ISSN 2278-3091
Volume 9, No.1.4, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse9091.42020.pdf
https://doi.org/10.30534/ijatcse/2020/9091.42020

Imputation methods for addressing missing data of monthly rainfall in Yogyakarta, Indonesia

Shazlyn Milleana Shaharudin[1], Sri Andayani[2], Kismiantini3, Nikenasih Binatari[4], Agusta Kurniawan[5]
Muhammad Afdal Ahmad Basri[6], Nurul Hila Zainuddin[7]
[1,6,7]Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Malaysia
[2,4]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Indonesia
[3]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Indonesia
[5]Climatology Station Class IV Sleman, Meteorological Climatological and Geophysical Agency, Indonesia
shazlyn@fsmt.upsi.edu.my
andayani@uny.ac.id
kismi@uny.ac.id
agusta.kurniawan@bmkg.go.id
nikenasih@uny.ac.id.
D073909@siswa.upsi.edu.my
nurulhila@fsmt.upsi.edu.my

ABSTRACT

Rainfall data are the most significant values in hydrology and climatology modelling. However, the datasets are prone to missing values due to various issues. This study aspires to impute the rainfall missing values by using various imputation methods such as Replacing by Mmean (RM), Nearest Neighbor (NN), Random Forest (RF), Non-linear Interactive Partial Least-Square (NIPALS) and Markov Chain Monte Carlo (MCMC). Monthly rainfall datasets from 24 rainfall stations in Yogyakarta, Indonesia were used in this study. The datasets were then used for bootstrapping to obtain an estimate of the within-imputation standard errors for each imputed dataset. The performances of five methods were evaluated using root mean square method (RMSE). The experimental results showed that the RF-Bootstrap (RF-B) approach was attained as the most satisfying fitting for missing rainfall data in Yogyakarta, Indonesia.

Key words: MCMC, Missing value, nearest neighbor, NIPALS, random forest, replace by mean, bootstrap

1. INTRODUCTION

Rainfall are the most significant variables in climatology and hydrological modelling. However, missing data of rainfall are common problems in climatic series due to many circumstances. Most of the statistical analyses require a complete data rather than the data with missing values. In order to conduct statistical analysis, the crucial problem of missing data forces researchers to choose either imputing data or discarding missing values method to be used [1]. However, simply discard the missing data is not a reasonable practice, as valuable information may be lost and inferential power compromised [2]. Thus, the imputing missing data is the most suitable and more practical way to proceed.

According to [3], three types of missing data were Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). In real life situation, the MCAR is practically used. In hydrological data, especially in the case of missing rainfall datasets, it is classified as MCAR since the data in a particular area does not affect the occurrence of missing in rainfall datasets of an area [4]. Missing data were inserted completely at random (MCAR) in the following percentages: 10%, 20%, 30%, 40%, 50% and 60% of the total of instances [3]. The problem of missing values in meteorological series is particularly significant in developing countries where gauging stations are scarce and degree of missingness is large [5]. Basically, missing data can be caused by human errors in collecting and managing the datasets, natural disaster, and machinery defects on site [6]. Thus, ignoring missing data can eventually lead to partial and biased results in data analysis [7].

In order to handle the type of input data, this study introduces five methods of imputation. The motivation is to make as few assumptions as possible about structural aspects of the data due to large missing values, where (> 5% - 10%) is categorized as large missing values [8] . According to [9], the solution to the problem is a real challenge, when a large proportion (30% or more) of data is missing. In recent years, many methods have been applied to fill gaps in precipitation series such as Arithmetic Mean, Normal Ratio, Inverse Distance Weighting, Weighted Linear Regression, Multiple Linear Regression and Probabilistic method [10]-[14]. In the field of hydrologic modelling, it is extremely important to use the most efficient method to achieve the best rainfall valuation. In this study, several imputation techniques were proposed including using Replace by Mean, Nearest Neighbor, Markov Chain Monte Carlo (MCMC), Nonlinear Iterative Partial Least Squares (NIPALS) and Random Forest (RF) method [15]-[19]. Random Forest is capable of efficiently incorporating a large number of both

continuous and categorical variables, as a result of sub-sampling predictor variables at each node when constructing regression trees [20].Additionally, these algorithms have desirable properties which are their capability of handling diversified types of missing data.

