Mathematical Sciences and Informatics Journal
Vol. 2, No. 2, Nov. 2021, pp. 26-34
http://www.mijuitmjournal.com
DOI : 10.24191/mij.v2i2.13776

Modelling the Distribution of Rainfall Amount

Isnewati Ab Malek
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Negeri Sembilan Branch
Seremban Campus, Negeri Sembilan, Malaysia
isnewati@uitm.edu.my

Nuratikah Kassim
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Negeri Sembilan Branch
Seremban Campus, Negeri Sembilan, Malaysia
2017404792@uitm.edu.my, atikahkassim07@gmail.com

Nur Dini Athirah Kamal Ariffin
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Negeri Sembilan Branch
Seremban Campus, Negeri Sembilan, Malaysia
dini.athirah@yahoo.com

Siti Norafifah Abd Jalil
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Negeri Sembilan Branch
Seremban Campus, Negeri Sembilan, Malaysia
afifahjalil98@gmail.com

Haslinda Ab Malek
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Negeri Sembilan Branch
Seremban Campus, Negeri Sembilan, Malaysia
haslinda8311@uitm.edu.my

Article Info

Article history:

Received may 10, 2021
Revised Aug. 15, 2021
Accepted Sept. 22, 2021

Keywords:

Rainfall amount
Weibull distribution
Akaike Information Criterion
Bayesian Information Criterion
Anderson-Darling test

ABSTRACT

The amount of rainfall from seven selected stations in Negeri Sembilan used to find the best fit model employing four continuous distributions: Exponential, Gamma, Weibull, and Normal. This study analyzed distributions of monthly rainfall amounts for ten years from the year 2010 until 2019. Parameters for each distribution were estimated using the maximum likelihood method. Also, the model selection technique based on Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) have been used in examining the best fitting distribution among the four distributions. The Anderson-Darling goodness-of-fit test was applied to all the distributions to determine if the data set follows the specified distribution or not. The test indicates that the Weibull and Normal distribution can be used to model the rainfall amount in Negeri Sembilan. Hence, the Weibull distribution is the best model in describing the rainfall amounts since it has the best fit among all criteria of both AIC and BIC as the distribution has the lowest value compared to other distributions.

Corresponding Author:
Name: Isnewati Ab Malek
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Negeri Sembilan Branch,
Seremban Campus, Negeri Sembilan, Malaysia.
email: isnewati@uitm.edu.my

1. Introduction

Shortage or excessive water gives a positive and negative impact on humans and nature. For example, the advantages of having extreme rainfall are the rain, which is very helpful in keeping the dew balance in the atmosphere and balances the temperature on the Earth. Most of the surface of the earth is covered with water, and most of the water cannot be drunk by humans. Ninety-seven per cent of Earth's water is salty seawater that is useless to most plants and animals living on the ground. That is why rain and snow are vital to life on Earth [1]. Approximately 90 per cent of the

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

world's water supply relies on rainwater. Rain offers a renewable resource that, if it is appropriately managed, can fulfil the food production needs. People must realize that there is nothing in life on its own.

However, if not managed properly, natural disaster such as floods and landslides also can occur because of the excessive water supply. Heavy rainwater can create a flood situation that can impact both individuals and communities. The effects of floods and landslides include loss of human life, property damage, and harm to crops and other plants. Moreover, excessive rain or flood situation also causes many water-borne diseases, which ultimately causes the death of many people. In some areas, infrastructure such as roads and bridges are damaged due to floods or landslides that can disrupt economic activities. Besides, infrastructure damage can cause long-term disruption to clean water, power, education, and health care. Therefore, rainfall analysis is an essential tool for the design of water related infrastructure because it can be used to predict future flood magnitudes for a given magnitude and frequency of extreme rainfall events [2].

Malaysia is an Asian state, which is in the north of the equator. Malaysia also experiences tropical weather every year in which all twelve months have an average temperature of warmer than 18^{o}C (64^{o}F). In tropical weather, there are only two seasons, which is a wet season and a dry season. Malaysia's weather and climate enjoy a monsoon season however this varies based on the destination. The Northeast (Perlis, Kedah, Penang, and Perak) is experiencing the wettest season from November to March, while the Southwest (Negeri Sembilan, Melaka, and Johor) is experiencing their monsoon season from May to September. Statistically, Peninsular Malaysia can be differentiated into eight distinct rainfall regions [3].

