Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-020-01784-2

ORIGINAL PAPER

Risk assessment of extreme air pollution based on partial duration series: IDF approach

Nurulkamal Masseran[1] · Muhammad Aslam Mohd Safari[1]

Abstract
The occurrences of extreme pollution events have serious effects on human health, environmental ecosystems, and the national economy. To gain a better understanding of this issue, risk assessments on the behavior of these events must be effectively designed to anticipate the likelihood of their occurrence. In this study, we propose using the intensity–duration–frequency (IDF) technique to describe the relationship of pollution intensity (i) to its duration (d) and return period (T). As a case study, we used data from the city of Klang, Malaysia. The construction of IDF curves involves a process of determining a partial duration series of an extreme pollution event. Based on PDS data, a generalized Pareto distribution (GPD) is used to represent its probabilistic behaviors. The estimated return period and IDF curves for pollution intensities corresponding to various return periods are determined based on the fitted GPD model. The results reveal that pollution intensities in Klang tend to increase with increases in the length of time between return periods. Although the IDF curves show different magnitudes for different return periods, all the curves show similar increasing trends. In fact, longer return periods are associated with higher estimates of pollution intensity. Based on the study results, we can conclude that the IDF approach provides a good basis for decision-makers to evaluate the expected risk of future extreme pollution events.

Keywords Extreme pollution event · Environmental modeling · Peak over threshold · Pollution risk assessment

1 Introduction

Air pollution endures as one of the most critical problems faced by the world. In Urban areas particularly, air pollution is closely associated with many economic activities, the tremendous number of vehicles, various forms of ongoing infrastructure development, and high population densities mean that the problem of air pollution is getting worse every year and has become alarming in its effect on human health and environmental ecosystems (Gulia et al. 2015; Kumar et al. 2013; Xu et al. 2016). In fact, as economic growth and the steady urbanization of rural areas continues throughout the world, issues related to air quality must be addressed, controlled, and well managed to maintain a healthy and sustainable ecosystem (Jayasooriya et al. 2017; Kumar et al. 2015; Masseran 2017; Zhang et al. 2016).

To gain a better understanding of the problems related to air pollution, it is necessary to correctly observe, monitor, and analyze the behaviors of pollutant variables. By doing so, information about unhealthy or extreme pollution events can be evaluated and anticipated as a basis for air-quality risk management. In Malaysia, the Department of Environment (DOE) has recorded data and supervised the development of the air pollutant index (API) to provide measures of air quality at particular times. The API includes information about five major pollutant variables, including carbon monoxide (CO), ozone (O_3), particulate matter less than 10 μm in size (PM_{10}), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2). Figure 1 shows a schematic of the DOE data recorded hourly from which the API is determined, based on the highest value of these five pollutants (Al-Dhurafi et al. 2018a, b). The DOE classifies several API values that provide information about air quality and its associated health effects, as shown in

✉ Nurulkamal Masseran
kamalmsn@ukm.edu.my

[1] Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

Published online: 05 March 2020

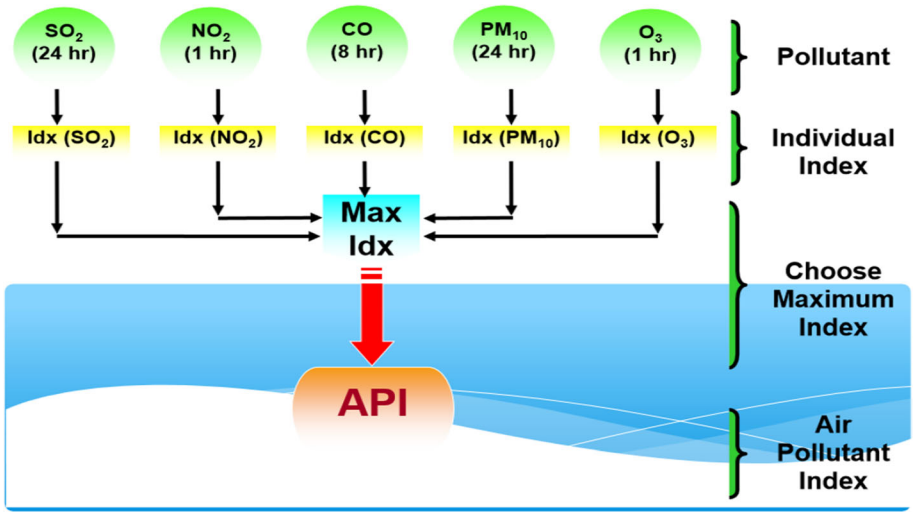

Fig. 1 API determination at a particular time (Department of Environment 1997)

Table 1 API values corresponding to air quality and health effects

Pollution index	Status	Health effect
0–50	Good	Low pollution with no ill effect on health
51–100	Moderate	Moderate pollution that poses no ill effect on health
101–200	Unhealthy	Worsens the health conditions of high-risk individuals with heart and lung complications
201–300	Very unhealthy	Worsens the health conditions and lowers tolerances to physical exercise in individuals with heart and lung complications. Affects public health
> 300	Hazardous	Hazardous to high-risk individuals and public health in general

Table 1 (Alyousifi et al. 2018; Department of Environment 1997).

As shown in Table 1, higher API values indicate poorer air quality, and a high API value also indicates the occurrence of an extreme pollution event. Along with a high API index, extreme pollution events can negatively affect human health and disrupt the economic activities and environmental ecosystems of the affected countries. For example, in 2005 and 2013, the government of Malaysia declared a state of emergency in several areas due to extreme API values, specifically due to the occurrence of haze events (Sahani et al. 2014). Thus, risk assessment regarding the behavior of extreme pollution events should be designed to reflect the likelihood or probability of these events. To achieve this goal, the method known as intensity–duration–frequency (IDF) can be used to describe the relationship of pollution intensity (i) to the duration (d) and return period (T) (Mohymont et al. 2004; Van de Vyver 2015; Willems 2000).

IDF is a popular statistical approach used as a risk assessment tool for extreme events. Generally, an IDF curve can be obtained using two different approaches, i.e., the annual maximum series (AMS) and partial duration series (PDS) methods. The AMS approach is based on the maximum value of a series each year (annual extreme event), so only one extreme event per year will be used to construct the IDF curve. This approach leads to the loss of information contained in other large-sample values in a given period (Li et al. 2014; Pickands 1975). In fact, as reported by Xia et al. (2012), hazardous phenomena are not only those that are maximum in effect, but include those that are second or third largest that may lead to extreme events for different time periods within a year. Since the AMS approach selects only one maximum point per year, it reduces the quality of the data analyzed with respect to events of interest. In addition, since the AMS approach focuses only on the probability of annual exceedances, it cannot be used to determine the probability of multiple extreme events occurring within a year (Vrban et al. 2018).

To overcome the limitations of the AMS method, the other approach is the peak-over-threshold (POT) method, which retains for use all peak values that "exceed" a certain threshold level u. In frequency analysis, the POT method is also known as a PDS. The POT method allows for the consideration of a more rational selection of events as "extreme," which enables the inclusion of a wider range

of extreme events in the analysis (Lang et al. 1999). In fact, a wider range of selected observations as extreme events will provide more precise inferences, particularly with respect to the accuracies of the parameters and quantile estimates (Khaliq et al. 2006). For example, Vrban et al. (2018) show that the use of PDS provides expected outcomes that have 4–10% greater intensities than those provided by the AMS approach. Thus, the authors conclude that the PDS approach has additional merits and they recommend its use rather than the AMS. In addition, Karim et al. (2017) found the PDS to provide better frequency estimates than AMS, particularly for the case of risk assessments on small or medium disaster events such as floods. Thus, in this study, rather than AMS, we used the PDS to construct IDF curves to describe the behaviors of extreme events based on air pollution data. Several authors have described the importance of the information about the intensity, duration and frequency on air pollution event, for example, see Dale et al. (2001), Lui et al. (2016) and Yoo et al. (2014). However, no proper statistical approach has been used to analyze the data of intensity, duration and frequency on air pollution event. Thus, by adopting the technique of IDF curve using POT approach, we believe it could be a valuable knowledge for the researchers in the field of air pollution studies.

2 Study area and data

Klang, which is located in Peninsular Malaysia at a latitude of 101° 26′ 44.023 E and longitude of 3° 2′ 41.701 N, is one of Malaysia's large cities, with a land area of approximately 573 km^2 and a dense population. Klang is the home of important industrial and economic interests for Malaysia as it is a center for import and export activities operating out of Port Klang. In fact, Klang has been recognized as the 13th busiest trans-shipment port and the 16th busiest container port in the world (Al-Dhurafi et al. 2018c). However, the rapid development of Klang with respect to urban commercial and industrial areas in recent decades has elevated its risk of atmospheric pollution (Azmi et al. 2010). Therefore, given the importance of its industrial activities, it is crucial to monitor and evaluate API behaviors for Klang. Figure 2 show the map of Peninsular Malaysia with the Klang location (Google 2019).

In this study, as data, we used the hourly API index for the period January 1, 1997 to December 31, 2016, which has a small percentage of missing values in a random pattern. To estimate these missing values, we used the single imputation method based on the average of the last known and next known observations. This method is easy to implement and is reported to provide a good result for random missing data (Masseran et al. 2013).

