MATEMATIKA, 2013, Volume 29, Number 1b, 43-52
©Department of Mathematical Sciences, UTM

Determination of Best-fit Distribution and Rainfall Events in Damansara and Kelantan, Malaysia

[1]Ho Ming Kang and [2]Fadhilah Yusof
[1,2]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Johor
e-mail: [1] ryanho1025@gmail.com, [2]fadhilahy@utm.my

Abstract This paper presents a study to determine the best-fit distribution to represent the rainfall process in Damansara and Kelantan, Malaysia. Three probability density functions, namely Wakeby distribution, Generalized Extreme Value function (GEV) and two-parameter Weibull distributions are selected and compared. The parameters of the distributions are estimated using L-moments method while the best-fit distribution is determined by using Kolmogorov-Smirnov goodness-of-fit test. In addition, weighted-average algorithm which is based on the probability values from the stations in Damansara and Kelantan is used to identify the occurrence of wet and dry events in the rainfall data. The impact of different distributions used in the determination of rainfall events is evaluated by making comparison between the actual and the reconstructed rainfall data. The results indicate that the Wakeby distribution is the best-fit distribution to explain the rainfall patterns in Damansara and Kelantan. However, Wakeby, GEV and Weibull distributions perform equally well in the estimation of wet and dry events in Damansara and Kelantan.

Keywords Wakeby Distribution; Generalized Extreme Values; Weibull Distributions; L-moments; Weighted Average; Kolmogorov-Smirnov Goodness-of-fit Test.

2010 Mathematics Subject Classification 46N60, 92B99

1 Introduction

Malaysia is a country with unique rainfall patterns and characteristics than other countries in the world. In particular, Malaysia deals with two distinct monsoon seasons that are southeast monsoon that occurs from May to September, and northwest monsoon that occurs from November to February. The rapid changes in the climate have consistently increased the number of extreme floods in Malaysia, especially in the highly populated areas. The high loss in economics especially in agriculture sectors have significant impacts to the country development and that inspires the researchers to investigate and model the rainfall process to further understand the changes and the characteristics of rainfall in the past. The rainfall processes that have always been investigated include the estimation of the rainfall distribution and the identification of wet or dry events on a particular day. This is important to be highlighted as it will not only used in planning the water resources, but it can also be use to improve the sensitivity of the rainfall systems and managements in a country.

The selection of the best-fit distribution of the rainfall process is always the main interest in the study of hydrology. Hanson and Vogel [1] stated that the best distribution to represent daily precipitation in United States is Pearson Type-III distribution while Kappa distribution is best to describe the patterns of wet-day daily rainfall. Sharma and Bhagwan Singh [2] investigated that lognormal and gamma distribution are the best fit probability distribution for annual and monsoon period, and generalized extreme value distribution is the best for weekly period in India. Modarres [3] stated that 3-parameter of lognormal distribution is the best to describe the rainfall pattern in Iran. Meanwhile [4] showed that 2-parameter of gamma distribution is the most fitted to the daily rainfall.

In Malaysia, there is also e several studies investigated on the distribution of rainfall, either hourly, daily or annually. Fadhilah et. al. [5] showed that the best-fit distribution for hourly rainfall amount in Wilayah Persekutuan, Malaysia is Mixed-Exponential distribution while the study in [6] used Generalized Pareto distribution on the hourly rainfall data of five stations in Peninsular Malaysia. Dan'azumi et. al. [7] divided the data according to six hour times and used generalized pareto (GP), exponential, beta and gamma distributions to model the hourly rainfall intensity of Peninsular Malaysia. The results indicated that GP is the best to describe the pattern of hourly rainfall intensity. Meanwhile in the study of [8], they also found that Mixed Exponential is the most appropriate distribution to represent the daily rainfall amount in Peninsular Malaysia. In addition, they investigated that mixed lognormal is another favourable distribution to describe the patterns of daily rainfall amount in Malaysia [9]. For annually maximum rainfall data in Peninsular Malaysia, [10] stated that generalized extreme value (GEV) distribution is the more suitable to be used. However, [11] presented that generalized logistic and generalized pareto distributions are the most frequently distribution for annual data of maximum daily rainfall in Peninsular Malaysia. Moreover, [12] fitted Generalized Extreme Value (GEV) distribution to the data that consists of annual maximum data of extreme rainfall in Alor Setar, Kedah for the analysis.

