Carpathian Journal of Earth and Environmental Sciences, February 2020, Vol. 15, No. 1, p. 103 - 112; DOI:10.26471/cjees/2020/015/113

ESTIMATION OF MISSING DAILY RAINFALL DURING MONSOON SEASONS FOR TROPICAL REGION: A COMPARISON BETWEEN ANN AND CONVENTIONAL METHODS

Yi Xun TAN[1], Jing Lin NG[1] & Yuk Feng HUANG[2]

[1]Department of Civil Engineering, Faculty of Engineering, Technology and Built Environment, UCSI University, 56000 Cheras. Kuala Lumpur, Malaysia, e-mail: sean_tanyixun@hotmail.com; ngjl@ucsiuniversity.edu.my

[2] Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Selangor, Malaysia, e-mail: huangyf@utar.edu.my

Abstract: An incomplete rainfall data series could affect the reliability of related hydrological modelling. As Malaysia experiences a tropical climate with sensational variations, the availability of complete rainfall series is important for climate change assessments and water resources management. In this study, the estimation of missing rainfall data was carried out using the approach that covers an artificial neural network (ANN) and other conventional methods. These conventional methods were the inverse distance weighting method (IDW), the linear regression (LR) method, the normal ratio (NR) method and the ordinary kriging (OK) method. The performances of the estimation methods were evaluated by the goodness of fit tests, namely the mean absolute error (MAE), mean bias error (MBE), mean square error (MSE), scaled mean square error (SMSE) and the linear correlation coefficient (LCC). From the results, ANN was found to be the overall best estimation method. ANN resulted in lower values for MAE, MSE and SMSE, less biasedness for the MBE and the highest correlation for LCC. From the conventional method list, the OK was selected as the better option. Overall, ANN was more efficient approach to the estimation of missing rainfall data for the Kelantan River Basin in tropical Malaysia.

Keywords: Missing rainfall data, ANN, Conventional methods, tropical climate, Kelantan River Basin

1. INTRODUCTION

Rainfall is a highly significant piece of hydrological component that triggers off the whole chain of hydrological responses in the system and provides information thereafter for all kinds of analysis. As such, beholding detailed knowledge of the rainfall data serves as prerequisite which allows decision making to be carried out for water resources management, climate change assessment, agricultural planning and ecological studies (Huang et al., 2016, Ademe et al., 2019, Lian et al., 2019, Tan et al., 2019). However, in practice, rainfall data often suffer from shortage of consecutive data or missing data (Kang & Yusof, 2012a, Teegavarapu et al., 2018, Jahan et al., 2019). The missing rainfall data could have been contributed by systematic errors or even random errors. Systematic errors arise due to measurement of instruments during the experimental observations. On the other hand, random errors arise due to unforeseen and unpredictable changes that are influenced by surrounding environmental conditions during the experimental observations. All these errors eventually cause the data collected not only to become invalid and deviates from the original purpose of collecting it.

Several methods are used in estimating the missing rainfall data and this depends on the accessibility of rainfall data from neighbouring stations, length of the missing gaps, length of available rainfall data, computational burden and the climate characteristics of study area etc. In general, the conventional methods such as the normal ratio method, inverse distance weighting method, regression based method and the arithmetic average method, are widely applied to estimate the missing rainfall data, especially when the missing gaps are small. Ismail et al., 2017, compared the performances of arithmetic average method, normal ratio method, inverse distance method and coefficient of correlation method in estimating the missing rainfall and streamflow. They found that

inverse distance method and normal ratio method were chosen as the best method for rainfall and streamflow data, respectively. Kizza et al., 2012, used two spatial interpolation methods namely universal kriging method and inverse distance weighting method to estimate missing rainfall data at Lake Victoria Basin. From the cross validation of the generated dataset, it was evident that the universal kriging method shows better performance than the inverse distance weighting method. Teegavarapu et al., 2009, found that a fixed functional set genetic algorithm method (FFSGAM) indicates a better estimation when compared to the traditional IDW method. The FFSGAM was influenced by different topographic locations of the rain gauges as the information of spatial correlations and Euclidean distances are combined. A study was carried out by De Silva et al., 2007, to compare the arithmetic mean method, normal ratio method, inverse distance method, and the newly proposed aerial precipitation ratio method. They found that the arithmetic mean method and the aerial precipitation ratio method were more applicable for upcountry wet zone and mid-country wet zone.

While most of the conventional methods are well known for their simplicity in estimating missing rainfall, there are some criticisms. For instances, the assumption of linearity between variables fall into questions when Simolo et al., 2009, mentioned that weighting method and regression based method were found to overestimate the number of rainy days and underestimate the heavy rainfall events. Further on, the assumption of linear relationship between the observed rainfall and neighbouring stations by the conventional methods may not be true (Mwale et al., 2012, Dubey, 2013). Studies by Teegavarapu, 2014 and Teegavarapu et al., 2012 indicated the limitations of multiple regression method was negative estimates. Another issue in the inverse distance weighting method is the arbitrary selection of the neighbouring stations where the condition is not always true (Di Piazza et al., 2011).