3 PDS based on POT approach

The distribution of rare events can usually be well described by the use of a model based on the extreme value theory, particularly when the topic of interest is the risk of extreme events (Reiss and Thomas 2007). In this study, our focus is the POT of extreme pollution events. As described above, the API provides an indicator of the level of air pollution at a particular time. High API values indicate occurrences of extreme air pollution events. Let $Y_1, Y_2, \ldots, Y_n$ represent independent and identically distributed (iid) random variables of the hourly API. The distribution of API data is governed by some unknown density function F. As shown in Table 1, a high API value that exceeds the unhealthy level ($API \geq 100$) is considered to be a pollution event. Mathematically, this phenomenon could represent a conditional event that is larger than some threshold u, and its conditional exceedance distribution function (cedf) $F^{[u]}$ can be written as follows:

$$\begin{aligned} F^{[u]}(y) &= \Pr(Y \leq y | Y > u) \\ &= \frac{\Pr\{Y \leq x, Y > u\}}{\Pr\{Y > u\}} \\ &= \frac{F(y) - F(u)}{1 - F(u)}, \quad y \geq u \end{aligned} \tag{1}$$

Equation (1) is also known as a conditional excess distribution function with the left endpoint of $F^{[u]}$, i.e., $\alpha\left(F^{[u]}\right) = \inf\left(y : F^{[u]}(y) > 0\right)$ equal to 0. In this study, a specific threshold $u = 100$, which correspond to the occurrence of an unhealthy API is the event of interest to be evaluated. Let $X = Y - u$ represent API data that are higher than the threshold u. Thus, k data points that exceed the threshold u, as represented by the vector $\mathbf{X} = (x_1, x_2, \ldots x_k)$, will have the empirical cumulative density function (ecdf) $\hat{F}_k(\mathbf{x}; .)$, as given by the following equation:

$$\hat{F}_k(\mathbf{x}; .) = \left(\hat{F}_k(\mathbf{x}; .)\right)^{[u]} \tag{2}$$

The ecdf in Eq. (2) provides the same information as the cedf as in Eq. (1) (Reiss and Thomas 2007). The parametric form of the ecdf in Eq. (2) can be determined from a limiting distribution of the normalized values exceeding the threshold. As that threshold approaches the endpoint of the variable, the ecdf in Eq. (2) will follow GPD (Pickands 1975; Ribatet 2007). The GPD defines the probability for a given random variable $X = Y - u$, where X is the event that exceeds the threshold u. The distribution function for the GPD is given as follows:

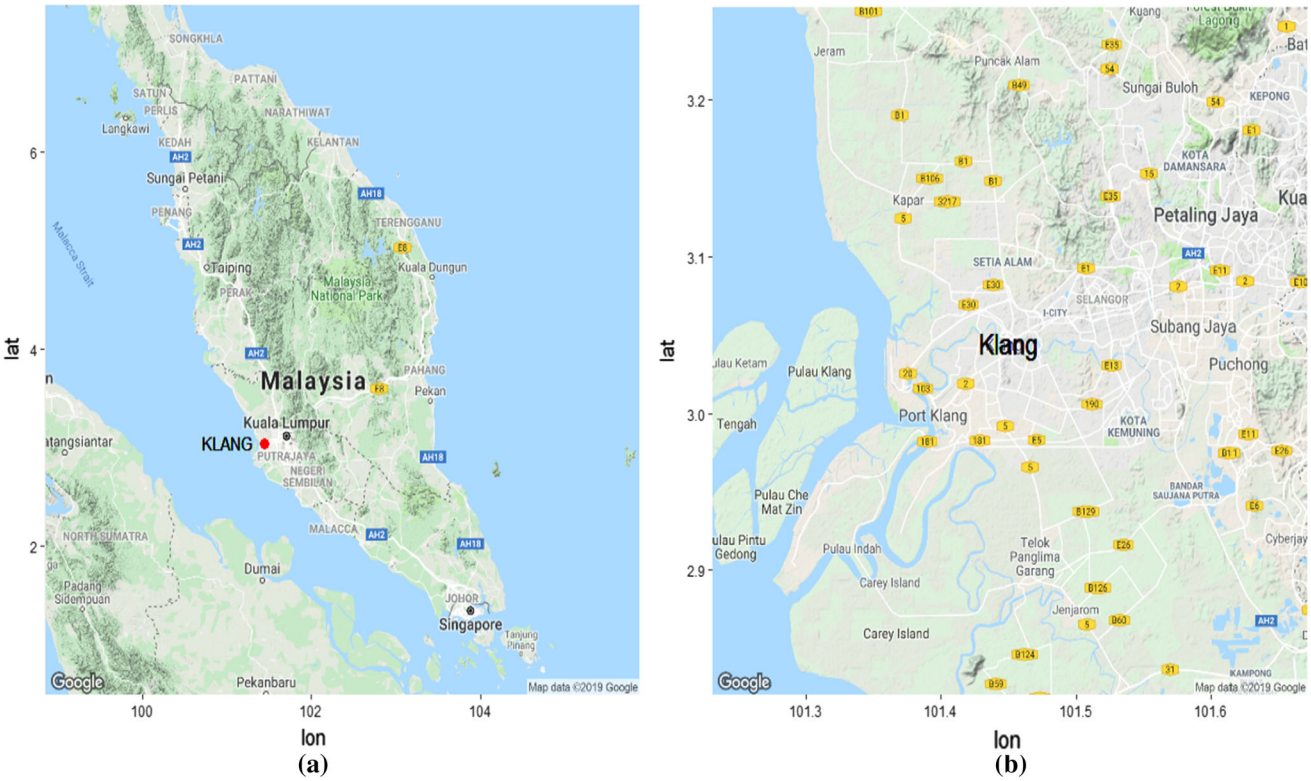

Fig. 2 **a** Map of Peninsular Malaysia with the Klang location identified by a red dot. **b** Map of Klang. *Source*: Google (2019)

$$F(y) = P(Y \le y | Y > u) = \begin{cases} 1 - \left(1 + \xi\left(\frac{y-u}{\sigma}\right)\right)^{-1/\xi} & if \quad \xi \neq 0 \\ 1 - e^{-\left(\frac{y-u}{\sigma}\right)} & if \quad \xi = 0 \end{cases} \tag{3}$$

which is equivalent to the following:

$$G(x) = P(X \le x) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-1/\xi} & if \quad \xi \neq 0 \\ 1 - e^{-\frac{x}{\sigma}} & if \quad \xi = 0 \end{cases} \tag{4}$$

where $x \ge 0, \quad 1 + \frac{\xi y}{\sigma} > 0$. The terms u, ξ, and σ are threshold, shape, and scale parameters, respectively (Southworth and Heffernan 2014). The shape parameter ξ describes the tail behavior of the distribution for extreme data. For $\xi = 0$, the GPD exhibits the properties of a medium-sized tail, which approximates an exponential distribution. The GDP exhibits the properties of a short-tail distribution if $\xi > 0$, which approximates a Pareto type-II model. Otherwise, the GPD exhibits the properties of long-tailed behavior if $\xi < 0$ and approximates an ordinary Pareto distribution (Masseran et al. 2016). In fact, Ben-Zvi (2009) mentioned that the value of shape parameter significantly reflect the magnitudes of predicted rare events.

Next, to estimate the GPD parameter, Husler et al. (2011) proposed a solution based on the maximum likelihood and goodness-of-fit approach. The log-likelihood function of the GPD model can be written as follows:

$$\log(L) = \sum_{i=1}^{k} \left(\log\left(\theta/\xi\right) - \left(1 + \frac{1}{\xi}\right) \log(1 + \theta x_i) \right) \tag{5}$$

where $\theta = \xi/\sigma$. The parameter θ in Eq. (5) can be estimated using the following equation:

$$\frac{\sum_{i=1}^{k} \hat{H}(x_i)}{k} = \frac{1}{2} \tag{6}$$

where $\hat{H}(x_i) = 1 - (1 + \theta x_i)^{-\frac{1}{\xi}}$ with the condition that $\hat{H}(x_i)$ can be roughly regarded as uniformly distributed on [0, 1]. Then, the estimators for both (θ, ξ) are given by the following equations:

$$\hat{\xi} = \frac{\sum_{i=1}^{k} \log\left(1 + \hat{\theta} x_i\right)}{k} \tag{7}$$

$$\frac{\sum_{i=1}^{k} \left(1 + \hat{\theta} x_i\right)^{-\frac{1}{\hat{\xi}}}}{k} = \frac{1}{2} \tag{8}$$

However, as we can see, there is no analytical solution provided by the estimators in Eqs. (7) and (8). Thus, an iterative procedure must be used to obtain a final solution for parameters $\hat{\theta}$ and $\hat{\xi}$.

