ASEAN Engineering Journal

Full Paper

A COMPARATIVE STUDY TO ADDRESS MISSING DATA FROM HISTORICAL RAINFALL DATA IN KELANTAN AND PAHANG, MALAYSIA

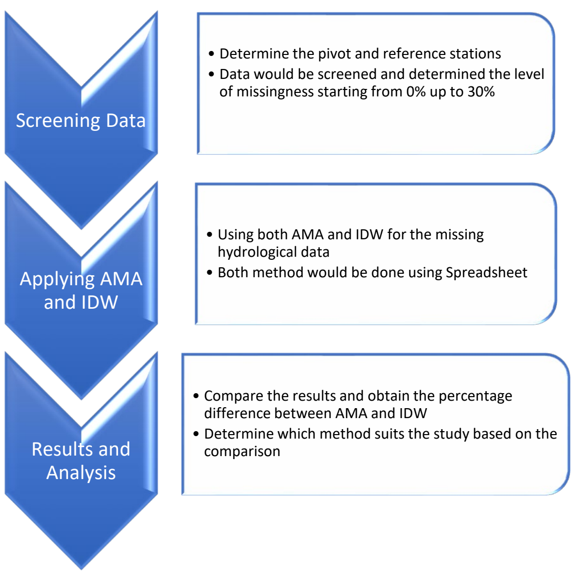
Shah Aiman Iqbal Sahrol*, Fauzi Baharudin

Faculty of Civil Engineering, UiTM Shah Alam, 40450, Shah Alam, Selangor, Malaysia

Article history
Received
07 August 2025
Received in revised form
27 November 2025
Accepted
17 March 2026
Published online
31 August 2026

*Corresponding author
2023281174@isiswa.uitm.edu.my

Graphical abstract



Abstract

Climate change poses a critical threat to Malaysia, particularly in the context of increasing flood risks. The country is experiencing more frequent and intense rainfall events, leading to heightened incidences of flooding. These climatic shifts can be attributed to global warming and changing weather patterns. Malaysia's vulnerability to climate-induced floods necessitates urgent and comprehensive adaptation strategies. However, if there are any lacking or missing data to be analyze, hence the next step could be jive. Addressing missing data considers to be the most important part to do before proceeding to further hydrological procedure such as statistical comparison or Intensity-Duration-Frequency (IDF). Arithmetic Mean Average and Inverse Distance Weighing method would be use in this study to address missing data. The site location for this study were in Kelantan and Pahang. The stations involved in addressing the missing data are Station Pertanian Lundang, Stor JPS Kota Bharu and Mardi Kubang Keranji in Kelantan while Station Sungai Lembing, Bukit Goh and Sungai Panching Selatan in Pahang were selected. The range of data for all the stations were from 2013 to 2023 depends on which station have missing data throughout the years. The results shown that the percentage difference or compatibility between the two methods varies from 0 up to 50 percent difference. This study considers to be significance as this method would be the first step to do before proceeding to another hydrological method. Authorities such as Department of Irrigation and Drainage (DID) would be beneficiary to this study as they will use the missing rainfall data for further hydrological steps.

Keywords: Climate change, Missing data, rainfall data, arithmetic mean average, inverse distance weighing

1.0 INTRODUCTION

Climate change is a global phenomenon that has far-reaching implications for ecosystems and communities around the world. This happened when there are increment of greenhouse gases in atmosphere [1]. As the Industrial Revolution began, the Carbon Dioxide emission increases up to one third. In the context of Malaysia, one of the most pressing and visible consequences of climate change is the increasing frequency and severity of floods. Blessed with its diverse geography and tropical climate, Malaysia is particularly vulnerable to the impacts of climate change. The changing climate patterns, characterized by more intense rainfall and rising sea levels, contribute to a heightened risk of flooding in various regions of the country[2].

United Nations Framework Convention on Climate Change (UNFCC) in 1992 describes Climate change as any change of climate that is attributed directly or indirectly to human activity that alters the composition of the global atmosphere and that is in addition to natural climate variability observed over comparable periods. The impacts of climate change-induced high rainfall intensity in Malaysia which are multifaceted, affecting both urban and rural areas [3]. Rapid urbanization, deforestation, and inadequate infrastructure influence the vulnerability of communities to flooding. In urban centers like Kuala Lumpur, the combination of heavy rainfall and poor drainage systems often leads to flash floods, disrupting daily life and causing damage to property and infrastructure[4]). In rural areas, where agriculture plays a crucial role in livelihoods, floods can result in crop failures, loss of livestock, and economic hardships for local communities [2].

The climate of Malaysia is classified as tropical, and it is marked by high temperatures and heavy rainfall throughout the entire year. It is of the utmost importance to investigate the high rainfall intensity in Malaysia because of the impact it has on a variety of areas of the surrounding environment, infrastructure, and society. It is possible for intense rainfall events to result in flash floods, landslides, and other natural disasters, which presents substantial issues for urban planning, water resource management, and disaster preparedness[5][6] .However, due to missing historical rainfall data screened from the secondary data obtained from authorities such as Department of Irrigation and Drainage (DID) and Meteorology Malaysia, the issue needs to be resolved first before proceeding to any relevant hydrological studies. Rainfall data may be missing due to some reasons such as equipment failure, human errors and data transmissions. Having missing data may reduce the accuracy of hydrological modeling, trend detection and design of infrastructures to overcome any natural disasters. Many places reported forecasting gradual warming of 1.8 to 4.0 degree Celsius with average of 3.0 degree Celsius during next century [2]. Additionally, international collaboration and concerted efforts to mitigate greenhouse gas emissions are essential to tackle the root causes of climate change. The effects of climate changes can be found in all parts of the world including Malaysia. The expected changes that could be describes are changes in precipitation pattern, increasing risk of drought, increasing ocean level globally and increasing frequencies of bad weather.[7]

Norliyana et al., did AMA in addressing missing data in Terengganu, Malaysia but however using Root Mean Square Error (RMSE) in order to determine the level of suitability for the data to be used [8]. [9] used three main methods in addressing missing hydrological data which were Arithmetic Mean Average method, Normal Ratio Method and Inverse Distance Weighing method (IDW). The location of the study was done at mountainous area in Sri Lanka, India. Other than that, there was a study made by Wuthinwongyothin [10] where the AMA, IDW and Coefficient of Correlation (CC) took place. As previous researches have done the research at Terengganu, this study would take place in Kelantan and Pahang to covers up the East Coast area of Malaysia Peninsular. This study avoids in using many states as there are so many human error or missing data at other states makes the data to be less dependable in using it the researcher compares the three method and concluded that there is only a small variation of results depending on the different percentage missing amounts. Hence induced that the infilling methods are not sensitive to the percentage of missing values based on [11]. These researches shared the same objective in addressing the missing hydrological data but none of them would compare the percentage the difference as it will be use in this study. However, missing data causes climate change assessment tougher. Incomplete or missing rainfall data may cause problem to detect long-term trends due to climate change. In example, if extreme rainfall years are missing, the calculation of Climate Change Factor (CCF) would be underestimated and may causes the authority to be unprepared for extreme rainfall events.

