Arabian Journal of Geosciences (2022) 15: 1369
https://doi.org/10.1007/s12517-022-10555-y

ORIGINAL PAPER

Prediction of river discharge of Kesinga sub-catchment of Mahanadi basin using machine learning approaches

Shreya Nivesh[1] · Diksha Negi[2] · Pradeep Singh Kashyap[3] · Sejal Aggarwal[4] · Balraj Singh[5] · Bhagwat Saran[3] · Pragati N. Sawant[6] · Parveen Sihag[7,8]

Received: 23 July 2021 / Accepted: 26 June 2022 / Published online: 4 August 2022

Abstract

River discharge is a relevant ingredient of the hydrological cycle for a wide scale of utilizations and evaluation of water assets, plan of water-related designs and flood admonitory and relief plans. The predictive discharge of the basin using the machine learning approaches is therefore significant for managing water resources and the prevention of flooding control. This investigation evaluated the viability of several machine learning methods, M5P tree, Random forest, Regression tree, reduced error pruning tree, Gaussian process and support vector machine, to predict the basin discharge of the Kesinga basin. Various statistical measures, i.e. correlation coefficient, mean absolute error, root mean square error, Willmott's index, Nash–Sutcliffe efficiency coefficient, Legates and McCabe's index and normalized root mean square, error were utilized to assess the performance of the developed model. The presentation of random forest and M5P models was found to be the best when compared with the regression tree, reduced error pruning tree, Gaussian process and support vector machine–based models. Overall RF-based model gave the best results among all applied models for predicting water discharge for the Kesinga basin with the coefficient of determination (R^2) values of 0.978 and 0.890 for the training and testing stages, respectively. The main significance of soft computing techniques is that they help users solve real-world problems by providing approximate results that conventional and analytical models cannot solve.

Keywords Water discharge · M5P · Random forest · Regression tree · Reduced error pruning · Gaussian process · Support vector machine

Responsible Editor: Broder J. Merkel

✉ Parveen Sihag
parveen12sihag@gmail.com; parveen.e11725@cumail.in

Shreya Nivesh
shreyanivesh@gmail.com

Diksha Negi
dikshan220@gmail.com

Pradeep Singh Kashyap
pskashyap@gmail.com

Sejal Aggarwal
sejal.170501023@muj.manipal.edu

Balraj Singh
balrajzinder@gmail.com

Bhagwat Saran
saran.bhagwat007@gmail.com

Pragati N. Sawant
pragatisawant2012@gmail.com

[1] ICAR-Mahatma Gandhi Integrated Farming Research Institute, East Champaran 845429, Bihar, India

[2] School of Biological and Environmental Sciences, Shoolini University, Solan, India

[3] Department of Soil and Water Conservation Engineering, College of Technology, Govind Ballabh Pant University of Agriculture and Technology, Pantnagar 263145, Uttarakhand, India

[4] Department of Architecture and Design, Manipal University, Jaipur, Rajasthan, India

[5] Department of Civil Engineering, Panipat Institute of Engineering and Technology, Panipat, India

[6] Department of Soil and Water Conservation Engineering, Mahatma Phule Krishi Vidyapeeth, Ahmednagar-413722, Rahuri, Maharashtra, India

[7] Civil Engineering Department, Chandigarh University, Mohali, Punjab, India

[8] University Centre for Research & Development, Chandigarh University, Mohali, Punjab, India

Introduction

Rivers play a very important role in our Earth's climate system. They ensure the link between both the atmosphere and the ocean (Vörösmarty et al. 2000). In many parts of the world, rivers are the only available water source holding up the local socio-economic development (Sullivan 2002). Hence, monitoring and quantification of river flow becomes essential not only for its sustainable management but also to estimate the future possible conditions (Zakharova et al. 2020). Hydrology and water resource management require the quantity of streamflow. The information on streamflow propagation speed and time for streams to pass downstream is analytical for flood prediction, supply tasks and watershed displaying (Brakenridge et al. 2012). Consequently, there is a requirement for long haul, ceaseless, spatially reliable and promptly accessible streamflow data. Streamflow is presently recorded at waterway measuring stations, although access to data is intermittent or non-existent, especially in developing nations, and they are under restrictive control in advanced countries (Calmant and Seyler 2006). These difficulties limit the research that requires waterway release information. A basic factor in assessing waterway release lies in the capacity to practically estimate spatial pressure-driven factors (i.e. width, depth and speed) and additionally to set up the connections between them (Mersel et al. 2013). The ground perception technique is the most exact proportion for streamflow. Ground waterway release is attained by assessing the pressure-driven attributes of stream channels including width, depth and speed (Stutter et al. 2021). These gauge discharge estimates form the backbone of human water management decisions and hydrologic science. Variability in rainfall and potential evaporation are the primary reasons for annual variation in the surface river discharge from a basin (Chien et al. 2013). Increasing human activities such as dam constructions and operations, land use/land cover (LULC) change, surface and groundwater extractions, mining, etc. have resulted in the changes in river discharges (Destouni et al. 2013). Furthermore, the relation between river discharge and LULC changes varies depending upon the location and size of basins, land management, elevation and LULC types (Li et al. 2001).

By then, fluctuated contemplate to various investigations in regard to Mahanadi waterway through the utilization of hydro-climatic factors like temperature, precipitation and streamflow (Rao 1993, 1995; Gosain et al. 2006; Raje and Mujumdar 2009; Asokan and Dutta 2008; Ghosh et al. 2010). Gosain et al. (2006) assessed that due to the variation in environment, the severity of floods turns viral, and this also causes an impact on the Mahanadi River basin. Ghosh et al. (2010) examined the pattern in Mahanadi under a future environment situation and noticed a declining pattern in the drift of Mahanadi at Hirakud.

In the most recent couple of years, different soft computing techniques like random forest, support vector machine, artificial neural network, Gaussian process and M5P model tree are effectively executed in engineering and water asset issues (Singh et al. 2022, 2021, 2019; Bhoria et al. 2021; Sepahvand et al. 2021; Sihag et al. 2020; Pandhiani et al. 2020). Garg et al (2022) used two soft computing techniques, artificial neural network and genetic programming, in the prediction of the streamflow and found that both soft computing techniques work well. Sridharam et al. (2021) implemented soft computing techniques (layered recurrent neural network, coactive neuro-fuzzy inference system and cascade forward back propagation neural network) and got reliable results in the prediction of streamflow. Muhammad Adnan et al. (2019) also investigated the potential of soft computing techniques and found them suitable for the prediction of the discharge of a river. Hence, the soft computing techniques are the technique which can be used in the discharge prediction. Also, these techniques solve the real-life problem in an efficient way which is very hard to analyse using conventional methods. Keeping this in view, the current study focuses on the analysis of different soft computing techniques and change points in various hydrometeorological variables specifically rainfall, evapotranspiration, inflow discharge (inflow), percolation, groundwater, surface runoff, water yield, potential ET and discharge in Kesinga basin.

Methodology

The data sets used for the present study include hydrological data (surface runoff, inflow and discharge), meteorological data (rainfall, evapotranspiration, potential evapotranspiration) and groundwater data (percolation, water yield contribution and groundwater) on monthly basis for the years 1990–2004. The meteorological data (1990–2004) were collected from the India Meteorological Department (IMD), Pune. Hydrological data (1990–2004) were collected from the Central Water Commission (CWC), Mahanadi and Eastern Rivers division, Bhubaneswar, Odisha. Rainfall and runoff data were recorded at Kasinga gauging station as shown in Fig. 1. Groundwater data (1990–2004) were collected from the Central Groundwater Board, South Eastern Region Bhubaneshwar, Odisha. The evapotranspiration data and missing observed data of streamflow were supplemented by the outputs of the SWAT hydrological model.

The area selected for the research is the Kesinga sub-catchment of the Mahanadi basin. The Mahanadi River is one of the main streams in India which establishes freshwater supply

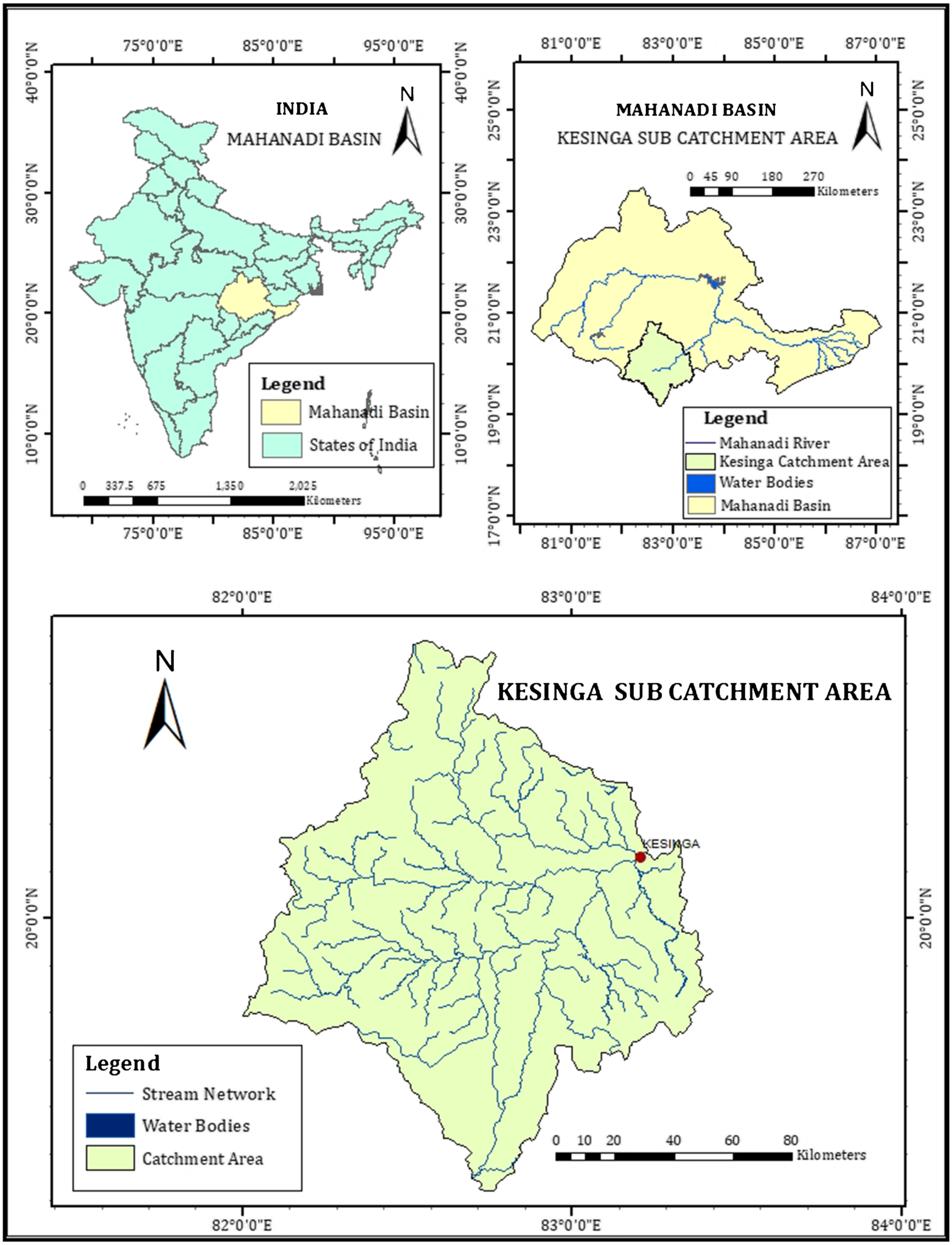


Fig. 1 Kesinga catchment area of Mahanadi basin situated in Odisha state

for irrigation, commercial and household use in the watershed (Agarwal et al. 2019). The Kesinga sub-catchment covers an approximate area of about 11,855 km^2, which expands from east longitudes of 82′21°–83′24° and north latitudes of 19′15°–20′44°. Most of the sub-catchment is located in the Kalahandi district of Odisha which has a population density of 50–100 people for each square kilometre. The elevation of the area is 187 m and the land corresponds to a flatter topography. Throughout the region, the product of the constant progression of the water stream is fine and medium-textured soil. Such soil types are productive and quite appropriate for husbandry. Kesinga basin is rich in its water resources which comprise multiple reservoirs, dams, barrages, wells, etc. In the study area, the higher temperature is felt in May and the lowest temperature in December. In summer, the temperature goes from 25 to 40 °C, and in winter, the temperature ranges between 11 and 27 °C. Maximum rainfall is observed in the monsoon season from June to September.

