Thailand Statistician
July 2021; 19(3): 642-658
http://statassoc.or.th
Contributed paper

Extreme Value Analysis of PM_{10} Concentration in Thailand

Kuntalee Chaisee* and Kamonrat Suphawan
Data Science Research Center, Department of Statistics, Faculty of Science,
Chiang Mai University, Chiang Mai, Thailand.
*Corresponding author; e-mail: kuntalee.chaisee@cmu.ac.th

Received: 24 July 2020
Revised: 13 March 2021
Accepted: 1 April 2021

Abstract

This research aims to analyze the extreme values of the air pollutants, in particular, PM_{10} concentration in Thailand. Due to the limitation of data, we restrict our attention to 23 air quality monitoring stations in Thailand. The daily PM_{10} concentration data from 2008 to 2019 are used to analyze and are divided into two types; 24-hour averages and daily maxima. The Peak Over Threshold (POT) approach is used to assess the risk of air pollutants; hence the Generalized Pareto Distribution (GPD) is used to fit the data. One of the challenging issues in POT is the choice of threshold. In this work, we combine the mean residual life plot and the goodness of fit test methods to determine the threshold. The maximum likelihood estimation and the bootstrap method are used to deal with parameter estimation in GPD and uncertainty quantification. We then estimate the return levels, which present extreme predictive events in terms of the values expected to exceed average once every return period. The results show that daily PM_{10} concentration at station 24t in Saraburi, 73t in Chiang Rai, and 36t in Chiang Mai have very high predictive extreme values. Many stations located in the north of Thailand also have relatively high levels. Consequently, the northern region is most likely to encounter high exposures to PM_{10}.

Keywords: Extreme values, air pollution, PM_{10}, GPD, POT, return levels.

1. Introduction

In the past several years, atmospheric particulate matter (PM) has been in the spotlight and considered to be a global environmental issue because it can contribute to health problems, predominantly to the respiratory and cardiovascular systems such as lung cancer and cardiopulmonary diseases (Pope et al. 2002, Brook et al. 2010). PM is often divided based on size; PM_{10} refers to the fine particles with a diameter of less than 10 micrometers and smaller fine particles with a diameter of less than 2.5 micrometers commonly known as $PM_{2.5}$. These two fine particles can come from various sources such as power plants, motor vehicles, forest fires, agricultural burning. In Thailand, according to the report by the Pollution Control Department Ministry of Natural Resources and Environment in Pollution Control Department (2019), it has been revealed that the major causes of particulate matter are different in a different part of Thailand. In the northern region of Thailand, the major causes are open burning and forest fires. In contrast, the major cause of Thailand's central

region is the diffusion of particulate matter from cement plants, lime plants, stone crushing plants, quarries in the area, and nearby, as well as traffic congestion, transportation, and logistics activity in the area where roads are damaged. The primary causes in the largest urban area, Bangkok and vicinity, are vehicles in addition to meteorological conditions of not-circulating and no wind speed. According to the WHO air quality guidelines (World Health Organization 2006), the 24-hour average should not exceed 25 micrograms per cubic meter ($\mu g/m^3$) for $PM_{2.5}$ and 50 micrograms per cubic meter for PM_{10}. Nevertheless, it has been reported in Pollution Control Department (2019) that the 24-hour average of $PM_{2.5}$ was in the range of 3-133 micrograms per cubic meter and 2-303 micrograms per cubic meter for PM_{10}. These maximum concentrations are considered very unhealthy. The data analysis of air pollution is usually focused on a few extreme situations. It substantially impacts human health and well-being. Over the past years, a study of PM in Thailand has mainly focused on the health effect (Vichit-Vadakan et al. 2001, Viroj 2008, Pothirat et al. 2019). The results have shown a significant association between exposure to PM and health effects. Nevertheless, few studies have paid attention to giving an important piece of information of PM_{10} in terms of extreme values awareness in Thailand using extreme value analysis. Extreme value analysis has been widely used to assess the risk of rare events, especially in an environmental disaster such as flood frequency analysis, extreme temperature, finance, and insurance (Ragulina and Reitan 2017, Osman et al. 2015, Gilli and Këllezi 2006). There are two conventional approaches for extreme value analysis, the block maxima method, where the model is the Generalized Extreme Value (GEV) distribution. Another is the Peak Over Threshold (POT), where the model is Generalized Pareto Distribution (GPD). The GEV for modeling extremes of a (time) series of observations is based on the maximum or minimum values of these observations within a certain period size or a time block. The GPD, on the other hand, uses the observations over a high threshold to model the extremes. Both methods have some difficulties; modeling using GEV requires a suitable time block, using too long periods gives only a few values. At the same time, too short periods lead to biases. GPD needs a good choice of threshold to get an optimal balance between bias and variance (Coles 2001). They are often used in environmental data such as wind gusts, precipitation, and air pollution (Brabson and Palutikof 2000, Engeland et al. 2004, Martins et al. 2017). Nonetheless, the GPD is more attainable than the GEV for extreme values of the air pollution data (Masseran et al. 2016, Gyarmati-Szabó et al. 2017, AL-Dhurafi et al. 2018). This research focuses on assessing the risk of extreme values of PM_{10} daily concentrations in Thailand using the POT method. The future estimates of extreme values are manifested in terms of return levels obtained from estimated parameters in the GPD with selected thresholds.