Therefore, the specific objectives for this study are to address missing rainfall data obtained from secondar data retrieved from DID and Meteorology Malaysia. The second objective is to apply Arithmetic Mean Average method (AMA) and Inverse Distance Weighing Method (IDW) for different categories of missing data. The last research objective would be to compare and determine which method suits best in resolving the missing data before proceeding to any relevant hydrological study.

2.0 METHODOLOGY

This section will describe methods and the instruments used in conducting this study, including the process of preliminary checking the missing data, determining the methods as well as addressing the missing data. Figures 1 and Figure 2 show the stations involved in Kelantan and Pahang used in this study. Table 1 shown the lists of the stations involved in this study. The stations are labelled with yellow pin and circled in red. All of the stations might not share the same basin or catchment but however it is still reliable to be used as it was not strictly required. However, it is recommended for the stations to be in the same basin to improved accuracy and hydrological consistency. Due to lack of stations in the same basin, the study would only consider the nearest station to be used in addressing the missing hydrological data[12].

The stations involved in Pahang are Station Sungai Lembing, Station Bukit Goh and Station Sungai Panching Selatan. On the other hand, the stations involved in Kelantan are Station JPS Kota Bharu, Station Mardi Kubang Keranji and Station Pertanian Lundang. The pivot stations which will be addressed the missing hydrological data are Station Sungai Lembing and Station Stor JPS Kota Bharu accordingly by states. There are only two states to be use in this study as there are lack of data availability throughout the years due to gauge or human error. The secondary data that would be use in this study were obtained through relevant authorities which are JPS Malaysia and Meteorology Malaysia. There is a previous researcher which covers part of East Coast side of Malaysia Peninsular such as Terengganu [8] hence in complying the total East Coast, Pahang and Kelantan would be use in this study to caters it. The location of the study can be seen in Figure 3 where Kelantan and Pahang were highlighted.

Screening criteria in this study would varies from 0 to 10 percent and 10 percent to 30 percent depends on degree of missingness of the data. 0 to 10 percent of missing data would induce that the data series is still considered to be statistically complete and representative of the true rainfall pattern. This statement was supported by research done by [13] mentioning that AMA performs well when the data are minimal among nearby stations. While for 10 to 30 percent of missing data would use IDW as the missing portion could distort rainfall statistics if not treated carefully. IDW considered to be reliable as it gives higher weights to nearby stations and lower weights to stations with greater distance producing a more realistic estimation. [14] mentioned in his study that IDW performs well for datasets where missing data are moderate which are 10 to 30 percent of missingness due to weighting based on distance helps maintaining spatial continuity.

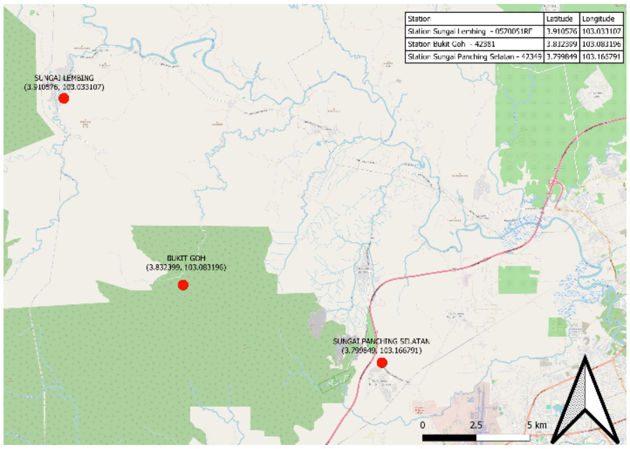


Figure 1 Stations involved in Pahang State, Malaysia.

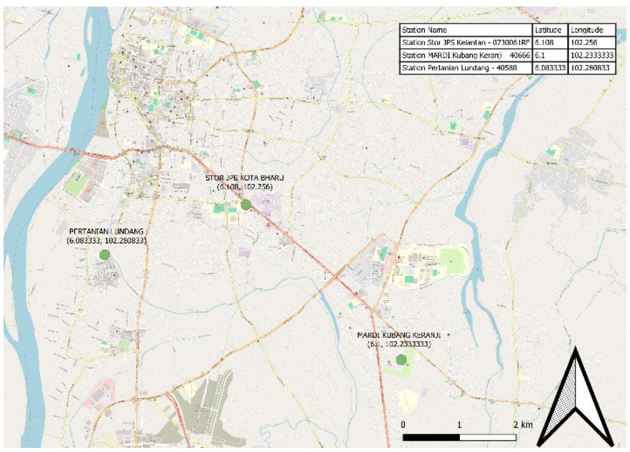


Figure 2 Stations involved in Kelantan State, Malaysia.

