Extreme Value Distribution for Prediction of Future PM\textsubscript{10} Exceedences

Noor Faizah Fitri Md Yusof, Nor Azam Ramli, Ahmad Shukri Yahaya
Clean Air Research Group
Environmental and Sustainable Development Section
School of Civil Engineering, Universiti Sains Malaysia
Engineering Campus, 14300 Nibong Tebal, Pulau Pinang, Malaysia.
noorfaizah@eng.usm.my

Abstract—Central fitting distribution (CFD) such as Weibull, gamma and lognormal distribution can give a good result for fitting the mean concentration of air pollutants data. However, it cannot precisely fit the high concentration region. Therefore, extreme value distributions (EVD) that are Gumbel and Frechet distributions were used in this research to fit the high particulate event in Seberang Perai, Penang from 2002 to 2006 to reduce the predicting error. The cfd (Weibull, gamma and lognormal distributions) and evd (Frechet and Gumbel distributions) were used to fit the daily maximum concentration. The best distribution that can fit the data was selected based on performance indicators. Furthermore, the exceedences of a critical PM\textsubscript{10} concentration over the Malaysian Ambient Air Quality Guidelines were estimated using the best distributions. The results of performance indicators show that the extreme value distribution gives better fit to the actual high PM\textsubscript{10} concentration than the central fitting distribution. The exceedences over a high particulate event were successfully predicted. In 2002, the exceedences is 291 days, 224 days in 2003, 151 days in 2004, 156 days in 2005 and 9 days in 2006.

Keywords—central fitting distribution; Gumbel distribution; Frechet distribution; method of moments; daily maximum concentration

I. INTRODUCTION

In recent years, statistical analysis and probability distributions have been used widely in the analysis of air pollution data to understand the current situation of air quality and to predict future air quality. There are many types of probability distributions that have been used to fit air pollutant data, such as the Weibull distribution [1,2], the gamma distribution [3,4] and the lognormal distribution [2,5]. Weibull, gamma and lognormal distributions can give a good result for fitting the mean concentration of air pollutants data. However, it cannot precisely fit the pollutants data when the concentration is high. Extreme value distribution is usually used for fitting the high concentration of air pollutants data.

The extreme value theory (EVT) which is used in storm, flood, wind, sea waves, and earthquake estimation, dates back to the pioneering works by Frechet in 1927 and Fisher and Tippet in 1928 [6]. This theory was extensively developed by Gumbel in 1958 following the extremal type theorem originated by Genedenko in 1943. The EVT concerns probability calculations and the statistical inference associated with the extreme values of random processes [7].

The EVT has also been widely used in wind speed [9,10], health [7,11] and air pollution studies [11,3].

Gumbel distribution was applied to fit CO data in India in order to predict violations of air quality standards at urban road intersections [8]. The results showed that the Gumbel distribution gave satisfactory performance for predictions of extreme air pollution events. In addition, the extreme value theory was successfully used to fit the monthly maximum data and high concentration data of air pollutant concentration over a specific percentile in China [3]. Furthermore, research conducted in Switzerland to model indoor radon distributions in using EVT showed that the EVT is relevant in areas characterized by high mean concentrations, while lognormal distributions seem to be more relevant in small or medium concentration areas [12].

Therefore, the aim of this study is to compare between the central fitting distribution (CFD) that is the Weibull, gamma and lognormal distributions with the extreme value distributions (EVD) that are Gumbel and Frechet distributions.

II. EXPERIMENTAL SET UP

A. Area of Study

The station selected for this research is Seberang Perai (SP). SP is situated in the north part of Peninsular Malaysia and is categorized as an industrial area. Therefore, PM10 concentration is expected to originate mostly from the industrial emission as well as the vehicles emission. Penang is a small state in Peninsular Malaysia but the estimated population density is high. For every square kilometer, there are 1274 inhabitants in Penang. Fig. 1 displays the location of the monitoring site in SP and its description is as in Table 1.
B. The Data

The air quality monitoring stations in Malaysia are strategically located in residential, urban, and industrial areas to detect any significant change in the air quality which may be harmful to human health and the environment. SP station is located in a heavily industrialized area of Seberang Perai. The samples of PM10 were collected by using continuous particulate monitor BAM 1020 (Met One Instruments, Inc.). This instrument automatically measures and records hourly PM10 concentration levels (in milligrams or micrograms per cubic meter) using the industry proven principle of beta-ray attenuation. The data recorded are regularly subjected to standard quality control processes and quality assurance procedures by the Department of Environment (DoE), Malaysia. In order to achieve the aim of this study, the maximum daily PM10 concentrations were selected for year 2002 until 2006. Therefore, the total number of data for 1 year is 365.

