PAPER • **OPEN ACCESS**

Comparison of Missing Rainfall Data Treatment Analysis at Kenyir Lake

To cite this article: Azreen Harina Azman *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1144** 012046

Comparison of Missing Rainfall Data Treatment Analysis at Kenyir Lake

Azreen Harina Azman[1*], Nurul Nadrah Aqilah Tukimat[1] and M A Malek[2]

[1] Faculty of Civil Engineering Technology, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Pahang Darul Makmur, Malaysia
[2] Institute of Sustainable Energy (ISE), Universiti Tenaga Nasional, Malaysia
*Corresponding author: azreenharinaazman@gmail.com

Abstract. Rainfall is one of the frequent data used in weather-related studies. Sometimes the data have missing information that needs the treatment to make sure the data can be useful, complete and reliable. There are many methods in treating missing data suggested by previous studies. The best selected method to estimate missing rainfall data in different regions may vary depending on the rainfall pattern and spatial distribution. Therefore, this paper discussed and compared 3 different methods in missing data treatment. The selected methods are Expectation Maximization (EM), Inverse Distance Weighted (IDW) and Multiple Imputation (MI). After analysis, the best method is IDW based on root mean square error (RMSE), mean absolute error (MAE), correlation coefficient (r) and percentage of error (% of error) values. The IDW method has RMSE, MAE values and the lowest % of error values. In addition, the r value of IDW method is highest compared to EM and MI method. MI method recorded the highest values of RMSE, MAE and % of error with the lowest r value that proved MI method is the least accurate method to use in missing data treatment. After all methods were implemented, it proved that the IDW method is the best way to treat missing data because the analysis shows monthly rainfall distribution for 4 treatment stations in line to 3 missing data stations compared to EM and MI methods.

1. Introduction

The observational datasets are very useful to estimate the return periods of extreme events. Rainfall is one of the components in the hydrological cycle that is frequently used in weather-related studies. Rain gauges are provided in several places to record and monitor rainfall data but some problems may arise and contribute to loss of rainfall data such as gauge damage, human error, extreme weather and measurement errors. The lacks of information in periodic climate data limit its use [1].

All missing data must be treated first before running any test to ensure that data is complete, homogeneous and reliable. Various methods can be used to treat missing data but the selection is based on suitability and accuracy. According to [2] and [3], the critical issue and most important stage in meteorological data analysis are filling the gaps in daily weather data before the data can be used in further analysis. Analysis of precipitation is complicated because it deals with space and time. The accurate estimation in missing data analysis is a difficult task when dealing with long time series and rain gauge distribution [4]. This situation can be critical when involving a huge amount of data records with low quality.

There are many methods in treating missing data suggested by previous researchers. Some of the first approaches are to remove records and to replace it with mean or mode. Removing records with lost value as the rest of the records leads to skew in data. The best selected method to estimate missing rainfall data in different regions may vary depending on the rainfall pattern and spatial distribution. Treating missing data is very important for high-risk areas such as dams and areas that are exposed to extreme weather. Rainfall data is very important to identify rainfall distribution especially heavy rainfall event to prevent floods, spillovers and the worst is the dam structures failure.

2. Study Area

Terengganu experiences northeast monsoon between November to January every year and received heavy rainfall during this period. Kenyir Lake is located in the district of Hulu Terengganu in Terengganu state. The reservoir catchment area is 2600km^2. Figure 1 shows the study area. Thirty years (1988–2017) daily rainfall records from 7 stations used in Statistical Downscaling Model (SDSM) but only 3 of the stations have the complete data while other 4 stations got the missing values. All 7 stations scattered in Kenyir Lake area observed and measured by Electricity Power Provider.

