SEMARAK ILMU PUBLISHING

Semarak International Journal of Machine Learning

Journal homepage:
https://semarakilmu.my/index.php/sijml/index
ISSN: 3030-5241

Machine Learning-Based Approach for Filling Gaps in Streamflow Data

Jing Lin Ng[1,2,*], Aik Hang Chong[2], Jin Chai Lee[2], Nur Ilya Farhana Md Noh[1], Muyideen Abdulkareem[2], Deprizon Syamsunur[2], Ramez A. Al-Mansob[3], Majid Mirzaei[4], Siaw Yin Thian[5]

[1] School of Civil Engineering, College of Engineering, Universiti Teknologi Mara (UiTM), 40450 Shah Alam, Selangor, Malaysia
[2] Department of Civil Engineering, Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur, 56000, Malaysia
[3] Department of Civil Engineering Department, International Islamic University Malaysia, Gombak, 53100, Malaysia
[4] Department of Civil, Construction, and Environmental Engineering, University of Alabama, Tuscaloosa, AL, USA
[5] Water Resources and Climate Change Research Centre, National Hydraulic Research Institute of Malaysia (NAHRIM), Seri Kembangan, Selangor, 43300, Malaysia

ARTICLE INFO

Article history:
Received 12 January 2025
Received in revised form 12 February 2025
Accepted 12 March 2025
Available online 21 March 2025

Keywords:
kNN; Machine Learning Models; CART; missing streamflow data; estimation method; Naïve Bayes (NB)

ABSTRACT

The lack of streamflow data can significantly impact the flood prediction capacity of various Malaysian agencies, including the National Disaster Management Agency (NADMA). To address this issue, we investigated the use of machine learning methods to estimate missing streamflow data in eleven stations in Peninsular Malaysia. We compared the performance of three machine learning methods (Naive Bayes, k-Nearest Neighbors model, and Multiple Classification and Regression Tree) with five conventional methods (coefficient of correlation, Arithmetic Average Method, Inverse Distance Weighting Model, Linear Interpolation, and Normal Ratio) using statistical approach such as Coefficient of Correlation (R), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). We conducted homogeneity tests using the Pettitt test, Buishand Range (BR) test, Standard Normal Homogeneity Test (SNHT), and Von Neumann Ratio (VNR) test to determine the quality of the data after the data collection was completed. The results of the homogeneity tests showed that the streamflow data series were not randomly distributed. Our results indicated that the machine learning approach outperformed conventional methods in estimating missing streamflow data. The Naive Bayes approach, in particular, was the most successful, using only a modest quantity of training data to properly forecast the outcomes. Our study's contribution is the application of machine learning algorithms to estimate missing streamflow data, and our findings might help Malaysian flood control efforts. Overall, our findings show that machine learning approaches have the potential to improve the accuracy of streamflow data prediction, which is critical for successful flood control.

1. Introduction

Streamflow datasets are often incomplete for various reasons, such as damage to the measuring station. These missing values reduce the power and accuracy of statistical analysis methods, which can lead to biased estimates of relationships between variables. Furthermore, the lack of streamflow data caused by these missing values has greatly impacted the flood forecasting

* *Corresponding author.*
E-mail address: jinglin.ng787@gmail.com

https://doi.org/10.37934/sijml.5.1.4663a

capabilities of some agencies in Malaysia, such as the National Disaster Management Agency (NADMA). Engineers and hydrologists may make inaccurate assumptions about these incomplete datasets, and even a small error or missing value can significantly impact the engineering results and analysis. As a result, engineers and hydrologists have a hard time determining how much water is available to prevent flooding in different locations.

According to Hamzah *et al.,* [10], streamflow datasets are often incomplete due to long-term exposure of physical sensing equipment to multiple risks such as maintenance or technical problems, adverse weather conditions, equipment failures, and human error during data entry, and data corruption due to storage mechanical failure. They found that streamflow data was missing, probably because the stations relied heavily on automated data acquisition systems with many sensing devices. Hence, sensor readings are often incomplete over long periods of time.

Handling missing values is a crucial responsibility. Several conventional methods for estimating missing streamflow data include the coefficient of correlation (CC) method, the Angstrom method, crowd-sourcing method, the inverse distance (IDW) method, and the normal ratio (NR) method [7,17,31]. In a study by Ismail *et al.,* [17] that compared the accuracy of these methods in Terengganu, Malaysia, the normal ratio method was found to be the best method, with the lowest values of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) compared to the other three methods. It is recommended to increase the number and distance of adjacent stations to improve the accuracy of the method.

Conventional methods are still widely used in missing data estimation. However, they may lead to biased and inaccurate results under certain conditions. Machine learning models such as ANN are more suitable for a wide range of sample data than conventional methods. According to another study by Tan *et al.,* [31], missing data estimation was carried out using ANN and other traditional methods, namely the inverse distance weighting method (IDW), the ordinary kriging method (OK), the linear regression method (LR), and the normal ratio method (NR) in Kelantan River Basin, Malaysia. The findings showed that the ANN was the best overall estimating approach, obtaining lower mean absolute error, mean square error, highest linear correlation, and less bias compared with the traditional methods. Hence, they concluded that the ANN method was a more efficient method for estimating missing data in the Kelantan River Basin of Malaysia. The process of using computers to learn information without explicit instructions is known as machine learning. The adaptive neuro-fuzzy inference system (ANFIS) is based on how the human brain performs categorization, recognition, and identification. In a study by Sharma *et al.,* [35], several machine learning models were used to estimate missing values in a station located in Eastern Bhutan, including the k-Nearest Neighbours model (kNN) and Multiple Linear Regression Model (MLRM). The kNN model with bootstrapping technique was used to estimate missing data, resulting in very low data bias and standard error, and the data was accepted for further study. The multiple linear regression model (MLRM) was also used to estimate the data, and the estimated data fairly represented the overall data variability.

Missing data is a common problem in engineering research and can significantly influence the conclusions drawn from the data. It can degrade research performance, produce biased estimates, and lead to invalid conclusions. Therefore, the objectives of this study were to investigate different machine learning models for estimating missing streamflow, evaluate the performances of different estimation methods, and determine the most appropriate method for estimating missing streamflow data. The study results will contribute to a better understanding of estimation methods and their performance for estimating missing streamflow data.

