Modelling extreme precipitation: an application to two selected rainfall stations in Malaysia

Aida Adha Mohd Jamil[1*], *Rossita* Mohamad Yunus[2], and *Yong Zulina* Zubairi[3]

[1]Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Selangor, Malaysia
[2]Institute of Mathematical Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia
[3]Mathematics Division, Centre for Foundation Studies in Science, University of Malaya, 50603 Kuala Lumpur, Malaysia

Abstract. Statistical models of rainfall have been applied in the understanding of the rainfall past trends, identifying for any anomalies, and making projections of future climate change in Malaysia. Herein, we analyse the rainfall data of 7-year period using the gamma and beta regression models to fit Malaysian extreme precipitation events of two stations, each in the West Coast region and the East Coast region, with extreme precipitation calendar date (in the angular form) as the predictor of the models. While the significance test as the p-value is much less than 0.05, it shows that there is a significant relationship between the climatology response variables. The deviance residual plot will be used to check the goodness of fit for diagnostic checking. The results show the models are useful in highlighting the latest trends and projections of climate change in Malaysia.

1 Introduction

The use of statistical models of rainfall has been applied worldwide to give a better understanding about the rainfall pattern and its characteristics. This process involves the understanding of the past trends, identifying for any anomalies, and making projections of future climate change in Malaysia. There are many studies in the literature focusing on fitting rainfall data with a distribution, such as the Normal, Log-normal, Gamma and Weibull [1, 2]. Based on [3], the rainfall events can be modelled as a Poisson process whereas the intensity of each rainfall event is Gamma distributed. By assuming rainfall arrives in forms of storms following a Poisson process, and the current intensity at each arrival time increases by a random amount based on Gamma distribution. Therefore, rainfall volume and occurrence could be modelled by using the classical regression model with the assumption that the response variable is normally distributed. However, the response variable is not always normally distributed in real data.

Malaysia is located near the equator and well known for its hot and humid tropical rainfall climate all the year. Peninsular Malaysia and East Malaysia (Malaysian Borneo) are two major parts of country that located in the same latitudes and are influenced by wind, El Niño

* Corresponding author: aidaadha@utar.edu.my

effect, and monsoon seasons. Malaysia has temperature average from 25ºC to 35ºC and faces two monsoon winds seasons. Monsoon is a significant wind system that changes its direction according to seasons. The Northeast Monsoon (NEM) is coming from China and the North Pacific, occurs between October and March has attributed heavy rainfall to East Coast of Malaysia [4]. On the other hand, West Coast of Malaysia is affected by the Southwest Monsoon (SWM) from the deserts of Australia that occurs between May and September as shown in Fig. 1.

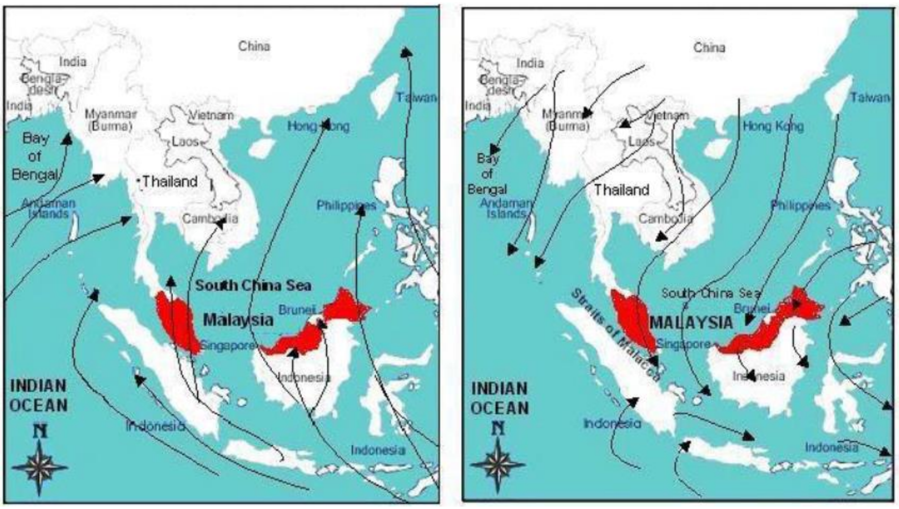


Fig. 1. Southwest (left) and northeast (right) monsoons around South-East Asia [5].

