International Journal of Engineering & Technology, 7 (4.18) (2018) 204-212

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

The Improvement of Missing Rainfall Data Estimation During Rainy Season at Ampang Station

Nurul Aishah Rahman, Sayang Mohd Deni*, Norazan Mohamed Ramli, Norshahida Shaadan

Centre of Statistical and Decision Science Studies, Advanced Analytics Engineering Centre (AAEC),
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, 40450, Selangor, Malaysia
**Corresponding author E-mail: sayang@tmsk.uitm.edu.my*

Abstract

The availability of rainfall data plays a significant role in water related sectors. The presence of missing values could produce biasness in the results of data analysis. Several methods have been used to estimate the missing values such as simple arithmetic average, normal ratio method, inverse distance weighting method, correlation coefficient weighting method and geographical coordinate. However, the estimated values produced by the imputation methods had been used scarcely considered rainfall pattern during the estimation process. To fill the gap, the generalized linear model (GLM) was used to assess the performance of the imputation methods at a target station namely Ampang station. The experimentation was conducted using real data set for the period from 1975 to 2014. Neighbouring rainfall stations with 25km and 35km away from the target station and the duration during rainy and non-rainy period were considered in assessing the capability of the model. This study aims to assess the performance of GLM methods in comparison with the current existing techniques in estimating the missing values. Based on the mean absolute error and root mean squared error, the results have shown that the application of GLM able to produce better and accurate rainfall data estimation.

Keywords: *Daily rainfall data; Generalized linear model; Imputation method; Missing data.*

1. Introduction

Analyzing rainfall data, such as identifying the rainfall characteristics and pattern, is very crucial and highly in demand. The outcomes of the research will be used to be applied in various purposes including flood monitoring, data generation and prediction of extreme weather events such as landslides. However, the rainfall data are often encountering with the missing problem due to several factors such as malfunction of instruments[5, 27, 28], power failure in recording the rainfall[1, 3, 8] and absence of observer[4, 20, 22]. Most of the researchers have concluded that these factors could interrupt the continuity and the consistency of the rainfall data. In addition, the presence of missing values could lead to biased results and may cause obstacles in analyzing the findings. Hence, estimating the missing rainfall data is important in order to obtain more reliable results.

Various imputation methods have been proposed and developed by some of the previous researchers to estimate the missing values in daily rainfall data. The imputation method is a process of replacing the missing values with the reasonable values. However, tendency of having errors during the estimation process cannot be avoided, for example, in the case of ignoring the differences between complete and incomplete cases and loss of sample size. Hence, it may distort the distribution of the variable when dealing with large proportion of missing values, underestimate the variability of the data and consequently, the efficiency of the methods is questionable.

The estimation of the missing values in daily rainfall data have been discussed in the literature. Although modern and sophisticated methods were widely used in the rainfall estimation, these methods have a complex mathematical formulation, required intensive calculations and computational cost which results in difficulty to be implemented. Therefore, it was decided that a simpler imputation method be used rather than the sophisticated methods. The simple arithmetic average (SAA), normal ratio method (NRM), inverse distance weighting method (IDW), correlation coefficient weighting method (CCW) and geographical coordinate (GC) are the five imputation methods that will be used to estimate the missing rainfall data in this study.

It is known that SAA can be categorized as one of the simplest imputation method where the missing data were imputed by averaging the data from neighbouring stations[14, 30]. Meanwhile, the NRM and IDW were also known as the most popular methods in estimating the missing values in daily rainfall data. In the estimation of the missing values using NRM, the weighting factor was obtained by taking into consideration the amount of rainfall from the neighbouring sites will be used[15, 16, 25]. Alternatively, IDW was also known as one of the robust method in the estimation of missing values as this method will take into consideration the distance between the neighboring stations and the target stations[2, 6, 9].

In addition, the other two methods namely CCW and GC, were proposed as the alternative of some existing imputations methods. The strength of the correlation between the neighboring and the target station will be measured and consider as the weighting factor for CCW method. The researchers have recommended CCW as one of the best method when imputing missing rainfall data due to its superiority in producing the estimation results[25, 26]. Meanwhile, GC resembled IDW where it also considers a more significant role of the neighbouring stations in estimating the missing data at target station[15, 21]. This method utilizes the longitude and latitude between the target and the neighbouring stations in estimating the missing data instead of using distance as weighting factor.

Although these methods were found to be the most efficient when estimating the missing rainfall data, the estimated values produced by these methods usually do not consider the behavior and the monsoon season of daily rainfall in the estimation of the missing values. Normally, the pattern of Malaysian daily rainfall influenced by the monsoon seasons which could be considered as one of the important factors in helping to provide a more precise and accurate estimation. Moreover, analyzing the pattern of long records of daily rainfall amount will also provide solution, forecasting and monitoring for most of the disastrous events such as floods and landslides[31-35]. Thus, this study will consider the pattern of daily rainfall when estimating missing data using generalized linear model (GLM).

GLM provides a more realistic representation of day to day variability with non-negative and non-Gaussian approach[12]. This model allows the application of regression analysis when the data originates from an exponential family distribution rather than a Gaussian. Hence, GLM were used to model the rainfall data with gamma distributions and Fourier series as the link function and smoothing technique, respectively[10, 24]. This model is often chosen since it has capability in fitting both unimodal and bimodal seasonal pattern[11]. Generally, rainfall in Malaysia has seasonal variation where the parameter of the rainfall amount and occurrence of the rain changing throughout the year. Rainfall amount and occurrence are the two types of stochastic model of rainfall. The first type is a model of rainfall amount which simulates on rainy days only. Meanwhile, the second type is a model of rainfall occurrence that simulates a sequence on rainy and dry days[7, 23]. In this study, modelling the rainfall amount on the rainy days will be considered and used to estimate the missing values.

In the estimation procedure, separate parameters of the rainfall amount were derived for each month of the year to handle the seasonal variation. However, it will cause a large number of parameters to be estimated. Thus, Fourier series were used to smooth the model parameters, as this can best describe the rainfall pattern and its temporal variation. In addition, gamma distribution is known as a proper distribution in representing the rainfall data with the sense that is no negative value of rainfall. Moreover, the seasonal fluctuations of parameters in gamma distribution can be described using the Fourier series.

