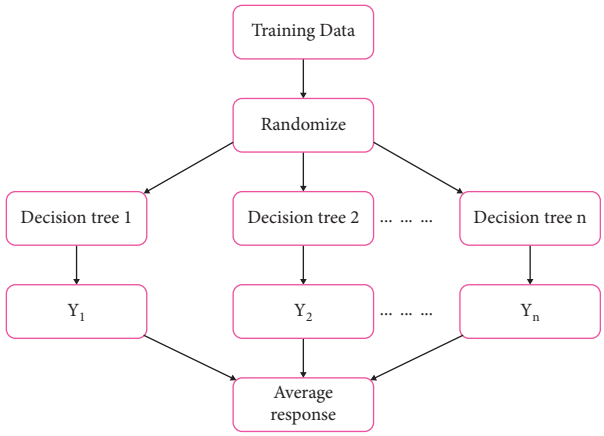

Figure 3: The flowchart of the quantile regression forest algorithm.

boosting technique and, hence, it is very useful for classification and regression problems [62]. The boosting structure is primarily a constructive scheme of ensemble formation that involves successively adding new weak base models that are trained according to the calculated error of the previous whole ensemble model for each iteration, and these base learners generate only a slightly lower error rate compared to random guessing. The boosting method family is based on a constructive strategy in which the learning mechanism will fit new models sequentially to produce a more precise estimation of the response variable. Figure 4 shows the structure of gradient boosting machine regression model.

The approach of the GBM model can be illustrated in several steps as follows:

(i) The GMB is initialized to minimize the loss function with a constant value.

(ii) The negative gradient of the cost function is estimated in each iterative training process as the residual value in x_i model (current one).

(iii) A new regression tree will be trained to fit the residual obtained from the second step.

(iv) In this step, the residual is updated and the current regression tree is added to the previous model.

(v) The algorithm of GBM is still iteratively trained and the maximum iterations number (selected by the user) is reached.

The mathematical expressions and brief description of applied GMB algorithm are shown below [63].

3.5. Statistical Evaluation Metrics. The four applied models have been compared and assessed to select the best models for predicting monthly evaporation. There are five statistical criteria, root mean square error (RMSE), mean absolute error (MAE), correlation coefficient (R), mean absolute percentage error (MAPE), and modified index of agreement (Md), which were used to assess the models' performances for training and testing phases. The mathematical expressions of these parameters are illustrated below [64]:

$$\begin{aligned}
\mathrm{MAE} &= \sum_{i=1}^{n} \frac{\left|\mathrm{EP}_{obi} - \mathrm{EP}_{smi}\right|}{n}, \\
\mathrm{RMSE} &= \sqrt{\sum_{i=1}^{n} \frac{\left(\mathrm{EP}_{obi} - \mathrm{EP}_{smi}\right)^2}{n}}, \\
\mathrm{MAPE} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\left|\mathrm{EP}_{obi} - \mathrm{EP}_{smi}\right|}{\mathrm{EP}_{obi}}, \\
R &= \frac{\sum_{i=1}^{n} \left(\mathrm{EP}_{obi} - \overline{\mathrm{EP}_{ob}}\right)\left(\mathrm{EP}_{smi} - \overline{\mathrm{EP}_{sim}}\right)}{\sum_{i=1}^{n} \left(\mathrm{EP}_{obi} - \overline{\mathrm{EP}_{ob}}\right)^2 \sum_{i=1}^{n} \left(\mathrm{EP}_{smi} - \overline{\mathrm{EP}_{sim}}\right)}.
\end{aligned} \tag{11}$$

In the above equations, EP_{obi} and EP_{smi} are the actual and predictive monthly evaporation values at i – th record, respectively. $\overline{\mathrm{EP}_{ob}}$ and $\overline{\mathrm{EP}_{sim}}$ are the mean observed and predicted monthly evaporation values and n is the number of records Algorithm 1.

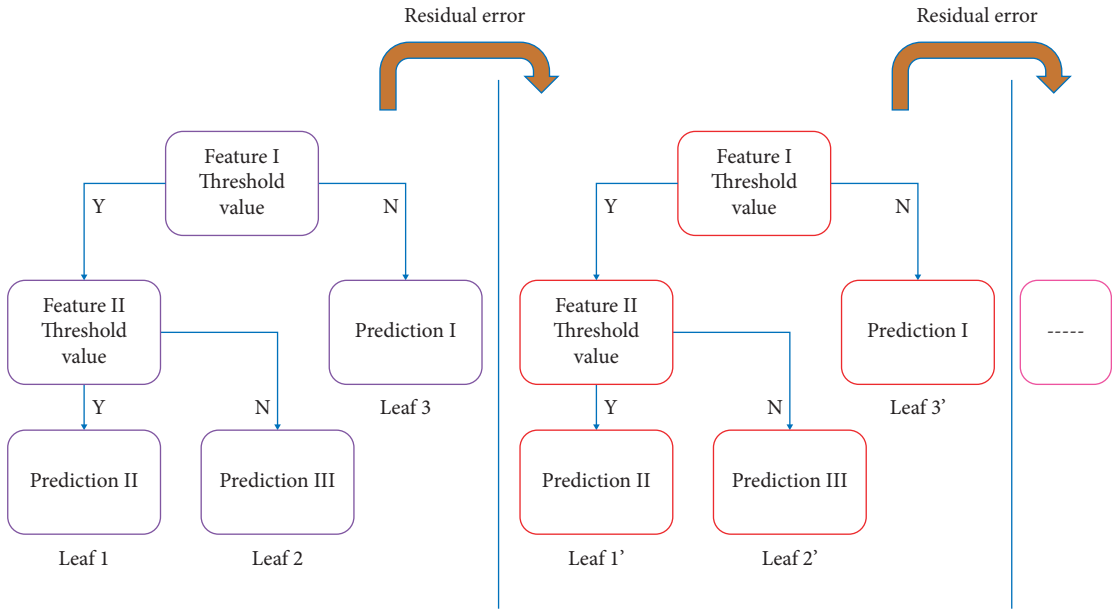

Figure 4: The structure of gradient boosting machine regression model.

