Applied Water Science (2024) 14:99
https://doi.org/10.1007/s13201-024-02135-0

ORIGINAL ARTICLE

Streamflow prediction using support vector regression machine learning model for Tehri Dam

Bhanu Sharma[1] · **N. K. Goel**[1]

Received: 3 October 2023 / Accepted: 5 February 2024

Abstract
Accurate and reliable streamflow prediction is critical for optimising water resource management, reservoir flood operations, watershed management, and urban water management. Many researchers have published on streamflow prediction using techniques like Rainfall-Runoff modelling, Time series Models, Data-driven models, Artificial intelligence, etc. Still, there needs to be generalised method practise in the real world. The resolution of this issue lies in selecting different methods for a particular study area. This paper uses the Support vector regression machine learning model to predict the streamflow for the Tehri Dam, Uttarakhand, India, at the Daily and Ten Daily time steps. Two cases are considered in predicting daily and ten daily time steps. The first case includes four input variables: Discharge, Rainfall, Temperature, and Snow cover area. The second case comprises only three input variables: Rainfall, Temperature, and Snow cover area. Radial Kernel is used to overcome the space complexity in the datasets. The K-fold cross-validation is suitable for prediction as it averages the prediction error rate after evaluating the SVR model's performance on various subsets of the training data. The streamflow data for daily and ten daily time steps have been collected from 2006 to 2020. The calibration period is from 2006 to 2016, and the validation period is from 2017 to 2020. Nash Sutcliffe Efficiency (NSE) and Coefficient of determination (R^2) are used as the accuracy indicator in this manuscript. The lag has been observed in the daily prediction time series when three input variables are considered. For other scenarios, the respective model shows excellent results at both the temporal scale and the parametres, which play a vital role in prediction. The study also enhances the effect on the potential use of input parametres in the machine learning model.

Keywords Machine learning · Optimization · Support vector regression · Calibration period · Validation period

Introduction

The streamflow process is considered a vital component of the complex hydrological cycle and is difficult to predict accurately (Zhang et al. 2016; Loaiciga et al. 2018; Ireson et al. 2015; Nourani et al. 2014). It is invariably affected by Precipitation, Temperature, evapotranspiration, snow cover area, land use pattern, and drainage basin (Adnan et al. 2019). The accurate and reliable forecast of streamflow processes is critical in the design, planning, optimisation, utilisation, and management of water resources (Adnan et al. 2018; Keshtegar et al. 2016; Riahi-Madvar et al. 2021; Khosravi et al. 2022; Senthil Kumar et al. 2017). Streamflow prediction models, also known as hydrological models or runoff models, are used to anticipate how much water will flow in rivers and streams over time. These models are critical tools in hydrology and water resource management because they expect river discharge, which is essential for various applications such as flood forecasting, water resource planning, and environmental management. Streamflow prediction models are classified into two categories; each category has its unique technique and level of complexity (Solomatine and Ostfeld 2008): (i) a Physically based model and (ii) a Data-driven model. A variety of data are needed for physically based models, including information on human activity, land use, physiographic features of the drainage basin, and the volume, intensity, and distribution of rainfall (Ochoa-Tocachi et al. 2022; Teutschbein et al. 2018). In contrast, a mathematical relationship (linear or nonlinear)

✉ Bhanu Sharma
bsharma@hy.iitr.ac.in

N. K. Goel
nkgoel@hy.iitr.ac.in

[1] Department of Hydrology, IIT Roorkee, Roorkee, Uttarakhand 247667, India

Published online: 13 April 2024

is established between streamflow and its constraints (Rainfall, Temperature, snow cover, etc.) [Zhang et al. (2021), Yaseen et al. (2015)]. Elshorbagy et al. (2010) studied the data-driven model in simulating hydrological components like evapotranspiration, soil moisture, and rainfall-runoff using neural networks, genetic programming, evolutionary polynomial regression, support vector machines, K-nearest neighbours, and multiple linear regression. They discovered that data-driven models can be successfully used in hydrological applications. The traditional linear models do not capture the non-linearity and non-stationarity of hydrological applications (Afan et al. 2016; Yaseen et al. 2015; Yadav et al. 2022; Imrie et al. 2000). In hydrological time-series forecasting, the linear models like moving average (MA), autoregressive (AR), autoregressive moving average (ARMA), and autoregressive integrated moving average (ARIMA) have found widespread use (Wu et al. 2009; Wu and Chau 2010; Valipour et al. 2013, Valipour 2015). To overcome the shortcomings of traditional models, researchers have concentrated on building machine learning-based models (Yaseen et al. 2015; Adnan et al. 2019).

The modelling and prediction of streamflows have seen extensive use of machine learning techniques over the past 20 years on a global scale (Granata et al. 2016; Elebeltagi et al. 2018; Hadi and Tombul 2018; Yaseen et al. 2015; Al-Sudani et al. 2019; Rasouli 2020; Malik et al. 2020). Huang et al. (2019) used the Bayesian model averaging (BMA), Artificial Neural Network (ANN), and Support Vector Machine (SVM) to predict the Monthly runoff for Huang Zhuang station in the Hanjiang River basin, China. The study suggested that ANN and SVM models performed best. Rahmani-Rezaeieh et al. (2019) predicted daily streamflow in the Shahrchay River Basin, Iran, using Ensemble Gene Expression Programming (EGEP). Rezaie-Balf et al. (2019) used Random Forest Regression (RFR) to model the daily streamflow at the Bilghan, Siira, and Gachsar stations in Iran. Hussain and Khan (2020) have used the Support vector regression (SVR), Multilayer Perceptron (MLP), and Random Forest (RF) models to predict the monthly flow of the Hunza River, Pakistan, and found that the RF model outperformed other models in the basin. Pandhiani et al. (2020) have used the Random Forest and Artificial Neural Network data-driven models for monthly streamflow prediction in Malaysia's Berman and Tualang rivers and concluded that both models work well for the study area.

The present research aims to predict the daily and ten daily time series of streamflow at Tehri Dam in Uttarakhand, India. The novelty is considering using the Support Vector Regression (SVR) for streamflow prediction at the Daily and Ten Daily temporal scales. The input parametres (Discharge, Rainfall, Temperature, and Snow cover area) are used in the model to predict the streamflow at Tehri Dam. The SVR was trained for a period from 2006 to 2016 and was validated from 2017 to 2020. The calibrated parametres for SVM have been finalised using a K-fold cross-validation approach. The prediction accuracy is assessed over observed streamflow through NSE (Nash Sutcliffe Efficiency) and R^2 (Coefficient of Determination). It is worth mentioning that the performance of the proposed SVR model is examined for the first time in the Tehri Catchment at daily and ten daily streamflow series.

Study area

Tehri Dam is located at the confluence of the Bhagirathi and Bhilangana Rivers in the Uttarakhand state of India. It is an earthen rockfill dam with a height of 260.5 m (Elevation 839.50 m above MSL). It has an installed capacity of 1000 MW. The Tehri project was commissioned in 2006 and provides water for irrigation to Uttar Pradesh (UP) and Uttarakhand states. It also provides drinking water to nearly seven million people of UP and Uttarakhand. It has a gross and live storage of 3540 and 2615 MCM (Million Cubic Metres). The dam is designed to pass the Probable Maximum Flood (PMF) of 15,540 Cumecs. The PMF is catered by three Chute spillways (5500 Cumecs), two left bank shaft spillways (3650 Cumecs), and two ungated spillways (3850 Cumecs). The Maximum Flood Level (MFL) and Full Reservoir Level (FRL) are 839.50 m and 830 m, respectively (Figs. 1 and 2).

Methodology

The input variables of a machine learning model are the fragments of Information that the model utilises to produce predictions and decisions.

The selection of input variables is an essential phase in developing a machine learning model since the quality and relevance of these parametres significantly impact the model's performance. The choice of input parametres should align with your problem statement, the nature of your data, and the machine learning algorithms you intend to use. It is often an iterative process that involves refining the feature set based on the model's performance and domain knowledge. The input variables for the Support Vector Regression model are as follows:

Discharge data

Observing discharge data from hydroelectric power plants is crucial to their operation and management. The information of this data is essential for ensuring the efficient and safe operation of the power plants and managing

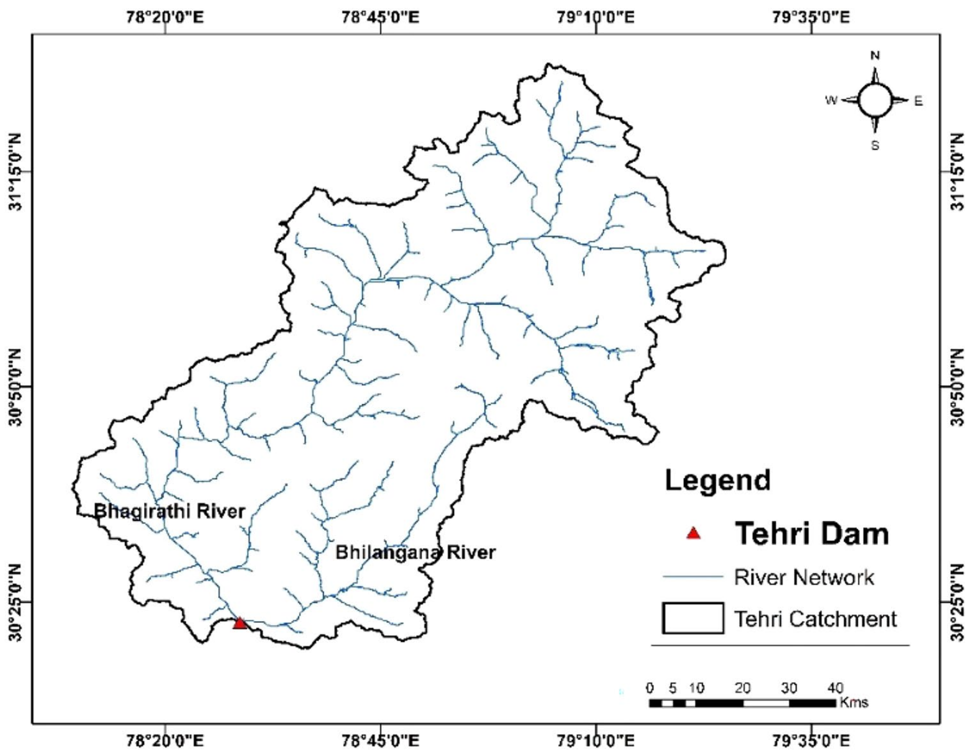


Fig. 1 Location map of Tehri catchment and major rivers

Fig. 2 Front View of Tehri Dam

downstream water resources. The Daily and Ten daily discharge data have been observed by the THDC India Limited officials since 2006. The data from 2006 to 2020 are used in the present manuscript. The Calibration and validation periods are taken from 2006 to 2016 and 2017 to 2020, respectively (Fig. 3).

