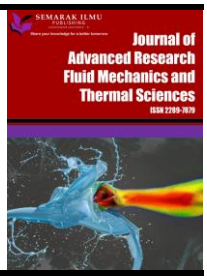
SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Fluid Mechanics and Thermal Sciences

Journal homepage:
https://semarakilmu.com.my/journals/index.php/fluid_mechanics_thermal_sciences/index
ISSN: 2289-7879



Assessing the Suitability of Probability Distribution Models for Streamflow Analysis in Peninsular Malaysia

Jing Lin Ng[1,*], Norashikin Ahmad Kamal[1], Nur Ilya Farhana Md Noh[1], Jin Chai Lee[2], Ruzaimah Razman[3], Jurina Jaafar[1], Deepak Tirumishi Jada[1], Majid Mirzaei[4]

[1] School of Civil Engineering, College of Engineering, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia
[2] Department of Civil Engineering, Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur, Malaysia
[3] Faculty of Civil and Built Environment, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia
[4] Department of Environmental Science and Technology, University of Maryland, College Park, MD 20742, United States

ARTICLE INFO

Article history:
Received 7 August 2024
Received in revised form 4 November 2024
Accepted 14 November 2024
Available online 30 November 2024

Keywords:
Probability distribution; streamflow analysis; goodness-of-fit tests; GEV distribution; Peninsular Malaysia

ABSTRACT

Streamflow analysis is indeed a need for Malaysia which always threaten by extreme streamflow events including flooding. However, the determination of the most suitable probability distribution for streamflow analysis has been the major concern and received considerable critical attention recently. The purpose of this study is to analyse the performances of several distributions in fitting the streamflow data collected form 11 streamflow stations in Peninsular Malaysia. and select the best performed distribution eventually. Normal, Generalized Extreme Value (GEV), three-parameter Gamma (G3), two-parameter Gamma (G2), three-parameter Weibull (W3), two-parameter Weibull (W2), three-parameter Log-Normal (LN3) and two-parameter Log-Normal (LN2) were used in this study to fit the maximum, minimum, and mean streamflow with monthly and seasonal time scales by using Easyfit software. The performances of distributions were evaluated through constructing Probability-Probability (P-P) Plot, Cumulative Distribution Function (CDF) and conducting goodness-of-fit tests including Chi-square (X^2) Test, Anderson-Darling (AD) Test and Kolmogorov-Smirnov (KS) Test. All the results were tabulated for comparison and the most suitable probability distribution was selected. Overall, the results indicated that the GEV distribution was the best fit distribution for most of the stations and most of the data series or time scales, while the LN3 distribution appeared to be the second-best distribution. These findings add to a growing body of literature on the selection of probability distribution especially for the streamflow analysis in Peninsular Malaysia. Besides, these findings provide significant information and practical knowledge in supporting the future development of streamflow-related plans or management including the flood and drought mitigation plans.

1. Introduction

Streamflow is the amount of water that flows in a stream or river over a specific time period. It is a critical resource for human being as well as the environment which directly affects stream

* *Corresponding author.*
E-mail address: jinglin.ng@uitm.edu.my

https://doi.org/10.37934/arfmts.124.2.175191

ecosystems. Extreme streamflow in both magnitude and frequency may lead to unfavourable hydrological events such as flood and drought. As one of the efforts to address this issue, streamflow analysis is performed to understand the characteristics and patterns of the streamflow. This would allow the users or researchers to estimate and describe the streamflow data accurately, and eventually come out with suitable hydrological planning.

During the streamflow analysis, the parameters necessary to suit the probability distribution must be estimated. Plenty of parameter estimation methods are available in the market. Previous studies on parameter estimation methods for the streamflow analysis have primarily concentrated on L-moments (LM) method, LH-moments (LHM) method, methods of moments (MOM) and maximum likelihood estimation (MLE). The findings of Zadeh *et al.,* [18] as well as those of Khan *et al.,* [15] showed that the LM method produced the lowest bias estimates when the sample size and skewness of the data were small. On the other hand, the results from the study of Piyapatr *et al.,* [22] showed that the LHM method produced better result in high quantile estimation as compared to LM method. This idea was in an agreement with the finding of Shabri [26] which concluded that LHM was better and more reliable to be used for the upper part distributions' estimation. Besides, Hasan [13] found that MOM method was the most efficient method in his study to analyse the flood frequencies in the northeast of Iraq since it was simpler to derive. While for the MLE, it had shown to be very accurate in estimating the parameters of Generalized Pareto Distribution (GPD) in the study of Gharib *et al.,* [12]. Additionally, when the sample size was large, the estimation provided by MLE method was more accurate when compared to the LM and maximum product of spacing (MPS) methods [15]. Taken together all these previous findings, the suitability and accuracy of the parameter estimation method are strongly depending on the characteristics of the data including the size, skewness and existence of outliers.

Next, probability distribution models are essential in describing streamflow data, mainly for extreme flood investigations, drought analysis as well as the studies on water system management. Generally, probability distribution is a statistical function that provides the probabilities of every possible value for a random variable within a particular range [14]. Similar to parameter estimation methods, plenty of probability distribution models are available to be used. Selecting the most suitable probability distribution model to a collection of regionally gathered data is always a challenging task among researchers. Previous studies on parameter estimation methods in streamflow analysis have primarily concentrated on three-parameters lognormal (LN3), log-Pearson type III (LP3) and generalised extreme value (GEV). The findings from the studies of Langat *et al.,* [17] and Eris *et al.,* [9] found that the LN3 distribution were able to described the maximum and minimum flows well. Also, the LN3 distribution was selected to be the best distribution in describing the maximum streamflow in Air Putih [8]. Then, from the findings of Bhat *et al.,* [3] in analysing the flood frequency of River Jhelum, LP3 had shown to be the better distribution model with smaller standard deviation and hence, more reliable and accurate. While for the GEV distribution, Chikobvu and Chifurira [6] found that GEV had a good match to the minimum annual average rainfall in Zimbabwe. This finding is consistent with those of Eris *et al.,* [9] who concluded that GEV was the most effective function to describe the streamflow data at Seyhan Basin. Additionally, the GEV distribution was shown to be the most effective functions in modelling Tana River's yearly minimum and mean flows [17]. Considering all these findings, the choice of the most suitable distribution is strongly influenced by geographic and climatic conditions, along with the characteristics of collected streamflow data. Therefore, a detailed statistical analysis, considering these factors, is essential for accurately selecting appropriate distribution models in a specific study area.

