Investigation of the best fit probability distribution for annual maximum rainfall in Kelantan River Basin

J L Ng1, S Y Yap1, Y F Huang2, N I F Md Noh1, R A Al-Mansob1 and R Razman1

1Department of Civil Engineering, Faculty of Engineering, Technology and Built Environment, UCSI University, Kuala Lumpur, Malaysia
2 Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Selangor, Malaysia

E-mail: ngjl@ucsiuniversity.edu.my

Abstract. Flooding is one of the natural disasters that happens annually in Malaysia. Flooding is induced by the extreme rainfall event and this can cause severe impacts in terms of environment, society, health and safety. Consequently, investigating the best fit probability distribution for annual maximum rainfall can provide the fundamental ideas for government departments and relevant authorities to mitigate the flooding problems. The aim of this study is to investigate the best fit probability distribution in describing the characteristics of annual maximum rainfall for the period of 1994 to 2013 at the Kelantan River Basin. The Gamma, Gumbel, Generalized Extreme Value and Log Pearson Type-III distributions were fitted to the historical rainfall series. Three goodness of fit tests, namely the Kolmogorov-Smirnov, the Anderson-Darling and the Chi-Square tests were used to evaluate the probability distributions. The performances of each probability distribution generated by the goodness of fit tests were compared. Overall, the Generalized Extreme Value distribution seems to be the best fit probability distribution for the annual maximum rainfall at most of the rainfall stations except for the RPS Kuala Betis station that had the Log Pearson Type-III distribution as its best fit.

1. Introduction

The annual maximum rainfall is the highest of the values that occur within each year of the record period. The annual maximum rainfall can be described as an extreme occasion with hazard duration for a river basin with instant effect to agriculture, soil conservation, drainage and so on [1, 2]. The information of the annual maximum rainfall is very important for flood mitigation and water resources management.

Malaysia had experienced variability in rainfall for the past 10 years, especially at Kelantan. This had caused severe floods in the state of Kelantan every year during the monsoon season. Malaysia usually has two monsoon seasons, namely the Southwest Monsoon and the Northeast Monsoon. There are more rainfalls in Kelantan during the Northeast Monsoon compared to the period of the Southwest Monsoon. Flooding in Kelantan during by the monsoon season is a serious issue. Floods cause property damage and significant loss of life every year in Kelantan River Basin [3, 4, 5]. Flooding occurs when the river loses its ability to hold the volumes of water which overflows into the low-lying areas and adversely affecting the human settlements located after the buffer zone.

Rainfall analysis using probability distribution and annual maximum rainfall can be used to enhance the management of water resources. The rainfall analysis provides useful information for
flood prevention/mitigation. The information is also useful for the planning and designing of water resources engineering systems such as reservoirs, flood controls and drainage designs. In addition, the information resulting from the rainfall analysis also provide fundamental information for impending heavy rainfall events. Water resources planner, farmers and urban engineers require reliable rainfall data for the management and implementation of water resource strategies.

The probability distribution is fundamental to all experiments that provide the probabilities of occurrence in various samples [6, 7, 8]. The probability distribution is used to predict how often certain values of a variables phenomenon may occur and to assess the reliability of the prediction. There are several probability distributions that are applied to continuous random variables, an example of such a variable is the annual maximum rainfall. A continuous random variable can be expressed by not only one probability distribution. For evaluating heavy rainfall events in an area, a best fit probability distribution function (PDF) should be selected in hydrology [9]. Bhakar et al. [10] stated that probability and frequency analysis of rainfall data enables us to determine the expected rainfall at various chances. Upadhyaya and Singh [11] reported that different types of probability distribution can forecast precipitation precisely for different return periods. The Normal, Log-Normal (LN), Gumbel, Gamma 2 (G2), Pearson Type-III (P3), Log-Pearson Type-III (LP3) and Weibull distributions are the most common PDFs used in hydrology [12]. In hydrology, the commonly goodness of fit tests are the Kolmogorov–Smirnov, the Chi-Square, Filliben and the Anderson–Darling [13]. When a best fit probability distribution is obtained with certain values for its parameters, this distribution is then used to generate the corresponding random variables during the simulation.

