RESEARCH

Application of hybrid machine learning-based ensemble techniques for rainfall-runoff modeling

Gebre Gelete[1,2]

Abstract

The main aim of this study was to develop hybrid machine learning (ML)-based ensemble modeling of the rainfall-runoff process in the Katar catchment, Ethiopia. This study used four single ML models, namely the general regression neural network (GRNN), long short-term memory neural network (LSTM), extreme learning machine (ELM) and Hammerstein-Weiner (HW) for modeling the rainfall-runoff process. Subsequently, two strategies were followed to improve the performance of the single models. In the first strategy, simple average ensemble (SAE), weighted average ensemble (WAE), Hammerstein-Weiner ensemble (HWE) and Neuro-fuzzy ensemble (NFE) were developed using the results of the single models. A hybrid Boosted Regression Tree (BRT) ensemble was developed in the second strategy to enhance the single models' modeling accuracy. The study used ten years (2008–2017) of data for calibration and validation of the developed models. The performances of the developed models were assessed using root mean square error (RMSE), percent bias (PBIAS), mean absolute error (MAE) and Nash-Sutcliffe coefficient efficiency (NSE). The results of single ML models showed that the LSTM model gave the best prediction performance with NSE $=0.933$ and RMSE $=5.308$ m^3/s in the validation phase. For ensemble modeling, the best result was obtained by NFE increasing the performance of HW, GRNN, LSTM and ELM models by 3.35%, 13.25%, 2.57% and 19.9%, respectively. Evaluation of the hybrid BRT models showed that all the hybrid models provide reliable modeling performance with LSTM-BRT demonstrating better predictive accuracy (NSE $=0.981$, RMSE $=1.999$ m^3/s and PBIAS $=0.75$%). In general, the result of this study proved the promising influence of ensemble techniques and hybrid BRT models for rainfall-runoff modeling.

Keywords Ensemble technique · Hybrid model · Katar catchment · Machine learning · Rainfall-runoff

Introduction

Accurate modeling of the rainfall-runoff process has been a hot topic in hydrology research as it plays a critical role in water resources management (Adnan et al. 2021). Rainfall-runoff is a nonlinear and complex process influenced by humidity, soil moisture, topography, rainfall, groundwater (Mohammadi et al. 2022). Modeling the rainfall-runoff process is vital in planning and optimizing water resources such

as reservoir operations, watershed planning, irrigation water management, and navigation (Noori and Kalin 2016). It also plays an important role in preventing and mitigating natural disasters such as floods, droughts, and other extreme events (Asadi et al. 2019).

Due to the increasing of water resource depletion around the globe, accurate modeling of runoff is becoming increasingly important and requires more accurate models (Adnan et al. 2021). However, accurate modeling of the rainfall-runoff processes is a difficult task due to the temporal and spatial dynamics of the process exhibiting nonlinearity, complex properties and chaotic perturbations (Nourani et al. 2021a, b). So far different modeling approaches which are broadly categorized as physically-based and data-driven models have already been applied to hydrologic modeling (Young et al. 2017). The physically based models such as the Soil and Water Assessment Tool (SWAT) require extensive data and time consuming (Adnan et al. 2019; Nourani et al.

✉ Gebre Gelete
    gegelete04@gmail.com

1   Faculty of Civil and Environmental Engineering, Near East University, 99138 Nicosia/TRNC, Mersin-10, Turkey

2   College of Agriculture and Environmental Science, Arsi University, 193 Asela, Ethiopia

2021a, b). This type of model also has a limitation in achieving accurate hydrological modeling (Young et al. 2017). When accurate modeling is more important than understanding of the physics of the problem, data driven models (e.g., machine learning), are preferred to physically-based models because of their few input data requirements (Nourani et al. 2019). Machine learning (ML)-based models have successfully been used for accurate modeling non-stationary and non-linear rainfall-runoff process (Adnan et al. 2019; Cai et al. 2021; Koch and Schneider 2022). The ML models are calibrated and tested for a specific type of data and locations. In data-scarce regions, common hydrologic forecasting models cannot accurately predict runoff because of the less dense distribution of rain gauges (Mehr et al. 2015). In this cases, ML models can precisely simulate runoff using single station or cross-station streamflow data (Nourani et al. 2021a, b).

The ML-based models used for hydrologic process modeling includes Feed forward neural network (FFNN), support vector machine (SVM), Adaptive Neuro-Fuzzy Inference System (ANFIS), general regression neural network (GRNN), boosted regression tree (BRT), long short-term memory neural network (LSTM), Hammerstein-Weiner (HW) and ELM. Artificial neural networks (ANNs) such as FFNN are known to handle highly nonlinear and complex problems, but the selection of lags as input is difficult and the accuracy of the model depends on the choice of the optimal network structure (Kratzert et al. 2018). Recurrent neural networks (RNNs), on the other hand, take into account the sequential order of inputs and therefore capture temporal dynamics (Bhattacharjee and Tollner 2016). However, RNNs have the problem that the gradients vanish and explode when trained with long-term delays (Yin et al. 2021). The LSTM model has emerged as an alternative technique to solve these limitations of FFNN and RNN. According to (Kratzert et al. 2018), the strength of the LSTM model lies in its ability to learn long-term dependencies between the input and output of the network. The LSTM cell is able to make the decision whether to pass the information to the next cell or to prevent the flow of information from the last to the next cell (Nourani et al. 2022). Previous studies has applied LSTM model for different hydrologic modeling such as rainfall-runoff (Kratzert et al. 2018; Yin et al. 2021), suspended sediment concentration (Kaveh et al. 2021) and evaporation (Lakmini Prarthana Jayasinghe et al. 2022), ground water level (Nourani et al. 2022). The other ML model, Hammerstein-Wiener (HW) model is known for the identification of nonlinear systems. Although the HW model has not been used for rainfall-runoff modeling, it has shown great potential in rainfall prediction (Pham et al. 2019), dissolved oxygen concentration (Abba et al. 2020) and turbidity of water (Gaya et al. 2017). The other ML model used in this study was generalized regression neural network (GRNN). The GRNN is a type of ANNs that uses only backpropagation algorithm

for training (Mehr et al. 2015). The GRNN has been used in different field such as stream flow (Cai et al. 2021), dissolved oxygen (Abba et al. 2020) and suspended sediment prediction (Khan et al. 2019) and showed acceptable result.

Among the various forms of ANNs, single-layer FFNN have been successfully used for modeling hydrologic processes (Nourani et al. 2020a, b; Nourani et al. 2021a, b). However, single hidden layer FFNNs trained based on the gradient method have the limitation of local convergence, requirement of long training time and overfitting (Niu et al. 2019a, b). Therefore, Extreme Learning Machines (ELM) model, a new learning paradigm, has recently been proposed that drastically reduces the limitation imposed by the high computational time requirements to train the single-layer feed-forward neural network (Taormina and Chau 2015). The potential of ELM can be attributed in particular to its high learning speed and generalization capability. Hence, the ELM model has been widely applied in hydrological and environmental studies (Li et al. 2019; Mundher et al. 2016). The other ML-based model used in different hydro-climate modeling is BRT. The BRT model combines different regression trees through boosting technique (Malik et al. 2022). BRT is a robust and powerful prediction model, enhancing the accuracy of a single model by combining multiple independent models. Although it is rarely utilized for rainfall-runoff modeling, the BRT model has shown reliable performance in different fields such as reference evapotranspiration (Malik et al. 2022), landslide susceptibility mapping (Park and Kim 2019; Youssef et al. 2016) and mapping flush flood susceptibility (Abedi et al. 2022).

The reliability of ML models has been demonstrated by previous studies for hydrologic modeling in different watersheds around the world. Despite the reliable results of these ML-based models, it is evident that there is no single model that has been proven to be the most accurate for all types of datasets (Nourani et al. 2021a, b). The characteristics of the data, such as size, linearity, normality and magnitude affect the predictive power of the model (Abba et al. 2020). For example, when modeling a time series, one model may more accurately simulate the maximum values, while the other may predict the lower values well (Sharghi et al. 2018). Thus, combining several models using different ensemble techniques could improve the modeling performance (Phukoetphim et al. 2016). In this regard, Nourani et al. (2021a, b) used an ensemble of SVM, FFNN, ANFIS and multilinear regression (MLR) using different ensemble techniques in suspended sediment load estimation. The result showed that the ensemble modeling led an improved modeling accuracy. Improving the accuracy of modeling has already been applied in various hydrologic process such as rainfall-runoff simulation (Nourani et al. 2021a, b), suspended sediment load estimation (Himanshu et al. 2017; Nourani et al. 2021a, b), evapotranspiration modeling (Nourani et al. 2019) and

earth fill dam seepage analysis (Sharghi et al. 2018). However, to the best of the authors' knowledge, no previous studies have employed the HW and Boosted Regression Tree (BRT) techniques in the context of rainfall-runoff modeling. Additionally, there is a lack of research that has explored the application of hybrid ensemble BRT with ML models for rainfall-runoff modeling. Thus, the objective of this study was; (i) to compare the performance of different ML models namely HW, GRNN, ELM and LSTM in modeling rainfall-runoff; (ii) to improve the overall modeling accuracy using four ensemble techniques namely neuro-fuzzy ensemble (NFE), simple average ensemble (SAE) and weighted average ensemble (WAE) and HWE (in scenario 1) and hybrid BRT ensemble with ML models in scenario 2. In this study, some robust modeling techniques that are relatively new to the application of rainfall-runoff modeling were considered. For chaotic, dynamic, and complex systems such as rainfall-runoff, single models often lead to unreliable predictions. To overcome the limitations of single models in modeling complex problems, hybrid models have nowadays attracted the attention of researchers. To this end, the hybrid BRT models was developed in this study to improve the predictive performance of single models. The use of the hybrid ensemble BRT with the ML-based models (HW-BRT, LSTM-BRT, ELM-BRT and GRNN- BRT) for rainfall-runoff has not been reported in the literature.

Based on the literature review, several studies examined the applicability of ML models in hydrological process simulation. However, new algorithms for hydrologic modeling must be used to enable optimum decision-making. In addition, most ML-based models are complicated and thus their calibration involves high computational costs. Recently, ML models such as GRNN, ELM, HW and LSTM have gradually gained acceptance in various water management applications due to their robustness, simplicity, and high modeling efficiency in processing large amounts of data compared to other ML methods. Thus, this study evaluated the applicability and accuracy of the relatively simple LSTM, ELM, HW and GRNN models as well WAE, NFE, SAE, HWE ensemble techniques and the hybrid BRT model for rainfall-runoff in Katar catchment, Ethiopia.

Material and methods

Description of the study area

The study area, Katar catchment is a sub-watershed of the Ethiopian Central Rift Valley basin covering an area of 3293 km^2. Topographically the study area is located at longitude 38.899° to 39.41°E and latitude 7.359° to 8.165°N (see Fig. 1). The main river, Katar River and its tributaries flow into Lake Ziway which is the source of income for the fishing community, sources of water supply and irrigation for the community in the catchment. It is characterized by its complex topography in which the elevation varies between 1635 m and 4167 m above mean sea level. The catchment consists of six main soil types, namely Luvisols, Cambisols, Fluvisols, Vertisols, and Leptosols. The climate of the study area is semi-arid to sub-humid with an annual average temperature ranging between 16 and 20 °C. The minimum and maximum precipitation values of the area, based are 729.6 mm and 1231.7 mm, respectively. The dry season lasts from October to May and the rainy season from June to September (and accounts for about 70% of the precipitation). There are six meteorological stations in the study area: Arata, Assela, Bekoji Kulumsa, Ogolcho, and Sagure whose location is shown in Fig. 1. The area has one hydrometry station (Abura) at the outlet of the catchment, which recorded a maximum discharge of 152.033 m^3/s in August and a minimum discharge of 0.106 m^3/s in January.

Data type and source

In this study, ten years (2008–2017) of daily rainfall and runoff data of the Katar River catchment at Abura Station were used for model calibration and validation. The Theisen polygon average of rainfall from the six stations were used. The required climatic data (temperature and rainfall) were obtained from the national meteorologic agency whereas the runoff data were collected from the Ethiopian ministry of water and energy. The data used were divided into two subsets, with the first 70% of the data (2008–2014) used for calibration and the remaining 30% (2015–2017) used for validation of the developed models. The descriptive statistics of the data is shown in Table 1.

In addition to the descriptive statistics, the time series of the runoff and rainfall data is shown in Fig. 2.

