Hydrological Sciences Journal

ISSN: 0262-6667 (Print) 2150-3435 (Online) Journal homepage: www.tandfonline.com/journals/thsj20

Multiple imputations by chained equations for recovering missing daily streamflow observations: a case study of Langat River basin in Malaysia

Fatimah Bibi Hamzah, Firdaus Mohamad Hamzah, Siti Fatin Mohd Razali & Ahmed El-Shafie

To cite this article: Fatimah Bibi Hamzah, Firdaus Mohamad Hamzah, Siti Fatin Mohd Razali & Ahmed El-Shafie (2022) Multiple imputations by chained equations for recovering missing daily streamflow observations: a case study of Langat River basin in Malaysia, Hydrological Sciences Journal, 67:1, 137-149, DOI: 10.1080/02626667.2021.2001471

To link to this article: https://doi.org/10.1080/02626667.2021.2001471

HYDROLOGICAL SCIENCES JOURNAL
2022, VOL. 67, NO. 1, 137–149
https://doi.org/10.1080/02626667.2021.2001471

Multiple imputations by chained equations for recovering missing daily streamflow observations: a case study of Langat River basin in Malaysia

Fatimah Bibi Hamzah[a,b], Firdaus Mohamad Hamzah[a], Siti Fatin Mohd Razali[a] and Ahmed El-Shafie[c]

[a]Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Malaysia; [b]Faculty of Computing and Multimedia, Kolej Universiti Poly-Tech Mara Kuala Lumpur, Taman Shamelin Perkasa, Malaysia; [c]Department of Civil Engineering, Faculty of Engineering, Universiti Malaya, Malaysia

ABSTRACT

Missing values in hydrological studies are a common issue for hydrologists, especially in statistical analyses as a complete dataset is required. This work evaluates the performance of the multiple imputations by chained equations (MICE) approach to predicting recurrence in streamflow datasets. To evaluate and verify the effectiveness of the MICE approach in treating missing streamflow data, complete historical daily streamflow data from 2012 to 2014 were used. Later, MICE methods coupled with multiple linear regression (MLR) were applied to restore streamflow rates in Malaysia's Langat River basin from 1978 to 2016. The best estimation methods are validated with tests such as adjusted R-squared (Adj R^2), residual standard error (RSE), and mean absolute percentage error (MAPE). The findings revealed that the classification and regression tree (CART) method combined with MLR outperformed the other approaches tested, with the highest Adj R^2 value and the lowest RSE and MAPE values observed regardless of missing conditions.

ARTICLE HISTORY
Received 12 July 2020
Accepted 15 October 2021

EDITOR
A. Castellarin

ASSOCIATE EDITOR
A. Castellarin & A. Requena

KEYWORDS
missing data; multiple imputations; chained equations; streamflow; CART; MICE

1 Introduction

One of the challenges often faced in hydrological research is missing values in a dataset. Despite the introduction of various missing data reconstruction approaches over the years, the issue of missing values that limit hydrological analysis due to the occurrence of natural disasters or improper operation and battery drainage of equipment (Mwale *et al.* 2012) remains unresolved (Mispan *et al.* 2015, Tencaliec *et al.* 2015, Hamzah *et al.* 2020). Technicalities, bad weather, device failures or tool errors during the information-gathering process, operator fault upon data entry, calibration mode and/or damage of data as a result of malfunctioning storing machinery as well as budget reductions have caused difficulties in extended hydrometric data construction and organization and, at times, gaps in the dataset arise (Johnston 1999, Gao 2017, Tencaliec 2017, Gires *et al.* 2021). Missing data are particularly observed in remote catchments where equipment failures are repaired only after significant delays following extreme events, which can be crucial for hydrological frequency analysis (Ahn 2021). The consequences of employing this kind of data are uncertainty and low efficiency of water resource management systems (Adeloye 1996).

As stated by Gill *et al.* (2007), it is an accepted practice to disregard observations with missing variable values at any given time for hydrological modelling, even though only one of the independent variables is missing. Usually, incomplete data are marked and discarded from both model construction and subsequent model testing and verification. However, this method indicates the lack of appropriate treatment of missing data that may result in bias and/or the loss of significant information, which may influence the interpretation of data, the analytical efficiency, and the scientific findings (Gill *et al.* 2007, Zhao and Long 2016, Semiromi and Koch 2019, Nor *et al.* 2020). Harvey *et al.* (2012) agreed that even very small data gaps may rule out the significant computation of essential summary statistics and hydrological indexes, such as monthly runoff totals or *n*-day minimum flows, hence restraining the analysis and explication of past flow variability. Therefore, reconstruction and treatment of missing data should be first carried out in the data preparation process, where the approach to be employed is influenced by the pattern and mechanism of the missing data (Plaia and Bondì 2006, Ahmat Zainuri *et al.* 2015, Kamaruzaman *et al.* 2017).

Little and Rubin (2002) reported that there are three types of missing data: (i) missing completely at random (MCAR), (ii) missing at random (MAR), and (iii) missing not at random (MNAR). The missing data mechanism is referred to as MCAR, and it is completely independent of the values of any variables in the dataset, whether it is missing or observed. Meanwhile, MAR can be described as the root of missing data, which is not related to the missing values but might be correlated to the observed values of other variables; and MNAR observations are not missing at random, nor are MCAR or MAR. Streamflow data imputation using the MAR assumption was performed by Gill *et al.* (2007). However, in reference to the definition by Little and Rubin (2002), missing value in a streamflow study can be determined as MCAR due to the existence of missingness in the streamflow data of an area that is not influenced by the data in that area or any other areas. A recent study by Moritz and Bartz-Beielstein (2017) revealed that MCAR and MAR imputation for time-series data were almost the same.

CONTACT Firdaus Mohamad Hamzah fir@ukm.edu.my Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600 UKM, Malaysia

Over the past few years, there has been a growing interest in reconstructing missing streamflow data using several statistical approaches (Regonda *et al.* 2013). To address the missing values problem, various data estimation methods have been proposed and extensively discussed in the literature. They range from the most basic traditional statistical methods, such as filling in missing values for given variables with mean or median values, or stations at other locations, to advanced computational techniques. Among the statistical approaches designed for reconstructing missing data, multiple imputations (MI) can be performed in a variety of circumstances using existing software packages, and it allows the researcher to apply standard complete-data analysis directly to the imputed dataset. MI can be performed by reconstructing the missing values with draws from some predictive model n times, where the obtained n completes the dataset which can be used for the analysis. The idea is to replace each missing item with two or more plausible values that represent a distribution of possibilities.

A recognized technique in performing MI is sequential regression modelling, also defined as multiple imputations by chained equations (MICE) (Su *et al.* 2011, van Buuren and Groothuis-Oudshoorn 2011). Conditional models for all variables with missing data can be indicated using the algorithm developed by Stef van Buuren. It is simpler to identify the conditional models than a plausible joint distribution of data (van Buuren 2007). Nevertheless, there is no joint distribution that matches a set of specified conditional distributions, hence there is a possibility that this process creates illogical infilling models (Gelman and Speed 1999). Regardless of this limitation, the technique is broadly used considering its flexibility and relative simplicity of implementation (van Buuren and Groothuis-Oudshoorn 2011).

Apart from MICE, various other studies on missing data imputation methods have been conducted. Recent work by Norazizi and Deni (2019) presented three methods to reconstruct missing rainfall data: artificial neural network (ANN), bootstrapping and expectation-maximization algorithm, and MICE. The findings indicated that ANN is the most preferred method, followed by MICE and then the bootstrapping and expectation-maximization algorithm method. A similar study was conducted by Zvarevashe *et al.* (2019) using the MICE approach to reconstruct missing rainfall and temperature data. The MICE approach was selected since it does not assume a normal distribution of the data and assumes data MAR. Earlier, Plaia and Bondì (2006) examined several infilling methods for environmental pollution datasets such as mean imputation, last and next observation carried forward/backward, the MICE approach, and a newly introduced single imputation method called the site-dependent effect method.

Many studies have shown better prediction performance using the MICE approach for classification or prediction models (Donders *et al.* 2006, Schmitt *et al.* 2015, Chhabra *et al.* 2017). Donders *et al.* (2006) also preferred the MICE method over single imputation for a clinical dataset since the latter resulted in very small estimated standard errors, whereas MICE provided good estimated standard errors and confidence intervals. In another study, the MICE pattern was reported to be tied to the size of the dataset when compared with other imputation methods using the MCAR assumption (Schmitt *et al.* 2015). According to Chhabra *et al.* (2017), the power of the MICE method lies in obtaining smaller standard errors and narrower confidence intervals where more accurate predicted values can be obtained, thus minimizing the bias and inefficiency considerably. Additionally, combining MICE methods with machine learning and genetic algorithms was suggested in the study to further limit the bias and inefficiency.

