Environmental Science

http://www.aimspress.com/journal/environmental

AIMS Environmental Science, 11(4): 593–609.
DOI: 10.3934/environsci.2024029
Received: 07 June 2024
Revised: 13 July 2024
Accepted: 17 July 2024
Published: 26 July 2024

Research article

Analysis of data splitting on streamflow prediction using random forest

Diksha Puri[1], Parveen Sihag[2], Mohindra Singh Thakur[3], Mohammed Jameel[4], Aaron Anil Chadee[5] and Mohammad Azamathulla Hazi [5,*]

[1] School of Environmental Science, Shoolini University, Solan, Himachal Pradesh, 173229, India (dikshapuri10@gmail.com)

[2] Department of Civil Engineering, Chandigarh University, Punjab 140301, India (parveen12sihag@gmail.com)

[3] Department of Civil Engineering, Shoolini University, Solan, Himachal Pradesh, 173229, India (mohindrasinghthakur@shooliniuniversity.com)

[4] Department of Civil Engineering, King Khalid University, Abha, Saudi Arabia (jamoali@kku.edu.sa)

[5] Department of Civil and Environmental Engineering, University of the West Indies, Trinidad.

* **Correspondence:** Email: hazi.azamathulla@sta.uwi.edu.

Abstract: This study is focused on the use of random forest (RF) to forecast the streamflow in the Kesinga River basin. A total of 169 data points were gathered monthly for the years 1991–2004 to create a model for streamflow prediction. The dataset was allotted into training and testing stages using various ratios, such as 50/50, 60/40, 70/30, and 80/20. The produced models were evaluated using three statistical indices: the root mean square error (RMSE), the mean absolute error (MAE), and the correlation coefficient (CC). The analysis of the models' performances revealed that the training and testing ratios had a substantial impact on the RF model's predictive abilities; models performed best when the ratio was 60/40. The findings demonstrated the right dataset ratios for precise streamflow prediction, which will be beneficial for hydraulic engineers during the water-related design and engineering stages of water projects.

Keywords: streamflow prediction; Kesinga basin; data separation ratios; random forest; 60/40

1. Introduction

Water is an essential natural resource that sustains life, ecosystems, and society ([1]). Due to the overexploitation of this natural resource, many parts of the world have been experiencing deteriorating situations such as rivers having little to no discharge [2–5], ecosystems being degraded [6], reduced water table levels [7], and decreasing lakes and wetland areas [8]. The hydrological time series is becoming increasingly essential for the effective distribution, management, and planning of water resources [9]. Furthermore, human activity and socioeconomic growth, in addition to climate change, have an impact on hydrological processes such as temperature, evaporation, and precipitation [10–13]. As a result, nonlinear and time-varying hydrological time series are a constant [14]. The intricate nonlinearity, high irregularity, and multi-scale irregularity of hydrological time series make forecasting an intimidating task. Hydrological time series prediction has been the subject of numerous studies [15], but a thorough knowledge of hydrological processes is still lacking. Particularly for complex time series, the existing forecasting methods still have low forecast accuracy. Highly precise streamflow predictions are necessary for hydrological application operations and planning on a range of time scales, including daily, weekly, and monthly. While long-term forecasting, such as weekly and monthly flow, is essential for planning hydropower generation, reservoir release scheduling, irrigation management, river sediment transport, and several others, real-time streamflow forecasting, or hourly and daily flow, is crucial for flood control and mitigation [16–18].

In the past few decades, data-driven machine-learning approaches have included support vector machines (SVMs), neural networks (NNs) [19–21], fuzzy logic [22–24], and hybrid ML techniques, which include a blend of deep learning algorithms, the nonlinear autoregressive network with exogenous inputs, multilayer perceptron, and random forest [25]. Additionally, grey wolf optimization (GWO)-integrated AI models [26] have drawn a lot of attention for streamflow forecasting applications. For instance, a study investigated and compared the effect of different dataset ratios in the prediction of groundwater using NN methods, observing that dataset size and model choice had significant effects on output [27]. Researchers working with ML algorithms have used training/testing ratios of 70/30, 80/20, and 90/10 for producing datasets [28–30]. ML-based models have been established and applied effectively and efficiently to solve a lot of real-world challenges [31–35]. The primary benefit of ML is its ability to subjectively analyze an infinite quantity of data and produce accurate results and evaluations. However, the quality of the data as well as how it is utilized determines the outcome [36]. Thus, in order to choose an appropriate data splitting for improved ML-based modeling, it is necessary to evaluate the impact of data splitting on the execution of soft computing models. This study investigates how data splitting affects the performance of random forest (RF) in predicting streamflow. The primary goal of the research was to develop hydrological models using monthly hydrological data and apply different training/testing ratios, namely 50/50, 60/40, 70/30, and 80/20. The objective was to analyze the impact of these ratios on the RF approach for forecasting river flow in the Kesinga sub-catchment of the Mahanadi basin.

2. Materials and methods

2.1. Modeling approach

2.1.1. Random forest (RF)

RF is the ML model based on the ensemble method of learning [37–39]. RF is a highly preferred ensemble learning method in ML research and practice, having been introduced by Breiman L [37], and successfully performing classification and regression tasks. The combination of decision trees is the basic idea underlying RF; it creates variety among the trees and minimizes overfitting by training each tree on an arbitrary subset of input characteristics and a different subset of the training data. To build an RF, a decision tree is created using a randomly selected subdivision of the training data by bootstrapping [40]. For each split, a selection of the input features is randomly considered, coined as the "random subspace" approach, to ensure each tree is different from the others. Adjusting the number of decision trees to be generated (Ntree) is a significant parameter to consider. Increasing the number of trees often improves the performance of RF. Since overfitting has little impact on RF, Ntree can be scaled relatively high. A subset of data can be used, known as out-of-bag (OOB) samples, for estimating the performance of the RF without needing a separate validation set. When performing RF regression, the individual trees are grown using the CART algorithm without pruning. Each tree uses different random subsets of features for splitting at every node and is grown to its maximum depth. The estimates of individual trees are combined to generate the final regression output, typically taking an average. A significant aspect of RF is the ability to determine variable importance. Variable importance is computed by evaluating how much a variable contributes to the improvement of prediction performance when it is randomly permuted. This approach can be used to evaluate the relevance of multiple features in the dataset and to select the most crucial variables for a particular task [41]. In conclusion, with their robustness, effectiveness, and interpretability, RFs are powerful and versatile ensemble learning models that have gained popularity. Figure 1 illustrates the working of RFs.

