Journal of Engineering and Applied Sciences 11 (11): 2360-2365, 2016
ISSN: 1816-949X
© Medwell Journals, 2016

Rainfall Data Analysis in Langat River Basin Using Hyfran-Plus

[1]Khairi Khalid, [2]Mohd Fozi Alib, [3]Nurul Fatin Manc, [2]Nor Faiza Abd Rahmanb,
[1]Ahmad Amzari Yaccob, [1]Nur Asmaliza Mohd Noor and [1]Siti Hawa Rosli
[1]Faculty of Civil Engineering, UiTM Pahang, 26400 Bandar Jengka, Pahang, Malaysia
[2]Faculty of Civil Engineering,
UiTM Shah Alam, 40450 Shah Alam, Selangor, Malaysia
[3]JKR Malaysia, Menara PJD, Jalan Tun Razak, 50582 Kuala Lumpur, Malaysia

Abstract: The selection of the best-fit distribution of the rainfall data is always highlighted by the researchers in the hydrology study and the information is used for planning and designing of various water resource projects. The issue is critical to Malaysia and not excluded to the Langat River Basin since the country is experiencing two distinct monsoon seasons namely Southeast and Northwest Monsoon that brings heavy rainfall and cause floods. The study is attempted to test the goodness-of-fit of the rainfall data and to determine the best-fit distribution to explain the rainfall process in the study area. The reliability of rainfall data was determined using the independence, stationarity and homogeneity tests. Three probability distribution functions were utilized in the study; namely Normal Distributions, Log-Pearson Type 3 and Generalized Extreme Value (GEV) distribution. The parameters of the distributions were estimated by using the method of maximum likelihood. The best-fit distribution to explain the rainfall process is determined by using Chi-squared test. For Langat River Basin, the reliability of rainfall data of all the four rainfall stations is independent, stationary and homogeneous. It was found that GEV distribution is the best-fit distribution to explain the rainfall process in Langat River Basin followed by Log-Pearson Type 3 and normal distribution.

Key words: Langat river basin, Chi-squared test, GEV distribution, rainfall, stationary

INTRODUCTION

The science of hydrology deals with the occurrence, circulation and distribution of water of the earth and earth's atmosphere. In any study of water resources discipline, it is needed to understand the hydrology in the watershed (Subramanya, 2006). Rainfall is one of the important components which affects the whole process in the hydrologic cycle (Mohammad *et al*., 2005). Too much of rainfall will cause flooding while too little of rainfall will cause drought. Malaysia is experiencing heavy rainfall during the monsoonal seasons. In inter-monsoon periods of April to May and August to October, Malaysia faces intense rainstorms which causing flash floods in major towns. These two phenomena affected almost 4.9 million people and it is reported that the average annual flood damage is estimated at RM1 billion (Abdullah, 2004).

In the hydrology analysis regarding the planning and designing of various water resources project, the quantification process of rainfall is a must. It is necessary for the use of proper design for hydraulic structures such as dams, culverts, highways, sewage disposal, bridges and many more. There are several studies conducted in Malaysia in order to investigate the rainfall distribution, either hourly, daily or annually. It is found that the Wakeby distribution is the best-fit distribution to explain the rainfall process in Damansara and Kelantan (Ho and Yusof, 2013). Fadhilah *et al*. (2007) identified that the best-fit distribution for hourly rainfall amount in Wilayah Persekutuan, Malaysia is Mixed-Exponential Distribution. Other study showed that the Generalized Extreme Value (GEV) distribution is the more suitable to be used to represent annually maximum rainfall data in Peninsular Malaysia (Zalina *et al*., 2002).

The study attempted to highlight the rainfall data analysis in the upper part of Langat River Basin using HYFRAN-PLUS Software.

Rainfall data analysis: It is necessary to check the data for continuity and consistency before using the rainfall data in any hydrology analysis. This is due to the fact that the recorded data might be erroneous due to wind effects, changes in station environment, errors while observing the data and many more. Furthermore, the checking is conducted in order to test the validity of the rainfall data itself.

There are various analysis that can be conducted to analyze the rainfall data, namely the independence test, stationarity test, homogeneity test, consistency test, basic statistics and frequency distribution analysis (Rao and Kao, 2006). In the study, the data screening has been analyzed using independence, stationarity and

Corresponding Author: Khairi Khalid, Faculty of Civil Engineering, UiTM Pahang, 26400 Bandar Jengka, Pahang, Malaysia

homogeneity test. In all the tests, the level of significance was expressed as a p-value. The most commonly used p-value for statistical analysis is 5% and is applied in the study (Roy, 2013; Nury and Alam, 2014).