Although considerable research has been carried out on missing value imputation using the MICE approach in different experimental settings, only a few studies have investigated the reconstruction of missing streamflow data using the MICE approach. Other methods used for missing streamflow data imputation include the Bayesian regression model, as reported by Devineni *et al.* (2013) for rebuilding the average summer streamflow at five gauges in the Delaware River basin via eight regional tree-ring chronologies. Also, Multiple classification and regression tree (CART) or random forest approaches for missing streamflow record imputation have been recommended in some studies (Vezza *et al.* 2010, Erdal and Karakurt 2013, Karakurt *et al.* 2013, Tyralis *et al.* 2019) and the findings indicated the CART model outperformed the models derived by other classification methods concerning explained variance. Erdal and Karakurt (2013) used 35 years of measured data from the Karsßköy observation station on the Oruh River in Turkey (1968–2002) to generate three attribute combinations based on previous monthly stream flows to forecast current streamflow values. Tyralis *et al.* (2019) examine classification and regression applications in water resources, highlighting the potential of the original method and its variations, and evaluate the extent to which this method is used in a variety of applications. With the help of a reference complete dataset from another catchment, Baddoo *et al.* (2021) studied the efficacy of various missing value imputation algorithms on real-world field data with missing data whose actual values are unknown. The MICE package's approach was used to recover missing data from univariate time-series data using MI and fully conditional specification. Nevertheless, for streamflow imputation, comparison among the MICE approaches has not been performed yet, specifically the use of predictive mean matching (PMM), stochastic regression imputation (SRI), and multiple linear regression with bootstrap imputation (BOOT) to reconstruct missing streamflow data.

In hydrological research, establishing models involving multiple variables is challenging because the relationship between the variables might be interactional and nonlinear, and detecting these complexities can be an arduous task with no assurance of success. Furthermore, many variables have distributions that are difficult to capture using standard parametric models. As a result, the first goal of this study was to examine the accuracy of the MICE R-package (van Buuren and Groothuis-Oudshoorn 2011) using conditional models such as PMM, SRI, CART, BOOT, and Bayesian linear regression imputation for imputation in estimating missing flow records. Second, the performance of imputation methods in conjunction with the MLR model will be evaluated in forecasting daily streamflow values. MICE imputation could also be used to effectively impute missing streamflow data without the need

for information from neighbouring monitoring stations. The findings of this study are expected to help in the discovery of the best and finest approaches for the data imputation method, which allows for the reconstruction of complete daily streamflow datasets.

2 Area of study

The Langat River basin, depicted in Fig. 1, was chosen as the study site. The river basin is located in southern Selangor and northern Negeri Sembilan, specifically at latitude 2°40′152" to 3°16′15"N and longitude 101°19′20" to 102°1′10"E, over an area of 2394.38 km^2. This river basin, Malaysia's most urbanized river basin, is thought to compensate for the benefits of "spillover" development from Klang Valley (Noorazuan *et al.* 2003). It is one of the most important raw water resources for drinking water and other activities such as recreation, industrial applications, fishing, and agriculture (Juahir *et al.* 2008). Over the last four decades, this water source has served roughly half of Selangor's population, or approximately 1.2 million people within the basin, and has been a source of hydropower and flood control (Juahir *et al.* 2011, Puah *et al.* 2016, Mohamad Hamzah *et al.* 2019). The Langat basin was chosen as one of the major areas for economic growth in Selangor because it is home to Kuala Lumpur International Airport, West Port at Klang, the Multimedia Super Corridor (MSC), and Putrajaya (Juahir *et al.* 2010). Four sub-basins of Langat River – Kajang, Dengkil, Lui, and Semenyih – were examined in this study.

The Langat basin is influenced 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 (Yang *et al.* 2011, Memarian *et al.* 2012). There are four flow rate gauging stations in the Langat River basin: Dengkil and Kajang at Langat River, Kg. Rinching at Semenyih River, and Kg. Lui at Lui River. Table 1 depicts the characteristics of the sub-basins associated with Langat basin gauging stations governed by the Department of Irrigation and Drainage (DID).

The high-dimensional data used in this study were obtained from the DID, Ampang, Selangor, between 1978 and 2016. There were 12.5% missing values among the 56 980 data points. Widaman (2006) defines moderate data as datasets with 10–25% missing values. According to Bennett (2001), if the percentage of missing data exceeds 10%, the statistical analysis is likely to be biased. A large number of time-series observations were required to obtain a precise outline of the streamflow patterns (Tencaliec *et al.* 2015). Aside from this, the reliability of a frequency estimator of a long time series of data is extremely valuable in data analysis because it is strongly associated with sample size.

3 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. The method used for this study was a cross-validation technique for data from the year 2012 to 2014 to examine the competence of infilling methods. The period 2012–2014 was selected as the baseline due to the availability of complete data for this period. The missing daily streamflow data were simulated at random

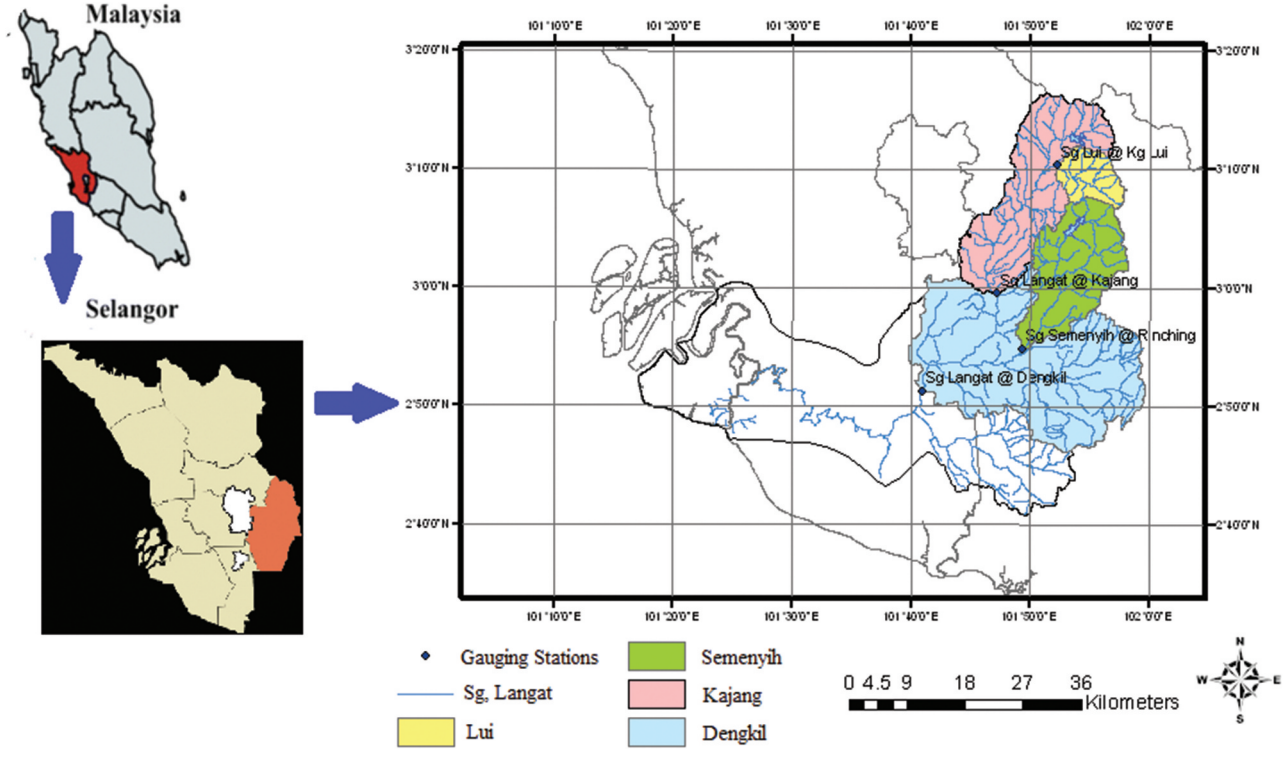

Figure 1. Map of the Langat River basin.

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

Sub-basin	Hulu Langat	Hulu Langat	Semenyih	Lui
Station number	2816441	2917401	2918401	3118445
Station name	Langat River at Dengkil	Langat River at Kajang	Semenyih River at Kg. Rinching	Lui River at Kg. Lui
District	Hulu Langat	Hulu Langat	Hulu Langat	Hulu Langat
River	Langat	Langat	Semenyih	Lui
Latitude (N)	02°59′34″	02°59′40″	02°54′55″	03°10′25″
Longitude (E)	101°47′13″	101°47′10″	101°49′25″	101°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			

Note: Data were obtained from the Department of Irrigation and Drainage (DID) of Malaysia (2018).