2. Methodology

2.1 Study Area

The study focused on the state of Peninsular Malaysia, also known as West Malaysia, located at 3.97°N latitude and 102.43°E longitude. The total area of Peninsular Malaysia is approximately 132,265 square kilometers, which is nearly 40% of the total area of Malaysia. Peninsular Malaysia had high temperature, high humidity, and heavy rainfall along the year. Typically, the climate of Peninsular Malaysia is affected by winds from the southwest monsoon, which occurs from June to October, while the northeast monsoon occurs from November to March. The transition period of these two monsoons is the inter-monsoon period from March to May and September to October, which brings strong convective rain to many parts of the peninsular [21].

2.2 Data Collection

The historical streamflow data used in this study were purchased from the Department of Irrigation and Drainage (DID). The geographic coordinates of all 11-streamflow station selected from each of the 11 states in Peninsular Malaysia are shown in Figure 1. Meanwhile, the details of each selected stations were shown in Table 1.

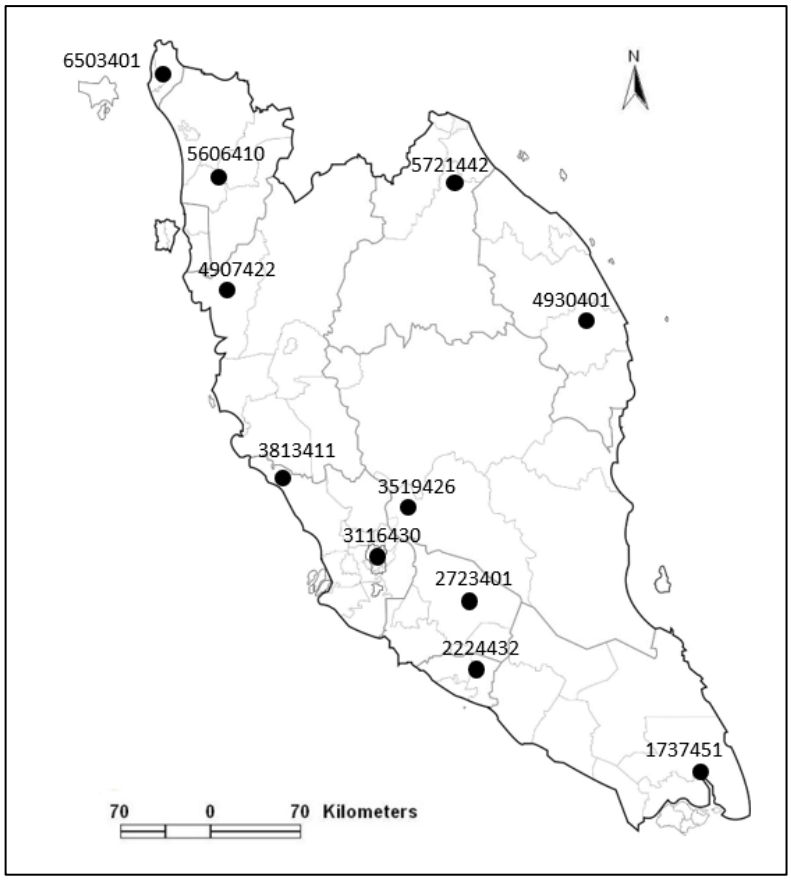

Fig. 1. Geographic coordinates of all 11-streamflow station selected

Table 1
Information of all streamflow station

Station Code	Station name	Study period	Duration	Longitude	Latitude
1737451	Sungai Johor at Rantau Panjang	1972~1992	20 years	01° 46' 50"E	103° 44' 45"N
5606410	Sungai Muda at Jambatan Syed Omar	1974~1994	20 years	05° 36' 35"E	100° 37' 35"N
5721442	Sungai Kelantan at Jambatan Guillemard	1973~1993	20 years	05° 45' 45"E	102° 09' 00"N
2224432	Sungai Kesang at Chin Chin	1960~1980	20 years	02° 17' 25"E	102° 29' 35"N
2723401	Sungai Kepis at Jambatan Kayu Lama	1979~1999	20 years	02° 42' 20"E	102° 21' 20"N
3519426	Sungai Bentong at Kuala Marong	1970~1990	20 years	03° 30' 45"E	101° 54' 55"N
6503401	Sungai Arau at Ladang Tebu Felda	1984~2004	20 years	06° 30' 10"E	100° 21' 05"N
3116430	Sungai Klang at Jambatan Sulaiman	1995~2015	20 years	03° 08' 20"E	101° 41' 50"N
4930401	Sungai Berang at Menerong	1998~2018	20 years	04° 56' 20"E	103° 03' 45"N
3813411	Sungai Bernam at Jambatan Skc	1984~2004	20 years	03° 48' 27"E	101° 21' 70"N
4907422	Sungai Kurau at Bt. 14 Jalan Taiping	1975~1995	20 years	04° 58' 40"E	100° 46' 50"N

2.3 Estimation of Missing Streamflow Data using Conventional Method
2.3.1 Normal Ratio Method (NR)

The NR method is weighted according to the ratio average of the accessible data of the target station to the i_{th} adjacent station. The estimated missing values, P_t are given by (Hamzah *et al.,* [10] 2020):

$$P_t = \frac{1}{n}\sum_{i=1}^{n}\frac{N_t}{N_i}x_i \tag{1}$$

where N_i is the total streamflow to each adjacent station and N_t is the total streamflow to the target station. This method should only be used if the normal streamflow data for any adjacent station exceeds 10% of the station under consideration [10].

2.3.2. Inverse Distance Method (IDW)

The IDW method is based on the idea of weighting the distance between the target station and the adjacent stations [10]. The formula of IDW method can express as:

$$P_t = \frac{\sum_{i=1}^{n}\frac{x_i}{N_i}}{\sum_{i=1}^{n}\frac{1}{d_{it}}} \tag{2}$$

where P_t is missing values, the i_{th} adjacent station and d_{it} is the distance between the target station.

2.3.3 Coefficient of Correlation Method (CC)

In the CC method, the distance is replaced by the correlation coefficient between the target and the adjacent station as the weighted value. The missing values, P_t are estimated as:

$$P_t = \frac{\sum_{i=1}^{n} x_i r_{it}}{\sum_{i=1}^{n} r_{it}} \quad (3)$$

where i_{th} adjacent stations and the r_{it} is the correlation coefficient for daily time series data between targets [10].