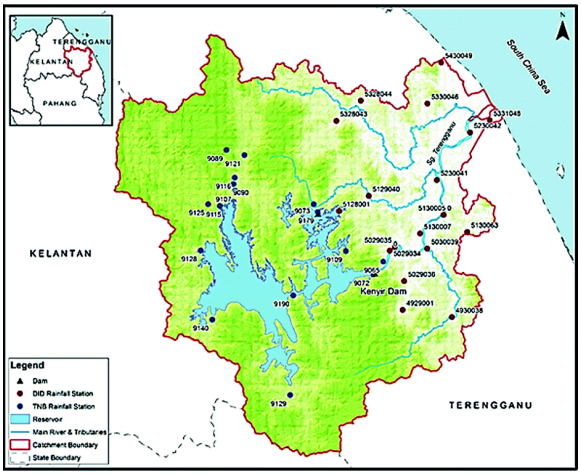


Figure 1. Kenyir Lake.

3. Methodology

The flow chart of the methodology is shown in Figure 2. To run SDSM, all the rainfall stations must have full data but 4 out of 7 stations got missing values. This study compared 3 selected methods which are EM, IDW and MI. One control station is picked to test the performance of 3 selected methods. Ten full data nearby stations used to run all the methods. Every method has its own features and speciality. The values of statistical tests determine the accuracy of each method in treating the missing data at the study area. The selected statistical tests are RMSE, MAE, r and % of error. The best method from the results of statistical tests will be applied in treating missing data in 4 stations mentioned earlier. Then, the data distribution after missing treatment will be compared with 3 full data stations to make sure that the results of the treated stations are in line with the observed data.

