

Statistical modelling of extreme rainfall in Peninsular Malaysia

Wei Lun Tan^{1*}, Woon Shean Liew¹, and Lloyd Ling²

¹Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Kajang 43000, Malaysia

²Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Kajang 43000, Malaysia

Abstract. Flash floods are known as one of the common natural disasters that cost over billions of Ringgit Malaysia throughout history. Academically, an extreme rainfall model is effective in modelling to predict and prevent the occurrence of flash floods. This paper compares four probability distributions, namely, exponential distribution, generalized extreme value distribution, gamma distribution, and Weibull distribution, with the rainfall data of 10 stations in peninsular Malaysia. The period of the data is from 1975 to 2008. The comparison is based on the descriptive and predictive analytics of the models. The determination of the most effective model is through Kolmogorov-Smirnov, Anderson-Darling, and chi-square test. The result shows that generalized extreme value is the most preferred extreme rainfall model for the rainfall cases in Peninsular Malaysia.

1 Introduction

In Peninsular Malaysia, massive floods frequently happened because of extreme rainfall during the monsoon season, especially the northeast monsoon season. Hence, understanding extreme rainfall characteristics are vital to disaster risk management. An efficient model will enable the Malaysian ministry to determine the probability of a flood occurrence.

The probability distribution of extreme rainfall plays a vital role in predicting flash floods. There are a few existing studies on the best-fit distribution of extreme rainfall. Syafrina and Norzaida [1] tested and compared the performance of gamma and Weibull in a weather generator model. Advanced weather generator (AWE-GEN) model is employed to model the rainfall at an hourly scale. They concluded that the gamma distribution provided a better result for the hourly rainfall data. On the other hand, Kumar et al. [2] conducted in Uttarakhand, India showed different results. The best-fit distribution is applied, with the comparison performed based on the goodness-of-fit test. Weibull distribution outperformed other distributions while the chi-square and log-Pearson are the next best distributions to be used. The usage of Theil-Sen's slope estimator, Mann-Kendall (MK) and modified Mann-Kendall (MMK) were discussed by Prabhakar et al. [3]. These estimators were applied for trend analysis of rainfall data for a long period. The changing point for long-term rainfall

*Corresponding author: tanwl@utar.edu.my

time series is investigated by using standard normal homogeneity test (SNHT) and Mann–Whitney–Pettitt (MWP) test in Odisha, India. The result showed a decreasing trend of rainfall beyond the year 1945.

Extreme rainfall events are ranked according to Weibull's method in the study done by Sabarish et al. [4], while the chi-square and Kolmogorov-Smirnov tests are used to investigate the suitability of distribution in Tiruchirappalli City, South India. The results showed that log-Pearson type III distribution is suitable for estimating rainfall amounts at various probability levels. A daily rainfall disaggregation model was adopted by Paola et al. [5] to evaluate the IDF curves of rainfall. The IDF curves were obtained using the probability distribution of Gumbel, and a short duration of rainfall data that are less than 24 hours have been obtained by using two different models of disaggregation in the historical rainfall data in African. The result showed that the effect of climate change affects the frequency of extreme events.

Ten commonly used probability distributions for extreme rainfall were considered in the study made by Nguyen and Nguyen [5], and a further investigation is made by Nguyen et al. [6] with the same distribution tested in Ontario, Canada. The results showed that generalized extreme value (GEV), generalized normal and Pearson type III (PE3) are the best overall distributions that provide the best goodness-of-fit and sturdy quantile extrapolations. Besides, GEV distribution is more preferred as compared to two other best overall distributions. Kar et al. [7] used a regional approach based on L-moments to estimate hourly rainfall frequency estimation and goodness-of-fit measure in Jeju Island, Korea. The study showed that Gumbel and GEV distributions are considered more reliable and successful models for the studied area. This study showed that the model is suitable and can be implicated in other areas with similar characteristics, limited rainfall data and steep land slope.

Smith [8] applied a weighted least-squares regression to measure Barker's rainfall trends in Southeast Texas. The result failed to demonstrate less frequent, more extreme annual rainfall events occurring now than occurred in the past. Mehr et al. [9] developed and applied a novel classification-forecasting model, namely binary GP (BGP), for teleconnection studies between sea surface temperature (SST) variations and maximum monthly rainfall (MMR) events in the northwest of Iran. A few limitations were found throughout the studies. One of the limitations is that the model is only suitable for maximum monthly rainfall forecasting, and there will be binary classification issues using genetic programming.