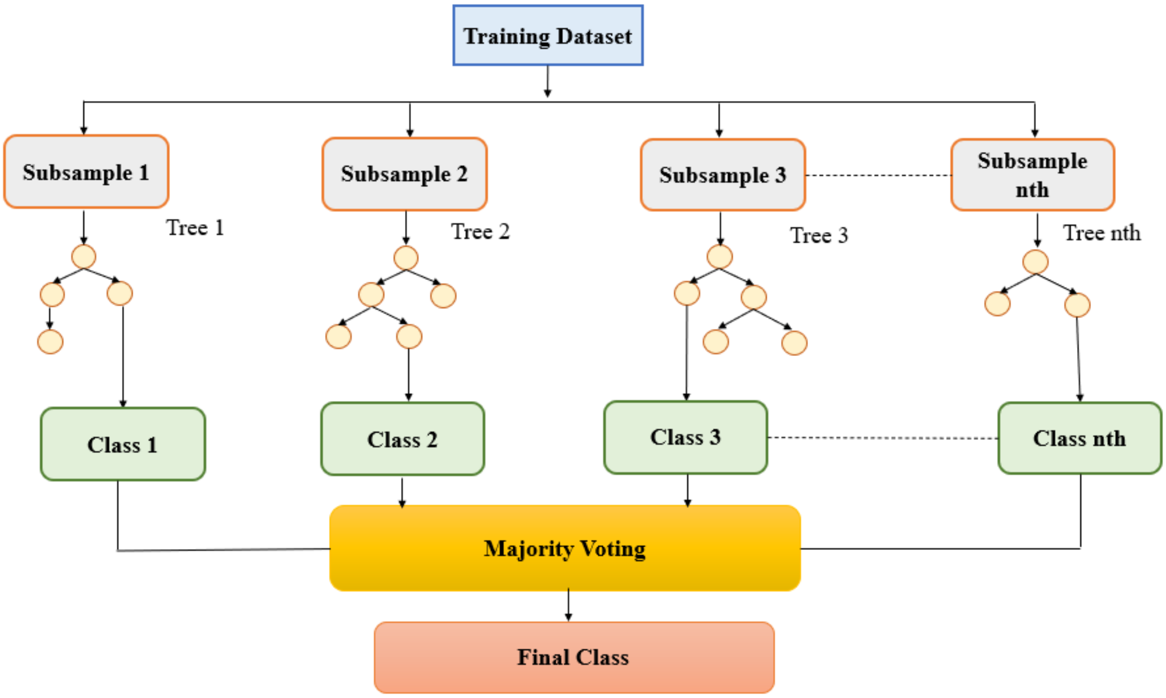
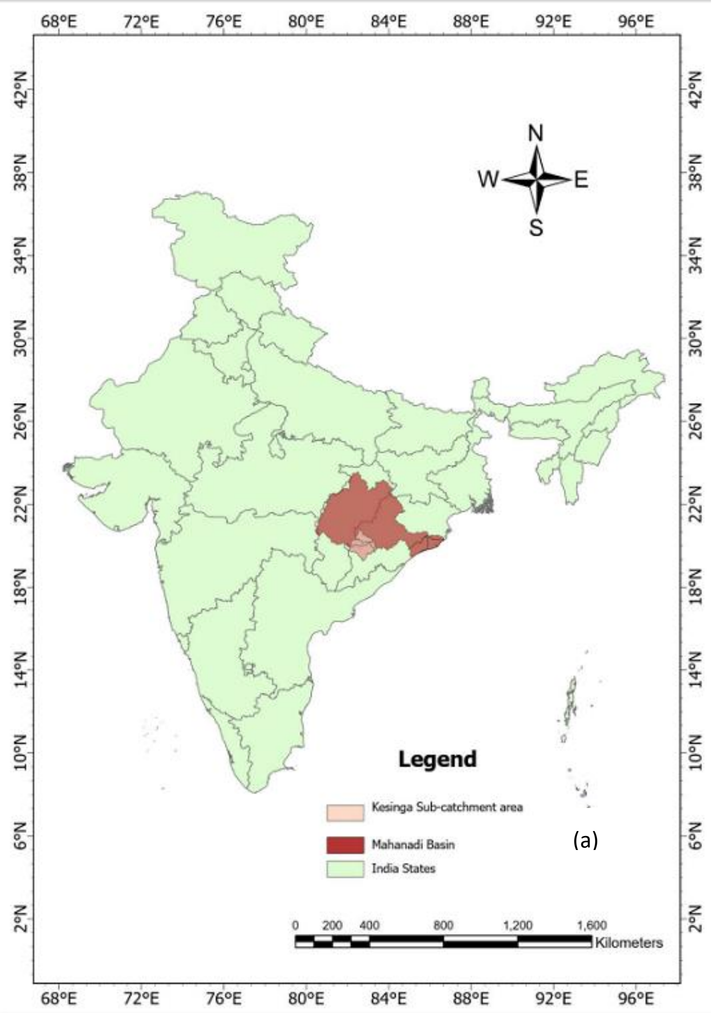


Figure 1. Schematic representation of random forest.

2.2. Methodology and dataset

To develop a model for the prediction of streamflow, a total of 169 data points were collected on a monthly basis for the years 1991–2004, from the Central Water Commission (CWC), Mahanadi and Eastern Rivers division, Bhubaneswar, Odisha. The total sample was then randomly chosen and segregated into two subsets, i.e., training and testing, with a 50/50, 60/40, 70/30, and 80/20 ratio. The models were generated using WEKA 3.9.5 software. The input characteristics were assessed to predict the aeration efficiency outcome using statistical measures like correlation coefficient (CC), mean absolute error (MAE), and root mean square error (RMSE). Figure 2 shows a schematic representation of the study site whereas, Figure 3 shows the flowchart for the current study.

In developing a model, choosing the input variable is the most important step. Samples of 12 distinct input data combinations were grouped, as shown in Table 2, in order to accomplish this. The number of variables that were chosen by looking at the monthly river flow is shown in the input M-1 to M-12. Models for streamflow prediction were developed using the same combinations of inputs. The details of the models are shown in Table 1.



