Available online at www.CivileJournal.org

Civil Engineering Journal

(E-ISSN: 2476-3055; ISSN: 2676-6957)

Vol. 7, No. 09, September, 2021

A Comparison of Multiple Imputation Methods for Recovering Missing Data in Hydrological Studies

Fatimah Bibi Hamzah [1, 2*], Firdaus Mohd Hamzah [1*], Siti Fatin Mohd Razali [1], Hafiza Samad [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.*

Received 02 May 2021; Revised 31 July 2021; Accepted 11 August 2021; Published 01 September 2021

Abstract

Missing data is a common problem in hydrological studies; therefore, data reconstruction is critical, especially when it is crucial to employ all available resources, even incomplete records. Furthermore, missing data could have an impact on statistical analysis results, and the amount of variability in the data would not be fittingly anticipated. As a result, this study compared the performance of three imputation methods in predicting recurrence in streamflow datasets: robust random regression imputation (RRRI), k-nearest neighbours (k-NN), and classification and regression tree (CART). Furthermore, entire historical daily streamflow data from 2012 to 2014 (as training dataset) were utilised to assess and validate the effectiveness of the imputation methods in addressing missing streamflow data. Following that, all three methods coupled with multiple linear regression (MLR), were used to restore streamflow rates in Malaysia's Langat River Basin from 1978 to 2016. The estimation techniques effectiveness was evaluated using metrics inclusive of the Nash-Sutcliffe efficiency coefficient (CE), root-mean-square error (RMSE), and mean absolute percentage error (MAPE). The results confirmed that RRRI coupled with MLR (RRRI-MLR) had the lowest RMSE and MAPE values, outperforming all other techniques tested for filling missing data in daily streamflow datasets. This indicates that the RRRI-MLR is the best method for dealing with missing data in streamflow datasets.

Keywords: Missing Data; Streamflow; Robust Regression; CART; k-NN; MLR.

1. Introduction

Missing data in hydrological models is a prevalent problem owing to natural disasters, improper operation, and battery drainage, which restrict hydrological analysis [1, 2] and has remained unsolved regardless of advancements in missing data imputation techniques over the years [3]. Missing data reconstruction is crucial, especially in an event where all available resources, including partial information, must be used. The lack of particular data can pose severe problems in hydrological studies, resulting in uncertainty and low efficiency of water resource systems [4-6].

Even minor data breaches can prohibit the computation of significant summary statistics and hydrological indexes, such as monthly runoff totals or n-day minimum flows, restricting analysis and explanation of historical flow variability [7]. Water development system planning, hydraulic structure design, and water resource management are all hampered

* Corresponding author: bibi@kuptm.edu.my; fir@ukm.edu.my

http://dx.doi.org/10.28991/cej-2021-03091747

by these gaps and breaks [8]. Additionally, extra expenses may be spent if a modelling system or decision support system eventually demands the utilisation of this measured data. As a result of these disadvantages, gaps must be filled, and the handling of missing data should be prioritised in the data preparation procedure.

The most convenient method for dealing with missing data is to delete the entire observations with partial data and analyse the remaining complete data [6]. On the other hand, deleted data may result in discontinuous data, resulting in information loss and skewed conclusions. In recent years, however, various data estimation approaches have been proposed and widely debated in relevant literature to solve this challenge. These approaches range from simple classic statistical methods such as replacing mean, median, or alternative location stations for each missing value to advanced computational techniques.

For example, Hirsch (1979) and Wallis et al. (1991) [9, 10] addressed reconstruction methods for daily data utilising data from neighbouring stations. In another study, the missing consecutive streamflow was reconstructed using the k-NN algorithm, and the impact of increasing the value of k on the results was also demonstrated [11]. Recent work by Cheng and Syu (2019) [12] compared the k-NN average (kNN-AVG) algorithm to the backpropagation neural network (BPNN) in wireless positioning systems. Due to BPNN's learning capacity, the study discovered it improved area positioning accuracy more than kNN-AVG. Meanwhile, in Worland et al. (2018) [13] study, the performance of eight machine learning models (including k-NN) and four baseline models to forecast hydrological low-flow indices in ungauged basins in South Carolina, Georgia, and Alabama, USA, was examined. The study found that machine learning models produced much-reduced cross-validation errors than baseline models. Another study used four different approaches to substitute missing meteorological data: linear interpolation, mode imputation, k-NN, and multivariate imputation by chain equations (MICE). When the k-NN method was used to the test data, the prediction performance provided results closest to the original data with no missing values. The prediction model's performance remained steady even when the missing data rate rose [14].

Several recent studies have proposed techniques for filling in missing hydrological data utilising CART approaches or random forests derived from CART for missing streamflow records imputation. According to these studies, the CART model outperformed other classification methods in terms of explained variance [15-17]. Similarly, in Erdal et al. (2013) [18] study, CART was utilised for monthly streamflow forecasting, with a support vector regression (SVR) model serving as the benchmark model. CART was demonstrated to outperform SVR in both the training and testing stages.

Regression methods have long been utilised in statistical approaches to reconstruct missing streamflow data [19]. MICE, a well-known technique for performing multiple imputations, employs sequential regression modelling as well [20, 21]. The aim is to simulate flow at one gauge as a function of flow at another or several gauges. In Beauchamp et al. (1989) [19], regression and time series methods were utilised to synthesise and predict streamflow at a downstream gauge over an upstream gauge in California. According to the study, either method produced logically good estimations and projections of the flow at the downstream gauge. In another study, an MLR model was employed to estimate streamflow in the Wainganga River. According to the study, the method employed was one of the simplest and fastest ways to compute runoff [22]. Furthermore, Gyau-Boakye and Schultz (1994) study [23] examined ten well-known techniques, including interpolation, recursive models, autoregressive models, regression, and non-linear models and concluded that the interpolation and multiple regression models fared well. A detailed outline of techniques used in hydrology for the reconstruction of missing data, including an applied comparison of simple and multiple regression models was presented in Harvey et al. (2012) [7]. The study revealed that when multiple input variables are included, accuracy improves.

However, no research on the reconstruction of missing streamflow data utilising effective RRRI techniques had been conducted before. As a result, this study's objectives were twofold: (1) to reconstruct missing flow data from the Langat River Basin using RRRI from the statistical field in comparison to machine learning techniques: k-NN and CART; and (2) to evaluate the performance of imputation methods coupled with the MLR model in forecasting future daily streamflow values. The findings of this study are likely to aid in the development of the best techniques for data imputation that allow for the reconstruction of entire daily streamflow datasets.