Generally, different areas with different rainfall characteristics affect the choice of appropriate distribution to be used. Hence, it is necessary to analyze the extreme rainfall characteristics to determine the best-fit distribution for extreme rainfall.

2 Data

This study focuses on 10 rainfall stations (refer to Table 1) from Peninsular Malaysia for the year 1975 to the year 2008. The northeast monsoon from November until February is considered in this study. All the historical rainfall data were obtained from the Department of Irrigation and Drainage Malaysia. A peak-over-threshold (POT) approach with thresholds of 90th percentile and 95th percentile is applied to obtain a list of extreme rainfall data to fit into some selected probability distributions.

3 Methodology

First, the rainfall data will start by applied the POT. POT is one of the many methods used in extreme value analysis by looking at the extreme values from the given data that exceed a particular threshold value. First, by applying POT, all the zero rainfall are withdrawn from

Table 1. Geographical coordinates of the 10 selected rainfall stations in Peninsular Malaysia during the period 1975-2008.

NO	DISTRICT	LONGITUDE	LATITUDE
1	JPS Wilayah Persekutu	101.68	3.16
2	Ipoh	101.10	4.57
3	PekanMerlimau	102.43	2.15
4	LadangBenut	103.35	1.84
5	Arau	100.27	6.43
6	BukitBerapit	100.48	5.38
7	Mersing	103.83	2.45
8	KotaBharu	102.28	6.17
9	LadangBoh	101.43	4.45
10	GuaMusang	101.97	4.88

the data. Then, the rainfall amount which exceeds a certain threshold will be included into the model. The thresholds used in this study are determined by the 90th percentile and 95th percentile. The extreme rainfall data are fitted to four probability distributions.

3.1 Probability distribution function

In this study, the rainfall data are cleansed using POT to obtain a list of extreme rainfall data for fitting into the four probability distributions. There are exponential distribution, GEV distribution, gamma distribution, and Weibull distribution.

3.1.1 Exponential distribution

The probability density function (pdf) for exponential distribution is shown as follow:

$$f(x) = \lambda e^{-\lambda x} \tag{1}$$

where λ represents the rate. The maximum likelihood estimator (MLE) of λ is given by,

$$\hat{\lambda} = \frac{1}{\bar{x}} \tag{2}$$

where \bar{x} denotes the sample mean,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \tag{3}$$

in which the MLE represents the reciprocal of the sample mean.

3.1.2 Generalized extreme value distribution

The pdf for GEV distribution [10] is shown as follow:

$$f(x) = \frac{1}{\sigma} \left[1 - \xi \left(\frac{x - \mu}{\sigma} \right) \right]^{\frac{1-\xi}{\xi}} e^{-\left[1 - \xi \left(\frac{x - \mu}{\sigma} \right) \right]^{\frac{1}{\xi}}}, \xi \neq 0 \tag{4}$$

where μ, σ and ξ represent the location, scale and shape of the distribution function, respectively. The log-likelihood function [11] is given by,

$$L(\mu, \sigma, \xi) = -n \log(\sigma) - (1 - \xi) \sum_{i=1}^n y_i - \sum_{i=1}^n e^{-y_i} \tag{5}$$

where

$$y_i = -\frac{1}{\xi} \log \left[1 - \xi \left(\frac{x_i - \mu}{\sigma} \right) \right] \tag{6}$$

The MLE's of μ, σ , and ξ are those values that maximize the likelihood function, subject to the following constraints:

$$\sigma > 0 \tag{7}$$

$$\xi \leq 1 \tag{8}$$

$$x_i < \mu + \frac{\sigma}{\xi} \text{ if } \xi > 0 \tag{9}$$

$$x_i > \mu + \frac{\sigma}{\xi} \text{ if } \xi < 0 \tag{10}$$

A constraint $\xi \leq 1$ is imposed because the likelihood can be made infinite and cause the MLE to not exist when $\xi > 1$.

