Bayesian Changepoint
Analysis of the Extreme Rainfall Events

Wan Zawiah Wan Zin,
Abdul Aziz Jemain and Kamarulzaman Ibrahim
School of Mathematical Sciences, Faculty of Science and Technology
University Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

Abstract: Problem statement: This study assesses recent changes in extremes of annual rainfall in Peninsular Malaysia based on daily rainfall data for 50 rain-gauged stations over the period 1975-2004. **Approach:** Eight indices that represent the extreme events are defined and analyzed, which are extreme Dry Spell (XDS), extreme Rain-Sum (XRS), extreme wet-day intensities at 95% and 99% percentiles (I95 and I99), proportion of extreme rainfall amount to the total rainfall amount (R95& R99) and frequency of extreme wet-day at 95 and 99% percentiles (N95 and N99). Bayesian approach based on a single shifting model is used to investigate the change in the mean level of these extreme rainfall indices. The detection on whether the change has occurred or not is analyzed followed by the estimation of the location of change point. **Results:** The results of the analysis showed that half of the stations considered displayed significant changes. The analysis also found that in general, the changes occurred in the early 90s. More than 75% of the stations which recorded significant changes are situated on the west coast of the peninsula. **Conclusion/Recommendations:** The west coast of Peninsular Malaysia displays more significant changes in trend especially at stations located in urban areas compared to the east coast of the peninsula. In terms of the Bayesian methods used, the existence of any outlier in the data series may influence the result since the analysis is based on mean value which is very sensitive to any outlier.

Key words: Bayesian method, changepoint, time series, hydrology

INTRODUCTION

The phenomenon of extreme precipitation events, which include extreme rainfall and extremely long spell of dry days (drought), are among the most disruptive of atmospheric phenomena. These events may cause significant damage to agriculture, ecology and infrastructures, disruption towards daily activities, accidents and loss of life. In Peninsular Malaysia, the phenomenon of unpredictable rainfall events which increases in frequency lately has brought about damages costing millions of Malaysian ringgit. The increase in massive flood cases including flash flood and landslides in the recent 10 years is believed to be due to the increase in rainfall intensities. On the other hand, prolonged dry condition has forced the local authorities to impose water rationing, resulting in negative impact to daily life. Apart from that, in agricultural areas, crops can suffer damage from excess rainfall as well as extreme dry spells.

In this context, this research provides an insight to the possible change in the rainfall extreme and extreme dry spell for the past 30 years as measured by 8 extreme indices. This is because, any changes in extreme rainfall trend brings great implication to engineering, insurance, town planning and any activities that assumed that climate has been stable for the last few decades. For example, the design of drainage, bridge, retaining wall and dam systems depends on the expected rain amount received during certain time duration. Earlier research on the changes in these extreme rainfall indices at the same area using data from only 8 rain-gauged stations can be found from Zin *et al*. (2010). In their article, the extreme rainfall indices derived from rainfall data at these 8 rain-gauged stations for the period of 35 years have been analysed for any change in trend. The statistical methods used in that study is classical approach where the changes in trend were tested using Mann-Kendall test and linear regression while the change detection points were identified using Pettitt test.

The data: The data consisting of daily rainfall amount from 50 rain gauge stations in Peninsular Malaysia from 1975 to 2004 have been obtained from Malaysian Meteorological Services and Drainage and Irrigation Department.

