ESTEEM Academic Journal
Vol. 12, No. 1, June 2016, 1-10

PREDICTION OF PM_{10} EXTREME CONCENTRATIONS IN SELECTED INDUSTRIAL MONITORING STATIONS IN MALAYSIA USING EXTREME VALUE DISTRIBUTION (EVD): CLASSICAL AND BAYESIAN APPROACHES

Hasfazilah Ahmat[1], Ahmad Shukri Yahaya[2] and Nor Azam Ramli[3]

[1]Department of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Pulau Pinang, 13500 Permatang Pauh, Pulau Pinang, MALAYSIA.
[2,3]Clean Air Research (CARE) Group, School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 Pulau Pinang, MALAYSIA
[1]hasfazilah.ahmat@gmail.com; [2]ceshukri@usm.edu.my; [3]ceazam@usm.edu.my

ABSTRACT

Where air pollution control is concerned, the rare (extreme) event is typically more significant than the common event. Conventionally, a theory developed to address questions relating to the distribution of extremes is the Extreme Value Theory (EVT). The Bayesian approach offers a more comprehensible framework in incorporating all of the uncertainties involved in the prediction process using the conventional methods. To evaluate the performances of the classical and Bayesian approaches using non-informative priors in estimating the parameters of the Generalized Extreme Value (GEV) and to attain the best model to predict PM_{10} concentrations level. The daily maximum monitoring records of PM_{10} from January 2000 to December 2012 furnished by the Department of Environment, Malaysia were used in this study. 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. This study revealed that the Bayesian GEV with non-informative prior gave the best estimators for PM_{10} concentrations in three industrial monitoring stations and it could be applied in the PM_{10} analysis to predict the exceedances of future extreme concentrations of PM_{10} in these stations.

Keywords: air pollution; Extreme Value Distribution; Bayesian; PM_{10}; Generalized Extreme Value; Prediction.

1. INTRODUCTION

The attempt to curb air pollution within the country itself is not sufficient since air pollution is a trans-boundary issue. The trans-boundary pollution resulting from land and forest fires has been a regular and recurring phenomenon among the ASEAN countries since the late 1980s. The worst incidence was in the 1997/1998 high particulate event which most affected Malaysia along with its neighbouring countries such as Brunei Darussalam, Singapore, parts

p-ISSN 1675-7939; e-ISSN 2289-4934

of Southern Thailand and Indonesia. Billions of dollars were lost from the negative impacts on the tourism and transportation industry, the productive and aesthetic values of the environment as well as the health of the people (Sulaiman et al., 2003). The incidences of high particulate events are generally associated with the presence of PM_{10} which bring together negative effects to human health, environment and economy (Bowman & Johnston, 2005; Mott et al., 2005; Ostermann & Brauer, 2001; Sastry, 2002; Vedal & Dutton, 2006; Yadav et al., 2003). High concentrations of PM_{10} is generally associated with poor visibility and air quality conditions which has been established by Ilyas et al. (2010), McKenzie et al. (2006), Yadav et al. (2003) and Tsai (2005). The economic loss of health impacts due to exposure to PM_{10} generally involve huge amount of cost (Hedley, 2009; Kim et al., 2007; Othman et al., 2014; Sun et al., 2013).

Conventionally, a theory developed to address questions relating to the distribution of extremes is the Extreme Value Theory (EVT) (Finkenstadt & Rootzen, 2001). It develops techniques and models for describing the unusual (extremes) rather than the usual phenomenon (Kotz & Nadarajah, 2000). Where air pollution control is concerned, the rare event is typically more significant than the common event. On the other hand, Coles et al. (2003) indicated that the Bayesian approach offers a more comprehensible framework in incorporating all of the uncertainties involved in the prediction process using the conventional methods. In general, the uncertainties of the parameter is as a result of lack of data and that the selection of the probability distribution does not describe the data perfectly (Chung & Kim, 2013).

In this regard, this paper discusses two objectives in mind, firstly, to evaluate the performances of the classical and Bayesian approaches using non-informative priors in estimating the parameters of the Generalized Extreme Value (GEV) and secondly, to attain the best model to predict PM_{10} concentrations level in industrial monitoring stations in Bukit Rambai, Melaka, Nilai, Negeri Sembilan and Pasir Gudang, Johor which are located in Peninsular Malaysia.