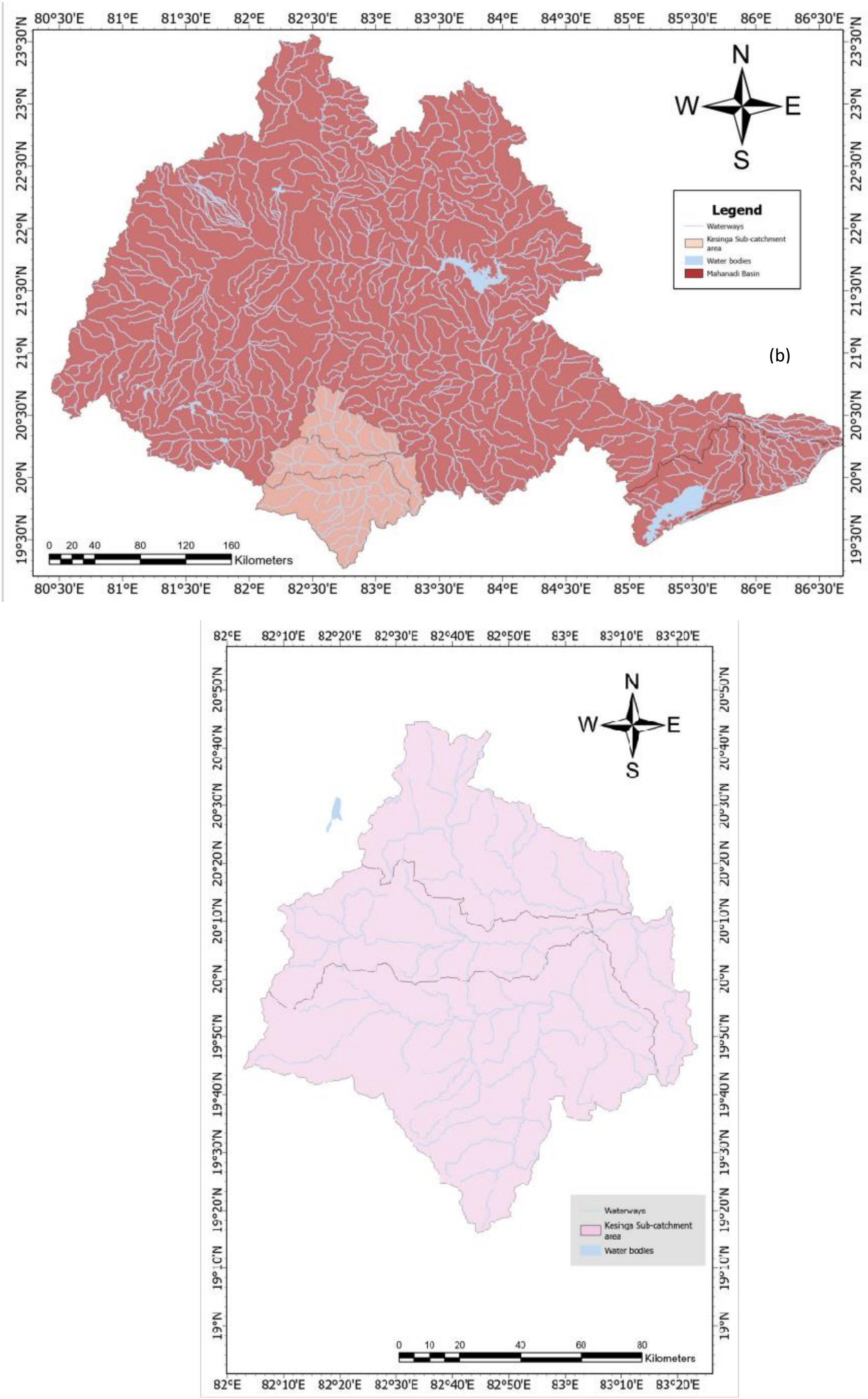

N
W
E
S
Legend
Waterways
Kesinga Sub-catchment area
Water bodies
Mahanadi Basin
(b)
0 20 40 80 120 160
Kilometers
80°30'E 81°E 81°30'E 82°E 82°30'E 83°E 83°30'E 84°E 84°30'E 85°E 85°30'E 86°E 86°30'E
23°30'N 23°N 22°30'N 22°N 21°30'N 21°N 20°30'N 20°N 19°30'N
Waterways
Kesinga Sub-catchment area
Water bodies
0 10 20 40 60 80
Kilometers
82°E 82°10'E 82°20'E 82°30'E 82°40'E 82°50'E 83°E 83°10'E 83°20'E
20°50'N 20°40'N 20°30'N 20°20'N 20°10'N 20°N 19°50'N 19°40'N 19°30'N 19°20'N 19°10'N 19°N

Figure 2. Study site: (a) Indian map showing the states; (b) Mahanadi Basin; (c) Kesinga sub-catchment.

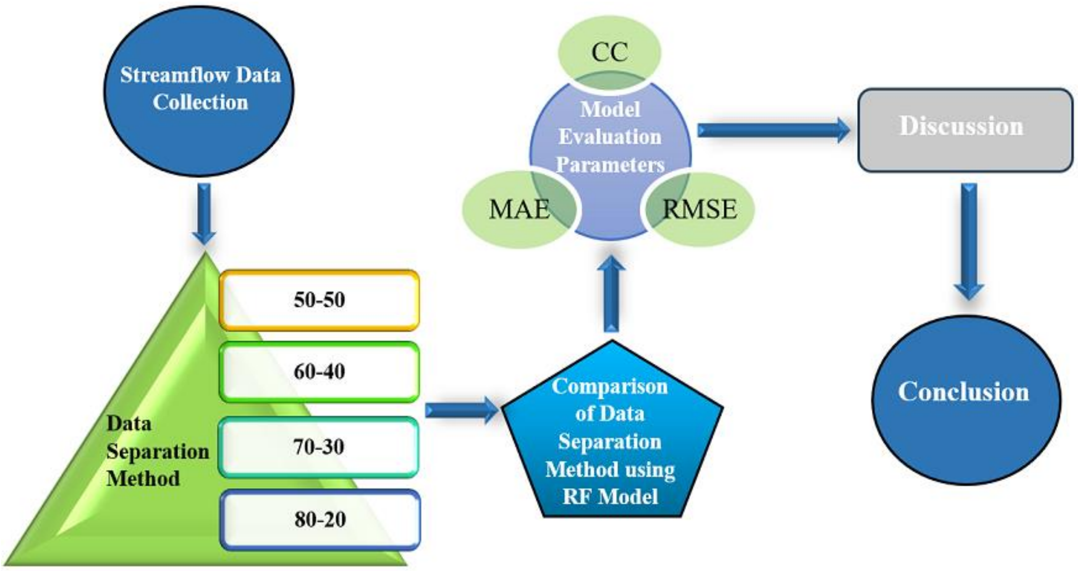

Streamflow Data Collection
Data Separation Method
50-50
60-40
70-30
80-20
Comparison of Data Separation Method using RF Model
CC
Model Evaluation Parameters
MAE
RMSE
Discussion
Conclusion

Figure 3. Flowchart of the methodology adopted in the current study.

2.3. Performance evaluation parameters

The precision of the models utilized for E_{20} at different jet geometries in an open channel was assessed using three statistical measurements: correlation coefficient (CC), mean absolute error (MAE), and root mean square error (RMSE). The calculations for CC, MAE, and RMSE are outlined in Eqs (1–3):

$$\mathrm{CC} = \frac{\sum_{i=1}^{N}(o_i - \bar{o})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{N}(o_i - \bar{o})^2 \sum_{i=1}^{N}(p_i - \bar{p})^2}} \tag{1}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|p_i - o_i| \tag{2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2} \tag{3}$$

Where o = observed values

$\underline{o}$ = average of observed values

p = predicted values

$\underline{p}$= average of predicted values

N = number of observations

Table 1. Model structure with input parameter combinations.