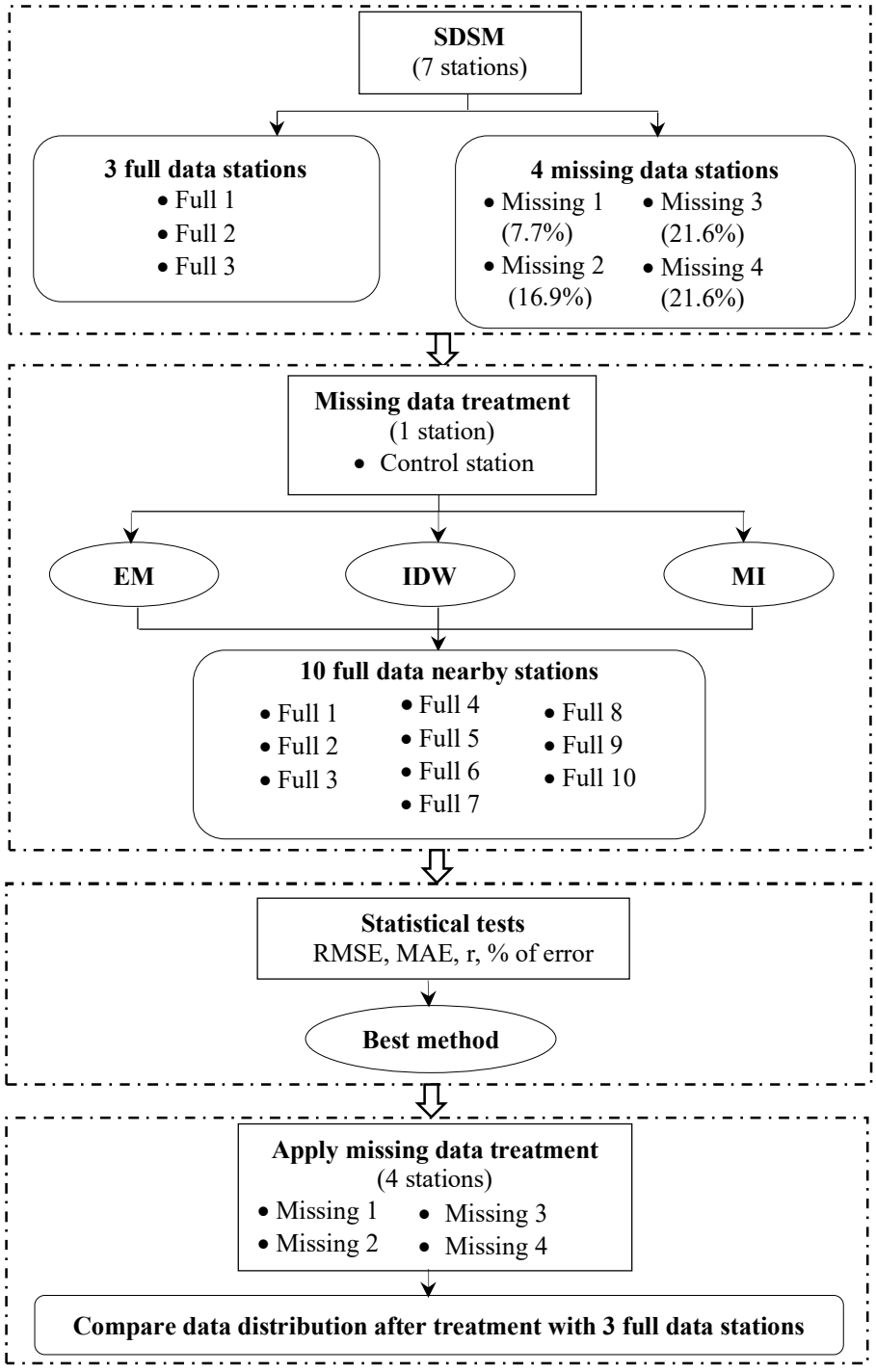


Figure 2. Flow chart of the methodology.

3.1. Expectation Maximization (EM)

EM is a suitable technique which frequently applied for analyzing data in handling lost data because of huge amount of datasets. In EM algorithm, the amount of the most probable variable value is depending on other variables [5]. EM involves missing values observation in E-step created in multiple entries while the regression model is built in M-step. The method provides an unbiased estimation of the missing values [6] and the formula as in equation (1).

$$F = \frac{\sum_{k=1}^{N}(x_{k+1} - Fx_k)}{\sum_{k=1}^{N} x_k^2} \quad (1)$$

where x_k and x_{k+1} are scalar estimates, N is scalar measurements.

3.2. Inverse Distance Weighted (IDW)

IDW is the missing data treatment method that is related to the nearest station where the smoothing of the rainfall distribution is reducing as distance increasing [7]. IDW is among the simplest method available and most commonly used. The assumption is depending on the values of closest stations (typically 10 to 30). Equation (2) is the formula for IDW method which uses the observed values at nearest stations in estimating the missing values.

$$V_o = \frac{\sum_{i=1}^{n}(V_i/D_i)}{\sum_{i=1}^{n}(1/D_i)} \quad (2)$$

where V_o is the assessed value of the missing data, V_i is the value of same parameter at i^{th} nearest station, D_i is the distance between the station with missing data and the i^{th} nearest station [8].

3.3. Multiple Imputation (MI)

MI is a robust method that measures the uncertainty associated with the estimation [6]. It is a filling method that affords valid statistical interpretations. This method using the standard procedures of regression, then combine imputing results to obtain final result [9]. The missing data at the target station is the average value of imputed data [10]. According to previous researchers, 3 to 5 imputed data sets are sufficient [8] and estimated using equation (3).

$$P_x = \frac{\sum_{i=1}^{k} I_i(P_x, P_i)}{k} \quad (3)$$

where P_x is missing data and P_i is imputations data.

3.4. Statistical Analysis

The statistical tests used to compare the effectiveness the performance of the selected methods are RMSE, MAE, r and % of error. r, RMSE and MAE are the most commonly used and accepted in many statistical analysis [11-12]. The equations are stated as follow:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}} \quad (4)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \quad (5)$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|x_i - y_i|}{n} \quad (6)$$

$$\%\ \text{of error} = \frac{|y - x|}{x} \text{ x } 100 \quad (7)$$

where x is the observed value, y represents the computed value and n denotes the number of data observations.

4. Results and Analysis

Table 1 shows the list of rainfall stations in the study where 3 of the stations are full while another 4 have missing data. The minimum missing days is 28 days (7.7%) for Missing 1 station and maximum missing days are 79 days (21.6%) for Missing 3 and Missing 4 stations.

Table 1. List of rainfall stations in the study.