2. MATERIALS AND METHODS

2.1 Study Area and Monitoring Records

The daily maximum monitoring records of PM_{10} from January 2000 to December 2012 furnished by the Department of Environment, Malaysia were used in this study. Data collections were done through a continuous monitoring by Alam Sekitar Sdn. Bhd. (ASMA) from three monitoring stations in the west coast of Peninsular Malaysia using Beta Attenuation Method (BAM). Bukit Rambai, Nilai and Pasir Gudang are classified as industrial since all the locations are located in rapidly growing towns with heavily industrialized areas that are affected by heavy traffic and seasonal high particulate events (Azid et al., 2015; Mohamed Noor et al., 2011; Yap & Hashim, 2013). Pasir Gudang monitoring station is located at a residential area within two kilometres of Pasir Gudang industrial area with various types of industries such as steel melting, fertilizers manufacturing, cement production, edible oil refinery, electroplating and a Tenaga National Berhad power generating plant. It is identified as the most polluted area in Johor (Lee et al., 2012).

p-ISSN 1675-7939; e-ISSN 2289-4934

2.2 Methods

2.2.1 Software

The IBM SPSS (Statistical Package for the Social Sciences) version 18 was used to obtain the descriptive statistics of the data. The R language, which provides a wide range for data manipulation, calculation and graphical displays was used to determine the trend of the PM_{10} concentrations of the monitoring stations (Albert, 2009). Matlab® version 11, a programming language for numerical computation, visualization, and programming package for engineers, was utilised to estimate parameters of the distributions and the performance indicators (Mathworks, 2015). OpenBUGS323 is an alternative to WinBUGS (**B**ayesian inference **U**sing **G**ibbs **S**ampling), a free downloadable software specifically designed for the Bayesian analysis of complex statistical models using Markov Chain Monte Carlo (MCMC) methods (Ntzoufras, 2009) was used to perform all the simulations of the Bayesian model.

2.2.2 Parameter Estimates

i. Extreme Value Distribution (EVD)

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. The method of estimations discussed in this paper is the method of Maximum Likelihood Estimator (MLE).

ii. Bayesian approach

The adoption of the Bayesian approach requires a likelihood distribution, the Generalized Extreme Value (GEV) which uses priors of three parameters, namely: location, μ, scale, σ and shape, λ parameters that are intended to represent beliefs about parameters, prior to the availability of data. This study adopts the non-informative priors of the GEV distribution which were assumed to be uniformly distributed with all means equal to zero and variances equal to 200, 50 and 10, respectively. The distribution was set at 1000 for convergence towards the equilibrium and was simulated at the value of 10,000 with 10 replicates.

2.3 Performance Indicators

Six performance indicators were used to select the best model to represent the concentrations records. The accuracy measures are the Prediction Accuracy (PA), Coefficient of Determination (R^2) and Index of Agreement (IA) of which the accuracy value is between 0 and 1. As the value approaches 1, the model is said to be appropriate. The error measures used are the Root Mean Square Error (RMSE), the Normalized Absolute Error (NAE) and the Mean Absolute Error (MAE) (Junninen et al., 2004; Yahaya & Ramli, 2008). As opposed to the accuracy measures, as the value of error measures approaches 0, the model is deemed to be the best model. 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-dependant (Ji & Gallo, 2006).

p-ISSN 1675-7939; e-ISSN 2289-4934

3. RESULTS AND DISCUSSIONS

3.1 Statistical Characteristics of PM_{10}

Table 1 presents the descriptive statistics of PM_{10} concentration for the monitoring stations. The unit of measurement of the concentrations is microgram per cubic meter ($\mu g/m^3$). The missing values were not included in the analysis since the missing value only constituted ≤ 1% of the total data. All the three average readings of the PM_{10} concentrations were well above the stipulated Malaysia Ambient Air Quality Guidelines (MAAQG) for the yearly average of 50 $\mu g/m^3$ (Department of Environment Malaysia, 2014). The highest average among the monitoring stations was that of Bukit Rambai's with 73.29 $\mu g/m^3$. All the records were skewed to the right - above 1, an indication of the existence of the extreme concentrations during the 2000 – 2012 period with Nilai indicating the highest (3.03). It is interesting to note that the maximum concentrations were recorded in Nilai with the reading of 344$\mu g/m^3$. The analysis indicates low variability in the monitoring records since all the coefficients of variation were less than or equal to 0.34.

Table 1: Descriptive statistics of daily maximum of PM_{10} concentration.