The independence test of Wald-Wolfowitz is used in this study. This is a non-parametric test of a null hypothesis for a two-valued data sequence that comes from the same population. Stationarity of a data series was conducted using the Mann-Kendall (MK) test. The MK statistical test has been used widely in identifying the monotonic trends in hydro-meteorological data, namely rainfall, streamflow and temperature (Suhaila *et al*., 2010). The Wilcoxon Test was performed in order to evaluate the homogeneity of the data series. The test is useful to determine if the measurements of the data are taken at the same time with the same instruments and environment.

Given p is a significant value whereby if it is <0.05 ($p<0.05$), the null hypothesis can be rejected at a significance level of 5%. In this study, the results of independence test were generated by HYFRAN-PLUS Software.

Probability Distribution Function (PDF) and goodness-of-fit test: The probability distribution is applied in a hydrological study to analyze the rainfall data for the computation of expected rainfall of a given frequency (Dawood, 2009). It is defined as the statistical analysis of a random variable. The most frequently used PDF including normal and log-normal distribution, Pearson Type 3 and Log-Pearson Type 3. Besides the above method, the fitting of these frequency distributions will be carried out by using the method of maximum likelihood to identify the probability of the events occurring. The three PDF used in the study namely Normal Distributions, Log-Pearson Type 3 and Generalized Extreme Value (GEV) distribution and the selection of the PDF from recommended by the previous studies for annual maximum rainfall estimates (Daud *et al*., 2002; Singh *et al*., 2012; Bhat *et al*., 2013).

Normal distribution: The normal distribution or also known as Gaussian distribution is applied to asymmetrically distributed data. Besides, it is also referred as the bell curve due to the bell shape of distribution (Kwaku and Duke, 2007). The normal distribution can be specified by two parameters, namely mean (μ) and standard deviation (σ). The data will fall between two real numbers with non-zero over the entire line. The probability distribution function of normal distribution can be determined by using the Eq. 1 for $-\infty \leq \times \leq = \infty$:

$$n = \frac{1}{(\sigma\sqrt{2}\pi)} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^{2}\right] \tag{1}$$

Where:

n = Number of observation
σ = Standard deviation
μ = Mean

Log-pearson type 3 distribution: The Log-Pearson Type 3 is extensively used for hydrological projects in USA (Ewemoji and Ewemooji, 2011). The concept of this distribution is to transform the variate into logarithmic form and then the transformed data is further can be analyzed. The series of Z variates as given:

$$z = \log x \tag{2}$$

where, x is variate of a random hydrologic series. For this Z series, the values of variate X of a random hydrologic series with a return period T is given by:

$$Z_T = \bar{Z} + K_z \sigma_z \tag{3}$$

Where:

K_z = A frequency factor that is a function of recurrence interval T and the coefficient skew Cs
σ_z = Standard deviation of the Z-variate sample
C_s = Coefficient of skew of variate Z

$$\sqrt{(\sum((Z - \bar{Z})\hat{}2/(N-1)))} \tag{4}$$

as:

$$= (N\sum(z - \bar{z})\hat{}3/((N - 1(N - 2)([\sigma\_z)]\hat{}\,3 \tag{5}$$

Where:

$\bar{z}$ = Mean of the z values
N = Number of years of records

GEV distribution: The theory of extreme value was first developed in 1927 for independent and identically distributed random variables (Shukla *et al*., 2012). The simplest three forms of the extreme value distribution are given by the Gumbel, Frechet and Weibull families or also known as type 1-3, respectively. However, due to the problem of del termining which of the distributions should be used to analyze a data set, the GEV was developed. There are three parameters of GEV namely scale, location and shape. The PDF of GEV as given in HYFRAN-PLUS Software is as follows:

$$\begin{gathered} f(x) = 1/\sigma[1 - k/\sigma(x - \mu)]\hat{}\,(1/k - 1) \\ \exp\left\{-[1 - k/\sigma(x - \mu)]\hat{}\,(1/k)\right\} \end{gathered} \tag{6}$$