Floods and droughts are natural or environmental phenomena that commonly occurred in Peninsular Malaysia. These two phenomena happened due to unexpected high and low streamflow

respectively. Things get worse when these environmental phenomena brought lots of negative impacts to the environment and the human being including loss of life, stoppage of water supply, damage of personal assets, damage of public assets and damage of infrastructure as well as decrement of the agricultural yield. For instance, in December 2014, several states in Peninsular Malaysia experienced severe flooding in which Kelantan is the worst affected state. More than 200,000 people were affected with 21 people died in this disaster [5]. Hence, streamflow analysis needs to be carried out for developing proper hydrological plans as the response to these issues. On the other hand, several distributions are available with different parameter estimation procedures, merits, and suitability to fit in different types of data, causing the selection of a suitable statistical probability distribution in streamflow analysis to be a challenging task. Therefore, this study aims to investigate and evaluate the performances of several probability distributions in fitting the streamflow data in Peninsular Malaysia, and eventually select the most suitable probability distribution model to be used for the future streamflow analysis in Peninsular Malaysia. With this, the negative effects of the issues stated above are expected to be minimised.

2. Methodology

2.1 Study Area and Description

Peninsular Malaysia which also known as West Malaysia, is situated along the tropics with latitude in the range from 1°N to 7°N while longitude is in between 99°E to 105°E. It covers an area of 130,598 km^2 which contains mountainous, wetlands and coastal zones. One of the significant topographic characteristics of Peninsular Malaysia is the existence of backbone in which rippling mountains are at the centre and they are bordered by relatively flat coast from three sides (east, west, and south). Based on Köppen climate classification, Peninsular Malaysia is classified as a tropical country with excess rainfall ranging from 1950mm to 4000mm annually, uniform temperature ranging from 23 °C to 32 °C and constantly high humidity [19,24].

In general, the climate of Peninsular Malaysia is highly influenced by seasonal wind flow during two different monsoons, namely Southwest Monsoon (SWM) and Northeast Monsoon (NEM) [21]. NEM which normally happens from November to February and the eastern parts of Peninsular Malaysia tend to experience heavy rainfall and eventually flooding. The western parts are mostly free from this influence because they are well protected by the mountainous topography. SWM normally happens from May to August and relatively drier period will be experienced in the whole Peninsular Malaysia.

The streamflow data recorded from eleven stations at Peninsular Malaysia were acquired from Malaysian Meteorological Department (MMD). The data consisted of annual streamflow amount for a duration of 30 years in every station with exception at Station Sungai Kepis (25 years) and Station Sungai Arau (22 years) due to the incompleteness of the data in a continuous 30-years-period. Prior to the software analysis, the daily streamflow data was summed up to monthly and seasonal data. In addition, the monthly and seasonal data were computed for the respective maximum, minimum and mean streamflow. As a result, there would have eight data series to work on including the monthly streamflow, maximum monthly streamflow, minimum monthly streamflow, mean monthly streamflow, seasonal streamflow, maximum seasonal streamflow, minimum seasonal streamflow and lastly mean seasonal streamflow. The geographical locations for all the 11 stations are shown in Figure 1, while the detail list of stations involved were provided in Table 1.

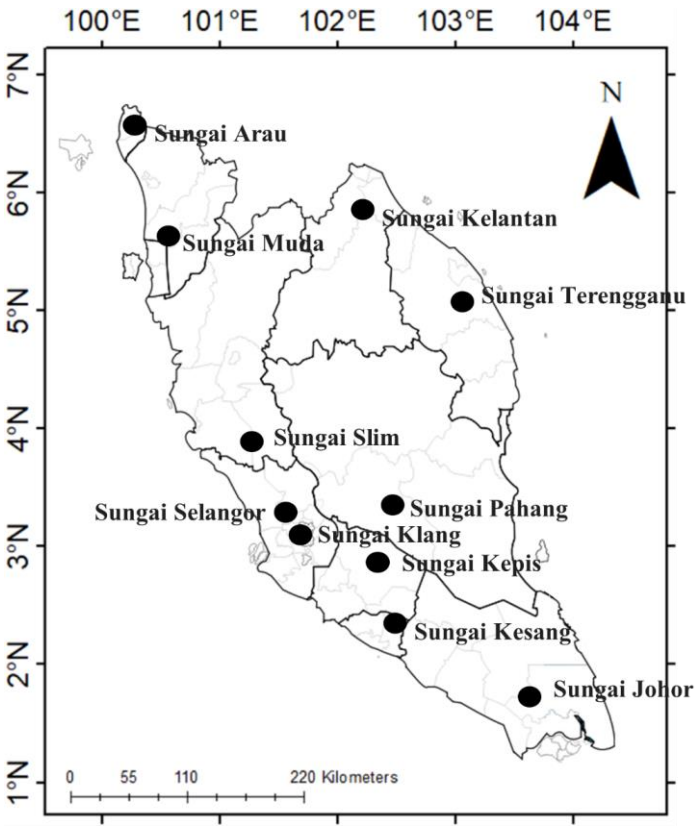


Fig. 1. Geographical location of Peninsular Malaysia and the selected meteorological stations involved in this study

Table 1
Detail list of meteorological stations involved in this study

Station Code	Station Name	Study Period	Duration (years)	Latitude	Longitude
1737451	Sungai Johor at Rantau Panjang	1990~2019	30	01° 46' 50"N	103° 44' 45"E
5606410	Sungai Muda at Jambatan Syed Omar	1990~2019	30	05° 36' 35"N	100° 37' 35"E
1737451	Sungai Johor at Rantau Panjang	1990~2019	30	01° 46' 50"N	103° 44' 45"E
5606410	Sungai Muda at Jambatan Syed Omar	1990~2019	30	05° 36' 35"N	100° 37' 35"E
5721442	Sungai Kelantan at Jambatan Guillemard	1990~2019	30	05° 45' 45"N	102° 09' 00"E
2224432	Sungai Kesang at Chin Chin	1990~2019	30	02° 17' 25"N	102° 29' 35"E
2723401	Sungai Kepis at Jambatan Kayu Lama	1984~2008	25	02° 42' 20"N	102° 21' 20"E
3424411	Sungai Pahang at Temerloh	1990~2019	30	03° 26' 40"N	102° 25' 45"E
3814416	Sungai Slim at Slim River	1990~2019	30	03° 49' 35"N	101° 24' 40"E
6503401	Sungai Arau at Ladang Tebu Felda	1998~2019	22	06° 30' 10"N	100° 21' 05"E
3414421	Sungai Selangor at Rantau Panjang	1990~2019	30	03° 24' 10"N	101° 26' 35"E
5130432	Sungai Terengganu at Kampung Tanggol	1990~2019	30	05° 08' 15"N	103° 02' 45"E
3116430	Sungai Klang at Jambatan Sulaiman	1990~2019	30	03° 08' 20"N	101° 45' 50"E

2.2 Data Screening and Estimating Missing Data

There were some missing data in the streamflow data set collected which may be due to technical issue, human error or natural causes. As recommended in the study of Ismail *et al.,* [16], Normal Ratio Method was applied to estimate those missing data by using the equation below:

$$P_m = \frac{1}{n}\sum_{i=1}^{n}\frac{N_m}{N_i}P_i \quad (1)$$

where P_m represents the assumed value of the missing data at the target station, n represents the number of nearby stations, P_i represents the collected data at i-th nearby stations, N_m represents the annual streamflow amount at the target station and N_i represents the annual streamflow amount at the i-th nearby station.