	Bukit Rambai	**Nilai**	**Pasir Gudang**
N valid	4694	4724	4734
Missing	55	25	15
Mean	73.29	65.75	55.63
Median	70.00	62.00	54.00
Std. Deviation	22.24	22.54	17.87
Skewness	1.87	3.03	1.66
Kurtosis	9.33	23.17	5.84
Minimum	22.00	23.00	20.00
Maximum	268.00	344.00	192.00
Coefficient of variation	0.30	0.34	0.32

3.2 Parameter Estimates and Performance Indicators

Table 2 provides the goodness-of-fit for classical and Bayesian GEV. In comparing the classical and the Bayesian approaches, the GEV Bayesian approach with the non-informative (NI) uniform prior was the best distribution for all the industrial monitoring stations with the smallest errors (NAE, RMSE and MAE) and the highest accuracy measures (PA, R^2 and IA). The indicators obtained were excellent with more than 97% in PA, IA and R^2 for all stations. The error measures recorded from the analyses were all below 5 - the closest to zero.

p-ISSN 1675-7939; e-ISSN 2289-4934

Table 2: Performance indicators for classical and Bayesian GEV.

Monitoring Stations	Method	Performance Indicators						Overall
		NAE	RMSE	MAE	PA	R^2	IA	
Bukit Rambai	GEV – EVD	0.3891	29.1232	28.5176	0.9735	0.9473	0.7325	GEV - NI
	GEV – NI	**0.0226**	**3.3652**	**1.6569**	**0.9887**	**0.9771**	**0.9943**	
Nilai	GEV – EVD	28.1464	48159.98	4029.47	0.3796	0.1421	0.0014	GEV - NI
	GEV – NI	**0.0233**	**4.8523**	**1.5322**	**0.9770**	**0.9541**	**0.9877**	
Pasir Gudang	GEV – EVD	**0.0126**	1.6076	**0.7014**	0.9959	0.9914	0.9979	GEV - NI
	GEV – NI	0.0127	**1.6054**	0.7062	**0.9960**	**0.9915**	**0.9980**	

3.3 Prediction

The selection of the best distribution to represent each of the monitoring stations was determined using the ranking of the performance indicators. Following the estimation of parameters and the selection of the best model based on the performance indicators, Probability Density Functions (PDF) and Cumulative Distribution Functions (CDF) of the GEV for all the monitoring stations were plotted. Figure 1 demonstrates that the plots of the distribution have long tails to the right indicating the existence of the extreme concentrations in those years under review. The distributions obtained fitted very well with the observed concentrations in all locations.

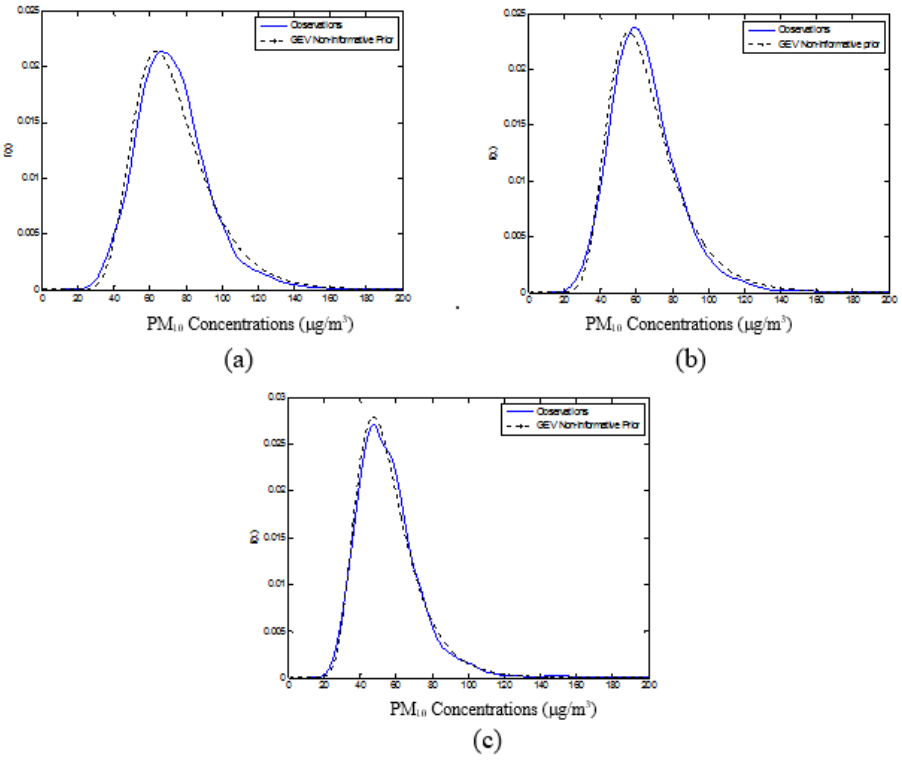

Figure1: Probability Density Function (PDF) of GEV for (a) Bukit Rambai, (b) Nilai, and (c) Pasir Gudang.

p-ISSN 1675-7939; e-ISSN 2289-4934

From the plots of CDF as shown in Figure 2, the probabilities of the concentrations exceeding the levels of MAAQG of 150 μg/m^3 were estimated.