In this study, an ensemble strategy for missing value imputation is proposed, which provides an alternative solution to the problem of how to choose the optimum imputation tool for different application. The main contributions in this study are as follows: (1) Identify the best method to overcome the issue of large missing value, and (2) Identify the validity of reducing Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) by using the Bootstrap algorithm. The data imputation was performed using Replace by Mean, Nearest Neighbor (NN), Markov Chain Monte Carlo (MCMC), Nonlinear Iterative Partial Least Squares (NIPALS) and Random Forest (RF) methods for monthly rainfall data in Yogyakarta, Indonesia. The outcomes from this study are expected to contribute towards finding the best and finest method for data imputation technique which enables the reconstruction of complete rainfall datasets. However, challenges remain when applying these methods for estimating up to 60% missing values in data set.

2. RESULTS METHOD

2.1 Replace by mean

According to [21] replace by mean are single-value imputation methods which is estimate each missing value might have been and replace it with a single value in the data set.Advantage of using this technique is that it is not unduly complicated and can be easily implemented in most common statistical packages.

The mean and the standard deviation of each variable were calculated using the actual complete data and compared with their counterparts after replacing the missing values in each condition. The absolute difference in mean and standard deviation was used to evaluate the effectiveness of both procedures. However, accurately reproducing means and standard deviations is just one criterion for determining the efficiency of a method. Another criterion which is the accuracy in estimating parameters must also be used as part of statistical analysis. It comprises changing every missing value in series (k), k=1, …, d by the respective component mean. The formula can be written as:

$$P_x = \frac{\sum_{i=1}^{n} P_i}{n} \quad (1)$$

where P_x is the observed rainfall data, $P_i$is …. and$n$ is the number of rainfall days.

2.2 Nearest neighbor

Another effective method to fulfill the missing data is nearest neighbor imputation algorithm. It is also significant that the use of different kernel functions can improve the performance of k-nearest neighbor (KNN) imputation when predicting missing rainfall data [22]. Every missing value on a number of records is substituted by a value acquired in the overall records set from interrelated cases [23]. KNN method is proven to be reliable and practical to treat missing hydrological data through the application of KNN imputation conducted by [24]. The first step of nearest neighbor imputation is by assuming n observations gathered on p covariates. The corresponding $n \times p$ data matrix is given by $\mathbf{X}=(x_{is})$, where x_{is} denotes the ith observation of the sth variable. Let $\mathbf{O}=(o_{is})$ denotes the corresponding $n \times p$ matrix of dummies with entries given below,

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

The distances between two observations of x_i and x_j is calculated by using the Lq-metric from the observed data,

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

Where $m_{ij} = \sum_{s=1}^{p} I(o_{is} = 1) I(o_{js} = 1)$denotes the number of valid components in the distances computation.

2.3 Markov Chain Monte Carlo (MCMC)

A Markov Chain Monte Carlo (MCMC) algorithm is another useful method for missing data imputation. This algorithm is highly illustrative example of incomplete data [25].According to [26], the expectation-maximization (EM) is a technique that replaces missing data by figuring the maximum probable valuations for MCMC process. The MCMC method is based on Bayesian inference with missing data by obeying several steps as follows [27].

1) imputation step. Estimate the mean and covariance matrix, then simulate the missing values for each observation.
2) posterior step (P-step). Simulate the mean vector and covariance matrix from the imputed step.

2.4 Non-linear interactive partial least-square (NIPALS)

The non-linear interactive partial least-square (NIPALS) is another imputation method that can used in principal component analysis (PCA) problem with missing values. This approach was implemented in commercial chemometric software with varying degree of correctness for the general PCA problem with missing values [28]. In most data analyses require complete data. The complete data may be obtained in the simplest way by removing any rows with missing values, but this approach can lead to large amount of data loss or undesirable bias [29]. A better approach can be conducted by replacing the missing value with the sample mean.. Given a rectangular dataset with the size of $n \times p$, the algorithms of NIPALS work as follows,

1) define a matrix for the ith observed value in the jth variable, $\mathbf{X} = \{x_{ij}\}, 1 \le i \le n, 1 \le j \le p$.

