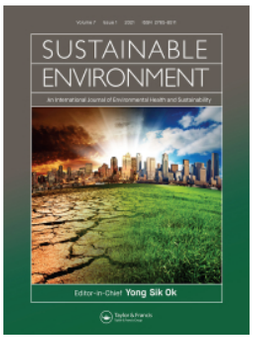


Cogent Environmental Science

ISSN: 2331-1843 (Online) Journal homepage: www.tandfonline.com/journals/oaes20

Imputation methods for recovering streamflow observation: A methodological review

Fatimah Bibi Hamzah, Firdaus Mohd Hamzah, Siti Fatin Mohd Razali, Othman Jaafar & Norhayati Abdul Jamil |

To cite this article: Fatimah Bibi Hamzah, Firdaus Mohd Hamzah, Siti Fatin Mohd Razali, Othman Jaafar & Norhayati Abdul Jamil | (2020) Imputation methods for recovering streamflow observation: A methodological review, Cogent Environmental Science, 6:1, 1745133, DOI: 10.1080/23311843.2020.1745133

To link to this article: https://doi.org/10.1080/23311843.2020.1745133

Published online: 06 Apr 2020.

Hamzah et al., *Cogent Environmental Science* (2020), 6: 1745133
https://doi.org/10.1080/23311843.2020.1745133

cogent environmental science

ENVIRONMENTAL MANAGEMENT & CONSERVATION | REVIEW ARTICLE

Imputation methods for recovering streamflow observation: A methodological review

Fatimah Bibi Hamzah[1,2*], Firdaus Mohd Hamzah[2], Siti Fatin Mohd Razali[2], Othman Jaafar[2] and Norhayati Abdul Jamil[1]

Received: 16 October 2019
Accepted: 05 March 2020

*Corresponding author: Fatimah Bibi Hamzah, Faculty of Computing and Multimedia Kolej Universiti Poly-Tech Mara Kuala Lumpur, Jalan 6/91, Taman Shamelin Perkasa, Kuala Lumpur 56100, Malaysia
E-mail: bibi@gapps.kptm.edu.my

Reviewing editor:
Fei Li, Zhongnan University of Economics and Law, China

Additional information is available at the end of the article

Abstract: Missing value in hydrological studies is an unexceptional riddle that has long been discussed by researchers. There are various patterns and mechanisms of "missingness" that can occur and this may have an impact on how the researcher should treat the missingness before analyzing the data. Supposing the consequence of missing value is disregarded, the outcomes of the statistical analysis will be influenced and the range of variability in the data will not be appropriately projected. The aim of this paper is to brief the patterns and mechanism of missing data, reviews several infilling techniques that are convenient to time series analyses in streamflow and deliberates some advantages and drawback of these approaches practically. Simplest infilling approaches along with more developed techniques, such as model-based deterministic imputation method and machine learning method, were discussed. We conclude that attention should be given to the method chosen to handle the gaps in hydrological aspects since missing data always result in misinterpretation of the resulting statistics.

Subjects: Hydrology; Surface Hydrology; Civil, Environmental and Geotechnical Engineering; Hydrology

Keywords: missing data; imputation; deletion; simple; model-based; machine learning

ABOUT THE AUTHOR

Fatimah Bibi Hamzah, Ph.D candidate at Universiti Kebangsaan Malaysia, Faculty of Engineering and Built Environment cum a Senior Lecturer at Kolej Universiti Poly-Tech MARA Kuala Lumpur. Master degree in applied statistics and Bachelor's degree in pure mathematics from UPM, Malaysia. Area of research interest includes mathematical problems in engineering, hydrology, multi-criteria decision analysis and developing methods to overcome missing data and outliers' problems.

PUBLIC INTEREST STATEMENT

Missing data becomes one of the most common problems in almost all research areas. There are many excuses for why data may be missing. Generally, if the consequence of missing value is not taken into consideration, the outcomes of the statistical analysis will be influenced and the amount of variability in the data will not be appropriately projected, thus leading to invalid conclusions. This manuscript reviews the problems and types of missing data, along with the techniques for handling missing data. We summarize the patterns and mechanism of missing data, review several infilling techniques that are convenient to time series analyses and deliberates some advantages and disadvantages of these approaches practically. The goals of this review paper are to raising awareness among the use of various imputation techniques in the hydrologic context. The paper concludes with recommendations for the handling of missing data.

1. Introduction

Streamflow data plays a significant pose in the hydrological functioning of watersheds. The availability of complete quality streamflow data of sufficiently long duration is considered as a major requirement in defining the flow characteristics, estimate the occurrence of the extreme event such as high or low flow and rectify problems like landslides, droughts and even floods (Tencaliec et al., 2015). Nevertheless, studies on streamflow generally are faced with the missing data issue and it is not rare to obtain data that are damaged with lapses, featured with the dubious quality and short periods (Norliyana et al., 2017). Every so often, the information is just not accessible and in many cases, incomplete observations, missing or outliers complicate dramatically the work of the analysts (Kim et al., 2015).

The dataset is often not complete, may occur due to several reasons such as not continuous data recording or lost in storage. According to (Žliobaite et al., 2014), statistical models for monitoring in environmental studies strongly depend on automatic data acquisition systems which employ numerous physical sensing devices. Frequently, sensor readings are incomplete for long lengths of time, though model outputs necessity to be uninterruptedly accessible in real-time (Santosa et al., 2014). The physical sensing device is laid bare to several risks due to acute ecological circumstances, subjection to physical destruction, and battery drainage (Tencaliec et al., 2015). Due to technical or maintenance issues, not optimal weather conditions, instrumental failures or apparatus errors throughout the data collection, human error during data entry, calibration process and/or a damage of data due to malfunctioning storing machinery, extended hydrometric data construction and organization become a hard task and, in time, gaps in the data set arise (Gao, 2017; Johnston, 1999; Peña-angulo et al., 2019; Tencaliec, 2017).

Despite Malaysia, such conditions are more common in developing countries and the repercussion is a large proportion of unpredictability in the measured features of water operation systems and eventually its weak execution (Kim et al., 2015; N.A. Rahman et al., 2017). According to (Fontaine, 1982) in a study of the USGS stream-gaging program in Maine, an average of 5.6 percent of the stream-gages are malfunctioned during the ten year period of study, while a 1983 study of Virginia stream-gages records a total breakdown rate of 1.95 percent which leads to missing data (Carpenter, 1985). (Shields & Sanders, 1986) reported that the lengthiest extent of inaccessible streamflow data was 16 consecutive months stretch throughout the seven years programmed observation period of the log. A study of 58 USGS Alabama stilling well and bubbler-gage stations reveals a total breakdown rate of 4 to 11 percent of the years 1979 through 1983 (Meadows & Jeffcoat, 1990), and a nationwide study of 1,009 stream-gages for the period 1948 to 1988 reports that five percent or more of streamflow data may be pointed as missing (Wallis et al., 1991).

These missing value(s) in the time series usually lessens the ability and the accuracy of statistical analysis approaches (Roth et al., 1999) and initiate biased estimations of the relationship among variables (Pigott, 2001). Both issues—lessening in ability and bias of estimates—may cause imprecise assumptions in the exploration of a set of data that consists of incomplete data (Graham, 2009). Due to this fact, reconstructing and missing data treatment should become the first priority in the data preparation procedure. According to (Tencaliec, 2017), the imputation of missing streamflow data is an issue investigated since decades ago and, up till these days, it persists to be a challenge. In the literature, the computation of missing data also called imputation (Schneider, 2001), reconstruction (Kim & Pachepsky, 2010), and infilling (Goa, 2017; Harvey et al., 2012). Imputing means substituting to individually missing data with a reasonable value (single imputation) or a vector of reasonable values (multiple imputations) but the aim is not to substitute all missing value but then to retain the features of their dispersal and associations among different variables (Rubin, 1976).

Several researchers have studied imputing missing streamflow data with various statistical methods (Regonda et al., 2013). Numerous data estimation methods have been suggested to

tackle the problem and are profoundly discussed in the literature; from the simple traditional statistical method such as substituting each missing value for given variables with mean, median or other location stations to the advanced techniques. Among these, (Kim et al., 2015), who employed Soil and Water Assessment Tool (SWAT), and two machine learning (ML) approaches, Artificial Neural Network (ANN) and Self Organizing Map (SOM) to substitute wrong values and fix missing streamflow data, conclude that the ML models were mostly advance on obtaining high flows, whereas SWAT was preferable on representing low flows. Further references, like the studies of (Santosa et al., 2014) suggested that the method derives from information entropy principles have the potential to be developed as the methods to be used to forecast the missing monthly average discharge. Most recent studies by (Norliyana et al., 2017) recommend the use of a normal ratio method for reconstructing the missing data in streamflow compared to the arithmetic average, inverse distance technique and coefficient of correlation technique or methods that imply artificial neural network as presented in (Elshorbagy et al., 2002; Gill et al., 2007; Mispan et al., 2015).