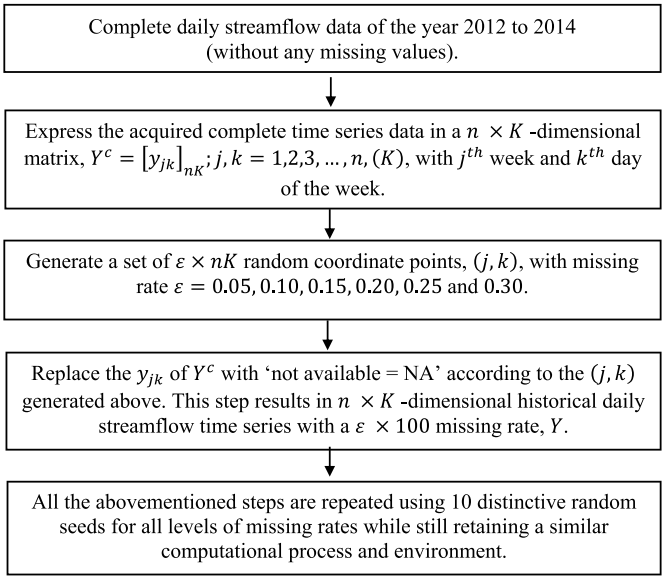

Figure 2. The procedure for introducing the missing data into the complete time series.

and extracted from the entire time-series data. Figure 2 depicts the procedure for incorporating missing data into the complete time series.

Initially, all missing values are filled in by MICE methods with replacement from the observed values, as described in White and Wood (2011). The first variable with missing values, say x_1, is regressed on all other variables $x_2, x_3, \ldots x_k$, 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, say x_2, is then regressed on all other variables $x_1, x_3, \ldots x_k$, 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 stabilize the results, the procedure is typically repeated for a number of 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. Then, the adjusted R-squared (Adj R^2), residual standard error (RSE), and mean absolute percentage error (MAPE) were calculated for each of the five predicted values. Later, MICE methods combined with MLR were used to restore streamflow rates in Malaysia's Langat River basin from 1978 to 2016.

3.1 Imputation methods

The MI method is a novel approach to dealing with missing data issues. The MI method replaces each missing value with multiple viable solutions. With the help of infilling techniques, the incomplete dataset is converted into a complete dataset that can then be analysed using any standard analysis method (van Buuren 2007). When compared to single imputation, this method accounts for the uncertainty of missing value estimation (Hamzah *et al.* 2021). The method generates a number of datasets from which parameters of interest can be estimated (Chhabra *et al.* 2017). The variance estimated in this manner is less likely to be underestimated when compared to a single imputation.

In this study, five MICE methods (PMM, SRI, BLR, CART, and BOOT) were compared as the conditional models for imputation in estimating missing flow records. Figure 3 illustrates the main steps used in MICE as suggested by van Buuren and Groothuis-Oudshoorn (2011).

The advantage of MICE is that the outcomes are calculated over relatively few iterations. As reported by van Buuren and Groothuis-Oudshoorn (2011) and Müller *et al.* (1997), five iterations are generally sufficient. Figure 4 summarizes the MICE algorithm's procedure for filling multivariate missing data.

The reconstruction method was performed by generating a prediction model for the target variable with missing values by all other variables. The response variable is the variable upon imputation, and the other relevant variables are independent variables. Equation (1) depicts the regression equation:

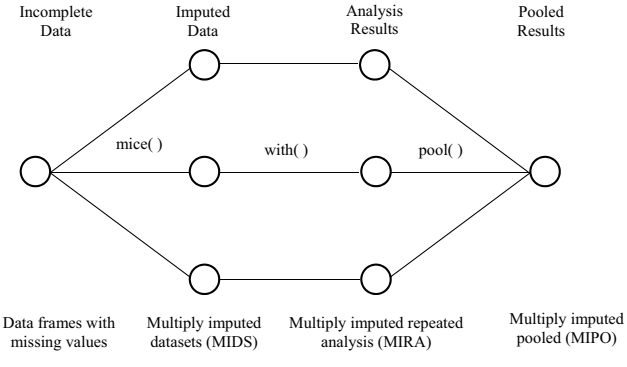

Figure 3. Schematic illustration of how MICE works.

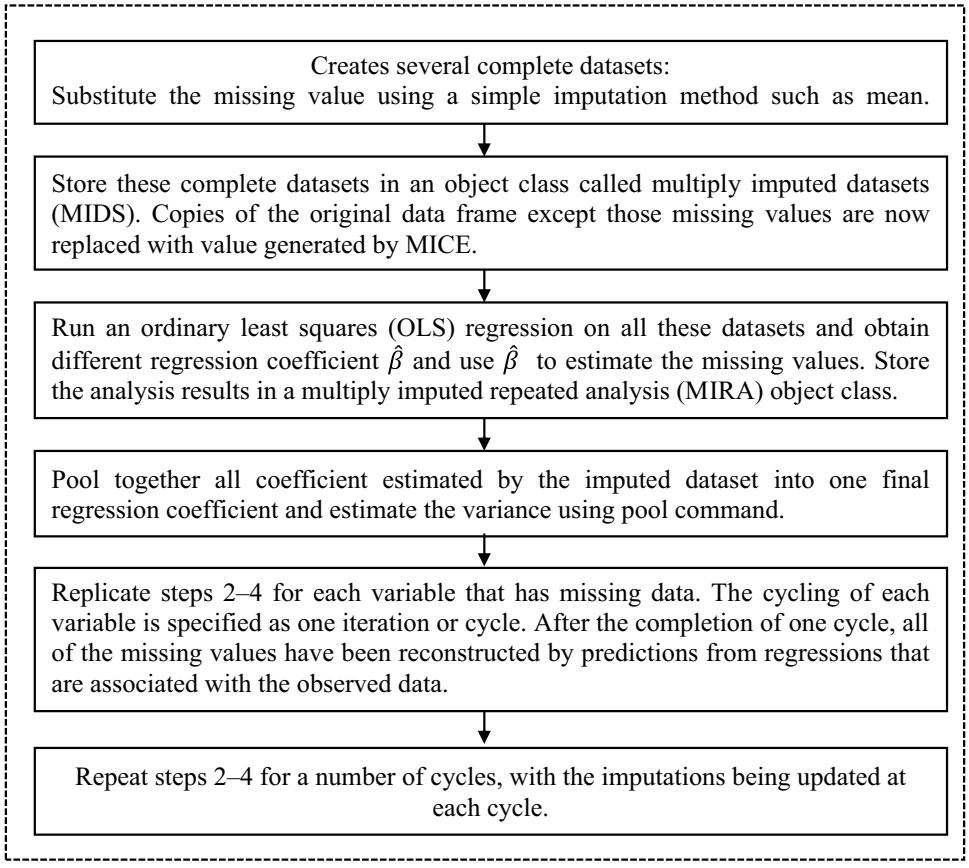

Figure 4. The procedure of the MICE algorithm.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \quad (1)$$

Let each of the k independent variables, $x_1,\ x_2, \ldots,\ x_k$, have n levels. Then, x_{ij} represents the i^{th} level of the j^{th} independent variable x_j, and $y_1,\ y_2, \ldots,\ y_n$, have n levels. Thus, n-tuples of observations were assumed to follow the same model, which were expressed as the following Equations (2) to (5).

$$y_1 = b_0 + b_1 x_{11} + b_2 x_{12} + \ldots + b_k x_{1k} + e_1 \quad (2)$$

$$y_2 = b_0 + b_1 x_{21} + b_2 x_{22} + \ldots + b_k x_{2k} + e_2 \quad (3)$$

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik} + e_i \quad (4)$$

$$\ldots\ldots\ldots\ldots$$

$$y_n = b_0 + b_1 x_{n1} + b_2 x_{n2} + \ldots + b_k x_{nk} + e_n \quad (5)$$

Equation (1) was reformatted to the following Equation (6):

$$y = X\beta + \varepsilon \quad (6)$$

where X is a $(n \times k)$ matrix of n observation on k independent variables $X_1,\ X_2, \ldots,\ X_k$, y is a $(n \times 1)$ vector of n observations of the study variable, β is a $(k \times 1)$ vector of regression coefficient and ε is the $(n \times 1)$ vector of disturbances.

Using matrix notation, these n equations can be written as the following Equation (7):

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1n} \\ 1 & x_{21} & x_{22} & \ldots & x_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} \quad (7)$$

The first column in matrix X corresponds to β_0, and the regression coefficient was expressed as Equation (8):

$$\beta = (X'X)^{-1}X'y \quad (8)$$

where X' is the transpose matrix of X.

