Jurnal
Teknologi

Full Paper

PREDICTION OF PM_{10} EXTREME CONCENTRATIONS IN URBAN MONITORING STATIONS IN SELANGOR, MALAYSIA USING THREE PARAMETERS EXTREME VALUE DISTRIBUTIONS (EVD)

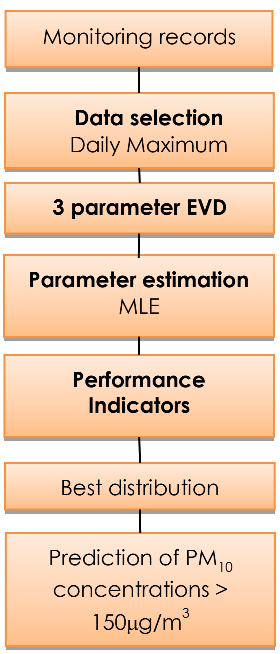
Article history
Received
15 July 2015
Received in revised form
1 October 2015
Accepted
25 October 2015

Hasfazilah Ahmat*, Ahmad Shukri Yahaya, Nor Azam Ramli

Clean Air Research (CARE) Group, School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Pulau Pinang, Malaysia

*Corresponding author
hasfazilah.ahmat@gmail.com

Graphical abstract



Abstract

Objective: The purpose of the study was to determine the best distribution to predict the extreme concentrations of PM_{10} using the three parameters Weibull, Generalized Extreme Value (GEV) and Generalized Pareto Distribution (GPD). *Methodology:* The study used maximum concentration of daily PM_{10} data recorded in the period of 2000-2012 in Klang and Shah Alam, Selangor. The parameters of all distributions were estimated using the method of Maximum Likelihood Estimator (MLE). The goodness of fit of the distribution was determined using performance indicators namely; the accuracy measures and error measures. The best distribution was selected based on the highest accuracy measures and the smallest error measures. *Results:* The findings showed that the three parameters GEV was the best fit for daily maximum concentration for PM_{10} in these two stations. The result also demonstrated that the predicted number of days in which the concentration of PM_{10} exceeded the Malaysia Ambient Air Quality Guideline (MAAQG) for daily concentrations of 150µg/m³ were more than 85% in compliance of the actual number of days. Hence, the GEV can be used for the prediction of the PM_{10} extreme concentrations.

Keywords: Extreme value theory (EVT); Extreme value distribution (EVD); Weibull; Generalized Extreme Value (GEV);Generalized Pareto Distribution (GPD); PM_{10}; air pollution; prediction

Abstrak

Objektif: Tujuan kajian ini adalah untuk menentukan taburan yang terbaik untuk meramalkan kepekatan ekstrem PM_{10} menggunakan tiga parameter Weibull, Nilai Melampau Teritlak (GEV) dan Taburan Pareto Teritlak (GPD). *Metodologi:* Kajian menggunakan kepekatan maksimum data harian PM_{10} yang dicatatkan dalam tahun 2010 - 2012 di Klang dan Shah Alam, Selangor. Parameter untuk semua taburan dianggarkan menggunakan kaedah Penganggar Kebolehjadian Maksimum (MLE). Kebaikan penyesuaian taburan ditentukan dengan menggunakakan penunjuk prestasi iaitu pengukuran kejituan dan pengukuran ralat. *Dapatan:* Hasil kajian menunjukkan bahawa tiga parameter GEV adalah taburan terbaik untuk kepekatan maksimum harian bagi PM_{10} di kedua-dua stesen. Hasil kajian juga menunjukkan bahawa anggaran bilangan hari kepekatan PM_{10} melebihi Garis Panduan Kualiti Udara Ambien Malaysia (MAAQG) bagi kepekatan harian PM_{10} iaitu 150 µg/m³ lebih daripada 85% dalam pematuhan jumlah sebenar hari. Oleh itu, GEV boleh digunakan untuk peramalan kepekatan melampau bagi PM_{10}.

Kata kunci: Teori Nilai Melampau (EVT);Taburan Nilai Melampau (ECD);Weibull;Nilai Melampau Teritlak (GEV);Taburan Pareto Teritlak (GPD);PM_{10};pencemaran udara; peramalan

1.0 INTRODUCTION

The existence of air pollutants in the form of gaseous, liquid or fine particles suspended in air is generally termed as air pollution. It takes place as a result of natural processes (volcanic activity, forest fires and windblown dust) along with anthropogenic activity. Anthropogenic activity can further be categorized into mobile and stationary sources. Motor vehicles either public or private, are examples of mobile sources, whilst stationary sources refers to industrial processes with examples such as power plants, incinerators, cements factories and iron and steel mills. In general, air pollutant in Malaysia is transported in the wind direction pattern of the northeast monsoon and southwest monsoon [1]. Previous studies showed that it can generate damaging effects to human health, crops and environment [2–4].

Reference [5] states that PM_{10}, which is identified as coarse particle and filterable particulate matter with aerodynamic diameter less than 10 micrometer (μm), continues to be the predominant pollutant in several areas in Malaysia that can cause unhealthy conditions during the dry period, as a result of the southwest monsoon. Several unhealthy days in some locations in Malaysia during this period were due to trans-boundary pollution. Particulate Matter (PM) is among the common six criteria pollutants [6].

For a decade, power plants and industrial have been the major sources of PM_{10} in Malaysia. However in 2012, there was a significant increase of more than 50% in motor vehicles as the major contributor of PM_{10} where Malaysia recorded significant increase in number of registered vehicles and active vehicles on road.

A guideline published by [8] as presented in Table 1 pointed out that the increase in mortality of around 2.5% for each 50 μg/m^3 increment in the daily concentration was expected. Thus, it was expected 5% increase in daily mortality as the effect of the increase of PM_{10} concentration of 150μg/m^3. The findings of a study conducted by [9] in Beijing also indicated that daily mortality has significant associations with PM_{10}.

