Intercomparing the robustness of machine learning models in simulation and forecasting of streamflow

Parthiban Loganathan and Amit Baburao Mahindrakar

ABSTRACT

The intercomparison of streamflow simulation and the prediction of discharge using various renowned machine learning techniques were performed. The daily streamflow discharge model was developed for 35 observation stations located in a large-scale river basin named Cauvery. Various hydrological indices were calculated for observed and predicted discharges for comparing and evaluating the replicability of local hydrological conditions. The model variance and bias observed from the proposed extreme gradient boosting decision tree model were less than 15%, which is compared with other machine learning techniques considered in this study. The model Nash–Sutcliffe efficiency and coefficient of determination values are above 0.7 for both the training and testing phases which demonstrate the effectiveness of model performance. The comparison of monthly observed and model-predicted discharges during the validation period illustrates the model's ability in representing the peaks and fall in high-, medium-, and low-flow zones. The assessment and comparison of hydrological indices between observed and predicted discharges illustrate the model's ability in representing the baseflow, high-spell, and low-spell statistics. Simulating streamflow and predicting discharge are essential for water resource planning and management, especially in large-scale river basins. The proposed machine learning technique demonstrates significant improvement in model efficiency by dropping variance and bias which, in turn, improves the replicability of local-scale hydrology.

Key words | Cauvery river basin, climate change, hydrological model, machine learning, streamflow

Parthiban Loganathan
Amit Baburao Mahindrakar (corresponding author)
School of Civil Engineering,
VIT University,
Vellore,
Tamilnadu 632014,
India
E-mail: *amahindrakarlab@gmail.com*

HIGHLIGHTS

- The credibility of machine learning models in representing the regional-scale hydrology is performed.
- Evaluation to prioritize model selection for river basin management.
- Season-based approach in evaluating model performance in local hydrology.
- Hydrological indices were inter-compared for high-, medium-, and low-flow zones.
- Outcome delivers valuable suggestions to decision-makers in the planning of future water resources.

doi: 10.2166/wcc.2020.365

INTRODUCTION

The human population makes use of global runoff up to 54% for various purposes such as consumption, extraction, and instream flow needs (Andreadis *et al.* 2007). Moreover, the estimation of global streamflow is highly uncertain because of limitations in observation and reachability. The simulation and forecasting of streamflow is a primary necessity in water resource planning and management (Hashim *et al.* 2016; Kersbergen 2016; Adnan *et al.* 2019b; Choubin *et al.* 2019). The forecasting of river flow with higher accuracy is essential for early hazard mapping and management which benefits a huge population and socio-economic activities (Wu & Chau 2013; Taormina & Chau 2015; Hussain & Khan 2020; Shamshirband *et al.* 2020). Further, the forecast will help in minimizing potential risks of flood and droughts, water supply for urban areas, irrigation planning for agricultural purposes, and also hydro-power projects (Londhe & Charhate 2010; Fotovatikhah *et al.* 2018; Adnan *et al.* 2020a; Homsi *et al.* 2020). An important issue in hydrological streamflow time-series prediction has been a greater concern in the past few decades.

Numerous models were proposed for forecasting and simulating the river discharge in various parts of the globe, especially data-driven models which had an upper hand over physical conceptual models due to their ease and computational efficiency (Wu *et al.* 2009; Diop *et al.* 2018; Adnan *et al.* 2019a; Alizamir *et al.* 2020). However, it is difficult to find a model that performs equally well for low-, medium-, and high-flow zones. Thus, the forecasting of streamflow becomes more complex and makes it difficult to create a real-time early warning system (Rezaie-Balf & Kisi 2018; Yaseen *et al.* 2019b; Adnan *et al.* 2020b; Li *et al.* 2020). In this concern, there is a need for a new forecasting approach that will be effective as well as efficient in predicting reliable and accurate data. In recent times, several researchers suggested that machine learning models predict streamflow with various significant approaches (Rezaie-balf *et al.* 2017; Kaya *et al.* 2019; Keum *et al.* 2020; Tikhamarine *et al.* 2020). These learning algorithms are data-driven models with the ability to learn the local environment and respond based on the scenarios with high accuracy.

In recent decades, various machine learning algorithms were proposed by researchers for predicting streamflow with decent performance. Previous studies suggested renowned machine learning techniques such as generalized linear model (GLM; Asong *et al.* 2016), partial least-squared regression (PLS; Matulessy *et al.* 2015), neural network (NNET; Coulibaly *et al.* 2005), K-nearest neighbor (KNN; Devak *et al.* 2015; Sekhar *et al.* 2018), and principle component regression (PCR; Sahriman *et al.* 2014), which are better for representing the local hydrological process. However, most of the machine learning techniques perform well in forecasting during the training period but fail to do the same in the testing period (Ghorbani *et al.* 2018; Yuan *et al.* 2018; Naganna *et al.* 2019; Yaseen *et al.* 2019a). The trick of handling bias and high variance in streamflow is still not resolved which clearly shows the overfitting issues associated with machine learning algorithms.

Though there are numerous machine learning techniques which perform better in streamflow projection, research scientists are facing issues in handling the drawbacks and improvising the model performance. Unfortunately, no technique overcomes all the drawbacks as we are still exploring methods to accurately model the local hydrological process. The present study proposes the Extreme Gradient Boosting Decision Tree (EXGBDT) approach for comparing its performance with other traditional models and validating it through the evaluation of various hydrological indices. The present study aims to predict river discharge with the help of daily weather parameters such as precipitation, average temperature, maximum temperature, and minimum temperature. The intercomparison of data-driven hydrological models was performed with renowned machine learning techniques and a proposed method to attain a low bias and variance in monthly streamflow prediction.

Numerous hydrological studies over the Indian subcontinent have previously been performed (Kale *et al.* 2010; Bhuvaneswari *et al.* 2013; Bhave *et al.* 2018; Arulbalaji & Padmalal 2020). However, most of the studies focused on the sub-basin-level and station-level discharge prediction. The current study deals with a large-scale river basin named Cauvery river basin located in southern peninsular India, which has frequent flood and drought issues. The study basin is one of the essential rivers in the southern part of India which provides water supply to a huge urban community for domestic use and enormous agricultural land area for irrigation purposes. Therefore, it is essential to model the streamflow and forecast the discharge pattern throughout the tributaries of the river basin. It is essential to build an individual model that performs equally well at low-, medium- and high-discharge stations to reduce the computational burden. Thus, an intercomparison of various machine learning model performances is carried out to select an optimum model and validated through the evaluation of multiple hydrological indices. The key objectives of the present study are (1) to improve the quality of observed hydrological time-series data by handling missing values and (2) to develop a hydrological model to perform equally well at low-, medium-, and high-flow zones at a large-scale river basin.

STUDY AREA

Geography

The current study was conducted over the Cauvery river basin, which is located over the southern peninsular region of the Indian subcontinent. The basin extends over 75°27′E to 79°54′E and 10°9′N to 13°30′N and lies over three states and one union territory. The river originates in Karnataka and meets the sea at Tamil Nadu passing through Kerala and Pondicherry. The total drainage area of the basin is 85,626 km^2, and the overall length of the river is 802 km. The boundary map representing the extent of the Cauvery river basin is presented in Figure 1. The river is confined by the Western Ghats and the Eastern Ghats on the west and east, respectively. The key portion of the river basin is concealed with cultivated land and forest, and it is also known as the rice bowl of South India. The water depletion in the basin has increased by up to 40% in the past few decades (Raju *et al.* 2013; Madolli *et al.* 2015). The risk of drought is high during the dry seasons, and the risk of flood is high during monsoon seasons in the basin area.