2. Materials and Methods

This paper aims to apply the POT to estimate the future extreme values of PM_{10} concentrations expressed in return levels using generalized Pareto distribution (GPD). In this section, we present the data used in the analysis and an overview of the methods. We describe the principal concepts of modeling extreme values using GPD, selecting threshold, parameter estimation methods in GPD, and return level estimation.

2.1. Study area and data

The data used in this research is PM_{10} hourly concentration collected by the Pollution Control Department, Air Quality and Noise Management Bureau, The Ministry of Natural Resources and Environment of Thailand. Currently, there are 70 air monitoring stations located across the country. Unfortunately, there are many missing data in many stations. As a result, we restrict our attention to

stations where PM_{10} data are complete or nearly complete from January 2008 - June 2019. The list of the stations is shown in Table 1, and the location is shown in Figure 1. In this work, we separate the analysis of the data into two directions; the analysis for 24-hour average and the maximum or the peak of hourly concentrations. The 24-hour averages are commonly used to report the air quality, whereas the maxima of hourly concentration reflect health impacts better (Lin et al. 2017, Zikova et al. 2017).

Table 1 List of 23 selected air quality monitoring stations

Code	Region	Province	Latitude	Longitude
02t	Central	Bangkok	13.733	100.488
05t	Central	Bangkok	13.666	100.606
10t	Central	Bangkok	13.780	100.646
14t	Central	Samut Sakhon	13.705	100.316
17t	Central	Samut Prakan	13.652	100.532
24t	Central	Saraburi	14.686	100.872
26t	West	Ratchaburi	13.533	99.815
27t	Central	Samut Sakhon	13.550	100.264
30t	East	Rayong	12.672	101.276
32t	East	Chon Buri	13.119	100.919
35t	North	Chiang Mai	18.841	98.970
36t	North	Chiang Mai	18.791	98.988
40t	North	Lampang	18.283	99.660
43t	South	Phuket	7.885	98.391
44t	South	Songkhla	7.021	100.484
46t	Northeast	Khon Kaen	16.445	102.835
47t	Northeast	Nakhon Ratchasima	14.980	102.098
59t	Central	Bangkok	13.783	100.541
61t	Central	Bangkok	13.770	100.615
63t	South	Yala	6.546	101.283
67t	North	Nan	18.789	100.776
69t	North	Phrae	18.129	100.162
73t	North	Chiang Rai	20.427	99.884