The trend of the annual average levels of PM_{10} concentration in the ambient air between 2000 and 2013 was in compliance with the Malaysian Ambient Air Quality Guidelines (MAAQG) of 50μg/m^3 as shown in Figure 1. Though the averages were lower than the annual permissible value, there were incidences during 2000 – 2013 whereby the concentrations of PM_{10} occasionally exceeded the stipulated hourly MAAQG value of 150μg/m^3.

Table 1 WHO air quality guidelines and interim targets for PM based on 24-hour concentrations (source : [7])

	PM_{10} (ug/m^3)	**Basis for selected level**
Interim target-1 (IT-1)	150	Based on published risk coefficients from multi-centre studies and meta-analyses (about 5% increase of short-term mortality over the AQG value)
Interim target-2 (IT-2)	100	Based on published risk coefficients from multi-centre studies and meta-analyses (about 2.5% increase of short-term mortality over the AQG value)
Interim target-3 (IT-3)*	75	Based on published risk coefficients from multi-centre studies and meta-analyses (about 1.2% increase of short-term mortality over the AQG value)
Air quality guideline (AQG)	**50**	Based on relationship between 24-hour and annual PM_{10} levels.

In the period of 2004 – 2013, Malaysia experienced several short spell of high particulate event as the result of the local peat land and trans-boundary smoke fires from the neighbouring country particularly during the dry period of between May to September [10]. Researches had long attracted to studies involving natural phenomena such as rain, wind speed, air pollution, ocean wave heights and erosion [3, 4] using extreme value theory (EVT). Though hydrological studies utilized the EVT the most, the application in air pollution is irrefutably important.

A review of the literature shows that the Extreme Value Distribution (EVD) does not appear to have been employed for the analysis of any air pollution data until late 1970s. Reference [13] provides a comprehensive review of the principles and underlying assumptions of extreme value statistics in the application of the EVD in air pollution data. EVD is widely used method for assessing and estimating the concentrations of air pollutions [14]–[23]. Reference [24] had applied EVT in their air pollution study.

Measure of central tendency is a single value that attempts to explain a set of data by recognizing the central position within that set of data. Mean concentration of air pollution data may be best fitted using the method of central fitting. However, it does not fit precisely for high concentration or extremes of air pollution data [25]. Extreme PM_{10} concentrations need to be modelled by suitable statistical distributions that give the best inferences of the behaviour of the extremes. Such application in air pollution studies for PM_{10} is still not widely explored particularly in Malaysia. Reference [25–27] had applied EVD in PM_{10} studies in Malaysia.

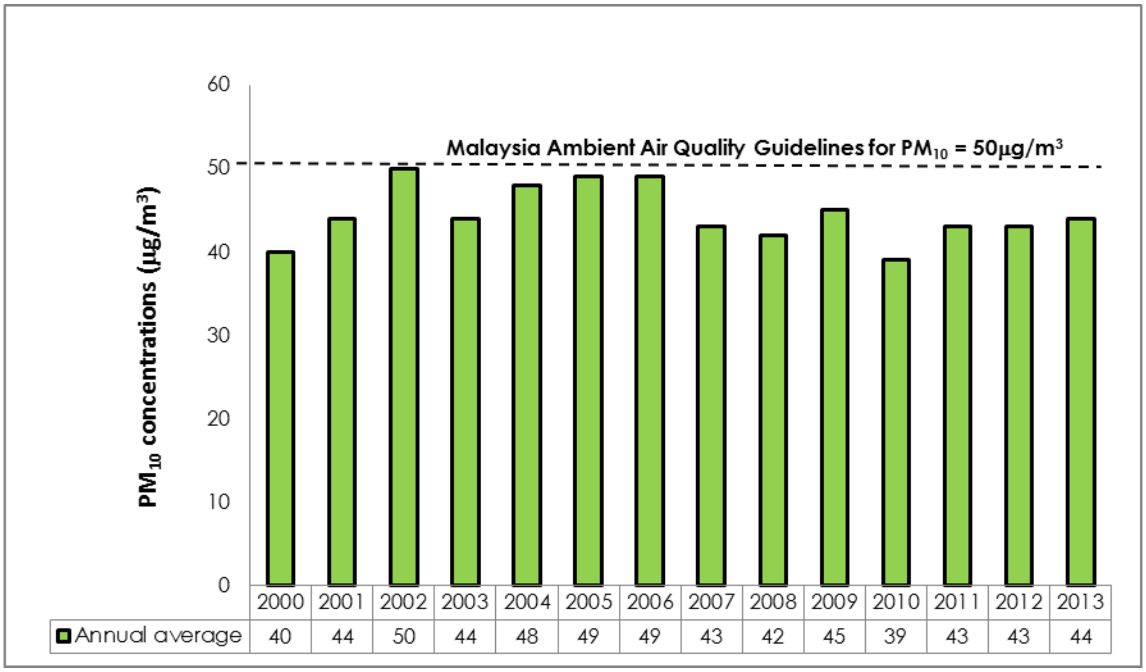


Figure 1 Annual average of PM_{10} concentrations in Malaysia from 2000 – 2013 (Source : [5])

This study was undertaken with the objective of determining the best model to predict the extreme concentrations of PM_{10} using the three parameters Weibull, Generalized Extreme Value (GEV) and Generalized Pareto Distribution (GPD).

