International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I8P201
Volume 69 Issue 8, 1-10, August, 2021

Bidirectional Recurrence Neural Network Imputation For Recovering Missing Daily Streamflow Data

Fatimah Bibi Hamzah[#1,2], Firdaus Mohd Hamzah[#1], Siti Fatin Mohd Razali[#1], Juanita Zainudin[#2]

[#1]*Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi Selangor, Malaysia.*
[#2]*Faculty of Computing and Multimedia, Kolej Universiti Poly-Tech Mara Kuala Lumpur, Jalan 6/91, Taman Shamelin Perkasa, 56100 Kuala Lumpur, Malaysia.*

[1]bibi@kuptm.edu.my / bibi@gapps.kptm.edu.my, [2]fir@ukm.edu.my, [3] fatinrazali@ukm.edu.my

Abstract *— Missing value in hydrological research is common, and there is a growing interest to recover missing streamflow data as accurate information is required for various purposes. Due to missing data limitations, this study aims to evaluate the performance of the RNN-based method compared to the non-RNN based imputation methods to predict recurrence in a streamflow dataset. In this study, daily streamflow datasets from Malaysia's Langat River Basins were used. Following that, the datasets were fed into the Multiple Linear Regression (MLR) model. The validation of the best estimation methods was performed based on the estimation error, using methods such as Nash-Sutcliffe Efficiency Coefficient (CE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE). The findings revealed that the RNN-based method coupled with MLR (BRNN-MLR) outperformed all the approaches examined for filling missing values in streamflow datasets, with the highest CE value and lowest MAPE and RMSE value regardless of any missing data conditions.*

Keywords — *BRNN, imputation, MICE, Missing data, streamflow, MLR.*

I. INTRODUCTION

Streamflow is considered one of the primary variables used to describe the hydrologic function of water bodies. It provides critical records for water resource management and climate change surveillance, either as a signal of prior hydrological variability or as a contributor to water's future behavior[1]–[3]. Numerous research and operational applications that are water resource management and planning, anticipating extreme floods or droughts, forecasting streamflow, and analyzing climate irregularity, necessitate reliable time-series data [4]. Streamflow research, on the other hand, is difficult due to missing data limitations, and it is common to get erroneous data of questionable quality and for short periods [5], [6]. Furthermore, complete information is not always available due to incomplete observations, missing data, or outliers, which have a significant impact on hydrological research work [7], [8].

Statistical models play an important role in hydrological research when forcing inputs and model parameters are required [9], [10]. Nonetheless, these statistical models used for tracking purposes in environmental studies are heavily reliant on automatic data acquisition systems that necessitate a large number of physical sensing devices [11] that are vulnerable to damage due to extreme environmental conditions, physical destruction, and battery drain [4]. In turn, this fails control stations, matchless measurements, or manual data entry procedures that can cause errors, inaccurate calibration processes and/ or data damage caused by malfunctioned storing machinery, extended hydrometric data construction, and organization besides increased gaps in a dataset [8], [12]–[15].

In general, missing value(s) in time-series data represent a loss of information, which can lead to incorrect summary data explanations or untrustworthy scientific analysis. As a result, reconstruction and missing data treatment should be prioritized during the data preparation procedure. However, the selection of missing data handling techniques is dependent on the missing data trend and mechanism [16], which affects the statistical output.

There are three types of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [17]. The missing data mechanism is referred to as MCAR, and it is completely independent of the values of any variables in a dataset, whether they are missing or observed. Meanwhile, MAR is the root of missing data that is not associated with missing values but may be associated with observed values of other variables.MNAR observations are not missing at random, nor are MCAR or MAR. Missing value in the streamflow study is determined as MCAR due to the existence of missingness in the streamflow data of an area not influenced by the data in that area or any other area. [18] described streamflow data imputation using the MAR assumption. According to [19], the MCAR and MAR imputation for time-series studies are nearly identical.

Various infilling approaches for the missing value of the engineering database have been proposed and debated in the literature, ranging from the simple traditional statistical method to advanced techniques. Of all the available techniques, the mean imputation and deletion technique is

the most commonly used technique to reconstruct the missing meteorological and hydrological data [5], [8], [15], [20], [21]. Recently, [22] presented six missing data imputations to reconstruct missing rainfall data: hot-deck, k-nearest neighbors (KNNI), weighted k-nearest-neighbors (WKNNI), multiple imputations (MI), linear regression (LR), and simple average method (SAM). It was concluded that KNNI, WKNNI, and hot-deck methods resulted in good missing data filling regardless of any percentage of missing data. In another research by [23], missing rainfall and temperature data were reconstructed using multiple imputations by chained equations (MICE) approaches due to its simplicity where normal data distribution was not assumed and data missing at random (MAR) was assumed. Earlier, [24] investigated several infilling methods in environmental pollution datasets, ranging from simple ones like mean imputation and last and next observation carried forward/backward to advanced MICE approaches. Besides that, a new single imputation method called the Site-Dependent Effect was introduced by the authors. Another imputation method recognized as the random forest method for water resources was recently reviewed by [25]. Despite various imputation methods that were introduced, the simple traditional method is still applied. [26] used multiple regression, the random forest method, and machine learning approaches to simulate monthly streamflow in five highland rivers that are highly seasonal, and they discovered that the random forest method was the best imputation method for their study.

Many recent studies [27]–[31] have demonstrated improved prediction performance by using recurrence neural networks (RNN) approaches for classification or prediction models in hydrology and related fields. The recurrent components are coupled with the classification or regression component, which improves statistical analysis accuracysignificantly[32]. [31]introduced the Short-Term Long Memory (LSTM) network model for flood forecasting, which used daily discharge and rainfall as input data, and the model's predictive ability was claimed to be impressive. The finding is in line with the research by [27], who used RNN to forecast river flow. Previously, [29]employed RNN to simulate solid transport in sewer systems during storm events,[30]created a recurrent sigma-P neural network model for Hong Kong rainfall forecasting. [28] claimed that RNN outperformed other ANN architectures for predicting watershed runoff.

There has been no research on the reconstruction of missing streamflow data using effective RNN approaches to date. As a result, the goal of this study is to assess missing daily streamflow data using an RNN-based (bidirectional recurrence neural network (BRNN)) approach. The goal of this study is twofold: first, to reconstruct the missing flow data from the Langat River basin using RNN-based (BRNN) rather than non-RNN based methods such as arithmetic average (AA), hot-deck(HD), multivariate imputation by chain equations (MICE), k-nearest neighbor (KNN) and random forest (RF). Second, the performance of imputation methods in conjunction with the Multiple Linear Regression (MLR) model will be evaluated in forecasting future daily streamflow values. This study's findings are expected to contribute to the discovery of the best and finest approaches for the data imputation method, which allows for the reconstruction of complete daily streamflow data sets.