2) assume **X**has a rank of a, then decompose **X** as $\boldsymbol{X} = \sum_{h=1}^{a} t_h P'_h$, where $t_h = (t_{h1},\ldots, t_{hi},\ldots,t_{hn})'$ and $P_h = (P_{h1},\ldots, P_{hj},\ldots,P_{hp})'$ are the principal factors and principal components, respectively.
3) estimate the missing value using the NIPALS algorithm to the cell $(i,)$ as:

$$x_{ij} = \sum_{l=1}^{k} t_{li} p_{lj} \quad (4)$$

where k ($k \leq a$) is established by a cross-validation. The implementation of NIPALS algorithm is straightforward, based on simple linear regressions.

2.5 Random forest

Random forest is an effective tool in prediction. Random forests (RFs) are very flexible and powerful ensemble classifiers based on decision trees which were firstly developed by Breiman (2001) [30]. In addition, the framework gives insight into the ability of the random forest to predict in terms of strength of the individual predictors and their correlations. The random forest can be applied for classification, regression, and unsupervised learning [31]. By assuming that $\mathbf{X} = (X_1, X_2, \ldots X_P)$ as a $n \times p$-dimensional data matrix and X_S as an arbitrary variable with missing values at entries $i_{mis}^{(s)} \subseteq \{1,\ldots,n\}$, the rainfall dataset can be separated into two categories: 1) the observed values of variable X_S are denoted by $y_{obs}^{(s)}$, 2) the missing values of variable $X_S$are denoted by $y_{mis}^{(s)}$.

An initial guess for the missing value in **X** can be determined by using mean or other imputation method. The variables $X_S, s = 1,\ldots,p,$ are sorted based on the total missing values beginning with the smallest amount. For each variable X_S, the missing values are imputed by fitting an Random Forest with response $y_{obs}^{(s)}$ and predictors $x_{obs}^{(s)}$; and predicting the missing values of $y_{mis}^{(s)}$ by applying trained Random Forest to $x_{mis}^{(s)}$. The imputation procedure is then repeated until a stopping criterion is satisfied.

2.6 Bootstrap Approach

The bootstrap approach was introduced by [32] in 1979 and is well known for its functional in reducing the inherent variance and bias in sample set of data. Theoretically, a small variance indicates that the estimate from the specified imputation method is very close to the true value of the expected value [33].

Assuming $\mathrm{X_{k,i}} = \mathrm{x_{1,1}},\ldots,\mathrm{x_{24,n}}$ is a sample set of variables data with $k = 1,..,24$ variables and $i = 1,..,n$ observations, where *n* is a sample size of each variable, i.e. *n*=601. The X_{ki} data matrix is used to obtain a group of bootstrap replication data through sampling with replacement,$x_{1,i}^{B(t)}$where *B(t)* refers to a number of bootstrap replication. It is common to employ a 1000 replication in bootstrapping, *t* = 1,…,1000 [34]. For example, the replication for the first variable can be written as follows:

$$\boldsymbol{x}_{\mathbf{1,i}}^{\boldsymbol{B(1)}} \begin{bmatrix} \boldsymbol{x}_{\mathbf{1,1}}^{\boldsymbol{B(1)}} & \ldots & \boldsymbol{x}_{\mathbf{1,1}}^{\boldsymbol{B(1000)}} \\ \vdots & & \vdots \\ \boldsymbol{x}_{\mathbf{1,601}}^{\boldsymbol{B(1)}} & \ldots & \boldsymbol{x}_{\mathbf{1,601}}^{\boldsymbol{B(1000)}} \end{bmatrix} \quad (5)$$

A set of bootstrap samples can be obtained by calculating the average of each row of $x_{1,i}^{B(t)}$ matrix as given below:

$$x_{1,i}^{B} = \frac{\sum_{i=1}^{t} x_{1,i}^{B(t)}}{T} \quad (6)$$

where T represent the total number of bootstrap replications i.e. T = 1000. In terms of matrix notation, the bootstrap sample of the first variable can be rewritten as follows:

$$\boldsymbol{x}_{\mathbf{1},i}^{\boldsymbol{B}} = \boldsymbol{x}_{\mathbf{1,1}}^{\boldsymbol{B}},\ldots,\boldsymbol{x}_{\mathbf{1,601}}^{\boldsymbol{B}} \quad (7)$$

The bootstrap approach is eventually useful for increasing accuracy or validity of the statistical estimation. In order to investigate the validity of the estimate, the RMSE and MAE estimation is used and explained in next section.