2.3 Selection of Parameter Estimation Method

The Maximum Likelihood Estimation (MLE) method was chosen for parameter prediction in this study due to its desirable properties, including high consistency and efficiency in handling historically collected streamflow data and accuracy with large sample sizes. MLE maximizes the likelihood of sample data by finding values that maximize the likelihood function (LF), representing the probability of observed sample series occurrences. For independent events, the LF is obtained by multiplying the probability density functions of observed data. Using the derivative of the logarithm of the likelihood function (LLF) is a common practice for convenience [17].

The LF and LLF for a two-parameter probability distribution were specified as below:

$$LF = \prod_{i=1}^{n} F\ (x_i; \gamma, \alpha) \quad (2)$$

$$LLF = \sum_{i=1}^{n} \ln[F\ (x_i; \gamma, \alpha)] \quad (3)$$

While the LF and LLF for a three-parameter probability distribution were specified as below:

$$LF = \prod_{i=1}^{n} F\ (x_i; \gamma, \alpha, \beta) \quad (4)$$

$$LLF = \sum_{i=1}^{n} \ln[F\ (x_i; \gamma, \alpha, \beta)] \quad (5)$$

where n is the number of observations, F(x) is the cumulative distribution function (CDF), while α, γ, and β indicates the scale parameter, location parameter and shape parameter respectively.

2.4 Selection of Probability Distribution Models

This study evaluates and compares eight recommended probability distributions—Normal, Generalized Extreme Value (GEV), three-parameter Gamma (G3), two-parameter Gamma (G2), three-parameter Weibull (W3), two-parameter Weibull (W2), three-parameter Log-Normal (LN3), and two-parameter Log-Normal (LN2)—for streamflow analysis. Evaluation is based on their Cumulative Distribution Function (CDF) and Probability Density Function (PDF).

(a) Normal Distribution

Normal distribution is a probability distribution with a bell-shaped curve that graphically interpretating the data [1]. It consists of two parameters which are variance and mean. For a stochastic variable x with normal distribution, the expression of PDF is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \quad (6)$$

while its expression of CDF is:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]\right) dx \quad (7)$$

where σ represents the standard deviation, σ^2 represents the variance and μ represents the mean of normal distribution.

(b) Generalized Extreme Value (GEV) Distribution

Generalized Extreme Value (GEV) Distribution is a probability distribution that model the maxima of long sequences of stochastic variables [4]. It consists of three parameters which are location, scale and shape parameters. For a stochastic variable x with GEV distribution, the expression of PDF is:

$$f(x) = \alpha^{-1} exp[-(1-\beta)y - \exp(-y)] \quad (8)$$

with conditions:

$$y = -\beta^{-1} \log\left\{1 - \frac{\beta(x-\gamma)}{\alpha}\right\}, \beta \neq 0 \quad (9)$$

$$y = \frac{(x-\gamma)}{\alpha}, \beta = 0 \quad (10)$$

while its expression of CDF is:

$$F(x) = \exp[-exp - y] \quad (11)$$

where α, γ, and β indicates the scale parameter, location parameter and shape parameter respectively.

(c) Two-parameter Gamma (G2) Distribution

Two-parameter Gamma (G2) Distribution is a probability distribution that widely used in environmental analysis, clinical trials, reliability analysis and signal processing [25]. It consists of two parameters which are scale and shape parameters. For a non-negative variable x with G2 distribution, the expression of PDF is

$$f(x) = \frac{x^{\beta-1}}{\alpha^{\beta}\Gamma(\beta)} \exp\left(-\frac{x}{\alpha}\right) \quad (12)$$

with conditions

$$\Gamma(\beta) = \int_0^\infty x^{\beta-1} e^{-x}\, dx \tag{13}$$

while its expression of CDF is

$$F(x) = \frac{\Gamma_{x/\alpha}(\beta)}{\Gamma(\beta)} \tag{14}$$

where α indicates the scale parameter, β indicates the shape parameter and Γ indicates the gamma function.

(d) Three-parameter Gamma (G3) Distribution

Three-parameter Gamma (G3) Distribution is a type of Gamma distribution that has an additional parameter aside from the shape and scale parameter in G2 distribution, which is location parameter [25]. For a non-negative and stochastic variable x with G3 distribution, the expression of PDF is:

$$f(x) = \frac{(x-\gamma)^{\beta-1}}{\alpha^{\beta}\Gamma(\beta)} \exp\left(-\frac{x-\gamma}{\alpha}\right) \tag{15}$$

with conditions

$$\Gamma(\beta) = \int_0^\infty x^{\beta-1} e^{-x}\, dx \tag{16}$$

while its expression of CDF is

$$F(x) = \frac{\Gamma_{(x-\gamma)/\alpha}(\beta)}{\Gamma(\beta)} \tag{17}$$

where Γ indicates the gamma function, β indicates the shape parameter, γ indicates the location parameter and α indicates the scale parameter.

(e) Two-parameter Weibull (W2) Distribution

Weibull Distribution is a high flexibility probability distribution which able to model continuous data, right-skewed data and left-skewed data [28]. Scale and shape parameters are the parameters for two-parameter Weibull (W2) Distribution. For a stochastic variable x with W2 distribution, the expression of PDF is:

$$f(x) = \left(\frac{\beta}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta-1} exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right] \tag{18}$$

while its expression of CDF is:

$$F(x) = 1 - exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right] \tag{19}$$

where β and α indicates the shape and scale parameters respectively.

(f) Three-parameter Weibull (W3) Distribution

Three-parameter Weibull (W3) Distribution is a type of Weibull distribution that has an additional parameter aside from the shape and scale parameter in W2 distribution, which is location parameter [28]. For a stochastic variable x with W3 distribution, the expression of PDF is:

$$f(x) = \left(\frac{\beta}{\alpha}\right)\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} exp\left[-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right] \tag{20}$$

while its expression of CDF is

$$F(x) = 1 - exp\left[-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right] \tag{21}$$

where β indicates the shape parameter, α indicates the scale parameter and γ indicates the location parameter.