II. AREA OF STUDY

This study was carried out at Langat basin (Fig. 1), which is located to the south of Selangor and north of Negeri Sembilan, specifically between the latitudes of 2° 40'M 152" N to 3° 16'M 15" N and longitudes of 101° 19'M 20" E to 102° 1'M 10" E with a range of 2,394.38 km^2. This river basin, Malaysia's most urbanized river basin, is thought to compensate for the benefits of spill-over development from Klang Valley[33], [34]. It is an important raw water resource for drinking water and other activities such as recreation, industrial uses, fishing, and agriculture[35]. Over the last four decades, these water sources have served roughly half of Selangor's population, or approximately 1.2 million people within the basin, and have served as a source of hydropower and flood control[36]–[38]. The Langat basin was chosen as one of the major areas for economic growth in Selangor because it contains Kuala Lumpur International Airport, West Port at Klang, the Multimedia Super Corridor (MSC), and Putrajaya[39].In terms of hydrometeorology, the Langat basin is influenced by two types of monsoons, namely the northeast and southwest monsoons, which occur from November to March and May to September, respectively [40], [41].

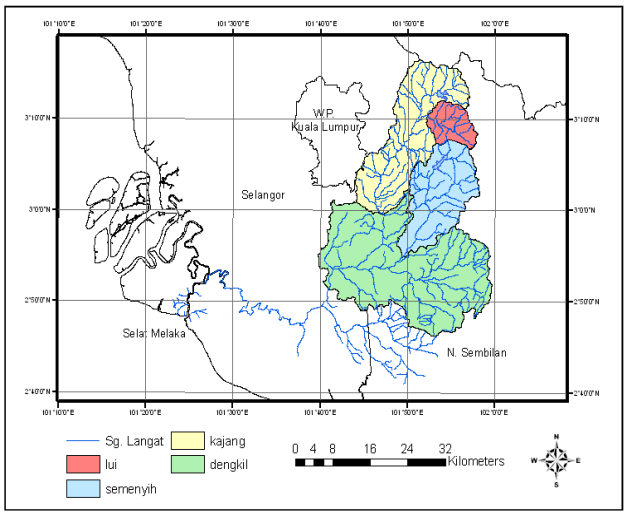

Fig. 1. Map of the Langat River Basin.

The Langat River Basin has four flow rate gauging stations: Dengkil and Kajang (Langat River), Kg. Rinching (Semenyih River), and Kg. Lui (Lui River). The characteristics of sub-basins associated with Langat Basin gauging stations are shown in Table 1.

TABLE 1. OVERVIEW OF THE LANGAT BASIN'S SUB-BASINS AND GAUGING STATIONS.

Sub-Basin	Hulu Langat	Hulu Langat	Semenyih	Lui
Station No.	2816441	2917401	2918401	3118445
Station name	Langat River at Dengkil	Langat River at Kajang	Semenyih River at Kg. Rinching	Lui River at Kg. Lui
River	Langat	Langat	Semenyih	Lui
Area (km^2)	1251.4	389.4	236.0	68.4
Location in the basin	Lower catchment	Middle catchment	Middle catchment	Upper catchment
Latitude	02o 59' 34"	02o 59' 40"	02o 54' 55"	03o 10' 25"
Longitude	101o 47' 13"	101o 47' 10"	101o 49' 25"	101o 52' 20"

The high-dimensional data used in this study were obtained from the Department of Irrigation and Drainage (DID), Ampang, Selangor, between 1978 and 2016. There were 12.4 percent missing values among the 56,980 data points. [42]defines moderate data as datasets with 10 to 25% missing values, whereas [43]asserts that if the percentage of missing data exceeds 10%, the statistical analysis will be skewed. To obtain an accurate outline of the streamflow patterns, a large number of time series observations were required [4]. Aside from that, since it is strongly related to sample size, the reliability of a frequency estimator of a long time series data is extremely useful in data analysis.

III. RESEARCH METHODOLOGY

This section is divided into two main subsections. Approaches for estimating missing data will be discussed in the first subsection. Meanwhile, assessing the performance of the methods used will be explained in the second subsection.

A. Imputation Methods

Both RNN- and non-RNN-based methods were compared in this study to determine which technique is best suited to impute missing values in streamflow data sets. For RNN-based methods, BRNN was introduced, while the non-RNN based imputation methods include: AA, HD, MICE, KNN, and RF. The datasets were then fed into the MLR model to determine the best methods for dealing with missing data when imputation values were combined with modeling.

***a) Bidirectional Recurrence Neural Network*:** RNN is a deep neural network architecture that uses feedback connections from its units to learn temporal patterns in sequential data. Mean imputation was used to initialize the missing values, and the values were updated using the feedback connection while the network was trained to learn the classification task. The missing values were modified in the previous iteration as a function of the missing input and the weighted sum of a set of recurrent links from the other units (hidden and missing) to the missing unit with a unit delay. The RNN approach used by[44]is depicted in Fig. 2.

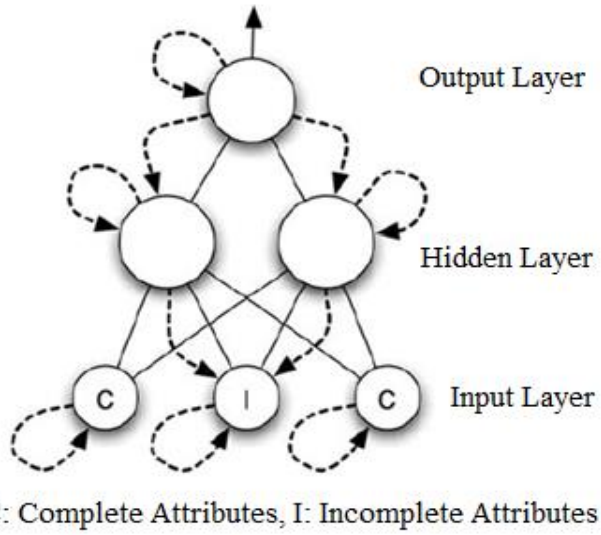

Fig. 2. The architecture of the recurrent networks.

In RNN, the hidden layer imputation (Fig. 2) is also referred to as a recurrent layer and a regression layer. A recurrent neural network was used to achieve the recurrent component, and a fully connected network was used to achieve the regression component. A typical recurrent network can be represented as follows:

$$h_t = \sigma(W_h h_{t-1} + U_h x_t + b_h) \quad (1)$$

where σ is the sigmoid function, W_h , U_h and b_h are parameters, and h_t Is the previous time-hidden step's state. In this research, since x_t may contain missing values, x_t was not used as the direct input as in Eq.1. Instead, a complement input x_t^c derived from the algorithm when x_t is missing was used. Fig.3 and 4summarise and illustrate the RNN algorithm procedure for filling multivariate missing data.