The limitations of the conventional methods have led to increased interest in data driven methods, one of which is the artificial neural network (ANN). ANN has been successfully applied by many researches due to its ability to represent the non-linear characteristics of rainfall and do not require prior information of the underlying process. Londhe et al., 2015, had estimated missing rainfall data in India using ANN and ANN showed its capability by giving low values of error and high values of correlation coefficients. According to the study conducted by Kuligowski and Barros, 1998, the application of ANN using the backpropagation technique had the overall smallest error in estimating missing rainfall data at Middle Atlantic region of the United States. Similar findings were observed in the study carried out by Dubey & Hardaha, 2019. The results indicate that ANN is the most suitable method for the estimation of missing rainfall data in Madhya Pradesh, India. Additionally, Dastorani et al., 2010, compared ANN as adaptive neuro-fuzzy inference system (ANFIS) with two other conventional methods, the NR method and correlation method. The study revealed that ANN was the more efficient method in estimating missing rainfall data in Iran when compared to traditional methods.

Since the actual and complete rainfall data cannot be determined at ease, therefore, the estimation of missing rainfall data in this study is significant to achieve a complete and accurate rainfall data. This is especially useful for Malaysia as the rainfall stations with available complete rainfall series are limited. Malaysia is a tropical country that experiences four monsoon seasons, namely the Southwest Monsoon (SWM), the Northeast Monsoon (NEM), and the two inter-monsoon seasons (INM1 and INM2). Malaysia is often affected by the NEM when from November to February as it receives high volume of rainfall. This often results in disastrous floods and other unfavourable predicaments for the people of Kelantan, loss of precious life and even damages to infrastructure (Ng et al., 2019). The availability of complete rainfall datasets is necessary for the flood control works, climate change assessment and water resources planning. The main objective of this study is to compare the capabilities of artificial neural network (ANN) and other conventional methods, namely inverse distance weighting (IDW), linear regression (LR), normal ratio (NR), and ordinary kriging (OK) in estimating missing rainfall data during monsoon seasons. The performances of the estimation methods are then evaluated with goodness of fit tests. Based on the performance from the goodness of fit tests, the most appropriate method in estimating missing rainfall data can be determined.

2. STUDY AREA

At the north-eastern part of Peninsular Malaysia, almost 80% of the state is covered by the Kelantan River Basin, which makes up an approximate area of 12,000 km^2 (Tan et al., 2017). The Kelantan River Basin originates from the Gunung Tahan and the Titiwangsa ranges which is about 284km long in total. The Kelantan River Basin is located at longitude of 101°15́ E to 102°45́ E and latitude of 4°30́ N to 6° N, as shown in Figure 1. The main inflow to Kelantan River Basin is mainly from the rivers near Kuala Krai, namely; the Galas and the Lebir rivers as it stretches over the coastal plains before pouring into the South China Sea (Pradhan et al., 2009).

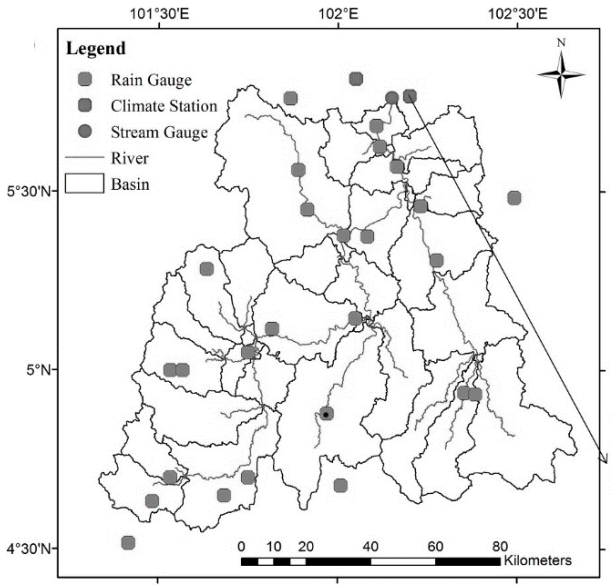

Figure 1. Location of Kelantan River Basin

Table 1. Study area details

Station Code	Station Name	Abbreviation	Record Period	Latitude & Longitude
40431	Pos Blau	PBL	1978-2013	04° 39' N 101° 41' E
40432	RPS Kuala Betis	RKB	1974-2013	04° 42' N 101° 45' E
40433	Pos Hau	PH	1978-2013	04° 42' N 101° 32' E
40470	Pos Lebir	PL	1978-2013	04° 56' N 102° 23' E
40516	Pos Gob	PG	1978-2013	05° 17' N 101° 38' E
40547	Mardi Jeram Pasu	MJP	1984-2013	05° 48' N 102° 20' E
40502	Pos Bihai	PBH	1978-2013	06° 02' N 102° 07' E
40512	Pos Wias	PW	1978-2013	06° 05' N 102° 17' E
48615	Kota Bharu	KB	1954-2013	06° 10' N 102° 18' E
48616	Kuala Krai	KK	1986-2013	05° 32' N 102° 12' E

Daily rainfall dataset were identified and obtained from the Malaysian Meteorological Department (MMD). The geographical coordinates and the available record periods of the 10 rainfall stations of the Kelantan River Basin are described in Table 1. The rainfall dataset available for analysis comprises of a record of minimum 10 continuous years of rainfall readings. These rainfall dataset are then extracted for checking its trend and homogeneity.

