Jurnal Teknologi, 46(C) Jun 2007: 49–58

FITTING THE BEST-FIT DISTRIBUTION FOR THE HOURLY RAINFALL AMOUNT IN THE WILAYAH PERSEKUTUAN

FADHILAH. Y[1]*, ZALINA MD.[2], NGUYEN V-T-V.[3], SUHAILA S.[4] & ZULKIFLI Y.[5]

Abstract. In determining the best-fit model for the hourly rainfall amounts for the twelve stations in the Wilayah Persekutuan, four distributions namely, the Exponential, Gamma, Weibull and Mixed-Exponential were used. Parameters for each distribution were estimated using the maximum likelihood method. The best-fit model was chosen based upon the minimum error produced by the goodness-of-fit tests used in this study. The tests were justified further by the exceedance probability plot. The Mixed-Exponential was found to be the most appropriate distribution in describing the hourly rainfall amounts. From the parameter estimates for the Mixed-Exponential distribution, it could be implied that most of the hourly rainfall amount recorded were received from the heavy rainfall even though there was a high occurrences of light rainfall.

Keywords: Hourly rainfall amount, goodness-of-fit test, exceedance probability, maximum likelihood

Abstrak. Dalam mengenal pasti model yang terbaik untuk mewakili taburan jumlah hujan bagi data selang masa satu jam di 12 stesen di Wilayah Persekutuan empat taburan digunakan iaitu Taburan Eksponen, Gamma, Weibull dan Gabungan Eksponen. Parameter-parameter dianggar menggunakan kaedah kebolehjadian maksimum. Model yang terbaik dipilih berdasarkan nilai minimum yang diperolehi daripada ujian-ujian kebagusan penyuaian yang digunakan dalam kajian ini. Ujian ini dipertahankan lagi dengan plot kebarangkalian dilampaui. Taburan Gabungan Eksponen di dapati paling baik untuk mewakili taburan jumlah hujan dalam selang masa satu jam. Daripada anggaran parameter bagi taburan Gabungan Eksponen ini, boleh diterjemah bahawa jumlah hujan tertinggi yang direkodkan diperolehi daripada hujan yang dikategorikan sebagai hujan lebat, walaupun hujan renyai-renyai berlaku lebih kerap.

Kata kunci: Jumlah hujan dalam selang masa sejam, ujian kebagusan penyuaian, kebolehjadian maksimum

1.0 INTRODUCTION

Much research has dealt with stochastic modeling of daily totals [1] but less efforts on shorter timescales (e.g., hourly). The most prevalent approach is based upon the so-called conceptual (or physically based) models that involved storms and rain-cells [2].

[1&4] Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 Skudai, Johor.
* Corresponding author: Email: fadhilah@mel.fs.utm.my
[2] College Science and Technology, UTM *City Campus.* Jalan Semarak, 54100, Kuala Lumpur.
[3] McGill University, Quebec, Canada.
[5] Institute of Environmental and Water Resources Management, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor.

Even though there is a wealth of information available in hourly rainfall data, modeling of hourly data is not as active as the daily data due to shortage of data records. Nevertheless, data of shorter range such as hourly has many applications in hydrology such as in modeling rainfall-runoff transformation, forecasting flood-flow hydrograph, designing storm water drainage system and designing storms and flood frequency. Hence, using the hourly data in modeling of rainfall is also important for the hydrological studies.

In understanding the characteristics of the rainfall pattern, one of the method used is through modeling using various distributions. Several methods have been presented in the literature for modeling rainfall amounts on daily scales. The most common approach is to assume that rainfall amounts on successive days are independent and fit some theoretical distribution to the rainfall amounts [1,3]. However, there is no attempt so far to extend the method to the hourly rainfall amounts. Hence, this study will explore the methods done by [1] and Woolhiser and Roldan [3] on the hourly rainfall amounts in the Wilayah Persekutuan. The best fitting distributions for the hourly rainfall amounts based upon several criteria of goodness-of-fit tests is to be determined. Four theoretical distributions used include the Exponential, the Weibull, the Gamma and the Mixed-Exponential.