Most recent reviews studies by Gao et al. (2018), summarize and compare several widely known techniques employed for imputing missing hydrological datasets, as well as the arithmetic mean imputation, principal component analysis, regression-based approaches, and multiple imputation approaches. In (Gupta & Srinivasan, 2011), Two-Directional Exponential Smoothing (TES) been used in order to forecast the missing data for a raw streamflow dataset, while Exponentially Weighted Moving Average (EWMA)was used for a forecast by utilizing the identified data values of prior two seasons. The initial stage in the TES technique is to create the complete data set employing the Average Nearest Observation (ANO) approach (Huo, 2005). The ANO technique will substitute the missing values with the mean of the adjacent pre-existing and the subsequent observation i.e. the values are projected by a weighted mean of the adjoining observations with more weight given to the nearer observation (Gupta & Srinivasan, 2011). Their results show that both the approaches estimated the data inside and outside the period space with excellent outcomes.

However, based on (Tabachnick & Fidell, 2014) suitable technique for recovering missing data rest on the patterns and mechanisms of the missingness and, in fact, both—patterns and mechanisms—have a greater influence on the results than the proportion of the missing data itself. The method chosen has a significant influence on the accurateness of any hydrological model output. The details of the patterns and mechanisms of the missingness and the imputation method are described in the following study.

The paper aims to confer an overview of various infilling techniques that are suitable for time series analysis in streamflow. According to (Gill et al., 2007), it is becoming a common practice for a researcher in pre-processing of data for use in hydrological modeling to disregard observations with any missing variable values at any given time frame, although it is only one of the independent variables that are missing. Mostly, these rows of data are considered poor and would not be used in either model building or subsequent model testing and verification. This is not assuredly an ideal approach for handling the missing data since the important facts might be mislaid once deficient rows of data are dropped. Even very small data-breaches in this study may result in terribly distinct analysis outcomes (Tencaliec et al., 2015). Therefore, works on these issues is a crucial exercise in hydrological analyses and the goals of this review paper are to raising awareness among the use of various infilling practices in hydrological context.

2. Percentage of missingness

The percentage of missingness is straight off linked to the value of statistical inferences. Up until now, there is no ascertained cut-off based on the literature concerning on tolerable proportion of missing rate in streamflow data as convincing statistical inferences. Schafer (1997) stated that a missing rate of 5 percent or a lesser amount of it, is insignificant. Bennett (2001) asserted that whenever the percentage of missing data exceeding 10 percent, then the statistical analysis is

Table 1. Missing data rule of thumb according to (Widaman, 2006)

Percentage of data missing	Categories
1%—2%	Negligible
5%—10%	Minor
10%—25%	Moderate
25%—50%	High
>50%	Excessive

expected to be biased. Dong & Peng (2013) agreed that data missing by 20 percent is a common thing in research while (Widaman, 2006) pigeon-holed missing data according to the percentage of the missingness as outlined in Table 1.

Moreover, the percentage of missing rate is not the only principle through which a researcher weighs the missing value issue Tabachnick and Fidell (2014) affirmed that the missing data mechanisms and the patterns have a major influence on research outcomes compared with the percentage of missing data.

3. Patterns of missingness

Even though a wide range of reconstruction methods to handle missing data problems have been developed, the method individually suffers from various restraints and may not perform practically well under some circumstances (Gao et al., 2018). One reason for this is that most of these methods make a presumption about how the missing values are dispersed within the data set. It is worthwhile to know the underlying missingness pattern and mechanism before deciding which method to use to handle missing data (Tabachnick & Fidell, 2014).

Collins et al. (1991) first described and divided the pattern of missingness into two groups: general (random) and special patterns including univariate missing data, unit nonresponse, and monotone missing data. The general or random pattern of missingness is exactly as the name implies, where the missing data occur in any of the variables in any position. If there is only one variable with missingness while the other variables are completely recorded, the pattern is called univariate missing data. Additionally, when the multivariate pattern is detected, means that the missing value arises in more than one variable. According to (Ingsrisawang & Potawee, 2012), the unit nonresponse pattern has missing values on a block of variables for the same set of cases, and the remaining of the variables are all complete. The pattern is said to be monotone whenever the observations are ordered and item k is missing, and all $k+1, \ldots, n$ cases are also missing. Figure 1 illustrate missing data patterns discussed above.

4. Types of missing data

The mechanism by which data is missing is very important when determining the efficacy and appropriateness of imputation strategies (Sakke et al., 2016). Kamaruzaman et al. (2017) suggested that the type of mechanism missing data should be interpreted prior to the imputation process because the effectiveness of an imputation technique depends entirely on their assumptions. In spite of that, it is impossible to carry out a proper test to examine whether the presumptions made are valid. Ignoring this missingness might not only lose efficiency yet also induce biased outcomes as well as deceptive inferences (N.A. Rahman et al., 2017).

Little and Rubin (2002) classified three possible ways that data may go missing: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). MCAR describes data where the gaps are distinct from any of the variables in the dataset. In any event, the missing values probably correlated to other observed values, yet not to missing values, in that case, the missingness assumed to be MAR. Missing data which are dependent on the

Figure 1. Missing data patterns: (*a*) General (random); (*b*) Univariate missing data; (*c*) Unit non-response; (*d*) Monotone missing data.

(a).

Item	Variables						
	v_1	v_2	v_3	v_4	…	v_{k-1}	v_k
1	NA				NA		
2		NA	NA				NA
3				NA		NA	
…		NA			NA		NA
n − 1						NA	
n	NA			NA	NA		

(b).

Item	Variables						
	v_1	v_2	v_3	v_4	…	v_{k-1}	v_k
1							
2							
3							
…							NA
n − 1							NA
n							NA

(c).

Item	Variables						
	v_1	v_2	v_3	v_4	…	v_{k-1}	v_k
1							
2							
3							
…					NA	NA	NA
n − 1					NA	NA	NA
n					NA	NA	NA

(d).

Item	Variables						
	v_1	v_2	v_3	v_4	…	v_{k-1}	v_k
1							
2							
3							NA
…						NA	NA
…					NA	NA	NA
…				NA	NA	NA	NA
n − 1			NA	NA	NA	NA	NA
n		NA	NA	NA	NA	NA	NA

observed value is also known as MNAR. Missing data can be presented in the form of a probabilistic process that describes the association among the measured variables and the probability of missing value (Gao et al., 2018). Each is discussed below.

4.1. Missing completely at random (MCAR)

In MCAR there is no methodical form on the way the data is missing. The missingness occurs entirely at random which rests on neither on observed nor on missing values. Simply put, the probability for an observation being missing is independent of both the values of other variables as well as the value of the observation itself. In the univariate time series such as streamflow datasets, there are no other variables existent except time as implicit variables. In this case, we can conclude that in MCAR the probability for a certain observation being missing is independent of the point of time of this observation in the series.

$$P\left(m|Y_{(observed)},\ Y_{(missing)}\right) = P(m) \tag{1}$$

4.2. Missing at random (MAR)

Similar to MCAR, in MAR probability for an observation being missing is also independent of the value of the observation itself but not on other variables. For univariate time series where there are no other variables than time (implicitly given), the probability for an observation being missing in MAR is dependent on the point in time of this observation in the series.

$$P\left(m|Y_{(observed)},\ Y_{(missing)}\right) = P\left(m|Y_{(observed)}\right) \tag{2}$$

4.3. Missing not at random (MNAR)

MNAR observations are not missing in a random manner. The data are neither MCAR nor MAR. The probability for an observation becomes missing dependent on other variables such as point of time for univariate time series.

$$P\left(m|Y_{(observed)},\ Y_{(missing)}\right) = P\left(m|Y_{(observed)},\ Y_{(missing)}\right) \tag{3}$$

Based on the definition of (Little & Rubin, 2002), missing value in the streamflow study is determined as MCAR as of the episode of missingness in the streamflow data of an area not influenced by the data in that area or any area. There is also a study on streamflow data imputation using the MAR assumption (Gill et al., 2007). However, (Moritz & Bartz-Beielstein, 2017) have stated that imputation MCAR and MAR for univariate time series study is nearly the same.

5. Imputation Methods

Imputation is a term used to replace or substitute the missing values by some predictable which said to be plausible values in the dataset (Ahmat Zainuri et al., 2015). It utilizes observed supporting information for cases with missing value to maintain high accuracy. In this paper, the imputation method classified into four categories; deletion technique, single imputation, model-based deterministic imputation method, and machine learning technique according to the characteristic of each method discussed.

5.1. Deletion technique

Deletion (list-wise and pair-wise) approaches are the standard settings for missing data problems in most statistical software packages, and these methods are most likely the elementary approaches in recovering missing data (Gao et al., 2018; Marsh, 1998). According to (Gill et al., 2007; Kabir et al., 2019), deletion technique are among the popular and the most common method used in the treatment of missing value in hydrological field. However, in the study practiced using time series data, deletion methods may cause the data to become discontinuous.

5.1.1. List-wise deletion

McDonald et al. (2000) defines list-wise deletion as the removal of each and every case (observations) whichever has a missing value in at least one of the selected variable. The primary advantage of list-wise deletion is fast, simple to understand and apply, and hence has set off as a preselected option for analysis in most statistical software since there is no special computational methods are required (Dong & Peng, 2013). Certain researchers claim that it possibly will lead to bias in the estimation of the parameters (Dong & Peng, 2013; Honaker & King, 2010; Marsh,

1998; McKnight, 2007). In the nature of big data which power is not a problem, and the presumption of MCAR is fulfilled, the list-wise technique might be a practical method, but for a small sample, or the presumption of MCAR is not fulfilled, the list-wise technique is not an ideal approach (Gao et al., 2018). If the data are not MCAR then there be lacking comparability over time points, which would introduce extremely biased outcomes.

