Environ Monit Assess (2011) 180:573–588
DOI 10.1007/s10661-010-1806-8

Statistical analysis of PM_{10} concentrations at different locations in Malaysia

Nurulilyana Sansuddin · Nor Azam Ramli · Ahmad Shukri Yahaya · Noor Faizah Fitri MD Yusof · Nurul Adyani Ghazali · Wesam Ahmed Al Madhoun

Received: 19 July 2010 / Accepted: 22 November 2010 / Published online: 7 December 2010
© Springer Science+Business Media B.V. 2010

Abstract Malaysia has experienced several haze events since the 1980s as a consequence of the transboundary movement of air pollutants emitted from forest fires and open burning activities. Hazy episodes can result from local activities and be categorized as "localized haze". General probability distributions (i.e., gamma and log-normal) were chosen to analyze the PM_{10} concentrations data at two different types of locations in Malaysia: industrial (Johor Bahru and Nilai) and residential (Kota Kinabalu and Kuantan). These areas were chosen based on their frequently high PM_{10} concentration readings. The best models representing the areas were chosen based on their performance indicator values. The best distributions provided the probability of exceedances and the return period between the actual and predicted concentrations based on the threshold limit given by the Malaysian Ambient Air Quality Guidelines (24-h average of 150 $\mu g/m^3$) for PM_{10} concentrations. The short-term prediction for PM_{10} exceedances in 14 days was obtained using the autoregressive model.

Keywords Gamma distribution · Log-normal distribution · PM_{10} · Autoregressive (AR) model

N. Sansuddin · N. A. Ramli (✉) · A. S. Yahaya · N. F. Fitri MD Yusof · N. A. Ghazali · W. A. Al Madhoun
Clean Air Research Group, Environmental and Sustainable Development Section, School of Civil Engineering, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong Tebal, Pulau Pinang, Malaysia
e-mail: ceazam@eng.usm.my

Introduction

Haze events have become common phenomena in urban and industrial areas in Malaysia due to increasing quantities of pollutants emitted into the atmosphere by local anthropogenic sources (Malaysian Meteorological Service 2008). Malaysia has experienced several haze events since the 1980s. The 1997 haze episode is regarded as the most severe in Malaysian history. Nevertheless, this type of haze does not occur continuously, and the levels are generally below the necessary limit to evoke emergency status. Furthermore, haze events occur and pose an impact only in the local level; thus, they are generally categorized as "localized haze" (Ramli et al. 2001).

Data monitoring and ambient air-quality studies show that some of the air pollutants in several large cities in Malaysia are increasing with time and are not always at acceptable levels (Malaysian Ambient Air Quality Guidelines (MAAQG) Department of Environment (DoE Malaysia) 2002). The industrialization policy has started to impose costs in terms of pollution and the degradation of the urban environment. In line with the need for regional harmonization and for easy comparison with countries in the region, DoE Malaysia revised its index system in 1996 and adopted the Air Pollutant Index (API). The guidelines in the MAAQG form the basis for the interpretation of the API (Table 1. The listed concentration values are regarded as the "safe levels" (Awang et al. 2000). As the main pollutant involved in this research, the average threshold limit concentration for particulate matter (PM_{10}) is at 150 μg/m^3 for a 24-h period.

A country with a tropical climate such as Malaysia experiences uniform temperature and continuous high humidity. Seasons in this country are distinguished according to the changes of wind flow patterns and rainfall. The wind throughout the country is generally light and variable because the country is located near the equator. However, uniform periodic changes in the wind flow patterns describe the country's four seasons: northeast monsoon (November to March), transitional period (April to May), southwest monsoon (June to September), and another transitional period (October to November) (Malaysian Meteorological Department 2008).

Fuller and Murphy (2006) stated that the forest clearing fires are tightly linked to the monsoon systems in the region. Furthermore, Hyer and Chew (2010) reported that accidental fires can be easily started in the forest, especially in Kalimantan, during the dry season. In the case of the haze event in 1997, the particulates came from the biomass burning to clear vegetated (forest and grassland) areas in Indonesia. Wild fires significantly increase the input of organic aerosol components in the atmosphere (Abas et al. 2004).

The levels of particulate matter, particularly PM_{10}, are remarkably high around the cities during the dry season (Kim Oanh et al. 2006). In addition, the concentration of air particulate matter in Malaysia is influenced by the southwest monsoon wind and the occurrence of biomass burning (Abas et al. 2004). During non-haze periods, the level of particulate matter is influenced mostly by motor vehicles and industries (Afroz et al. 2003).

Soleiman et al. (2003) reported that a heavily industrialized urban area in Malaysia such as Klang Valley experienced severe haze episodes during the early August 1990, October 1991, and August–October 1994. There are two kinds of haze in the Klang Valley region: (1) shallow localized haze that arises from trapping the pollutants from anthropogenic emissions and (2) dense haze due either to the injection of suspended ash particles from large-scale forest fires and open burning in Indonesia or a combination of both. Haze occurs during the southwest monsoon season because low-level winds over Malaysia are generally southwesterly and the weather is generally dry. This paper focuses on industrial areas (i.e., Johor Bahru and Nilai) that have reported high readings of PM_{10} concentrations during normal periods and haze events, which are influenced by monsoonal variations. This research also includes Kota Kinabalu and Kuantan as representatives for residential areas.

Wang and Zhang (2009) stated that ambient air pollution in China is closely associated with industrialization and urbanization. This means that the urban population is likely to be the primary group exposed to high levels of ambient air pollution. Similarly, DoE Malaysia (2008) claimed that industries (including power stations), motor vehicles, and open burning activities remain the major sources of air pollution in the country. In 2008, 22,971 industrial sources were recorded, with the highest number of stationary sources located in Johor (35.4%) (DoE Malaysia 2008).

Many statistical methods have been developed to analyze datasets. However, environmental moni-

Table 1 Malaysian ambient air quality guidelines

Pollutant	Averaging time	Malaysia guideline	
		Ppm	μg/m^3
Particulate matter	24 h		150
(PM_{10})	1 year		50
Carbon monoxide	1 h	30	35
(CO)	8 h	9	10
Nitrogen dioxide	1 h	0.17	320
(NO_2)	24 h	0.04	
Sulphur dioxide	1 h	0.13	350
(SO_2)	24 h	0.04	105
Ozone (O_3)	1 h	0.10	200
	8 h	0.06	120

Department of Environment, Malaysia (2002) toring records are frequently asymmetrical and skewed to the right (i.e., with a long tail towards high concentrations); therefore, the validities of the procedures are questionable (Gilbert 1987). At present, only a few studies on distribution fitting have been conducted in Malaysia (e.g., Ramli and Ibrahim 2003; Sedek et al. 2006; Fitri et al. 2009) to understand the trends in PM_{10} concentrations. Air pollution monitoring in Malaysia was reported recently by Fitri et al. (2009) for PM_{10} and Ghazali et al. (2010) with regard to ozone predictions.

With regard to the air pollution in Malaysia, Fitri et al. (2009) and Juneng et al. (2009) observed the air pollutant concentrations during monsoon periods. They reported that the high concentration of pollutant is reduced drastically during the subsequent months when the winds change. It is signifying the beginning of the rainy season over the part of Peninsular Malaysia and Borneo, based on southwest monsoon (June–September) and northeast monsoon (November–March).

Emission levels and meteorological conditions influence the concentrations of air pollutants. When the parent probability distribution of air pollutants is chosen correctly, the specific distribution can be used to predict the mean concentration and probability of exceeding critical concentration (Lu and Fang 2003). Selection of appropriate probability models for the data is an important step in environmental data analysis. These probability models may serve as the basis for estimating the parameters to meet the evolving information needs of environmental quality management (Singh et al. 2001). In this research, probability distributions are used together with the maximum likelihood estimation (MLE). Mage and Ott (1984) stated that MLE provides the best estimate of the parameters; thus, this research uses MLE to estimate the distribution parameters simulating the PM_{10} concentrations. Moreover, several different approaches to time series modeling, such as autoregressive (AR) model, moving average (MA) model, and the combination of two autoregressive moving average (ARMA) models, are used in this research. The time series forecasting is an important analysis tool to extrapolate the time series for future prediction.

