THE BEST FIT PROBABILITY DISTRIBUTION MODEL FOR THE ESTIMATION OF EXTREME RAINFALL IN LIMBANG, SARAWAK

T. Tahir[1], A. M. Hashim[2], H. Takaijudin[1*], K. Wan Yusof[1], M. Osman[1]

[1]Department of Civil and Environmental Engineering, Universiti Teknologi PETRONAS, Perak, Malaysia
[2]Global Water Consultants Sdn. Bhd., Kuala Lumpur, Malaysia

Email: husna_takaijudin@utp.edu.my

ABSTRACT

In Malaysia, the increment of annual rainfall patterns is causing frequent floods, mainly in Sabah and Sarawak. Limbang river basin was selected as a case study due to it was facing of high-risk flooding problem mainly during the transition of climate. This study was aimed to estimate the frequency of rainfall under various return periods and to identify the best fit model probability distribution of annual maximum rainfall based on twenty-four hours sample in Limbang. The three statistical models were used, which are Gumbel, Log-Pearson type III, and Log-Normal. Based on the goodness of fit tests, Chi-Square, Kolmogorov Smirnov test, and the Log-Normal was found to be the best fit model for the station of Panduran. The Log-Pearson type III was found to be the best-fit distribution model for the rest of the stations, which occupies almost more than 90%. The maximum values of expected rainfall were calculated using the best fit probability distributions and could be used by a design engineer in the future.

Keywords: *flood, return period, Gumbel distribution*

INTRODUCTION

The planning and building of various projects viz, urban drainage work and dams, the water resources system and the flood damage prevention, need abundant information on the return period of heavy rainfall intensities. So, it is pertinent to check the trend of rainfall frequency; therefore, a statistical approach is applied to available data to obtain the desired parameters of chosen trend distribution [1]. Then with the help of this technique, the return periods for a future event are extracted out, and that will give an idea about the future scenario as well as its mode of intensity [2]. This is why it is crucial to work on early frequency analysis in order to estimate the havoc caused by heavy rainfall, and it will also help to design a more accurate and efficient hydraulic structure.

Although the maximum annual rainfall in Limbang reaches up to 4500 mm, there is a large gap between the occurrences of rainfalls and are not uniformly distributed transiently and spatially [3]. Transiently there are two main seasons that receive more rainfall as compared to the rest of the year, and these continuous and heavy spells are highly responsible for flash flood as well as seasonal flood. Spatially, it is a coastal area, and naturally, it receives more rainfall. All these causes are made an area more crucial to do early work on extreme rainfall frequency analysis. According to our best information and review literature, Nowadays, Malaysia experiences more frequent extreme rainfall. Recently, it was reported that an extraordinary flood happened in most of the area in Pahang state in January 2021. This was due to heavy down pours (approximately 280 mm) for more 12 hours duration, which lead to the increment of water level at Sungai Pahang [4]. According to Othman [5], the highest extreme rainfall in 44 years (1970-2014) was 158.80 mm rainfall which was happened in 1988. Hence, it can be concluded that, there was an increment trends happened after 33 years in 2021 which approximately about 43% increment.

The statistical approach of the extreme value theory is the most appropriate method to determine the behaviour of extreme rainfall [6]. There are several kinds of techniques that have been developed to analyse the frequency of rainfall viz, Gumbel, Pearson, Log-Pearson [7], Log-Normal, and so on. However, the selection of an appropriate model still quite a difficult task in engineering practices as there is no specific covenant that one can rely on and used in rainfall frequency analysis. Moreover, the characteristics of accessible rainfall data also very relatable in the selection of a suitable model. Therefore, it is pertinent to perform the analysis with a bunch of distributions method in order to measure the precise extreme occasions of rainfall [8]-[10].

A recent study was done to find out the possibility of extreme events of rainfall for various intervals in the northern areas of Pakistan. They applied four models of distribution. The suitability of the results obtained from the models was verified by applying the Chi-Square, Kolmogorov Smirnov test, and Anderson Darling test. The test results indicated that Log-Pearson III was measured intervals more precisely in most of the region except on one station where Log-Normal results were more authentic [9].