In addition, it is important to consider the return period with respect to the degree of the API intensity as it relates to extreme pollution events. The return period indicates roughly how frequently pollution intensities higher than unhealthy levels have occurred in the past. This information can be used to predict how likely extreme pollution events are to occur in the future. However, the interpretation of return period differs in the AMS and PDF approaches. For PDS, a return period is the average period between extreme events that exceed a threshold u within a particular period of time (Vrban et al. 2018), which is derived based on knowledge of a distribution function, which is expressed as follows (Coles 2001):

$$P(Y > y|Y > u) = \left[1 + \xi\left(\frac{y-u}{\sigma}\right)\right]^{-\frac{1}{\xi}} \tag{9}$$

Let $\zeta_u = P(Y > u) = \frac{k}{n}$, with k being the number of data points y_i that exceed the threshold u. Then, Eq. (9) can be simplified as follows:

$$P(Y > y) = \zeta_u\left[1 + \xi\left(\frac{y-u}{\sigma}\right)\right]^{-\frac{1}{\xi}} \tag{10}$$

Therefore, we can compute the API level that was exceeded on average once every m series of observations as follows:

$$\frac{1}{m} = \zeta_u\left[1 + \xi\left(\frac{y_u-u}{\sigma}\right)\right]^{-\frac{1}{\xi}} \tag{11}$$

Specifically, for $u = 100$, which is the threshold for an unhealthy air pollution event, Eq. (11) can be simplified as follows:

$$\hat{y}_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right] \tag{12}$$

To obtain the expected the return level that corresponds to a particular return period, the return period formula can be easily manipulated (Zhou et al. 2012), as shown in the following equation:

$$u = G^{-1}\left(1 - \frac{1}{P_R(u)}\right) \tag{13}$$

where G^{-1} is the quantile function of the GPD model. Apart from that, the efficiency and robustness of the expected return level can be evaluated based on its standard error or confidence interval (CI) which derived by a delta method given as

$$Var(\hat{y}_m) \approx \nabla y_m^T V \nabla y_m \tag{14}$$

where ∇y_m^T obtained as

$$\begin{aligned}\nabla y_m^T &= \left[\frac{\delta y_m}{\delta \zeta_u}, \frac{\delta y_m}{\delta \sigma}, \frac{\delta y_m}{\delta \xi}\right] \\ &= \left\{\sigma m^{\xi}\zeta_u^{\xi-1}, \xi^{-1}\left[(m\zeta_u)^{\xi} - 1\right], -\sigma\xi^{-2}\left[(m\zeta_u)^{\xi} - 1\right]\right. \\ &\quad \left. + \sigma\xi^{-1}(m\zeta_u)^{\xi}\log(m\zeta_u)\right\}\end{aligned} \tag{15}$$

while the variance–covariance matrix of V is obtained as

$$V = \begin{bmatrix} \frac{\zeta_u(1-\zeta_u)}{n} & 0 & 0 \\ 0 & v_{1,1} & v_{1,2} \\ 0 & v_{2,1} & v_{2,2} \end{bmatrix} \tag{16}$$

where $v_{i,j}$ denoted the covariance term for parameters σ and ξ (Coles 2001; Li et al. 2014). The value of the parameters σ and ξ are estimated using Eqs. (7) and (8), while the value of ζ_u is equal to $P(Y > u) = \frac{k}{n}$, with k being the number of data points y_i that exceed the threshold u. Then, the 95% CI is given as

$$\hat{y}_m \pm 2\sqrt{Var(\hat{y}_m)} \tag{17}$$

In order to ensure the efficiency of the estimated return levels, its values should be within the 95% CI which described by Eq. (17). The smaller CI will provide a less uncertainties on the expected return level in which implies a better efficiency and robustness.

4 Declustering of POT events

Pollution events always occur in sequences over time. An extreme air pollution event is likely to be followed by another of several hours or days (Gyarmati-Szabo et al. 2017; Zidek et al. 2005). Thus, API values that exceed the unhealthy threshold u will occur in clusters. This scenario violates the assumption of independence required by the GPD model. To address this problem, we used a declustering technique to filter dependent consecutive API values that exceed the threshold u. Declustering is performed by selecting only the POT/PDS data with r-minimum separation (run length) from each other. Thus, the maximum excess POT/PDS data in each cluster of pollution events can be identified and the selected data identified as cluster maxima will at least exhibit approximately independent behavior. As suggested by Karim et al. (2017), r = 240 h are sufficient for use as the minimum number of separation hours between two extreme events.

5 Adopting the IDF relationship for extreme pollution based on POT approach

The IDF relationship is typically represented as a curve with the duration as the abscissa and intensity as the ordinate, with a series of curves, one for each return period. Based on the IDF curves, we can obtain the expected intensity of pollution (i) and its frequency (T) and duration (d). Generally, this relationship can be described by the mathematical function given in the following equation:

$$i(d,T) = \frac{a(T)}{b(d)} \tag{18}$$

The IDF relationship described in Eq. (18) has the advantage of a separable functional dependence of i on T and d. The general formula for the function of $b(d)$ is $b(d) = (d+\theta)^{\eta}$. The function of the numerator term $a(T)$ must be determined based on a suitable probability model that can effectively describe the behavior of maximum intensities $I(d)$ in the data. Thus, its general function can be written as $a(T) = G_X^{-1}\left(1 - {}^1\!/_T\right)$, where G_X^{-1} is the quantile function of probability model. Simply stated, the random variable of intensity, $I(d)$, which represents the extreme pollution data derived from certain duration hours d, will follow some particular distribution model $G_{I(d)}(i,d)$ (Koutsoyiannis et al. 1998). In fact, based on Eq. (18), the distribution of $a(T) \approx I(d)b(d)$ is actually the intensity rescaled by the term $b(d)$, which can be expressed as follows:

$$\begin{aligned} P[I(d) \le i] &= P[I(d)b(d) \le ib(d)] \\ &= P(X \le x) \end{aligned} \tag{19}$$

Therefore, to express the IDF relationship, for each $[t_j, t_j + d]$ time interval, the average intensity of the API over the interval d will be computed as follows:

$$I_j(d) = \frac{API_j(d)}{d} \tag{20}$$

where $API_j(d)$ is the sum of API values available for the time interval d. In this study, as time intervals, we used six different hours of duration, d = 1, 2, 4, 6, 12, and 24 to determine the average pollution intensity $I_j(d)$. Next, POT events of extreme pollution are determined by the average intensity value that exceed the unhealthy threshold $u = 100$, which is given as $I(d) = I_j(d) - 100$.

Based on the IDF curve, key information that must be obtained are the return pollution-intensity level of $I(d)$ that corresponds to the return period T and duration d, which is denoted as $i_T(d)$ (Van de Vyver 2015). As noted above, for any pollution event with duration d, $G_{I(d)}(i,d) = P\{I(d) \le i\}$ indicates the probability distribution of $I(d)$. Thus, the return period T associated with the return pollution-intensity level $i_T(d)$ can be expressed as follows:

$$T = \frac{1}{1 - G_{I(d)}(i_T; d)} \tag{21}$$

where $G_{I(d)}(i_T; d)$ is represented by the GPD model in Eq. (4). In addition, based on the IDF formula and GPD model, the function $a(T)$ can be written as follows:

$$a(T) = G^{-1}\left(\frac{T-1}{T}\right) = u + \frac{\sigma}{\xi}\left\{\left(\frac{T-1}{T}\right)^{-\xi} - 1\right\} \tag{22}$$

A complete mathematical IDF relationship for a pollution event can be determined by combining the information provided by Eq. (22) with the empirical function $b(d) = (d+\theta)^{\eta}$, which can be expressed as shown in the following equation:

$$i(d,T) = \frac{u + \frac{\sigma}{\xi}\left\{\left(\frac{T-1}{T}\right)^{-\xi} - 1\right\}}{(d+\theta)^{\eta}} \tag{23}$$

where u, ξ and σ are threshold, shape, and scale parameters of the GPD model, which are estimated using Eqs. (7) and (8). The parameters θ and η are constants that determine the magnitude of the IDF curve for each defined duration. Thus, the simplified version the IDF formula can be written as follows:

$$i(d,T) \propto \frac{u\xi + \sigma\left\{\left(\frac{T-1}{T}\right)^{-\xi} - 1\right\}}{\xi(d)} \tag{24}$$

6 Results and discussion

Before making our detailed analysis, we performed a preliminary statistical analysis. Figure 3 shows the time series for the observed data corresponding to various air-quality thresholds. In the figure, we can see that the API fluctuates around the mean of 132.71, with a standard deviation of 52.42, which clearly shows the volatility of the data. Specifically, when extreme pollution events occur, the API values tend to increase up to 500 units, which is a hazardous level of air quality. In fact, the POT events with the threshold u = 100 that occurred over the past 20 years (1997–2016) comprise about 2.86% of the data. Thus, it is worthwhile to develop a statistical model and evaluate its IDF relationship for the purposes of risk assessment and management.