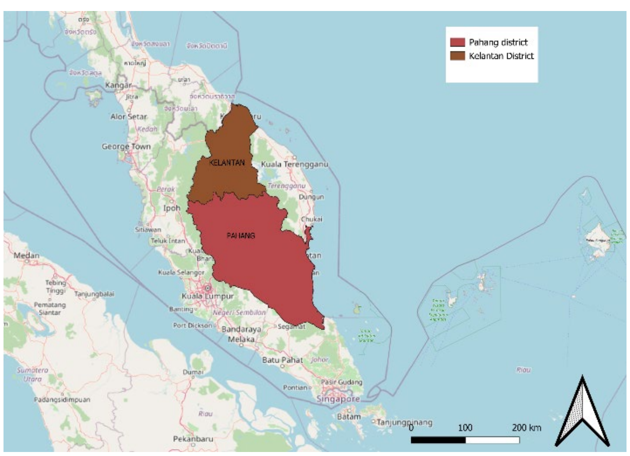


Figure 3 Kelantan and Pahang in Peninsular Malaysia

2.1 Screening Data

This study was conducted in two states of Malaysia which are Kelantan and Pahang. These states are located at the central and the east coast part of Malaysia Peninsular. The rainfall data were obtained from Department of Irrigation and Drainage Malaysia (DID) and Meteorology Malaysia. Researchers usually use secondary data when primary data collection is unavailable due to time, cost or lack of equipment. Secondary data can broaden the scope of analysis, especially when access to primary data is limited or unethical. Figure 4 shows the overall flowchart while processing the data.

The rainfall data obtained from the authorities has its lacking and missing of data makes it vulnerable to be use further in hydrological analysis. The process of screening was done by counting the days of missing data in a year and converted into percentages. For the year that has less than 10% of missing data, a simple AMA method was used. While for the year that varies from 10% to 30% of missing data, an IDW considering the distance between stations were preferred. However, it is acceptable to use AMA method for 10% to 30% missing data if the nearby stations have highly correlated rainfall patterns and are in similar climatic and topographic zones [7] [15][16]

Other than that, it can be acceptable when the distribution of missing data is random or spread across time rather than large continuous gaps. Although it is acceptable to be use on higher missing percentage of hydrological data, it does not consider distance or elevation making it less accurate and reliable when topographic effects are significant. Topographic effects are considered in IDW method when the distance was included in the calculation. On the other hand, IDW method seems to be suitable for low percentage missing data as well as it can maintain the original data structure without introducing much bias

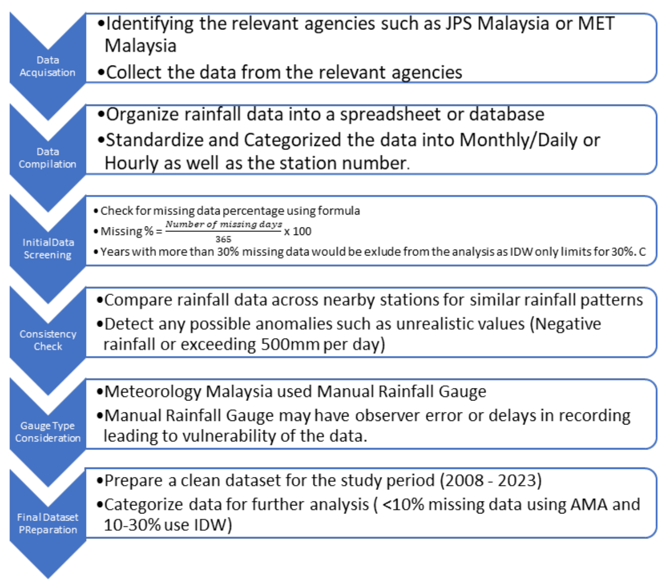


Figure 4 Flowchart of Processing Data

Table 1 Stations involved in all states in Malaysia

State	Station	Station Number
Pahang	Sungai Lembing	0570051RF
Pahang	Bukit Goh	42381
Pahang	Sungai Panching Selatan	42349
Kelantan	Pertanian Lundang	40588
Kelantan	Stor JPS Kota Bharu	0730061RF
Kelantan	Mardi Kubang Keranji	40666

Table 2 Missing Data Percentages for Kelantan

Year	Percentage of Missing Data (%)
2008	20.27
2009	0.272
2010	7.95
2011	0.55
2012	0
2013	3.562
2014	1.37
2015	0.55
2016	4.932
2017	1.644
2018	1.37
2019	1.096
2020	0
2021	0
2022	0
2023	0

Table 3 Missing Data Percentages for Pahang

Year	Percentage of Missing Data (%)
2008	23.56
2009	0
2010	0.55
2011	0
2012	0
2013	21.92
2014	0
2015	0
2016	0
2017	7.945
2018	0
2019	0
2020	0
2021	0
2022	0
2023	0

2.2 Arithmetic Mean Average Method

In this study, AMA was done for the year that has less than 10% of missing data. AMA offers a simple and efficient estimate of the central tendency with a dataset, particularly when the data distribution is symmetrical and measured at the interval or ratio [17]. According to [13], the effective application of the AMA method requires complete dataset from at least three surrounding stations on the same dates as the missing observations to ensure similar topographical characteristics can be achieved.

This study only utilized two nearby stations in which the stations met the criteria of spatial proximity and topographical similarity to the pivot station, hence maintaining the validity of the AMA application. The pivot station, where data are missing was compared with reference stations to generate the necessary values using the AMA. The method calculates the average of the available data using (1) where n represents the number of stations included in the computation. As highlighted by Smith et al. [18], the AMA method is particularly valuable for comparing data across multiple groups and monitoring temporal changes making it well-suited to be use before using it for further hydrological studies.

$$\textit{Arithmetic Mean}=\frac{\sum x}{n} \quad (1)$$

$\sum X$ = Summation of Observed value at station
n = Total number of stations used

IDW is an interpolation technique used to estimate values at unsampled locations based on values from known data points. There are a few conditions like AMA method previously such as at least 3 to 5 nearby stations are needed and the data is evenly distributed.

However, it is acceptable to only use 2 nearby stations if the stations are reasonably close to the target location. In this study, there are only 2 reliable stations can be used as there many lacking and missing of data at other stations. Considering that the other non-included stations are reasonably far in distance from the pivot stations, it can be considered that the outcome of this method will probably be less accurate Inverse Distance Weighing method could effectively estimate missing historical rainfall data, even with a limited number of nearby stations. Methods such as interpolation (Kriging) can also be used to increase the accuracy of the data [18]

The steps of this method started with identifying the location with missing data. Hence, determine the nearby stations where two or more nearby stations could be used. The distances between stations and the pivot station in straight line needs to be calculated with the assistance of spreadsheet. The *d* in the IDW Equation (Equation 2) would be fill with the distance obtained from Google Earth Pro. The Equation can be described where $Z(x_0)$ is the estimated value at the unknown location, $Z(x_i)$ is the observed value at location *i* , $d(x_0), (x_i)$is the distance between the unknown and known points, and *P* is the power parameter

$$Z(x_0) = \frac{\sum_{i=1}^{n} \frac{Z\,(x_i)}{d\,(x_0, x_i)^p}}{\sum_{i=1}^{n} \frac{1}{d(x_0\ , x_i)^p}} \quad (2)$$