5.1.2. Pair-wise deletion

Pair-wise deletion also known as an improvement version of list-wise deletion in which it helps conserves significantly more information by minimalizing the number of observations dropped (Marsh, 1998). The advantage of using this method is it is convenient and not as much of an influence for the MCAR and MAR dataset, along with the suitable procedures are counted in as covariates (Honaker & King, 2010). According to (Croninger & Douglas, 2005), one of the most important disadvantages of the pair-wise deletion technique is the poor conformity of various analyses since the number of cases differs among distinct pair-wise comparisons. This technique is suitable only when the data consist of a proportionately insignificant percentage of observations with missing data.

5.2. Single imputation

Single imputation signifies that the missing data is substituted by a value. In this method, the sample size is retrieved. However, the reconstructed values are presumed to be the actual values one might have been observed when the data would have been complete (Plaia & Bondì, 2006). This method is very convenient since it generates a complete dataset easily (Hasan et al., 2017). The drawbacks of the single imputation approaches are that they reconstruct the same missing value every single time. Consequently, a statistical analysis, which treats incorrect values entirely as similar as observed values, will consistently depreciate the variance, even presumptuous that the exact cause for the missingness is known. It is impossible for the single imputation to signify any extra variation that rises when the reasons for missing data are unidentified (Gómez-Carracedo et al., 2014). The idea of using a single imputation method to recover the missing value in hydrological been widely used by a variety of authors as in (Ben Aissia et al., 2017; Gao et al., 2018; Kabir et al., 2019; Norliyana et al., 2017; N.A. Rahman et al., 2017). However, despite the simplicity, the researchers agreed that reconstruct the missing value using the same "number" do not reflect the variation that would probably occur if the variables were observed. The actual figures possibly vary from the imputed. Thus, the variance of those same variables is underestimated.

5.2.1. Arithmetic average method

According to (Schneider, 2001), the simplest way to reconstruct the missing value is to substitute every missing value with the average of the observed values for that variable. The average of the variable is calculated through the non-missing values and is employed to reconstruct the missing value of that variable. This method generally used to impute the missing meteorological and hydrological data (Norliyana et al., 2017). The missing data of streamflow are obtained by the average of selected nearby stations around the target station or the date on the same day with different years. The estimated missing value is given by

$$\hat{y}_t = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

where $\hat{y}_t$ is the projected value of the missing data at the t target station or date, x_i is the observed data at the i^{th} nearby stations or the date of the same with different years and n is the count of nearby stations or count of years. This method presumes the data is MCAR but is not suggested as it can result in depreciating the variance.

5.2.2. Median imputation

Median imputation substitutes missing values with the median value of the observed values of the same variable (McKnight, 2007). Median imputation does not require normality assumptions but instead skewed (Gómez-Carracedo et al., 2014). Despite it is simple to understand and apply, it can

affect the correlation between variables and the probability distribution of the imputed variable. This method also depreciates the variance of the estimators (Gao et al., 2018).

5.2.3. Hot-deck method

This technique requires substituting missing data with values based on the current dataset or matching covariates (Chen & Shao, 2001). The process detects any data which are akin to data observed. This approach is convenient as such quite straightforward and upholds the appropriate quantification levels of the logged covariates (Aljuaid & Sasi, 2017). In other words, the imputed values under hot-deck imputation will have a similar dispersal form as the observed data (Rubin, 1976). It is normally less biased than an arithmetic average method or deletion technique methods and presumes that the missing data are MAR (Blend & Marwala, 2008). Weakness is about multiple paralleling algorithms occasionally required to be used in such a way to suit the data.

5.2.4. Last observation carried forward (LOCF)

The LOCF is one form of hot-deck imputation. This method is usually used for longitudinal data which are detected to be missing at a specific "visit" or at any given time for a certain entry. The researcher would then drag the last obtainable value forward i.e. from the last visit or time point, and use this value to substitute the missing values. Some values might be used several times for infilling if more than one missing value occurs in a row, and others may not be used at all. The most crucial challenge of this technique is that the researcher is inferring that there will be no shift from one visit to the next. It is sensible if the data are MCAR, otherwise, the results can be extremely biased (Gao et al., 2018).

5.2.5. Cold-deck method

The cold-deck imputation is very much alike to the hot-deck method in its approach with the exception of the tactic for weighing subject uniformity is derived from external information prior knowledge instead of the information obtainable in the existing dataset (Kumar et al., 2017). In hydrometeorological, researchers may decide to reconstruct the missing value out of their knowledge of prior research that observed at related variables to estimate individual data. A prominent weakness of this technique is it hooked on the quality of the available external information (Nishanth & Ravi, 2013).

5.2.6. Linear interpolation

Linear interpolation is among the simplest methods to impute missing data which made up of drawing a straight line separating observed values before and after the gap and then estimating missing values by interpolation. The linear interpolation method is quick and easy to use and may be adequate for well-resolved data. (Fleig et al., 2011) was used this method in univariate regional hydrological frequency analysis, but (Ben Aissia et al., 2017) believe that it was not used to estimate missing value in multivariate hydrological frequency analysis. To interpolate the value of M_2, the following formula is used.

N_1	M_1
N_2	M_2
N_3	M_3

$$M_2 = \frac{(N_2 - \mathrm{N})(M_3 - M_1)}{(N_3 - N_1)} + M_1 \tag{5}$$

5.3. Model-based deterministic imputation method

Model-based methods will produce more precise imputations provided the model presumptions are contented. However, the struggle on model-based methods is that those presumptions are frequently unverifiable in practical terms and consequently it may not be easy to specify a suitable model-based infilling technique to impute the missingness (Nishanth & Ravi, 2013). A good model-based method would operate efficiently for various options of underlying data distributions and missing mechanisms. (Norliyana et al., 2017) studied daily and monthly streamflow estimating using a model-based deterministic imputation method. Many statistical indicators have been utilized for evaluating performance such as root mean square error (RMSE), Mean Absolute Error (MAE) and Correlation Coefficient (R) tests. The NR method is found to be the best estimation technique for streamflow data in their studies. A comparison between the performance of regression and time-series methods were also conducted by (Beauchamp et al., 1989). The results indicated that both methods present fairly good estimates and forecasts of the flow at the Foresta Bridge gage. However, the author discovered that the regression model gives a significant amount of autocorrelation in the residuals and this lead to the conclusion of the standard errors and confidence limits on the estimated flow from the regression model would not be appropriate and could be misleading.

5.3.1. Normal ratio method (NR)

NR approach is weighted on the basis of the ratio average of the accessible data among the target station and the i^{th} neighbouring station (N.A. Rahman et al., 2017). The simultaneous streamflow data at the neighboring stations are weighted by the ratios of whole streamflow data in the target as well as neighboring stations. The estimated missing value is given by

$$\hat{Y}_t = \frac{1}{n}\left[\left(\sum_{i=1}^{n}\frac{N_t}{N_i}\right)y_i\right] \tag{6}$$

where N_t is total streamflow in the target station while N_i is total streamflow for each neighboring station. This method is used only if any neighboring stations have the normal streamflow data which exceeded more than 10% of the considered station (De Silva et al., 2007).

5.3.2. Inverse distance method (ID)

The ID is based on an idea of distance weighting among the target station and neighboring station. The closest stations are better correlated with the target station compared to further stations. The estimated missing value is given by

$$\hat{Y}_t = \frac{\sum_{i=1}^{n} y_i/d_{it}}{\sum_{i=1}^{n} 1/d_{it}} \tag{7}$$

with d_{it} is the distance between the target station and the i^{th} neighbouring station. The main advantage of ID is it's easy to use or simple and results in sensible outcomes for various options of data. It is all right with results over and above the range of meaningful values (Chen & Liu, 2012). In contrast, there are several disadvantages, in which the ID approach is very sensitive to the weighting function and can be exaggerated by uneven data points distribution (Caruso & Quarta, 1998).

According to (Norliyana et al., 2017), the ID technique is the most frequently used in the computation of rainfall and streamflow missing data. This technique gives good results for missing data analysis, provided, the researcher has the data of neighboring stations of the same period (Teegavarapu & Chandramouli, 2005). However, based on (Chen & Liu, 2012) using data of neighboring stations of the same period, is valid for climatic parameters like temperature and rainfall, but for data of river flow, it's unwise to use data of a neighboring station that controls another watershed due to the surfaces, slopes, the permeability or the overall morphology are not the same. They propose to use the ID method only if the researcher has two adjacent stations on the same river.

5.3.3. The coefficient of correlation method (CC)

The CC method is influenced by the success of the ID method (N.A. Rahman et al., 2017). This method is used by replacing the distance with the coefficient of correlation between the target and neighboring station, as the weighting value. The missing value is estimated by

$$\hat{Y}_t = \frac{\sum_{i=1}^{n} y_i r_{it}}{\sum_{i=1}^{n} r_{it}} \tag{8}$$

where r_{it} is the coefficient of correlation of day-to-day time series data among the target and the i^{th} neighbouring stations. The advantages of the CC method are that it is easy to work out and it's easy to interpret. However, the correlation coefficient does not imply causality, that is it may show that two stations are strongly correlated, but it doesn't mean that they are responsible for each other (Norliyana et al. 2017).