Station Name	Code	% of missing
Full 1	F1	0
Full 2	F2	0
Full 3	F3	0
Missing 1	M1	7.7 (28 days)
Missing 2	M2	16.9 (62 days)
Missing 3	M3	21.6 (79 days)
Missing 4	M4	21.6 (79 days)

Ten selected rainfall stations used in those 3 missing data treatment methods are listed in Table 2 and shown in Figure 3. The values of RMSE, MAE, r and % of error are used to compare 3 selected methods and determine the best method to treat the missing data. All the statistical values are listed in Table 3. The data were analyzed according to the number of missing days which are 28, 62 and 79 days.

Table 2. Locations of the selected rainfall stations used in missing data treatment methods.

Station Name	Code	Latitude	Longitude
Full 1	F1	-	-
Full 2	F2	-	-
Full 3	F3	-	-
Full 4	F4	5.057	102.932
Full 5	F5	4.968	102.970
Full 6	F6	5.143	102.844
Full 7	F7	4.940	103.057
Full 8	F8	4.636	102.953
Full 9	F9	4.835	103.194
Full 10	F10	5.306	102.856

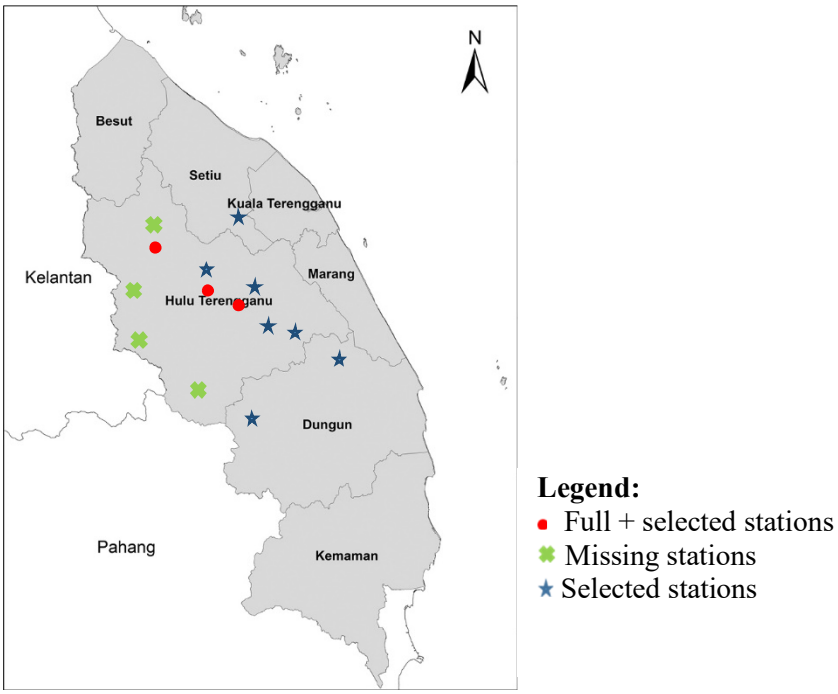


Figure 3. Locations of all rainfall stations in the study.

Table 3. Statistical values of every method.

Method	Missing days	RMSE	MAE	r	% of error
EM	28	6.46	4.52	0.60	50.14
	62	15.85	8.30	0.00	8.04
	79	5.38	2.98	0.63	94.64
IDW	28	6.56	3.48	0.53	17.41
	62	10.01	4.61	0.74	48.47
	79	4.25	1.43	0.73	3.33
MI	28	11.66	9.16	0.25	92.52
	62	17.05	11.81	0.12	64.67
	79	14.13	10.56	0.31	147.67

Based on Figure 4, the IDW method has the lowest RMSE, MAE and % of error values. IDW method has the smallest RMSE and MAE values compared to EM and MI methods. The smaller the RMSE and MAE, the more accurate the formula are [13]. The lower % of error between observed data on the ground and predicted data using missing treatment methods proved that IDW is the most suitable method in treating the missing data. The graph clearly shows that the r-value of IDW method is the highest compared to EM and MI methods. The larger value of r shows the higher correlation between the data set of observed data and predicted data after the missing data treatment. The results for 79 missing days are the best to prove that IDW is the best method among others. The RMSE value is 4.25 and MAE is 1.43 which are the lowest compared to EM and MI methods. Meanwhile, the r value is 0.73, among the highest value. In terms of % of error, 3.33% is the lowest and best value. MI method recorded the highest values of RMSE, MAE and % of error with the lowest r value that proved MI method is the least accurate method to use in missing data treatment.