Due to the high frequency of flood occurs in Malaysia, exposure of the models related to the extreme events is studied. For this purpose, GEV distribution is selected for the evaluation since many studies have proven its suitability to represent the rainfall patterns in Malaysia. In addition, Wakeby distribution is chosen as it is always related and commonly used in the study of flood analysis [13]. Besides, they are compared with two-parameter Weibull distributions because Weibull is belongs to the family of extreme value distributions and frequently involved in extreme events and weather forecasting. To evaluate the performance of the distributions and to determine the best describe of the daily rainfall process, Kolmogorov Smirnov (KS) goodness-of-fit test is applied. The reason to select KS as the evaluation criteria is because it treats the observation individually in order to avoid the loss of information due to grouping. It can also provide high accuracy even with small sample size.

This study will adopt the method by [14] where the weighted-average algorithm from the probability distribution of the surrounding stations will be constructed in order to get a complete replica of a target station estimated probability. The wet and dry events in the target data are then identified by obtaining the threshold for each day. The impact of the distributions used in identifying the dry and wet events in the rainfall process is observed and compared with the observed rainfall events.

2 Study Area

Data from two river basins namely Sg. Damansara and Sg. Kelantan are obtained from the Drainage and Irrigation Department (DID) of Malaysia. Kelantan river basins are located in the eastern part of Peninsular Malaysia with annual mean rainfall from 2,032mm to 2,540mm. North-east monsoon is the main factor that contributes to heavy rainfall to Kelantan that occurs from November to February. The major economic activities in Kelantan are agricultural based, mainly the cultivation of paddy, rubber, oil palm and tobacco. Meanwhile Damansara river basin is the most important urban centre within Klang Valley. The catchment has undergone rapid development in the last 50 years whereby agriculture lands were converted into townships consisting of commercial and industrial areas. Damansara experienced heavy rainfall during the inter-monsoon season which usually occurs from March to April and from September to October. Table 1 shows the stations selected from Damansara and Kelantan. The mean, standard deviation and the geographical location of the stations are also included. In this study, the major criteria for the selection of the both river basins is because they can represent the major geographic and rainfall regime and are likely to have different level of sensitivities to climate change impacts.

Table 1 The geographical coordinates of the stations in Damansara.

No.	Station	Mean (mm)	Standard Deviation (mm)	Latitude	Longitude
1.	Sri Aman	4.1250	10.7541	3 06’12”N	101 37’46”E
2.	Sri Permata	3.9338	11.0609	3 06’07”N	101 36’51”E
3.	Kompleks Fas	3.8162	9.9049	3 06’33”N	101 38’08”E
4.	Sea Park	3.0368	10.2657	3 07’20”N	101 38’07”E

Table 2 The geographical coordinates of the stations in Kelantan

No.	Station	Mean (mm)	Standard Deviation (mm)	Latitude	Longitude
1.	Brook	5.5846	11.4268	4 40’35”N	101 29’04”E
2.	Blau	3.9632	12.9426	4 46’00”N	101 45’25”E
3.	Gunung Gagau	6.5699	18.2478	4 45’25”N	102 39’20”E
4.	Hau	5.7434	13.8957	4 49’00”N	101 31’60”E

3 Probability Density Functions

In this study, three distributions namely as Wakeby distribution, Generalized Extreme Value (GEV) distribution and two-parameter Weibull distribution are used. Note that *X* is a random variable that represents the rainfall amount.

3.1 Wakeby Distribution

The inverse distribution function is defined as,

$$x(F) = \xi + \frac{\alpha}{\beta}[1-(1-F)^{\beta}] - \frac{\gamma}{\delta}[1-(1-F)^{-\delta}] \quad (1)$$

where ξ , α , β , γ and δ are the parameters and the domain is: $\xi \le x \le \infty$ if $\delta \ge 0$ and $\gamma > 0$; $\xi \le x \le \xi + \alpha/\beta + \gamma/\delta$ if $\delta < 0$ or $\gamma = 0$. The distribution will be reduced to Generalized Pareto if α or γ is equal to 0.