2. Area of Study

The Langat River Basin is located in the southernmost state of Selangor and upstate of Negeri Sembilan, especially between latitudes 2o 40'M 152" N to 3o 16'M 15" N and longitudes 101o 19'M 20" E to 102o 1'M 10" E in the western part of Peninsular Malaysia, as seen in Figure 1. The basin comprised an area of approximately 2,394.38 km2, with a major river channel stretching for around 141 km. The river runs southward towards the lower mainland and westwards towards the coast of the Selangor state, with its mouth is located in the Straits of Malacca [24]. This river basin, Malaysia's most urbanised river basin, is considered to compensate for the advantages of Klang Valley spill-over development [25, 26]. It is a critical raw water resource for drinking water as well as other activities such as recreation, industrial usage, fishing, and agriculture [27]. This study looked at four Langat River sub-basins (Kajang, Dengkil, Lui, and Semenyih).

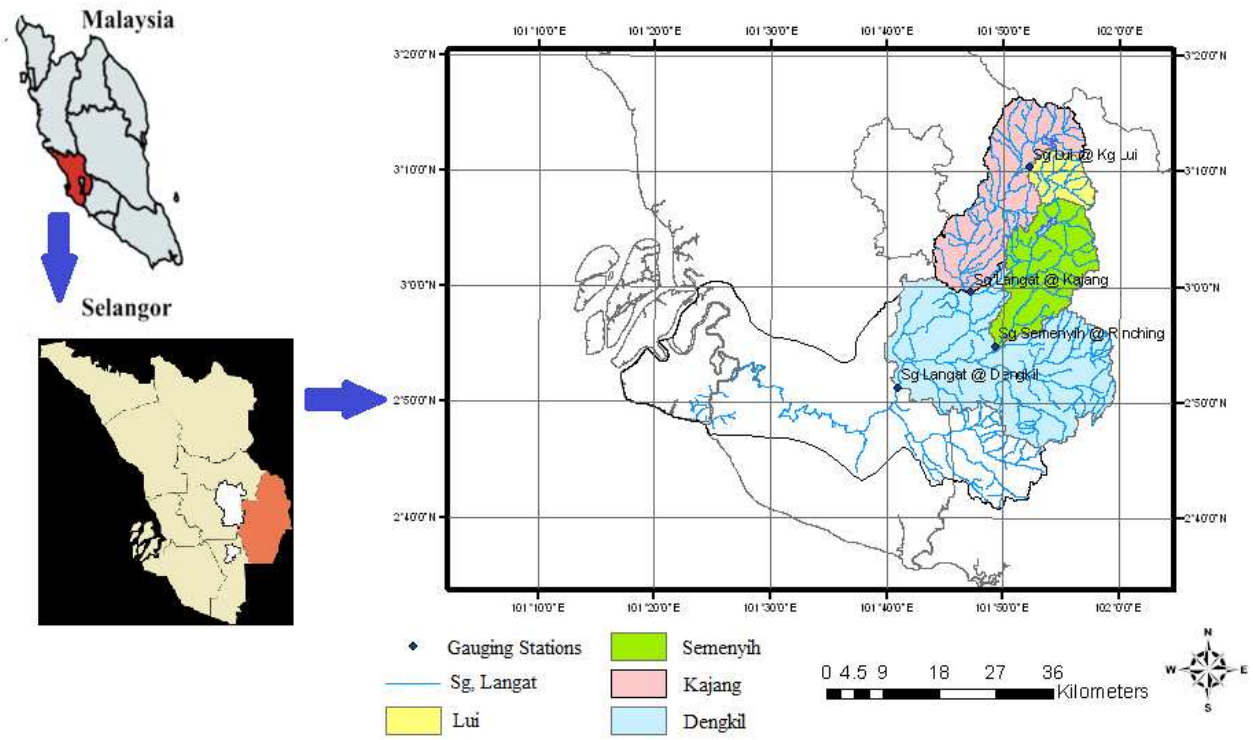

Figure 1. Map of Langat River Basin

The Langat Basin is impacted by two types of monsoons in terms of hydrometeorology: the northeast and southwest monsoons, which occur from November to March and May to September, respectively [28, 29]. The southwest monsoon, which blows over the Malacca Strait has the most impact on the climate of the basin [30]. The Langat River Basin has four flow rate gauging stations: Dengkil and Kajang at Langat River, Kg. Rinching at Semenyih River, and Kg. Lui at Lui River. The characteristics of sub-basins connected with Langat Basin gauging stations regulated by the Department of Irrigation and Drainage (DID) are depicted in Table 1.

Table 1. Overview of the sub-basins allied with gauging stations of the Langat Basin

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
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”
Area (km^2)	1251.4	389.4	236	68.4
Period of Data Availability (with missing data)	1978 -2016			
Period of Data Availability (without missing data)	2012 - 2014			

Notes: Data obtained from Malaysia's DID.

High-dimensional data obtained from Malaysia’s DID, Ampang, Selangor, recorded from 1978 to 2016 was utilized in this study. The streams have been monitored constantly and recorded in m^3/s as daily mean flow rates. Of the 56,980 data points, 12.5% had missing values. Datasets containing 10 to 25% missing values are classified as moderate data [31]. Furthermore, as stated in Bennett (2001) study [32], if the percentage of missing data exceeds 10%, the statistical analysis is likely to be skewed. A large number of time series data are necessary to get a precise outline of the streamflow patterns [33]. Aside from that, the reliability of a frequency estimator for a lengthy time series dataset is extremely valuable in data analysis since it is closely related to sample size.

3. Research Methodology

This section is split into two major subsections. The techniques for estimating missing data are discussed in the first subsection. Meanwhile, the second subsection describes how the performance of the methods utilised is evaluated. This study employed a cross-validation methodology on data from 2012 to 2014 to assess the competency of infilling methods. This period was chosen as the baseline due to the availability of comprehensive data. The missing daily streamflow data were recovered from the entire time series data after being simulated at random. Figure 2 depicts the flowchart of imputation and the procedure for integrating missing data into the entire time series.