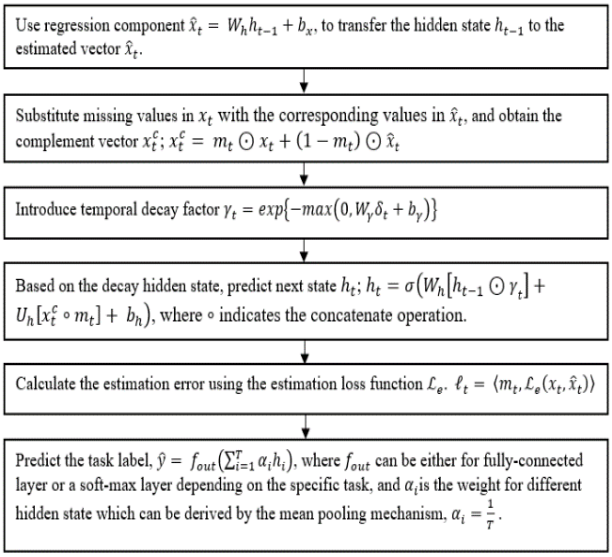

Fig. 3. The procedure of the RNN algorithm is suggested by [32].

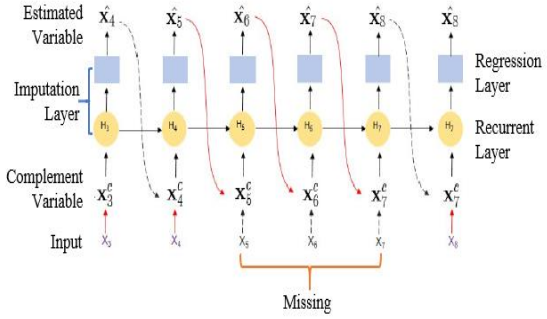

Fig. 4. The unidirectional imputation procedure of the RNN algorithm.

For time-series $X = \{x_1, x_2, x_3, \ldots x_{10}\}$, with missing values $x_5, x_6$and x_7, predicted on the recurrent dynamics, at each time-step, an estimation of $\hat{x}_t$ Can be obtained based on the previous step*t - 1*. The estimation error for*t = 1,2,3,4* can be calculated in the first four steps using the estimation loss function. However, since the true values were missing at *t = 5,6,7*, a delayed error for $\hat{x}_{t=5,6,7}$ at the 8th step can be obtained. Estimated missing value errors can be postponed until the next observation. In this case, the error of $\hat{x}_5$ can be postponed until x_8 is observed. Because of the error delay, the model converges slowly and inefficiently, as well as resulting in a bias exploding problem [45], in which the model was fed incorrect values from the previous sequential prediction, causing rapid amplification of most of the hydrology components.

Because of RNN's limitations, the BRNN, as shown in Fig. 5, was proposed as an improved version. The BRNN algorithm addresses the aforementioned issues by utilizing bidirectional recurrent dynamics for the given time series data. Each value in the time series dataset can be derived from the reverse direction by another fixed arbitrary function in addition to the forward direction. When using BRNN in the backward direction of the time series dataset, the estimation of $\hat{x}_4$inversely depends on the $\hat{x}_5$ *to* $\hat{x}_7$. As a result, the 5th step error is propagated not only to the 8th forward step (which is far from the current position) but also to the 4th backward step, which is closer.

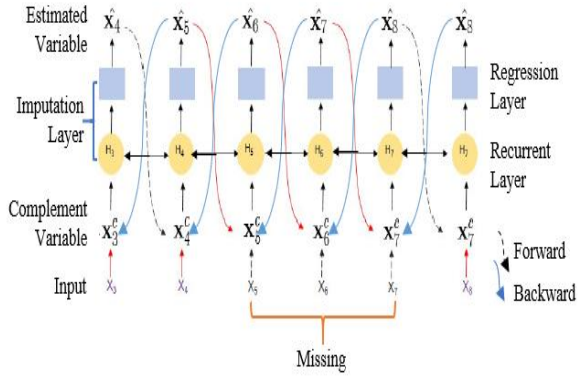

Fig. 5. The bidirectional imputation procedure of the RNN algorithm.

Formally, the BRNN algorithm performs the RNN as shown in steps 1 through 5 of Fig. 5 in both forward and backward directions. An estimation sequence of $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_T\}$ and the loss sequence $\{\ell_1, \ell_2, \ldots, \ell_T\}$ It can be derived for the forward direction. Similarly, another estimation sequence of $\{\hat{x}'_1, \hat{x}'_2, \ldots, \hat{x}'_T\}$ and loss sequence $\{\ell'_1, \ell'_2, \ldots, \ell'_T\}$ It can be obtained in the backward direction. The consistency loss was used to ensure that the forecast in any step was unchanging in both directions:

$$\ell_t^{cons} = Discrepancy(\hat{x}_t, \hat{x}'_T) \quad (2)$$

Where the mean squared error was used as the discrepancy in this study. The total estimation loss was calculated by adding the forward ℓ_t , the backward ℓ'_t And the consistency ℓ_t^{cons} Losses. In the tth step, the mean of $\hat{x}_t$ and $\hat{x}'_t$Is the final estimate.

b) Arithmetic Average Method (AA): The simplest way to reconstruct the missing value, according to[46], is to replace each missing value with the average of the observed values for the specific variable. The non-missing values were used to compute the variable's average, and the mean was used to reconstruct the missing value of the specific variable. In other words, for each missing value in the series $X^{(i)}, i = 1,2,\ldots,n$, the corresponding of the respective component is substituted. Eq. 3 was used to calculate the estimated missing value:

$$\hat{y}_t = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (3)$$

where $\hat{y}_t$ is the estimated value of the missing data at the *t* target station or date, $x_i$is the observed data (non-missing values) or the same date with different years, and *n* is the number of days or years.

c) Hot-Deck Method (HD): This method entails replacing missing data with values from the existing dataset or matching covariates [47]. Any data that is similar to the observed data can be detected by the process. This study's methodology was based on the last observation carried forward (LOCF). This technique is typically used to fill in missing longitudinal data at a specific 'visit' or at any given time for a specific entry [48]. The most recent obtainable value, i.e., from the most recent visit or time point, will be carried forward and used to replace the missing values. If multiple missing values occur in sequence, the monitoring may be used several times for infilling, and other values may not be used at all.

d) Multivariate Imputation by Chain Equations (MICE):[49]created the MICE algorithm, which uses the Markov Chain Monte Carlo method, with the state space containing all imputed values. Since relevant procedures are commonly included in standard statistical software packages [21], recent advances in computational power have enabled multiple imputations. MICE anticipates data loss at random (MAR). It is assumed that the observed data determines the likelihood of a missing variable.MICE functions with multiple regression models and conditionally models each missing value based on the observed (non-missing) values. In other words, it constructs a series of regression (or other appropriate) models based on its 'method' parameter to provide multiple

values in place of one missing value [50]. Each missing variable is classified as a dependent variable in this case, while the remaining data in the record is classified as an independent variable. Fig.6 depicts the procedure. It consists of three steps: data imputation, data analysis, and data pooling.