It could be noticed that very few studies have been conducted on imputation of missing values by considering the pattern and model of daily rainfall amount. Thus, the main purpose of this study is to estimate the missing values in daily rainfall at Jabatan Pengairan Saliran Ampang (JPS Ampang), by considering the pattern and model of rainfall amount with the application of GLM. Five imputation methods including SAA, NRM, IDW, CCW and GC will be used to evaluate the capability of GLM in producing more reliable estimation of the missing values in daily rainfall amount at the target stations. In addition, two different distances of the neighboring stations which situated within 25km and 35km from the target station will be considered. The estimation procedure will also take into consideration the duration during rainy and non-rainy period at different levels of missingness (i.e. 5%, 10%, 15% and 20%). Missingness is defined as the data that are missing from a rainfall series. For example, if 5% of rainfall data is missing, then the remaining portions of 95% will be used to calculate the correlation or average between the target station and its neighbouring.

In order to pursue the main purpose of this study, the contents of this paper were organized as follows. A brief introduction on the missing rainfall estimation and the objectives of this study were presented first. Then, a brief description of the target station, JPS Ampang and its neighbouring stations that were involved in this study was provided. The theoretical of the imputation methods and the rainfall modelling, GLM were described in the next section, followed by the performance criteria. Finally, the analysis and results, general remark, conclusions and recommendations were presented.

2. Material and methods

2.1. Study area and data description

Peninsular Malaysia lies entirely in the equatorial zone which is situated in the northern latitude between 1^{o} and 6^{o} N and the eastern longitude from 100^{o} to 103^{o} E. There are two types of monsoons that influence the climate of the country, namely, the Southwest monsoon (May to August) and the Northeast monsoon (November to February). The data used in this study were collected from the database of the Drainage and Irrigation Department (DID), for the 40 years of records that ranged from 1975 to 2014. However, some of the data not in a good quality due to the missing values, the stations are closed and lack of lengthy records. Generally, problems with missing data had forced the researchers to limit the selection of the rainfall station in their study. This problem frequently arises in real world application and can be critical since it could affect the study analysis.

Five rainfall stations in Selangor region as shown in Table 1 were selected for the illustration purposes. These stations were chosen due to the rapid development as well as urbanization in the areas. JPS Ampang station was selected as the target station due to the importance of having completeness in daily rainfall data at the area. Meanwhile, the data from the neighbouring stations are also considered in estimating the missing data. The specific locations of the target station and its corresponding neighbouring are also displayed in Fig. 1.

Table 1: General Description of the Target Station (*) and Neighbouring Stations in Selangor

Station	Station Name	Coordinate		Distance (km)
		Longitude	Latitude	
1	**JPS Ampang (*)**	**101.75**	**3.16**	**0**
2	Sek. Keb. Kg. Lui	101.87	3.17	13.38
3	Sawah Sg. Lui	101.91	3.17	17.82
4	Lalang Sg. Lui	101.91	3.14	17.92
5	Bandar Tasik Kesuma	101.87	3.00	31.87

A bimodal pattern can be observed for the rainfall in Selangor area. The likely cause of the form the rainfall patterns take is the monsoonal flow that contributes to the heavy rainfall in Selangor region at different times of the year. Becoming one of the city that have rapid development of industry as well as the transportation, JPS Ampang is the most suitable station to be chosen as the target one. Furthermore, this target station also affected by serious flash floods which happened may be due to insufficient drainage infrastructure. However, prolonged and frequent rainfall could be considered as one of the most contributing factors to flash floods during heavy rain. Due to this problem, the government had allocated RM10.5 million for flood mitigation work, upgrading the drainage systems and retention ponds. Thus, a forecast on the rainfall using complete data could help in reducing the losses and provide warnings to people.