Corresponding Author: Wan Zawiah Wan Zin, School of Mathematical Sciences, Faculty of Science and Technology,
University Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia Tel: 03-8921 5790 Fax: 03-8921 3289
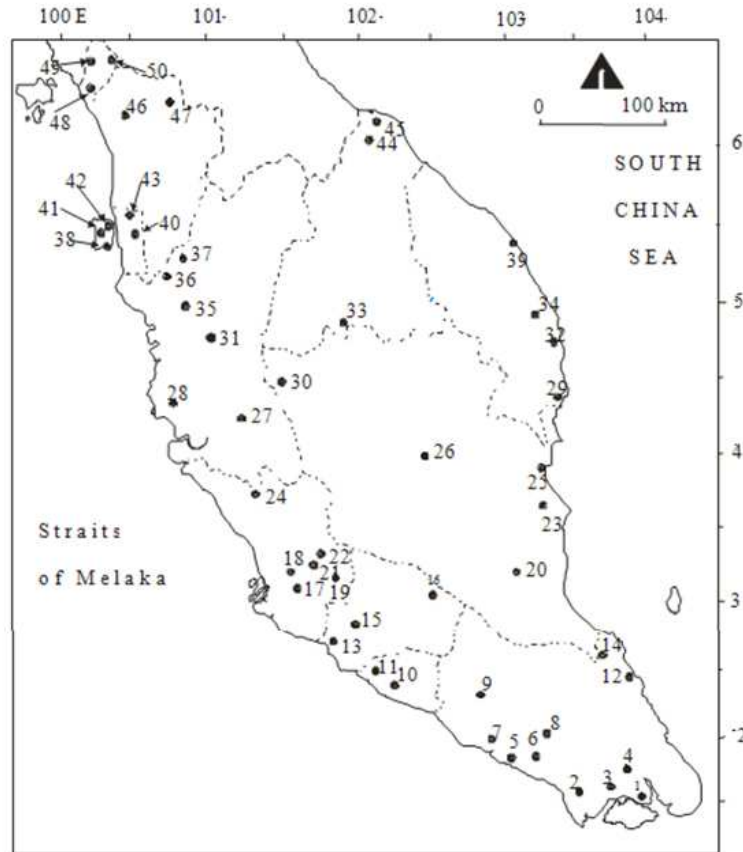
Fig. 1: Location of stations used in this study

All the 50 stations, numbered from 1 to 50, are located at various places throughout Peninsular Malaysia, as shown in Fig. 1. These stations could represent the overall trend for extreme rainfall for the peninsular. We consider the data from 1975 to 2004 because this is the longest period for which a complete set of data is available for all stations. As described Hosking and Wallis (2005), the problem of availability of a large set of data is not uncommon when the analysis is based on the annual series. This situation is common in developing countries where long records are often unavailable; however, studies need to be done for various planning purposes such as for construction of infrastructures.

Peninsular Malaysia experiences a tropical climate due to its location with respect to the equator and the influence of monsoon seasons. It lies in the equatorial zone, situated in the northern latitude between 1°N and 6°N and the eastern longitude from 100°E and 103°E. Throughout the year, the peninsular experiences a wet and humid condition with daily temperature ranges from 25.5-35°C. The two monsoons that contribute to rainy seasons are the Southwest monsoon, occurring in May until September and the Northeast monsoon which occurs from November until March. The latter monsoon brings about heavier rainfall in the peninsular, with the worst affected areas in the east and south. In between these two monsoons are the inter-monsoon periods, occurring in March-April and September-October, bringing intense convective rain to many areas in the peninsula. On the other hand, the driest period for the peninsula usually occurs during the Southwest monsoon with the northern part, on the average, experiencing relatively long spells. Reports on potential water rationing are common during the dry seasons. Apart from the two monsoons, the weather in Peninsular Malaysia is also influenced by two climatic phenomenon, known as El Nino and La Nina which are believed to cause abnormal weather conditions. In Malaysia, El Nino phenomena, which causes drier condition, has occurred 12 times with the worst in 1982-83 and 1997-98. The worst weather condition caused by La Nina in this country is in 1998-2000, resulting in an increase in the daily rainfall amount.

MATERIALS AND METHODS

Methodology: In this research, the change point analysis focuses on the detection of change in the mean level of the time series data and the location of the change point. Several classical methods in change point analysis include the non-parametric Wilcoxon test, the Student-t test and the sequential Mann-Kendall test. The alternative for these classical approaches is Bayesian approach which takes into consideration prior information, the model of the shift assumed and observed data into forming a posterior distribution to model the associated analysis. The statistical inferences on the unknown parameters with regards to change point location can then be made based on this posterior distribution. Bayesian approach in change identification problem for mean level in time series data have been used by previous researchers such as by Smith (1975), Lee and Heighinian (1977), Booth and Smith (1982), Perreault *et al.* (2000a; 2000b) and Kim *et al.* (2009). The Bayesian approach in this study is based on a single shifting model and takes into consideration non-informative prior distributions on the unknown change point.