2.3.4. Arithmetic Average Method (AM)

The AM method is the easiest way to fill in the streamflow data. The missing data for streamflow was recovered by averaging selected adjacent stations around the target station or by using the same day date in different years to fill in the gaps [10]. The estimated missing values, P_t are given by:

$$P_t = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (4)$$

where *xi* is the observed data at the i_{th} nearby stations and P_t is the estimated value of the missing data at the *t* target station or date and *n* is the count of nearby stations or number of years.

2.3.5. Linear Interpolation Method (LI)

LI method is also one of the easiest ways to estimates data values from two data points in a one-dimensional data sequence that are adjacent to the point that needs to be interpolated. The missing values, P_t are estimated as [20]:

$$P_t(x) = P(x_0) + \frac{P(x_1)+P(x_0)}{x_1-x_0}\ (x - x_0) \quad (5)$$

where x_1 and x_0 are the known values of the independent variable, x is the independent variable and P_t is the dependent variable value of the independent variable x value.

2.4 Estimation of Missing Streamflow Data Using Machine Learning Models

2.4.1 K-Nearest-Neighbor Imputation (K-NN)

KNN is the method to calculate the distance between the missing point and other complete point. For KNN, finding and selecting the right k value is a key step in this process, which affects the accuracy and model enhancement. The better the estimation of the missing value, the smaller the k value. The following formula was used in this study to calculate Euclidean distance, D which is one of the most widely used distance measures [8]:

$$D(x, y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (6)$$

where yi a case from the streamflow data sample and xi is query point.

2.4.2 Multiple Classification and Regression Tree (CART)

CART is a well-known machine learning algorithm classification. CART learns fast and predicts fast. They are also generally accurate for a wide range of missing data. Each node in the tree contains a splitting rule for the CART issue that is specified by minimising the relative error (RE) [8]:

$$RE(d) = \sum_{l=0}^{L}(y_l - \overline{y_L})^2 + \sum_{r=0}^{R}(y_r - \overline{y_R})^2 \quad (7)$$

where y_l and y_r are the left and right partitions, respectively, with L and R observations of y in each, and respective means $\overline{y_L}$ and $\overline{y_R}$.

2.4.3 Naive Bayes (NB)

NB is widely used algorithm due to its simplicity in estimating the missing data. NB often works much better than expected in the most complex real-world situations because the algorithm is based on posterior probabilities that combine prior experience and event likelihood. According to Bayes' theorem, it shows how to calculate the posterior probability, $P(c|x)$ [24]:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (8)$$

where *P(c)* = Class prior probability; *P(c|x)* = Posterior probability of class (c, target) given predictor (x, attributes); *P(x)* = Predictor prior probability; and *P(x|c)* = Likelihood, which is the probability of predictor given class.

2.5 Performance Evaluation of Estimation Methods

The performances of all the proposed methods for estimating missing streamflow data were evaluated by using three metrics: the RMSE, the MAE and the R. MAE measures the average magnitude of the errors between the estimated and observed streamflow values. It indicates how close the estimates are to the true values. RMSE is a measure of the overall difference between the estimated and observed streamflow values, taking into account both the magnitude and direction of the errors. It is particularly sensitive to large errors, as they are squared in the calculation. R measures the strength of the linear relationship between the estimated and observed streamflow values. Based on the overall result, the method that acquires lowest value of RMSE and MAE, and the highest of R will be selected as the best estimation method. The equation formula of the three metrics was defined as below.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | X_i - Y_i | \quad (9)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2} \quad (10)$$

$$R = \frac{\mathrm{N}\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{N(\sum {X_i}^2) - (\sum X_i)^2}\sqrt{N(\sum {Y_i}^2) - (\sum Y_i)^2}} \quad (11)$$

where Y_i is the estimated data, X_i is the observed streamflow data, and N is the number of observations [28].

2.6 Homogeneity Test

Homogeneity testing was performed to determine whether a series can be considered homogenous. The four methods used to test the homogeneity of streamflow data were the Buishand Range test (BR), standard Normal homogeneity test (SNHT), Pettitt test, and Von Neumann Ratio test (VNR). Details of the various homogeneity tests applied in this study are presented below.

2.6.1 Standard Normal Homogeneity Test (SNHT)

The SNHT is used to check the non-homogeneity in meteorological data series. To compare the mean of the first *n* observations with the mean of the remaining observations with *n* data points, a statistic (Tk) is applied [13]:

$$T_k = kZ_1^2 + (n-k)Z_2^2 \tag{12}$$

$$Z_1 = \frac{1}{k}\sum_{i=1}^{k} \frac{(x_i - \bar{x})}{\sigma x} \tag{13}$$

$$Z_2 = \frac{1}{n-k}\sum_{i=k+1}^{d} \frac{(x_i - \bar{x})}{\sigma x} \tag{14}$$

where, $\bar{x}$ is the mean, σx is the standard deviation of the series and x_i is the observed value. T_k reaches its maximum value if the break is at year *k*.

2.6.2 Buishand Range Test

BR test can be used for variables that follow any type of distribution and it is based on the adjusted partial sum deviation from the mean. In this test, the adjusted partial sum S_k is defined as:

$$S_k^* = \sum_{t=1}^{k}(x_t - \bar{x}),\ k = 1,2,3,4,\dots,n \tag{15}$$

where $\bar{x}$ is the sample mean, *n* is the number of records in the time series, and x_t is the observed value.

2.6.3 Pettitt Test

The Pettit test is a widely used non-parametric test developed by Pettit (1979) to assess the occurrence of sudden changes in climate records. The test statistics X_E for this test may be defined as follows:

$$X_E = \max |\ X_k\ | \quad , \quad 1 \le k \le n \tag{16}$$

$$X_k = 2\sum_{i=0}^{n} r_i - k(n+1) \qquad k = 1,\dots,n \tag{17}$$

Where n is the number of years and r_i is the rank of the i_{th} observation is used to calculate the statistics [1].

2.6.4 Von Neumann Ratio Test

VNR test is used to determine overall non-homogeneity present in the data series. The von Neumann ratio, N is defined as:

$$N = \frac{\sum_{i=1}^{n-1}(x_t - x_{t+1})^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \quad (18)$$

where $\bar{x}$is the sample mean with sample size *n* and x_i is the observed value.

3. Results & Discussion

3.1 Estimate of Five Percent Missing Streamflow Data

The result of estimation methods based on R, MAE, and RMSE with five percent of missing values for streamflow data were shown in Table 2. A low RMSE value indicates that the estimation method can estimate data close to the actual data, resulting in high accuracy. It was observed that the Naïve Bayes (NB) method achieved the lowest RMSE values for six out of eleven stations. Therefore, when compared to those other methods of estimation, the NB method was the method that provided the most accurate results.