3.1.3 Gamma distribution

The pdf for gamma distribution is shown as follow,

$$f(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} \tag{11}$$

where k and θ represent the shape and scale of the distribution, respectively. The relationship between the coefficient of variation (τ) and mean (μ) of this distribution can be described as,

$$k = \tau^{-2} \text{ and } \theta = \frac{\mu}{k} \tag{12}$$

$$\mu = k\theta \text{ and } \tau = k^{-\frac{1}{2}} \tag{13}$$

The MLE's of k and θ are the solutions of the simultaneous equations:

$$\hat{k} = \frac{1}{n} \sum_{i=1}^n \log(x_i) - \log(\bar{x}) = \psi(\hat{k}) - \log(\hat{k}) \tag{14}$$

$$\hat{\theta} = \frac{\bar{x}}{\hat{k}} \tag{15}$$

where ψ denotes the digamma function, and \bar{x} denotes the sample mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \tag{16}$$

3.1.4 Weibull distribution

The pdf for Weibull distribution is shown as follow:

$$f(x) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k} \tag{17}$$

where λ and k represent scale and shape, respectively. The MLE's of λ and k are the solutions of the simultaneous equations:

$$\hat{\lambda} = \left[\frac{1}{n} \sum_{i=1}^n x_i^{\hat{k}} \right]^{\frac{1}{\hat{k}}} \tag{18}$$

$$\hat{k} = \frac{n}{\left\{ \left(\frac{1}{\hat{\lambda}}\right)^{\hat{k}} \sum_{i=1}^n [x_i^{\hat{k}} \log(x_i)] \right\} - \sum_{i=1}^n \log(x_i)} \tag{19}$$

3.2 Goodness-of-fit test

The goodness-of-fit test used in this study are Kolmogorov-Smirnov (*K-S*), Anderson-Darling (*A-D*), and chi-square test, with the significant level of 5%.

3.2.1 Kolmogorov-Smirnov test

The *K-S* test compares the empirical distribution function ($F_n(x)$) with a specified cumulative distribution function ($F(x)$). The equation for computing the Kolmogorov-Smirnov statistic (D_n) is:

$$D_n = \max |F_n(x) - F(x)| \tag{20}$$

where the equation is used to compute the distance between the two functions, $F_n(x)$ and $F(x)$. The larger the value of the test statistics, the higher the inconsistency between the observed data.

3.2.2 Anderson-Darling test

The *A-D* Test is the modified version of the *K-S* test that give higher weight on the tails of the tested distributions. The equation for the *A-D* test statistics is:

$$AD = -n - \frac{1}{n} \sum_{i=1}^n (2i - 1) \left[\ln(F(x_i)) + \ln(1 - F(x_{(n+1-i)})) \right] \tag{21}$$

where $x_{(1)}$ to $x_{(n)}$ is the ordered sample of size n from smallest to largest, and $F(x)$ is the cumulative distribution function for the specified distribution. A null hypothesis is rejected if the *AD* is greater than the critical value of AD_{α} with the given significant level of α .

3.2.3 Chi-square test

The chi-square test is used to check the suitability of a specific distribution by observing the sample's frequency. By using O as the “observed count” and E as the “expected count”, the equation to calculate chi-square is:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{22}$$

The null hypothesis for the test claim that there is no significant difference between the observed and expected frequencies whereas, the alternative hypothesis claims that they are different.