Climate

The Cauvery river basin is known for its tropical and sub-tropical climate zones where the north-west region is colder than the rest of the basin. The basin has four seasons, namely winter (December to February), summer (March to June), south-west monsoon (July to September), and north-east monsoon (October to November) (Bhuvaneswari *et al.* 2013; Madolli *et al.* 2015). The basin remains dry during summer and winter, which contributes a longer period of the year, and the monsoon season brings rainfall to the entire basin (Solaraj *et al.* 2010; Bhave *et al.* 2018). April is the hottest month, whereas January is the coldest

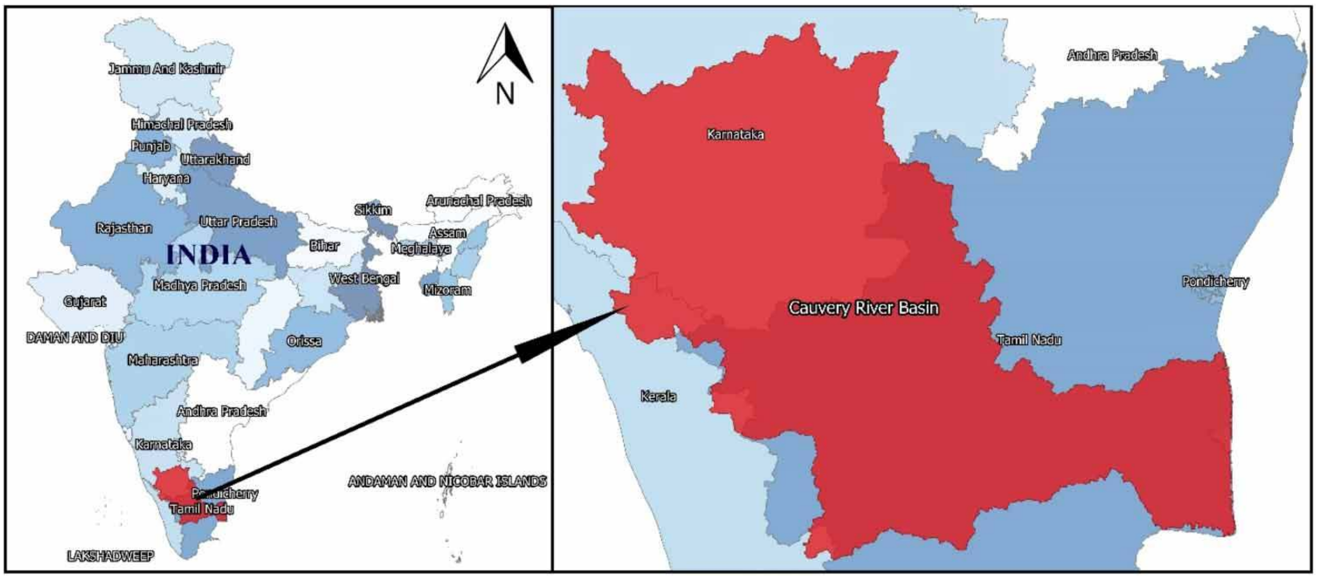

Figure 1 | Cauvery river basin boundary map.

month of the whole basin, and the average monthly temperature ranges from 18 to 33 °C (Nadu & Nadu 1981; Sunil *et al.* 2010). The basin is further classified into the upper, middle, and lower basins for a better comparison of climate variability and river flow discharge patterns within the basin. Further, it will help compare the different flow patterns in high-, medium-, and low-flow regions.

DATASETS

Observed data

Meteorological data are essential for predicting the streamflow of the river basin. Meteorological datasets include daily precipitation (rainfall) and temperature (minimum, maximum, and average). There are three main organizations in India which record meteorological parameters which are (1) India Meteorological Department (IMD), (2) Central Water Commission (CWC), and (3) Indian Space Research Organization (ISRO) Automatic Weather Stations. CWC has established 35 stations located in the basin to recognize the atmospheric and river dynamics relationship. The hydro-meteorological and river flow data from these 35 daily observed stations positioned in the Cauvery river basin from 1951 to 2015 are collected. The description of the observation stations situated in the Cauvery river basin is presented in Table 1. The classification of the Cauvery river basin, observation stations, and river line is mapped in Figure 2. The details of station numbers provided in Figure 2 are explained in Table 1.

The historical observed data for the study area is collected concerning 35 observation stations from 1950 to 2015. Further, the entire time-series data are divided into the calibration period (1950–2000) and the validation period (2001–2015) for better consideration and evaluation of the model performance. The selected weather parameters and their short name, description, and units are presented in Table A1 (Appendix). The classification of the Cauvery basin into the upper, middle, and lower basins for a better comparison of river flow discharge patterns within the basin is illustrated in Figure 3. The framework adopted in this study is presented in the following section.

Table 1 | Cauvery river basin observation stations description

S. No.	Station	Station ID	Latitude	Longitude	S. No.	Station	Station ID	Latitude	Longitude
Upper Cauvery river vasin					Middle Cauvery river basin				
1	Akkihebbal	1	12°36′10″	76°24′3″	1	Biligundulu	4	12°10′48″	77°43′48″
2	Bendrehalli	3	12°2′8″	77°0′53″	2	E-Managalam	6	11°1′59″	77°53′31″
3	Chunchunkatte	5	12°30′25″	76°18′0″	3	Hogenakkal	8	12°7′15″	77°47′7″
4	K.M.Vadi	9	12°20′32″	76°17′15″	4	Kanakpura	10	12°32′41″	77°25′37″
5	Kollegal	12	12°11′17″	77°5′59″	5	Kodumudi	11	11°5′5″	77°53′18″
6	Kudige	13	12°30′6″	75°57′40″	6	Kudlur	14	11°50′26″	77°27′45″
7	M.H.Halli	15	12°49′9″	76°8′2″	7	Musiri	17	10°56′40″	78°26′1″
8	Sakleshpur	24	12°57′8″	75°47′12″	8	Muthankera	18	11°50′49″	76°7′15″
9	T.Narasipur	28	12°13′54″	76°53′29″	9	Nalammaranpatti	19	10°52′54″	77°59′3″
10	Thimmanahalli	33	12°58′56″	76°2′16″	10	Nellithurai	21	11°17′17″	76°53′29″
Lower Cauvery river basin					11	Savandapur	25	11°31′22″	77°30′24″
1	Annavasal	2	10°58′21″	79°45′27″	12	Sevanur	26	11°33′16″	77°42′52″
2	Gopurajapuram	7	10°51′4″	79°48′0″	13	T.Bekuppe	27	12°30′58″	77°26′15″
3	Menangudi	16	10°56′55″	79°42′19″	14	T.K.Halli	29	12°25′0″	77°11′33″
4	Nallathur	20	10°59′28″	79°47′18″	15	Thengumarahada	31	11°34′21″	76°55′8″
5	Peralam	22	10°58′10″	79°39′38″	16	Thevur	32	11°31′42″	77°45′6″
6	Porakudi	23	10°54′13″	79°42′27″	17	Thoppur	34	11°56′18″	78°3′18″
7	Thengudi	30	10°54′56″	79°38′21″	18	Urachikottai	35	11°28′43″	77°42′0″