Rainfall data

The India Meteorological Department (IMD) provides gridded rainfall data at a spatial resolution of 0.25° by 0.25° degrees (Pai et al. 2014). This data is used for various meteorological and climatological applications, including weather forecasting, climate monitoring, and hydrological studies. The data can be downloaded from IMD's Pune website. The rainfall data for the Tehri catchment has been downloaded and divided into ten elevation zones. There is a significant variation in the Rainfall as the elevation increased in the Himalaya region (Singh and Bengtsson 2004; Sen Roy and Balling 2004; Goswami et al. 2006; Rajeevan et al. 2008; Roy et al. 2009; Krishnamurthy et al. 2009; Guhathakurta et al. 2011). To account for the variation of Rainfall with elevation is introduced in the model as the input variable.

Temperature data

The India Meteorological Department (IMD) provides temperature data for various purposes, including weather forecasting, climate monitoring, research, agriculture, health, and energy management applications. Temperature data from IMD is valuable for understanding climate patterns and trends in different regions of India. India Meteorological Department (IMD) Daily gridded temperature data (1° × 1°) (Srivastava et al. 2009) is used in the present manuscript. Temperature data from IMD is typically available in digital

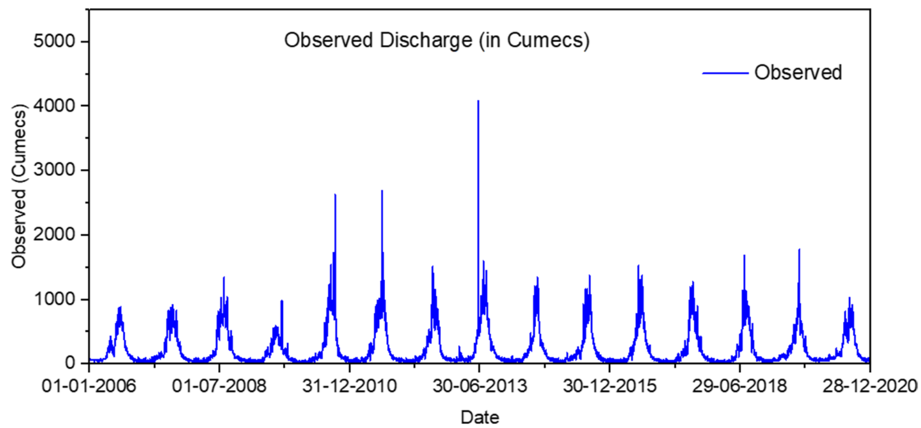


Fig. 3 Time series of observed discharge Data (2006–2020)

formats, such as text files (CSV or ASCII), NetCDF (Network Common Data Form), or other standard formats commonly used in meteorological and climatological data. The Catchment is divided into five Elevation Zones, and Temperature is calculated respectively.

The temperature and rainfall data from weather stations are point measurements; however, spatially distributed datasets are required for more systematic and detailed analysis (Kormos et al. 2018; Behnke et al. 2016). Therefore, high-resolution gridded meteorological datasets are preferred in climate modelling and hydrological processes studies, and the same has been applied in the present study (Caldwell et al. 2009; Walton et al. 2015).

Snow cover data

At a temporal resolution of 8 days, the Snow Cover Area for Tehri Catchment was derived using MODIS/Terra Snow Cover 8-Day L3 Global 500 m SIN Grid, Version 5. This dataset monitors and maps snow cover on Earth's surface. Researchers and government agencies use it to track changes in snow cover extent over time, which can provide insights into climate trends and seasonal variations (Coops et al. 2006).

Two cases have been considered for predicting Discharge for the daily and ten daily temporal scales: (1) Discharge data and the three input variables (Rainfall, Temperature, and snow cover Data) are used (2) Discharge data is not considered. The K-fold cross-validation technique is used to compute the optimum paraMetres of the model. Accurate hydro-system modelling requires systematic integration of factors, time series decomposition, data regression, and error suppression.

The proposed streamflow forecasting framework consists of Model selection, Time series decomposition, model training, model learning, optimum parameter estimation, error computation and error correction. Nash Sutcliffe Efficiency (NSE) and coefficient of determination (R^2) are performance indicators. The calibration and validation periods are 2006–2016 and 2017–2020, respectively. These performance indicators are used to assess the predictive skill of the machine learning model computed on each year's time scales in the present study.

Support Vector Regression works in high or infinite-dimensional space and generates a hyper-plane or collection of hyper-planes. According to intuition, the hyper-plane in each class farthest from the nearest training data points achieves a meaningful separation since, generally speaking, the wider the margin, the smaller the classifier's generalisation error. It functions well in high-dimensional spaces and may behave differently based on the kernel, a collection of mathematical operations. Many different types of functions are referred to by terminology like linear, polynomial, radial basis function (RBF), sigmoid, and others. The SVR algorithm can be summed up as follows: A suitable kernel function must be chosen, the regularisation parameter-C must be assigned a value, the quadratic programming (QP) problem must be resolved, and the discriminant process must be built using the support vectors (Fig. 4).

Support vector regression

SVR is a data training/fitting technique. The essence of SVR is to transfer the original problem into solving a quadratic programming problem, and it can theoretically obtain the global optimum result of the problem. The computing rate of SVM is significantly faster than that of other techniques.

Overview of basic SVM for regression Suppose the sample data for training is $\{X_i, y_i\}$, where $i = 1, 2, ..., l$, X_i is the input, and y_i is the output. The aim of SVM for regression is to find a function of this form:

$$\mathrm{Bold}y_i = W.X_i + b$$

where W is a hyperplane, and b is the offset. The regression SVM will use a penalty function:

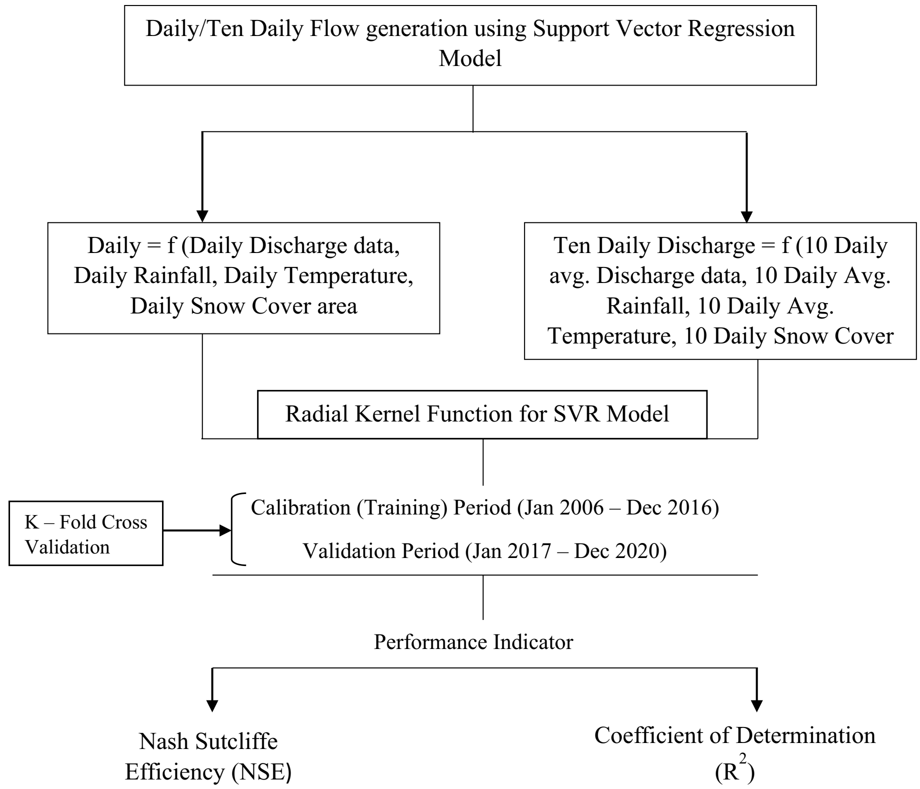


Fig. 4 Framework for forecast using support Vector regression

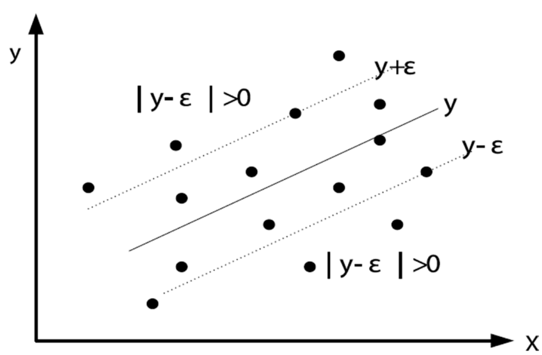


Fig. 5 SVM for regression with ε—insensitive tube

$$\begin{cases} \left|y_i - \left(W.X_i + b\right)\right| \leq \in, & \textit{not allocating a penalty} \\ \left|y_i - \left(W.X_i + b\right)\right| > \in, & \textit{allocating a penalty} \end{cases}$$

Referring to Fig. 5, the region bound by $y_i \pm$ e is called an e-insensitive tube. The goal of this problem can be written according to:

$$\text{audaciousMin}\left[\frac{1}{2}||W||^2 + C\sum_{i=1}^{l} L^{\in}(X_i, y_i, f)\right]$$

where the $L^{\in}(X_i, y_i, f)$ is defined as:

$$L^{\in}\left(X_i, y_i, f\right) = max(0, \left|f\left(X_i\right) - y_i\right| - \in)$$

And as the existence of fitting errors, the slack variables ξ^+ and ξ^- are introduced, then the model form of SVM for regression will be as follows:

$$\text{Min}\left[\frac{1}{2}||W||^2 + C\sum_{i=1}^{l}(\xi^+ + xi_i^-)\right]$$

Subject to: $\left(W.X_i + b\right) - y_i \leq \in + \xi^+$
$y_i - \left(W.X_i + b\right) \leq \in + \xi^-$
$\xi^+ > 0, \xi^- > 0$
$i = 1,2,3\ldots.,l$

The corresponding dual problem can be derived using the now standard techniques:

$$\text{Max}\left[\sum_{i=1}^{l}\left(\alpha_i^+ - \alpha_i^-\right).y_i - \in \sum_{i=1}^{l}(\alpha_i^+ - \alpha_i^-) - \frac{1}{2}\sum_{i,j}^{i}\left(\alpha_i^+ - \alpha_i^-\right).\left(\alpha_j^+ - \alpha_j^-\right).X_i.X_j\right]$$