C. The Methodology

Fig. 2 illustrates the flow to obtain the best distribution that can represent PM10 daily maximum concentration data. First, the input data was prepared by selecting the maximum concentration for each day in year 2002 to 2006. This data were then used to be fitted with the cfd and evd. For cfd, Weibull, gamma and lognormal distribution were applied, meanwhile for evd, Gumbel and Frechet distribution were used. Method of moments (MoM) was used to estimate parameters for the cfd and evd. In order to select the best distribution that can fit well the input data, performance indicators (PI) that are mean absolute error (MAE), normalized absolute error (NAE), prediction accuracy (PA), coefficient of determination (R²) and index of agreement (IA) were used. For MAE and NAE, values that are closer to zero indicate the best distribution. Conversely, for PA, R² and IA, values closer to 1 indicate the best distribution. Result of PI for cfd and evd were compared and the best distribution was selected. Finally, the predicted exceedences were estimated with using the best distribution.

The two parameters Weibull, gamma and lognormal cumulative distribution function (cdf) and probability density function (pdf) with parameters $\alpha$ and $\beta$ is given in Table 2 [3]. The distributions parameters $\alpha$ and $\beta$ were estimated with using the method of moments (MoM).

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|l|l|}
\hline
Site & Location & Coordinate & State Area (km$^2$) & Population Density (inhabitant/km$^2$) \\
\hline
Seberang Perai, Penang & Industrial area & N 05° 23.4704 E 100° 23.1977 & 1,031 & 1274 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|l|l|l|}
\hline
Distribution & Cdf equations, $F(x)$ & Pdf equations, $f(x)$ \\
\hline
Weibull & $1 - \exp \left[ - \left( \frac{x}{\beta} \right)^\alpha \right], x > 0, \alpha > 0, \beta > 0$ & $\frac{\alpha}{\beta} \left( \frac{x}{\beta} \right)^{\alpha-1} \exp \left[ - \left( \frac{x}{\beta} \right)^\alpha \right], x > 0, \alpha > 0, \beta > 0$ \\
\hline
Gamma & $\int_0^\infty \frac{1}{\beta^\alpha \Gamma(\alpha)} \left( \frac{x}{\beta} \right)^{\alpha-1} \exp \left[ - \frac{x}{\beta} \right] dx, x \geq 0, \alpha > 0, \beta > 0$ & $\frac{1}{\beta^\alpha \Gamma(\alpha)} \left( \frac{x}{\beta} \right)^{\alpha-1} \exp \left[ - \frac{x}{\beta} \right], x \geq 0, \alpha > 0, \beta > 0$ \\
\hline
Lognormal & $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\ln x-\beta} e^{-\frac{(x-e^x)^2}{2\alpha^2}} dx, x > 0, \alpha > 0, \beta > 0$ & $\frac{1}{\alpha x \sqrt{2\pi}} \exp \left[ - \frac{1}{2} \left( \frac{\ln(x) - \beta}{\alpha} \right)^2 \right], x > 0, \alpha > 0, \beta > 0$ \\
\hline
\end{tabular}
\end{table}

A. The MoM for Weibull Distribution

The average or expectation of a function of a random variable \( x \) can be found by weighting the function by its density or mass function. This procedure is called the method of moments [6]. Since the basic Weibull model has two parameters, estimation of the parameters can be obtained using the sample mean and sample variance.

Using the expression for the mean and variance, \( \beta \) was obtained by the solution of Equation (1).

\[
\frac{s^2}{\bar{x}} = \frac{\Gamma\left(1 + \frac{2}{\beta}\right)}{\Gamma\left(1 + \frac{1}{\beta}\right)^2} - 1
\]

\( \alpha \) is then calculated by the following equation:

\[
\alpha = \frac{\bar{x}}{\Gamma\left(1 + \frac{1}{\beta}\right)}
\]

B. The MoM for Gamma Distribution

The equations for this method are as follows [13]:

\[
\beta = \frac{1}{cv^2}
\]

Where \( cv = \frac{s}{\bar{x}} \) (the coefficient of variation)

\( \alpha \) is the solution of:

\[
\alpha \beta = \bar{x}
\]

The values of \( \alpha \) and \( \beta \) in this distribution can also be calculated by Equation (5) and (6).