2.0 METHODOLOGY

2.1 Study Area and Monitoring Records

Figure 2 describes the flow of research methodology. The daily maximum concentration records of PM_{10} for the period of 2000 – 2012 were furnished by the Department of Environment, Malaysia. The records were collected through a continuous monitoring by Alam Sekitar Sdn. Bhd. (ASMA) using Beta Attenuation Method [5] from the air monitoring station in Klang and Shah Alam in the west coast of Peninsular Malaysia as shown in Figure 3.

The Klang Valley area experiences continuous exposures to the problem of air quality due to its geographical location and as one of the most urbanized areas following rapid population growth, industrial and commercial activities [28]. The unhealthy days in this area were caused by PM_{10} mainly from trans-boundary pollution during the Southwest Monsoon [5]. Reference [29] stated that a high density of vehicles contributes to high concentrations of PM_{10} in Klang and Shah Alam.

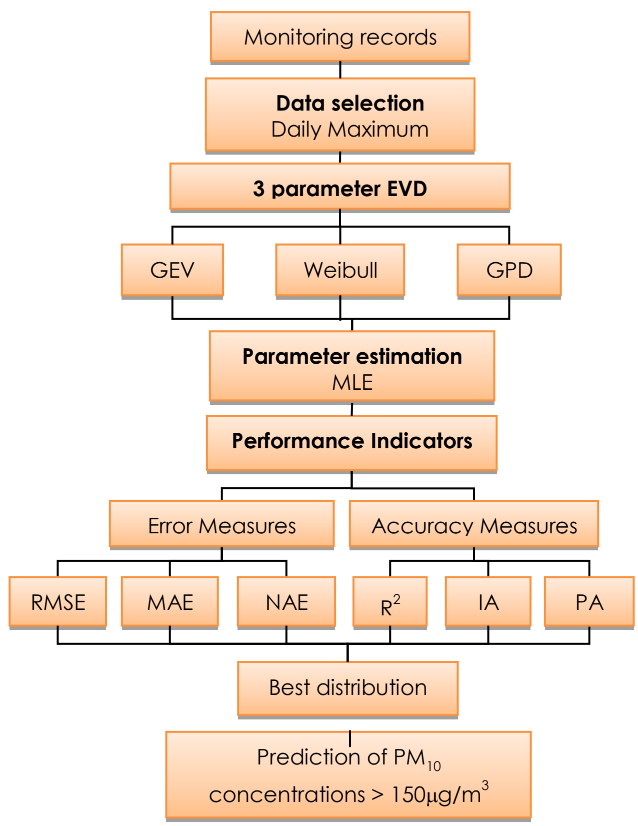


Figure 2 Flow of research methodology

2.2 Extreme Value Distribution (EVD)

This research analyzed the PM_{10} records using the three parameters extreme value distributions, namely: Generalized Extreme Value (GEV), Weibull and Generalized Pareto Distribution (GPD). The EVD of type III was formulated by Waloddi Weibull (1887-1979), a Swedish engineer and scientist well-known for his work on strength of materials and fatigue analysis. It was not until 1951 that the distribution was known in

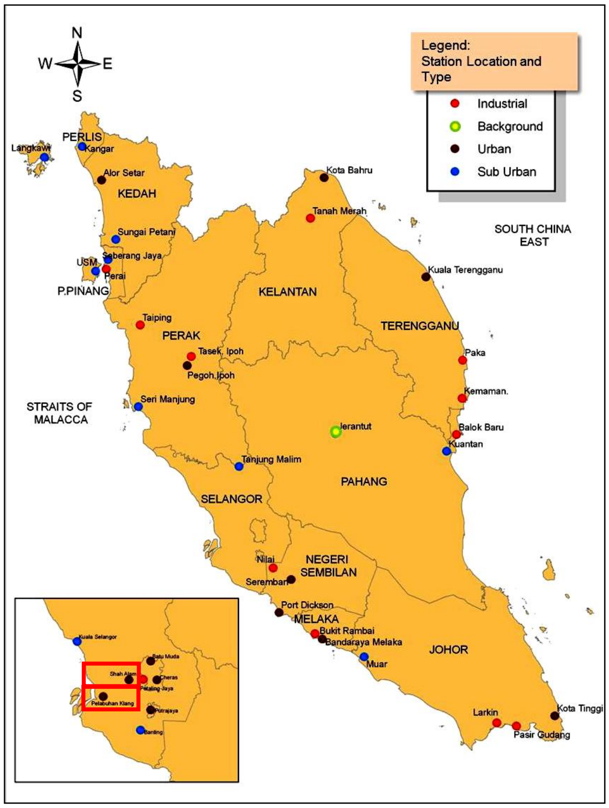


Figure 3 Location of continuous monitoring stations in Peninsular Malaysia (source : [5])

serious way for modeling the statistical variations of the data which then named as Weibull distribution after his name [33].

A single Generalized Extreme Value (GEV) distribution is as a result of a combination of a family of continuous probability distribution which was proposed for statistical stability. The GEV which was popularised by Jenkinson in 1955 [34] uses three parameters, namely: location, scale and shape. Location parameter, μ (mu) determining the shifting of a distribution in a specified direction on the horizontal axis. The dispersion of the distribution is measured by the scale parameter, σ (sigma) and it indicates where the concentration of the distribution lies. The distribution is said to expanse or more spread out if the scale parameter increases. Shape parameter, λ (lambda) on the other hand, affects the shape of distributions and tails of the distribution. Skewness governs the shape parameter as to where the majority of data concentrated, thus, creates the tail(s) of distribution [35].

The Generalized Pareto Distribution (GPD), which is a special case of both exponential and Wakeby distribution was introduced by Pickands in 1975 as a distribution of the sample of excesses (or exceedances) above a sufficiently high threshold [36]. It has been used widely by many scientists particularly in meteorological studies.