Methodology

A review of the literature shows that several studies have been conducted using data intelligence algorithms and have yielded promising results in modeling nonlinear systems. To this end, the current study employed four ML-based namely HW, ELM, GRNN and LSTM models to estimate the rainfall-runoff process. The study was conducted in three steps. First, nonlinear sensitivity analysis was conducted to identify the most relevant input that has a significant influence on the output. In the second step, the rainfall-runoff process was modeled using HW, ELM, GRNN and ELM models. In the third step, the overall efficiency was improved in two scenarios. In the first scenario, four novel ensemble techniques such as SAE, WAE, HWE and NFE were developed. In the second scenario a hybrid BRT ensemble was employed in
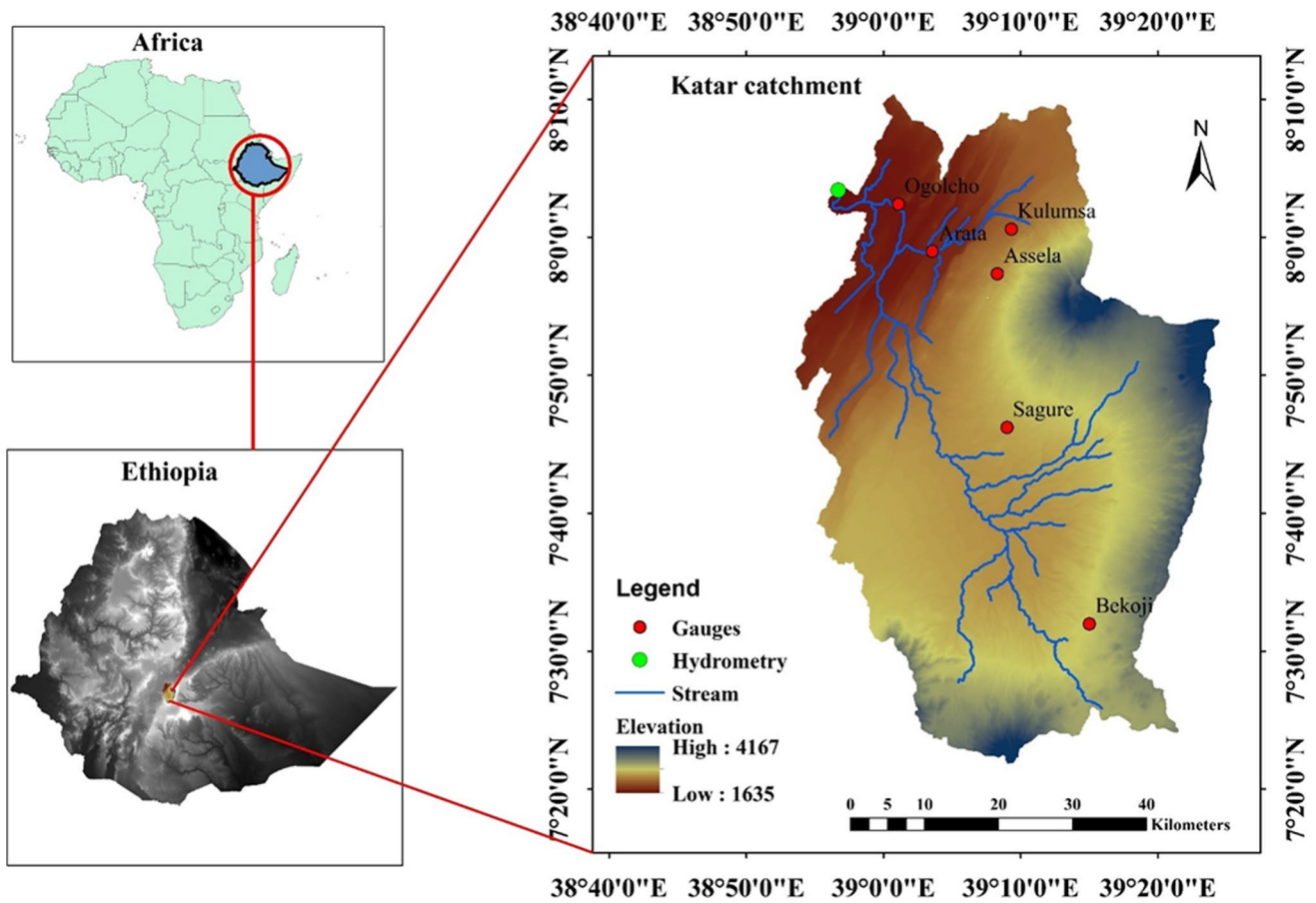
Fig. 1 The map of the Katar catchment

order to boost the accuracy of individual models. The schematic of the proposed meteorology is shown in Fig. 3.

Long short-term memory (LSTM)

The LSTM model structure is a special type of RNN developed to solve the limitations of traditional RNN in learning long-term dependencies (Kratzert et al. 2018). LSTM can regulate and store information over time. This makes this model well-suited for learning long-term memory

and dependencies effects (Koch and Schneider 2022). The study by (Bengio et al. 1994) showed that the traditional RNN can hardly remember sequences longer than 10. For modeling daily runoff, this would mean that only the meteorological data of the last 10 days can be used as input for predicting the next day's runoff. This period is too short considering that the memory of catchments, including snow, groundwater and glacier reservoirs, has a lag of up to several years between rainfall and runoff. All RNNs take the form of chained repetitive neural network

Table 1 The rainfall and runoff data descriptive statistics

Data type	Period	Min	Mean	Max	Standard deviation	Coefficient of variation
Rainfall	Whole	0	2.2767	71.2	4.8467	2.1288
(mm)	Calibration	0	2.1277	71.2	4.658	2.1891
	Validation	0	2.6243	52.4	5.2445	1.9984
Runoff (m^3/s)	Whole	0.106	12.192	152.033	19.47834	1.5976
	Calibration	0.106	11.848	152.033	19.043	1.6073
	Validation	0.115	12.995	126.779	20.4354	1.57255

Fig. 2 Time series of rainfall and runoff
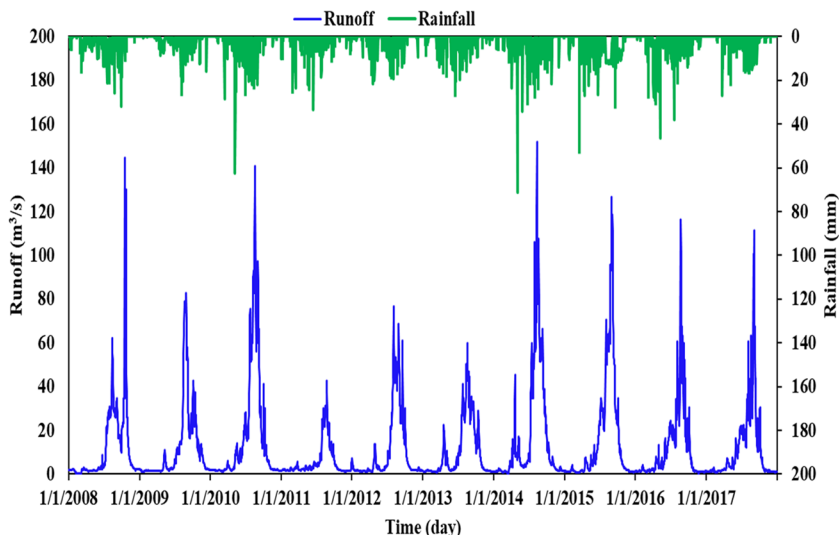
modules. The LSTM model, where special memory cells are used to store information, also have this chain similarly structured (Liu et al. 2019; Zhang et al. 2018).

According to Kratzert et al. (2018), the LSTM structure has a memory cell (C_t) that stores the information, and three circled letters (gates) that control the information flow within the LSTM cell (Fig. 3). The first gate, introduced by Gers et al. (2000), is the forget gate which controls which elements and to what extent the cell state vector C_{t-1} will be

forgotten. The internal LSTM model structure is shown in Fig. 4.

In the internal LSTM model cell, the i, f and O represents the input, forget and output gates, respectively. In addition, the h_t and C_t represent the hidden state and the cell state at time t, respectively. The mathematical expression for the different gate and cell states of the LSTM model is given in the following equations (Kratzert et al. 2018):

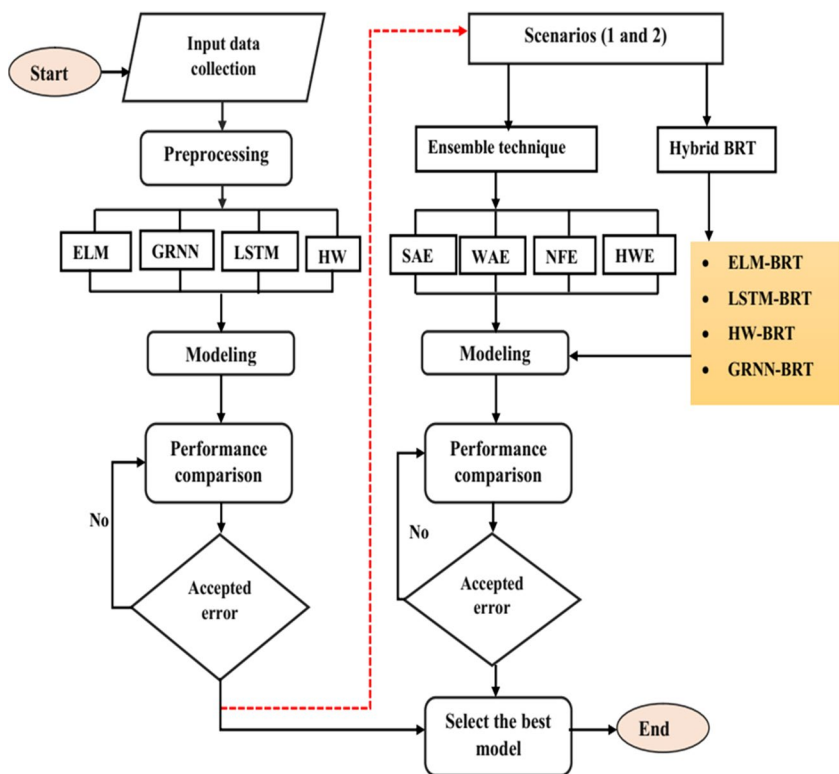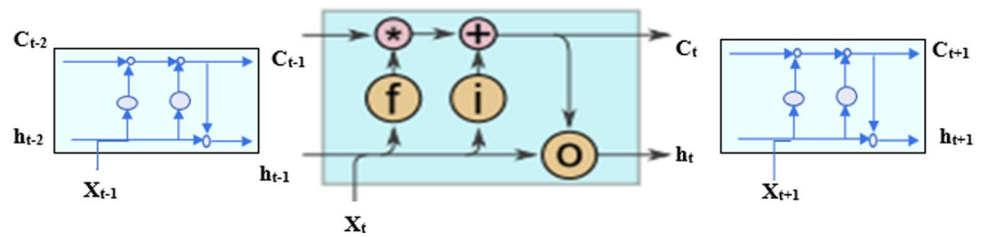**Fig. 3** The general schematic of the proposed methodology

Fig. 4 The LSTM model cell internals

Input gate

$$i_t = \sigma\left(W_i X_i + U_i h_{t-1} + b_i\right) \tag{1}$$

Where i_t, σ, W_i, X_i, U_i, h_{t-1} and b_i are the input gate vector having value between 0 and 1, sigmoidal function, weight linking input gate with inputs, input, weights from the input, output from the previous time step and bias vector, respectively.

Forget gate

$$f_t = \sigma\left(W_f X_t + U_f h_{t-1} + b_f\right) \tag{2}$$

Where f_t, W_f, U_f, and b_f are the resulting of vector forget gate having value between 0 and 1, weight forget gate with inputs, weights from the input and bias vector, respectively.

Output gate

$$O_t = \sigma\left(W_0 X_t + U_o h_{t-1} + b_o\right) \tag{3}$$

Where f_t, W_f, U_f, and b_f are the resulting of vector forget gate having value between 0 and 1, weight forget gate with inputs, weights from the input and bias vector, respectively.

Cell state

For the cell state the potential update vector is calculated from the last hidden state (h_{t-1}) and current input (x_t) as:

$$\overline{C}_t = \tanh\left(W_c X_t + U_c h_{t-1} + b_c\right) \tag{4}$$

Where $\overline{C}_t$ and tanh represents the cell state at the previous time with value in between −1 and 1 and hyperbolic tangent, respectively.