3.2 Generalized Extreme Value Dsitribution

The probability and cumulative density function is defined as,

$$f(x) = \frac{1}{\alpha}\left[1 - k\left(\frac{x-u}{\alpha}\right)\right]^{1/k-1} \exp\left[-\left(1 - k\left(\frac{x-u}{\alpha}\right)\right)^{-1/k}\right] \quad (2)$$

$$F(x) = \exp\left\{-\left[1 - k\left(\frac{x-u}{\alpha}\right)\right]^{1/k}\right\} \quad (3)$$

where *k*, *u* and *α* > 0 are the shape, location and scale parameters.

3.3 Two-parameter Weibull Distribution

The probability and cumulative density function is defined as,

$$f(x) = \frac{b}{a}\left(\frac{x}{a}\right)^{b-1} e^{-\left(\frac{x}{a}\right)^{b}} \quad (4)$$

$$F(x) = 1 - \exp\left[-\left(\frac{x}{a}\right)^{b}\right] \quad (5)$$

where *a* and *b* are scale and shape parameters.

4 L-Moments

The L-moments introduced by [15] are defined as,

$$\lambda_r = \int_0^1 x(F) P_{r-1}^{*}(F)\,dF, \; r = 1,2,.... \quad (6)$$

where

$$P_r^{*}(F) = \sum_{k=0}^{r} p_{r,k}^{*} F^k \qquad \text{and} \qquad p_{r,k}^{*} = (-1)^{r-k}\binom{r}{k}\binom{r+k}{k}.$$

λ_r is a linear combination of the expected order statistics, $P_r^{*}(F)$ is the shifted Legendre polynomial. The first five L-moments are

$$\lambda_1 = \int x.dF$$
$$\lambda_2 = \int x.(2F - 1)dF$$
$$\lambda_3 = \int x.(6F^2 - 6F + 1)dF$$
$$\lambda_4 = \int x.(20F^3 - 30F^2 + 12F - 1)dF$$
$$\lambda_5 = \int x.(70F^4 - 140F^3 + 90F^2 - 20F + 1)dF$$

As explained in [15], λ_2 is a measure of scale of a random variable. Hence it is normally used to standardise other higher moment λ_r , $r \geq 3$ to ensure they are independent unit of measurement of the random variable. Therefore the L-moments ratio, also known as scaled L-moments, is defined as,

$$\tau_r = \frac{\lambda_r}{\lambda_2}, \quad r = 3,4,.... \tag{7}$$

where τ_3 and τ_4 are L-skewness and L-kurtosis respectively.

5 Determination of Wet and Dry Events in a Data

The identification of the rainfall events occurs in the data is carried out by first estimating the parameters of the distributions for each day. In particular, the cumulative density functions (CDF) is used and bootstrap resampling technique is used by taking 500 samples from the surrounding station. Next, all the rainfall values of the surrounding stations are standardized and replaced by the corresponding probability values while the zero and missing values are remained unchanged. Then, a complete replicate data for the target station is constructed by using weighted average algorithm as in Eq. (11). Specifically, it is a method that use the probability values from the surrounding stations and the information related to elevation, distances from surrounding station to the target station, and angular separation will also used in the computation. The weight of the surrounding stations is calculated as follows:

$$w_i^d(x,y) = \exp\left(-\frac{d_i^2(x,y)}{c_d}\right) \text{ and } w_i^e(x,y) = \exp\left(-\frac{e_i^2(x,y)}{c_e}\right) \tag{8}$$

$$w_i^{ang}(x,y) = 1 + \frac{\sum_{j \neq i} w_j^d(x,y) w_j^e(x,y)(1-\cos\theta_{(x,y)}(j,i))}{\sum_{j \neq i} w_j^d(x,y) w_j^e(x,y)} \tag{9}$$

$$c_d = \frac{\tilde{d}^2}{\ln 2} \qquad \text{and} \qquad c_e = \frac{\tilde{e}^2}{\ln 2} \tag{10}$$

where i and j are the surrounding stations, (x,y) is the location of target station, $d_i(x, y)$ and $e_i\,(x, y)$ are indicated as distance and elevation difference between target station to the i-th surrounding station, respectively, $\tilde{d}$ and $\tilde{e}$ are the maximum distance and elevation difference of the surrounding stations. The weight of each surrounding station is the product of the equations above, which can be written as,