Input: Train data
Data includes $T = \{(x_1, y_1), (x_2, y_2)\ldots..(\text{xn, yn}), x\& y \in R$
Loss function $L(y, f(x))$.
Output: regression tree $\widehat{f}(x)$;
(i) Initialization $f_{o\,x} = argmin_c \sum_{i=1}^{N} (y_i, c)$
(ii) $for\, m = 1\, To\, M$
(a) For i T N, calculating the residuals.$r_{mi} = -[\partial L(y, f(x_i))/\partial f(x_i)]_{f(x)=f_{m-1(x)}}$
(b) Train a regression tree in order to fit the computed residual (r_{mi}) and then obtain the leaf node area regarding m_{th} tree R_{mj}, $j = 1, 2, 3, .., J$;
(c) For $j = 1$ To J compute$C_{mj} = \text{argmin}_c \sum_{xi \in R_{mj}} L(y_i, f_{m-1}(x_i) + c)$
(d) Update the current model.$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I(x \in R_{mj})$
(iii) Obtain the final additive model $\widehat{f}(x) = f_M(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I(x \in R_{mj})$

Algorithm 1: Gradient boosting machine model.

4. Results and Discussion

In this study, four machine learning modeling approaches have been developed to select the best model for predicting monthly evaporation. The four models (RF, ELM, GBM, and GPR) are trained and validated using climate data collected from two different locations in Iraq. About seventy percent of available data were used for calibration and the other thirty percentage used for validating the predictive models. The used models in this study have been assessed by different statistical criteria as well as graphical presentations.

For the case study, the performances of the applied models through the training phase are summarized in Table 1. The given statistics showed that all the models provided a good similarity between predicted evaporation and predicted values except GPR ($R = O.938$, and Md = 0.967). Furthermore, it can be observed that the GBM generated fewer error forecasts compared to other models (MAE = 14.170, RMSE = 23.092, MAPE = 0.095, $R = 0.987$, and Md = 0.993). However, the performances of the ELM and RF models were very similar. However, it can be said that there was a slight advantage in favor of the ELM model. This model provided smaller values of MAE and MAPE compared to the ORF model. Table 2 provides a significant analysis of the models' performances through the training phase for the second case study. Based on the obtained results, the GBM model showed an excellent ability to predict the monthly evaporation, providing lowest estimated errors (MAE = 13.645, RMSE = 20.509, and MAPE = 0.058) and highest prediction accuracy ($R = 0.994$, WI = 9.997). The second and third best models were ELM and QRF, respectively. However, the GPR was considered the worst predicted model because it gave the highest values of RMSE, MAPE, and MAE. It can be concluded that, through the training stage, the GPR was noticed to have a poor accuracy of both case studies. However, the GBM model has a robust

Table 1: The evaluations of the predictive models through training phase: first case study in Diyala state.

Model	MAE	RMSE	MAPE	R	Md
ELM	26.803	44.588	0.170	0.951	0.973
GPR	32.157	49.524	0.231	0.938	0.967
QRF	27.095	43.214	0.186	0.953	0.975
GBM	14.170	23.092	0.095	0.987	0.993

Table 2: The evaluations of the predictive models through training phase: second case study in Erbil state.

Model	MAE	RMSE	MAPE	R	Md
ELM	31.595	45.222	0.123	0.968	0.983
GPR	37.719	49.512	0.184	0.961	0.980
QRF	26.690	39.933	0.103	0.975	0.987
GBM	13.645	20.509	0.058	0.994	0.997

performance in the simulation of the evaporation rate for both case studies according to the obtained statistical parameters.

To assess the prediction accuracy of the applied models for the two case studies, boxplot diagrams were established to visually show the similarity of the prediction values with the observed evaporation rates. The performances of the four applied models to predict the monthly evaporation rate for both cases studies are graphically illustrated in Figures 5 and 6, respectively. The clearest observation that can be reported was the inability of the GPR model to generate an acceptable accuracy of evaporation estimations. Moreover, this model could not provide a satisfactory prediction especially for higher and lower values of evaporation. However, both figures illustrated that the GBM was superior because the calculated median for that model was very close to the actual value. Additionally, it successfully managed to simulate the higher and lower values of evaporation compared to other models.

Although success has been attained in the monthly evaporation using the GBM model during the training phase, it is very essential to evaluate the proposed model with testing dataset. As is well known, the training results may provide misleading assessment because the model is trained using known input and third corresponding targets [65]. Besides, the testing phase is very crucial in assessing the quality of the predictive models and, hence, the models' abilities would be assessed very well in terms of generalization and avoiding overfitting [66].