3.1.1 Predictive mean matching

PMM is an enticing method offered for missing value substitution for quantitative variables (Chhabra *et al.* 2017). The PMM method uses the algorithm shown in Fig. 4; however, in contrast to many imputation approaches, the linear regression was not used to develop the imputed values. Instead, a metric for matching cases with missing data that are similar to the present data was discovered. A predictive distance δ_{hj} was computed, which is defined as a measure of match quality. For all j, the h observations minimizing $|\delta_{hj}|$ were designated according to Equation (9):

$$\delta_{hj} = \alpha^{mis} z_j - \alpha^{obs} z_h \quad (9)$$

In this case, let h index observations with x observed and j index observation with x missing value. For all h, the linear predictor $\alpha^{obs}z_h$ was calculated, and for all j, the linear predictor $\alpha^{mis}z_j$ was calculated. Observed values around the linear-predicted value were selected as the donor pool. Often, the donor pool is set to consist of k candidate donors which were randomly selected.

The main question to solve in PMM is how many cases (k) need to be in each matching set. There are three extensive methods for setting out the donor pool. The first is to use a fixed number of donors k, where $k = 5$ is specified when using some software. In general, individual cases with incomplete data on x will be paired to the five complete cases that have the closest predicted values. One of the five complete cases is selected randomly and its x value is assigned to the missing data case. The second method is to define δ_{max} where any h for $|\delta_{hj}| < \delta_{max}$ is in the donor pool for j, which is known as "caliper matching". The third method is to use $k = n_h$, the number of observations for which x is observed, with the possibility of selecting the observed value with a small d_{hj}.

3.1.2 Stochastic regression imputation

The SRI method is analogous to regression imputation, in which missing values are estimated by regressing other related variables in the same dataset with a random residual value (Jamil 2012). In other words, SRI entails introducing random error into the regression prediction obtained through regression imputation. Equation (10) shows the estimated missing value:

$$\dot{y} = \hat{\beta}_0 + X_{mis}\hat{\beta}_1 + \dot{\in} \tag{10}$$

where $\dot{\in}$ is drawn at random from the normal distribution as $\dot{\in} \sim N(0, \hat{\sigma}^2)$. Any values drawn are denoted by a dot above the symbol.

The SRI can lower the bias with an additional stage of augmenting for a separately predicted score with a residual term which is normally distributed with a mean of zero and a variance equal to the residual variance from the regression of the predictor on the outcome. Zero mean is important as a non-bias condition and variance should be set equal to the error variance. Using the assumption $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2 I$, the distribution of ε_i, conditional on x_i, satisfies the properties for all values of X where x_i denotes the i^{th} row of X, as demonstrated in Equations (11) and (12).

Let $p(\varepsilon_i|x_i)$ be the conditional probability density function of ε_i given x_i and $p(\varepsilon_i)$ be the unconditional probability density function of ε_i. Then,

$$\begin{aligned} E(\varepsilon_i|x_i) &= \int \varepsilon_i p(\varepsilon_i|x_i) d\varepsilon_i \\ &= \int \varepsilon_i p(\varepsilon_i) d\varepsilon_i \\ &= E(\varepsilon_i) \\ &= 0 \end{aligned} \tag{11}$$

and

$$\begin{aligned} E(\varepsilon_i^2|x_i) &= \int \varepsilon_i^2 p(\varepsilon_i|x_i') d\varepsilon_i \\ &= \int \varepsilon_i^2 p(\varepsilon_i) d\varepsilon_i \\ &= E(\varepsilon_i^2) \\ &= \sigma^2 \end{aligned} \tag{12}$$

In the case that ε_i and x_i' are independent, then $p(\varepsilon_i|x_i') = p(\varepsilon_i)$. We chose to include a random normal deviate scaled by the estimated streamflow's standard error.

3.1.3 Bayesian linear regression imputation

In BLR, linear regression is expressed by probability distribution instead of point estimates. The response, y, is assumed to be drawn from a probability distribution rather than being computed as a single value (Kim and Lee 2009). The BLR model is as follows:

$$\dot{y} = \dot{\beta}_0 + X_{mis}\dot{\beta}_1 + \dot{\in}, \tag{13}$$

where $\dot{\in} \sim N\left(0, \dot{\sigma}^2\right)$ and $\dot{\beta}_0$, $\dot{\beta}_1$ and $\dot{\sigma}^2$ are random draws from their posterior distribution of data. The matrix notation is expressed as the following Equation (14):

$$y \sim N(X\beta, \sigma^2 I) \tag{14}$$

where $y = (y_1,\ y_2, \ldots,\ y_n)'$, $\beta = (\beta_1,\ \beta_2, \ldots,\ \beta_k)'$, and X is a $(n \times k)$ matrix of n observation on k explanatory variables $x_i = (x_{i1},\ x_{i2}, \ldots,\ x_{ik})$.

In addition to the response derived from a probability distribution, the model parameters are also expected to be derived from the distribution. The model parameters' posterior probability is dependent on the training inputs and outputs. Assume the parameter's standard non-informative prior probability is

$$P(\beta, \sigma^2) \propto 1/\sigma^2 \tag{15}$$

The posterior probability of the model parameters is then provided by

$$\begin{aligned} &P(\beta, \sigma^2|y) = P(\beta, \sigma^2|y)P(\sigma^2|y) \\ &\beta|\sigma^2, y \sim N\left(\widehat{\beta}, \sigma^2 V_\beta\right) \\ &\sigma^2|y \sim Inv - Gamma\left([n-k]/2, [n-k]s^2/2\right) \\ &\beta|y = t_{n-k}\left(\hat{\beta}, s^2 V_\beta\right) \end{aligned} \tag{16}$$

The ordinary least squares estimator (OLSE) of β is obtained by minimizing $(y - X\beta)'(y - X\beta)$ with respect to β as shown in Equation (17):

$$\hat{\beta} = (X'X)^{-1}X'y \tag{17}$$

and an estimator of σ^2 is obtained as

$$\sigma^2 = \frac{1}{n-k}\left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right) \tag{18}$$

and the variance of the OLS estimate β is

$$V_\beta = \sigma^2(X'X)^{-1} \tag{19}$$

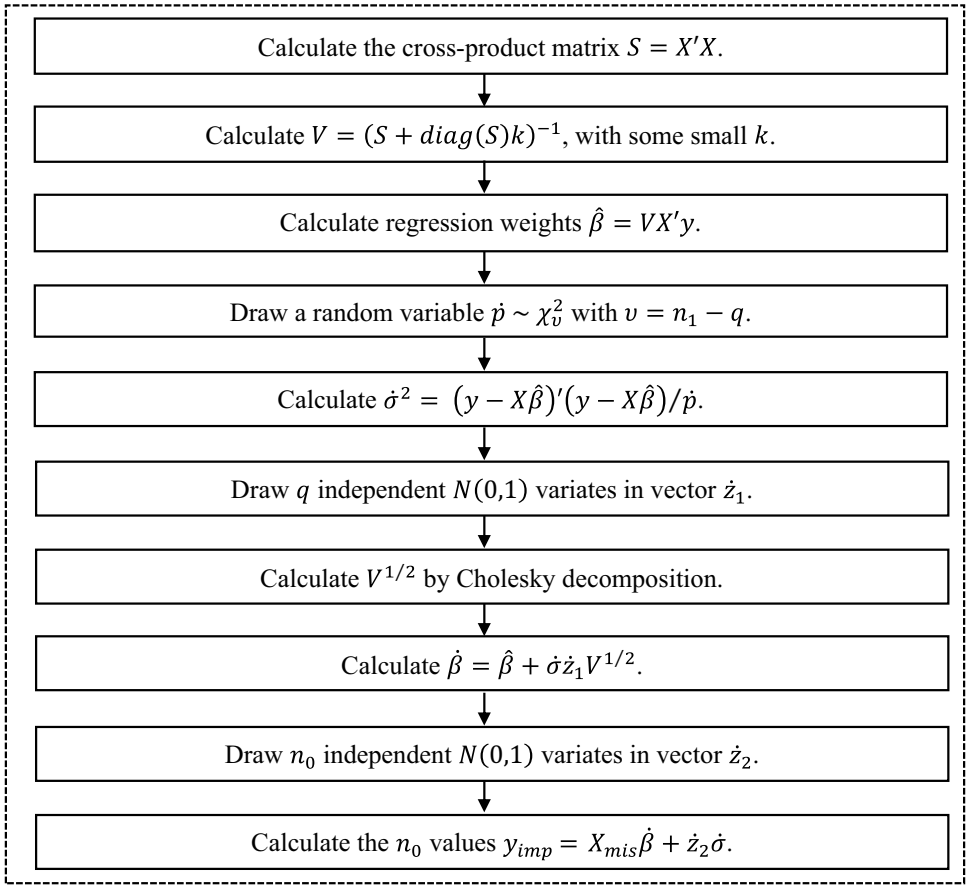

Figure 5. The procedure of the Bayesian imputation algorithm.