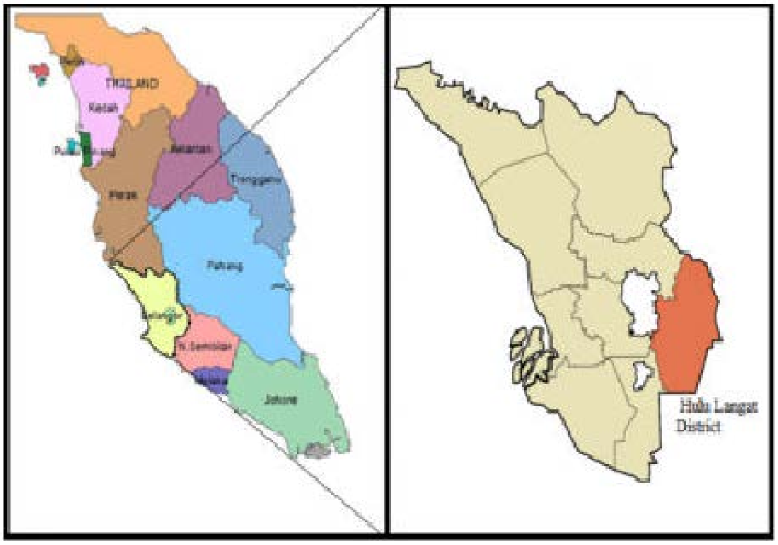

Fig. 1: Langat river basin in the Hulu Langat district of the selangor map

Where:

σ = Scale parameter
k = Shape parameter
μ = Location parameter

Goodness-of-fit test: In order to evaluate the quality of the fitted distributions, the goodness-of-fit test was conducted. It represents the statistical hypothesis used to evaluate if the input data is an independent sample from a particular distribution. The goodness-of-fit test can be measured using Chi-Square test (Oseni and Agoola, 2012). The Chi-square test is one of the goodness-of-fit tests to compare the input data histogram with the fitted distribution. The data was first divided into k class intervals whereby in the study $k \approx \sqrt{n}$ where n is the number of total recorded years. The average number of values in any group should be >5. The goodness-of-fit tests were conducted at 5% level of significance. The Chi-square is given by:

$$x^2 = \sum_{(j=1)}^{k} (o_j - [e_j)]^2 / e_j \qquad (7)$$

Where:

o_j = Observed frequency in the class interval j
e_j = Expected frequency from the theoretical distribution

From the Chi-square test (χ^2), it can be concluded that the if the observed frequencies are close to the corresponding expected frequencies, it representing a good fit or otherwise, it is a poor fit. The hypothesis made is that a good fit leads to acceptance of H_0 whereas the data is said not to follow the specified distribution for a poor fit that leads to a rejection.

Study area: Langat River Basin occupies the south and south-eastern parts of Selangor and a small portion of Negeri Sembilan and Wilayah Persekutuan. The mainstream, Langat River stretches for 180 km and has a total catchment area of 2271 km^2. The study only focused on the upper part of Langat River Basin with a catchment area of 331 km^2 and the main streamflow station of the study area is located in Kajang town in the District of Hulu Langat as in Fig. 1.

MATERIALS AND METHODS

The main objective of this study is to determine the best-fit distribution to represent the rainfall patterns in Langat River Basin, Malaysia. The study started with a data collection of the rainfall and followed by data preparation, PDF analysis and goodness-of-fit test. Finally, the best PDF was selected based on the tested data from the results of the goodness-of-fit test.

The rainfall data was obtained from the Malaysia Department of Irrigation and Drainage (DID), Ampang Branch records for Langat River Basin in 30 consecutive years, starting from 1981-2010. According to the DID records, there is a total of 22 rainfall stations for the upper part of Langat River Basin. Only four stations fulfill the requirement for analysis, namely RTM Kajang station, SK Kg. Sg. Lui station, TNB Pansun station and Ldg Dominion station. The missing values were calculated by using normal ratio method (Silva *et al*., 2007).

The HYFRAN-PLUS or Hydrological Frequency Analysis PLUS DSS is a tool developed by Canadian Developer used to fit statistical distributions (Water Resources Publication). The advantages of this tool are that it can be used for analyzing extreme events and can perform basic analysis of any time series of Independent and Identically Distributed (IID) data.