$$w_i(x,y) = w_i^d(x,y) w_i^e(x,y) w_i^{ang}(x,y) \tag{11}$$

Threshold to differentiate the wet or dry event in the data is identified. This threshold is estimated by the similar method used in obtaining the parameter estimates. In this step, the target station is entirely reconstructed after the application of weighted-average algorithm. In particular, the rainfall values associated with the corresponding reconstructed probability of the target station are obtained. Then, a bootstrap sampling with at least 400 non-missing data centred on each day is used and reordered the subsample to decreasing values. This is followed by the identification of the number of wet days, N in the original data and this N is then applied to the subsample. Lastly, the probability associated to the N-th days of the subsample is set as the threshold. This procedure is repeated for all days in the year and a complete replica of target data is expressed in terms of binary, i.e. 1 if the reconstructed probability is above or equal to the threshold, otherwise 0.

6 Results and Discussion

Table 3 and 4 illustrate the results of Kolmogorov Smirnov (KS) goodness-of-fit test in Damansara and Kelantan respectively and the ranking is based on the minimum error produced by the test. Among the distributions tested to the daily rainfall data, Wakeby is the fittest distribution to the rainfall process in Damansara and Kelantan. The error produced by Wakeby is the most minimum as compared to the Generalized Extreme Values (GEV) and Weibull distribution. However in year 2000, station Sri Aman in Damansara and station Brook and Gunung Gagau showed slightly different rainfall pattern than other stations in which the GEV distribution is the best to explain the rainfall behaviour. It is justified that the rainfall pattern may vary across the year and thus needed to examine the rainfall patterns every year. However, the Wakeby distribution is completely dominant after the year 2000 and consequently it can be concluded as the best to describe the rainfall patterns in Damansara and Kelantan.

It can be further verified by looking at Figure 1 where station Sri Aman which from Damansara is chosen as an example. The figure shows the comparison of the cumulative probability distributions used in Sri Aman, Damansara from 2000 to 2004. From Figure 1a, the shape of GEV is the closest to the shape of the actual rainfall (indicated by blue colour line). After the year of 2000, the pattern of the rainfall is changed and the shape of the actual rainfall process is solely represented by Wakeby distribution which is shown in Figure b, c, d and e. Therefore, it is clearly that Wakeby distribution is the most appropriate distribution as compared to GEV and Weibull to represent the rainfall process in Damansara.

Table 3 The results of KS test of stations in Damansara

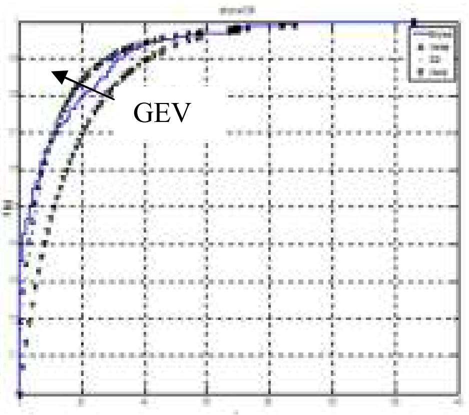
Station	Distribution	Year				
		2000	2001	2002	2003	2004
Sri Aman	WKY	0.1922	**0.3163**	**0.3292**	**0.2642**	**0.3113**
	GEV	**0.1891**	0.3430	0.3564	0.2899	0.3381
	WBL	0.3547	0.5656	0.5797	0.4963	0.5566
Sri Permata	WKY	0.2315	0.5486	0.3119	0.4008	0.2196
	GEV	0.2580	0.5722	0.3385	0.4274	0.2366
	WBL	0.4565	0.8563	0.5644	0.6715	0.4325
Kompleks Fas	WKY	**0.1980**	**0.2408**	**0.3009**	**0.2361**	**0.2379**
	GEV	0.1998	0.2672	0.3278	0.2619	0.2640
	WBL	0.3750	0.4754	0.5479	0.4698	0.4706
Sea Park	WKY	**0.3528**	**0.2328**	**0.3179**	**0.2408**	**0.2268**
	GEV	0.3802	0.2574	0.3445	0.2669	0.2532
	WBL	0.6055	0.4645	0.5744	0.4752	0.4509