The EVT is gradually applied in the study of PM_{10} in

Table 2 The Probability Density Function (PDF) and Cumulative Distribution Function (CDF) of the distributions

Distribution	PDF and CDF	Source
GEV	$PDF: \ f(x;\lambda,\sigma,\mu)=\frac{1}{\sigma}\left[1+\lambda\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\lambda-1}\exp\left\{-\left[1+\lambda\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\lambda}\right\}$ $CDF: \ F(x;\lambda,\sigma,\mu)=\begin{cases}\exp\left\{-\left[1+\lambda\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\lambda}\right\} & \lambda\neq 0 \\ & \text{and for } 1+\lambda\left(\frac{x-\mu}{\sigma}\right)>0 \\ \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} & \lambda=0\end{cases}$	[30]
Weibull	$PDF: f(x;\lambda,\sigma,\mu)=\frac{\lambda}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{\lambda-1}\exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{\lambda}\right]$ for $x\geq\mu; \quad \sigma,\lambda>0$ $CDF: \ F(x;\mu,\sigma,\lambda)=1-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)^{\lambda}\right]$ for $x\geq 0, \lambda>0, \sigma>0$	[31]
GPD	$PDF: \ f(x;\lambda,\sigma,\mu)=\frac{1}{\sigma}\left[1-\lambda\left(\frac{x-\mu}{\sigma}\right)\right]^{1/\lambda-1}, \mu\neq 0$ $CDF: \ F(x;\lambda,\sigma,\mu)=\left\{1-\left[1-\lambda\left(\frac{x-\mu}{\sigma}\right)\right]^{1/\lambda}\right.$, $\lambda\neq 0$ for $x\geq\mu$ when $\lambda\geq 0$ and $x<\mu-\sigma/\lambda$ and when $\lambda<0$	[32]

Notation : μ - location parameter ; σ - scale parameter ; λ - shape parameter ; x - observed values the recent years. The majority of the PM_{10} studies utilised the Weibull distribution though the Gumbel distribution is the most prevalent distribution in the studies of other air pollutants. The GPD is not commonly applied in the air pollution studies. The characteristic of the GPD is to investigate exceedances over high thresholds rather than average over certain period of time [16]. A huge gap was observed in the application of the GPD in air pollution studies may be due to the selection of hourly/daily/monthly/annual data rather than concentrations above certain threshold. Both of the GEV and GPD were not commonly used due to their complexities and many parameters have to be considered and the interpretation of the results might be difficult compared to the simple model [37].

The application of GEV and GPD were explored in this study since both were mainly used in other pollutants except PM_{10}. All the PDFs and CDFs of the distributions are presented in Table 2.

2.3 Parameter Estimates

There are several methods to estimate parameters for each EVD, however there is no consensus about which is the most appropriate. The appropriateness of the methods shall be determined by the performance indicators or error measures. All the parameters of the distributions in this study were estimated using the method of Maximum Likelihood Estimators (MLE). The formulae are different for every distribution which are listed in Table 3.

2.4 Performance Indicators

Six performance indicators were used to select the best distribution to represent the data. The accuracy measures were the prediction accuracy (PA), coefficient of determination (R^2) and Index of Accuracy (IA). The accuracy value is between 0 and 1 and as the value approaches 1, the model is appropriate. On the other hand, as the value of error

Table 3 Formulae for parameter estimates

Distribution	**Formulae**	Source
GEV	$\frac{1}{\sigma}\sum_{i=1}^{n}\left(\frac{1-\lambda-(1-(\lambda/\sigma)(x_i-\mu))^{1/\lambda}}{(1-(\lambda/\sigma)(x_i-\mu))}\right)=0$ $-\frac{n}{\sigma}+\frac{1}{\sigma}\sum_{i=1}^{n}\left[\frac{1-\lambda-(1-(\lambda/\sigma)(x_i-\mu))^{1/\lambda}}{(1-(\lambda/\sigma)(x_i-\mu))}\left(\frac{(x_i-\mu)}{\sigma}\right)\right]=0$ $-\frac{1}{\lambda^2}\sum_{i=1}^{n}\left[\ln(1-(\lambda/\sigma)(x_i-\mu))\{1-\lambda-[1-(\lambda/\sigma)(x_i-\mu)]^{1/\lambda}\}\right.$ $\left.+\frac{1-\lambda-[1-(\lambda/\sigma)(x_i-\mu)]^{1/\lambda}}{(1-(\lambda/\sigma)(x_i-\mu))}\lambda\left(\frac{x_i-\mu}{\sigma}\right)\right]=0$	**[40]**
Weibull	$\sigma=\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^{\lambda}\right]^{1/\lambda}$ $\quad \frac{1}{\lambda}-\frac{\sum_{i=1}^{n}(x_i-\mu)^{\lambda}\ln(x_i-\mu)}{\sum_{i=1}^{n}(x_i-\mu)^{\lambda}}+\frac{1}{n}\sum_{i=1}^{n}\ln(x_i-\mu)=0$ $\frac{\lambda-1}{\lambda}\sum_{i=1}^{n}(x_i-\mu)^{-1}-n\frac{\sum_{i=1}^{n}(x_i-\mu)^{\lambda-1}}{\sum_{i=1}^{n}(x_i-\mu)^{\lambda}}=0$	**[41]**
GPD	$\mu=x_1$ $\quad \sigma=\frac{\lambda}{\exp(n\lambda)-1}(x_n-\mu)$ $\sum_{i=1}^{n}\left(\exp(n\lambda)+\frac{x_n-x_i}{x_i-\mu}\right)^{-1}=\frac{n}{\exp(n\lambda)-1}-\frac{1}{\lambda\exp(n\lambda)}$	**[42]**

Notation : μ - location parameter ; σ - scale parameter ; λ - shape parameter ; n – number of observations ; x_i - observed values measures approaching 0, the model is deemed to be the best model. The error measures used in this study were the root mean squared error (RMSE), the normalized absolute error (NAE) and the mean absolute error (MAE) [22, 23]. The accuracy measures have the advantages that they are dimensionless and bounded between 0 and 1, that is independent of the unit of data while the error measures are scale and unit-dependent [39].