The MAE value was a measure the accuracy of method performs compared to another. If the value of MAE is lower, this indicates that the data produced by the estimation method are closer to the actual data. The NB had obtained the lowest MAE values in the evaluation of the method of performance MAE at stations 1737451, 2224432, 5721442, 6503401, 3519426, and 3813411, respectively. The NB method had the lowest MAE values for six of eleven stations. Therefore, NB method was the most accurate method in estimation missing data.

A greater value for R in the positive direction suggests more accurate data estimates and results. It is possible that this is one of the factors that contributes to how accurate the estimation method is. Among the stations 1737451, 2224432, 5721442, 6503401, and 3519426, it was determined that the NB had the greatest R values of 6.02, 1.04, 12866.63, 0, and 0.84, respectively. Five out of eleven stations had the greatest R values using the NB method. Consequently, the NB method obtained greater R values at most of the stations, showing that it was the most accurate method for estimating missing data.

3.2 Estimate of Ten Percent Missing Streamflow Data

The result of estimation methods based on R, MAE, and RMSE with ten percent of missing values for streamflow data were shown in Table 2. It was discovered that the NB method had the lowest RMSE values among the ten percent of streamflow data that was missing, at stations 1737451, 2224432, 5721442, 2723401, 6503401, 3519426, and 3813411. The NB achieved the lowest RMSE values for seven out of eleven stations. Therefore, when compared to the other methods, the NB method gave the most accurate result.

In the evaluation of the methods of performance MAE at the station, the NB also had received the lowest MAE values at stations 1737451, 2224432, 5721442, 2723401, 6503401, 3519426, and 3813411. The MAE values were the lowest for seven out of eleven stations when estimated using the NB method. Hence, the NB method was the method that was the most successful and accurate in estimating the missing data.

Furthermore, the NB had the highest R values of 19.84, 2, and 6.46 out of all the stations that were located at stations 1737451, 2224432, and 3519426, respectively. The NB method had the highest R values for three out of eleven stations. Therefore, the NB method obtained the highest number of possible stations, indicating that it was the most accurate method for estimating missing streamflow data.

3.3 Estimate of Fifteen Percent Missing Streamflow Data

The result of estimation methods based on R, MAE, and RMSE, with fifteen percent of missing values for streamflow data, is shown in Table 2. It was found that the NB method had the lowest RMSE values among the fifteen percent of streamflow data that were missing. These values were obtained as 0.23, 1538.73, 0.12, 0.47, and 0.34 at stations 1737451, 5721442, 2723401, 6503401, and 3519426, respectively. The NB method produced the best results in terms of RMSE for five out of eleven stations. As a result, the estimates produced by the NB method were the most accurate when compared with the results obtained by other estimation methods.

Additionally, the NB method acquired the lowest MAE values at stations 1737451, 5721442, 2723401, 6503401, and 3519426. When applying the NB method to estimate the MAE, five out of eleven stations had the lowest MAE values. Therefore, the NB method was the most effective and precise in estimating the missing data.

Moreover, the NB method was determined to have the greatest R values at stations 6503401, 3519426, 4930401, and 3813411. In four out of the eleven stations, the NB method produced the highest values of R. Therefore, the NB method received the most potential stations, indicating that it was the most accurate method for estimating the missing streamflow data.

Overall, the results suggest that the NB method is the most accurate and effective method for estimating missing streamflow data.

Table 2
Comparison of estimation methods based on R, RMSE and MAE with various percentages of missing values for streamflow data

Station	Method	5%			10%			15%		
		RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE	R
1737451	KNN	3.16	2.24	0.31	4.56	2.63	5.84	5.60	2.50	35.95
	CART	2.72	1.92	1.78	4.19	2.42	6.38	3.58	1.60	**54.42**
	NB	**0.21**	**0.15**	**6.02**	**0.11**	**0.06**	**19.84**	**0.23**	**0.10**	50.77
	LI	11.36	8.03	-6.70	9.23	5.33	-30.98	3.72	1.66	-134.82
	NR	2.94	2.08	-4.29	6.48	3.74	-23.35	12.85	5.75	-99.37
	IDW	5.03	3.56	-7.13	1.51	0.87	-37.02	5.79	2.59	-164.66
	AM	13.91	9.84	-7.96	10.81	6.24	-41.20	3.37	1.51	-186.45
	CC	32.99	23.32	-9.80	30.11	17.38	-50.54	20.82	9.31	-238.66
2224432	KNN	0.45	0.32	0.74	0.37	0.37	1.81	0.75	0.34	3.01
	CART	0.16	0.12	0.50	0.23	0.23	1.16	**0.37**	**0.16**	1.86
	NB	**0.07**	**0.05**	**1.04**	**0.18**	**0.18**	**2.00**	2.00	0.89	1.32
	LI	7.73	5.47	0.20	5.60	5.60	0.60	12.50	5.59	0.93
	NR	1.01	0.72	0.71	0.71	0.71	1.60	1.27	0.57	2.73
	IDW	14.39	10.18	0.73	12.06	12.06	1.93	25.10	11.23	**3.24**
	AM	10.04	7.10	0.61	8.10	8.10	1.64	16.80	7.52	2.73
	CC	8.53	6.03	0.32	6.29	6.29	0.85	13.05	5.84	1.39
5721442	KNN	393.17	278.01	4494.79	696.27	401.99	292114.18	2054.87	918.97	8162594.80
	CART	405.02	286.40	-3097.67	652.26	376.58	425694.08	1895.70	847.78	6559051.33
	NB	**53.03**	**37.50**	**12866.63**	**83.77**	**48.37**	623549.02	**1538.73**	**688.14**	5410179.06
	LI	1196.72	846.21	96110.36	1683.72	972.10	747603.41	3132.98	1401.11	15884903.34
	NR	877.79	620.69	108150.67	839.31	484.57	**1737129.08**	1877.21	839.51	5464458.00
	IDW	1707.52	1207.40	94779.83	2361.13	1363.20	1711614.68	4296.18	1921.31	36690684.79
	AM	1667.16	1178.86	49567.37	2303.74	1330.07	1414750.31	4152.65	1857.12	39111173.65
	CC	1617.02	1143.41	7566.47	2241.77	1294.29	1069684.35	3993.21	1785.82	**42554357.90**
2723401	KNN	**0.34**	**0.24**	**0.01**	0.59	0.34	**0.14**	1.31	0.59	2.45
	CART	1.39	0.99	0.00	1.92	1.11	0.07	2.80	1.25	3.02
	NB	1.13	0.20	0.01	**0.12**	**0.07**	0.06	**0.12**	**0.05**	**5.85**
	LI	1.14	0.81	-0.11	0.56	0.32	-1.11	0.17	0.08	-1.25
	NR	2.19	1.55	-0.55	5.09	2.94	-3.12	6.44	2.88	0.17
	IDW	2.18	1.54	-0.50	0.26	0.15	-4.04	0.38	0.17	0.82
	AM	1.79	1.27	-0.50	0.35	0.20	-4.08	0.37	0.16	0.68
	CC	2.59	1.83	-0.51	0.90	0.52	-4.00	1.17	0.52	0.93
4907422	KNN	0.57	0.40	-0.27	1.50	0.87	21.33	3.13	1.40	52.70