*WKY = Wakeby, GEV = Generalized Extreme Value, WBL = Weibull

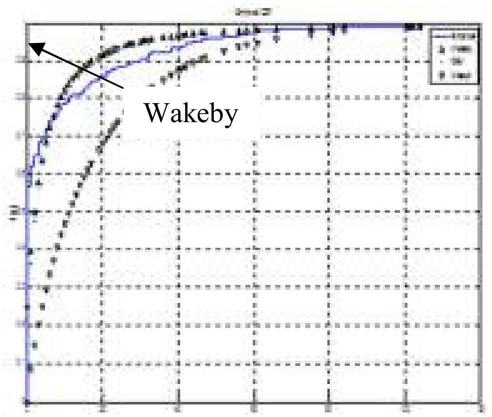
Table 4 The results of KS test of stations in Kelantan

Station	Distribution	Year				
		2000	2001	2002	2003	2004
Brook	WKY	0.2118	**0.2373**	**0.2270**	**0.2066**	**0.2059**
	GEV	**0.1975**	0.2614	0.2471	0.2283	0.2118
	WBL	0.3863	0.4735	0.4484	0.3886	0.3741
Blau	WKY	**0.4038**	**0.4563**	**0.3586**	**0.4368**	**0.4391**
	GEV	0.4305	0.4826	0.3857	0.4641	0.4670
	WBL	0.6712	0.7326	0.6110	0.6986	0.7014
Gunung Gagau	WKY	0.1939	0.2135	**0.2709**	**0.3188**	**0.3647**
	GEV	**0.1861**	**0.1946**	0.2970	0.3444	0.3929
	WBL	0.3176	0.3470	0.5014	0.5526	0.6047
Hau	WKY	**0.2029**	**0.2769**	**0.2014**	**0.4604**	0.1903
	GEV	0.2187	0.3017	0.2084	0.4856	**0.1682**
	WBL	0.3793	0.5146	0.3677	0.7489	0.3003

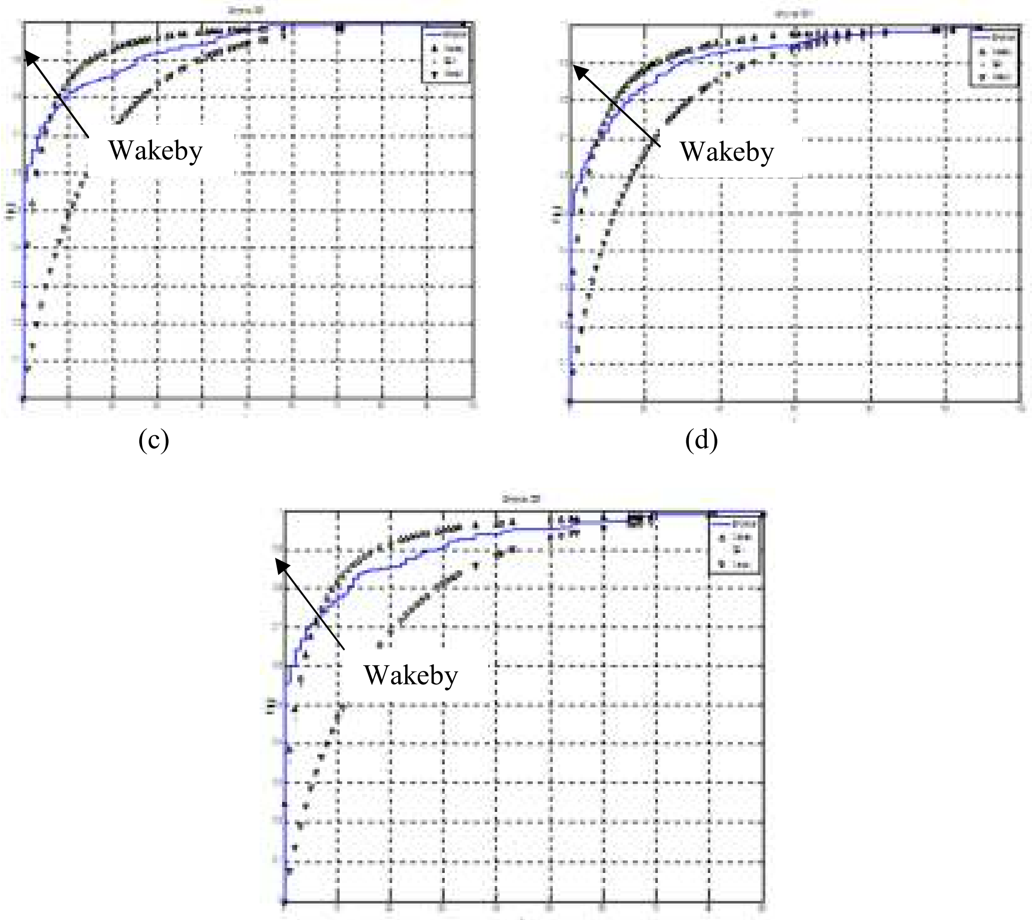
*WKY = Wakeby, GEV = Generalized Extreme Value, WBL = Weibull