Performance criteria

There is a necessity for the assessment of the performance of the models for analysing the data with model evaluation utilising various methods. There are various statistical methods to evaluate the performance of the developed model using observational and computational values of the model. In this analysis, the correlation coefficient, mean absolute error, root mean square error, Willmott's index, scattering index, Nash–Sutcliffe efficiency index, normalized root mean square error and Legates and McCabe's index are performance assessment indices that are carried out in the present investigation to assess the fitting capability of the techniques.

Mean absolute error (MAE)

It is the measure of error between observations that express the same phenomenon. The MAE is calculated as follows.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|A - B| \tag{1}$$

Root mean square error (RMSE)

Root mean square error is generally calculated to determine numeric evaluation. RMSE is calculated as in the following:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A - B)^2} \tag{2}$$

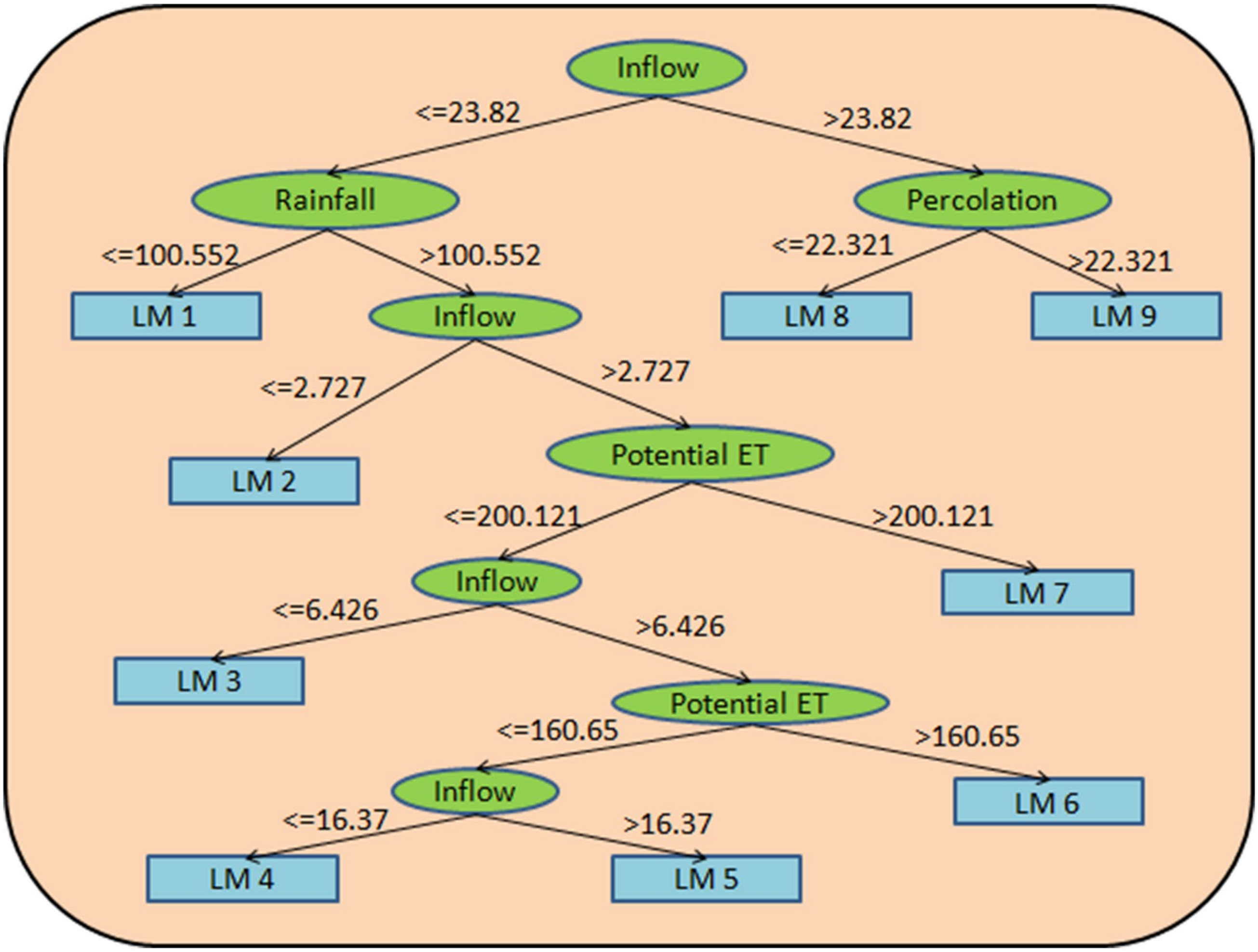


Fig. 2 Structure of M5P unpruned model used in this study

Coefficient correlation

It is used to analyse the performance of any model using a numerical value. The CC is given as:

$$\mathrm{CC} = \frac{\sum_{i=1}^{N}\left(A-\overline{A}\right)(B-\overline{B})}{\sqrt{\sum_{i=1}^{N}(A-\overline{A})^{2}}\sqrt{\sum_{i=1}^{N}(B-\overline{B})^{2}}} \tag{3}$$

Nash–Sutcliffe efficiency coefficient (NSE)

It is implemented to examine the predictive power of the models. It is expressed by (Nash and Sutcliffe, 1970):

$$\mathrm{NSE} = 1 - \left[\frac{\sum_{i=1}^{N}(A-B)^{2}}{\sum_{i=1}^{N}(\overline{A}-\overline{B})^{2}}\right] \tag{4}$$

Willmott's index (WI)

It is a standardized measure of the degree of model prediction error which varies between 0 and 1. It is expressed as (Willmott 1981):

$$\mathrm{WI} = 1 - \left[\frac{\sum_{i=1}^{N}(B-A)^{2}}{\sum_{i=1}^{N}\left(\left|B-\overline{A}\right|+\left|A-\overline{A}\right|\right)^{2}}\right] \tag{5}$$

Normalized root mean square error (NRMSE)

It is expressed as a percentage, where lower values indicate less residual variance.

$$\mathrm{NRMSE} := \frac{1}{\overline{B}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(A-B)^{2}} \tag{6}$$

Legates and McCabe's Index

It is utilized for measuring the accuracy of the model.

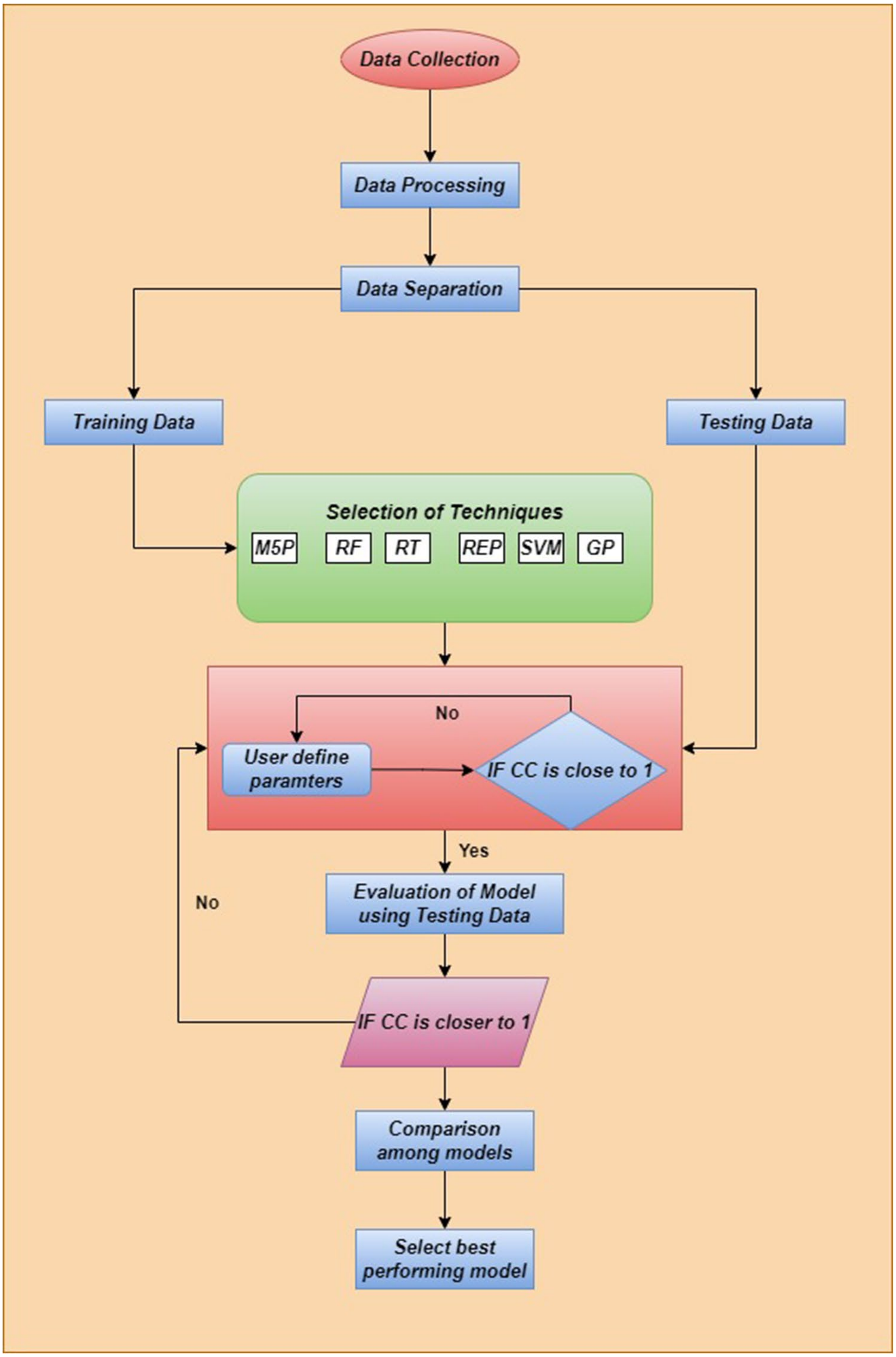
$$\mathrm{LMI} = 1 - \lceil\frac{\sum_{i=1}^{N}|A-B|}{\sum_{i=1}^{N}\left|B-\overline{B}\right|}\rceil \tag{7}$$

where A is the actual values, B is the predicted values, N is the number of observations and $\overline{B}$ is the average predicted value.