As a result, this study aims to determine the characteristic of monthly rainfall amount and find the best fitting distribution of rainfall amount in Negeri Sembilan.

2. Literature Review

Many studies used gamma distribution in rainfall analysis. Based on the previous research, it is common to use a gamma distribution to fit data with values distributed at intervals (0, ∞). The distribution of gamma is widely used to model monthly and daily non-zero rainfall levels [4–7].

In the other studies done by [8], four distributions, Exponential, Gamma, Weibull, and Mixed-Exponential, were used to determine the best fit model for the hourly rainfall amounts in Wilayah Persekutuan. The Maximum Likelihood Method used to estimate parameters for each distribution. The best fit model was selected based on the minimum goodness of fit test error. In describing the amount of hourly rainfall, the mixed exponential found to be the most appropriate distribution.

Ng et al. [4] carried out a study of the maximum annual rainfall in the Kelantan River Basin, Malaysia. The study included fitting the annual series with six probability distributions, namely Beta, Gamma, Gumball, Generalized Extreme Value (GEV), Type III Log-Pearson, and three Weibull parameters. To identify the best fitting distribution, Kolmogorov-Smirnov, Anderson Darling and Chi-Squared tests were applied and the evaluation was based on the total test score. Overall, by giving the highest test scores from all the trials, the GEV distribution is the most appropriate distribution for the examined set.

An analysis is done by [9], the lognormal, skew-normal and mixed lognormal distributions are among the normal transform distributions that are proposed and tested to identify the optimal model for daily rainfall amount in several rain gauge stations in Malaysia. Based on the goodness of fit test, the mixed lognormal is found to be the most appropriate distribution for describing the daily rainfall amount in Malaysia. In 2010, Suhaila et al. [10] conducted research about spatial patterns and trends of daily rainfall regime in Peninsular Malaysia during the Southwest and Northeast Monsoons. A variety of distributions were suggested and evaluated to determine the best method for the 33 years of annual rainfall in Peninsular Malaysia. Exponential, Gamma, Weibull, and Lognormal distributions were the mixed distributions that evaluated in the analysis. The best model is chosen based on the Akaike Information Criterion (AIC) lowest value. The Mixed Lognormal distribution has usually been selected as the best model for most of Peninsular Malaysia's rain gauge stations.

According to [11], the performance of four probability distributions, namely Mixed-Exponential, Gamma, Weibull, and Generalized Pareto, are evaluated and compared in terms of rainfall intensity. Root Mean Square Error (RMSE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Kolmogorov-Smirnov test and Anderson Darling test are used to conduct performance assessments between the distribution. The performance of both Mixed-Exponential and Generalized Pareto is very similar and both in Peninsular Malaysia are equally good in terms of rain intensity.

Furthermore, [12] reported that modelling of rainfall is crucial to determine the possible impacts of climate change. Advanced Weather Generator (AWE-GEN) has been shown to produce precipitation information in the temperate climate zone, with gamma distribution in the model reflecting rainfall intensity. Nevertheless, some reports questioned the inclusion of gamma distribution in a tropical climate like Malaysia. Weibull is therefore suggested to use a massive tail distribution in the analysis. Based on the goodness of fit test and Root Mean Square Error (RMSE), the output of gamma and Weibull distributions was compared. The result shows that gamma is the better distribution for simulating precipitation at rainfall stations located on the Northern Coast's outer parts, whereas Weibull is the better distribution for stations located in the Northern Coast's inner parts.

The study fitting the statistical distributions to the daily rainfall amount in Peninsular Malaysia was conducted by [13]. The gamma, Weibull, Mixed Exponential and Kappa distributions have been tested on the data set. Based on goodness-of-fit tests, the Mixed Exponential best describes the distribution of daily rainfall amount in Peninsular Malaysia.

Therefore, this study was conducted to find the best fitting distribution of rainfall amount in Negeri Sembilan. In the modelling of rainfall distribution, it is crucial to explore well what types of distribution can describe the rainfall trend. Thus, several distributions were proposed for this study which is Normal, Exponential, Gamma and Weibull distributions.