(g) Two-parameter Log-Normal (LN2) Distribution

Log-Normal Distribution is a probability distribution that distributed the log of stochastic variable normally. Literally, LN2 consists of two parameters which shape and scale parameters. For a stochastic variable x with LN2 distribution, the expression of PDF is

$$f(x) = \frac{1}{x\beta_Y\sqrt{2\pi}} exp\left[-\frac{1}{2\beta_Y{}^2}(\ln(x) - \mu_Y)^2\right] \tag{22}$$

while its expression of CDF is

$$F(x) = \frac{1}{\beta_Y\sqrt{2\pi}}\int_0^x \left(\frac{1}{x} exp\left[-\frac{1}{2\beta_Y{}^2}(\ln(x) - \mu_Y)^2\right]\right) dx \tag{23}$$

where

$$\beta_Y = \sqrt{\ln\left[\frac{\beta^2+\mu^2}{\mu^2}\right]} \tag{24}$$

$$\mu_Y = \ln\left[\frac{\mu^2}{\sqrt{\beta^2+\mu^2}}\right] \tag{25}$$

$$\mu = \log(\alpha) \tag{26}$$

where β indicates the shape parameter, α indicates the scale parameter, σ^2 indicates the variance and μ indicates the mean of LN2 distribution.

(h) Three-parameter Log-Normal (LN3) Distribution

Three-parameter Log-Normal (LN3) Distribution is a type of Log-Normal distribution that has an additional parameter aside from the shape and scale parameter in LN2 distribution, which is location parameter. For a stochastic variable x with LN3 distribution, the expression of PDF is:

$$f(x) = \frac{1}{(x-\gamma)\beta_Y\sqrt{2\pi}} exp\left[-\frac{1}{2\beta_Y{}^2}(\ln(x-\gamma)-\mu_Y)^2\right] \quad (27)$$

while its expression of CDF is:

$$F(x) = \frac{1}{\beta_Y\sqrt{2\pi}}\int_0^x \left(\frac{1}{x-\gamma} exp\left[-\frac{1}{2\beta_Y{}^2}(\ln(x-\gamma)-\mu_Y)^2\right]\right) dx \quad (28)$$

where

$$\beta_Y = \sqrt{\ln\left[\frac{\beta^2+\mu^2}{\mu^2}\right]} \quad (29)$$

$$\mu_Y = \ln\left[\frac{\mu^2}{\sqrt{\beta^2+\mu^2}}\right] \quad (30)$$

$$\mu = \log(\alpha) \quad (31)$$

where α indicates the scale parameter, γ indicates the location parameter, β indicates the shape parameter, σ^2 indicates the variance and μ indicates the mean of LN3 distribution.

Evaluation of Performances – Application of Goodness-of-fit (GoF) Tests

The difference between the theoretical distributions with empirical distributions was computed by three GoF Tests namely Kolmogorov-Smirnov (KS) Test, Anderson-Darling (AD) Test and Chi-Square (X^2) Test. At a 95% significance level of confidence, a value of test statistic from all three tests obtained that exceed the respective critical value would cause the rejection on the previously done hypothesis.

(a) Kolmogorov–Smirnov (KS) Test

Kolmogorov–Smirnov (KS) Test is used to compare the distribution generated through the given data with a hypothetical probability distribution through determining the highest vertical gap between the theoretical and empirical results of CDF [7]. This highest vertical distance is termed as KS test statistic. With n as the sample size and with data arrangement following an ascending order, $X_1 < X_2 < \ldots < X_n$, the expression of KS test statistic for each ordered value is:

$$S_n(x) = 0\ ;if\ X < X_1 \quad (32)$$

$$S_n(x) = \frac{k}{n}\ ;if\ X_k \leq X < X_{k+1} \quad (33)$$

$$S_n(x) = 1\ ;if\ X > X_n \quad (34)$$

where $S_n(x)$ represents the empirical cumulative distribution function (CDF) and k is the ordered of the data set.

$$D_n = max|F_x(x) - S_n(x)| \tag{35}$$

$$P(D_n \leq D_n^{\alpha}) = 1 - \alpha \tag{36}$$

where α represents the significance level and D_n^{α} represents the critical value.

(b) Anderson–Darling (AD) Test

Anderson–Darling (AD) Test was proposed by Anderson and Darling in year 1954 to refine the KS test by emphasizing more onto the tails of distribution or the outliers [11]. Hence, AD test is more suitable to evaluate the models which can best fit the maximum streamflow as compared to KS test. With n as the sample size and with data arrangement following an ascending order, X1 < X2 < … < Xn, the AD test statistic is expressed as below:

$$A^2 = -\sum_{i=1}^{n}\left[\frac{(2i-1)\{\ln F_X(x_i)+\ln[1-F_X(x_{n+1-i})]\}}{n}\right] - n \tag{37}$$

where A^2 represents the AD test statistic and $F_X(x_i)$ represents the CDF of the specified distribution with i = 1, 2, …, n.

(c) Chi Square (X^2) Test

Chi Square (X^2) Test is used to evaluate the correlation between two or more different categorical variables even for nonnumeric variables as well as to examine the fitness of the observed data to the expected distribution [20,27]. The steps in calculating the Chi-Square goodness of fit test consist of arranging and assigning the number of observations, n into a set of M cells, followed by mathematical calculations based on the formula below:

$$X^2 = \sum_{i=1}^{M}\frac{(O_i-E_i)^2}{E_i} \tag{38}$$

where O_i represents the observed frequency in the i-th cell, E_i represents the predicted frequency in the same i-th cell and M represents the number of intervals.

Selection of the Best Probability Distribution Model

Based on the tabulations, the three grading results from all the GoF tests for each of the station were added up to form a total grading value. Probability distribution model with the lowest total grading value was identified to be the most suitable probability distribution model to be used in describing the streamflow data in Peninsular Malaysia.

3. Results

3.1 Goodness f Fit Using Assessment-Based Graphs

After fitting the data with distributions in EasyFit (data analysis software), several graphical functions can be generated, while two of them namely Probability-Probability (P-P) Plot and Cumulative Distribution Function (CDF) were generated in this study to assist in evaluating the performances of all distributions. Generally, the results obtained from all 11 stations showed high similarity in terms of the graph patterns. Figure 2 and Figure 3 show the CDF and P-P Plot of six different stations for the maximum monthly streamflow data.