The POT data obtained from the original observed data in Fig. 3 could also be referred to as pollution intensities with 1-h durations. To develop an IDF relationship for an extreme pollution event, POT data for various durations must be determined. As described in Sect. 4 and Eq. (20),

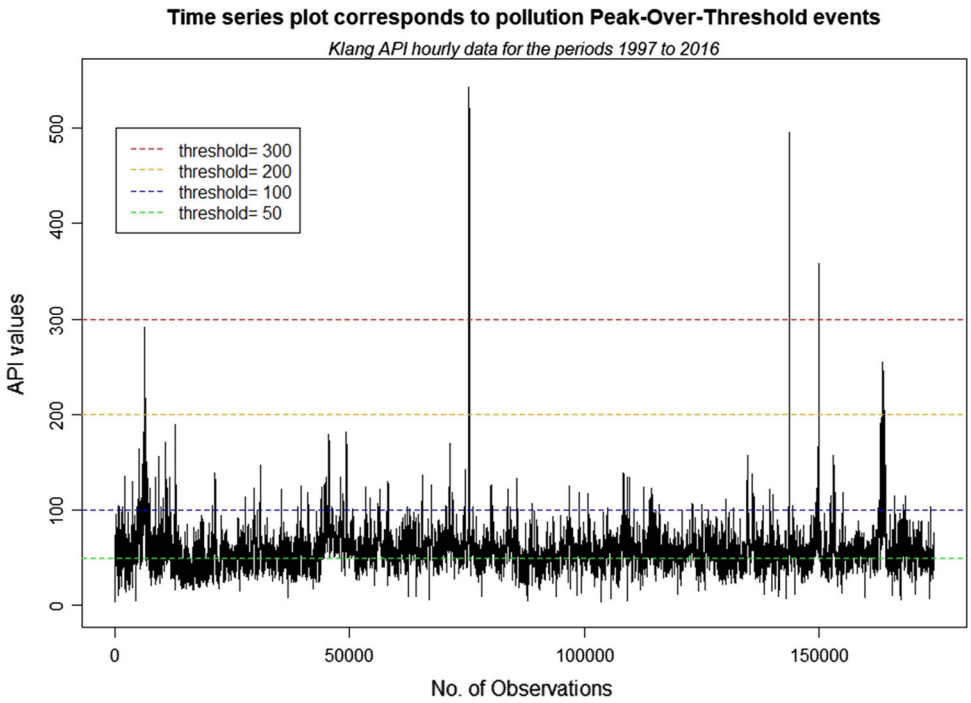

Fig. 3 Time series data corresponding to various air-quality thresholds

the average pollution intensity $I_j(d)$ corresponding to six different hours of duration d = 1, 2, 4, 6, 12, and 24 were determined based on the sum of the air pollution index data available for each of the duration hours d. Next, we extracted POT events with a threshold u = 100, $I(d) = I_j(d) - 100$ from the overall data. Table 2 presents descriptive statistics for a pollution intensity corresponding to a POT (u = 100) for the six different hours of duration, in which the mean of the pollution intensities to increase from 132.71 to 134.71 as the durations increases from 1 to 24 h. The standard deviations are similar for all the duration hours, ranging between 52.42 and 52.51, and the maximum values decrease from 543.0 to 524.3 as the duration increases. In addition, the proportion of pollution intensity values greater than threshold u = 100 decrease as the hours of duration increase.

Next, to gain a deeper understanding of POT events related to the unhealthy threshold for air quality u = 100, we used the GPD to describe its probabilistic behaviors. Based on the GPD model, we can estimate the return periods for unhealthy pollution intensities. In addition, using the GPD model, the IDF curves obtained can be used to estimate the behaviors of future extreme pollution intensities.

6.1 Thresholds assessment

In term of air pollution data, a threshold referring to degree of air quality severity at a particular time, as shown in Table 1. When a certain thresholds is exceeded, this generates public alerts or corrective measures (Smith 1984), especially with respect to pollution events that are always a matter of concern. In this study, since API values greater than 100 are the minimum level designated for unhealthy air quality, it is important to evaluate the risk of this scenario. In this study, we fixed the GPD threshold at u = 100 to indicate the minimum level of unhealthy air quality. However, a number of issues must be addressed in the selection of the threshold u = 100. Generally, the choice of a suitable threshold u for GPD modeling requires a compromise between two competing requirements. If the

Table 2 Descriptive statistics for pollution intensities corresponding to the threshold u = 100

Duration (h)	Mean	Standard deviation	Max	Proportion (API > 100)
1	132.71	52.42	543.0	0.0286
2	133.22	52.75	541.0	0.0280
4	133.37	52.85	535.5	0.0278
6	133.90	53.13	534.0	0.0272
12	133.37	52.63	534.0	0.0273
24	134.71	52.51	524.3	0.0255

choice of threshold is very low, the variance of the parameters will decrease, but the asymptotic basis of the GPD model will be violated, which leads to bias. If the threshold is too high, the corresponding selected sample points will be low, and although the asymptotic properties of the GPD are valid, this scenario will lead to the problem of high parameter variance (Vrban et al. 2018). Thus, it is very important to choose a suitable threshold u for use in GPD modeling. As noted above, we chose $u = 100$ as a fixed threshold in this study, which indicates the minimum level for an unhealthy air pollution event. Thus, it provides a meaningful interpretation of the problem under study. Then, to evaluate the suitability of the fixed threshold in providing a good fitted of GPD model to our data, several graphical tools such as Mean Residual Life (MRL) plot, fitted density plot and also the PP-plot will be used.

The optimal selection of the best threshold value is quite subjective. As mentioned by Scarrott and MacDonald (2012), more than one suitable threshold with different inferred tail behaviors could exist. The issue is whether the threshold provides a balanced trade-off between bias and parameter variance that ensures that the GPD can be used as an accurate approximation model of POT data. In our case, we fixed the unhealthy air-quality threshold for a pollution event to $u = 100$. Although this fixed threshold has a rational basis for air quality analysis, we needed to examine whether it can support valid GPD modeling on pollution data. If this threshold is higher than it should be, then the number of data points exceeding this threshold will be small, so the variance of GPD parameters will be large, which implies unfavorable final results. If this threshold is lower than it should be, most of the data points will be defined as POT data, which will lead to biased results from the GPD modeling. Therefore, it is very important to assess the suitability of our threshold $u = 100$. To do so, we used the mean excess plot method, also known as the plot of the mean residual life (MRL). The estimates of MRL values that exceed the threshold are described by the following equation:

$$E[X - u|X > u] = \hat{M}(u) = \frac{\sum_{i=1}^{n}(X_i - u)I_{[X_i > u]}}{\sum_{i=1}^{n} I_{[X_i > u]}}, \quad u \geq 0 \tag{25}$$

Based on Eq. (25), we can plot the MRL using threshold u against the mean of the excess over the threshold ($E[X - u|X > u]$) for a range of u values. A GPD model is valid for any threshold that generates linearity in the MRL plot (Begueria 2005; Davison and Smith 1990). In other words, the MRL function in Eq. (25) should be linear for threshold u, which is expressed as follows:

$$E[X - u|X > u] = \frac{\sigma}{1 - \xi} + \frac{\xi}{1 - \xi}u \tag{26}$$

For details describing the mean excess plot and its properties, please refer to Ghosh and Resnick (2010).

Figure 4 shows the MRL plots for pollution intensity data corresponding to six different hours of duration. Based on these plots, for the threshold $u = 100$, we can see that the linearity described in Eq. (26) is fulfilled for all 6 h of duration. However, if the threshold value is increased to $u = 200$ (a very unhealthy level of air quality) or $u = 300$ (a hazardous level of air quality), the property of linearity are not satisfied by any of the pollution intensities for different hours of duration. This result implies that the threshold $u = 100$, which refers to an unhealthy level of air quality, is a valid threshold for use in GPD modeling.

6.2 Fitted GPD model and IDF curves

Based on the results presented above, the threshold $u = 100$ is suitable to be used for fitting GPD model on the PDS/POT series. However, before that, it's important to justify the fulfillment of the statistical assumptions on PDS series. The most important assumptions the independence properties. As described previously, to address the problem dependency on PDS series, we use the method of declustering which filtering the dependent consecutive API values for 240 h separation length (cluster). The filtered PDS series based on this technique will exhibit independent behavior. However, to provide a stronger argument on the independent properties, a statistical test based on autocorrelation coefficients is used. Figure 5 shows the result of autocorrelation plot on each PDS series.

The POT series are significantly independent at the $\alpha = 0.05$ significance level if the absolute values of the calculated autocorrelation coefficients of different lag times are not larger than the critical value of $1.96/\sqrt{n}$ (Douglas et al. 2000; Yang et al. 2010; Li et al. 2014). Based on Fig. 5, it is found that the assumption of independent are fulfilled for all the PDS series. Apart from that, it's also important to ensure the number of exceedance data to be satisfied with the Poisson distribution. Table 3 shows the result of goodness-of-fit checking on the Poisson distribution assumption. Based on Table 3, all the statistics χ^2 are smaller which implies a high p values. The null hypothesis assumes that is no significant difference between the empirical data and the Poisson distributed data. Since the χ^2 statistics for all duration hours are exceed the 0.05 significance level, thus this hypothesis is failed to be rejected.