(a)



(b)



(c) (d)

(e)

Figure 1 Comparison of Wakeby, GEV and Weibull distributions of the rainfall process in Sri Aman, Damansara on a)2000, b)2001, c)2002, d)2003, e)2004

The determination of wet and dry events in the rainfall data by weighted-average algorithm is shown in Table 5 and 7. The comparison between the number of wet events occurs in the actual rainfall data and reconstructed data for a station from Damansara and Kelantan respectively is examined. In addition, the percentage of hitting between the reconstructed data and the actual wet events in the target data is also presented. In comparison, the accuracy provided by Wakeby distribution is slightly higher than GEV and Weibull distributions. Overall, the percentage of hitting reaches about 60% although different distributions used in this study. Therefore, it can be concluded that the number of wet events in the reconstructed data, either with Wakeby, GEV or Weibull distributions, do not differ much when compared to the actual number of wet days of the data. This also indicated that the method do not violate the number of rainfall events in the data even though different distributions are used in the analysis. Even if the Wakeby distribution is the best fit distribution as compared to GEV and Weibull distributions, the performances in determining the rainfall events in a data are almost equal for all the three distributions.

Table 5 Comparison of number of wet events (percentage of hitting) in station Sri Aman of Damansara by using Wakeby, GEV and Weibull distribution

Dist.	Year									
	2000		2001		2002		2003		2004	
	Obs.	Pred. (%)	Obs.	Pred. (%)	Obs.	Pred. (%)	Obs.	Pred. (%)	Obs.	Pred. (%)
WKL	194	193 (65.89)	159	162 (78.96)	153	151 (78.60)	136	144 (60.74)	137	137 (73.46)
GEV	194	194 (62.29)	159	161 (78.90)	153	151 (78.60)	136	140 (58.90)	137	137 (73.46)
WBL	194	194 (65.22)	159	162 (75.68)	153	156 (78.90)	136	140 (50.74)	137	136 (74.43)

*WKY = Wakeby, GEV = Generalized Extreme Value, WBL = Weibull

Table 6 Comparison of number of wet events (percentage of hitting) in station Brook of Kelantan by using Wakeby, GEV and Weibull distribution

Dist.	Year									
	2000		2001		2002		2003		2004	
	Obs.	Pred. (%)	Obs.	Pred. (%)	Obs.	Pred. (%)	Obs.	Pred. (%)	Obs.	Pred. (%)
WKL	194	195 (50.41)	119	145 (52.45)	159	156 (79.94)	193	192 (61.92)	199	186 (71.23)
GEV	194	195 (50.96)	119	149 (52.83)	159	149 (75.22)	193	192 (60.82)	199	184 (70.68)
WBL	194	196 (50.68)	119	140 (51.70)	159	136 (75.58)	193	192 (60.82)	199	183 (70.96)