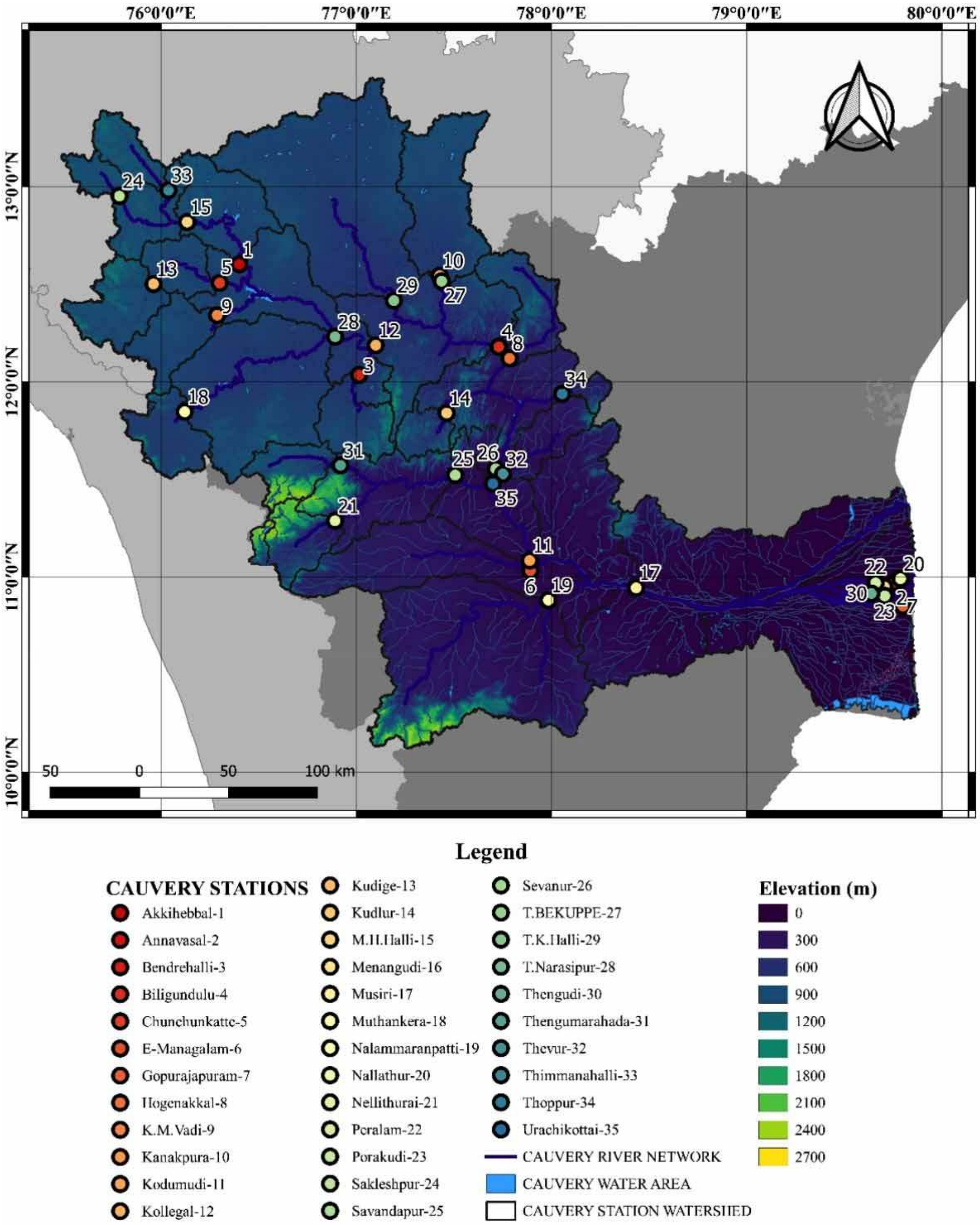

Figure 2 | Cauvery river basin elevation and observation stations.

METHODOLOGY

The proposed framework for building a data-driven hydrological model for simulating and forecasting streamflow in the Cauvery river basin is represented in Figure 4. The initial steps involve the collection of data for the study area which includes meteorological data and discharge data. The station-wise observed weather parameters (pr, tas, tasmax, and tasmin) are collected for the assigned baseline period of 1951–2005. For the same baseline period, the observed streamflow data for 35 stations along the Cauvery river basin are extracted. The collected discharge data are imputed for missing values using the weather data. Further, the collected data are divided into calibration (75%) and validation (25%)

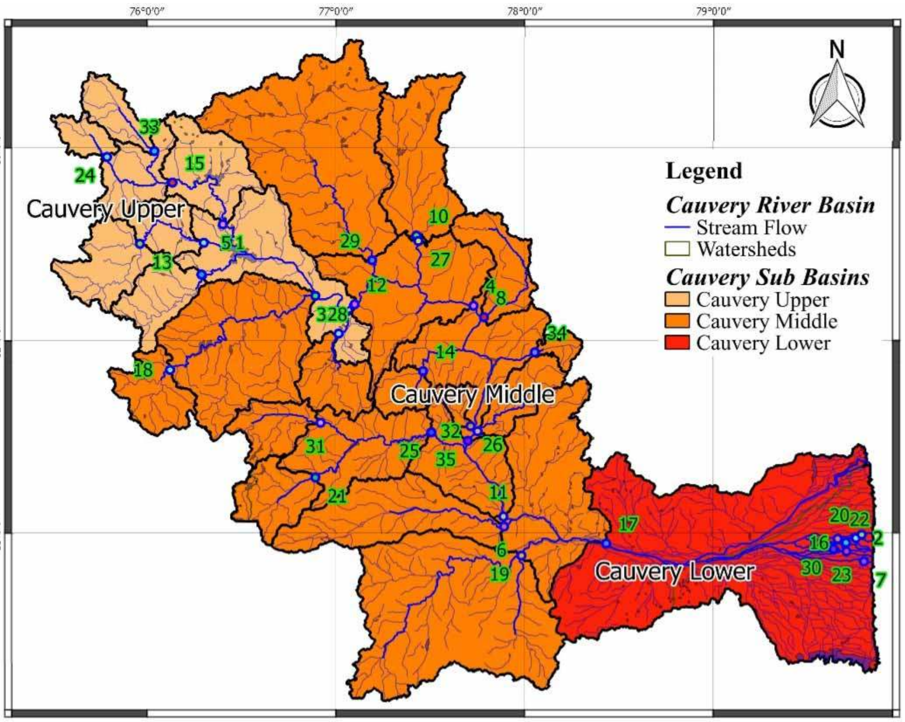

Figure 3 | Cauvery river basin classification.

datasets, i.e. 1951–1990 and 1991–2005, respectively. Later, the data-driven models are built using the selected machine learning models and proposed models for comparison of performance. The performance of the various models is evaluated by various performance evaluation parameters such as normalized root-mean-squared error (NRMSE %), percentage bias (PBIAS %), Nash–Sutcliffe efficiency (NSE), and coefficient of determination (R^2) for both calibration and validation periods. Further, the better performing model is selected based on the evaluation and hydrological indices which are calculated to compare with the actual observed data.