For P-P Plot, it is worth to mention that an approximately linear plot in P-P Plot indicates that the distribution best fits the data sample, and the suitability of several fitted distributions can be determined through direct visualization on a P-P plot. By referring to the Figure 2, the plots created from the data points of the GEV, LN3 and W3 distributions were close to the linear line, indicating good performance in fitting the maximum monthly streamflow data. On the contrary, it is apparent that the data points of the normal distribution were scattered far away from the linear line for majority of stations. This indicated that it exhibited poor performance. Besides, as the P-P plots of most of the stations showed, the LN2, W2, G2 and G3 distributions were the moderately well in fitting the maximum monthly streamflow data because the gaps between their data points and the linear line were moderately small.

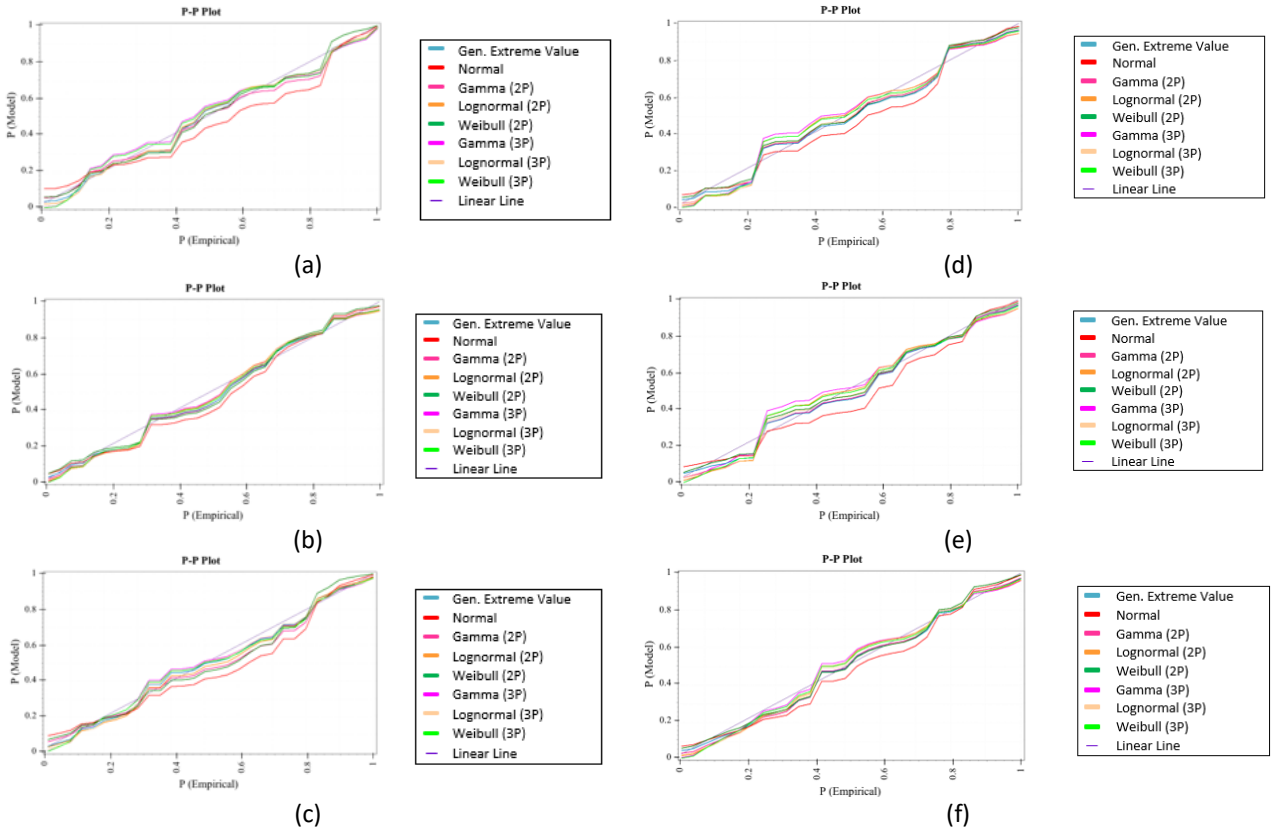


Fig. 2. P-P Plots of maximum monthly streamflow fitting by using the eight selected distribution models at (a) Johor, (b) Kedah, (c) Kelantan, (d) Melaka, (e) Negeri Sembilan and (f) Pahang

