SciVerse ScienceDirect

Physics Procedia 22 (2011) 318 – 324

Physics Procedia

2011 International Conference on Physics Science and Technology (ICPST 2011)

Modelling of PM_{10} concentration for industrialized area in Malaysia: A case study in Shah Alam

Norazian Mohamed N.[a]*, M.M.A. Abdullah[a], Cheng-yau Tan[b], N.A. Ramli[c], A.S. Yahaya[c], N.F.M.Y.Fitri[c]

[a]*Department of Environmental Engineering, Universiti Malaysia Perlis, Kompleks Pengajian Jejawi 3, 02600 Jejawi, Perlis, Malaysia*
[b]*Institute of Postgraduate Studies, Universiti Malaya, 50603 Kuala Lumpur, Malaysia*
[c]*Clean Air Research Group, Environmental and Sustainable Development Section, School of Civil Engineering, UniversitiSains Malaysia, Engineering Campus, NibongTebal, 14300, Pulau Pinang, Malaysia.*

Abstract

In Malaysia, the predominant air pollutants are suspended particulate matter (SPM) and nitrogen dioxide (NO_2). This research is on PM_{10} as they may trigger harm to human health as well as environment. Six distributions, namely Weibull, log-normal, gamma, Rayleigh, Gumbel and Frechet were chosen to model the PM_{10} observations at the chosen industrial area i.e. Shah Alam. One-year period hourly average data for 2006 and 2007 were used for this research. For parameters estimation, method of maximum likelihood estimation (*MLE*) was selected. Four performance indicators that are mean absolute error (*MAE*), root mean squared error (*RMSE*), coefficient of determination (R^2) and prediction accuracy (*PA*), were applied to determine the goodness-of-fit criteria of the distributions. The best distribution that fits with the PM_{10} observations in Shah Alamwas found to be log-normal distribution. The probabilities of the exceedences concentration were calculated and the return period for the coming year was predicted from the cumulative density function (*cdf*) obtained from the best-fit distributions. For the 2006 data, Shah Alam was predicted to exceed 150 μg/m^3 for 5.9 days in 2007 with a return period of one occurrence per 62 days. For 2007, the studied area does not exceed the MAAQG of 150 μg/m^3.

Keywords: Particulate matter; Probability distributions; Performance indicators; Exceedences; Return Period

*Corresponding author. Tel.: +6012 5075020; fax:+604 9798636.

Email address: norazian@unimap.edu.my

• Introduction

Exponential development of science and technology nowadays has lead to the rapid growing industrialization which is the major sources of various environmental pollutions, especially air pollution. Airpollutants, specifically particulate matter (PM) smaller than about 10 micrometers, referred as PM_{10}, have received extensive attention, due to its capability to settle in the bronchi and lungs and cause health problems [1]. Malaysian Ambient Air Quality Guidelines (MAAQG) were issued and target values for annual and daily mean mass concentrations for various air pollutant were established to control and reduce air pollutant levels in the atmosphere. Monitoring data and studies on ambient air quality show that some of the air pollutants in several large cities are increasing with time and are not always at acceptable levels according to the MAAQG. There are very limited data and case studies on air pollution in our country. Most of the air modeling using probability distribution is only applied in foreign countries.

1875-3892 © 2011 Published by Elsevier B.V. Selection and/or peer-review under responsibility of Garry Lee.
Open access under CC BY-NC-ND license. doi:10.1016/j.phpro.2011.11.050

Statistics are important in the analysis and interpretation of data in which the outcomes from the analysis can be utilized as prediction tools that have become the major aim in environmental engineering [2]. There are many statistical procedures to analyze various environmental data sets which are frequently asymmetrical and skewed to the right (that is with long tail towards high concentrations) [3]. Many types of probability distributions have been used to fit air pollutant concentrations including Weibull distribution [4], lognormal distribution [5], gamma distribution [6], Rayleigh distribution [7],Gumbel distribution [8] and Frechetdistribution . Lu [9] and Chen *et al.* [10] have studied the goodness-of-fit for selected probability distributions by using several performance indicators such as mean absolute error (*MAE*), root means error (*RMSE*), index of agreement (d_2), bias (*B*), normalized absolute error (*NAE*), prediction accuracy (*PA*) and coefficient of determination (R^2).The goals of this research were to study the statistical characteristics of the observed data, as well as to select the best-fit distribution in order to predict the exceedences and return period of the PM_{10} critical concentration.