Methods

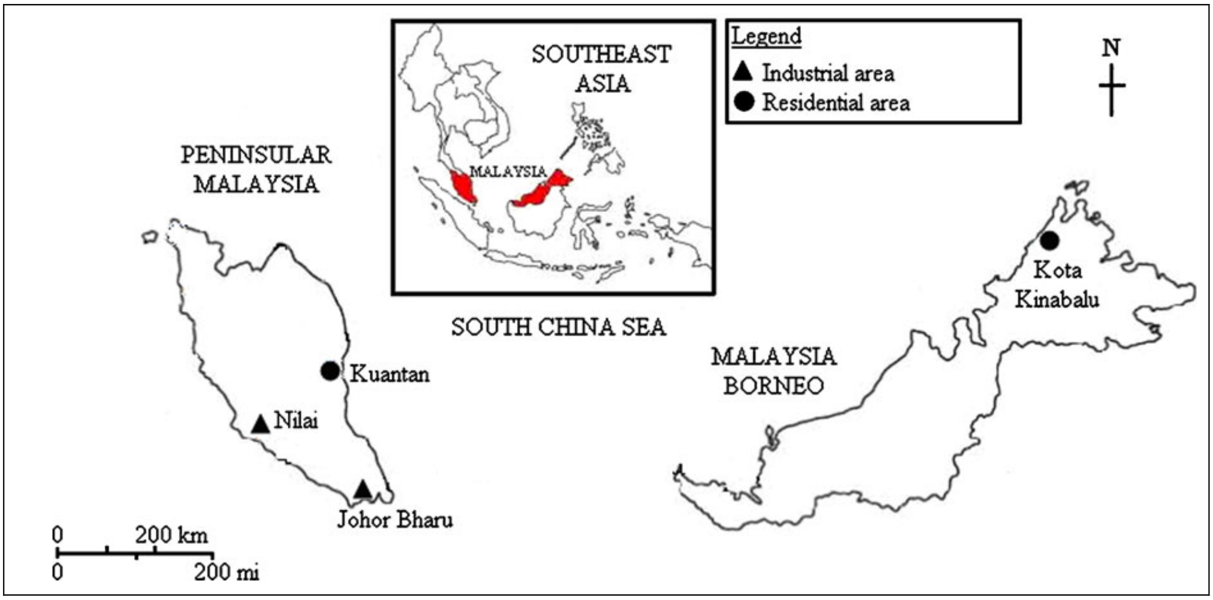
In this research, monitoring records were obtained from DoE Malaysia for four monitoring stations: Johor Bahru, Kota Kinabalu, Kuantan, and Nilai. Hourly PM_{10} concentrations for 2002 were obtained as well.

Figure 1 shows the location of the four monitoring stations. These stations are divided into two groups: industrial (Johor Bahru and Nilai) and residential (Kuantan and Kota Kinabalu). Johor Bahru and Nilai are situated in the South Peninsular Malaysia, whereas Kuantan is situated in Pahang, which is at the East Coast of Peninsular Malaysia, and Kota Kinabalu is situated in Sabah, in the island of Borneo. These monitoring stations can also be divided into two parts horizontally: northern (Kuantan and Kota Kinabalu) and southern (Johor Bahru and Nilai).

These stations were selected due to their location differences. They are also expected to be highly polluted due to industrialization, rapid development, and rapid economic growth accompanied by population growth. They may also be affected by the transboundary pollution from neighboring countries.

Data collection

Secondary monitoring records for PM_{10} concentrations were used in this research. The records were used to determine the best models and verification processes for the specific analyses. The PM_{10} concentration records are reliable since they have undergone a process of quality assurance and quality control established by the standards provided by DoE Malaysia. The hourly average PM_{10} concentrations were collected by Alam Sekitar Sdn. Bhd. The secondary monitoring records from 2000 to 2003 for the four selected monitoring stations were monitored using the federal equivalence method instrument known as beta attenuation mass (BAM). The hourly average of PM_{10} concentrations were used to find the best model for probability distribution analysis. These monitoring records were also converted to daily average of PM_{10} concentrations for the time series analysis.



Monitoring stations	***Coordinates***	***Station name***
Johor Bahru	N 01^{o} 29.815 E 103^{o} 43.617	Sekolah Menengah Vokasional Perdagangan
Kota Kinabalu	N 05^{o} 53.623 E 116^{o} 02.596	Sekolah Menengah Kebangsaan Putatan
Kuantan	N 03^{o} 49.138 E 103^{o} 17.817	Sekolah Kebangsaan Indera Mahkota
Nilai	N 02^{o} 49.3001 E 101^{o} 48.6894	Taman Semarak (Phase II)

Fig. 1 Map of the four chosen monitoring stations in Malaysia

Probability distributions

In this research, two probability distributions were needed for forecasting the PM_{10} concentrations: gamma and log-normal distribution. The MLE for the gamma distribution is given in Eqs. 1 and 2 (Evans et al. 2000):

$$\ln(\lambda) - \psi(\lambda) = \ln\left(\frac{\bar{x}}{g}\right) \tag{1}$$

$$\sigma\lambda = \bar{x} \tag{2}$$

where $\psi(\lambda)$ represents the digamma function and g represents the geometric sample mean and formulated as

$$g = \prod_{i=1}^{n} x_i^{1/n}$$

The symbol λ represents a shape parameter and σ represents a scale parameter.

Determining the maximum likelihood estimators of the log-normal distribution with parameters μ and σ uses the same procedure as the normal distribution. The symbol μ represents a location parameter, and σ represents a scale parameter. Equations 3 and 4 show the MLE formula for log-normal distribution given by Evans et al. (2000).

$$\sigma = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\ln(x_i) \tag{3}$$

$$\mu = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(\ln(x_i) - \sigma)^2 \tag{4}$$

Exceedances and return period

The performance and usability of the derived probability density function (pdf) were evaluated in terms of the predictions of exceedances based on their importance. The actual frequency of exceedances can be determined by simply counting the number of cases exceeding the threshold limit for PM_{10} concentrations, as given by the MAAQG, and dividing this number by the number of available data. Expected exceedances were calculated using data collected from both the actual and previous years (2003).

The frequency distribution of PM_{10} concentrations is very useful in the development of an effective control strategy. When the specific probability function of PM_{10} is known, predicting the required emission reduction, the frequency of exceedances of the threshold limit from MAAQG, and the return period is easy. The return period, $R(X_c)$ defined as the average number of averaging periods or records between exceedances of the given critical concentration, X_c can be calculated as

$$R\,(X_c) = \left(\text{Probability of exceedances}\right)\left(365\text{ days}\right)$$

where the unit of return period is days.

Time series modeling

Time series analysis involves the statistical methodology of the analysis of a sequence of data. The basis of the modeling time series consists of three phases: identification, estimation, and testing (Cooray 2008).

Identification phase

The identification phase consists of plotting the series, examining the main features of the graph, and determining if there is a trend, a seasonal component, any apparent sharp changes in behavior, and any outlying observations. If there are trends and seasonal components, these components are removed to obtain the stationary residuals. To accomplish this task, the application of a preliminary transformation to the data may be necessary. There are several ways in which trend and seasonality can be removed, such as by estimating the components and subtracting them from the data or by differencing the data. For example, Janacek (2001) stated the replacement of the original series $\{x\}$ by $\left\{x_t = x_y - x_{t-d}\right\}$ for some positive integer d. Regardless of the method, the aim is to produce a stationary series. To determine whether the models can represent the data, the most appropriate models are used to fit the residuals using the autocorrelation function (ACF) and partial autocorrelation function (PACF).