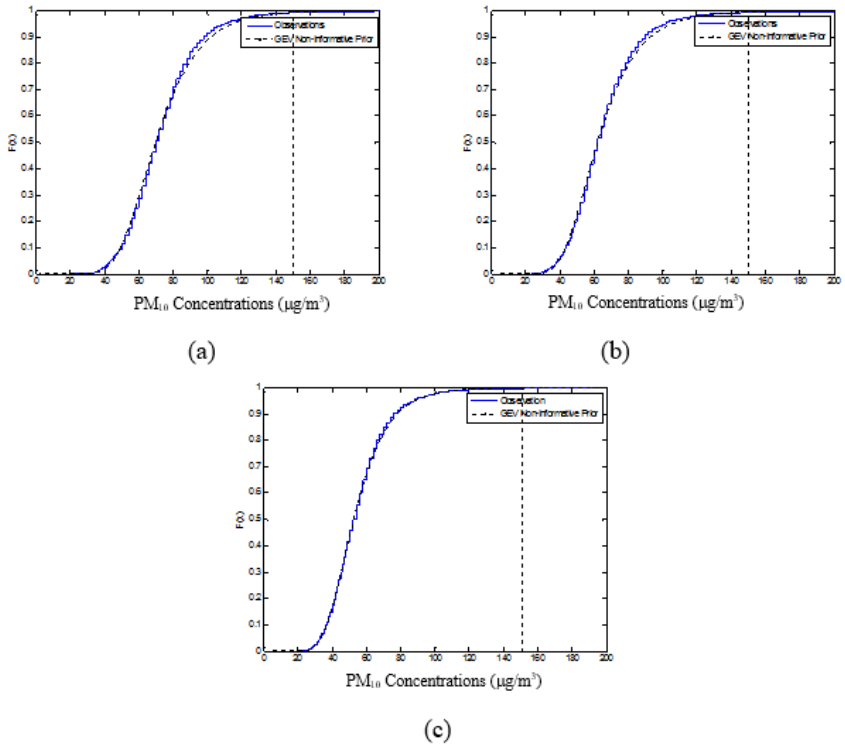

Figure 2: Cumulative Distribution Function (CDF) of GEV for for (a) Bukit Rambai, (b) Nilai and (c) Pasir Gudang.

Table 3 depicts the probability of PM_{10} concentrations exceeded MAAQG with the predicted and the actual number of days.

Table 3: Comparison of predicted and actual number of days with concentrations > 150 μg/m^3.

Stations	Probability concentrations exceeding 150 μg/m^3	Predicted	Actual	% compliance	Return Period
Bukit Rambai	0.0069	33	36	92	3
Nilai	0.0045	22	38	58	2
Pasir Gudang	0.0014	15	15	100	5

The Bayesian approach estimates the number of exceedances that have more than 92% compliance with the actual number of exceedances in Bukit Rambai and Pasir Gudang. The estimation was particularly good in Pasir Gudang with 100% compliance with the actual exceedances. The prediction is particularly important for Pasir Gudang due to the fact that the findings of the annual maximum trend indicated an increasing trend from 2000 – 2012 in this location.

The return period for every location is the reciprocal of the probability of exceedances (Selaman et al., 2007). In summary, it can be estimated that every year, Bukit Rambai will

p-ISSN 1675-7939; e-ISSN 2289-4934

experience an average of 3 days of extreme concentrations above 150μg/m^3 while Pasir Gudang is projected to have 5 days of high concentrations.

4. CONCLUSION

This paper discusses the best distribution/model to predict 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 three monitoring stations in the west coast of Malaysia. The two approaches, namely, the classical and Bayesian non-informative and informative priors were selected to fit the monitoring records. Both of the approaches utilized the GEV distribution. All the daily maximum concentrations records without missing values from 2000 – 2012 were used to analyse the efficiency of the GEV using these two different approaches. The analysis of three accuracy measures, namely, PA, R^2 and IA and three error measures – NAE, RMSE and MAE were obtained to indicate the efficiency or the performance indicators of the distributions.