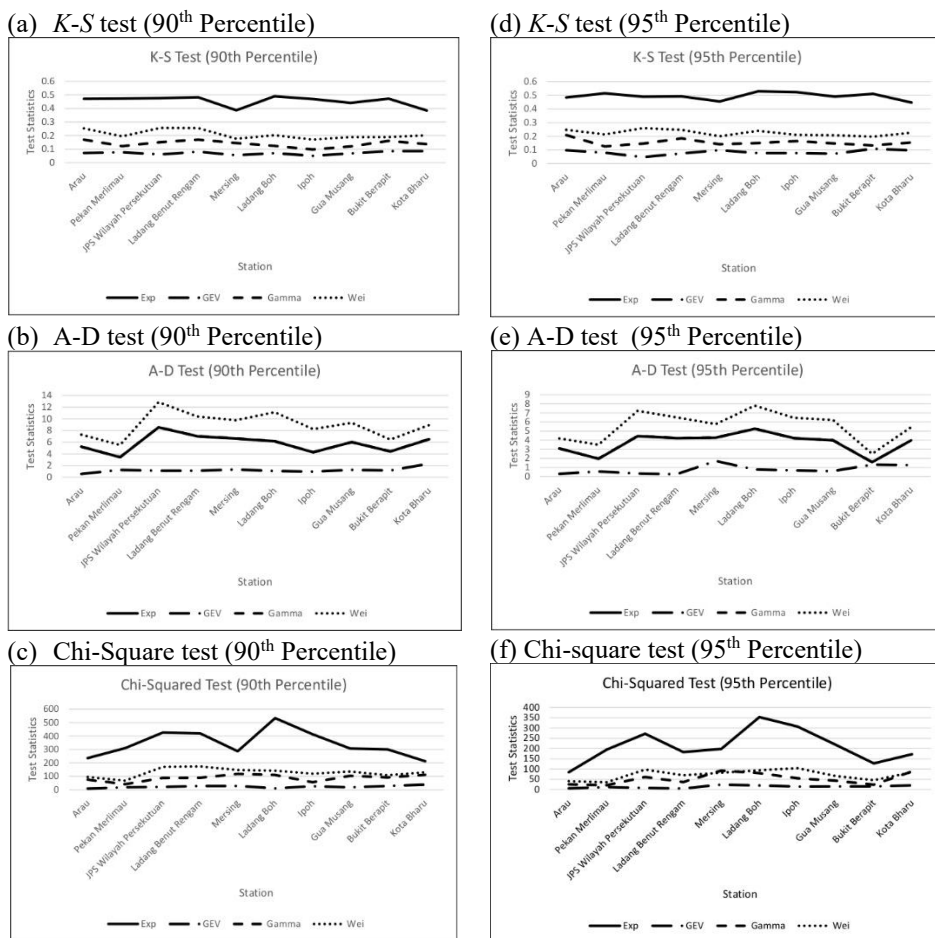


Fig. 1. Test statistics for (a)-(c) threshold value of 90th percentile and (d)-(f) threshold value of 95th percentile.

4 Results and discussions

First, we start with the data cleansing. All the zero rainfall are withdrawn from the data. After the data cleansing, the POT is applied to obtain the extreme rainfall data using the 90th

percentile and 95th percentile thresholds. The parameters for each distribution are estimated for both thresholds using maximum likelihood estimator (MLE).

Table 2 and Table 3 show the estimated parameters for exponential distribution, gamma distribution, Weibull distribution, and GEV distribution for all the rainfall stations for the 90th and 95th percentile thresholds. After all of the estimated parameters are obtained, a good-of-fit test is used to determine the best fit distribution for all the rainfall stations.

The test statistics of all the selected goodness-of-fit tests for exponential distribution (Exp), GEV, Gamma distribution (Gamma) and Weibull distribution (Wei) are as shown in Fig. 1.

Fig. 1 shows that GEV distribution is the best overall result from all 3 of the goodness-of-fit tests for both the 90th and 95th percentile thresholds. Therefore, it can be concluded that GEV distribution is the best fit distribution for the extreme rainfall for the 10 selected rainfall stations.

The quantile-quantile ($Q-Q$) plots are adopted into the extreme rainfall to further visualize the suitability of the selected distribution. The Fig. 2 and Fig. 3 show the $Q-Q$ plots of the four distributions with the threshold of 90th percentile and 95th percentile for 5 selected rainfall stations in Peninsular Malaysia. Fig. 1 and Fig. 2 show the $Q-Q$ plots that the extreme rainfall data fit into GEV distribution the best, with the majority of the data fall around the straight line. In contrast, the gamma distribution and the Weibull distribution are the second-best feasible choice with similar $Q-Q$ plots. The exponential distribution would be the least favorable distribution to be chosen for fitting the extreme rainfall model.

5 Conclusion

The fitting distribution for extreme rainfall event is crucial in hydrology studies. The best-fit distribution can be used in hydrology model such as rainfall runoff model. In this study, 10 selected rainfall stations over Peninsular Malaysia during northeast monsoon season from year 1975 until year 2008 were fitted to four probability distributions. The four distributions are exponential distribution, gamma distribution, Weibull distribution and GEV. The three-difference goodness-of fit tests were used to the model performance assessment. The goodness-of fit tests, $K-S$ test, $A-D$ test, and chi-square tests have indicated that GEV distribution is the best-fit for all the 10 selected rainfall stations. The suitability of the selected probability distribution with the extreme rainfall data is visualized through the quantile-quantile plots. Comparing all the quantile-quantile plots, GEV distribution shows the best at fitting the extreme rainfall data compared to other probability distribution. These results have shown the same agreement with the results obtained from the goodness of fit tests. It can be concluded that the GEV distribution is a best-fit probability distribution for the extreme rainfall event in Peninsular Malaysia. In future study, the GEV distribution can be used to predict extreme rainfall event.