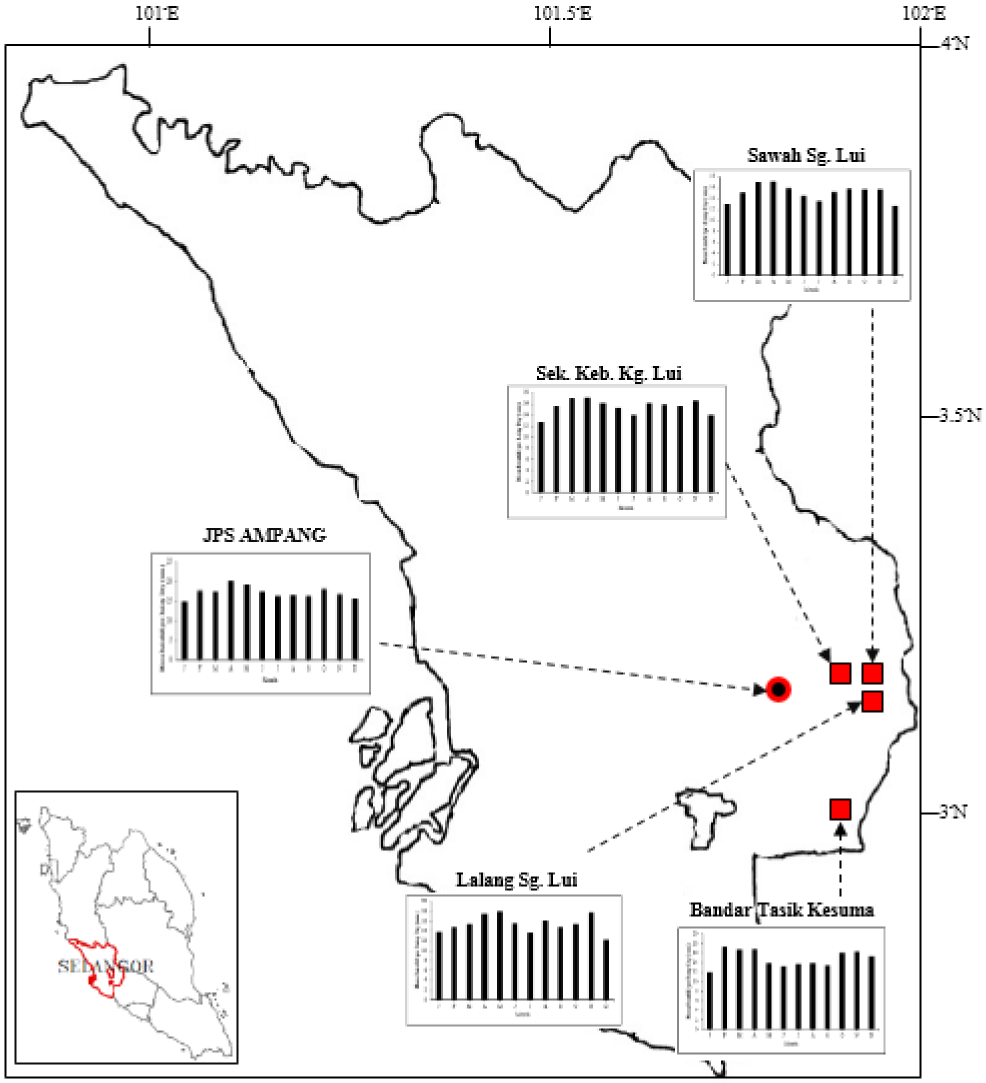

Fig. 1: Location of the Rainfall Stations in Selangor

2.2. Imputation method for estimating missing data

In this study, a target station, T and several neighbouring stations, N were selected for the analysis purposes. The target station was selected based on the importance of complete rainfall data towards that area. However, several problems were detected when selecting the neighbouring stations such as lengthy of record on missing data as well as the distance between the neighbouring and target station. Many researchers in the literature agreed that the use of three to four closest rainfall stations as the neighbouring stations were enough when estimating the missing data. Therefore, the neighbouring stations that were located within radius of 25km and 35km from the target station were selected in this study.

The missing data in the neighbouring stations were identified and imputed. Using the average value between available data from another neighbouring stations, the missing values were filled. Only then, four different levels of missingness, 5%, 10%, 15% and 20% were artificially created randomly in the target station based on complete data. In addition, rainy and non-rainy period that occurred in the target station were also considered in this study. Five imputation methods, SAA, NRM, IDW, CCW and GC were used in estimating the missing rainfall data. The difference between the estimated values and the actual data will then be quantified using two error indices, mean absolute error (MAE) and root mean squared error (RMSE).

However, previously, the estimated values obtained by the imputation methods have not considered the rainfall pattern while estimating the missing data. To handle this situation, a generalized linear model (GLM) was fitted to the estimated values of each method. Based on the analysis of deviance that is produced after the fitting process, the best harmonics is determined based on the reduction of deviance and the p-value. The missing values were estimated using the fitted values of the chosen harmonics. Finally, the performance of the revised estimated values was compared with the actual rainfall data in order to verify whether GLM could produce better results. This study also carried out a preliminary analysis where GLM was fitted to the actual rainfall of the target station before applied to the estimated values in order to evaluate the effectiveness of the GLM.

2.2.1. Simple arithmetic average (SAA)

Using this method, the missing rainfall data in target station will be imputed by the average of rainfall amount from the neighbouring stations. SAA method is also known as mean imputation.

$$\hat{Y}_{m_k t} = \frac{1}{N}\left(Y_{m_k 1} + Y_{m_k 2} + \ldots + Y_{m_k N}\right) = \bar{Y}_{m_k} \qquad (1)$$

where

$m_k t$ = subscripts of missing value recorded at target.

$\hat{Y}_{m_k t}$ = missing value at target station.

$\bar{Y}_{m_k \cdot}$ = average of value of neighbouring stations, N

$Y_{m_k 1} + Y_{m_k 2} + \ldots + Y_{m_k N}$ = actual rainfall data at neighbouring stations.

2.2.2. Normal ratio method (NRM)

In NRM, the missing data will be imputed by the helps of weighting factor. The actual rainfall data at neighbouring stations is weighted by the ratios of total rainfall data in target and neighbouring stations.

$$\hat{Y}_{m_k t} = \frac{1}{N}\left[\left(\frac{T_{.t}}{T_{.1}}\right)Y_{m_k 1} + \left(\frac{T_{.t}}{T_{.2}}\right)Y_{m_k 2} + \ldots + \left(\frac{T_{.t}}{T_{.N}}\right)Y_{m_k N}\right] \tag{2}$$

where

$T_{.t}$ = total rainfall in target station.

$T_{.1}, T_{.2}, \ldots T_{.N}$ = total rainfall for each neighbouring stations.

2.2.3. Inverse distance weighting method (IDW)

IDW method is based on a concept of distance weighting. The actual rainfall data at neighbouring stations is weighted by the distance between target and neighbouring stations.

$$\hat{Y}_{m_k t} = \frac{d_{1t}^{-2}}{\sum_{i=1}^{N} d_{it}^{-2}} Y_{m_k 1} + \frac{d_{2t}^{-2}}{\sum_{i=1}^{N} d_{it}^{-2}} Y_{m_k 2} + \ldots + \frac{d_{Nt}^{-2}}{\sum_{i=1}^{N} d_{it}^{-2}} Y_{m_k N} \tag{3}$$

where

$d_{1t}^{-2}, d_{2t}^{-2}, \ldots, d_{Nt}^{-2}$ = distance between the neighbouring and target station.

2.2.4. Correlation coefficient weighting method (CCW)

For this imputation method, the role of the distance as weighting factors will be replaced by correlation coefficient between target and neighbouring stations.

$$\hat{Y}_{m_k t} = \frac{r_{1t}}{\sum_{i=1}^{N} r_{it}} Y_{m_k 1} + \frac{r_{2t}}{\sum_{i=1}^{N} r_{it}} Y_{m_k 2} + \ldots + \frac{r_{Nt}}{\sum_{i=1}^{N} r_{it}} Y_{m_k N} \tag{4}$$

where

$r_{1t}, r_{2t}, \ldots, r_{Nt}$ = correlation coefficients of rainfall data between target and neighbouring station.

2.2.5. Geographical coordinate (GC)

GC method is more likely to IDW where the weight coefficient will be determined by the geographic coordinates of the neighbouring stations.

$$\hat{Y}_{m_k t} = \frac{W_1}{\sum_{i=1}^{N} W_i} Y_{m_k 1} + \frac{W_2}{\sum_{i=1}^{N} W_i} Y_{m_k 2} + \ldots + \frac{W_N}{\sum_{i=1}^{N} W_i} Y_{m_k N}$$
$$; \mathrm{W}_i = \frac{1}{x_i^2 + y_i^2} \tag{5}$$

where

$W_1, W_2, \ldots, W_N$ = weight coefficient for each neighbouring station.

$x_1, x_2, \ldots, x_N$ = longitude for neighbouring station.

$\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{N}$ = latitude for neighbouring station.