\[
\alpha = \frac{s^2}{\bar{x}}
\]

\[
\beta = \left(\frac{\bar{x}}{s}\right)^2
\]

C. The MoM for Lognormal Distribution

In this method, \( \alpha \) and \( \beta \) were obtained directly from Equation (7) and (8), [13].

\[
\alpha = \sqrt{\ln(s^2 + (\bar{x})^2) - 2 \ln(\bar{x})}
\]

\[
\beta = \ln(\bar{x}) - \frac{\alpha^2}{2}
\]

IV. THE EXTREME VALUE DISTRIBUTION (EVD)

A. The Gumbel Distribution

The Gumbel distribution was extensively developed and applied to flood flows by Gumbel in 1954 and 1958. This distribution results from any underlying distribution of the xi’s of the exponential type [6].

The probability density functions for the Gumbel distribution is as follows [6]:

\[
f(x) = \frac{1}{\beta} \exp\left[\frac{-x - \delta}{\beta} - \exp\left(\frac{-x - \delta}{\beta}\right)\right], \quad -\infty < x < \infty, -\infty < \delta < \infty, \beta > 0
\]

The cumulative distribution function for the Gumbel distribution is as follows;

\[
F(x) = \exp\left[-\exp\left(\frac{x - \delta}{\beta}\right)\right], \quad x > 0, -\infty < \delta < \infty, \beta > 0
\]

The location parameter, \( \delta \) is the mode of the distribution [6],

\[
\frac{df(x)}{dx} = 0 \text{ for } x = \delta
\]

The parameter \( \beta \) is a measure of dispersion, and it only depends on the variance of \( X_{max} \). The moment generating function is found to be;

\[
M_{X_{max}}(x) = \exp(\delta x) \Gamma \left(1 - \beta x\right), \ x < 1/ \beta
\]

Therefore, the mean, \( \text{E}(X_{max}) \) and variance, \( \text{Var}(X_{max}) \) of \( X_{max} \) are as follows;

\[
\text{E}(X_{max}) = \mu = \delta + n_e \beta
\]

\[
\text{Var}(X_{max}) = \sigma^2 = \frac{\pi^2 \beta^2}{6}
\]

Where \( ne = 0.5772 \) (Euler constant)

As a result, from Equation (13) and (14), \( \beta \) and \( \delta \) were obtained by the following equations;

\[
\beta = \frac{\sqrt{6}}{\pi} \sigma
\]

and,

\[
\delta = \mu - n_e \beta = \mu - \frac{n_e \sqrt{6}}{\pi} \sigma
\]

B. The Frechet Distribution

The Frechet distribution was first developed and applied to flood flows by Frechet in 1927. The probability density function of the Frechet distribution is as follows [6];

\[
f(x) = \frac{\alpha}{\beta} \left(\frac{\beta}{x}\right)^{\alpha+1} \exp\left[-\left(\frac{\beta}{x}\right)^{\alpha}\right], \quad x > 0, \alpha > 0, \beta > 0
\]

The cdf form of the Frechet distribution is as shown below;
The scale and shape parameters ($\alpha$ and $\beta$) in the Frechet distribution was also estimated by using the method of moment (MoM).

In this method, the coefficient of variation needs to be identified first. The coefficient of variation ($cv$) is the ratio of the sample standard deviation to the sample mean. The formula to estimate the standard deviation and sample mean are as in Equation (19) and (20), respectively;

Standard deviation,

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}}$$  \hspace{1cm} (19)

Mean, $$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n}$$  \hspace{1cm} (20)

cv = $\frac{s}{\bar{x}}$

$\alpha$ is obtained by the following equation;

$$cv = \frac{1}{\sqrt{\Gamma\left(1 - \frac{2}{\alpha}\right)}} - 1, \alpha > 2$$  \hspace{1cm} (21)

$\beta$ is the solution of;

$$\beta = \frac{\bar{x}}{\Gamma\left(1 - \frac{1}{\alpha}\right)}$$  \hspace{1cm} (22)