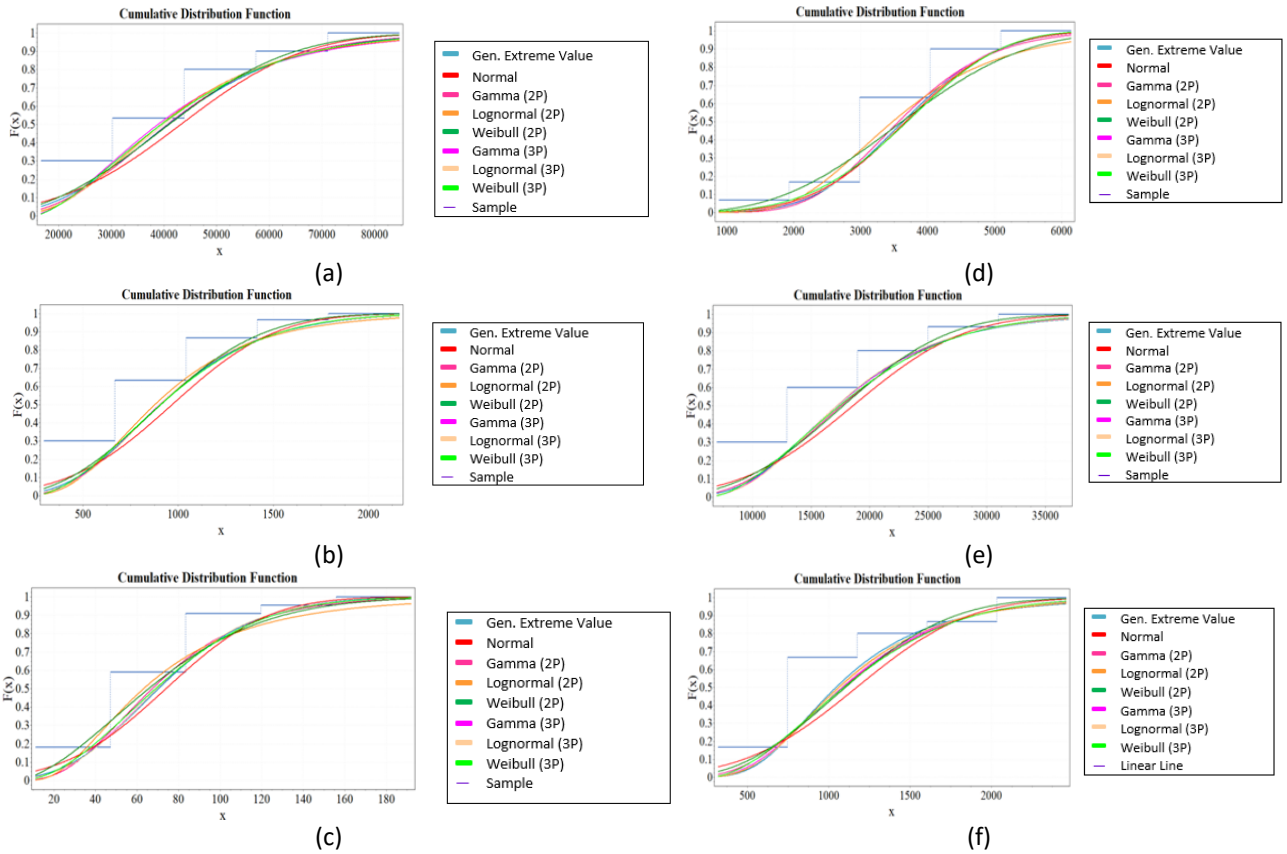


Fig. 3. CDF graph of maximum monthly streamflow fitting by using the eight selected distribution models at (a) Pahang, (b) Perak, (c) Perlis, (d) Selangor, (e) Terengganu and (f) Wilayah Persekutuan Kuala Lumpur

For CDF, only continuous distributions were used in this study, and hence the empirical CDF displayed will be in a stepped discontinuous line while the theoretical CDF will display as continuous curve. The smaller the gaps between the empirical (sample) and theoretical CDF, the better the performance of fitting the data sample. As can be seen from the Figure 3, the CDF curves of the GEV, LN3 and W3 distributions were the closet to the step lines of empirical CDF, indicating good performance in fitting the maximum monthly streamflow data. On the contrary, it is apparent that the CDF curves created by the normal and LN2 distributions were far away from the steps lines of empirical CDF for majority of stations, indicating that they offered poor performance. Next, the W2, G2 and G3 distributions showed similar gaps to the step lines which were at the average level. Hence, they performed moderately in fitting the maximum monthly streamflow data.

3.2 Goodness of Fit Using GoF Test-Based Analysis

As mentioned in the methodology, the GoF tests' statistics were tabulated for comparison through the algorithm: The lower the value of GoF tests' statistic, the better the result. For every test, each distribution model was assigned with a grade number ranging from 1 to 8, in which grade 1 representing the best-performed distribution model, and grade 8 representing the worst-performed distribution model. All the grading results from the three GoF tests for every station were added up to form a total grading value as shown in Table 2 and Table 3. The suitability of the eight probability distributions were examined and ranked through the algorithm: The lower the grading value, the more suitable the distribution.

Based on the GoF analysis, the GEV and LN3 distributions were found to be the most suitable distributions with high accuracy in fitting all the data series for this research while the normal

distribution was the least fit distribution. The remaining distributions which were LN2, G2, G3, W2 and W3 had the moderate grading results based on the three GoF tests, indicating that they were the moderate distributions for streamflow fitting in Peninsular Malaysia. Additionally, it is interesting to note that the grading results of a three-parameters distribution of one kind was better than the two-parameters distribution of the same kind in most of the data series.

Table 2
Ranking of distributions based on the total goodness-of-fit grading (Month)

Types of Distribution	Monthly Streamflow		Maximum Monthly Streamflow		Minimum Monthly Streamflow		Mean Monthly Streamflow	
	Total Grade	Rank	Total Grade	Rank	Total Grade	Rank	Total Grade	Rank
Normal	229	8	223	8	176	7	182	8
GEV	62	1	88	1	76	1	101	1
G2	204	7	143	5	169	6	157	5
G3	134	4	141	4	136	4	164	7
W2	194	6	165	6	167	5	162	6
W3	175	5	133	3	133	3	152	4
LN2	116	3	168	7	202	8	149	3
LN3	78	2	133	2	130	2	121	2

Table 3
Ranking of distributions based on the total goodness-of-fit grading (Season)

Types of Distribution	Seasonal Streamflow		Maximum Seasonal Streamflow		Minimum Seasonal Streamflow		Mean Seasonal Streamflow	
	Total Grade	Rank	Total Grade	Rank	Total Grade	Rank	Total Grade	Rank
Normal	223	8	214	8	187	8	197	8
GEV	95	2	96	1	102	1	94	1
G2	170	6	151	5	164	5	144	4
G3	135	4	124	3	130	3	132	3
W2	209	7	187	7	174	6	192	7
W3	142	5	145	4	117	2	144	4
LN2	129	3	155	6	182	7	153	6
LN3	84	1	118	2	134	4	127	2

4. Discussion

The results of this study indicate that, based on both graphical assessments (P-P Plot and CDF) and Goodness of Fit (GoF) tests analysis, the Generalized Extreme Value (GEV) distribution was the best fit distribution model that well representing most of the streamflow data series in Peninsular Malaysia except for the seasonal streamflow data that had three-parameter Lognormal (LN3) distribution as its best fit. It is interesting to note that the GEV and LN3 distributions appeared to be the top two best fit distributions in all the eight data series considered in this study. There are several possible explanations for this result. First, the GEV distribution is a complex distribution that consists of three parameters. It offers a higher flexibility in capturing, considering and modelling more climatic characteristics that subsequently results in having the ability to capture the data for extreme events [20]. Similarly, the LN3 distribution also consists of three parameters, causing it to be more flexible. Additionally, the LN3 distribution is robust to outliers or extreme values, allowing it to perform well in modelling the streamflow data in Peninsular Malaysia that contains extreme values due to the tropical climate. Next, the GEV distribution itself consists of three different kinds of extreme value

distributions including Weibull, Fréchet and Gumbel. This allows the GEV to provide a better analysis for different data series after comparing the performances of these three distributions. The findings are consistent with those of Ng *et al.,* [20] who found GEV distribution was the best fit probability distribution for the annual maximum rainfall at Kelantan, which is one of the states of Peninsular Malaysia. Additionally, these results also match those observed in earlier studies elsewhere in the world. For instance, Farooq *et al.,* [10] concluded that the GEV distribution was the best distribution to be used for flood frequency analysis in Pakistan since it provided the best estimation and fitting in all the four studied stations at Swat River, Pakistan. In Kenya, the finding of Langat *et al.,* [17] also found that the GEV distribution had the best performance in fitting the annual minimum and annual mean streamflow at Tana River, Kenya. Furthermore, the finding regarding on the suitability of the LN3 distribution is also in agreement with Badyalina's *et al.,* [2] findings which showed that the LN3 distribution was the optimum distribution in fitting the yearly peak flow of Segamat River at Johor. Therefore, the GEV distribution is suggested to simulate the streamflow data in Peninsular Malaysia.