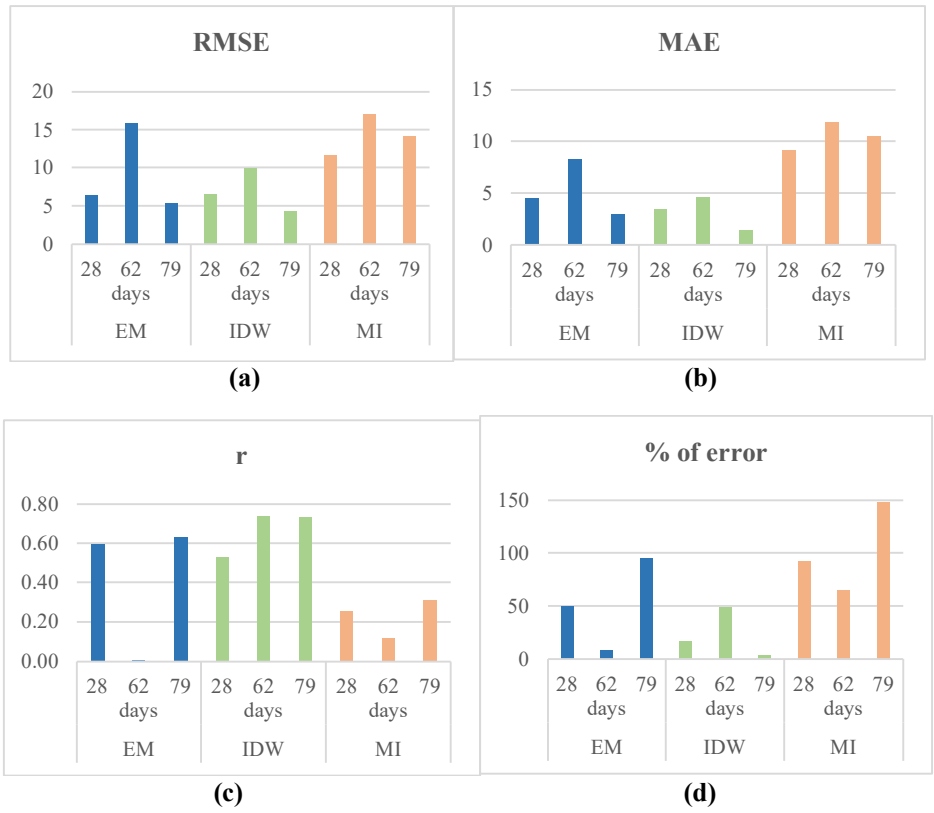


Figure 4. Graph of statistical values of every method used in missing data treatment.

According to the values of RMSE, MAE, r and % of error, the IDW method is the best method to treat missing data in this study. The comparison of average values between observed data and after treatment methods shown in Figure 5. The mean values after the IDW method for 28 and 79 missing days are very close to observed ground data. There is a gap between the mean value of ground data and after the IDW method for 62 missing days but still acceptable when compared to EM and MI method.