Next, Table 4 shows the parameter estimates of the fitted GPD model for pollution intensities and different hours of duration. The shape parameter of a GPD determines the characteristic of the PDS distribution and also the rate of increase of the physical variable as its exceedance

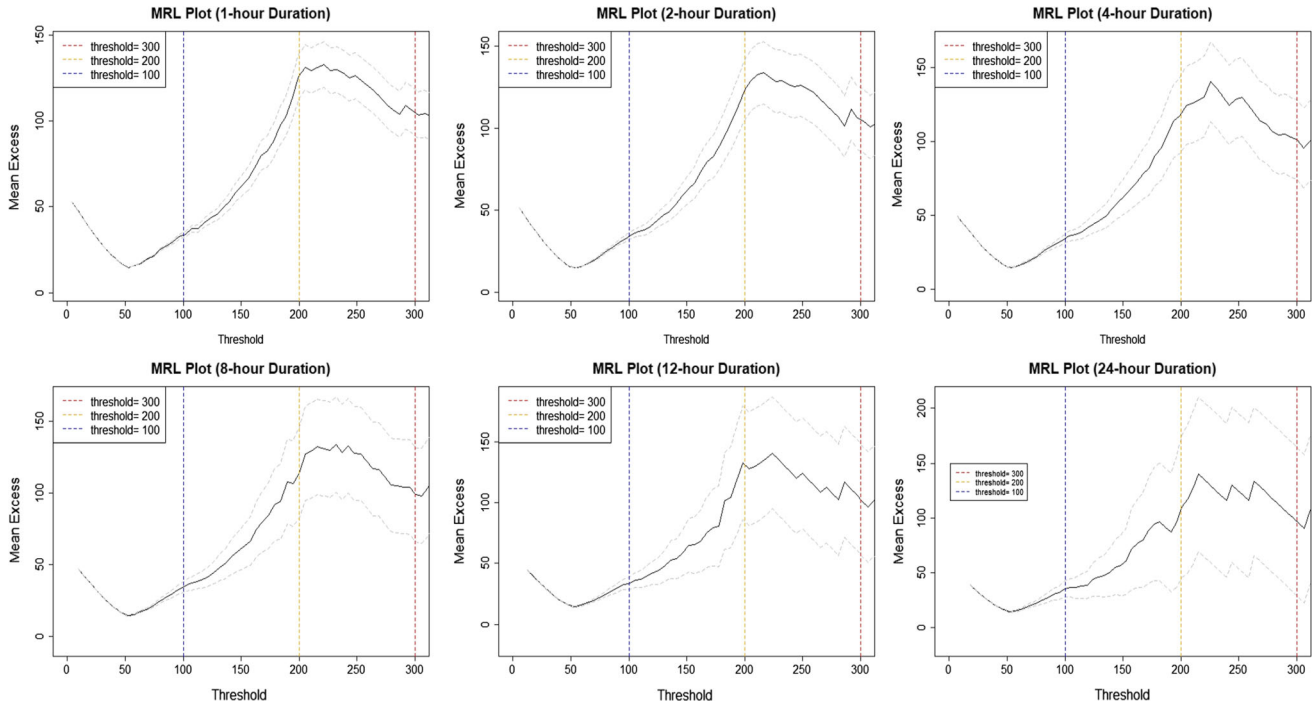

Fig. 4 Assessment of threshold suitability based on MRL plots

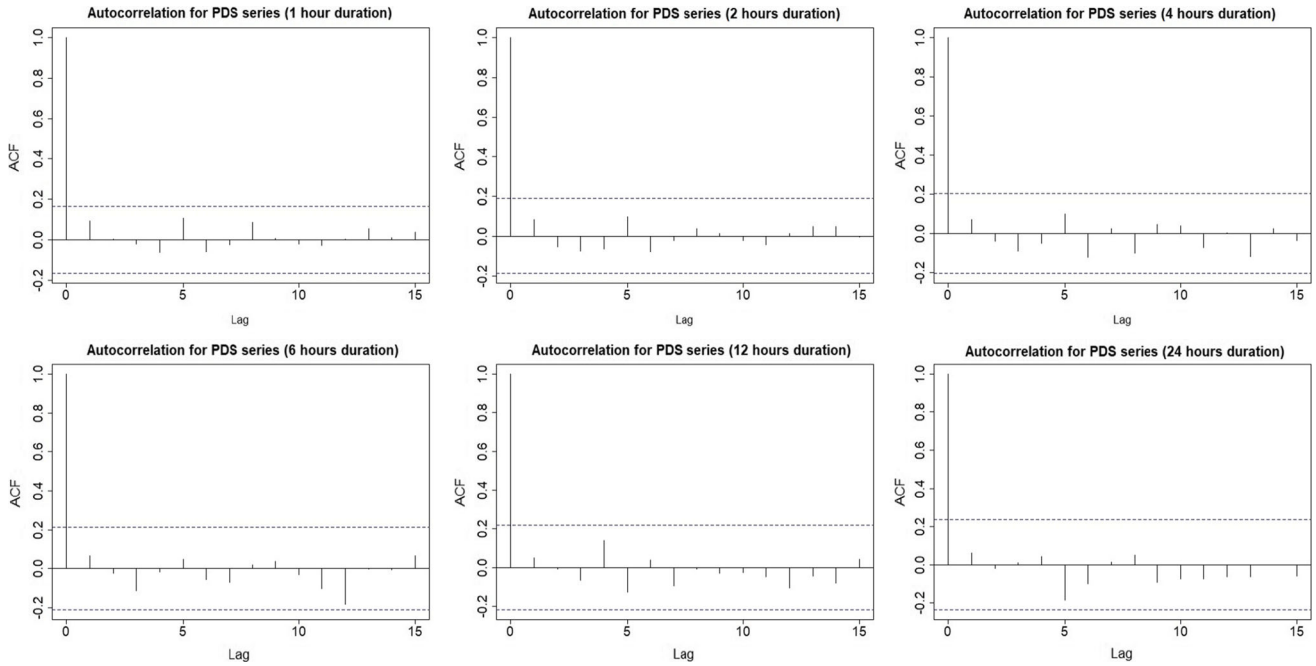

Fig. 5 Autocorrelation plot on each PDS series

probability decreases. From Table 4, it is found that the value of parameter shape $\xi > 0$ for all duration hours. Thus, we can conclude that the PDS series data for all duration hours will follow a Pareto type-II model, which indicates short-tail properties in the distribution of data. The largest absolute difference between the values of shape parameter is between 12 h and 1 h duration which is about 0.085. However, there is no general trend is found for these differences. Apart from that, the scale parameters for durations of 6-h and 24-h are found to be larger than other

Table 3 Goodness-of-fit checking on the Poisson distribution assumption

Duration (h)	χ^2 statistic	Degree of freedom	p value
1	3.683	10	0.961
2	6.439	9	0.695
4	4.192	8	0.839
6	3.476	7	0.838
12	6.143	8	0.631
24	3.813	7	0.801

Table 4 Estimated parameters of GPD model

Duration (h)	Shape (ξ)	Scale (σ)
1	0.376	21.677
2	0.459	21.155
4	0.456	23.549
6	0.391	27.061
12	0.461	23.172
24	0.379	26.679

duration hours which indicates that the PDS series on durations of 6-h and 24 h have a more spread out distribution as compared with the others.

Apart from that, Fig. 6 shows a density plot of the fitted generalized extreme value distribution for the different hours of duration. These density plots show that the GPD model can provide a good approximation of the pollution intensity data for all six durations. In addition, Fig. 7 shows a P–P plot for each fitted GPD model A P–P plot is a probability plot for assessing how closely the empirical distribution of PDS series will agree the fitted GPD model. This is done by plotting the empirical cumulative distribution function of a PDS series versus a GPD cumulative distribution function. Then the plot is compare with a straight line of 45°. If both distributions are equal, then the plot will falls on this line. A large deviation from a comparison line indicates a difference between the distributions. In this study, it is found that the PP-plot shows in Fig. 7 indicates that the empirical quantiles of the PDS series (pollution intensity) can be matched with those of the GPD model. Since the PP plot does not indicate any irregular behaviors, all the data points lie within the 95% confidence interval. Thus, we can conclude that the GPD model provides a good approximation of the pollution intensity data for all different hours of duration. However, for a longer return period, a larger CI will be occur as the return periods increase up to 10 years. This scenario implies the efficiency and robustness of the estimated return level for air pollution intensities will be decreased for longer return periods.

Next, based on the GPD model, Table 5 and Fig. 8 show the estimated returns of pollution-intensity levels corresponding to various return periods. The estimates for various return levels for the fitted GDP are calculated to provide an interpretation about how often pollution intensities larger than a certain value have occurred in the past and may occur in the future. Based on Table 5, as the length of time between return periods increase, it is clear that the pollution intensities will also increase for all different durations (h). As the duration periods increase, the pollution intensities decrease when the intervals between return periods is short (2 and 3 years). For a longer return–period intervals (4 and 5 years), the estimated pollution intensities exhibit inconsistent behavior. While Fig. 8 shows the trend of changes for pollution intensities (return level) with the return period increasing. In this figure, the empirical return levels, the estimated return level and their 95% CIs derived from delta method are expressed by the circles, the solid line and the dashed lines, respectively. It found that, for all duration hours, the empirical and the estimated return levels are within the 95% CIs of its estimated GPD return levels. This result provide an agreement that the GPD is suitable model for extreme air pollution intensities.

Apart from that, based on the GPD model, we also plotted IDF curves for various hours of duration and return periods, as shown in Fig. 9. For all return periods, the IDF curves show a similar trend. However, the magnitudes of the IDF curves differ for each return period, with long return periods having a higher estimated pollution intensity. This finding is in agreement with the results shown in Table 5 and Fig. 8, where all the IDF lines increase more steeply from 1- to 6-h of duration. For 6–24 h, however, the increasing trend of pollution intensity is less steep. For a short duration (1 h), the pollution intensities range from 117.16/h for a return period of 2 years to 179.38/h for a longer return period of 10 years.