The authors are grateful to the Drainage and Irrigation Department for providing the rainfall data. The work is funded by UTAR Research Fund Vote 6200/TG1 awarded by Universiti Tunku Abdul Rahman.

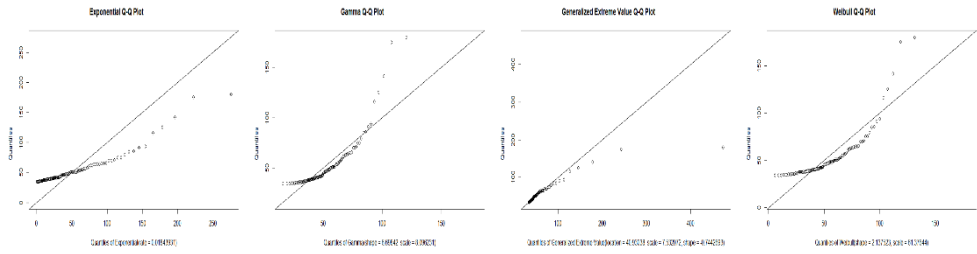
Table 2. Estimated parameters for the threshold value of the 90th percentile.

Area Name	Exponential Distribution
Arau	$\lambda = 0.018439$
Pekan Merlimau	$\lambda = 0.021381$
JPS Wilayah Persekutuan	$\lambda = 0.017493$
Ladang Benut Rengam	$\lambda = 0.016446$
Mersing	$\lambda = 0.009248$
Ladang Boh	$\lambda = 0.025871$
Ipoh	$\lambda = 0.018443$
Gua Musang	$\lambda = 0.019780$
Bukit Berapit	$\lambda = 0.017975$
Kota Bharu	$\lambda = 0.008088$
Area Name	Generalized Extreme Value Distribution
Arau	$\mu = 40.930375, \sigma = 7.532072, \xi = -0.744269$
Pekan Merlimau	$\mu = 37.383810, \sigma = 8.180785, \xi = -0.469732$
JPS Wilayah Persekutuan	$\mu = 44.763463, \sigma = 8.514707, \xi = -0.607598$
Ladang Benut Rengam	$\mu = 47.661947, \sigma = 8.684274, \xi = -0.625206$
Mersing	$\mu = 75.960770, \sigma = 25.187830, \xi = -0.525310$
Ladang Boh	$\mu = 32.172431, \sigma = 6.156170, \xi = -0.384057$
Ipoh	$\mu = 44.762656, \sigma = 9.880943, \xi = -0.324186$
Gua Musang	$\mu = 38.448843, \sigma = 9.302372, \xi = -0.553023$
Bukit Berapit	$\mu = 43.942123, \sigma = 8.937590, \xi = -0.573548$
Kota Bharu	$\mu = 81.994504, \sigma = 26.346240, \xi = -0.706841$
Area Name	Gamma Distribution
Arau	$k = 6.698420, \theta = 8.096231$
Pekan Merlimau	$k = 9.573792, \theta = 4.885238$
JPS Wilayah Persekutuan	$k = 8.118586, \theta = 7.041046$
Ladang Benut Rengam	$k = 8.122050, \theta = 7.486347$
Mersing	$k = 4.511983, \theta = 23.965971$
Ladang Boh	$k = 12.155040, \theta = 3.179993$
Ipoh	$k = 11.016434, \theta = 4.921835$
Gua Musang	$k = 7.030617, \theta = 7.190917$
Bukit Berapit	$k = 9.365283, \theta = 5.940380$
Kota Bharu	$k = 3.922303, \theta = 31.523720$
Area Name	Weibull Distribution
Arau	$k = 2.137523, \lambda = 61.376439$
Pekan Merlimau	$k = 2.746517, \lambda = 52.400516$
JPS Wilayah Persekutuan	$k = 2.207335, \lambda = 64.272211$
Ladang Benut Rengam	$k = 2.274670, \lambda = 68.503700$
Mersing	$k = 1.945663, \lambda = 122.803088$
Ladang Boh	$k = 2.950212, \lambda = 43.044884$
Ipoh	$k = 2.972338, \lambda = 60.523314$
Gua Musang	$k = 2.368762, \lambda = 57.123778$
Bukit Berapit	$k = 2.769707, \lambda = 62.415913$
Kota Bharu	$k = 1.760565, \lambda = 140.128654$