3.0 RESULTS AND DISCUSSION

3.1 Screening Process

Screening process required the secondary data to be screened with the aid of spreadsheet to ensure consistency and accuracy prior to analysis. During this stage, the presence of missing data was identified and quantified before going to the next stage. To determine the percentage of missing data for each year, the total number of missing daily records was divided by the total number of days in a year. This calculation provided a clear metric of data completeness, enabling the classification of datasets based on the proportion of missing values. The year considered to be use in this study would be from 2016 to 2023 as the previous year has so many missing data makes it unreliable to be use further in this study. For the year that has less than 10% of missing data, a simple AMA method would be used. This method considered to be straightforward and efficient allowed for quick estimation of missing rainfall data by averaging observations from nearby stations with similar geographic characteristics. While for the year that varies from 10% to 30%, an IDW Method considering the distance between stations would be done. The IDW Method can be use as it counts spatial relationships between stations by assigning greater weights to stations located closer to the pivot station thus producing more reliability in using the data for further hydrological procedures.

3.2 Arithmetic Mean Average Method

AMA method was used to address simple missing data varies from 1% to 10% in this study. AMA method is mathematically straightforward as the pivot station is estimated by simply averaging the values of surrounding reference stations. This requires minimal computational effort and can be carried out using spreadsheet making it less vulnerable to be miscalculated[13]. For a year that has more missing data, a more robust method such as IDW method would be use in this study as AMA does not require detailed spatial weighting calculations or correlation models makes it less vulnerable to be wrong where distance is unavailable to be use. Table 2 shown the missing data addressed at Station Sungai Lembing in 2017 where there are 7.945% missing data. As the numbers of missing data were divided by the total days over the year (365 days), an amount of 7.945% were obtained.

Throughout the year 2017 in Pahang, Only February and March have missing data makes it reliable to be use in this study. AMA method requires nearby complete data station to be use as main reference. Using the general equation in (1) as in-situ calculation, the data in Sungai Panching Selatan and Felda Bukit Goh were sum up and was divided with two to get the average value. AMA method plays an important role in quality control and gap-filling of climatic datasets

These steps would be done accordingly to the years that has less than 10% missing data throughout the years. These steps are also need to be done to another two states in Malaysia (Kelantan and Pahang). Considering the screening method to be done first, then the data would be classified either it is suitable for AMA method or IDW method. AMA method assumed to be useful in long-term climate and water resource assessments. Findings indicate that agencies involved in the development of climatic records and Intensity-Duration-Frequency (IDF) curves usually use AMA method to pre-process data or filling gaps before applying any further statistical analysis such as Gumbel Distributions and Log-Pearson method. [19] Tables 2 and 3 above shown different condition of missing data varies from 0% to 30% missing data where the researcher determined which method to be use based on the methodology of the study.

Table 4 Root Mean Square Error (RMSE) done to AMA vs IDW in Pahang

Pahang	2008	2013	2017
IDW	2.20621	26.24226	9.240081
AMA	3.419919	20.09213	7.074602

Table 5 Root Mean Square Error (RMSE) done to AMA vs IDW in Kelantan

Kelantan	IDW	AMA
2008	13.00	10.79
2010	7.21	4.90
2013	3.78	2.75
2014	0.98	0.85
2016	1.31	1.13
2017	0.07	0.06
2018	0.99	0.85
2019	18.06	15.56

Table 6 Example of Addressing Missing Hydrological Data at Station Sg. Lembing, Pahang in 2017 using AMA Method

Day	Month	Station Sg. Lembing	Station Sg. Panching Selatan	Station Felda Bukit Goh
1	2	0	0	0
2	2	0	0	2.4
3	2	30	8.4	0
4	2	14	2.3	0
5	2	39	26.3	0
6	2	23.65	4.6	42.7
7	2	1.25	0	2.5
8	2	0	0	0
9	2	0.2	0.4	0
10	2	0	0	0
11	2	1.05	2.1	0
12	2	0.85	1.7	0
13	2	11.65	9.2	14.1
14	2	10.95	10.9	11
15	2	8.8	10.4	7.2
16	2	2.7	0.4	5
17	2	0	0	0
18	2	1.35	2.7	0
19	2	0.65	1.3	0
20	2	15.35	0	30.7
21	2	1.45	2.9	0
22	2	1.3	2.6	0
23	2	0	0	0
24	2	5.25	0.7	9.8
25	2	15.35	30.7	0
26	2	14.75	29.5	0
27	2	37.4	42.6	32.2
28	2	13.85	5.9	21.8

Example of AMA calculation would be done using Equation (1) where n is the number of station and x is the total of rainfall for observed station for each particular day. As the total of reference stations in this study would be 2, hence the n value for this study would be constant. As an example, the AMA for rainfall event on 6th of February would be 23.65 mm. The same step would be used throughout the years.

$$\frac{(4.6+42.7)}{2} = 23.65 \text{ mm}$$

3.3 Inverse Distance Weighing Method

IDW method is a method widely used in estimating missing hydrological and meteorological data [20]. The fundamental principle of IDW is that data from locations closer to the missing point are more similar and therefore more reliable than those who has more distance away. IDW method is a method that is sensitive to distance assumption. Distances need to be carried out properly to avoid any miscalculation of addressing missing hydrological data. IDW assumes that the influence of a reference station decreases strictly with distance. However, rainfall is not always spatially uniform as local topography, elevations and wind patterns may cause nearby stations to have different rainfall patterns. For a complex terrain, IDW methods may be risky to be use considering the limitations mentioned previously. Therefore, the distance between stations is very important for this method to succeed. The distance between stations were determined with the aid of Google Earth Pro and two pins were used to measure the distance directly.

