GENERALIZED PARETO DISTRIBUTION FOR EXTREME TEMPERATURES IN PENINSULAR MALAYSIA

Nur Hanim Mohd Salleh and Husna Hasan

School of Mathematical Sciences, Universiti Sains Malaysia,11700 USM, Pulau Pinang, Malaysia

*For correspondence; Tel. + (60) 174005636, E-mail: nurhanim_sh@yahoo.com

*For correspondence; Tel. + (60) 46533969, E-mail: husnahasan@usm.my

ABSTRACT: *A Peak over Threshold (POT) method with Generalized Pareto Distribution (GPD) model has been fitted to the daily maximum temperatures data for the period of 1994-2013 in Peninsular Malaysia. A suitable choice of threshold was selected by evaluating the mean residual life plot and the parameter stability plot. The issue of dependence series of temperature excesses above the threshold was dealt by conducting the declustering procedure. The parameters were estimated using Maximum Likelihood Estimator (MLE) and the return level was determined. The adequacy of the fitted model was supported by the diagnostic plots. From the analysis, it was found out that selected threshold value for all stations ranges from 33°C to 35.75°C. In general, the maximum temperatures at all stations increase steadily for higher and higher return period. It is expected that a maximum temperature event will re-emerge within the next 25 to 100 years for all stations except Senai.*

Keywords: Generalized Pareto distribution, peak over threshold, extremal index, maximum temperature, threshold, return level

1. INTRODUCTION

The study of extremes events such as very high temperature, rainfall, earthquakes, storms and other extreme events is importance in natural science. Throughout the year, extreme events happen around the world and cause devastating impact including loss of life, disturbance in the ecosystem, and damage to the infrastructure and agricultural activities. In recent years, extreme maximum temperature events have received much attention due to their relationship with human thermal comfort and biodiversity. Extreme temperature can be understood as a rare event which occurs outside the usual range of adaptability. The characteristics of temperature extremes are varied spatially because of the natural factors (such as longitude, latitude or wind speed) and anthropogenic factors (like deforestation, land conversion or carbon dioxide emission).

Establishing a probability distribution which provides a good fit to temperature extremes has been a subject interest in the fields of meteorology and other related fields. Some of the researchers analyzed the extreme temperature trends by using climate indices while others applied Extreme Value Theory (EVT) to model the extreme temperature events. The EVT becomes one of the most important statistical disciplines for the applied sciences which require estimation of the probability of events that are more extreme than any that have already observed [1]. There are two methods used to identify the movements of the extreme values which are block maxima with Generalized Extreme Value (GEV) approach and Peaks over Threshold (POT) with Generalized Pareto distribution (GPD) approach [2]. GPD arises as the limit distribution for the excess over a threshold and it tends asymptotically to the GEV distribution for a sufficiently high threshold [3].

In Malaysia, the previous studies on the extreme temperatures were often based on statistical analysis of block maxima using GEV distribution [4][5]. The GEV approach models only one observation per block (for example, annual or monthly maximum) which makes it a wasteful approach if other data on extremes are available. The procedure of blocking is better to be avoided if an entire time series of daily observations is available as suggested by Coles [1]. Therefore, the GPD approach is proposed in this research to analyze the temperature values which exceed a fixed threshold at ten meteorological stations in Peninsular Malaysia. Despite the threshold selection difficulties, the GPD approach has advantages over GEV approach as it permits the consideration of more extremes cases per year [6].

2. DATA AND STUDY AREA

Malaysia is a developing country which is located in Southeast Asia. This country has two distinct parts which are Peninsular and East Malaysia. The Peninsular Malaysia is made up of eleven states and two federal territories. With a total area of 131,794 square kilometers, the climate of Peninsular Malaysia is classified as hot and humid throughout the year [7]. The highest recorded temperature is 40.1°C, observed on 9th April 1998 at Chuping and the lowest recorded temperature is 7.8°C, observed on 1st February 1978 at Cameron Highlands [8].