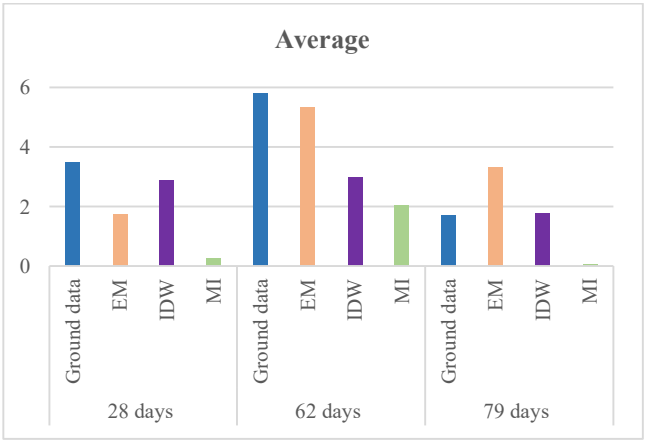
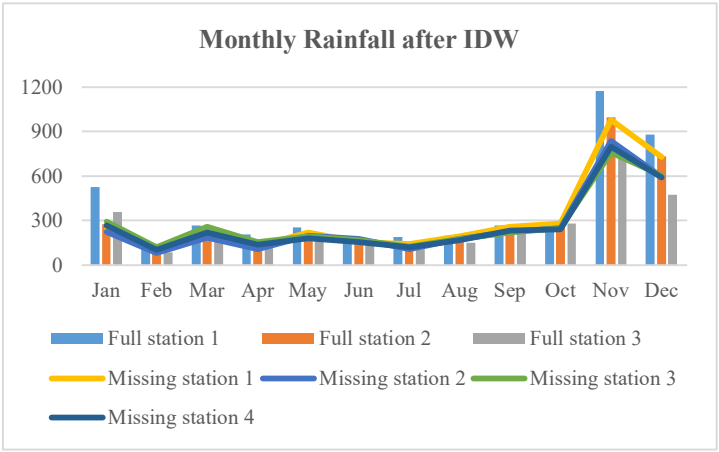


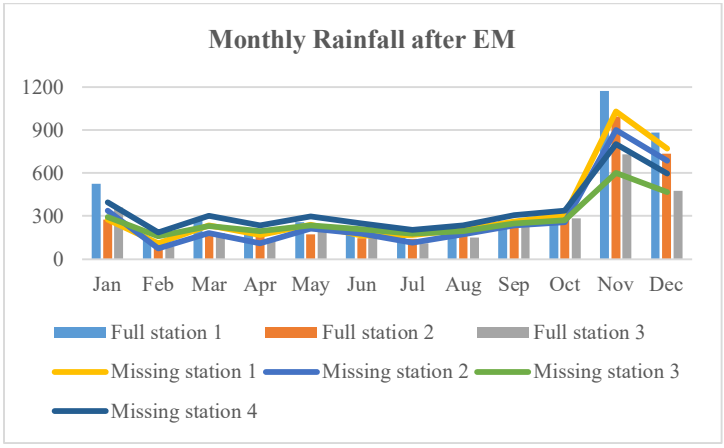
Figure 5. Comparison of average values between observed data and after treatment methods.

IDW method was used to treat the missing data for 4 stations which are Missing 1, Missing 2, Missing 3 and Missing 4. Figure 6(a) clearly shows the monthly rainfall for all 4 treated stations are identical with another 3 full data stations. The graph proved that IDW method is the best method to treat missing data because the monthly distribution of rainfall for 4 treated stations are in line with 3 non-missing data stations compared to EM and MI methods. Figure 6(b) and (c) shows that values of treated data using EM and MI methods have a gap between months especially in November and December.

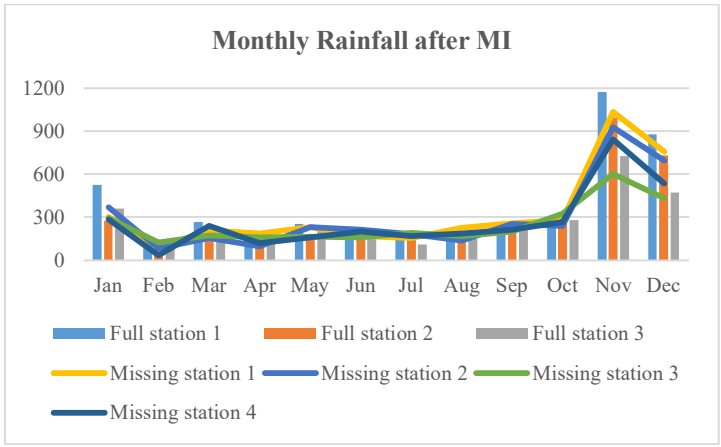


(a)

Figure 6. The monthly rainfall after IDW method for four missing stations.