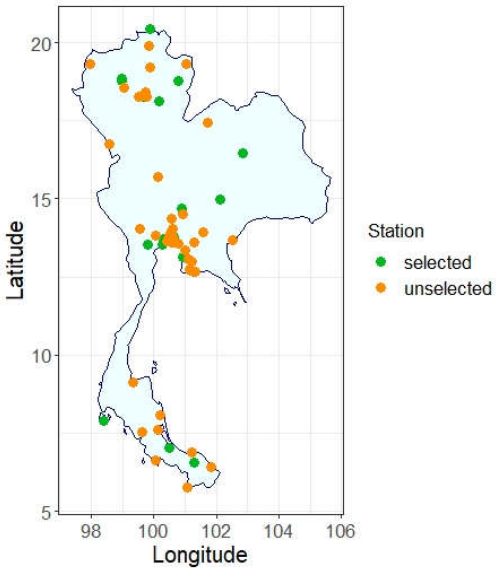

Figure 1 Locations of air quality monitoring stations in Thailand

2.2. Generalized Pareto distribution

The use of the Generalized Pareto Distribution (GPD) for modeling excesses over a high threshold is justified by arguments on the asymptotic behavior of the data (Coles 2001, Pickands 1975, Leadbetter 1983). Let $X_1,\ldots,X_n$ be identically distributed random variables with distribution function F. For $x>0$, $P(X-u\le x \mid X>u)$ is called the distribution function of exceedances over the threshold u. For $Y=X-u$, where $X>u$, the distribution of the exceedances $Y_j=X_i-u$ such that i is the index of the j^{th} exceedance, $j=1,\ldots,n_u$, the distribution of $Y_1,\ldots,Y_{n_u}$ can be approximated by the GPD given by

$$G(y \mid u,\sigma_u,\xi)=\begin{cases}1-\left(1+\xi\left(\dfrac{y}{\sigma_u}\right)\right)^{-1/\xi}, & \xi\neq 0\\ 1-\exp\left(-\left(\dfrac{y}{\sigma_u}\right)\right), & \xi=0,\end{cases}$$

where σ_u is the scale parameter and ξ is the shape parameter.

2.3. Threshold selection

Selecting threshold is very crucial in extreme value analysis because a high threshold provides the convergence towards the extreme value theory and reduced bias. However, a too high threshold leads to high variance of the estimated parameters as there will be fewer data exceeding the threshold. Typically, the threshold was chosen before fitting, but there are no best and promising methods. Several approaches for threshold selection in extreme value modeling have been established (Dupuis 1999, Thompson et al. 2009, Scarrott and MacDonald 2012). The standard practice for determining threshold is to compromise between bias and variance. One of the common methods is to use the Mean Residual Life (MRL) plot, also known as the mean excess plot. It is based on the mean of excesses over the threshold given by

$$E(X-u_0 \mid X>u_0)=\frac{\sigma_{u_0}}{1-\xi},$$

provided the shape parameter $\xi<1$ and the scale parameter σ_{u_0} for the exceedances over the threshold u_0. The idea is, if the GPD is valid for excesses of the threshold u_0, then it should be valid for all thresholds $u>u_0$ subjected to the suitable scale parameter σ_u. By the extreme value theory (Coles 2001),

$$E(X-u \mid X>u)=\frac{\sigma_u}{1-\xi}=\frac{\sigma_{u_0}+\xi_u}{1-\xi}.$$

The mean excess is expecting to change linearly with u for all $u>u_0$. Another method used in this paper is the Goodness-of-Fit (GoF) method. It is one of the statistical hypothesis testing methods to determine distribution. The null hypothesis of the tests is that the data follows the distribution of interest. Suppose that we want to test if $X_1,\ldots,X_n$ is a random sample from a continuous distribution with cumulative distribution function $F(x)$ with the parameter $\theta=(\sigma_u,\xi)$. We estimate parameter θ by $\hat{\theta}$ and then compute the test statistics using Cramer-van Mises, W^2, and Anderson-Darling, A^2,

$$W^2=\frac{1}{12n}+\sum_{i=1}^{n}\left(F(x_i)-\frac{2i-1}{2n}\right),$$

$$A^2=-n-\sum_{i=1}^{n}\frac{2i-1}{n}\left(\log[F(x_i)]+\log[1-F(x_{n+1-i})]\right),$$

where x_i are in ascending order (Chen and Balakrishnan 1995, Choulakian and Stephens 2001). To determine if the data follows the distribution, we often compare the P-values based on the computed statistics to the pre-specified significance level α. In this work, we initially use the MRL plots to locate plausible thresholds then use the GoF to test if the data above those thresholds follow the GPD distribution with the level of significance of 0.05.