Model	Input combinations
M-1	$Z_x = f(a_{x\text{-}1})$
M-2	$Z_x = f(a_{x\text{-}1}, a_{x\text{-}2})$
M-3	$Z_x = f(a_{x\text{-}1}, a_{x\text{-}2}, a_{x\text{-}3})$
M-4	$Z_x = f(a_{x\text{-}1}, a_{x\text{-}2}, a_{x\text{-}3}, a_{x\text{-}4})$
M-5	$Z_x = f(a_{x\text{-}1}, a_{x\text{-}2}, a_{x\text{-}3}, a_{x\text{-}4}, a_{x\text{-}5})$
M-6	$Z_x = f(a_{x\text{-}1}, a_{x\text{-}2}, a_{x\text{-}3}, a_{x\text{-}4}, a_{x\text{-}5}, a_{x\text{-}6})$
M-7	$Z_x = f(a_{x\text{-}1}, a_{x\text{-}2}, a_{x\text{-}3}, a_{x\text{-}4}, a_{x\text{-}5}, a_{x\text{-}6}, a_{x\text{-}7})$
M-8	$Z_x = f(a_{x\text{-}1}, a_{x\text{-}2}, a_{x\text{-}3}, a_{x\text{-}4}, a_{x\text{-}5}, a_{x\text{-}6}, a_{x\text{-}7}, a_{x\text{-}8})$
M-9	$Z_x = f(a_{x\text{-}1}, a_{x\text{-}2}, a_{x\text{-}3}, a_{x\text{-}4}, a_{x\text{-}5}, a_{x\text{-}6}, a_{x\text{-}7}, a_{x\text{-}8}, a_{x\text{-}9})$
M-10	$Z_x = f(a_{x\text{-}1}, a_{x\text{-}2}, a_{x\text{-}3}, a_{x\text{-}4}, a_{x\text{-}5}, a_{x\text{-}6}, a_{x\text{-}7}, a_{x\text{-}8}, a_{x\text{-}9}, a_{x\text{-}10})$
M-11	$Z_x = f(a_{x\text{-}1}, a_{x\text{-}2}, a_{x\text{-}3}, a_{x\text{-}4}, a_{x\text{-}5}, a_{x\text{-}6}, a_{x\text{-}7}, a_{x\text{-}8}, a_{x\text{-}9}, a_{x\text{-}10}, a_{x\text{-}11})$
M-12	$Z_x = f(a_{x\text{-}1}, a_{x\text{-}2}, a_{x\text{-}3}, a_{x\text{-}4}, a_{x\text{-}5}, a_{x\text{-}6}, a_{x\text{-}7}, a_{x\text{-}8}, a_{x\text{-}9}, a_{x\text{-}10}, a_{x\text{-}11}, a_{x\text{-}12})$

3. Results

3.1. 50/50 ratio

There were two user-defined parameters in RF regression-based models. First, there are k trees to be produced, and then there are m variables to be used in order to create a tree. It was discovered that 1 was the ideal value of k to obtain the optimal model, the M-11. Table 2 indicates that the M-11 RF model outperformed the other models. The CC, MAE, and RMSE performance values for the 50/50 separation method are displayed in the table. The table shows that M-11 has the lowest errors, the maximum CC value of 0.737, MAE = 94.468, and RMSE = 137.116. Using the M-11 RF model, Figure 4a, b shows the variation between the experimental and anticipated values of streamflow for the Kesinga basin.

Table 2. Performance evaluation parameters using different data ratios.

Separation		Training			Testing		
	Models	CC	MAE	RMSE	CC	MAE	RMSE
50/50	M-1	0.956	44.433	72.055	0.731	88.207	135.771
	M-2	0.974	35.728	56.577	0.679	95.683	147.251
	M-3	0.978	31.683	52.374	0.659	99.981	152.112
	M-4	0.976	34.474	55.117	0.639	108.467	158.318
	M-5	0.973	36.592	57.664	0.663	109.138	153.918
	M-6	0.976	36.581	56.497	0.658	110.903	154.807
	M-7	0.977	35.824	56.448	0.660	112.143	155.481
	M-8	0.977	35.319	56.607	0.661	113.066	155.373
	M-9	0.973	36.727	58.628	0.684	109.428	151.055
	M-10	0.976	35.643	56.626	0.725	100.850	141.984
	M-11	0.980	32.744	51.772	0.737	94.468	137.116
	M-12	0.984	28.970	46.125	0.708	90.424	141.166
60/40	M-1	0.963	34.482	60.529	0.653	87.942	161.539

	M-2	0.972	31.465	55.33	0.736	77.978	141.503
	M-3	0.975	32.451	53.434	0.798	73.571	125.408
	M-4	0.974	33.030	53.853	0.809	70.598	122.587
	M-5	0.976	33.808	53.628	0.799	74.699	125.404
	M-6	0.976	34.342	55.438	0.804	76.240	124.710
	M-7	0.976	35.895	56.491	0.805	76.619	124.995
	M-8	0.976	34.553	55.873	0.789	79.412	128.837
	M-9	0.977	34.535	56.214	0.791	79.737	128.685
	M-10	0.977	35.143	57.584	0.767	82.673	134.468
	M-11	0.978	32.779	55.071	0.822	72.491	119.448
	M-12	0.976	30.904	53.845	0.844	69.367	111.976
70/30	M-1	0.958	38.752	66.829	0.634	94.160	168.215
	M-2	0.975	30.038	53.587	0.745	76.024	144.438
	M-3	0.978	29.235	51.378	0.748	76.440	138.842
	M-4	0.977	30.009	51.736	0.744	76.912	139.626
	M-5	0.979	29.430	50.074	0.738	77.396	141.009
	M-6	0.981	29.861	49.188	0.739	79.247	140.047
	M-7	0.982	32.660	51.282	0.732	83.627	141.799
	M-8	0.9812	31.882	51.640	0.757	76.758	136.167
	M-9	0.982	30.907	50.207	0.743	79.559	139.651
	M-10	0.983	29.610	49.000	0.711	83.688	147.698
	M-11	0.984	27.606	45.975	0.688	92.821	151.125
	M-12	0.983	26.557	45.066	0.714	84.671	145.412
80-20	M-1	0.968	31.072	56.058	0.467	106.620	196.171
	M-2	0.976	27.919	50.385	0.606	88.491	172.731
	M-3	0.978	27.755	50.202	0.681	87.882	157.734
	M-4	0.980	27.790	49.072	0.688	85.108	154.887
	M-5	0.979	28.926	50.378	0.666	85.769	159.080
	M-6	0.982	28.887	49.219	0.675	85.142	156.360
	M-7	0.980	27.790	49.072	0.688	85.108	154.887
	M-8	0.981	30.202	49.489	0.617	90.404	166.448
	M-9	0.981	29.153	48.553	0.635	92.515	162.900
	M-10	0.981	27.142	48.077	0.623	94.394	165.182
	M-11	0.981	26.395	48.111	0.679	91.854	154.940
	M-12	0.979	26.924	48.047	0.708	95.575	150.163

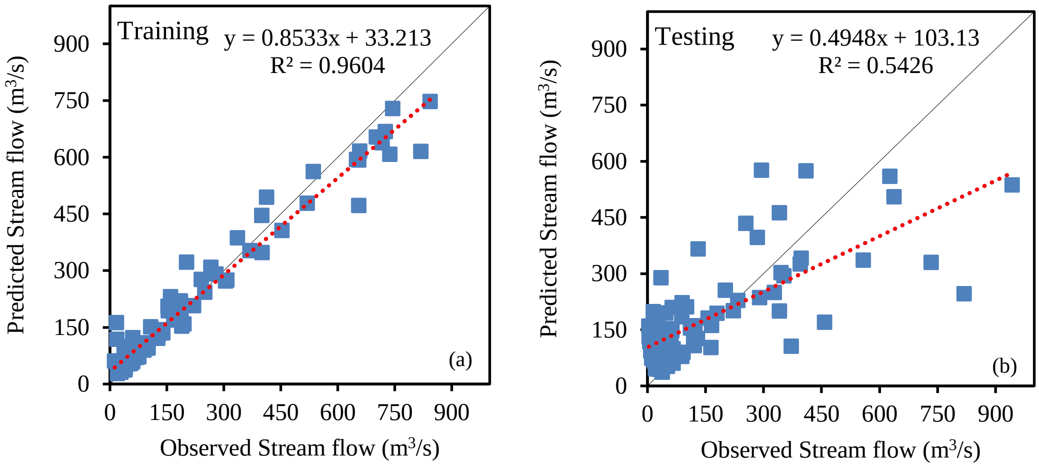


Figure 4. M-11 RF model prediction for 50/50 data separation during (a) training and (b) testing stage.