For the annual and monsoon seasons, the gamma distribution was discovered as the best fit probability distribution among the others [14]. Bhakar et al. [10] reported that the Gamma distribution was the best fit probability distribution tested with the Chi-square value when compared to other distributions. The rainfall data of 110 years were collected from Colombo to analyze the annual maximums of rainfall. Varathan et al. [15] discovered that the Gumbel distribution was the best fit probability distribution. One of the most popular distribution used for extreme rainfall data is the Generalized Extreme-Value (GEV) distribution [16, 17, 18]. To investigate the annual maximum rainfall in Peninsular Malaysia, the GEV distribution is a good choice because it has good descriptive and predictive abilities. Based on the estimation of rainfall in northern regions of Pakistan, the LP3 distribution was found to be the best fit probability distribution for all the rainfall stations in that region. This best-fit probability distribution can be used to determine the maximum values of predicted rainfall.

In Malaysia, there is a dearth of research on the characteristics of annual maximum rainfall. Lacking of sufficient or correct information in describing the characteristics of annual maximum rainfall can lead to inaccurate weather analysis. Accordingly, this study aims to evaluate the performances of various probability distributions in representing the annual maximum rainfall. The significances of this study is to contribute the research knowledge of describing the characteristics of annual maximum rainfall in a tropical basin such as the Kelantan River Basin. Furthermore, this study provides valuable information for community when making decisions concerning future development and public investment. Moreover, this study also provides insight into rainfall variability to help improve the range of climate prediction.

2. Study Area

Kelantan River Basin is located at the north-eastern part of Peninsular Malaysia. The location is between latitudes 4°40' and 6°12' North, and longitudes 101°20' and 102°20' East. The Kelantan River Basin is also ringed by parts of southern Thailand to the north, Perak to the west, Pahang to the south and Terengganu to the southeast. The total length of the Kelantan River Basin is roughly 248km long and covers an area of 13,100km² occupying more than 85% of the Kelantan state. The river has a maximum length and breadth of 150km and 140km respectively. The Kelantan River comprises of a
few regions, which are Kuala Krai, Tanah Merah, Pasir Mas and Kota Bharu. Distribution of rainfall stations in Kelantan is illustrated in Figure 1 and the details of study area are shown in Table 1.

![Figure 1. Distribution of rainfall stations in study area.](image)

### Table 1. Details of rainfall stations used in the study area.

| Station Code | Station Name          | Record Period | Duration (Years) | Latitude  | Longitude |
|--------------|-----------------------|---------------|------------------|-----------|-----------|
| 40431        | Pos Blau              | 1994-2013     | 20               | 04° 39' N | 101° 41' E|
| 40432        | RPS Kuala Betis       | 1994-2013     | 20               | 04° 42' N | 101° 45' E|
| 40433        | Pos Hau               | 1994-2013     | 20               | 04° 42' N | 101° 32' E|
| 40470        | Pos Lebir             | 1994-2013     | 20               | 04° 56' N | 102° 23' E|
| 40516        | Pos Gob               | 1994-2013     | 20               | 05° 17' N | 101° 38' E|
| 40547        | Mardi Jeram Pasu      | 1994-2013     | 20               | 05° 48' 46" N | 102° 20' 40" E |
| 40663        | Pusat Pertanian Pasir Mas | 1994-2013 | 20               | 06° 02' N | 102° 07' E |
| 40666        | Mardi Kubang Keranji  | 1994-2013     | 20               | 06° 05' N | 102° 17' E |
| 48615        | Kota Bharu            | 1994-2013     | 20               | 06° 10' N | 102° 18' E |
| 48616        | Kuala Krai            | 1994-2013     | 20               | 05° 32' N | 102° 12' E |

Data used in the analysis were annual maximum rainfall series for 20 years duration recorded from 10 rainfall stations that were located all over the areas in Kelantan. These data were obtained from the Malaysia Meteorology Department (MMD). Preliminary screening was carried out to check the quality of rainfall data. Missing data of more than 10% are eliminated to ensure that the quality is good and future analysis will be improved.