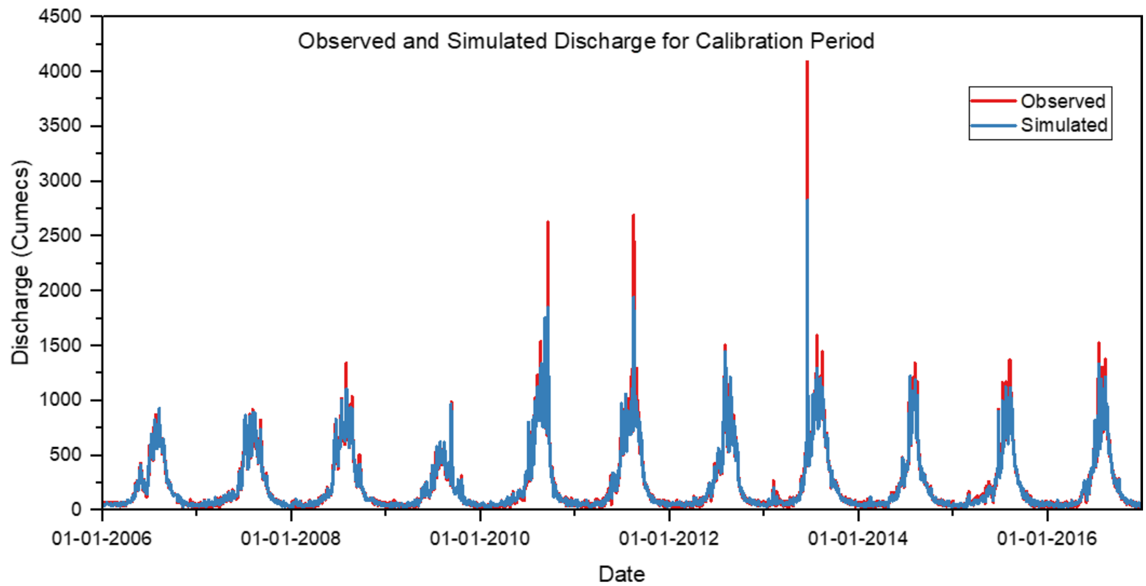


Fig. 6 Graph between observed and predicted (simulated) discharge for the calibration period

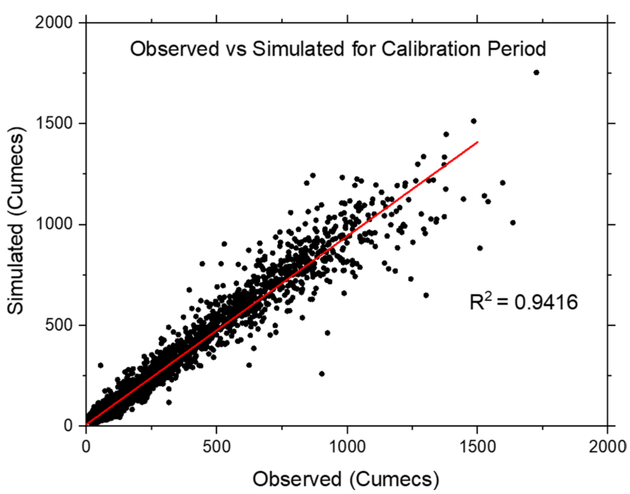


Fig. 7 Scatter plot for calibration period (2006–2016)

Subject to: $0 \leq \alpha_i^+ \leq C, 0 \leq \alpha_i^- \leq C$

$$\sum_{i=1}^{l} (\alpha_i^+ - \alpha_i^-) = 0$$

Solve this problem with a quadratic programming method, and then we can acquire the regression function of the system.

Results and discussion

The present manuscript represents the streamflow prediction using the Support vector regression machine learning models. The first part of this section describes the results of the Support vector regression (SVR) model for the Daily streamflow prediction. The second part of this section explains the results of the Support vector regression (SVR) model for the Ten Daily streamflow prediction.

Daily streamflow prediction using support vector regression model

Support vector machine uses the maximum margin algorithm, where, for a hyperplane, the algorithm searches for the most significant separating margin between the observed data for obtaining the optimal function that fits the observation. The algorithm uses a kernel to solve this nonlinear optimisation problem to get the most accurate hyperplane.

For this case, we use a radial kernel calibrated by adjusting cost ‘c’ and gamma ‘g’. A grid search method is applied for Calibration, where a combination of values of the hyperparaMetres is checked.

Now, for each combination of the hyperparametres, a K-fold cross-validation was performed (Anguita et al. 2009). The data is divided into ‘k’ subsets (4). *k*-1 subsets are used for training the model, and the remaining one for validation, for which an average error for k-trails was computed. This method helps us identify the paraMetres suited for more than one subset.

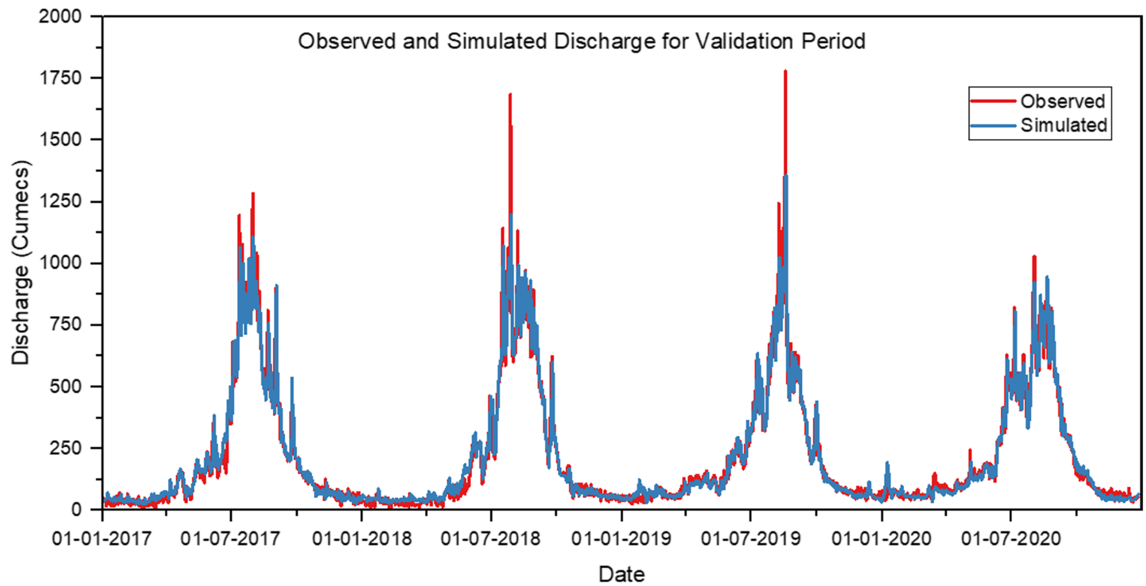


Fig. 8 Graph between observed and simulated discharge for the validation period

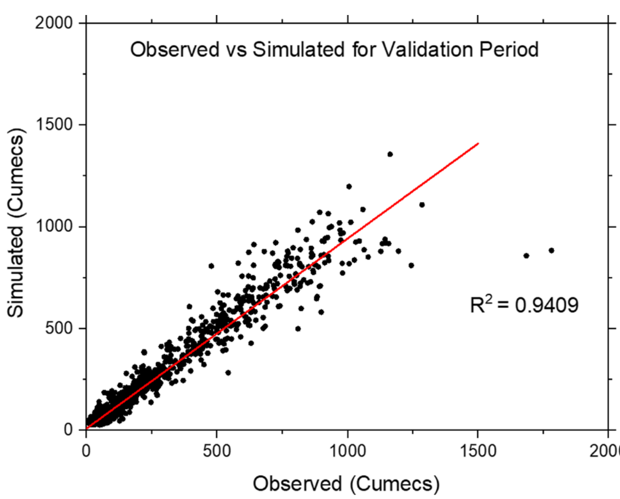


Fig. 9 Scatter plot for validation period (2017–2020)

The two cases have considered: (i) Four Input Variables are considered (Discharge Rainfall, Temperature, Snow cover area), (ii) Three Input Variables are considered (Rainfall, Temperature, Snow cover area).

Streamflow prediction when four variables are considered (discharge, rainfall, temperature, snow cover area)

In this case, four input parametres [Discharge (Q_t), Rainfall (R_t), Temperature (T_t), and Snow Cover Area (SCA_t)] have been considered. The model is trained using hyperparametres for the calibration and validation period. The Nash Sutcliffe efficiency (NSE) and Coefficient of Determination (R^2) are performance indicators. The NSE is 96.75 and 95.57 for the calibration and validation period. The coefficient of determination (R^2) for observed and simulated discharge for the Calibration and validation period is 0.9416 and 0.9578, respectively. The scatter plots have been plotted for all the discharge data from 2006 to 2020. The model fits the observed data well. The model shows high efficiency in the prediction of daily discharge. It has been observed in 2009, 2013, 2018, and 2019 that the model is unsuitable for predicting high discharges, but the overall efficiency of the model is excellent. The model efficiency (NSE & R^2) is also calculated for each year's data; NSE and R^2 range from 80.45 to 97.18 and 0.8965 to 0.9723, respectively. The model shows high performance in predicting discharge at a daily time scale (Figs. 6, 7, 8, 9 and 10; Table 1 and 2).