The posterior probability distribution is proper if $n > k$ and $rank(X) = k$. The procedure for using the BLR algorithm to reconstruct missing data is summarized in Fig. 5.

The algorithm employs a ridge parameter k to avoid difficulties with singular matrices. This number shall be fixed to a non-negative value narrow to zero, e.g. $k = 0.0001$. Larger k may be required for some data. A larger value of k, such as $k = 0.1$, is more likely to cause a systematic bias towards the null, and should be avoided.

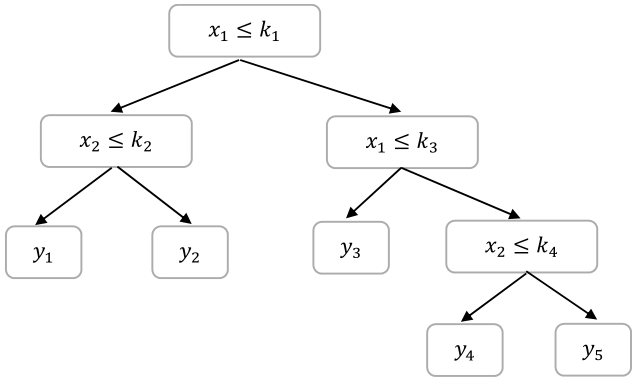

Figure 6. Multiple classification and regression tree structure.

3.1.4 Multiple classification and regression tree

The CART method, as introduced by Breiman *et al.* (1984), is one of a prominent class of machine learning algorithms using a concept shown in Fig. 6. To divide the sample CART models demand predictors, and cut points in the predictors were used. The cut points were used to divide the sample into larger homogeneous subsamples. The dividing operation reiterated on both subsamples enabled a series of splits that sets out a binary tree (Erdal and Karakurt 2013). Each vertex in the tree has a splitting rule, which is determined by minimizing the relative error (RE), which represents the sum of squares of the split for the regression problem:

$$RE(d) = \sum_{l=0}^{L} \left(y_l - \bar{y}_L\right)^2 + \sum_{r=0}^{R} \left(y_r - \bar{y}_R\right)^2 \qquad (20)$$

where y_l and y_r are the left and right partitions, respectively, with L and R observations of y in each, with respective means $\bar{y}_L$ and $\bar{y}_R$. The decision rule d is a point in the estimator variable x that specifies the left and right branches. The partitioning rule that minimizes the RE was then used to construct a tree vertex.

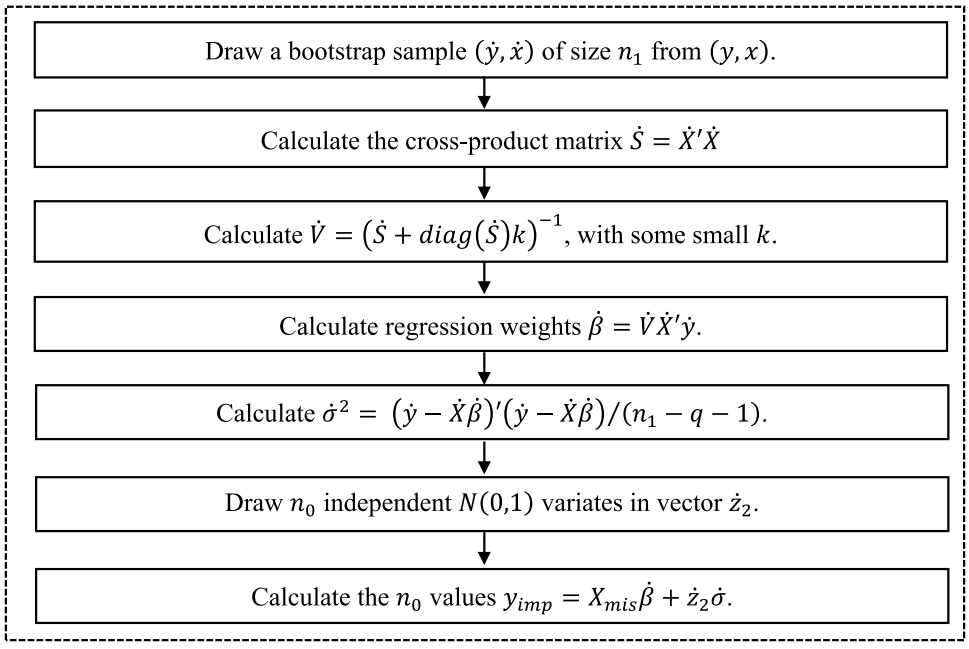

Figure 7. The procedure of the bootstrap imputation algorithm.

3.1.5 Multiple linear regression with bootstrap imputation classification

The bootstrap is a common technique used in quantifying variability by re-sampling the data (Chhabra *et al.* 2017). It applies any test or metric rooted in random sampling with substitution. The estimated missing value is predicted using Equation (21):

$$\dot{y} = \dot{\beta}_0 + X_{mis}\dot{\beta}_1 + \dot{\in} \tag{21}$$

where $\dot{\in} \sim N\left(0, \dot{\sigma}^2\right)$ and $\dot{\beta}_0$, $\dot{\beta}_1$ and $\dot{\sigma}^2$ are the least-squares estimates computed after a bootstrap sample was selected from the observed data. The procedure used by the BOOT algorithm for the reconstruction of multivariate missing data is summarized in Fig. 7.

An algorithm as in Fig. 7 calculates imputations by drawing a bootstrap sample from the fill-out part of the data and then estimates the least squares given the bootstrap sample as a "draw" that embeds sampling variability into the parameters (Heitjan and Rubin 1990). In comparison to the Bayesian technique, the bootstrap approach avoids the Cholesky decomposition, and it is not necessary to draw from the χ^2 distribution.

3.1.6 Multiple linear regression

Following the replacement of all missing values with various techniques, the datasets in their entirety are analysed 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 (Van Loon and Laaha 2015). The MLR model is a popular statistical method in many fields, including hydrology (Campozano *et al.* 2014, Carey and Paige 2016). The MLR model parameter is expressed as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i(\beta), i = 1, \ldots, N \tag{22}$$

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 Estimation of the methods' performance

Verification of the influence of missing data imputation on the streamflow dataset was performed using three performance criteria. The Adj R^2, RSE, and MAPE were calculated to evaluate imputation methods. The error measures the deviation between the estimated values and their corresponding observed values. The Adj R^2 value is the R^2 value adjusted for the number of independent variables in the model. The Adj R^2 values range from 0 to 1 and indicate the strength of the relationship among observations and estimates, where the higher value estimates the best performance of estimation methods. If the Adj R^2 approaches zero, the model performance is believed to be inadmissible or poor. In contrast, the model prediction is believed to be perfect if the values are close to one (Mispan *et al.* 2015, Rahman *et al.* 2015). Meanwhile, lower RSE and MAPE values correspond to better performance of the estimation methods. These statistics are calculated following Equations (22)–(24):

$$Adj\ R^2 = \bar{R}^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right] \tag{23}$$

$$RSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n-k-1}} \tag{24}$$

Table 2. Error of streamflow reference model.

Year	Adj R^2	RSE	MAPE
2012–2014	0.653	0.346	0.468

Table 3. The performance of six different percentages of missing data compared based on Adj R^2.

Methods	Missing data rate					
	5%	10%	15%	20%	25%	30%
PMM	0.615	0.681	0.657	0.671	0.675	0.668
SRI	0.634	0.668	0.629	0.652	0.667	0.672
BLR	0.631	0.670	0.643	0.661	0.666	0.673
CART	**0.637**	**0.691**	**0.669**	**0.681**	**0.685**	**0.684**
BOOT	0.606	0.651	0.621	0.640	0.650	0.656

Note: Bold values indicate a good model.

Table 4. The performance of six different percentages of missing data compared based on RSE.

Methods	Missing data rate					
	5%	10%	15%	20%	25%	30%
PMM	0.383	0.319	0.342	0.328	0.324	0.331
SRI	0.365	0.331	0.370	0.347	0.332	0.327
BLR	0.368	0.329	0.356	0.338	0.333	0.326
CART	**0.362**	**0.318**	**0.330**	**0.318**	**0.314**	**0.315**
BOOT	0.373	0.348	0.348	0.359	0.349	0.344

Note: Bold values indicate a good model.