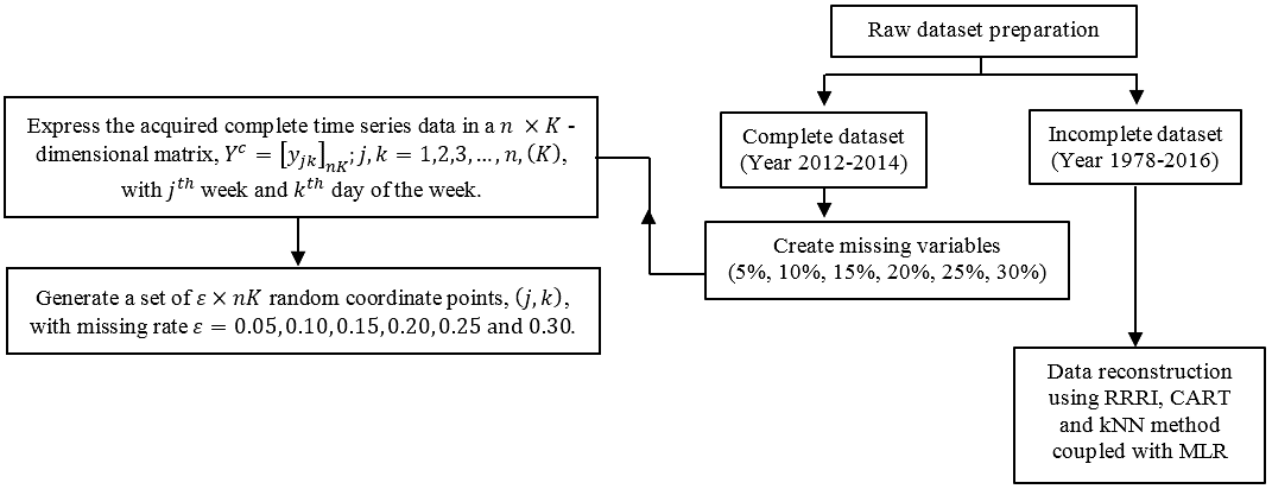

Figure 2. Flowchart of imputation and the process for incorporating missing data into the entire time series

3.1. Imputation Methods

This study compared three methods of imputation to determine the best fit technique to impute missing values in the streamflow datasets. The RRRI method is a statistical methodology, whereas k-NN and CART are machine learning techniques. Initially, multiple imputation methods were used to fill up all missing values with replacement from the observed values. In general, the first variable with missing values (x_1) is regressed on all other variables ($x_2, x_3, \dots x_k$), but only on individuals with the observed x_1. Missing values in x_1 are replaced with simulated draws from x_1's posterior predictive distribution. The next variable with missing values (x_2) is then regressed on all other variables ($x_1, x_3, \dots x_k$), restricted to individuals with the observed x_2, and using the imputed values of x_1. Missing values in x_2 are again replaced by draws from x_2's posterior predictive distribution. The process is repeated for each variable with missing values in turn - this is referred to as a cycle. To stabilise the results, the procedure is typically repeated for several cycles (e.g. 10 or 20) to produce a single imputed dataset, and the entire procedure is repeated m times to produce m imputed datasets. Following the missing values imputation process, the CE, RMSE, and MAPE for each of the three predicted values were then computed. Finally, all three methods were employed in conjunction with MLR to restore streamflow rates in Malaysia's Langat River Basin from 1978 to 2016.

3.1.1. *k*-Nearest-Neighbor Imputation (*k*-NN)

The machine learning-based k-NN imputation, also known as distance function matching, is a donor approach in which the donor is chosen by minimising a specified 'distance' and the mean is utilized as an imputation estimate [34, 35]. It represents the local estimate approach, which predicts using just neighbouring states. Local estimators are regularly believed to give good results in chaotic time series [11]. The missing values are based on a set number of cases, one of which is very certainly the instance of interest [36]. This procedure involves calculating an appropriate distance measure, with the distance defined by the auxiliary variables. This study employed the Euclidean distance, one of the most prominent methods for measuring distance, and the formula is as follows:

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

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

The first step in creating predictions using the k-NN method is determining the value of k. According to Yang (1999) [37], a larger value of k provides greater weight to accuracy and is more stable since it reduces total noise, but there is no assurance. The Elshorbagy et al. (2002) study [11], on the other hand, stated that the smaller the k value, the better the estimation of the missing value. The most commonly used rule of thumb is that k equals the square root of the

number of points in the training dataset [38]. As a result, this study opted to have a maximum number of k that is less than or equal to the square root of the size of the training dataset utilized. This yielded far superior results than 1NN, which merely nominated to the class of its nearest neighbour.

After determining the value of k, predictions can be made using the k-NN method. The imputation procedure by the nearest neighbour can be summarized for k neighbours as follows:

Assuming that there are m observations on n covariates, $X = x_{is}$ denotes the corresponding $m \times n$ matrix, where x_{is} represent the i^{th} observation of the s^{th} variable. Let $O = o_{is}$ represent the corresponding $m \times n$ dummy matrix with the following entries:

$$o_{is} = \begin{cases} 1 & if\ x_{is}\ was\ observed \\ 0 & for\ missing\ value \end{cases} \quad (2)$$

The L_q metric for the data observed can be used to compute the distances between two observations x_i and x_j, which are represented in the data matrix by rows. The distance is then calculated as follows:

$$d_q(x_i, x_j) = \left[d_{ij} \sum_{s=1}^{n} |x_{is} - x_{js}|^q 1(o_{is} = 1) I(o_{js} = 1) \right]^{1/q} \quad (3)$$

where $d_{ij} = \sum_{s=1}^{p} 1(o_{is} = 1) I(o_{js} = 1)$ represents the number of valid components in the computation of distances. Because parallel views were utilised to conceptualise distances, nearest neighbours were employed.

3.1.2. Multiple Classification and Regression Tree (CART)

CART [39] is a well-known classification of machine learning algorithms [39] that employs the concept depicted in Figure 3. CART models need predictors as well as cut-points in predictors used to split the sample. The cut-points split the sample into larger, homogenous subsamples. The splitting procedure is repeated on both subsamples, allowing a succession of splits to form a binary tree [18]. For regression problems, each node in the tree has a splitting rule defined by minimising the relative error (RE), which is similar to minimising the sums-of-squares of the split:

$$RE(d) = \sum_{l=0}^{L} (y_l - \bar{y}_L)^2 + \sum_{r=0}^{R} (y_r - \bar{y}_R)^2 \quad (4)$$

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 decides which branches go left and which go right. The partitioning rule that minimises the RE is then used to construct the tree node. Figure 3 shows an example of a CART framework.