2.7 Root mean square error (RMSE), Mean absolute error (MAE), and Nash-sutcliffe efficiency coefficient (CE)

The performance of the imputation methods was evaluated using root mean square root (RMSE) and mean absolute error (MAE). The RMSE is a standard statistical metric to measure the performance of model studies in meteorology, air quality and climate [35]. The RMSE, MAE and CE has been used to assess model performance for many years but here is no consensus on the most appropriate metric for model errors. In the field of geosciences, the RMSE is often used as a standard metric for model errors [36]. CE value of 1 are pursued in the best performance models. The mean absolute error (MAE) is another useful measure widely used in model evaluations. The lower the RMSE and MAE, the more accurate the evaluation is[37].The RMSE, MAE and CE formula is given below,

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 / n} \quad (8)$$

$$\mathrm{CE} = 1 - \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 / \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \quad (9)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|(y_i \tilde{y}_i)^2| \quad (10)$$

where y_i is the observed rainfall data and $\tilde{y}_i$ is the predicted rainfall data in Yogyakarta, Indonesia.

2.8 Research Approach

Briefly, this paper focuses on two statistical strategies in identify the extreme missing value of rainfall patternsin the city of Yogyakarta, Indonesia,

a) Identify the best method to overcome the issue of large missing values.

b) Identify the validity of reducing Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) by using the Bootstrap algorithm.

This approach is as in figure 1. In order to achieve these two statistical strategies, Replace by Mean, NN, MCMC, NIPALS and RF imputation methods are used in this study. This approaches to find out the best imputation method from the large missing value and reduced the value of the imputation by using the bootstrap algorithm.

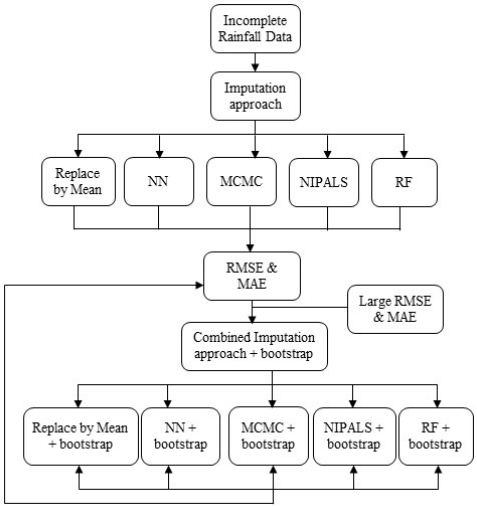

Figure 1: The flow of the proposed model

2.9 Study Area and Data

This study of rainfall focused on the long narrow Yogyakarta covering 7° S latitude and 110° E longitude. The monthly rainfall data from 24 stations over Yogyakarta were obtained from Meteorological, Climatological and Geophysical Agency and represented in Figure 2. Eventually, the rainfall data are incomplete with missing values starting from year 1970 to 2019 with 58.1% of total missing values. According to [38], the dataset containing 50-60% of missing values are regarded as the high degree of missingness in precipitation time series. Other than that, the reliability of a long time series data frequency estimator is highly valuable since it strongly associates with sample size in data analysis. Figure 2 shows the geographical coordinates of 24 rainfall stations chosen from the area of Yogyakarta, Indonesia.

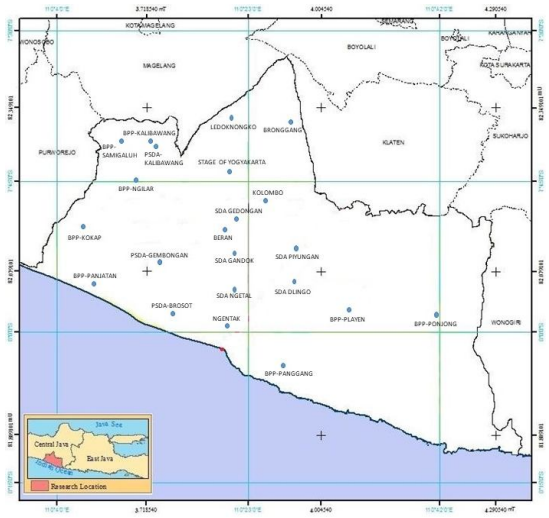

Figure 2: The geographical coordinates of 24 rainfall stations chosen from the area of Yogyakarta, Indonesia.