Table 1 Features of the data set

	Units	Data set	Mean	Standard deviation	Minimum	Maximum	Kurtosis	Skewness	Confidence level (95.0%)
Rainfall	(mm)	Training	116.36	157.46	0	694.2	1.67	1.56	27.11
		Testing	134.11	211.21	0	873.2	3.13	1.92	61.33
Inflow	(m^3/sec)	Training	42.93	68.97	0.183	370	8.85	2.88	11.87
		Testing	50.59	68.97	0.745	287.4	3.33	1.97	20.03
Evapotranspiration	(mm)	Training	68.78	47.87	0	164.2	−1.27	0.13	8.24
		Testing	64.62	48.61	0	148.8	−1.41	0.21	14.11
Percolation	(mm)	Training	31.03	66.95	0	333.9	6.23	2.58	11.53
		Testing	41.97	90.59	0	369.7	5.65	2.51	26.30
Groundwater	(mm)	Training	2.57	8.19	0	50.13	13.85	3.66	1.41
		Testing	1.62	5.82	0	30.09	18.13	4.26	1.69
Surface runoff	(mm)	Training	15.84	41.47	0	258.3	12.35	3.36	7.14
		Testing	26.59	68.56	0	367.7	13.66	3.49	19.91
Potential ET	(mm)	Training	161.06	56.47	72.92	304.1	−0.34	0.83	9.72
		Testing	167.97	60.35	92.05	314.3	−0.21	0.97	17.52
Water yield contribution	(mm)	Training	45.38	67.82	0.99	418.6	9.03	2.80	11.67
		Testing	63.07	88.62	1.4	454.1	7.72	2.54	25.73
Discharge	(m3/sec)	Training	188.35	227.17	3.29	916.2	1.65	1.64	39.11
		Testing	218.82	230.45	9.664	943.2	1.71	1.48	66.97

Fig. 3 Flow diagram of the methodology



Modelling approaches

M5P model

This model was initially established by Quinlan (1992) which is a combined type of decision tree learning process for both linear and nonlinear regression algorithms. The decision tree recommends a correlation between measured input data and rational learning output data which is relevant for categorized statistical inputs and outputs. The Model Tree algorithm assigns a one-dimensional function with output units as well as assigns

Table 2 The performance of M5P, RF, RT, REP Tree, GP_RBF, GB_PUK, SVM_RBF and SVM_PUK

Approaches	CC	MAE	RMSE	WI	NSE	LMI	NRMSE
	Training data set						
M5P	0.9644	34.3398	59.9091	0.9813	0.9299	0.8004	0.3181
RF	0.9893	20.3488	34.6361	0.9938	0.9766	0.8818	0.1839
RT	0.9999	1.9571	3.2966	0.9999	0.9998	0.9886	0.0175
REP Tree	0.9508	41.1466	70.0975	0.9743	0.9041	0.7609	0.3722
GP_RBF	0.9378	56.6186	82.5343	0.9591	0.8670	0.6710	0.4382
GP_PUK	0.9596	48.8750	73.0509	0.9675	0.8958	0.7160	0.3879
SVM_RBF	0.9565	32.8319	68.3398	0.9751	0.9088	0.8092	0.3628
SVM_PUK	0.9954	8.9104	21.7823	0.9977	0.9907	0.9482	0.1156
	Testing data set						
M5P	0.9465	51.2383	82.9280	0.9697	0.8678	0.7147	0.3790
RF	0.9438	44.4497	75.5598	0.9748	0.8902	0.7525	0.3453
RT	0.8900	83.5595	122.3565	0.9340	0.7121	0.5347	0.5592
REP Tree	0.8720	74.1214	115.4103	0.9413	0.7439	0.5873	0.5274
GP_RBF	0.9087	49.3153	99.1282	0.9567	0.8110	0.7254	0.4530
GP_PUK	0.8658	55.5909	119.5222	0.9370	0.7253	0.6904	0.5462
SVM_RBF	0.8929	59.3595	105.1593	0.9512	0.7873	0.6695	0.4806
SVM_PUK	0.8913	57.5504	104.8255	0.9516	0.7887	0.6795	0.4790

a multivariate regression model to each spatial domain by dividing and categorising the complete data storage into various input spaces. Furthermore, rather than discrete classes, the M5P technique is compact with constant class issues while handling tasks with very high dimensions. M5P is therefore not just quick and easy but is also a robust and appropriate method for predicting and modelling huge amounts of data. The structure of M5P unpruned model developed in this study is shown in Fig. 2.

Random forest (RF)

The random forest technique has been suggested by Breiman (1999) and has been used to produce an estimate which usually carried an organization of various trees. Every tree represents a specific categorization and also the vote categorization. The RF approach prefers a category that has optimum votes throughout the forest. The tree is fully developed unless the training set contains the number of *N* cases. *N* cases at random with the substitutes of the original information could be the input data set to fully mature the tree. Variable *m* is selected randomly from the input variables *K* for the best partition so that the value of *m* is not more than *K* and should be consistent. The tree is grown to its maximum possible without pruning. The set of data within each circumstance is handed down to each tree to arrange a new data set. Modelling a single tree is extremely complicated and sensitive, as small changes in the training data set often result in large variations in individual tree classifications, leading to a low accuracy rate (Breiman 1996). However, RF is relatively quick to achieve results and can be readily assimilated if there is the requirement of less computational time.

Random tree (RT)

The random tree is an algorithm-based technique that addresses both classification and regression problems. RT algorithm was originally established by Aldous (1993). Regression tree is an array or organizing of tree estimation methods; the method applied by the regression tree method includes a data set that collects information through input, categorizes it as a singular branch of a tree and eventually accepts the vote. They are instructed to use different training data set but similar elements. The development of certain data sets is made from total data using the bootstrap process. In each node of each tree, a subset of the parameter is used to derive the maximal split. Each node grows or develops a new subset, and furthermore, newly grown trees are not pruned.

Reduced error pruning (REP Tree)

The REP tree method is a speedy classification tree logic technique that uses the concept of computer technology–selected features with randomness and decrease variance inaccuracy (Quinlan, 1987). The

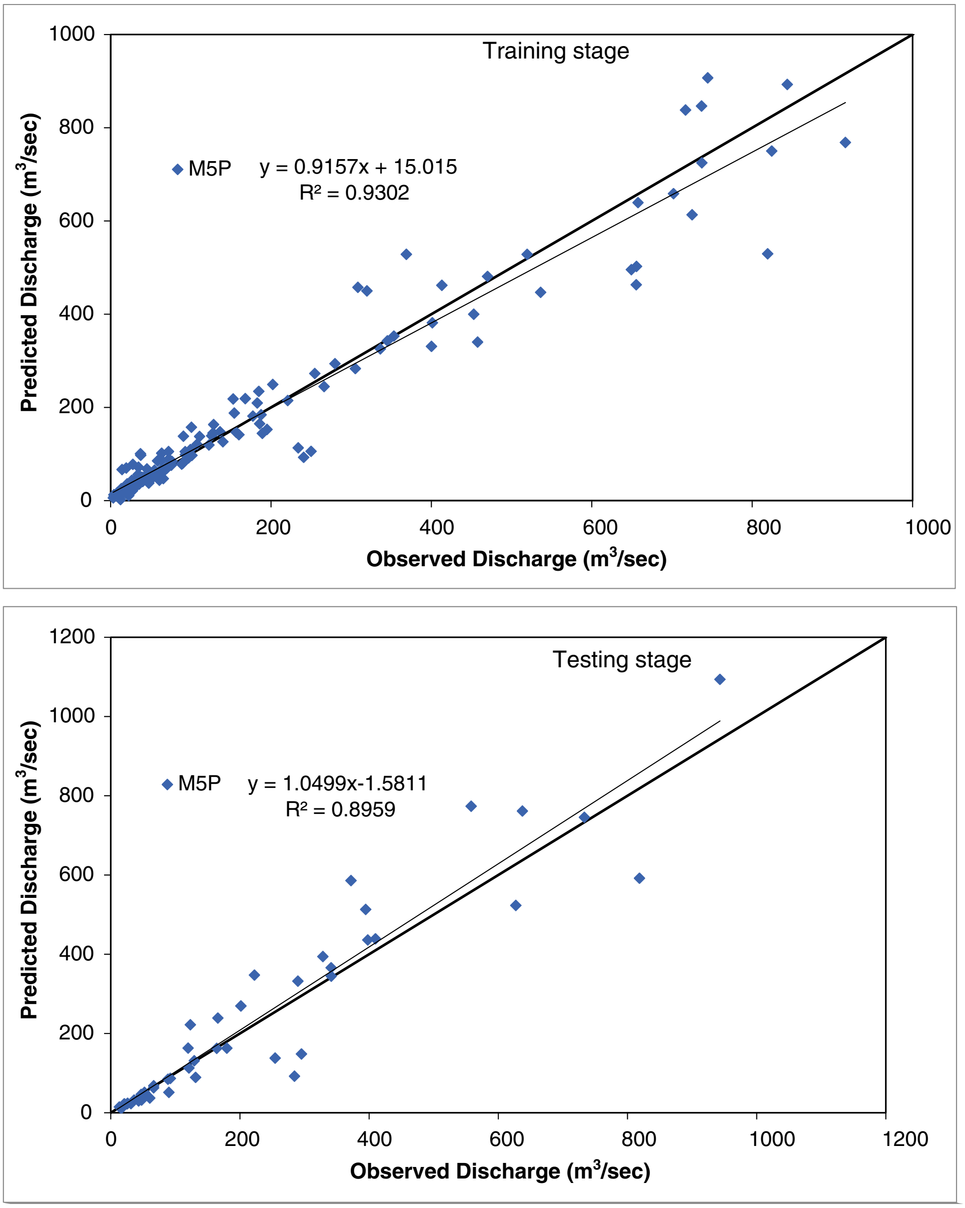


Fig. 4 The performance of the M5P model for training and testing stages

REP tree uses the logistic regression algorithm as well as generates numerous trees in several calculation procedures by which the simplest tree was taken out of all the produced trees. REP tree has been capable of generating a flexible and straightforward modelling procedure by observing training data sets whenever the outcome will be huge and the complication of the tree's internal structure is reduced. During this method, the pruning algorithm has taken under consideration the backward overfitting complexity and attempts to urge the minimum version of the best precision tree logic using the post-pruning algorithm (Quinlan 1987; Chen et al. 2019). It selects values for numeric attributes only once (Kalmegh, 2015).