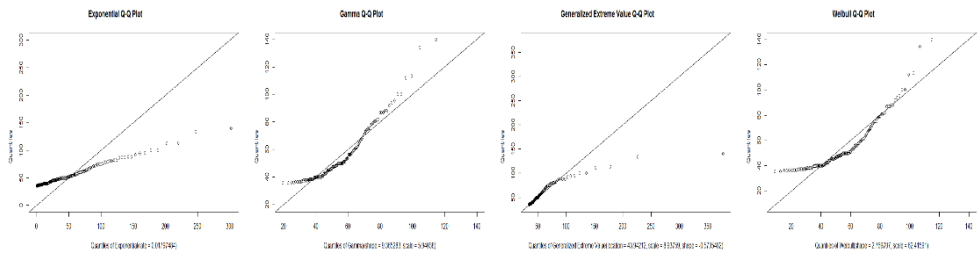
Table 3. Estimated parameters for threshold value of 95th percentile.

Area Name	Exponential Distribution
Arau	$\lambda = 0.014254$
Pekan Merlimau	$\lambda = 0.017004$
JPS Wilayah Persekutuan	$\lambda = 0.013866$
Ladang Benut Rengam	$\lambda = 0.012805$
Mersing	$\lambda = 0.006810$
Ladang Boh	$\lambda = 0.021215$
Ipoh	$\lambda = 0.014897$
Gua Musang	$\lambda = 0.015335$
Bukit Berapit	$\lambda = 0.014271$
Kota Bharu	$\lambda = 0.005824$
Area Name	Generalized Extreme Value Distribution
Arau	$\mu = 54.291925, \sigma = 9.036329, \zeta = -0.738278$
Pekan Merlimau	$\mu = 50.112093, \sigma = 8.048719, \zeta = -0.413445$
JPS Wilayah Persekutuan	$\mu = 58.789960, \sigma = 10.279920, \zeta = -0.496930$
Ladang Benut Rengam	$\mu = 64.075697, \sigma = 10.835179, \zeta = -0.471028$
Mersing	$\mu = 107.601563, \sigma = 22.834922, \zeta = -0.833175$
Ladang Boh	$\mu = 40.020977, \sigma = 4.943260, \zeta = -0.628783$
Ipoh	$\mu = 57.293432, \sigma = 7.766114, \zeta = -0.536825$
Gua Musang	$\mu = 52.904782, \sigma = 9.540268, \zeta = -0.539583$
Bukit Berapit	$\mu = 58.120913, \sigma = 10.195953, \zeta = -0.517098$
Kota Bharu	$\mu = 123.074747, \sigma = 27.001167, \zeta = -0.836290$
Area Name	Gamma Distribution
Arau	$k = 7.793649, \theta = 9.001875$
Pekan Merlimau	$k = 16.005415, \theta = 3.674432$
JPS Wilayah Persekutuan	$k = 9.899709, \theta = 7.285099$
Ladang Benut Rengam	$k = 9.962612, \theta = 7.838902$
Mersing	$k = 6.885273, \theta = 21.327076$
Ladang Boh	$k = 17.059947, \theta = 2.762971$
Ipoh	$k = 17.812432, \theta = 3.768701$
Gua Musang	$k = 10.808510, \theta = 6.033190$
Bukit Berapit	$k = 14.213285, \theta = 4.929905$
Kota Bharu	$k = 5.835814, \theta = 29.424126$
Area Name	Weibull Distribution
Arau	$k = 2.367737, \lambda = 79.236446$
Pekan Merlimau	$k = 3.357202, \lambda = 64.953610$
JPS Wilayah Persekutuan	$k = 2.354248, \lambda = 80.810852$
Ladang Benut Rengam	$k = 2.497066, \lambda = 87.689650$
Mersing	$k = 2.417365, \lambda = 166.005290$
Ladang Boh	$k = 3.396004, \lambda = 52.002488$
Ipoh	$k = 3.661288, \lambda = 73.918180$
Gua Musang	$k = 2.894411, \lambda = 72.942719$
Bukit Berapit	$k = 3.457429, \lambda = 77.581044$
Kota Bharu	$k = 2.102253, \lambda = 194.569794$