2.4. Parameter estimation

Several parameter estimation methods can be applied to estimate the GPD parameters. Conventional methods for parameter estimation in GPD are the maximum likelihood method, method of moments and probability-weighted moments method, and Bayesian method (Hosking and Wallis 1987, Grimshaw 1993, Worms and Worms, 2012). In this work, the parameters of the GPD can be estimated by the maximum likelihood method. Suppose $y_1,\ldots,y_n$ be n_u sequence of excesses of a threshold u. For $\xi\neq 0$, the log-likelihood can be written as

$$l(\sigma,\xi)=\begin{cases}-n_u\log\sigma-(1+1/\xi)\sum_{i=1}^{n_u}\log\left(1+\xi\frac{y_i}{\sigma}\right), & \xi\neq 0\\ -n_u\log\sigma-(1/\sigma)\sum_{i=1}^{n_u}y_i, & \xi=0,\end{cases}$$

provided $(1+\xi y_i/\sigma)>0$ for $i=1,\ldots,n_u$.

2.5. Return level estimation

One of the most common interests of extreme value analysis is the evaluation of return levels. The interest is the return level x_N, which is exceeded once every N year. Let ζ_u is the probability of the event of exceedance over the threshold u. The level x_m that is exceeded on average once every m observation is the solution of the following equation

$$\zeta_u \left[1+\xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{m}.$$

This can be expressed as follows

$$x_m = \begin{cases} u+\frac{\sigma}{\xi}((m\zeta_u)^{\xi}-1), & \text{for } \xi \neq 0 \\ u+\sigma\log(m\zeta_u), & \text{for } \xi = 0. \end{cases}$$

x_m is called the m observation return level. To compute the N-year return level, let n_y be the number of observations per year, so $n_y = m/N$ and hence the N-year return level is

$$x_N = \begin{cases} u+\frac{\sigma}{\xi}((Nn_y\zeta_u)^{\xi}-1), & \text{if } \xi \neq 0 \\ u+\sigma\log(Nn_y\zeta_u), & \text{if } \xi = 0, \end{cases}$$

where threshold u follows the binomial distribution with probability ζ_u (Coles 2001).

2.6. Uncertainty of the return level estimates

In general, an estimated parameter comes with an uncertainty that is usually measured by a standard error or/and a confidence interval. In this work, we then present the estimated return levels along with confidence intervals. There are several options to construct confidence intervals for return levels, such as the normal and the log-normal methods, the bootstrap method, and the profile likelihood method (Glotzer et al. 2017). As a result, we choose the bootstrap method that proceeds by resampling the data with replacement from the given samples, relying on computer simulations. The interval is determined by the $100(\alpha/2)$ and $100(1-\alpha/2)$, for $0<\alpha<1$, quantiles of the bootstrap distribution of the exceedance probability.

2.7. Statistical packages

This study uses the R programming with several packages; “extRemes,” “texmex”, and “gnFit” (R Core Team 2019, Gilleland and Katz 2016) to obtain the estimated parameters and return levels, to produce the mean residual life plots, and to test the distribution, respectively.

3. Results

3.1. Descriptive statistics

The PM_{10} concentrations are recorded hourly, so we divide the data into two types; the 24-average data and the daily maximum data, the hourly maximum of the day. Figures 2-3 show the PM_{10} data recorded over time (2008-2019). The time series plots reveal missing values in some stations. The summary of statistics in Tables 2-3 shows that station 24t located in Saraburi has a remarkably high level of PM_{10}. Stations 35t, 36t, 40t, 67t, 69t, and 73t located in the north of Thailand also have high concentrations compared to the stations 43t, 44t, and 63t located in the south. It is because the southern region of Thailand has a long wet season than the other parts of Thailand.