The cell stat (C_t), using the result of Eq. 4 is updated as:

$$C_t = f_t{}^\circ C_{t-1} + i_t{}^\circ \overline{C}_t \tag{5}$$

The new hidden state (h_t) is then computed by combining the results of output gate and the cell state as:

$$h_t = \tanh\left(C_t\right) O_t \tag{6}$$

Hammerstein-Weiner Model (HW)

HW is one of the back-box models developed for nonlinear system identification (Gaya et al. 2017). It consists of a set of interconnected parallel static nonlinear blocks and linear dynamics. The HW model intersection is considered a suitable representation with a more understandable and precise relationship to the nonlinear and linear system compared to the traditional ANN (Zhang et al. 2017).

In addition, the HW model incorporates a simple and flexible procedure for determining the specifications of parameters for nonlinear models and can effectively capture the physical knowledge of system properties (Pham et al. 2019). In the HW model, the linear dynamical system is embedded between two nonlinear blocks (Abba et al. 2020). In this model, the nonlinear model is converted into a piecewise linear function and then transformed to a nonlinear output function. According to Abinayadhevi and Prasad (2015), the distinct linear and nonlinear blocks, offer the HW model the advantage of stability analysis being solely dependent on the linear part, which can be readily assessed. However, when employing general non-linearities within the Hammerstein model, the overall model performance tends to reduced. The general HW model structure consists of three blocks: a static nonlinear input block, a linear dynamic block in the middle and another static nonlinear output block (Pham et al. 2019) (Fig. 5).
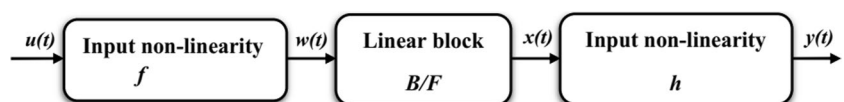
In the structure of HW model, the nonlinear model is converted into a piecewise linear function and then transformed to a nonlinear output function. In the HW:

$$w(t) = f(u(t)), \text{ nonlinear function for the input data} \tag{7}$$

$$x(t) = w(t)B/F, \text{ linear transfer function} \tag{8}$$

$$y(t) = f(x(t)) \tag{9}$$
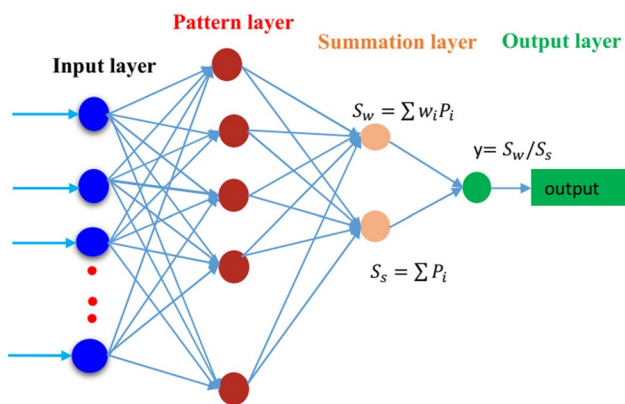
Fig. 5 The structure of the HW model

Fig. 6 The schematic diagram of GRNN model (Ji et al. 2017)

where u(t) is the input, y(t) is the output of the system, f is nonlinear function for the input, h nonlinearity for the output, w(t)and x(t) are internal variables that define the input and output in the linear block, respectively.

Generalized regression neural network (GRNN)

The GRNN is a type of radial basis function (RBF) that drives an estimator using only training data with back-propagation algorithm. The GRNN model has attractive and important property of being self-learning and capable of handling complex nonlinear problems. Modeling with GRNN can be done accurately without using large datasets (Ji et al. 2017). The GRNN model has the capability of solving problems regarding approximation, smooth functions, and can also provide consistent and accuracy prediction (Heddam 2014). Moreover, the GRNN model can be successfully used to solve various nonlinear and linear problems and can make accurate predictions without the need for large samples (Alilou and Yaghmaee 2015). The GRNN model is able to produce consistent predictions when the training data set is large and the estimation error approaches zero,
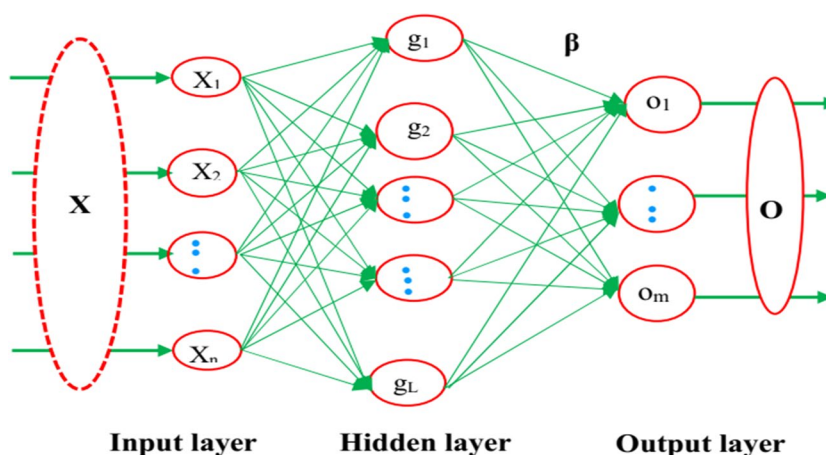
with only minor constraints on the function (Ji et al. 2017). Therefore, this model is characterized by a high learning speed, which leads to excellent results in the field of hydrological and environmental modeling. The GRNN structure comprises four layers: the input, pattern, summation and output layers as shown in Fig. 6. In the first layer, the number of input units are equal to parameter numbers. The input and pattern layers are fully connected, with each unit representing a training pattern and the output being a measure of the distance from the input to the stored patterns. The summation layer connects each pattern layer to two neurons (the S- and D-summation neurons). The D and S summation neurons compute the unweighted and weighted output of the pattern layer (Mehr et al. 2015).

Extreme learning Machine (ELM)

The ELM model is an improved training method proposed to address the shortcomings of the conventional FFNN with a single hidden layer, which is a gradient-based method (Niu et al. 2019a, b). According to Niu et al. (2019a, b), in ELM model, the theoretical basis is that if all the hidden node's activation function are noticeably differentiable, a simple linear system can be used to directly assign the weights of the hidden-output neurons after randomly determining the bias and the weights of the hidden and input neurons. In comparison to current training methods for conventional FFNN, the ELM model can mitigate a number of problems such as slow learning and stopping criteria while maintaining satisfactory generalization ability (Mundher et al. 2018). The limitation of ELM is that it is sensitive to selection of number of hidden neurons. The general sketch of the ELM model is shown in Fig. 7.

Assuming that the ELM network contains an input layer with n nodes, a hidden layer with L nodes, and an output layer with m nodes, the output of ELM model with regard to N samples $(x_i, t_i) \in R^m \times R^n$ can be written as:

Fig. 7 The structure of the ELM model

$$O_i = \sum_{i=0}^{L} \beta_i . g\left(w_i . x_i + b_i\right), i = 1, 2, \ldots \ldots, N \qquad (10)$$

Where O_i, $t_i \in R^m$ and $x_i \in R^n$, β_i, w_i, g, $w_i . x_i$ and b_i are the ELM output, the output and input vector of the ith sample, weight linking hidden node and output layer, weights linking input layer and hidden node, activation function, inner product of x_i and w_i, and the ith hidden neuron bias, respectively.

The assumption in ELM model is that all data samples can be approximated by the network output with zero error and thus the following equation always holds:

$$H\beta = T \qquad (11)$$

Where H and T are the hidden layer's output matrix with regard to N samples and the output of N data samples, respectively.

Data preprocessing and Model Evaluation Criteria

When modeling with black box models, both input and output data should first be normalized to bring all variables into the same range before feeding them into the models. This will ensure that all data receive equal attention, remove dimension and avoid data with small values being overshadowed by those in the upper number range (Nourani et al. 2019). In addition, data normalization simplifies numerical calculations in the model, which in turn increases modeling accuracy and reduces the time required to determine the local/global minimum. In this study, normalization of the data was performed using Eq. 12 to bring the value between 0 and 1:

$$X_n = \frac{X_i - X_{min}}{X_{max} - X_{min}} \qquad (12)$$

Where X_n, X_{max}, X_i and X_{min} represents the normalized, maximum, actual and minimum value of the dataset.

The accuracy of climatological and hydrological process simulation models must be assessed in both calibration and validation phases. To better evaluate the predictive performance of the models, at least one statistical error and one goodness-of-fit measure should be used (Nourani et al. 2018). Thus, in this study, Nash-Sutcliffe efficiency (NSE), mean absolute error (MAE), root mean square error (RMSE), percent bias (PBIAS) and coefficient of determination (R^2) were used to evaluate the accuracy of the proposed single, hybrid and ensemble models based on the recommendation of (Moriasi et al. 2015). The NSE usually takes values between -∞ and 1 and measures how well the computed value matches the actual runoff value. A perfect match between the computed and the actual data exists when the NSE value is 1. The closer the NSE is to 1, the more accurate the prediction. The RMSE, one of the

statistical error measures, was used in this study. Its value ranges from 0 to +∞, and a perfect model gives a RMSE value of 0. The MAE measures the actual error difference between the predicted and observed values by ignoring the influence of negative values. Low MAE value indicates accurate model prediction. The PBIAS indicator measures the average tendency of modeled value to be smaller or larger than the observed value. A positive PBIAS value indicates that the model tends to overestimate the observed values, while a negative PBIAS value indicates that the model tends to underestimate the observed values (Jimeno-Sáez et al. 2018). A value of zero indicates that the model has no bias and its predictions are unbiased. Coefficient of determination (R^2) describes the degree of association or collinearity between the predicted values and the observed values. It ranges from 0 to 1, with higher values indicating a better fit between the model and the data.

$$NSE = 1 - \frac{\sum_{i=1}^{N}\left(Q_{ob} - Q_{pr}\right)^2}{\sum_{i=1}^{N}\left(Q_{ob} \overline{-Q_{ob}}\right)^2} \quad -\infty < NSE \leq 1 \qquad (13)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{ob} - Q_{pr}\right)^2} \quad 0 \leq RMSE < \infty \qquad (14)$$

$$PBIAS = \left[\frac{\sum_{i}^{n}\left(Q_{ob} - Q_{pr}\right)100}{\sum_{i}^{n} Q_{ob}}\right] \quad -\infty \leq PBAS < \infty \qquad (15)$$

$$R^2 = \frac{\left(\sum_{i=1}^{n}\left(Q_{ob}\overline{-Q_{ob}}\right)\left(Q_{pre}\overline{-Q_{pr}}\right)\right)^2}{\sum_{i=1}^{N}\left(Q_{ob}\overline{-Q_{pr}}\right)^2 \sum_{i-1}^{n}\left(Q_{pre}\overline{-Q_{pr}}\right)^2} \quad 0 \leq R2 < 1 \qquad (16)$$

Ensemble techniques

Given similar datasets, one model may perform better than others, and if different datasets are used, the results of the models would be completely different. To take advantage of each model without neglecting the general nature of the data, the ensemble technique can be developed, which uses the output of each model as input, assigning a certain importance to each model using an arbitrator to offer the output (Kiran and Ravi 2008; Nourani et al. 2021a, b). The capability of ensemble technique in improving the overall prediction has been proven in several areas including hydrologic process, water resource, environment, regression and classification (Sharghi et al. 2018). The ensemble technique is a type of ML that is used to combine the results of multiple models to improve the final predictive performance (Elkiran et al. 2019). The main goal of the ensemble technique is to obtain more reliable and accurate estimates than would be possible by a single model. According to Kiran and Ravi (2008), ensemble techniques are categorized as linear and nonlinear

ensemble technique. The linear ensemble technique includes the linear ensembles by weighted median, weighted average and simple average. In the nonlinear ensemble technique, in the other hand, black-box models are used as nonlinear kernels to obtain an ensemble result. Abba et al. (2020) also divided ensemble techniques into two categories, heterogeneous and homogeneous ensembles. When ensemble unit consists of different learning algorithms, it is called heterogeneous, but when it consists of the same learning algorithms, it is defined as homogeneous. Nourani et al. (2018) recommended heterogeneous ensemble technique to achieve the prediction accuracy and overcome the model diversity. Therefore, in this study, in scenario 1 two linear i.e., SAE and WAE and two nonlinear ensemble techniques, such as neuro fuzzy ensemble (NFE) and Hammerstein-Weiner ensemble (HWE) were used to model the rainfall-runoff process of Katar catchment.