Gaussian process regression (GP)

The vector method (Gaussian process) is an artificial machine learning approach that allows computing systems to adapt and strengthen their skills. GP regression is depending on the premise that adjoining observations must share data with each other, and it is a strategy that refers directly above the spatial domain. Furthermore, Gaussian process also involves the generalization of the Gaussian kernel. The Kernel-based regression vector and Gaussian distribution matrix are presented in the form of mean and covariance. Based on the probability theorem, GP regression models are capable of making predictions about unknown input data, while at the same time, they can also provide predictive accuracy which

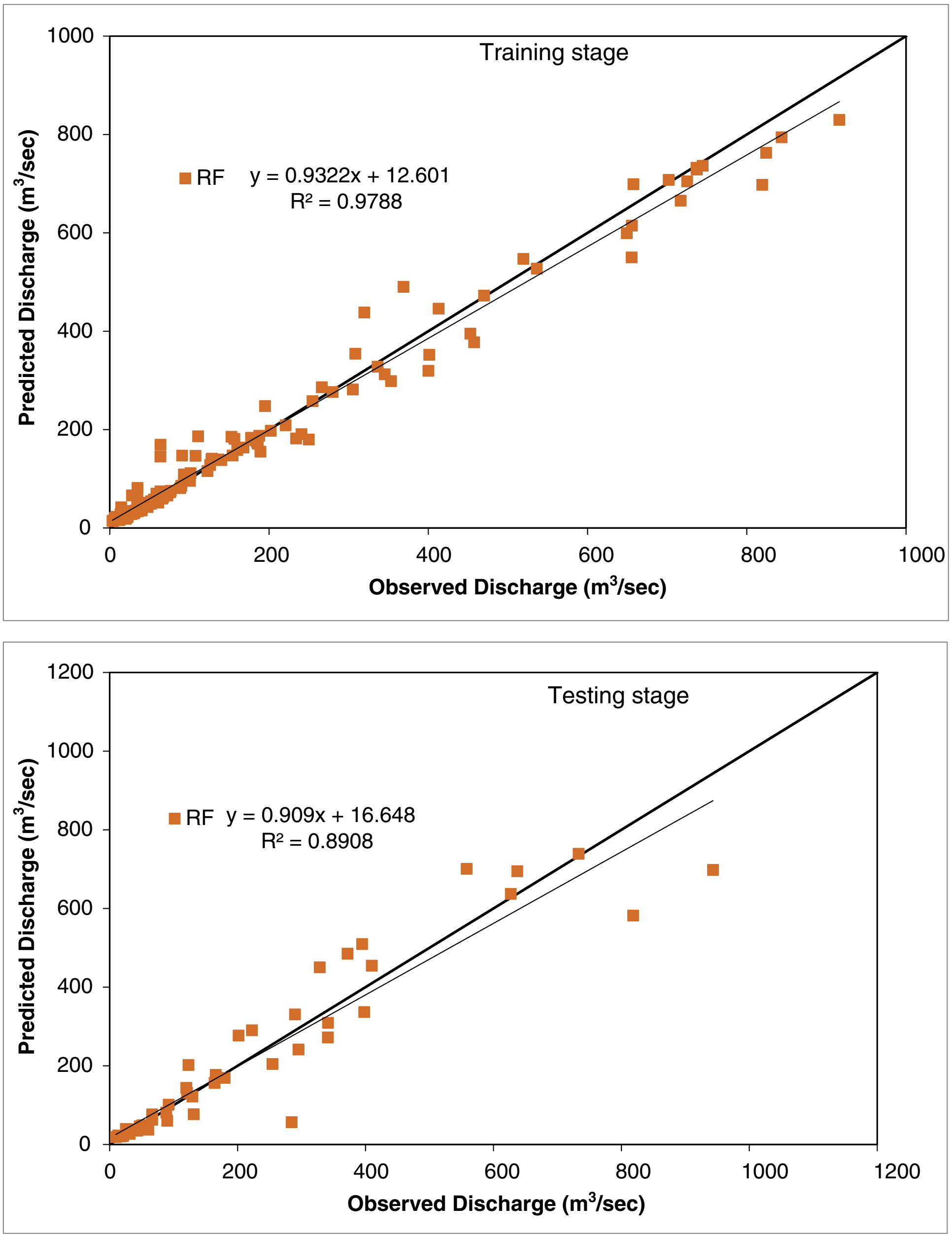


Fig. 5 The performance of the RF model for training and testing stages

greatly increases the statistically significant results of the predictive model. A Gaussian process is inclusion of multiple random variables that extend it to infinite dimensionality; hence, the processes are based on multivariate Gaussian distributions. Since the emergence of this technique in the last few years, it has been broadly used in varied research areas of chemistry, medicine, construction, etc.

Support vector machines (SVM)

Cortes and Vapnik (1995) were the first ones to propose SVM which uses a classification and regression approach which is based on the theory of statistical learning. The SVM concept is based on the ideal segment of courses. From divisible courses, SVM prefers one with the least error of generalization with an unlimited number of linear classifiers or sets a higher rate of return on the errors acquired from systemic risk assessment. The highest range between both the two classes could indeed be derived from the specified hyper lane, and the total of the hyper lane intervals from the nearest point of the two courses may set the highest range between the two classes. Hyper lanes are explained as a set of points whose dot product with vectors in that space is constant. The basic

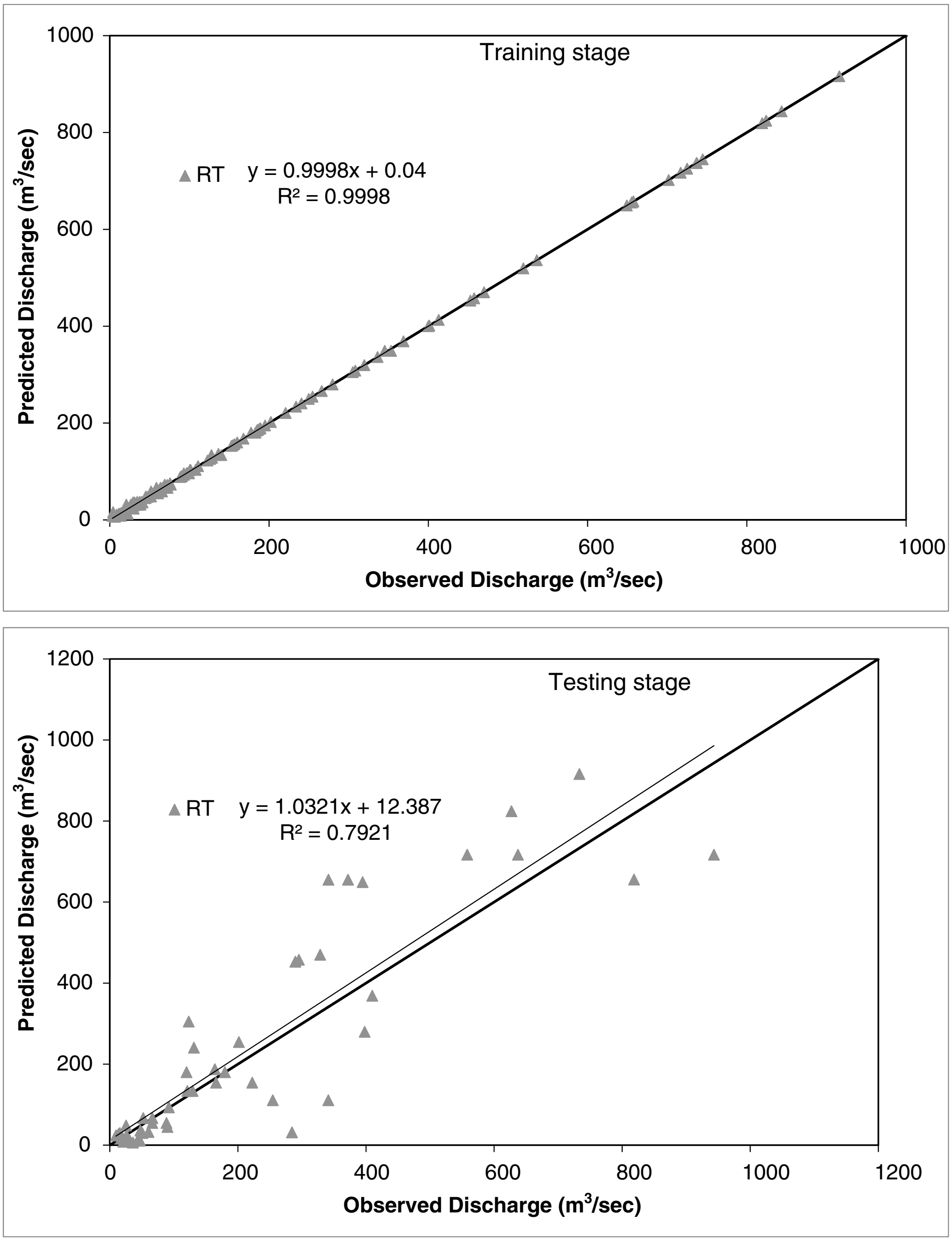


Fig. 6 The performance of the RT model for training and testing stages

approach of SVM is therefore to assemble sets of data from the interface region to the inexhaustible function region by designing a series of hyper-planes so that regression, categorization and other difficulties can be made easier in the function region. The vector machine method system provides the kernel function scheme.

Data set

The total data set containing 180 observations from the Kesinga basin was divided randomly into two categories of training and testing. Training data is the larger group which contains 70% of the total data, while testing data is the smaller group which contains the rest 30% of the total data. Different input variables used in this study are rainfall, evapotranspiration, inflow, percolation, groundwater, surface runoff, water yield and potential ET and the output parameter is discharge (*Q*) of the river. The features of both training and testing data sets are listed in Table 1. The complete flow diagram of the methodology is shown in Fig. 3. In this figure all steps explained clearly from data collection to best model selection for the prediction of river discharge.