The assessment process of the applied models through the testing phase for the first case study that was in Diyala state is exhibited in Table 3. The superiority of the GBM model in estimating the monthly evaporation compared to other models has been easily noted in the table. More specifically, the GBM model was found to produce a satisfactory estimate with RMSE of 28.478, MAE of 21.541, MAPE of 0.181, R of 0.976, and Md of 0.987. However, the QRF provided the worst prediction accuracy compared to the applied models. With respect to case study 2 which was in Erbil state, the performance of the GBM according

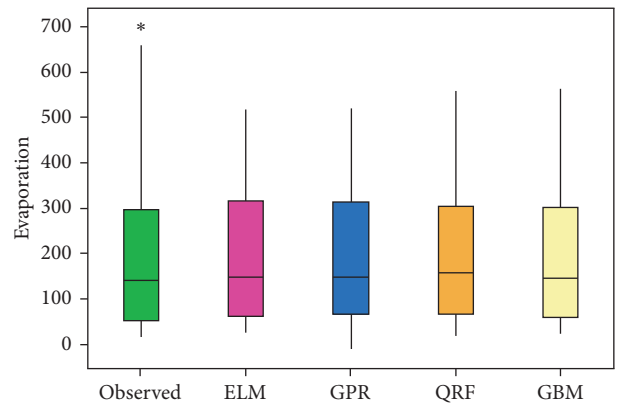

Figure 5: boxplot showing similarity between predictive and observed values for the first case study in Diyala state through training phase.

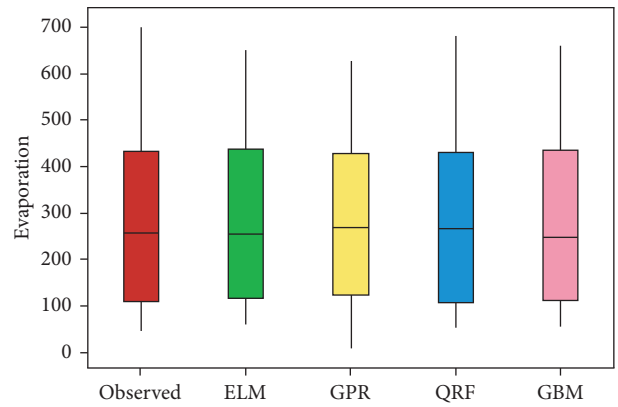

Figure 6: boxplot showing similarity between predictive and observed values for the second case study in Erbil state through training phase.

to Table 4 was also superior and provided fewer estimated errors (MAE = 26.368, RMSE = 35.345, and MAPE = 0.130) as well as higher values of R (0.985) and Md (0.989).

The reported results for both case studies showed that the GBM significantly outperformed the other machine learning models. The superiority of this model can be measured based on its capacity for reducing the MAE and RMSE for both stations during the testing phase (see Figure 7). The results showed for the case first case study a prediction enhancement in terms of MAE and RMSE by 7.17%, 21.01%; 16.51%, 15.74%; and 23.14%, 26.64%; during using GBM compared to ELM, GPR, and QRF, respectively. However, for the second case study in Erbil state, the prediction enhancement was improved in terms of reduction of MAE and RMSE by 10.88%, 9.24%; 15.24%, 5%; and 16.06%, 15.76%; respectively, compared to ELM, GPR, and QRF models.

The visualization assessment presented in Figures 8 and 9 proved that the estimated monthly evaporation rates for both stations by GBM through the testing phase were very close to the observed values. Moreover, the statistical parameters such as median and highest and lowest values

Table 3: The evaluations of the predictive models through testing phase: first case study in Diyala state.

Model	MAE	RMSE	MAPE	R	Md
ELM	23.204	36.053	0.171	0.963	0.978
GPR	25.801	33.799	0.216	0.966	0.982
QRF	28.026	38.818	0.200	0.959	0.974
GBM	21.541	28.478	0.181	0.976	0.987

Table 4: The evaluations of the predictive models through testing phase: second case study in Erbil state.

Model	MAE	RMSE	MAPE	R	Md
ELM	29.588	38.943	0.151	0.982	0.986
GPR	31.108	37.205	0.172	0.982	0.987
QRF	31.412	41.959	0.144	0.975	0.984
GBM	26.368	35.345	0.130	0.985	0.989

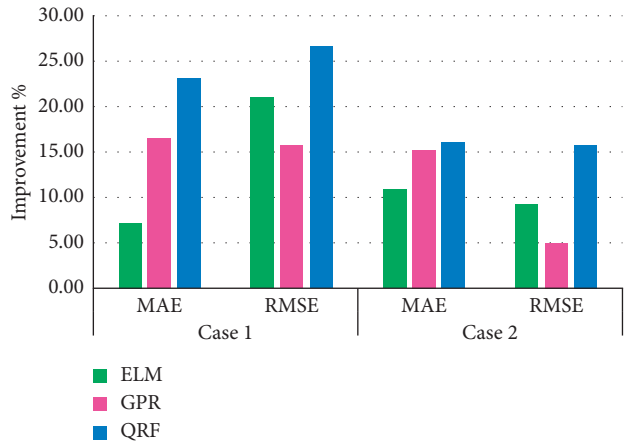

Figure 7: The superiority of the GBM over the ELM, GPM, and QRF in reducing the values of statistical parameters. Case 1 is in Diyala state, while case 2 is in Erbil state.