Table 1: The Geographical Coordinates of Meteorological Stations

Station	Location	Longitude	Latitude
CP	Northern	100° 16' E	6° 29' N
AS	Northern	100° 24' E	6° 12' N
BL	Northern	100° 16' E	5° 18' N
KB	Eastern	102° 18' E	6° 10' N
KT	Eastern	103° 06' E	5° 23' N
KLIA	Central	101° 42' E	2° 43' N
MC	Southern	102° 15' E	2° 16' N
MR	Southern	103° 50' E	2° 27' N
MS	Southern	103° 05' E	3° 03' N
SN	Southern	103° 40' E	1° 38' N

In this study, the daily maximum temperatures data obtained from Malaysian Meteorological Department were used. The data were recorded in Degree Celcius (°C) at ten meteorological stations in Peninsular Malaysia as listed in Table 1. Chuping (CP), Alor Setar (AS) and Bayan Lepas (BL) stations are located in the northern part of Peninsular Malaysia while other four stations, Malacca (MC), Mersing (MR), Muadzam Shah (MS) and Senai (SN) are located in the southern part of Peninsular Malaysia. There are two stations located at eastern region of Peninsular Malaysia which are Kota Bharu (KB) and Kuala Terengganu (KT) stations. Only one station is located at central part of Peninsular Malaysia that is Kuala Lumpur International Airport (KLIA) station. Nine of the stations except for KLIA have 20 years period of data that was observed from 1st January 1994 to 31st December 2013. The KLIA station has 15 years period of data that was recorded from 1st January 1999 to 31st December 2013.

As a tropical region, the thermal perceptions for Malaysia residents can be classified by the physiological equivalent temperature (PET) index as in Table 2. The range of temperature was proposed by Lin and Matzarakis [9] and it was then used by Makaremi et al. [10] and Hasan et al. [11] to study the thermal comfort condition in Malaysia and predict the future climate temperature respectively. With regards to the result obtained from both studies, the acceptable climatic condition in Malaysia correspond to the PET value was less than 34°C [10] and after a sufficiently long time, this country will still be experiencing slightly warm temperature with the range of 30°C to 34°C [11].

Table 2: Thermal Perception Classification

Category	Range of Temperature (°C)
Slightly Cool	(22,26)
Neutral	[26-30)
Slightly Warm	[30-34)
Warm	[34-38)

2. METHODOLOGY

In this analysis, Peak over Threshold method with Generalized Pareto Distribution approach is fitted to the extreme maximum temperatures data at the ten meteorological stations. The GPD is characterized by two parameters that are the scale, σ and the shape, ξ which measures variability and determines the tail behavior respectively [12]. For the random variable, $y = X - u$ conditional on $X > u$, the GPD is described by the following distribution function [1]

$$H(y) = \begin{cases} 1 - \left(1 + \dfrac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}, \xi \neq 0 \\ 1 - \exp\left(-\dfrac{y}{\sigma}\right), \xi = 0 \end{cases}$$

defined on $y : y > 0$ and $\left(1 + \dfrac{\xi y}{\tilde{\sigma}}\right) > 0$ with $\tilde{\sigma} = \sigma + \xi(u - \mu)$. u is a given high threshold and μ is the location parameter.

Next, the parameter estimation of GPD was conducted by using Maximum Likelihood Estimator (MLE). The log-likelihood is derived from the distribution function, $H(y)$ as follow:

$$\ell(\sigma, k) = -k \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \log\left(1 + \frac{\xi y_i}{\sigma}\right), \xi \neq 0$$

$$\ell(\sigma) = -k \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{k} y_i, \xi = 0$$

provided $\left(1 + \dfrac{\xi y_i}{\tilde{\sigma}}\right) > 0$ for $i = 1, \dots, k$. The k excesses over a threshold u are labeled as $y_{1,}y_{2,\dots,}y_k$.