3. METHODOLOGY

3.1. Validation of rainfall dataset

3.1.1. *Trend analysis*

Mann-Kendall (MK) Test is proposed to identify the existence of trend in rainfall data. In view of hydrology, the temperature and rainfall are among the climatic parameters used for trend analysis. The null hypothesis of no trend was checked against the alternative hypothesis for existence of either the increasing trend or decreasing trend, using the MK Test which is non parametric (Birara *et al.*, 2018, Jain & Kumar, 2012). Based on the daily rainfall dataset for all 10 rainfall stations, the total rainfall for every monsoon season (INM1, INM2, NEM and SWM), month and year were calculated. The MK Test Statistics, T, is defined as follow:

$$T = \sum_{i<j} sgn\,(x_j - x_k), \qquad (1)$$

where $sgn\,(x) = \begin{pmatrix} 1\ if\ x > 0 \\ 0\ if\ x = 0 \\ -1\ if\ x < 0 \end{pmatrix}$

3.1.2. *Homogeneity*

In order to ensure that the rainfall dataset from the 10 rainfall stations are taken at the same time using the same instrument under similar environmental conditions, the homogeneity of the data must be validated to identify the existence of variability, if any. The homogeneity test is carried out prior to conducting the analysis of rainfall series specifically in Malaysia. Homogeneity check was not applied for station metadata in the form of station history information due to unavailability of data. No other information on the type of land use, classifications of stations, controlling authority and summary of current station equipment were provided. Based on the data provided by the MMD, the basic metadata given were the coordinates (latitude and longitude) and the elevation of each stations. Depending on the objectives and the availability of data, many researchers carried out the homogeneity testing of rainfall at monthly, seasonal and yearly scales (Bickici Arikan & Kahya, 2019; Hadi et al., 2019; Ahmed et al., 2018; Basher et al., 2018; Javari, 2016; Kang & Yusof, 2012b). For this study, Pettit Test (PT) is proposed to test for the homogeneity of the rainfall dataset at monthly, seasonal and yearly scales from the 10 rainfall stations. Under the null hypothesis, the data are homogeneous when the annual values, Y_i of the testing variable Y are identically distributed and independent (where i represents the number of years starting from 1 to n years). Meanwhile, the data are inhomogeneous when the series consist of break

based on the alternative hypothesis. PT is applied based on their rank, r_i of the Y_i and the normality of the series is ignored as follow:

$$X_y = 2\sum_{i=1}^{y} r_i - y(n+1), \quad \text{where } y = 1, 2, 3, \ldots, n \tag{2}$$

The break occurs in year *k* when

$$X_k = \max|X_y| \tag{3}$$

3.2. Estimation of missing rainfall data

3.2.1. *Inverse distance weighting (IDW) method*

The inverse distance weighting (IDW) method is generally used and it is also known as the reciprocal distance weighting method (Teegavarapu *et al.*, 2009). In the estimation of missing rainfall data of an observation, θ_m using the observed value at other station, the weighting method is shown as follow:

$$\theta_m = \frac{\sum_{i-1}^{n} \theta_i d_{mi}^k}{\sum_{i-1}^{n} d_{mi}^k} \tag{4}$$

where *n* is the total number of stations observed used, θ_i is the observation at station *i*, d_{mi} is the distance between the position from station *i* to *m*, and the friction distance ranges between 1.0 to 6.0 is referred as *k*.

3.2.2. *Linear regression (LR) method*

The linear regression (LR) method is a function of the elevation of the rainfall stations that estimates missing rainfall data as follow:

$$\hat{z}(x_o) = f[q(x_o) = \alpha + \beta q(x_o) \tag{5}$$

where the α and β are two regression coefficients assumed to be constant over the study area. The two regression coefficients can be estimated from the rainfall dataset and elevation of the 10 rainfall stations using the ordinary least square method. It is given as follow:

$$\{z(x_i), q(x_i), \quad and\ i = 1, 2, 3, \ldots, N\} \tag{6}$$

In this case, the rainfall data at the observed point is expressed from the elevation at the observed point. Therefore, the records at the neighbouring rainfall stations shall not affect the rainfall data at the observed point. On the other hand, the robust regression method can also be used to determine the two regression coefficients (Di Piazza et al., 2011).

3.2.3. *Normal ratio (NR) method*

The normal ratio (NR) method is applicable when annual rainfall of the stations differs more than 10% of the rainfall stations under consideration. It is observed that the effect of each neighbouring rainfall station is weighed. Paulhus & Kohler, 1952, were among the first to introduce the NR method. The method was then modified by Young, 1992, for a higher precision of estimation (Sattari et al., 2017). A series of combination with parameters of different weights are considered for the estimated data as follow:

$$V_o = \frac{\sum_{i=1}^{n} W_i V_i}{\sum_{i=1}^{n} W_i} \tag{7}$$

where W_i is the weight of i^{th} nearest station defined as:

$$W_i = [R_i^2 \left(\frac{N_i - 2}{1 - R_i^2}\right)] \tag{8}$$

where R_i is the correlation coefficient between the target station and the i^{th} station neighbouring it, N_i is the total number of points used to derive the correlation coefficient.

3.2.4. *Ordinary kriging (OK) method*

The ordinary kriging (OK) method uses a particular function named a semi- variogram as different from the typical IDW method that uses the Euclidean distance to assess the weights, λ_i and measure the difference between the observations. The computation by halving the average squared between the components of the data pairs to obtain the experimental semi-variogram is the first step in kriging as follow:

$$\gamma(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} [z(x_i) - z(x_i + h)]^2 \tag{9}$$

where γ is the experimental semi-variance for the distance *h*, *m(h)* is the number of measured point pairs in the vector distance of class *h*. Both the distance and the direction are the functions of the experimental semi-variogram. When the field is isotropy and dependent only on the distance *h*, it becomes relatively easy to derive the functions (Di Piazza et al., 2011). The OK method is then followed by fitting the theoretical semi-variogram that shows spatial correlation pattern of rainfall in the study area to allow weight assessment. The fitting of the best theoretical semi-variogram is selected based on the ones with lowest residual sum of squares or highest correlation coefficient. Various theoretical semi-variogram with complete description can be obtained from Kitadinis, 1997.