### 3. Materials and methods

#### 3.1. Parameter estimation

The maximum likelihood estimation is a method that determines the parameter values of a model. The parameter values maximize the likelihood that the process described by the model produced the data
that were actually observed. The maximum likelihood estimate for the continuous probability distribution can be described as [19]:

$$\hat{\theta} \in \{\text{arg max } L(\theta; x)\}$$  \hspace{1cm} (1)

### 3.2. Fitting the probability distribution

Probability distribution is a concept related to the statistics, where the outcome of statistical experiments and their probabilities of occurrence are connected with the probability distributions. Four types of probability distributions were used, namely the Gamma, Gumbel, Generalized Extreme Value and the Log-Pearson Type-III to examine the rainfall data from Kelantan state in order to determine the best fit distribution.

The Gamma distribution is a widely used continuous probability distribution which is related to the beta distribution. The probability density function (PDF) and cumulative distribution function (CDF) of the Gamma distribution can be computed using the two formulas below [20]:

$$f(x) = \frac{(\frac{x-\mu}{\beta})^{\gamma-1} \exp\left(-\frac{x-\mu}{\beta}\right)}{\beta \Gamma(\gamma)}$$  \hspace{1cm} (2)

$$F(x) = \frac{\Gamma_x(\gamma)}{\Gamma(\gamma)}$$  \hspace{1cm} (3)

where \(\gamma\) is the shape parameter, \(\mu\) is the location parameter, \(\beta\) is the scale parameter and \(\Gamma\) is the gamma function.

The Gumbel distribution is also known as the Extreme Value Type-I distribution. The probability density function (PDF) and cumulative distribution function (CDF) of this distribution can be calculated using the two formulas below [15]:

$$f(x) = \frac{1}{\sigma} \exp\left(-\frac{x-\mu}{\sigma}\right) \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right)$$  \hspace{1cm} (4)

$$F(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right)$$  \hspace{1cm} (5)

where ‘\(\sigma\)’ and ‘\(\mu\)’ are the scale and location parameters respectively.

The Generalized Extreme Value (GEV) distribution is a well-known three parameter distribution for the maxima. The GEV distribution can be described as [21]:

$$F_{\xi, \mu, \sigma}(x) = \exp\left(-\left(1 + \xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}\right)$$  \hspace{1cm} (6)

$$\text{with } 1 + \xi\frac{x-\mu}{\sigma} > 0 \quad \xi \neq 0$$

where ‘\(\xi\)’, ‘\(\mu\)’ and ‘\(\sigma\)’ are shape, location and scale parameter respectively.

The Log Pearson Type-III (LP3) distribution is a probability distribution that is regularly used to fit the hydrological data in order to predict the design flood for a river. The probability density function
(PDF) and cumulative distribution function (CDF) of the LP3 distribution can be computed using the two formulas below [22]:

\[
f(x) = \frac{1}{x|\beta|\Gamma(\alpha)} \left( \frac{\ln(x) - y}{\beta} \right)^{\alpha-1} \exp \left( -\frac{\ln(x)}{\beta} \right)
\] (7)

\[
F(x) = \frac{\beta}{\Gamma(\alpha)} (\ln(x) - y)^{\alpha}
\] (8)

where \( \alpha \) is the shape parameter, \( \beta \) is the scale parameter and \( y \) is the location parameters.

3.3. Testing the goodness of fit

The Kolmogorov-Smirnov (K-S) test is a well-known goodness-of-fit test used to compare an empirical distribution function, \( \hat{F}_X \), with a specified distribution function \( F_Y \). The K-S test (D) can be calculated by using the formula below [23]:

\[
D = \max |\hat{F}_X(x) - F_Y(x)|
\] (9)

The Anderson-Darling (A-D) test is used to measure how well the data fits a specified probability distribution. This can be significant on selecting a probability distribution. The test statistic (\( A^2 \)) can be calculated from the formula below [24]:

\[
A^2 = -\sum_{i=1}^{n} \left[ (2i-1) \{ \ln F_X(x_i) + \ln [1 - F_X(x_{n+1-i})]/n \} - n \right]
\] (10)

where \( F_X \) is the cumulative distribution function of the specified distribution and \( x_i \) is the ordered data.