The formulae for all performance indicators are listed in Table 4.

2.5 Software

A software called MATLAB® ver. 11 [43], a high-level language and interactive environment for numerical computation, visualization, and programming package for engineers was used to estimate the parameters, calculate the performance indicators and plot the cumulative distribution functions (CDF) of the distributions. The IBM SPSS ver. 18, a statistical package for the social sciences was used to plot time series records of the two stations.

3.0 RESULTS AND DISCUSSION

3.1 Statistical Characteristics of PM_{10}

Table 5 depicts the descriptive statistics of the PM_{10} concentrations records for the period of 2000 - 2012. The average concentrations of PM_{10} recorded were 74.9µg/m^3 and 61.3µg/m^3 for Klang and Shah Alam respectively. The average values were slightly above the MAAQG for the annual average of 50 µg/m^3 [5]. The PM_{10} concentrations were skewed to the right (skewness > 0) which indicated the existence of the extreme concentrations of PM_{10} exist during the period under study.

The maximum concentrations recorded in Klang was higher (643 µg/m^3) than that of Shah Alam (587 µg/m^3). Both of the maxima were recorded in 2005 as represented by time series plot in Figure 4. The extreme concentrations in the year 2005 in Klang and Shah Alam normally occurred in the third quarter of the year. The following month recorded low concentrations as a result of the wind had blown all the particulates further northwards.

Box-and-Whisker plot in Figure 5 shows that extreme concentrations existed every year in 2000 – 2012.

3.2 Parameter Estimates and Performance Indicators

Table 7 lists the values for the location parameter, μ, scale parameter, σ and shape parameter, λ of the distributions.

Based on performance indicators in Table 6, the distributions were then ranked to represent the best distribution to predict the PM_{10} exceedances with concentrations more than 150µg/m^3. The best distribution was the three parameters GEV as it has the highest accuracy measures and the smallest error measure.

Table 4 Performance Indicators

Indicators	Equations
Prediction Accuracy (PA)	$PA = \sum_{t=1}^{n} \frac{(P_t - \overline{P})(O_t - \overline{O})}{(n-1)\sigma_P \sigma_O}$
Coefficient of Determination (R^2)	$1 - \frac{\sum_{t=1}^{n}(O_t - P_t)^2}{\sum_{t=1}^{n}(O_t - \overline{O})^2}$
Index of Accuracy (IA)	$1 - \frac{\sum_{t=1}^{n}(P_t - O_t)^2}{\sum_{t=1}^{n}\left(\lvert P_t - \overline{O}\rvert - \lvert O_t - \overline{O}\rvert\right)^2}$
Root Mean Square Error (RMSE)	$\sqrt{\frac{\sum_{t=1}^{n}(O_t - P_t)^2}{n}}$
Normalized Absolute Error (NAE)	$\frac{\sum_{t=1}^{n}(P_t - O_t)}{\sum_{t=1}^{n} O_t}$
Mean Absolute Error (MAE)	$\frac{\sum_{t=1}^{n}\lvert(O_t - P_t)\rvert}{n}$

Notation :
n = number of observations,
O_t = Observed values,

$\overline{O}$ = Mean of observed values = $\frac{1}{n}\sum_{i=1}^{n} O_t$,

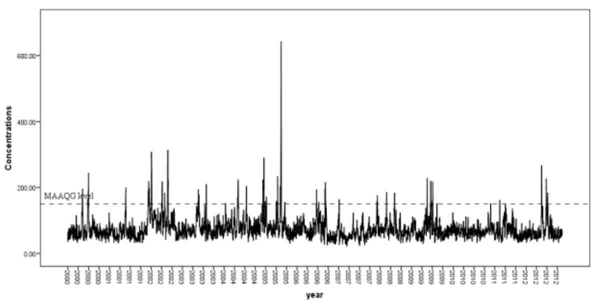
$\overline{O}$ = Mean of observed values = $\frac{1}{n}\sum_{i=1}^{n} O_t$,

P_t = Predicted values,

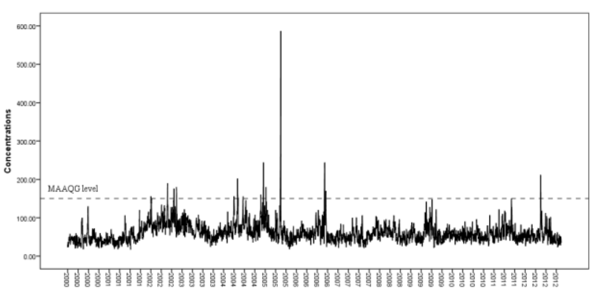
$\overline{P}$ = Mean of predicted values = $\frac{1}{n}\sum_{i=1}^{n} P_t$

Table 5 Descriptive statistics of PM_{10} ($\mu g/m^3$)

	Klang	Shah Alam
N valid	4741	4732
missing	8	17
Mean	74.88	61.33
Median	68.00	56.00
Std. Deviation	33.98	27.58
Variance	1154.69	760.37
Skewness	4.09	5.22
Kurtosis	40.32	76.24
Minimum	21.00	17.00
Maximum	643.00	587.00
Coefficient of variation	0.45	0.45



(a)



(b)