2.3. Generalized linear model (GLM) for modelling the daily rainfall data

Generalized Linear Model (GLM) with the linear function of gamma distribution was used to model the rainfall amounts on rainy days. Gamma distributions have been chosen as the best distribution in modelling the rainfall amounts since this distribution perform slightly better in term of efficacy than another distributions [19]. The idea was to express ln(μ(t)) which can be written as ln(μ(t)) = g(t) as linear function which involve with harmonic components. The model is GLM since g(t) is linear when parameters are unknown. The independent variables are the functions of time while dependent are the parameters from gamma distributions for rainfall amounts. Using Fourier series as the periodic function is a good approach in smoothing the model parameters. Based on the previous study, many researchers have applied Fourier series as the smoothing techniques [13, 17, 29]. It is known that Fourier series also allow for the time variations when smoothing the parameters where the fitted curves are able to connect at the beginning and end of the year[7]. Based on the previous study, the periodic seasonal fluctuations in gamma distribution can be explained using Fourier series. The Fourier series is expressed as follows:

$$g(t) = A_0 + \sum_{j=1}^{m} (A_j \sin(jt') + B_j \cos(jt')) \tag{6}$$

where

j = number of harmonic.
m = maximum harmonic required for the series.
A_j, *B_j* = parameter coefficient.
t' = π (t-183) /183.

The performance of the Fourier series in describing the rainfall pattern will depend on its deviance since it can measure the goodness of fit. Several models with different number of harmonics (i.e. one harmonics up to five harmonics) were compared to find the sufficient number of harmonics required in the study. Deviance can be classified into two components, 'between-day deviance' and 'within-day deviance'. The equation for 'between-day deviance' as follow:

$$D_B = 2\sum_t n(t)\ \left[\ln\hat{u}(t)\ -\ \ln\mu(t)\right] \tag{7}$$

where *û(t)* = predicted value of *μ(t)*.

Meanwhile, the equation of 'within-day deviance' as below:

$$D_W = 2\sum_t n(t)\left[\ln\mu(t)\ -\ \overline{\ln x}(t)\right]$$
$$; \overline{\ln x}(t) = \sum_{i=1}^{n(t)} \ln x_i(t) / n(t) \tag{8}$$

The deviance will determine the number of harmonics required in modelling the estimated values obtained by the imputation methods above. The sufficient harmonics in describing the rainfall pattern will be selected when there were no further harmonics that reduce the deviance significantly. In addition, when the probability value (p-value) were less than and equal to 0.01 (significance level), then it would be the maximum number of harmonics that best fit the model. However, in this study, parsimonious concept was applied where simplest model that can explained the rainfall pattern with as few parameters as possible was selected.

2.4. Performance measures criteria

The performances of the imputation methods were compared and evaluated using mean absolute error (MAE) and root mean squared error (RMSE) to identify whether it could provide more accurate estimation when considering GLM in the estimation of the missingness.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{N}\left|Y_{a_k t} - \hat{Y}_{m_k t}\right| \qquad (9)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{N}\left(Y_{a_k t} - \hat{Y}_{m_k t}\right)^2}{n}} \qquad (10)$$

where

$Y_{a_k t}$ = actual rainfall data.

$\hat{Y}_{m_k t}$ = estimated value.

$\overline{Y}_{.a}$ = average of actual rainfall data.

N = total number of neighbouring stations.

n = total number of observations.

3. Results and discussion

3.1. Descriptive statistics of daily and annual rainfall data at Selangor

The descriptive statistics of daily rainfall data at Selangor shows that the highest total rainfall amount of 101761.4 mm and the corresponding total rainy day of 5889 occurred at the target station, JPS Ampang. The maximum mean rainfall per rainy day and the highest rainfall standard deviation also recorded within this station with an amount of 17.3 mm and 14.5 mm, respectively. However, the maximum rainfall amount recorded at Sek. Keb. Kg. Lui station with an amount of 207.8 mm during the forty years. In addition, Lalang Sg. Lui station received lowest amount of rainfall compared to other stations with a value of 21547.3 mm.

Table 2: Descriptive Statistics for Daily and Annual Rainfall Data at Selangor

Station Name / Parameter (mm) (Daily)	**(i)**	**(ii)**	**(iii)**	**(iv)**	**(v)**	**(vi)**	**(vii)**
JPS Ampang	**7.0**	**14.5**	1.00	144.9	**101761.4**	**5889**	**17.3**
Sek. Keb. Kg. Lui	5.9	12.9	0.32	**207.8**	85806.8	5500	15.6
Sawah Sg. Lui	3.1	9.6	0.25	139.6	45915.7	3035	15.1
Lalang Sg. Lui	1.5	6.8	0.18	109.0	21547.3	1380	15.6
Bandar Tasik Kesuma	4.8	11.6	0.26	146.0	70539.1	4668	15.1

Station Name / Parameter (mm) (Annual)	**(i)**	**(ii)**	**(iii)**	**(iv)**	**(vii)**
JPS Ampang	**278.0**	5.6	1.00	**748.5**	**6306.7**
Sek. Keb. Kg. Lui	234.4	4.6	0.21	637.3	5639.5
Sawah Sg. Lui	125.5	5.7	0.11	458.8	5472.7
Lalang Sg. Lui	58.9	9.8	0.18	259.8	5552.8
Bandar Tasik Kesuma	192.7	4.7	0.17	546.9	5402.5

***(i)** = mean
(ii) = standard deviation
(iii) = correlation
(iv) = maximum rainfall
(v) = total rainfall
(vi) = total rainy day
(vii) = mean rainfall per rainy day

As illustrated on the descriptive statistics for annual rainfall, the station with the maximum amount of rainfall at Selangor was the target station, JPS Ampang which recorded an amount of 748.5 mm with a corresponding highest mean value of 278.04 mm, followed by Sek. Keb. Kg. Lui station. It also indicates that the highest mean rainfall per rainy day of 6306.7 mm occurred at JPS Ampang. After the process of transformation data of daily to annual rainfall, the variation of rainfall for the stations located at Selangor region were slightly reduced which is represented by the standard deviation in the table. Based on the correlation value, there was a relationship between the target and neighbouring stations, but it was not highly correlated which may be due to the location of the stations.

3.2. Preliminary analysis by fitting GLM to the daily rainfall data

The preliminary analysis in this study was performed to evaluate the performance of GLM when estimating the missing rainfall data. In performing this analysis, this model was fitted to the actual mean rainfall per rainy day of JPS Ampang. The performance of the GLM in describing the rainfall pattern was described by the analysis of deviance that were obtained after the fitting process. In this study, models with different number of harmonics (i.e. one harmonic up to five harmonics) were compared to find sufficient harmonics in modelling the mean rainfall per rainy day for this station. The model with best harmonics was determined based on the reduction in deviance as well as the probability value (p-value). Using the parameter estimates of the chosen harmonics, best fitted models were produced to estimate the missing rainfall data.