Linear ensemble

The SAE and WAE techniques are the most common model combination studies in hydrology and used as a reference to compare with the results of other ensemble technique such as neural ensemble. In the simple average ensemble (SAE) technique model combination is carried out by taking the arithmetic average of the runoff output of HW, GRNN, LSTM and ELM models as:

$$Q_t = \frac{1}{N} \sum_i^N Q_{ti} \tag{17}$$

Where Q_t, Q_{ti} and N represents the output of ensemble technique, the ith model output and number of single models (N-4).

Similarly, in the WAE technique, the ensemble output is calculated by assigning a unique different weight to each single model's output based on their relative importance. The ensemble prediction using WAE is calculated as:

$$Q_t = \sum_{i=1}^N w_i Q_{ti} \tag{18}$$

Where: w_i is the weight assigned to ith single model and calculated as:

$$w_i = \frac{NSE_i}{\sum_{i=1}^N NSE_i} \tag{19}$$

Where NSE is the Nash-Sutcliffe efficiency of the ith model.

Nonlinear ensemble technique

Nonlinear averaging, in the nonlinear ensemble technique, is performed by training nonlinear kernels such as ANN,

ANFIS and HW models using the runoff values of the individual models. Recently, the applicability of the nonlinear ensemble (e.g., neural network ensemble) has attracted attention in various fields of hydrological and environmental studies, e.g., vehicle traffic noise (Nourani et al. 2020a, b), seepage analysis of earth dam (Sharghi et al. 2018), suspended sediment load (Nourani et al. 2021a, b) and all have noted the better performance of the nonlinear ensemble over the single models. The above studies also recommended the use of other nonlinear kernels as an alternative for ensemble modeling. Therefore, this study proposes two nonlinear ensemble technique namely the neuro-fuzzy ensemble (NFE) and Hammerstein-Weiner ensemble (HWE) to improve the overall efficiency of rainfall-runoff modeling. The ANFIS model is a black-box model first developed by (Jang 1993) that combines the learning algorithm of neural networks (NN) and the reasoning capability of fuzzy interference systems (FIS). The hybrid between NN and FIS enables the ANFIS model to handle complex nonlinear hydrologic problems such as the rainfall-runoff process. Thus, in the NFE ensemble technique, the runoff values obtained from ELM, GRNN, HW and LSTM were fed into the ANFIS model input layer and the corresponding runoff value was determined by using different membership functions and a hybrid algorithm. The NFE technique has never been used as a model combination method in rainfall-runoff modeling. The NFE is chosen as a nonlinear ensemble because of its performance in previous studies in other fields such as suspended sediment load estimation(Nourani et al. 2021a, b) and particulate matter concentration prediction (Umar et al. 2021).

The other nonlinear ensemble technique used in this study was the HW ensemble (HWE) which is chosen because of its accuracy shown in modeling with a single model. Although the HW model has not previously been used in runoff modeling, this technique has shown tremendous potential in water resources research (Pham et al. 2019). The schematic of the proposed ensemble process is shown in Fig. 8.

Scenario 2: Hybrid Boosted Regression Tree (BRT) ensemble

The ANN model has been considered in previous studies as appropriate to deal with complex nonlinear problems. Nevertheless, recent studies have reported several difficulties and shortcomings in simulating hydrological processes with the traditional ANN and other ML-based models. As a result, researchers no longer rely solely on single ML-based models to capture the nonlinear nature of hydrologic processes. Therefore, higher prediction accuracy could be achieved by developing hybrid models (Abba et al. 2020). In this regard, this study proposed a novel hybrid BRT ensemble with four machine learning algorithms such as ELM, LSTM, GRNN, and HW in scenario 2.

Fig. 8 The schematic of the ensemble process
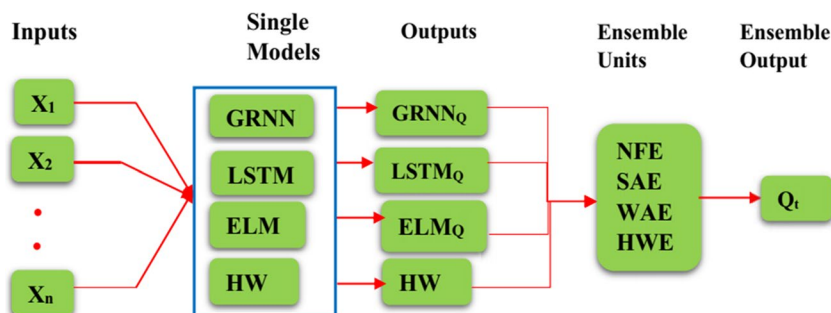
The BRT is a powerful ensemble tree for classification and prediction that is a combination of machine learning and statistical techniques. The BRT ensemble combines multiple models and merges them into a single model to boost the accuracy of each model in prediction problems (Youssef et al. 2016). The more advanced application of BRT is the simulation of natural phenomena where the input and output variables have a nonlinear relationship. Elimination of outliers or data transformation is not required in BRT before fitting the complex nonlinear relationship and establishing the interaction between predictor and output variables (Elith et al. 2008). The RBT model uses boosting and regression algorithm. Information is represented in decision trees in a way that is easy to visualize and intuitive, and has numerous other beneficial properties. According to Elith et al. (2008), surrogates are used by tree to modify the missing data in the input variable and are not sensitive to outliers. Boosting is a method used to increase the modeling accuracy based on the idea that it is easier to find many rough rules of thumb than a highly accurate single predictive rule. In the BRT model, fitting several trees overcomes the major limitation of the low predictive accuracy of single tree models (Youssef et al. 2016).

The RBT model itself, as an ensemble technique provides a suitable relationship between input and target variables. Hence, the proposed hybrid RBT with machine learning models applied in this study combines the best fit single model in the form of an ensemble tree. Although many hybrid models using optimization algorithms have been proposed to improve and evaluate the modeling accuracy, to the best of the author's knowledge, the hybrid combination of the ensemble method (i.e., RBT) with the ML-based models (HW-RBT, LSTM-RBT, ELM- RBT and GRNN-RBT) has not been used before.

Result and discussion

As mentioned earlier, this study has three objectives such as to develop four different ML models for rainfall-runoff modeling, improving the modeling accuracy using two nonlinear and two linear ensemble techniques, and finally to propose a hybrid BRT ensemble using the outputs of the single models. Thus, the results of sensitivity analysis, single models, ensemble techniques and the hybrid BRT model for rainfall-runoff modeling is presented in the following subsections.

Sensitivity analysis

Preprocessing of individual inputs is important in any hydrologic time series modeling because their selection can significantly affect the efficiency of individual models. An important step in hydrologic process modeling using ML-base techniques is the selection of appropriate input variables to feed to the various models (i.e., GRNN, ELM, HW and LSTM). The reason is that including large numbers of input parameters can lead to overfitting, which makes the modeling process complex and can lead to unrealistic results (Malik et al. 2022; Nourani et al. 2021a, b). Too few inputs, on the contrary, can reduce the accuracy of the modeling. In general, the factors (causal variables) for the rainfall-runoff process can be rainfall, temperature, lag runoff, and catchment characteristics. The input variables used in different studies often differ depending on the availability of data. Most studies used rainfall and lagged runoff as input variables (Adnan et al. 2019; Kisi et al. 2013), while some other studies also included additional factors such as evaporation or temperature as input variables (Hadi et al. 2019).).

Previous studies have shown that the current runoff value (Qt) are strongly correlated with its previous value (Kisi et al. 2012). Therefore, the inclusion of lag values in modeling can indirectly account for the influence of various factors affecting runoff formation. Pearson correlation has been used in previous studies to identify important inputs (Sharafati et al. 2020). Nourani et al. (2020a, b), however, criticized the applicability of the linear ensemble for the selection of dominant inputs in nonlinear hydrological processes. The strength of nonlinear sensitivity approach in selecting input variables has been demonstrated in several studies(Nourani et al. 2021a, b). Therefore, this study used the HW model was used for sensitivity analysis to determine the effects of input variables, i.e., rainfall (Pt, Pt-1, Pt-3) and lagged runoff (Qt-1, Qt-2, Qt-3, Qt-4, and Qt-5) on output (Qt). This

Table 2 Sensitivity analysis result

Input	RMSE (m³/s)	NSE	Rank
Qt-1	4.9699	0.865	1
Qt-2	6.7761	0.817	2
Qt-3	7.9649	0.772	3
Qt-4	8.9742	0.698	4
Pt	9.6859	0.554	5
Qt-5	14.493	0.493	6
Pt-1	16.345	0392	78
Qt-6	20.339	0.359	9
Pt-2	20.876	0.291	10
Pt-3	21.56	0.194	11

Table 4 Results of single models for rainfall-runoff modeling

Period	Performance measure	ELM	GRNN	HW	LSTM
Calibration	NSE	0.803	0.87	0.93	0.968
	R^2	0.803	0.877	0.932	0.972
	RMSE (m³/s)	8.452	6.925	5.046	3.41
	PBIAS (%)	−1.275	−0.33	−1.956	−4.19
	MAE (m³/s)	3.815	3.29	2.974	1.521
Validation	NSE	0.798	0.845	0.926	0.933
	R^2	0.802	0.855	0.927	0.937
	RMSE (m³/s)	9.088	7.794	5.546	5.308
	PBIAS (%)	2.64	4.2	−1.845	−4.41
	MAE (m³/s)	4.346	3.661	3.228	2.321

method is a single-input-single-output method in which one input variable at a time was fed into the HW model to simulate Qt. In this way, the relationship between the potential input variable and the output was determined without considering the influence of the other input variables and ranked based on their RMSE and NSE values, as shown in Table 2.

Table 2 shows that the inputs Q_{t-1} Qt-2, Q_{t-3} and Pt-1 had lowest RMSE value and were the most relevant inputs and ranked first, second, third and fourth, respectively. The inputs such as Q_{t-5}, P_{t-1}, Q_{t-6}, P_{t-2} and P_{t-3} were identified as less relevant and removed from the input combination set. Thus, the current runoff was modeled the by single ML-based models (GRNN, ELM, HW and LSTM) using different combinations of Q_{t-1} Qt-2, Q_{t-3}, Q_{t-4}, P_t and Q_{t-5} as input. Afterward, different input combinations with the dominant inputs were developed for predicting runoff at the current day (Q_t) using the four single models and only the best result obtained from each model were presented and discussed.

Results of the Single machine learning models

For each input combinations, the LSTM, GRRN, ELM and HW models were calibrated and validated and the performance of the models were evaluated using NSE, RMSE, R^2 and PBIAS and the best result is presented in this section. One of the most key tasks in modeling using ML models is to adjust hyperparameters to achieve maximum accuracy. It should be noted that the best structure of hyperparameters

for all four models was obtained by trial and error as shown in Table 3.

Performance comparison of the four single models were performed, and the results obtained are shown in Table 4. From the ML models, LSTM was found to be the best model for predicting rainfall-runoff, followed by HW, GRNN and ELM. The statistical performance indices of LSTM in terms of NSE, MAE, RMSE, and PBIAS was 0.968 and 0.933, 1.521(m³/s) and 2.321(m³/s), 3.41(m³/s) and 5.30 8(m³/s), −4.19 and −4.41, in the calibration and validation phase, respectively. The LSTM model provided the highest R^2 and NSE and the lowest RMSE and PBIAS compared to the other models. It improves the performances of HW, ELM and GRNN by 7.56%, 16.9% and 9.12% based on the validation phase NSE values. Based on the statistical performance guideline set by Jimeno-Sáez et al. (2018) and Moriasi et al. (2015), the LSTM model led very good modeling result. The superiority of LSTM model, as deep learning model, could be due to its high potential for extracting complex features from data due to their hierarchical structure compared to conventional ML models and achieve much better performance when the amount of data is sufficient (Rahimzad et al. 2021). In addition, according to Xiang et al. (2020), the high accuracy of the LSTM model could be due to its capability of considering between time series and capable of remembering information over long time of data such as seasonality, cyclical behavior and trends, which is not possible with the other computing models. After LSTM, HW model was

Table 3 Hyperparameters for single models

Model	Tuning parameter models
LSTM	Max epoch = 100; number of hidden units = 200; optimization = Adamax; learning rate = 0.01; Dropout rate = 0.4
ELM	Number of hidden nodes = 2; Activation function = sigmoid; Regularization parameter = 6;
HW	Nonlinearity = piecewise linear; max iteration = 50; tolerance = 0.001
GRNN	Hidden layers = 2; spread constant = 1; learning rate = 0.1

the next best model with the statistical performance indices in terms of NSE, MAE, RMSE, PBIAS and R^2 value of 0.926, 3.228m^2/s, 5.546 m^3/s, -1.845, and 0.92, respectively in the validation phase. The HW model provided the highest R^2 and NSE and the lowest RMSE and PBIAS compared to the GRNN and ELM. Indeed, the high performance of the HW model is not surprising, as it is an emerging nonlinear system identification model that has shown promising capabilities on highly complex problems (Abba et al. 2020; Gaya et al. 2017; Pham et al. 2019). The results show that the HW model was able to improve the prediction accuracy of the ELM, and GRNN models by 16% and 8.3%, respectively, based on the validation phase NSE value. Also, The PBIAS value of all the models were found to be negative in both the calibration and validation phases, which according to Moriasi et al. (2007), a negative PBIAS value indicates that the model is overestimating the observed values.