Extreme gradient boosting decision tree

The extreme gradient boosting method combines weak learners into a strong learner by performing multiple iterations. The main objective of the algorithm is to teach a model to predict the target by reducing the mean-squared error (MSE) of the prediction (Georganos *et al.* 2018), which can be represented in the common equation as follows:

$$\hat{y} = F(x) \tag{1}$$

where $\mathrm{MSE} = 1/n \sum_i (\hat{y}_i - y_i)^2$, $\hat{y}_i$ is the predicted value of $F(x)$, y_i is the observed value, and n is the number of samples in y. Consider a gradient boosting algorithm with N stages at each stage $n\ (1 \le n \le N)$ of gradient boosting. Where an imperfect model F_n for low n, this model can be simply represented as $\hat{y}_i = \bar{y}$ (mean of y). So, to improve F_n, the algorithm adds some new estimators, i.e. $h_n(x)$.

$$F_{n+1}(x) = F_n(x) + h_n(x) = y \quad \text{(or)} \quad h_n(x) = y - F_n(x) \tag{2}$$

Thus, the gradient boosting estimator will fit the residual. Further, F_{n+1} tries to specify the errors F_n.

$$L_{\mathrm{MSE}} = \frac{1}{2}(y - F(x))^2 \tag{3}$$

$$h_n(x) = -\frac{\partial L_{\mathrm{MSE}}}{\partial F} = y - F(x) \tag{4}$$

Thus, the gradient boosting could be generalized to a gradient descent algorithm for different loss and its gradient. In most of the supervised learning algorithms, the output variable y with the input variable x is represented as joint probability distribution $P(x,y)$, where the training set

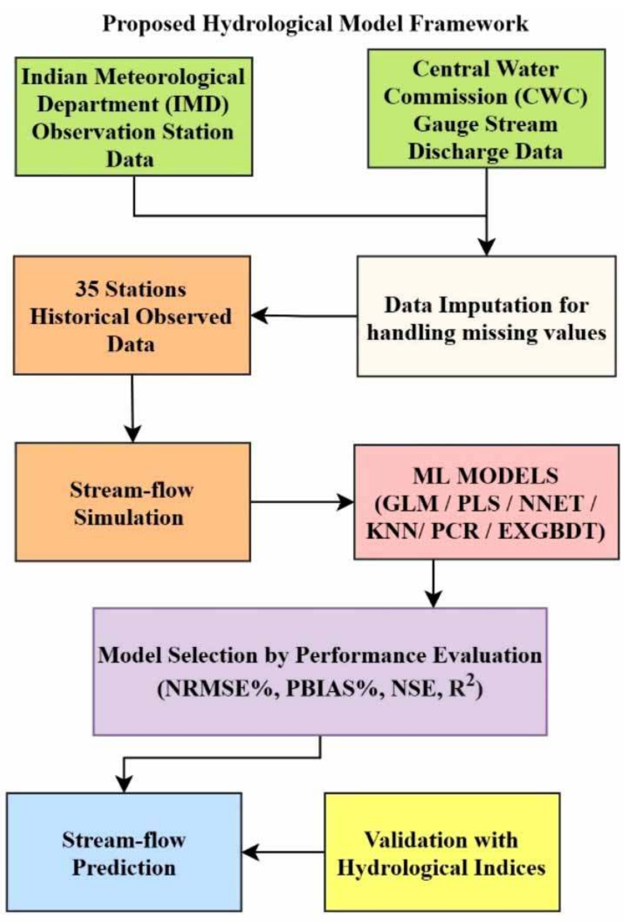

Figure 4 | Model selection and validation for streamflow prediction.

$\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with the known x value and the corresponding y value. The target is to find $\hat{F}(x)$ for a function $F(x)$ which reduces the loss with a loss function $L(y, F(x))$

$$\hat{F} = \arg\min_{F} \mathbb{E}_{x,y}[L(y, F(x))] \tag{5}$$

The gradient boosting method adopts known y and finds $\hat{F}(x)$ by a weighted sum of $h_i(x)$ from class H, known as weak learners:

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const} \tag{6}$$

For empirical risk minimization, the technique tries to find $\hat{F}(x)$ reduces loss function for the training data. It is attained by a base model with constant function $F_0(x)$, and additively increasing greedily:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \tag{7}$$

$$F_m(x) = F_{m-1}(x) + \underset{h_m \in \mathcal{H}}{\arg\min} \left[\sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h_m(x_i)) \right] \tag{8}$$

where $h_m \in H$ is a base learner function.

The complexity lies in the high computation requirement for optimizing loss function L for choosing the beat function h. Thus, a simplified approach is carried out by applying a steepest descent step to minimize the problem. Considering a continuous case where H is a set of arbitrary differentiable functions on R and the model can be updated as follows:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i)) \tag{9}$$

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} L(y_i, F_{m-1}(x_i))) \tag{10}$$

where the functions F_i derivatives are taken for $i \in \{1, \ldots, m\}$ and the step length is γ_m.

Hydrological indices

Insight into the streamflow model can be obtained by evaluating various hydrological statistics such as baseflow, high-, and low-spell statistics (Ladson *et al.* 2013; Ward 2013; Booker 2015). In this study, various hydrological indices which enlighten in-depth details of discharge at a selected basin were evaluated. Initially, the station-wise hydrological indices are calculated using the observed streamflow and later, these indices are compared with the simulated discharge to access the ability of the model in representing the local scenarios (Van Der Velde *et al.* 2013; Piras *et al.* 2016). Various hydrological indices considered and evaluated in this study using Hydrostats R package are given in Table A2 (Appendix).

RESULTS AND DISCUSSION

The intercomparison of machine learning models in the robustness of simulation and forecasting of streamflow is performed. The datasets are processed as mentioned in the framework and model results are presented concerning models' calibration and validation for a better understanding of model performance. The performance of selected models and their ability in representing the local conditions are discussed in the following sections.

Intercomparison of machine learning models

The historical station observed daily discharge and weather parameters considered in this study for the selected duration of 1951–2015 (65 years) are converted into time-series data. Further, the data are split into training data 1951–2000 (50 years) and testing data 2001–2015 (15 years). The interannual variability of observed data for precipitation and discharge for three different stations (Chunchunkatte, T.K.Halli, and Peralam from the upper, middle, and lower basins, respectively) from each sub-basin is presented in Figure 5. The plot clearly shows the annual trend precipitation and its respective discharge amount. There is a significant drop in discharge trend, especially in the lower Cauvery river basin over the past few decades. This is possibly due to rapid urbanization and amplified riverbed sand mining.

The streamflow for 35 observation stations is modeled using station observed precipitation, average, minimum, and maximum temperature data. The simulations were made using GLM, PLS, NNET, KNN, PCR, and the proposed EXGBDT model for the calibration period and predicted for the validation phase. The performance of each model is evaluated using the selected performance evaluation parameters and the observations are given in Table 2. The table compared the performance of each model at the calibration and validation phases. The evaluation parameters clearly state that the performance of models during the validation phase is slightly lower than the calibration. It is also evident that the proposed EXGBDT model performs exceptionally well compared to other machine learning models. The variance of the