Daily streamflow prediction when three input variables are considered (rainfall, temperature, snow cover area)

In this case, three input paraMetres [Rainfall (R_t), Temperature (T_t), and Snow Cover Area (SCA_t)] have been considered. The model is trained using hyperparametres for the calibration and validation period. The NSE is 86.68 and 75.85 for the calibration and validation period. The coefficient of determination (R^2) for observed and simulated discharge for the calibration and validation periods is 0.8617 and 0.7635, respectively. The scatter plot has

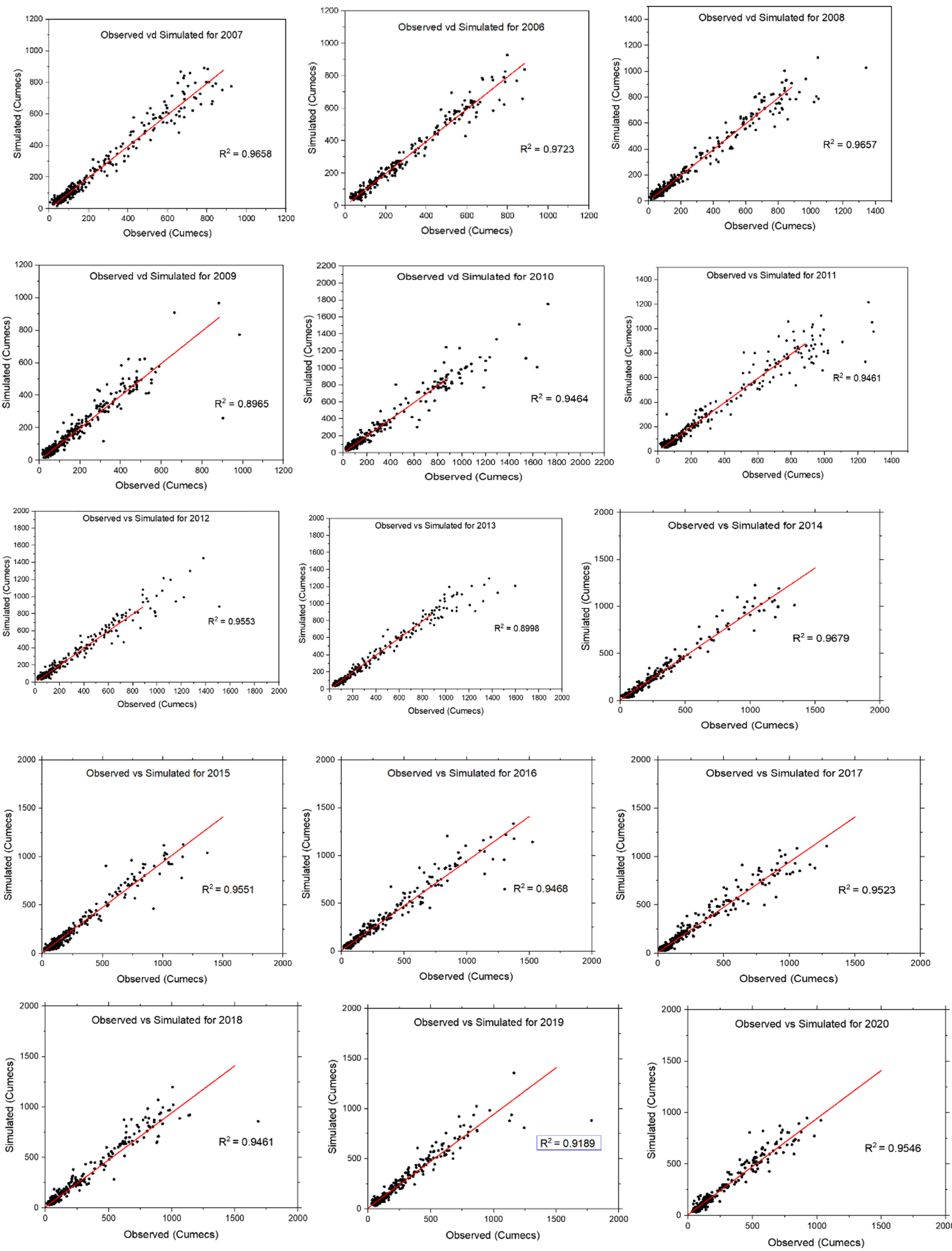


Fig. 10 Scatter plot for the period 2006–2020

Table 1 NSE and R^2 for calibration and validation period

Case-1 (four input variables have been considered for daily streamflow prediction)	Time	NSE	R^2
Calibration	Jan-2006 to Dec-2016	96.75	0.9416
Validation	Jan-2017 to Dec-2020	95.57	0.9578

Table 2 Year-wise NSE and R^2 for the period 2006–2020

Year	NSE	R^2
2006	97.18	0.9723
2007	96.51	0.9658
2008	96.53	0.9657
2009	89.25	0.8965
2010	94.4	0.9464
2011	94.18	0.9461
2012	95.52	0.9553
2013	88.45	0.8998
2014	96.54	0.9679
2015	95.34	0.955
2016	94.6	0.9468
2017	95.11	0.9523
2018	93.94	0.9419
2019	91.81	0.9189
2020	95.4	0.9546

been plotted for all the discharge data from 2006 to 2020 (Annexure I). The model fits the observed data well. The model shows high efficiency in the prediction of daily discharge. It has been observed in 2008, 2009, 2011, 2012, 2018, and 2019 (Annexure I) that the model is unsuitable for predicting high discharges, but the overall efficiency of the model is good. The model efficiency (NSE & R^2) is also calculated for each year's data; NSE and R^2 range from 62.24 to 89.61 and 0.6834 to 0.9168, respectively (Annexure I). The model shows high performance in predicting discharge at a daily time scale (Figs. 11, 12, 13 and 14; Table 3).

Ten daily streamflow prediction using support vector regression model

The two cases have been considered: (i) Four input Variables are considered (10 daily avg. discharge 10-daily average Rainfall, 10-daily average Temperature, 10-day Snow cover area) (ii) Three Variables are considered (10-daily average Rainfall, 10-daily average Temperature, 10-day Snow cover area).

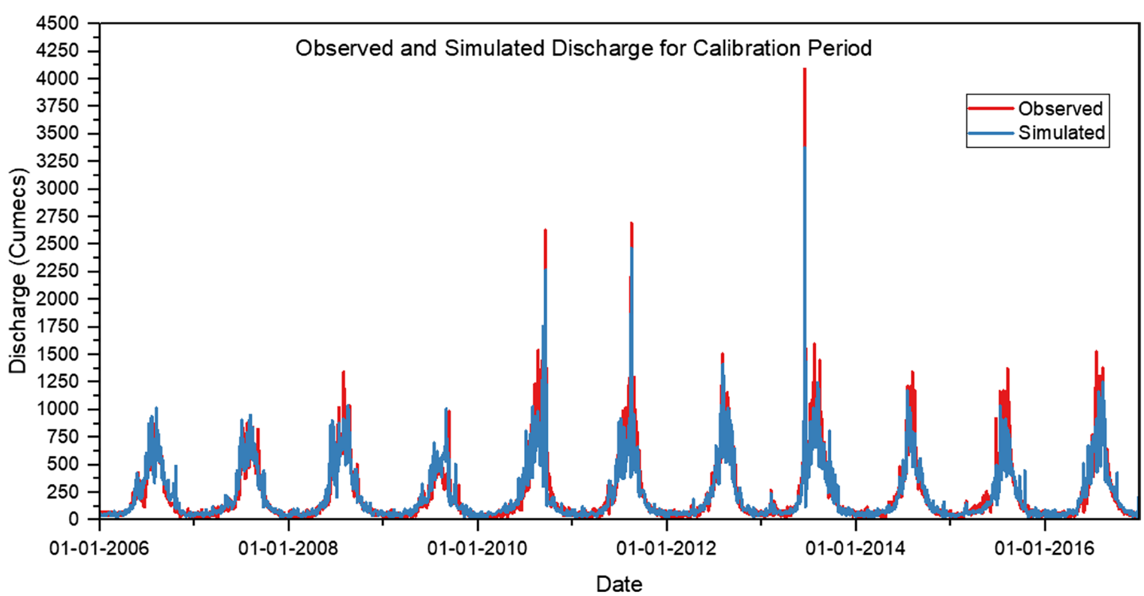


Fig. 11 Graph between observed and simulated discharge for the calibration period

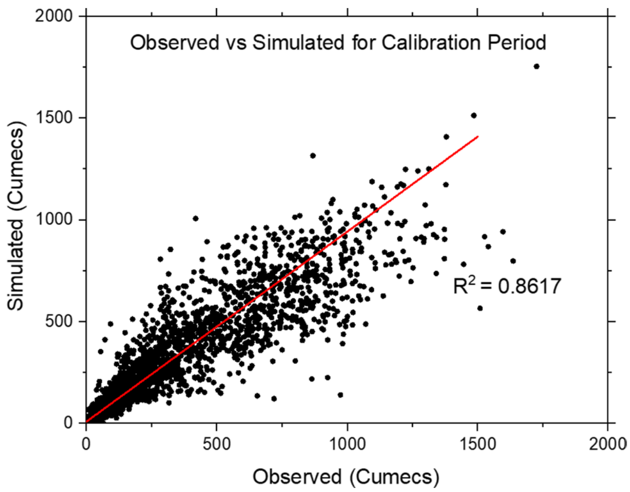


Fig. 12 Scatter plot for calibration period (2006–2016)

Ten daily streamflow predictions when four input variables are considered

In this case, four input variables [10 daily avg. Discharge (Q_t), ten daily avg. Rainfall (R_t), ten daily average. Temperature (T_t) and 10 daily Snow Cover Areas (SCA_t)] have been considered. The model is trained using hyperparametres for the calibration and validation period. The NSE is 96.77 and 95.60 for the calibration and validation period. The coefficient of determination (R^2) for the observed and simulated discharge periods is 0.9679 and 0.9561, respectively. The scatter plot has been plotted for all the discharge data from 2006 to 2020. The model fits the observed data well. The model shows high efficiency in the prediction of 10 daily discharges. The model efficiency (NSE & R^2) is also calculated for each year's data; NSE and R^2 range from 90.45 to 98.76 and 0.9337 to 0.9892, respectively (Annexure I). The model shows high performance in predicting discharge at ten daily temporal scales (Figs. 15, 16, 17 and 18; Table 4).

Streamflow prediction when three input variables are considered

In this case, three input variables [Rainfall (R_t), Temperature (T_t), and Snow Cover Area (SCA_t)] have been considered. The model is trained using hyperparaMetres for the calibration and validation period. The NSE is 88.22 and 92.52 for the calibration and validation period. The coefficient of determination (R^2) for observed and simulated discharge for the calibration and validation periods is 0.8827 and 0.9454, respectively. The scatter plot has been plotted for all the discharge data from 2006 to 2020 (Annexure I). The model fits the observed data well. The model shows high efficiency in the prediction of 10 daily discharges. The model efficiency (NSE & R^2) is also calculated for each year's data; NSE and R^2 range from 61.15 to 95.25 and 0.7735 to 0.9692,