3.2.5. *Artificial neural network (ANN) method*

Briefly, ANN functions quite similar to a human brain with a group of interconnecting neurons.

A nonlinear system can be generated without knowing the mechanism involve behind it as ANN is a self-adapting system. The ANN is useful in determining the patterns among the input and output generated as it works well for models with complex relationships. A training set containing the input pairs and output pairs of data must be modelled. When it is successfully modelled, the final weight vector of the ANN understands the problem of the computation and it is able to resolve the pattern. The Levenberg-Marquardt algorithm is used in this study. The typical structure of ANN is shown in Figure 2.

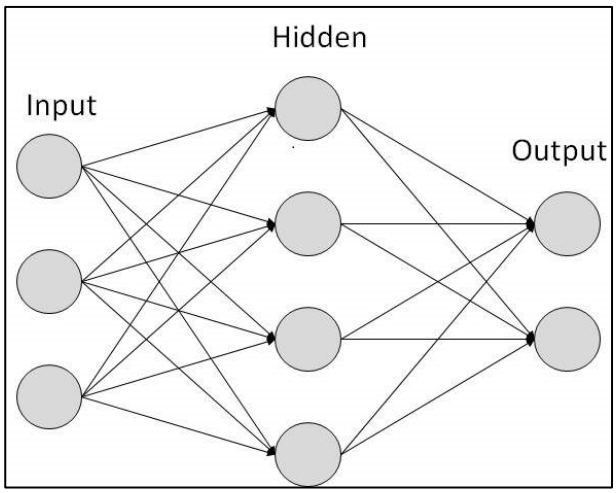

Figure 2. Typical structure of ANN

3.3. Evaluation of performance of methods applied

Based on the results obtained, the performance of each method, which comprises of different algorithms, can be assessed and evaluated from the validation subset using different indexes. The criterion for comparison between the values of estimated, $\hat{z}(x_i)$ and the measured rainfall, $z(x_i)$ are established to determine the strength of statistical relationship in the N_v points of validation subset (Di Piazza et al., 2011). The goodness of fit tests used to evaluate the performance of estimation methods includes the mean square error (MSE), the mean bias error (MBE), the mean absolute error (MAE), the scaled mean square error (SMSE) and the linear correlation coefficient (LCC). The goodness of fit tests was computed using the following set of equations:

The MSE measures the average square difference between the values of estimated and actual rainfall at the N_v validation points.

$$MSE = \frac{1}{N_v}\sum_{i=1}^{N_v}[z(x_i) - \hat{z}(x_i)]^2 \quad (10)$$

The MBE calculates the average bias between the values of estimated and actual rainfall at the N_v validation points.

$$MBE = \frac{1}{N_v}\sum_{i=1}^{N_v}[z(x_i) - \hat{z}(x_i) \quad (11)$$

The MAE measures the average in the amount of error obtained between the values of estimated and actual rainfall at the N_v validation points.

$$MAE = \frac{1}{N_v}\sum_{i=1}^{N_v}|z(x_i) - \hat{z}(x_i)| \quad (12)$$

The SMSE calculates the average square difference between the values of estimated and actual rainfall when divided by the actual rainfall at the N_v validation points.

$$SMSE = \frac{1}{N_v}\sum_{i=1}^{N_v}\left[\frac{z(x_i) - \hat{z}(x_i)}{z(x_i)}\right]^2 \quad (13)$$

The LCC measures the strength of a linear relationship between the values of estimated and actual rainfall at the N_v validation points.

$$LCC = \frac{1}{N_v}\sum_{i=1}^{N_v}\left[\frac{z(x_i) - \hat{z}(x_i)}{z(x_i)}\right]^2 \quad (14)$$

where z_m is the average of actual rainfall value, $\widehat{z_m}$ is the average estimated rainfall value, σ_e is the standard deviation for the value of actual rainfall and σ_o is the standard deviation for the value of estimated rainfall.

3.4. Selection of Best Estimation Method

From the cross validation mentioned above, a mean ranking can be derived from and for each approach used to validate the performance of the methods which comprises of different algorithms. Thus, the best estimation method can be determined based on the mean ranking obtained.

4. RESULTS AND DISCUSSION

4.1. Trend

The MK Test were carried out to identify the existence of trend in the rainfall data using XLSTAT statistical software. Using the MK Test, the rainfall data is free from trend if the p-value is greater than 0.05. Alternately, if the p-value is lesser than 0.05, the null hypothesis is rejected and the alternative hypothesis is accepted where trend is observed in the rainfall data analysed.

Table 2 shows that there were trends detected for the monthly rainfall at the Pos Lebir and Kota Bahru stations. Trends were also discovered for the INM1 (inter-monsoon) rainfall and INM2 rainfall at the Kota Bahru and Pos Blau stations, respectively. Two of the ten rainfall stations had trends for monthly rainfall which is equivalent to 20% of the rainfall data analysed. Subsequently, one of the ten rainfall stations for each of the INM1 and INM2 seasons were observed to have trend which amounts to 10% for each of the rainfall data analysed.