5.3.4. Regression method

Reconstruction using regression on either one or several variables may generate keener values. In this case, the researcher needs to fit a regression model by fixing the variable of interest as the response variable then another related variable as covariates (Gao et al., 2018). The coefficients are projected, followed by the estimation of missing values with the fitted model. Regression model fit with complete data given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots \ + \beta_n x_{ni} + \epsilon_i \tag{9}$$

and in a case k is missing y, the imputation is

$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_{1k} + \hat{\beta}_2 x_{2k} + \ \ldots \ + \hat{\beta}_n x_{nk} + \epsilon_k \tag{10}$$

where $\hat{\beta}_j$'s are estimates based on complete cases.

This method appears more sensible than that estimated with mean. Nevertheless, this approach rises the correlation coefficient among the variables and the variability of imputed data is underestimated. Another possible drawback of such a parametric method about the approach may be delicate to model misspecification of the regression model (Schenker & Taylor, 1996). In the case the regression model is not a good fit, the extrapolative power of the model might be poor (Little & Rubin, 2002).

5.3.5. Nearest-neighbor technique

The nearest-neighbor approach, also known as distance function matching, is a donor technique in which the donor is chosen by minimizing a fixed "distance" (Lee & Kang 2015; Chen & Shao 2001; Rajagopalan & Lall 1999; Yakowitz & Karlsson 1987; Kalton & Kish 1984). This process involves identifying an appropriate distance measure, where the distance is a function of the auxiliary variables. The observed unit with the shortest distance to the missing values is acknowledged and its value is used for infilling the missing item in proportion to the variable of concern. The beauty of the nearest neighbor technique is that truly observed values are utilized for reconstruction and it presents geographical effects, but the weakness is the outcome could be contingent on the selected order of the file. (Chen & Shao, 2001) demonstrates such the nearest-neighbor method, even though a deterministic method, estimates distribution precisely. The variance under the nearest-neighbor approach may be exaggerated in case some donors are used way more often than others (Bagus & Narinda, 2016).

5.3.6. Multiple imputations

Multiple imputations substitute every missing value by N probable values to generate a full dataset. The researcher further employs such "latest" datasets in the analysis as well as merges the outcomes toward a single summary outcome. This summary mirrors the additional discrepancy as a result of the missing data (Bertsimas et al., 2018). This technique doings well on

longitudinal data and is robust to violations of non-normality of the variables used in the analysis. A review of multiple imputations can be found in (Little & Rubin, 2002).

5.3.7. Expectation maximization (EM)

According to (Schneider, 2001), the EM techniques are a form of the maximum likelihood approach which offers estimates of the means, variances and covariance matrices which can be used to get consistent estimates of the parameters of interest. It is predicated on both; expectation and a maximization step, those are repeated several times until maximum likelihood estimates are obtained. For more detail, in the expectation step, the parameters such as mean, variance and covariance are estimated. Those estimates are then employed to develop a regression equation to forecast the missing value, and then the equations used in the maximization step to impute the missing values. The expectation step is then repeated with the new parameters, where the new regression equations are determined to impute the missing data. Both—the expectation and maximization steps are repeated until the system fixes when the covariance matrix for the subsequent iteration is essentially similar to that for the preceding iteration.

The advantage of EM reconstruction is that once the full data set with no missing values are reconstructed, a random error term for each reconstructed value is incorporated view to reflecting the unpredictability related to the infilling (Aljuaid & Sasi, 2017). On the other hand, the disadvantage of EM imputation is it takes a lot of time to converge, specifically when a large fraction of missing data occurs. This can result in biased parameter estimates and can underestimate the standard error (Gómez-Carracedo et al., 2014).

5.3.8. Spline interpolation

Spline interpolation is a piecewise polynomial function that is tailored through the sampled points. The spline represents a two-dimensional curve on a three-dimensional surface (Hutchinson & Gessler, 1994). Spline interpolation is often preferred to polynomial interpolation as it tries to avoid the oscillating effect that may be observed with polynomial functions of high degrees (Little & An, 2004). This method produces smooth and easily interpretable surfaces with low degree polynomials for the spline. The spline function is constrained at defined points (local technique). A specific number of neighboring values should be considered; therefore, the spline can adjust to local abnormalities without affecting the values of interpolation at other points on the global area. The constraints r are given by the degree m of the polynomial function

- if $r = 0$ there are no restraints.
- if $r = 1$ the function has to be continuous.
- when $r = m + 1$ the m^{th} derivative of the function has to be continuous for all points.

The spline for $m = 1$ is called linear, for $m = 2$ quadratic, and for $m = 3$ cubic.

5.4. Machine learning (ML) techniques

ML is a branch of artificial intelligence (AI) that employs statistical approaches to give computer systems the ability to study the data, without being explicitly programmed. The name machine learning was coined in 1959 by Arthur Samuel (Samuel, 1959). Many researchers agreed that the approaches relying on machine learning methods were the most suited for the reconstruction of missing values and usually lead to a significant improvement of prediction accurateness as against imputation methods based on statistical approaches (Krysanova & White, 2015; Minns & Hall, 1996; Varga et al., 2016). ML techniques were widely used by variety of authors to recover the missing value of streamflow as in (Allawi et al., 2018, 2017; Erdal & Karakurt, 2013; Kalteh et al., 2007; Karakurt et al., 2013; Kim & Pachepsky, 2010). Despite the fact that ML has been transformative in the hydrological field, effective ML is hard since detecting patterns is difficult and frequently not sufficient training data are accessible; consequently, many ML approaches often fail to give the expected value and most of all can suffer from different data biases (Worland et al., 2018).

5.4.1. Self-organizing map (SOM)

The SOM technique, also named characteristic map or Kohonen map, is the most broadly used of the ANN algorithms intended for unconfirmed patterns recognition applications (Kohonen et al., 1996). The ability of the SOM method in the estimation of missing univariate and multivariate hydrological data was proved in a number of studies such as in (Adeloye & Rustum, 2012; Mwale et al., 2012). It is a non-linear process for dimension lessening and illustration of data that could also be used for the missing data reconstruction (Miró et al., 2017). SOM describes a methodical two-dimensional discrete mapping, extending a set of data items onto a topologically ordered network. In other words, the SOM network captures the clusters of the available input data, which may be a starting point for the infilling of the missing value (Kalteh et al., 2007). This technique conserves the most significant association of the original data elements. This infers that, throughout the mapping, not much information is lost which makes the SOM approach a very good implement for estimation. However, for estimate values outside the range used for extrapolation, the SOM technique cannot be used. This is mostly on grounds that, as it is the case with most data-driven approaches, SOM is a very poor extrapolator (Adeloye et al., 2011). It has a restricted volume to forecast values that have not been observed in the past (Ben Aissia et al., 2017).

5.4.2. Decision trees

Decision trees make supervised groupings from categorical or discretized data. The decision tree algorithm computes the criterion for each potential split and opts for the one with the peak gain of purity. The algorithms stop splitting once a subset contains only objects fitting to the same class or a different standard is fulfilled. In the tree, a node signifies a subset, a branch characterizes a split, and the last nodes are called leaves (Han et al., 2011).

A popular implementation of decision trees is Quinlan's C4.5 (Quinlan & Kaufmann, 1994) algorithm, that by its nature manages missing values without taking them into account when calculating information gain. Thereby, as an infilling technique, C4.5 could be prepared once missing data are present in the predictor variables, rising the potential training set available in the multivariate missing value setting.

5.4.3. Bayesian networks

Bayesian networks pick up probabilistic associations among variables concisely by implementing conditional independence constraints (Uusitalo, 2007). They can be configured through a heuristic search using a Bayesian scoring function such as K2 (Cooper et al., 1992). Using Bayesian networks for the reconstruction of missing data has some benefits: One, it has greater efficiency compared to EM-based multiple imputation methods for a dataset with a large number of variables; two, it saves the joint probability distribution of the variables, one thing that methods like k-NN do not promise (Kim & Lee, 2009). Adversely, a lot of data is generally required to precisely learn a network, and discretization of all data is necessary except conditional probability densities are explicitly modeled and parameterized, often at the great computational expense (Chen & Pollino, 2012; John & Langley, 1995).

5.4.4. Soil and water assessment tool (SWAT)

The Soil and Water Assessment Tool (SWAT) is a continuous-time physically-based semi-distributed watershed-scale hydrologic model that can be used to simulate long-term impacts of climate, topography, soils, land use, and management operations on water, sediment, and chemical yields, without massive investments of resources (e.g., time, money and labor) (Arnold et al., 2012; Neitsch et al., 2011). SWAT practices a two-level disaggregation scheme; a preliminary sub-basin identification is conducted premised on topographic criteria, and then further discretization using land use and soil type considerations (Zeiger & Hubbart, 2018).