The most significant effect of climate change can cause rise of sea level, heavy rainfall, and lead to drought. Rainfall is essential for life in the earth, but frequent of extreme rainfall events may lead to serious flooding that cause massive loss in agriculture and fisheries. To anticipate the loss, statistical rainfall modelling has remarkable applications in decision-making in the face of bad predictions [2]. Prevention of damage to private and public property, and avoidance of health and ecological dangers are the impact of notable rainfall model.

It is noted that past research approaches fit generalized linear models (GLMs) to monthly or daily rainfall totals using potential predictors [6], however this method has limitation to fit extreme precipitation events. The purpose of this study is to analyse the rainfall data of 7-year period using regression models to fit Malaysian precipitation events more than 30 mm per day of two stations, each in the West Coast region and the East Coast region, with extreme precipitation calendar date in the angular form. Section 2 discusses the data used and preliminary results in this study. The methodology of extreme precipitation's predictors will be explored by using the log-gamma and beta regression models and how the best model is determined by using rainfall data in Section 3. Section 4 discusses the results based on the help of summary statistics in previous section. Concluding remarks are given in Section 5.

2 Data and preliminaries

The sample data is provided by Malaysian Meteorological Department and is located at 2 stations - Bayan Lepas in West Coast region, Muadzam Shah in West Coast, Malaysia (Fig. 2). Daily rainfall (mm) is the amount collected over 24-hour period at 0800 from 2008 to 2014. The studied data series from 1st May to 30th September for Station 1: Bayan Lepas and

1st October until 31st March for Station 2: Muadzam Shah by following the monsoon's period. Based on Malaysian Meteorological Department, the extreme precipitation happened when the total amount of rainfall exceeds 240mm per day. The red (danger) code warning will be issued to the residents. During the monsoon, the amount of rainfall is between 100mm to 200mm per day. The purpose of preliminary data analysis is to obtain the descriptive statistics to study the pattern and seasonality of rainfall data in Malaysia.

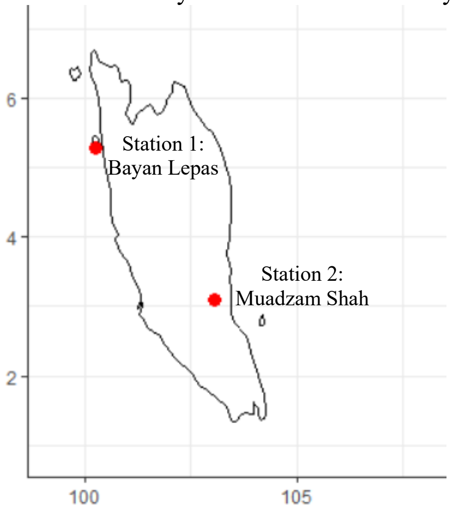


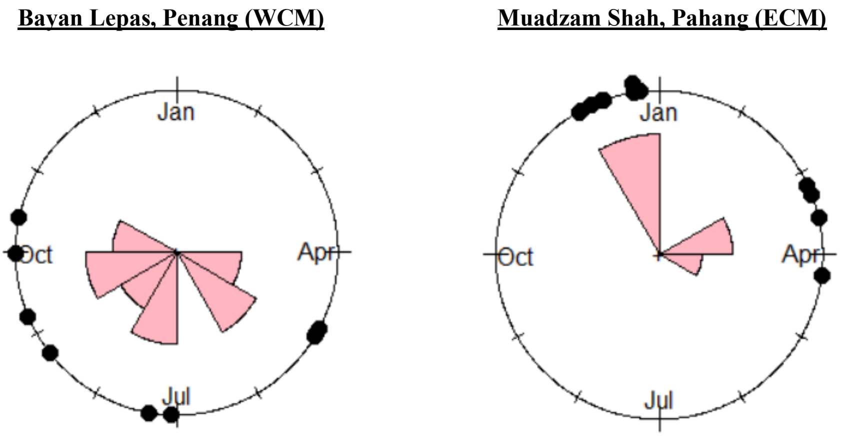
Fig. 2. The map location for 2 stations.