The distances were mainly important in this method for the method to be more accurate. By weighting observations based on inverse distance, IDW acknowledges spatial autocorrelation decreases with distance, an essential aspect for environmental variables such as rainfall. [21]. Mathematically, the estimated value at a missing point is calculated as a weighted average of known values, where the weights are determined by the inverse of the distance between each known station and the estimation point. However, IDW method has its own limitations. The method does not consider elevation, terrain barriers or wind direction which may influence weather patterns. As the name distance itself, the influence of each neighboring station is inversely proportional to its distance typically raised to a power commonly 1 or 2 to ensure that closer stations have more influence [22]

Table 3 shows the steps which were measuring the distance between stations and calculating the missing data using spreadsheet. In this example, the same station which Station Sungai Lembing was use as the pivot or known as missing station. The main stations to be use as reference in this study would be Station Sungai Panching Selatan and Station Bukit Goh. In expressing the general equation as mentioned before in the methodology section, the distance would be applied in the equation and further calculation was done with the aid of spreadsheet.

As mentioned by [21], the distances between stations are so important in this method hence using GPS coordinates (latitudes and longitudes) or mapping tools seems to be the best method in determining the distances. Although more stations indicate the results to be less vulnerable, but it was mentioned previously [23]that IDW method can be used effectively with as few as two stations, especially when they are the only nearby data sources. Different stations at different state would have different distances as well as the other parameter. In this station, the distance used was measured in meters. Figure 5 shown that the distance was measured in Meters before applying into the method. The power use in this study would be 1 as lower values of power indicates smooths effect of distance allowing farther stations to have more influence.

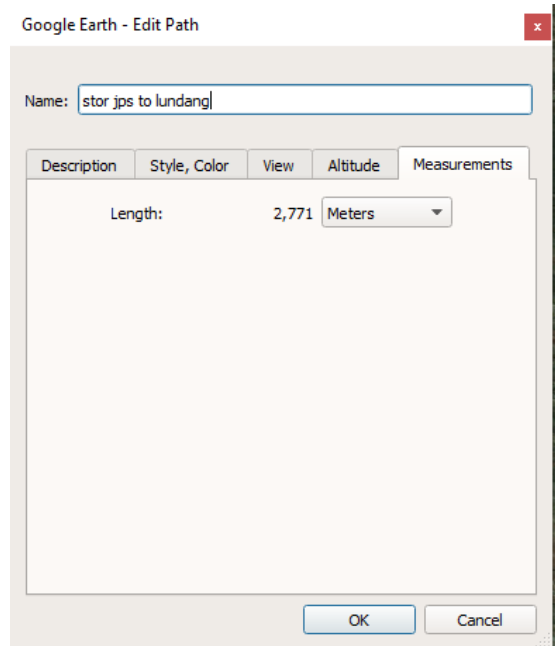


Figure 5 Distance between Station Stor JPS Kota Bharu and Station Pertanian Lundang, Kelantan

Table 7 Example of Addressing Missing Hydrological Data at Station Stor JPS Kelantan, Kelantan in 2008 using IDW Method

Day	Month	Station Stor JPS Kelantan	Station Mardi Kubang Keranji	Station Pertanian Lundang
1	7	0	2.7	0
2	7	0	0	20
3	7	1	51.8	0
4	7	0	0	0
5	7	0	0	0
6	7	0	0	0
7	7	3.5	46.7	0
8	7	0	0	23.9
9	7	0	0	0
10	7	4.6	17.3	0
11	7	0	0	0
12	7	0	0	0
13	7	2.4	8.9	0
14	7	0	0	0
15	7	18.2	7	22.2
16	7	0	0	0
17	7	24	90.7	0
18	7	22.9	0	31.1
19	7	0.6	2.3	0
20	7	32.9	34.4	32.3
21	7	0	0	0
22	7	0	0	0
23	7	0	0	0
24	7	2.7	10.2	0
25	7	0.0	0	0
26	7	8.9	22	4.2
27	7	4.2	0	5.7
28	7	0	0	0
29	7	0	0	0
30	7	0	0	0
31	7	0	0	0

Example of calculation is shown below using the Equation (2) where distance from each station were measured in this method. The distance from Station Stor JPS Kelantan and Mardi Kubang Keranji is approximately 4772 meter while the distance from Station Stor JPS Kelantan to Station Pertanian Lundang is approximately 1717 meter. Based on rainfall event on 10th of July, the IDW value would be 4.6mm and then verified using RMSE to check the eligibility of the results. The same step would be used throughout the years with 10 to 30 percent of missing hydrological data.

$$\frac{\left(\frac{17.3}{4772}+\frac{0}{1717}\right)}{\frac{1}{4772}+\frac{1}{1717}} = \mathbf{4.6\ mm}$$

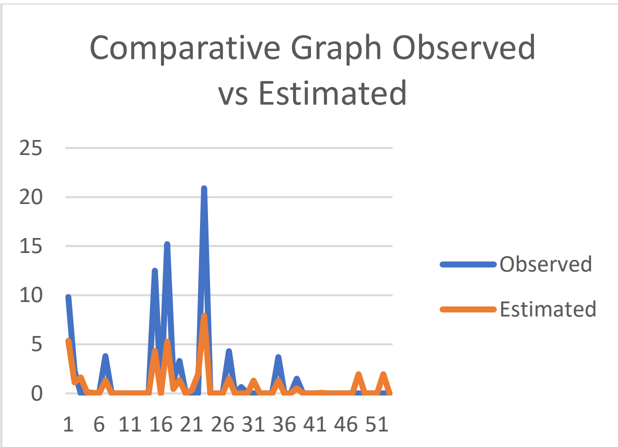


Figure 6 Comparative Graph Observed vs Estimated

Percentage Difference between Arithmetic Mean Average Method and Inverse Distance Weighing Method

Both methods were applied on the same year and the percentage differences were determined.