The Chi-Square (\( X^2 \)) test is used to determine if there is a relationship between two categorical variables. Arranging the number of observations \( N \) into a set of \( k \) cells is the first step and calculating the Chi-Square (\( X^2 \)) from the formula below is the second step of the Chi-Square goodness of fit test [20]:

\[
X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}
\] (11)

where \( O_i \) is observed frequency in the \( i \)th cell, \( E_i \) is expected frequency in the same cell and \( k \) is number of intervals.

4. Results and discussion

The methodology shown above was applied to the rainfall series collected from the 10 rainfall stations at the Kelantan River Basin. The study period was from 1994-2013 which is total of 20 years. The four probability distributions mentioned above were fitted to the annual maximum rainfall series of the 10 rainfall stations. The Gamma and the Gumbel distributions are simpler distributions consisting of two parameters while the Generalized Extreme Value and the Log-Pearson Type-III are complex distributions consisting of three parameters. This wide range of probability distributions can be used to explore and examine the best fit distribution for the annual maximum rainfall series.

The statistical parameters for the annual maximum rainfall data of the Pos Blau station are summarized in Table 2. The mean, variance, standard deviation, coefficient of variation, skewness,
kurtosis, maximum and minimum values are given. The parameters of probability distribution at the rainfall stations were calculated by the maximum likelihood estimation.

Table 2. Summary of statistics from the Pos Blau station.

| Statistical Parameters       | Pos Blau station |
|------------------------------|------------------|
| Mean                         | 107.695          |
| Variance                     | 1094.47          |
| Standard deviation           | 33.0828          |
| Coefficient of variation     | 0.30719          |
| Skewness                     | 0.69514          |
| Kurtosis                     | 0.24194          |
| Maximum value                | 186.4            |
| Minimum value                | 60.4             |

The three goodness of fit tests, namely Kolmogorov-Smirnov test (D), Anderson-Darling test ($A^2$) and Chi-Square test ($\chi^2$) for annual maximum rainfall data were used to evaluate the adequacy of four probability distributions. The results were ranked from 1 (best fit) to 4 (least fit) for all probability distributions. The selection of the best fit probability distribution is based on the lowest total test scores acquired by summing up the scores of each goodness of fit test.

The results of goodness of fit test at each rainfall stations are shown in Table 3. Those probability distributions having the same lowest test scores will be included in the selection of the best fit probability distribution. Based on the results of goodness of fit tests, the best-fit probability distribution for each station are shown in Table 4.

From Table 4, it can be observed that the Pos Lebir station has three best fit probability distributions compared to other stations. This is because the Gumbel, Generalized Extreme Value and the Log-Pearson Type-III distributions performed equally good at Pos Lebir station. For the RPS Kuala Betis station, the Log-Pearson Type-III distribution performed the best in describing the characteristics of annual maximum rainfall at the RPS Kuala Betis station. The Gamma and the Gumbel distributions seem to be less frequently being selected as the best fit probability distribution in this study due to its number of parameters. This is because these distributions only have two parameters which are not so flexible as compared to the Generalized Extreme Value and the Log-Pearson Type-III distributions. Thus, the ability of these distributions can only to capture less climatic characteristics.

Overall, the Generalized Extreme Value distribution is the best fit probability distribution for the annual maximum rainfall series at most of the stations at the Kelantan River Basin. The results are similar with some of the literature findings. For example, in Bangladesh, the Generalized Extreme Value distribution was found to be the best fit distribution for the annual maximum rainfall in the research by [9]. A possible explanation is that the Generalized Extreme Value distribution is a complex distribution which is more flexible on capturing more climatic characteristics. Furthermore, the Generalized Extreme Value distribution is able to capture the behaviour of extreme rainfall characteristics. Therefore, the Generalized Extreme Value distribution is recommended to simulate the annual maximum rainfall data at the Kelantan River Basin.
### Table 3. Results of goodness of fit tests.