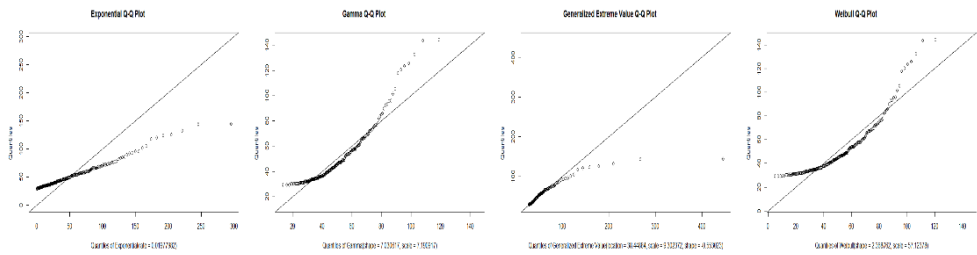
(a) Arau



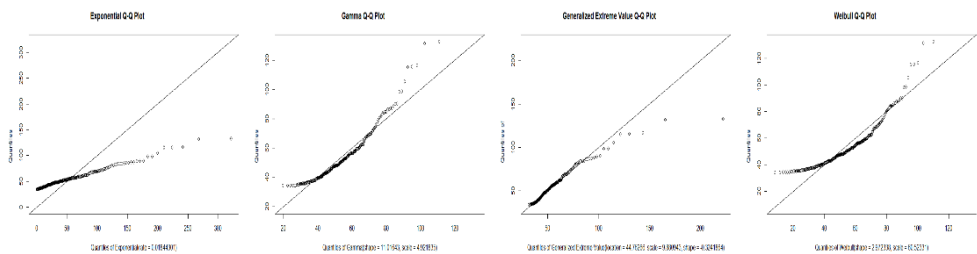
(b) Bukit Berapit



(c) Gua Musang



(d) Ipoh



(e) JPS W.P.

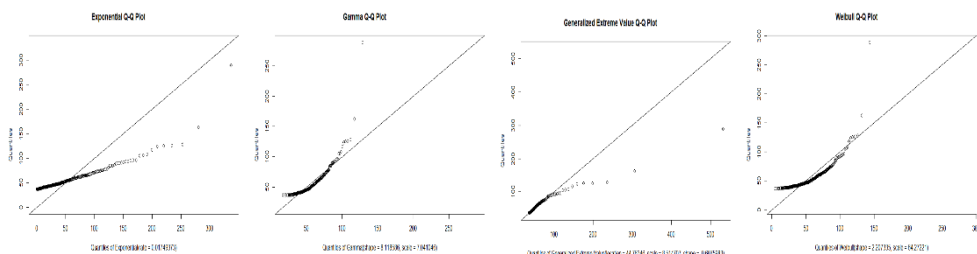
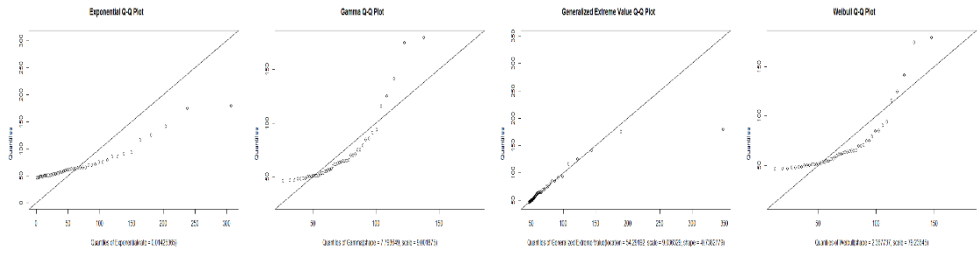
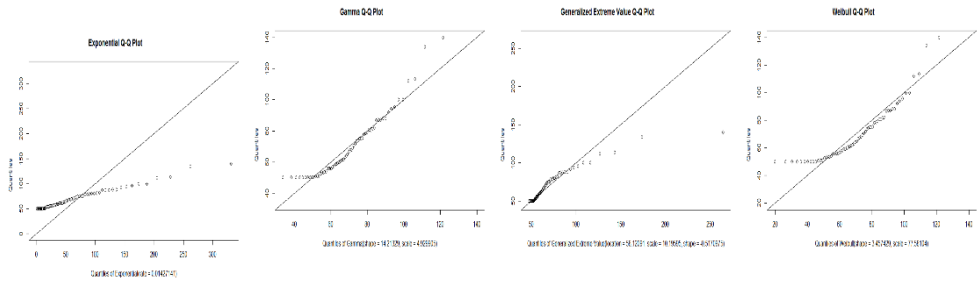


Fig. 2. *Q-Q* plots for selected rainfall with threshold of 90th percentile.