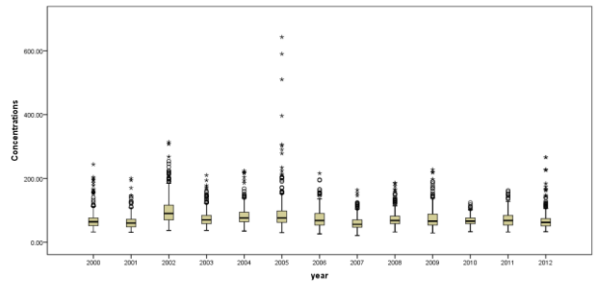
Figure 4 Time series plot for (a) Klang and (b) Shah Alam

3.3 PDF and CDF of the distribution

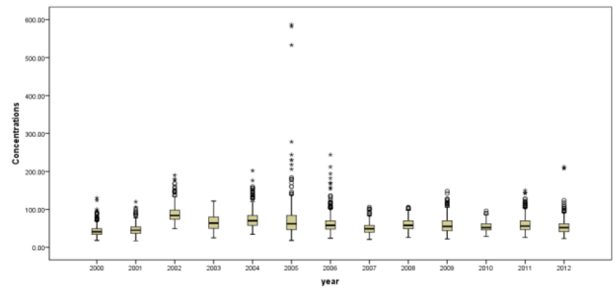
Figure 6 depicts the Probability Density Function (PDF) of the concentrations in Klang and Shah Alam. The GEV MLE fits the observations well in both monitoring stations with long tails to the right – an indication of existence of extreme concentrations in both of the monitoring stations.

Cumulative distribution functions (CDF) of the three-parameter GEV for Klang and Shah Alam are presented in Figure 7. The probabilities of the concentrations exceeding the levels of MAAQG of 150 $\mu g/m^3$ were predicted in Klang and Shah Alam.

Table 6 lists the probability, actual number and predicted number of days of concentrations exceeding 150 $\mu g/m^3$. From the table, the predicted numbers of days were 132 and 43 for Klang and Shah Alam respectively which were more than 85% in compliance with the actual number of days. Approximately, extreme concentrations occur 10 days and 3 days on average in Klang and Shah Alam yearly.

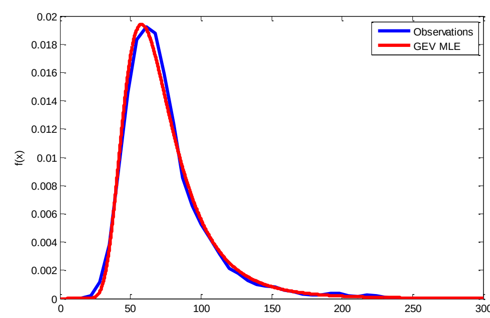


(a)

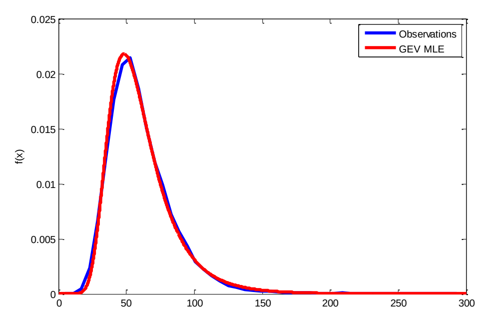


(b)

Figure 5 Box-and-Whisker Plot for (a) Klang and (b) Shah Alam

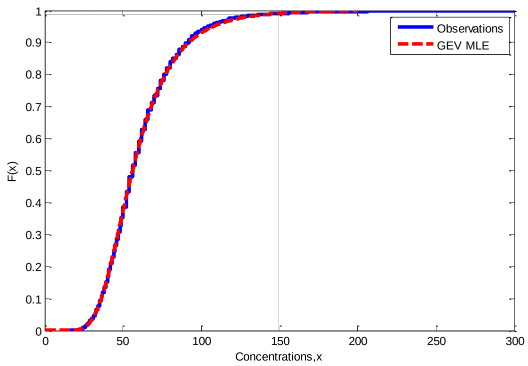


(a)

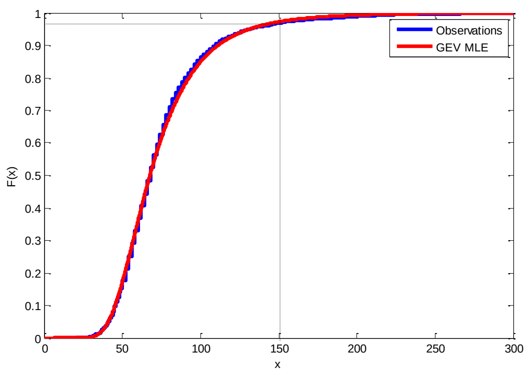


(b)

Figure 6 Probability density function (PDF) for (a) Klang and (b) Shah Alam



(a)



(b)

Figure 7 Cumulative Distribution Function (CDF) for (a) Klang and (b) Shah Alam

Table 6 Probability, actual number and predicted number of days which concentrations exceeding 150 μg/m^3

	Prob.	Actual number of days	Predicted number of days	% compliance
Klang	0.0277	155	132	85%
Shah Alam	0.0090	44	43	98%

4.0 CONCLUSION

In the study of air pollutions, the researchers focused on high concentrations of pollutants as it is detrimental to human health.

This is paper discussed the probability and the number of days of the extreme concentrations which exceeded the permissible value of PM_{10} concentrations of 150μg/m^3 in two monitoring stations in the west coast of Malaysia. Three parameters extreme value distributions, namely: Weibull, Generalized Extreme Value (GEV) and Generalized Pareto Distribution (GPD) were used to analyze the daily maximum of PM_{10} concentrations with the parameter estimator of MLE.