3. Methodology

3.1 Description of Data

The rainfall amount data obtained from the Department of Irrigation and Drainage (DID), Ampang. The data consist of 10 years from January 2010 until December 2019 in monthly form. There are fifty-two stations from all districts in Negeri Sembilan. However, only seven stations are selected. This is because there are missing values problem in the data from each station. Therefore, the station with the completed dataset value of each district is selected to represent the whole Negeri Sembilan. Table 1 shows the variable used in the study:

Table 1. Description of Variable

Variable Name	Type of Variable	Scale of Measurement	Description
Rainfall amount	Quantitative continuous	Ratio	The amount of rainfall by each station (mm)

3.2 Modelling of Rainfall Amount

The mean daily amount of rain, X is a continuous variable since the data has an infinite number of possible values in a selected range. The assumption of daily rainfall is independent of each other. Four continuous distributions proposed to model a rainfall amount in Negeri Sembilan. The chosen distributions are Exponential distribution, Gamma distribution, Weibull distribution and Normal distribution. The probability density functions, and the properties of the distributions presented in Table 2.

Table 2. Properties of the Distributions

Name of Distribution	Probability Density Function	Parameters
Exponential	$f(x) = \beta e^{-\beta x}, x \geq 0$	$\beta > 0, rate$
Gamma	$f(x) = \frac{\beta^{\alpha}}{\tau(\alpha)} X^{\alpha-1} e^{-\beta x}, x \geq 0$	$\alpha > 0, shape$ $\beta > 0, rate$
Weibull	$f(x) = \frac{\alpha(x)^{\alpha-1}}{\mu^{\alpha}} e^{-\left(\frac{x}{\mu}\right)^{n}}, x \geq 0$	$\alpha > 0, shape$ $\mu > 0, scale$
Normal	$f(x) = \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2e^2}(x-\mu)^2}, -\infty < X < \infty$	$\mu \in R, mean$ $\sigma^2 > 0, variance$

3.3 Maximum Likelihood Estimation

The most common parameter estimation method in rainfall analysis is the maximum likelihood. Maximum Likelihood Estimation (MLE) is one of the best methods to obtain a good point of estimator that was developed by Fisher in 1920s. Therefore, in the present study, the MLE method was used for parameter estimation. More details regarding methods on parameter estimation can be found in Myung (2003) [14]. A previous study by [15] also applied MLE to estimate the parameters for seven distributions to represent the rainfall event characteristics in the selected regions in the Peninsular Malaysia. For this study, the parameters of each distribution in Table 2 are estimated using the MLE. Let X be the mean daily amount of rain with probability density function $f(x, \theta)$. The q is an unknown parameter of the distribution. A continuous probability density function $f(x, \theta)$ were observed from the values $x_1, x_2, \ldots, x_n$ of a random sample of size n, the likelihood function of the sample is given by:

$$L(\theta) = f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta) \qquad (1)$$

The logarithmic likelihood function was obtained by taking the natural logarithmic in equation (1):

$$lnL(\theta) = ln \prod_{i=1}^{n} f(x_i, \theta) \qquad (2)$$

The MLE consists of maximizing the likelihood function with respect to the unknown parameter θ by differentiating equation (2) with respect to θ and set the derivative to zero.

$$\frac{\partial lnL(\theta)}{\partial \theta} = 0$$

The value of θ that maximizes the likelihood function as the maximum likelihood estimate of θ.

3.4 Model Selection Criteria

There are mainly two kinds of model selection techniques: hypothesis tests based on goodness-of-fit and information-based criteria. With respect to the distribution selection, one hypothesis test (Anderson-Darling) and two information-based criteria (AIC and BIC) are used in this study to achieve the objective.

Goodness-of-fit (GOF) test statistics are used for checking the validity and choosing the best-fit model among various distribution models for a specific data set [13]. The GOF tests that commonly used in statistics were the chi-square, Kolmogorov-Smirnov, Anderson-Darling and Shapiro-Wilk. The chi-square test was the most common of the GOF test. It can be used for discrete distribution like the binomial distribution and the Poisson distribution, while the Kolmogorov-Smirnov and Anderson-Darling goodness-of-fit tests can only be used for a continuous distribution. The Anderson-Darling test is used in this study and defined as:

H_0: The data follow a specified distribution.
H_1: The data do not follow the specified distribution.

When applying the Anderson-Darling test, if the probability value more than 0.05, the null hypothesis from a specified distribution is not rejected.