Another important issue in applying MLE is to achieve independence assumption. In stationary series, the extremes values tend to occur in clusters which can violate the independence assumption. The clusters can be described as the consecutive occurrences of above threshold. Thus, in order to obtain a set of threshold excesses that are approximately independence, the declustering method was adopted as proposed by Ferro and Segers [13].

This method involves estimating the extremal index, θ to measure the degree of clustering. The index value ranges from 0 to 1 which $\theta \to 1$ which indicates a weak dependence at extreme levels. Using this clustering method, a different run length, r is automatically selected as a function of the degree of clustering of extremes [14].

Moving on, the goodness of fit for the fitted GPD is determined by analyzing the diagnostic plots including probability, quantile, return level and density plots [15]. For the probability and quantile plots, the points of the plots which lie close to the diagonal line prove that the estimated GPD function is an adequate model for the data.

Finally, the return levels are estimated subsequent to the estimation of scale and shape parameters. As pointed by Coles [1], interpreting the extreme value models in terms of quantiles or return levels is more convenient compared to individual parameters. The values of return levels are important for extreme temperature design and management.

The *N*-year return level can be defined as the level expected to be exceeded once every *N* year. It is described by the following equation [1]

$$z_N = \begin{cases} u + \dfrac{\sigma}{\xi}\left[\left(Nn_y\zeta_u\right)^{\xi} - 1\right], \xi \neq 0 \\ u + \sigma \log\left(Nn_y\zeta_u\right), \xi = 0 \end{cases}$$

where N is the return period (year), n is the number of observation per year, and ζ_u is the probability of an individual observation exceeding the threshold u. Modeling of the temperatures data was performed using the *R* software and the extRemes package.

3. THRESHOLD SELECTION

Threshold selection is a critical part of a POT analysis using GPD approach. The threshold needs to be selected carefully as a very large threshold value would exclude too much data and lead to a high variance whereas a too small threshold value would violate the asymptotic basis of the model and lead to bias [1]. In this research, two threshold selection methods recommended by Coles [1] were applied.

These methods include mean residual life plot (also known as mean excess plot) and a complimentary technique of fitting GPD at a range of thresholds (parameter threshold stability plots). The mean residual life plot is simply the sample mean of the events above threshold minus the threshold, plotted against the threshold while the parameter stability plots are the plots of modified scale parameter, σ^* and the shape parameter, ξ against the threshold, u.

The main idea of the mean residual life plot is that the plot should be approximately linear in u, at levels of u for which the GPD is an appropriate model to approximate the excess distribution. In addition to the mean residual life plot, the modified scale parameter and the shape parameter plots are used to find the appropriate threshold u by selecting the lowest value of u where the two parameters estimate remain near-constant. Comparing the above two methods, the mean residual life plot is more difficult to automate in a sensible fashion compared to parameter threshold stability plots [16].

4. RESULT AND DISCUSSION

First of all, the descriptive statistics of the daily maximum temperatures at ten meteorological stations were analyzed. Table 3 shows the minimum (Min), maximum (Max) and mean values of the recorded temperatures for each station as well as their mean, variance (Var) and standard deviation (SD). The highest value of maximum temperature (40.1°C) was observed at Chuping station while the lowest value of maximum temperature (35.6°C) was observed at Bayan Lepas station. Comparing to the other stations, the variance and standard deviation are found to be higher at Muadzam Shah (Var = 4.492, SD = 2.119) and Chuping (Var = 3.835, SD = 1.958) stations which may indicate that the extreme temperatures are relatively more spread in both stations.