Table 8 Addressing Missing Data using AMA Method at Station Sungai Lembing, Pahang in 2017

Day	Month	Station Lembing Sg.	Station Panching Selatan Sg.	Station Felda Bukit Goh
1	2	0	0	0
2	2	0	0	2.4
3	2	30	8.4	0
4	2	14	2.3	0
5	2	39	26.3	0
6	2	23.65	4.6	42.7
7	2	1.25	0	2.5
8	2	0	0	0
9	2	0.2	0.4	0
10	2	0	0	0
11	2	1.05	2.1	0
12	2	0.85	1.7	0
13	2	11.65	9.2	14.1
14	2	10.95	10.9	11
15	2	8.8	10.4	7.2
16	2	2.7	0.4	5
17	2	0	0	0
18	2	1.35	2.7	0
19	2	0.65	1.3	0
20	2	15.35	0	30.7
21	2	1.45	2.9	0
22	2	1.3	2.6	0
23	2	0	0	0
24	2	5.25	0.7	9.8
25	2	15.35	30.7	0
26	2	14.75	29.5	0
27	2	37.4	42.6	32.2
28	2	13.85	5.9	21.8
1	3	22.95	22.4	23.5
2	3	25.5	40	11
3	3	13.4	0	26.8
4	3	0.55	1.1	0
5	3	3.6	7.2	0
6	3	3.5	0	7
7	3	0	0	0
8	3	0	0	0
9	3	0	0	0
10	3	0	0	0
11	3	0	0	0
12	3	0	0	0
13	3	0	0	0
14	3	0	0	0
15	3	5.5	0	0
16	3	3.5	0	0
17	3	15	0	0
18	3	0	0	0
19	3	0	0	0
20	3	0	1.8	0
21	3	48	49	9
22	3	0	0	24.8
23	3	0	0	0
24	3	0	0	0
25	3	0	31	0
26	3	0.5	0	0
27	3	2.5	0	1.8
28	3	16	0	0
29	3	2	0	15.5
30	3	10.5	47.4	0
31	3	9.5	6.2	36.6

Table 9 Addressing Missing Data using IDW Method at Station Sungai Lembing, Pahang in 2017

Day	Month	Station Lembing Sg.	Station Panching Selatan Sg.	Station Felda Bukit Goh
1	2	0	0	0
2	2	0	0	2.4
3	2	30	8.4	0
4	2	14	2.3	0
5	2	39	26.3	0
6	2	29.5	4.6	42.7
7	2	1.6	0	2.5
8	2	0	0	0
9	2	0.1	0.4	0
10	2	0	0	0
11	2	0.7	2.1	0
12	2	0.6	1.7	0
13	2	12.4	9.2	14.1
14	2	11	10.9	11
15	2	8.3	10.4	7.2
16	2	3.4	0.4	5
17	2	0	0	0
18	2	0.9	2.7	0
19	2	0.5	1.3	0
20	2	20	0	30.7
21	2	1	2.9	0
22	2	0.9	2.6	0
23	2	0	0	0
24	2	6.6	0.7	9.8
25	2	10.7	30.7	0

26	2	10.2	29.5	0
27	2	35.8	42.6	32.2
28	2	16.3	5.9	21.8
1	3	23.1	22.4	23.5
2	3	21.1	40	11
3	3	17.5	0	26.8
4	3	0.4	1.1	0
5	3	2.5	7.2	0
6	3	4.6	0	7
7	3	0	0	0
8	3	0	0	0
9	3	0	0	0
10	3	0	0	0
11	3	0	0	0
12	3	0	0	0
13	3	0	0	0
14	3	0	0	0
15	3	5.5	0	0
16	3	3.5	0	0
17	3	15	0	0
18	3	0	0	0
19	3	0	0	0
20	3	0	1.8	0
21	3	48	49	9
22	3	0	0	24.8
23	3	0	0	0
24	3	0	0	0
25	3	0	31	0
26	3	0.5	0	0
27	3	2.5	0	1.8
28	3	16	0	0
29	3	2	0	15.5
30	3	10.5	47.4	0
31	3	9.5	6.2	36.6

Table 10 Percentage Difference for AMA and IDW Method at Pahang in 2017

Month	Arithmetic Mean Average Method (AMA)	Inverse Distance Weighing Method (IDW)	Percentage Difference (%)
2	23.65	29.48	19.8
2	1.25	1.63	23.4
2	0	0.0	0
2	0.2	0.15	44.1
2	0	0.0	0
2	1.05	0.73	44.1
2	0.85	0.59	50.0
2	11.65	12.4	6.0
2	10.95	10.97	0.1
2	8.8	8.31	5.9
2	2.7	3.4	20.7
2	0	0.0	0
2	1.35	0.94	44.1
2	0.65	0.45	44.1
2	15.35	20.05	23.4
2	1.45	1.01	44.1
2	1.35	0.9	44.1
2	0	0	0
2	5.25	6.64	21.0
2	15.35	10.65	44.1
2	14.75	10.24	44.1
2	37.4	35.81	4.4
2	13.85	16.28	14.9
3	22.95	23.12	0.7
3	25.5	21.06	21.1
3	13.4	17.5	23.4
3	0.55	0.38	44.1
3	3.6	2.5	44.1
3	3.5	4.57	23.4

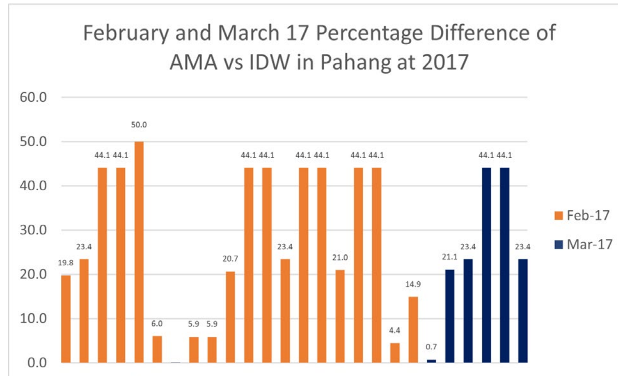


Figure 7 Barchart of February and March Percentage Difference of AMA vs IDW in Pahang at 2017

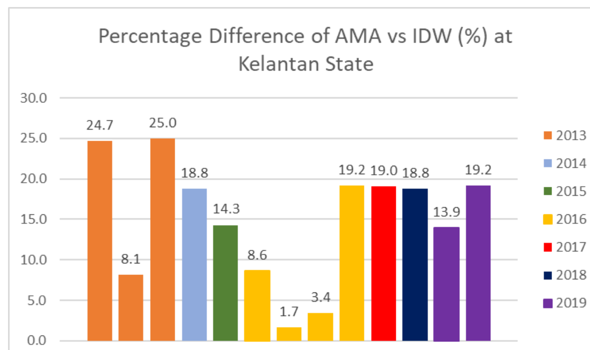


Figure 8 Barchart of Percentage Difference of AMA vs IDW in percentage at Kelantan State

Figure 7 shows the percentage difference of AMA and IDW at Station Sungai Lembing, Pahang. The bar in orange and dark blue indicates the percentage different for February and March in 2017 only. Only in Pahang where there is only a year of 2017 which can be used to address the missing data compared to the previous state, Kelantan where there are many years that have missing data.