From the findings, the average readings of the PM_{10} concentrations in the monitoring stations were well above the stipulated MAAQG for the yearly average of 50 μg/m^3 with the maximum readings recorded in Klang. The highest concentration recorded in 2005 was due to trans-boundary smoke from forest fires in Sumatera which was transported by South-westerly winds. The central region of Peninsular Malaysia was the most affected by the unfavourable weather conditions of hot and dry periods as the effect of South-westerly winds.

The GEV gave the best estimator for both of the monitoring stations with the smallest errors (NAE, RMSE and MAE) and the highest accuracy measures (PA, R^2 and IA) when compared to the other two distributions. The method gave the accuracy of more than 93% in PA, IA and R^2 for both stations and the smallest errors.

From the plots, the probabilities of the concentrations exceeded the levels of MAAQG of 150 μg/m^3 were estimated and the predicted numbers of day were calculated. The estimated number of days for Klang and Shah Alam were more than 85% in compliance with the actual number of days.

Since most of the literatures employed the other Extreme Value Distributions (EVD) such as the Gumbel, Frechet and Weibull for air pollution studies, this study explored the possibilities of GEV and GPD distributions to the concentrations of PM_{10}.

To conclude, the GEV MLE had an advantage over the Weibull and GPD, thus, it can be used to estimate future exceedances of PM_{10} extreme concentrations. As a result, it may help the policy makers in the respective field to plan suitable measures to curb the occurrence of PM_{10} extreme concentrations and eventually may reduce the effects on human health and environment.

Acknowledgement

The authors would like to thank the Ministry of Education, Malaysia for the scholarship granted, Universiti Teknologi MARA for the study leave, Department of Environment, Malaysia for providing the air quality data in this study and USM for the RUI grant: 814165.

Table 7 Parameters estimation and performance indicators

Station	Dist.	Parameters		Performance Indicators						Best Distribution
				NAE	PA	R^2	RMSE	IA	MAE	
Klang	3Weibull	μ	20.9990	0.2538	0.9735	0.9473	20.6289	0.9165	19.0028	GEV
		σ	60.7897							
		λ	1.7447							
	GEV	μ	60.4816	**0.0187**	**0.9907**	**0.9812**	**5.0077**	**0.9942**	**1.3990**	
		σ	19.1367							
		λ	0.1450							
	3GPD	μ	21.0000	0.3711	0.9483	0.8988	39.0770	0.8459	27.7855	
		σ	81.0916							
		λ	-0.1221							
Shah Alam	3Weibull	μ	16.0034	0.2732	0.9521	0.9061	18.8394	0.8963	16.7535	GEV
		σ	51.0523							
		λ	1.7820							
	GEV	μ	49.7773	**0.0154**	**0.9683**	**0.9373**	**7.0179**	**0.9821**	**0.9468**	
		σ	16.9558							
		λ	0.0938							
	3GPD	μ	17.0000	0.3694	0.9334	0.8709	33.79	0.8304	22.6544	
		σ	65.7384							
		λ	-0.1049							

References

[1] Dominick, D., Latif, M. T., Juahir, H., Aris, A. Z. and Zain, S. M. 2012. An assessment of influence of meteorological factors on PM_{10} and NO_2 at selected stations in Malaysia. *Sustain. Environ. Res.* 22(5): 305–315.

[2] Jamal, H. H., Pillay, M. S., Zailina, H., Shamsul, B. S., Sinha, K., Zaman Huri, Z., Khew, S. L., Mazrura, S., Ambu, S., Rasimah, A. and Ruzita, M. S. 2004. *A Study of Health Impact and Risk Assessment of Urban Air Pollution in the Klang Valley*. Malaysia, Kuala Lumpur.

[3] Kumar, A. and Goyal, P. 2011. Forecasting of daily air quality index in Delhi. *Sci. Total Environ.* 409(24) : 5517–5523.

[4] Sattar, M. R., Mat, R. and Puji, L. 2012. A Preliminary Survey of Air Quality in Makassar City South Sulawesi Indonesia. *J. Teknol. (Sciences Eng.*) 57(2012) : 123–136.

[5] Department of Environment Malaysia. 2014. *Malaysia Environmental Quality Report 2013*. Kuala Lumpur.

[6] Andracsek, R., Burns, P. E. and Mcdonnell. 2012. Flavors of Particulate Matter. *Power Eng.* 116 : 10.

[7] World Health Organization. 2005. *Malaysia Environmental Health Country Profile*.

[8] World Health Organization. 2005. *WHO Air Quality Guidelines For Particulate Matter, Ozone, Nitrogen Dioxide And Sulfur Dioxide*.

[9] Li, P., Xin, J., Wang, Y., Wang, S., Shang, K., Liu, Z., Li, G., Pan, X., Wei, L. and Wang, M. 2013. Time-series analysis of mortality effects from airborne particulate matter size fractions in Beijing. *Atmos. Environ.* 81 : 253–262.

[10] Department of Environment Malaysia. 2006. *Malaysia Environmental Quality Report 2005*. Kuala Lumpur.

[11] Kotz, S. and Nadarajah, S. 2000. *Extreme-Value Distributions : Theory and Applications*. London: Imperial College Press.

[12] Surman, P. G., Bodero, J. and Simpson, R. W. 1987. The prediction of the numbers of violations of standards and the frequency of air pollution episodes using extreme value theory. *Atmos. Environ.* 21(8): 1843–1848.

[13] Roberts, E. M. 1979. Review of Statistics of Extreme Values with Applications to Air Quality Data Part II : Applications. *J. Air Pollut. Control Assoc.* 29(7): 733–740.

[14] Roberts, E. M. 1979. Review of Statistics of Extreme Values with Applications to Air Quality Data Part I : Review. *J. Air Pollut. Control Assoc.* 29(6): 632–637.