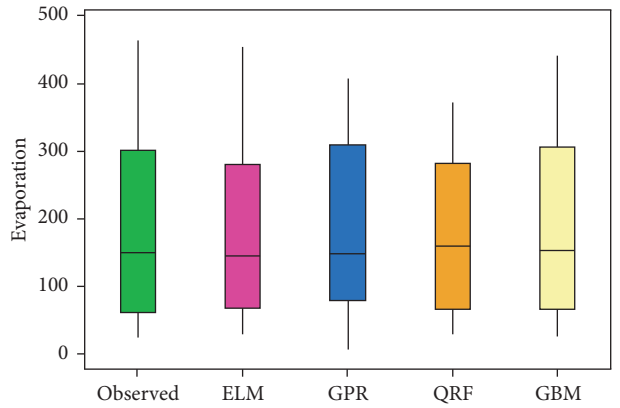

Figure 8: The boxplot showing similarity between predictive and observed values for the first case study in Diyala state through testing phase.

were noticed to be very similar to the actual values. However, these figures showed that the GPR model had a poor performance in both case studies compared to other models.

For further assessment, Taylor diagrams were created using the prediction values obtained from four models for both stations (see Figures 10 and 11). The advantage of

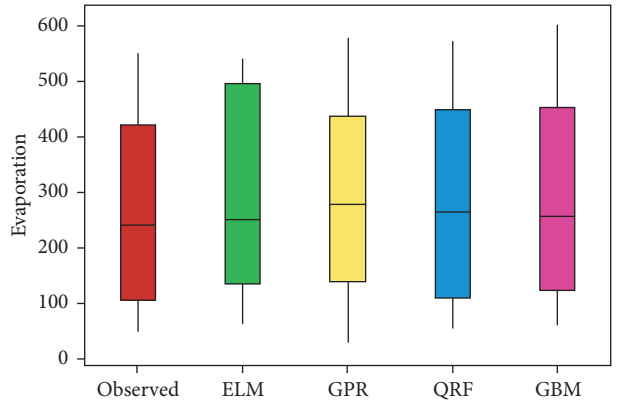

Figure 9: boxplot showing similarity between predictive and observed values for the second case study in Erbil state through testing phase.

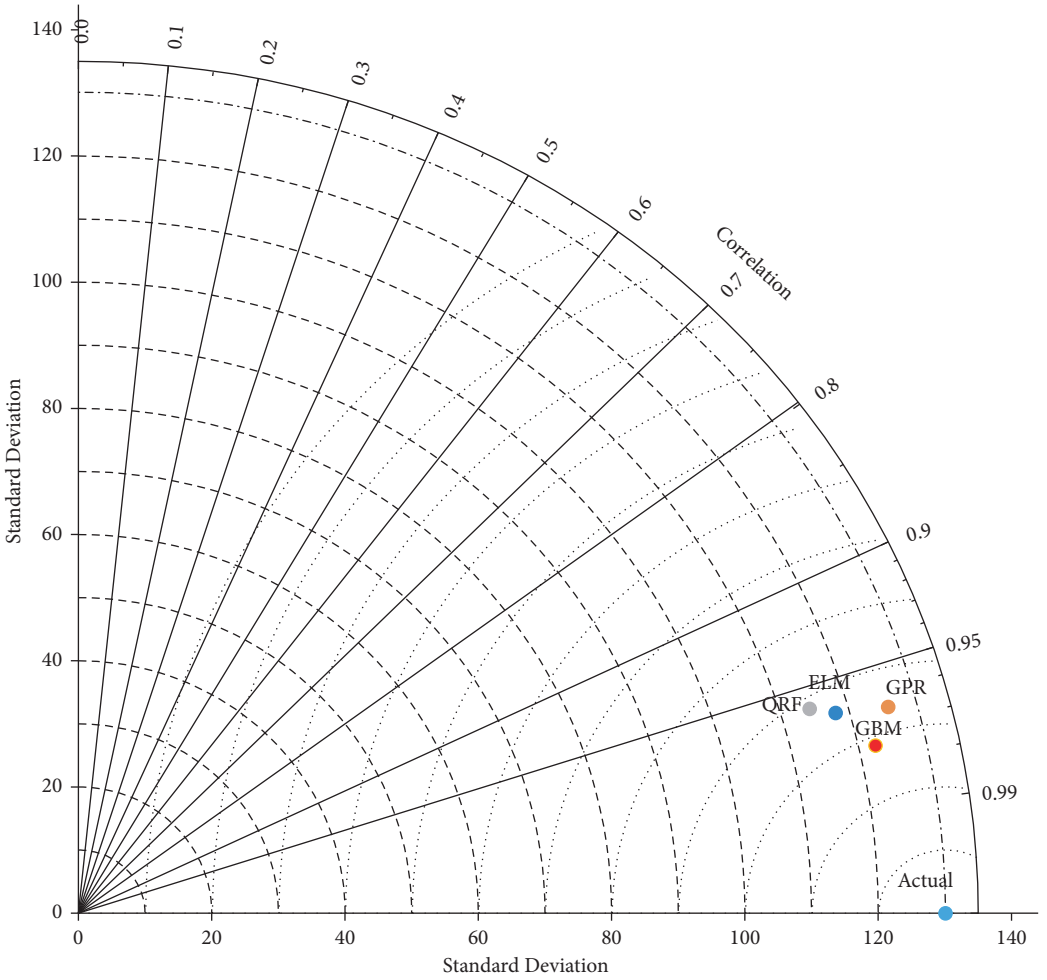

Figure 10: Taylor diagram was created to illustrate the similarity between observed and predicted values during the testing phase: case study 1 in Diyala state.