Table 1 shows descriptive statistics of the rainfall totals of the studied stations. Parameters $\bar{\theta}$, ρ, and csd describe accessible station-by-station summary of maximum daily precipitation variability for the 2008 to 2014 period throughout the year.

Table 1. Descriptive statistics of the daily rainfall for Bayan Lepas and Muadzam Shah stations.

Station	Mean (mm)	Percentage of days with no rainfall	Circular Statistics	
			Mean Direction, $\bar{\theta}$ (degree)	Standard Deviation
West Coast, Malaysia (WCM) Bayan Lepas, Penang	7.35	43.06	196.21 *(July)*	1.01
East Coast, Malaysia (ECM) Muadzam Shah, Pahang	6.89	51.70	8.53 *(Jan)*	0.87

Muadzam Shah recorded the higher percentages of days with no rainfall and has lower mean of rainfall compared to Bayan Lepas. Variation of rainfall patterns is observed due to have differences in monsoon experience. The rainy periods are very correlated with monsoon winds that hit different sections of the country throughout the year. The Southwest Monsoon (SWM) influences the rainfall of Bayan Lepas, a station located at the northern region of the west coast of Peninsular Malaysia. Meanwhile, Muadzam Shah is located at the east coast of Peninsular Malaysia where the rainfall climate is affected by direct exposure to the Northeast Monsoon (NEM) [2].


Bayan Lepas, Penang (WCM)

Muadzam Shah, Pahang (ECM)

Fig. 3. Circular plots for these two stations.

In the following analysis, we run the analysis by a year to investigate further if this pattern is prevalent annually. Here, we assume that the 365 days in a year is equivalent to a circle. The plots in the circle is represented the occurrence of amount of rainfall more than 100mm per day from 2008 until 2014. The rose diagram in Fig. 3 shows the occurrence of extreme rainfall for East Coast (ECM) and West Coast (WCM), respectively. Based on circular plots, its illustration that the extreme precipitation events occur during the monsoon season, whereas heavy rainfall from December until April can be observed in Muadzam Shah during the NEM, and from May until October in Bayan Lepas during the SWM.

3 Methodology

At the first stage of the regression model, the angular value, θ_i for extreme event "i", is computed in radians and the values are angles in the range of [0,2π) or [–π, π). The angular position of the date of occurrence (D) for extreme precipitation event "i" is defined as in [7]:

$$\theta_i = D_i\left(\frac{2\pi}{365}\right) \quad (1)$$

D = 1 for 1st of January and D = 365 for 31st of December (D = 366 for leap year). In terms of angular value in radians, 0 radian corresponds to 1st of January and 2π radians is 31st of December. The x and y coordinates of the mean extreme precipitation date throughout the year for a sample of n extreme precipitation events are calculated based on the following equation:

$$\bar{x} = \frac{\sum_{i=1}^{n} \cos(\theta_i)}{n} \quad (2)$$

$$\bar{y} = \frac{\sum_{i=1}^{n} \sin(\theta_i)}{n} \quad (3)$$

where $\bar{x}$ and $\bar{y}$ are the coordinate of the mean extreme precipitation date. The mean date of occurrence of n extreme precipitation events is represented by direction and obtained using:

$$\bar{\theta} = tan^{-1}\left(\frac{\bar{x}}{\bar{y}}\right) \quad (4)$$

Let $X_1, \ldots, X_n$ be a random sample of X's, [6] used the model between X and $\bar{\theta}$ for the regression of a linear variable on a circular variable for fixed $\theta_1, \ldots, \theta_n$, given by

$$X = a + b\cos(\theta_i - \bar{\theta}) + \epsilon \quad (5)$$

where ϵ has 0 mean and a variance σ^2, a and b are the intercept and slope parameters, respectively, and $\bar{\theta} \in (0, \pi]$ is the mean direction of θ_i , . . . , θ_n.