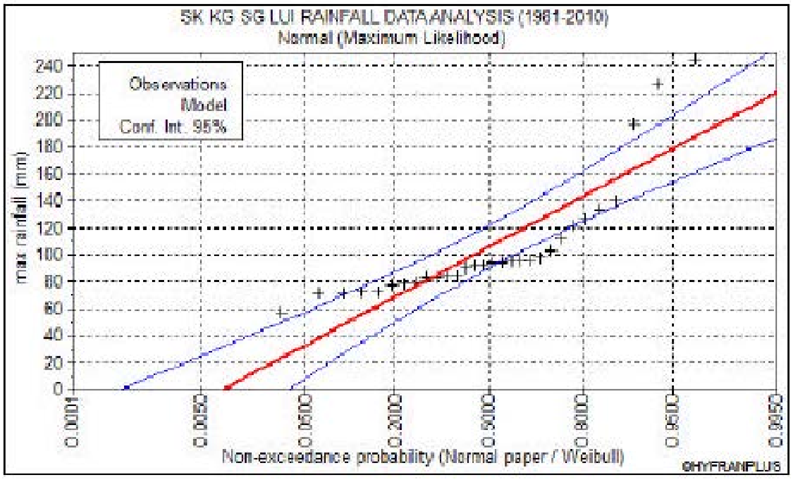

Fig. 2: A normal distribution for SK Kg. Sg. Lui Station

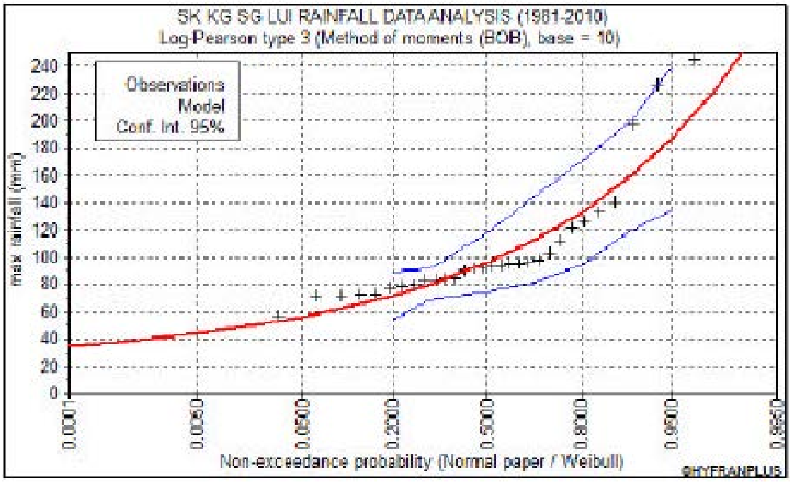

Fig. 3: A log-pearson Type 3 distribution for SK Kg. Sg. Lui Station

RESULTS AND DISCUSSION

Rainfall data analyzes were used to test and find the data validity. Data validity is used to check on the recorded data at the respected stations because its might have errors. The 30 year series of rainfall data from the stations in the Langat River Basin were used in the study. The statistic of each station of the Langat River Basin was automatically generated and computed by HYFRAN-PLUS Software as in Table 1.

The statistic values for independence, stationarity and homogeneity test were identified as U, K and W, respectively. The significant value (p) in the statistics measures whether the given hypothesis is probably true or not due to change. If the significant value that is $p<0.05$, the null hypothesis (H_0) can be rejected at a significant level of 5%. The significant value (p) that is >0.05 or 5% indicates that the data are independence.

The results summary of independence, stationarity and homogeneity test for four stations are as shown in Table 2-4, respectively. It can be clearly seen that all the p-values for independence, stationarity and homogeneity test are >0.05. Since all the data from the tests show that all the $p>0.05$, the rainfall data recorded in TNB Pansun Station, SK Kg. Sg. Lui Station, Ldg. Dominian Station and RTM Kajang Station are independent, stationary and homogeneous.

The samples of the graphic output for each of the PDF fitting of SK. Kg. Sg Lui Station is shown in Fig. 2-4, respectively. The graphs show the best-fitted PDF to describe the rainfall was indicated by the annual maximum rainfall data that lay within the lower and upper limit of control bands of 95% confidence intervals. As shown in Fig. 2 and 3, these figures demonstrate that some of the annual maximum rainfall data were not laid within the lower and upper limit of control bands set up which is 95%. Thus, the 95% confidence level was not