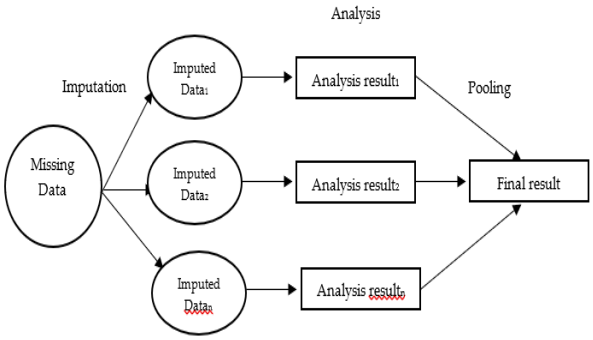

Fig. 6. MICE method mechanism.

MICE predicts missing data-first by using existing data from other variables. The imputed dataset is then generated by replacing missing values with predicted values. Iteratively, it generates multiply imputed datasets. Each dataset is then analyzed using standard statistical analysis techniques, with multiple analysis results provided.

The advantage of MICE is that the results are calculated after a few iterations, and in most cases, five iterations are sufficient [49], [51]. The MICE algorithm procedure for filling multivariate missing data is summarised as follows:

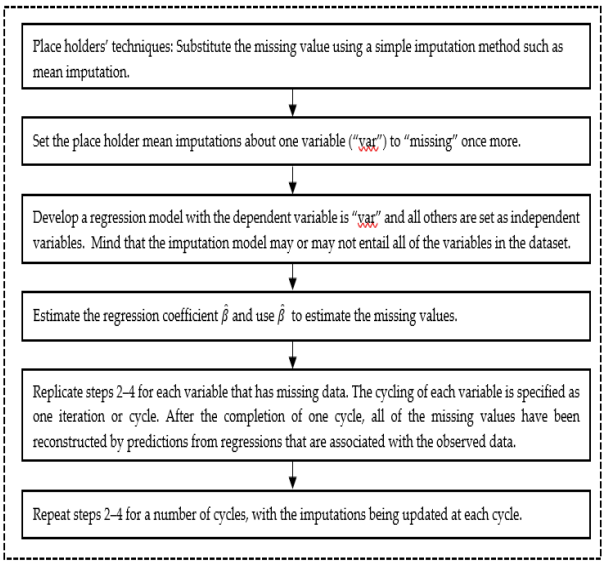

Fig. 7. The procedure of the MICE algorithm is suggested by [49], [51].

To summarise, MICE uses a divide and conquer approach to impute missing values in a data set's variables, focusing on one variable at a time. When one variable is chosen as the focus, MICE predicts missingness in that variable using all of the other variables in the data set (or a carefully chosen subset of these variables). The prediction is based on a regression model, the shape of which is determined by the nature of the focus variable.

e) K-nearest-neighbor imputation (KNN): KNN imputation is a machine learning technique that is also called distance function matching. It is a donor technique in which the donor is chosen by minimizing a fixed 'distance,' and their mean is used as an imputation estimate [47], [52]–[55]. KNN projections are based on the results of the k-neighbors closest to the missing value. This process computes a suitable distance measure, where the distance is determined by the auxiliary variables. The Euclidean Distance formula, as shown in Eq. 4, was used in this study.

$$D(x,y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad (4)$$

where x_i and y_i Are the query point and a case from the streamflow data sample, respectively.

The value of k- must be determined before making predictions using the KNN approach. [56] reported that a large k- value is more precise and relatively stable as it reduces overall noise, but there is no assurance. [57], on the other hand, argued that the lower the k- value, the better the estimation of missing observations. Because the rule of thumb is that k equals the square root of the number of points in the training dataset [58], a maximum number of k- was specified in this study to be no greater than the square root of the training dataset size, which produces better results than 1NN, which is usually assigned to the class of its nearest neighbor.

Predictions based on the KNN approach were made after the value of k- was chosen. The imputation process by the nearest neighbor for k-neighbors can be summarised as follows:

Let's say we have m observations on n covariates. $X = x_{is}$ denotes the corresponding $m \times n$ matrix, where x_{is}represents the i^{th} observation of the s^{th} variable. Let $O = o_{is}$ represent the corresponding $m \times n$ dummy matrix, which has the following entries:

$$o_{is} = \begin{cases} 1 & if\ x_{is}\ was\ observed \\ 0 & for\ missing\ value \end{cases} \qquad (5)$$

The L_q- metric for the data observed can be used to compute the distances between two observations x_i and x_jWhich are represented in the data matrix by rows. The distance is then applied as:

$$d_q(x_i, x_j) = \left[d_{ij} \sum_{s=1}^{n} |x_{is} - x_{js}|^q 1(o_{is} = 1) I(o_{js} = 1) \right]^{1/q} \qquad (6)$$

where $d_{ij} = \sum_{s=1}^{p} 1(o_{is} = 1) I(o_{js} = 1)$ Represent the number of valid components in the computation of distances. Since parallel views conceptualize distances, nearest neighbors were used.

f) RandomForests (RF): Random forests is a machine learning method similar to bagging (bootstrap aggregation

of multiple regression trees) that incorporates randomization to decorrelate the trees [59]. It is an extension of classification and regression trees, predictive models that recursively subdivide the data based on the predictor variable values. A large number of trees are built using bootstrapped training samples. Each tree "votes," and this vote is used to classify each variable based on the majority (mode) vote overall trees [60]. Bootstrap training samples are drawn from the original dataset *n* times with replacement. This will result in a new training set with the same number of observations as the original training set that is unique to each tree. Fig.8 depicts the random forest structure, which demonstrates the power of combining multiple decision trees into a single model. Demand predictors and cut-points in the predictors used to divide the sample in RF models. The cut-points divided the sample into larger, homogeneous subsamples. The dividing operation is reiterated on both subsamples, allowing a series of splits to form a binary tree [61]. For regression problems, each node in the tree has a splitting rule determined by minimizing the relative error (RE), which is equivalent to minimizing the sums-of-squares of the split:

$$RE(d) = \sum_{l=0}^{L}(y_l - \bar{y}_L)^2 + \sum_{r=0}^{R}(y_r - \bar{y}_R)^2 \qquad (7)$$

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 $\bar{y}_L$ and $\bar{y}_R$. The decision rule *d*is a point in some estimator variable *x* that determines which branches go left and which go right. The partitioning rule that minimizes the RE is then used to derive the tree node. Fig.8 depicts a random forest (RF) framework.