As the duration increases, the pollution intensities decrease for all return periods. For the longest duration (24 h), the intensity ranges from 121.15/h for a 2-year return period to 198.08/h for a 10-year return period. However, although the plot of the return period and IDF curve indicate that high-intensity pollution events pose a high environmental risk, these events occur less frequently than low-intensity pollution events. All these results pro-

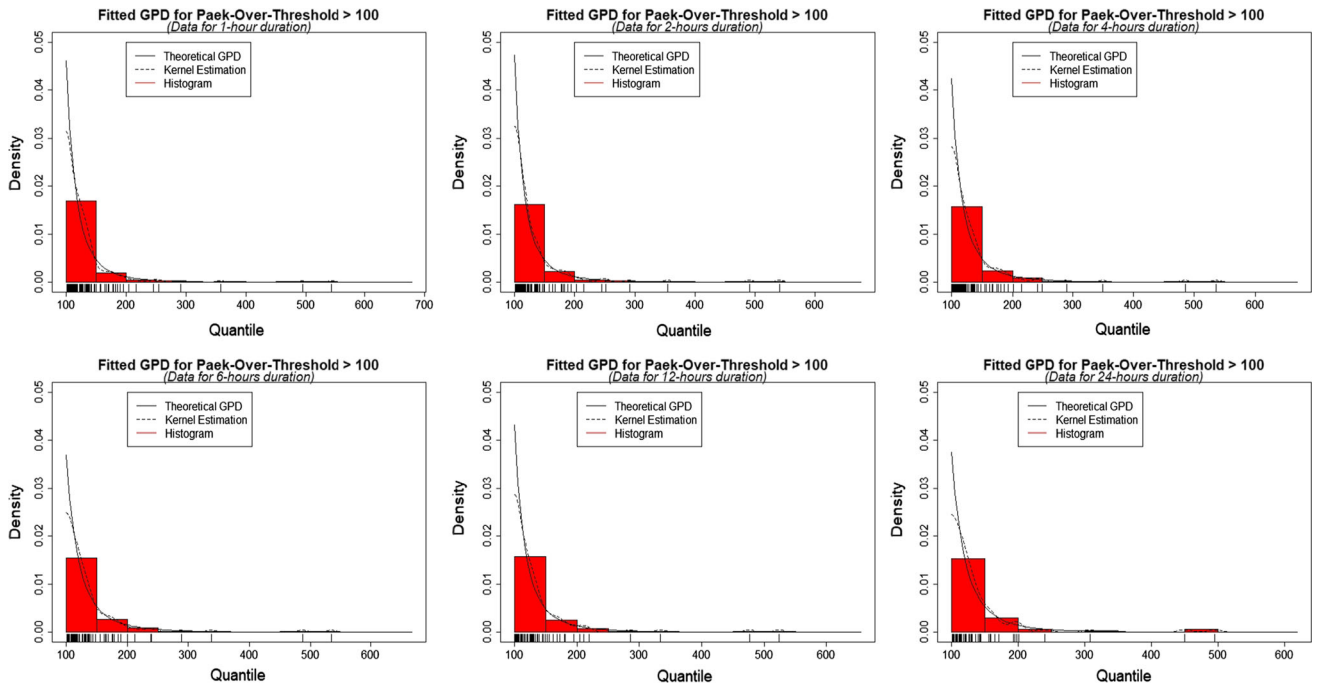

Fig. 6 Fitted GPD model for various durations (h)

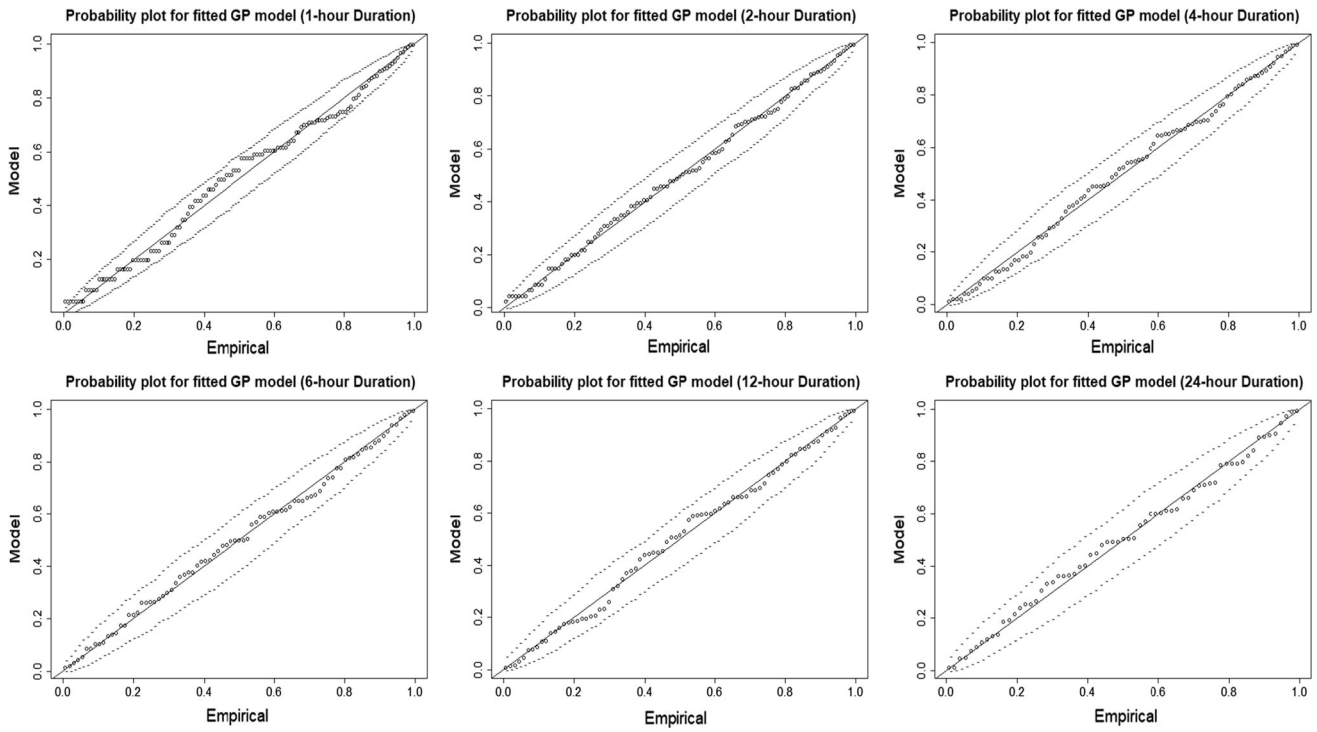

Fig. 7 PP plot for each fitted GPD model

vide useful information for the authorities in terms of pollution risk management and control. For example, in the current scenarios, for any hours of duration, the pollution intensities are estimated to increase in the future. Thus, additional prevention policies and stricter enforcement should be employed to manage the risk of extreme pollution, particularly for long-term environmental sustainability.

Table 5 Estimates of return pollution-intensity levels on the risk of pollution event

Duration (h)	Return levels estimates on extreme pollution events			
	2-years period	3-years period	4-years period	5-years period
1	723.48	835.54	926.06	1003.32
2	1000.09	1193.62	1354.49	1494.75
4	1092.86	1305.67	1482.47	1636.57
6	936.67	1092.64	1219.33	1327.92
12	1099.58	1316.20	1496.51	1653.86
24	869.22	1008.40	1120.92	1217.03