3.2. 60/40 ratio

The values in Table 2 represent the performance evaluation of M-1 to M-12 RF model based on the 60/40 ratio data separation method, showing the CC, MAE, and RMSE values. The M-12 RF model outperformed all other models with a CC value of 0.976 and 0.844, a MAE value of 30.904 and 69.367, and a RMSE value of 53.845 and 111.976 during training and testing stages, respectively. Figure 5 shows the graphical representation of actual and predicted values of M-12 RF model during both training and testing stages.

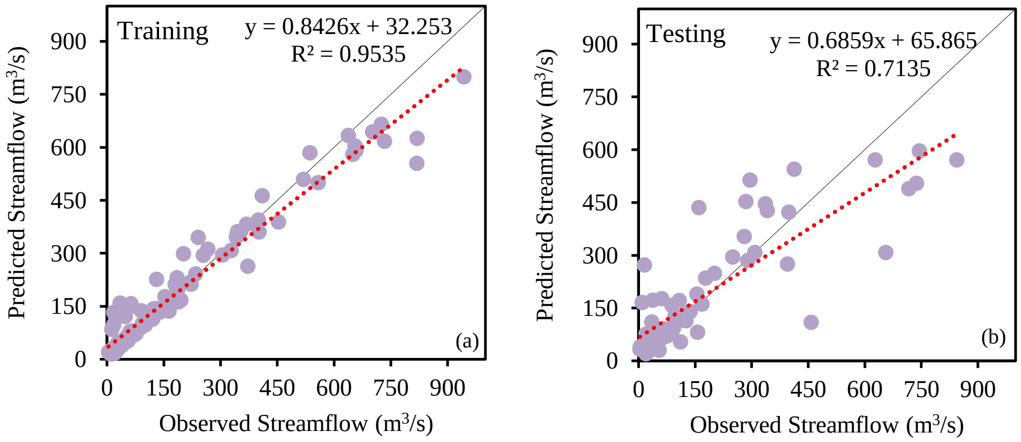


Figure 5. M-12 RF model prediction for 60/40 data separation during (a) training and (b) testing.

3.3. 70/30 ratio

The evaluation index values given in Table 2 show that the M-8 RF model is superior when data is segregated in a 70/30 ratio. The CC value achieved by the M-8 RF model was 0.9812 in the training stage and 0.757 during the testing stage. The MAE and RMSE values during training and testing stages are 31.882, 51.640, 76.75, and 136.167, respectively. The graphical representation of actual and predicted streamflow with 70/30 data is shown in Figure 6. The input parameter to obtain the best model with 70/30 ratio is shown in Table 3.

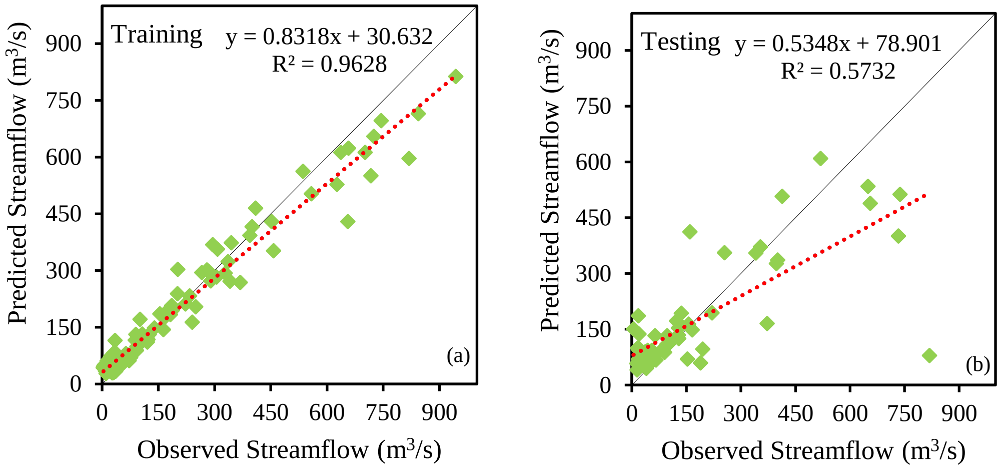


Figure 6. M-12 RF model prediction for 70/30 data separation during (a) training and (b) testing stage.

3.4. 80/20 ratio

From the evaluation index values shown in Table 2, it is observed that the highest CC during the testing stage was obtained by the M-12 RF model. During the training stage, the CC value obtained was 0.979; the MAE and RMSE values were 26.924 and 48.047, respectively. The CC, MAE, and RMSE values during the testing stage were 0.708, 95.575, and 150.163, respectively. The prediction of streamflow with 80/20 data separation is shown in Figure 7a, b.

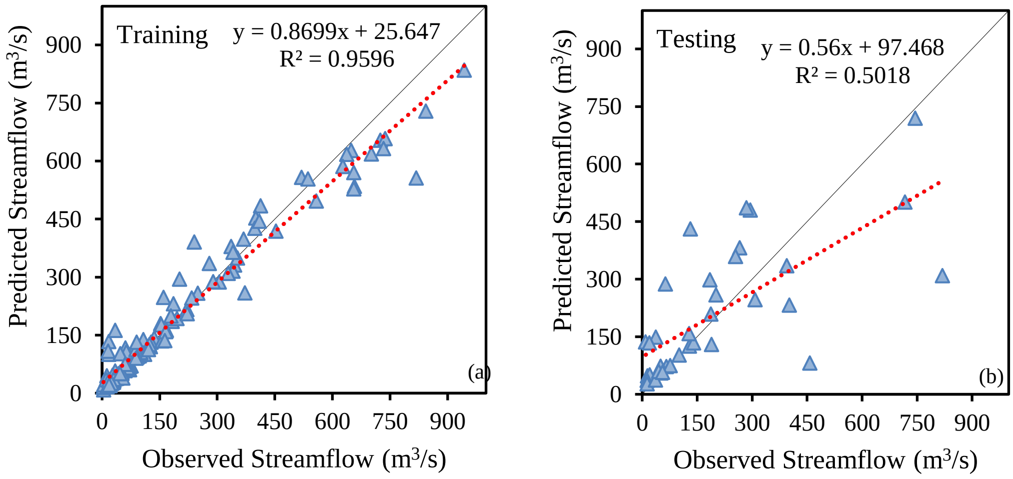


Figure 7. M-12 RF model prediction for 80/20 data separation during (a) training and (b) testing stage.

Table 3. Major factors using RF model.