| Station         | Distribution model | Rank | Score |
|-----------------|--------------------|------|-------|
|                 |                    | I    | II    | III   |
| Pos Blau        | Gamma              | 4    | 4     | 4     | 12    |
|                 | Gumbel             | 2    | 3     | 3     | 8     |
|                 | GEV                | 3    | 1     | 1     | 5     |
|                 | Log Pearson type III | 1    | 2     | 2     | 5     |
| RPS Kuala Betis | Gamma              | 1    | 1     | 4     | 6     |
|                 | Gumbel             | 4    | 4     | 3     | 11    |
|                 | GEV                | 3    | 3     | 2     | 8     |
|                 | Log Pearson type III | 2    | 2     | 1     | 5     |
| Pos Hau         | Gamma              | 1    | 3     | 4     | 8     |
|                 | Gumbel             | 3    | 4     | 3     | 10    |
|                 | GEV                | 4    | 1     | 1     | 6     |
|                 | Log Pearson type III | 2    | 2     | 2     | 6     |
| Pos Lebir       | Gamma              | 3    | 4     | 2     | 9     |
|                 | Gumbel             | 1    | 3     | 3     | 7     |
|                 | GEV                | 2    | 1     | 4     | 7     |
|                 | Log Pearson type III | 4    | 2     | 1     | 7     |
| Pos Gob         | Gamma              | 3    | 3     | 4     | 10    |
|                 | Gumbel             | 4    | 4     | 2     | 10    |
|                 | GEV                | 2    | 1     | 2     | 5     |
|                 | Log Pearson type III | 1    | 2     | 3     | 6     |
| Mardi Jeram Pasu| Gamma              | 3    | 3     | 4     | 10    |
|                 | Gumbel             | 4    | 4     | 2     | 10    |
|                 | GEV                | 1    | 1     | 1     | 3     |
|                 | Log Pearson type III | 2    | 2     | 3     | 7     |
| Pusat Pertanian Pasir Mas | Gamma | 1 | 3 | 4 | 8 |
|                 | Gumbel             | 4    | 4     | 1     | 9     |
|                 | GEV                | 3    | 1     | 2     | 6     |
|                 | Log Pearson type III | 2    | 2     | 3     | 7     |
| Mardi Kubang Keranji | Gamma | 3 | 3 | 4 | 10 |
|                 | Gumbel             | 4    | 4     | 2     | 10    |
|                 | GEV                | 1    | 1     | 3     | 5     |
|                 | Log Pearson type III | 2    | 2     | 1     | 5     |
| Kota Bharu      | Gamma              | 4    | 4     | 3     | 11    |
|                 | Gumbel             | 3    | 3     | 4     | 10    |
|                 | GEV                | 1    | 1     | 1     | 3     |
|                 | Log Pearson type III | 2    | 2     | 2     | 6     |
| Kuala Krai      | Gamma              | 3    | 3     | 2     | 8     |
|                 | Gumbel             | 4    | 4     | 4     | 12    |
|                 | GEV                | 1    | 1     | 1     | 3     |
|                 | Log Pearson type III | 2    | 2     | 3     | 7     |

*Rank I, II and III denoted by Kolmogorov Smirnov test, Anderson Darling test and Chi - Squared test respectively.*

*Bolded values refer to the best fit probability distribution.*
Table 4. Best fit probability distribution for each station.

| Station              | Best fit probability distribution                                      |
|----------------------|-------------------------------------------------------------------------|
| Pos Blau             | Generalized Extreme Value                                               |
| RPS Kuala Betis      | Log Pearson type III                                                     |
| Pos Hau              | Generalized Extreme Value                                               |
| Pos Lebir            | Gumbel & Generalized Extreme Value & Log Pearson type-III               |
| Pos Gob              | Generalized Extreme Value                                               |
| Mardi Jeram Pasu     | Generalized Extreme Value                                               |
| Pusat Pertanian Pasir Mas | Generalized Extreme Value                              |
| Mardi Kubang Keranji | Generalized Extreme Value                                               |
| Kota Bharu           | Generalized Extreme Value                                               |
| Kuala Krai           | Generalized Extreme Value                                               |

5. Conclusion
The focus of this study was to investigate the best fit probability distribution for the annual maximum rainfall in tropical Peninsular Malaysia, specifically for the Kelantan River Basin. The Gamma, Gumbel, Generalized Extreme Value and the Log Pearson Type-III distributions were fitted to describe the annual maximum rainfall at the Kelantan River Basin. Their performances were evaluated by using goodness of fit tests. It was found that the Generalized Extreme Value distribution is the best fit distribution to characterize the annual maximum rainfall series as it offers the advantage of being efficient and flexible. This study was limited by the availability of recorded data at best was 20 years. A longer period of historical rainfall data is better and recommended (more than 20 years), and a search for other appropriate catchments with such data is envisaged.