V. RESULTS AND DISCUSSION

A. Time Series Plot

Fig. 3 shows the time series plot of daily maximum PM10 concentration at the study site for year 2002 to 2006 and Table 3 shows the descriptive statistics. Malaysia that is located at the equatorial, experience uniformed temperature, high humidity and copious rainfall. Changes of wind flow patterns determine the seasons in this country. The wind over the country is generally light and variable. However, there are some uniform periodic changes in the wind flow patterns that describe the four seasons experienced by the country namely, northeast monsoon (November to March), transitional period (April to May), southwest monsoon (June to September), and another transitional period (October to November). PM10 concentration during southwest monsoon is expected to be high in the area of study as the effect of dry weather condition. Transboundary sources aggravate this situation. A study conducted on chemical characterization of the haze in Brunei had found that particulate matter was the dominant pollutants during haze episodes [14]. In the case of haze event in 1997, the particles come from biomass burning in Indonesia for clearing vegetated (forest and grassland) areas. The wild fires significantly increase the input of organic aerosol components to the atmosphere [15]. Therefore PM10 concentration is high during haze episodes.

| TABLE 3 THE DESCRIPTIVE STATISTICS FOR DAILY MAXIMUM OF PM10 CONCENTRATION AT SEBERANG PERAI FOR 2002 TO 2006 |
|---|---|---|---|---|---|
| Descriptive Statistics | 2002 | 2003 | 2004 | 2005 | 2006 |
| N | 365 | 365 | 365 | 365 | 365 |
| Mean | 128 | 145 | 161 | 144 | 80 |
| Median | 122 | 139 | 159 | 154 | 74 |
| Std. Deviation | 41 | 37 | 40 | 71 | 28 |
| Variance | 1682 | 1375 | 1639 | 5105 | 787 |
| Minimum | 64 | 17 | 68 | 31 | 41 |
| Maximum | 540 | 362 | 421 | 425 | 222 |
| Range | 476 | 345 | 353 | 394 | 181 |

The maximum concentration was recorded during the southwest monsoon (Jun to September) in 2002, 2005 and 2006 as indicated in Fig. 2. In 2003 and 2004, the maximum concentration occurs in the month of February and January, respectively, which falls under the northeast monsoon. However, Table 4 explains that the maximum mean concentration occurs either during the southwest monsoon or during the transition period. Therefore, the maximum concentration in 2003 and 2004 does not represent the overall situation in that year and this might be because of outliers due to ad hoc event, e.g. open burning.

The time series plot for 2005 shows a sudden decrease of PM10 concentration caused by relocation of monitoring site by the DoE Malaysia. The new station is located at about 2 km radius from the old station. The mean for PM10 daily maximum concentration in 2005 for January until July is 196 µg/m³, and from August to December is only 72 µg/m³. The effect of relocation of the monitoring stations is clearly seen after 2005 when PM10 concentration was reduced significantly in 2006. The ranges of PM10 daily maximum concentration from 2002 to 2005 were between 345µg/m³ to 476µg/m³, but in 2006 the range had been reduced to only 181µg/m³.

| TABLE 4 THE MONSOONAL MEAN OF DAILY MAXIMUM PM10 CONCENTRATION |
|---|---|---|---|---|---|
| MONSOON | Year |
|---|---|---|---|---|---|
| 1 | 2002 | 2003 | 2004 | 2005 | 2006 |
| 1 | 118.34 | 133.38 | 151.76 | 142.43 | 72.38 |
| 2 | 122.18 | 141.66 | 161.13 | 175.67 | 74.68 |
| 3 | 141.98 | 160.60 | 169.46 | 150.23 | 89.88 |
| 4 | 132.94 | 144.58 | 178.00 | 62.35 | 90.52 |

*1 Northeast monsoon, 2 Transition (Apr-May), 3 Southwest monsoon, 4 Transition (Oct)
Fig. 3 The time series plot for daily maximum of PM$_{10}$ concentration at Seberang Perai for 2002 to 2006

Table 5 Parameters for Weibull, Gamma, Lognormal and Extreme Value Distribution Using MoM

<table>
<thead>
<tr>
<th>Parameters</th>
<th>CFD</th>
<th>EVD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weibull</td>
<td>Gamma</td>
</tr>
<tr>
<td>2002</td>
<td>3.46</td>
<td>142</td>
</tr>
<tr>
<td>2003</td>
<td>4.42</td>
<td>159</td>
</tr>
<tr>
<td>2004</td>
<td>4.53</td>
<td>177</td>
</tr>
<tr>
<td>2005</td>
<td>2.12</td>
<td>162</td>
</tr>
<tr>
<td>2006</td>
<td>3.13</td>
<td>89.6</td>
</tr>
</tbody>
</table>