2.0 CASE STUDY

Wilayah Persekutuan(WP) is the most progressive state in Malaysia. Many major cities in Malaysia are situated within or nearby this state. WP is undergoing extensive developments and can be considered to be the most advanced state in terms of facilities and infrastructures. It is also known to be a flash-flood prone area due to the high occurrence of convectional rainfalls and large areas of impervious surface. The recent study in this area has shown that the convective rainfall occurrences are rapid during the inter monsoon periods, that is in April to May and October [4]. Hence, determining the best probability distribution in modeling the hourly wet amounts in this area would provide some insight knowledge for future studies in convective rainfall and flood.

There are 13 rainfall stations located in the vicinity. Historical rainfall data of every 15 minutes and daily amount are supplied by Department of Irrigation and Drainage {DID) Selangor for this study. The 15 minutes data are then aggregated to become hourly data. For this study, twelve stations were chosen based on the completeness of the data. The study period ranges from 1981-1991 with most stations having a ten-year period hourly data. This period was chosen because the number of missing data was at a minimum as compared to the more recent period.

3.0 MODELING OF RAINFALL AMOUNTS

According to the World Meteorological Organization (WMO), a wet day is defined as a day with a rainfall amount above a fixed threshold of 0.1 mm. This threshold will hold in this study with amount of greater than or equal to 0.1 to be identified as wet hours. The sequence of rainfall amounts on wet hours is also considered as the intensity process [5].

The maximum likelihood method that is claimed to being a minimum variance unbiased estimator is used in estimating the parameters of the distributions. However, the method of moments is still being used to set up the initial points of the maximum likelihood method.

Four distributions for the rainfall amounts and their probability density functions along with the log likelihood functions are as follows. The $f(x)$ represents the probability density function and $\log L$ represents the likelihood function.

(i) The Exponential distribution with parameter λ represents mean while x represents the hourly rainfall amounts.

$$f(x) = \frac{1}{\lambda} e^{\frac{-x}{\lambda}}, \qquad x, \lambda > 0$$

$$\log L = \sum_{i=1}^{n} \log \left[\frac{1}{\lambda} e^{\frac{-x}{\lambda}} \right]$$

(ii) The Weibull distribution with two-parameters, namely α and β to represent shape and scale parameters respectively while x represents the hourly rainfall amounts.

$$f(x) = \frac{\alpha}{\beta} \left(\frac{x}{\beta} \right)^{\alpha - 1} e^{-\left(\frac{-x}{\lambda} \right)^{\alpha}}, \qquad \alpha > 0, \beta > 0, x > 0$$

$$\log L = \sum_{i=1}^{n} \log \left[\frac{\alpha}{\beta} \left(\frac{x_i}{\beta} \right)^{\alpha - 1} e^{-\left(\frac{-x_i}{\lambda} \right)^{\alpha}} \right]$$

(iii) The Gamma distribution with two-parameters, namely α and β to represent shape and scale parameters respectively while x represents the hourly rainfall amounts.

$$f(x) = \frac{1}{\Gamma(\alpha) \beta^{\alpha}} x^{\alpha - 1} e^{\frac{-x}{\beta}}, \qquad \alpha > 0, \beta > 0, x > 0$$

$$\log L = \sum_{i=1}^{n} \log \left[\frac{1}{\Gamma(\alpha) \beta^{\alpha}} x_i^{\alpha - 1} e^{\frac{-x_i}{\beta}} \right]$$

(iv) The Mixed-Exponential distribution is a weighted average of two one-parameter exponential distributions. As a result, the mixture distribution has three parameters, with α represents the mixing probability and β_1 and β_2 represent the scale parameters, while x represents the hourly rainfall amounts.

$$f(x) = \left(\frac{\alpha}{\beta_1}\right) e^{\frac{-x}{\beta_1}} + \left(\frac{1-\alpha}{\beta_2}\right) e^{\frac{-x}{\beta_2}}$$
$$x > 0, 0 \le \alpha \le 1, 0 < \beta_1 < \beta_2$$
$$\log L = \sum_{i=1}^{n} \log \left[\left(\frac{\alpha}{\beta_1}\right) e^{\frac{-x_i}{\beta_1}} + \left(\frac{1-\alpha}{\beta_2}\right) e^{\frac{-x_i}{\beta_2}} \right]$$

4.0 GOODNESS OF FIT TESTS

In determining the best-fit distributions, five quantitative methods are used in this study. In addition, graphical representation will also be presented to justify further the goodness-of-fit tests used.