• Data and methods

2.1. Data set

• The datasets consisted of PM_{10} concentration on a time-scale of one per hour (hourly averaged) for 2006 and 2007 in Shah Alam, Selangor. It is an industrialized area with high population and traffic density with the weak prevailing winds was recordedcausing the air contaminants to stagnate [11].

2.2.Probability distributions

• Six theoretical distributions, namely Weibull, gamma, log-normal, Rayleigh, Gumbel and Frechet distributions are used to fit the entire measured PM_{10} data [12,13]. For parameters estimation, method of maximum likelihood estimation (*MLE*) was selected.

2.3Performance indicators

• Four performance indicators (PI) that are mean absolute error (*MAE*), root mean squared error (*RMSE*), coefficient of determination (R^2) and prediction accuracy (*PA*), were applied to determine the goodness-of-fit criteria as to judge which type of parent distribution is the most appropriate to represent the PM_{10} pollutant concentration [14].

2.4Exceedences and return period

• Once the best-fit distribution is determined, the cumulative distribution function (cdf) of the fitted distribution was used to calculate the exceedence, or the probability that the event is equalled or exceeded in computed period. The reciprocal of the exceedance probability was calculated so to obtain the return period (also known as the recurrence interval) of the event.

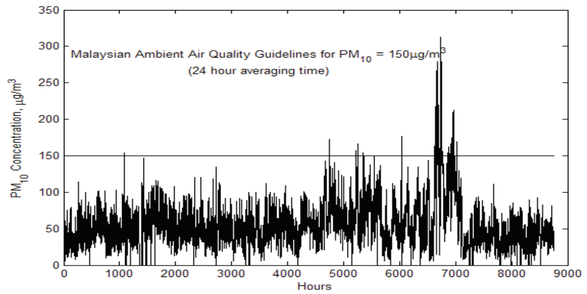
• Results and discussion

• *3.1Data description*

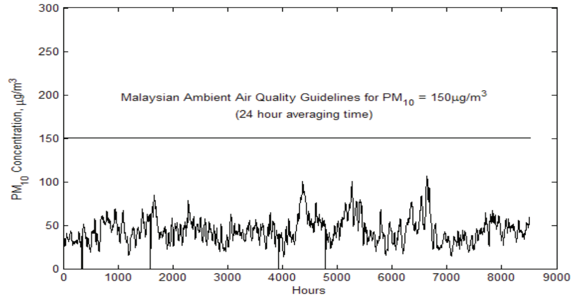
- Table 3.1 gives the summary of the descriptive statistics for PM_{10} hourly data of Shah Alam for 2006 and 2007. The mean values for the area in both years are higher than their respective median which indicates that the pollutants distributions are positively skewed (also called right-skewed). This means most of the data is concentrated on the left of the figure with few high values. The maximum value for Shah Alam for 2006 was 313.0 decreased to below MAAQG limit of 150 µg/m^3 in 2007.Fig. 1.indicates the PM_{10} concentrations had small exceedences starting from the end of September to the mid of October due to the hazeepisodes in Malaysia in 2006 [15]. No exceedenceabove 150 µg/m^3 was observed for 2007 as shown in Figure 1.

3.2Probability distributions

Table 2 shows the parameter estimates of the six distributions for 2006 and 2007. All the estimates have been obtained using maximum likelihood estimators (MLE).It is observed that the value of scale parameter, β, is always larger than the corresponding shape parameter, α, indicating the computations are done right.



PM_{10} concentration in Shah Alam (2006)



PM_{10} concentration in Shah Alam (2007)

Fig.1.Time series plot of PM_{10}concentration in Shah Alam for 2006 and 2007

Table 1.Descriptive statistics for PM_{10} concentration

		Shah Alam	
		2006	**2007**
Valid Data		8598	8462
Missing Data		162	298
Mean		55.7	44.5
Median		50.0	42.8
Standard Deviation		30.7	14.6
Mode		48.0	41.7
Variance		942.8	212.6
Skewness (Standard Error)		2.13 (0.03)	0.86 (0.03)
Kurtosis (Standard Error)		8.79 (0.05)	1.23 (0.05)
Minimum Value		6.0	13.2
Maximum Value		313.0	106.9
Range		307.0	93.7
Percentiles	25	36.0	34.2
	50	50.0	42.8
	75	69.0	52.4