using Taylor diagram is to assess the comparable models with the actual data using three statistical parameters (standard deviation, root mean square error, and correlation coefficient). Besides, the equivalent evaporation rates obtained from each model and the actual values were assigned on a polar diagram. It can be seen from

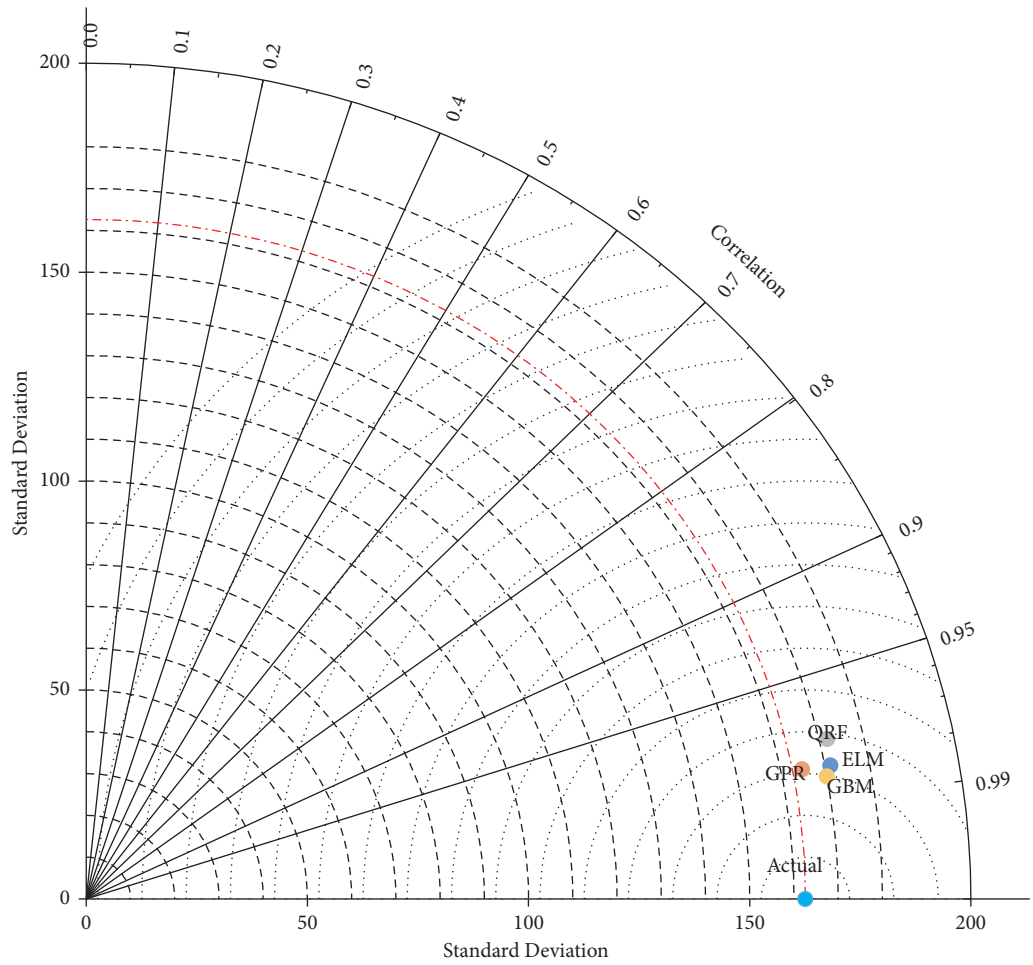

Figure 11: Taylor diagram was created to illustrate the similarity between observed and predicted values during the testing phase: case study 2 in Erbil state.

figures related to both stations that the location of the GBM model was closer to the actual values than other comparable models.

5. Conclusions

As the evaporation rate is a significant element in the hydrological cycle, its process in nature is very complicated and stochastic. In this paper, the capability of artificial intelligence models such as ELM, QRF, GBM, and GPR has been evaluated in the prediction of monthly evaporation over two stations located in Diyala and Erbil states, Iraq. The input parameters include metrological data such as sunshine hours, minimum and maximum temperature, wind speed, and relative humidity. The models were assessed using different statistical criteria as well as graphical plots. The findings of this study revealed that the GBM modeling approach has an excellent performance in the prediction of the monthly rate of evaporation over two stations with minimum forecasting errors. However, the QRF models showed the poorest performance compared with other applied models. All in all, the achieved results proved that the suggested predictive model (GBM) showed an optimistic technique for these regions; thus, it may assist local stakeholders in the management of water resources.

6. Recommendations

The recommendations for future research can be illustrated as follows:

(i) This study recommends the use of the adopted model GBM to estimate the monthly evaporation rates and investigate over several stations located in the middle and southern parts of Iraq. This study showed that the GBM model showed a good prediction accuracy in areas located in the eastern and northern parts of Iraq. Thus, it is very important to investigate the ability of this model in estimating evaporation in another regions.