After going through both the graphical and GoF Test assessments, the results indicated that the normal distribution was the least fit distribution in most of the streamflow data series in Peninsular Malaysia except for the minimum monthly streamflow data that had two-parameter Lognormal (LN2) distribution as its least fit. However, the normal distribution was still the second least fit for minimum monthly streamflow which confirmed that it was the overall least fit distribution in this present study. It seems possible that these results are due to the characteristic of streamflow in Peninsular Malaysia which is highly stochastic. As mentioned in the literature review, Peninsular Malaysia opposed tropical climate and big variability of rainfall patterns. The northeast monsoon season (starts from November to February) and the southwest monsoon season (starts from May to August) would be experienced every year. Hence, extreme streamflow might occur, causing the streamflow data to be highly spread and not normally distributed. Similar concept can be applied to the LN2 distribution as well since it does not include the location parameter, and it considered the logarithm of the random variable to be normally distributed. Hence, it is not suggested to be used for streamflow data fitting in Peninsular Malaysia. These findings further support the conclusion of Langat *et al.,* [17] which found that the normal and LN2 distributions were the least fitting.

Results showed that the G2, G3, W2 and W3 distributions were the moderate distributions with average fitting accuracy in all the streamflow data series. The rankings of these distributions were different in different data series. It is difficult to explain this result, but it might be related to the idea and finding of Langat *et al.,* [17] in which they found that different distributions may be suitable for different data series such as the minimum streamflow, mean streamflow, and maximum and maximum streamflow even the data were collected from the same site since the frequency of each data series are unique [23]. However, it is worth to note that the performance of a three-parameters distribution of one kind was better than the two-parameters distribution of the same kind in most of the data series in this study. This finding can be explained by using the same concept mentioned before in which more parameters involved will provide a better fitting performance as it has more flexibility and more ability in capturing climatic characteristics [20].

The combination of this study's findings provides some support for the potential stakeholders including the local authority and engineers by providing a guidance in selecting the probability distribution for estimating the streamflow based on their suitability. The findings of maximum streamflow have important implications for developing water systems infrastructure and flood mitigation strategies. Besides, the findings of minimum and mean streamflow help in understanding the hydrological drought and developing the water resource management for irrigation as well as agriculture. However, more research on this topic needs to be undertaken before the plans and strategies to deal with the streamflow are more clearly understood and well executed.

5. Conclusion

The purpose of the current study was to investigate the most suitable probability distribution models for streamflow in Peninsular Malaysia. A total of eight distributions including the normal, GEV, G2, G3, W2, W3, LN2 and LN3 distributions were implemented to fit the streamflow data. Once all the distributions done fitting the data, the performance of each distribution in describing the streamflow data was evaluated graphically (P-P Plot and CDF) and statistically through Goodness of Fit (GoF) Tests (KS test, AD Test and Chi-squared Test). Following this, the overall results were tabulated and ranked, and the best suited probability distribution models can be determined.

In conclusion, the study identifies the Generalized Extreme Value (GEV) distribution as the best fit for representing most streamflow data series in Peninsular Malaysia. However, for seasonal streamflow data, the three-parameter Lognormal (LN3) distribution is the most suitable. These distributions consistently yield the lowest grading values for most stations and data series. The study highlights the enhanced flexibility of three-parameter distributions, better capturing climatic characteristics compared to two-parameter distributions. These findings are crucial for guiding the selection of suitable probability distributions in streamflow estimation worldwide. Additionally, they provide insights into the hydrological climates and characteristics of Peninsular Malaysia, aiding in flood and drought mitigation, as well as water resource management for irrigation and agriculture.

To improve the accuracy in future findings, below are some recommendations to be considered. First, involve the streamflow stations in East Malaysia and increase the numbers of stations considered so that findings can represent the entire Malaysia as a whole. Next, it is suggested to select the stations with complete data set with at least 30-year study period to ensure the precision and reliability of the results. Furthermore, the candidates of the distributions involved is suggested to be increased especially with high parameter distributions to maximise the comparison of their performances in choosing the best fitting distribution for streamflow data.

Acknowledgement

The authors would like to express gratitude to School of Civil Engineering, College of Engineering, Universiti Teknologi MARA (UiTM) for the financial support. The authors also would like to thank Kurita Water and Environment Foundation (KWEF) for providing funding under KURITA Overseas Research Grant 2024 (Grant number: 24Pmy098). Special appreciation goes to the Malaysia Meteorological Department (MMD) for providing the weather data.

References

[1] Atangana, Abdon, and José Francisco Gómez-Aguilar. "A new derivative with normal distribution kernel: Theory, methods and applications." *Physica A: Statistical Mechanics and Its Applications* 476 (2017): 1-14. https://doi.org/10.1016/j.physa.2017.02.016

[2] Badyalina, Basri, Nurkhairany Amyra Mokhtar, Nur Amalina Mat Jan, Nor Hidayah Hassim, and Haslenda Yusop. "Flood frequency analysis using L-moment for Segamat river." *Matematika* (2021): 47-62. https://doi.org/10.11113/matematika.v37.n2.1332

[3] Bhat, M. Sultan, Akhtar Alam, Bashir Ahmad, Bahadur S. Kotlia, Hakim Farooq, Ajay K. Taloor, and Shabir Ahmad. "Flood frequency analysis of river Jhelum in Kashmir basin." *Quaternary International* 507 (2019): 288-294. https://doi.org/10.1016/j.quaint.2018.09.039

[4] Boudrissa, Naima, Hassen Cheraitia, and Lotfi Halimi. "Modelling maximum daily yearly rainfall in northern Algeria using generalized extreme value distributions from 1936 to 2009." *Meteorological Applications* 24, no. 1 (2017): 114-119. https://doi.org/10.1002/met.1610

[5] Buslima, F. S., R. C. Omar, Tajul Anuar Jamaluddin, and Hairin Taha. "Flood and flash flood geo-hazards in Malaysia." *International Journal of Engineering & Technology* 7, no. 4.35 (2018): 760-764. https://doi.org/10.14419/ijet.v7i4.35.23103