Table 3: Analysis of Deviance for Modelling Actual Daily Rainfall Data at JPS Ampang

Source	**Degree of freedom**	**JPS Ampang**	
		Deviance	**P-value**
Between Day	365	624.30	
1 Harmonic	2	28.75	0.00
2 Harmonics	**2**	**19.69**	**0.00**
3 Harmonics	2	4.25	0.26
4 Harmonics	2	2.31	0.48
5 Harmonics	2	0.64	0.81
Residual	355	568.70	
Within Days	5524		
Total	5889		

Based on the analysis of deviance in Table 3, the results indicate that Fourier series with two harmonics were required to model the mean rainfall per rainy day at JPS Ampang. No further harmonics were required since the deviance reduced significantly, and the p-value was less than 0.01 when two harmonics were applied. In addition, there were 5889 rainy days in the record period of this station.

Table 4: Parameter Estimates (standard error) of the Fourier Series with Two Harmonics

A_0	**A_1**	**B_1**	**A_2**	**B_2**
2.93770	-0.06646	0.06349	-0.05339	-0.05831
(0.01575)	(0.02180)	(0.02269)	(0.02210)	(0.02239)

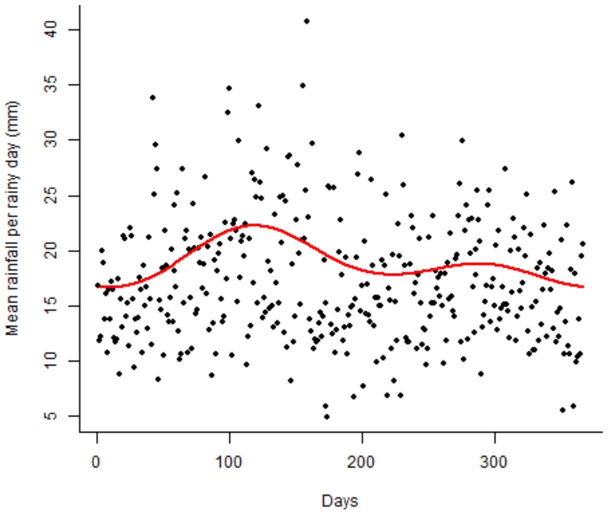

Fig. 2: Actual and Fitted Mean Rainfall per Rainy Day at JPS Ampang

The parameter estimates along with the standard error were displayed in Table 4. Meanwhile, the results in Fig. 2 support that the actual rainfall data was fitted well by the two harmonics. Then, the estimation by using the best fitted model from the chosen harmonics will be determined. Moreover, five imputation methods that have mentioned before were also considered to compare the estimation with GLM.

Fig. 3 presents a spider chart to compare graphically multiple quantitative variables. The error produced by each method is shown in terms of points on the chart. A point closer to the centre of the wheel indicates a lower value of error and vice versa. From the figure, it can be concluded that GLM outperforms the other imputation methods by producing the least MAE and RMSE values, regardless of any level of missingness and distance. After evaluating the effectiveness of the GLM in the preliminary analysis, this model will be fitted to the estimated values obtained by the five imputation methods described in the next section. The performance of the revised estimated values that were obtained after fitting process will be evaluated to see how GLM could improve the estimation process.

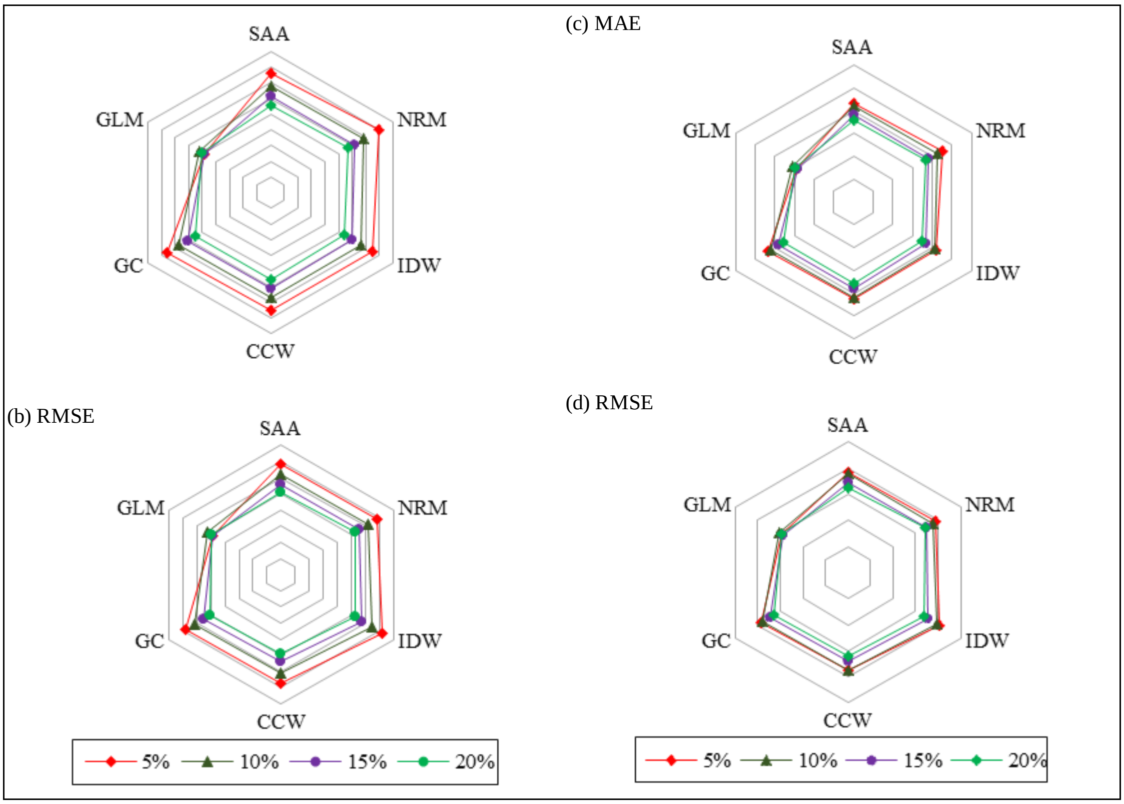

Fig. 3: The Estimation of Missingness within (a-b) 25km and (c-d) 35km of Neighbouring Stations based on MAE and RMSE

3.3. The performance of the revised estimated values in estimating the missing data using 25km and 35km of neighbouring stations

In analysing the performance of GLM in producing the best estimation results, four different levels of missingness (i.e. 5%, 10%, 15% and 20%) and two distances of neighbouring stations (i.e. 25km and 35km) were considered. In this study, the best method without GLM was analyzed, followed by the best method with GLM which represent the performance of the revised estimated values. The values that are printed in bold indicate the least error for method without GLM and method with GLM, with respect to different levels of missingness and distance. Then, a comparison between the performance of the imputation method without and with GLM was made. This comparison indicates overall performance of the imputation method when estimating the missing data.