Acknowledgements
The authors are appreciative of and thankful to the Malaysian Meteorological Department (MMD) for providing the daily rainfall data. The authors also would like to acknowledge the sincere appreciation towards the financial support from UCSI University through Pioneer Scientist Incentive Fund (PSIF) with project code Proj-2019-In-FETBE-065.

References
[1] Douka M T S Karacostas E Katragkou and C Anagnostopoulou 2017 Annual and Seasonal Extreme Precipitation Probability Distributions at Thessaloniki Based Upon Hourly Values in Perspectives on Atmospheric Sciences (Springer, Cham) 521–527
[2] Li H D Wang V P Singh Y Wang J Wu J Wu J Liu Y Zou R He and J Zhang 2019 Journal of Hydrology 571 114–131
[3] Ng J L S Abd Aziz Y F Huang A and Wayayok M D 2015 Jurnal Teknologi 76 1–6
[4] Ng J L S Abd Aziz Y F Huang A Wayayok M K and Rowshon S 2016 Stochastic Environmental Research & Risk Assessment 31 (9) 2215-2233 19
[5] Kayode J S M H Arifin M K A Kamarudin A Hussin M N M Nordin and N Roslan 2019 Arabian Journal of Geosciences 12
[6] Ng J L S Abd Aziz Y F Huang A Wayayok and M D Rowshon, Analysis of annual maximum rainfall in Kelantan, Malaysia, Acta Hortic, 1152, 11–17 (2017)
[7] Ng J L S Abd Aziz Y F Huang M Mirzaei A Wayayok and M K Rowshon 2019 Journal of Earth System Science 128
[8] Yuan J K Emura C Farnham M A Alam 2018 Urban Climate 24 276–286
[9] Khudri M and F Sadia European Journal of Scientific Research 2013 103 391–404
[10] Bhakar S A Bansal and N Chhajed 2008 Journal of the Institution of Engineers (India): Agricultural Engineering Division 89 14–16
[11] Upadhaya A and S Singh 1998 Indian Journal of Soil Conservation 26 193-201
[12] Aksoy H 2000 Turkish Journal of Engineering and Environment Sciences 24 419-428
[13] Ben-Zvi A 2009 Journal of Hydrology 367 104-114
[14] Sharma M and J Singh 2010 3 40-49
[15] Varathan N K Perera and N Wikramanayake 2010 Statistical modeling of extreme daily rainfall in Colombo (University of Moratuwa Sri Lanka: MSc Thesis)
[16] Schaefer M 1990 Water Resources Research 26 119-131
[17] Rulfová Z A Buishand M Roth and J Kyselý 2016 Journal of Hydrology 534 659–668
[18] Hazarika S P Borah A Prakash 2019 Atmospheric Environment 202 53–63
[19] Martins E and J Stedinger 2011 Water Resources Research 36 737-744
[20] Alghazali N O and D A H Alawadi 2014 Civil and Environmental Research 6 40-46
[21] Porto de Carvalho J R E D Assad A F de Oliveira and H Silveira Pinto 2014 Weather Clim. Extrem. 5 7–15
[22] Amin M Izwan A Alazba 2016 Open Life Sciences 11 432-440
[23] Guizani M A Rayes B Khan and A Al-Fuqaha 2010 Network modeling and simulation: a practical perspective (United States: John Wiley & Sons) 1st ed
[24] Anderson T 2010 Anderson-Darling tests of goodness-of-fit in International Encyclopedia of Statistical Science (Berlin, Heidelberg: Stanford University Press, Springer)