	CART	**0.04**	**0.03**	**0.66**	**0.36**	**0.21**	**23.44**	**0.51**	**0.23**	**62.09**
	NB	0.71	0.50	0.00	1.15	0.67	0.00	3.13	1.40	0.00
	LI	13.75	9.73	0.56	16.28	9.40	5.30	24.91	11.14	22.47
	NR	0.15	0.11	0.52	0.33	0.19	20.57	1.07	0.48	51.16
	IDW	10.56	7.47	0.53	20.10	11.61	21.72	31.45	14.07	54.79
	AM	12.49	8.83	0.48	23.86	13.78	21.62	34.88	15.60	52.49
	CC	12.22	8.64	0.52	21.00	12.12	18.57	30.20	13.51	45.75
6503401	KNN	1.41	1.00	0.00	1.73	1.00	0.00	2.24	1.00	0.00
	CART	1.13	0.80	0.00	1.35	0.78	0.00	1.66	0.74	0.00
	NB	**0.40**	**0.29**	**0.00**	**0.42**	**0.24**	**0.00**	**0.47**	**0.21**	**0.00**
	LI	14.14	10.00	0.00	16.60	9.58	0.00	20.46	9.15	0.00
	NR	0.43	0.30	0.00	0.51	0.30	0.00	0.65	0.29	0.00
	IDW	37.73	26.68	0.00	44.59	25.74	0.00	55.27	24.72	0.00
	AM	18.38	13.00	0.00	21.75	12.56	0.00	26.98	12.07	0.00
	CC	6.17	4.36	0.00	7.32	4.23	0.00	9.11	4.08	0.00
5606410	KNN	**2.97**	**2.10**	2.67	**0.46**	**0.27**	393.00	**0.54**	**0.24**	660.44
	CART	5.42	3.83	4.89	5.24	3.03	711.80	7.57	3.39	1233.57
	NB	6.36	4.50	27.10	2.31	1.33	291.84	10.29	4.60	424.62
	LI	30.05	21.25	30.36	46.33	26.75	90.02	54.78	24.50	31.71
	NR	15.40	10.89	**34.50**	4.43	2.56	1886.64	23.32	10.43	2621.12
	IDW	46.27	32.72	26.11	57.30	33.08	1796.92	69.86	31.24	2643.99
	AM	47.85	33.83	17.14	58.89	34.00	**1974.10**	71.55	32.00	**2853.74**
	CC	46.42	32.83	28.13	59.46	34.33	1550.88	73.20	32.73	2292.32
3519426	KNN	2.33	1.65	1.08	1.73	1.00	1.10	0.91	0.41	-2.37
	CART	1.42	1.01	0.07	1.33	0.77	4.83	0.55	0.24	6.28
	NB	**0.42**	**0.30**	**0.84**	**0.25**	**0.14**	**6.46**	**0.34**	**0.15**	**18.68**
	LI	237.93	168.25	-0.26	297.26	171.62	-1.54	417.17	186.56	-14.62
	NR	1.23	0.87	-0.68	0.49	0.28	-0.59	0.96	0.43	-0.68
	IDW	348.21	246.22	-0.39	415.08	239.64	1.68	523.97	234.33	4.62
	AM	344.50	243.60	-0.39	410.54	237.02	1.70	518.73	231.99	4.59
	CC	336.09	237.65	-0.40	401.12	231.59	1.58	505.98	226.28	4.53
4930401	KNN	**0.12**	**0.08**	**0.23**	**0.02**	**0.01**	**0.69**	1.40	0.63	104.79
	CART	0.80	0.57	0.01	1.17	0.67	0.24	**0.48**	**0.22**	81.31
	NB	0.06	0.05	0.11	0.18	0.11	0.61	0.89	0.40	**173.16**
	LI	38.53	27.24	0.22	44.39	25.63	0.68	63.61	28.45	-1.00
	NR	4.70	3.33	0.22	5.60	3.24	0.64	5.65	2.53	54.32
	IDW	42.39	29.97	0.18	49.64	28.66	0.53	79.96	35.76	95.16
	AM	47.47	33.57	0.20	55.33	31.95	0.58	88.60	39.62	93.71

	CC	76.46	54.06	0.13	90.42	52.20	0.40	159.42	71.30	114.53
3116430	KNN	9.96	7.05	-16.31	10.22	5.90	-18.24	15.21	6.80	-26.44
	CART	**0.86**	**0.60**	40.81	**0.83**	**0.48**	72.79	**0.23**	**0.10**	135.21
	NB	7.46	5.28	59.74	11.83	6.83	90.15	8.27	3.70	47.22
	LI	16.58	11.73	-18.51	25.99	15.01	-109.66	31.92	14.27	-163.39
	NR	1.30	0.92	65.35	4.13	2.38	166.26	1.79	0.80	461.36
	IDW	26.66	18.85	75.30	33.94	19.60	150.67	45.82	20.49	429.20
	AM	26.26	18.57	77.33	33.05	19.08	163.78	44.95	20.10	461.17
	CC	34.11	24.12	**79.51**	42.17	24.35	**200.73**	55.81	24.96	**550.88**
3813411	KNN	16.18	11.44	-40.76	22.96	13.26	269.01	**7.68**	**3.43**	-465.14
	CART	18.63	13.18	**109.47**	23.23	13.41	**667.74**	8.51	3.80	-66.05
	NB	**8.06**	**5.70**	56.36	**11.80**	**6.81**	448.25	23.70	10.60	**3906.79**
	LI	43.26	30.59	-54.39	61.03	35.23	-50.54	69.09	30.90	472.53
	NR	9.09	6.43	42.21	20.30	11.72	20.76	16.84	7.53	915.10
	IDW	70.20	49.64	64.25	94.80	54.73	-40.86	110.71	49.51	1232.53
	AM	55.61	39.32	71.39	78.53	45.34	-222.34	89.43	39.99	804.60
	CC	45.97	32.50	76.91	67.76	39.12	-260.47	75.59	33.80	817.16

The bolded value represents the method that obtained the highest accuracy.