Different model performance measures have been used by different researchers, including statistical, graphical, or a combination of them. According to Harmel et al. (2014), it is important to use a combination of both statistical and graphical performance measures to get a robust assessment of model performance. This is because, some statistical performance measures, such as the NSE, can perform well even when low values are poorly fitted (Moriasi et al. 2015). In such cases, graphical measures can provide additional evidence of where the model performance is inadequate. Thus, this study used time series, scatter plot, flow duration curve (FDCs) and Taylor diagram to evaluate the performance of the developed models. In this regard, Fig. 9 shows the scatter plot and time series variation between observed and predicted runoff by single ML models.

As can be seen from Fig. 9g., the LSTM model appears to have better performance in runoff prediction, especially during dry periods of the year. In the LSTM model, the predicted hydrograph was much smoother and fits the general hydrograph trends (Fig. 9h) as well as the data points were closer to 1:1 line (Fig. 9g). However, the LSTM model as shown in the time series plot, failed to accurately catch the peak flows. Similar results of LSTM model's inefficiency in catching the peak flow have been reported in other studies (Rahimzad et al. 2021). The time series and scatter plot of HW model (Fig. 9e and f) shows that the data points are hardly deviate from the observed value and line. The time series plot of HW model also shows a consistent pattern with the observed hydrograph especially in peak flows, which is a significant performance compared to the other competing models. The series plot of ELM (Fig. 9b) and GRNN (Fig. 9d) models contained some noises especially in large flows. In addition, the two models (ELM and GRNN) generate more points deviating from the observed runoff values and with larger distances

to the ideal line compared to the LSTM and HW models. The statistical measure and graphical illustration results suggest that especially the LSTM and also HW models are able to capture the complex nonlinear nature of the rainfall-runoff process in both calibration and validation phase. The superiority of the LSTM model over the other competing models obtained in this study is supported by previous studies (Kaveh et al. 2021; Rahimzad et al. 2021; Yun et al. 2021). The HW model is applied in this study for the first time for rainfall-run off modeling also showed a promising performance. The HW model performance in modeling nonlinear problems have been reported in previous studies conducted in other fields (Abba et al. 2020; Pham et al. 2019) In general, the ELM model was found to be the least accurate model in both the calibration and validation phases.

Comparative evaluation was performed for the best individual models in the validation phase based on a two-dimensional Taylor diagram, as shown in Fig. 10. Taylor diagram summarizes and highlights several statistical indices such as standard deviation (SD) and correlation coefficient (r) between the actual and predicted values (Taylor 2001). From the figure, it can be seen that the better goodness of fit of runoff modeling was achieved with LSTM model with a value of r = 0.968, followed by HW (r = 0.962), GRNN (r = 0.924) and ELM (r = 0.875).

As stated by Jimeno-Sáez et al. (2018), the statistical performances, scatter and time series plots could not give an explicit performance comparison at different value intervals. This problem can be solved by applying flow duration curves (Tibangayuka et al. 2022). In this study, in addition to the statistical measures used to evaluate the performance of the models, the results were also analyzed using flow duration curves (FDCs) to visually compare and illustrate the differences between the measured and predicted runoff by single models. This provides a graphical representation of the model's performance and can help identify specific segments of flow where the model may be over or under estimating the runoff (Jimeno-Sáez et al. 2018). Thus, this study developed the FDCs for the single model as shown in Fig. 11. Pfannerstill et al. (2014) developed a method to improve the evaluation of the models' performance by dividing the FDCs into different segments. In this study, to evaluate the performance of the models at different phases of the hydrograph, the FDCs were divided into five segments as very low (>Q95), low (Q70-Q95), medium (Q20-Q70), high (Q5-Q20) and very high (<Q5) flows based on the recommendation of Jimeno-Sáez et al. (2018) and Pfannerstill et al. (2014); where Q_p denotes the runoff with probability of exceedance (%). As shown in Fig. 10, LSTM model provide better performance in low and very flow segment, GRNN model led better performance in medium flow

Fig. 9 Scatter and time series plot for single models in the validation phase
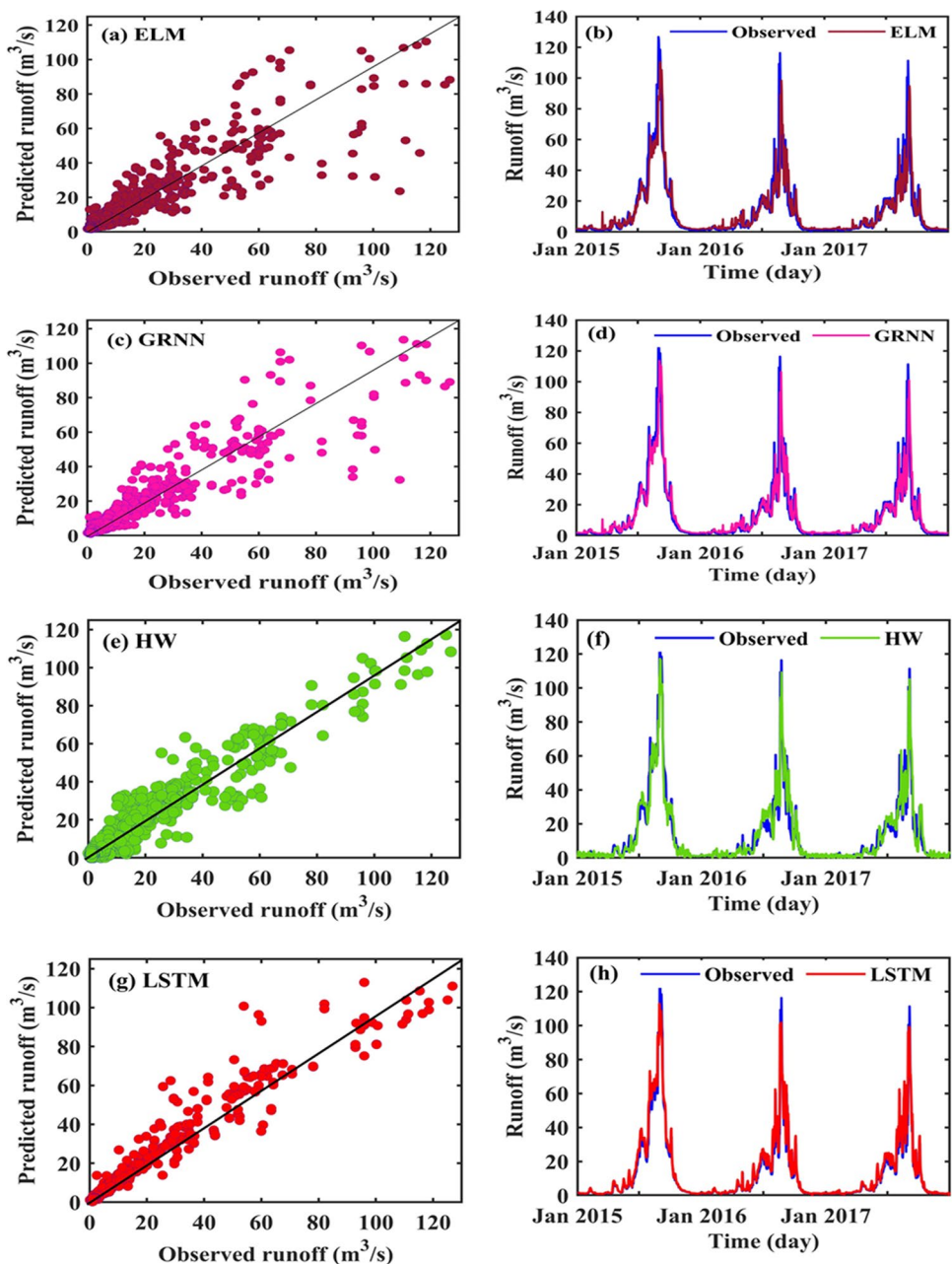
segment, LSTM and ELM provide better performance in low and very low flows. For example, the observed Q95 value was 1.0012m³/s while the predicted values by ELM, LSTM, HW and GRNN was 1.81 m³/s, 1.0896 m³/s, 0.46 m³/s and 1.592 m³/s, respectively. This indicates LSTM model predicts the Q95 runoff better than the others. Similarly, the observed Q20 value was 21.643 m³/s while the predicted values by ELM, LSTM, HW and GRNN was 20.526 m³/s, 24.452m³/s, 25.84 m³/s and 20.756 m³/s, respectively, indicating better modeling efficiency of GRNN and ELM. The observed Q5 value was 60.032 m³/s while the predicted values by ELM, LSTM, HW and GRNN was 54.268 m³/s,

64.696 m³/s, 59.767 m³/s and 54.6 m³/s, showing the superiority of HW model at this point.

A detail performance analysis of each model at different segment of the FDCs was also made based on the RMSE value (as shown Table 5). The RMSE values in the table indicate that the low and very low flows was better modeled using LSTM model. The GRNN was better at modeling medium flows (RMSE = 0.795 m³/s) and HW model was better at modeling very high flows with RMSE value of RMSE = 5.692 m³/s. Based on the RMSE value, the ELM model provides a better modeling result compared to the HW and LSTM models in medium and high flows. This
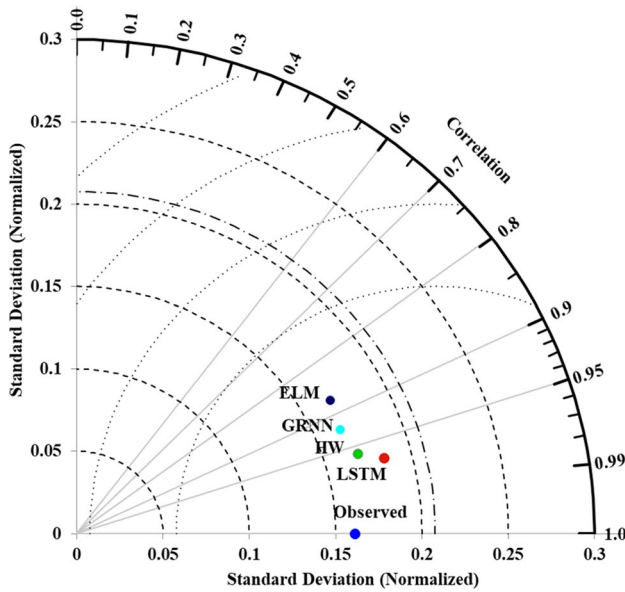
Fig. 10 Taylor diagram for single models in the validation phase

Table 5 The RMSE values (m³/s) of individual models in each hydrograph segment

Hydrograph phase	Flow range (m³/s)	HW	LSTM	ELM	GRNN
Very low flow	[0.115–0.995]	0.62	0.091	0.864	0.623
Low flow	[0.998–1.479]	0.239	0.144	0.831	0.614
Medium flow	[1.479–21.643]	1.6	1.291	0.997	0.795
High flow	[21.643–59.96]	3.391	4.131	2.805	2.151
Very high flows	[60.032–126.779]	5.692	6.301	8.385	7.485

the results of multiple models to obtain a more accurate estimate of the runoff.

Result of ensemble technique (scenario 1)

As stated earlier, the ensemble techniques presented in this study aim to improve the accuracy of the individual models (i.e., ELM, LSTM, GRNN and HW). To this end, the advantages of individual models are combined and the best results of each single models were considered as the subsequent input parameters for the ensemble technique. For the SAE and WAE techniques, modeling was conducted using Eqs. (17) and (18), respectively. Whereas, the two nonlinear ensemble models (HWE and NFE) were modeled by using a similar approach as the respective individual models. The results of the ensemble techniques are shown in Table 6. From the results, all linear ensemble techniques and HWE have higher performance than the individual models, with the exception of LSTM. The NFE ensemble technique improved all the models although the improvement for LSTM model was not significant. This leads to the conclusion that for modeling the rainfall-runoff process in Katar catchment, the ensemble technique

shows that even the least accurate model provides better result at a certain segment of the hydrograph.