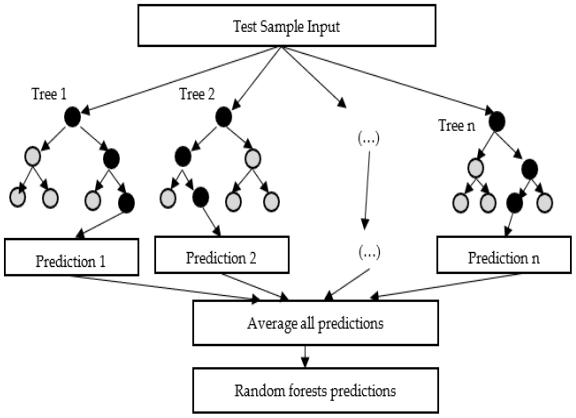

Fig. 8. Random forest structure

A random forest, according to [62], can handle mixed data types and, as a non-parametric method, is likely to produce non-linear (regression) and interactive effects. Assume that $X = (X_1, X_2, \dots, X_n)$ is a $m \times n$ - dimensional data matrix.The streamflow dataset in this study could be divided intwocataegories for arbitrary variables X_s with missing values at entries, $i_{mis}^{(s)} \subseteq \{1, \dots, m\}$, : $y_{obs}^{(s)}$denotes the observed values of the variable X_s, while $y_{mis}^{(s)}$ denotes the missing values of the variable X_s.

To begin, mean or other imputation methods are used to generate the first guess for the missing values in *X*. The variables $X_s, s = 1, \dots, p$ are then ordered by the number of missing values, starting with the smallest. Missing values are reconstructed for each variableX_s by first fitting an RF with a response$y_{obs}^{(s)}$and predictors $x_{obs}^{(s)}$and then predicting missing values $y_{mis}^{(s)}$ by applying the trained RF to $y_{mis}^{(s)}$. The imputation process is repeated until a stopping criterion is met.

g) Multiple Linear Regression (MLR): Following the replacement of all missing values with various techniques, the entire datasets are analyzed using MLR to determine the finest approaches for dealing with missing data in daily streamflow datasets. Regression analysis is a statistical technique that examines the relationship between at least two quantitative variables and their expected variables [63]. The MLR model is a popular statistical method in many disciplines, including hydrological data [64], [65]. The MLR model parameter is expressed as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i(\beta), \quad i = 1, \dots, N \qquad (8)$$

where Y_i is the response variable's value, $\beta_0, \beta_1, \beta_2 and \beta_k$are unknown constants, X_y is the predictor variable's value, and ε_i Is the random error.

B. Estimation Method Performance

Four performance criteria were used in this study. To evaluate the imputation methods, the CE, RMSE, and MAPE were calculated. The error is calculated as the variation between the estimated and observed values. The CE is a well-known index used to weigh the predictive power of hydrological models. The most effective performance models aim for a CE value of one (1). The RMSE and MAPE, a standard statistical metric used to evaluate model performance in meteorology, air quality, and climate research studies, are also used in this study to assess model performance. The RMSE statistic, which is a measure of the difference between predicted and observed values, provides information on short-term efficiency. Another useful measure that is popular in model evaluations is the MAPE. MAPE is a measure of the average difference between predicted and observed values, and it can provide insight into the models' long-term performance. The lower the RMSE and MAPE values, the better the model's long-term performance. The following equations are used to compute these statistics:

$$CE = 1 - \frac{\sum_{i=1}^{n}(x_i - \tilde{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} \qquad (9)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - \hat{x}_i|}{x_i} \qquad (10)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{x}_i - x_i)^2}{n}} \qquad (11)$$

where x_i is the observed streamflow data, $\hat{x}_i$ Is the estimated value, *n* is the sample size, and *k*is the number of independent variables in the regression equation over daily Langat River Basin streamflow datasets.

IV. RESULTS AND DISCUSSION

As stated earlier in this paper, the purpose of this research is to determine which methods, between RNN and non-RNN, are considered to be the best imputation methods for recovering missing streamflow data. The models were first tested for all four sub-basins using data from 1978 to 2016. The results were then computed as a mean of the results of each imputation method. Each method's performance was evaluated using CE, RMSE, and MAPE. When the difference between the estimated and observed values is small, RMSE and MAPE will produce the smallest value. Meanwhile, CE values can range from -∞ to 1, with values greater than 0.5 considered acceptable. The technique with the highest CE value, as well as the lowest RMSE and MAPE value, will be selected as the best fit method. Table 2 provides the results of the overall performance of the methods in reconstruction data from 1978 to 2016.

The results showed that the BRNN method produced the smallest RMSE with the highest CE. CE values, on the other hand, revealed that all imputation methods yield acceptable results, with values close to one. Based on Table 2, BRNN performed the best. Meanwhile, among the other methods, AA was the poorest imputation method for daily streamflow data in Malaysia's Langat River Basin, with the lowest CE and highest RMSE. Table 2 also revealed that the KNN imputation method has a lower RMSE and higher CE values than the other four methods, putting the model on par with BRNN.

TABLE 2. RMSE AND CE VALUES FOR SIX IMPUTATION METHODS ON AVERAGE.

Method	RMSE	CE
AA	38.091	0.358
HD	30.950	0.542
BRNN	**27.716**	**0.790**
MICE	28.588	0.670
KNN	28.096	0.767
RF	28.247	0.727

Notes: A better model is in bold.

After the missing values have been filled in, the MLR model will be used to analyze the entire dataset in this study. When imputation values were combined with modeling, the MLR model was used to determine the best methods for dealing with missing data. MAPE and RMSE were used to assess the performance of imputation methods when combined with the MLR model.

The RMSE and MAPE values for each statistical approach for imputing missing values of daily streamflow data in Malaysia's Langat River Basin coupled with an MLR model are shown in Table 3. As a result, when combined with a regression model, the final results showed that BRNN-MLR is the best statistical method for imputing missing values in daily streamflow data with the lowest RMSE and MAPE of 20.789 and 0.261, respectively when compared to other approaches.