*WKY = Wakeby, GEV = Generalized Extreme Value, WBL = Weibull

7 Conclusion

The determination of the best-fit distribution to represent the rainfall process and the identification of wet and dry events in stations of Damansara and Kelantan are discussed in this paper. Three distributions namely Wakeby, Genralized Extreme Value (GEV) and two-parameter Weibull have been used on the daily rainfall data from 2000 to 2004. The selection of the best-fit distribution is done by examining the minimum error produced by the Kolmogorov-Smirnov (KS) goodness-of-fit test. Based on the results of KS goodness-of-fit test, Wakeby distribution is the most suitable to describe the rainfall patterns in the stations of Damansara and Kelantan as the error produced is the minimum. This is followed by the GEV distribution and finally by the Weibull distribution. The identification of dry and wet events in the rainfall data is also done to evaluate the effects of using different distributions. The results pointed out that the impacts of different probability distributions used are not significant in determining the rainfall events in the data using weighted-average algorithm. This is justified when Wakeby performs equally well with

GEV and Weibull distributions in determining the wet and dry events in the rainfall data eventhough Wakeby is the most fitted distribution that can describe the rainfall pattern in Damansara and Kelantan.

Acknowledgements

The authors are grateful to the Malaysian Meteorological Department for providing the daily precipitation data. The work is financed by RU Grant Vot 06J18 provided by Universiti Teknologi Malaysia.

References

[1] Hanson, L.S. and Vogel, R. The Probability Distribution of Daily Rainfall in the United States. *World Environment and Water*. 2008: 1-12.

[2] Sharma, M.A. and Bhagwan Singh, J. Use of Probability Distribution in Rainfall Analysis. *New York Science Journal*. 2010. 3(9): 40-49.

[3] Modarres, R. Regional Precipitation Climates of Iran. *Journal of Hydrology (NZ)*. 2006. 45(1): 13-27.

[4] Aksoy, H. Use of Gamma Distribution in Hydrological Analysis. *Turk. J. Engin. Environ. Sci.* 2000. 24: 419-428.

[5] Fadhilah, Y., Zalina, M. D., Nguyen, V-T-V, Suhaila, S., Zulkifli, Y. Fitting the Best-Fit Distribution for the Hourly Rainfall Amount in the Wilayah Persekutuan. *Jurnal Teknologi*. 2007. 46(C): 49-58.

[6] Ling, W. S. Y. and Ismail, N. Analysis of T-Year Return Level for Partial Duration Rainfall Data. *Sains Malaysiana*. 2012. 41(11): 1389-1401.

[7] Dan'azumi, S., Shamsudin, S., Aris, A. Modeling the Distribution of Rainfall Intensity using Hourly Data. *American Journal of Environmental Sciences*. 2010. 6(3): 238-243.

[8] Suhaila, J. and Jemain, A. A. Fitting the Statistical Distributions to the Daily Rainfall Amount in Peninsular Malaysia. *Jurnal Teknologi*. 2007a . 46(C): 33-48.

[9] Suhaila, J. and Jemain, A. A. Fitting daily rainfall amount in Malaysia using the normal transform distribution. *Journal of Applied Sciences*. 2007b. 7(14): 1880-1886.

[10] Zalina, M. D., Amir, H. M. K., Mohd Nor Mohd Desa, Nguyen, V-T-V. Statistical analysis of at-site extreme rainfall processes in Penisular Malaysia. In *Regional Hydrology: Bridging the Gap Between Research and Practice (Proceeding of the Fourth International FRIEND Conference*, March. Cape Town, South Africa: FRIEND. 2002.

[11] Wan Zin, W.Z. and Jemain, A.A. The best fitting distribution of annual maximum rainfall in Peninsular Malaysia based on methods of L-moment and LQ-moment. *Theor. Appl. Climatiol*. 2009. 96: 337-344.

[12] Eli, A., Shaffie, M., Wan Zin, W.Z. Preliminary study on bayesian extreme rainfall analysis: A case study of Alor Setar, Kedah, Malaysia. *Sains Malaysiana*. 2012. 41(11): 1403-1410.

[13] Rao, A.R. and Hamed, K.H. *Flood Frequency Analysis*. United States: 2012. 2000.

[14] Simolo, C., Brunetti, M., Maugeri, M., Nanni, T. Improving estimation of missing values in daily precipitation data by a probability density function-preserving approah. *Int. J. Climatol*. 2009. DOI: 10.1002/joc.1992.

[15] Hosking, J.R.M. L-Moments: Analysis and estimation of distribution using linear combinations of order statistics*. J. R. Statist. Soc. B.* 1990. 52(1): 105-124.