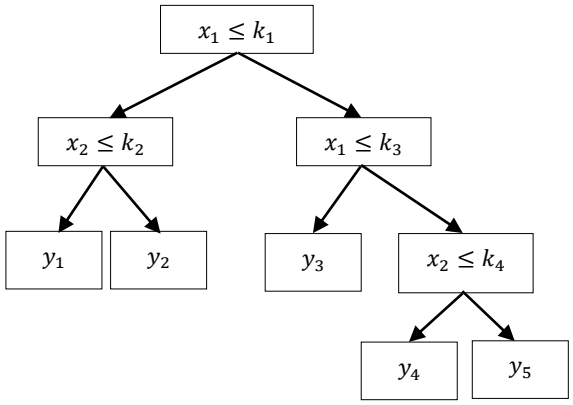

Figure 3. Multiple CART structure

According to Breiman et al. (1984) [39], a random forest can handle diverse data types, and, being a non-parametric method, non-linear (regression) and interaction effects are expected. Assuming $X = (X_1, X_2, \ldots, X_n)$ is a $m \times n$-dimensional data matrix, for an arbitrary variable X_s that includes missing values at entries $i_{mis}^{(s)} \subseteq \{1, \ldots, m\}$, the streamflow dataset could be split into two categories: $y_{obs}^{(s)}$ denotes the observed values of variable X_s, while $y_{mis}^{(s)}$ denotes the missing values of variable X_s.

To begin, mean or other imputation methods are used to make an initial estimation for the missing values in X. The variables $X_s, s = 1, \ldots, p$ are then sorted by the number of missing values, beginning with the smallest. Missing values are reconstructed for each variable X_s by first fitting a CART with response $y_{obs}^{(s)}$ and predictors $x_{obs}^{(s)}$, and then predicting the missing values $y_{mis}^{(s)}$ by applying the trained CART to $x_{mis}^{(s)}$. The imputation procedure is performed until a stopping criterion is reached.

3.1.3. Robust Random Regression Imputation (RRRI)

RRRI is a less stringent variant of least squares regression that operates with looser assumptions. It offers considerably better estimations of the regression coefficients when the data are ambiguous. There are numerous good examples of robust methods in the literature, the most often used: the M-estimator, the least median of squares (LMS) estimator, and the least trimmed sum of squares (LTS) S-estimator, and the MM-estimator. This study adopted the high breakdown and high-efficiency MM-estimator proposed by Yohai (1987) [40]. The following is a simple linear regression model:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad ; i = 1,2,3, \dots, n \tag{5}$$

where y is the response variable, x is the regressor, α is the intercept and β is the slope, and ε_i is the random error term.

Assuming that the k^{th} $(k = m+1, \dots, s)$ cases in the response variable y are missing, the MM-estimator is used to fit the regression line for the available cases $(i = 1,2,3, \dots, m)$. The estimated regression for the available case is given by

$$\hat{y}_i = \hat{\alpha}^*_{MM} + \hat{\beta}^*_{MM} x_i \qquad ; i = 1,2,3, \dots, m \tag{6}$$

The estimated parameters in (6) and the x value corresponding to the missing y value are then utilised to reconstruct missing y values as predicted values. The following is proposed RRRI:

$$\hat{y}_k = \hat{\alpha}^*_{MM} + \hat{\beta}^*_{MM} x_k + \hat{\varepsilon}^*_k \qquad ; k = m+1, \dots, s \tag{7}$$

where the random error term $\hat{\varepsilon}^*_k \sim N(0, S)$, S = MSE of the residuals from (6), is added to the predicted values according to Little and Rubin (2002) [41].

3.1.4. Multiple Linear Regression (MLR)

Following the replacement of all missing values with multiple techniques, the complete dataset is analysed using MLR to determine the best 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 predicted variables [42]. The MLR model is a widely used statistical technique in many fields, including hydrological data [43, 44]. The following is how the MLR model parameter is expressed:

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

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.

3.2. Performance of the Estimation Methods

In this study, three performance metrics were utilised to assess imputation methods: CE, RMSE, and MAPE. The CE index is a well-known metric for assessing the prediction power of hydrological models. The most effective performance models aim for a CE value of one (1). Meanwhile, RMSE is a common statistical metric used to assess model performance in meteorology, air quality, and climate research investigations. The RMSE statistic, which is a measure of the difference between predicted and observed values, offers information about short-term efficiency. Another useful measure used extensively in model evaluations is MAPE. Especially, MAPE provides an insight into the average deviation of the predicted values from the observed values and the long-term performance of these models. The lower the values of RMSE and MAPE values, the better findings of the long-term model. The following formulae can be used to compute these statistics:

$$CE = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 / n} \tag{10}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \tilde{y}_i|}{y_i} \tag{11}$$

where y_i is the observed streamflow data, $\tilde{y}_i$ is the predicted value, $\bar{y}_i$ indicates the average streamflow data, n represents the sample size and k represents the number of independent variables in the regression equation over daily streamflow datasets from the Langat River Basin.

4. Results and Discussion

This study was conducted to find the best imputation method for calculating missing streamflow data comparing k-NN, CART, and RRRI. The models were first evaluated without missing values on the training dataset from the year 2012 to 2014. The simulation process was performed in the following flow: a conventional training dataset was generated using the missing data rates (i.e. 5, 10, 15, 20, 25, and 30%), and the missing values were substituted with new values, acquired using each of the previously mentioned imputation methods. The error was computed by subtracting the trained model's predicted value from the reference model's predicted value and the data acquired using the missing value replacement method. The trained model with the original training data and test data with no missing values was used as the reference model. Each method's performance was evaluated using CE, RMSE, and MAPE. When the gap between the estimated and observed values is minimal, RMSE and MAPE will provide the smallest value. Meanwhile, CE values can vary from $-\infty$ to 1, with values larger than 0.5 deemed acceptable. The method with the greatest CE value and the lowest RMSE and MAPE values was chosen as the best. Table 2 shows the prediction model errors, whereas Tables 3-5 display the deviation results.