(i) The mean and median absolute difference between the hypothesized distribution $F(x)$ and the empirical distribution, $F_n(x)$.

$$\text{Mean} = \frac{\sum_{i=1}^{n} \left| F_n(x_i) - F\left(x_i, \hat{\theta}\right) \right|}{n}$$

$$\text{Median} \left| F_n(x_i) - F\left(x_i, \hat{\theta}\right) \right|$$

(ii) Kalmogorov-Smirnov(KS) test calculates the maximum difference between the hypothesized distribution and empirical distribution.

$$D^+ = \max\{i/n - Z_i\}, \qquad D^- = \max\{Z_i - (i-1)/n\}$$
$$KS = \max\{D^+, D^-\}$$

(iii) Cramer-Von-Mises (CVM) calculates the squared difference between $F(x)$ and $F_n(x)$.

$$W^2 = \sum_{i=1}^{n} \{Z_i - (2i-1)/2n\}^2 + \frac{1}{12n}$$

(iv) Anderson-Darling (AD) test calculates the squared difference between $F(x)$ and $F_n(x)$, and divided them by the weight function $[F(x)(1 - F(x))]^{-1}$.

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[(2i-1)\log Z_i + (2n+1-2i)\log(1-Z_i) \right]$$

(v) Akaike Information Criterion (AIC) is derived by minimizing the Kullback Leibler distance between the proposed model and true one. The best model is the one having the smallest AIC. The AIC is given by:

$$\begin{aligned} \text{AIC} &= -2 \log(\text{maximum likelihood}) + 2k \\ &= -2\text{MLL} + 2k \end{aligned}$$

in which k denotes the number of parameters and MLL is the maximum likelihood value.

(vi) A graphical representation that calculates the exceedance probability is used to justify further the above test. The horizontal axis represents the wet hours amounts and the vertical axis represents the compliments of $F(x)$ and $F_n(x)$ that is $[1 - F(x)]$ and $[1 - F_n(x)]$. This plot will display every wet hours data distinctly.

5.0 RESULTS AND DISCUSSION

The summary of the descriptive statistics for the stations used in this study is shown in Table 1. The overall statistics show that the rainfall variability in the studied stations is quite homogenous where the means and variances are ranged from 3 mm/h to 4.3 mm/h and from 5.5 mm/h to 7.3 mm/h respectively. The coefficient of variations for all stations seem to be fairly similar with the highest is 1.794 and the lowest is 1.533.

Table 1 Descriptive statistics of the rainfall amounts for the Wilayah Persekutuan

Station no.	Station names	Duration	Mean	Stdev	CV	Skewness	Kurtosis	No.of wet hours	Max. amount
3015001	Puchong Drop	1982-1990	3.997	7.17	1.794	3.712	18.814	4057	82.10
3116005	Sek. Ren. Taman Maluri	1981-1990	3.663	6.337	1.73	3.768	20.946	6466	92.50
3116006	Ladang Edinburgh Site 2	1981-1990	3.68	6.249	1.698	3.808	20.234	5598	72.70
3216001	Kampung Sg. Tua	1981-1990	3.98	6.102	1.533	3.28	14.504	6074	69.60
3216004	SMJK Kepong	1982-1991	4.3	7.346	1.708	3.736	19.277	4328	75.50
3217001	Km 16 Gombak	1981-1990	3.359	5.682	1.692	3.815	20.313	7102	58.20
3217002	Empangan Genting Kelang	1981-1990	3.145	5.495	1.747	3.901	20.313	6819	57.70
3217003	Km 11 Gombak	1981-1990	3.779	6.318	1.672	3.789	21.830	5551	92.90
3217004	Kpg. Kuala Saleh	1981-1990	4.16	7.046	1.694	3.682	18.399	4549	72.30
3217005	Gombak Damsite	1982-1991	3.768	6.753	1.792	3.771	18.954	3447	70.10
3317001	Air Terjun Sg. Batu	1985-1994	4.042	6.732	1.666	3.524	16.986	5279	69.70
3317004	Genting Sempah	1981-1990	3.018	5.272	1.747	4.2	27.805	7484	83.00