3.1 Gamma GLM

Assessing the rainfall predictors in the model is a main concern since the rainfall data have skewed distributions. The only positive values in gamma distribution is an advantage in climatological variables as the lowest limit value of rainfall is equal to zero [8]. In this study, the GLM with gamma distributions is fitted as the continuous random variables y fits to the probability density function of

$$f(y_i|\alpha_i,\ \varepsilon_i) = \frac{{\varepsilon_i}^{\alpha}}{\Gamma(\alpha_i)} y^{\alpha_i - 1} e^{-\ \varepsilon_i y_i}; y_i, \alpha_i, \varepsilon_i \geq 0 \tag{6}$$

with α_i is an inverse-scale parameter, ε_i is a shape parameter, and $\Gamma(\alpha)$ is a gamma function. It can be shown that the distribution belongs to the exponential family by re-arranging the density for the Gamma as follows:

$$f(y_i|\alpha_i,\ \varepsilon_i) = exp\left\{\left(\frac{y_i\left(-\frac{\varepsilon_i}{\alpha_i}\right)+\ln\,(\,\varepsilon_i)}{\frac{1}{\alpha_i}}\right) + (\alpha_i - 1)\ln(y_i) - \ln\,(\Gamma(\alpha_i))\right\} \tag{7}$$

with the canonical parameter, θ for the gamma family is $-\ {}^{1}\!/_{\mu}$. The mean of distribution is $E(Y) = \mu$ and variance is $Var(Y) = \mu^2$ [9].

3.2 Beta regression model

The beta distribution is unpopular the model of choice for fitting the rainfall continuous data. The response variables need to be in interval $(0,1)$ by normalization process. However, the flexibility gives in terms of the variety of shapes to accommodate the data fails to account for asymmetries between variables. The probability density function of a beta distributed random variables y parameterized in terms of its mean μ and a precision parameter ϕ is given by

$$f(y_i|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\big((1-\mu)\phi\big)} {y_i}^{\mu\phi - 1}(1 - y_i)^{(1-\mu)\phi - 1};$$
$$0 < y_i < \ 1,\ \ 0 < \mu < \ 1, \phi > 0, \tag{8}$$

where $\Gamma(\,.\,)$ denotes the gamma function, $E(Y) = \mu$ and $Var(Y) = \frac{\mu(1-\mu)}{1+\phi}$ [10].

A link function g_1 will connect the covariate vector $\mathbf{x}_i$ to random sample $Y_1, \ldots, Y_n\ of\ Y$ that maps the mean interval (0,1) onto the real line as $g_1(\mu) = {\mathbf{x}_i}^{\top}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the vector of regression parameters, and the first element of $\boldsymbol{\beta}_0$ is the intercept. In gamma GLM analysis, the log link function, $g(\mu) = \ln \mu = {\mathbf{x}_i}^{\top}\boldsymbol{\beta}$ with $\mu = \ E(Y)$ is used. The relation rainfall model between linear response variables and a circular explanatory variable is developed using this link function. Amount of rainfall (mm), mean of wind speed (m/s), and solar radiation (MJm^{-2}) are predictors of extreme precipitation will be explored in this study. For diagnostic checking, deviance residual plot is used to check the goodness of fit. All residuals are positioned on or near to the straight line and no large deviations were observed are suggesting well-fitting models [11]. R programming language is used for data analysis and graphical displays.

4 Results and discussions

Rainfall models with weather predictors are valuable to explore the trends of rainfall in Malaysia. The correlation values and patterns in the Fig. 4 illustrates the behaviour of the coefficients associated with the meteorological predictors of extreme precipitation.