Table 2. Error of streamflow reference model

Year	CE	RMSE	MAPE
2012 – 2014	0.653	0.346	0.468

Table 3. Performance of six different percentages of missing data compared based on CE

Methods	Missing data rate					
	5%	10%	15%	20%	25%	30%
RRR	0.643	0.699	0.669	0.683	0.701	0.712
k-NN	0.644	0.698	0.669	0.682	0.697	0.692
CART	0.637	0.661	0.659	0.661	0.665	0.664

Table 4. Performance of six different percentages of missing data compared based on RMSE

Methods	Missing data rate					
	5%	10%	15%	20%	25%	30%
RRR	0.292	0.291	0.299	0.302	0.305	0.307
k-NN	0.306	0.313	0.307	0.317	0.310	0.313
CART	0.362	0.318	0.330	0.318	0.314	0.315

Table 5. Performance of six different percentages of missing data compared based on MAPE

Methods	Missing data rate					
	5%	10%	15%	20%	25%	30%
RRR	0.411	0.415	0.416	0.499	0.461	0.419
k-NN	0.458	0.435	0.422	0.501	0.464	0.427
CART	0.476	0.450	0.427	0.501	0.467	0.427

The consistency of each imputation method was determined using gap analysis, as indicated by reduced gaps between training and validation results. According to the aforementioned data (Tables 2 to 5), the RRRI method had the highest CE and the lowest RMSE and MAPE values. Conversely, CART was the worst imputation method, having the lowest CE and the greatest RMSE and MAPE values. Despite this, CE values indicated that all imputation methods gave acceptable results, with values near to one and deviating from the training set by less than 10%. Nevertheless, RMSE produced somewhat lower values than the training sets as the missing data rate increased. Meanwhile, MAPE measured the magnitude of the error in percentage terms, and the values varied slightly depending on the mean difference between the observed known outcome values and the values predicted by the model. As the rate of missing data increased, so did the error between the reference and validation models with missing data imputation. This indicated a small error when training data were used with no missing values. Even if the missing data rate was only 30%, the training model would have followed the pattern of the remaining 70% training data rather than the 30% missing data. Consequently, a substantial error went unnoticed despite the existence of missing values in the training data.

Later, the model was evaluated for all four sub-basins using a dataset spanning the years 1978 to 2016. The findings were then computed as an average of each imputation method results. Table 6 shows the overall performance of each method in the reconstruction of data from 1978 to 2016. The RRRI method produced the lowest RMSE and greatest CE values, according to the results. Despite this, CE values indicated that all imputation methods gave acceptable results, with values near to one. Based on the findings, it is possible to infer that the RRRI method gave the best results, whereas k-NN was the worst imputation method for daily streamflow data in Malaysia's Langat River Basin due to the lowest CE and greatest RMSE values among the other methods.

Table 6. Average RMSE and CE values for three imputation methods

Method	RMSE	CE
RRR	8.099	0.900
k-NN	28.096	0.767
CART	10.484	0.854

Notes: The best method is in bold.

Following the completion of the missing values, the MLR model was used to analyze the whole dataset in this study and find the best approaches for dealing with missing data when imputation values were coupled with modelling. The performance of imputation methods coupled with an MLR model was evaluated using MAPE and RMSE. Table 7 shows 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 the MLR model. RRRI-MLR exhibited the lowest RMSE and MAPE values of 12.157 and 0.216, respectively, when compared to the other methods. The final results showed that RRRI is the best statistical approach for imputing missing values in daily streamflow data when coupled with a regression model. Table 7 further indicated that the CART-MLR imputation method had low RMSE and MAPE values, putting it on par with the RRRI-MLR model.

Table 7. The outcomes for MLR combined with imputation methods.

Method	RMSE	MAPE
k-NN-MLR	23.784	1.397
CART-MLR	15.778	0.487
RRR-MLR	12.157	0.216

Notes: The best method is in bold.

Finally, visual inspection/ analytic data were drawn up with observed and predicted values for the kNN-MLR, CART-MLR, and RRRI-MLR models. Figure 4 depicts the results of three imputation methods for replacing 7,124 missing daily streamflow data points in Malaysia's Langat River Basin and shows the comparable patterns of imputed daily streamflow values from all three methods. All models, for instance, responded with comparable peaks and durations in streamflow events.

It can be concluded that the RRRI-MLR method achieved the best performance, whereas k-NN-MLR performed the least. This was not surprising for the k-NN method from the aspects of the most popular methods and the methodological simplicity; the method is a lazy learner, unable to learn anything from the training data. The training data was instead used for classification, which can result in poor algorithm generalization and outliers susceptibility [45]. This was consistent with a recent finding in Miró et al. (2017) [46], where advanced linear methods produced far superior results than traditional methods, such as k-NN. To predict a new instance label, the k-NN algorithm searches the data for the k closest neighbours to the new instance and sets the predicted class label to be the most prevalent label among the k closest neighbouring points. In each prediction, the algorithm must compute and sort the distance and data for a number of training instances, which might be time-consuming. Another possible consequence of changing the k value is the change in the class label [47].

On the other hand, the CART method performed better than k-NN when coupled with MLR. The CART method is recognised for its simplicity, robustness, ability to handle multicollinearity and skewed distributions, and adaptability to interactions and non-linear relationships [21]. This conclusion was consistent with previous studies [17, 18], which found that the CART model outperformed other classification algorithms in terms of explained variance. Regardless of how flexible and interpretable CART is, it is critical to understand how it works to establish and cross-validate suitable tuning parameters, such as tree depth or split number [48]. The CART technique can also be time-intensive when applied to large datasets.