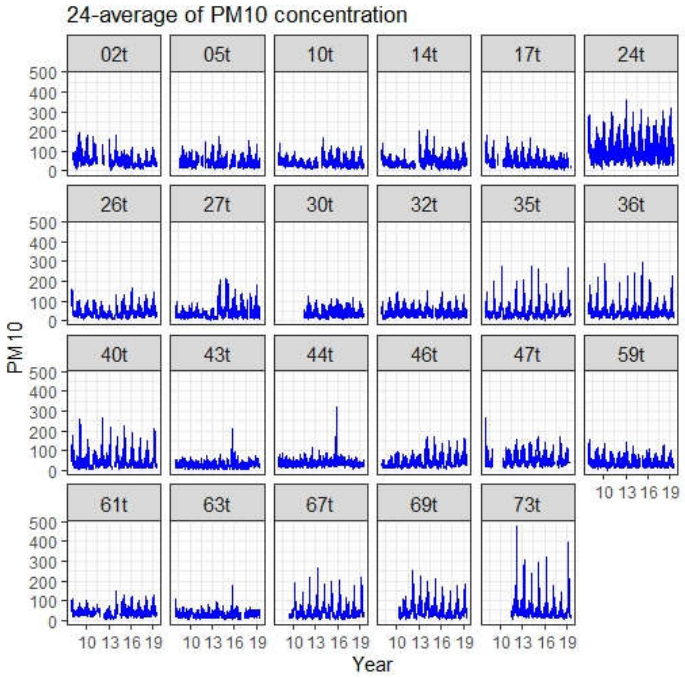

Figure 2 Yearly plots of PM_{10} concentration: 24-average data

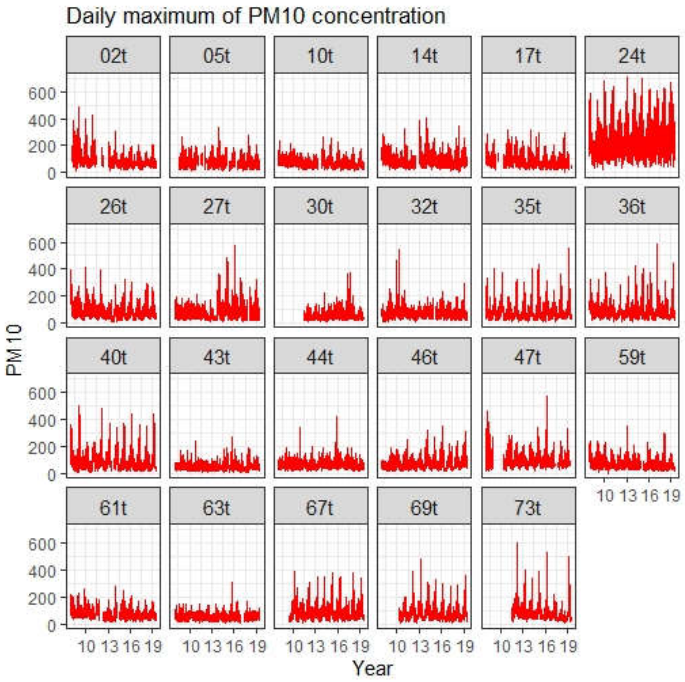

Figure 3 Yearly plots of PM_{10} concentration: daily maximum data

Table 2 Summary of statistics of the 24-average of PM_{10} concentration in Thailand