TABLE 5 PARAMETERS FOR WEIBULL, GAMMA, LOGNORMAL AND EXTREME VALUE DISTRIBUTION USING MoM
Starting from August 2005, the daily maximum concentrations decrease when the station was transferred to a new site. This is also because of relocation of sampling site. Furthermore, it is obvious that the decreasing trend in the data are not the indication of overall air quality of the study area. However, it does not indicate that air quality in the study area is improving. The reason for this is the data that were used is the daily maximum data, not the daily average data. Hence, the maximum concentration might occur only once a day, due to ad hoc event such as open burning. Therefore, the exceedences or number of days that PM10 concentration is more than MAAQG is 291 days. The probability of exceedences for 2003 is 0.6148 \((F(x>150) = 0.6148)\) and the number of days that exceed MAAQG is 224 days. For 2004, the probability of exceedences is 0.4153 \((F(x>150) = 0.4153)\) with 151 days exceed MAAQG.

For 2005 and 2006, the probability of exceedences is 0.4286 \((F(x>150) = 0.4286)\) and 0.0254 \((F(x>150) = 0.0254)\) respectively. Thus, the number of days that exceed MAAQG is 156 days for 2005 and 9 days for 2006.

The exceedences obtained in this study show a decreasing trend. However, it does not indicate that air quality in the study area is improving. The reason for this is the data that were used is the daily maximum data, not the daily average data. Hence, the maximum concentration might occur only once a day, due to ad hoc event such as open burning. Hence, the data are not the indication of overall air quality of the study area. Furthermore, it is obvious that the decreasing trend is also because of relocation of sampling site.

pdf plots by using method of moment for EVD show almost similar distribution with long tail to the right (Fig. 5). This pattern indicates that there are dominant sources that contribute to high PM10 concentration in Seberang Prai and increasing with time. However, this is not peculiar as it is known that Seberang Prai is a heavily industrialized area and this sector is believed to have become a major contributor to air pollution problem.

### Table 6 Goodness of fit criteria

<table>
<thead>
<tr>
<th>Year</th>
<th>Distributions</th>
<th>MAE</th>
<th>NAE</th>
<th>PA</th>
<th>(R^2)</th>
<th>IA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>CFD</td>
<td>9.1733</td>
<td>0.0716</td>
<td>0.8898</td>
<td>0.7874</td>
<td>0.9417</td>
</tr>
<tr>
<td></td>
<td>EVD</td>
<td>4.9803</td>
<td>0.0389</td>
<td>0.9431</td>
<td>0.8845</td>
<td>0.9706</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.8310</td>
<td>0.0377</td>
<td>0.9625</td>
<td>0.9213</td>
<td>0.9793</td>
</tr>
<tr>
<td>2003</td>
<td>CFD</td>
<td>6.4169</td>
<td>0.0443</td>
<td>0.9568</td>
<td>0.9104</td>
<td>0.9779</td>
</tr>
<tr>
<td></td>
<td>EVD</td>
<td>3.0576</td>
<td>0.0211</td>
<td>0.9827</td>
<td>0.9604</td>
<td>0.9912</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.2677</td>
<td>0.0157</td>
<td>0.9871</td>
<td>0.9691</td>
<td>0.9934</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.0950</td>
<td>0.0145</td>
<td>0.9889</td>
<td>0.9726</td>
<td>0.9943</td>
</tr>
<tr>
<td>2004</td>
<td>CFD</td>
<td>5.4840</td>
<td>0.0379</td>
<td>0.9776</td>
<td>0.9505</td>
<td>0.9876</td>
</tr>
<tr>
<td></td>
<td>EVD</td>
<td>9.3759</td>
<td>0.0581</td>
<td>0.9629</td>
<td>0.9221</td>
<td>0.9763</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.8737</td>
<td>0.0172</td>
<td>0.9825</td>
<td>0.9601</td>
<td>0.9906</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.5561</td>
<td>0.0158</td>
<td>0.9870</td>
<td>0.9689</td>
<td>0.9932</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.8473</td>
<td>0.0176</td>
<td>0.9889</td>
<td>0.9725</td>
<td>0.9943</td>
</tr>
<tr>
<td>2005</td>
<td>CFD</td>
<td>6.1556</td>
<td>0.0381</td>
<td>0.9776</td>
<td>0.9505</td>
<td>0.9877</td>
</tr>
<tr>
<td></td>
<td>EVD</td>
<td>13.4575</td>
<td>0.0936</td>
<td>0.9780</td>
<td>0.9512</td>
<td>0.9889</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15.9995</td>
<td>0.1113</td>
<td>0.9683</td>
<td>0.9324</td>
<td>0.9838</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19.9595</td>
<td>0.1388</td>
<td>0.9505</td>
<td>0.8984</td>
<td>0.9744</td>
</tr>
<tr>
<td>2006</td>
<td>CFD</td>
<td>3.0751</td>
<td>0.0384</td>
<td>0.9858</td>
<td>0.9776</td>
<td>0.9877</td>
</tr>
<tr>
<td></td>
<td>EVD</td>
<td>7.0631</td>
<td>0.0881</td>
<td>0.9303</td>
<td>0.8607</td>
<td>0.9639</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.7665</td>
<td>0.0583</td>
<td>0.9682</td>
<td>0.9323</td>
<td>0.9838</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.3117</td>
<td>0.0413</td>
<td>0.9829</td>
<td>0.9608</td>
<td>0.9912</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.0801</td>
<td>0.0384</td>
<td>0.9846</td>
<td>0.9641</td>
<td>0.9921</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.5359</td>
<td>0.0316</td>
<td>0.9945</td>
<td>0.9836</td>
<td>0.9953</td>
</tr>
</tbody>
</table>