The best fit model determined by evaluating the value of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). It was found from related research that information-based criteria (AIC and BIC) can help to identify the best probability model in certain situations [17]. AIC and BIC are considered as measures of relative quality among all the distributions for each station. The model fitting AIC and BIC are defined as in Figure1.

$$AIC = 2k - 2\ln(L)$$

$$BIC = k\ln(n) - 2\ln(L)$$

where,

k =number of parameters in the model,

L=maximum likelihood value and n=number of data.

Figure 1. Definition of AIC and BIC

Therefore, the AIC and BIC of the less parameter would return a smaller value with the same value of maximum likelihood. In other words, the minimum index of AIC and BIC, among others of the model was preferred as the adequate fit to the data [16-18].

4. Results and Discussion

4.1 The Overall Characteristic of Rainfall Amount

Figure 2 shows the average monthly rainfall amount by each station selected in this study from January 2010 until December 2019. Negeri Sembilan consists of seven districts which are Seremban, Kuala Pilah, Rembau, Jelebu, Jempol, Tampin and Port Dickson

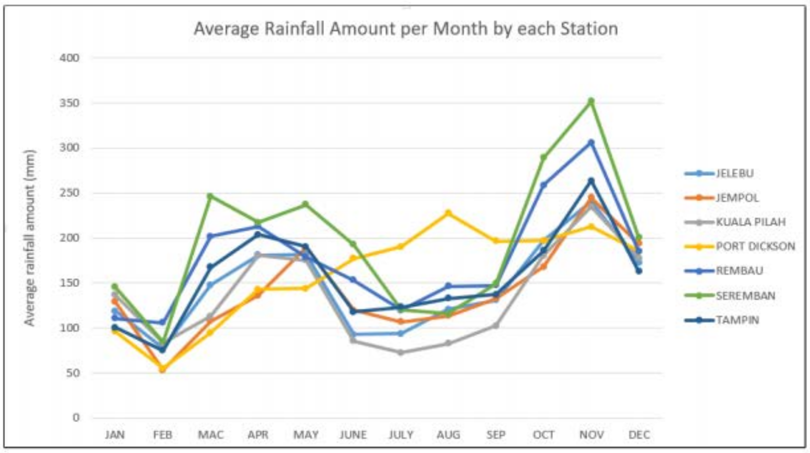

Figure 2. The average monthly rainfall amount by each station

The result shows that the highest average rainfall amount is Seremban station in November with 350 mm per month while the lowest is in February which is Port Dickson and Jempol station with 50 mm per month. Moreover, November indicates the highest average rainfall amount for all station except Port Dickson since the average for Port Dickson is in August. The rainfall variability analysis shows dry and wet conditions. The variability of annual rainfall in Negeri Sembilan is changing due to climate change.

The line graph from Figure 3 illustrates the characteristic amount of rainfall on average, maximum and minimum from January to December in Negeri Sembilan. The black line represents the average amount of rainfall for each month. The red line indicates the maximum amount of rainfall, and the blue line indicates the minimum value of amount rainfall recorded for each month within 10 years. Negeri Sembilan receives most of its rain from October to November in a year.

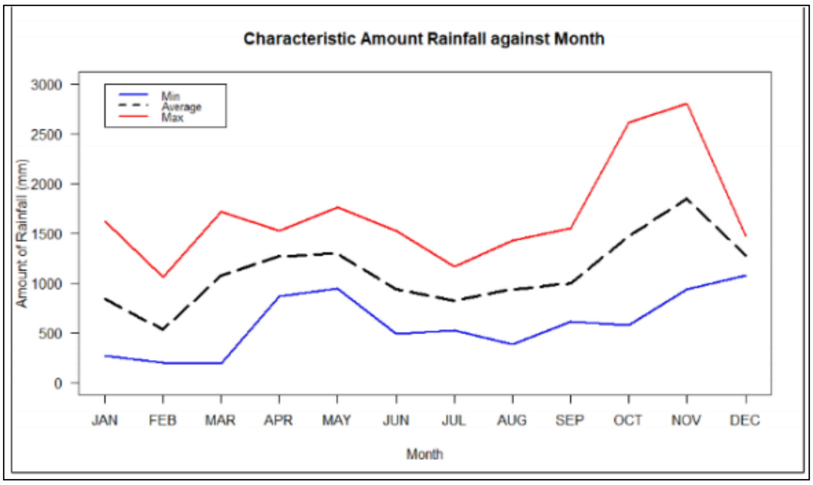

Figure 3. The characteristic of rainfall in Negeri Sembilan

From May until September, it experienced a gradual decrease while starting from September both maximum amount lines and the average amount of rainfall sharply grew to 2806.5 mm and 1852.91mm respectively in November. The minimum line’s amount of rainfall in Mac was rapidly surged to 866 mm in April and continued with a gradual increase to 946 mm in May.