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

Methods	Missing data rate					
	5%	10%	15%	20%	25%	30%
PMM	0.515	**0.450**	0.456	0.501	0.471	0.439
SRI	0.858	0.655	0.688	0.901	1.089	1.289
BLR	0.554	0.541	0.876	0.862	0.836	0.838
CART	**0.476**	**0.450**	**0.427**	**0.501**	**0.467**	**0.427**
BOOT	0.630	0.897	1.106	1.130	1.171	0.905

Note: Bold values indicate a good model.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i-\hat{x}_i|}{x_i} \quad (25)$$

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.

4 Results and discussion

Evaluation of MICE methods to identify the best imputation technique for recovering missing streamflow data is carried out in this study. The models were first tested on the training dataset covering the period 2012–2014 without missing values. 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 replaced with new values obtained using each MICE method discussed earlier. The error was calculated by subtracting the predicted value of the trained model from the predicted value of the reference model and the data obtained using the missing value replacement method. The model trained with the original training data and test data with no missing values is referred to as the reference model. The smaller the difference between the estimated and observed values, the smaller the RSE and MAPE values. If the estimated value matches the observed value, the Adj R^2 value will be close to one. The best-fit method will be chosen based on the highest Adj R^2 value and the lowest RSE and MAPE values. Table 2 shows the prediction model errors, while Tables 3–Table 5 show the deviation results.Tables 4

Gap analysis is used to determine which imputation method is more consistent, as evidenced by smaller gaps between training and validation results. From the results summarized in Tables 2–Table 5, it can be observed that the CART method produced the highest Adj R^2 with the lowest RSE and MAPE values. Meanwhile, BOOT was the worst imputation method for daily streamflow data in Malaysia's Langat River basin, with the lowest Adj R^2 and highest RSE and MAPE values. Adj R^2 values, on the other hand, revealed that all of the imputation methods yield acceptable results, with values close to one and differing by less than 10% from the training set (Pham 2019), whereas RSE produces slightly lower values than the training sets as the missing data rate increases. Meanwhile, the MAPE measures the magnitude of the error in percentage terms, and the values vary slightly depending on the mean difference between the observed known outcome values and the values predicted by the model.

As can be observed, model accuracies do not decrease as the missing data rate increases. A possible explanation for the efficiency gain with the MICE method is that it is able to make better use of the available information by accommodating nonlinearities among the predictors (Islam Khan and Hoque 2020). The MICE method is recognized for its simplicity, robustness, ability to handle multicollinearity and skewed distributions, and flexibility to suit interactions and nonlinear relations (van Buuren and Groothuis-Oudshoorn 2011). With increasing missing data rates, the error between the reference and validation models with missing data imputation grows. This indicated a small error when training data was used with no missing values. Even if the missing data rate was only 30%, the training model followed the pattern of the remaining 70% training data rather than the 30% missing data. As a result, despite the presence of missing values in the training data, a significant error went unnoticed.

From the results obtained, the CART method resulted in better performance in comparison to the other four methods; PMM, SRI, BLR, and BOOT. CART resulted in the highest Adj R^2 value, as can be seen in Table 3. Amongst the five methods, the poorest performance was found when the BOOT method was used, where the lowest Adj R^2 was determined irrespective of the percentage of missing values. and Table 5 present the performance indicator concerning the RSE and MAPE, respectively. The best performance with the lowest RSE was noted when the CART method was used, followed by the PMM method. This finding is in agreement with previous studies (Vezza *et al.* 2010, Erdal and Karakurt 2013, Karakurt *et al.* 2013, Tyralis *et al.* 2019), where the CART model outperformed other classification algorithms in terms of explained variance.

The results from Erdal and Karakurt (2013) indicate that the CART model is a promising technique for monthly streamflow forecasting and yields better results than the other models evaluated. Karakurt *et al.* (2013) conclude that a classification- and regression-based model slightly outperformed the conventional ANN, with R^2 values of 0.8998 and 0.8942, respectively. The CART approach was also studied by Vezza *et al.* (2010),who used a one-way analysis of variance to estimate the explained variance for the CART classification, yielding a result of 69%. This finding leads to the conclusion that the CART approach is an excellent classification method able to find distinct groups in terms of both low-flow catchment response and catchment characteristics. The CART model may also provide variable importance metrics, which distinguishes it from the general class of data-driven models that are only focused on predictive modelling (Tyralis *et al.* 2019).

The BOOT approach, on the other hand, gave the poorest performance and was also the most time-consuming for large datasets. Meanwhile, the SRI method, which is often regarded as a conservative and safe approach to dealing with missing data, underestimated the variance as the rate of missing data increased. This is because the SRI method can produce implausible results. Variables in streamflow data are typically delimited to specific intervals (e.g. remain positive), and the SRI method cannot reconstruct missing data based on such constraints. The BLR imputation method, similarly, assumes that a random error has a similar mean for all variables in the distribution, resulting in extremely small or large errors for the imputed values. Overall, the BLR method performed poorly compared to the PMM and CART approaches. In contrast, the PMM approach was not affected greatly by missing data up to 30%, which could be due to the imputation predicated based on valid values observed elsewhere. It generates imputed values that are considerably more similar to actual values that use "borrow" concepts from individuals with real data (Schenker and Taylor 1996). The spread of imputed values hence lies between the minimum and the maximum of the observed values. Imputation will not occur outside of the observed data range, avoiding issues with pointless imputations (e.g. the non-positive value of streamflow). Despite the fact that Dong and Peng (2013) suggested PMM as one of the best and most practical imputation methods for continuous missing variables, it lacks a theoretical foundation and has no explicit formulation as an optimization problem (Bertsimas

Table 6. Adj R^2, RSE, and MAPE values for three imputation methods on average.

Method	Adj R^2	RSE	MAPE
PMM	0.691	0.442	0.543
SRI	0.601	0.479	1.265
BLR	0.638	0.458	1.222
CART	**0.725**	**0.387**	**0.487**
BOOT	0.553	0.511	1.289

Note: Bold values indicate a good model.

Table 7. The results for MLR when combined with imputation methods.

Method	Adj R^2	RSE	MAPE
PMM-MLR	0.628	0.500	0.759
SRI-MLR	0.584	0.534	1.801
BLR-MLR	0.561	0.551	1.172
CART-MLR	**0.777**	**0.479**	**0.591**
BOOT-MLR	0.312	0.562	1.884

Note: Bold values indicate a good model.

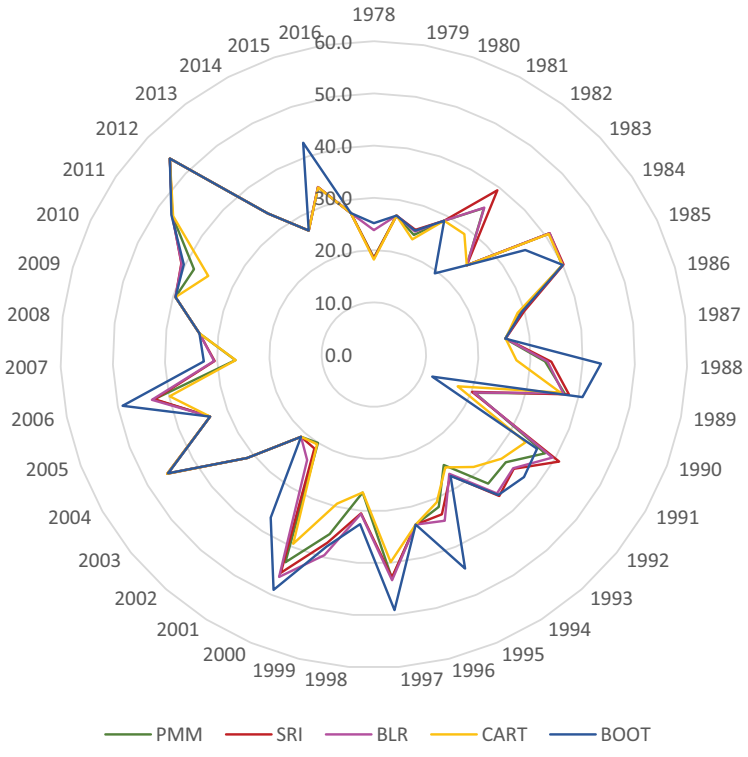

Figure 8. Visualization of imputed values using PMM, SRI, BLR, CART, and BOOT.

et al. 2018). The performance of the five imputation methods in terms of MAPE was comparable to their Adj R^2 and RSE. The CART method outperformed the other methods studied, by virtue of its lowest RSE and MAPE values and highest Adj R^2, regardless of any missing data.

The models were later validated using data from 1978 to 2016 for all four sub-basins. The results were then computed as an average of the results of each imputation method's outcome. Table 6 displays the results of the overall performance of the methods in the reconstruction of data from 1978 to 2016. Based on Table 6, CART performed the best. Meanwhile, BOOT was the worst imputation method for daily streamflow data in Malaysia's Langat River basin, with the lowest Adj R^2 and highest RSE and MAPE. Table 6 also shows that the PMM imputation method has a higher Adj R^2 and lower RSE and MAPE values than the other four methods, putting it on par with CART.