SWAT is an open-source tool and detailed online documentation, user groups, video tutorials, international conferences, and a unique literature database are available. This all makes the tool user-friendly, which can explain, at least partly, the fact that it is one of the best known and most

commonly used tools to develop water quality models at the watershed scale (Gassman et al., 2010; Refsgaard et al., 2010; Varga et al., 2016). The tool is continuously improved, supported by the core development team and as a response to shortcomings demonstrated by the many users (Neitsch et al., 2011). This results in the development of new tools, e.g., GIS interface tools, pre- and post-processing tools and statistical evaluation tools (Gassman et al., 2010). SWAT is also proven to be the most practical and flexible tool for a variety of applications, watershed scales, and environmental conditions (Gassman et al., 2014; Krysanova & White, 2015; Tuppad et al., 2011). Moreover, the semi-distributed structure makes the model computationally efficient and enables us to generate spatially explicit outputs. The tool is really suitable for large, complex watersheds (Gassman et al., 2014).

However, every tool has its shortcomings and these are often linked with its advantages. The constant improvements, for example, have led to difficult code and a high number of parameters, requiring expertise to run the model and complicating the calibration process (Gassman et al., 2010; Vigerstol & Aukema, 2011). Additionally, the tool is highly data-intensive. Although SWAT is said to run on readily available input data this is not always the case, especially in developing countries. Certainly, the data accuracy and precision might be an issue, as expressed by the rule "garbage in is garbage out" (Oosthuizen et al., 2018).

5.4.5. Exponentially weighted moving average (EWMA)

The EWMA is frequently utilized to a time-ordered sequence of random variables (Perry, 2011). It is a statistic for surveillance of the progression which averaging the data in a way that gives less and less weight to data as they are further removed in time (Čisar & Čisar, 2011).In other words, more recent observations have greater weight on the variance.

This method introduces lambda, which is called the smoothing parameter. The value of λ must be less than one and usually set between 0.2 to 0.3 (Perry, 2011), even though this choice is slightly arbitrary. By the choice of weighting factor λ, the EWMA control technique can be made sensitive to a small or gradual drift in the process. The equation for estimated missing value was established by Roberts as described in (Roberts, 1959) and given by

$$EWMA_t = \lambda Y_t + (1-\lambda)EWMA_{t-1}; \qquad for\ t = 1, 2, 3, \ \ldots\ , n \tag{11}$$

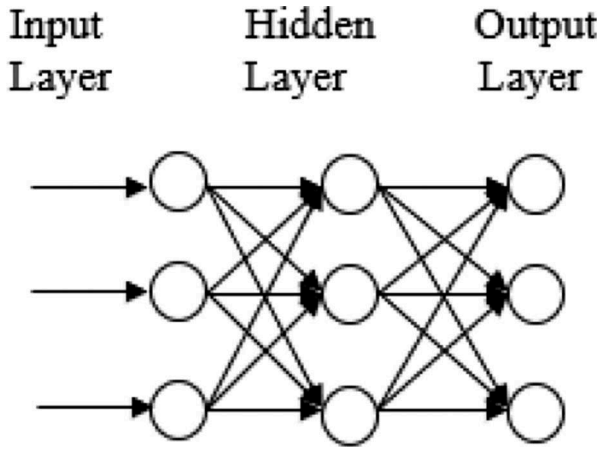
where $EWMA_0$ is the mean of historical data, Y_t is the observation at time t and n is the number of observations to be observed together with $EWMA_0$ while $0<\lambda \leq 1$ is a constant that assesses the depth of memory.

A possible weakness of an EWMA chart with a small λ is that a very big abrupt change in a parameter may not be sensed rapidly since data gained right away after the shift will be averaged with in-control data obtained prior to the shift. The consequence is that the in-control data have a propensity to mask the effect of the shift. Another possible drawback with an EWMA chart is that the EMWA statistic may be in a disadvantageous standing when the shift occurs. For instance, once there is an upward shift in a process parameter, the EWMA statistic may occur to be close to the lower control limit immediately before the shift, and in this case, it may take the EWMA statistic a relatively long time to reach the upper control limit. This latter problem is usually referred to as the "inertia problem" of the EWMA chart (Spiring, 2007; Woodall & Mahmoud, 2005; Yashchin, 1987, 1993).

5.4.6. Artificial neural network (ANN)

ANN is one of the main tools employed in machine learning which is inspired by human brain systems. ANN are mathematical models that are capable of learning complex relationships (Tsintikidis et al., 1997). This method involves input and output layers, along with (in most cases) a hidden layer contains units that alter the input hooked on something that the output layer can use. This process is illustrated in Figure 2. Although ANNs do not need to be fully connected, they often are.

Figure 2. A fully connected neural network with one hidden layer.



(Jain & Indurthy, 2003) claimed that among the black-box models, the ANN model has accessed broader pertinency, because the functional system sandwiched among the input variable and the output is not necessary to be outlined from scratch, and it entangles minimum understanding of the fundamental process to model such hydrologic issues. A number of research on the application of ANN in modeling hydrologic input-output associations were discussed in the literature as in (Hipel, 1995; Hsu et al., 1995; Jeong & Kim, 2009; Kim & Pachepsky, 2010; Minns & Hall, 1996; N.F.A. Rahman et al., 2015; Regonda et al., 2013; Smith & Eli, 1995; Somwanshi & Chaware, 2014; Zeiger & Hubbart, 2018).

The ANN model, even with only three layers, results in a nonlinear relationship between input and output with a very large number of connections between the input layer, to the hidden layer, and so on to the output layer (Eash et al., 2013). Resulting from the existence of such nonlinearity, the ANN model is very sensitive to the values of input neurons. Clearly, if the input values are subjected to large errors, then the functional form, which the ANN evolves at the training stage, may perform poorly at the validation stage (Elshorbagy et al., 2002). This calls for the careful selection of input neurons for such complex studies.

5.4.7. Two-directional exponential smoothing (TES)

The TES method was developed by (Huo et al., 2010) to substitute missing data. The TES technique is subject to an appropriate Exponential Smoothing (ES) technique and was established by using Holt's linear trend algorithm system. The TES process estimates missing values founded on the autocorrelations of the time-recorded data for the fact that the missing values arise at non-random periods (Gupta & Srinivasan, 2011). This method is intended to signify mutually forward and reverse autocorrelations in the time series, which can lessen the variance triggered by diverse directions. The initial stage in the TES approach is to reconstruct the complete data set employing the average nearest observation (ANO) method (Huo, 2005).

The ANO approach will substitute the missing values with the mean of the nearest previous and the subsequent observation. In other words, the values are projected by a weighted average of the nearest observations with higher weight given to the nearer observation. After the data set is reconstructed using the ANO technique, the missing values are forecast using Holt's linear trend approach, both in a forward and backward way. The Holt's Linear Trend algorithm can be denoted as

$$F_{t+k} = a_t + b_t k \tag{12}$$

$$a_t = \delta Y_t + (1-\delta)(a_{t-1} + b_{t-1}) \tag{13}$$

$$b_t = \gamma(a_t - a_{t-1}) + (1-\gamma)b_{t-1} \tag{14}$$

where F_{t+k} is an estimated value at period $t+k$; Y_t is the actual value at time t; a_t and a_{t-1} are intercepts at time t and $t-1$ respectively; b_t and b_{t-1} are the slopes at time t and $t-1$ respectively; δ and γ are smoothing constants that are between 0 and 1 (Gupta & Srinivasan, 2011; Moahmed et al., 2014). The smoothing constants govern the weight specified to the latest previous observations and thus control the degree of smoothing or averaging. Values close to 1 give weightage to more current data and close to 0 allocate the weights to reflect data from the more detached previous data.

6. Conclusions and discussion

Imputation methods lessen the loss of information which may introduce suboptimum outcomes and later to misleading conclusions concerning, as an example, the risk estimation of an extreme event. Several techniques had long been suggested in the literature for managing missing value together with the option of a suitable approach hang on, in particular, on the missing data pattern, and mechanism. The plainest technique, but in parallel, the minimal powerful in recovering missing value is the deletion technique. It is based on simply removing the variable(s) that have some missing values. The deletion method results in an unreasonable loss of valuable data. It is a well-known fact in statistics that smaller sample sizes reduce the statistical power and precision of standard statistical procedures (Little & Rubin, 2002). A simulation by (Raaijmakers, 1999) proved that the statistical power is reduced between 35% (with 10% missing data) and 98% (with 30% missing data) by using deletion techniques. Regardless of their weak points, most of the statistical software packages offers deletion methods as a default options to overcome the missing data problems, and these methods classified as the simplest approaches in treating missing data (Marsh, 1998). However, this is surely not the best possible way of dealing with missing data problems in hydrological fields.

On the other hand, in the single imputation approach, in preference to totally removing the pattern, researchers have to infill (compute) value for it and use the following complete data set for the forecast model. Single imputation methods generate a single replacement for each missing value with suitable values prior to the actual analysis of the data. Even though in the strong MCAR presumption event, these methods misrepresent the resulting parameter estimates (Gao, 2017). Particularly, they attenuate the standard deviation and the variance of estimates obtained from analyses of single imputed variables since the imputed values are identical and at the center of the distribution which reduces the variability of the data (Little & Rubin, 2002).

A more complex group of methods in reconstructing the hydrological data set is the model-based deterministic imputation method. This group of the method produces more precise imputations compared to the deletion technique and single imputation method. The main idea behind this technique is utilizing details from all observations with complete values in the variables of interest to reconstruct the missing values which are intuitively appealingly. In a hydrological context, the presumption of the model-based deterministic approaches was well established but seemed to be too restrictive (Machiwal & Jha, 2008).