[15] Horowitz, J. and Barakat, S. 1979. Statistical analysis of the maximum concentration of an air pollutant: Effects of autocorrelation and non-stationarity. *Atmos. Environ.* 13(6): 811–818.

[16] Smith, R. L. 1989. Extreme Value Analysis of Environmental Time Series : An Application to Trend Detection in Ground-Level Ozone. *Stat. Sci.* 4(4): 367–393.

[17] Dasgupta, R. and Bhaumik, D. K. 1995. UPPER AND LOWER TOLERANCE LIMITS OF ATMOSPHERIC OZONE LEVEL AND Extreme Value Distribution. *Sankhya Indian J. Stat.* 57(B(2)) : 182–199.

[18] Kuchenhoff, H. and Thamerus, M. 1995. Extreme value analysis of Munich air pollution data. *Sonderforschungsbereich.* 386(4): 1–24.

[19] Lu, H. 2002. The statistical characters of PM_{10} concentration in Taiwan area. *Atmos. Environ.* 36: 491–502.

[20] Lu, H.-C. and Fang, G.-C. 2003. Predicting the exceedances of a critical PM10 concentration—a case study in Taiwan. *Atmos. Environ.* 37: 3491–3499.

[21] Hurairah, A., Ibrahim, N. A., Bin Daud, I. and Haron, K. 2005. An application of a new extreme value distribution to air pollution data. *Manag. Environ. Qual. An Int. J.* 16(1): 17–25.

[22] Reyes, H. J., Vaquera, H. and Villasenor, J. A. 2010. Estimation of trends in high urban ozone levels using the quantiles of (GEV). *Environmetrics.* 21 : 470–481.

[23] Quintela-del-Río, A. and Francisco-Fernández, M. 2011. Analysis of high level ozone concentrations using nonparametric methods. *Sci. Total Environ.* 409: 1123–1133.

[24] Sharma, P., Chandra, A., Kaushik, S. C., Sharma, P. and Jain, S. 2012. Predicting violations of national ambient air quality standards using extreme value theory for Delhi city. *Atmos. Pollut. Res.* 3: 170–179.

[25] Md Yusof, N. F. F., Ramli, N. A., Yahaya, A. S., Sansuddin, N., Ghazali, N. A. and Al Madhoun, W. 2011. Central Fitting Distributions and Extreme Value Distributions for Predicion of High PM_{10} Concentration. *IEEE*. 6263–6266.

[26] Ahmat, H., Yahaya, A. S. and Ramli, N. A. 2014. Prediction of PM_{10} Extreme Concentrations using Three Parameters Extreme Value Distributions (EVD). *Sixth International Conference on Postgraduate Education (ICPE-6)*. Melaka. 17-18 December 2014.

[27] Ahmat, H., Yahaya, A. S., Ramli, N. A., Ul-saufie, A. Z. and Abdul Hamid, H. 2015. Analysis of PM_{10} using Extreme Value Theory. *ESTEEM Acad. J.* 11(Special Issue. 1): 135–143.

[28] Yap, X. Q. and Hashim, M. 2013. A robust calibration approach for PM_{10} prediction from MODIS aerosol optical depth. *Atmos. Chem. Phys.* 13: 3517–3526.

[29] Azmi, S. Z., Latif, M. T., Ismail, A. S., Juneng, L. and Jemain, A. A. 2010. Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. *Air Qual. Atmos. Heal.* 3: 53–64.

[30] Aryal, G. R. and Tsokos, C. P. 2009. On the transmuted extreme value distribution with application. *Nonlinear Anal. Theory, Methods Appl.* 71(12) : e1401–e1407.

[31] [31] Georgopoulos, P. and Seinfeld, J. 1982. Statistical distributions of air pollutant concentrations. *Environ. Sci.* 16(7): 401A–416A.

[32] Singh, V. P. and Guo, H. 1995. Parameter estimation for 3-parameter generalized pareto distribution by the principle of maximum entropy (POME). *Hydrol. Sci. J.* 40(2) : 165–181.

[33] Bury, K. 1999. *Statistical Distributions In Engineering*. Cambridge University Press.

[34] Bali, T. G. 2003. The generalized extreme value distribution. *Econ. Lett.* 79 : 423–427.

[35] Millington, N., Das, S. and Simonovic, S. P. 2011. *The Comparison of GEV, Log-Pearson Type 3 and Gumbel Distributions in the Upper Thames River Watershed under Global Climate Models.* Ontario, Canada.

[36] Bermudez, P. de Z. and Kotz, S. 2010. Parameter estimation of the generalized Pareto distribution—Part II. *J. Stat. Plan. Inference*. 140: 1374–1388.

[37] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M. 2004. Methods for imputation of missing values in air quality data sets. *Atmos. Environ.* 38 : 2895–2907.

[38] Yahaya, A. S. and Ramli, N. A. 2008. *Modelling of Carbon Monoxide Concentration in Major Towns in Malaysia : A Case Study in Penang, Kuching and Kuala Lumpur*. Nibong Tebal.

[39] Ji, L. and Gallo, K. 2006. An Agreement Coefficient for Image Comparison. *Photogramm. Eng. Remote Sens.* 72(7) : 823–833.

[40] Martins, E. and Stedinger, J. 2000. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resour. Res.* 36(3) : 737–744.

[41] Rinne, H. 2008. *The Weibull Distribution: A Handbook*. Florida: CRC Press.

[42] Abd-el-hakim, N. S. and Sultan, K. S. 2004. Maximum likelihood estimation from record-breaking data for the generalized Pareto distribution. *Int. J. Stat.* LXII(3) : 377–389.

[43] Mathworks. 2015. *MATLAB Programming Fundamentals*. 8.5 ed. MA: The MathWorks, Inc.