Table 2.Parameter estimates

Distributions	2006		2007	
	α	β	α	β

Weibull	1.933	63.06	3.167	49.622
Gamma	3.921	14.216	4.603	9.67
Log-Normal	0.524	3.888	0.327	3.743
Rayleigh	-	44.999	-	33.118
Gumbel	21.34	42.959	12.0	37.774
Frechet	1.739	37.3	2.936	35.756

Fig.2 shows the cumulative distribution (cdf) plots for PM_{10} concentration in Shah Alam where six distributions were plotted and compared with the observed distribution.cdf plots from For 2006, gamma, log-normal and Gumbel distributions have better fit on the PM_{10} observation data in Shah Alam. For 2007, log-normal distribution has better fit than Weibull, gamma, and Gumbel which also fit well the observed data. Frechet distribution overestimates at concentration less than 27 µg/m³ and underestimates after that indicating the worst fitting.

- *3.3Performance indicators*

- Table 3 shows the smallest value for *MAE* is given by the Gumbel distribution (2006) and that of *RMSE* is given by log-normal distribution which also indicates the highest values for R^2 and *PA*, i.e. 0.989 and 0.995 respectively. Gamma distribution (2007) gives the smaller value of *MAE* that is 0.5521 compared to log-normal distribution which is 0.6631. However, log-normal distribution indicates smaller value of *RMSE* that is 1.0561 compared to gamma distribution, 1.0938. The highest value for R^2 and *PA* are both given by log-normal distribution with the value of 0.9953 and 0.9978 respectively. Based on the analysis of these results, log-normal distribution fits the databetter than Gumbel distribution (2006) and gamma distribution (2007) in representing the PM_{10} concentration in Shah Alam for both 2006 and 2007.

- *3.4Exceedences and return period*

The distribution that fits the PM_{10} concentration in Shah Alam is log-normal for both 2006 and 2007. From Fig.4, the probability that the PM_{10} concentration for 2006 equal or less than 150 µg/m³ is 0.9839 [that is, $P\{X \leq 150\} = 0.9839$] and the probability that the concentration greater than 150 µg/m³ is 0.0161 [that is, $P\{X > 150\} = 0.0161$]. There will be 5.9 days where the PM_{10} concentrations in 2007 exceed 150 µg/m³. Hence the return period for 2007 is once per 62 days. Meanwhile, the probability that the concentration in Shah Alam for 2007 greater than 150 µg/m³ is 0 [that is, $P\{X > 150\} = 0$]. This shows the PM_{10} concentrations for the whole year stay below 150 µg/m³. There is no return period predicted for concentration above 150 µg/m³ in 2008.

- **Conclusion**

- Six distributions namely Weibull, log-normal, gamma, Rayleigh, Gumbel and Frechet distributions were chosen to model the PM_{10} observations inShah Alam, Selangor. One-year period hourly average data for 2006 and 2007 were used. Method of maximum likelihood estimation (*MLE*) was selected for parameters estimation. Four performance indicators specifically mean absolute error (*MAE*), root mean squared error (*RMSE*), coefficient of determination (R^2) and prediction accuracy (*PA*), were applied to determine the goodness-of-fit criteria of the distributions. The best distribution that fits with the PM_{10} observationsin Shah Alamwas found to be log normal distribution. The probabilities of the exceedences concentration were calculated and the return period for the coming