The first and the most important step in any time series analysis is to plot the observations against time. This time plot graph (time series plot) shows important features of the series, such as trend, seasonality, outliers, and discontinuities. The plot is vital both to describe the data and to help in formulating a sensible model. In this case, a simplified assumption can be made, such as if the series is stationary or can be made stationary. Stationarity denotes that the series is much the same over all time periods. More specifically, the broad statistical properties are similar if the time origin changes (Janacek 2001). The augmented Dickey–Fuller test (ADF) (Hamilton 1994; Yaffee and McGee 2000; Carmona 2004) can also be used a unit-root test (Carmona 2004). Given observations $x_1, \ldots, x_n$ of the data, the least squares estimates of ϕ_1 and of the variance σ are given in either model by the following:

$$\hat{\phi}_1 = \frac{\sum_{i=1}^{n} x_i\, x_{i-1}}{\sum_{i=1}^{n} x_{i-1}^2}, \tag{5}$$

and

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \hat{\phi}_1\, x_{i-1}\right)^2 \tag{6}$$

where $x_0 = 0$ by convention. Under conditions of residual serial correlation, the test statistic for the ADF test is given in Eq. 7 (Hamilton 1994; Yaffee and McGee 2000; Carmona 2004):

$$DF = \frac{\hat{\phi}_1 - 1}{\sqrt{\hat{\sigma}^2}} \tag{7}$$

The ADF test uses the following regression:

$$ADF = \alpha_0 + \mathrm{p}_1\, y_{t-1} + \sum_{j=2}^{p-1} \beta_j \nabla y_{t-j} + e_t \tag{8}$$

where α_0 is the drift component and e_t is the independent and homogeneous error term. The null hypothesis (H_0) noted as the time series data is non-stationary and can be rejected under two conditions: (1) ADF value is smaller than the ADF critical value and (2) the significance level (p) is smaller or equal to 0.05.

The AR model is implemented in this research. When the value of a series at a current time period is a function of its immediately previous value plus an error, the underlying generating mechanism is called an autoregressive process. For example, for a lag of one time period, the nature of this relationship may be expressed as follows (Yaffee and McGee 2000):

$$\begin{aligned} Y_t &= \phi_1 Y_{t-1} + e_t \\ &= \phi_1 L Y_t + e_t \\ \text{or} & \\ (1 - \phi_1 L)\, Y_t &= e_t \end{aligned} \tag{9}$$

When the output is a regression of the immediately previous output plus an error term, the portion of the previous rating carried over to the rating at time t is designated as ϕ_1. This kind of relationship is called a first-order autoregressive process and is designated as AR (1). The series is regressed upon a previous value of itself plus a random error. However, if the effect is carried over for two time periods, the autoregressive relationship is represented as follows:

$$\begin{aligned} Y_t &= \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t \\ &= \left(\phi_1 L + \phi_2 L^2\right) Y_t + e_t \end{aligned} \tag{10}$$

In this formula, the current value is a function of its two previous ratings, a so-called second-order autoregressive relationship, or AR (2).

The sample autocorrelation function (sample ACF) of the data is an important tool to assess the degree of dependence in the data and to select a data model reflecting this characteristic. If the data are realized values of a stationary time series $\{x_t\}$, then the sample ACF provides an ACF estimate of $\{x_t\}$. For example, a sample ACF that is close to zero for all lags suggests that an appropriate model for the data may be independent and has an identically distributed (iid) error. Note that the iid error theory is perhaps the simplest model for a time series where there is no trend or seasonal component. Moreover, the observations are independent and have iid random variables with zero mean.

Identification of the model can be defined by the lag behavior. A simple concept given by Janacek (2001), as shown in Table 2, can be used as guide. If the model appropriately has an MA(q) process, the autocorrelations are zero after lag q. However, for an AR(p) model, the decay is exponential. For mixed ARMA(p, q) model, the correlations are expected to tail off after lag ($p - q$).

Estimation phase

In order to fit an AR model, the following steps need to be completed. The first step is the computation of the Akaike information criterion (AIC) and Schwartz bayesian criterion (SBC) to determine the model order. The second step is the residual computations to decide if the model fitting is complete or if it needs to be pursued further. Model identification tools such as ACF and PACF are used only for identifying adequate models to represent the monitoring records. For a given data set, when there are multiple adequate models, the selection criterion is normally based on summary statistics from residuals computed from a fitted model or on forecast errors calculated from the sample forecast output. The latter is often accomplished by using the first portion of the series for model construction and the remaining portion as a holdout period for forecast evaluation. AIC and SBC are two model selection criteria based on residuals. Jalil and Mahmud (2009) claimed that the SBC is known as the parsimonious model (a simple model) in selecting the smallest possible lag length, whereas AIC is known for selecting the maximum relevant lag length.

Table 2 Behavior of the auto and partial correlations

Process (Model)	Autocorrelation function (ACF)	Partial autocorrelation function (PACF)
AR(p)	Exponential decay	Zero after lag p
MA(q)	Zero after lag q	Exponentially decay
ARMA(p, q)	Exponential decay after lag ($q - p$)	Decay after lag ($p - q$)

Testing phase

Typically, the goodness-of-fit of a statistical model to a data set is judged by comparing the observed values with the corresponding predicted values obtained from the fitted model. If the fitted model is appropriate, then the residuals behave in a manner consistent with the model. With most statistical models, this is generally performed by referring to the residuals, which are generally defined as follows (Janacek 2001):

$$e_i = y_i - \hat{y}_i \tag{11}$$

where e_i is the residual, y_i is the observed monitoring records, and $\hat{y}_i$ is the predicted monitoring records.

For a univariate time series model, the fitted value is the one-step-ahead forecast so that the residual is the one-step-ahead forecast error. If the model is "good," then the residuals are expected to be "random" and "close to zero," and model validation usually consists of plotting residuals in various ways for verification. Two obvious steps are to plot the residuals as time plot and to calculate the residual correlograms using ACF and PACF. The time plot reveals any outliers and any obvious autocorrelation or cyclic effects. Rather than referring to the residual autocorrelations one at a time, the Jarque–Bera test can simply be performed. The Jarque–Bera test is one of the simplest statistical tests for testing the normality of sufficiently large samples. Jarque–Bera is a test statistic to determine whether the series has normal distribution. The test statistic measures the difference of the skewness and kurtosis of the series with those from the normal distribution (Jarque and Bera 1980). The statistic is computed as:

$$\text{Jarque} - \text{Bera} = \frac{N-k}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \tag{12}$$

where S is the skewness, K is the kurtosis, and k represents the number of estimated coefficients used to create the series.

Under the null hypothesis of a normal distribution, the Jarque–Bera statistic is distributed as χ^2 with a degree of freedom of 2. The reported probability or p value is the probability that a Jarque–Bera statistic exceeds (in absolute value) the observed value under the null hypothesis. The null hypothesis to determine the normality of the model can be rejected when the probability or significance level (p) is smaller or equal to 0.05.