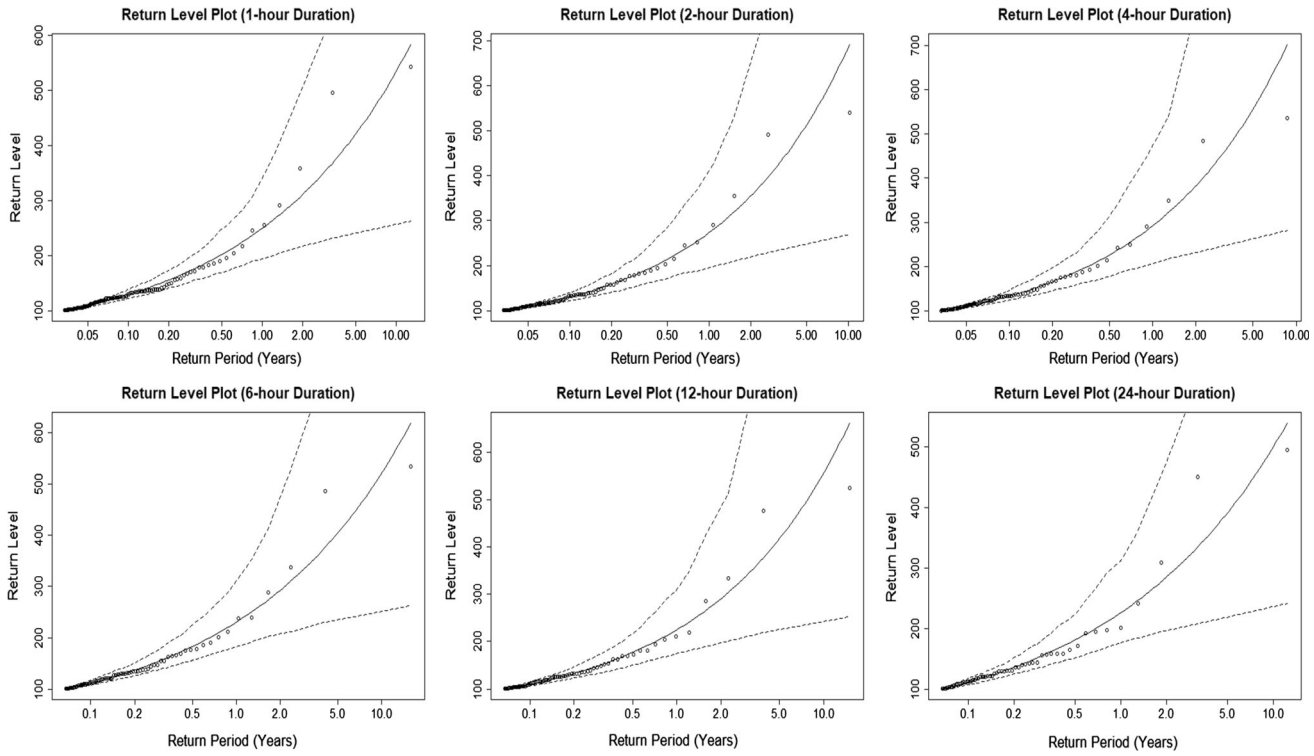

Fig. 8 Plots of estimated API intensities against return period

6.3 Evaluating the performance of IDF model

In order to evaluate the IDF model performance, the bootstrap method is used. Bootstrap method is a class of Monte Carlo approach which provide a way to evaluate the performance of fitted statistical model based on resampling technique (Rizzo 2008). Based bootstrap method, the PDS series will be treated as finite population, and the random samples will be generated repeatedly from it to re-estimate the IDF model. Then, based on the bootstrap replicates of IDF model, the estimate standard error and biases of IDF model will be determined to describe about its performance. Let $\hat{i}(d,T)^{(1)}, \hat{i}(d,T)^{(2)}, \ldots, \hat{i}(d,T)^{(B)}$ be the bootstrap replicates of IDF model, then the bootstrap estimate of standard error for IDF curve can be written as follow

$$\hat{\sigma}\left(\hat{i}(d,T)\right) = \sqrt{\frac{\sum_{b=1}^{B}\left(\hat{i}(d,T)^{(b)} - \bar{i}(d,T)\right)^2}{B-1}} \tag{27}$$

where $\bar{i}(d,T) = \frac{\sum_{b=1}^{B}\hat{i}(d,T)^{(b)}}{B}$. According to Efron and Tibshirani (1993), the number of replicates, $B = 50$ is sufficient to provide a good estimates of standard error. Then, based on the standard deviation, $100(1-\alpha)\%$ bootstrap confidence interval for IDF can be obtained as

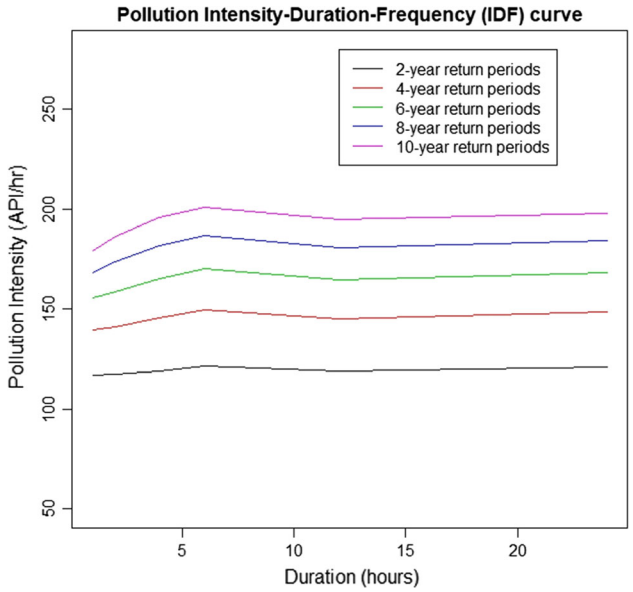

Fig. 9 IDF curves of extreme API intensities

$$\hat{i}(d,T) \pm z_{\alpha}\hat{\sigma}\left(\hat{i}(d,T)\right) \tag{28}$$

Apart from that, the bias of IDF model is measure as

$$\begin{aligned} bias\left(\hat{i}(d,T)\right) &= E\left[\left(\hat{i}(d,T)\right) - (i(d,T))\right] \\ &= E\left[\hat{i}(d,T)\right] - i(d,T) \end{aligned} \tag{29}$$

The sample mean, $\bar{i}(d,T)$ of the bootstrap replicates $\hat{i}(d,T)^{(1)}, \hat{i}(d,T)^{(2)}, \ldots, \hat{i}(d,T)^{(B)}$ is unbiased for its expected value $E\left[\hat{i}(d,T)\right]$, while $i(d,T) = \hat{i}(d,T)$ is the estimate which compute from the original data (Rizzo 2008). Thus, Eq. (29) can be simplify as

$$bias\left(\hat{i}(d,T)\right) = \bar{i}(d,T) - \hat{i}(d,T) \tag{30}$$

The IDF models (different duration hour) with a smaller value of standard error indicates a better performance. In parallel with that, the IDF models with a smaller value of bias indicates a better accuracy with a less of potential

Table 6 Bootstrap results on IDF model performance

Duration (h)	Return period (T) (years)	IDF estimate	Bias	Standard error	Confidence interval
1	2	117.165	0.133	1.772	(113.709, 120.701)
	4	139.441	0.420	3.809	(132.011, 146.870)
	6	155.431	0.760	5.723	(144.270, 166.591)
	8	168.349	1.131	7.708	(153.316, 183.381)
	10	179.377	1.522	9.732	(160.397, 198.356)
2	2	117.264	0.413	2.546	(174.412, 184.342)
	4	140.996	0.526	5.635	(168.388, 190.365)
	6	158.809	0.441	8.260	(163.269, 195.484)
	8	173.617	0.317	10.826	(158.265, 200.488)
	10	186.528	0.190	13.384	(153.276, 205.477)
4	2	119.197	0.661	2.955	(173.613, 185.140)
	4	145.530	1.598	6.692	(166.326, 192.427)
	6	165.265	2.406	9.809	(160.249, 198.504)
	8	181.653	3.172	12.777	(154.460, 204.293)
	10	195.930	3.917	15.688	(148.784, 209.969)
6	2	121.545	0.190	3.354	(172.836, 185.917)
	4	149.797	0.327	7.281	(165.177, 193.576)
	6	170.242	0.507	10.650	(158.608, 200.146)
	8	186.844	0.754	13.949	(152.175, 206.578)
	10	201.071	1.053	17.224	(145.789, 212.964)
12	2	118.924	0.383	2.350	(174.793, 183.960)
	4	144.973	0.765	5.136	(169.361, 189.393)
	6	164.548	1.126	8.251	(163.286, 195.467)
	8	180.830	1.533	11.798	(156.369, 202.384)
	10	195.034	1.986	15.589	(148.977, 209.777)
24	2	121.148	0.504	3.300	(172.941, 185.812)
	4	148.650	0.917	7.579	(164.597, 194.156)
	6	168.423	1.195	11.284	(157.371, 201.382)
	8	184.414	1.458	14.864	(150.391, 208.362)
	10	198.076	1.729	18.388	(143.519, 215.234)

errors. Table 6 show the bootstrap results on the IDF model performance.

Based on Table 6, the measured bias for all IDF estimates are found to be low which indicate that the potential of errors generated by IDF model is low. However, the bias of IDF model for all duration hours are found to be increase as the return period increase except for the IDF model with the duration of 2-h. This results implies that the IDF model provide a more accurate estimation of pollution intensity for short-term compare to a long-term return period. Apart from that, the standard error provide the information about the magnitude of variability for each IDF estimate. The performance of the IDF model with low duration hours (1 h, 2 h and 4 h) provide a less variability compare to IDF model with a higher duration hours (8 h, 12 h and 24 h). Besides that, the IDF estimate for short-term period (2, 4 and 6 years) also having less variability compare to a longer return period (8 and 10 years). In overall, we can conclude that the best IDF model is the model with low duration hours which better to be used for a short-term estimate of return period.

7 Conclusion

In this study, we proposed the use of the IDF approach as an alternative tool for evaluating the risk of extreme air pollution indexes. Data from the city of Klang, Malaysia was used as a case study. The steps for developing the IDF curve involved the determination of a PDS for extreme pollution events using the POT method. Then, we determined the GPD to indicate the probabilistic behaviors of the PDS data. Based on the GPD model, we estimated the pollution intensities corresponding to various return periods. The results showed that the pollution intensities in Klang tend to increase with increases in the length of time between return periods. In addition, based on the GPD model, the IDF curve showed similar increasing trends for different return periods. However, the magnitude of the IDF curves differed for different return periods, with long return periods associated with higher pollution-intensity estimates. Overall, based on the study results, we conclude that the IDF approach provides a good basis for decision-makers to assess the expected risk of future extreme pollution events.