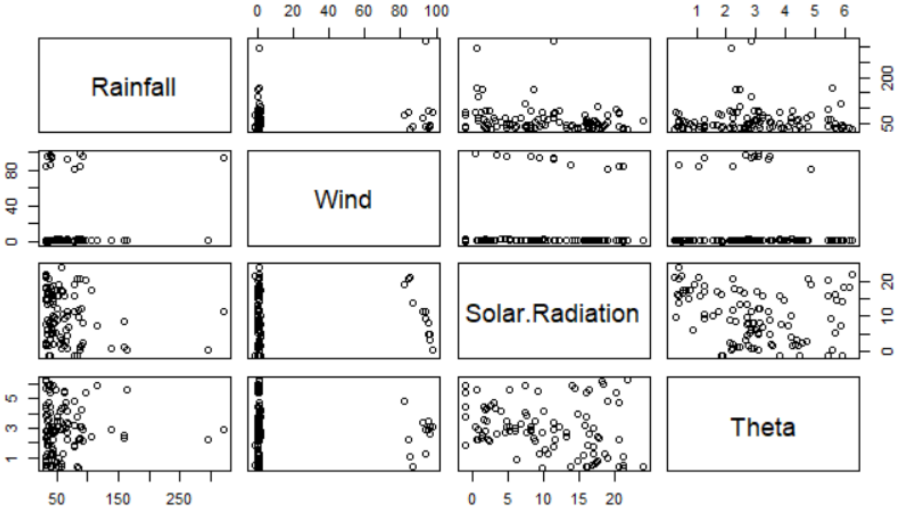


Fig. 4. Scatterplots matrix for rainfall (mm), theta (θ_i, radian), mean of wind speed (m/s), and solar radiation (MJm^{-2}) at Muadzam Shah, Pahang.

As expected, non-significant correlation between each predictor, which is commonly occurred for the right skewed rainfall data. Besides that, the pattern and variation of rainfall data are always assumed to be continuous with time and found to be seasonal for each monsoon winds seasons throughout the years. Therefore, the alternative approach must develop to fit this situation. The importance of understanding the pattern of rainfall is emphasized in order to develop the new statistical model of the occurrence (dry/wet days) and amount (rainfall totals on wet days) of daily rainfall data with a set of predictors simultaneously. 7-year period of rainfall amount, mean of wind speed, and solar radiation influenced of the southwest and northeast monsoons were analysed by using the log gamma and beta regression models to fit Malaysian extreme precipitation events of two stations extreme precipitation calendar date in the angular form. As a result, the following 3 univariate models were compared as below:

Table 2. Details of fitted models.

Model	Response Variables	Explanatory Variables
1	Amount of Rainfall (mm)	cos $(\theta_i - \bar{\theta})$
2	Mean of Wind Speed (m/s)	cos $(\theta_i - \bar{\theta})$
3	Solar Radiation (MJm^{-2})	cos $(\theta_i - \bar{\theta})$

where θ_i is angular position of the date of occurrence for extreme precipitation event "i", and $\bar{\theta}$ is mean date of occurrence of n extreme precipitation in angular form.