Separation method	Best model	Number of seeds (k) to get the optimum model
50-50	M-11	k= 1
60-40	M-12	k= 3
70-30	M-8	k= 3
80-20	M-12	k= 3

3.5. Comparative results of data separation methods

In the aforementioned sections, the best RF model was observed for each data segregation method. The comparison among each data segregation method is also needed to provide insightful knowledge on which training and testing datasets need to be set to get the optimum result. For the intercomparison among data separation methods, the streamflow series hydrographs are shown in Figure 8. The actual and predicted streamflow with each data separation method is shown with colored dashed lines, which indicate consistent trends and static positions of each series. It is observed that the trend followed by the 60/40 data series is consistent with the actual data series. Figure 9 shows the box plot, which also confirms that the 60/40 data segregation method is the best. The lower and higher quartiles of the box are used to represent the 25th and 75th percentile values, while the box's median is used to represent the 50th percentile values. A vertical line matching the boxes indicates variation outside of the top and bottom quartiles. In Figure 9, the box plot for the prediction stage indicates that the width of the upper and bottom end of the boxes in the 60/40 data separation method is nearly identical to that of the actual values. Figure 10 also indicates that the 60/40 data separation method gives the highest CC during the testing stage, among others.

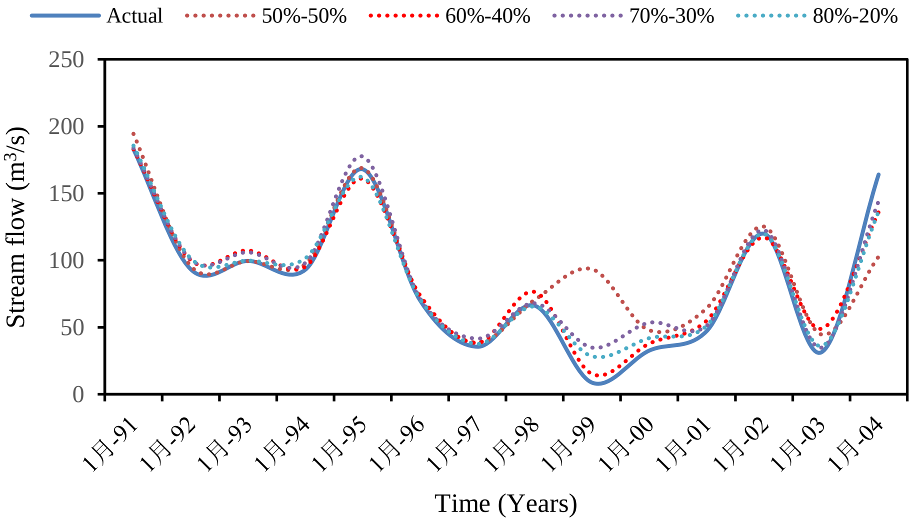


Figure 8. Prediction of data separation using a hydrograph for RF.

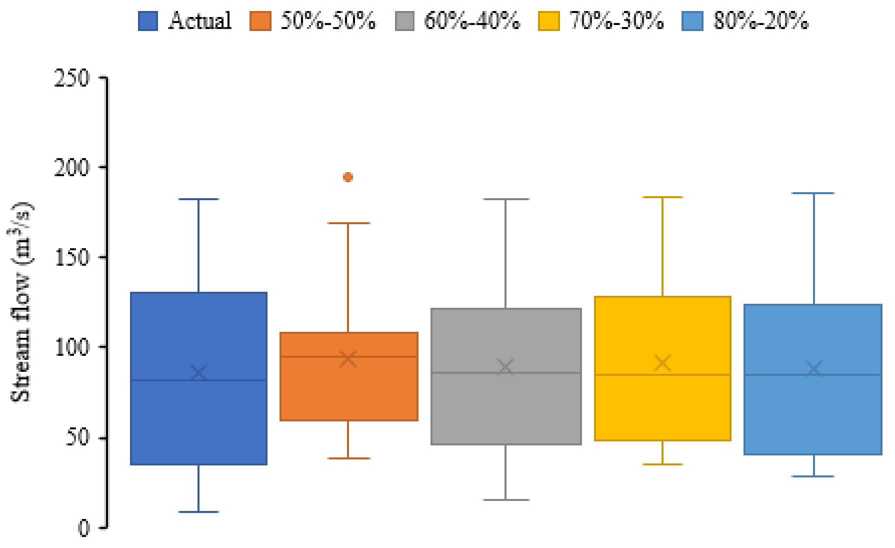


Figure 9. Box plot for actual and predicted values using a data separation method for the testing stage.

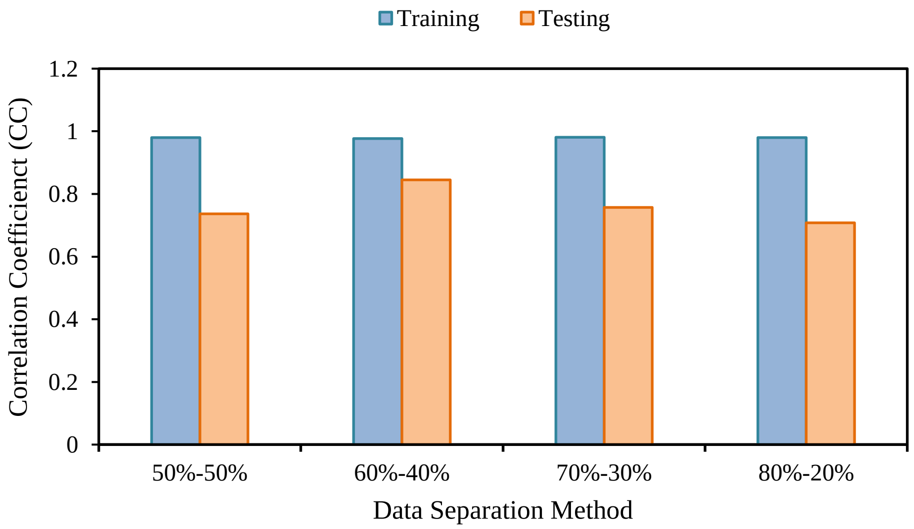


Figure 10. Comparison among data separation me thods based on CC values during the testing stage.

4. Discussion

Before soft computing models, mathematical models were employed to predict streamflow. The results produced by mathematical models were quite satisfactory. The main disadvantage of using conventional methods for time series prediction is that it requires a lot of time [42]. Thus, research has been exploring AI-driven models to reduce the time required for time series prediction. To develop and test the AI models, the dataset needs to be segregated into training and testing datasets. This paper explores four dataset split ratios, i.e., datasets separated into training and testing datasets in a 50:50, 60:40, 70:30, and 80:20 ratio. The extreme learning machine model's (ELM) exceptional flexibility in adapting to variations in training settings and data sizes is evident. The findings demonstrated that when the testing data size falls between 50% and 25% of the total number of data observations, the ELM model requires fewer input variables. But, in certain situations, the model needs extra input features, especially when the training data comprises 80%–90% of the total dataset [43]. Huang F, et al [44] investigated the computer-assisted diagnosis of ECG using a least-square support-vector machine (LS-SVM). The data was divided into three performance measures, i.e., 50/50, 70/30, and 80/20 training and testing. The classification precision obtained was 100% for all the training-to-testing ratios. Many researchers suggested using a training/testing set ratio of 70/30 or 80/20 to create datasets for landslide susceptibility issues [45–47]. Many other researchers have investigated dataset splitting in various other fields [48–51]). A study by Kulkarni S [52] found that when using ML models with an 80/20 training/testing dataset, the ANN-FF model was 100% accurate in classifying lung cancer. The performance was reduced to 96% with a 70/30 data split and further to 94% as the data was split into 60/40. But, as the data was split into 50/50, the accuracy rose to 98%. The SVM accomplished an accuracy of 100% for 50/50 training and testing datasets.