4. Discussions

In this study, missing streamflow data were estimated using all eight estimation methods for each streamflow station located on the Peninsular Malaysia with similar climatic conditions. The streamflow data were evaluated at each station based on three different percentages of missing data: 5%, 10%, and 15%, respectively. The performance of the estimation methods was evaluated using measures such as RMSE, MAE, and R. The higher the total points, the better the performance of the methods. The overall ranking of the methods was presented in Table 3 to Table 5 based on the various percentages of missing data.

According to the overall ranking for estimation methods, it was discovered that NB was the most effective method, while CART and KNN were classified as the second and third most effective options, respectively. NB showed advantages over other estimation methods as it provides a higher degree of accuracy and enables the use of incomplete data for estimation. The results were comparable to those found in studies carried out by Bayhaqy *et al.,* [2] and Jackins *et al.,* [12], in which NB was found to have the best performance when used to estimate missing streamflow datasets in Tokopedia and Bukalapak, Indonesia. This is because NB offers extremely accurate and specific predictions. NB is a form of machine learning that takes into account the Bayesian approach while doing calculations involving probability. It is not necessary to have a lot of data before starting the estimation process. This method is still extremely straightforward and functions excellently and quickly even when applied to most of the missing data estimation.

Additionally, conventional methods such as NR, LI, AM, IDW, and CC were compared with machine learning methods. The NR method was identified as the fourth best method for estimating missing streamflow data and the best among conventional methods, as shown in Table 3 to Table 5. The result was similar to past studies by Ismail *et al.,* [17] and Hamzah *et al.,* [10], in which the NR method was determined to be the most accurate estimation method, especially when compared to other conventional methods. NR produced results that were superior and had RMSE and MAE values that were significantly lower. NR was determined by the ratio average of the accessible data at the target station to those at the neighboring station. The data collected at neighboring stations were usually identical to the data collected at the target station because neighboring stations commonly possessed the same characteristics and properties as the target station. Therefore, NR produces results that are very close to the target station. Subsequently, the results will be more accurate.

Table 3
Overall ranking for missing 5% estimation methods

Overall Ranking Missing 5%								
Method	KNN	CART	NB	LI	NR	IDW	AM	CC
RMSE	7	7	8	4	5	2	2	3
MAE	7	7	8	4	5	2	2	3
R	4	8	8	3	5	8	2	1
Total	18	22	**24**	11	15	12	6	7

The bolded value represents the method that obtained the highest accuracy.

Table 4
Overall ranking for missing 10% estimation methods

Overall Ranking Missing 10%								
Method	KNN	CART	NB	LI	NR	IDW	AM	CC
RMSE	6	7	8	4	5	3	2	1
MAE	6	7	8	4	5	3	2	1
R	8	3	8	1	5	5	4	1
Total	20	17	**24**	9	15	11	8	3

The bolded value represents the method that obtained the highest accuracy.

Table 5
Overall ranking for missing 15% estimation methods

Overall Ranking Missing 15%								
Method	KNN	CART	NB	LI	NR	IDW	AM	CC
RMSE	6	8	7	4	6	1	3	2
MAE	6	8	7	4	6	1	3	2
R	3	3	6	1	2	8	6	5
Total	15	19	**20**	9	14	10	12	9

The bolded value represents the method that obtained the highest accuracy.

4.1 Homogeneity Test

After estimating the missing data, homogeneity testing of the time series data must be performed to ensure the quality of the data. This test is essential as it can identify changes in behavior over the length of a time series. The Pettitt test, the BR test, the VNR test, and the SNHT were used to check the homogeneity of the streamflow data. If the null hypothesis had a p-value greater than 0.05 at the significance level of 5 percent, the streamflow data were considered to be homogeneous.

The homogeneity test results for the daily time series can be found in Table 6 and Table 7, respectively. The results obtained from each of the four unique tests fall into one of three categories: useful, doubtful, or suspect [32]. The categorization was determined to be "useful" when either one or none of the p-values in the homogeneity tests were lower than 0.05. It was determined to be "doubtful" when the series rejected two out of four homogeneity tests, and it was determined to be "suspect" when three or all homogeneity tests were rejected.

Based on the daily time series, it was discovered that four out of eleven streamflow stations were categorized as useful, whereas seven out of eleven streamflow stations were categorized as doubtful. The seven doubtful stations indicate that there may be inhomogeneity in the data, and these stations can still be used for further analysis as long as the data are carefully processed [19]. Furthermore, there were not even one of the suspect stations available in the daily time series. Although, the data should not contain any suspect stations as the data will be inhomogeneous.

In addition, only one homogeneity test was found to be rejected by the majority of the stations, the VNR test. Since the VNR test assumes that the time series is considered a random distribution, thus the results showed that the streamflow data were not randomly distributed.

Table 6
Summary of homogeneity tests for daily time series

Station Code	p-value Pettitt	SNHT	BRT	VNR test	Results
1737451	0.959	**0.002**	0.970	**< 0.0001**	Doubtful
2224432	**0.040**	0.099	0.068	**0.010**	Doubtful
5721442	0.265	**0.000**	0.581	**< 0.0001**	Doubtful
2723401	0.967	**< 0.0001**	0.965	**< 0.0001**	Doubtful
4907422	0.177	0.277	0.234	**0.000**	Useful
6503401	0.967	0.100	0.995	**0.001**	Useful
5606410	0.111	0.158	0.117	**< 0.0001**	Useful
3519426	0.908	**< 0.0001**	0.979	**< 0.0001**	Doubtful
4930401	0.057	**0.012**	0.133	**< 0.0001**	Doubtful
3116430	0.099	0.307	0.194	**0.008**	Useful
3813411	0.967	**0.000**	0.959	**< 0.0001**	Doubtful

The bolded value represents the heterogeneous time series.