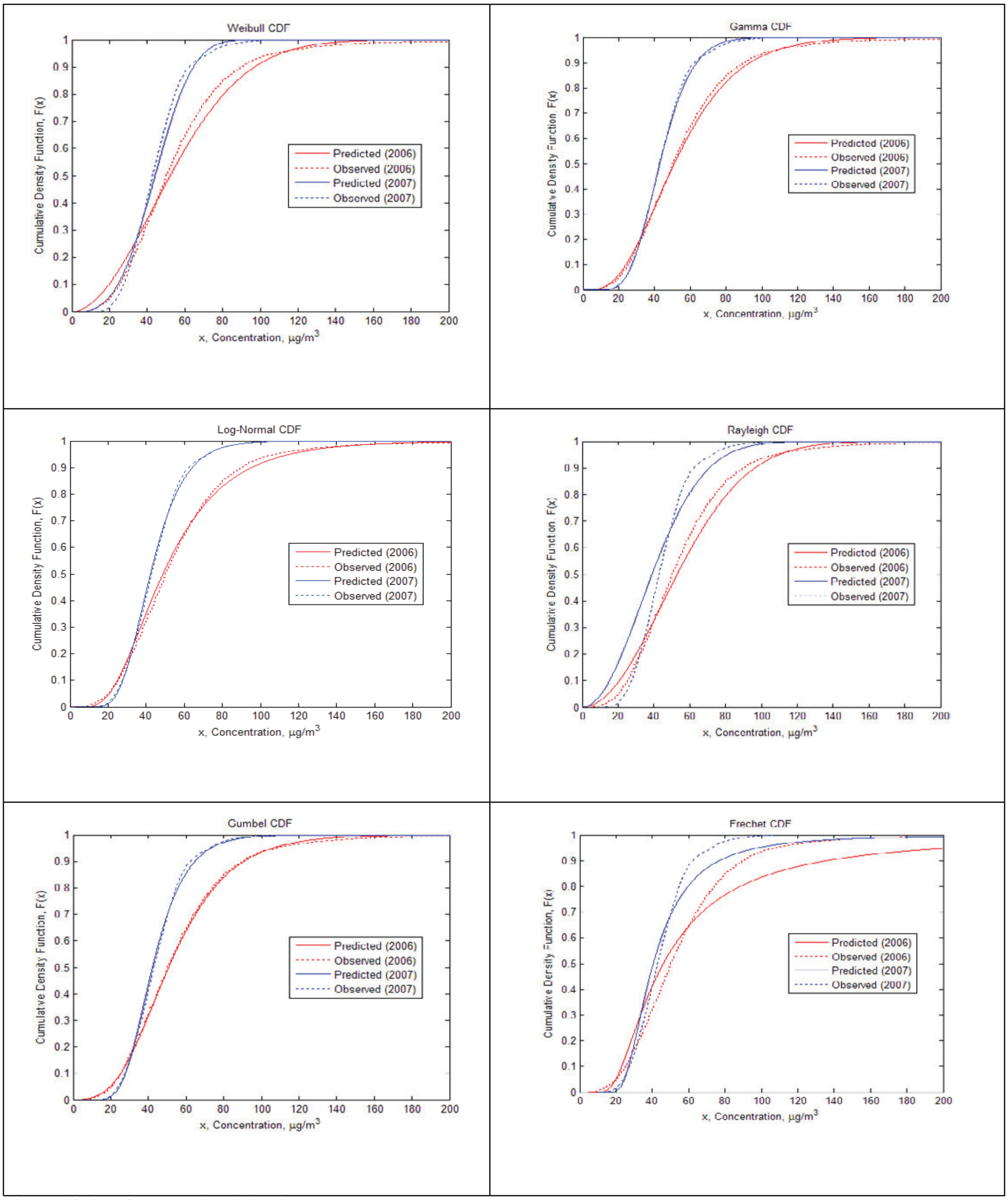
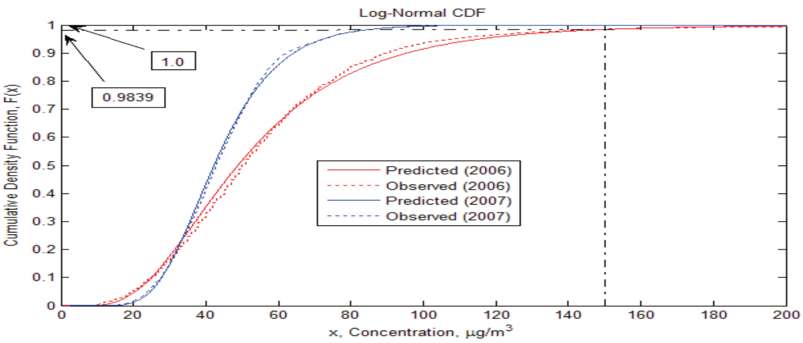


Fig.2 .cdf plots for PM10 in Shah Alam

Table 3.Performance Indicators value for PM_{10} concentration in Shah Alam

Distributions	Performance Indicators			
	Mean Absolute Error (*MAE*)	Root Mean Squared Error (*RMSE*)	Coefficient of Determination (R^2)	Prediction Accuracy (*PA*)