The findings of the current study imply that a training/testing ratio of 60/40 was ideal for both model testing and training. The current study's findings are consistent with the research conducted by AlOmar M. K, et al [43,53]. It was also observed that, with an increase in the percentage of training datasets, the MAE and RMSE error values also increased.

5. Conclusions

The present study aimed to predict the streamflow of the Kesinga basin by using dataset splitting (training/testing), namely at 50/50, 60/40, 70/30, and 80/20 ratios. The RF model was used to predict the streamflow. Validation and comparison results showed that the 60/40 split provided the highest accuracy. The highest CC (0.844) was obtained by the M-12 RF model during the testing stage with a 60/40 dataset split. The MAE and RMSE values during the testing stage acquired by the same model were 69.367 and 111.976, respectively. The 60/40 ratio was followed by the 70/30 ratio, with CC, MAE, and RMSE values of 0.7571, 76.7581, and 136.167, respectively, during the testing stage.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors duly acknowledge with thanks the contribution of the authors from literature whose research articles have been cited in this study. This research did not receive any specific support from public, commercial, or nonprofit funding organizations. All the information was gathered from published works that are in the public domain and properly cited. The datapoints were collected on a monthly basis for the years 1991–2004, from the Central Water Commission (CWC), Mahanadi and Eastern Rivers division, Bhubaneswar, Odisha. Diksha Puri:Conceptualization, experimentation, methodology, formal analysis, models development, writing original draft Parveen Sihag: Reviewing and editing, revision recommendations M.S. Thakur: Supervision, approval of final draft., Mohammad Jameel: proof reading editing, Aaron Chadee: English check; Hazi Azamathulla: Supervision draft Editing.

Conflict of interest

The authors declare they did not have any conflict of interest in the creation of this article.

References

1. Yang D, Yang Y, Xia, J (2021) Hydrological cycle and water resources in a changing world: A review. *Geogr Sustain* 2: 115–122. https://doi.org/10.1016/j.geosus.2021.05.003
2. Liang, S, Ge, S, Wan, L., & Zhang, J. (2010). Can climate change cause the Yellow River to dry up? *Water Resour Res* 46 https://doi.org/10.1029/2009WR007971
3. L Mampitiya, N Rathnayake, Y Hoshino et al. (2024). Forecasting PM10 Levels in Sri Lanka: A Comparative Analysis of Machine Learning Models. *J Hazard Mater Adv* 13: 1–10. https://doi.org/10.1016/j.hazadv.2023.100395
4. HI Tillekaratne, IMSP Jayawardena, V Basnayaka, et al. (2023) Hydro-meteorological disaster incidents and associated weather systems in Sri Lanka. *J Environ Informatics Lett* 10: 89–103. https://doi.org/10.3808/jeil.202300119
5. M Fuladipanah, A Shahhosseini, N Rathnayake, et al. (2024) In-depth simulation of rainfall-runoff relationships using machine learning methods. *Water Pract Technol* (In-Press). https://doi.org/10.2166/wpt.2024.147
6. Palmer M, Ruhi A (2019) Linkages between flow regime, biota, and ecosystem processes: Implications for river restoration. *Science* 365: eaaw2087. https://doi.org/10.1126/science.aaw2087
7. Bierkens, MF, Wada, Y (2019) Non-renewable groundwater use and groundwater depletion: a review. *Environ Res Lett* 14: 063002. https;//doi.org/ 10.1088/1748-9326/ab1a5f
8. Zhou Y, Ma J, Zhang Y, et al. (2019) Influence of the three Gorges Reservoir on the shrinkage of China's two largest freshwater lakes. *Global Planet Change* 177: 45–55. https://doi.org/10.1016/j.gloplacha.2019.03.014
9. Adamowski J. F (2008) Development of a short-term river flood forecasting method for snowmelt driven floods based on wavelet and cross-wavelet analysis. *J* Hydrol 353: 247–266. https://doi.org/10.1016/j.jhydrol.2008.02.013
10. Vorosmarty CJ, Green P, Salisbury J, et al. (2000) Global water resources: vulnerability from climate change and population growth. *Science* 289: 284–288. https://doi.org/10.1126/science.289.5477.284
11. Hanson RT, Newhouse MW, Dettinger, MD (2004) A methodology to asess relations between climatic variability and variations in hydrologic time series in the southwestern United States. *J* Hydrol *287*: 252–269. https://doi.org/10.1016/j.jhydrol.2003.10.006
12. Yang C, Lin Z, Yu Z, et al. (2010) Analysis and simulation of human activity impact on streamflow in the Huaihe River basin with a large-scale hydrologic model. *J Hydrometeorol* 11: 810–821. https://doi.org/10.1175/2009JHM1145.1
13. Makumbura RK, Rathnayake U (2022) Variation of Leaf Area Index (LAI) under changing climate: Kadolkele mangrove forest, Sri Lanka, Advances in Meteorology. https://doi.org/10.1155/2022/9693303
14. Labat D, Ababou R, Mangin A (2000) Rainfall–runoff relations for karstic springs. Part II: continuous wavelet and discrete orthogonal multiresolution analyses. *J Hydrol* 238: 149–178. https://doi.org/10.1016/S0022-1694(00)00322-X