Several important findings can be drawn based on the error produced by method without GLM in Table 5. IDW performs the best at 25km except in the case of 5% of missing data (see CCW) based on the MAE values. A similar performance pattern can be observed based on the RMSE values where IDW outperforms other methods within this distance regardless to any level of missingness. Meanwhile, CCW perform better for the distance of 35km based on the two error indices. Then, GLM was fitted to the estimated values produced by the imputation methods.

Table 5: Performance of the Imputation Method without and with GLM within 25km and 35km of Neighbouring Stations at JPS Ampang

Error Measures/ Percentage of missingness/ Distance/ Method			**25km**						**35km**				
			SAA	**NRM**	**IDW**	**CCW**	**GC**		**SAA**	**NRM**	**IDW**	**CCW**	**GC**
Method without GLM	**MAE**	**5%**	7.6022	8.0012	7.4707	**7.4697**	7.6019		6.8521	6.8065	7.2521	**6.6909**	6.8516
		10%	6.7542	6.8189	**6.6467**	6.6892	6.7540		6.2097	6.1845	6.4731	**6.1214**	6.2093
		15%	6.1070	6.1427	**5.9583**	6.0735	6.1068		5.5412	5.5896	5.7861	**5.4116**	5.5408
		20%	5.5616	5.7398	**5.3988**	5.5050	5.5613		5.0710	5.3172	5.2451	**4.9281**	5.0706

	RMSE	**5%**	8.6416	8.9817	**8.3513**	8.5108	8.6413		7.6662	7.7901	8.0566	**7.5443**	7.6656
		10%	8.4128	8.4688	**8.2033**	8.3653	8.4125		7.5663	7.5513	7.9391	**7.5235**	7.5658
		15%	7.5848	7.6110	**7.3535**	7.5663	7.5845		6.8916	6.9679	7.1341	**6.8146**	6.8912
		20%	7.1395	7.3347	**6.9266**	7.1371	7.1392		6.5179	6.8380	6.7290	**6.4381**	6.5176
Method with GLM	MAE	**5%**	5.0168	5.0478	**5.0049**	5.0125	5.0168		4.9735	4.9968	4.9922	**4.9690**	4.9734
		10%	5.2867	5.3007	**5.2743**	5.2828	5.2867		5.2455	5.2594	5.2619	**5.2419**	5.2455
		15%	4.7844	**4.7657**	4.7741	4.7806	4.7844		4.7461	**4.7326**	4.7622	4.7389	4.7460
		20%	4.8917	**4.8280**	4.8802	4.8854	4.8917		4.8476	**4.7943**	4.8659	4.8365	4.8476
	RMSE	**5%**	6.0057	6.0403	**5.9927**	6.0001	6.0057		5.9617	5.9864	5.9800	**5.9564**	5.9617
		10%	6.1892	6.2057	**6.1748**	6.1842	6.1892		6.1419	6.1571	6.1608	**6.1376**	6.1419
		15%	5.6684	**5.6492**	5.6558	5.6649	5.6684		5.6275	**5.6157**	5.6434	5.6207	5.6275
		20%	5.8143	**5.7594**	5.8019	5.8106	5.8143		5.7678	**5.7290**	5.7873	5.7585	5.7678

After the fitting process, the deviance was reduced significantly when two harmonics were applied which indicates that the model was fitted very well for both distances. In addition, the p-value of two harmonics was less than 0.01, indicating that no further harmonics required. Using the parameter estimates of the two harmonics, best fitted were produced to estimate the missing data.

Several conclusions can be drawn for the performance of the imputation method with GLM in Table 5. IDW and NRM consistently show best performance when estimating the missing data at 25km, by having the lowest MAE and RMSE values. A similar performance pattern was observed in CCW and NRM at 35km. Meanwhile, the results also show that there were decreases in the value of MAE and RMSE for each of the imputation method at JPS Ampang as the missing percentage increases. This may be due to the existence of high variation in the estimation of missingness. In addition, the neighbouring stations that located within 25km from JPS Ampang were enough to be considered in the estimation process since there were not many differences can be observed between the two distances.

Based on the results in Table 5, a further analysis was performed by comparing the performance of the imputation method without and with GLM. It can be concluded that the imputation method with GLM consistently perform better since these methods consistently produce much lower MAE and RMSE values for all levels of missingness. Thus, it could be suggested that the estimation using GLM with 25km of neighbouring stations will produced more accurate and better estimation results.

3.4. The performance of the revised estimated values in estimating the missing data during rainy and non-rainy period

Generally, Peninsular Malaysia receives high amount of rainfall during the months of October to December. The missingness were created randomly within the last 92 days whereas the rest of the year were assumed to be non-rainy period. The analysis on missingness during rainy period is important since it could help in predicting the flood occurrence that occurred due to the heavy rainfall. Meanwhile, the analysis on missingness during non-rainy period could help in determining the success of tourism sector in Malaysia since the climatic conditions influence the destination choice.

Table 6: Performance of the Imputation Method without and with GLM at JPS Ampang During Rainy Period and Non-Rainy Period