Station 3217003 shows the highest maximum hourly rainfall amount with a value of 92.50 mm/hour and the lowest maximum is Station 3217001 with a value of 57.70 mm/hr. Station 3217001 received the maximum number of wet hours and Station 3015001 received the least. However, all stations are positively skewed and the skewness values are very much the same.

Using the good-of-fit tests that have been discussed in Section 4.0 the best-fit distribution is chosen based on the minimum error. The distributions are ranked according to these criteria. Table 2 shows the results of the tests.

Among the four distributions tested, Mixed-Exponential was found to be the best fitting distributions for all the stations studied where almost all the criteria of goodness-of-fit tests resulted in the minimum error for the Mixed-Exponential. These are followed by the Weibull, the Gamma, and finally the Exponential distributions.

The above results can be verified further by presenting the graphical representations through the plot of the exceedence probability. Due to the limited space, Figure 1 only shows the exceedance probability for Stations 3217001 and 3317004. From both graphs, it clearly shows that the Mixed-Exponential plot is the nearest to the observed plot. The same happened to the other Stations where the Mixed-Exponential plot is the nearest to the observed plot. Therefore, we may conclude that the Mixed-Exponential distribution is the best in describing the hourly rainfall amounts in the Wilayah Persekutuan.

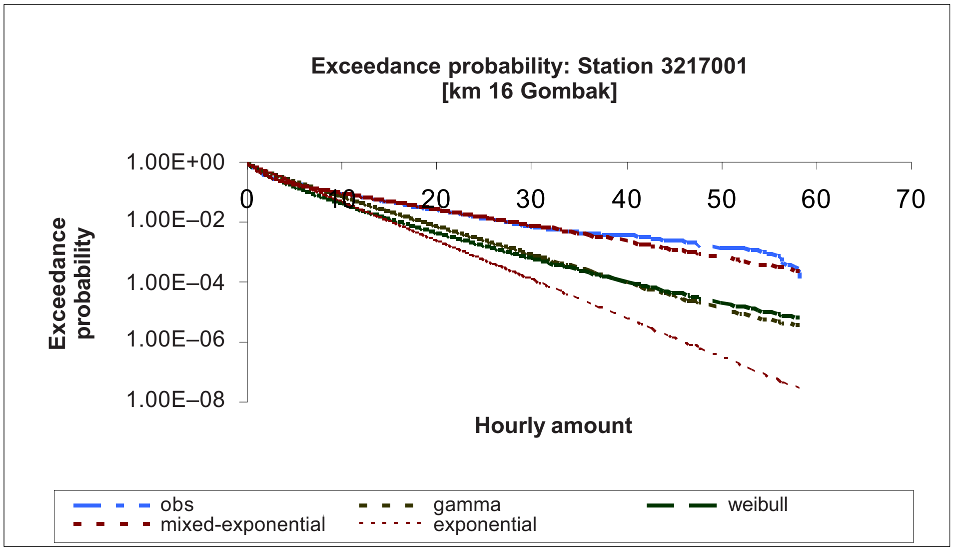
The estimated parameters of the Mixed-Exponential distribution are shown in Table 3. The mixing probability that indicated the percentage of variation of the hourly rainfall amounts in the Wilayah Persekutuan has shown an approximate value of between 0.6 to 0.7. The weighted average of two exponential distributions in the mixed-