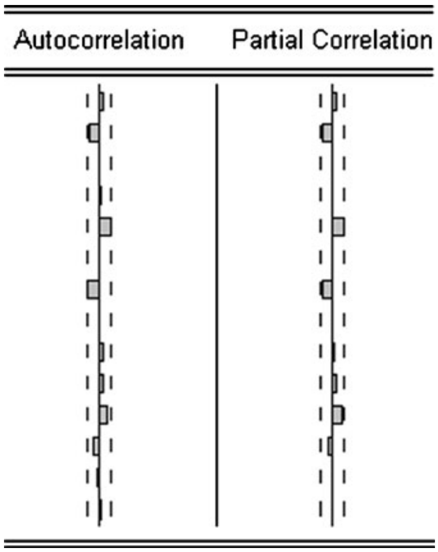


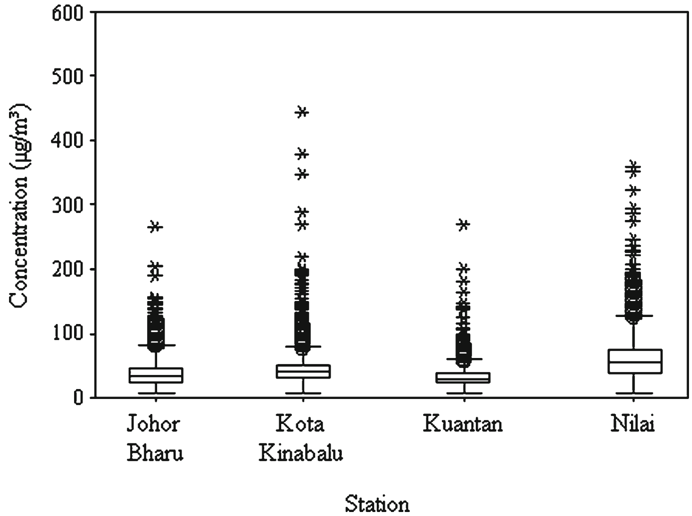
Fig. 2 The sample of ACF and PACF plots

By referring to the ACF and PACF of the residuals, the selected time series model can be confirmed as the best model. The dotted lines in the plots (Fig. 2) of the autocorrelations are the approximated two standard error lines computed as $\pm\frac{2}{\sqrt{T}}$, where T represents the record length (number of points or lags) of the time series being analyzed. If the autocorrelation is within these lines, it is not significantly different from zero at (approximately) the 5% significance level.

Results and discussions

The box plot is a simple graphical display ideal for comparisons. Figure 3 shows the box plot and descriptive statistics for PM_{10} concentrations in 2002 for the four monitoring stations. The mean is sensitive to very large or very small observations. In other words, when the mean shifts toward the direction of skewness in which the mean value is higher than the median, the pollutants data are skewed to the right.

Fig. 3 Box plot and descriptive statistics for 2002



	Johor Bahru	*Kota Kinabalu*	*Kuantan*	*Nilai*
N	8760	8760	8760	8760
Maximum	266	445	270	359
Mean	37.03	44.15	30.74	59.12
Median	34	40	29	55
Std. dev.	19.66	21.85	13.77	28.59
Skewness	1.48	3.42	2.20	1.54
Kurtosis	8.75	34.78	22.14	9.66

The maximum concentrations exceeded the threshold limit for PM_{10} concentration (150 μg/m^3) at all sites, with the highest concentration recorded in Kota Kinabalu at 445 μg/m^3 and the smallest in Johor Bahru at 266 μg/m^3. The highest mean concentration was in Nilai at 59.12 μg/m^3, whereas Kuantan had the smallest mean concentration at 30.74 μg/m^3. The highest mean concentrations were recorded at an industrial area (Nilai), whereas the smallest recorded was at a residential area (Kuantan). In sum, the industrial areas (Johor Bahru and Nilai) showed high values of standard deviations, as expected. These results are likely because of the high particulate events that occurred in the neighboring country, including Kota Kinabalu station. Moreover, Kuantan, located in the residential area, had the smallest standard deviation.

The highest value of skewness for 2002 shows that Kota Kinabalu experienced high particulate events. This is likely due to the haze event that occurred that year as an effect of the transboundary movement of air pollutants emitted from forest fires and open burning activities in Indonesia, which is near East Malaysia, coupled with emissions from local sources. Furthermore, the standard deviation indicates the possibility of extreme concentrations (when standard deviation is high, the variability is likewise high, indicating extreme concentrations). These high concentrations are due to air pollutant sources, such as transboundary pollution, industrial activities, traffic emissions, and open burning near the monitoring stations.

Based on information from DoE Malaysia (2002), several unhealthy air quality days were recorded at various locations. These were caused by the transboundary pollution (land and forest fires) in June to September occurring in Sumatra and Kalimantan, Indonesia, coupled with the southwesterly winds, which contributed to the deterioration of air quality in the West Coast Peninsular Malaysia. Figure 4 shows there are distinct peaks during the southwest monsoon in almost all of the involved stations. Although Johor Bahru and Nilai supposedly experienced wet season during the southwest monsoon (June to September), the spikes during that period show higher

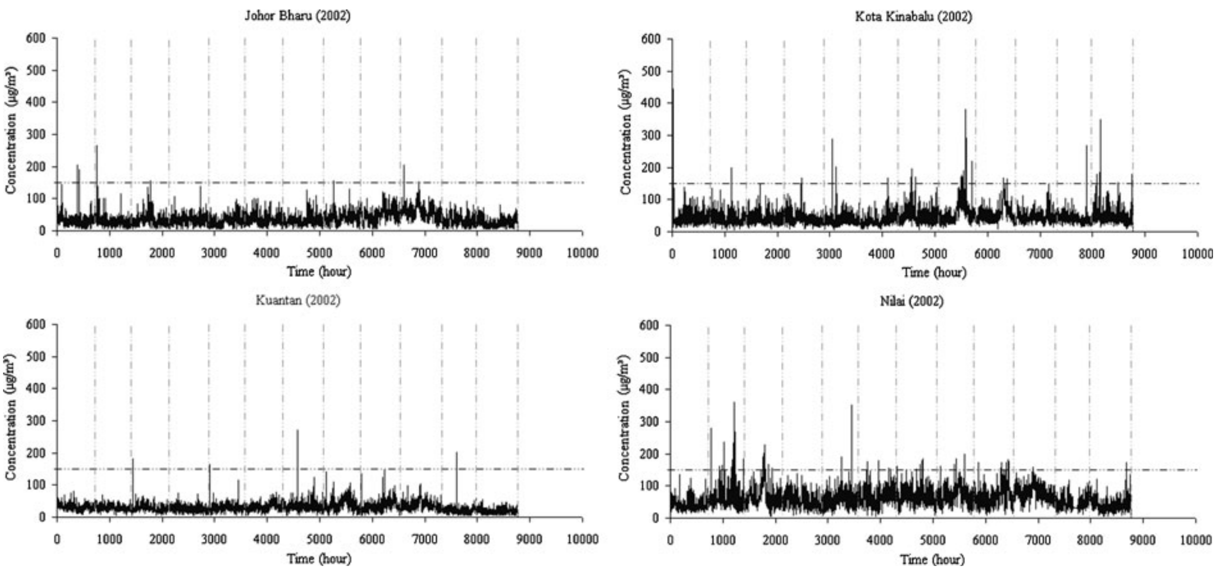


Fig. 4 Time series plots using hourly PM_{10} concentrations during 2002

concentration readings according to the high particulate events in the neighboring country. Similar to the high particulate event in 2002, Fig. 4 shows that some concentration readings exceeded the MAAGQ levels in the residential areas (Kota Kinabalu and Kuantan).

Furthermore, during the monsoon seasons, the wet removal processes are most effective and the aerosol concentration (particulate matter) reaches its minimum. The land remains wet, the sky is generally cloudy, and diurnal variation in temperatures is the lowest (Pillai and Moorthy 2001), which results in the weakening of the continental sources. However, the increased wind speed and its steady onshore direction are favorable for generation and advection of more a greater amount of sea spray aerosol, resulting in the high PM_{10} concentration readings (Pillai et al. 2002).