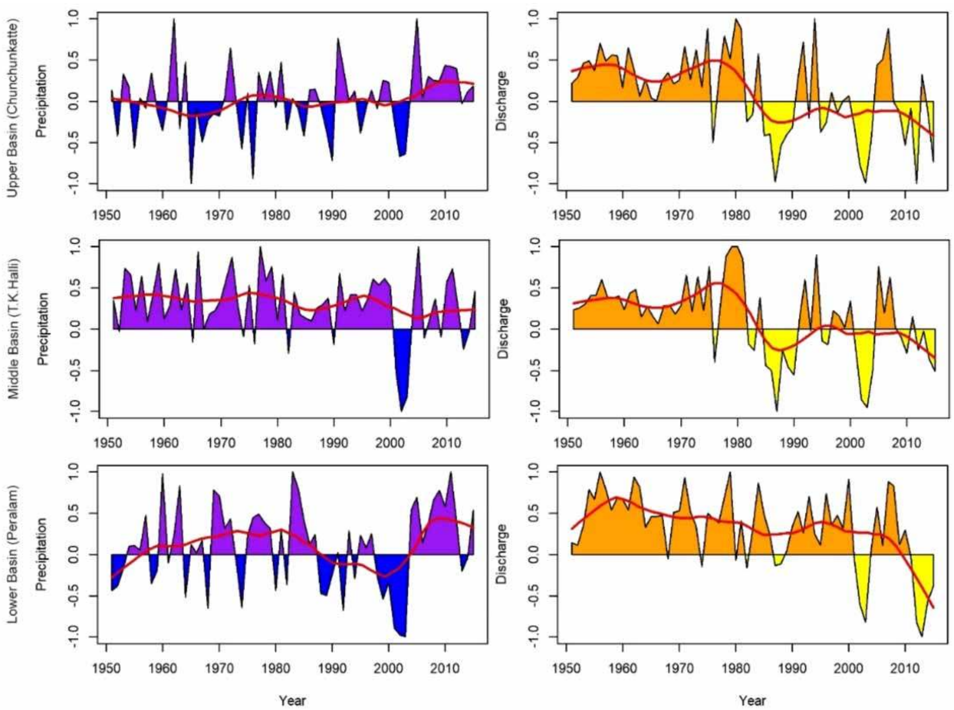

Figure 5 | Interannual variability of precipitation and streamflow.

Table 2 | Intercomparison of performance evaluation

		Calibration						Validation					
Sub-basin	**PEP**	**GLM**	**PLS**	**NNET**	**KNN**	**PCR**	**EXGBDT**	**GLM**	**PLS**	**NNET**	**KNN**	**PCR**	**EXGBDT**
Upper basin	NRMSE	12.1	12.3	13.0	13.5	14.1	6.4	26.2	28.0	29.9	34.0	31.4	13.1
	PBIAS	0.0	0.0	0.5	1.9	0.0	0.0	14.7	16.2	17.1	20.1	17.3	4.0
	NSE	0.7	0.7	0.7	0.8	0.6	0.9	0.4	0.3	0.3	0.4	0.3	0.8
	R^2	0.7	0.7	0.7	0.8	0.6	0.9	0.5	0.4	0.4	0.4	0.4	0.8
Middle basin	NRMSE	16.6	19.1	19.1	19.1	21.6	10.2	28.6	30.5	32.4	37.6	33.0	15.5
	PBIAS	0.0	0.0	0.2	0.9	0.0	0.0	12.3	12.0	12.8	13.5	13.4	1.4
	NSE	0.6	0.5	0.5	0.6	0.4	0.8	0.3	0.3	0.3	0.3	0.3	0.7
	R^2	0.6	0.5	0.5	0.6	0.4	0.8	0.4	0.3	0.3	0.3	0.3	0.7
Lower basin	NRMSE	8.2	8.0	9.4	11.8	9.4	3.9	17.1	18.3	20.0	23.6	20.7	11.1
	PBIAS	0.0	0.0	0.0	0.2	0.0	0.0	16.9	20.4	19.4	18.0	22.3	5.6
	NSE	0.7	0.7	0.7	0.8	0.6	0.9	0.5	0.5	0.5	0.5	0.4	0.8
	R^2	0.7	0.7	0.7	0.8	0.6	0.9	0.6	0.6	0.6	0.6	0.5	0.8

proposed model for the testing period is around 15% throughout the basin and bias is reduced to less than 6%. Further, the R^2 and NSE values are above 0.7, illustrating the model efficiency. The plot showing the intercomparison of streamflow simulation outcomes from various machine learning models is given in Figure 6. The monthly hydrograph of considered models was compared for sample high-, medium-, and low-flow stations from upper, middle, and lower basins. The hydrographs show a close association of EXGBDT model simulation, especially in peaks and fall throughout various discharge ranges.

The EXGBDT model is selected due to its advantages over other machine learning models for predicting the streamflow discharge at the Cauvery river basin. The model is built to simulate the discharge using training data and the same model is used to predict the discharge for the testing period. The outcome is signified in Figure 7 which illustrates the significance of the model at both calibration and validation phases. Further, the ability of the model in representing the local conditions is evaluated through various hydrological indices in the following section.

Hydrological indices

The comparison of hydrological indices for observed and modeled discharges over the Cauvery river sub-basins is given in Table 3. The daily discharge data are used to calculate these indices. The table gives the percentage of the variance between observed and model data at each index considering 35 stations. The percentage variance shows that the model is performing well in representing the baseflow statistics such as mean and median daily flow, mean baseflow volume, and index. Similarly, the model signifies high-spell and low-spell statistics with an acceptable variance in all sub-basins. The assessment of performance evaluation parameters and the evaluation of hydrological indices suggest that the proposed model is better at representing the local conditions. Consequently, the model can be suggested for forecasting future discharge projection for river basin-scale studies.

SUMMARY AND CONCLUSIONS

The intercomparison of streamflow simulation and prediction models using various machine learning techniques was conducted. A large-scale river basin located in southern peninsular India named Cauvery with frequent floods and drought problems was considered in this study. The daily streamflow discharge model was developed for 35 stations located in the basin using the daily observed precipitation, average, maximum, and minimum temperature. The performance of various machine learning models was evaluated and compared for model selection. Later, various hydrological indices were calculated for observed and