15. Coulibaly P, Burn DH (2004) Wavelet analysis of variability in annual Canadian streamflows. *Water Resour Res 40*. https://doi.org/10.1029/2003WR002667
16. Guven A (2009) Linear genetic programming for time-series modelling of daily flow rate. *J Earth Syst Sci* 118: 137–146. https://doi.org/10.1007/s12040-009-0022-9
17. Yaseen ZM, El-Shafie A, Jaafar O, et al. (2015) Artificial intelligence-based models for streamflow forecasting: 2000–2015. *J Hydrol* 530: 829–844. https://doi.org/10.1016/j.jhydrol.2015.10.038
18. SP Hemakumara, MB Gunathilake, U Rathnayake (2023) Flow alterations due a constructed reservoir in the Menik Ganga basin, Sri Lanka. *Discover Water* 3: 1–15. https://doi.org/10.1007/s43832-023-00049-7
19. Ghimire S, Yaseen ZM, Farooque AA, et al. (2021). Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. *Sci Rep* 11: 17497. https://doi.org/10.1038/s41598-021-96751-4
20. Liu D, Jiang W, Mu L, et al. (2020) Streamflow prediction using deep learning neural network: case study of Yangtze River. *IEEE access* 8: 90069–90086. https://doi.org/10.1109/ACCESS.2020.2993874
21. Arsenault R, Martel JL, Brunet F, et al. (2023) Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. *Hydrol Earth Syst Sci* 27: 139–157. https://doi.org/10.5194/hess-27-139-2023
22. Tabbussum R, Dar AQ (2021) Comparison of fuzzy inference algorithms for stream flow prediction. *Neural Comput Appl* 33: 1643–1653. https://doi.org/10.1007/s00521-020-05098-w
23. Üneş F, Demirci M, Zelenakova M, et al. (2020) River flow estimation using artificial intelligence and fuzzy techniques. *Water* 12: 2427. https://doi.org/10.3390/w12092427
24. Mohammadi B, Linh NTT, Pham QB, et al. (2020) Adaptive neuro-fuzzy inference system coupled with shuffled frog leaping algorithm for predicting river streamflow time series. *Hydrol Sci J* 65: 1738–1751. https://doi.org/10.1080/02626667.2020.1758703
25. Di Nunno F, de Marinis G, Granata, F. (2023) Short-term forecasts of streamflow in the UK based on a novel hybrid artificial intelligence algorithm. *Sci Rep* 13: 7036. https://doi.org/10.1038/s41598-023-34316-3
26. Tikhamarine Y, Souag-Gamane D, Ahmed AN, et al. (2020) Improving artificial intelligence models accuracy for monthly streamflow forecasting using grey Wolf optimization (GWO) algorithm. *J Hydrol* 582: 124435. https://doi.org/10.1016/j.jhydrol.2019.124435
27. Seidu J, Ewusi A, Kuma JSY, et al. (2023) Impact of data partitioning in groundwater level prediction using artificial neural network for multiple wells. *Int J River Basin Ma* 21: 639–650. https://doi.org/10.1080/15715124.2022.2079653
28. Jahanpanah E, Khosravinia P, Sanikhani H, et al. (2019) Estimation of discharge with free overfall in rectangular channel using artificial intelligence models. *Flow Meas Instrum* 67: 118–130. https://doi.org/10.1016/j.flowmeasinst.2019.04.005
29. Demir S, Sahin EK (2022) Comparison of tree-based machine learning algorithms for predicting liquefaction potential using canonical correlation forest, rotation forest, and random forest based on CPT data. *Soil Dyn Earthq Eng* 154: 107130. https://doi.org/10.1016/j.soildyn.2021.107130
30. Ebtehaj I, Bonakdari H, Safari MJS, et al. (2020) Combination of sensitivity and uncertainty analyses for sediment transport modeling in sewer pipes. *Int J Sediment Res* 35: 157–170. https://doi.org/10.1016/j.ijsrc.2019.08.005

31. Zhang W, Zhang R, Wu C, et al. (2020) State-of-the-art review of soft computing applications in underground excavations. *Geosci Front* 11: 1095–1106. https://doi.org/10.1016/j.gsf.2019.12.003
32. Xu Z, Sheykhahmad FR, Ghadimi N, et al. (2020) Computer-aided diagnosis of skin cancer based on soft computing techniques. *Open Med* 15: 860–871. https://doi.org/10.1515/med-2020-0131
33. Al-Janabi S, Mohammad M, Al-Sultan A (2020) A new method for prediction of air pollution based on intelligent computation. *Soft Comput* 24: 661–680. https://doi.org/10.1007/s00500-019-04495-1
34. Wang F, Chun W, Cui, Y (2022) Urban water resources allocation and low-carbon economic development based on soft computing. *Environ Technol Inno* 28: 102292. https://doi.org/10.1016/j.eti.2022.102292
35. Luan C, Liu R, Peng S (2021) Land-use suitability assessment for urban development using a GIS-based soft computing approach: A case study of Ili Valley, China. *Ecol Indic* 123: 107333. https://doi.org/10.1016/j.ecolind.2020.107333
36. Asteris PG, Apostolopoulou M, Armaghani DJ, et al. (2020). On the metaheuristic models for the prediction of cement-metakaolin mortars compressive strength. 1*:* 063.
37. Breiman L (2001) Random forests. *Mach Learn* 45: 5–32. https://doi.org/10.1023/A:1010933404324
38. Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2: 18–22.
39. Scornet E, Biau G, Vert JP (2015). Consistency of random forests. https://doi.org/10.1214/15-AOS1321
40. Mohanty S, Roy N, Singh SP, et al. (2019). Estimating the strength of stabilized dispersive soil with cement clinker and fly ash. *Geotech Geol Eng* 37: 2915–2926. https://doi.org/10.1007/s10706-019-00808-1
41. Breiman L (1996) Bagging predictors. *Mach Learn* 24: 123–140. https://doi.org/10.1007/BF00058655
42. Egawa, T, Suzuki K, Ichikawa Y, et al. (2011, July) A water flow forecasting for dam using neural networks and regression models. In *2011 IEEE Power and Energy Society General Meeting* (1–6). IEEE. https://doi.org/10.1109/PES.2011.6038925
43. AlOmar M. K, Khaleel F, AlSaadi A. A, et al. (2022) The influence of data length on the performance of artificial intelligence models in predicting air pollution. *Adv Meteorol* 2022. https://doi.org/10.1155/2022/5346647
44. Polat K, Akdemir B, Güneş S (2008) Computer aided diagnosis of ECG data on the least square support vector machine. *Digit Signal Process 18*: 25–32. https://doi.org/10.1016/j.dsp.2007.05.006
45. Bui D. T, Pradhan B, Lofman O, et al. (2012) Landslide susceptibility mapping at Hoa Binh province (Vietnam) using an adaptive neuro-fuzzy inference system and GIS. *Comput Geosci* 45: 199–211. https://doi.org/10.1016/j.cageo.2011.10.031
46. Huang F, Yin K, Huang J, et al. (2017) Landslide susceptibility mapping based on self-organizing-map network and extreme learning machine. *Engineering Geology* 223: 11–22. https://doi.org/10.1016/j.enggeo.2017.04.013
47. Pham B. T, Tien Bui D, Pourghasemi H. R, et al. (2017) Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. *Theor Appl Climatol* 128: 255–273. https://doi.org/10.1007/s00704-015-1702-9

48. Khasawneh N, Fraiwan M, Fraiwan L. (2023) Detection of K-complexes in EEG signals using deep transfer learning and YOLOv3. *Cluster Comput* 26: 3985–3995. https://doi.org/10.1007/s10586-022-03802-0
49. Kaur R, Kumar R, Gupta, M (2022) Predicting risk of obesity and meal planning to reduce the obesity in adulthood using artificial intelligence. *Endocrine* 78: 458–469. https://doi.org/10.1007/s12020-022-03215-4
50. Ikram R. M. A, Dai H. L, Ewees A. A, et al. (2022) Application of improved version of multi verse optimizer algorithm for modeling solar radiation. *Energy Rep* 8: 12063–12080. https://doi.org/10.1016/j.egyr.2022.09.015
51. Shirzadi A, Solaimani K, Roshan M. H, et al. (2019) Uncertainties of prediction accuracy in shallow landslide modeling: Sample size and raster resolution. *Catena* 178: 172–188. https://doi.org/10.1016/j.catena.2019.03.017
52. Kulkarni S (2023, November) Impact of Various Data Splitting Ratios on the Performance of Machine Learning Models in the Classification of Lung Cancer. In *Proceedings of the Second International Conference on Emerging Trends in Engineering (ICETE 2023)* (223: 96). Springer Nature. https://doi.org/10.2991/978-94-6463-252-1_12
53. Kisi O, Mirboluki A, Naganna S. R, et al. (2022) Comparative evaluation of deep learning and machine learning in modelling pan evaporation using limited inputs. *Hydrol Sci J* 67*:* 1309–1327. https://doi.org/10.1080/02626667.2022.2063724