Table 2. Results of Mann-Kendall Test for monthly, seasonal and yearly rainfall.

Station	Monthly	INM1	INM2	NEM	SWM	Yearly
PBL	0.488	0.858	0.020	1.000	0.371	0.721
PL	0.002	0.152	0.721	0.283	0.858	0.371
KB	0.039	0.049	0.152	0.371	1.000	0.210
KK	0.635	0.721	0.474	0.858	0.152	0.592
MJP	0.308	0.283	0.152	0.371	0.107	0.283
PG	0.079	0.592	1.000	0.371	0.721	0.107
PH	0.118	0.283	0.074	0.210	0.474	0.074
RKB	0.421	0.592	0.283	1.000	0.592	1.000
PBH	0.075	0.283	0.371	0.474	1.000	0.474
PW	0.408	0.152	0.721	0.721	0.592	1.000

- Grey shaded value indicated that trend was detected at the rainfall station at 95% significance level.

The presence of trend may be due to the changes in the long term rainfall over the Kelantan River Basin. The results of trend analysis are dependent on several factors such as the period of rainfall data and the geographic location of the rainfall station (Jain & Kumar, 2012). In this study, the rainfall data were only selected for a period of 10 years, although more historical data were available to ensure consistency among all other rainfall stations with lesser historical data.

Additionally, trend can be identified at the rainfall stations due to localized rainfall. The localized rainfall occurs where the rainfall observed are focused and concentrated at one particular rainfall station. However, the rainfall event may not occur at the neighbouring rainfall stations (Ng et al., 2018 and Lian et al., 2019). Even though trend were recorded, the rainfall data are still fit for continuing studies on the estimation of missing rainfall data.

4.2. Homogeneity

The PT were carried out to validate the variability of the rainfall data using the XLSTAT statistical software. Table 3 shows the inhomogeneous monthly rainfall at the Pos Lebir and Pos Bihai stations. Similarly, the INM1 rainfall and INM2 rainfall were detected to be inhomogeneous at the Pos Gob and Pos Hau stations, respectively.

From Table 3, 2 out of 10 rainfall stations are inhomogeneous for monthly rainfall which is equivalent to 20% of the rainfall data analysed. Subsequently, 1 out of 10 rainfall stations for each of the INM1 and INM2 seasons were observed to be inhomogeneous which amounts to 10% for each of the rainfall data analysed.

Table 3. Parameters of probability distribution at rainfall stations

Station	Monthly	INM1	INM2	NEM	SWM	Yearly
PBL	0.618	0.165	0.192	0.421	0.431	0.963
PL	0.002	0.088	0.407	0.094	0.159	0.905
KB	0.311	0.085	0.268	0.267	0.596	0.153
KK	0.103	0.747	0.424	0.409	0.267	0.414
MJP	0.707	0.750	0.425	0.444	0.911	0.740
PG	0.300	0.590	0.015	0.750	0.966	0.274
PH	0.478	0.015	0.331	0.770	0.576	0.320
RKB	0.875	0.976	0.254	0.957	0.396	0.587
PBH	0.039	0.911	0.713	0.752	0.426	0.267
PW	0.978	0.408	0.950	0.413	0.580	0.089

- Grey shaded value indicated that trend was detected at the rainfall station at 95% significance level.

In the PT, the null hypothesis, H0, has one or more distributions at the same location parameter for variable T whereas the alternative hypothesis, Ha, is a two tailed test where time exists when the variables changes the location parameter. The rainfall data analysed can be further grouped into three different classes where Class A is classified as useful, Class B is doubtful and Class C is suspect respectively.

Relocations of rainfall stations and changes in measurement routines could lead to inhomogeneity in rainfall data. Although the four rainfall stations mentioned earlier were inhomogeneous and classified in “Class B” as doubtful due to variation in the rainfall data, the rainfall data are still fit for carrying out further studies on the estimation of missing rainfall data if handled with care (Wijngaard et al., 2003, Ng et al., 2015).

4.3. Selection of best estimation method

The overall points from each goodness fit test from Table 4 to Table 8 were added and the overall rankings are shown in Table 9. The conventional methods such as the IDW, LR, NR and OK were compared with the ANN. The OK method was identified as the best conventional method for the estimation of missing rainfall data. The findings are found to be similar to those of studies carried out by Di Piazza et al., 2011, where the OK method provided the

best performance to complete the rainfall series in Sicily, Italy.

Based on Table 4 to Table 8, the results indicated that the estimation of missing rainfall data were more accurate using the OK method. The OK method is a modern tool that applies the geostatistical methods to create a surface network based on the coordinates of the neighbouring rainfall stations. Among the few types of network available in Kriging (ordinary and universal), the OK disregards the variable of elevation, z in the exponential model used in the estimation of missing rainfall data.

In the creation of a surface network, the OK method uses the secondary attributes of the rainfall data (coordinates of rainfall station) which are more densely sampled with the primary attributes of the rainfall data based on the application of the geostatistical methods. It is known that the OK method has fewer computational problems due to simpler calculations since the exponential model is more restrictive in its assumptions (Yang et al., 2015).

During the estimation of missing rainfall data, the geostatistical method considers the spatial pattern in the observed rainfall data. Goovaerts, 2000, and Di Piazza et al., 2011, has shown that the OK method is able to estimate a better result of missing rainfall data since the conventional methods does not take the spatial pattern for the observed rainfall into considerations.