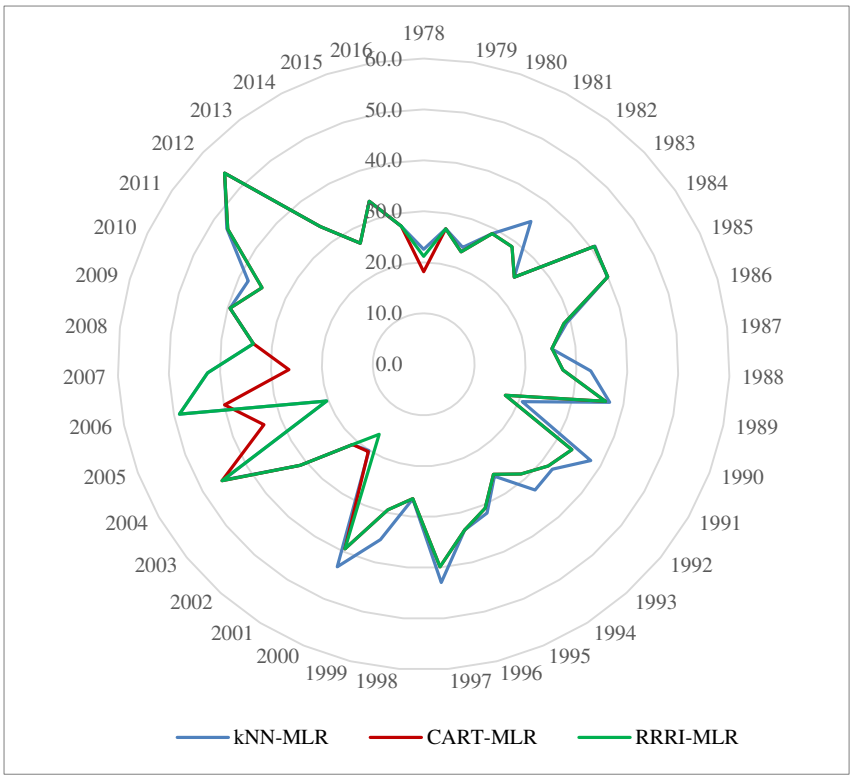

Figure 4. kNN-MLR, CART-MLR and RRRI-MLR data imputation results for 7,124 missing streamflow data

Comparatively, the results revealed that the robust technique of this study prevailed over the investigated non-robust approaches. The obvious reason for this is that RRRI with MM-estimator has high asymptotic efficiency, about 95% compared to ordinary least squares (OLS) under Gauss-Markov assumptions, and a high breakdown (about 50% ~ n/2) [49]. RRRI add a random error term in which without adding this term, for the same value of independent variables, it will result in the same response which is not true in reality. Furthermore, when the error percentage is mirrored by the missing data proportion, the RRRI technique produces a lower error than the k-NN and CART techniques. These simulations demonstrate unequivocally that the RRRI technique is the most effective missing data imputation method for reconstructing missing streamflow data.

5. Conclusion

Missing data is a frequent constraint of hydrological research and usually leads to misinterpretations of statistical output and hydrological modelling techniques. Therefore, method performance evaluation is necessary to reduce the impact of missing data. Several techniques have been proposed in the literature to manage missing data. However, a suitable approach to be used as the missing data trend and mechanism are still unclear. Researchers typically exclude observations with missing data or replace them through naive methods, such as the mean or mode of all other observations, because of convenience, even though these methods are statistically significantly worse. In this study, a novel statistical approach to treating and estimating missing data in streamflow datasets was proposed. The findings demonstrated that when coupled with MLR, the proposed method, RRRI with MM-estimator, gave the best results. The RRRI-MLR method outperformed k-NN and CART on all three performance metrics (CE, RMSE, and MAPE), with a higher adjusted CE and lower RMSE and MAPE. This result indicates that the RRRI technique has the lowest variance between the reference model and the prediction model with missing data imputation. As a result, using the RRRI technique to address missing streamflow data can produce the best results. As such, this method should be classified among suitable contenders for managing missing data in streamflow datasets.

5.1. Limitations and Directions for Future Research

This study was based on k-NN, CART, and RRRI performance as conditional models for imputing missing flow records. Four gauging stations were used to analyze the data matrices of the Langat River Basin. However, other critical streamflow characteristic contributors, such as rainfall, temperature, topography, or other study area parameters were not examined due to data limitations. Ignoring such parameters may result in erroneous data prediction. In future studies, the performance of the proposed imputation method should be compared to other imputation methods like support vector machines and artificial neural networks. In order to examine the robustness of results, a sensitivity analysis may also be beneficial with several methods for managing missing data.

6. Declarations

6.1. Author Contributions

Conceptualization, F.M.H.; methodology, F.B.H., F.M.H., S.F.R., and N.H.S.; validation, F.B.H., F.M.H., S.F.R. and N.H.S.; formal analysis, F.B.H.; writing—original draft preparation, F.B.H.; writing—review and editing, F.M.H., S.F.R., and N.H.S.; visualization, F.B.H.; supervision, F.M.H., and S.F.R. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

Restrictions apply to the availability of these data. Data was obtained from Department of Irrigation and Drainage (DID), Ampang, Selangor, Malaysia.

6.3. Funding

This research was funded by a KUPTM research grant (Ref: URG/1219/FCOM/FP01125(11)).

6.4. Conflicts of Interest

The authors declare no conflict of interest.

7. References

[1] Mwale, F.D., A.J. Adeloye, and R. Rustum. “Infilling of Missing Rainfall and Streamflow Data in the Shire River Basin, Malawi-A Self Organizing Map Approach.” Physics and Chemistry of the Earth, Parts A/B/C 50–52 (2012): 34–43. doi:10.1016/j.pce.2012.09.006.

[2] Hamzah, Fatimah Bibi, Firdaus Mohd Hamzah, Siti Fatin Mohd Razali, Othman Jaafar, and Norhayati Abdul Jamil. “Imputation Methods for Recovering Streamflow Observation: A Methodological Review.” Edited by Fei Li. Cogent Environmental Science 6, no. 1 (January 1, 2020): 1745133. doi:10.1080/23311843.2020.1745133.

[3] Mispan, M. R., N. F. A. Rahman, M. F. Ali, K. Khalid, M. H. A. Bakar, and S. H. Haron. "Missing river discharge data imputation Approach using artificial neural network." ARPN J. Eng. Appl. Sci. 10, no. 22 (December 2015): 10480-10485.

[4] Adeloye, Adebayo J., Rabee Rustum, and Ibrahim D. Kariyama. “Kohonen Self-Organizing Map Estimator for the Reference Crop Evapotranspiration.” Water Resources Research 47, no. 8 (August 2011): 1-19. doi:10.1029/2011wr010690.

[5] Adeloye, Adebayo J. “An Opportunity Loss Model for Estimating the Value of Streamflow Data for Reservoir Planning.” Water Resources Management 10, no. 1 (February 1996): 45–79. doi:10.1007/bf00698811.