In February 2017, the chart shows a considerable number of data points exceeding 40% with 50% as the peak indicates that the AMA produce significantly different estimations compared to IDW method. However, there are certain data points in February also exhibit much lower differences such as 5.9% and 6% showing that under certain conditions or for specific dates, both methods produced similar results.

Figure 8 shown the percentage difference of AMA and IDW for Station Stor JPS Kota Bharu, Kelantan. As Kelantan has many missing historical data throughout the year, hence the barchart combines all the percentage difference into one figure. The legend at the right side of the barchart indicates which colour represents the year. In the year 2013, the chart displays the highest discrepancy with a percentage error at 25% and another value 24.7% indicating a considerable difference between AMA and IDW estimations during that year. IDW may have captured localized rainfall more effectively due to its distance-based weighing approach.

While for the year 2016, the lowest percentages error is observed with values as minimal as 1.7% and 3.4% indicating a very close agreement between the two methods. This may occur because under certain hydrological conditions, AMA can perform as good as IDW. A moderate discrepancy observed in 2015 and 2018 with values ranging from 13.9% to 18.8% produced a relative performance of AMA and IDW is context-dependent based on a few aspects. In strengthening the confidence of the addressed missing data, RMSE method was done to measure which method suits well for every years/stations. The result can be seen in Table 2 and Table 3.

A lower percentage differences in the results shown that rainfall patterns among the surrounding stations are relatively homogenous suggesting that both methods provide similar estimation of missing data. This often occurs during periods of consistent regional rainfall or when the surrounding station located at a similar topographic area and climatic zones. While for higher percentage differences, this may occur due to monsoon variability or natural phenomena such as *La Nina*. Years with extreme rainfall event may cause higher discrepancies due to increased spatial heterogeneity in rainfall distribution.

The application of AMA and IDW methods in addressing missing rainfall data provides certain significant implications for hydrological studies (Tables 8 and 9). AMA method is advantageous due to its simplicity and efficiency especially if the missing data percentage is minimal (0 to 10 percent of missingness). While for IDW method offers greater spatial precision by assigning weights to nearby stations based on their stations[14].

From a climate modelling perspective, the consistency of reconstructed rainfall data plays crucial role in validating Regional Climate Models (RCMs) and Global Climate Models (GCMs). Failure in addressing missing data may lead to significant bias in trend analysis and rainfall projections. In terms of flood risk management, both methods are important as the rainfall data would be use in the development of Intensity-Duration-Frequency (IDF) curves and designing flood mitigation structures. If missing data are not accurately addressed, under or overestimation may lead to inappropriate infrastructure design as well as increasing flood vulnerability.

4.0 CONCLUSION

The study has highlighted that the two method which are AMA and IDW method has their own percentage difference. The sample calculation of Addressing Missing data in Pahang and Kelantan can be seen from Table 6. Based on the results, a clear comparison between the AMA and the IDW Method can be observed. The IDW method consistently provided alternative values that were either higher or lower than those derived using the arithmetic mean, with varying degrees of percentage difference across different dates. These differences highlight the sensitivity of each method to spatial variability and proximity of nearby stations. Based on Root Mean Square Error (RMSE) method done previously on the results, it can be shown that AMA is a better and safer options if the distance is not considered into calculation. The lowest RMSE compared between AMA and IDW, result shown that only 1 out of 11 figures indicates that IDW are lesser in value of RMSE.

Referring to Table 10 with Percentage Difference for AMA and IDW at Pahang, It is evident that on several occasions, the difference between two methods was significant which was exceeding 30% particularly for days with low rainfall. This may indicate that the amounts of exceeding 30% occurs when AMA method may overestimate values due to its inability to consider spatial distributions mainly because of no distance and topographical aspects included. From all the data extracted, it can be mentioned that there are a few major errors with a max of 50 percent which indicates that it is a major issue to be used in this study for both methods. This may occur as AMA is a direct method to address missing data and only reliable to be use if there are no distances can be obtained between the stations. However, it was mentioned before that IDW is more accurate due to involvement of spatial distributions where distances and topographical was considered. From Figure 6 and Figure 7, the result shown a consistent amount of percentages error not exceeding 50 percent which can deduce that both methods can be use as long as the missingness of data does not exceed 50 percent. However, referring to Tables 4 and 5, the best method was determined using RMSE. Kelantan station have a lesser amount of RMSE on AMA compared to IDW as well as in Pahang. In example, Kelantan station in 2013 shown that the RMSE for IDW is 3.78 and 2.75 for AMA. There is no significant difference throughout the years/stations as can be seen in Tables 4 and 5. The difference of RMSE was considered to be so small that both methods can be use on the stations depending on parameters availability (Weighting Distance and Topography).

Implications for future hydrological studies founds that both methods should be context-driven depends on the extent of missing data. Future hydrological studies should include a combination of both AMA and IDW depending on rainfall uniformity and data availability which in this study is limited. Using more validation method such as Root Mean Square Error (RMSE), Nash-Sutcliffe Efficiency (NSE) and Coefficient of Correlation (CC) can further increase the confidence in replacing the missing datasets. From policy and practical relevance, authorities such as Department of Irrigation and Drainage (DID) and Meteorology Malaysia may use this study and must first acknowledge the missing threshold from 0 to 30 percent. Using this method can improve data quality and improve consistency across regional studies in Malaysia. This result may affect Malaysia's flood management or climate adaptation where authorities may use the data for further hydrological studies. Authorities can design infrastructures to overcome or reduce flooding prone in Malaysia especially at the East Coast of Malaysia Peninsular.

Overall, this study show that AMA method is simpler and effective in homogenous settings, the IDW method provides a more spatially realistic estimation as it involved distances and topographical aspects. The consistency of the high amount of percentage difference indicates that IDW method reinforces its importance for spatially sensitive hydrological modelling especially for cases that has distances parameter available to be included. For recommendations, the data can be re-evaluate using any other interpolation method such as regression and kriging method. Both methods are relatively proven where AMA simply calculates the average of nearby stations assuming equal influence while IDW assigns weights based on distance assuming closer stations contribute more reliable estimates.