In addition, the spatial pattern of the observed rainfall data was also influenced by the weights obtained from the semi-variogram used in the exponential model. The finding is similar to the study conducted by Ly et al., (2011), where the estimated rainfall data using the OK method shows a greater correlation than the IDW method. Hence, the OK method was selected as the best conventional method.

Based on Table 9, ANN was selected as the overall best estimation method. The findings are similar to the studies conducted by Dubey & Hardaha, 2019 where the ANN performed the best in the estimation of missing rainfall data in Madhya Pradesh, India. In the application of ANN, it was trained using the Levenberg-Marquardt algorithm which was one of the built in function in MATLAB R2016a software. According to the study conducted by Dubey, 2013, the relationship between the rainfall observed and the neighbouring stations shall not be linear. Nevertheless, the application of ANN has the ability to construct a non-linear equation to estimate the missing rainfall data using the Levenberg-Marquardt algorithm (Dubey & Hardaha, 2019). Therefore, the ANN produced the highest accuracy and least error in the estimation of missing rainfall data.

Additionally, it is to note that the application of ANN in the estimation process differs from the conventional methods (Kuligowski & Barros, 1998). ANN uses the approach of local in space that applies different equations to estimate the missing rainfall data at the targeted station. Then, the ANN is further enhanced by using the improved rainfall data from the selected time series. In comparison, the known values of the rainfall data for the conventional methods are assumed to be missing. The conventional methods apply the approach of global in space which selects the rainfall data at the targeted station. Then, the rainfall data are improved using the rainfall data selected from a single time series. After that, the estimated rainfall data are compared to the actual value which were initially assumed to be missing. A significantly lesser amount of computation time is required for the conventional methods since the rainfall data at the targeted station are optimized in whole (global in space). However, the estimated rainfall data are less accurate for the conventional methods which applies the approach of local in time.

Table 4. Summary of performance for monsoon seasons based on MAE

Season	Estimation Method	Overall Points
INM1	*IDW*	37
	LR	39
	NR	35
	OK	29
	ANN	10
INM2	*IDW*	38
	LR	39
	NR	35
	OK	28
	ANN	10
NEM	*IDW*	29
	LR	46
	NR	36
	OK	29
	ANN	10
SWM	*IDW*	39
	LR	40
	NR	34
	OK	27
	ANN	10

As for the approach used by ANN, it has numerous advantages over its disadvantages. Using the scheme of global in time, the rainfall data can be optimized and performed via offline. However, this action requires development and progression using a large list of rainfall data. As for the scheme of local in space, this approach is more reliable and accurate as no involvement of correlograms are observed. With the absence of the empirical spatial relationship,

the signals in the rainfall data inside the neural network are more efficient in the estimation of missing rainfall data. However, the disadvantage of this scheme includes greater computation time since the rainfall data at the targeted station is being analysed individually.

Table 5. Summary of performance for monsoon seasons based on MBE

Season	Estimation Method	Overall Points
INM1	*IDW*	30
	LR	42
	NR	38
	OK	29
	ANN	11
INM2	*IDW*	34
	LR	43
	NR	30
	OK	31
	ANN	12
NEM	*IDW*	32
	LR	42
	NR	32
	OK	31
	ANN	13
SWM	*IDW*	33
	LR	45
	NR	32
	OK	30
	ANN	10

Table 6. Summary of performance for monsoon seasons based on MSE

Season	Estimation Method	Overall Points
INM1	*IDW*	30
	LR	42
	NR	38
	OK	29
	ANN	11
INM2	*IDW*	34
	LR	43
	NR	30
	OK	31
	ANN	12
NEM	*IDW*	32
	LR	42
	NR	32
	OK	31
	ANN	13
SWM	*IDW*	33
	LR	45
	NR	32
	OK	30
	ANN	10

Table 7. Summary of performance for monsoon seasons based on SMSE

Season	Estimation Method	Overall Points
INM1	*IDW*	35
	LR	33
	NR	38
	OK	34
	ANN	10
INM2	*IDW*	40
	LR	34
	NR	36
	OK	30
	ANN	10
NEM	*IDW*	35
	LR	36
	NR	35
	OK	34
	ANN	10
SWM	*IDW*	44
	LR	28
	NR	36
	OK	32
	ANN	10

Table 8. Summary of performance for monsoon seasons based on LCC

Season	Estimation Method	Overall Points
INM1	*IDW*	35
	LR	19
	NR	37
	OK	39
	ANN	20
INM2	*IDW*	25
	LR	19
	NR	33
	OK	25
	ANN	48
NEM	*IDW*	29
	LR	20
	NR	38
	OK	36
	ANN	27
SWM	*IDW*	28
	LR	36
	NR	24
	OK	22
	ANN	40

Table 9. Overall ranking of estimation methods

Estimation Methods	Total Points	Overall Ranking
IDW	672	3
LR	733	5
NR	681	4
OK	607	2
ANN	307	1

5. CONCLUSIONS

This study was aimed at investigating and subsequently determining the most appropriate method for estimating missing daily monsoonal rainfall data for the Kelantan River Basin, Malaysia. Five different estimation methods such as the conventional methods (IDW, LR, NR and OK) and the ANN were investigated. These five estimation methods were then applied to estimate the missing rainfall data at rainfall stations in the Kelantan River Basin, for the annual four different monsoon seasons. Using the estimated rainfall obtained, these data were compared with the actual rainfall and then evaluated for goodness of fit with tests consisting of a set of separate individual error-related parameters namely; MAE, MBE, MSE, SMSE and LCC.