The results of this analysis show that different models can lead to different performances at different segments in the hydrograph. Therefore, more accurate modeling of the rainfall-runoff process can be achieved by combining the results of individual models. For this purpose, ensemble modeling was developed in two scenarios in this study. In the first scenario, WAE, SAE, NFE and HW were used to combine the results of the individual models (LSTM, ELM, HW and GRNN). In the second scenario, a hybrid BRT ensemble technique was used to improve the overall modeling performance. These ensemble methods and hybrid BRT combine

Fig. 11 FDCs of single models in the validation phases
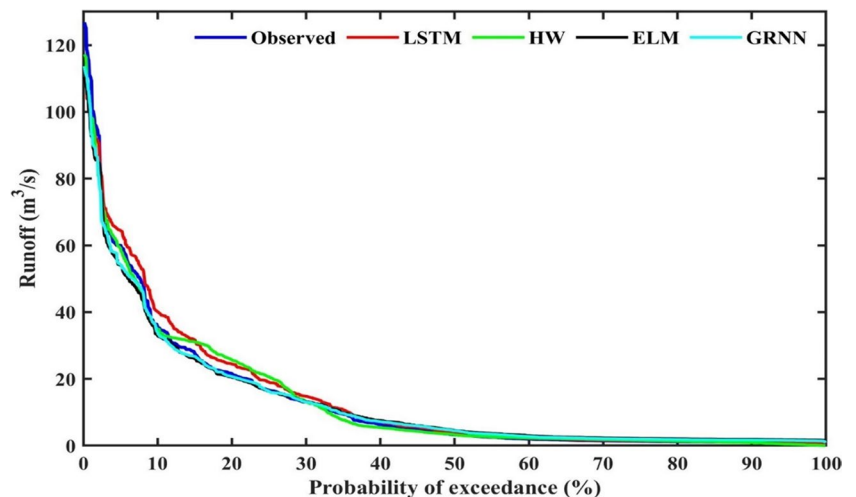
Table 6 The result of ensemble techniques (scenario 1)

Period	Hybrid models	R^2	RMSE (m³/s)	MAE (m³/s)	NSE	PBIAS (%)
Calibration	SAE	0.932	5.004	2.289	0.931	−2.76
	WAE	0.937	4.824	2.212	0.936	−2.964
	HWE	0.941	4.662	2.594	0.94	4.43
	NFE	0.973	3.155	1.343	0.973	−0.271
Validation	SAE	0.912	6.091	2.874	0.911	−2.388
	WAE	0.931	5.442	2.765	0.929	−2.554
	HWE	0.93	5.45	2.918	0.931	2.75
	NFE	0.957	4.232	1.867	0.957	−0.797

provides the most reliable result. The reason for the lower performance of the ensemble techniques compared to the LSTM model could be because of the weakness of the other individual models in the ensemble unit. According to Abba et al. (2020), the modeling performance of ensemble techniques depends on the accuracy and efficiency of each of the individual models. For example, in the SAE technique, the athematic average of all individual model's is used to generate the ensemble output. According to Nourani et al. (2020a, b), linear averaging gives values higher than the least performing models and lower than the most accurate model. In WAE technique, on the other hand, assigns weights to the outputs of individual model based on their relative importance to enhance the predictive accuracy. These phenomena may prove to be a weakness for prediction accuracy improvement of the ensemble techniques. Some studies have shown that ensemble techniques perform less than a single model (Abba et al. 2020). The statistical performance indices in Table 6 shows that the WAE ensemble technique has slightly better accuracy compared to SAE. The NFE technique provides the best result with the value of NSE = 0.973 and 0.957, MAE = 1.3432.918 m³/s and 2.918 m³/s, PBIAS = -0.271 and 0.797 and RMSE = 3.155 m³/s and 4.232 m³/s, respectively in calibration and validation phases. The NFE technique improved the performance of ELM, LSTM, GRNN and HW models by 19.9%, 2.57%, 13.25% and 3.35%, respectively, based on the validation phase NSE value. The

NFE technique has not been used in rainfall-runoff study, but its robustness has been demonstrated in several studies conducted in other field (Nourani et al. 2020a, b; Nourani et al. 2021a, b; Umar et al. 2021). The superiority of NFE could be due to that it is hybrid learning algorithm that combines the advantages of both ANN and FIS and able to handle the nonlinear and complex rainfall-runoff process.

Hybrid BRT ensemble result (scenario 2)

In the scenario 1 of ensemble modeling, the best ensemble technique (NFE) couldn't show a significant improvement for LSTM (best single) model. Thus, this study developed a hybrid of the RBT ensemble (LSTM-BRT, ELM-BRT, HW-BRT, and GRNN-BRT) to enhance the accuracy and compare with the ensemble techniques (senarion1) discussed in Section 3.3. Similar to the single and ensemble models, the prediction accuracy of the developed hybrid models was evaluated using NSE, R^2, RMSE, MAE, and PBIAS. The results of the hybrid BRT models in both the calibration and validation phase are shown in Table 7. As can be seen in Table 7, LSTM-RBT outperformed all the other hybrid models with the highest value of NSE = 0.987 and 0.978, R^2 = 0.987 and 0.981, and the lowest values of RMSE = 2.999 m³/s and 2.199 m³/s, MAE = 1.034 m³/s and 1.438 m³/s and PBIAS of 0.117% and 0.75% in the calibration and validation phases, respectively. Based on the statistical performance indices, it was proved the superiority

Table 7 The result of hybrid BRT models

	Hybrid models	R^2	RMSE (m³/s)	MAE (m³/s)	NSE	PBIAS (%)
Calibration	HW-BRT	0.979	3.142	1.314	0.976	4.4
	GRNN-BRT	0.926	5.175	2.484	0.926	2.085
	LSTM-BRT	0.987	2.199	1.034	0.987	0.117
	ELM-BRT	0.898	7.335	3.834	0.872	−2.26
Validation	HW-BRT	0.97	3.707	2.449	0.967	0.385
	GRNN-BRT	0.93	5.506	2.861	0.927	0.5
	LSTM-BRT	0.981	2.999	1.438	0.978	0.75
	ELM-BRT	0.87	7.471	3.709	0.866	2.76

of all hybrid models despite the better prediction accuracy of LSTM-RBT. The hybrid LSTM-BRT model increased the accuracy of the LSTM, HW, GRNN and ELM model by 4.82%, 6.59%, 16.8% and 23.68%, respectively based on the NSE value in the validation phase. The ELM-BRT, HW-BRT, and GRNN-BRT increased the performance (based on NSE value) of their counterpart in the single model by 8.52%, 9.7% and 4.43%, respectively in the validation phase.

Figure 13 shows the scatter plots and time series between the observed and predicted values of the four-hybrid ensemble models. From Fig. 13, it can be seen that the agreement between the observed and predicted runoff values was achieved in the following order: LSTM-BRT > HW-BRT > GRNN-BRT > ELM-BRT. Moreover, the performance of the hybrid models developed in the study are considered very good in modeling the rainfall-runoff modeling based on the guidelines set by Moriasi et al. (2015), as the NSE, R^2 and PBIAS value for all models is greater than 0.80, 0.85 and less than ±5, respectively in both calibration and validation phases.

Generally, the comparison of modeling accuracy between the ensemble and hybrid models (i.e., Scenario 1 and 2) showed that the best hybrid BRT ensemble model (i.e., LSTM-RBT) outperformed all four-ensemble techniques (WAE, SAE, NFE and HWE). According to Abba et al. (2020), using ensemble of different models like BRT with other highly promising nonlinear models such as ELM, LSTM, GRNN and HW can further enhance the robustness of the BRT model. This is because, combining multiple models helps to reduce the variance and bias of the predictions, resulting in a more accurate and robust model. Another reason could be due to the BRT model uses an ensemble of different decision tree models, which allows it to fit complex nonlinear relationships and automatically address interaction effects between the predictions. This results in high accuracy for the BRT model. Similarly, the statistical performance comparison results showed that HWE and NFE techniques were superior to the other three hybrid BRT morsels (i.e., ELM-BRT, LSTM-BRT and GRNN-BRT). A more detailed comparison of the observed and predicted runoff values using the single models, ensemble techniques and the hybrid BRT models revealed the importance of using both the ensemble techniques and the hybrid BRT models to increase the prediction performance of the single models. Similar to the single and ensemble models, the performance of the hybrid models was also compared using scatter and time series plots as shown in Fig. 13. From the figure, the scatter plot of predicted hybrid BRT models specially LSTM-BRT and HW-BRT modes were very close to the actual runoff values and the 1:1 line showing their accurate prediction. The ELM-BRT model provide the worst performance compared to the other ensemble techniques. As shown in Fig. 12a, the data points of ELM-BRT model

are more scattered and far from the diagonal line specially in large runoff values. Also, Fig. 13b shows the ELM-BRT model follow the pattern of the hydrograph well during low flow periods, but not in the high flow period of the flow of the years.

Fig. 12a-h provides visual representation using scatter and time series plots of the four different ensemble techniques (SAE, WAE, HW-E and NFE) to compare their performance in the validation phase. The time series plot can show the deviation of ensemble techniques from the actual value at different time. The figure shows that NFE led the best agreement and pattern with observed runoff and catches the high and low flows with high accuracy. In the linear ensemble techniques (SAE and WAE), the data points deviate the corresponding observed values and unable to catch the peak flows. The ensemble results of this study were compared with findings by Nourani et al. (2020a, b), Nourani et al. (2021a, b) and Umar et al. (2021) and found fair similarity.

The performance of ensemble technique and hybrid BRT models were also compared using box-plot as shown in Fig. 14. The boxplot in Fig. 14 investigates the variability of observed runoff as compared to those obtained from ensemble and hybrid models using their quartile and interquartile ranges (IQR). Figure 14a depicts that better consistent result was found between the variability of runoff value predicted by NFE (IQR = 14.92 m^3/s) and observed value (IQR = 15.19 m^3/s). Among the hybrid BRT models (Fig. 14b), the most consistent value was obtained by HW-BRT model (IQR = 15.08 m^3/s).

Conclusions

In the current study, the performance of four ML models such as ELM, LSTM, HW and GRNN were evaluated for modeling of the rainfall-runoff process in Katar catchment, Ethiopia. Subsequently, to improve the accuracy of the single models, four different ensemble techniques such as SAE, WAE, HWE and NFE (Scenario 1) as well as hybrid BRT models (Scenario 2) were used separately. The modeling performance of the developed models was evaluated and compared using statistical performance indices (NSE, R^2, MAE, PBIAS and RMSE), FDCs, scatter plots, time series and Tylor diagram. Prior to the modeling, a sensitivity analysis between input variables and the output variable (Qt) was performed using a nonlinear input variable selection method. In summary, the results of the single ML models showed that the LSTM model performed best in modeling the rainfall-runoff process, followed by the HW, GRNNN, and ELM models. The LSTM model had