Acknowledgement

This research is fully supported by Ministry of Higher Education (MOHE) via MyBrain2.0 scholarship. The authors fully acknowledged Ministry of Higher Education (MOHE) and Universiti Teknologi MARA (UiTM) for the approved fund which makes this important research viable and effective.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper

References

1] M. Filonchyk, M. P. Peterson, L. Zhang, V. Hurynovich, and Y. He, 2024, "Greenhouse gases emissions and global climate change: Examining the influence of CO2, CH4, and N2O," *Science of The Total Environment*, 935: 173359, DOI: 10.1016/J.SCITOTENV.2024.173359.

[2] H. A. Rahman, 2018. "CLIMATE CHANGE SCENARIOS IN MALAYSIA: ENGAGING THE PUBLIC," *International Journal of Malay-Nusantara Studies*, 1(2): 65-78,

[3] A. H. A. Tarmizi, S. N. Rahmat, A. T. A. Karim, and N. N. A. Tukimat, 2019, "Climate change and its impact on rainfall," *International Journal of Integrated Engineering*, 11(1): 170–177, DOI: 10.30880/ijie.2019.11.01.020.

[4] M. Y. Safiah Yusmah *et al.*, 2020, "Understanding urban flood vulnerability and resilience: a case study of Kuantan, Pahang, Malaysia," *Natural Hazards*, 101(2): 551–571, DOI: 10.1007/s11069-020-03885-1.

[5] N. D. Mohd Idris and N. Abd Majid, 2024, "An Analysis of Malaysia's Flood Vulnerability Study Using the Geographical Information System Technique," *International Journal of Academic Research in Business and Social Sciences*, 14(7): 288-302, DOI: 10.6007/ijarbss/v14-i7/21831.

[6] B. C. C. Xian, C. W. Kang, M. A. Wahab, M. R. R. M. A. Zainol, and F. Baharudin, 2021. "Evaluation of low impact development and best management practices on peak flow reduction using SWMM," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd,. DOI: 10.1088/1755-1315/646/1/012045.

[7] Z. Sa'adi *et al.*, 2024, "Spatiotemporal assessment of rainfall and drought projection for integrated dam management in Benut River Basin, Malaysia under CMIP6 scenarios," *Environmental Challenges*, 15: 100892. DOI: 10.1016/J.ENVC.2024.100892.

[8] W. Norliyana Wan Ismail, W. Zawiah Wan Zin, and W. Ibrahim, 2017. "Estimation of rainfall and stream flow missing data for Terengganu, Malaysia by using interpolation technique methods,", *Penerbit UTM Press*,

[9] H. P. G. M. Caldera, V. R. P. C. Piyathisse, and K. D. W. Nandalal, 2016. "A Comparison of Methods of Estimating Missing Daily Rainfall Data," *Engineer: Journal of the Institution of Engineers, Sri Lanka, 49*(4), 1–8. The Institution of Engineers, Sri Lanka,

[10] S. Wuthiwongyothin, 2021. C. Kalkan, and J. Panyavaraporn, "Evaluating Inverse Distance Weighting and Correlation Coefficient Weighting Infilling Methods on Daily Rainfall Time Series." *SNRU Journal of Science and Technology, 13*(2): 71–79. Sakon Nakhon Rajabhat University,

[11] M. Az-Zuhri Azman, R. Zakaria, N. Fadhilah, and A. Radi, "Estimation of Missing Rainfall Data in Pahang Using Modified Spatial Interpolation Weighting Methods." *AIP Conference Proceedings* 1643(1):65-72

[12] M. L. Tan and X. Yang, 2020. "Effect of rainfall station density, distribution and missing values on SWAT outputs in tropical region," *Journal of Hydrology,* 584: 124660, *Amsterdam*, DOI: 10.1016/J.JHYDROL.2020.124660.

[13] R. Berndtsson, 2016. *"Chow's Handbook of Applied Hydrology,"* McGraw Hill Education, 2nd Edition,

[14] K. N. Dirks, J. E. Hay, C. D. Stow, and D. Harris, 1998, "High-resolution studies of rainfall on Norfolk Island: Part II: Interpolation of rainfall data," *Journal of Hydrology.* 208(3–4): 187–193, Jul. *Amsterdam,* DOI: 10.1016/S0022-1694(98)00155-3.

[15] G. E. A. P. A. Batista and M. C. Monard, 2003, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence*, 17(5–6): 519–533, DOI: 10.1080/713827181.

[16] A. Aieb, K. Madani, M. Scarpa, B. Bonacorso, and K. Lefsih, 2019, "A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria," *Heliyon*, 5(2): e01247, DOI: 10.1016/J.HELIYON.2019.E01247.

[17] S. Epps *et al.*, 2008 "School Psychology Book Series Working With Parents of Noncompliant Children: A Guide to Evidence-Based Parent Training for Practitioners and Students.", *American Psychological Association (APA).*

[18] A. S. Suchithra and S. Agarwal, 2020. *IDF Curve Generation for Historical Rainfall Events Proceeding of National Conference on Emerging Trends in Civil Engineering IDF Curve Generation for Historical Rainfall Events*. Retrieved October 2025.

[19] R. S. V. Teegavarapu and V. Chandramouli, 2005, "Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records," Journal of Hydrology. 312(1–4): 191–206. *Amsterdam*, DOI: 10.1016/J.JHYDROL.2005.02.015.

[20] I. A. Nalder and R. W. Wein, 1998, "Spatial interpolation of climatic Normals: test of a new method in the Canadian boreal forest," *Agricultural and Forest Meteorology*, 92(4): 211–225, DOI: 10.1016/S0168-1923(98)00102-6.

[21] G. Y. Lu and D. W. Wong, 2008, "An adaptive inverse-distance weighting spatial interpolation technique," *Computers & Geosciences.* 34(9): 1044–1055. DOI: 10.1016/J.CAGEO.2007.07.010.

[22] S. S. P. Shen, P. Dzikowski, G. Li, and D. Griffith, 2001, "Interpolation of 1961-97 Daily Temperature and Precipitation Data onto Alberta Polygons of Ecodistrict and Soil Landscapes of Canada." Retrieved October 2025.