A more powerful group of methods is the machine learning techniques. Many researchers conclude that the imputation approaches relying on machine learning algorithms predominated imputation techniques based on statistical procedures in the prediction of missing streamflow data (see e.g.; Krysanova & White, 2015; Minns & Hall, 1996; Varga et al., 2016). Generally, machine learning techniques are significantly more flexible than the standard statistical models and can capture higher-order interactions between the data, which leads better predictions. However, the predictions are made on the basis of complex relationships between the data, thus, the interpretability of the outcomes is sometimes harder, even though there are tools to pull out the knowledge required by these models. As a result, these alternative models are often criticized.

From these facts, one may conclude that a good imputation method would work well for various options of underlying data distributions and missing mechanisms. Generally, imputation in streamflow datasets often lacks a clear conceptual framework and a sound selection of methods depending on the statistical properties of the respective observable and the respective research question. Existing imputation techniques therefore have room for further improvement. As discussed earlier, the researcher with missing data issues in their studies has numerous choices once determine how to handle this common issue. More attention should be given to the missing data in the design and performance of the studies and in the analysis of the resulting data. Application of the sophisticated statistical analysis techniques should only be performed after the maximal efforts have been employed to reduce missing data in the design and prevention techniques. It may also be valuable to perform a sensitivity analysis using different methods in managing the missing data in order to measure the robustness of the outcomes and take into account other critical streamflow characteristic contributors like rainfall, temperature, topography or other parameters of the study area.

Funding
This work was supported by the Kolej Universiti Poly-Tech MARA Kuala Lumpur Micro-Grant.

Competing Interests
The authors declares no competing interests.

Author details
Fatimah Bibi Hamzah[1,2]
E-mail: bibi@gapps.kptm.edu.my
ORCID ID: http://orcid.org/0000-0003-2757-6141
Firdaus Mohd Hamzah[2]
E-mail: fir@ukm.edu.my
Siti Fatin Mohd Razali[2]
E-mail: fatinrazali@ukm.edu.my
Othman Jaafar[2]
E-mail: ojaafar@gmail.com
Norhayati Abdul Jamil[1]
E-mail: hayati@gapps.kptm.edu.my
[1] Faculty of Computing and Multimedia, Kolej Universiti Poly-Tech Mara Kuala Lumpur, Kuala Lumpur, 56100, Malaysia.
[2] Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi, Selangor, Malaysia.

Author Contributions
writing—original draft preparation, Fatimah Bibi Hamzah; writing—review and editing, Firdaus Mohamad Hamzah, Siti Fatin Mohd Razali, Othman Jaafar and Norhayati Abdul Jamil.

Conflicts of Interest
The authors declare no conflict of interest.

Citation information
Cite this article as: Imputation methods for recovering streamflow observation: A methodological review, Fatimah Bibi Hamzah, Firdaus Mohd Hamzah, Siti Fatin Mohd Razali, Othman Jaafar & Norhayati Abdul Jamil, *Cogent Environmental Science* (2020), 6: 1745133.

References

Adeloye, A. J., & Rustum, R. (2012). Self-organising map rainfall-runoff multivariate modelling for runoff reconstruction in inadequately gauged basins. *Hydrology Research*, *43*(5), 603. https://doi.org/10.2166/nh.2012.017

Adeloye, A. J., Rustum, R., & Kariyama, I. D. (2011). Kohonen self-organizing map estimator for the reference crop evapotranspiration. *Water Resources Research*, *47*(8), 1–19. https://doi.org/10.1029/2011WR010690

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

Aljuaid, T., & Sasi, S. (2017). Proper imputation techniques for missing values in data sets. *2016 International Conference on Data Science and Engineering (ICDSE)*. https://doi.org/10.1109/ICDSE.2016.7823957.

Allawi, M. F., Jaafar, O., Hamzah, F. M., Abdullah, S. M. S., & El-Shafie, A. (2018). Review on applications of artificial intelligence methods for dam and reservoir-hydro-environment models. *Environmental Science and Pollution Research*, *25*(14), 13446–13469. https://doi.org/10.1007/s11356-018-1867-8

Allawi, M. F., Jaafar, O., Mohamad Hamzah, F., Mohd, N. S., Deo, R. C., & El-Shafie, A. (2017). Reservoir inflow forecasting with a modified coactive neuro-fuzzy inference system: A case study for a semi-arid region. *Theoretical and Applied Climatology*, *134*(1–2), 545–563.. https://doi.org/10.1007/s00704-017-2292-5

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., Santhi, C., Harmel, R. D., Van Griensven, A., Van Liew, M. W., Kannan, N., & Jha, M. K. (2012). SWAT: Model use, calibration, and validation. *American Society of Agricultural and Biological Engineers*, *55*(4), 1491–1508. https://doi.org/10.13031/2013.42256

Bagus, I., & Narinda, G. (2016). Missing value imputation using KNN method optimized with memetic algorithm. *E-Proceeding Engineering* (pp. 1098–1105).

Beauchamp, J. J., Downing, D. J., & Railsback, S. F. (1989). Comparison of regression and time-series methods for synthesizing missing streamflow records. *Water Resources Bulletin*, *25*(5), 961–975. https://doi.org/10.1111/jawr.1989.25.issue-5

Ben Aissia, M. A., Chebana, F., & Ouarda, T. B. M. J. (2017). Multivariate missing data in hydrology – Review and applications. *Advances in Water Resources*, *110* (2017), 299–309. https://doi.org/10.1016/j.advwatres.2017.10.002

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. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

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

Blend, D., & Marwala, T. (2008). Comparison of data imputation techniques and their impact. *Scientific Commons*, *7*. Retrieved from https://www.researchgate.net/publication/23625477_

Comparison_of_Data_Imputation_Techniques_and_their_Impact/citations

Carpenter, D. H. (1985). *Cost-effectiveness of the federal stream-gauging program in Virginia*. Towson.

Caruso, C., & Quarta, F. (1998). Interpolation methods comparison. *Computers & Mathematics with Applications*, *35*(12), 109–126. https://doi.org/10.1016/S0898-1221(98)00101-1

Chen, F. W., & Liu, C. W. (2012). Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy and Water Environment*, *10*(3), 209–222. https://doi.org/10.1007/s10333-012-0319-1

Chen, J., & Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, *96*(453), 260–269. https://doi.org/10.1198/016214501750332839

Chen, S. H., & Pollino, C. A. (2012). Good practice in Bayesian network modelling. *Environmental Modelling & Software*, *37*(2012), 134–145. https://doi.org/10.1016/j.envsoft.2012.03.012

Čisar, P., & Čisar, S. M. (2011). Optimization methods of EWMA statistics. *Acta Polytechnica Hungarica*, *8*(5), 73–87. Retrieved from https://www.semanticscholar.org/paper/Optimization-Methods-of-EWMA-Statistics-%C4%8Cisar-Cisar/1f98b8a64d18489b513716c9277a146dac7e915a

Collins, A. J., Fox, J., & Long, J. S. (1991). Modern methods of data analysis. *The Statistician*, *40*(4), 458. https://doi.org/10.2307/2348744

Cooper, G. E., Herskovits, E., & Edu, F. (1992). A Bayesian method for the induction of probabilistic networks from data. *Journal of Machine Learning*, *9*(4), 309–347. https://doi.org/10.1007/BF00994110

Croninger, R. G., & Douglas, K. M. (2005). Missing data and institutional research, New Dir. *New Directions for Institutional Research*, *2005*(127), 33–49. https://doi.org/10.1002/ir.154

De Silva, R. P., Dayawansa, N. D. K., & Ratnasiri, M. D. (2007). A comparison of methods used in estimating missing rainfall data. *Journal of Agricultural Science*, *3*(2), 101. https://doi.org/10.4038/jas.v3i2.8107

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data method for researchers. *Springer Plus*, *2*(1), 222. https://doi.org/10.1186/2193-1801-2-222

Eash, D. A., Barnes, K. K., & Veilleux, A. G. (2013). *Methods for estimating annual exceedance-probability discharges for streams in iowa, based on data through water year 2010*. U.S. Geological Survey Scientific Investigations Report 2013–5086, 63 p. with appendix.

Elshorbagy, A., Simonovic, S. P., & Panu, U. S. (2002). Estimation of missing streamflowdata using principles of chaos theory. *Journal of Hydrology*, *255*(1–4), 123–133. https://doi.org/10.1016/S0022-1694(01)00513-3

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

Fleig, A. K., Tallaksen, L. M., Hisdal, H., & Hannah, D. M. (2011). Regional hydrological drought in north-western Europe: Linking a new regional drought area index with weather types. *Hydrological Processes*, *25*(7), 1163–1179. https://doi.org/10.1002/hyp.7644

Fontaine, R. A. (1982). *Cost effective stream-gaging strategies for Maine*. Augusta.

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

Gao, Y., Merz, C., Lischeid, G., & Schneider, M. (2018). A review on missing hydrological data processing. *Environmental Earth Sciences*, *77*(2), 47. https://doi.org/10.1007/s12665-018-7228-6

Gassman, P. W., Arnold, J. G., Srinivasan, R., & Reyes, M. (2010). The worldwide use of the SWAT model: Technological drivers, networking impacts, and simulation trends. *21st century watershed technology: Improving water quality and environment transacation ASABE* (pp. 21–24). https://doi.org/10.1088/0953-8984/6/4/007.