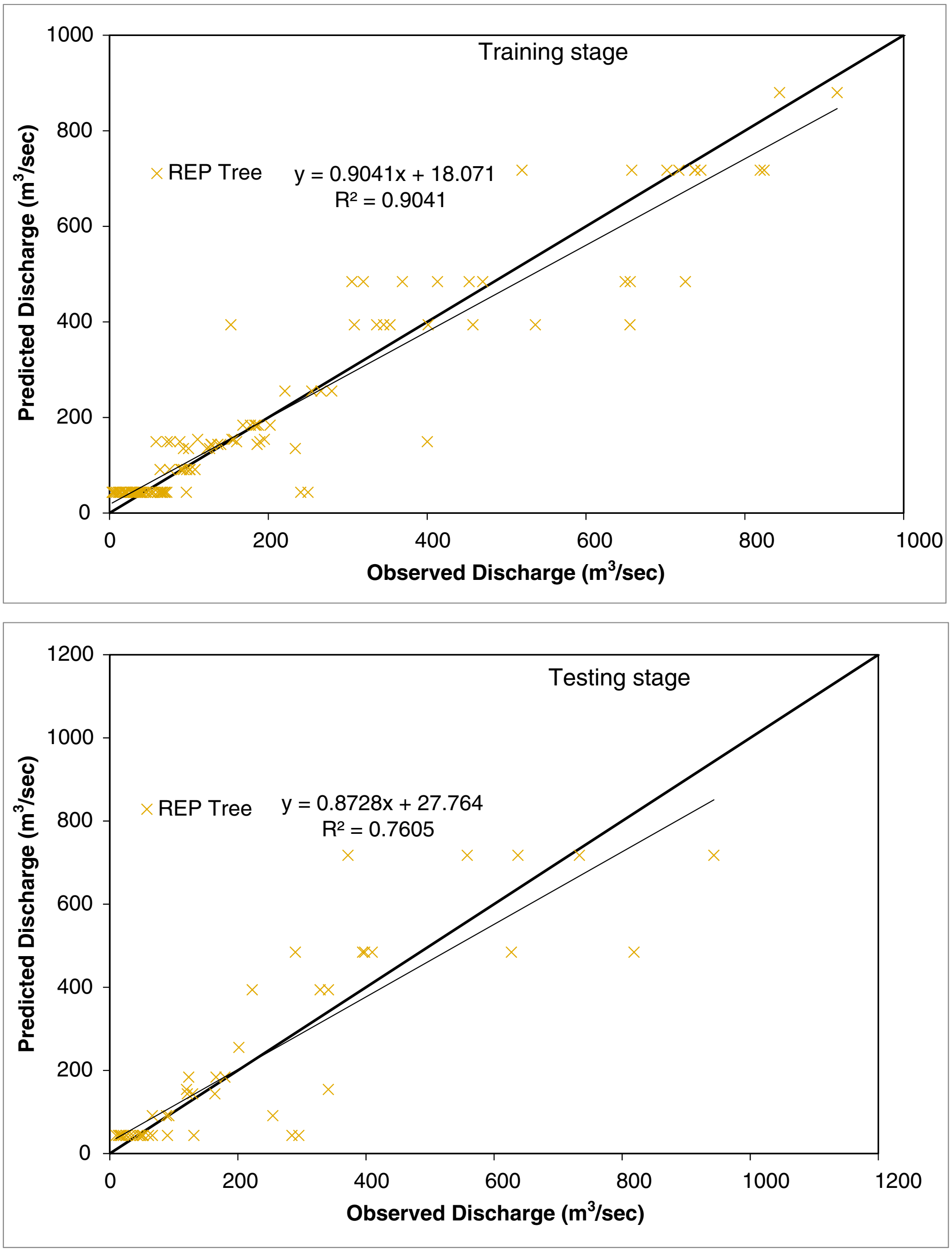


Fig. 7 The performance of the REP tree model for training and testing stages

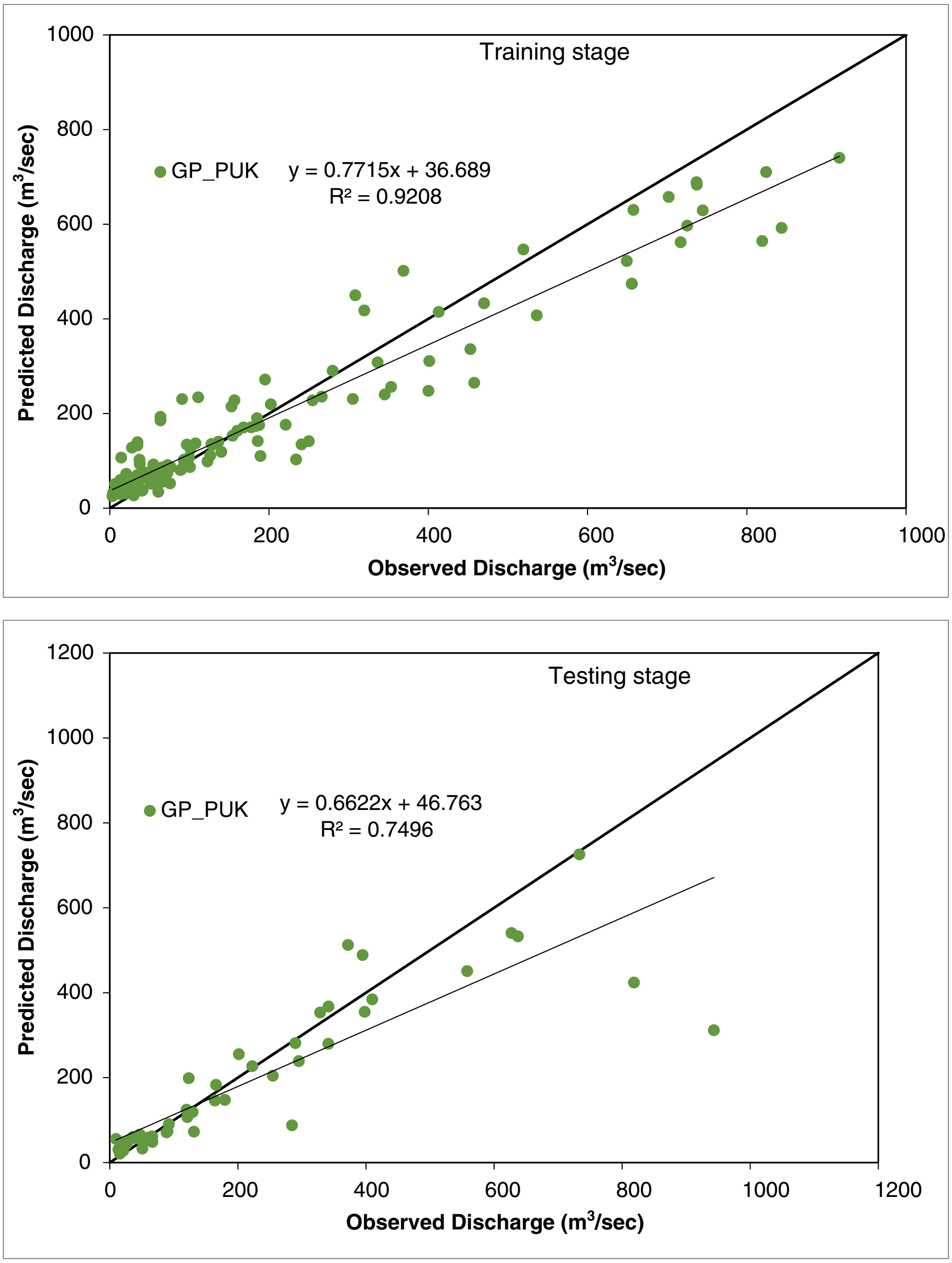
Results and discussions

The effectiveness of the soft computing techniques in predicting the discharge of the river in Kesinga basin is tested by several soft computing techniques, viz. random forest (RF), M5P, radial basis Kernel function–based Gaussian process (GP_RBF), Pearson VII kernel function–based Gaussian process (GP_PUK), reduced error pruned (REP tree), radial basis kernel function–based support vector machine (SVM_RBF), Pearson VII kernel function–based support vector machine (SVM_PUK) and random tree (RT).

Result of M5P, RF, REP Tree, RT

The performance of four models to predict Kesinga basin discharge for both training and testing stages using various performance assessment indices is shown in Table 2. The preparation of M5P, RF, REP tree and RT models is a trial-and-error process. The numbers of manual trials were done to find the maximum value of user-defined variables of M5P, RF, REP tree and RT. Scatter plot among observed and predicted discharge for training and testing stages using M5P, RF, RT and REP Tree based models are shown in Figs. 4, 5, 6

Fig. 8 The performance of the GP_PUK model for training and testing stages

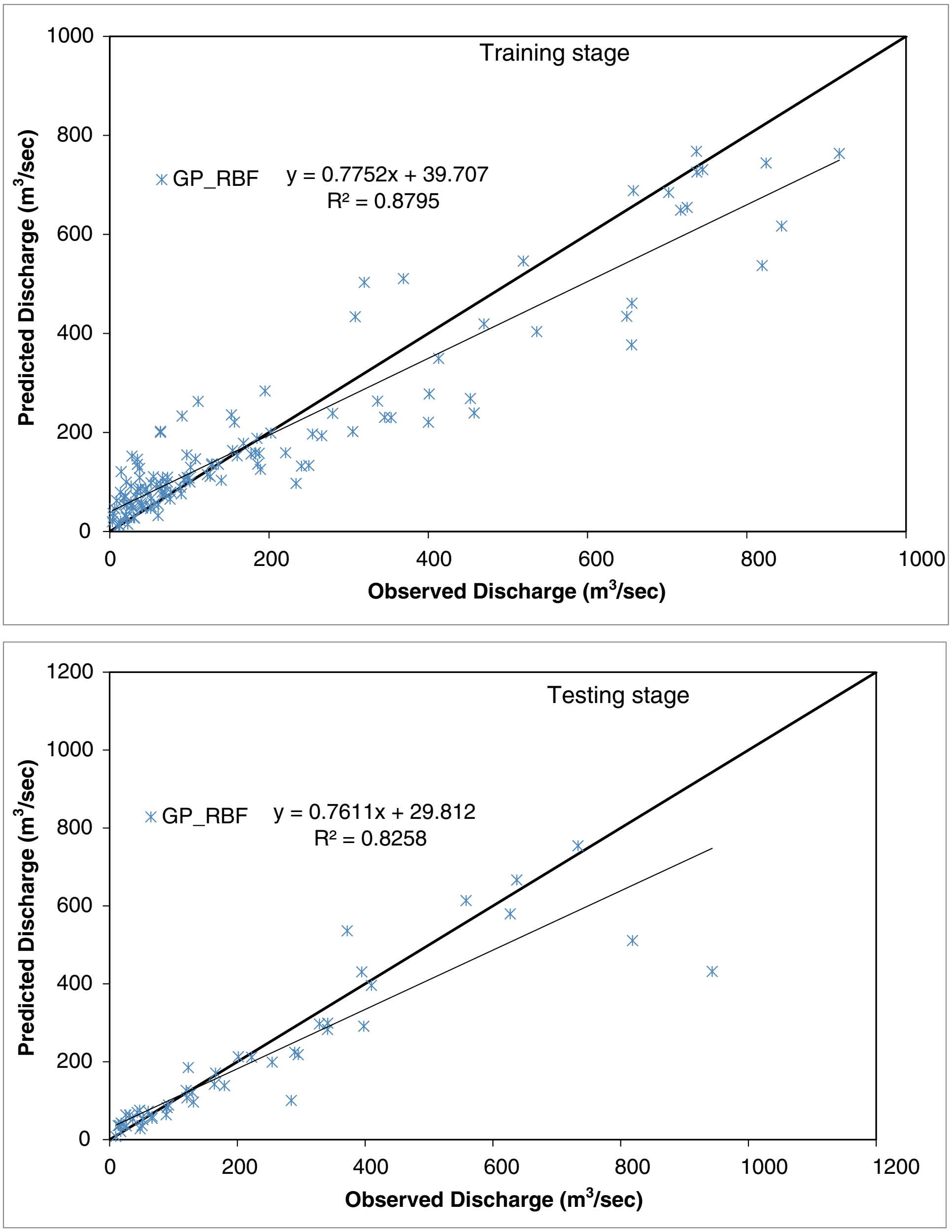


and 7, respectively. These figures indicates that the performance of M5P model is better than RF, RT and REP Tree based models with $R^2 = 0.8959$ for testing stage. The performance of M5P, RF, RT, REP Tree are listed in Table 2. For the ideal model, maximum value of WI, NSE and LMI and minimum value of CC, MAE, RMSE and NRMSE were considered. Out of these four models, M5P and RF show comparable results. The result of CC in M5P model (0.9465) is performing better than random forest (0.9438) but in other cases of MAE (44.4497), RMSE (75.5598), WI (0.9748), NSE (0.8902), LMI (0.7525) and RMSE (0.3453). Random forest shows best results by having minimum value in MAE, RMSE and NRMSE and maximum value in WI, NSE and LMI.