3. RESULTS AND DISCUSSION

This section discusses the results of imputation methods for monthly rainfall datasets from January 1970 until December 2019. The imputation methods were applied to 24 stations in Yogyakarta, Indonesia. The experiments were conducted for each station, using all five imputation methods. The results were then calculated as an average result, representing each imputation method. The RMSE and MAE was calculated in each method. The method with the smallest RMSE and MAE value was selected as the best technique for filling out the missing data from monthly rainfall datasets. The experimental results for each imputation method are showed in Table 1.

Table 1 indicates the average RMSE and MAE values for the five methods. The results showed that the smallest RMSE and MAE was obtained from the RF method. It is noted that higher RMSE and MAE values are obtained in this study because of a high model variance when the rainfall data is large [39]. In addition, the largest RMSE and MAE obtained in this study due to the high proportion of missing data (>50%). Meanwhile the NIPALS was the worst imputation method for monthly rainfall data in Yogyakarta, Indonesia as the results showed that the NIPALS has the highest RMSE among other methods.

Table 1: Average RMSE values for five imputation methods

Method	RMSE	MAE
MCMC	217.14	148.7526
NIPAL	29756.62	1482.937
NN	174.19	97.71053
MEAN	169.23	121.837
RF	**153.96***	**92.25785***

***indicate the best results**

The next step in this study, once the missing values have been filled in, is to use bootstrapping to obtain an estimate of the within-imputation standard error for each

method by rearranging the completed imputing missing data. The bootstrap models were used to identify the best approaches for handling missing data. To evaluate the performance of the five imputation methods coupled with bootstrap model, the RMSE and MAE was used respectively.

Table 2 presents the RMSE and MAE values for each method coupled with bootstrap for imputing the missing values of daily rainfall data in Yogyakarta, Indonesia. It can be observed that the RF method has the lowest RMSE and MAE of 7.9551 and 0.286298 compared to other methods. Thus, the final results suggested that Random Forest coupled with bootstrap is the best statistical method for imputing the missing values of daily rainfall data in Yogyakarta, Indonesia. Table 2 also showed that the imputation method of NN coupled with bootstrap has relatively small RMSE and MAE, so it can be used as an alternative.

Table 2: Average RMSE values for five imputation methods

Method	RMSE	MAE
MCMC-Bootstrap	9.4341	0.348005
NIPAL-Bootstrap	127.5058	2.117418
NN-Bootstrap	8.2134	0.296501
MEAN-Bootstrap	8.6984	0.319056
RF-Bootstrap	**7.9551***	**0.286298***

***indicate the best results**

With these results, the RF was considered the best model to fill the gaps in the missing values and the bootstrap algorithm was valid to reduce the RMSE and MAE value of monthly rainfall data in Yogyakarta, Indonesia.

4. CONCLUSION

The search of the most efficient method for imputation of the missing rainfall values are the most important. In this study, we have tackled the added difficulty of estimating extremely large amount of missing values (58.1%). Five imputation methods of Replace by Mean, NN, MCMC, NIPALS and RF were used and compared to obtain the most appropriate technique for filling the missing data in monthly rainfall data in Yogyakarta, Indonesia. Bootstrap algorithm coupled with imputation methods were used to reduce the value of RMSE and MAE. The results showed that the Random Forest-Bootstrap (RF-B) was the best method for single imputation when estimating extremely large amount of missing values. Therefore, finding the best method for data imputation is crucial to improve any predictive modelling. To conclude, the use of various imputation techniques based on the characteristics of rainfall is endorsed, and further studies should be explored with different methodologies and datasets.

ACKNOWLEDGEMENTS

This study was conducted under the Research Scheme of International Cooperation Research Program Study through Vote No. B/236/UN35.21/TU/2020 offered by the Ministry of Education and Culture Universitas Negeri Yogyakarta, Institute of Research and Community Service.

REFFERENCES

[1] Siti Mariana Che Mat Nor, Shazlyn MilleanaShaharudin, Shuhaida Ismail, Nurul Hila Zainuddin, Mou Leong Tan. **A comparative study of different imputation methods for daily rainfall data in east-coast Peninsular Malaysia,** *Bulletin of Electrical Engineering and Informatics.* vol. 9, no.2, pp. 635-643, April 2020.

[2] F. Tang and H. Ishwaran. **Random forest missing data algorithms,** *Statistical Analysis and Data Mining: The ASA Data Science Journal.* vol. 10, no.6, pp. 363–377, June 2017.