[6] Chikobvu, Delson, and Retius Chifurira. "Modelling of extreme minimum rainfall using generalised extreme value

distribution for Zimbabwe." *South African Journal of Science* 111, no. 9-10 (2015): 01-08. https://doi.org/10.17159/sajs.2015/20140271
[7] Conover, William Jay. *Practical nonparametric statistics*. Vol. 350. John Wiley & Sons, 1999.
[8] Deraman, Wan Husna Aini Wan, Noor Julailah Abd Mutalib, and Nur Zahidah Mukhtar. "Determination of return period for flood frequency analysis using normal and related distributions." In *Journal of Physics: Conference Series*, vol. 890, no. 1, p. 012162. IOP Publishing, 2017. https://doi.org/10.1088/1742-6596/890/1/012162
[9] Eris, Ebru, Hafzullah Aksoy, Bihrat Onoz, Mahmut Cetin, Mehmet Ishak Yuce, Bulent Selek, Hakan Aksu et al. "Frequency analysis of low flows in intermittent and non-intermittent rivers from hydrological basins in Turkey." *Water Supply* 19, no. 1 (2019): 30-39. https://doi.org/10.2166/ws.2018.051
[10] Farooq, Muhammad, Muhammad Shafique, and Muhammad Shahzad Khattak. "Flood frequency analysis of river swat using Log Pearson type 3, Generalized Extreme Value, Normal, and Gumbel Max distribution methods." *Arabian Journal of Geosciences* 11 (2018): 1-10. https://doi.org/10.1007/s12517-018-3553-z
[11] Farrell, Patrick J., and Katrina Rogers-Stewart. "Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test." *Journal of Statistical Computation and Simulation* 76, no. 9 (2006): 803-816. https://doi.org/10.1080/10629360500109023
[12] Gharib, Amr, Evan G. R. Davies, Greg G. Goss, and Monireh Faramarzi. "Assessment of the combined effects of threshold selection and parameter estimation of Generalized Pareto Distribution with applications to flood frequency analysis." *Water* 9, no. 9 (2017): 692. https://doi.org/10.3390/w9090692
[13] Hasan, Ihsan F. "Flood Frequency Analysis of Annual Maximum Streamflows at Selected Rivers in Iraq." *Jordan Journal of Civil Engineering* 14, no. 4 (2020).
[14] Hayes, Adam. "Probability Distribution: Definition, Types, and Uses in Investing." *Investopedia*. May 14, 2022. https://www.investopedia.com/terms/p/probabilitydistribution.asp#:~:text=A%20probability%20distribution%20depicts%20the,deviation%2C%20skewness%2C%20and%20kurtosis.
[15] Khan, Muhammad Shafeeq ul Rehman, Zamir Hussain, and Ishfaq Ahmad. "Effects of L-moments, maximum likelihood and maximum product of spacing estimation methods in using pearson type-3 distribution for modeling extreme values." *Water Resources Management* 35 (2021): 1415-1431. https://doi.org/10.1007/s11269-021-02767-w
[16] Ismail, Wan Norliyana Wan, Wan Zawiah Wan Zin, and Wan Ibrahim. "Estimation of rainfall and stream flow missing data for Terengganu, Malaysia by using interpolation technique methods." *Malaysian Journal of Fundamental and Applied Sciences* 13, no. 3 (2017): 214-218. https://doi.org/10.11113/mjfas.v13n3.578
[17] Langat, Philip Kibet, Lalit Kumar, and Richard Koech. "Identification of the most suitable probability distribution models for maximum, minimum, and mean streamflow." *Water* 11, no. 4 (2019): 734. https://doi.org/10.3390/w11040734
[18] Zadeh, Shabnam Mostofi, Martin Durocher, Donald H. Burn, and Fahim Ashkar. "Pooled flood frequency analysis: a comparison based on peaks-over-threshold and annual maximum series." *Hydrological Sciences Journal* 64, no. 2 (2019): 121-136. https://doi.org/10.1080/02626667.2019.1577556
[19] Ng, Jing Lin, Yuk Feng Huang, Sheng Kwan Tan, Jin Chai Lee, Nur Ilya Farhana Md Noh, and Siaw Yin Thian. "Comparative evaluation of various parameter estimation methods for extreme rainfall in Kelantan River Basin." *Theoretical and Applied Climatology* 155, no. 3 (2024): 1759-1775. https://doi.org/10.1007/s00704-023-04723-7
[20] Ng, J. L., S. Y. Yap, Y. F. Huang, NIF Md Noh, R. A. Al-Mansob, and R. Razman. "Investigation of the best fit probability distribution for annual maximum rainfall in Kelantan River Basin." In *IOP Conference Series: Earth and Environmental Science*, vol. 476, no. 1, p. 012118. IOP Publishing, 2020. https://doi.org/10.1088/1755-1315/476/1/012118
[21] Noor, Muhammad, Tarmizi Ismail, Eun-Sung Chung, Shamsuddin Shahid, and Jang Hyun Sung. "Uncertainty in rainfall intensity duration frequency curves of peninsular Malaysia under changing climate scenarios." *Water* 10, no. 12 (2018): 1750. https://doi.org/10.3390/w10121750
[22] Piyapatr, Busababodhin, Chiangpradit Monchaya, Phoophiwfa Tossapol, Jeong-Soo Park, Do-ove Manoon, and Guayjarernpanishk Pannarat. "LH-Moments of the Wakeby Distribution applied to Extreme Rainfall in Thailand." *Malaysian Journal of Fundamental and Applied Sciences* 17, no. 2 (2021): 166-180. https://doi.org/10.11113/mjfas.v17n2.2005
[23] Rahman, Ayesha S., Ataur Rahman, Mohammad A. Zaman, Khaled Haddad, Amimul Ahsan, and Monzur Imteaz. "A study on selection of probability distributions for at-site flood frequency analysis in Australia." *Natural Hazards* 69 (2013): 1803-1813. https://doi.org/10.1007/s11069-013-0775-y
[24] Ramli, Suzana, Siti Nur Nabillah Ibrahim, and Gusti Agung Putu Eryani. "Flood Discharge For Ungauged Catchment At Teriang River, Pahang By Using HEC-HMS." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 33, no. 3 (2023): 39-50. https://doi.org/10.37934/araset.33.3.3950
[25] Ramos, Pedro L., Dipak K. Dey, Francisco Louzada, and Eduardo Ramos. "On posterior properties of the two

parameter gamma family of distributions." *Anais da Academia Brasileira de Ciências* 93, no. suppl 3 (2021): e20190826. https://doi.org/10.1590/0001-3765202120190826

[26] Shabri, Ani. "Comparisons of the LH Moments and the L Moments." *Matematika* (2002): 33-43.

[27] Turhan, Nihan Sölpük. "Karl Pearson's Chi-Square Tests." *Educational Research and Reviews* 16, no. 9 (2020): 575-580. https://doi.org/10.5897/ERR2019.3817

[28] Wais, Piotr. "Two and three-parameter Weibull distribution in available wind power analysis." *Renewable Energy* 103 (2017): 15-29. https://doi.org/10.1016/j.renene.2016.10.041