Code	Min	Max	Median	Mean	SD	Quantiles		
						90%	95%	99%
02t	5.3	193.4	46.1	52.4	26.6	87.5	105.9	140.8
05t	4.0	172.5	35.2	40.8	20.9	68.6	81.0	111.1
10t	3.7	169.0	34.0	38.1	19.2	62.7	74.4	104.2
14t	4.0	210.5	34.8	40.8	23.7	70.4	88.5	129.0
17t	5.0	180.1	35.3	41.2	21.6	69.7	84.2	117.2
24t	10.9	357.9	86.4	98.3	49.8	165.4	196.7	246.2
26t	3.7	167.2	34.9	41.4	24.2	75.6	91.7	120.5
27t	4.3	216.0	29.6	38.4	27.7	75.2	97.5	144.6
30t	3.0	125.3	29.0	33.5	16.5	56.4	67.7	88.0
32t	5.0	155.0	33.1	37.9	18.7	63.8	76.0	99.1
35t	3.8	274.8	34.0	44.1	32.4	85.1	108.8	169.7
36t	5.3	296.0	36.9	46.3	31.5	84.2	110.7	168.0
40t	3.0	265.3	35.0	45.5	33.0	92.7	110.0	154.1
43t	4.3	209.4	24.0	26.2	12.6	40.9	47.6	66.1
44t	8.1	322.5	35.8	37.2	14.5	51.5	59.0	77.5
46t	5.8	171.3	35.2	42.2	25.7	76.6	94.8	129.6
47t	8.0	264.7	45.2	52.2	26.3	86.8	101.6	135.1
59t	2.0	159.3	32.2	37.8	19.9	66.6	78.7	102.2
61t	2.0	149.4	34.1	37.7	19.0	62.9	75.2	101.9
63t	3.8	178.6	26.5	28.6	12.4	43.8	49.6	64.0
67t	3.2	263.5	30.4	41.2	32.0	82.3	110.7	160.7
69t	3.2	249.0	35.4	46.7	33.1	92.5	111.7	163.6
73t	5.5	479.1	37.3	54.0	52.7	100.0	163.3	288.0

3.2. Return levels

The data in each station are quite different, so it is challenging to justify the threshold. We aim to use the same threshold in every station as long as it follows the GPD with not too little data over the threshold. In this work, we initially use the MRL plots to provide some sensible thresholds for each station. Then we justify the thresholds together with the estimated parameters of the GPD model fitted to the PM_{10} data over the selected threshold by using the GoF tests to see if the data above those thresholds follow the GPD. The MRL plots for the two data sets are shown in Figures 4-5. For the 24-average data, the GoF tests suggest that the data above the threshold 90 follow the GPD in most stations except 24t and 73t. Due to the very high values of data in these stations, we need to use a higher threshold. As a result, 120 is chosen to be the threshold for these two stations. For the daily maximum data, we choose 150 to be the optimal threshold for most stations except 24t, 36t, and 73t use 180 as the threshold.

The estimated parameters of GPD over the selected thresholds using maximum likelihood estimation, exceedance numbers, and the return levels are shown in Tables 4-7. Nonetheless, the 24-average data in stations 30t, 43t, 44t, and 63t have not many numbers of exceedances when the threshold is 90, so we exclude these stations as the estimated parameters obtained from small data are non-informative because of high variance. For daily maximum data, the high values of PM_{10} are more available even though higher thresholds are applied. Therefore, we include stations 30t and 44t

in the study. The return levels of the daily maximum data are much higher than the 24-average data. We also illustrate the return levels on the locations of the stations shown in Figures 6-7.

We estimate not only the return levels but also its uncertainty using 95% confidence intervals using the bootstrap method shown in Figures 8-9. The confidence intervals of the 24-average data are relatively narrower than the daily maximum data. Generally speaking, it means that they have less uncertainty than the other result. It is understandable as the daily maximum is more likely to fluctuate from day-to-day.

Table 3 Summary of statistics of the daily maximum of PM_{10} concentration in Thailand