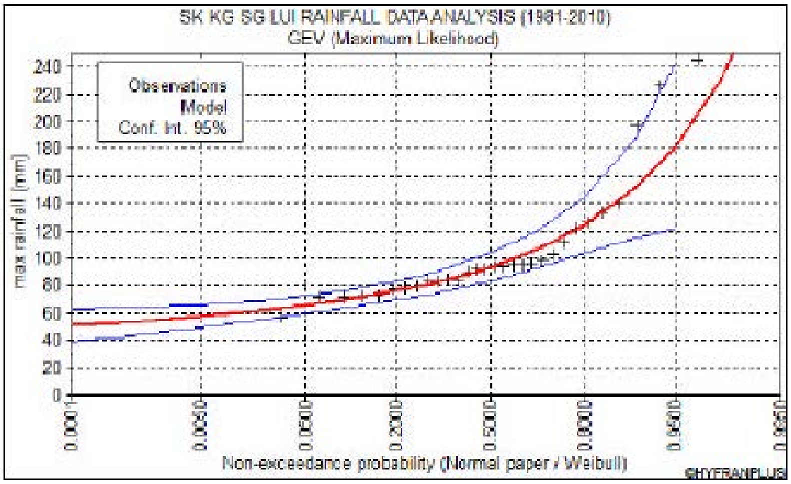

Fig. 4: A GEV distribution for SK Kg. Sg. Lui station

Table 1: The statistical value

		Statistical value			
Station No.	Station name	Min. (mm)	Max. (mm)	SD	Median
3218101	TNB Pansun	41	162	32.8	104
3118102	SK. KG. SG. Lui	57	245	44.4	93.5
3018107	LDG. Dominion	65	143	18.5	96
2917001	RTM Kajang	39.5	185	34.8	104

Table 2: The independence test

		Independence test (Waldwolfowitz)	
Station No.	Station name	U	p-values
3218101	TNB Pansun	1.110	0.265
3118102	SK. KG. SG. Lui	0.332	0.740
3018107	LDG. Dominion	0.179	0.858
2917001	RTM Kajang	0.894	0.371

Table 3: The stationarity test

		Stationarity test (Mann-kendall)	
Station No.	Station name	K	p-values
3218101	TNB Pansun	0.482	0.630
3118102	SK. KG. SG. Lui	0.428	0.669
3018107	LDG. Dominion	0.339	0.735
2917001	RTM Kajang	1.160	0.246

Table 4: The homogeneity test

		Stationarity test (Mann-(Wilcokson)	
Station No.	Station name	W	p-values
3218101	TNB Pansun	1.330	0.184
3118102	SK. KG. SG. Lui	0.728	0.467
3018107	LDG. Dominion	0.727	0.467
2917001	RTM Kajang	1.330	0.184

Table 5: The Chi-squared test

	Probability distribution function fitting					
	Normal		Log-Pearson Type III		GEV	
Station name	χ^2	p-values	χ^2	p-values	χ^2	p-values
TNB Pansun	1.73	0.7847	5.00	0.1718	1.7300	0.6295
SK. KG. SG. Lui	21.33	0.0003	13.87	0.0031	7.3300	0.0620
LDG. Dominion	1.27	0.8670	1.73	0.6295	1.7300	0.6295
RTM Kajang	4.07	0.3971	3.60	0.3080	3.6000	0.3080
AVE.	7.10	-	6.05	-	3.5975	-

achieved. However, as shown in Fig. 4, it is clearly seen that all the annual maximum rainfall data lie within the lower and upper limit of control bands which indicates that a 95% confidence intervals were achieved for GEV PDF. Therefore, GEV shows the best-fit distribution for SK Kg. Sg Lui Station. The same output was computed in the other stations.

A summary of Chi-Square test for each of the PDF is shown in Table 5. The p-value for SK. Kg. Sg. Lui Station is 0.0003 for Normal Distribution, 0.0031 for Log-Pearson Type 3 which both results are <0.05. Therefore, the null hypotheses can be rejected at a significance level of 5% and also indicates that the underlying distribution of this sample is neither Normal nor Log-Pearson Type 3. The p-value is 0.0620 for GEV PDF that is >0.05. Thus, the null hypothesis was accepted at a significance level of 5% and also indicates that the underlying distribution of this sample is GEV. The same analysis was performed on the other three stations by using the same methods to identify the best-fit distribution. The smallest average value of Chi-Square test for Langat River Basin is 3.5975 for GEV Distribution which indicates that GEV Distribution is the best-fit pdf to describe rainfall pattern in Langat River Basin.

CONCLUSION

In conclusion, the rainfall data analysis in Langat River Basin using HYFRAN-PLUS has yielded the acceptable results. All the rainfall data are independent, stationary and homogeneous. The rainfall data were fitted into three probability distribution functions. The best fitted PDF give the smallest value of Chi-square value χ^2 which is 3.5975, indicates that the GEV fitting is the best PDF for the annual maximum data taken from the four stations in the upper part of Langat River Basin.