Result of GP_RBF, GP_PUK

To predict Kesinga basin discharge, the performance of the Gaussian process (GP_RBF and GP_PUK) for both

Fig. 9 The performance of the GP_RBF model for training and testing stages



training and testing stages using performance assessment indices is shown in Table 2. The preparation of GP_RBF and GP_PUK models is a trial-and-error process. The Scatter plot among observed and predicted discharge using GP_PUK and GP_RBF are shown in Figs. 8 and 9 respectively. These figures indicates that the performance of GP_RBF is better than GP_PUK based model with R^2 = 0.8258. For the ideal model, maximum value of WI, NSE and LMI and the lower value of CC, MAE, RMSE and NRMSE were considered. Although these models show considerable results on the bases of CC value, still GP_RBF is the best model based on the model assessment pattern. Result of CC in GP_RBF model (0.9087) is performing better than GP_PUK (0.8658) but in other cases of MAE (49.3153), RMSE (99.1282), WI (0.9567), NSE (0.8110), LMI (0.7254) and RMSE (0.4530). GP_PUK performs best results by having minimum value in MAE, RMSE and NRMSE and maximum value in WI, NSE and LMI.

Result of SVM_RBF and SVM_PUK

Presentation of support vector machine (SVM_RBF and SVM_PUK) to predict Kesinga basin discharge for both training and testing stages using performance assessment indices is depicted in Table 2. The preparation of SVM_RBF

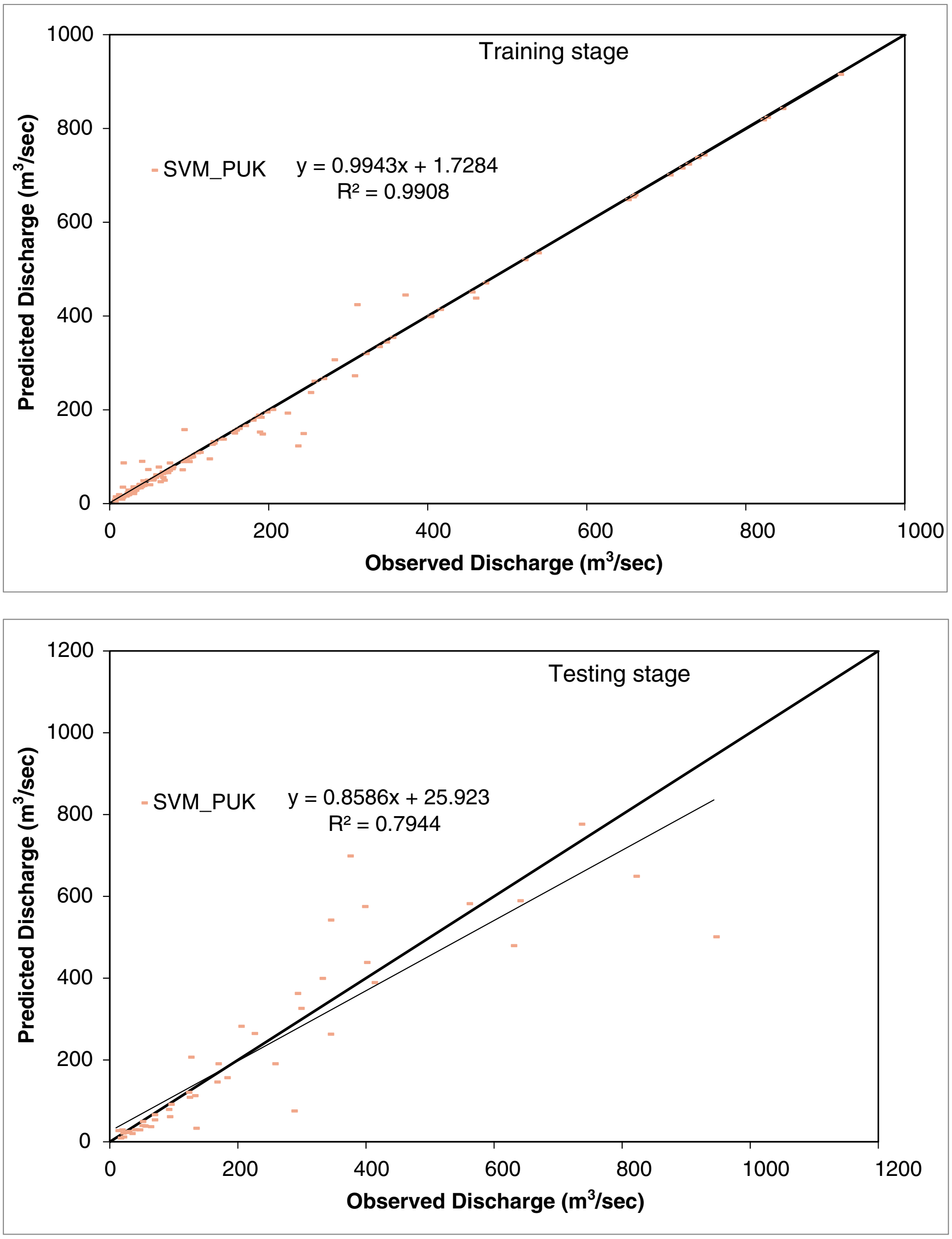


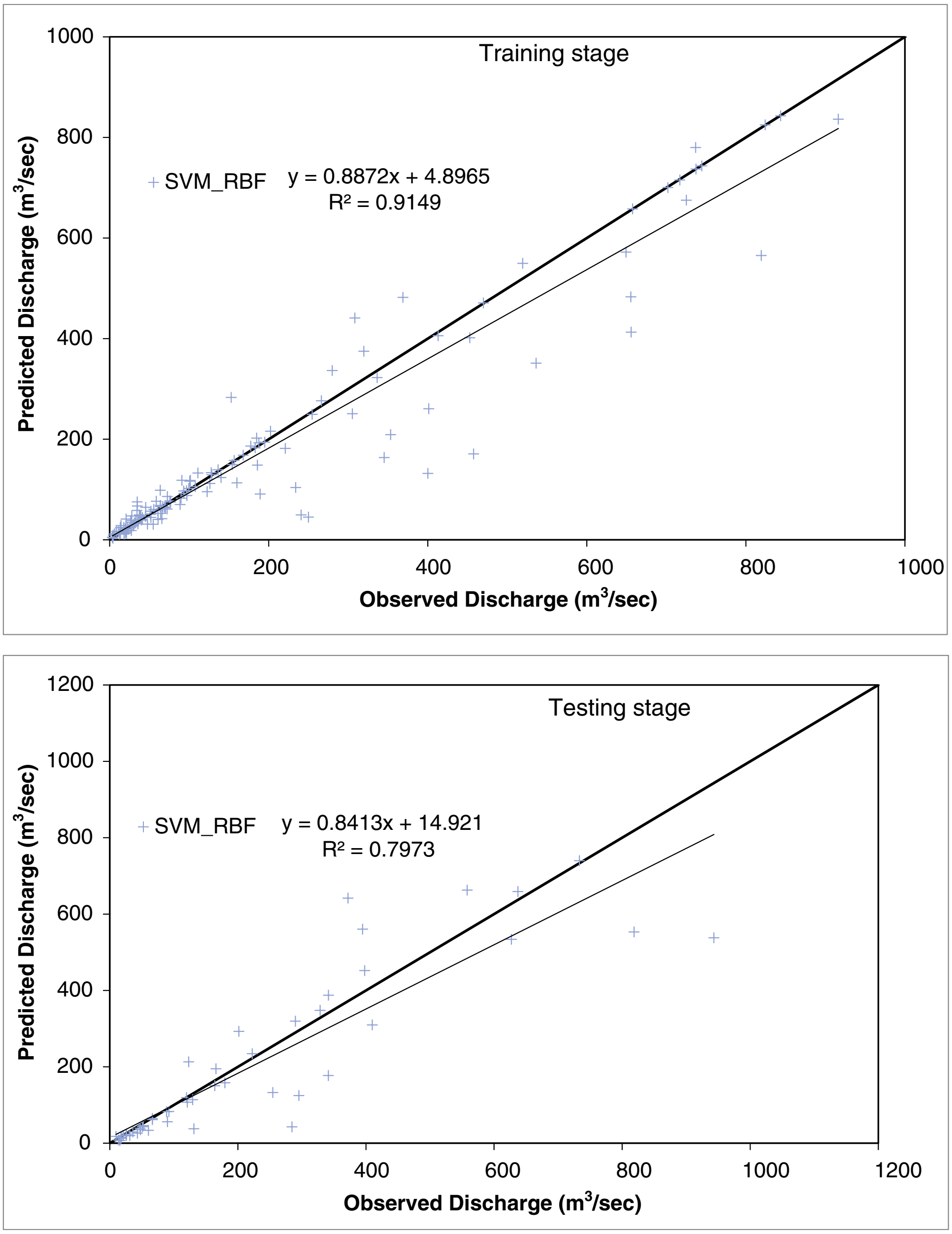
Fig. 10 The performance of the SVM_PUK model for training and testing stages

and SVM_PUK models is a trial-and-error process. Several manual trials were done to discover the maximum value of user-defined parameters of SVM_RBF and SVM_PUK. The Scatter plot among observed and predicted discharge using SVM_PUK and SVM_RBF are shown in Figs. 10 and 11 respectively. These figures indicates that the performance of SVM_RBF is better than SVM_PUK based model with $R^2 = 0.7973$ for testing stage. For the ideal model, maximum value of WI, NSE, CC and LMI and minimum value of MAE, RMSE and NRMSE were considered. SVM_PUK performance assessment indices shows CC (0.8913), MAE (57.5504), RMSE (104.8255), WI (0.9516), NSE (0.7887), LMI (0.6795) and RMSE (0.4790). Based on these outcomes, SVM_PUK can be concluded as the best model.