TABLE 3. THE RESULTS FOR MLR WHEN COMBINED WITH IMPUTATION METHODS.

Method	RMSE	MAPE
AA-MLR	31.024	0.601
HD-MLR	39.451	0.953
BRNN-MLR	**20.789**	**0.261**
MICE-MLR	30.387	0.534
KNN-MLR	23.784	0.397
RF-MLR	29.937	0.454

Notes: A better model is in bold

In conclusion, the BRNN-MLR method presented the best performance, whereas HD-MLR with the LOCF approach had the poorest among them. This could be due to the reason that the LOCF technique has no shift from one visit to the other, and this is sensible for MCAR data sets. On the other hand, although the AA method is known for its simplicity of methodology, the method does not make use of the primary correlation structure of the data and thus performed poorly. This proved that a popular method does not necessarily mean the best method. Meanwhile, MICE which is often regarded as a conservative and safe approach to handle missing data, also underestimated the variance. MICE is based on a more complex algorithm, and its performance is related to the dataset size of the dataset. Performance of MICE is fast and efficient when small datasets were used while the performance decreases with large datasets and results in time-intensive analysis.

RF, on the other hand, outperforms MICE, AA, and HD. However, as stated by [25], no algorithm is perfect, and thus RF should not be used to solve all types of problems. Regardless of its flexibility and interpretability, it is necessary to fully understand how RF works to set and cross-validate appropriate tuning parameters such as tree depth or split number [60]. With imbalanced data, the RF cannot extrapolate beyond the training range and cannot fit the model adequately.

In comparison to the AA, HD, MICE, and RF methods, the KNN is a machine learning approach that performed slightly well but turned out to be time-intensive once applied to the large dataset. This finding was in agreement with a recent finding that the KNN model's performance in monthly streamflow prediction was highly satisfying and capable of producing more accurate predictions [66], [67]. In contrast, [68] claimed that advanced linear methods produced far superior results when compared to more traditional methods such as the KNN. The KNN algorithm will search the data for the k- closest neighbors to the new instance and set the predicted class label as the most common label among the k- closest neighboring points to predict the label of a new instance[69]. The algorithm must compute the distance and sort all of the training data at each prediction, which takes time if there are a large

number of training examples. Changing the k- the value may also result in a different predicted class label [70].

In contrast, all findings show that the proposed RNN technique outperforms the non-RNN approaches investigated. When compared to other methods, BRNN-MLR had significantly lower RMSE and MAPE. The obvious reason for this is that the BRNN duplicates the RNN processing chain, allowing inputs to be processed both forward and backward in time. This enables a BRNN to consider future context, as well as BRNN, add a random error term in which the same value of independent variables without this term will result in the same response, which is not true in reality. Furthermore, when the error proportion is mirrored by the missing data proportion, the error derived from the BRNN technique is comparatively low when compared to that derived from the non-RNN techniques. These simulations demonstrate unequivocally that the BRNN technique is the most effective missing data imputation method for reconstructing missing streamflow data.

Finally, for visual inspection, the observed and predicted values for all models were plotted. Fig.9 depicts the results of six imputation methods used to replace 7124 missing daily streamflow data points in Malaysia's Langat River Basin. Fig.9 depicts how the imputed values of daily streamflow data from all six methods followed similar trends. All models, for example, reacted to streamflow events with similar magnitude peaks and times.

Findings from this study showed that the RNN-based method significantly outperformed other approaches. With the lowest MAPE and RMSE and the highest CE value, the BRNN outperformed the other methods tested. This shows the error derived from the BRNN technique was lower than that compared to the AA, HD, MICE, KNN, and RF techniques since the error rate was mirrored by the missing data rate. Furthermore, delayed gradients for missing values in both the forward and backward directions improve the accuracy of missing value imputation [32].

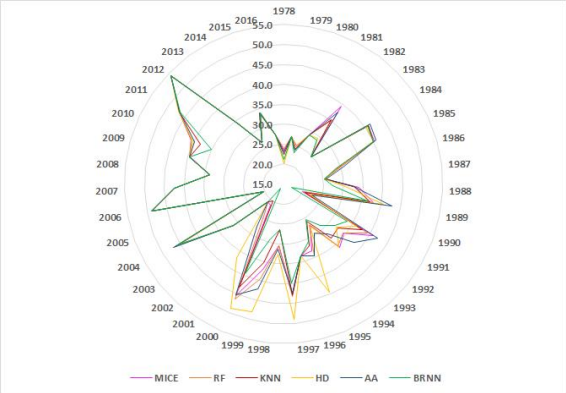

Fig.9. MICE, RF, KNN, HD, AA, and BRNN data imputation results for 7124 missing streamflow data.

Conclusively, these simulations demonstrated that the BRNN technique coupled with MLR is the best suited missing data imputation method for reconstructing missing streamflow data.

V. CONCLUSION

Missing data is a common limitation in hydrological research and usually results in misinterpretation of statistical output and modeling approaches in hydrology. Therefore, there is a need for method performance evaluation to reduce the impact of missing data in hydrological research. Several techniques were suggested in the literature to manage missing data. However, a suitable approach to be used as the missing data trend and mechanism are still unclear. Due to convenience, researchers generally discard observations with missing data or substitute with a naive method such as the mean of all other observations and hot-deck (last/ next observation carry forward/ backward), although these methods have significant statistical shortcomings.

More sophisticated approaches, such as the RNN-based method, the BRNN, have been shown to improve the accuracy of missing value imputation and reduce the statistical issues caused by naive imputation approaches. According to the findings of this study, the BRNN performed consistently regardless of the presence of missing values. When compared to the AA, HD, MICE, KNN, and RF methods, all three performance indicators agreed that the BRNN method is among the best, with a higher CE and lower MAPE and RMSE values. This finding indicated that the BRNN method had the smallest difference between the reference model and the prediction model with missing data imputation. Thus, the BRNN is recommended for processing missing streamflow data. In conclusion, the outcome of this study significantly contributes to the accurate infilling of missing data in streamflow data sets.

REFERENCES

[1] P. Dobriyal, R. Badola, C. Tuboi, and S. Ainul., A review of methods for monitoring streamflow for sustainable water resource management, Appl. Water Sci., 7(2017) 2617–2628.

[2] M. H R and P. V N., Impact of Climatological Parameters on Reference Crop Evapotranspiration Using Multiple Linear Regression Analysis, Int. J. Civ. Eng., 2(1) (2015) 21–24.

[3] J. Odiero, B. T. I Ong'or, and M. N. Edward., Rainfall-Runoff Nexus in Mid-block of Yala Catchment, Int. J. Civ. Eng., 5(10) (2018) 6–16.