Fig. 12 Scatter and time series plot of observed and predicted runoff by ensemble techniques in the validation phase
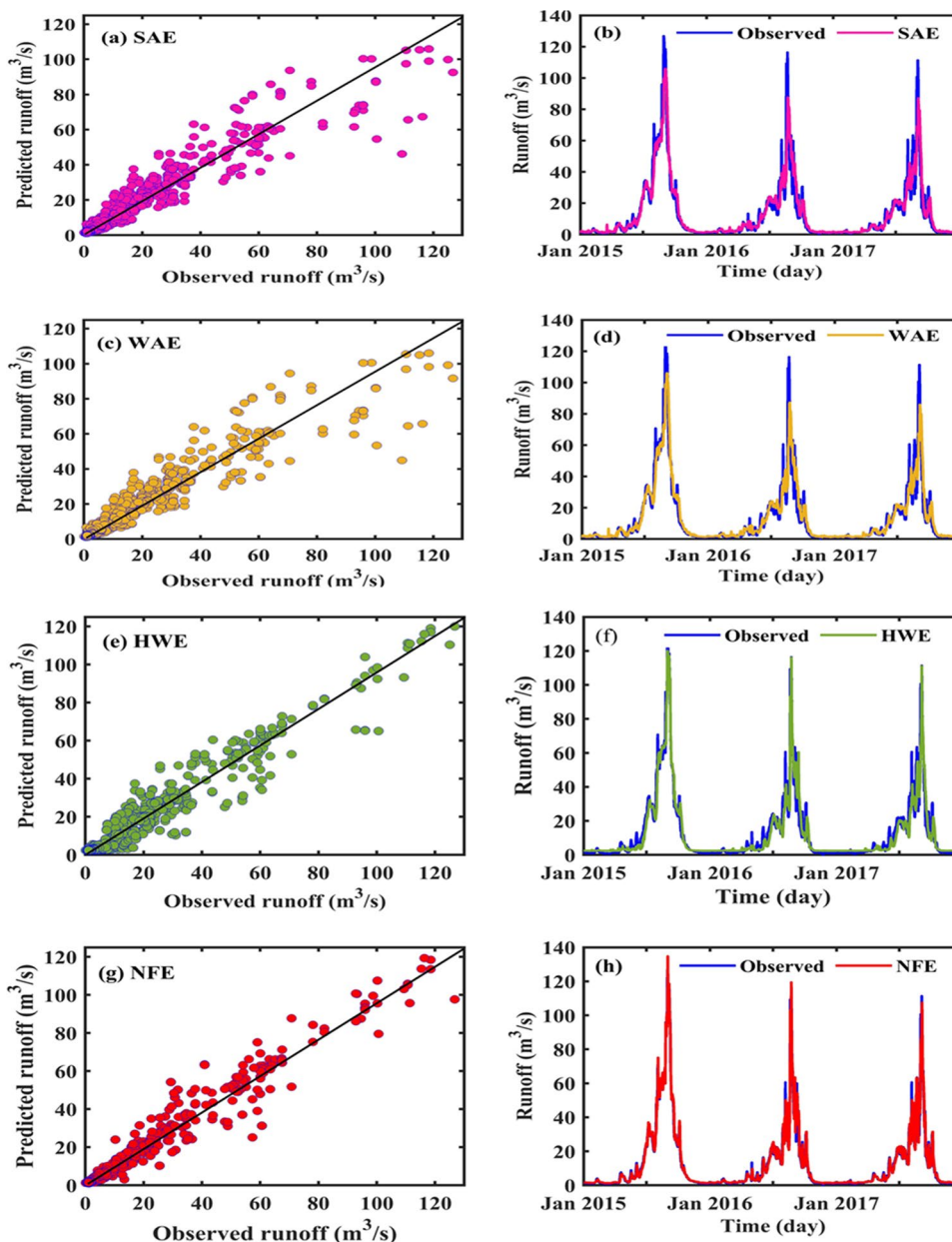
a performance accuracy that was 10.41%, 0.76%, 16.92%, 2.4% and 0.4% higher than the GRNN, HW, ELM, SAE and WAE models based on the NSE value in the validation phase. Among the four ensemble techniques, the NFE ensemble had the best performance in both calibration and validation phases. The NFE technique improved the performance of ELM, HW and GRNN by reducing the RMSE value by 52.41%, 22%, and 44.5%, respectively, in the validation phase. Also, the best ensemble technique in scenario 1 (NFE), had 2.793%, 3.014%, and 5.05% higher NSE value than the HWE, WAE and SAE technique, respectively, in the validation phase. Moreover, the developed hybrid models (LSTM-BRT, GRNN-BRT, ELM-BRT and

HW-BRT) all had better performance than their counterpart single models with the superior performance obtained by the LSTM-BRT model. The overall results of this study show the promising effect of the ensemble techniques and specially hybrid BRT models for predicting the rainfall-runoff process in the Katar catchment in Ethiopia. Future studies should evaluate the proposed hybrid ensemble techniques for other hydrologic processes modeling. Moreover, the study also recommended the use of other alternatives such as new deep learning models and optimization algorithms in conjunction with the promising ensemble models (e.g., random forest) in future studies to improve the overall modeling accuracy.

Fig. 13 Scatter and time series plot of observed and hybrid model's runoff in the validation phase
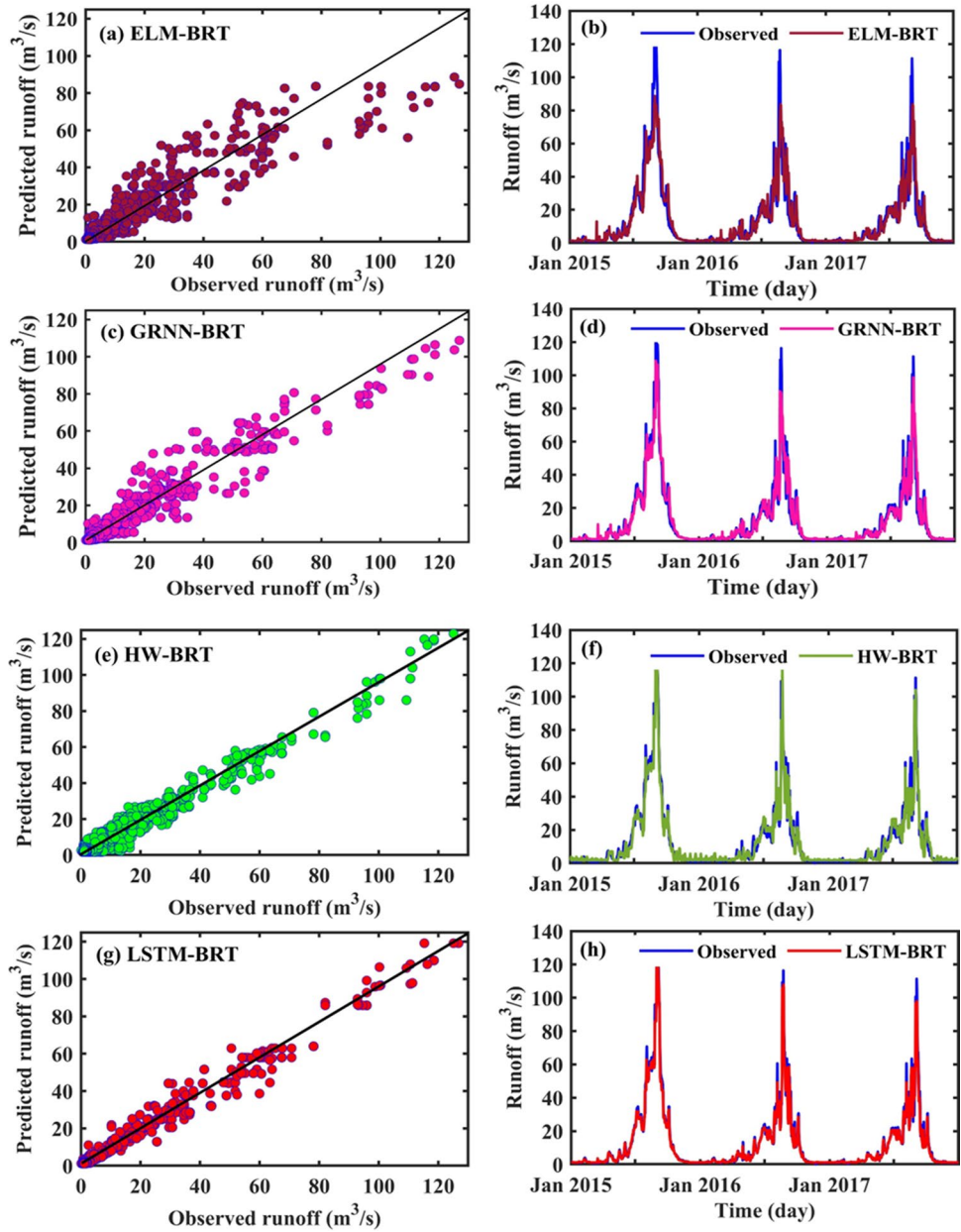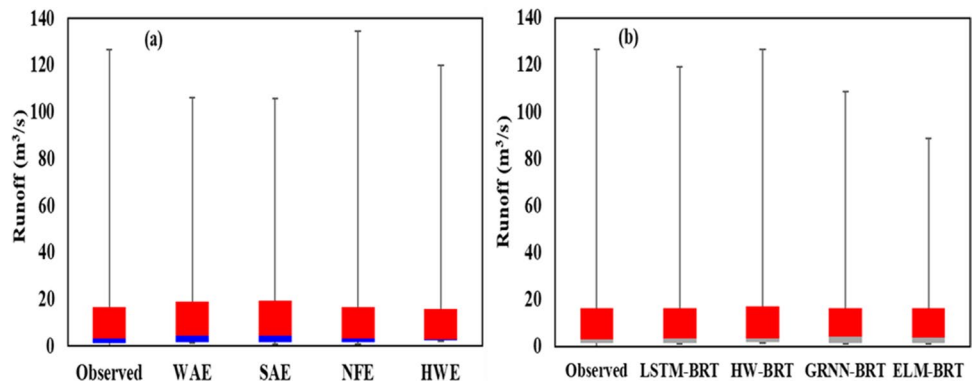

Fig. 14 The comparison of the box plot for (**a**) the ensemble technique and (**b**) hybrid BRT models

Declarations

References

Abba SI, Linh NTT, Abdullahi J, Ali SIA, Pham QB, Abdulkadir RA, Costache R, Nam VT, Anh DT (2020) Hybrid machine learning ensemble techniques for modeling dissolved oxygen concentration. IEEE Access 8:157218–157237. https://doi.org/10.1109/ACCESS.2020.3017743

Abedi R, Costache R, Shafizadeh-Moghadam H, Pham QB (2022) Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees. Geocarto Int 37(19):5479–5496. https://doi.org/10.1080/10106049.2021.1920636

Abinayadhevi, P, Prasad, SJS (2015) Identification of pH process using Hammerstein-Wiener model. Proceedings of 2015 IEEE 9th International Conference on Intelligent Systems and Control, ISCO 2015, 1–5. https://doi.org/10.1109/ISCO.2015.7282297

Adnan RM, Liang Z, Trajkovic S, Zounemat-Kermani M, Li B, Kisi O (2019) Daily streamflow prediction using optimally pruned extreme learning machine. J Hydrol 577(July):123981. https://doi.org/10.1016/j.jhydrol.2019.123981

Adnan RM, Petroselli A, Heddam S, Santos CAG, Kisi O (2021) Short term rainfall-runoff modelling using several machine learning methods and a conceptual event-based model. Stoch Env Res Risk A 35(3):597–616. https://doi.org/10.1007/s00477-020-01910-0

Alilou VK, Yaghmaee F (2015) Application of GRNN neural network in non-texture image inpainting and restoration. Pattern Recogn Lett 62:24–31. https://doi.org/10.1016/j.patrec.2015.04.020

Asadi H, Shahedi K, Jarihani B, Sidle R (2019) Rainfall-Runoff Modelling Using Hydrological Connectivity Index and Artificial Neural Network Approach. Water 11(2):212. https://doi.org/10.3390/w11020212

Bengio Y, Simard P, Frasconi P (1994) Learning Long-Term Dependencies with Gradient Descent is Difficult. IEEE Trans Neural Netw 5(2):157–166. https://doi.org/10.1109/72.279181

Bhattacharjee NV, Tollner EW (2016) Improving management of windrow composting systems by modeling runoff water quality dynamics using recurrent neural network. Ecol Model 339:68–76. https://doi.org/10.1016/j.ecolmodel.2016.08.011

Cai, QC, Hsu, TH, Lin, JY (2021) Using the general regression neural network method to calibrate the parameters of a sub-catchment. Water (Switzerland), 13(8). https://doi.org/10.3390/w13081089

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. J Anim Ecol 77(4):802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Elkiran G, Nourani V, Abba S (2019) Multi-step ahead modelling of river water quality parameters using ensemble artificial intelligence-based approach. J Hydrol 577(June):123962. https://doi.org/10.1016/j.jhydrol.2019.123962

Gaya MS, Zango MU, Yusuf LA, Mustapha M, Muhammad B, Sani A, Tijjani A, Wahab NA, Khairi MTM (2017) Estimation of turbidity in water treatment plant using hammerstein-wiener and neural network technique. Indon J Electric Eng Comput Sci 5(3):666–672. https://doi.org/10.11591/ijeecs.v5.i3.pp666-672

Gers FA, Schmidhuber J, Cummins F (2000) Learning to Forget: Continual Prediction with LSTM. Neural Comput 12(10):2451–2471

Hadi SJ, Abba SI, Sammen SSH, Salih SQ, Al-Ansari N, Yaseen MZ (2019) Non-linear input variable selection approach integrated with non-tuned data intelligence model for streamflow pattern simulation. IEEE Access 7:141533–141548. https://doi.org/10.1109/ACCESS.2019.2943515

Harmel RD, Smith PK, Migliaccio KW, Chaubey I, Douglas-Mankin KR, Benham B, Shukla S, Muñoz-Carpena R, Robson BJ (2014) Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: A review and recommendations. Environ Model Softw 57:40–51. https://doi.org/10.1016/j.envsoft.2014.02.013