	2006	2007	2006	2007	2006	2007	2006	2007
Weibull	4.758	2.174	8.296	2.839	0.927	0.967	0.963	0.983
Gamma	2.253	0.552	6.555	1.094	0.958	0.994	0.979	0.997
Log-Normal	2.009	0.663	3.267	1.056	0.989	0.995	0.995	0.998
Rayleigh	4.686	6.829	8.541	7.88	0.923	0.991	0.961	0.996
Gumbel	1.514	0.823	6.111	1.39	0.969	0.994	0.985	0.997
Frechet	25.063	6.641	133.26	21.241	0.528	0.74	0.727	0.86



•

•

Fig.3.Estimation of exceedences above MAAQG (150 µg/m^3) in Shah Alam for 2006 and 2007 using log-normal cdf plot

year was predicted. For the 2006 data, Shah Alam was predicted to exceed 150 µg/m^3 for 6 days in 2007 with a return period of one occurrence per 62 days. However, the PM_{10} observations for 2007 do not exceed the MAAQG of 150 µg/m^3.

• Acknowledgement

This project was granted by Fundamental Research Grant Scheme (FRGS 9003-00272). We also wish to thank Department of Environment (DoE) Malaysia for the data.

• References

[1] Nawrot, T.S., Torfs, R., Fierens, F., De Henauw, S., Hoet, P.H. and Van Kersschaever, G. (2007). Stronger Associations Between Daily Mortality and Fine Particulate Air Pollution in Summer Than in Winter: Evidence From A Heavily Polluted Region in Western Europe. Epidemiology and Community Health, 61, pp. 146-149.

[2] Gilbert, O. R., *Statistical Method for Environmental Pollution Monitoring*. Van Nostrand Reinhold Company Inc., New York 1987.

[3] Wang, X. and Mauzerall, D. L., Characterizing Distributions of Surface Ozone and its Impact on Grain Production in China, Japan and South Korea: 1990 and 2020, *Atmos Environ*, 2004**38** (74), 4383-4402.

[4] Hadley, A. and Toumi, R., Assessing Changes to the Probability Distribution of Sulphur Dioxide in the UK Using Lognormal Model, *AtmosEnviron*, 2002**37** (24), 455-467.

[5] Singh, P., Simultaneous Confidence Intervals for the Successive Ratios of Scale Parameters,*J Stat Plan Infer*, 2004**36** (3), 1007-1019.

[6] Celik, A. N., A Statistical Analysis of Wind Power Density Based on TheWeibull and Rayleigh Models at The Southern Region of Turkey.*Renew Energ*,2003**29**, 593-604.

[7] Phien, H. N., A Computer Assisted Learning Package for Flood Frequency Analysis with the Gumbel Distribution. *AdvEngSoftw*, 1989 **11**, 206-212.

[8] Caleyo,F., Velázquez, J.C., Valor, A., and Hallen, J.M.,Probability distribution of pitting corrosion depth and rate in

underground pipelines: A Monte Carlo study. *CorrosSci*, 2009**51**, 1925-1934.

[9] Lu, H. C., Estimating the Emission Source Reduction of PM_{10} in CentralTaiwan. *Chemosphere*, 2003**54**, 805-814.

[10] Chen, J.L., Islam, S. and Biswas, P., Nonlinear Dynamics of Hourly Ozone Concentrations: Nonparametric Short Term Prediction, 1998, *Atmos Environ*,**32**, 1839-1848.

[11] Department of Environment, Malaysia *Malaysia Environmental Quality Report*. Department of Environment, Ministry of Natural Resources and Environment, Malaysia 2007.

[12] Evans, M. Hastings, N. and Peacock, B., *Statistical Distributions.* 3rd Edition, Wiley, New York 2000.

[13] Kottegoda, N. T. and Rosso, R.,*Statistic, Probability and Reliability for Civil and Environmental Engineers.* McGraw-Hill, Singapore 1998.

[14] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M., Methods for Imputation of Missing Values in Air Quality Data Sets. *AtmosEnviron*,2004. **38**, 2895-2907.

[15] Department of Environment, Malaysia, *Malaysia Environmental Quality Report*, Department of Environment, Ministry of Natural Resources and Environment, Malaysia 2006.