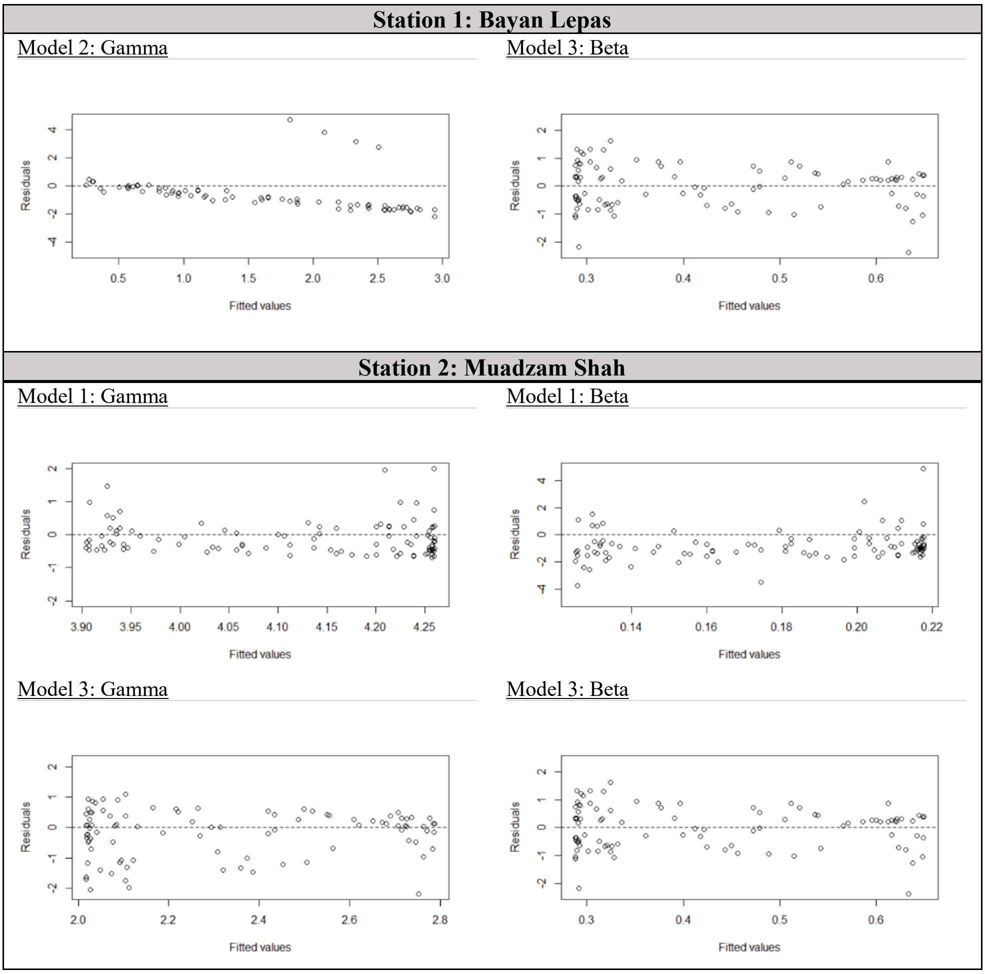
Results of the models fitted to the rainfall data from the analysed stations are shown in Table 3. The significance of the individual predictors on daily rainfall amounts in angular form was assessed using respective level of significance. As seen in Table 3, Model 1 and Model 3 are significant for gamma and beta distribution for Station 2 at East Coast (ECM). Meanwhile, only Model 2 for gamma distribution and Model 3 for beta distribution are significant for Station 1 at West Coast (WCM).

Table 3. Comparison of the parameter of regression models for Station 1: Bayan Lepas and Station 2: Muadzam Shah.

Model	Station	Distribution					
		Gamma			Beta		
		Intercept $\boldsymbol{\beta}_0$	Estimate $\boldsymbol{\beta}_1$	p-value	Intercept $\boldsymbol{\beta}_0$	Estimate $\boldsymbol{\beta}_1$	p-value
Model 1 Amount of Rainfall ~ cos $(\theta_i - \bar{\theta})$	1	4.083	-0.0225	0.791	-0.9338	-0.0527	0.799
	2	4.0818	0.1784	**0.0501***	-1.6126	0.3340	**0.0206***
Model 2 Mean of Wind Speed ~ cos $(\theta_i - \bar{\theta})$	1	1.5894	-1.3545	**0.0030***	-1.6188	-0.0589	0.779
	2	2.2876	0.5886	0.108	-1.3361	0.2113	0.188
Model 3 Solar Radiation ~ cos $(\theta_i - \bar{\theta})$	1	2.5717	0.18027	0.0782	-0.5267	0.3967	**0.0500***
	2	2.4035	-0.3858	**1.34 e-05***	-0.1422	-0.7624	**5.75 e-07***

Bold p-value indicates the ideal model, and no superscript means not significant.