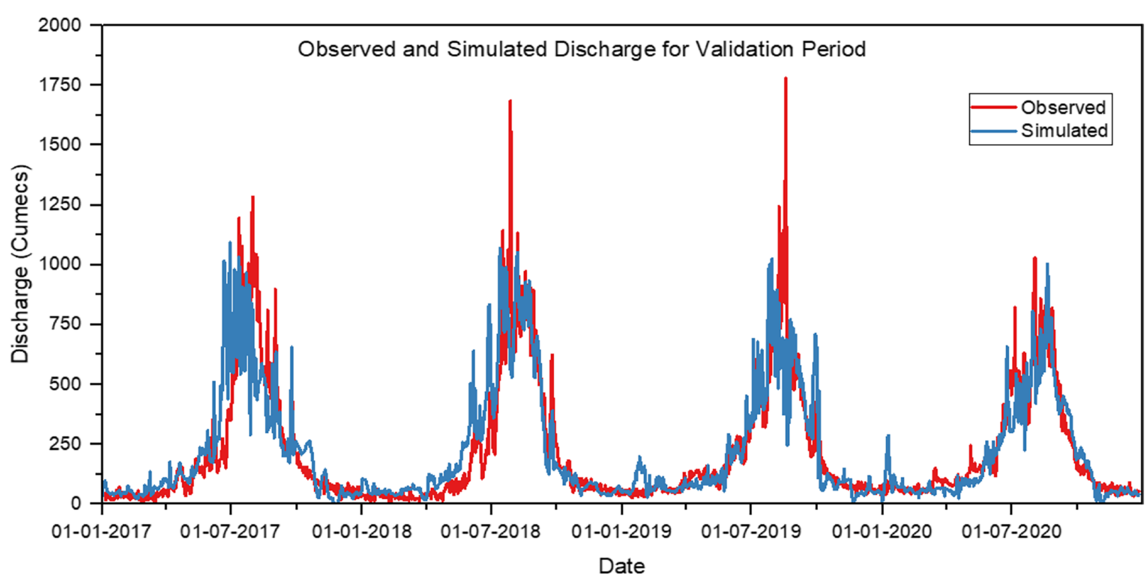


Fig. 13 Graph between observed and simulated discharge for the validation period

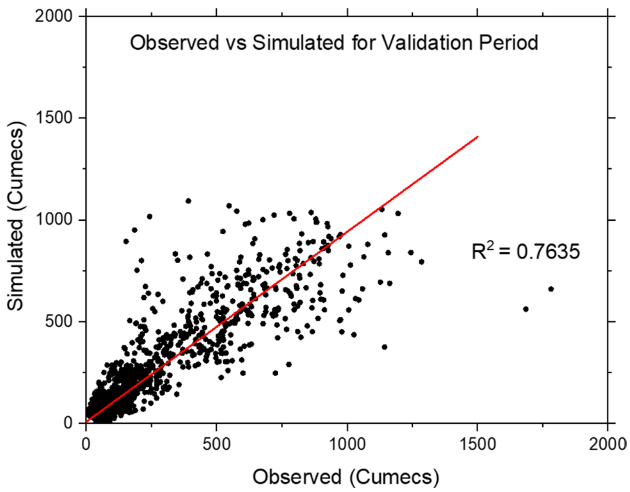


Fig. 14 Scatter plot for validation period (2017–2020)

theory that applies the structural minimisation principle. A nonlinear classifier or regression line can be found using the kernel function known as the radial basis kernel in the machine learning model. The model exhibits excellent efficiency when applying the specific model to the Daily and Ten Daily time series whilst considering various input variables. The prediction effectiveness is evaluated using the two performance indicators, NSE and R^2.

Cross-validation (CV) is sometimes referred to as a resampling method because it requires fitting the same statistical way several times using various subsets of the data. The data set will be divided into two parts for cross-validation: a first part for training the model and a second for evaluating it. The prediction error will be estimated to determine the model's accuracy. The k-fold cross-validation calculates the average prediction error rate after evaluating the SVR model's performance on various subsets of the training data. The data is divided into *k* folds randomly to begin the procedure (Fig. 5). The preferred type of SVR model is then provided in sequence to the *k*-onefold once *k* iterations of training and testing have been completed (Yoon et al. 2017). The first fold is utilised in the first iteration to test the model, whilst the remaining folds are used to train the model. The second fold is used as the testing set, and the remaining folds are the training set in the second iteration. This process is repeated until all of the k folds have been used as the testing set. After the model has been developed in a training phase, it will be checked on the test dataset. The forecast error will be calculated after that. *K*-fold cross-validation

respectively (Annexure I). The model shows high performance in predicting discharge at ten daily time scales (Figs. 19, 20, 21 and 22; Table 5).

A data fitting-based machine learning technique called a support vector regression (SVR) was first presented by Vapnik (1995). Numerous sectors, including streamflow prediction and water resources, have effectively used this approach. Dibike et al. (2001) described the first application of the SVR model to water-related topics and rainfall-runoff modelling. The support vector regression (SVR) is an effective learning system based on bounded optimisation

Table 3 NSE and R^2 for calibration and validation period

Case-2 (three input variables have been considered for daily streamflow prediction)	Time	NSE	R^2
Calibration	Jan-2006 to Dec-2016	86.68	0.8617
Validation	Jan-2017 to Dec-2020	75.85	0.7635

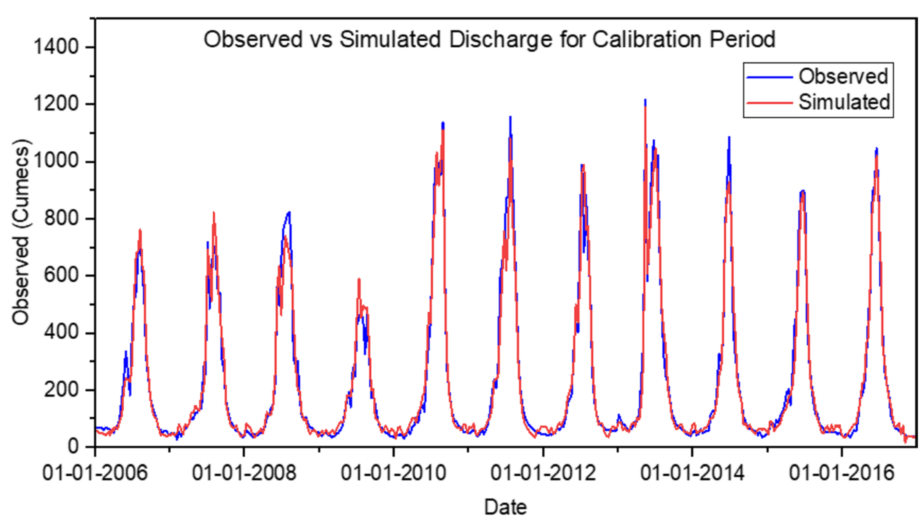


Fig. 15 Graph between observed and simulated ten daily discharges for the calibration period

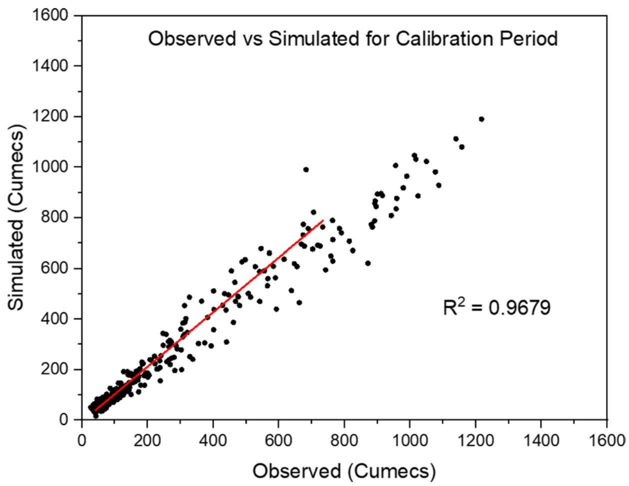


Fig. 16 Scatter plot for calibration period (2006–2016)

(CV) is reliable for assessing a model's correctness. The benefit of k-fold CV is that it consistently provides estimates of the test error rate that are more accurate (Juahir et al. 2011). A smaller value of K is inappropriate since it is more biassed. Larger K values, however, can lead to increased variance even though they are less biassed. These values have been shown empirically to yield test error rate estimates that suffer neither excessively high bias nor very high variance (Huang et al. 2015).

The SVR model for daily streamflow prediction considering four input variables (Q_{t-1}, R_{t-1}, T_{t-1}, SCA_{t-1}) shows excellent efficiency. There is no lag between the observed and Predicted time series. The NSE and R^2 are computed at a yearly time scale for observed and predicted discharge, which shows excellent efficiency. The model for daily streamflow prediction having three input variables does not work well because lag is present in the observed and predicted time series. However, the overall efficiency is good. The SVR model for ten daily streamflows considering four input variables shows good efficiency as the NSE and R^2 are 96.77 and 95.60 for the calibration and validation period. The NSE and R^2 are computed at a yearly time scale for observed and simulated discharge series, which also shows excellent efficiency. The discharge data are a guiding variable in prediction at daily and ten daily time scales. The ten daily streamflow predictions considering three variables (R_{t-1}, T_{t-1}, SCA_{t-1}) show good efficiency.

Conclusions

In this research work, Daily and Ten daily streamflows are predicted using the Support Vector Regression (SVR) Machine learning Model. Two combination of Input variables have been used in generation of daily and Ten daily Streamflow (i) Prediction (Q_t) considering four input

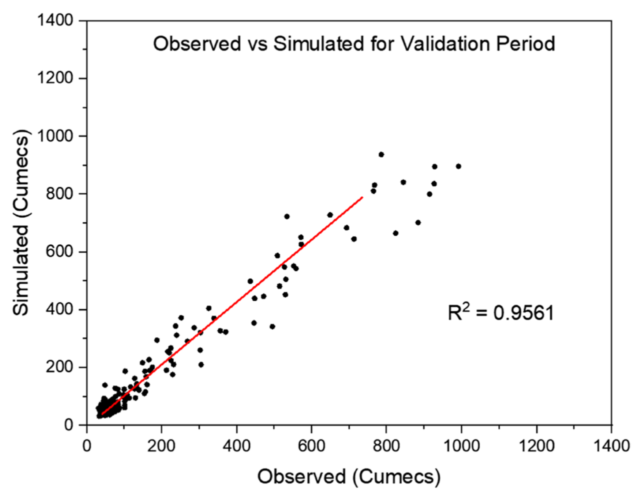


Fig. 18 Scatter plot for validation period (2017–2020)

Fig. 17 Graph between observed and predicted discharge for the validation period

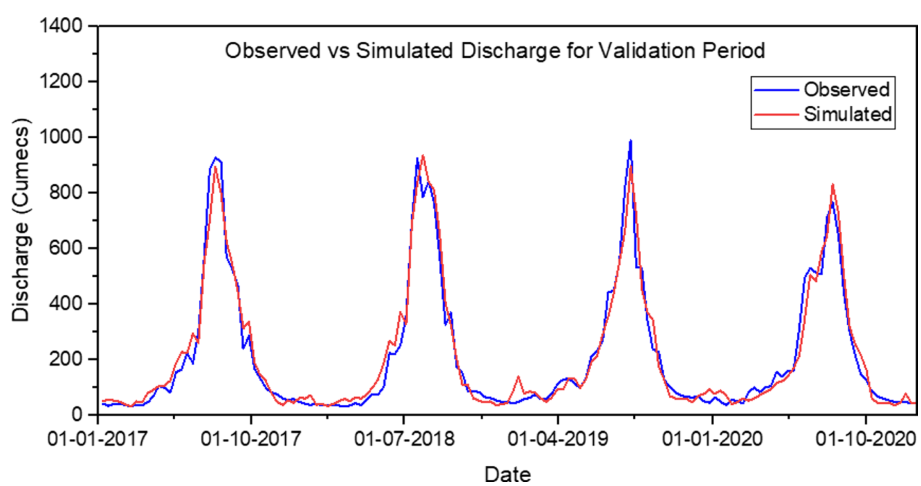


Table 4 NSE and R^2 for calibration and validation period

Case-1 (four input variables have been considered for daily streamflow prediction)	Time	NSE	R^2
Calibration	Jan-2006 to Dec-2016	96.77	0.9679
Validation	Jan-2017 to Dec-2020	95.60	0.9561