Table 2 The ranking of the distributions using AIC and goodness of fit tests

No.	Stations	AIC	KS	CVM	AD	Means	Median
1	3015001	MEX	MEX	MEX	MEX	MEX	MEX
		WE	GM	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP
2	3116005	MEX	GM	MEX	MEX	MEX	MEX
		WE	MEX	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP
3	3116006	MEX	MEXP	MEX	MEX	MEX	MEX
		WE	GM	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP

4	3216001	MEX	MEX	MEX	MEX	MEX	MEX
		WE	GM	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP
5	3216004	MEX	GM	MEX	MEX	MEX	MEX
		WE	MEX	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP
6	3217001	MEX	MEX	MEX	MEX	MEX	MEX
		WE	GM	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP
7	3217002	MEX	MEX	MEX	MEX	MEX	MEX
		WE	GM	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP
8	3217003	MEX	MEX	MEX	MEX	MEX	MEX
		WE	GM	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP
9	3217004	MEX	MEX	MEX	MEX	MEX	MEX
		WE	GM	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP
10	3217005	MEX	MEX	MEX	MEX	MEX	MEX
		WE	GM	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP
11	3317001	MEX	MEX	MEX	MEX	MEX	MEX
		WE	GM	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP
12	3317004	MEX	MEX	MEX	MEX	MEX	MEX
		WE	GM	WE	WE	WE	WE
		GM	WE	GM	GM	GM	GM
		EXP	EXP	EXP	EXP	EXP	EXP

MEX=MIXED-EXPONENTIAL; WE=WEIBULL; GM=GAMMA; EXP=EXPONENTIAL; AIC=AIKAKE INFORMATION CRITERION; KS=KOLMOGOROV-SMIRNOV; CVM=CRAMER-VON-MISES; AD=ANDERSON-DARLING



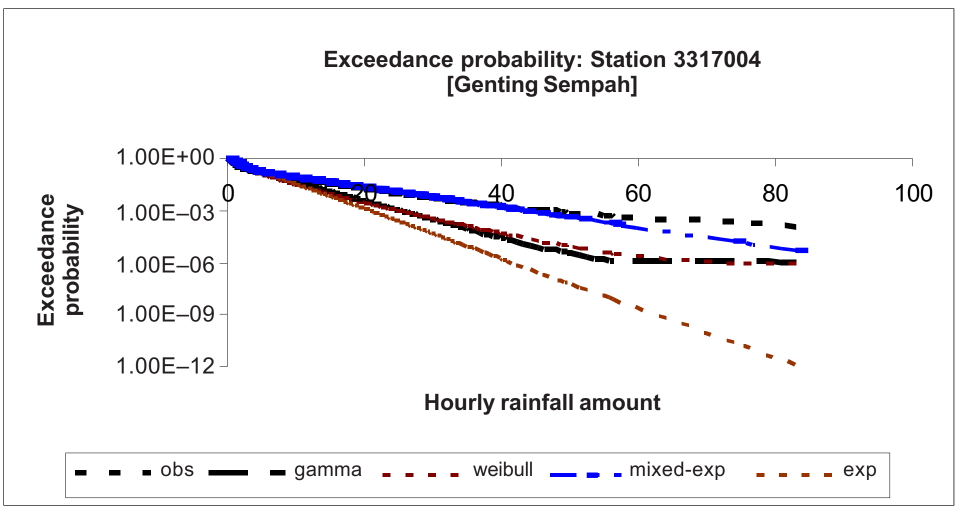


Figure 1 Exceedance probability for Stations 3217001 and 3317004

exponential distributions may refer to the two types of rainfall, namely "light" or "heavy". Hence, it can be interpreted that between 60% and 70% of the hourly rainfall series in the Wilayah Persekutuan is contributed by the light rain. Therefore the remainder is being contributed by the heavy rain. This is true due to the higher frequency of light rain for the hourly data.