Probability distributions

The hourly PM_{10} concentrations of the four monitoring stations were divided using gamma and log-normal distributions. Gamma distribution was used for the industrial areas (Johor Bahru and Nilai), whereas the log-normal distribution was used for the residential areas (Kota Kinabalu and Kuantan). The parameters for the distributions were estimated using MLE. The best distribution was selected according to the three types of goodness-of-fit criteria (also known as performance indicators): normalized absolute error (NAE), coefficient of determination (R^2), and root mean square error ($RMSE$). The best distribution was used to determine the prediction of future long-term PM_{10} concentrations at each station. The distribution parameters were estimated using MLE from the provided data of PM_{10} concentrations over a 5-year period. Numerical calculations were implemented using the MATLAB software package to estimate the distribution parameters. The performance indicators (also known as the goodness-of-fit statistics) describe how well the distribution fits a set of observations. Measures of the values typically summarize the discrepancy between observed values and the expected values under the distributions in question.

The first step in such modeling studies is to identify the statistical distribution using the goodness-of-fit. This research used three types of goodness-of-fit criteria to find the best distribution to represent the monitoring stations. The 2002 performance statistics for the distributions at the four monitoring stations are summarized in Table 3.

The indicators must comply with several conditions. When the adequacy measure, such as R^2, is closer to 1, the model is more appropriate to

Table 3 The performance indicators value for four monitoring stations

Station	Performance indicator	Gamma	Log-normal	Best distribution
Johor Bahru	*NAE*	0.0126	0.0482	Gamma
	R^2	0.9896	0.9866	
	RMSE	2.0189	3.4542	
Kota Kinabalu	*NAE*	0.0635	0.0366	Log-normal
	R^2	0.8976	0.9451	
	RMSE	7.2659	5.4872	
Kuantan	*NAE*	0.0252	0.0437	Log-normal
	R^2	0.9551	0.9715	
	RMSE	2.9241	2.5022	
Nilai	*NAE*	0.0136	0.0330	Gamma
	R^2	0.9895	0.9832	
	RMSE	3.7744	3.6195	

simulate the experimental data. However, smaller error measures, such as *NAE* and *RMSE*, indicate that the fitted theoretical distribution is more adequate.

The best distribution representing each monitoring station can be identified according to the goodness-of-fit criteria results (Table 3).

The pdf and cdf graphs were plotted using the parameter values according to the best distribution for each monitoring station, as given in Table 3. Figures 5 and 6 show the plots of the PM_{10} concentrations using the distribution selected through the performance indicators criteria. The pdf plot in Fig. 5 shows the distributions for Johor Bahru and Nilai, which are skewed to the right. This is probably caused by the high particulate events in 2002.

However, the pdf and cdf plots shown in Fig. 5 depict a concentration of PM_{10} that does not exceed the threshold limit given by MAAQG (150 μg/m^3), despite the fact that they represent Johor Bahru and Nilai, monitoring stations that are located in industrial areas. From Fig. 6, both stations skewed to the right for PM_{10} concentrations in 2002. As reported by the DoE Malaysia (2002), Kota Kinabalu did not experience any unhealthy days in 2002, despite the fact that there were haze events that year. Moreover, the air quality in Kota Kinabalu was better by more than 60% during 2002 compared to Kuantan. This explains the long thin tail of the plots indicating the existence of PM_{10} concentrations much higher than the average.

Figure 6 also shows the pdf and cdf plots of the PM_{10} concentrations for Kuantan using log-normal distribution. The tails of the pdf plot tended to the right, indicating that there were high PM_{10} concentrations during 2002; however, these concentrations did not exceed the threshold limit. The pdf plots in Fig. 6 illustrate that the density for 2002 was low; however, the long tail signifies that high particulate events had occurred. However, according to the DoE Malaysia (2002), the air quality in the East Coast of Peninsular Malaysia, including Kuantan, remained generally satisfactory because the mean of the PM_{10} concentrations is 30.74 μg/m^3 (Fig. 6).

Predicted exceedances can be estimated by the probability of PM_{10} concentration higher than the MAAQG level (150 μg/m^3). The predicted exceedances were calculated from the previous year

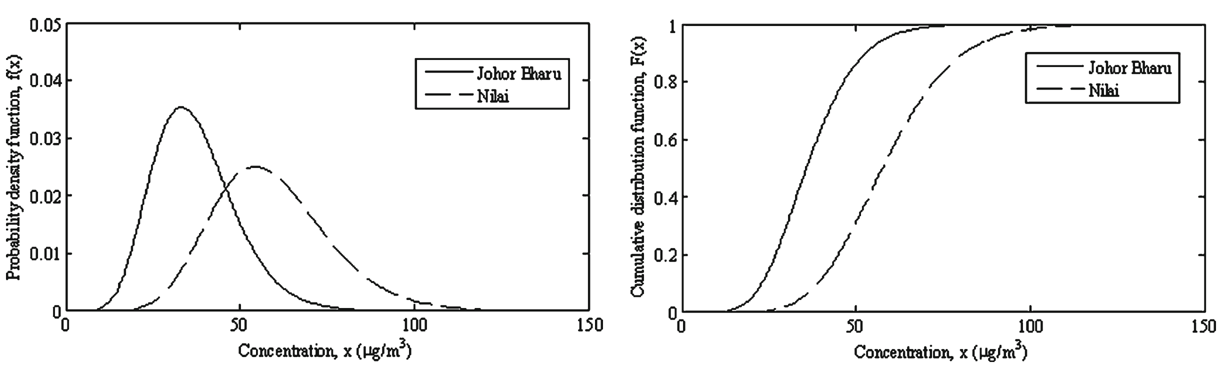


Fig. 5 The pdf and cdf plots for industrial areas using gamma distribution

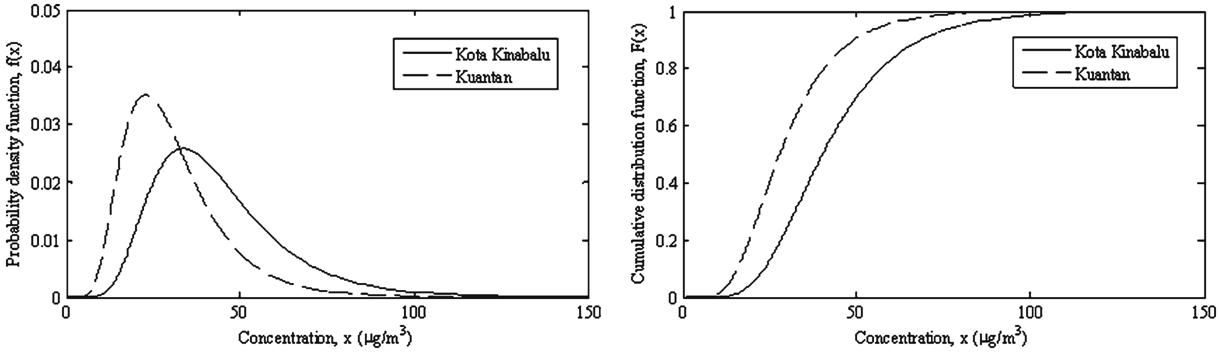


Fig. 6 The pdf and cdf plots for residential areas using log-normal distribution

before being compared to the actual year using the best distributions (Table 4).

Table 4 shows the difference between the predicted and actual exceedances in units of percent for 2003. The exceedances are acceptable, with less than a 1-day difference, for both industrial areas (Johor Bahru and Nilai) and residential areas (Kota Kinabalu and Kuantan) due to the high detection of PM_{10}. DoE Malaysia (2004) reported that the identified main sources of PM_{10} were industrial activities, motor vehicle emissions, and transboundary pollution that occurred during the southwesterly monsoon. Regardless of the transboundary pollution in 2002, Johor Bahru and Nilai stations recorded constant high PM_{10} concentration observations due to intensive industrial activities in those areas. Moreover, Kota Kinabalu and Kuantan showed below a 1-day difference between the actual and predicted exceedances.