From the findings, the average readings of the PM_{10} concentrations in all the monitoring stations were well above the stipulated MAAQG for the yearly average of 50 μg/m^3 with the maximum readings recorded in Nilai. 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 Bayesian GEV with non-informative priors gave the best estimators for all the monitoring stations with the smallest errors (NAE, RMSE and MAE) and the highest accuracy measures (PA, R^2 and IA) as compared to the MLE. 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 numbers of days for two of the monitoring stations were close to that of the actual numbers of days. To conclude, the Bayesian non-informative priors had an advantage over the MLE in all the monitoring stations under study since it provided better performance indicators in estimating the numbers of day that exceeded the specified levels of MAAQG of 150 μg/m^3 for daily concentrations.

Though Coles et al. (2003) and Kery (2010) affirmed that the Bayesian approach is superior than the conventional methods, but again, there is no single distribution and estimators that can best fit the concentrations in various locations in Malaysia since the appropriateness of the selected distribution depends on the selection of data, the availability of the size of the observations, the variability of the data and the choice of the specific test statistic as discussed by Soukissian and Tsalis (2015). The Bayesian approach might be suitable to represent data in these three locations, but it might be otherwise for the other locations. This study proposed that the distribution be used for the estimation of future exceedances of PM_{10} in these locations. 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.

p-ISSN 1675-7939; e-ISSN 2289-4934

ACKNOWLEDGMENT

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

REFERENCES

Albert, J. (2009). *Bayesian Computation with R*. (R. Gentleman, K. Hornik, & G. Parmigiani, Eds.) (Second Edition.). Oho, USA: Springer.

Azid, A., Juahir, H., Ezani, E., Toriman, M. E., Endut, A., Abdul Rahman, M. N., Yunus, K., Kamarudin, M. K. A., Che Hasnam, C. N., Mohd Saudi, A. S., & Umar, R. (2015). Identification Source of Variation on Regional Impact of Air Quality Pattern Using Chemometric. *Aerosol and Air Quality Research, 15*, 1545–1558.

Bowman, D. M. J. S., & Johnston, F. H. (2005). Wildfire Smoke, Fire Management, and Human Health. *EcoHealth*, *2*(1), 76–80.

Chung, E.-S., & Kim, S. U. (2013). Bayesian Rainfall Frequency Analysis with Extreme Value using the Informative Prior Distribution. *KSCE Journal of Civil Engineering, 17*(6), 1502–1514. doi:10.1007/s12205-013-0189-0

Coles, S., Pericchi, L. R., & Sisson, S. (2003). A Fully Probabilistic Approacha to Extreme Rainfall Modeling. *Journal of Hydrology, 273*, 35–50.

Department of Environment Malaysia. (2014). *Malaysia Environmental Quality Report 2013*. Kuala Lumpur.

Finkenstadt, B., & Rootzen, H. (Eds.). (2001). *Extreme Values in Finance, Telecommunications and the Environment*. Florida, United States of America: Chapman & Hall/CRC.

Hedley, A. J. (2009). The Current Avoidable Burden of Health Problems , Community Costs and Harm to Future Generations. *Journal of Toxicology and Environmental Health, Part A, 71*(9), 544–554.

Ilyas, S. Z., Khattak, A. I., Nasir, S. M., Qurashi, T., & Durrani, R. (2010). Air Pollution Assessment in Urban Areas and its Impact on Human Health in the city of Quetta, Pakistan. *Clean Technologies and Environmental Policy, 12*(3), 291–299.

Ji, L., & Gallo, K. (2006). An Agreement Coefficient for Image Comparison. *Photogrammetric Engineering & Remote Sensing, 72*(7), 823–833.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for Imputation of Missing Values in Air Quality Data Sets. *Atmospheric Environment, 38*, 2895–2907.

Kery, M. (2010). *Introduction to Winbugs for Ecologists*. Switzerland: Academic Press.

p-ISSN 1675-7939; e-ISSN 2289-4934

Kim, S. Y., O'Neill, M. S., Lee, J. T., Cho, Y., Kim, J., & Kim, H. (2007). Air Pollution, Socioeconomic position, and Emergency Hospital visits for Asthma in Seoul, Korea. *International Archives of Occupational and Environmental Health, 80*, 701–710.