B. Parameters for CFD and EVD

Table 5 displays parameters obtained for CFD and EVD using MoM. The best distribution that can fit the daily maximum concentration was selected based on goodness of fit criteria in Table 6.

From the result of goodness of fit with using mean absolute error (MAE), normalised absolute error (NAE), Prediction Accuracy (PA), coefficient of determination \((R^2)\) and Index of Agreement (IA) in Table 6, it is clear that the EVD fits the high concentration better than the CFD, except for 2005. For MAE and NAE, value closer to zero indicates a better fit with the Gumbel distribution. It shows that the EVD can fit the actual data very well except for the year 2005 where the best distribution is Weibull. This is due to inconsistency of data recorded when Seberang Perai station was relocated as mentioned in 3.1 The Time Series Plot. PM10 concentration was estimated. For the year 2002, the probability for PM10 concentration more than 150\(\mu g/m^3\) is equal to 0.7973 \((F(x>150) = 0.7973)\). Therefore, the exceedences or number of days that PM10 concentration is more than MAAQG is 291 days. The probability of exceedences for 2003 is 0.6148 \((F(x>150) = 0.6148)\) and the number of days that exceed MAAQG is 224 days. For 2004, the probability of exceedences is 0.4153 \((F(x>150) = 0.4153)\) with 151 days exceed the MAAQG.

For 2005 and 2006, the probability of exceedences is 0.4286 \((F(x>150) = 0.4286)\) and 0.0254 \((F(x>150) = 0.0254)\) respectively. Thus, the number of days that exceed MAAQG is 156 days for 2005 and 9 days for 2006.

The exceedences obtained in this study show a decreasing trend. However, it does not indicate that air quality in the study area is improving. The reason for this is the data that were used is the daily maximum data, not the daily average data. Thus, the maximum concentration might occur only once a day, due to ad hoc event such as open burning. Hence, the data are not the indication of overall air quality of the study area. Furthermore, it is obvious that the decreasing trend is also because of relocation of sampling site.

C. cdf, pdf and Probability of Exceedences

Fig. 4 illustrates the cdf plot for 2002 to 2006 with using the best distributions. It shows that the EVD can fit the actual data very well except for the year 2005 where the best distribution is Weibull. This is due to inconsistency of data recorded when Seberang Perai station was relocated as mentioned in 3.1 The Time Series Plot. PM10 concentration decrease when the station was transferred to a new site. Starting from August 2005, the daily maximum concentrations exceed MAAQG only twice (Fig. 3).

From Fig. 4, the exceedences of PM10 concentration that is more than the Malaysian Ambient Air Quality Guideline (MAAQG) was estimated. For the year 2002, the probability
Fig. 4 cdf plots for 2002 to 2006

Cumulative Distribution Function, F(x)

PM$_{10}$ concentration ($\mu$g/m$^3$)

--- Theoretical distribution

--- Observed data
VI. CONCLUSIONS

In environment, particularly air pollution, researchers are more concern on high pollutants concentration because it can affect human health as well as the ecosystem. This study proved that EVD gives better fit than the CFD for the daily maximum PM10 concentration from 2002 to 2006 except for 2005 when there is inconsistency of data recorded due to relocation of sampling site. Therefore, the prediction for future air quality for high PM10 concentration was more accurate. It was found out that the exceedences or number of days when PM10 concentration is over the MAAQG for 2002 is 291 days, 2003 is 224 days, 2004 is 151 days, 2005 is 156 days and 2006 is 9 days.

REFERENCES