4.2 The Parameters Using Maximum Likelihood Estimation

Table 3 shows the result of the estimation of the parameters using maximum likelihood estimation. Each of the distributions requires different parameters to be estimated. The result is shown below:

Table 3. Summary of the Parameter Estimates

Distributions	**Rate** (β)	**Shape** (α)	**Scale** (η)	**Mean** (μ)	**Standard Deviations** (σ)
Exponential	0.009005			0.009	0.00008109
Gamma	0.040574	4.505171		0.1828	0.00741654
Weibull		2.440696	125.15114	2.4296	0.02475513
Normal				111.0473	48.3433

The rainfall distribution for every station was distinctive considering its land, topographic and atmosphere changes. Consequently, it was hard to characterize the limit of light, moderate and overwhelming precipitation for each station. The parameters calculated for every distribution are using the maximum likelihood estimator. From the results, this study found that the normal distributions can straightforwardly give an intuitive understanding of some aspect distribution with mean and standard deviation. Besides, Exponential distribution requires rate parameters while Gamma distribution requires shape and rate parameters, in the interim, Weibull dissemination requires shape and scale parameters.

The understanding of parameters passes on the conveyance of values in modelling rainfall data which permits the stability of rainfall. The ability of all distributions and the estimated parameters to fit the empirical distribution of value is examined using the Anderson-darling goodness-of-fit test.

Table 4 shows the result of the goodness-of-fit test for all distributions using the Anderson-darling test. Anderson-darling goodness-of-fit test is used to select the best fitting of distribution in this study.

Table 4. Results Goodness-of-fit test

Distributions	Parameter Value	Adjusted Value	Probability
Exponential	16.90825	16.99280	0.0000
Gamma	1.195284	1.195284	<0.005
Weibull	0.421867	0.429570	>0.25
Normal	0.362923	0.365248	0.4364

Based on the results shown in Table 4, a 5% level of significance for Anderson-darling goodness-of-fit tests indicates that Normal and Weibull models are found to be the most suitable model among the four probability distributions tested in describing monthly rainfall amounts. The Normal and Weibull distributions are chosen based on the probability value that is 0.4364 and 0.25. Since the probability value is more than 0.05, the data follow the specified distributions.

4.3 The Best-Fitting Distribution Using Model Selection Criterion

To validate the result in Table 4, further analysis using the model selection criteria as a comparison was performed, and the result is shown in Table 5. Akaike information criterion and Bayesian information criterion used to select the best model. As mentioned earlier, the lower the value of AIC and BIC, the distribution is said to be the best fit model.

Table 5: Results of AIC and BIC

Distributions	AIC	BIC
Exponential	1372.390	1375.177
Gamma	1275.535	1281.110
Weibull	1270.069	1275.644
Normal	1275.344	1280.919

The table shows that the Weibull distribution has the best fit among all criteria of both AIC and BIC as the distribution has the lowest values. Followed by Normal distribution as the second-lowest value with 1275.344 for AIC and 1280.919 for BIC. This finding has the same result as Norzaida et al. [12], which indicated that the distribution of Weibull was the best distribution for stations located in the inner sections of the northern coast of Peninsular Malaysia.

5. Conclusion

The monthly rainfall amounts were tested to model the distribution and the best fitting distributions evaluated based on the lowest value of AIC and BIC. Based on the Anderson-Darling goodness-of-fit test, the most suitable distribution for modelling the rainfall amounts was the Weibull distribution since the probability value is more than 0.05 and has the smallest value of AIC and BIC. Thus, Weibull distribution recommended as the best model in describing the rainfall amounts in Negeri Sembilan. For future study, it is recommended to combine stochastic modelling and distributions in obtaining the best-fitted model for rainfall amount.