Acknowledgements The author is indebted Malaysian Department of Environment for providing air pollution data. This research would not be possible without the sponsorship from the Universiti Kebangsaan Malaysia (Grant Number DIP-2018-038).

References

Al-Dhurafi NA, Masseran N, Zamzuri ZH (2018a) Compositional time series analysis for air pollution index data. Stoch Environ Res Risk Assess 32:2903–2911

Al-Dhurafi NA, Masseran N, Zamzuri ZH, Safari MAM (2018b) Modeling the air pollution index based on its structure and descriptive status. Air Qual Atmos Health 11(2):171–179

Al-Dhurafi NA, Masseran N, Zamzuri ZH, Razali AM (2018c) Modeling unhealthy air pollution index using a peaks-over-threshold method. Environ Eng Sci 35(2):101–110

Alyousifi Y, Masseran N, Ibrahim K (2018) Modeling the stochastic dependence of air pollution index data. Stoch Environ Res Risk Assess 32(6):1603–1611

Azmi SZ, Latif MT, Ismail AS, Juneng L, Jemain AA (2010) Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. Air Qual Atmos Health 3:53–64

Begueria S (2005) Uncertainties in partial duration series modeling of extremes related to the choice of threshold value. J Hydrol 303:215–230

Ben-Zvi A (2009) Rainfall intensity–duration–frequency relationships derived from large partial duration series. J Hydrol 367:104–114

Coles S (2001) An introduction to statistical modeling of extreme values. Springer, London

Dale VH, Joyce LA, McNulty S, Neilson RP, Ayres MP, Flannigan MD, Hanson PJ, Irland LC, Lugo AE, Peterson CJ, Simberloff D, Swanson FJ, Stocks BJ, Wotton BM (2001) Climate change and forest disturbances: climate change can affect forests by altering the frequency, intensity, duration, and timing of fire, drought, introduced species, insect and pathogen outbreaks, hurricanes, windstorms, ice storms, or landslides. Bioscience 51(9):723–734

Davison A, Smith R (1990) Models for exceedances over high thresholds. J R Stat Soc Ser B 52:393–442

Department of Environment (1997) A guide to air pollutant index in Malaysia (API). Ministry of Science, Technology and the Environment, Kuala Lumpur, Malaysia. https://aqicn.org/images/aqi-scales/malaysia-api-guide.pdf

Douglas EM, Vogel RM, Kroll CN (2000) Trends in floods and low flows in the United States: impact of spatial correlation. J Hydrol 240:90–105

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall/CRC, Boca Raton

Ghosh S, Resnick SA (2010) A discussion on mean excess plots. Stoch Process Appl 120:1492–1517

Google (2019) Source: https://maps.googleapis.com/maps/api/geocode/json?address=Klang%2CSelangor&key=xxx

Gulia S, Nagendra SMS, Khare M, Khanna I (2015) Urban air quality management—a review. Atmos Pollut Res 6:286–304

Gyarmati-Szabo J, Bogachev LV, Chen H (2017) Nonstationary POT modelling of air pollution concentrations: statistical analysis of the traffic and meteorological impact. Environmetrics 28(5):e2449-1–e2449-15

Husler J, Li D, Raschke M (2011) Estimation for the generalized Pareto distribution using maximum likelihood and goodness of fit. Commun Stat Theory Methods 40:2500–2510

Jayasooriya VM, Ng AWM, Muthukumaran S, Perera BJC (2017) Green infrastructure practices for improvement of urban air quality. Urban For Urban Green 21:34–47

Karim F, Hasan M, Marvanek S (2017) Evaluating annual maximum and partial duration series for estimating frequency of small magnitude floods. Water 9:481

Khaliq MN, Ouarda TBMJ, Ondo J-C, Gachon P, Bobee B (2006) Frequency analysis of sequence of dependent and/or non-

stationary hydro-meteorological observations: a review. J Hydrol 329:534–552
Koutsoyiannis D, Kozonis D, Manetas A (1998) A mathematical framework for studying rainfall intensity–duration–frequency relationships. J Hydrol 206:118–135
Kumar P, Jain S, Gurjar BR, Sharma P, Khare M, Morawska L, Britter R (2013) New directions: can a "blue sky" return to Indian megacities? Atmos Environ 71:198–201
Kumar P, Morawska L, Martani C, Biskos G, Neophytou M, Di Sabatino S, Bell M, Norford N, Britter R (2015) The rise of low-cost sensing for managing air pollution in cities. Environ Int 75:199–205
Lang M, Ouarda TBMJ, Bobee B (1999) Towards operational guidelines for over-threshold modeling. J Hydrol 225:103–117
Li Z, Li C, Xu Z, Zhou X (2014) Frequency analysis of precipitation extremes in Heihe River basin based on generalized Pareto distribution. Stoch Environ Res Risk Assess 28(7):1709–1721
Lui JC, Mickley LJ, Sulprizio MP, Dominici F, Yue X, Ebisu K, Anderson GB, Khan RFA, Bravo MA, Bell ML (2016) Particulate air pollution from wildfires in the Western US under climate change. Clim Change 138:655–666
Masseran N (2017) Modeling fluctuation of PM10 Data with existence of volatility effect. Environ Eng Sci 34(11):816–827
Masseran N, Razali AM, Ibrahim K, Zaharim A, Sopian K (2013) Application of the single imputation method to estimate missing wind speed data in Malaysia. Res J Appl Sci Eng Technol 6(10):1780–1784
Masseran N, Razali AM, Ibrahim K, Latif MT (2016) Modeling air quality in main cities of Peninsular Malaysia by using a generalized Pareto model. Environ Monit Assess 188(1):65-1–65-12
Mohymont B, Demarée GR, Faka DN (2004) Establishment of IDF-curves for precipitation in the tropical area of central Africa—comparison of techniques and results. Nat Hazards Earth Syst Sci 4:375–387
Pickands J (1975) Statistical inference using extreme order statistics. Ann Stat 3:119–131
Reiss R-D, Thomas M (2007) Statistical analysis of extreme values: with application to insurance, finance, hydrology and other fields. Die Deutsche Bibliothek, Berlin
Ribatet M (2007) POT: modelling peak over a threshold. R News 7:33–36
Rizzo ML (2008) Statistical computing with R. Chapman and Hall/CRC, Boca Raton
Sahani M, Zainon NA, Wan Mahiyuddin WR, Latif MT, Hod R, Khan MF, Tahir NM, Chan C-C (2014) A case-crossover analysis of forest fire haze events and mortality in Malaysia. Atmos Environ 96:257–265
Scarrott C, MacDonald A (2012) A review of extreme value threshold estimation and uncertainty quantification. REVSTAT Stat J 10:33–60
Smith RL (1984) Threshold methods for sample extremes. Stat Extrem Appl NATO ASI Ser 131:621–638
Southworth H, Heffernan JE (2014) texmex: statistical modelling of extreme values. R package version 2.1
Van de Vyver V (2015) Bayesian estimation of rainfall intensity–duration–frequency relationships. J Hydrol 529:1451–1463
Vrban S, Wang Y, McBean EA, Binns A, Gharabaghi B (2018) Evaluation of stormwater infrastructure design storms developed using partial duration and annual maximum series models. J Hydrol Eng 23(12):04018051
Willems P (2000) Compound intensity/duration/frequency-relationships of extreme precipitation for two seasons and two storm types. J Hydrol 233:189–205
Xia J, Du H, Zeng S, She D, Zhang Y, Yan Z, Ye Y (2012) Temporal and spatial variations and statistical models of extreme runoff in Huaihe River Basin during 1956–2010. J Geogr Sci 22(6):1045–1060
Xu Q, Li X, Wang S, Wang C, Huang F, Gao Q, Wu L, Tao L, Guo J, Wang W, Guo X (2016) Fine particulate air pollution and hospital emergency room visits for respiratory disease in urban areas in Beijing, China, in 2013. PLoS ONE 11(4):e0153099
Yang T, Shao QX, Hao Z-C, Chen X, Zhang Z, Xu C-Y, Sun L (2010) Regional frequency analysis and spatio-temporal pattern characterization of rainfall extremes in the Pearl River Basin, China. J Hydrol 380:386–405
Yoo J-M, Lee YR, Kim D, Jeong MJ, Stockwell WR, Kundu PK, Oh SM, Shin DB, Lee SJ (2014) New indices for wet scavenging of air pollutants (O_3, CO, NO_2, SO_2, and PM_{10}) by summertime rain. Atmos Environ 82:226–237
Zhang H, Wang S, Hao J, Wang X, Wang S, Chai F, Li M (2016) Air pollution and control action in Beijing. J Clean Prod 112(2):1519–1527
Zhou S-M, Deng Q-H, Lui W-W (2012) Extreme air pollution events: modeling and prediction. J Cent South Univ Technol 19:1668–1672
Zidek JV, Shaddick G, White R, Meloche J, Chatfield C (2005) Using a probabilistic model (pCNEM) to estimate personal exposure to air pollution. Environmetrics 16:481–493