The estimated total means and the percentages of the estimated mean contributed by both scale parameters are also calculated. The estimated total means for each

Table 3 The estimated parameters for the mixed-exponential distribution

Station no.	Station names	Mixing probability (α)	Scale 1 (β_1)	Scale 2 (β_2)	Estimated mean
3015001	Puchong Drop	0.6772	1.137 [19%]*	9.996 [81%]*	3.997
3116005	Sek. Ren. Taman Maluri	0.6504	1.077 [19%]*	8.474 [81%]*	3.663
3116006	Ladang Edinburgh Site 2	0.6261	1.108 [19%]*	7.985 [81%]*	3.68
3216001	Kampung Sg. Tua	0.6218	1.44 [23%]*	8.154 [77%]*	3.977
3216004	SMJK Kepong	0.6302	1.253 [18%]*	9.48 [82%]*	4.295
3217001	Km 16 Gombak	0.687	1.193 [24%]*	8.114 [76%]*	3.359
3217002	Empangan Genting Kelang	0.702	1.114 [25%]*	7.93 [75%]*	3.145
3217003	Km 11 Gombak	0.6433	1.211 [21%]*	8.409 [79%]*	3.778
3217004	Kpg. Kuala Saleh	0.6482	1.313 [20%]*	9.4 [80%]*	4.158
3217005	Gombak Damsite	0.6477	1.002 [17%]*	8.853 [83%]*	3.768
3317001	Air Terjun Sg. Batu	0.6245	1.178 [18%]*	8.804 [82%]*	4.042
3317004	Genting Sempah	0.6998	1.066 [25%]*	7.57 [75%]*	3.019

*The percentages in the brackets refer to the estimated means of the hourly rainfall amounts contributed by both scales

station are very close to the observed values given in Table 3. The results also show that from 75% to 82% of the total estimated mean of the hourly rainfall amount for all stations was contributed by the heavy rainfall while the remainder by the light rainfall. Hence, the mean of the hourly rainfall amount in the Wilayah Persekutuan is dominated by the heavy rainfall.

6.0 CONCLUSION

The distribution of the hourly rainfall amounts in the Wilayah Persekutuan is best described by the Mixed-Exponential distribution. The Weibull and the Gamma distribution are ranked second and third respectively, and the last in the ranking is the Exponential distribution. These are based on the good-of-fit tests performed on the studied station, as discussed in Section 4.0.

From the estimated parameters of the Mixed-Exponential distribution obtained, it could be interpreted that between 60% and 70% of the wet hourly series in the Wilayah Persekutuan is contributed by the light rainfall and the remainder by the heavy rainfall. However, the total estimated mean shows that about 80% is attributed to heavy rainfall. This implies that most of the rainfall amount recorded in the study area is received from heavy rains even though there is a higher occurrence of light rainfall. The hourly duration used indicates short duration heavy rainfall has a large impact on the rain amount received and potentially is the main contribution to flash flood events. These would indeed provide grounds for further studies on convective rainfall and flash floods.

REFERENCES

[1] Todorovic, P. and D.Woolhiser. 1975. A Stochastic Model of n-day Precipitation. *J. Applied Meteorology*. 14(1): 17-24.

[2] Woolhiser, D. A. and J. Roldan. 1982. Stochastic Daily Precipitation Models 2. A Comparison of Distributions of Amounts. *Water Resources Research*. 18(5): 1461-1468.

[3] Rodriguez-Iturbe, I., D. R. Cox, and V. Isham. 1987a. Some Models for Rainfall based on Stochastic Point Processes. *Proc. Roy. Soc. London Ser. A*. 410(1839): 269-288.

[4] Nordila, A., Y. Zulkifli, and M. D. Zalina. 2006. Characterization of Convective Rain in Klang Valley. In Proceedings of National Conference on Water for Sustainable Development towards a Developed Nation by 2020. Port Dickson, Negeri Sembilan, MALAYSIA. 13-14 July 2006.

[5] Katz, R. W. 1995. Precipitation as a Chain-dependent Process. *Journal of Applied Meteorology*. 16: 671-676.