Time series modeling

The most important step in time series modeling is to determine the trend and seasonality in the observations. The time series plots show that there were no trends or seasonality to the observations. First, a check for stationarity was made with the time series plot using raw data, which is useful for detecting the lack of stationarity in a time series. The applied method consists of plotting observations against time, as shown in Fig. 7; in order have a first insight into the data structure. Moreover, the ADF unit root test, as shown in Table 4, was used also to identify the stationarity of the data observations. The ADF needs to follow the hypothesis given earlier.

Figure 7 shows the time series plots representing Johor Bahru, Kota Kinabalu, Kuantan, and Nilai in 2002 using the daily average of the PM_{10} concentrations. The time series plot used to check the stationarity shows no trend and seasonality for the observation data in all stations. The plots also show that most of the concentrations were higher than 150 $\mu g/m^3$, which is the acceptable level given by the MAAQG.

In statistics, an ADF is a test for a unit root in a time series sample. The ADF test is used for a larger and more complicated set of time series

Table 4 The predicted and actual exceedences for 2003

Station	Probability of exceedances, Prob $(X \geq 150)$	Predicted exceedances (day)	Actual exceedances (day)	Difference of predicted and actual exceedances (day)
Johor Bahru	0.000125	0.0	0.3	0.3
Kota Kinabalu	0.000866	0.3	1.3	1.0
Kuantan	0.000087	0.0	0.1	0.1
Nilai	0.006121	2.2	1.8	0.4

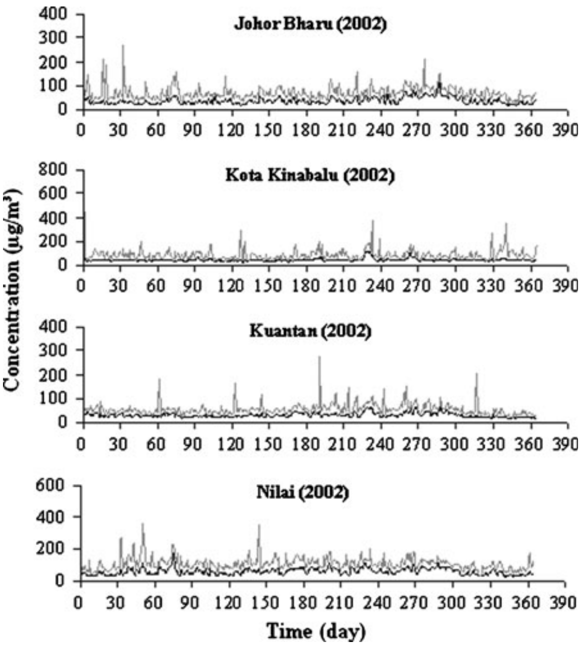


Fig. 7 Time series plot for four stations using daily PM_{10} concentrations. Note: Daily maximum (*gray solid line*), Daily average (*black solid line*)

models with a negative number (Greene 2003). Thus, the smaller the ADF value compared to the ADF critical value, the stronger the rejections of the null hypothesis (H_0). As shown in Table 5, the ADF values for Johor Bahru, Kota Kinabalu, Kuantan, and Nilai stations were smaller compared with the ADF critical values at 1%, 5%, and 10%. Moreover, the significance level (p value) associated with the test statistic gave a smaller value than 0.05 rejecting the null hypothesis.

Thus, the PM_{10} concentrations for Johor Bahru, Kota Kinabalu, Kuantan, and Nilai stations were confirmed to be stationary after the ADF test. Moreover, as stated earlier, the seasonality exhibits variations annually (Chatfield 2004). These variations consume the annual fluctuation that coincide with periods of the year and may be additive or multiplicative. However, this fluctuation is not shown in the time series plots in this research.

After verifying that the series was stationary, the observed data were used to examine the ACF and PACF. The potential models were identified using the ACF and PACF, as shown in Fig. 8, which shows that the best time series model to represent each station is AR(p) with $p = 1$ or ARMA(p, q) with p and $q = 1$, respectively. This can be seen from the given lags in the ACF, which decay exponentially after Lag 1. This is further strengthened with the situation shown by the PACF, in which the PACF approached the zero value quickly after Lag 1.

Generally, the ARMA model order can be determined by examining the decay trends of the ACF and PACF of the stationary series. However, the ACF and PACF values do not always provide a clear indication of the suitability of the selected ARMA model (in this case, ARMA has a p and q equal to 1). As shown in Fig. 8, the decay properties did not possess obvious tail-off patterns within the less lag number. Inspection of the latter graphs suggests that a first-order autoregressive model, [AR (1)] or [ARMA (1, 1)] can be applied to Johor Bahru, Kota Kinabalu, Kuantan, and Nilai, as shown in Fig. 8.

AIC and SBC were used to ensure that the ACF and PACF plots selected the most suitable models. Table 6 shows the best time series models corresponding with the values of AIC and SBC and the model equations for Johor Bahru, Kota Kinabalu, Kuantan, and Nilai.

The time series models were chosen based on the SBC criterion because SBC statistics emphasize parsimony of the model compared to AIC (Christian and Chrisian 2002). In contrast, the simplest model needs to be considered based on

Table 5 The ADF and ADF critical values

Station	ADF	p-value	ADF critical value		
			1%	5%	10%
Johor Bahru	−6.7584				
Kota Kinabalu	−6.2535	< 0.001	−3.4481	−2.8692	−2.5709
Kuantan	−6.1945				
Nilai	−7.5051				

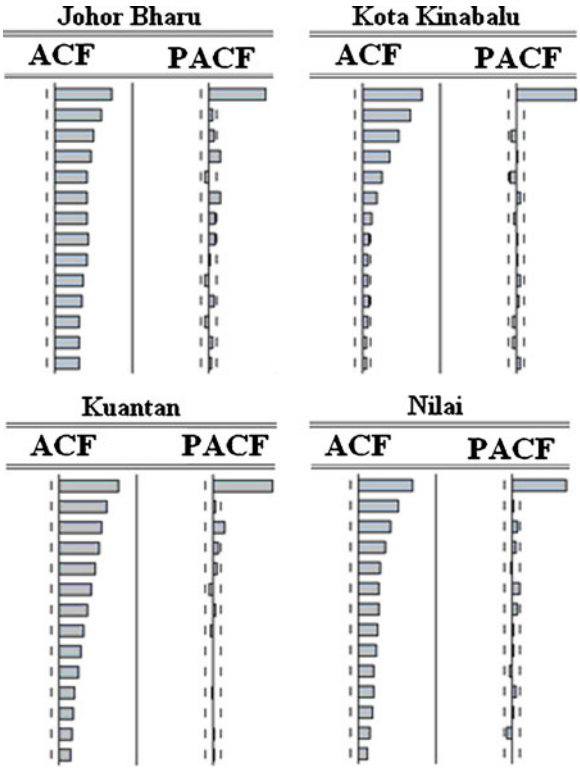


Fig. 8 The ACF and PACF for four monitoring stations

the principle of parsimony, as recommended by Box and Jenkins (1976), who stated that a simpler model with a small order must follow the requirement $p + q \leq 5$. Thus, in this research, models have been fitted such that $p + q \leq 5$, producing 20 possible models, with the best model being chosen based on the smallest AIC and SBC values. Based on the time series analyses, the best time series models are given in Table 6. All stations can be modeled by AR (1).

After selecting the best models using AIC and SBC, the adequacy of the best time series model was checked by referring the residuals using the ACF, PACF, and Jarque–Bera test. Goyal et al. (2006) noted that the relative success of the statistical models for any specific period can be measured in terms of residuals. The study of residual is very important in deciding the adequacy of the statistical model. For the best performance of the model, residuals need to be random and follow normal distribution with zero mean and constant variance.