[3] Gustavo E. A. P. A. Batista and Maria Carolina Monard. **A Study of K-Nearest Neighbour as an Imputation Method,**2003.

[4] M. A.Ben Aissia, F. Chebana, andT. B. M. J. Ouarda. **Multivariate missing data in hydrology – Review and applications**, *Advances in Water Resources.* vol. 110, pp. 299–309, December 2017. https://doi.org/10.1016/j.advwatres.2017.10.002

[5] E. Nkiaka, N. R. Nawaz, and J. C. Lovett.**Evaluating global reanalysis precipitation datasets with rain gauge measurements in the Sudano-Sahel region: case study of the Logone catchment, Lake Chad Basin,***Meteorological Applications.* vol. 24, no.1, pp. 9–18, December 2016.

[6] Wai Yan Lai, Kuok King Kuok, Shirley Gato-Trinidad, Derrick, andKuoXiong Ling. **A Study on Sequential K-Nearest Neighbor (SKNN) Imputation for Treating Missing Rainfall Data,***International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 3, pp. 363-368, 2019. https://doi.org/10.30534/ijatcse/2019/05832019

[7] O. Harel, and X. H. Zhou,**Multiple imputation: review of theory, implementation and software**, *Statistics in Medicine*. vol. 26, pp. 3057–3077, January 2007.

[8] R. Lo Presti, E. Barca, and G. Passarella. **A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy),** *Environmental Monitoring and Assessment*,vol. 160, no.1-4, pp. 1–22, December 2008.

[9] Hamzeh M. Dodeen. **Effectiveness of Valid Mean Substitution in Treating Missing Data in Attitude Assessment,** *Assessment & Evaluation in Higher Education*,vol. 28, no.5, pp. 505-513, May 2003.

[10] L. Campozano, E. Sanchez, A. Aviles and E. Samaniego. **Evaluation of Infilling Methods for Time Series of Daily Precipitation and Temperature: The Case of the Ecuadorian Andes,***MASKANA*, vol 5, no.1, pp.101-15, June 2014.

[11] M. Hasan and B. Croke.**Filling Gaps in Daily Rainfall Data: A Statistical Approach,***In 20th*

International Congress on Modelling and Simulation, Adelaide, Australia, December 2013.

[12] V. T. Chow, D. R. Maidment and L. W. Mays. ***Applied Hydrology***, McGraw-Hill Book Company, Singapore, 1988.

[13] V. P. Singh. ***Elementary Hydrology,*** Prentice Hall of India, New Delhi, 1994.

[14] F. W. Chen and C. W. Liu.**Estimation of the Spatial Rainfall Distribution using Inverse Distance Weighting (IDW) in the Middleof Taiwan,***Paddy Water Environment*,vol. 10, pp. 209-222, February 2012.

[15] L. Breiman.***Machine Learning***, vol. 45, no.1, pp. 5–32, 2001.

[16] W. R. Gilks, S. Richardson, and D. J. Spiegalhalter.***Markov Chain Monte Carlo in Practice,*** Chapman & Hall, London, 1996.

[17] L. R. Beard and A. J. Fredrich.**Hyrologic frequency analysis,** *Hydrologic Engineering Methods for Water Resources Development*, 1975

[18] P. Geladi, and B. R. Kowalski. **Partial least-squares regression: a tutorial**, *Analytica Chimica Acta*,vol. 185, pp. 1–17, November 2001.

[19] C. Wang, N. B. Chang and G. T. Yeh. **Copula-based flood frequency (COFF) analysis at the confluences of river systems,** *Hydrological Processes,*vol. 23, no.10, pp. 1471-1486, March 2009.

[20] P. Wilkes, S. Jones, L. Suarez, A. Mellor,W. Woodgate, M. Soto-Berelov, andA. Skidmore. **Mapping Forest Canopy Height Across Large Areas by Upscaling ALS Estimates with Freely Available Satellite Data,***Remote Sensing*, vol. 7, no. 9, pp. 12563–12587, September 2015.

[21] P. Li, E. A. Stuart, andD. B. Allison.**Multiple Imputation: A Flexible Tool For Handling Missing Data,***JAMA Guide to Statistics and Methods*, vol. 314, no. 18, pp. 1966, November 2015.
https://doi.org/10.1001/jama.2015.15281

[22] W. Y. Lai, andK. K. Kuok.**A Study on Bayesian Principal Component Analysis for Addressing Missing Rainfall Data,** *Water Resources Management*, vol. 33, no.8, pp. 2615-2628, June 2019.