Table 3: Descriptive Statistics of the Daily Maximum Temperatures

Station	Min	Max	Mean	Var	SD
AS	24.4	39.1	32.67	3.330	1.825
BL	25.1	35.6	31.64	1.657	1.287
CP	23.7	40.1	32.8	3.835	1.958
KB	23.8	36.4	31.28	3.004	1.733
KT	23.8	35.8	31.34	2.944	1.716
KLIA	24.2	37.2	32.08	2.229	1.493
MC	24.4	38.0	32.11	2.319	1.523
MR	23.6	36.2	31.10	3.358	1.833
MS	23.3	37.3	32.43	4.492	2.119
SN	23.4	37.2	31.87	3.169	1.780

As presented in Table 4, the possible threshold, u was obtained by evaluating the mean residual life plot and parameter threshold stability plots carefully. Figure 1 and Figure 2 illustrate the mean residual life plot and parameter stability plots respectively for Kota Bharu station as an example to explain the threshold selection procedure. From our observation, the plots appear to be consistent and stable approximately at u = 33.75°C, indicating u = 33.75°C as the most suitable threshold for Kota Bharu station.

Overall, the selected threshold value for all stations ranges from 33°C (Mersing) to 35.75°C (Alor Setar). Comparing with the thermal perceptions classification for the tropical region [9], it is found that the obtained temperature threshold falls into slightly warm (30°C -34°C) and warm categories (34°C -38°C).

Table 4: Threshold Choice with Extremal Index Value

Station	u (°C)	θ	Thermal Perception Category
AS	35.75	0.108	Warm
BL	33.50	0.133	Slightly Warm
CP	35.25	0.088	Warm
KB	33.75	0.214	Slightly Warm
KT	33.50	0.108	Slightly Warm
KLIA	34.50	0.233	Warm
MC	33.25	0.216	Slightly Warm
MR	33.00	0.138	Slightly Warm
MS	35.00	0.194	Warm
SN	34.50	0.199	Warm