[4] P. Tencaliec, A. Favre, C. Prieur, and T. Mathevet., Reconstruction of missing daily streamflow data using dynamic regression models, Water Resour. Res. Am. Geophys. Union,51(12) (2015) 9447–9463.

[5] W. Norliyana, W. Ismail, W. Zawiah, W. Zin, and W. Ibrahim., Estimation of rainfall and streamflow missing data for Terengganu, Malaysia by using interpolation technique methods, Malaysian J. Fundam. Appl. Sci., 13(3) (2017) 213–217.

[6] M. N. Sediqi et al., Spatio-Temporal Pattern in the Changes in Availability and Sustainability of Water Resources in Afghanistan, Sustainability, 11(17) (2019) 5836.

[7] M. Kim, S. Baek, M. Ligaray, J. Pyo, M. Park, and K. H. Cho., Comparative studies of different imputation methods for recovering streamflow observation, Water (Switzerland), 7(12) (2015) 6847–6860.

[8] F. B. Hamzah, F. M. Hamzah, S. F. M. Razali, O. Jaafar, and N. A. Jamil., Imputation methods for recovering streamflow observation : A methodological review, Cogent Environ. Sci., 6(2020) 1745133.

[9] C. Deng and W. Wang., Runoff predicting and variation analysis in upper Ganjiang Basin under projected climate changes, Sustainability, 11(18) (2019) 5885.

[10] A. Temesgen., Quantifying model uncertainty to improve

streamflow prediction geba cathment, upper tekeze River basin, Ethiopia, Int. J. Hydrol., 6(3) (2020) 48–53.

[11] I. Žliobaite, J. Hollmén, and H. Junninen., Regression models tolerant to massively missing data: A case study in solar-radiation nowcasting, Atmos. Meas. Tech., 7(12) (2014) 4387–4399.

[12] Y. Gao., Dealing with missing data in hydrology - Data analysis of discharge and groundwater time-series in Northeast Germany, Freie Universität Berlin, Germany, (2017).

[13] C. A. Johnston., Development and evaluation of infilling methods for missing hydrologic and chemical watershed monitoring data, Virginia Polytechnic Institute and State University, (1999).

[14] P. Tencaliec, Developments in statistics applied to hydrometeorology: imputation of streamflow data and semiparametric precipitation modeling, Universite Grenoble Alpes, (2017).

[15] N. Ahmat Zainuri, A. Aziz Jemain, and N. Muda., A comparison of various imputation methods for missing values in air quality data, Sains Malaysiana, 44(3) (2015) 449–456.

[16] I. F. Kamaruzaman, W. Z. Wan Zin, and N. Mohd Ariff., A comparison of a method for treating missing daily rainfall data in Peninsular Malaysia, Malaysian J. Fundam. Appl. Sci., no. Special Issue on Some Advances in Industrial and Applied Mathematics, (2017) 375–380.

[17] R. J. A. Little and D. B. Rubin, Statistical analysis with missing data, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., (2002).

[18] M. K. Gill, T. Asefa, Y. Kaheil, and M. McKee, Effect of missing data on the performance of learning algorithms for hydrologic predictions: Implications to an imputation technique, Water Resour. Res., 43(7)(2007) 1–12.

[19] S. Moritz and T. Bartz-Beielstein., imputeTS: Time series missing value imputation in R,R J., 9(1) (2017) 207–218.

[20] G. Kabir, S. Tesfamariam, J. Hemsing, and R. Sadiq., Handling incomplete and missing data in water network database using imputation methods, Sustain. Resilient Infrastruct., 00(0) (2019) 1–13.

[21] Y. Gao, C. Merz, G. Lischeid, and M. Schneider., A review on missing hydrological data processing, Environ. Earth Sci., 77(2) (2018) 47.

[22] A. Aieb, K. Madani, M. Scarpa, B. Bonaccorso, and K. Lefsih., A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria, Heliyon5(27) (2019) 01247.

[23] W. Zvarevashe, S. Krishnannair, and V. Sivakumar, Analysis of rainfall and temperature data using ensemble empirical mode decomposition, Data Sci. J., 18(46) (2019) 1–9.

[24] A. Plaia and A. L. Bondì., Single imputation method of missing values in environmental pollution data sets, Atmos. Environ., 40(
8) (2006) 7316–7330.

[25] H. Tyralis, G. Papacharalampous, and A. Langousis., A brief review of random forests for water scientists and practitioners and their recent history in water resources, Water, 11(910) (2019) 1–37.

[26] J. E. Shortridge, S. D. Guikema, and B. F. Zaitchik., Machine learning methods for empirical streamflow simulation : a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds,(2016) 2611–2628.

[27] D. N. Kumar, K. S. Raju, and T. Sathish., River flow forecasting using recurrent neural networks, Water Resour. Manag., 18(2004) 143–161.

[28] J. Anmala, B. Zhang, and R. S. Govindaraju., Comparison of ANNs and empirical approaches for predicting watershed runoff, J. Water Resour. Plan. Manag., 126(3) (2000) 156–166.

[29] N. Gong, T. Denoeux, and J. L. Bertrand-Krajewski, Neural networks for solid transport modeling in sewer systems during storm events., Water Sci. Eng.,33(9) (1996). 85–92.

[30] T. W. S. Chow and S. Y. Cho., Development of a recurrent sigma-Pi neural network rainfall forecasting system in Hong Kong., Neural Comput. Appl., 5(2) (1997) 66–75.

[31] X. H. Le, H. V. Ho, G. Lee, and S. Jung., Application of long short-term memory (LSTM) neural network for flood forecasting., Water (Switzerland),11(2019) 1387.

[32] W. Cao, H. Zhou, D. Wang, Y. Li, J. Li, and L. Li., BRITS: Bidirectional recurrent imputation for time series, in 32nd International Conference on Neural Information Processing Systems, (2018) 6776–6786.

[33] M. . Noorazuan, R. Rainis, H. Juahir, and N. Jaafar., GIS Application in Evaluating Land Use-Land Cover change and its Impact on Hydrological Regime in Langat River Basin, Malaysia, Proc. Conf. MapAsia (Malaysia, Kuala Lumpur).,(2003).

[34] W. H. M. Wan Mohtar, S. A. Bassa Nawang, and M. N. S. Rahman., Statistical Analysis in Fluvial Sediments of Selangor Rivers: Downstream variation in grain size distribution, J. Kejuruter., S(1)(2017) 37–45.

[35] H. Juahir, T. M. Ekhwan, S. M. Zain, M. B. Mokhtar, Z. Jalaludin, and I. K. M. Jan., The Use of Chemometrics Analysis as a Cost-effective Tool in Sustainable Utilisation of Water Resources in the Langat River Catchment, Am. J. Agric. Environ. Sci.,4(2) (2008) 258–265.