[23] D. J. Stekhoven andP. Buhlmann. **MissForest--non-parametric missing value imputation for mixed-type data,** *Bioinformatics*, 28, no.1, 112-118, January 2012.

[24] H. Lee and K. Kang. **Interpolation of Missing Precipitation Data Using Kernel Estimations for Hydrologic Modeling,***Advances in Meteorology*, vol. 2015, pp. 1-12, October 2015.

[25] L. Knorr-Held.**Analysis of Incomplete Multivariate Data**, **J. L. Schafer, Chapman & Hall, London, 1997,** *Statistics in Medicine.* vol.19, no.7, 1006–1008, March 2000.

[26] R. J. A. Little and D. B. Rubin. ***Statistical analysis with missing data 2nd edition,***Wiley, New York: pp. 4 – 22, 2019.

[27] NorhazlinaSuhaimi, Nurul Adyani Ghazali, Muhammad Yazid Nasir, Muhammad IzwanZariq Mokhtar, and Nor Azam Ramli. **Markov Chain Monte Carlo Method for Handling Missing Data in Air Quality Datasets**, *Malaysian Journal of Analytical Sciences*, vol. 21, no.3, pp. 552-559, April 2017.

[28] B. Grung, and R. Manne. **Missing values in principal component analysis**, *Chemometrics and Intelligent Laboratory Systems*, vol. 42, no. 1-2, pp. 125–139, August 1998.

[29] K. A. Severson, M. C. Molaro, andR. D.Braatz. **Principal Component Analysis of Process Datasets with Missing Values,** *Processes*, vol. 5, no.4, pp. 38, July 2017.

[30] G. Ibarra-Berastegi, J. Saénz, A. Ezcurra, A. Elías, J. Diaz Argandoña, and I. Errasti.**Downscaling of surface moisture flux and precipitation in the Ebro Valley (Spain) using analogues and analogues followed by random forests and multiple linear regression,***Hydrologyand Earth System Science*, vol. 15, pp. 1895–1907, June 2011.

[31] J. H. Yang, C. H. Cheng, and C. P. Chan. **A Time-Series Water Level Forecasting Model Based on Imputation and Variable Selection Method,** *Computational Intelligence and Neuroscience*, pp. 1–11, November 2017.

[32] J. F.Kiviet. **On bias, inconsistency, and efficiency of various estimators in dynamic panel data models,***Journal of econometrics.* vol. 68, no. 1, pp. 53-78, July 1995.

[33] B. Efron, and R. J. Tibshirani. **An Intoduction to the Bootstrap,** New York, London: Chapman & Hall, 1993.

[34] B. Efron. **Bootstrap methods: another look at the jacknife,***The Annals of Statistics,*vol. 7, no. 1, pp. 1-26, December 1979.

[35] T. Chai and R. R. Draxler, **Root Mean Square Error (RMSE) or Mean Absolute Error (MAE),***Geoscientific Model Development*, vol. 7, no. 1, pp. 1247-1250, June 2014.

[36] N. H. Savage, P. Agnew, L. S. Davis, C. Ordóñez, R. Thorpe, C. E. Johnson, and M. Dalvi. **Air quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial evaluation,** *Geoscientific Model Development*,vol. 6 no. 2, pp. 353–372, March 2013.

[37] Wan Norliyana Wan Ismail and Wan Zawiah Wan Zin@Wan Ibrahim. **Estimation of rainfall and stream flow missing data for Terengganu, Malaysia by using interpolation technique methods**, *Malaysian Journal of Fundamental and Applied Sciences,*vol. 13, no. 3, pp. 213-217, August 2017.

[38] H. Aguilera, C. Guardiola-Albert, and C. Serrano-Hidalgo. **Estimating extremely large amounts of missing precipitation data,** *Journal of Hydroinformatics,*vol. 22, no.36, pp. 578-591, May 2020.
https://doi.org/10.2166/hydro.2020.127

[39] N. S.Altman. **An introduction to kernel and nearest-neighbornonparametric regression,** *The American Statistician*. vol. 46, no. 3, pp. 175-185, February 2012.