Gassman, P. W., Sadeghi, A. M., & Srinivasan, R. (2014). Applications of the SWAT model special section: Overview and insights. *Journal of Environmental Quality*, *43*(1), 1. https://doi.org/10.2134/jeq2013.11.0466

Gill, M. K., Asefa, T., Kaheil, Y., & McKee, M. (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. https://doi.org/10.1029/2006WR005298

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

Gómez-Carracedo, M. P., Andrade, J. M., López-Mahía, P., Muniategui, S., & Prada, D. (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems*, *134*(2014), 23–33. https://doi.org/10.1016/j.chemolab.2014.02.007

Graham, J. W. (2009). Missing data analysis : Making it work in the real world. *Annual Review of Psychology*, *60*(1), 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Gupta, P., & Srinivasan, R. (2011). Missing data prediction and forecasting for water quantity data. *International Conference on Modeling, Simulation and Control IPCSIT Vol.10. 10* (pp. 98–102). https://doi.org/10.7763/IPCSIT.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

Harvey, C. L., Dixon, H., & 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. https://doi.org/10.2166/nh.2012.110

Hasan, H., Ahmad, S., Osman, B. M., Sapri, S., & Othman, N. (2017). A comparison of model-based imputation methods for handling missing predictor values in a linear regression model: A simulation study. *Proceedings of the 24th national symposium mathematical sciences American institute of physics*. https://doi.org/10.1063/1.4995930.

Hipel, K. W. (1995). Stochastic and statistical methods in hydrology and environmental engineering. *Stochastic Hydrology and Hydraulics*, *9*(1), 1–11. https://doi.org/10.1007/BF01581755

Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, *54*(2), 561–581. https://doi.org/10.1111/j.1540-5907.2010.00447.x

Hsu, K., Gupta, H. V., & Sorooshian, S. (1995). Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research*, *31*(10), 2517–2530. https://doi.org/10.1029/95WR01955

Huo, J. (2005). *Application of statistical methods and process models for the design and analysis of activated*

sludge Wastewater Treatment Plants (WWTPs). University of Tennessee.

Huo, J., Cox, C. D., Seaver, W. L., Robinson, R. B., & Jiang, Y. (2010). Application of two-directional time series models to replace missing data. *Journal of Environmental Engineering*, *136*(4), 435–443. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000171

Hutchinson, M. F., & Gessler, P. E. (1994). Splines - more than just a smooth interpolator. *Geoderma*, *62*(1–3), 45–67. https://doi.org/10.1016/0016-7061(94)90027-2

Ingsrisawang, L., & Potawee, D. (2012). Multiple imputation for missing data in repeated measurements using MCMC and copulas. *International MultiConference of engineering and computer science II* (pp. 1–5).

Jain, A., & Indurthy, S. K. V. P. (2003). Comparative analysis of event-based rainfall-runoff modeling techniques—Deterministic, statistical, and artificial neural networks. *Water*, *8*(2), 93–98. https://doi.org/10.1061/(ASCE)1084

Jeong, D. I., & Kim, Y. O. (2009). Combining single-value streamflow forecasts - A review and guidelines for selecting techniques. *Journal of Hydrology*, *377*(3–4), 284–299. https://doi.org/10.1016/j.jhydrol.2009.08.028

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (pp. 338–345). Morgan Kaufmann. Retrieved October 26, 2018, from http://web.cs.iastate.edu/~honavar/bayes-continuous.pdf

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

Kabir, G., Tesfamariam, S., Hemsing, J., & Sadiq, R. (2019). Handling incomplete and missing data in water network database using imputation methods. *Sustainable and Resilient Infrastructure*, *00*, 1–13. https://doi.org/10.1080/23789689.2019.1600960

Kalteh, A. M., Hjorth, P., & Berndtsson, R. (2007). Review of the Self-Organizing Map (SOM) approach in water resources: Analysis, modelling and application. *Environmental Modelling & Software*, *23*(7), 835–845. https://doi.org/10.1016/j.envsoft.2009.01.008

Kalton, G., & Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics - Theory Methods.*, *13*(16), 1919–1939. https://doi.org/10.1080/03610928408828805

Kamaruzaman, I. F., Wan Zin, W. Z., & Mohd Ariff, N. (2017). A comparison of method for treating missing daily rainfall data in Peninsular Malaysia. *Malaysian Journal of Fundamental & Applied Sciences*, *13*(4–1), 375–380. https://doi.org/10.11113/mjfas.v13n4–1.781

Karakurt, O., Erdal, H. I., Namli, E., Yumurtaci-Aydogmus, H., & Turkkan, Y. S. (2013). Comparing ensembles of decision trees and Neural networks for one-day-ahead stream flow predict. *Science Park*, *1*(17), 1–12. https://doi.org/10.9780/23218045/1172013/41

Kim, J., & Pachepsky, Y. A. (2010). Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *Journal of Hydrology*, *394*(3–4), 305–314. https://doi.org/10.1016/j.jhydrol.2010.09.005

Kim, M., Baek, S., Ligaray, M., Pyo, J., Park, M., & Cho, K. H. (2015). Comparative studies of different imputation methods for recovering streamflow observation. *Water (Switzerland)*, *7*(12), 6847–6860. https://doi.org/10.3390/w7126663

Kim, S. U., & 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. https://doi.org/10.1007/s12205-010-0087-7

Kohonen, T., Oja, E., Simula, O., Visa, A., & Kangas, J. (1996). Engineering applications of the self-organizing map. *Proceedings of the IEEE*, *84*(10), 1358–1383. https://doi.org/10.1109/5.537105

Krysanova, V., & White, M. (2015). Advances in water resources assessment with SWAT—an overview. *Hydrological Sciences Journal*, *60*(5), 1–13. https://doi.org/10.1080/02626667.2015.1029482

Kumar, A. S., Kumar, A., Krishnan, R., Chakravarthi, B., & Deekshatalu, B. L. (2017). Soft computing in remote sensing applications. *Proceedings of the National Academy of Sciences, India Section A: Physical Sciences 87*(4): 503–517. https://doi.org/10.1007/s40010-017-0431-0

Lee, H., & Kang, K. (2015). Interpolation of missing precipitation data using kernel estimations for hydrologic modeling. *Advances in Meteorology*, *2015(5)*, 12. https://doi.org/10.1155/2015/935868

Little, R., & An, H. (2004). Robust likelihood-based analysis of multivariate Data with missing values. Statistica Sinica, *14*(3), 949–968. U.S. Geological Survey Scientific Investigations Report 2013–5086, 63 p. with appendix. Retrieved from https://www.jstor.org/stable/24307424?seq=1

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). A JOHN WILEY & SONS, INC. https://doi.org/10.1002/9781119013563

Machiwal, D., & Jha, M. K. (2008). Comparative evaluation of statistical tests for time series analysis : Application to hydrological time series. *Hydrological Sciences Journal*, *53*(2), 353–366. https://doi.org/10.1623/hysj.53.2.353

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositivedefinite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling A Multidisciplinary Journal*, *5*(1), 22–36. https://doi.org/10.1080/10705519809540087

McDonald, R. A., Thurston, P. W., & Nelson, M. R. (2000). A Monte Carlo study of missing item methods. *Organisational Research Methods*, *3*(1), 71–92. https://doi.org/10.1177/109442810031003

McKnight, P. E. (2007). *Missing data : A gentle introduction*. Guilford Press.

Meadows, E. A., & Jeffcoat, H. H. (1990). *Water resources publication for Alabama, 1857-1990*. Denver: U.S. Geological Survey Books and Open-File Reports Federal Center.

Minns, A. W., & Hall, M. J. (1996). Artificial neural networks as rainfall- runoff models. *Hydrological Sciences Journal*, *41*(3), 399–417. https://doi.org/10.1080/02626669609491511

Miró, J. J., Caselles, V., & Estrela, M. J. (2017). Multiple imputation of rainfall missing data in the Iberian Mediterranean context. *Atmospheric Research*, *197* (2017), 313–330. https://doi.org/10.1016/j.atmosres.2017.07.016

Mispan, M. R., Rahman, N. F. A., Ali, M. F., Khalid, K., Bakar, M. H. A., & Haron, S. H. (2015). Missing river discharge data imputation Approach using artificial neural network. *Journal of Agricultural and Biological Science*, *10*(22), 10480–10485. Retrieved from http://www.arpnjournals.org/jeas/research_papers/rp_2015/jeas_1215_3088.pdf

Moahmed, T. A., El Gayar, N., & Atiya, A. F. (2014). Forward and backward forecasting ensembles for the estimation of time series missing data. *IAPR workshop on artificial neural networks in pattern recognition* (pp. 93–104). https://doi.org/10.1007/978-3-642-12159-3.