[6] Mariana Che Mat Nor, Siti, Shazlyn Milleana Shaharudin, Shuhaida Ismail, Nurul Hila Zainuddin, and Mou Leong Tan. “A Comparative Study of Different Imputation Methods for Daily Rainfall Data in East-Coast Peninsular Malaysia.” Bulletin of Electrical Engineering and Informatics 9, no. 2 (April 1, 2020): 635-643. doi:10.11591/eei.v9i2.2090.

[7] Harvey, Catherine L., Harry Dixon, and Jamie Hannaford. “An Appraisal of the Performance of Data-Infilling Methods for Application to Daily Mean River Flow Records in the UK.” Hydrology Research 43, no. 5 (April 12, 2012): 618–636. doi:10.2166/nh.2012.110.

[8] Tfwala, Samkele S., Yu-Min Wang, and Yu-Chieh Lin. “Prediction of Missing Flow Records Using Multilayer Perceptron and Coactive Neurofuzzy Inference System.” The Scientific World Journal 2013 (2013): 1–7. doi:10.1155/2013/584516.

[9] Hirsch, Robert M. “An Evaluation of Some Record Reconstruction Techniques.” Water Resources Research 15, no. 6 (December 1979): 1781-1790. doi:10.1029/wr015i006p01781.

[10] Wallis, James R., Dennis P. Lettenmaier, and Eric F. Wood. “A Daily Hydroclimatological Data Set for the Continental United States.” Water Resources Research 27, no. 7 (July 1991): 1657–1663. doi:10.1029/91wr00977.

[11] Elshorbagy, Amin, S.P. Simonovic, and U.S. Panu. “Estimation of Missing Streamflow Data Using Principles of Chaos Theory.” Journal of Hydrology 255, no. 1–4 (January 2002): 123–133. doi:10.1016/s0022-1694(01)00513-3.

[12] Cheng, Chia-Hsin, and Siang-Jhih Syu. “Improving Area Positioning in ZigBee Sensor Networks Using Neural Network Algorithm.” Microsystem Technologies 27, no. 4 (January 22, 2019): 1419–1428. doi:10.1007/s00542-019-04309-2.

[13] Worland, Scott C., William H. Farmer, and Julie E. Kiang. “Improving Predictions of Hydrological Low-Flow Indices in Ungaged Basins Using Machine Learning.” Environmental Modelling & Software 101 (March 2018): 169–182. doi:10.1016/j.envsoft.2017.12.021.

[14] Kim, Taeyoung, Woong Ko, and Jinho Kim. “Analysis and Impact Evaluation of Missing Data Imputation in Day-Ahead PV Generation Forecasting.” Applied Sciences 9, no. 1 (January 8, 2019): 204. doi:10.3390/app9010204.

[15] Vezza, Paolo, Claudio Comoglio, Maurizio Rosso, and Alberto Viglione. “Low Flows Regionalization in North-Western Italy.” Water Resources Management 24, no. 14 (May 6, 2010): 4049–4074. doi:10.1007/s11269-010-9647-3.

[16] Karakurt, Onur, Halil Ibrahim Erdal, Ersin Namli, Hacer Yumurtaci-Aydogmus, and Yusuf Sait Turkkan. “Comparing Ensembles Of Decision Trees And Neural Networks For One-Day-Ahead Stream Flow Predict.” Science Park 1, no. 17 (November 1, 2013): 43–54. doi:10.9780/23218045/1172013/41.

[17] Tyralis, Hristos, Georgia Papacharalampous, and Andreas Langousis. “A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources.” Water 11, no. 5 (April 30, 2019): 910. doi:10.3390/w11050910.

[18] Erdal, Halil Ibrahim, and Onur Karakurt. “Advancing Monthly Streamflow Prediction Accuracy of CART Models Using Ensemble Learning Paradigms.” Journal of Hydrology 477 (January 2013): 119–128. doi:10.1016/j.jhydrol.2012.11.015.

[19] Beauchamp, J.J., D.J. Downing, and S.F. Railsback. “Comparison of Regression and Time-Series Methods for Synthesizing Missing Streamflow Records.” Journal of the American Water Resources Association 25, no. 5 (October 1989): 961–975. doi:10.1111/j.1752-1688.1989.tb05410.x.

[20] Su, Yu-Sung, Andrew Gelman, Jennifer Hill, and Masanao Yajima. “Multiple Imputation with Diagnostics (mi) inR: Opening Windows into the Black Box.” Journal of Statistical Software 45, no. 2 (2011): 31. doi:10.18637/jss.v045.i02.

[21] Buuren, Stef van, and Karin Groothuis-Oudshoorn. “Mice: Multivariate Imputation by Chained Equations inR.” Journal of Statistical Software 45, no. 3 (2011): 1-67. doi:10.18637/jss.v045.i03.

[22] Schilling, Keith E., and Calvin F. Walter. “Estimation of Streamflow, Base Flow, and Nitrate-Nitrogen Loads in Iowa Using Multiple Linear Regression Models.” Journal of the American Water Resources Association 41, no. 6 (December 2005): 1333–1346. doi:10.1111/j.1752-1688.2005.tb03803.x.

[23] Gyau-Boakye, P., and G. A. Schultz. “Filling Gaps in Runoff Time Series in West Africa.” Hydrological Sciences Journal 39, no. 6 (December 1994): 621–636. doi:10.1080/02626669409492784.

[24] Ebrahimian, Mahboubeh, Ahmad Ainuddin Nuruddin, Mohd Amin Mohd Soom, Alias Mohd Sood, Liew Ju Neng, and Hadi Galavi. “Trend Analysis of Major Hydroclimatic Variables in the Langat River Basin, Malaysia.” Singapore Journal of Tropical Geography 39, no. 2 (February 27, 2018): 192–214. doi:10.1111/sjtg.12234.

[25] Noorazuan, M. H., Ruslan Rainis, Hafizan Juahir, S. M. Zain, and Nazari Jaafar. "GIS application in evaluating land use-land cover change and its impact on hydrological regime in Langat River basin, Malaysia." In 2nd annual Asian Conference of Map Asia, (February 2003): 14-15.

[26] Wan Mohtar, Wan Hanna Melini, Siti Aminah Bassa Nawang, and Mohd Noor Shafique Rahman. “Statistical Analysis in Fluvial Sediments of Selangor Rivers: Downstream Variation in Grain Size Distribution.” Jurnal Kejuruteraan S, no. 1 (July 1, 2017): 37–45. doi:10.17576/jkukm-s-01-06.