(b)



(c)

Figure 6. The monthly rainfall after IDW method for four missing stations(*Cont…)*

5. Conclusions

Three selected methods, EM, IDW and MI were used to treat four missing daily rainfall stations in Kenyir Lake. The minimum missing days is 28 days (7.7%) while the maximum missing days is 79 days (21.6%). After analysis, the best method is IDW based on RMSE, MAE, r values and percentage error values. The IDW method has RMSE, MAE values and the lowest percentage error values. Besides, the r value of IDW method is highest compared to EM and MI method. After the IDW method was implemented, it proved that the IDW method is the best way to treat missing data because the analysis shows monthly rainfall distribution for 4 treatment stations in line to 3 missing data stations compared to EM and MI methods. The results are in line with the statement of the DID staff stating that the method used by the DID to treat missing data is IDW. However, the DID does not treat missing data to maintain the reliability of existing data.

6. References

[1] Barrios A, Trincado, G and Garreaud R 2018 Alternative approaches for estimating missing climate data: application to monthly precipitation records in South-Central Chile. *For. Ecosyst* **5** 28 1-10

[2] Fenta Mekonnen, D and Disse, M 2018 Analyzing the future climate change of Upper Blue Nile River basin using statistical downscaling techniques. *Hydrol. Earth Syst. Sci.* **22** 2391–2408

[3] Hassan M M and Croke B F W 2013 Filling gaps in daily rainfall data: a statistical approach. *Proc. 20th International Congress on Modelling and Simulation* (Adelaide)

[4] Simolo C, Brunetti M, Maugeri M and Nanni T 2010 Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach. *Int. J. Climatol* **30** 1564–1576
[5] Khalifeloo M H, Mohammad M and Heydari M 2015 Application of different statistical methods to recover missing rainfall data in the Klang River catchment. *Int. J. Innov. Sci. Math.* **3** 2347–9051
[6] Presti R L, Barca E and Passarella G 2010 A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environ. Monit. Assess.* **160** 1–22
[7] Kamaruzaman I F, Wan Zin W Z and Mohd Ariff N 2017 A comparison of method for treating missing daily rainfall data in Peninsular Malaysia *Mal. J. Fund. Appl. Sci.* **2017** 375-380
[8] Sattari M, Rezazedah-Joudi A and Kusiak A 2016 Assessment of different methods for estimation of missing data in precipitation studies. *Hydrol. Res.* **48** 1032-1044.
[9] Aieb A, Madani K, Scarpa M, Bonaccorso B and Lefsih K 2019 A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria. *Heliyon* **5** e01247.
[10] Radi, N, Zakaria R and Azman M 2015 Estimation of missing rainfall data using spatial interpolation and imputation methods *AIP Conference Proceedings* **1643** 42
[11] Amirabadizadeh M, Ghazali A H, Huang Y F and Wayayok A 2016 Downscaling daily precipitation and temperatures over the Langat River Basin in Malaysia: a comparison of two statistical downscaling approaches *Int. J. Water Res. Environ. Eng.* **8** 120–136
[12] Bennett N D, Croke B. F, Guariso G, Guillaume J H, Hamilton S H, Jakeman A J, Marsili-Libelli S, Newham L T, Norton J P and Perrin C 2013 Characterizing performance of environmental models *Environ. Model. Softw.* **40** 1–20
[13] Nguyen, L and Ho T H T 2018 Fetal weight estimation in case of missing data *International Technology and Science Press* **1** 45-65

Acknowledgement
This research is supported by Universiti Malaysia Pahang, Ministry of Higher Education (MHE) RDU1901141 (Ref: FRGS/1/2019/TK01/UMP/02/1), Malaysian Meteorological Department (MMD), and Drainage and Irrigation Department (DID).