A ranking system was derived and it clearly indicated that the ANN was the overall better estimation method. This is due to its ability in analysing the rainfall data individually and expressing the spatial pattern of the rainfall data in Kelantan River Basin. The ANN has proved to become a far more credible estimation method than the other conventional methods.

From the ranking system, OK method was selected as the best conventional method amongst the IDW, LR, NR and OK. Although, these conventional methods are commonly used in the estimation of missing rainfall data, however, these conventional methods do not have the ability to address the spatial relationship of the rainfall data, excepting for the OK method, which eventually having outperformed the IDW, LR and NR methods in estimating missing rainfall data.

In conclusion, the consistency and continual rainfall data were among the factors for a viable hydrological studies and analysis such as weather forecasting. Beholding the knowledge of rainfall distribution around the globe, the design of hydrological structure can be more cost effective and optimum in the mitigation efforts of floods in the Kelantan state. By applying the ANN estimation method, one is able to acquire and estimate the possible missing rainfall data with better accuracy which is of great significance in this study.

Acknowledgement

The authors would like to express their sincere gratitude to the Malaysian Meteorological Department (MMD) for providing the record of daily rainfall data.

REFERENCES

Ademe, F., Kibret, K., Beyene, S., Getinet, M. & Mitikee, G. 2019. *Rainfall analysis for rain-fed farming in the Great Rift Valley basins of Ethiopia*. Journal of Water and Climate Change.

Ahmed, K., Shahid, S., Ismail, T., Nawaz, N. & Wang, X. J. 2018. *Absolute homogeneity assessment of precipitation time series in an arid region of Pakistan.* Atmosfera, 31(3), 301–316.

Basher, M. A., Stiller-Reeve, M. A., Saiful Islam, A. K. M. & Bremer, S. 2018. *Assessing climatic trends of extreme rainfall indices over northeast Bangladesh.* Theoretical and Applied Climatology, 134(1–2), 441–452.

Bickici Arikan, B. & Kahya, E. 2019. *Homogeneity revisited: analysis of updated precipitation series in Turkey.* Theoretical and Applied Climatology, 135(1–2), 211–220.

Birara, H., Pandey, R. P. & Mishra, S. K. 2018. *Trend and variability analysis of rainfall and temperature in the Tana basin region, Ethiopia*. Journal of Water and Climate Change, **9**(3), 555-569

Dastorani, M. T., Moghadamnia, A., Piri, J. & Rico-Ramirez, M. 2010. *Application of ANN and ANFIS models for reconstructing missing flow data*. Environmental monitoring and assessment, 166(1-4), 421-434.

De Silva, R. P., Dayawansa, N. D. K. & Ratnasiri, M. D. 2007. *A comparison of methods used in estimating missing rainfall data*. Journal of agricultural sciences, 3(2).

Di Piazza, A., Conti, F. L., Noto, L. V., Viola, F. & La Loggia, G. 2011. *Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy*. International Journal of Applied Earth Observation and Geoinformation, 13(3), 396-408.

Dubey, M. 2013. *Development of Artificial Neural Network for Estimation of Missing Rainfall Data*. Doctoral dissertation, JNKVV, Jabalpur.

Dubey, M. & Hardaha, M.K. 2019. *Application of Standard Models and Artificial Neural Network for Missing Rainfall Estimation.* Int. J. Curr. Microbiol. App. Sci. 8(01), 1564-1572.

Goovaerts, P. 200). *Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall*. Journal of hydrology, 228(1-2), 113-129.

Hadi, S. J., Hadi, A. J., Ismail, K. S. & Tombul, M. 2019. *Homogeneity and Trend Analysis of Rainfall and Streamflow of Seyhan Basin (Turkey).* In: Chaminé H., Barbieri M., Kisi O., Chen M., Merkel B. (eds) Advances in Sustainable and Environmental Hydrology, Hydrogeology, Hydrochemistry and Water Resources. Advances in Science, Technology & Innovation (IEREK Interdisciplinary Series for Sustainable Development). Springer, Cham.

Huang, Y. F., Mirzaei, M. & Amin, M. Z. M. 2016. *Uncertainty quantification in rainfall intensity duration frequency curves based on historical extreme precipitation quantiles*. Procedia Engineering, 154, 426-432.

Jahan, F., Sinha, N.C., Rahman, M.M., Rahman, M.M., Mondal, M.S.H. & Islam, M.A. 2019. *Comparison of missing value estimation techniques in rainfall data of Bangladesh*. Theor. Appl. Climatol. 136, 1115–1131.

Jain, S. K. & Kumar, V. 2012. *Trend analysis of rainfall and temperature data for India*. Current Science(Bangalore), 102(1), 37-49.

Javari, M. 2016. *Trend and Homogeneity Analysis of Precipitation in Iran*. Climate, 4(3), 44.

Kang, H. M. & Yusof, F. 2012a. *Application of self-organizing map (SOM) in missing daily rainfall data in Malaysia.* International Journal of Computer Applications, 48(5).

Kang, H. M. & Yusof, F. 2012b. *Homogeneity Tests on Daily Rainfall Series in Peninsular Malaysia.* International Journal of Contemporary Mathematical Sciences, 7(1), 9–22.

Kitadinis, P. 1997. *Introduction to Geostatistics: Applications in Hydrogeology*. Cambridge University Press, UK.