This paper aims to estimate the frequency of extreme rainfall under various return period and later to assess the best probability models that fit to the extreme rainfall analysis for Limbang river basin. In a study, the person used Gumbel, Log-Gumbel, Pearson, Log-Pearson, Normal and Log-Normal statistical investigation over 20 different stations having 54-year data. The statistical investigation was applied individually in every station to accomplish the frequency analysis of rainfall. After predicting the peak rainfall, the better test coefficient of determination, chi-square, and Fisher's test was applied to the sepredicted rainfall frequencies. The results indicated that Log-Pearson type III was best among all others [7],[11].

METHODOLOGY

Site Description

The Limbang river basin, with an estimated area of 4200 km^2 is in the northeast of Sarawak, Malaysia, with a longitude of 4°45′0″ North and latitude 115°0′0″ towards East. It is on the list of one of the largest river catchments in Sarawak. Limbang river basin was selected as a study area since it was facing a high risk of flooding due to the transition in climate. The region in cluded tourist spots, land agriculture, rainforest and so on. With little variation, the atmosphere Limbang can be divided into four separate seasons as it is a tropical region. Two of these four seasons are monsoon seasons, with the other two being inter-monsoon seasons.

Data Collection

It is necessary to have precipitation data in the long period of time mainly for extreme rainfall analysis which offers more reliable result [7]. The daily observed rainfall data for the nominated stations were collected from the Department of Irrigation and Drainage Sarawak (DID). The rainfall data were recorded in (mm) each day for the whole year. The rainfall data were selected for the period of 28 years from 1976-2003. Firstly, missing data is completed with the help of a climtol tool in R studio in order to obtain comprehensive data [12]. Nine stations were nominated in different locations of Limbang for the purpose of this study. The geographical representation of these stations is shown in Figure 1. These are Lubai, Lubok, Medamit, Merabu, Napir, Panduran, Rutho, Setuan, Tegari, Ukong, and ULU. The extreme events summary statistic is mentioned in Table 1. Secondly, the data were plotted or arranged according to Welbull's theory to estimate the return period [13]. Then different frequency analysis methods were applied to find out the return period of rainfall and are frequency for the intervals of 2, 5, 10, 25, 50, 75, 100, 500, and 1000. These are Gumbel, Log–Pearson type III, and Log-Normal, which were identified as the most appropriate approaches for extreme rainfall analysis. Moreover, to find out the best from these three applied models' goodness of fit test Chi-Square (x2), Kolmogorov Smirnov test and were applied.

Frequency Analysis

a) *Gumbel Distribution*: The Gumbel distribution method was applied in a selected station to find out the frequency analysis for the selected return intervals. It is the most frequently used technique in frequency analysis. The technique is simple as compared to the other techniques. This expression of the Gumbel to calculate rainfall RT (mm) for the return interval of T in (year) is given by:

$$R_t = R_{avg} + KS \tag{1}$$

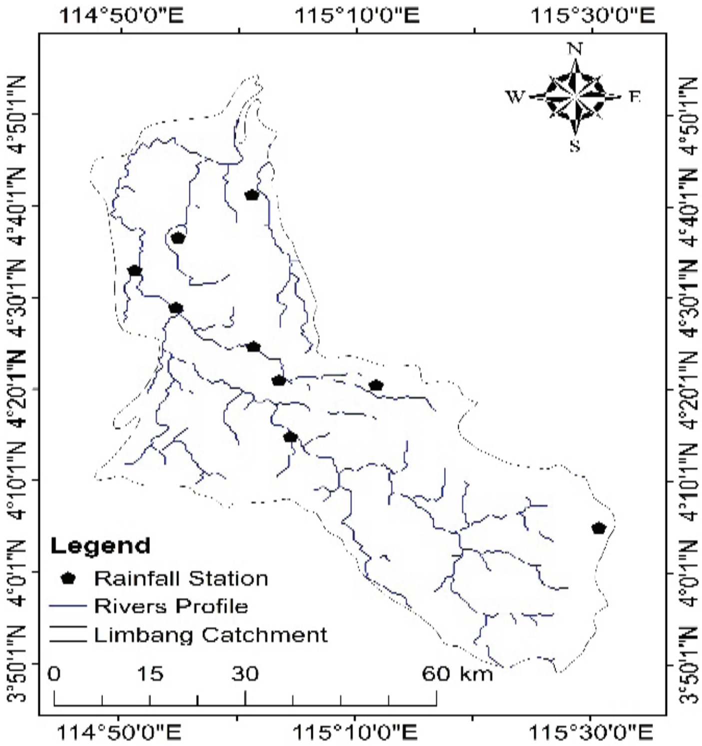

Figure 1 The geographical location of stations

Table 1 Summary of statistics of extreme daily rainfall

Station Names	Mean value (mm)	Standard deviation (mm)	Skewness coefficient	Coefficient of variation	Maximum (mm)	Minimum (mm)
Lubai	97.20	40.41	1.76	0.41	247	63.1
Lubok	89.33	17.91	0.27	0.20	123	62.8
Medamit	106.10	22.46	-0.005	0.21	162.3	68.2
Merabu	98.60	23.28	0.26	0.24	143.6	62.2
Panduran	121.77	27.65	-0.20	0.23	183.5	72.5
Rutho	75.04	17.90	-0.04	0.23	118.3	40.1
Setuan	83.52	23.72	0.50	0.28	153	48
Ukong	122.78	21.09	0.04	0.17	170.5	90
ULU	109.83	24.32	0.23	0.22	175.9	68.2

where K is the frequency factor for Gumbel and given as:

$$K = -\frac{\sqrt{6}}{\pi}\left[0.5772 + \ln\left[\ln\left[\frac{T}{T-1}\right]\right]\right] \quad (2)$$

While R_{avg} is the extreme average of rainfall related to a definite duration. So arithmetic average used in Equation 1.

$$R_{avg} = \frac{1}{n}\sum_{i=1}^{n} R_i \quad (3)$$

where R_i is the individual value of extreme rainfall and n represents the total number of years recorded. To calculate the standard deviation for Gumble the following equation isused

$$S = \left[\frac{1}{n}\sum_{i=1}^{n}(R_i - R_{avg})^2\right]^{1/2} \quad (4)$$

where S represents the standard deviation of R data, so, when the standard deviation (S) is multiplied by the frequency factor (K) and then the addition of

R_{avg} in it. This will give the frequency of rainfall for the selected period.

b) *Log-Pearson type*: Log-Pearson type III (LPT III) model is used to compute the frequency of rainfall concerning the time period to estimate the intensity of rainfall for the selected duration. In this technique, the logarithm is applied to the observed rainfall. Then mean, maximum, and standard deviation are computed with the help of these logarithmic data. The basic expressions which are required to interpret this model are given by:

$$R_t^* = R_{avg}^* + K_t S^* \tag{5}$$

$$R_{avg} = \frac{1}{n} \sum_{i=1}^{n} R^* \tag{6}$$

$$R^* = \log(R_t) \tag{7}$$

$$S = \left[\frac{1}{n-1} \sum_{i=1}^{n} (R_t^* = R_{avg}^*)^2 \right]^{1/2} \tag{8}$$

where R_t^*, R_{avg}^*, and S^* are explained earlier in the above-mentioned method, the only difference is that these rely on logarithmically converted values. While the Pearson factor for frequency is not calculated directly as in the previous case, for this, it depends on the coefficient of skewness (C_S) and the return period (*T*).