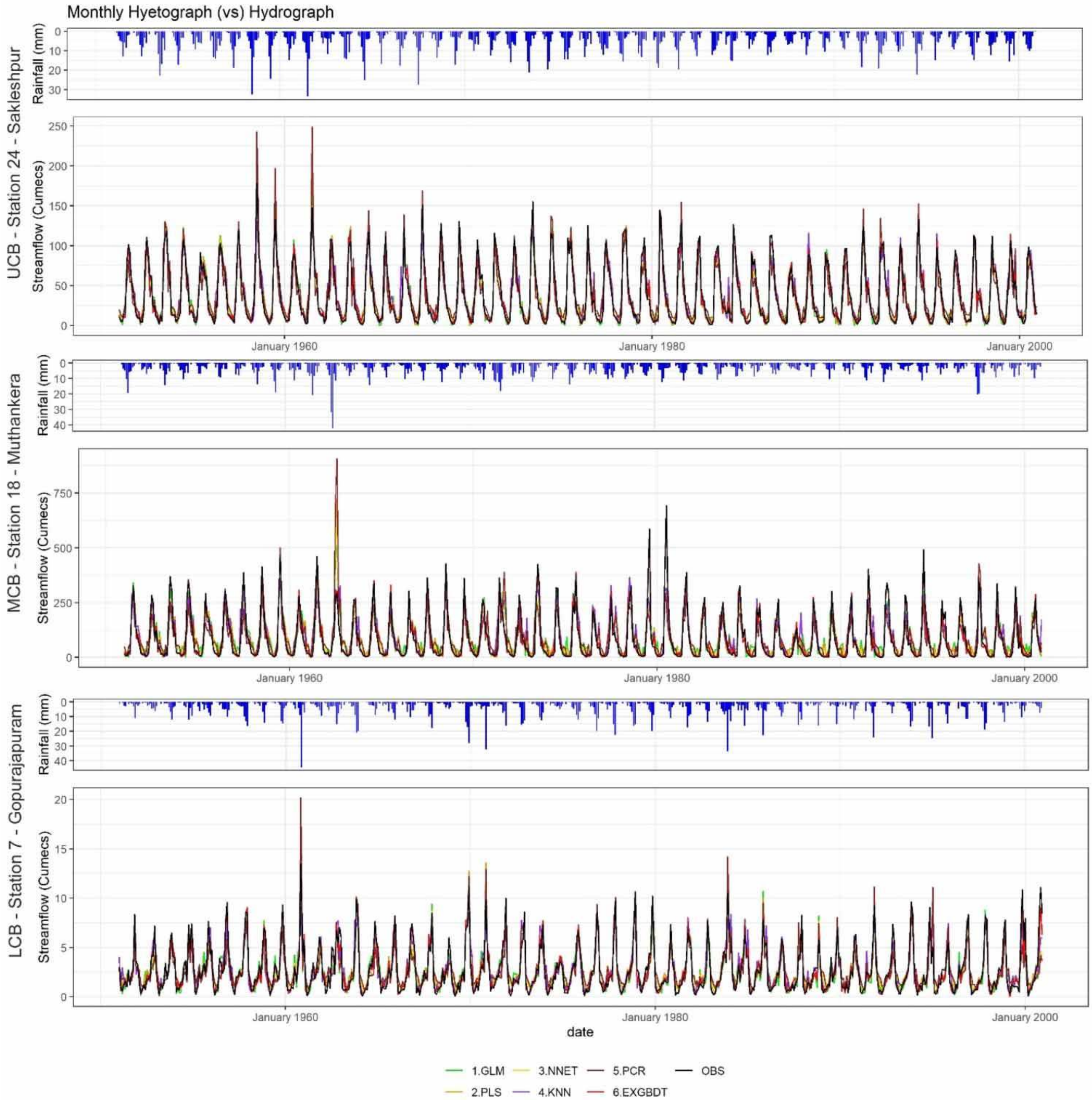

Figure 6 | Intercomparison of streamflow simulation by machine learning models.

predicted discharges for comparing and evaluating the replicability of local conditions.

The following conclusions were drawn from the study:

(1) The model variance and bias of the EXGBDT are less than 15 and 5%, respectively, throughout the basin, which is the least compared with other machine learning techniques considered in this study.
(2) The NSE and R^2 values are above 0.7 for both the training and testing phases which demonstrate the effectiveness of the model's performance.

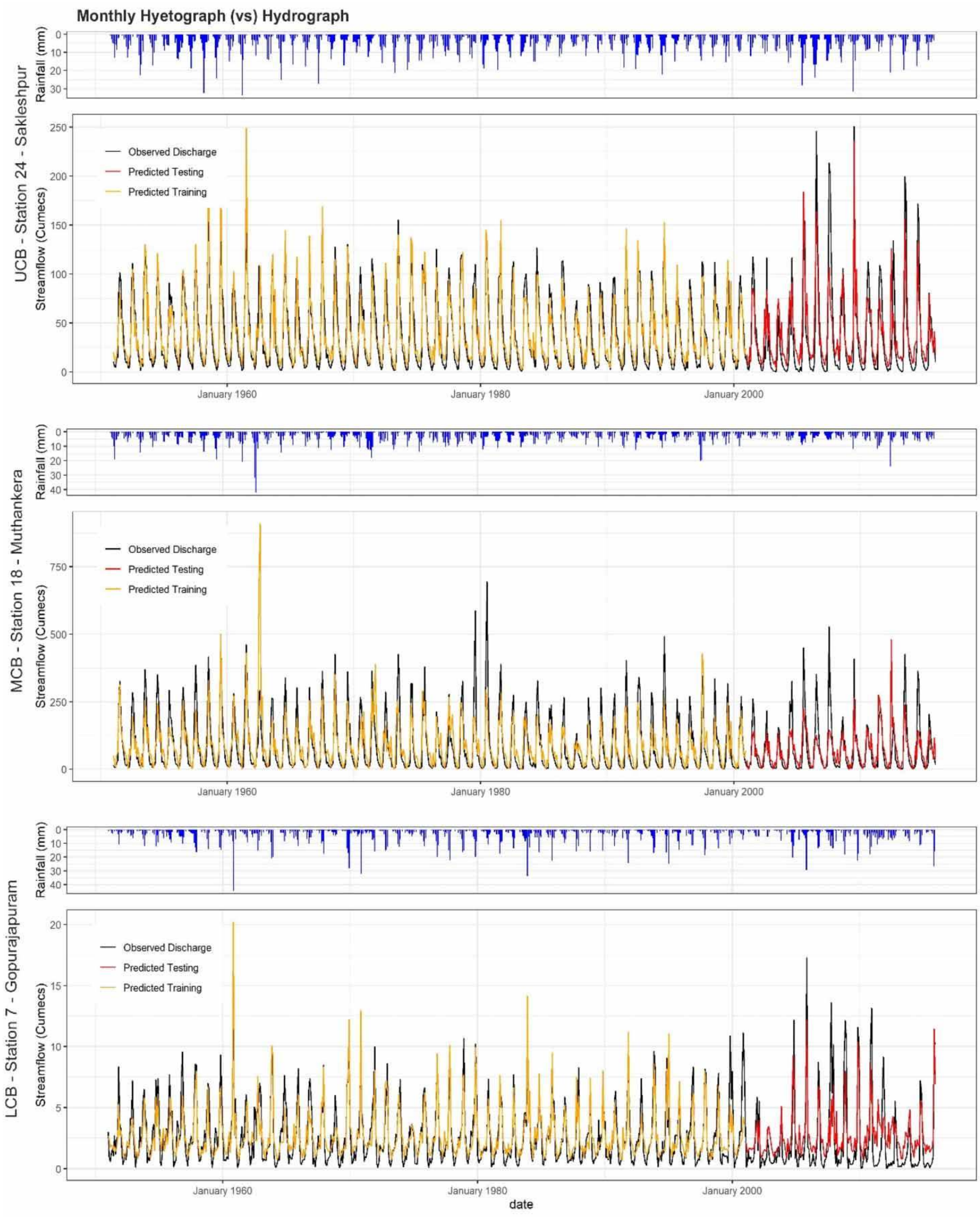

Figure 7 | Streamflow prediction using the EXGBDT model.

Table 3 | Comparison of hydrological indices for observed vs. modeled data over sub-basins

		NRMSE %		
S. No	**Index ID**	**Upper**	**Middle**	**Lower**
Baseflow statistics				
1	MDF	0.7	0.9	1.1
2	Q50	6.4	7.6	7.9
3	mean.bf	5.5	10.1	17.5
4	mean.bfi	13.7	34.3	21.4
High-spell statistics				
5	high.spell.threshold	2.0	4.0	4.7
6	n.events	10.0	19.5	12.3
7	spell.freq	9.8	19.3	11.4
8	avg.high.spell.dur	12.5	14.1	15.9
9	avg.spell.peak	1.0	1.2	2.8
10	sd.spell.peak	23.8	22.6	6.3
11	avg.rise	23.5	18.5	8.0
12	avg.fall	21.5	18.2	7.9
13	avg.max.ann	28.4	24.7	4.6
14	ann.max.timing	10.5	7.6	38.1
15	ann.max.timing.sd	19.9	12.8	31.8
Low-spell statistics				
16	low.spell.threshold	13.3	16.3	29.1
17	avg.min.ann	23.6	33.0	25.1
18	ann.min.timing	28.8	15.3	21.4
19	monthly.cv	11.1	12.2	6.4
20	flow.threshold	32.6	29.4	3.7

(3) The comparison of monthly observed and model-predicted discharges during the validation period illustrates the model's ability in representing the peaks and fall in high-, medium-, and low-flow zones.
(4) The assessment and comparison of hydrological indices between observed and predicted discharges illustrate the model's ability in representing the baseflow, high-flow, and low-flow statistics.