Heddam S (2014) Generalized regression neural network (GRNN)-based approach for colored dissolved organic matter (CDOM) retrieval: case study of Connecticut River at Middle Haddam Station, USA. Environ Monit Assess 186(11):7837–7848. https://doi.org/10.1007/s10661-014-3971-7

Himanshu SK, Pandey A, Yadav B (2017) Ensemble wavelet-support vector machine approach for prediction of suspended sediment load using hydrometeorological data. J Hydrol Eng 22(7):05017006

Jang JSR (1993) ANFIS : Adaptive-Network-Based Fuzzy Inference System. IEEE Trans Syst Man Cybern 23(3):665–685. https://doi.org/10.1109/21.256541

Ji X, Shang X, Dahlgren RA, Zhang M (2017) Prediction of dissolved oxygen concentration in hypoxic river systems using support vector machine: a case study of Wen-Rui Tang River. China Environ Sci Pollut Res 24(19):16062–16076. https://doi.org/10.1007/s11356-017-9243-7

Jimeno-Sáez, P, Senent-Aparicio, J, Pérez-Sánchez, J, Pulido-Velazquez, D (2018) A comparison of SWAT and ANN models for daily runoff simulation in different climatic zones of peninsular Spain. Water (Switzerland), 10(2). https://doi.org/10.3390/w10020192

Kaveh K, Kaveh H, Bui MD, Rutschmann P (2021) Long short-term memory for predicting daily suspended sediment concentration. Eng Comput 37(3):2013–2027. https://doi.org/10.1007/s00366-019-00921-y

Khan MYA, Hasan F, Tian F (2019) Estimation of suspended sediment load using three neural network algorithms in Ramganga River catchment of Ganga Basin. India Sustain Water Resource Manag 5(3):1115–1131. https://doi.org/10.1007/s40899-018-0288-7

Kiran RN, Ravi V (2008) Software reliability prediction by soft computing techniques. J Syst Softw 81(4):576–583. https://doi.org/10.1016/j.jss.2007.05.005

Kisi O, Dailr AH, Cimen M, Shiri J (2012) Suspended sediment modeling using genetic programming and soft computing techniques. J Hydrol 450–451:48–58. https://doi.org/10.1016/j.jhydrol.2012.05.031

Kisi O, Shiri J, Tombul M (2013) Modeling rainfall-runoff process using soft computing techniques. Comput Geosci 51:108–117. https://doi.org/10.1016/j.cageo.2012.07.001

Koch J, Schneider R (2022) Long short-term memory networks enhance rainfall-runoff modelling at the national scale of Denmark. GEUS Bull 49:1–7. https://doi.org/10.34194/geusb.v49.8292

Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M (2018) Rainfall-runoff modelling using Long Short-Term Memory (LSTM)

networks. Hydrol Earth Syst Sci 22(11):6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Lakmini Prarthana Jayasinghe WJM, Deo RC, Ghahramani A, Ghimire S, Raj N (2022) Development and evaluation of hybrid deep learning long short-term memory network model for pan evaporation estimation trained with satellite and ground-based data. J Hydrol 607(December 2021):127534. https://doi.org/10.1016/j.jhydrol.2022.127534

Li X, Sha J, Wang Z (2019) Comparison of daily streamflow forecasts using extreme learning machines and the random forest method. Hydrol Sci J 64(15):1857–1866. https://doi.org/10.1080/02626667.2019.1680846

Liu P, Wang J, Sangaiah AK, Xie Y, Yin X (2019) Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment. Sustain 11(2058):1–14. https://doi.org/10.3390/su11072058

Malik A, Jamei M, Ali M, Prasad R, Karbasi M, Yaseen ZM (2022) Multi-step daily forecasting of reference evapotranspiration for different climates of India: A modern multivariate complementary technique reinforced with ridge regression feature selection. Agric Water Manag 272(March):107812. https://doi.org/10.1016/j.agwat.2022.107812

Mehr AD, Kahya E, Şahin A, Nazemosadat MJ (2015) Successive-station monthly streamflow prediction using different artificial neural network algorithms. Int J Environ Sci Technol 12(7):2191–2200. https://doi.org/10.1007/s13762-014-0613-0

Mohammadi, B, Safari, MJS, Vazifehkhah, S (2022) IHACRES, GR4J and MISD-based multi conceptual-machine learning approach for rainfall-runoff modeling. In Scientific Reports (Vol. 12, Issue 1). https://doi.org/10.1038/s41598-022-16215-1

Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Trans ASABE 50(3):885–900. https://doi.org/10.13031/2013.23153

Moriasi DN, Gitau MW, Pai N, Daggupati P (2015) Hydrologic and water quality models: Performance measures and evaluation criteria. Trans ASABE 58(6):1763–1785. https://doi.org/10.13031/trans.58.10715

Mundher Z, Jaafar O, Deo RC, Kisi O, Adamowski J, Quilty J, Elshafie A (2016) Stream-flow forecasting using extreme learning machines : A case study in a semi-arid region in Iraq. J Hydrol 542:603–614. https://doi.org/10.1016/j.jhydrol.2016.09.035

Mundher Z, Mohammed Y, Allawi F, Yousif AA, Jaafar O, Mohamad F, Ahmed H (2018) Non-tuned machine learning approach for hydrological time series forecasting. Neural Comput & Applic 30(5):1479–1491. https://doi.org/10.1007/s00521-016-2763-0

Niu W, Feng Z, Feng B, Min Y, Cheng C (2019a) Comparison of Multiple Linear Regression, Artificial Neural Network, Extreme Learning Machine, and Support Vector Machine in Deriving Operation Rule of Hydropower Reservoir. Water 11(1):88. https://doi.org/10.3390/w11010088

Niu W, Feng Z, Zeng M, Feng B, Min Y (2019b) Forecasting reservoir monthly runoff via ensemble empirical mode decomposition and extreme learning machine optimized by an improved gravitational search algorithm. Appl Soft Comput J 82:105589. https://doi.org/10.1016/j.asoc.2019.105589

Noori N, Kalin L (2016) Coupling SWAT and ANN models for enhanced daily streamflow prediction. J Hydrol 533:141–151. https://doi.org/10.1016/j.jhydrol.2015.11.050

Nourani V, Elkiran G, Abba SI (2018) Wastewater treatment plant performance analysis using artificial intelligence – an ensemble approach. Water Sci Technol 78(10):2064–2076. https://doi.org/10.2166/wst.2018.477

Nourani V, Elkiran G, Abdullahi J (2019) Multi-station artificial intelligence based ensemble modeling of reference evapotranspiration using pan evaporation measurements. J Hydrol 577(June):123958. https://doi.org/10.1016/j.jhydrol.2019.123958

Nourani V, Elkiran G, Abdullahi J (2020a) Multi-step ahead modeling of reference evapotranspiration using a multi-model approach. J Hydrol 581(October 2019):124434. https://doi.org/10.1016/j.jhydrol.2019.124434

Nourani V, Gökçekuş H, Umar IK (2020b) Artificial intelligence based ensemble model for prediction of vehicular traffic noise. Environ Res 180(October 2019):108852. https://doi.org/10.1016/j.envres.2019.108852

Nourani, V, Gokcekus, H, Gelete, G (2021a) Estimation of Suspended Sediment Load Using Artificial Intelligence-Based Ensemble Model. Complexity, 2021(Article ID 6633760), 19. https://doi.org/10.1155/2021/6633760

Nourani V, Gökçekuş H, Gichamo T (2021b) Ensemble data-driven rainfall-runoff modeling using multi-source satellite and gauge rainfall data input fusion. Earth Sci Inf 14(4):1787–1808. https://doi.org/10.1007/s12145-021-00615-4

Nourani V, Khodkar K, Gebremichael M (2022) Uncertainty assessment of LSTM based groundwater level predictions. Hydrol Sci J 67(5):773–790. https://doi.org/10.1080/02626667.2022.2046755

Park, S, Kim, J (2019) Landslide susceptibility mapping based on random forest and boosted regression tree models, and a comparison of their performance. Appl Sci (Switzerland), 9(5). https://doi.org/10.3390/app9050942

Pfannerstill M, Guse B, Fohrer N (2014) Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. J Hydrol 510:447–458. https://doi.org/10.1016/j.jhydrol.2013.12.044

Pham QB, Abba SI, Usman AG, Linh NTT, Gupta V, Malik A, Costache R, Vo ND, Tri DQ (2019) Potential of Hybrid Data-Intelligence Algorithms for Multi-Station Modelling of Rainfall. Water Resour Manag 33(15):5067–5087. https://doi.org/10.1007/s11269-019-02408-3

Phukoetphim P, Shamseldin, Asaad Y, Adams, Keith (2016) Multi-model Approach Using Neural Networks and Symbolic Regression to Combine the Estimated Discharges of Rainfall-Runoff Models. J Hydrol Eng 21(8):1–18. https://doi.org/10.1061/(ASCE)HE.1943-5584

Rahimzad, M, Moghaddam Nia, A, Zolfonoon, H, Soltani, J, Danandeh Mehr, A, Kwon, HH (2021) Performance Comparison of an LSTM-based Deep Learning Model versus Conventional Machine Learning Algorithms for Streamflow Forecasting. In Water Resources Management (Vol. 35, Issue 12, pp. 4167–4187). https://doi.org/10.1007/s11269-021-02937-w

Sharafati A, Seyed H, Asadollah SB, Motta D, Yaseen ZM (2020) Application of newly developed ensemble machine learning models for daily suspended sediment load prediction and related uncertainty analysis. Hydrol Sci J 65(12):1–21. https://doi.org/10.1080/02626667.2020.1786571

Sharghi E, Nourani V, Behfar N (2018) Earthfill dam seepage analysis using ensemble artificial intelligence based modeling. J Hydroinf 20(5):1071–1084. https://doi.org/10.2166/hydro.2018.151

Taormina R, Chau K (2015) Data-driven input variable selection for rainfall – runoff modeling using binary-coded particle swarm optimization and Extreme Learning Machines. J Hydrol 529:1617–1632. https://doi.org/10.1016/j.jhydrol.2015.08.022

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 106(D7):7183–7192. https://doi.org/10.1029/2000JD900719

Tibangayuka N, Mulungu DMM, Izdori F (2022) Evaluating the performance of HBV, HEC-HMS and ANN models in simulating streamflow for a data scarce high-humid tropical catchment in Tanzania. Hydrol Sci J 67(14):1–14. https://doi.org/10.1080/02626667.2022.2137417

Umar IK, Nourani V, Gökçekuş H (2021) A novel multi-model data-driven ensemble approach for the prediction of particulate matter concentration. Environ Sci Pollut Res 28(36):49663–49677. https://doi.org/10.1007/s11356-021-14133-9

Xiang Z, Yan J, Demir I (2020) A Rainfall-Runoff Model With LSTM-Based Sequence-to-Sequence Learning. Water Resour Res 56(1):1–17. https://doi.org/10.1029/2019WR025326

Yin H, Zhang X, Wang F, Zhang Y, Xia R, Jin J (2021) Rainfall-runoff modeling using LSTM-based multi-state-vector sequence-to-sequence model. J Hydrol 598(October 2020):126378. https://doi.org/10.1016/j.jhydrol.2021.126378

Young CC, Liu WC, Wu MC (2017) A physically based and machine learning hybrid approach for accurate rainfall-runoff modeling during extreme typhoon events. Appl Soft Comput J 53:205–216. https://doi.org/10.1016/j.asoc.2016.12.052

Youssef AM, Pourghasemi HR, Pourtaghi ZS, Al-Katheeri MM (2016) Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region. Saudi Arabia Landslides 13(5):839–856. https://doi.org/10.1007/s10346-015-0614-1

Yun D, Abbas A, Jeon J, Ligaray M, Baek SS, Cho KH (2021) Developing a deep learning model for the simulation of micro-pollutants in a watershed. J Clean Prod 300:126858. https://doi.org/10.1016/j.jclepro.2021.126858

Zhang D, Skullestad E, Lindholm G, Ratnaweera H (2018) Hydraulic modeling and deep learning based flow forecasting for optimizing inter catchment wastewater transfer. J Hydrol 567:792–802. https://doi.org/10.1016/j.jhydrol.2017.11.029

Zhang X, Zwiers FW, Li G, Wan H, Cannon AJ (2017) Complexity in estimating past and future extreme short-duration rainfall. Nat Geosci 10(4):255–259. https://doi.org/10.1038/ngeo2911