Therefore, in order to measure the frequency factor first, it is important to determine the coefficient of skewness C_S. The C_S is determined with the help of the following formula:

$$C_S = \frac{n \sum_i^{ni} (P_i^* - P_{avg}^*)^3}{(n-1)(n-2)(S^*)^3} \tag{9}$$

The values of is available in a number of hydrological books, for instance. Then interpolation is applied with the known values of C_S in order to extract the for LPT III for the desired return period. The required extreme event for the desired return period will be calculated by taking the antilog of the Equation 7.

c) *Log-Normal*: Log-Normal (LN) distribution is a distribution of indiscriminate variables by simply applying the customarily distributed logarithm. This method has been widely used to identify the frequency of rainfall for the selected time period. Therefore, in order to calculate the maximum frequency of rainfall by applying the Log-Normal method, the following equation is used:

$$R_t = R_{avg} + K_t C_V \tag{10}$$

$$R_{avg} = \frac{1}{n} \sum_{i=1}^{n} R^* \tag{11}$$

$$R^* = \log(R_t) \tag{12}$$

where C_V is the coefficient of variance, and it is obtained by dividing the standard deviation *S*, with the mean value of the data, as in our case is rainfall. The expression for *S* is the same as mentioned above in the section. The frequency analysis followed the universal procedure that was previously adopted by a number of the researcher. By following this method, the mathematical expressions were developed for all the above-said models that were used in this study, which are presented in Table 2.

Table 2 The mathematical equations for an extreme event

Station	Gumbel	Log-Normal	Log-Pearson type III
Lubai	$R_t = 97.20 + 40.41k$	$R_t^* = 1.96 + 0.13k$	$R_t^* = 1.96 + 0.13k^*$
Lubok	$R_t = 89.33 + 17.01k$	$R_t^* = 1.94 + 0.085k$	$R_t^* = 1.94 + 0.085k^*$
Medamit	$R_t = 106.10 + 22.46k$	$R_t^* = 2.01 + 0.09k$	$R_t^* = 2.01 + 0.091k^*$
Merabu	$R_t = 98.60 + 23.28k$	$R_t^* = 1.98 + 0.099k$	$R_t^* = 1.98 + 0.099k^*$
Panduran	$R_t = 121.77 + 27.65k$	$R_t^* = 2.07 + 0.10k$	$R_t^* = 2.07 + 0.10k^*$
Rutho	$R_t = 95.04 + 17.90k$	$R_t^* = 1.86 + 0.10k$	$R_t^* = 1.86 + 0.10k^*$
Setuan	$R_t = 83.52 + 23.72k$	$R_t^* = 1.90 + 0.11k$	$R_t^* = 1.90 + 0.11k^*$
Ukong	$R_t = 122.78 + 21.09k$	$R_t^* = 2.08 + 0.07k$	$R_t^* = 2.08 + 0.07k^*$
ULU	$R_t = 109.83 + 24.32k$	$R_t^* = 2.03 + 0.09k$	$R_t^* = 2.03 + 0.09k^*$

Probability distribution models have been applied to three (3) statistical tests in order to evaluate the best-fit models at individual rainfall station. Chi-square (χ_2), Kolmogorov Smirnov, and coefficient of determination (*R2*) are used as statistical measures. In line with the normal, the statistical tests were performed, which aligned with the standard procedure [14]-[16]. The evaluation of the probability distribution models was based on the cumulative test score gathered from all the studies. Each distribution model is awarded test scores ranging from zero to three (0-3) according to the criterion that the distributions with the highest total score for the data of a particular station or it is selected as the best distribution models. Typically, a score of three (3) is conferred to the distribution best supported by a test, the next best is given a score of two (2) and so on in descending order. The overall ranks of each distribution were assembled for each test group by adding the individual point ranks at everyone of the nine (9) stations by following the same process as the other stations [17]-[18].

RESULTS AND DISCUSSION

Twenty-eight years rainfall data were gathered for nine (9) station in Limbang river basin. Based on these data, only extreme rainfall data was chosen for this study. Statistical tools (goodness of fit tests) were applied to determine the best-fit probability function. Table 3 indicates the results of the statistical