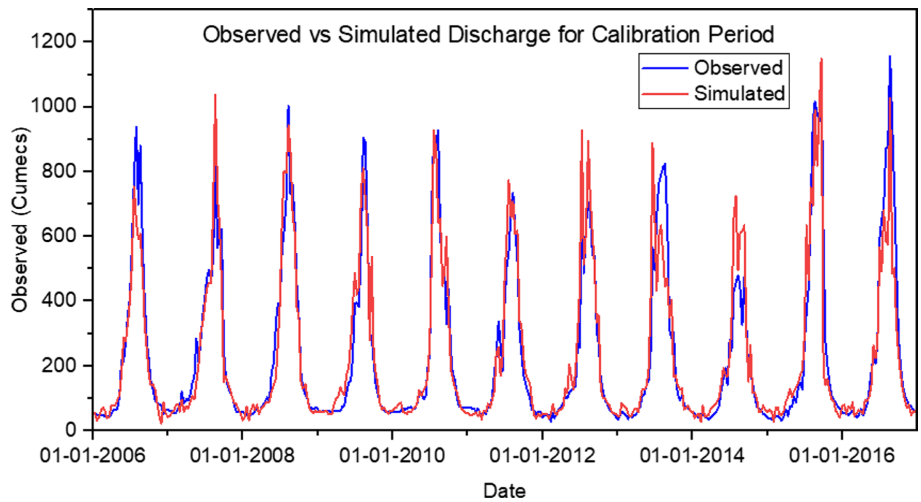


Fig. 19 Graph between observed and simulated discharge for the calibration period

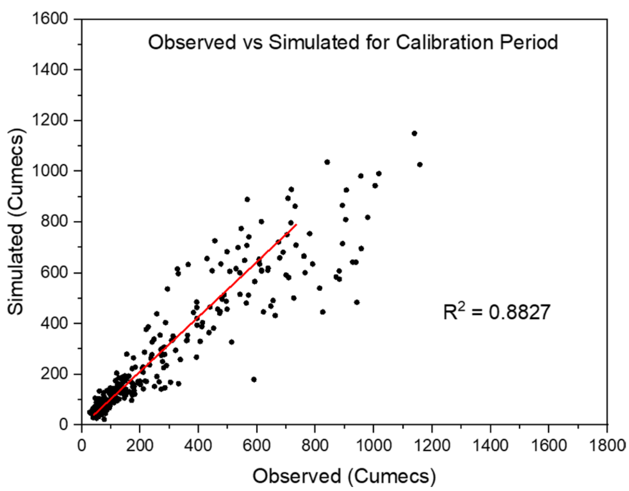


Fig. 20 Scatter plot for validation period (2017–2020)

variables {Discharge (Q_{t-1}), Rainfall (R_{t-1}), Temperature (T_{t-1}), Snow Cover Area (SCA_{t-1})} (ii) Prediction (Q_t) considering three input variables {Rainfall (R_{t-1}), Temperature (T_{t-1}), Snow Cover Area (SCA_{t-1})}. It is very tedious and time-consuming to select the input variables in modelling complex hydrological processes (Moghaddamnia et al. 2009; Kakaei Lafdani et al. 2013; Mahmoodzadeh et al. 2016; Malik et al. 2019b). The output (Q_t) is evaluated considering different sets of input parametres using *K*-fold cross-validation. 75% of the data is used for Calibration and 25% for validation. The results revealed that the SVR approach is reliable and efficient for streamflow prediction. Using the Radial kernel function helped obtain the high dimensionality, resulting in the expected outcomes from the study. The choice of kernel defines the promising results for the Support vector Regression model. The parameter cost '*c*' and gamma '*g*' are adjusted to optimise the hyperparametres, and the approach was presented by Cherkassky and Ma (2004). The quality of SVR models depends on the proper setting of SVR hyper-parametres. The two performance indicators, Nash Sutcliffe efficiency (NSE) and Coefficient of Determination (R^2) were used in the study to evaluate the efficiency of the prediction. The two-performance indicator shows excellent prediction quality and states that the SVR technique can be successfully used for nonlinear applications in Hydrology. After fuzzy and artificial neural networks, the SVR is the most promising development in the hydrological field. SVR is suitable for other purposes such as rainfall runoff, streamflow prediction and sediment yield forecasting, evaporation and evapotranspiration forecasting, Lake and reservoir water level prediction, Flood forecasting, Drought forecasting, Groundwater level prediction, Soil moisture estimation, Groundwater quality assessment Cherkassky and Ma (2004). The SVR touches on the many facets of computational hydrology. The framework can be a foundation for future researchers to build more exact hybrid mechanisms and expand the use of support vector regression approaches in complex hydrological prediction.

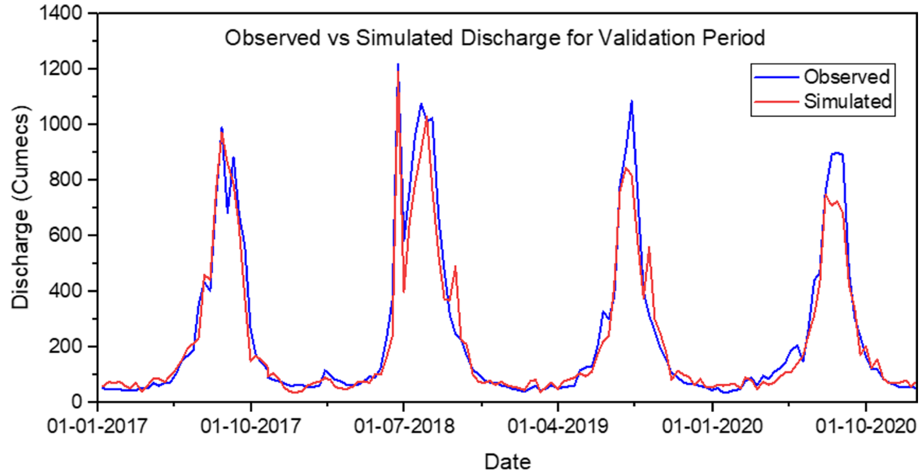


Fig. 21 Graph between observed and predicted discharge for the calibration period

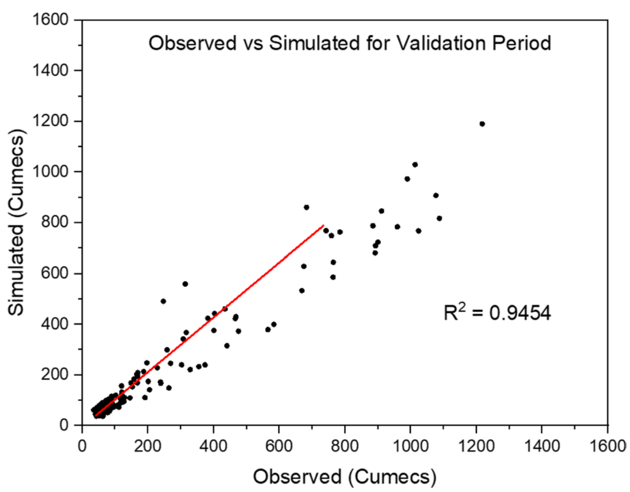


Fig. 22 Scatter plot for validation period (2017–2020)

Table 6 Year-wise NSE and R^2 for the period 2006–2020

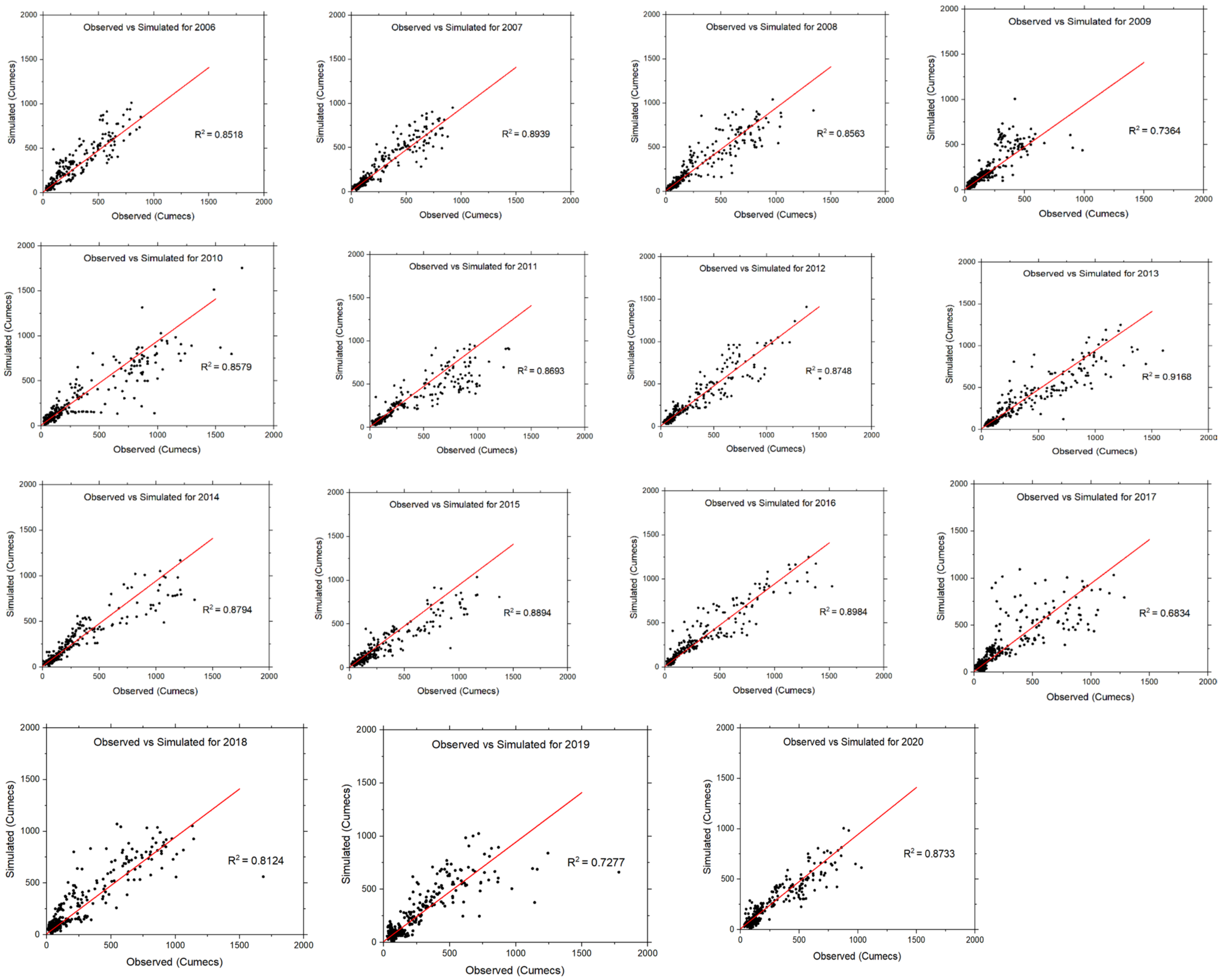
Year	NSE	R-squared (R^2)
2006	81.67	0.8518
2007	88.73	0.8939
2008	62.24	0.8563
2009	83.57	0.7364
2010	82.57	0.8579
2011	87.43	0.8693
2012	89.32	0.8748
2013	88.52	0.9168
2014	86.76	0.8704
2015	82.68	0.8834
2016	189.61	0.8984
2017	67.54	0.6834
2018	79.66	0.8124
2019	72.07	0.7277
2020	86.74	0.8733