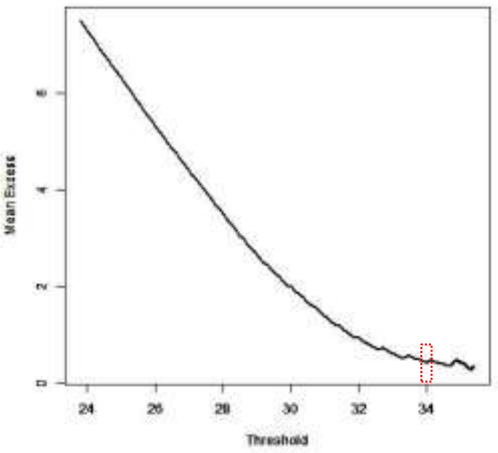


Figure (1) Mean Residual Life Plot for Kota Bharu Station

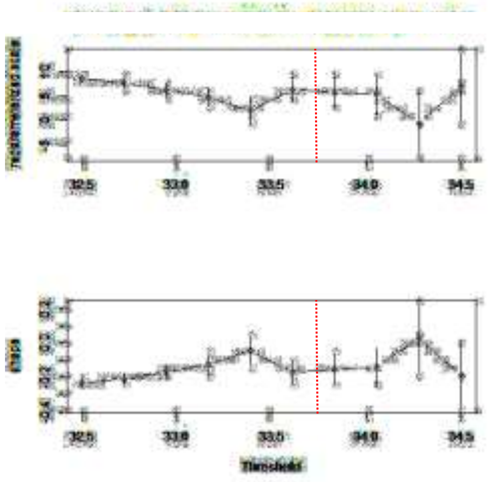
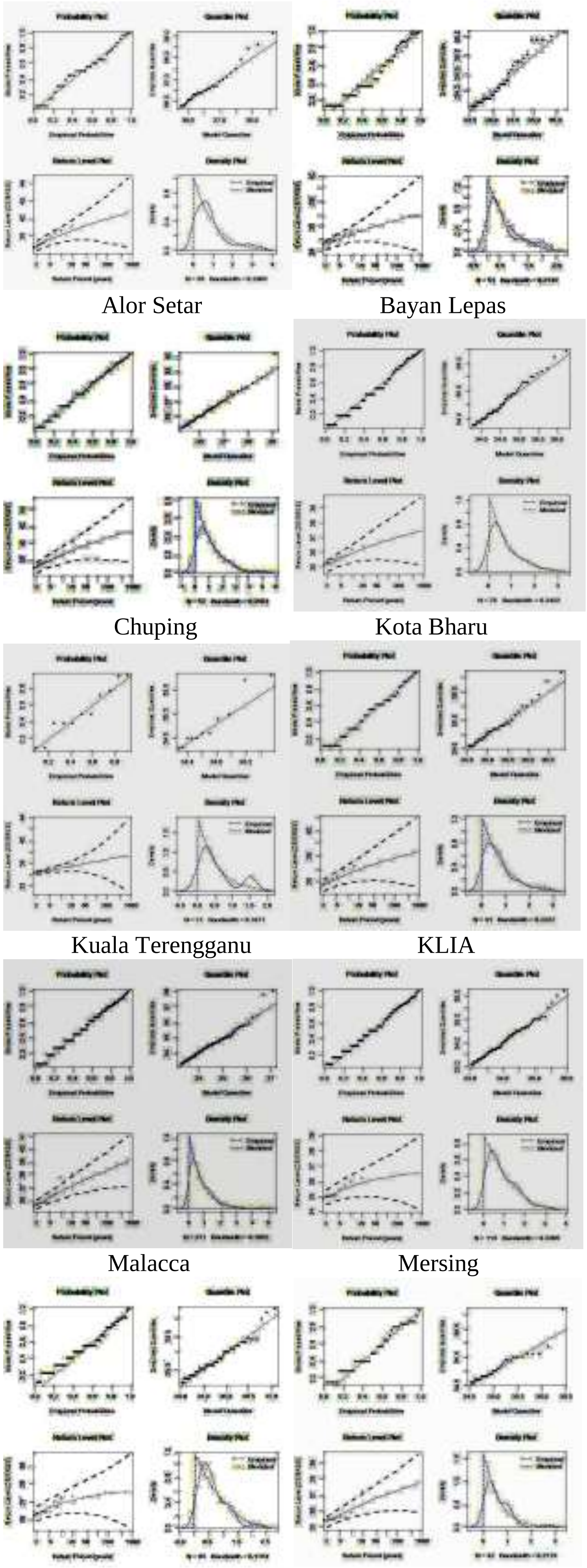


Figure (2) Parameter Stability Plot for Kota Bharu Station



Alor Setar
Bayan Lepas
Chuping
Kota Bharu
Kuala Terengganu
KLIA
Malacca
Mersing
Muadzam Shah
Senai

Figure (3) Diagnostic Plots for All Stations

The declustering procedure was conducted by taking into account the value of the extremal index, θ. As mentioned in the above, this index value is required to provide independent clusters. In this study, the extremal index value estimated for each station is $\leq$ 0.5 which indicate that the dependence increases due to large observations that cluster together. Knowing the threshold and the extremal index value, the parameter estimates together with the corresponding standard errors (se) are presented in Table 5. As pointed above, the estimation was conducted using MLE. All stations have negative shape parameter indicating that the fitted distribution has an upper bound. The standard errors for the scale parameter are in the interval [0.063, 0.240] and in the intervals [0.051, 0.190] for the shape parameter.

Table 5: Parameter Estimation

Station	σ (se)	ξ (se)
AS	0.968 (0.240)	-0.105 (0.183)
BL	0.712 (0.168)	-0.131 (0.190)
CP	1.211 (0.229)	-0.115 (0.129)
KB	0.790 (0.142)	-0.149 (0.136)
KT	0.738 (0.148)	-0.183 (0.141)
KLIA	0.797 (0.177)	-0.137 (0.159)
MC	0.796 (0.063)	-0.054 (0.051)
MR	1.149 (0.129)	-0.285 (0.073)
MS	0.839 (0.111)	-0.295 (0.081)
SN	0.617 (0.119)	-0.100 (0.116)

The adequacy of the models was revealed by evaluating the diagnostics plots demonstrated as in Figure 3. The points on the probability plot are near-linear for all stations while for the quantile plot, the goodness of fit seems unconvincing as some of the points do not lie close to the diagonal line. However, the points on the return level plot lie between the confidence limit which provides a satisfactory representation of validity of the fitted model. As a consequence of the negative shape parameter value for all stations, the return level curve from the return level plot asymptotes to a finite level. Lastly, the modeled density plot of most stations seems consistent with the distribution of excess. Therefore, all four diagnostic plots support the adequacy of the fitted model.