Table 7
Summary of homogeneity tests result for daily time series

Class	Daily
Useful	4
Doubtful	7
Suspect	0

4. Conclusions

In conclusion, the estimation of missing streamflow data using various estimation methods was successfully performed in Peninsular Malaysia. A total of eleven streamflow stations' data were selected as input data from the DID, with the main objective of determining the most suitable method in estimating missing streamflow data. This was achieved by comparing the performance of the estimation methods based on measures such as RMSE, MAE, and R. The machine learning methods were the KNN, CART, and NB, while the conventional methods were the AM, CC, LI, NR, and IDW. All the estimation methods were used to estimate the missing streamflow data to determine the best method. Therefore, it was found that NB was the most appropriate method according to the overall ranking for estimation methods in terms of the lowest values of RMSE and MAE, and the highest value of R. The estimated data from the NB methods were much closer to the actual data, resulting in higher accuracy.

The findings of estimation of missing streamflow data using various estimation methods are essential to help engineers be aware of how streamflow data sets can improve their planning or urban design. As soon as the entire data set is available, the precision of the analyses performed by the engineers will also increase, resulting in more exact designs.

In future studies, it is suggested to increase the number of neighboring stations. This is because the more data available from neighboring stations, the more accurate and reliable results will be produced by the estimation method. Other recommendations from this study are that estimation methods should be tested with different streamflow station datasets. Finally, other machine learning models can also be applied to the study to obtain the best method with the highest accuracy.

Contribution of Study

This study contributes to engineers and hydrologists who rely on streamflow data by improving their urban planning. Recommendation for future research were provided as well, such as testing

estimation methods with different streamflow station datasets and applying other machine learning models to obtain the best method with the highest accuracy.

Acknowledgement
The authors would like to express their gratitude to the Ministry of Higher Education Malaysia for funding this research project through Fundamental Research Grant Scheme (FRGS) with project code: FRGS/1/2021/TK0/UCSI/03/3. The authors would like to acknowledge the sincere appreciation towards the financial support from Centre of Excellence for Research, Value Innovation and Entrepreneurship (CERVIE), UCSI University.

References

[1] Ahmed, Khalid, Saba Shahid, Tariq Ismail, Naseem Nawaz, and X-jun Wang. "Absolute homogeneity assessment of precipitation time series in an arid region of Pakistan." Atmósfera 31, no. 3 (2018): 301-316. https://doi.org/10.20937/atm.2018.31.03.06.

[2] Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018). Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes. 2018 International Conference on Orange Technologies (ICOT). https://doi.org/10.1109/icot.2018.8705796

[3] Chhabra, Gurjeet, Varun Vashisht, and Jyoti Ranjan. "A comparison of multiple imputation methods for data with missing values." Indian Journal of Science and Technology 10, no. 19 (2017): 1-7. https://doi.org/10.17485/ijst/2017/v10i19/110646.

[4] Chiu, Po-Chi, Azizul Selamat, Oldrich Krejcar, and Kuok Kwee Kuok. "Missing Rainfall Data Estimation Using Artificial Neural Network and Nearest Neighbor Imputation." In SoMeT, 132-143, September 2019. https://doi.org/10.3233/FAIA190044.

[5] Feng, X., Li, S., Yuan, C., Zeng, P., & Sun, Y. (2018). Prediction of Slope Stability using Naive Bayes Classifier. KSCE Journal of Civil Engineering, 22(3), 941–950. https://doi.org/10.1007/s12205-018-1337-3

[6] Galopo, J. P., & S. Perez, E. (2021). Predicting Student Program Completion Using Naïve Bayes Classification Algorithm. International Journal of Modern Education and Computer Science, 13(3), 57–67. https://doi.org/10.5815/ijmecs.2021.03.05

[7] Goh, E. H., Ng, J. L., Huang, Y. F., & Yong, S. L. S. "Performance of potential evapotranspiration models in Peninsular Malaysia." Journal of Water and Climate Change 12, no. 7 (2021): 3170-3186.

[8] Hamzah, Fauziyah Binti, Fatin Mohamad Hamzah, Siti Farhanah Mohd Razali, and Hasmahani Samad. "A comparison of multiple imputation methods for recovering missing data in Hydrological Studies." Civil Engineering Journal, 9, no. 3 (2021): 387-397. https://doi.org/10.28991/cej-2021-03091747.

[9] Hamzah, Fauziyah Binti, Fatin Mohamad Hamzah, Siti Farhanah Mohd Razali, and Ahmed El-Shafie. "Multiple imputations by chained equations for recovering missing daily streamflow observations: A case study of Langat River basin in Malaysia." Hydrological Sciences Journal 67, no. 1 (2022): 137-149. https://doi.org/10.1080/02626667.2021.2001471.

[10] Hamzah, Fauziyah Binti, Fatin Mohamad Hamzah, Siti Farhanah Mohd Razali, Osman Jaafar, and Norshuhada Abdul Jamil. "Imputation methods for recovering streamflow observation: A Methodological Review." Cogent Environmental Science 6, no. 1 (2020): 1745133. https://doi.org/10.1080/23311843.2020.1745133.

[11] Ismail, Wan Wardatulakmar, Wan Zin Wan Yusoff, and Wan Ibrahim. "Estimation of rainfall and stream flow missing data for Terengganu, Malaysia by using interpolation technique methods." Malays. J. Fundam. Appl. Sci 13 (2017): 214-218. https://doi.org/10.11113/mjfas.v13n3.578.

[12] Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2020). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing, 77(5), 5198–5219. https://doi.org/10.1007/s11227-020-03481-x

[13] Jaiswal, R.K., A.K. Lohani, and H.L. Tiwari. "Statistical Analysis for Change Detection and trend assessment in climatological parameters." Environmental Processes 2, no. 4 (2015): 729-749. https://doi.org/10.1007/s40710-015-0105-3.

[14] Kamwaga, S., Mulungu, D. M. M., & Valimba, P. (2018). Assessment of empirical and regression methods for infilling missing streamflow data in Little Ruaha Catchment Tanzania. Physics and Chemistry of the Earth, Parts A/B/C, 106, 17–28. https://doi.org/10.1016/j.pce.2018.05.008

[15] Lai, W. Y. (2019). A Study on Sequential K-Nearest Neighbor (SKNN) Imputation for Treating Missing Rainfall Data. International Journal of Advanced Trends in Computer Science and Engineering, 8(3), 363–368. https://doi.org/10.30534/ijatcse/2019/05832019

[16] Li, L., Liu, Y., Wei, T., & Li, X. (2020). Exploring Inter-Sensor Correlation for Missing Data Estimation. IECON 2020: The 46th Annual Conference of the IEEE Industrial Electronics Society. Retrieved from https://doi.org/10.1109/IECON43393.2020.9254904

[17] Low, C. Y., M. I. Solihin, C. K. Ang, W. H. Lim, and G. Hayder. "Towards Estimating Rainfall Using Cellular Phone Signal." In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, November 2022, pp. 1-6. IEEE.