ACKNOWLEDGEMENTS

The project was funded by the Exploratory Research Grant Scheme (ERGS), Ministry of Education, Universiti Teknologi MARA (UiTM), Malaysia and SWAT Network of Malaysia. Special thanks to Malaysia Department of Irrigation and Drainage for providing the related data needed to carry out the project.

REFERENCES

Abdullah, K., 2004. Floods in Malaysia power point slides. Department of Irrigation and Drainage, Kuala Lumpur, Malaysia.

Bhat, B.A., N.A. Rather and T.A. Rather, 2013. On a class of probability distributions with application using rainfall data of Kashmir Valley. Int. J. Emerging Technol. Adv. Eng., 3: 493-499.

Daud, Z.M., A.H.M. Kassim, M.N.M. Desa and V.T.V. Nguyen, 2002. Statistical analysis of at-site extreme rainfall processes in peninsular Malaysia. Int. Assoc. Hydrol. Sci. Publ., 274: 61-68.

Dawood, A.S., 2009. Probability analysis of extreme monthly rainfall in Mosul City. North Iraq Marsh Bull., 4: 60-74.

Ewemoje, T.A. and O.S. Ewemooje, 2011. Best distribution and plotting positions of daily maximum flood estimation at Ona River in Ogun-Oshun River Basin, Nigeria. Agric. Eng. Int. CIGR. J., 13: 1-13.

Fadhilah, Y., M.D. Zalina, V.T.V. Nguyen, S. Suhaila and Y. Zulkifli, 2007. Fitting the best-fit distribution for the hourly rainfall amount in the Wilayah Persekutuan. J. Teknol., 46: 49-58.

Ho, M.K. and F. Yusof, 2013. Determination of best-fit distribution and rainfall events in damansara and Kelantan, Malaysia. Math., 29: 43-52.

Kwaku, X.S. and O. Duke, 2007. Characterization and frequency analysis of one day annual maximum and two five consecutive days' maximum rainfall of Accra, Ghana. ARPN. J. Eng. Appl. Sci., 2: 27-31.

Mohammad, F., M. Ota, J. Roushan, B. Mna and M. Mubarak *et al*., 2005. Determination of probability distribution for data on rainfalland flood levels in Bangladesh. J. Inst. Eng., 66: 61-72.

Nury, A.H. and M.J.B. Alam, 2014. Analysis and prediction of time series variations of rainfall in North-Eastern Bangladesh. Br. J. Appl. Sci. Technol., 4: 1644-1656.

Oseni, B.A. and F.J. Ayoola, 2012. Fitting the statistical distribution for daily rainfall in Ibadan, based on Chi-square and Kolmogorov-Smirnov Goodness-of-Fit Tests. West Afr. J. Ind. Acade. Res., 7: 93-100.

Rao, A.R. and S.C. Kao, 2006. Statistical analysis of Indiana rainfall data, joint tansportation research program technical reports. MA Thesis, Purdue University, West Lafayette, Indiana.

Roy, M., 2013. Time series, factors and impacts analysis of rainfall in north-eastern part in Bangladesh. Int. J. Sci. Res. Publ., 3: 1-7.

Shukla, R.K., M. Trivedi and M. Kumar, 2012. On the proficient use of GEV distribution: A case study of subtropical monsoon region in India. Comput. Sci. Series, 8: 81-92.

Silva, R.P.D., N.D.K. Dayawansa and M.D. Ratnasiri, 2007. A comparison of methods used in estimating missing rainfall data. J. Agric. Sci., 3: 101-108.

Singh, B., D. Rajpurohit, A. Vasishth and J. Singh, 2012. Probability analysis for estimation of annual one day maximum rainfall of Jhalarapatan Area of Rajasthan, India. Plant Arch., 12: 1093-1100.

Subramanya, K., 2006. Engineering Hydrology. 3rd Edn., Tata McGraw-Hill, New Delhi.

Suhaila, J., S.M. Deni, W.Z. Wan Zin and A.A. Jemain, 2010. Trends in peninsular Malaysia rainfall data during the southwest monsoon and Northeast monsoon seasons: 1975-2004. Sains Malaysiana, 39: 533-542.

Zalina, M.D., M.N.M. Desa, V.V. Nguyen and A.H.M. Kassim, 2002. Selecting a probability distribution for extreme rainfall series in Malaysia. Water Sci. Technol., 45: 63-68.