Comparative results of M5P, RF, RT, REP Tree, GP_RBF, GP_PUK, SVM_RBF and SVM_PUK

Figure 12 shows the comparison of the models used in the present study for the prediction of Kesinga basin discharge. Random forest is outperforming among all applied models. Based on performance assessment indices, the output of CC in M5P model (0.9465) is performing better than random forest (0.9438), but

Fig. 11 The performance of the SVM_RBF model for training and testing stages



is best in terms of MAE (44.4497), RMSE (75.5598), WI (0.9748), NSE (0.8902), LMI (0.7525) and RMSE (0.3453). Descriptive statistics of observed and predicted values using various oft computing techniques are listed in Table 3. Figure 13 indicates the box plot for observed and predicted values of discharge using soft computing techniques. Taylor diagram for the assessment of the soft computing based models for the prediction of discharge is shown in Figure 14. This figure indicates that M5P model works better than other applied models. Red solid circle is closer to hollow black circle with maximum value of CC. Overall the performance of RF model is also suitable for the prediction of discharge.

Comparison of results with multilinear regression (MLR)

Finally, a comparison of results is done with multilinear regression which is a simplified method of soft computing. MLR is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. MLR equation generated from the current data set is presented in Eq. 8. Figure 15 shows a comparison of the results with the MLR which suggested that the advanced soft computing techniques gave the good results than MLR. The value of CC for MLR is 0.7781 which is less than all other soft computing techniques.

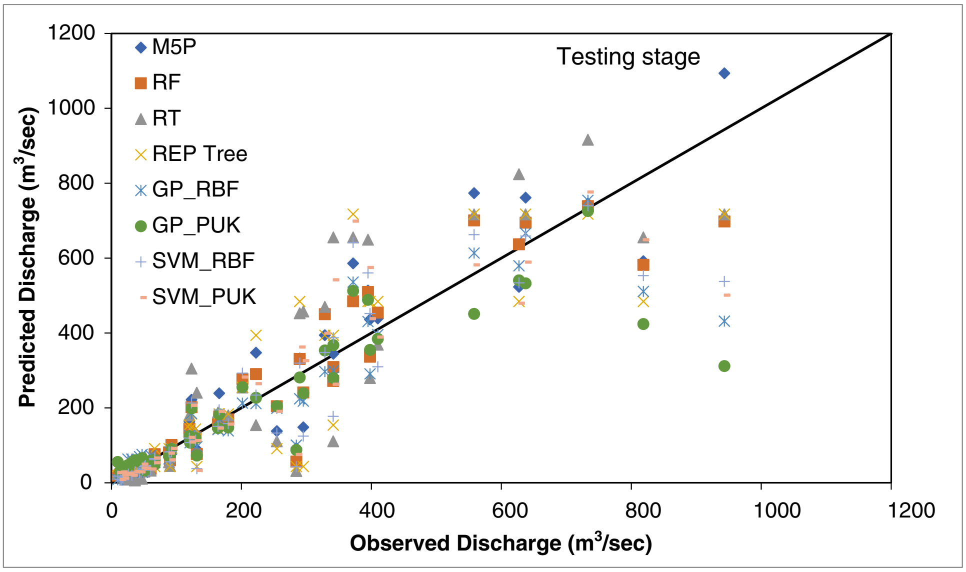


Fig. 12 The performance of the comparison of soft computing model

Table 3 Descriptive statistics (quantitative data)

Statistic	Observed	M5P	RF	RT	REP Tree	GP_RBF	GP_PUK	SVM_RBF	SVM_PUK
Minimum	9.6640	−3.8590	18.9230	6.4070	43.4690	8.1520	20.9030	6.9240	9.1040
Maximum	943.2000	1093.6890	739.0220	916.2000	717.5560	754.3330	725.8310	739.8630	776.3950
1st quartile	47.5900	35.8490	40.2030	31.1555	43.4690	58.4475	55.9833	34.9188	35.8120
Median	126.2500	122.1555	126.6605	122.3750	91.1170	114.3355	113.1045	110.7065	110.6845
3rd quartile	331.6000	352.0595	314.5735	389.8000	394.0220	284.8580	288.9635	312.2945	369.1775
Mean	218.8207	228.1570	215.5630	238.2374	218.7401	196.3603	191.6761	199.0218	213.7949
IQR	284.0100	316.2105	274.3705	358.6445	350.5530	226.4105	232.9803	277.3758	333.3655

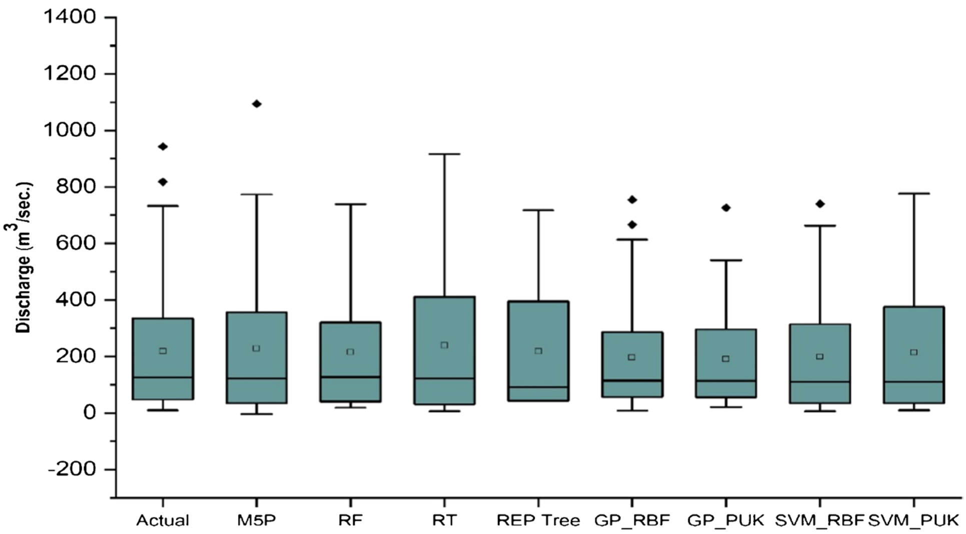


Fig. 13 Box plot for actual and predicted values using M5P, RF, RT, REP tree, GP and SVM for the testing stage

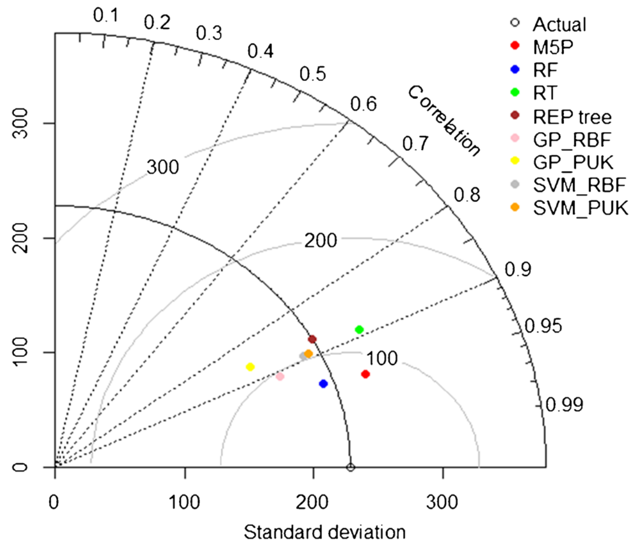


Fig. 14 Taylor diagram of various soft computing techniques for the testing stage

$$\begin{aligned} \textit{Discharge} =& 0.74\ \text{Rainfall} - 1.08\ \text{inflow} + 0.59\ \text{Evapotranspiration} \\ & - 1.03\ \text{Percolation} + 8.64\ \text{Groundwater} - 1.13\ \text{Surface Runoff} \\ & - 0.51\ \text{Potential ET} + 2.93\ \text{Water Yield Contribution} + 96.10 \end{aligned} \tag{8}$$

The present study aims at evaluating the prediction of annual water discharge of Kesinga sub-catchment of Mahanadi basin, India. In this research, different soft computing models, M5P, random forest, regression tree, reduced error pruning, Gaussian process (GP_RBF, GP_PUK) and support vector machine (SVM_RBF, SVM_PUK), are used. Based on performance assessment indices, random forest performs the best among all other models.

Ghorbani et al. (2016) performed a study in which four modelling techniques were reported to provide evidence for an appropriate method for forecasting discharge data. Different soft computing models, viz. support vector machines (SVM), rating curve (RC), artificial neural networks (ANNs) and multiple linear regression (MLR), were used. This research reveals that the ANN, SVM and RC models display a clear edge over the MLR model in forecasting discharge values, which may be explained by their nonlinear mathematical formulations. In Ghorbani et al. (2016), SVM and ANN show comparable results but perform better than the rating curve and multiple linear regression. He et al. (2014) performed the modelling technique, viz. support vector machine (SVM), artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS), on small river basins of semi-arid mountainous with complex topography by predicting the river flow performance. Support vector machine was found to be outperforming than artificial neural network and adaptive neuro-fuzzy inference system. This research has served to establish the excellent performance of the RF and the M5P techniques over the other approaches (RT, REP Tree, GP_RBF, GP_PUK, SVM_RBF and SVM_PUK). The RF and M5P is the tree-based approach that gave the edge to these approaches over other approaches. Among the RF and M5P approaches, the RF are the superior one for the prediction of the discharge. Therefore, it may be decided that RF is the ideal machine learning approach in the prediction of the Kesinga basin discharge by using different soft computing techniques.