Code	Min	Max	Median	Mean	SD	Quantile		
						90%	95%	99%
02t	9.0	489.0	76.0	86.2	43.6	140.0	172.0	233.0
05t	6.0	336.0	65.0	72.9	34.9	119.0	142.0	184.4
10t	7.0	263.0	62.0	67.5	30.1	107.0	126.0	169.4
14t	6.0	408.0	73.0	82.1	40.6	132.0	162.0	224.6
17t	5.0	316.0	66.0	76.4	39.4	128.0	152.0	214.0
24t	21.0	705.0	197.0	217.3	103.7	356.0	417.0	544.1
26t	5.0	410.0	70.0	80.6	44.0	140.0	163.0	220.5
27t	9.0	576.0	59.0	74.9	51.5	140.0	180.0	264.8
30t	3.0	370.0	47.0	55.5	30.4	93.0	109.0	154.9
32t	5.0	542.0	58.0	66.3	34.2	109.0	128.1	176.9
35t	8.0	557.0	61.5	78.9	55.6	146.0	187.0	293.7
36t	10.0	590.0	73.0	85.1	49.9	143.0	177.0	275.0
40t	3.0	505.0	71.0	86.2	60.8	163.3	201.0	310.0
43t	8.0	270.7	40.0	44.8	21.8	66.0	79.0	133.6
44t	16.0	423.6	59.0	62.5	25.1	90.0	104.0	144.0
46t	12.0	353.3	64.0	73.6	40.0	126.0	148.0	214.0
47t	8.0	569.0	84.0	93.9	44.7	146.0	173.1	258.9
59t	2.0	349.0	60.0	68.2	33.5	112.0	134.0	179.0
61t	2.0	275.0	63.0	67.5	32.2	108.0	128.0	173.7
63t	7.0	306.7	46.0	49.3	20.4	72.0	83.0	117.3
67t	7.0	388.0	66.0	80.2	49.9	143.0	176.0	255.3
69t	5.0	479.2	62.0	76.6	51.9	142.0	174.0	274.0
73t	11.0	605.0	71.0	88.7	67.6	151.0	222.0	391.4