In contrast to the classical approach, the Bayesian methods take into consideration the parameters of a model as random variables represented by a statistical distribution (prior distribution) rather than fixed values. The Bayesian methods allow the integration of statistical analysis through the prior distribution with the most current information based on the observations into a posterior distribution. In other words, the prior distribution reflects beliefs about the parameters prior to experimentation; the posterior distribution provides an updated belief about the parameters after sample data is obtained. The analysis involved includes getting the mean values before and after the shift, the amount of change and the variation in observations. In this study, two related problems will be analysed that is on the detection of the change point and the estimation of the change point.

There are several definitions used to describe the extreme precipitation events. In this study, eight extreme indices were examined based on daily rainfall data at 50 stations. Some of the indices chosen are standard extreme precipitation indices as defined by The Expert Team on Climate Change Detection, Monitoring and Indices (ETCCDMI).

The ETCCDMI indices considered in this study are maximum dry spell, maximum cumulative rain-sum, extreme intensities, extreme frequencies and extreme proportions (Table 1). A wet day is defined as a day with a rainfall amount of at least 1 mm. The dry spell index calculates the longest dry spell (rainfall amount less than 1mm) recorded each year. The extreme rain-sum index measures the greatest cumulative rainfall amount received during a wet spell in a year. This index is considered as flooding usually occurs when infrastructure is unable to accommodate the amount of excess water during prolonged and continuous rainy days. The 95th and 99th percentiles are selected as the thresholds to represent extreme events. Extreme proportion measures the proportion of annual rainfall amount to the total amount of annual rainfall received. Extreme frequency is a count of rainfall events per year which equal to or above the long-term (1975-2004) mean of 95th and 99th percentiles. We will refer to rainfall exceeding 95th percentile as very wet days and exceeding the 99th percentile as extremely wet days.

Consider a set of hydrological time series $X = \{x_1, x_2, ..., x_n\}$ whose due to some exogenous factors, the first and second segments fluctuate, around two different mean levels, denoted by μ_1 and μ_2 respectively but with the same variance, σ^2. Assume that the single change point has occurred at time point τ and the time series follow a Normal distribution Eq. 1:

$$X_i \sim N(\mu_1, \sigma^2), i = 1, ..., \tau$$
$$X_i \sim N(\mu_2, \sigma^2), i = \tau + 1, ..., n \qquad (1)$$

where $N(\mu_k, \sigma^2)$ represents a Normal distribution with density function given as Eq. 2:

$$f(\mu_k, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma^2}\right), x \in \mathbb{R} \qquad (2)$$

The parameters τ, μ_1, μ_2 and σ^2 represent the change point, the mean before and after the shift and the variance of the series, respectively. The prior distribution of μ_1 and μ_2 are assumed to be the same Normal distribution, denoted by Eq. 3 and 4:

$$\mu_1 \sim N(\mu_0, \sigma_0^2) \qquad (3)$$

$$\mu_2 \sim N(\mu_0, \sigma_0^2) \qquad (4)$$

With large σ_0^2, these distributions will approach the non-informative prior distributions.

The variance of the series σ^2, is assumed constant and estimated by Eq. 5:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}; \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (5)$$

Table 1: Rainfall indices with their definition and units. "Events" in this table refers to the daily rainfall rate

Definitions and units	Indicator name and ID
Maximum number of consecutive dry days (days)	Dry spell (XDS)
The maximum cumulative total rainfall collected during a wet spell in a year (mm)	Extreme Rain-Sum (XRS)
Average intensity of events greater than or equal to the 95th i.e. average four wettest events (mm)	Extremely wet day intensity (I99)
Percentage of total rainfall from events greater than or equal to the 95th percentile (%)	Very wet day proportion (R95)
Percentage of annual total rainfall from events greater than or percentile, i.e. average eighteen wettest events (mm)	Very wet day intensity (I95)
Average intensity of events greater than or equal to the 99th percentile, equal to the 99th percentile (%)	Extremely wet dayproportion (R99)
Frequency of daily rainfall exceeding the 1971-2005 mean 95th percentile (days)	Very wet days (N95)
Frequency of daily rainfall exceeding the 1971-2005 mean 99th percentile (days)	Extremely wet days (N99)