Simulating streamflow and predicting discharge are essential for water resource planning and management especially in large-scale river basins. The proposed machine learning technique demonstrates significant improvement in model efficiency by dropping variance and bias, which in turn improves the replicability of local-scale hydrology. The present study considered streamflow discharge simulation of individual station projection and performance. However, simulation based on stream order is not performed in this study which can be considered as the future direction in improvement of the model performance.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Adnan, R. M., Liang, Z., Yuan, X., Kisi, O., Akhlaq, M. & Li, B. 2019a Comparison of LSSVR, M5RT, NF-GP, and NF-SC models for predictions of hourly wind speed and wind power based on cross-validation. *Energies* **12**. https://doi.org/10.3390/en12020329.

Adnan, R. M., Malik, A., Kumar, A., Parmar, K. S. & Kisi, O. 2019b Pan evaporation modeling by three different neuro-fuzzy intelligent systems using climatic inputs. *Arab. J. Geosci.* **12**. https://doi.org/10.1007/s12517-019-4781-6

Adnan, R. M., Chen, Z. & Yuan, X. 2020a Reference evapotranspiration modeling using new heuristic methods. *Entropy* **22**, 547. https://doi.org/10.1007/978-1-349-00594-9_11.

Adnan, R. M., Liang, Z., Heddam, S., Zounemat-Kermani, M., Kisi, O. & Li, B. 2020b Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. *J. Hydrol.* **586**, 124371. https://doi.org/10.1016/j.jhydrol.2019.124371.

Alizamir, M., Kisi, O., Muhammad Adnan, R. & Kuriqi, A. 2020 Modelling reference evapotranspiration by combining neuro-fuzzy and evolutionary strategies. *Acta Geophys.* **68**, 1113–1126. https://doi.org/10.1007/s11600-020-00446-9.

Andreadis, K. M., Clark, E. A., Lettenmaier, D. P. & Alsdorf, D. E. 2007 Prospects for river discharge and depth estimation through assimilation of swath-altimetry into a raster-based hydrodynamics model. *Geophys. Res. Lett.* **34**, 1–5. https://doi.org/10.1029/2007GL029721.

Arulbalaji, P. & Padmalal, D. 2020 Sub-watershed prioritization based on drainage morphometric analysis: a case study of Cauvery River Basin in South India. *J. Geol. Soc. India* **95**, 25–35. https://doi.org/10.1007/s12594-020-1383-6.

Asong, Z. E., Khaliq, M. N. & Wheater, H. S. 2016 Projected changes in precipitation and temperature over the Canadian Prairie Provinces using the generalized linear model statistical downscaling approach. *J. Hydrol.* **539**. https://doi.org/10.1016/j.jhydrol.2016.05.044.

Bhave, A. G., Conway, D., Dessai, S. & Stainforth, D. A. 2018 Water resource planning under future climate and socioeconomic uncertainty in the Cauvery river basin in Karnataka, India. *Water Resour. Res.* **54**, 708–728. https://doi.org/10.1002/2017WR020970.

Bhuvaneswari, K., Geethalakshmi, V. & Lakshmanan, A. 2013 Rainfall scenario in future over cauvery basin in India. *Indian J. Sci. Technol.* **6**, 4966–4970. https://doi.org/10.17485/ijst/2013/v6i7/34350.

Booker, D. J. 2015 *Hydrological Indices for National Environmental Reporting*. NIWA Client Report CHC2015-015 39.

Choubin, B., Solaimani, K., Rezanezhad, F., Habibnejad Roshan, M., Malekian, A. & Shamshirband, S. 2019 Streamflow regionalization using a similarity approach in ungauged basins: application of the geo-environmental signatures in the Karkheh river basin, Iran. *Catena* **182**, 104128. https://doi.org/10.1016/j.catena.2019.104128.

Coulibaly, P., Dibike, Y. B. & Anctil, F. 2005 Downscaling precipitation and temperature with temporal neural networks. *J. Hydrometeorol.* **6**, 483–496. https://doi.org/10.1175/JHM409.1.

Devak, M., Dhanya, C. T. & Gosain, A. K. 2015 Dynamic coupling of support vector machine and K-nearest neighbour for downscaling daily rainfall. *J. Hydrol.* **525**, 286–301. https://doi.org/10.1016/j.jhydrol.2015.03.051.

Diop, L., Bodian, A., Djaman, K., Yaseen, Z. M., Deo, R. C., El-shafie, A. & Brown, L. C. 2018 The influence of climatic inputs on stream-flow pattern forecasting: case study of Upper Senegal River. *Environ. Earth Sci.* **77**, 1–13. https://doi.org/10.1007/s12665-018-7376-8.

Fotovatikhah, F., Herrera, M., Shamshirband, S., Chau, K. W., Ardabili, S. F. & Piran, M. J. 2018 Survey of computational intelligence as basis to big flood management: challenges, research directions and future work. *Eng. Appl. Comput. Fluid Mech.* **12**, 411–437. https://doi.org/10.1080/19942060.2018.1448896.

Georganos, S., Grippa, T., Vanhuysse, S., Lennert, M., Shimoni, M. & Wolff, E. 2018 Very high resolution object-based land use–land cover urban classification using extreme gradient boosting. *IEEE Geosci. Remote Sens. Lett.* **15**, 607–611. https://doi.org/10.1109/LGRS.2018.2803259.

Ghorbani, M. A., Khatibi, R., Karimi, V., Yaseen, Z. M. & Zounemat-Kermani, M. 2018 Learning from multiple models using artificial intelligence to improve model prediction accuracies: application to River Flows. *Water Resour. Manag.* **32**, 4201–4215. https://doi.org/10.1007/s11269-018-2038-x.

Hashim, R., Roy, C., Motamedi, S., Shamshirband, S., Petković, D., Gocic, M. & Lee, S. C. 2016 Selection of meteorological parameters affecting rainfall estimation using neuro-fuzzy computing methodology. *Atmospheric Res.* **171**, 21–30. https://doi.org/10.1016/j.atmosres.2015.12.002.

Homsi, R., Shiru, M. S., Shahid, S., Ismail, T., Harun, S. B., Al-Ansari, N., Chau, K. W. & Yaseen, Z. M. 2020 Precipitation projection using a CMIP5 GCM ensemble model: a regional investigation of Syria. *Eng. Appl. Comput. Fluid Mech.* **14**, 90–106. https://doi.org/10.1080/19942060.2019.1683076.

Hussain, D. & Khan, A. A. 2020 Machine learning techniques for monthly river flow forecasting of Hunza River, Pakistan. *Earth Sci. Inform.* 939–949. https://doi.org/10.1007/s12145-020-00450-z.

Kale, V. S., Achyuthan, H., Jaiswal, M. K. & Sengupta, S. 2010 Palaeoflood records from upper Kaveri River, Southern India: evidence for discrete floods during holocene. *Geochronometria* **37**, 49–55. https://doi.org/10.2478/v10003-010-0026-0.

Kaya, C. M., Tayfur, G. & Gungor, O. 2019 Predicting flood plain inundation for natural channels having no upstream gauged stations. *J. Water Clim. Change* **10**, 360–372. https://doi.org/10.2166/wcc.2017.307.

Kersbergen, A. M. 2016 *Skill of A Discharge Generator in Simulating low Flow Characteristics in the Rhine Basin*. *Thesis*.