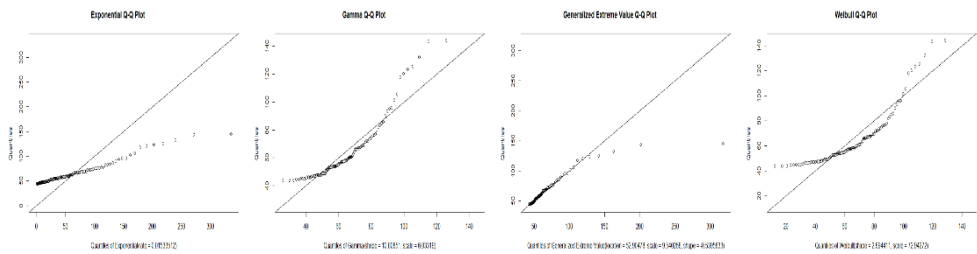
(a) Arau



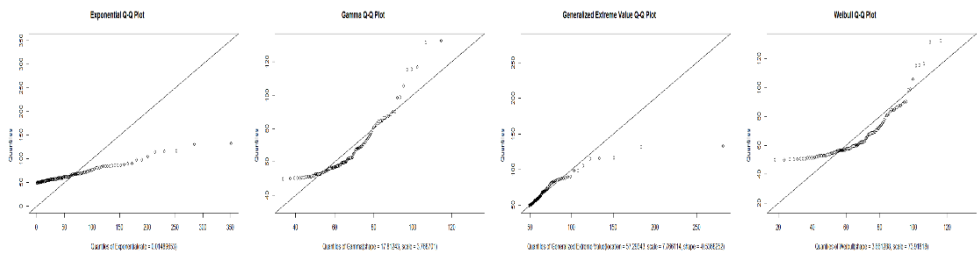
(b) Bukit Berapit



(c) Gua Musang



(d) Ipoh



(e) JPS W.P.

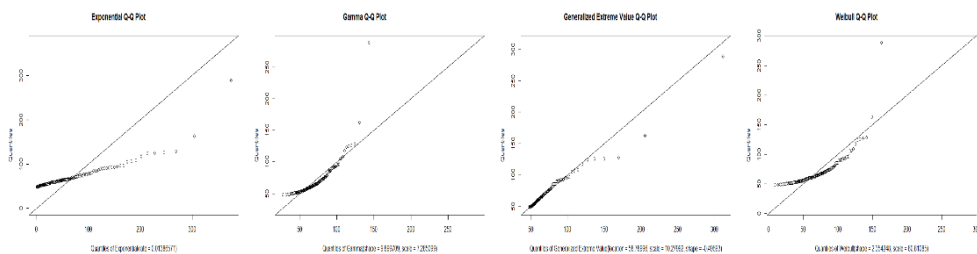


Fig. 3. *Q-Q* plots for selected rainfall station with threshold of 95th percentile.

References

1. A.H. Syafrina and A. Norzaida, *Comm. App. Math. Comp. Sci.* **8**, 241 – 251 (2017)
2. V. Kumar, S. Khan, Jahangeer, *Appl. Water Sci.* **7**, 4765 – 4776 (2017)
3. A.K. Prabhakar, K.K. Singh, A.K. Lohani, S.K. Chandniha, *Appl. Water Sci.* **9**, 1 – 15 (2019)
4. R.M. Sabarish, R. Narasimhan, A.R. Chandhru, C.R. Suribabu, J. Sudharsan, and S. Nithiyantham, *Appl. Water Sci.* **7**, 1033 – 1042 (2017)
5. F.D. Paola, M. Giugni, M.E. Topa, E. Bucchignani, *Springerplus* **3**, 1 – 18 (2014)
6. V.T.V. Nguyen and T.H. Nguyen, *Procedia Eng.* **154**, 624 – 630 (2016)
7. K.K. Kar, S.K. Yang, J.H. Lee, F.K. Khadim, *Geoenviron. Disasters* **4**, 1 – 13 (2017)
8. R.K. Smith, *Springerplus* **4**, 1 – 11 (2015)
9. A.D. Mehr, V. Nourani, B. Hrnjica, A. Molajou, *J. Hydrol.* **555**, 397 – 406 (2017)
10. V.P. Singh, *Water Sci. Technol. Libr.* **30**, 169 – 183 (1998)
11. S. El Adlouni, T. B. M. J. Ouarda, X. Zhang, R. Roy, B. Bobee, *Water Resour. Res.* **43**, W03410 (2007)