After a sample of time series X is observed, the posterior distribution of the mean levels μ_1 and μ_1 can be determined using Bayes theorem Eq. 6 and 7:

$$\mu_1 \mid X_\tau \sim N\left(\mu_1^*, \sigma_1^{*2}\right); \mu_2 \mid X_{\tau+1} \sim N\left(\mu_2^*, \sigma_2^{*2}\right) \tag{6}$$

where:

$$\mu_1^* = \frac{n^* . \mu_0 + \sum_{i=1}^{\tau} x_i}{n^* + \tau}, \qquad \sigma_1^{*2} = \frac{\sigma^2}{n^* + \tau}$$

$$\mu_2^* = \frac{n^* . \mu_0 + \sum_{i=\tau+1}^{n} x_i}{n^* + (n-\tau)}, \qquad \sigma_2^{*2} = \frac{\sigma^2}{n^* + (n-\tau)}$$

$$n^* = \frac{\sigma^2}{\sigma_0^2}.$$

The likelihood function for τ, μ_1 and μ_1 can be derived using this formula,

$$L\left(X \mid \tau, \mu_1, \mu_2\right) = \prod_{i=1}^{\tau} f(x_1) \prod_{i=\tau+1}^{n} f(x_2)$$

$$= \prod_{i=1}^{\tau} \frac{1}{\sqrt{2\pi\sigma}^2} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma^2}\right) \tag{7}$$

$$\prod_{i=\tau+1}^{n} \frac{1}{\sqrt{2\pi\sigma}^2} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma^2}\right)$$

Using the Bayes theorem, the posterior distribution of the change point location, τ is Eq. 8:

$$\pi\left(\tau \mid X, \mu_1, \mu_2\right) = \frac{L\left(X \mid \tau, \mu_1, \mu_2\right).p(\tau)}{\sum_{j=1}^{n-1} L\left(X \mid j, \mu_1, \mu_2\right).p(j)} \tag{8}$$

where p(j) represents the prior distribution of the change point location, τ, assumed to follow a uniform distribution, that is $p(j) = 1/(n-1), j = 1, ..., n-1$. Thus Eq. 9:

$$\pi\left(\tau \mid X, \mu_1, \mu_2\right) = \frac{L\left(X \mid \tau, \mu_1, \mu_2\right)}{\sum_{j=1}^{n-1} L\left(X \mid j, \mu_1, \mu_2\right)} \tag{9}$$

The justification on whether a shift has occurred or not can be checked using Bayes factor, B Eq. 10:

$$B = \frac{p}{1-p} . \frac{1}{\pi\left(\tau = n \mid X, \mu_1, \mu_2\right)} \sum_{j=1}^{n-1} \pi\left(\tau = j \mid X, \mu_1, \mu_2\right) \tag{10}$$

p is a constant such that $0 \le p \le 1$.

The calculation in this procedure may not be expressed in a simple form but it can be estimated by using Monte Carlo Markov Chain approach.