Keum, H. J., Han, K. Y. & Kim, H. I. 2020 Real-time flood disaster prediction system by applying machine learning technique. *KSCE J. Civ. Eng.* **24**, 2835–2848. https://doi.org/10.1007/s12205-020-1677-7.

Ladson, A. R., Brown, R., Neal, B. & Nathan, R. 2013 A standard approach to baseflow separation using the Lyne and Hollick filter. *Austral. J. Water Resour.* **17**, 25–34. https://doi.org/10.7158/13241583.2013.11465417.

Li, Y., Liang, Z., Hu, Y., Li, B., Xu, B. & Wang, D. 2020 A multi-model integration method for monthly streamflow prediction: modified stacking ensemble strategy. *J. Hydroinformatics* **22**, 310–326. https://doi.org/10.2166/hydro.2019.066.

Londhe, S. & Charhate, S. 2010 Comparaison de techniques de modélisation conditionnée par les données pour la prévision des débits fluviaux. *Hydrol. Sci. J.* **55**, 1163–1174. https://doi.org/10.1080/02626667.2010.512867.

Madolli, M. J., Kanannavar, P. S. & Yaligar, R. 2015 Impact of climate change on precipitation for the upper Cauvery river basin, Karnataka State. *Int. J. Agric. Sci. Res. IJASR* **5**, 99–104.

Matulessy, E. R., Wigena, A. H. & Djuraidah, A. 2015 Quantile regression with partial least squares in statistical downscaling for estimation of extreme rainfall. *Appl. Math. Sci.* **9**, 4489–4498. https://doi.org/10.12988/ams.2015.53254.

Nadu, T. & Nadu, T. 1981 References: assessment of drought and flood in Cauvery delta zone with a special reference to Tamil Nadu rice research institute, Aduthurai. *Madras Agric. J.* **95**, 457–461.

Naganna, S. R., Deka, P. C., Ghorbani, M. A., Biazar, S. M., Al-Ansari, N. & Yaseen, Z. M. 2019 Dew point temperature estimation: application of artificial intelligence model integrated with nature-inspired optimization algorithms. *Water Switz.* **11**, 1–17. https://doi.org/10.3390/w11040742.

Piras, M., Mascaro, G., Deidda, R. & Vivoni, E. R. 2016 Impacts of climate change on precipitation and discharge extremes through the use of statistical downscaling approaches in a

Mediterranean basin. *Sci. Total Environ.* **543**, 952–964. https://doi.org/10.1016/j.scitotenv.2015.06.088.

Raju, K. V., Somashekar, R. K. & Prakash, K. L. 2013 Spatio-temporal variation of heavy metals in Cauvery River basin. *Proc. Int. Acad. Ecol. Environ. Sci.* **3**, 59–75.

Rezaie-Balf, M. & Kisi, O. 2018 New formulation for forecasting streamflow: evolutionary polynomial regression vs. extreme learning machine. *Hydrol. Res.* **49**, 939–953. https://doi.org/10.2166/nh.2017.283.

Rezaie-balf, M., Naganna, S. R., Ghaemi, A. & Deka, P. C. 2017 Wavelet coupled MARS and M5 model tree approaches for groundwater level forecasting. *J. Hydrol.* **553**, 356–373. https://doi.org/10.1016/j.jhydrol.2017.08.006.

Sahriman, S., Djuraidah, A. & Wigena, A. H. 2014 Application of principal component regression with dummy variable in statistical downscaling to forecast rainfall. *Open J. Stat.* **04**, 678–686. https://doi.org/10.4236/ojs.2014.49063.

Sekhar, S., Pijush, R., Deo, R. & Ntalampiras, S. 2018 *Studies in Big Data – Big Data in Engineering Applications*. Springer Nature, Cham, Switzerland.

Shamshirband, S., Hashemi, S., Salimi, H., Samadianfard, S., Asadi, E., Shadkani, S., Kargar, K., Mosavi, A., Nabipour, N. & Chau, K. W. 2020 Predicting standardized streamflow index for hydrological drought using machine learning models. *Eng. Appl. Comput. Fluid Mech.* **14**, 339–350. https://doi.org/10.1080/19942060.2020.1715844.

Solaraj, G., Dhanakumar, S., Rutharvel Murthy, K. & Mohanraj, R. 2010 Water quality in select regions of Cauvery Delta River basin, southern India, with emphasis on monsoonal variation. *Environ. Monit. Assess.* **166**, 435–444. https://doi.org/10.1007/s10661-009-1013-7.

Sunil, C., Somashekar, R. K. & Nagaraja, B. C. 2010 Riparian vegetation assessment of Cauvery River Basin of South India. *Environ. Monit. Assess.* **170**, 545–553. https://doi.org/10.1007/s10661-009-1256-3.

Taormina, R. & Chau, K. W. 2015 ANN-based interval forecasting of streamflow discharges using the LUBE method and MOFIPS. *Eng. Appl. Artif. Intell.* **45**, 429–440. https://doi.org/10.1016/j.engappai.2015.07.019.

Tikhamarine, Y., Souag-Gamane, D., Najah Ahmed, A., Kisi, O. & El-Shafie, A. 2020 Improving artificial intelligence models accuracy for monthly streamflow forecasting using grey Wolf optimization (GWO) algorithm. *J. Hydrol.* **582**, 124435. https://doi.org/10.1016/j.jhydrol.2019.124435.

Van Der Velde, Y., Lyon, S. W. & Destouni, G. 2013 Data-driven regionalization of river discharges and emergent land cover-evapotranspiration relationships across Sweden. *J. Geophys. Res. Atmospheres* **118**, 2576–2587. https://doi.org/10.1002/jgrd.50224.

Ward, G. H. 2013 *Hydrological Indices and Triggers, and Their Application to Hydrometeorological Monitoring and Water Management in Texas*. Tex. Water Dev. Board, Austin, TX, p. 254.

Wu, C. L. & Chau, K. W. 2013 Prediction of rainfall time series using modular soft computing methods. *Eng. Appl. Artif. Intell.* **26**, 997–1007. https://doi.org/10.1016/j.engappai.2012.05.023.

Wu, C. L., Chau, K. W. & Li, Y. S. 2009 Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. *Water Resour. Res.* **45**, 1–23. https://doi.org/10.1029/2007WR006737.

Yaseen, Z. M., Ebtehaj, I., Kim, S., Sanikhani, H., Asadi, H., Ghareb, M. I., Bonakdari, H., Wan Mohtar, W. H. M., Al-Ansari, N. & Shahid, S. 2019a Novel hybrid data-intelligence model for forecasting monthly rainfall with uncertainty analysis. *Water Switz.* **11**. https://doi.org/10.3390/w11030502

Yaseen, Z. M., Mohtar, W. H. M. W., Ameen, A. M. S., Ebtehaj, I., Razali, S. F. M., Bonakdari, H., Salih, S. Q., Al-Ansari, N. & Shahid, S. 2019b Implementation of univariate paradigm for streamflow simulation using hybrid data-driven model: case study in tropical region. *IEEE Access* **7**, 74471–74481. https://doi.org/10.1109/ACCESS.2019.2920916.

Yuan, X., Chen, C., Lei, X., Yuan, Y. & Muhammad Adnan, R. 2018 Monthly runoff forecasting based on LSTM–ALO model. *Stoch. Environ. Res. Risk Assess.* **32**, 2199–2212. https://doi.org/10.1007/s00477-018-1560-y.

First received 7 September 2020; accepted in revised form 10 November 2020. Available online 24 November 2020