(ii) The application of feature selection tool is very important to choose the most proper input variables, thus reducing the model complexity [13,67].

(iii) The GBM model is incorporated with novel bio-inspirated algorithms for enhancing its performance prediction, thereby producing much accurate predictions [68–70].

Data Availability

The data are available upon request from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] G. Konapala, A. K. Mishra, Y. Wada, and M. E. Mann, "Climate change will affect global water availability through compounding changes in seasonal precipitation and evaporation," *Nature Communications*, vol. 11, no. 3044, pp. 1–10, 2020.

[2] D. J. Short Gianotti, R. Akbar, A. F. Feldman, G. D. Salvucci, and D. Entekhabi, "Terrestrial evaporation and moisture drainage in a warmer climate," *Geophysical Research Letters*, vol. 47, no. 5, Article ID e2019GL086498, 2020.

[3] J. Shiri, "Evaluation of a neuro-fuzzy technique in estimating pan evaporation values in low-altitude locations," *Meteorological Applications*, 2018.

[4] S. Wang, J. Lian, Y. Peng, B. Hu, and H. Chen, "Generalized reference evapotranspiration models with limited climatic data based on random forest and gene expression programming in Guangxi, China," *Agricultural Water Management*, vol. 221, pp. 220–230, 2019.

[5] W. Ma, "Research progresses of flash evaporation in aerospace applications," *International Journal of Aerospace Engineering*, vol. 2018, Article ID 3686802, 15 pages, 2018.

[6] S. M. Vicente-Serrano, M. Bidegain, M. Tomas-Burguera et al., "A comparison of temporal variability of observed and model-based pan evaporation over Uruguay (1973-2014)," *International Journal of Climatology*, vol. 38, no. 1, pp. 337–350, 2018.

[7] J. Fan, B. Chen, L. Wu, F. Zhang, X. Lu, and Y. Xiang, "Evaluation and development of temperature-based empirical models for estimating daily global solar radiation in humid regions," *Energy*, vol. 144, pp. 903–914, 2018.

[8] W. Jing, Z. M. Yaseen, S. Shahid et al., "Implementation of evolutionary computing models for reference evapotranspiration modeling: short review, assessment and possible future research directions," *Engineering Applications of Computational Fluid Mechanics*, vol. 13, no. 1, pp. 811–823, 2019.

[9] Y. Yang, H. Su, and J. Qi, "A critical evaluation of the nonparametric approach to estimate terrestrial evaporation," *Advances in Meteorology*, vol. 2016, Article ID 5343718, 10 pages, 2016.

[10] O. Kisi, I. Mansouri, and J. W. Hu, "A new method for evaporation modeling: dynamic evolving neural-fuzzy inference system," *Advances in Meteorology*, vol. 2017, Article ID 5356324, 9 pages, 2017.

[11] C. M. Burt, A. J. Mutziger, R. G. Allen, and T. A. Howell, "Evaporation research: review and interpretation," *Journal of Irrigation and Drainage Engineering*, vol. 131, no. 1, pp. 37–58, 2005.

[12] Ö. Kişi, "Daily pan evaporation modelling using multi-layer perceptrons and radial basis neural networks," *Hydrological Processes*, vol. 23, no. 2, pp. 213–223, 2009.

[13] A. Malik, A. Kumar, S. Kim et al., "Modeling monthly pan evaporation process over the Indian central Himalayas: application of multiple learning artificial intelligence model," *Engineering Applications of Computational Fluid Mechanics*, vol. 14, no. 1, pp. 323–338, 2020.

[14] A. Malik, "Pan evaporation estimation in Uttarakhand and Uttar Pradesh States, India: Validity of an integrative data intelligence model," *Atmosphere*, 2020.

[15] H. Wang, H. Yan, W. Zeng, G. Lei, C. Ao, and Y. Zha, "A novel nonlinear Arps decline model with salp swarm algorithm for predicting pan evaporation in the arid and semi-arid regions of China," *Journal of Hydrology*, vol. 582, Article ID 124545, 2020.

[16] Z. M. Yaseen, A. M. Al-Juboori, U. Beyaztas et al., "Prediction of evaporation in arid and semi-arid regions: a comparative study using different machine learning models," *Engineering Applications of Computational Fluid Mechanics*, vol. 14, no. 1, pp. 70–89, 2019.

[17] R. Arunkumar and V. Jothiprakash, "Reservoir evaporation prediction using data-driven techniques," *Journal of Hydrologic Engineering*, vol. 18, no. 1, pp. 40–49, 2013.

[18] V. Nourani, M. Sayyah-Fard, M. T. Alami, and E. Sharghi, "Data pre-processing effect on ANN-based prediction intervals construction of the evaporation process at different climate regions in Iran," *Journal of Hydrology*, vol. 588, Article ID 125078, 2020.

[19] J. Tanny, "Evaporation from a small water reservoir: direct measurements and estimates," *Journal of Hydrology*, vol. 351, no. 1–2, pp. 218–229, 2008.

[20] L. Wang, O. Kisi, M. Zounemat-Kermani, and H. Li, "Pan evaporation modeling using six different heuristic computing methods in different climates of China," *Journal of Hydrology*, vol. 544, pp. 407–427, 2017.

[21] A. Guven and Ö. Kişi, "Daily pan evaporation modeling using linear genetic programming technique," *Irrigation Science*, vol. 29, no. 2, pp. 135–145, 2011.