[18] Meher, J. (2019). Missing Discharge Data Filling with Artificial Neural Network. i-Manager's Journal on Civil Engineering, 9(2), 24. https://doi.org/10.26634/jce.9.2.14657

[19] Ng, J. L., Tiang, S. K., Huang, Y. F., Noh, N. I. F. M., & Al-Mansob, R. A. (2021). Analysis of annual maximum and partial duration rainfall series. IOP Conference Series: Earth and Environmental Science, 646(1), 012039.

[20] Noor, N. M., Al Bakri Abdullah, M. M., Yahaya, A. S., & Ramli, N. A. (2014). Comparison of Linear Interpolation Method and Mean Method to Replace the Missing Values in Environmental Data Set. Materials Science Forum, 803, 278–281. https://doi.org/10.4028/www.scientific.net/msf.803.278

[21] Nor, S. M. C. M., Shaharudin, S. M., Ismail, S., Zainuddin, N. H., & Tan, M. L. (2020). A Comparative Study of Different Imputation Methods for Daily Rainfall Data in East-Coast Peninsular Malaysia. Bulletin of Electrical Engineering and Informatics, 9(2), 635-643. https://doi.org/10.11591/eei.v9i2.2090

[22] Norazizi, N. A., & Deni, S. M. (2019). Comparison of Artificial Neural Network (ANN) and Other Imputation Methods in Estimating Missing Rainfall Data at Kuantan Station. Communications in Computer and Information Science, 298-306. https://doi.org/10.1007/978-981-15-0399-3_24

[23] Panda, B. S., Misra, A., & Gantayat, S. S. (2019). Methods and Concepts of Data Mining Techniques to Impute Missing Data Information. Far East Journal of Electronics and Communications, 20(1), 41-54. https://doi.org/10.17654/ec020010041

[24] Roslin, N., Mustapha, A., Samsudin, N., & Razali, N. (2018). A Bayesian Approach to Prediction of Flood Risks. Semantic Scholar. Retrieved from https://doi.org/10.14419/ijet.v7i4.38.27750

[25] Sadek, F. M., Solihin, M. I., Heltha, F., Hong, L. W., & Rizon, M. "A Comparison of Machine Learning and Deep Learning in Hyperspectral Image Classification." In Enabling Industry 4.0 through Advances in Mechatronics: Selected Articles from iM3F 2021, Malaysia, edited by A. M. L. Tan, 221-235. Singapore: Springer Nature Singapore, 2022.

[26] Santosa, B., Legono, D., & Suharyanto. (2014). Prediction of Missing Streamflow Data Using the Principle of Information Entropy. Civil Engineering Dimension, 16(1), 40-45. https://doi.org/10.9744/ced.16.1.40-45

[27] Saplioglu, K., & Kucukerdem, T. S. (2018). Estimation of Missing Streamflow Data Using ANFIS Models and Determination of the Number of Datasets for ANFIS: The Case of Yeşilırmak River. Applied Ecology and Environmental Research, 16(3), 3583-3594. https://doi.org/10.15666/aeer/1603_35833594

[28] Sattari, M. T., Falsafian, K., Irvem, A., S, S., & Qasem, S. N. (2020). Potential of kernel and tree-based machine-learning models for estimating missing data of rainfall. Engineering Applications of Computational Fluid Mechanics, 14(1), 1078–1094. https://doi.org/10.1080/19942060.2020.1803971

[29] Shaharudin, S. M., Sri Andayani, Kismiantini, Nikenasih Binatari, Agusta Kurniawan5, Nurul Hila Zainuddin, & Muhammad Afdal Ahmad Basri. (2020). Imputation methods for addressing missing data of monthly rainfall in Yogyakarta, Indonesia. International Journal of Advanced Trends in Computer Science and Engineering, 9(1.4), 646–651. https://doi.org/10.30534/ijatcse/2020/9091.4

[30] Souza, G. R. de, Bello, I. P., Corrêa, F. V., & Oliveira, L. F. C. de. (2020). Artificial Neural Networks for filling missing streamflow data in Rio do Carmo Basin, Minas Gerais, Brazil. Brazilian Archives of Biology and Technology. Retrieved from https://doi.org/10.1590/1678-4324-2020180522

[31] Tan, Y. X., NG, J. L., & Huang, Y. F. (2020). Estimation of missing daily rainfall during monsoon seasons for tropical region: A comparison between Ann and conventional methods. Carpathian Journal of Earth and Environmental Sciences, 15(1), 103–112. https://doi.org/10.26471/cjees/2020/015/113

[32] Tiu, E. S. K., Huang, Y. F., Ng, J. L., AlDahoul, N., Ahmed, A. N., & Elshafie, A. (2021). An evaluation of various data pre-processing techniques with machine learning models for water level prediction. Natural Hazards, 110(1), 121–153.

[34] Turhan, E. (2021). A comparative evaluation of the use of artificial neural networks for modeling the rainfall–runoff relationship in water resources management. Journal of Ecological Engineering, 22(5), 166–178. https://doi.org/10.12911/22998993/135775

[35] Vasker Sharma. (2021). Imputing Missing Data in Hydrology using Machine Learning Models. International Journal of Engineering Research And, V10(01). https://doi.org/10.17577/ijertv10is010011

[36] Wiche, G. J., & Holmes, R. R. (2016). Streamflow Data. Flood Forecasting. Retrieved from https://doi.org/10.1016/B978-0-12-801884-2.00013-X

[37] Yilmaz, M. U., & Onoz, B. (2019). Evaluation of statistical methods for estimating missing daily streamflow data. Teknik Dergi. Retrieved from https://doi.org/10.18400/tekderg.421091

[39] Yilmaz, M. U., & Onoz, B. (2020). A comparative study of statistical methods for daily streamflow estimation at ungauged basins in Turkey. Water, 12(2), 459. https://doi.org/10.3390/w12020459

[40] Žliobaite, I., Hollmén, J., & Junninen, H. (2014). Regression models tolerant to massively missing data: A case study in solar-radiation nowcasting. Atmospheric Measurement Techniques, 7(12), 4387–4399. https://doi.org/10.5194/amt-7-4387-2014