RESULTS AND DISCUSSION

Table 2 shows the results on the trend identification using Bayesian method at 90% confidence level. Stations with significant shift in mean of extreme indices are indicated with bold letters in respective boxes for increasing trend and shaded boxes for decreasing trend. It is found that half of the stations studied showed significant trend for at least one extreme rainfall index with station 26 (Rumah Pam Paya Kangsar) and 40 (Bukit Berapit) having the most significant changes in the indices that is as much as 5 indices with significant change in trend. XDS index is found to be decreasing significantly while majority of other indices showed significant increase. The next part of the analysis that is the identification of the change point location is also carried out. Bayesian analysis to detect the change point location is performed on all stations with significant shift in trend as listed in Table 2.
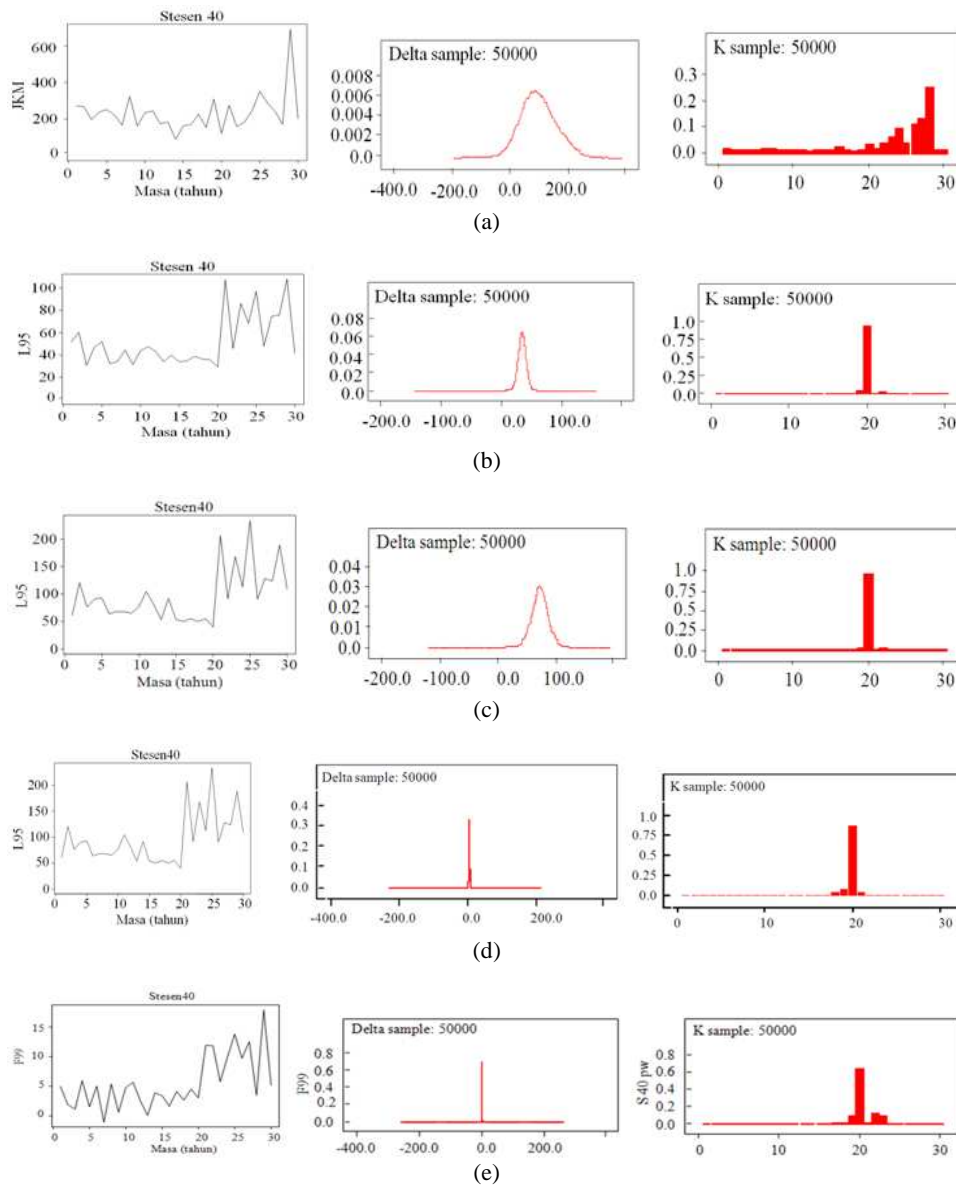
Fig. 2: Time series plot (above) and the posterior probability distribution of the changepoint location (below) for station 40 for extreme indices (a) XRS, (b) I95, (c) I99, (d) N95 dan (e) N99

As an example, the results for station 40, which is one of the two stations with most number of significant changes in extreme indices, are displayed in Fig. 2. For each pair of graphs, the top graph displays the time series plot of the extreme index data while the graph at the bottom of the pair shows the posterior probability plot for change point that is the probability that the change point occurs at a particular point. The largest value represents the point when the shift in trend is most likely to occur.

Table 2 also shows the year when the shift in trend for each station with the related indices. In terms of year when the changes occurred, it is found that most changes occurred at the end of 1980s to early 1990s. Despite that, there are several indices showing that changes occurred at the initial period when data was taken (before 1980) that is for XDS (station 11), I95 and I99 (station 30), R95 (station 11), R99 (station 27) and N95 (station 32). This may implies that the Bayesian method is more sensitive towards any change in the data although the number of data may be short.

Table 2: The year when the change point is detected for each station with significant trend. The bold letter refers to significantly increasing trend and the shaded box refers to significantly decreasing trend

Station	XDS	XRS	I95	I99	R95	R99	N95	N99
2	1993		1982	1983				
5		1996						
6				2003		2003		
7							1996	
8				2003		2003		
11	1997				1978			
13					1988	1994		
14	1983							
16	1984							
17		1997	1990	1991				
18			1986	1986			1987	
20							2002	
21	1991							
22				1992				1992
26		2002	2002	2002	2002			1999
27					1996	1977		
30			1975	1976				
32							1976	
33					1980			
40		1996	1994	1994			1994	1994
43			1995				1996	
44					1980			
46	1985							
47	1990							
50	1987					1983		

Nevertheless, the existence of data with relatively larger value compared to other data (outlier) may influence the analysis results as this analysis is based on the change in mean value that is rather sensitive towards any outlier in data.