Acknowledgements

The authors are grateful to acknowledge the Information Management Unit Hydrology Section, Department of Irrigation and Drainage Ampang because supplying the hydrological data and allowing us the opportunity to conduct this research.

References

[1] M. Fernandez-Raga, A. Castro, E. Marcos, C. Palencia, and R. Fraile, “Weather types and rainfall microstructure in Leon, Spain,” *International Journal of Climatology*, vol. 37, no. 4, pp. 1834–1842, 2016.

[2] F. Alahmadi, N. A. Rahman, and M. Abdulrazzak, “Evaluation of the best Fit distribution for Partial Duration series of daily rainfall in Madinah, western Saudi Arabia,” *Proceedings of the International Association of Hydrological Sciences*, vol. 364, pp. 159–163, 2014.

[3] C. Wong, J. Liew, Z. Yusop, T. Ismail, R. Venneker, and S. Uhlenbrook, “Rainfall characteristics and regionalization in Peninsular Malaysia based on a high resolution gridded data set,” *Water*, vol. 8, no. 11, p. 500, 2016.

[4] J. L. Ng, S. Abd Aziz, Y. F. Huang, A. Wayayok, and M. K. Rowshon, “Analysis of annual maximum rainfall in Kelantan, Malaysia,” *Acta Horticulturae*, no. 1152, pp. 11–18, 2017.

[5] J. Piantadosi, J. Boland, and P. Howlett, “Generating synthetic rainfall on various timescales—daily, monthly and yearly,” *Environmental Modeling & Assessment*, vol. 14, no. 4, pp. 431–438, 2008.

[6] R. E. Chandler and H. S. Wheater, “Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland,” *Water Resources Research*, vol. 38, no. 10, 2002.

[7] Z. R., H. P. G., P. J., B. J. W., and M. N. H., “Modelling catchment rainfall using sum of correlated gamma variables,” *Jurnal Teknologi*, vol. 63, no. 2, 2013.

[8] F. Y., Z. Md., N. V–T–V., S. S., and Z. Y., “Fitting the best–fit distribution for the hourly rainfall amount in the Wilayah Persekutuan,” *Jurnal Teknologi*, 2012.

[9] J. Suhaila, and A. A. Jemain, “Fitting daily rainfall amount in Malaysia using the normal transform distribution,” *Journal of Applied Sciences*, vol. 7, no. 14, pp. 1880–1886, 2007.

[10] J. Suhaila, S. M. Deni, W. Z. Wan Zin, and A. A. Jemain, “Spatial patterns and trends of daily rainfall regime in Peninsular Malaysia during the southwest and Northeast Monsoons: 1975–2004,” *Meteorology and Atmospheric Physics*, vol. 110, no. 1-2, pp. 1–18, 2010.

[11] A. H. Syafrina, M. D. Zalina, and A. Norzaida, “Climate projections of future extreme events in Malaysia,” *American Journal of Applied Sciences*, vol. 14, no. 3, pp. 392–405, 2017.

[12] A. Norzaida, M.R. Siti Musliha, M.D Zalina, A.H Syafrina, “Probability distributions comparative analysis in assessing rainfall process in time and space,” *International Journal of Civil Engineering and Technology*, vol. 8 no. 10, pp. 1679-1688, 2017.

[13] S. Jamaludin and A. A. Jemain, “Fitting the statistical distributions to the daily rainfall amount in peninsular malaysia,” *Jurnal Teknologi*, 2012.

[14] I. J. Myung, “Tutorial on maximum likelihood estimation,” *Journal of Mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.

[15] Z. Hassan, S. Shamsudin, and S. Harun, “Choosing the best Fit distribution for rainfall event characteristics based on 6h-ietd within Peninsular Malaysia,” *Jurnal Teknologi*, vol. 75, no. 1, 2015.

[16] M. Alam, C. Farnham, and K. Emura, “Best-fit probability models for maximum monthly rainfall in Bangladesh using gaussian mixture distributions,” *Geosciences*, vol. 8, no. 4, p. 138, 2018.

[17] G. Di Baldassarre, F. Laio, and A. Montanari, “Design flood estimation using model selection criteria,” *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 34, no. 10-12, pp. 606–611, 2009.

[18] F. Laio, G. Di Baldassarre, and A. Montanari, “Model selection techniques for the frequency analysis of Hydrological Extremes,” *Water Resources Research*, vol. 45, no. 7, 2009.