After the missing values were filled in, the MLR model was used to analyse the entire dataset in this study. The MLR model was used to identify the best approaches for dealing with missing data when imputation values were combined with modelling. Table 7 shows the performance of all five imputation methods in conjunction with the MLR model in forecasting streamflow rates in Malaysia's Langat River basin from 1978 to 2016. Based on Table 7, the CART-MLR method presented the best performance, whereas BOOT-MLR showed the poorest performance, among the five methods evaluated. Although the PMM-MLR approach performed slightly better compared to SRI-MLR, BLR-MLR and BOOT-MLR methods, the CART-MLR method outperformed other approaches. Finally, for visual inspection, the predicted values for all models were plotted. Figure 8 depicts the results for the five imputation methods used to replace 7124 missing daily streamflow data points in Malaysia's Langat River basin. Figure 8 depicts how the imputed values of daily streamflow data from all five methods followed similar trends. All models, for example, reacted to streamflow events with peaks of similar magnitude and timing.

Findings from this study show that the CART method coupled with MLR significantly outperformed the other methods tested, with the lowest RSE and MAPE and the highest Adj R^2 value. This shows the error derived with the CART technique was comparatively lower than that compared to the PMM, SRI, BLR, and BOOT techniques, since the error rate was mirrored by the missing data rate. CART, as a tree-based ensemble model, can reasonably increase its accuracy by generating many replica datasets and developing various models with lower bias and then integrating them in the construction of a higher-performing ensemble model (Erdal and Karakurt 2013, Tyralis *et al.* 2019). These findings are consistent with those reported in the literature, confirming the recommendations of the CART imputation method (De'ath and Fabricius 2000, Erdal and Karakurt 2013, Karakurt *et al.* 2013). Conclusively, these simulations demonstrate that the CART technique coupled with MLR is the best missing data imputation method for reconstructing missing streamflow data.

5 Limitations and directions for future research

This study is based on the performance of MICE as the conditional model for data imputation in estimating missing flow records, with several limitations. The data matrices of Langat River basin with four gauging stations were analysed. However, other critical factors that contribute to streamflow characteristics, such as rainfall, temperature, topography or other parameters of the study area, were not investigated due to unavailability of the data. Disregarding such parameters may lead to inaccuracy in predicting missing data, but the use of the MICE technique has provided a simple and fast way to estimate the missing data.

MICE is an increasingly popular method of analysis. However, like any powerful statistical technique, it must be used with caution. The main methodological limitation of the MICE procedure is that it lacks a clear theoretical rationale, and the conditional regression models may be incompatible. Future studies may include investigation of other imputation methods, such as nearest neighbours, principal component analysis, and artificial neural network, to equate the performance of the proposed imputation method in this study. It may also be beneficial to perform a sensitivity analysis using various methods for dealing with missing data in order to assess the robustness of the results.

6 Conclusion

Missing data always leads to misinterpretation of the statistical output, so the method used to fill the gaps in a dataset should be carefully considered. Several techniques for managing missing data have been proposed in the literature, and the choice of a suitable approach is still unclear, including the missing data pattern and the missing data mechanism. Imputation methods have reduced information loss, which could have resulted in sub-optimal outcomes and misleading conclusions, such as the risk estimation of an extreme event.

The results of this study showed that the CART method was consistently superior, regardless of the percentage of missing values. All three performance indicators agreed that the CART method is among the best, with a higher Adj R^2 and lower RSE and MAPE compared to the other MICE-introduced methods. The results also revealed that the CART method produced the smallest difference between the reference and prediction models with missing data imputation. As a result, the best results were obtained by processing missing streamflow data using CART in conjunction with MLR. Finally, this research contributes to the accurate filling of the missing streamflow dataset.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Geran Universiti Penyelidikan GUP-2020-013.

ORCID

Siti Fatin Mohd Razali http://orcid.org/0000-0003-2757-6141

References

Adeloye, A.J., 1996. An opportunity loss model for estimating the value of streamflow data for reservoir planning. *Water Resources Management*, 10 (1), 45–79. doi:10.1007/BF00698811.

Ahmat Zainuri, N., Aziz Jemain, A., and Muda, N., 2015. A comparison of various imputation methods for missing values in air quality data. *Sains Malaysiana*, 44 (3), 449–456. doi:10.17576/jsm-2015-4403-17.

Ahn, K.H., 2021. Streamflow estimation at partially gaged sites using multiple-dependence conditions via vine copulas. *Hydrology and Earth System Sciences*, 25 (8), 4319–4333. doi:10.5194/hess-25-4319-2021.

Baddoo, T.D., *et al.*, 2021. Comparison of missing data infilling mechanisms for recovering a real-world single station streamflow observation. *International Journal of Environmental Research and Public Health*, 18 (16), 8375. doi:10.3390/ijerph18168375.

Bennett, D.A., 2001. How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25 (5), 464–469. doi:10.1111/j.1467-842X.2001.tb00294.x.

Bertsimas, D., Pawlowski, C., and Zhuo, Y.D., 2018. From predictive methods to missing data imputation: an optimization approach. *Journal of Machine Learning Research*, 18 (2018), 1–39. Available from: http://jmlr.org/papers/v18/17-073.html

Breiman, L., *et al.*, 1984. *Classification and regression trees*. New York: Wadsworth Publishing.

Campozano, L., *et al.*, 2014. Evaluation of infilling methods for time series of daily precipitation and temperature: the case of the Ecuadorian Andes. *Maskana*, 5 (1), 99–115. doi:10.18537/mskn.05.01.07.

Carey, A.M. and Paige, G.B., 2016. Ecological site-scale hydrologic response in a semiarid rangeland watershed. *Rangeland Ecology and Management*, 69 (6), 481–490. doi:10.1016/j.rama.2016.06.007.

Chhabra, G., Vashisht, V., and Ranjan, J., 2017. A comparison of multiple imputation methods for data with missing values. *Indian Journal of Science and Technology*, 10 (19), 1–7. doi:10.17485/ijst/2017/v10i19/110646.

De'ath, G. and Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81 (11), 3178–3192. doi:10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2.

Devineni, N., *et al.*, 2013. A tree-ring-based reconstruction of Delaware River basin streamflow using hierarchical Bayesian regression. *Journal of Climate*, 26 (12), 4357–4374. doi:10.1175/JCLI-D-11-00675.1.

Donders, A.R.T., *et al.*, 2006. Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59 (10), 1087–1091. doi:10.1016/j.jclinepi.2006.01.014.

Dong, Y. and Peng, C.-Y.J., 2013. Principled missing data methods for researchers. *SpringerPlus*, 2 (1), 1–17. doi:10.1186/2193-1801-2-222.

Erdal, H.I. and Karakurt, O., 2013. Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. *Journal of Hydrology*, 477 (2013), 119–128. doi:10.1016/j.jhydrol.2012.11.015.

Gao, Y., 2017. *Dealing with missing data in hydrology - data analysis of discharge and groundwater time-series in Northeast Germany*. Germany: Freie Universität Berlin.

Gelman, A. and Speed, T.P., 1999. Corrigendum: characterizing a joint probability distribution by conditionals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61 (2), 483. doi:10.1111/1467-9868.00189.

Gill, M.K., *et al.*, 2007. Effect of missing data on performance of learning algorithms for hydrologic predictions: implications to an imputation technique. *Water Resources Research*, 43 (7), 1–12. doi:10.1029/2006WR005298.

Gires, A., Tchiguirinskaia, I., and Schertzer, D., 2021. Infilling missing data of binary geophysical fields using scale invariant properties through an application to imperviousness in urban areas. *Hydrological Sciences Journal*, 66 (7), 1197–1210. doi:10.1080/02626667.2021.1925121.

Hamzah, F.B., *et al.*, 2020. Imputation methods for recovering streamflow observation : a methodological review. *Cogent Environmental Science*, 6 (1), 21. doi:10.1080/23311843.2020.1745133.

Hamzah, F.B., *et al.*, 2021. A comparison of multiple imputation methods for recovering missing data in hydrological studies. *Civil Engineering Journal*, 7 (9), 1608–1619. doi:10.28991/cej-2021-03091747.

Harvey, C.L., Dixon, H., and Hannaford, J., 2012. An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK. *Hydrology Research*, 43 (5), 618–637. doi:10.2166/nh.2012.110.

Heitjan, D.F. and Rubin, D.B., 1990. Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85 (410), 304–314. doi:10.1080/01621459.1990.10476202.