Kizza, M., Westerberg, I., Rodhe, A. & Ntale, H. K. 2012. *Estimating areal rainfall over Lake Victoria and its basin using ground-based and satellite data*. Journal of Hydrology,464, 401-411.

Kuligowski, R. J. & Barros, A. P. 1998. *Using artificial neural networks to estimate missing rainfall data*. JAWRA Journal of the American Water Resources Association, 34(6), 1437-1447.

Lian, C. Y., Huang, Y. F., Ng, J. L., Mirzaei, M., Koo, C. H. & Tan, K. W. 2019. *A proposed hybrid rainfall simulation model: bootstrap aggregated classification tree–artificial neural network (BACT-ANN) for the Langat River Basin, Malaysia*. Journal of Water and Climate Change.

Londhe, S., Dixit, P., Shah, S. & Narkhede, S. 2015. *Infilling of missing daily rainfall records using artificial neural network*. ISH J. Hydraul. Eng. 21, 255–264.

Ly, S., Charles, C., & Degre, A. 2011. *Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the Ourthe and Ambleve catchments, Belgium.* Hydrology and Earth System Sciences, 15(7), 2259-2274.

Mwale, F.D., Adeloye, A.J. & Rustum, R. 2012. *Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi – A self organizing map approach*. Phys. Chem. Earth, Parts A/B/C 50–52, 34–43.

Ng, J. L., Aziz, S. A., Huang, Y. F., Mirzaei, M., Wayayok, A. & Rowshon, M.K. 2019. *Uncertainty analysis of rainfall depth duration frequency curves using the bootstrap resampling technique*. Journal of Earth System Science, 128 (5), 113.

Ng, J.L., Abd Aziz, S., Huang, Y.F., Wayayok, A. & Rowshon, M.D., 2015. *Homogeneity Analysis of Rainfall in Kelantan, Malaysia*. J. Teknol. 76.

Ng, J. L., Aziz, S. A., Huang, Y. F., Wayayok, A. & Rowshon, M.K. 2018. *Analysis of annual maximum rainfall in Kelantan, Malaysia*. In III International Conference on Agricultural and Food Engineering 1152, 11-18.

Paulhus, J. L., & Kohler, M. A. 1952. *Interpolation of missing precipitation records*. Monthly Weather Review, 80(8), 129-133.

Pradhan, B., Shafiee, M. & Pirasteh, S. 2009. *Maximum flood prone area mapping using RADARSAT images and GIS: Kelantan river basin*. International Journal of Geoinformatics, 5(2).

Sattari, M. T., Rezazadeh-Joudi, A. & Kusiak, A. 2017. *Assessment of different methods for estimation of missing data in precipitation studies*. Hydrology Research, 48(4), 1032-1044.

Simolo, C., Brunetti, M., Maugeri, M. & Nanni, T. 2009. *Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach*. Int. J. Climatol. 30, n/a-n/a.

Tan, K. W., Loh, P. N., & Huang, Y. F. 2019. *Development of climate hazards decision support system: A study of Cameron Highlands, Malaysia.* Carpathian Journal of Earth and Environmental Sciences, 14(2), 495-504, DOI:10.26471/cjees/2019/014/098.

Tan, M. L., Yusop, Z., Chua, V. P. & Chan, N. W. 2017. *Climate change impacts under CMIP5 RCP scenarios on water resources of the Kelantan River Basin, Malaysia*. Atmospheric research, 189, 1-10.

Teegavarapu, R. S., Tufail, M. & Ormsbee, L. 2009. *Optimal functional forms for estimation of missing precipitation data.* Journal of Hydrology, 374(1-2), 106-115.

Teegavarapu, R.S.V., Meskele, T. & Pathak, C.S. 2012. *Geo-spatial grid-based transformations of precipitation estimates using spatial interpolation methods*. *Comput. Geosci. 40, 28–39*.

Teegavarapu, R.S.V. 2014. *Missing precipitation data estimation using optimal proximity metric-based imputation, nearest-neighbour classification and cluster-based interpolation methods*. Hydrol. Sci. J. 59, 2009–2026.

Teegavarapu, R.S. V., Aly, A., Pathak, C.S., Ahlquist, J., Fuelberg, H. & Hood, J. 2018. *Infilling missing precipitation records using variants of spatial interpolation and data-driven methods: use of optimal weighting parameters and nearest neighbour-based corrections.* Int. J. Climatol. 38, 776–793.

Ismail, W. N. W., Zin, W. Z. W. & Ibrahim, W. 2017. *Estimation of rainfall and stream flow missing data for Terengganu, Malaysia by using interpolation technique methods*. Malaysian J. Fundam. Appl. Sci. 13.

Wijngaard, J. B., Klein Tank, A. M. G. & Können, G. P., 2003. *Homogeneity of 20th century European daily temperature and precipitation series*. International Journal of Climatology: A Journal of the Royal Meteorological Society, 23(6), 679-692.

Yang, X., Xie, X., Liu, D. L., Ji, F. & Wang, L. 2015. *Spatial interpolation of daily rainfall data for local climate impact assessment over greater Sydney region*. Advances in Meteorology.

Young, K. C. 1992. *A three-way model for interpolating monthly precipitation values*. Monthly Weather Review 120, 2561–2569.

Received at: 16. 10. 2019
Revised at: 02. 01. 2020
Accepted for publication at: 06. 01. 2020
Published online at: 15. 01. 2020