In terms of climatology, the two extreme El-Nino events occurring in 1982/83 and 1997/98 may be the contributing factor to the change in climate as detected by the Bayesian change point detection test. Apart from that, majority of the stations located at the west coast of the peninsula experience significant changes in the studied indices. In general, these areas experience rapid development in late 1980s to early 1990s. This factor may contribute to the obvious climate change compared to the east coast of the peninsula. Shaharuddin (1992; 2004) discovered that there exist effects from Urban Heat Island at big cities which influence the temperature change and directly cause an increase in rainfall intensity at these areas. This can be seen clearly at stations located in Selangor and Federal Territory (stations 17, 18 and 22).

CONCLUSION

As a whole, the west coast of Peninsular Malaysia displays more significant changes in trend compared to the east coast of the peninsula. The increase in significant trend at urban area as seen at station 17, 18 and 22 for extreme cumulative rainfall amount, extreme intensities and extreme frequency need to be viewed with caution as this area is a highly populated area. Many factors such as rapid township development, industrialisation, increase in the number of vehicles and population may influence the pattern of rainfall for this area where the change in trend is found to begin in the 1980s to early 1990s. Apart from that, climate phenomena such as El-Nino and La-Nina may play important role in determining the weather pattern in Peninsular Malaysia

In terms of the Bayesian methods used, the existence of any outlier in the data series may influence the result since the analysis is based on mean value which is very sensitive to any outlier. This situation may cause the Bayesian change point analysis to show significant change although in fact the other points in that respective station are actually consistent.

REFERENCES

Booth, N.B. and A.F.M. Smith, 1982. A Bayesian approach to retrospective identification of change-points. J. Econ. 19: 7-22. DOI: 10.1016/0304-4076(82)90048-3

Hosking, J.R.M. and J.R. Wallis, 2005. Regional Frequency Analysis: An Approach Based on L-Moments. Cambridge University Press, Cambridge, ISBN-10: 0521430453 pp: 224.

Kim, C., M.S. Suh and K.O. Hong, 2009. Bayesian changepoint analysis of the annual maximum of daily and subdaily precipitation over south korea. J. Climate, 22: 6741-6757. DOI: 10.1175/2009JCLI2800.1

Lee, A.S.F. and S.M. Heghinian, 1977. A shift of the mean level in a sequence of independent normal random variables: A Bayesian approach. Technometrics 19: 503-506. DOI: 10.2307/1267892

Perreault, L., J. Bernier, B. Bobee and E. Parent, 2000a. Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited. J. Hydrol. 235: 221-241. DOI: 10.1016/S0022-1694(00)00270-5

Perreault, L., J. Bernier, B. Bobee and E. Parent, 2000b. Bayesian change-point analysis in hydrometeorological time series. Part 2. Comparison of change-point models and forecasting. J. Hydrol. 235: 242-263. DOI: 10.1016/S0022-1694(00)00271-7.

Shaharuddin, A., 1992. Some effects of urban parks on air temperature variations in Kuala Lumpur, Malaysia. Proceedings of the 2nd Tohwa University International Symposium on Urban Thermal Environment, Sep. 7-10, Tohwa University, Fukuoka, Japan pp: 107-108.

Shaharuddin, A., 2004. Trends and variability of rainfall in Malaysia: A case of Kuala Lumpur and Kuantan, Pahang. Proceedings of the 3rd Annual Hawaii International Conference on Social Sciences, Jun. 16-19, Waikiki, Hawaii.

Smith, A.F.M., 1975. A Bayesian approach to inference about a change-point in a sequence of random variables. Biometrika 62: 407-416. DOI: 10.1093/biomet/62.2.407

Zin, W.Z.W, S. Jamaludin, S.M. Deni and A.A. Jemain, 2010. Recent changes in extreme rainfall events in Peninsular Malaysia: 1971-2005. Theoretical and Applied Climatology 99: 303-314. DOI: 10.1007/s00704-009-0141-x