Islam Khan, S. and Hoque, A.S.M.L., 2020. SICE: an improved missing data imputation technique background and related works. *Journal of Big Data*, 7 (1), 37. doi:10.1186/s40537-020-00313-w.

Jamil, J.M., 2012. *Partial least squares structural equation modelling with incomplete data: an investigation of the impact of imputation methods*. The University of Bradford.

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

Juahir, H., *et al.*, 2008. 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 (1), 258–265.

Juahir, H., *et al.*, 2010. Hydrological trend analysis due to land use changes at langat river basin. *EnvironmentAsia*, 3 (SPECIAL ISSUE), 20–31. doi:10.14456/ea.2010.61.

Juahir, H., *et al.*, 2011. Spatial water quality assessment of Langat River Basin (Malaysia) using environmetric techniques. *Environmental Monitoring and Assessment*, 173 (1–4), 625–641. doi:10.1007/s10661-010-1411-x.

Kamaruzaman, I.F., Wan Zin, W.Z., and Mohd Ariff, N., 2017. A comparison of method for treating missing daily rainfall data in Peninsular Malaysia. *Malaysian Journal of Fundamental and Applied Sciences*, 13 (4), 375–380. (Special Issue on Some Advances in Industrial and Applied Mathematics). doi:10.11113/mjfas.v13n4-1.781.

Karakurt, O., *et al.*, 2013. Comparing ensembles of decision trees and neural networks for one-day-ahead stream flow predict. *Scientific Research Journal*, 1 (15), 43–54. doi:10.9780/23218045/1172013/41.

Kim, S.U. and Lee, K.S., 2009. Regional low flow frequency analysis using Bayesian regression and prediction at ungauged catchment in Korea. *KSCE Journal of Civil Engineering*, 14 (1), 87–98. doi:10.1007/s12205-010-0087-7.

Little, R.J.A. and Rubin, D.B. 2002. Statistical analysis with missing data, hlm. 2nd Edisi. Hoboken, New Jersey: John Wiley & Sons, Inc. doi:10.1002/9781119013563.

Memarian, H., *et al.*, 2012. Trend analysis of water discharge and sediment load during the past three decades of development in the Langat basin, Malaysia. *Hydrological Sciences Journal*, 57 (6), 1207–1222. doi:10.1080/02626667.2012.695073.

Mispan, M.R., *et al.*, 2015. Missing river discharge data imputation approach using artificial neural network. *ARPN Journal of Engineering and Applied Sciences*, 10 (22), 10480–10485.

Mohamad Hamzah, F., Mohd Yusoff, S.H., and Jaafar, O., 2019. L-moment-based frequency analysis of high-flow at the Sungai Langat, Kajang, Selangor, Malaysia. *Sains Malaysiana*, 48 (7), 1357–1366. L-Moment-Based. doi:10.17576/jsm-2019-4807-05.

Moritz, M.S. and Bartz-Beielstein, T., 2017. imputeTS: time series missing value imputation in R. *The R Journal*, 9 (1), 207–218. doi:10.32614/RJ-2017-009.

Müller, K.-R., *et al*. 1997. Predicting time series with support vector machines. In: W. Gerstner, *et al*. (eds) *Artificial Neural Networks — ICANN'97. ICANN 1997. Lecture Notes in Computer Science*. Berlin Heidelberg: Springer. doi:10.1007/bfb0020283

Mwale, F.D., Adeloye, A.J., and Rustum, R., 2012. Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi - A self organizing map approach. *Physics and Chemistry of the Earth*, 50–52 (2012), 34–43. doi:10.1016/j.pce.2012.09.006.

Noorazuan, M., *et al*. 2003. GIS application in evaluating land use-land cover change and its impact on Hydrological regime in Langat River Basin, Malaysia. *In*: *Proceedings of the Conference MapAsia 2003*, Malaysia, Kuala Lumpur. February.

Nor, S.M.C.M., *et al*., 2020. A comparative study of different imputation methods for daily rainfall data in east-coast Peninsular Malaysia. *Bulletin of Electrical Engineering and Informatics*, 9 (2), 635–643. doi:10.11591/eei.v9i2.2090.

Norazizi, N.A.A. and Deni, S.M. 2019. Comparison of Artificial Neural Network (ANN) and other imputation methods in estimating missing rainfall data at Kuantan Station. *Soft Computing in Data* Science, *5th International Conference, SCDS 2019*, hlm, Singapore. Iizuka, Japan: Springer, 298–308. doi:10.1007/978-981-15-0399-3_24.

Pham, H., 2019. A new criterion for model selection. *Mathematics*, 7 (12), 12. doi:10.3390/MATH7121215.

Plaia, A. and Bondì, A.L., 2006. Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*, 40 (38), 7316–7330. doi:10.1016/j.atmosenv.2006.06.040.

Puah, Y.J., *et al*., 2016. River catchment rainfall series analysis using additive Holt – winters method. *Journal of Earth System Science*, 125 (2), 269–283. doi:10.1007/s12040-016-0661-6.

Rahman, N.F.A., *et al*., 2015. Semi distributed hydro climate model; The Xls2NCascii program approach for weather generator. *ARPN Journal of Engineering and Applied Sciences*, 10 (15), 6619–6622.

Regonda, S.K., *et al*., 2013. Short-term ensemble streamflow forecasting using operationally-produced single-valued streamflow forecasts - A Hydrologic Model Output Statistics (HMOS) approach. *Journal of Hydrology*, 497 (2013), 80–96. doi:10.1016/j.jhydrol.2013.05.028.

Schenker, N. and Taylor, J.M.G., 1996. Partially parametric techniques for multiple imputation. *Computational Statistics and Data Analysis*, 22 (4), 425–446. doi:10.1016/0167-9473(95)00057-7.

Schmitt, P., Mandel, J., and Guedj, M., 2015. A comparison of six methods for missing data imputation. *Journal of Biometrics & Biostatistics*, 6 (1), 1–6. doi:10.4172/2155-6180.1000224.

Semiromi, M.T. and Koch, M., 2019. Reconstruction of groundwater levels to impute missing values using singular and multichannel spectrum analysis: application to the Ardabil Plain, Iran. *Hydrological Sciences Journal*, 64 (14), 1711–1726. doi:10.1080/02626667.2019.1669793.

Su, Y.-S., *et al*., 2011. Multiple imputation with diagnostics (mi) in R: opening windows into the black box. *Journal of Statistical SoftwareSoftware*, 45 (2), 31. doi:10.18637/jss.v045.i02.

Tencaliec, P., *et al*., 2015. Reconstruction of missing daily streamflow data using dynamic regression models. *Water Resources Research, American Geophysical Union*, 51 (12), 9447–9463. doi:10.1002/2015WR017399.

Tencaliec, P., 2017. *Developments in statistics applied to hydrometeorology: imputation of streamflow data and semiparametric precipitation modeling*. Universite Grenoble Alpes.

Tyralis, H., Papacharalampous, G., and Langousis, A., 2019. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, 11 (5), 1–37. doi:10.3390/w11050910.

van Buuren, S., 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16 (3), 219–242. doi:10.1177/0962280206074463.

van Buuren, S. and Groothuis-Oudshoorn, K., 2011. mice : Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45 (3), 1–67. doi:10.18637/jss.v045.i03.

Van Loon, A.F. and Laaha, G., 2015. Hydrological drought severity explained by climate and catchment characteristics. *Journal of Hydrology*, 526, 3–14. doi:10.1016/j.jhydrol.2014.10.059

Vezza, P., *et al*., 2010. Low flows regionalization in north-western Italy. *Water Resources Management*, 24 (14), 4049–4074. doi:10.1007/s11269-010-9647-3.

White, I.R. and Wood, A.M., 2011. Multiple imputation using chained equations : issues and guidance for practice. *Statistics in Medicine*, 30 (4), 377–399. doi:10.1002/sim.4067.

Widaman, K.F., 2006. Missing Data: what to do with or without them. *Monographs of the Society for Research in Child Development*, 71 (1), 210–211. doi:10.1111/j.1540-5834.2006.00404.x.

Yang, H.H., *et al*., 2011. Analysis of hydrological processes of Langat River sub basins at Lui and Dengkil. *International Journal of the Physical Sciences*, 6 (32), 7390–7409. doi:10.5897/IJPS11.1036.

Zhao, Y. and Long, Q., 2016. Multiple imputation in the presence of high-dimensional data. *Statistical Methods in Medical Research*, 25 (5), 2021–2035. doi:10.1177/0962280213511027.

Zvarevashe, W., Krishnannair, S., and Sivakumar, V., 2019. Analysis of rainfall and temperature data using ensemble empirical mode decomposition. *Data Science Journal*, 18 (1), 1–9. doi:10.5334/dsj-2019-046.