Table 3 Best test score at individual stations

Station	Distribution Test	Chi-square	Kolmogorov Smirnov	R^2	Total
Lubai	Gumbel	1	1	1	3
	Log-Pearson III	3	2	3	8
	Log-Normal	2	3	2	7
Lubok	Gumbel	1	1	1	3
	Log-Pearson III	3	3	3	9
	Log-Normal	2	2	2	6
Medamit	Gumbel	1	1	1	3
	Log-Pearson III	3	3	2	8
	Log-Normal	2	2	3	7
Merabu	Gumbel	1	2	1	4
	Log-Pearson III	3	3	3	9
	Log-Normal	2	1	2	5
Panduran	Gumbel	1	2	2	5
	Log-Pearson III	2	3	1	6
	Log-Normal	3	1	3	7
Rutho	Gumbel	1	1	1	3
	Log-Pearson III	2	3	2	7
	Log-Normal	3	2	3	8
Setuan	Gumbel	1	1	1	3
	Log-Pearson III	3	3	3	9
	Log-Normal	2	2	2	6
Ukong	Gumbel	2	1	1	4
	Log-Pearson III	3	3	2	8
	Log-Normal	1	2	3	6
ULU	Gumbel	1	1	1	3
	Log-Pearson III	2	3	3	8
	Log-Normal	3	2	2	7

Table 4 Projection of extreme rainfall intensity (unit in mm/hr) based on various return period.

Station	Best model	Return Period in Years							
		2	5	10	25	50	75	100	500
Lubai	Log-pearson III	84.11	101.59	139.19	182.98	224.73	248.58	275.13	401.86
Lubok	Log-pearson III	86.89	104.56	123.42	137.03	146.97	151.74	156.73	174.64
Medamit	Log-pearson III	103.82	114.94	136.16	150.32	160.20	164.88	169.65	186.20
Merabu	Log-pearson III	96.59	106.86	129.74	146.60	159.04	165.09	171.40	194.34
Panduran	Log-Normal	118.72	132.66	159.68	177.97	190.89	196.99	203.28	225.15
Rutho	Log-pearson III	72.92	81.87	98.78	110.11	118.04	121.78	125.61	138.92
Setuan	Log-pearson III	78.70	90.48	114.28	133.93	149.31	157.16	165.42	197.46
Ukong	Log-pearson III	120.84	131.40	151.45	164.63	173.82	178.12	182.53	197.72
ULU	Log-pearson III	106.48	118.55	142.04	158.99	171..36	177.31	183.55	205.93

test score (the goodness of fit tests) and the best fit models, respectively. These findings were used to assess which of the probability model(s) best fit the extreme daily rainfall at each station. For example, highest rank was obtained using Log Pearson III for three (3) stations i.e. Lubok, Merabu and Setuan. Contradictly, the Gumble probability model shows the lowest rank for the selected station which indicated that this model has poor performance as a prediction tool. Assessment of the goodness-of-fit test results proved that in many cases, there was almost no distinction between the various distributions for the individual station. Besides, the good fit evaluation for all stations also showed that no one distribution ranked reliably best at all locations. Nonetheless, the overall ranks for the eight (8) stations combined show that Log-Pearson type III was ideal for depicting the extraordinary rainfall. Conversely, the overall rank of one (1) station indicates that Log-Normal was ideal to portray the extreme rainfall. This was explained that the Log-Pearson III and Log-Normal models might be appropriately used as the best prediction tools for extreme precipitation for Limbang area.

Table 4 shows the quantile of extreme rainfall estimation for the selected station based on various return periods. The projection was extended for 500 years ARI by using the the best-fit probability model which generated from Table 3. Based on the table, it shows that Lubai station was predicted to receive maximum rainfall intensity which indicated by the increment at a return period of 500 years. This was explained that the respective station are potentially exposed to the occurrence of flood. The other stations were expected to have consistent trends across a return period.

CONCLUSION

Based on the results, it was found that the best frequency distribution gathered for the extreme daily rainfall in Limbang was the log-Pearson type III distribution, which dominates roughly more than 90%. Three (3) goodness-of-fit tests i.e. Chi-square, Kolmogorov Smirnov, and coefficient of determination (R^2) were successfully used to determine the significance of the findings. The proposed assessment procedure has been successfully used to discover the best probability distributions that could provide accurate extreme rainfall estimation for Limbang. The frequency analysis suggest log-Pearson type III and Log-Normal distributions have the primary distribution pattern for this study site and should be used as a universal distribution model for the prediction of peak daily rainfall in Limbang. The projection of extreme rainfall shows a significant trend for Limbang which indicates consistent increment of rainfall pattern across the return period.