Table 5 NSE and R^2 for calibration and validation period

Case-2 (three input variables have been considered for daily streamflow prediction)	Time	NSE	R^2
Calibration	Jan-2006 to Dec-2016	88.22	0.8827
Validation	Jan-2017 to Dec-2020	92.52	0.9454

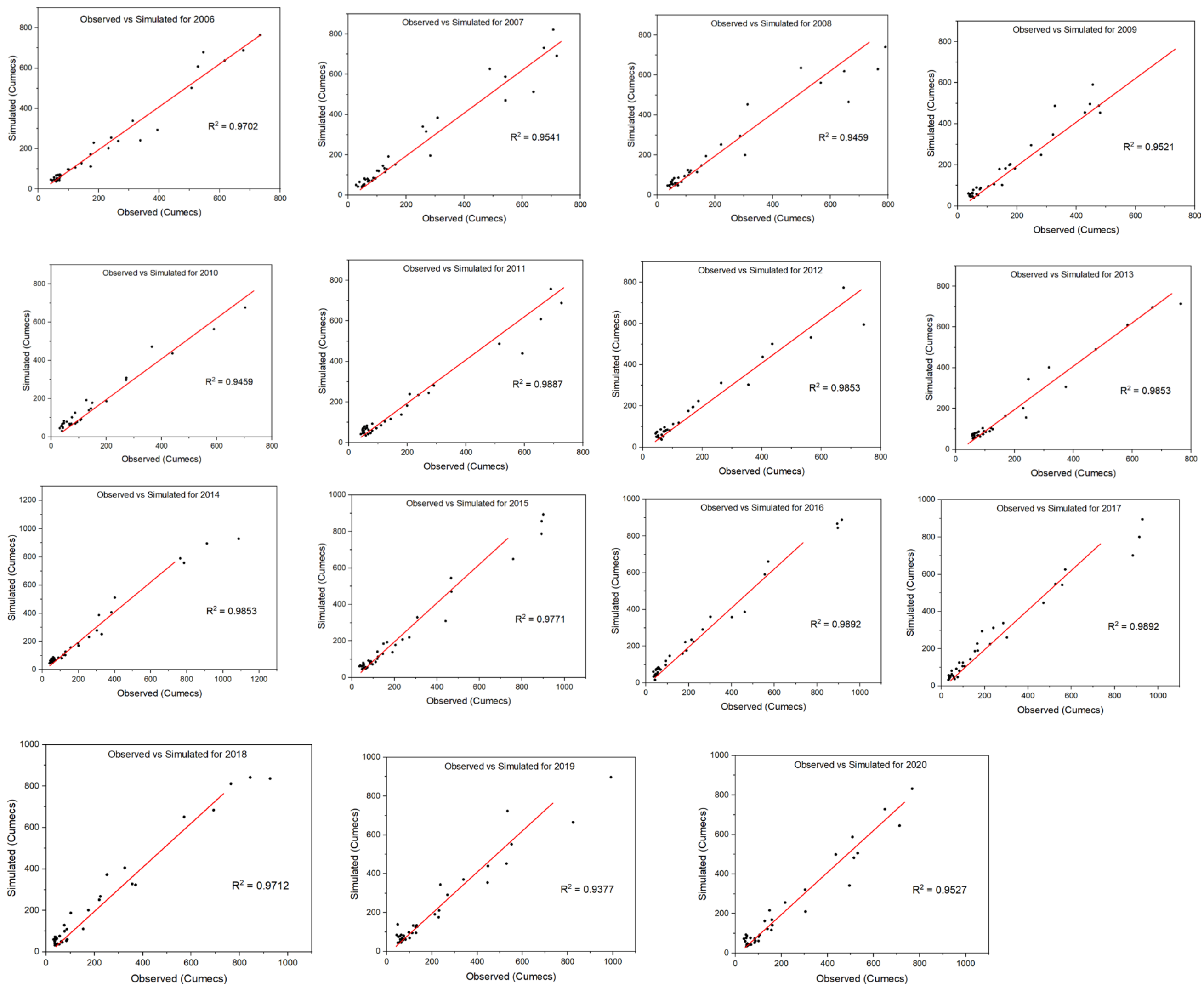
Annexure I

1. **Daily streamflow prediction considering three input variables (rainfall, temperature, snow cover area)**



See Table 6.

2. **10 daily streamflow prediction considering four input variables (10 daily avg. discharge, 10 daily average rainfall, 10 daily average temperature, 10-day snow cover area)**



See Table 7.

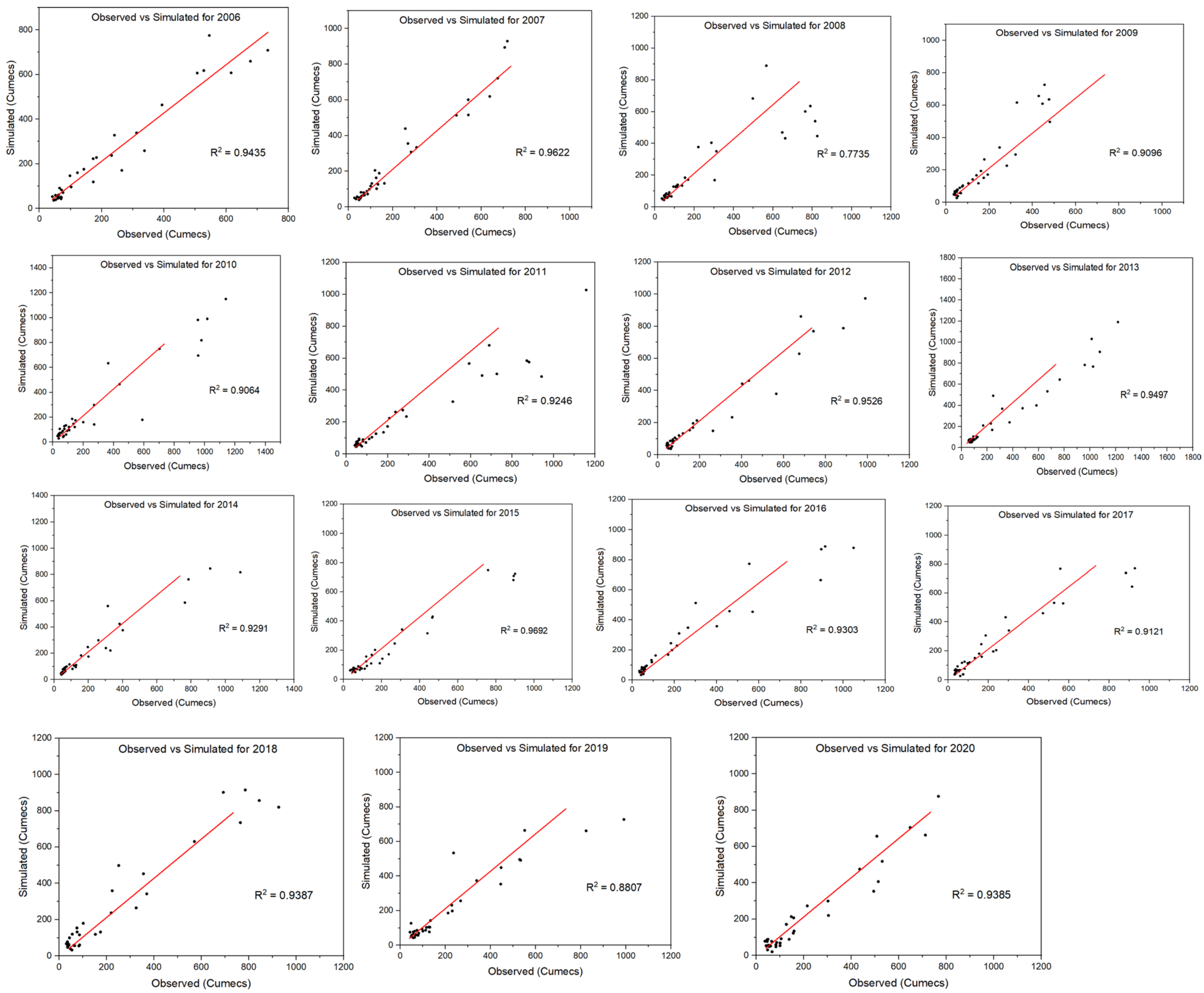
Table 7 Year-wise NSE and R^2 for the period 2006–2020

Year	NSE	R-squared (R^2)
2006	96.08	0.9702
2007	94.62	0.954
2008	93.78	0.945
2009	91.72	0.9521
2010	90.45	0.9459
2011	94.23	0.9887
2012	93.46	0.9853
2013	98.32	0.9853
2014	97.66	0.9781
2015	97.26	0.977
2016	98.76	0.9892
2017	96.41	0.9892
2018	96.5	0.9712
2019	93.67	0.9377
2020	95.02	0.9527

Table 8 Year-wise NSE and R^2 for the period 2006–2020

Year	NSE	R-squared (R^2)
2006	90.93	0.9435
2007	89.57	0.9622
2008	94.12	0.7735
2009	88.75	0.90.96
2010	92.25	0.9094
2011	91.97	0.9246
2012	91.35	0.9526
2013	77.07	0.9497
2014	61.15	0.9291
2015	90.57	0.9692
2016	85.09	0.9303
2017	95.25	0.9121
2018	93.33	0.9387
2019	91.89	0.8807
2020	93.23	0.9385

3. **10 daily streamflow prediction considering three input variables (10 daily average rainfall, 10 daily average temperature, 10-day snow cover area).**



See Table 8.