In the diagnostic phase, the ACF and PACF of the residuals were obtained. A time series is called white noise if the time series is a sequence of independent and identically distributed random variables with finite mean and variance. In particular, if the time series is normally distributed with mean zero and variance, the series is called Gaussian white noise. For a white noise series, all the ACFs are zero. In practice, if all sample ACFs are close to zero, then the series is a white noise series (Ruey 2005).

Figure 9 shows the autocorrelation and partial autocorrelation functions for the residuals in Johor Bahru, Kota Kinabalu, Kuantan, and Nilai. From Fig. 9, approximately 95% of the sample autocorrelations fell between the lines, computed as $\pm\frac{2}{\sqrt{T}}$, wherein T represent the lags. The sample autocorrelations were computed up to Lag 14. If more than one value fell outside the lines, or that one value fells further outside the lines, then the independent and iid hypothesis is rejected. Figure 9 shows that the autocorrelations were significantly different from zero. The correlogram indicates that the residuals of fitted AR (1) model for Johor Bahru, Kota Kinabalu, Kuantan, and Nilai were consistent with the realization of white noise and support the use of the models.

In addition, to check the normality of the models based on null hypothesis (H_0), which was closely related to the normal distribution, the residuals were analyzed using the Jarque–Bera test. Inevitably, Jarque–Bera can be used to rough check for the normality that provides the value

Table 6 The best time series model

Monitoring station	Best time series model	Year	Equation
Johor Bahru	AR (1)	2002	$y_t = 8.56 + 0.77y_{t-1} + \varepsilon_t$
Kota Kinabalu	AR (1)	2002	$y_t = 8.35 + 0.81y_{t-1} + \varepsilon_t$
Kuantan	AR (1)	2002	$y_t = 6.51 + 0.79y_{t-1} + \varepsilon_t$
Nilai	AR (1)	2002	$y_t = 16.08 + 0.73y_{t-1} + \varepsilon_t$

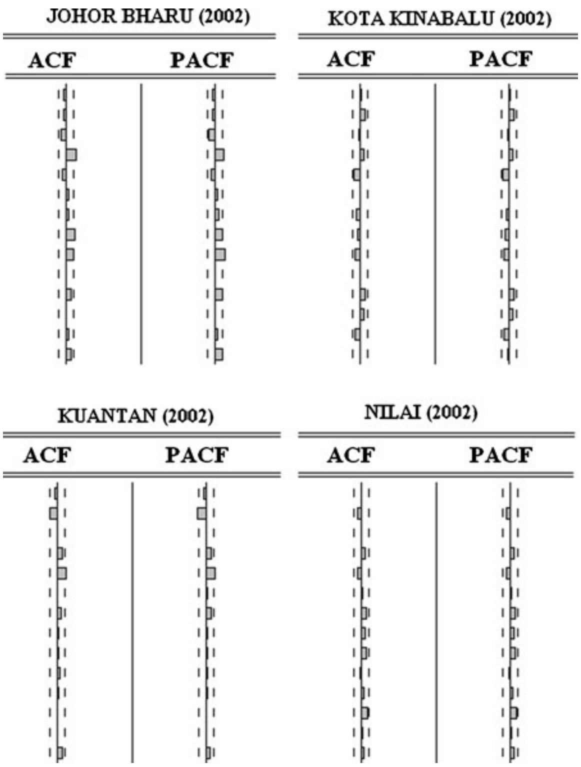


Fig. 9 The ACF and PACF using residuals for four monitoring stations

of probability (known as *p* value). Results of the Jarque–Bera measure the difference of the skewness and kurtosis of the series compared to those from the normal distribution. The *p* values, which were computed from Jarque–Bera test, were more than 0.5 for all sites. Normally in this case, a *p* value more than 0.05 intervals is taken as a strong evidence for the rejection of the null hypothesis.

Performance indicators such as *NAE*, R^2, and *RMSE* were used to verify the best time series models. These statistical indicators have been applied by several researchers (Ryan 1995; Comrie 1997; Wilks 1995; Goyal et al. 2006) and have been used to provide a general indication of the relationship between the observed and the predicted data. Fourteen-day predicted observations with the actual observations were used in the statistical analyses for verification of the best time series model.

The performance statistics for the time series models for four monitoring stations are summarized in Table 7. The performance measures

Table 7 Model performance indicators

Station	Performance indicators		
	NAE	R^2	RMSE
Johor Bahru	0.0814	0.8234	2.9028
Kota Kinabalu	0.0513	0.8280	3.1908
Kuantan	0.1084	0.8512	2.2749
Nilai	0.0903	0.8235	4.4965

reflect that the AR (1) model for Johor Bahru, Kota Kinabalu, Kuantan, and Nilai gave satisfactory results based on the imposed conditions.

Numerous studies on time series, especially using AR, MA, and ARMA models, in air pollutant modeling have been conducted, such as those by Portnov et al. (2009), Liang et al. (2009), Chelani and Devotta (2006), Chattopadhyay and Chattopadhyay (2009), Ballester et al. (2002), and Abdel-Aziz and Frey (2003). With regard to future prediction of air pollutant concentrations, these previous studies have only focused on other variables such as sulphur dioxide (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO), and ozone (O_3), rather than PM_{10}. Moreover, unlike this research, which predicts for daily concentrations 14 days ahead, most of the studies on predicting future air pollutant concentrations using AR, MA, and ARMA models predict only the hourly concentrations. A 14-day prediction of the PM_{10} concentrations is important to help the related responsible agencies to ensure the prevention and curb the adverse impact of PM_{10} on human health as early as possible.

Conclusion

Based on the results of the analysis carried out for PM_{10} concentrations in two types of monitoring stations (i.e., industrial (Johor Bahru and Nilai) and residential (Kota Kinabalu and Kuantan)), all stations experienced high particulate events in 2002. Two parent distributions in air pollution, namely, gamma and log-normal distributions were used to fit the PM_{10} concentrations using the MLE method. Results show that the gamma distribution is the best distribution to represent the industrial areas, whereas log-normal distribution is the best distribution to represent the residential areas.

Prediction of the PM_{10} exceedances (in unit of days) for the year ahead (2003) were estimated using the best distributions and were compared to the actual exceedances. Most of the differences between the predicted and actual values occurred because of the high particulate events in 2003; thus, the prediction made for the next year may have been made less accurate due to the occurrence of unexpected circumstances. The monitoring records of PM_{10} concentrations were used to find the best time series models to represent each involved station. Results indicate that the most simple time series model, AR (1), is the best model to represent all stations.

Acknowledgements This study was funded by the Universiti Sains Malaysia under the ST Grant No. 6013404. We express our appreciation to the Universiti Sains Malaysia and the Fellowship Scheme for providing the financial support to carry out this study. We also thank the Department of Environment of Malaysia for their support.

References

Abas, M. R., Oros, D. R., & Simoneit, B. R. T. (2004). Biomass burning as the main source of organic aerosol particulate matter in Malaysia during haze episodes. *Chemosphere, 55*, 1089–1095.

Abdel-Aziz, A., & Frey, H. C. (2003). Development of hourly probabilistic utility NO_x emission inventories using time series techniques: part I—univariate approach. *Atmospheric Environment, 37*, 5379–5389.

Afroz, R., Hassan, M. N., & Ibrahim, N. A. (2003). Review air pollution and health impacts in Malaysia. *Environmental Research, 92*(2), 71–77.

Awang, M. B., Jaafar, A. B., Abdullah, A. M., Ismail, M. B., Hassan, M. N., Abdullah, R., et al. (2000). Air quality in Malaysia: Impacts, management issues and future challenges. *Respiratory, 5*, 183–196.