ACKNOWLEDGMENT

This project is fully funded by national grant UKM-YSD with cost centre 015MC0-004. The authors would like to thank University Teknologi PETRONAS for providing the facilities and also thankful to the Department of

Irrigation and Drainage Sarawak for providing the required data to accomplish this study.

REFERENCES

[1] M. Nyatuame, V. Owusu-Gyimah & F. Ampiaw, Statistical analysis of rainfall trend for Volta Region in Ghana, *International Journal of Atmospheric Sciences*, 2014.

[2] B. Singh, *et al*., Probability analysis for estimation of annual one day maximum rainfall of Jhalarapatan area of Rajasthan, India, *Plant Archives*, 12, 2, pp. 1093-1100, 2012.

[3] D. S. N. A., binti Pg. Ali *et al.*, Prediction of climate change in Brunei Darussalam using statistical down scaling model, *Theoretical and Applied Climatology*, 133, 1, pp. 343-360, 2018.

[4] Alagesh, T., *Pahang records slight increase in flood victims despite promising weather*, in New Straits Times: Kuantan, Pahang, 2021.

[5] Othman, M.A., *et al*., Analysis of Trends of Extreme Rainfall Events Using Mann Kendall Test: A Case Study in Pahang and Kelantan River Basins, *Jurnal Teknologi*, 78, 9-4, 2016.

[6] G. Lazoglou, *et al*., A review of statistical methods to analyze extreme precipitation and temperature events in the Mediterranean region, *Theoretical and Applied Climatology*, 136, 1, pp. 99-117, 2019.

[7] S. El Adlouni & T. Ouarda, Frequency analysis of extreme rainfall events. *Rainfall: State of the science*, 191, pp.171-188, 2010.

[8] I. H. Elsebaie, Developing rainfall intensity–duration–frequency relationship for two regions in Saudi Arabia, *Journal of King Saud University-Engineering Sciences*, 24, 2, pp. 131-140, 2012.

[9] M. Amin, M. Rizwan & A. Alazba, A best-fit probability distribution for the estimation of rainfall in northern regions of Pakistan, *Open Life Sciences*, 11, 1, pp. 432-440, 2016.

[10] C. Lee, Application of rainfall frequency analysis on studying rainfall distribution characteristics of Chia-Nan plainarea in Southern Taiwan, *Crop Environ. Bioinf*, 2, 1, pp. 31-38, 2005.

[11] O. Olofintoye, B. Sule & A. Salami, Best-fit probability distribution model for peak daily rainfall of selected citiesin Nigeria, *New York Science Journal*, 2, 3, pp. 1-12 2009.

[12] J. A. Guijarro, Homogenization of climatic series with Climatol. Version 3.1.1. [Online]. Available: https://cran.r-project.org/web/packages/climatol/index.html, 2018.

[13] K. A. Berthe, B. Abdramane & S. Reichenbach, Gumbel Weibull distribution function for Sahel precipitation modeling and predicting: Case of Mali, *African Journal of Environmental Science and Technology*, 9, 5, pp. 405-412,2015.

[14] M. A. Karim & J.U. Chowdhury, A comparison of four distributions used in flood frequency analysis in Bangladesh, *Hydrological Sciences Journal*, 40, 1, pp. 55-66, 1995.

[15] S. Xeflide & D. Ophori, Return Period Analysis As A Tool For Urban Flood Prediction In The Accra Plains, Southern Ghana, *Journal of the International Association for Environmental Hydrology*, 16, 15, pp. 1-8, 2008.

[16] Z. Hassan & T. Ajibola, Statistical analysis of some meteorological variables data for Sokoto and its vicinity, *World Scientific News*, 24, pp. 8-17, 2015.

[17] L. Oyebande, Deriving rainfall intensity-duration-frequency relationships and estimates for regions with in adequate data, *Hydrological Sciences Journal*, 27, 3, pp. 353-367, 1982.

[18] L. S. Hanson & R. Vogel, The probability distribution of daily rainfall in the United States, in *World Environmental and Water Resources Congress 2008: Ahupua'A*, 2008.