Error Measures/ Percentage of missingness/ Period/ Method			**Rainy Period**						**Non-Rainy Period**				
			SAA	**NRM**	**IDW**	**CCW**	**GC**		**SAA**	**NRM**	**IDW**	**CCW**	**GC**
Method without GLM	MAE	**5%**	5.0806	**4.8341**	4.9375	5.2235	5.0804		4.8643	4.6155	**4.5520**	4.8341	4.8638
		10%	4.8504	4.8033	**4.7391**	4.9336	4.8503		5.1912	5.1794	**4.8391**	5.1244	5.1907
		15%	4.9681	5.0315	**4.8617**	5.1216	4.9679		5.3375	5.3695	**4.9773**	5.1469	5.3370
		20%	5.1106	5.2848	**4.9968**	5.2259	5.1104		4.9475	5.1095	**4.6079**	4.7149	4.9470
	RMSE	**5%**	6.2789	**5.8831**	6.2554	6.4934	6.2788		6.4712	6.3099	**6.2567**	6.4891	6.4709
		10%	5.8745	5.8106	**5.7576**	6.0542	5.8743		7.3713	7.3433	**7.1915**	7.4150	7.3711
		15%	6.1650	6.2826	**6.0518**	6.3445	6.1649		7.3684	7.4649	**7.2141**	7.3496	7.3682
		20%	6.4763	6.6439	**6.3366**	6.6142	6.4761		6.7738	7.0609	**6.5995**	6.7384	6.7735
Method with GLM	MAE	**5%**	3.7372	3.7571	**3.7311**	3.7351	3.7372		3.5834	3.6180	**3.5713**	3.5807	3.5834
		10%	4.4753	4.4897	**4.4599**	4.4703	4.4753		4.6060	4.6154	**4.6004**	4.6043	4.6060
		15%	4.3822	4.3723	**4.3717**	4.3814	4.3822		4.8051	**4.7770**	4.7885	4.8046	4.8051
		20%	4.3259	**4.2860**	4.3118	4.3245	4.3259		4.6074	**4.5292**	4.5883	4.6049	4.6074
	RMSE	**5%**	4.4621	4.4683	**4.4605**	4.4630	4.4621		4.1626	4.2015	**4.1484**	4.1596	4.1626

		10%	5.2810	5.2912	**5.2712**	5.2786	5.2810		5.9743	5.9843	**5.9634**	5.9700	5.9743
		15%	5.0957	5.0910	**5.0893**	5.0967	5.0957		6.1145	**6.0895**	6.0935	6.1130	6.1145
		20%	5.0114	**4.9942**	5.0033	5.0115	5.0114		5.7661	**5.6970**	5.7443	5.7626	5.7660

Therefore, in order to assess the performance of the estimation using GLM, this study considered both rainy and non-rainy period in doing analysis where it was done separately due to heavier rainfall recorded during rainy period and to avoid misleading results. In addition, the neighbouring stations that located within 25km from the target station were selected to perform the analysis since this distance was enough to be considered in the estimation process based the results of previous section.

Table 6 shows the performance of the imputation method without and with GLM during rainy period and non-rainy period at JPS Ampang. Although the MAE and RMSE values at 5% of missingness were relatively high for IDW during rainy period, this method constantly performs better for the rest of missingness. This method was also found to be the best during non-rainy period with the least MAE and RMSE values compared to other methods. Then, the missing data were imputed using the fitted values from the chosen harmonics in the previous section. These fitted values have considered GLM and the results were presented in Table 6 as the imputation method with GLM.

Based on the performance of the imputation method with GLM, it was revealed that IDW constantly shows good performance when estimating the missing data during rainy period at JPS Ampang. IDW also performs very well when estimating the missing data during the non-rainy period, followed by NRM. The findings also indicated that the estimation of the MAE and RMSE increases as the level of missingness increases for both periods. The estimation also showed much better results during rainy period compared to non-rainy period at JPS Ampang, with respect to MAE and RMSE values.

Then, a comparison was made between the performance of the imputation method without and with GLM to evaluate the potential on how GLM can improve the estimation process during both periods. Based on the least MAE and RMSE value, method with GLM was found to be better overall estimator by having a significant result compared to method without GLM regardless to any level of missingness and period.

4. Conclusion

This study aims to estimate the missing rainfall data using five imputation methods and to produce better estimation by fitting the estimated values obtained by the imputation methods using generalized linear model (GLM). To assess the performance of the revised estimated values after the fitting process, two different distance of the neighbouring stations and duration during rainy and non-rainy period were considered with respect to different levels of missingness.

The preliminary results indicated that GLM have potential in improving the estimation process since this model performs excellently when estimating the missing data at JPS Ampang. By fitting this model to the estimated values obtained by the imputation methods gave a new insight and improvement in the estimation process. The results of the study showed that the imputation method with GLM constantly perform better with the least error regardless to any level of missingness and distances. It is also recommended that neighbouring stations within 25km was enough to be considered in producing sufficient estimation of missingness. In addition, the imputation method with consideration of GLM constantly perform best at JPS Ampang, regardless to any period. Throughout this study, it can be concluded that considering GLM when estimating the missing data at JPS Ampang had not been in the lowest priority at all. This may be due to the capability of GLM in describing the rainfall pattern since it can fit to both unimodal or bimodal pattern. In other words, GLM have succeeded in producing more accurate and better missing rainfall data estimation. The work of this study can be extended by covering another state in Peninsular Malaysia and considering another imputation methods in order to strengthen the findings obtained in this existing study.

Acknowledgement

The authors wish to thank Malaysia Meteorological Department and Department of Irrigation and Drainage (DID) for the data and sponsorship from Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Ministry of Higher Education and Institute of Research Management and Innovation (IRMI), UiTM, 600-RMI/FRGS 5/3(58/ 2014).

References

[1] Abebe, A., Solomatine, D. and Venneker, R., "Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events", Hydrological Sciences Journal, Vol. 45, No. 3, pp. 425-436, 2000.

[2] Aly, A., Pathak, C., Teegavarapu, R. S., Ahlquist, J. and Fuelberg, H., "Evaluation of improvised spatial interpolation methods for infilling missing precipitation records", Proceedings World Environment Water Resources Congress, 2009.

[3] Bhattacharya, B., Shrestha, D. and Solomatine, D., "Neural networks in reconstructing missing wave data in sedimentation modelling", Proceedings of the XXXth IAHR Congress, 2003.

[4] Boke, A. S., "Comparative Evaluation of Spatial Interpolation Methods for Estimation of Missing Meteorological Variables over Ethiopia", Journal of Water Resource and Protection, Vol. 9, No. 08, pp. 945, 2017.

[5] Burhanuddin, S. N. Z. A., Deni, S. M. and Ramli, N. M., "Normal ratio in multiple imputation based on bootstrapped sample for rainfall data with missingness", International Journal of GEOMATE, Vol. 13, No. 36, pp. 131-137, 2017.

[6] Chen, F.-W. and Liu, C.-W., "Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan", Paddy and Water Environment, Vol. 10, No. 3, pp. 209-222, 2012.

[7] Coe, R. and Stern, R., "Fitting models to daily rainfall data", Journal of Applied Meteorology, Vol. 21, No. 7, pp. 1024-1031, 1982.

[8] Dastorani, M. T., Moghadamnia, A., Piri, J. and Rico-Ramirez, M., "Application of ANN and ANFIS models for reconstructing missing flow data", Environmental Monitoring and Assessment, Vol. 166, No. 1-4, pp. 421-434, 2010.