The estimated *N*-year return levels and 95% confidence intervals for *N* = 10, 25, 50, 100 and 125 return periods are presented in Table 6. The expected return period for the maximum temperature event to reappear varies among the stations. The estimation results exhibit that the maximum temperature for Bayan Lepas station is expected to reappear within the next 25 years return period. In contrast, the return level estimates for Alor Setar, Kota Bharu, Kuala Terengganu and KLIA stations show that the maximum temperatures are expected to re-enter their maximum temperature state within the next 50 years. Within the next 100 years, Chuping, Malacca, Mersing and Muadzam Shah stations are predicted to enter their maximum states.

Comparing the above results with our previous analysis using block maxima with GEV distribution approach, it is found that both GEV and GP approaches provide almost similar return level estimates for all stations. The differences between GEV and GP return level estimates are less than

0.4°C. Nevertheless, the expected return period of the maximum temperature to reappear are earlier using GEV compared to GP approach.

Table 6: Return Level Estimates for the Fitted Model

Station	Return Period, N				
	10	**25**	**50**	**100**	**125**
AS	38.15 (37.28, 39.01)	38.77 (37.52, 40.03)	**39.21** **(37.54,** **40.88)**	39.61 (37.45, 41.78)	39.74 (37.40, 42.08)
BL	35.38 (34.76, 36.00)	**35.78** **(34.83,** **36.73)**	36.06 (34.78, 37.33)	36.31 (34.67, 37.95)	36.38 (34.62, 38.15)
CP	38.54 (37.56, 39.53)	39.27 (37.84, 40.75)	39.77 (38.02, 41.52)	**40.23** **(38.04,** **42.42)**	40.37 (38.03, 42.71)
KB	35.93 (35.32, 36.55)	36.33 (35.46, 37.21)	**36.60** **(35.48,** **37.72)**	36.84 (35.45, 38.23)	36.91 (35.44, 38.39)
KT	35.28 (34.71, 35.85)	35.63 (34.87, 36.38)	**35.86** **(34.92,** **36.79)**	36.06 (34.92, 37.20)	36.11 (34.91, 37.33)
KLIA	36.62 (35.89, 37.35)	37.06 (36.04, 38.07)	**37.35** **(36.05,** **38.65)**	37.62 (36.00, 39.24)	37.70 (35.97, 39.43)
MC	36.68 (35.09, 37.28)	37.22 (36.42, 38.04)	37.62 (36.62, 38.63)	**38.00** **(35.79,** **39.22)**	38.12 (36.83, 39.42)
MR	35.69 (34.95, 36.42)	35.97 (35.03, 36.92)	36.14 (35.00, 37.28)	**36.28** **(34.91,** **37.66)**	36.32 (34.86, 37.78)
MS	36.90 (36.29, 37.50)	37.12 (36.34, 37.05)	37.25 (36.31 38.20)	**37.36** **(36.21,** **38.51)**	37.39 (36.17, 38.62)
SN	36.12 (35.60, 36.64)	36.52 (35.83, 37.20)	36.79 (35.94, 37.65)	37.05 (35.60, 38.11)	37.14 (36.00, 38.26)

5. CONCLUSION

In this research, the Peak over Threshold method with Generalized Pareto Distribution approach was applied to analyze the temperatures data for ten meteorological stations located in Peninsular Malaysia. The choice of threshold is crucial in forecasting possible extreme events. Thus, a careful examination of the mean residual life plot and the parameter stability plot was conducted to select the appropriate threshold. Based on the diagnostic plots, the GPD was adequately fitted to the temperature extremes. It can be observed from the return level estimation that the maximum temperatures at all stations increase steadily for higher and higher return period. Within the next 25 to 100 years, it is expected that a maximum temperature event will re-emerge for all stations except Senai. Future studies should incorporate other threshold selection procedures and consider modeling of the non-stationary GPD.