[27] Juahir, Hafizan, T. Mohd Ekhwan, Sharifuddin M. Zain, M. Mokhtar, J. Zaihan, and M. J. Ijan Khushaida. "The use of chemometrics analysis as a cost-effective tool in sustainable utilisation of water resources in the Langat River Catchment." American-Eurasian Journal of Agricultural & Environmental Sciences 4, no. 1 (2008): 258-265.

[28] Memarian, Hadi, Siva K. Balasundram, Jamal B. Talib, Alias M. Sood, and Karim C. Abbaspour. “Trend Analysis of Water Discharge and Sediment Load During the Past Three Decades of Development in the Langat Basin, Malaysia.” Hydrological Sciences Journal 57, no. 6 (June 26, 2012): 1207–1222. doi:10.1080/02626667.2012.695073.

[29] Hai Hwee Yang. “Analysis of Hydrological Processes of Langat River Sub Basins at Lui and Dengkil.” International Journal of the Physical Sciences 6, no. 32 (December 2, 2011): 7390–7409. doi:10.5897/ijps11.1036.

[30] Juahir, Hafizan, Sharifuddin M. Zain, Mohd Kamil Yusoff, T. I. Tengku Hanidza, A. S. Mohd Armi, Mohd Ekhwan Toriman, and Mazlin Mokhtar. “Spatial Water Quality Assessment of Langat River Basin (Malaysia) Using Environmetric Techniques.” Environmental Monitoring and Assessment 173, no. 1–4 (March 27, 2010): 625–641. doi:10.1007/s10661-010-1411-x.

[31] K. F. Widaman, “Best practices in quantitative methods for developmentalists: III. Missing data: What to do with or without them,” Monographs of the Society for Research in Child Development 71, no. 1, (April 2006): 210–211, doi: 10.1111/j.1540-5834.2006.00404.x.

[32] Bennett, Derrick A. “How Can I Deal with Missing Data in My Study?” Australian and New Zealand Journal of Public Health 25, no. 5 (October 2001): 464–469. doi:10.1111/j.1467-842x.2001.tb00294.x.

[33] Tencaliec, Patricia, Anne-Catherine Favre, Clémentine Prieur, and Thibault Mathevet. “Reconstruction of Missing Daily Streamflow Data Using Dynamic Regression Models.” Water Resources Research 51, no. 12 (December 2015): 9447–9463. doi:10.1002/2015wr017399.

[34] Lee, Hyojin, and Kwangmin Kang. “Interpolation of Missing Precipitation Data Using Kernel Estimations for Hydrologic Modeling.” Advances in Meteorology 2015 (2015): 1–12. doi:10.1155/2015/935868.

[35] Chen, Jiahua, and Jun Shao. "Jackknife Variance Estimation for Nearest-Neighbor Imputation." Journal of the American Statistical Association 96, no. 453 (March 2001): 260–269. doi:10.1198/016214501750332839.

[36] Aljuaid, Tahani, and Sreela Sasi. "Proper Imputation Techniques for Missing Values in Data Sets." International Conference on Data Science and Engineering (ICDSE) (August 2016). doi:10.1109/icdse.2016.7823957.

[37] Yang, Yiming. "An evaluation of statistical approaches to text categorization." Information retrieval 1, no. 1 (1999): 69-90. doi:10.1023/A:1009982220290.

[38] Gupta, Anjali, and Vijay Bhaskar Semwal. "Multiple Task Human Gait Analysis and Identification: Ensemble Learning Approach." Emotion and Information Processing (2020): 185–197. doi:10.1007/978-3-030-48849-9_12.

[39] Breiman, L., J. Friedman, R. Olshen, and C. Stone. "Classification and Regression Trees. New York: Wadsworth & Brooks." Pacific Grove, CA (1984).

[40] Yohai, Victor J. "High Breakdown-Point and High Efficiency Robust Estimates for Regression." The Annals of Statistics 15, no. 2 (June 1, 1987). doi:10.1214/aos/1176350366.

[41] Little, Roderick J. A., and Donald B. Rubin. "Statistical Analysis with Missing Data" (August 26, 2002). doi:10.1002/9781119013563.

[42] Van Loon, A.F., and G. Laaha. "Hydrological Drought Severity Explained by Climate and Catchment Characteristics." Journal of Hydrology 526 (July 2015): 3–14. doi:10.1016/j.jhydrol.2014.10.059.

[43] Carey, Austin M., and Ginger B. Paige. "Ecological Site-Scale Hydrologic Response in a Semiarid Rangeland Watershed." Rangeland Ecology & Management 69, no. 6 (November 2016): 481–490. doi:10.1016/j.rama.2016.06.007.

[44] Thanh, Nguyen Tien. "Evaluation of Multi-Precipitation Products for Multi-Time Scales and Spatial Distribution during 2007-2015." Civil Engineering Journal 5, no. 1 (January 27, 2019): 255. doi:10.28991/cej-2019-03091242.

[45] Khazaee Poul, Ahmad, Mojtaba Shourian, and Hadi Ebrahimi. "A Comparative Study of MLR, KNN, ANN and ANFIS Models with Wavelet Transform in Monthly Stream Flow Prediction." Water Resources Management 33, no. 8 (May 30, 2019): 2907–2923. doi:10.1007/s11269-019-02273-0.

[46] Miró, Juan Javier, Vicente Caselles, and María José Estrela. "Multiple Imputation of Rainfall Missing Data in the Iberian Mediterranean Context." Atmospheric Research 197 (November 2017): 313–330. doi:10.1016/j.atmosres.2017.07.016.

[47] Bertsimas, Dimitris, Colin Pawlowski, and Ying Daisy Zhuo. "From predictive methods to missing data imputation: an optimization approach." The Journal of Machine Learning Research 18, no. 1 (2017): 7133-7171.

[48] Chhabra, Geeta, Vasudha Vashisht, and Jayanthi Ranjan. "A Comparison of Multiple Imputation Methods for Data with Missing Values." Indian Journal of Science and Technology 10, no. 19 (June 29, 2017): 1–7. doi:10.17485/ijst/2017/v10i19/110646.

[49] Rana, Sohel, Ahamefule Happy John, and Habshah Midi. "Robust Regression Imputation for Analyzing Missing Data." 2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE) (September 2012). doi:10.1109/icssbe.2012.6396621.