To assess the best fitted model adequacy, Fig. 5 is plotted for model checking. It is depicted that there are no wildly deviant observations in the residual plots. The relationship between the observed and predicted responses looks reasonably linear and the fit is good. Thus, the assumption of homogeneous variance is satisfied. However, residuals for Model 2 (Gamma distribution) at Station 1 are not evenly distributed and clustering along vertical lines. So, we have a reason to suspect problems with the validity of our conclusion and this model is rejected as the significant model.


Station 1: Bayan Lepas

Model 2: Gamma

Model 3: Beta

Station 2: Muadzam Shah

Model 1: Gamma

Model 1: Beta

Model 3: Gamma

Model 3: Beta

Fig. 5. Residual plots for significant models at Station 1 and Station 2.

This study demonstrates that the amount of rainfall and solar radiation are the important climatological outcome variable during the occurrence of Northeast Monsoon (NEM) since the frequency of extreme rainfall at ECM from December has drastically above average. The contradiction happens at Station 1 indicates no seasonal peaks and has misleading indicator of the dominant South West Monsoon's (SWM) season, where the amount of rainfalls has relatively lower precipitation. WCM has not exposed to SWM and NEM directly due to barrier from North Sumatra, Indonesia [12] and the presence of Titiwangsa Range of Peninsular Malaysia, respectively. At the end of the process, it is concluded that the opposite condition happens for different station locations for each monsoon winds seasons throughout the years.

5 Conclusions

The amount of rainfall at East Coast, Malaysia (ECM) in December has significantly higher and mostly random around the year in areas of West Coast, Malaysia (WCM). The seasonal variation pattern on rainfall is observed on ECM stations; but imprecise of seasonal trend happens on WCM stations. Therefore, further works will include identifying suitable model that can accommodate seasonality component of Malaysian rainfall for both regions. The contradiction between the amount of rainfall at ECM and WCM calls for a better understanding on the rainfall pattern and its characteristics. From the present study, the best fitted models are useful in understanding the amount of rainfall, wind speed, and solar radiation are crucial predictors of extreme precipitation event.

The second author thanks the Fundamental Research Grant Scheme, Ministry of Education Malaysia for funding [Research Grant no FP083-2018A].

References

1. H. Aksoy, Turkish J. Eng. Environ. Sci. **24(6)**, 419 – 428 (2000)
2. R.M. Yunus, M.M. Hasan, N.A. Razak, Y.Z. Zubairi, P.K. Dunn, Int. J. Climatol. **37(3)**, 1391 – 1399 (2017)
3. N.C. Dzupire, P. Ngare, L. Odongo, J. Probab. Stat. 1–12 (2018)
4. J.B. Ooi, A. Zakaria, *Malaysia - Climate | Britannica.com* (2019)
5. S.G. D. Iya, M.B. Gasim, M. E. Toriman, M. G. Abdullahi, Floods in Malaysia: historical reviews, causes, effects and mitigations approach - Scientific Figure on ResearchGate (2014)
6. S. Kim, A. Sengupta, Commun. Stat.-Theor. M. **44(22),** 4772 – 4782 (2015)
7. N. Dhakal, S. Jain, S. Gray, M. Dandy, E. Stancioff, Water Resour. Res. **51(6)**, 4499 – 4515 (2015)
8. R.D. Markovic, Hydrol. Pap. **8**, (1965)
9. P. McCullagh, J.A. Nelder, Generalized linear models, 2nd ed. (Chapman & Hall, 1989)
10. C. Mollica, L. Tardella, Int. J. Numer. Meth. Fluids **33**, 3759 – 3771 (2014)
11. A. Pewsey, M. Neuhäuser, G.D. Ruxton, Circular statistics in R (Oxford University Press, Oxford and New York, 2013)
12. C.L. Wong, Z. Yusop, T. Ismail, J. Liew, R. Venneker, S. Uhlenbrook, Water (Switzerland) **8(11)**, 500 (2016)