6. REFERENCES

[1] Coles, S. “An Introduction to Statistical Modeling of Extreme Values” Great Britain: Springer (2001).

[2] Rahayu, A. “Identification of Climate Change with Generalized Extreme Value (GEV) Distribution Approach” *Journal of Physics*, **423**: 1-7 (2013).

[3] Martin, M.L., Luna, M.Y., Morata, A. and Fenoy, M. “Statistical Modelling of Extreme Pluviometric Events by Means of Generalized Pareto Distribution” *Congreso Nacional de Estadística e Investigación Operativa*, 718-727 (2003).

[4] Hasan, H., Mohd Salleh, N.H. & Kassim, S. “Stationary and Non-Stationary Extreme Value Modeling of Extreme Temperature in Malaysia” *AIP Conference Proceeding*, **1613**: 355-367 (2014).

[5] Hasan, H., Ahmad Radi, N.F. & Kassim, S. “Modeling Extreme Temperature Using Generalized Extreme Value (GEV) Distribution: A Case Study of Penang” *World Congress on Engineering*, **1**: 181-186 (2012).

[6] Tencer, B. and Rusticucci, M. "Analysis of Interdecadal variability of temperature events in Argentina applying EVT" *Atmosfera* **25**(4): 327-337 (2012)

[7] Zalina, M.D., Desa, M.N.M., Nguyen, V.T.V. and Kassim, A.H.M. “Selecting a probability distribution for extreme rainfall series in Malaysia” *Water Science and Technology*, **45**(2): 63-68 (2002).

[8] Malaysia Meteorological Department “General Climate Information” Retrieved from http://www.met.gov.my/en/web/metmalaysia/education/climate/generalclimateinformation (2017).

[9] Lin, T. and Matzarakis, A. “Tourism climate and thermal comfort in Sun Moon Lake, Taiwan” *Int J Biometeorol*, **52**: 281-290 (2008).

[10] Makaremi, N., Salleh, E., Jaafar, M.Z. and Ghaffarian Hoseini, A. “Thermal comfort conditions of shaded outdoor spaces in hot and humid climate of Malaysia” *Building and Environment*, **48**: 7-14 (2012).

[11] Hasan, H., Che Nordin, M.A. and Mohd Salleh, N.H. “Modeling Daily Maximum Temperature for Thermal Comfort in Northern Malaysia” *Advances in Environmental Biology*, **9**(26): 12-18 (2015).

[12] Zahid, M., Blender, R., Lucarini, V. and Bramati, M.C. “Return Levels of Temperature Extremes in Southern Pakistan” *Earth Syst. Dynam. Discuss, doi:10.5194/esd-2016-72, in review* (2017).

[13] Ferro, C.A.T. and Segers J. “Inference for Clusters Of Extreme Values” *J. Roy. Stat. Soc,* **65B**: 545-556 (2003).

[14] Acero, F. J., Garcia, J.A. and Gallego, M.C. “Peaks-over-Threshold Study of Trends in Extreme Rainfall over the Iberian Peninsula” *American Meteorological Society*, **24**: 1089-1105 (2011).

[15] Cueto, O.R.G., Soto, N.S., Nunez, M.Q., Benitez, O. and Limon, N.V. “Extreme temperature scenarios in Mexicali, Mexico under climate change conditions” *Atmosfera* **26**(4): 509-520 (2013).

[16] Campbell, B. Belenky, V. and Pipiras, V. “On the Application of the Generalized Pareto Distribution for Statistical Extrapolation in the Assessment of Dynamic Stability in Irregular Waves” *The 14th International Ship Stability Workshop (ISSW)*, 149-153 (2014).