[22] M. Allen, L. S. Pereira, D. Raes, and Smith, "Crop evaporation–Guidelines for computing crop water requirements–FAO Irrigation and drainage paper 56," *Food Agric. Organ. United Nations*, vol. 300, no. 9, p. 300, 1998.

[23] M. M. Hameed, M. K. Alomar, S. F. Mohd Razali et al., "Application of artificial intelligence models for evapotranspiration prediction along the southern coast of Turkey," *Complexity*, vol. 2021, Article ID 8850243, 20 pages, 2021.

[24] M. A. Ghorbani, R. C. Deo, S. Kim, M. Hasanpour Kashani, V. Karimi, and M. Izadkhah, "Development and evaluation of the cascade correlation neural network and the random forest models for river stage and river flow prediction in Australia," *Soft Computing*, vol. 24, no. 16, pp. 12079–12090, 2020.

[25] A. Ashrafzadeh, A. Malik, V. Jothiprakash, M. A. Ghorbani, and S. M. Biazar, "Estimation of daily pan evaporation using neural networks and meta-heuristic approaches," *ISH Journal of Hydraulic Engineering*, pp. 1–9, 2018.

[26] A. Elbeltagi, N. Azad, A. Arshad et al., "Applications of Gaussian process regression for predicting blue water footprint: case study in Ad Daqahliyah, Egypt," *Agricultural Water Management*, vol. 255, Article ID 107052, 2021.

[27] G. T. Patle, M. Chettri, and D. Jhajharia, "Monthly pan evaporation modelling using multiple linear regression and artificial neural network techniques," *Water Supply*, vol. 20, no. 3, pp. 800–808, 2019.

[28] A. S. Ponraj and T. Vigneswaran, "Daily evapotranspiration prediction using gradient boost regression model for irrigation planning," *The Journal of Supercomputing*, vol. 76, no. 8, pp. 5732–5744, 2019.

[29] A. Sharafati, R. Yasa, and H. M. Azamathulla, "Assessment of stochastic approaches in prediction of wave-induced pipeline scour depth," *Journal of Pipeline Systems Engineering and Practice*, vol. 9, no. 4, 2018.

[30] M. K. Goyal, B. Bharti, J. Quilty, J. Adamowski, and A. Pandey, "Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS," *Expert Systems with Applications*, vol. 41, no. 11, pp. 5267–5276, 2014.

[31] G. Tezel and M. Buyukyildiz, "Monthly evaporation forecasting using artificial neural networks and support vector machines," *Theoretical and Applied Climatology*, 2013.

[32] S. N. Qasem, S. Samadianfard, S. Kheshtgar et al., "Modeling monthly pan evaporation using wavelet support vector regression and wavelet artificial neural networks in arid and humid climates," *Engineering Applications of Computational Fluid Mechanics*, vol. 13, no. 1, pp. 177–187, 2019.

[33] A. Rakhee and A. Kumar, "Predictive modeling of pan evaporation using random forest algorithm along with features selection," in *Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering*, pp. 380–384, Uttarpradesh, India, January 2020.

[34] D. Althoff, R. Filgueiras, and L. N. Rodrigues, "Estimating small reservoir evaporation using machine learning models for the Brazilian savannah," *Journal of Hydrologic Engineering*, vol. 25, no. 8, Article ID 05020019, 2020.

[35] L. Qian, "A study of the conversion of different evaporation pans in South China based on the extreme learning machine model," *Hydrological Sciences Journal*, 2021.

[36] L. Dong, W. Zeng, L. Wu et al., "Estimating the Pan evaporation in northwest China by coupling CatBoost with bat algorithm," *Water*, vol. 13, no. 3, p. 256, 2021.

[37] M. Al-Mukhtar, "Modeling the monthly pan evaporation rates using artificial intelligence methods: a case study in Iraq," *Environmental Earth Sciences*, vol. 80, no. 1, 2021.

[38] A. Ghaemi, M. Rezaie-Balf, J. Adamowski, O. Kisi, and J. Quilty, "On the applicability of maximum overlap discrete wavelet transform integrated with MARS and M5 model tree for monthly pan evaporation prediction," *Agricultural and Forest Meteorology*, vol. 278, p. 107647, 2019.

[39] J. Chenoweth, "Impact of climate change on the water resources of the eastern Mediterranean and Middle East region: Modeled 21st century changes and implications," *Water Resource Research*, 2011.

[40] N. Al-Ansari, A. A. Ali, and S. Knutsson, "Present conditions and future challenges of water resources problems in Iraq," *Journal of Water Resource and Protection*, vol. 6, no. September, pp. 1066–1098, 2014.

[41] J. Lelieveld, "Climate change and impacts in the Eastern Mediterranean and the Middle East," *Climate Change*, 2012.

[42] O. T. Al-Taai and S. H. Hadi, "Analysis of the monthly and annual change of soil moisture and evaporation in Iraq," *Al-Mustansiriyah J. Sci.*vol. 29, no. 4, pp. 7–13, 2018.

[43] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, Massachusets, MA, USA, 2005.

[44] R. C. Deo and P. Samui, "Forecasting evaporative loss by least-square support-vector regression and evaluation with genetic programming, Gaussian process, and minimax probability machine regression: case study of brisbane city," *Journal of Hydrologic Engineering*, vol. 22, no. 6, p. 05017003, 2017.