Sensitivity study

To examine the most influential input variable, a sensitivity study was designed for the prediction of water discharge in the

Fig. 15 Comparison of results with MLR

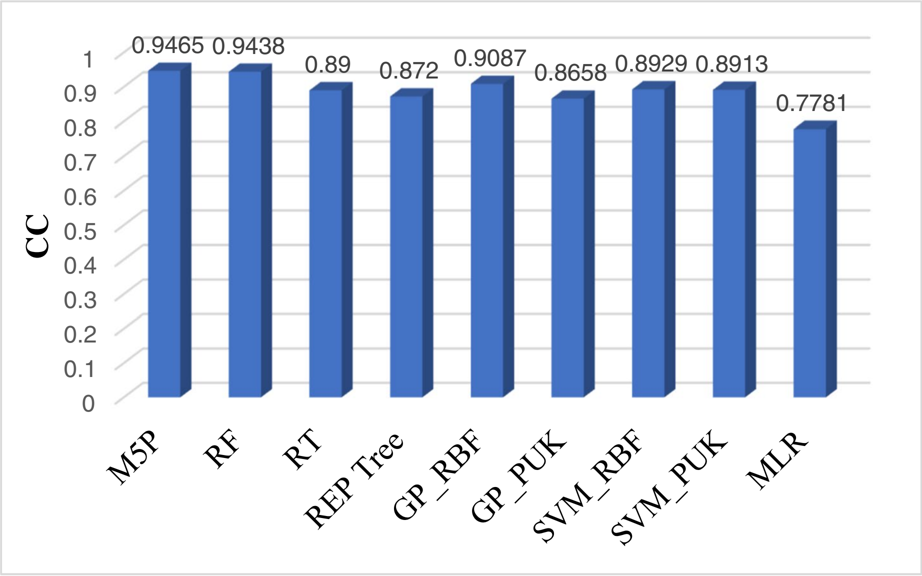


Table 4 Sensitivity study using best model (RF model)

Input combination								Target	RF model	
Rainfall(mm)	inflow (m3/sec)	Evapotranspiration (mm)	Percolation(mm)	Groundwater (mm)	Surface runoff(mm)	Potential ET(mm)	Water yield contribution (mm)	Discharge(m3/sec)	CC	RMSE
Y	Y	Y	Y	Y	Y	Y	Y	Y	0.9438	75.5599
N	Y	Y	Y	Y	Y	Y	Y	Y	0.9344	81.4147
Y	N	Y	Y	Y	Y	Y	Y	Y	0.9525	69.5505
Y	Y	N	Y	Y	Y	Y	Y	Y	0.9422	76.7077
Y	Y	Y	N	Y	Y	Y	Y	Y	0.9457	74.3876
Y	Y	Y	Y	N	Y	Y	Y	Y	0.9467	73.8692
Y	Y	Y	Y	Y	N	Y	Y	Y	0.9474	73.1401
Y	Y	Y	Y	Y	Y	N	Y	Y	0.9411	77.5447
Y	Y	Y	Y	Y	Y	Y	N	Y	0.9466	74.6483

basin. It was found that the random forest model outperformed other models selected in this research. Each input variable is removed one by one to quantify the impact of every variable on the output at a time and the outcomes were presented in the form of CC and RMSE for the test data set. From the result shown in Table 4, it has been observed that the most significant variable of Kesinga basin discharge is rainfall.

Conclusion

This study aimed at predicting the Kesinga basin discharge by using different soft computing techniques with various input variables. The primary focus of this study is comparing the performance of discharge of the Kesinga basin using M5P, RF, RT, REP Tree, GP_RBF, GP_PUK, SVM_RBF and SVM_PUK models. During testing and training, the performance of all applied models is reliable and significant for the prediction of Kesinga basin discharge data sets. RF and M5P shows comparable outcomes based on higher CC value of M5P and lower MAE and RMSE values. RF models performed better than all other models for the forecasting of discharge of the Kesinga basin with the coefficient of determination (R^2) values of 0.978 and 0.890 for the training and testing stages, respectively, using rainfall, inflow, evapotranspiration, percolation, groundwater, surface runoff, potential ET and water yield contribution. However, further expansion and exploration of the M5P, RF, RT, REP tree, GP and SVM models are required for the prediction of river discharge and sustainability of water resources management. Results of sensitivity analysis suggested that the most significant variable of Kesinga basin discharge is rainfall.

Declarations

Conflict of interest The authors declare no competing interests.

References

Agarwal P, Pal L, Alam MA (2019) Regional scale analysis of hydrometeorological variables in Kesinga sub-catchment of Mahanadi Basin. India Environmental Earth Sciences 78(15):1–25. https://doi.org/10.1007/s12665-019-8457-z

Aldous D (1993) The continuum random tree III. The Annals of Probability, 248–289. www.jstor.org/stable/2244761

Asokan SM, Dutta D (2008) Analysis of water resources in the Mahanadi River Basin, India under projected climate conditions. Hydrological Processes: an International Journal 22(18):3589–3603. https://doi.org/10.1002/hyp.6962

Bhoria S, Sihag P, Singh B, Ebtehaj I, Bonakdari H (2021) Evaluating Parshall flume aeration with experimental observations and advance soft computing techniques. Neural Comput Appl 33(24):17257–17271. https://doi.org/10.1007/s00521-021-06316-9

Brakenridge GR, Cohen S, Kettner AJ, De Groeve T, Nghiem SV, Syvitski JP, Fekete BM (2012) Calibration of satellite measurements of river discharge using a global hydrology model. J Hydrol 475:123–136. https://doi.org/10.1016/j.jhydrol.2012.09.035

Breiman L (1996) Bagging predictors. Machine Learning 24(2):123–140. https://doi.org/10.1007/BF00058655

Breiman L (1999) Random forests - random features. Technical Report 567. Statistics Department, University of California, Berkeley.

Calmant S, Seyler F (2006) Continental surface waters from satellite altimetry. CR Geosci 338(14–15):1113–1122. https://doi.org/10.1016/j.crte.2006.05.012

Chien H, Yeh PJF, Knouft JH (2013) Modeling the potential impacts of climate changeon streamflow in agricultural watersheds of the Midwestern United States. J Hydrol 491:73–88. https://doi.org/10.1016/j.jhydrol.2013.03.026

Chen W, Pradhan B, Li S, Shahabi H, Rizeei HM, Hou E, Wang S (2019) Novel hybrid integration approach of bagging-based fisher's linear discriminant function for groundwater potential analysis. Nat Resour Res 28(4):1239–1258. https://doi.org/10.1007/s11053-019-09465-w

Cortes C, Vapnik V (1995) Support-Vector Networks. Machine Learning 20(3):273–297. https://doi.org/10.1007/BF00994018

Destouni G, Jaramillo F, Prieto C (2013) Hydroclimatic shifts driven by human water use for food and energy production. Nat Clim Chang 3(3):213–217. https://doi.org/10.1038/nclimate1719

Garg V, Sambare RS, Thakur PK, Dhote PR, Nikam BR, Aggarwal SP (2022) Improving stream flow estimation by incorporating time delay approach in soft computing models. ISH J Hydraul Eng 28(sup1):57–68. https://doi.org/10.1080/09715010.2019.1676171

Ghorbani MA, Khatibi R, Goel A, FazeliFard MH, Azani A (2016) Modeling river discharge time series using support vector machine and artificial neural networks. Environ Earth Sci 75(8):685. https://doi.org/10.1007/s12665-016-5435-6

Ghosh S, Raje D, Mujumdar PP (2010) Mahanadi streamflow: climate change impact assessment and adaptive strategies. Curr Sci 1084–1091. https://www.jstor.org/stable/24111765

Gosain AK, Rao S, Basuray D (2006) Climate change impact assessment on hydrology of Indian river basins. Curr Sci 346–353. https://www.jstor.org/stable/24091868

He Z, Wen X, Liu H, Du J (2014) A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. J Hydrol 509:379–386. https://doi.org/10.1016/j.jhydrol.2013.11.054

Kalmegh SR (2015) Comparative analysis of weka data mining algorithm random forest, random tree and lad tree for classification of indigenous news data. Int J Emerg Technol Adv Eng 5(1):507–517. https://ijiset.com/vol2/v2s2/IJISET_V2_I2_63.pdf

Li G, Tang Z, Yue S, Zhuang K, Wei H (2001) Sedimentation in the shear front off the Yellow River mouth. Cont Shelf Res 21(6–7):607–625. https://doi.org/10.1016/S0278-4343(00)00097-2

Mersel MK, Smith LC, Andreadis KM, Durand MT (2013) Estimation of river depth from remotely sensed hydraulic relationships. Water Resour Res 49(6):3165–3179. https://doi.org/10.1002/wrcr.20176

Muhammad Adnan R, Yuan X, Kisi O, Yuan Y, Tayyab M, Lei X (2019) Application of soft computing models in streamflow forecasting. In Proceedings of the institution of civil engineers-water management (Vol. 172, No. 3, pp. 123–134). Thomas Telford Ltd. https://doi.org/10.1680/jwama.16.00075

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—A discussion of principles. J Hydrol 10(3):282–290. https://doi.org/10.1016/0022-1694(70)90255-6

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. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001463

Quinlan JR (1987) Simplifying decision trees. Int J Man Mach Stud 27(3):221–234. https://doi.org/10.1016/S0020-7373(87)80053-6

Quinlan JR (1992) Learning with continuous classes. In 5th Australian joint conference on artificial intelligence (Vol. 92, pp. 343–348). https://doi.org/10.1142/9789814536271

Raje D, Mujumdar PP (2009) A conditional random field–based downscaling method for assessment of climate change impact on multisite daily precipitation in the Mahanadi basin. Water Resour Res 45(10). https://doi.org/10.1029/2008WR007487

Rao PG (1993) Climatic changes and trends over a major river basin in India. Climate Res 2:215–223

Rao PG (1995) Effect of climate change on streamflows in the Mahanadi river basin. India Water International 20(4):205–212. https://doi.org/10.1080/02508069508686477

Sepahvand A, Singh B, Ghobadi M, Sihag P (2021) Estimation of infiltration rate using data-driven models. Arab J Geosci 14(1):1–11. https://doi.org/10.1007/s12517-020-06245-2

Sihag P, Angelaki A, Chaplot B (2020) Estimation of the recharging rate of groundwater using random forest technique. Appl Water Sci 10(7):1–11. https://doi.org/10.1007/s13201-020-01267-3

Singh A, Singh B, Sihag P (2021) Experimental Investigation and Modeling of Aeration Efficiency at Labyrinth Weirs. J Soft Comput Civ Eng 5(3):15–31. https://doi.org/10.22115/SCCE.2021.284637.1311

Singh B, Ebtehaj I, Sihag P, Bonakdari H (2022) An expert system for predicting the infiltration characteristics. Water Supply 22(3):2847–2862. https://doi.org/10.2166/ws.2021.430

Singh B, Sihag P, Deswal S (2019) Modelling of the impact of water quality on the infiltration rate of the soil. Appl Water Sci 9(1):15. https://doi.org/10.1007/s13201-019-0892-1

Sridharam S, Sahoo A, Samantaray S, Ghose DK (2021) Assessment of Flow Discharge in a River Basin Through CFBPNN, LRNN and CANFIS. In Communication Software and Networks (pp. 765–773). Springer, Singapore. https://doi.org/10.1007/978-981-15-5397-4_78

Stutter M, Baggaley N, Wang C (2021) The utility of spatial data to delineate river riparian functions and management zones: a review. Sci Total Environ 757:143982. https://doi.org/10.1016/j.scitotenv.2020.143982

Sullivan C (2002) Calculating a water poverty index. World Dev 30(7):1195–1210. https://doi.org/10.1016/S0305-750X(02)00035-9

Vörösmarty CJ, Fekete BM, Meybeck M, Lammers RB (2000) Global system of rivers: Its role in organizing continental land mass and defining land-to-ocean linkages. Global Biogeochem Cycles 14(2):599–621. https://doi.org/10.1029/1999GB900092

Willmott CJ (1981) On the validation of models. Phys Geogr 2(2):184–194. https://doi.org/10.1080/02723646.1981.10642213

Zakharova E, Nielsen K, Kamenev G, Kouraev A (2020) River discharge estimation from radar altimetry: assessment of satellite performance, river scales and methods. J Hydrol 583:124561. https://doi.org/10.1016/j.jhydrol.2020.124561