Acknowledgements The authors acknowledge the Department of Hydrology, IIT Roorkee to provide Article Processing Charges (APC). The authors also acknowledge THDCIL, Rishikesh to provide the Data for the study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Adnan MSG, Dewan A, Zannat KE, Abdullah AYM (2019) The use of watershed geomorphic data in flash flood susceptibility zoning: a

case study of the Karnaphuli and Sangu River basins of Bangladesh. Nat Hazards 99:425–448
Afan HA, El-shafie A, Mohtar WHMW, Yaseen ZM (2016) Past, present and prospect of an Artificial Intelligence (AI) based model for sediment transport prediction. J Hydrol 541:902–913
Al-Sudani ZA, Salih SQ, Yaseen ZM (2019) Development of multivariate adaptive regression spline integrated with differential evolution model for streamflow simulation. J Hydrol 573:1–12
Anguita D, Ghio A, Ridella S, Sterpi D (2009) K-fold cross validation for error rate estimate in support vector machines. In: DMIN, pp. 291–297.
Behnke R, Vavrus S, Allstadt A, Albright T, Thogmartin WE, Radeloff VC (2016) Evaluation of downscaled, gridded climate data for the conterminous United States. Ecol Appl 26(5):1338–1351
Caldwell P (2010) California wintertime precipitation bias in regional and global climate models. J Appl Meteorol Climatol 49(10):2147–2158
Cherkassky V, Ma Y (2004) Practical selection of SVM paraMetres and noise estimation for SVM regression. Neural Netw 17(1):113–126
Chou HK, Ochoa-Tocachi BF, Moulds S, Buytaert W (2022) Parameterizing the JULES land surface model for different land covers in the tropical Andes. Hydrol Sci J 67(10):1516–1526
Coops NC, Wulder MA, Iwanicka D (2009) Large area monitoring with a MODIS-based Disturbance Index (DI) sensitive to annual and seasonal variations. Remote Sens Environ 113(6):1250–1261
Cortes C, Vapnik V (1995) Support-Vector Networks Machine Learning 20:273–297
Dibike YB, Velickov S, Solomatine D, Abbott MB (2001) Model induction with support vector machines: introduction and applications. J Comput Civ Eng 15(3):208–216
Elbeltagi A, Di Nunno F, Kushwaha NL, de Marinis G, Granata F (2022) River flow rate prediction in the Des Moines watershed (Iowa, USA): a machine learning approach. Stoch Env Res Risk Assess 36(11):3835–3855
Elshorbagy A, Corzo G, Srinivasulu S, Solomatine DP (2010) Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology-Part 1: concepts and methodology. Hydrol Earth Syst Sci 14(10):1931–1941
Ghaemi A, Rezaie-Balf M, Adamowski J, Kisi O, Quilty J (2019) On the applicability of maximum overlap discrete wavelet transform integrated with MARS and M5 model tree for monthly pan evaporation prediction. Agric for Meteorol 278:107647
Goswami BN, Venugopal V, Sengupta D, Madhusoodanan MS, Xavier PK (2006) Increasing trend of extreme rain events over India in a warming environment. Science 314(5804):1442–1445
Granata F, Gargano R, De Marinis G (2016) Support vector regression for rainfall-runoff modeling in urban drainage: a comparison with the EPA's storm water management model. Water 8(3):69
Guhathakurta P, Sreejith OP, Menon PA (2011) Impact of climate change on extreme rainfall events and flood risk in India. J Earth Syst Sci 120:359–373
Hadi SJ, Tombul M (2018) Monthly streamflow forecasting using continuous wavelet and multi-gene genetic programming combination. J Hydrol 561:674–687
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, pp 770–778.
Huang H, Liang Z, Li B, Wang D, Hu Y, Li Y (2019) Combination of multiple data-driven models for long-term monthly runoff predictions based on Bayesian model averaging. Water Resour Manage 33:3321–3338
Hussain D, Khan AA (2020) Machine learning techniques for monthly river flow forecasting of Hunza River, Pakistan. Earth Sci Inf 13:939–949
Imrie CE, Durucan S, Korre A (2000) River flow prediction using artificial neural networks: generalisation beyond the calibration range. J Hydrol 233(1–4):138–153
Ireson AM, Barr AG, Johnstone JF, Mamet SD, Van der Kamp G, Whitfield CJ, ... Sagin J (2015) The changing water cycle: the Boreal Plains ecozone of Western Canada. Wiley Interdiscip Rev: Water 2(5):505–521
Juahir H, Zain SM, Yusoff MK, Hanidza TT, Armi AM, Toriman ME, Mokhtar M (2011) Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. Environ Monit Assess 173:625–641
Keshtegar B, Allawi MF, Afan HA, El-Shafie A (2016) Optimized river stream-flow forecasting model utilizing high-order response surface method. Water Resour Manage 30:3899–3914
Khosravi K, Golkarian A, Tiefenbacher JP (2022) Using optimized deep learning to predict daily streamflow: a comparison to common machine learning algorithms. Water Resour Manage 36(2):699–716
Kormos PR, Marks DG, Seyfried MS, Havens SC, Hedrick A, Lohse KA, ... Garen D (2018) 31 years of hourly spatially distributed air temperature, humidity, and precipitation amount and phase from Reynolds Critical Zone Observatory. Earth Syst Sci Data 10(2):1197–1205
Lafdani EK, Nia AM, Ahmadi A (2013) Daily suspended sediment load prediction using artificial neural networks and support vector machines. J Hydrol 478:50–62
Loaiciga HA, Valdes JB, Vogel R, Garvey J, Schwarz H (1996) Global warming and the hydrologic cycle. J Hydrol 174(1–2):83–127
Mahmoodzadeh A, Ghafourian H, Mohammed AH, Rezaei N, Ibrahim HH, Rashidi S (2023) Predicting tunnel water inflow using a machine learning-based solution to improve tunnel construction safety. Transp Geotech 40:100978
Malik A, Kumar A, Ghorbani MA, Kashani MH, Kisi O, Kim S (2019) The viability of co-active fuzzy inference system model for monthly reference evapotranspiration estimation: case study of Uttarakhand State. Hydrol Res 50(6):1623–1644
Malik A, Tikhamarine Y, Souag-Gamane D, Kisi O, Pham QB (2020) Support vector regression optimized by meta-heuristic algorithms for daily streamflow prediction. Stoch Env Res Risk Assess 34:1755–1773
Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Brief Bioinform 18(5):851–869
Moghaddamnia A, Gousheh MG, Piri J, Amin S, Han D (2009) Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. Adv Water Resour 32(1):88–97
Nourani V, Baghanam AH, Adamowski J, Kisi O (2014) Applications of hybrid wavelet–artificial intelligence models in hydrology: a review. J Hydrol 514:358–377
Pai DS, Rajeevan M, Sreejith OP, Mukhopadhyay B, Satbha NS (2014) Development of a new high spatial resolution (0.25× 0.25) long period (1901–2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. Mausam 65(1):1–18
Pandey P, Irulappan V, Bagavathiannan MV, Senthil-Kumar M (2017) Impact of combined abiotic and biotic stresses on plant growth and avenues for crop improvement by exploiting physio-morphological traits. Front Plant Sci 8:537
Pandhiani SM, Sihag P, Shabri AB, Singh B, Pham QB (2020) Time-series prediction of streamflows of Malaysian rivers using data-driven techniques. J Irrig Drain Eng 146(7):04020013
Rahmani-Rezaeieh A, Mohammadi M, Danandeh Mehr A (2020) Ensemble gene expression programming: a new approach for evolution of parsimonious streamflow forecasting model. Theoret Appl Climatol 139(1–2):549–564
Rajeevan M, Bhate J, Jaswal AK (2008) Analysis of variability and trends of extreme rainfall events over India using 104 years of gridded daily rainfall data. Geophys Res Lett 35(18):L18707

Rasouli A (2020) Deep learning for vision-based prediction: a survey. arXiv preprint arXiv:2007.00095

Riahi-Madvar H, Dehghani M, Memarzadeh R, Gharabaghi B (2021) Short to long-term forecasting of river flows by heuristic optimization algorithms hybridized with ANFIS. Water Resour Manage 35:1149–1166

Roy A, Chatterjee A, Tiwari S, Sarkar C, Das SK, Ghosh SK, Raha S (2016) Precipitation chemistry over urban, rural and high-altitude Himalayan stations in eastern India. Atmos Res 181:44–53

Sen Roy S, Balling RC Jr (2004) Trends in extreme daily precipitation indices in India. Int J Climatol: A Journal of the Royal Meteorological Society 24(4):457–466

Singh P, Bengtsson L (2004) Hydrological sensitivity of a large Himalayan basin to climate change. Hydrol Process 18(13):2363–2385

Solomatine DP, Ostfeld A (2008) Data-driven modelling: some past experiences and new approaches. J Hydroinf 10(1):3–22

Srivastava AK, Rajeevan M, Kshirsagar SR (2009) Development of a high resolution daily gridded temperature data set (1969–2005) for the Indian region. Atmospheric Sci Lett 10(4):249–254

Teutschbein C, Grabs T, Laudon H, Karlsen RH, Bishop K (2018) Simulating streamflow in ungauged basins under a changing climate: the importance of landscape characteristics. J Hydrol 561:160–178

Valipour M (2015) Long-term runoff study using SARIMA and ARIMA models in the United States. Meteorol Appl 22(3):592–598

Valipour M, Banihabib ME, Behbahani SMR (2013) Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. J Hydrol 476:433–441

Walton DB, Sun F, Hall A, Capps S (2015) A hybrid dynamical–statistical downscaling technique. Part I: development and validation of the technique. J Clim 28(12):4597–4617

Wu Z, Huang NE (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Anal 1(01):1–41

Wu CL, Chau KW, Fan C (2010) Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. J Hydrol 389(1–2):146–167

Yadav A, Chithaluru P, Singh A, Albahar MA, Jurcut A, Álvarez RM, ... Joshi D (2022. Suspended sediment yield forecasting with single and multi-objective optimization using hybrid artificial intelligence models. Mathematics 10(22):4263

Yaseen ZM, El-Shafie A, Jaafar O, Afan HA, Sayl KN (2015) Artificial intelligence-based models for stream-flow forecasting: 2000–2015. J Hydrol 530:829–844

Zaz SN, Romshoo SA, Krishnamoorthy RT, Viswanadhapalli Y (2019) Analyses of temperature and precipitation in the Indian Jammu and Kashmir region for the 1980–2016 period: implications for remote influence and extreme events. Atmos Chem Phys 19(1):15–37

Zhang YG, Tang J, Liao RP, Zhang MF, Zhang Y, Wang XM, Su ZY (2021) Application of an enhanced BP neural network model with water cycle algorithm on landslide prediction. Stoch Env Res Risk Assess 35:1273–1291