[36] H. Juahir et al., Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques, Environ. Monit. Assess., 173(2011) 1–4 625–641.

[37] F. Mohamad Hamzah, S. H. Mohd Yusoff, and O. Jaafar., L-Moment-Based Frequency Analysis of High-Flow at Sungai Langat , Kajang , Selangor , Malaysia,Sains Malaysiana, 48(7) (2019) 1357–1366.

[38] Y. J. Puah, Y. F. Huang, K. C. Chua, and T. S. Lee., River catchment rainfall series analysis using additive Holt-Winters method, J. Earth Syst. Sci.,2(2016) 269–283.

[39] H. Juahir, S. M. Zain, A. Z. Aris, M. K. Yusof, M. A. A. Samah, and M. Bin Mokhtar., Hydrological trend analysis due to land-use changes at langat river basin, EnvironmentAsia, 3(2020) 20–31(2010).

[40] H. Memarian, S. K. Balasundram, J. B. Talib, A. M. Sood, and K. C. Abbaspour. Trend analysis of water discharge and sediment load during the past three decades of development in the Langat basin, Malaysia, Hydrol. Sci. J., 57(6) (2012) 1207–1222.

[41] H. H. Yang, O. Jaafar, E.-S. A., and S. M. S. A, Analysis of hydrological processes of Langat River sub-basins at Lui and Dengkil, Int. J. Phys. Sci., 6(32) (2011) 7390–7409.

[42] K. F. Widaman., Missing Data: What to do with or without them,Monogr. Soc. Res. Child Dev., 71(1) (2006) 210–211.

[43] D. A. Bennett., How can I deal with missing data in my study? Aust. N. Z. J. Public Health, 25(5) (2001) 464–469.

[44] Y. Bengio and F. Gingras., Recurrent neural networks for missing or asynchronous data, in 8th International Conference on Neural Information Processing Systems, (1995) 395–401.

[45] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer., Scheduled sampling for sequence prediction with recurrent neural networks, Adv. Neural Inf. Process. Syst., 9(2015) 1171–1179.

[46] T. Schneider., Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values., J. Clim., 14(5) (2001) 853–871.

[47] J. Chen and J. Shao., Jackknife variance estimation for nearest-neighbor imputation, J. Am. Stat. Assoc., 96(453) (2001) 260–269.

[48] T. Aljuaid and S. Sasi., Proper imputation techniques for missing values in data sets, in 2016 IEEE International Conference on Data Science and Engineering ICDSE, (2016).

[49] S. van Buuren and K. Groothuis-oudshoorn., mice : Multivariate Imputation by Chained Equations in R., J. Stat. Softw., 45(3) (2011) 1–67.

[50] S. Islam Khan and A. Sayed Md Latiful Hoque., SICE: an improved missing data imputation technique Background and related works,7(37)(2020).

[51] K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, Predicting time series with support vector machines. In: Gerstner W., Germond A., Hasler M., Nicoud JD. (eds) Artificial Neural Networks — ICANN'97. ICANN 1997. Lecture Notes in Computer Science, 1327 (1997) Springer Berlin Heidelberg.

[52] H. Lee and K. Kang., Interpolation of missing precipitation data using kernel estimations for hydrologic modeling., Adv. Meteorol., 12(5) (2015).

[53] B. Rajagopalan and U. Lall., A k-nearest-neighbor simulator for

daily precipitation and other variables, Water Resour. Res., 35(10) (1999) 3089–3101.

[54] S. Yakowitz and M. Karlsson., Nearest neighbor methods for time series, with application to rainfall/runoff prediction, in Advances in the Statistical Sciences: Stochastic Hydrology, Dordrecht: Springer Netherlands, (1987) 149–160.

[55] G. Kalton and L. Kish., Some efficient random imputation methods, Commun. Stat. - Theory Methods, 13(16) (1984) 1919–1939.

[56] Y. Yang., An evaluation of statistical approaches to text categorization, Inf. Retr. Boston., 1(1999) 1–2 69–90.

[57] A. Elshorbagy, S. P. Simonovic, and U. S. Panu., Estimation of missing streamflow data using principles of chaos theory, J. Hydrol., 255 (2002) 1–4 123–133, 2002.

[58] A. B. Hassanat, M. A. Abbadi, A. A. Alhasanat, and G. A. Altarawneh., Solving the problem of the K parameter in the KNN Classifier using an ensemble learning approach, Int. J. Comput. Sci. Inf. Secur., 12 (2014) 33–39.

[59] L. Breiman., Random forests,Mach. Learn., 45 (2001) 5–32.

[60] G. Chhabra, V. Vashisht, and J. Ranjan., A comparison of multiple imputation methods for data with missing values., Indian J. Sci. Technol.,10(19) (2017) 1–7.

[61] H. I. Erdal and O. Karakurt., Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms., J. Hydrol., 477 (2013) 119–128.

[62] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. New York: Wadsworth Publishing, (1984).

[63] A. F. Van Loon and G. Laaha., Hydrological drought severity explained by climate and catchment characteristics, J. Hydrol., 526 (2015) 3–14.

[64] A. M. Carey and G. B. Paige., Ecological Site-Scale Hydrologic Response in a Semiarid Rangeland Watershed, Rangel. Ecol. Manag., 69(6) (2016) 481–490.

[65] L. Campozano, E. Sánchez, Á. Avilés, and E. Samaniego., Evaluation of infilling methods for time series of daily precipitation and temperature: The case of the Ecuadorian Andes., Maskana, 5(1) (2014) 99–115.

[66] A. K. Poul, M. Shourian, and H. Ebrahimi., A comparative study of MLR, KNN, ANN and ANFIS models with wavelet transform in monthly streamflow prediction, Water Resour. Manag., 33(2019) 2907–2923.

[67] S. C. Worland, W. H. Farmer, and J. E. Kiang., Improving predictions of hydrological low-flow indices in ungaged basins using machine learning,Environ. Model. Softw., 101 (2018) 169–182.

[68] J. J. Miró, V. Caselles, and M. J. Estrela., Multiple imputations of rainfall missing data in the Iberian Mediterranean context, Atmos. Res., 197 (2017) 2313–330.

[69] C. H. Cheng and S. J. Syu., Improving area positioning in ZigBee sensor networks using neural network algorithm, Microsyst. Technol., 27(4) (2021) 1419–1428.

[70] D. Bertsimas, C. Pawlowski, and Y. D. Zhuo., From predictive methods to missing data imputation: An optimization approach, J. Mach. Learn. Res., 18 (2018) 1–39.