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

Mwale, F. D., Adeloye, A. J., & 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. https://doi.org/10.1016/j.pce.2012.09.006

Neitsch, S., Arnold, J., Kiniry, J., & Williams, J. (2011). *SWAT theoretical documentation version 2009*. Texas Water Resources Institute. Texas: Texas A&M University System. https://doi.org/10.1016/j.scitotenv.2015.11.063

Nishanth, K. J., & Ravi, V. (2013). A computational intelligence based online data imputation method: An application for banking. *Journal of Information Processing Systems*, *9*(4), 633–650. https://doi.org/10.3745/JIPS.2013.9.4.633

Norliyana, W., Ismail, W., Zawiah, W., Zin, W., & Ibrahim, W. (2017). Estimation of rainfall and stream flow missing data for Terengganu, Malaysia by using interpolation technique methods. *Malaysian Journal of Fundamental & Applied Sciences*, *13*(3), 213–217. https://doi.org/10.11113/mjfas.v13n3.578

Oosthuizen, N., Hughes, D. A., Kapangaziwiri, E., Mwenge Kahinda, J.-M., & Mvandaba, V. (2018). Parameter and input data uncertainty estimation for the assessment of water resources in two sub-basins of the Limpopo River Basin. *Proceedings of the international association of hydrological sciences, copernicus publications on behalf of the international association of hydrological sciences* (pp. 11–16). https://doi.org/10.5194/piahs-378-11-2018.

Peña-angulo, D., Nadal-romero, E., González-hidalgo, J. C., Albaladejo, J., Andreu, V., Bagarello, V., Barhi, H., Batalla, R. J., Bernal, S., Bienes, R., Campo, J., Campo-bescós, M. A., Canatario-duarte, A., Cantón, Y., Casali, J., Castillo, V., Cerdà, A., Cheggour, A., Cid, P., Cortesi, N., … Zorn, M. (2019). Spatial variability of the relationships of runo ff and sediment yield with weather types throughout the Mediterranean basin. *Journal of Hydrology*, *571*(2019), 390–405. https://doi.org/10.1016/j.jhydrol.2019.01.059

Perry, M. B. (2011). The exponentially weighted moving average *Wiley encyclopedia of operations … Management Science* (pp. 1–9). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470400531.eorms0314.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, *7*(4), 353–383. https://doi.org/10.1076/edre.7.4.353.8937

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

Quinlan, J. R., & Kaufmann, M. (1994). *C4.5: Programs for machine learning*. Khawer Academic.

Raaijmakers, Q. A. W. (1999). Effectiveness of different missing data treatments in surveys with likert-type data: Introducing the relative mean substitution approach. Educational and Psychological Measurement, *59*(5), 725–748. https://doi.org/10.1177/00131644995950001

Rahman, N. A., Deni, S. M., & Ramli, N. M. (2017). Generalized linear model for estimation of missing daily rainfall data. *4th international conference on applied physics* (pp. 0800191–0800198). https://doi.org/10.1063/1.4981003.

Rahman, N. F. A., Ali, M. F., Mohd, M. S. F., Khalid, K., Haron, S. H., Kamaruddin, H., & Mispan, M. R. (2015). Semi distributed hydro climate model; The Xls2NCascii program approach for weather generator. *Journal of Agricultural and Biological Science*, *10*(15), 6619–6622. Retrieved from https://www.researchgate.net/publication/281264842_Semi_distributed_hydro_climate_model_The_Xls2NCascii_program_approach_for_weather_generator

Rajagopalan, B., & Lall, U. (1999). Ak-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Research*, *35*(10), 3089–3101. https://doi.org/10.1029/1999WR900028

Refsgaard, J. C., Storm, B., & Clausen, T. (2010). Système Hydrologique Europeén (SHE): Review and perspectives after 30 years development in distributed physically-based hydrological modelling. *Hydrology Research*, *41*(5), 355. https://doi.org/10.2166/nh.2010.009

Regonda, S. K., Seo, D. J., Lawrence, B., Brown, J. D., & Demargne, J. (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. https://doi.org/10.1016/j.jhydrol.2013.05.028

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, *1*(3), 239–250. https://doi.org/10.1080/00401706.1959.10489860

Roth, P. L., Switzer, F. S., III, & Switzer, D. M. (1999). Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques. *Organisational Research Methods*, *2*(3), 211–232. https://doi.org/10.1177/109442819923001

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Sakke, N., Ithnin, H., Ibrahim, M. H., Pah, T., & Syed, R. (2016). Hydrological drought and the sustainability of water resources in Malaysia : An analysis of the properties of the Langat Basin, Selangor. *Malaysian Journal of Society and Space*, *12*(7), 133–146. Retrieved from https://www.researchgate.net/publication/313497530_Kemarau_hidrologi_dan_kelestarian_sumber_air_di_Malaysia_Kajian_analisis_sifat_Lembangan_Langat_Selangor

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Santosa, B., Legono, D., & Suharyanto. (2014). Prediction of missing streamflow data using principle of information entropy. *Civil Engineering Dimension*, *16*(1), 40–45. https://doi.org/10.9744/ced.16.1.40-45

Schafer, J. L. J. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall.

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

Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, *14*(5), 853–871. https://doi.org/10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2

Shields, F. D., & Sanders, T. G. (1986). Water quality effects of excavation and diversion. *Journal of Environmental Engineering*, *112*(2), 211–228. https://doi.org/10.1061/(ASCE)0733-9372(1986)112:2(211)

Smith, J., & Eli, R. N. (1995). Neural-network models of rainfall-runoff process. *Journal of Water Resources Planning and Management*, *121*(6), 499–508. https://doi.org/10.1061/(ASCE)0733-9496(1995)121:6(499)

Somwanshi, P. D., & Chaware, S. M. (2014). A review on: Advanced Artificial Neural Networks (ANN) approach for IDS by layered method. *International Journal of Computer Science and Information Technologies*, *5*(4), 5129–5131. Retrieved from http://ijcsit.com/docs/Volume%205/vol5issue04/ijcsit2014050466.pdf

Spiring, F. (2007). Introduction to statistical quality control. *Technometrics*, *49*(1), 108–109. https://doi.org/10.1198/tech.2007.s465

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th ed.). Pearson Education Limited.

Teegavarapu, R. S. V., & Chandramouli, V. (2005). Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology*, *312*(1–4), 191–206. https://doi.org/10.1016/j.jhydrol.2005.02.015

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

Tencaliec, P., Favre, A., Prieur, C., & Mathevet, T. (2015). Reconstruction of missing daily streamflow data using dynamic regression models. *Water Resources Research: American Geophysical Union*, *51*(12), 9447–9463. https://doi.org/10.1002/2015WR017399

Tsintikidis, D., Haferman, J. L., Anagnostou, E. N., Krajewski, W. F., & Smith, T. F. (1997). A neural network approach to estimating rainfall from space-borne microwave data. *IEEE Transactions on Geoscience and Remote Sensing,35*(5), 1079–1093. http://doi.org/10.1109/36.628775

Tuppad, P., Mankin, K. R. D., Lee, T., Srinivasan, R., & Arnold, J. G. (2011). Soil and Water Assessment Tool (SWAT) hydrologic/water quality model: Extended capability and wider adoption. *American Society of Agricultural and Biological Engineers*, *54*(5), 1677–1684. https://doi.org/10.13031/2013.39856

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, *203*(3–4), 312–318. https://doi.org/10.1016/j.ecolmodel.2006.11.033

Varga, M., Balogh, S., & Csukas, B. (2016). GIS based generation of dynamic hydrological and land patch simulation models for rural watershed areas. *Information Processing in Agriculture*, *3*(1), 1–16. https://doi.org/10.1016/j.inpa.2015.11.001

Vigerstol, K. L., & Aukema, J. E. (2011). A comparison of tools for modeling freshwater ecosystem services. *Journal of Environmental Management*, *92*(10), 2403–2409. https://doi.org/10.1016/j.jenvman.2011.06.040

Wallis, J. R., Lettenmaier, D. P., & Wood, E. F. (1991). A daily hydroclimatological data set for the continental United States. *Water Resources Research*, *27*(7), 1657–1663. https://doi.org/10.1029/91WR00977

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. https://doi.org/10.1111/j.1540-5834.2006.00404.x

Woodall, W. H., & Mahmoud, M. A. (2005). The inertial properties of quality control charts. *Technometrics*, *47*(4), 425–436. https://doi.org/10.1198/004017005000000256

Worland, S. C., Farmer, W. H., & Kiang, J. E. (2018). Improving predictions of hydrological low-flow indices in ungaged basins using machine learning. *Environmental Modelling & Software*, *101*(2018), 169–182. https://doi.org/10.1016/j.envsoft.2017.12.021

Yakowitz, S., & Karlsson, M. (1987). Nearest neighbor methods for time series, with application to rainfall/runoff prediction. *Advances in the statistical sciences: Stochastic hydrology* (pp. 149–160). Springer Netherlands. https://doi.org/10.1007/978-94-009-4792-4_9.

Yashchin, E. (1987). Some aspects of the theory of statistical control schemes. *IBM Journal of Research and Development*, *31*(2), 199–205. https://doi.org/10.1147/rd.312.0199

Yashchin, E. (1993). Statistical control schemes: Methods, applications and generalizations. *International Statistical Review*, *61*(1), 41–66. https://doi.org/10.2307/1403593

Zeiger, S., & Hubbart, J. (2018). Assessing the difference between Soil and Water Assessment Tool (SWAT) simulated pre-development and observed developed loading regimes. *Hydrology*, *5*(2), 29. https://doi.org/10.3390/hydrology5020029

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

***Cogent Environmental Science* (ISSN: 2331-1843) is published by Cogent OA, part of Taylor & Francis Group.**