[45] M. Akbari, F. Salmasi, H. Arvanaghi, M. Karbasi, and D. Farsadizadeh, "Application of Gaussian process regression model to predict discharge coefficient of Gated Piano Key Weir," *Water Resources Management*, vol. 33, no. 11, pp. 3929–3947, 2019.

[46] P. Sihag, P. Jain, and M. Kumar, "Modelling of impact of water quality on recharging rate of storm water filter system using various kernel function based regression," *Modeling Earth Systems and Environment*, 2018.

[47] K. Roushangar and S. Shahnazi, "Prediction of sediment transport rates in gravel-bed rivers using Gaussian process regression," *Journal of Hydroinformatics*, 2019.

[48] B. T. Pham, H.-B. Ly, N. Al-Ansari, and L. S. Ho, "A comparison of Gaussian process and M5P for prediction of soil permeability coefficient," *Scientific Programming*, vol. 2021, Article ID 3625289, 13 pages, 2021.

[49] P. Koepernik and F. Pfaff, "Consistency of Gaussian process regression in metric spaces," *Journal of Machine Learning Research*, vol. 22, no. 244, pp. 1–27, 2021.

[50] Y. Wang and B. Chaib-draa, "An online Bayesian filtering framework for Gaussian process regression: application to global surface temperature analysis," *Expert Systems with Applications*, vol. 67, pp. 285–295, 2017.

[51] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[52] G. Huang, G.-B. Huang, S. Song, and K. You, "Trends in extreme learning machines: a review," *Neural Networks*, vol. 61, pp. 32–48, 2015.

[53] A. M. Araba, Z. A. Memon, M. Alhawat, M. Ali, and A. Milad, "Estimation at completion in civil engineering projects: review of regression and soft computing models," *Knowledge-Based Engineering and Sciences*, vol. 2, no. 2, pp. 1–12, 2021.

[54] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[55] S. A. Gyamerah, P. Ngare, and D. Ikpe, "Probabilistic forecasting of crop yields via quantile random forest and Epanechnikov Kernel function," *Agricultural and Forest Meteorology*, vol. 280, Article ID 107808, 2020.

[56] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, no. Jun, pp. 983–999, 2006.

[57] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, 2001.

[58] J. Zhou, E. Li, S. Yang et al., "Slope stability prediction for circular mode failure using gradient boosting machine approach based on an updated database of case histories," *Safety Science*, vol. 118, pp. 505–518, 2019.

[59] Z. Zhou, L. Zhao, A. Lin et al., "Exploring the potential of deep factorization machine and various gradient boosting models in modeling daily reference evapotranspiration in China," *Arabian Journal of Geosciences*, vol. 13, no. 24, p. 1287, 2021.

[60] M. Xenochristou, C. Hutton, J. Hofman, and Z. Kapelan, "Water demand forecasting accuracy and influencing factors at different spatial scales using a gradient boosting machine," *Water Resources Research*, vol. 56, no. 8, p. e2019WR026304, 2020.

[61] S. R. Naganna, B. H. Beyaztas, N. Bokde, and A. M. Armanuos, "On the evaluation of the gradient tree boosting model for groundwater level forecasting," *Knowledge-Based Engineering and Science*, vol. 1, no. 1, pp. 48–57, 2020.

[62] A. Sharafati, S. B. H. S. Asadollah, and A. Neshat, "A new artificial intelligence strategy for predicting the groundwater level over the Rafsanjan aquifer in Iran," *Journal of Hydrology*, 2020.

[63] M. Gong, Y. Bai, J. Qin, J. Wang, P. Yang, and S. Wang, "Gradient boosting machine for predicting return temperature of district heating system: a case study for residential buildings in Tianjin," *Journal of Building Engineering*, vol. 27, Article ID 100950, 2020.

[64] Z. M. Yaseen, "An insight into machine learning models era in simulating soil, water bodies and adsorption heavy metals:

review, challenges and solutions," *Chemosphere*, vol. 277, Article ID 130126, 2021.

[65] H. Tao, "Training and testing data division influence on hybrid machine learning model process: application of river flow forecasting," *Complexity*, 2020.

[66] M. M. Hameed, M. K. AlOmar, W. J. Baniya, and M. A. AlSaadi, "Incorporation of artificial neural network with principal component analysis and cross-validation technique to predict high-performance concrete compressive strength," *Asian Journal of Civil Engineering*, vol. 22, no. 6, pp. 1019–1031, 2021.

[67] A. Malik, A. Kumar, and O. Kisi, "Daily Pan evaporation estimation using heuristic methods with gamma test," *Journal of Irrigation and Drainage Engineering*, vol. 144, no. 9, p. 04018023, 2018.

[68] H. Tao, A. A. Ewees, A. O. Al-Sulttani et al., "Global solar radiation prediction over North Dakota using air temperature: development of novel hybrid intelligence model," *Energy Reports*, vol. 7, pp. 136–157, 2021.

[69] N. Arya Azar, N. Kardan, and S. Ghordoyee Milan, "Developing the artificial neural network–evolutionary algorithms hybrid models (ANN–EA) to predict the daily evaporation from dam reservoirs," *Engineering Computation*, 2021.

[70] A. Malik, "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," *Engineering Applications of Computational Fluid Mechanics*, vol. 15, no. 1, pp. 1075–1094, 2021.