Ballester, E. B., iValls, G. C., Carrasco-Rodriguez, J. L., Olivas, E. S., & Valle-Tascon, S. D. (2002). Effective 1-day ahead prediction of hourly surface ozone concentrations in Eastern Spain using linear models and neural networks. *Ecological Modelling, 156*(1), 27–41.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis—Forecasting and control revised*. San Francisco: Holden-Day Inc.

Carmona, R. A. (2004). *Statistical analysis of financial data in S-plus*. U.S.A.: Springer.

Comrie, A. C. (1997). Comparing neural network and regression models for RSPM forecasting. *Journal of Air and Waste Management Association, 47*, 513–529.

Chatfield, C. (2004). *The analysis of time series—An introduction* (6th Edn.). Florida: Chapman & Hall/CRC.

Chattopadhyay, G., & Chattopadhyay, S. (2009). Autoregressive forecast of monthly total ozone concentration—A neurocomputing approach. *Computer and Geosciences, 35*(9), 1925–1932.

Chelani, A. B., & Devotta, S. (2006). Air quality forecasting using a hybrid autoregressive and nonlinear model. *Atmospheric Environment, 40*, 1774–1780.

Christian, H., & Chrisian, H. (2002). Index forecasting and model selection. *International Journal of Intelligent Systems in Accountly, Finance and Management, 11*, 119–135.

Cooray, T. M. J. A. (2008). *Applied time series analysis and forecasting*. Oxford: Alpha Science International Ltd.

Department of Environment (DoE), Malaysia (2002). *Malaysia environmental quality report 2002*. Kuala Lumpur: Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia.

Department of Environment (DoE), Malaysia (2004). *Malaysia environmental quality report 2004*. Kuala Lumpur: Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia.

Department of Environment (DoE), Malaysia (2008). *Malaysia environmental quality report 2008*. Kuala Lumpur: Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia.

Evans, M., Hastings, N., & Peacock, B. (2000). *Statistical distributions* (3rd Edn.). New York: Wiley.

Fitri, M. D. N. F., Ramli, N. A., Yahaya, A. S., Sansuddin, N., Ghazali, N. A., & AlMadhoun, W. (2009). Monsoonal differences and probability distribution of PM_{10} concentration. *Environmental Monitoring Assessment, 163*(1–4), 655–667.

Fuller, D. O., & Murphy, K. (2006). The ENSO-fire dynamic in insular Southeast Asia. *Climatic Change, 74*(4), 435–455.

Ghazali, N. A., Ramli, N. A., Yahaya, A. S., Md Yusof, N. F. F., Sansuddin, N., & Al Madhoun, W. A. (2010). Transformation of nitrogen dioxide into ozone and prediction of ozone concentrations using multiple linear regression techniques. *Environmental Monitoring Assessment, 165*(1–4), 475–489.

Gilbert, O. R. (1987). *Statistical method for environmental pollution monitoring*. New York: Van Nostrand Reinhold Company Inc.

Goyal, P., Chan, A. T., & Jaiswal, N. (2006). Statistical models for the prediction of respirable suspended particulate matter in urban cities. *Atmospheric Environment, 40*, 2068–2077.

Greene, W. H. (2003). *Econometric analysis* (5th Edn.). New Jersey: Prentice Hall.

Hamilton, J. (1994). *Time series analysis*. New Jersey: Princeton University Press.

Hyer, E. J., & Chew, B. N. (2010). Aerosol transport model evaluation of an extreme smoke episode in Southeast Asia. *Atmospheric Environment, 44*(11), 1422–1427.

Jalil, A., & Mahmud, S. F. (2009). Enviroment Kuznets curve for CO_2 emissions—A cointegration analysis for China. *Energy Policy, 37*(12), 5167–5172.

Janacek, G. (2001). *Practical time series*. London: Arnold.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, heteroscedasticity and serial independence of regression residuals. *Economics Letters, 6*, 255–259.

Juneng, L., Latif, M. T., Tangang, F. T., & Mansor, H. (2009). Spatio-temporal characteristics of PM_{10} concentration across Malaysia. *Atmospheric Environment, 43*(30), 4584–4594.

Kim Oanh, N. T., Upadhyay, N., Zhuang, Y. H., Hao, Z. P., Murthy, D. V. S., Lestari, P., et al. (2006). Particulate air pollution in six Asian cities—Spatial and temporal distributions and associated sources. *Atmospheric Environment, 40*, 3367–3380.

Liang, W. M., Wei, H. Y., & Kuo, H. W. (2009). Association between daily mortality from respiratory and cardiovascular diseases and air pollution in Taiwan. *Environmental Research, 109*(1), 51–58.

Lu, H. C., & Fang, G. C. (2003). Predicting the exceedances of critical PM_{10} concentration—A case study in Taiwan. *Atmospheric Environment, 37*, 3491–3499.

Mage, D. T., & Ott, W. R. (1984). An evaluation of the method of fractiles moments and maximum likelihood for estimating parameters when sampling air quality data from a stationary lognormal distribution. *Atmospheric Environment, 18*, 163–171.

Malaysian Meteorological Department (2008) *Official website of Malaysia Meteorological Department (MMD): General Climate of Malaysia*. Available online at http://www.met.gov.my/english/education/climate/climated.html, accessed on 1 February 2008.

Pillai, P. S., & Moorthy, K. K. (2001). Aerosol mass—Size distributions at a tropical coastal station. Response to mesoscale and synoptic processes. *Atmospheric Environment, 35*, 4099–4122.

Pillai, P. S., Babu, S. S., & Moorthy, K. K. (2002). A study of PM, PM_{10} and $PM_{2.5}$ concentration at a tropical coastal station. *Atmospheric Research, 61*, 149–167.

Portnov, B. A., Dubnov, J., & Barchana, M. (2009). Studying the association between air pollution and lung cancer incidence in a large metropolitan area using a Kernel density function. *Socio-Economic Planning Sciences, 43*(3), 141–150.

Ramli, N. A., & Ibrahim, W. H. W. (2003). PM_{10} concentration measurements at four selected sites in Semenanjung Malaysia: A comparison between sites with different background. In *Proceeding environment 2003, Pulau Pinang* (pp. 236–238).

Ramli, N. A., Wathern, P., Ibrahim, W. W. H., & Hamidi, A. A. (2001). Some observations on respirable particulate matter during 1997 haze event in Malaysia. In H. L. Koh, & Y. A. Hassan (Eds.), *Proceeding of ecological and environmental modelling 2001, 3–4 September 2001* (pp. 226–234). USM Publisher. ISBN 983–861–245–6.

Ruey, S. T. (2005). *Analysis of financial time series* (2nd Edn.). Hoboken: Wiley.

Ryan, W. (1995). Forecasting severe ozone episodes in the Baltimore metropolitan area. *Atmospheric Environment, 29*, 2387–2398.

Sedek, J. N. M., Ramli, N. A., & Yahaya, A. S. (2006). Air quality predictions using lognormal distribution functions of particulate matter in Kuala Lumpur, Malaysia. *Journal of Environmental Management, 7*, 33–41.

Singh, K. P., Bartolucci, A. A., & Bae, S. (2001). Mathematical modeling of environmental data. *Journal of Mathematical and Computer Modeling, 33*, 793–800.

Soleiman, A., Othman, M., Samah, A. A., Sulaiman, N. M., & Radojevic, M. (2003). The occurrence of haze in Malaysia—A case study in an urban industrial area. *Pure and Applied Geophysics, 160*, 221–238.

Wang, Y., & Zhang, Y. S. (2009). Air quality assessment by contingent valuation in Ji'nan, China. *Journal of Environmental Management, 90*(2), 1022–1029.

Wilks, D. (1995). *Statistical methods in the atmospheric sciences—An introduction*. New York: Academic.

Yaffee, R. A., & McGee, M. (2000). *Introduction to time series analysis and forecasting with applications of SAS and SPSS*. New York: Academic.