[9] De Silva, R., Dayawansa, N. and Ratnasiri, M., "A comparison of methods used in estimating missing rainfall data", Journal of Agricultural Science, Vol. 3, No. pp. 101-108, 2007.

[10] Hanisah, S. and Suhaila, J., "Generalized linear models (GLMs) approach in modeling rainfall data over Johor area", AIP Conference Proceedings, 2014.

[11] Hashim, N. M., Deni, S. M., Shariff, S. S. R., Tahir, W. and Jani, J., Identification of Seasonal Rainfall Peaks at Kelantan Using Fourier Series, Proceedings of the International Symposium on Flood Research and Management (ISFRAM 2015), 2016.

[12] Hussain, Z., Mahmood, Z. and Hayat, Y., "Modeling the daily rainfall amounts of north-west Pakistan for agricultural planning", Sarhad J. Agric, Vol. 27, No. 2, pp. 313-321, 2010.

[13] Jimoh, O. and Webster, P., "Stochastic modelling of daily rainfall in Nigeria: intra-annual variation of model parameters", Journal of Hydrology, Vol. 222, No. 1-4, pp. 1-17, 1999.

[14] Kashani, M. H. and Dinpashoh, Y., "Evaluation of efficiency of different estimation methods for missing climatological data", Stochastic environmental research and risk assessment, Vol. 26, No. 1, pp. 59-71, 2012.

[15] Khorsandi, Z., Mahdavi, M., Salajeghe, A. and Eslamian, S., "Neural network application for monthly precipitation data reconstruction", Journal of Environmental Hydrology, Vol. 19, No. pp. 2011.

[16] Khosravi, G., Nafarzadegan, A. R., Nohegar, A., Fathizadeh, H. and Malekian, A., "A modified distance-weighted approach for filling annual precipitation gaps: application to different climates of Iran", Theoretical and Applied Climatology, Vol. 119, No. 1-2, pp. 33-42, 2015.

[17] Kottegoda, N., Natale, L. and Raiteri, E., "Some considerations of periodicity and persistence in daily rainfalls", Journal of Hydrology, Vol. 296, No. 1-4, pp. 23-37, 2004.

[18] Londhe, S., Dixit, P., Shah, S. and Narkhede, S., "Infilling of missing daily rainfall records using artificial neural network", Journal of Hydraulic Engineering, Vol. 21, No. 3, pp. 255-264, 2015.

[19] Mah, P. J. W. and Shitan, M., "Construction of a New Bivariate Copula Based on Rüschendorf Method", Applied Mathematical Sciences, Vol. 8, No. 153-156, pp. 7645-7658, 2014.

[20] McCullagh, P., "Generalized linear models", European Journal of Operational Research, Vol. 16, No. 3, pp. 285-292, 1984.

[21] Moeletsi, M. E., Shabalala, Z. P., De Nysschen, G. and Walker, S., "Evaluation of an inverse distance weighting method for patching daily and dekadal rainfall over the Free State Province, South Africa", Water SA, Vol. 42, No. 3, pp. 466-474, 2016.

[22] Mohtar, I. S. A., Tahir, W., Bakar, S. H. A. and Zuhari, A. Z. M., Use of Numerical Weather Prediction Model and Visible Weather Satellite Images for Flood Forecasting at Kelantan River Basin, Proceedings of the International Symposium on Flood Research and Management (ISFRAM 2014), 2015.

[23] Sadatinejad, S., Shayannejad, M. and Honarbakhsh, A., "Investigation of the efficiency of the fuzzy regression method in reconstructing monthly discharge data of hydrometric stations in Great Karoon River Basin", Journal of Agricultural Science and Technology, Vol. 12, No. pp. 111-119, 2010.

[24] Sattari, M.-T., Rezazadeh-Joudi, A. and Kusiak, A., "Assessment of different methods for estimation of missing data in precipitation studies", Hydrology Research, Vol. 48, No. 4, pp. 1032-1044, 2017.

[25] Suhaila, J. and Jemain, A. A., "A comparison of the rainfall patterns between stations on the East and the West coasts of Peninsular Malaysia using the smoothing model of rainfall amounts", Meteorological Applications, Vol. 16, No. 3, pp. 391-401, 2009.

[26] Suhaila, J., Sayang, M. D. and Jemain, A. A., "Revised spatial weighting methods for estimation of missing rainfall data", Journal of Atmospheric Sciences, Vol. 44, No. 2, pp. 93-104, 2008.

[27] Tahir, W., Abu Bakar, S. H. and Mohamad, M., Intense convective rain estimation using geostationary meteorological satellite, Advances in Geosciences, 2009.

[28] Tahir, W., Aminuddin, A. K., Ramli, S. and Jaafar, J., "Quantitative precipitation forecast using numerical weather prediction and meteorological satellite for Kelantan and Klang river basins", Jurnal Teknologi, Vol. 79, No. 1, pp. 45-53, 2017.

[29] Teegavarapu, R. S. and Chandramouli, V., "Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records", Journal of Hydrology, Vol. 312, No. 1, pp. 191-206, 2005.

[30] Teegavarapu, R. S. and Nayak, A., "Evaluation of long-term trends in extreme precipitation: Implications of in-filled historical data use for analysis", Journal of Hydrology, Vol. 550, No. pp. 616-634, 2017.

[31] Thevakaran, A. and Sonnadara, D., "Estimating missing daily temperature extremes in Jaffna, Sri Lanka", Theoretical and Applied Climatology, Vol. No. pp. 1-8, 2017.

[32] Wardah, T., Kamil, A., Hamid, A. S. and Maisarah, W., "Statistical verification of numerical weather prediction models for quantitative precipitation forecast", IEEE Colloquium on Humanities, Science and Engineering (CHUSER), 2011.

[33] Woolhiser, D. A. and Pegram, G., "Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models", Journal of Applied Meteorology, Vol. 18, No. 1, pp. 34-42, 1979.

[34] Yozgatligil, C., Aslan, S., Iyigun, C. and Batmaz, I., "Comparison of missing value imputation methods in time series: the case of Turkish meteorological data", Theoretical and Applied Climatology, Vol. 112, No. 1-2, pp. 143-167, 2013.

[35] Zaini, N., Malek, M. A. and Yusoff, M., "Application of computational intelligence methods in modelling river flow prediction: A review", International Conference on Computer, Communication, and Control Technology (I4CT 2015), 2015.