Kotz, S., & Nadarajah, S. (2000). *Extreme-Value Distributions : Theory and Applications.* London: Imperial College Press.

Lee, M. H., Abd. Rahman, N. H., Suhartono, Latif, M. T., Nor, M. E., & Kamisan, N. A. B. (2012). Seasonal ARIMA for Forecasting Air Pollution Index : A Case Study. *American Journal of Applied Sciences, 9*(4), 570–578.

Mathworks. (2015). *MATLAB Programming Fundamentals* (8.5 ed.). MA: The MathWorks, Inc. Retrieved from https://www.mathworks.com/help/pdf_doc/matlab/matlab_prog.pdf

McKenzie, D., O'Neill, S. M., Larkin, N. K., & Norheim, R. A. (2006). Integrating Models to Predict Regional Haze from Wildland Fire. *Ecological Modelling, 199*, 278–288.

Mohamed Noor, N., Tan, C. Y., Abdullah, M. M. A.-B., Ramli, N. A., & Yahaya, A. S. (2011). Modelling of PM10 Concentration in Industrialized Area in Malaysia : A Case Study in Nilai. In 2011 *International Conference on Environment and Industrial Innovation* (Vol. 13, pp. 18–22). Singapore: IACSIT Press.

Mott, J. A., Mannino, D. M., Alverson, C. J., Kiyu, A., Hashim, J., Lee, T., Falter, K., & Redd, S. C. (2005). Cardiorespiratory Hospitalizations associated with Smoke exposure during the 1997 Southeast Asian Forest Fires. *International Journal of Hygiene and Environmental Health, 208*, 75–85.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS.* (P. Giudici, G. H. Givens, & B. K. Mallick, Eds.). New Jersey, USA: John Wiley and Sons, Inc.

Ostermann, K., & Brauer, M. (2001). Air Quality During Haze Episodes and Its Impact on Health. In P. Eaton & M. Radojevic (Eds.), *Forest Fires and Regional Haze in Southeast Asia* (p. 26).

Othman, J., Sahani, M., Mahmud, M., & Ahmad, M. K. S. (2014). Transboundary Smoke Haze Pollution in Malaysia: Inpatient Health Impacts and Economic Valuation. *Environmental Pollution, 189*, 194–201.

Sastry, N. (2002). Forest fires, Air Pollution, and Mortality in Southeast Asia. *Demography, 39*(1), 1–23.

Selaman, O. S., Said, S., & Putuhena, F. J. (2007). Flood Frequency Analysis for Sarawak using Weibull, Gringorten and L-Moments Formula. *Journal - the Institution of Engineers, Malaysia, 68*(1), 43–52.

Soukissian, T. H., & Tsalis, C. (2015). The Effect of the Generalized Extreme Value Distribution Parameter Estimation Methods in Extreme Wind Speed Prediction. *Natural Hazards, 78*, 1777–1809.

p-ISSN 1675-7939; e-ISSN 2289-4934

Sulaiman, M., Ibarahim, H. R., & Hooper, M. (2003). *Management Of Haze ; An Asean Regional Perspective*. Retrieved from http://www.fire.uni-freiburg.de/summit-2003/3-IWFC/Papers/3-IWFC-002-Rosnani-Ibrahim.pdf

Sun, Z., An, X., Tao, Y., & Hou, Q. (2013). Assessment of Population Exposure to PM10 for Respiratory Disease in Lanzhou (China) and its Health-Related Economic Costs based on GIS. *BMC Public Health, 13*(1), 891.

Tsai, Y. I. (2005). Atmospheric Visibility Trends in an Urban Area in Taiwan 1961–2003. *Atmospheric Environment, 39*(30), 5555–5567.

Vedal, S., & Dutton, S. J. (2006). Wildfire Air Pollution and Daily Mortality in a Large Urban area. *Environmental Research, 102*(1), 29–35.

Yadav, A. K., Kumar, K., Hj Awang Kasim, A. M., Singh, M. P., Parida, S. K., & Sharan, M. (2003). Visibility and Incidence of Respiratory Diseases During the 1998 Haze Episode in Brunei Darussalam. *Pure and Applied Geophysics, 160*(1-2), 265–277.

Yahaya, A. S., & 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. Short Term Grant Report.

Yap, X. Q., & Hashim, M. (2013). A Robust Calibration Approach tor PM_{10} Prediction from MODIS Aerosol Optical Depth. *Atmospheric Chemistry and Physics, 13*, 3517–3526.

p-ISSN 1675-7939; e-ISSN 2289-4934