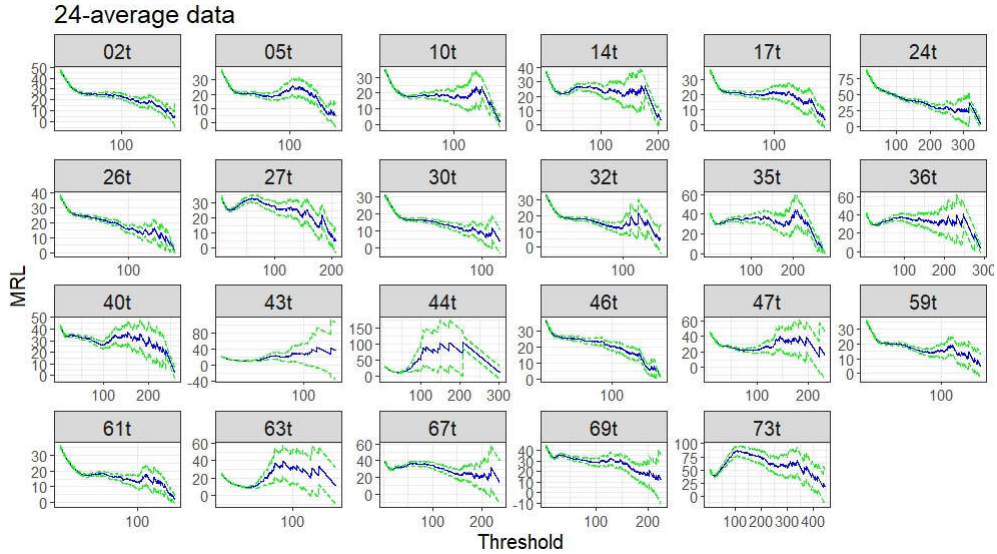

Figure 4 Mean residual life plots of PM_{10}: 24-average data

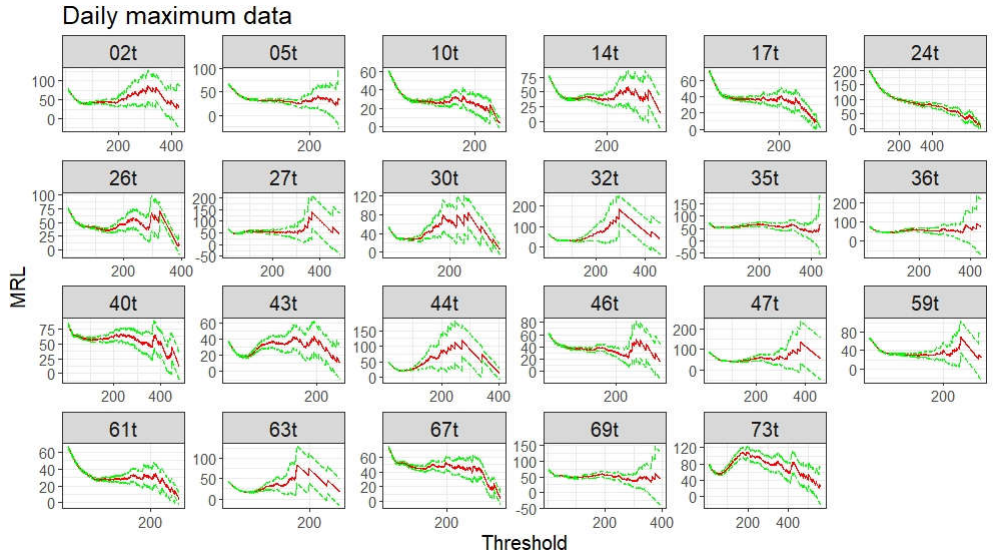

Figure 5 Mean residual life plots of PM_{10}: daily maximum data

Table 4 Parameter estimates, number of exceedances and return levels of 24-average data using threshold 90

Code	Parameters		p-value		Number of	Return levels			
	Scale	Shape	W^2	A^2	exceedances	2-year	5-year	10-year	20-year
02t	29.54	−0.20	0.94	0.93	293	173.75	184.42	191.28	197.25
05t	20.98	0.01	0.46	0.28	91	155.15	174.74	189.57	204.39
10t	19.46	−0.06	0.66	0.54	88	140.66	155.51	166.25	176.58
14t	29.04	−0.11	0.34	0.47	171	174.59	191.52	203.20	214.00
17t	21.09	−0.03	0.84	0.57	131	155.53	172.81	185.57	198.07
26t	20.57	−0.17	0.54	0.64	212	146.27	155.67	161.87	167.39
27t	31.92	-0.13	0.70	0.77	243	187.14	203.96	215.43	225.93
30t	-	-	-	-	-	-	-	-	-
32t	12.08	0.05	0.44	0.41	78	124.88	137.75	147.88	158.34
35t	36.53	−0.02	0.53	0.44	356	235.42	267.59	291.92	316.26
36t	40.73	−0.08	0.98	0.99	346	233.11	259.21	277.75	295.30
40t	25.61	0.06	0.03	0.04	428	218.57	250.26	275.47	301.78
43t	-	-	-	-	-	-	-	-	-
44t	-	-	-	-	-	-	-	-	-
46t	27.23	−0.23	0.67	0.53	234	158.58	167.79	173.56	178.46
47t	20.08	0.10	0.68	0.57	264	192.51	221.33	244.90	270.12
59t	13.29	0.07	0.89	0.90	90	131.59	146.86	159.07	171.86
61t	15.90	−0.09	0.29	0.34	74	128.62	139.69	147.50	154.85
63t	-	-	-	-	-	-	-	-	-
67t	41.06	−0.17	0.79	0.73	295	210.81	227.95	239.23	249.23
69t	29.70	−0.02	0.53	0.51	347	213.72	238.14	256.28	274.14

Table 5 Parameter estimates, number of exceedances and return levels of 24-average data using threshold 120

Code	Parameters		p-value		Number of	Return levels			
	Scale	Shape	W^2	A^2	exceedances	2-year	5-year	10-year	20-year
24t	49.28	−0.14	0.12	0.18	1,179	303.91	323.65	336.95	348.99
73	103.3	−0.20	0.95	0.97	205	403.43	441.28	465.57	468.65

Table 6 Parameter estimates, number of exceedances and return levels of daily maximum data using threshold 150

Code	Parameters		p-value		Number of	Return levels			
	Scale	Shape	W^2	A^2	exceedances	2-year	5-year	10-year	20-year
02t	44.42	0.04	0.25	0.17	250	342.85	390.77	428.09	466.37
05t	31.88	−0.04	0.29	0.23	116	244.89	271.33	291.33	311.33
10t	39.04	−0.22	0.74	0.72	65	225.68	244.54	256.52	266.83
14t	41.54	−0.02	0.60	0.33	241	307.09	341.46	366.84	391.70
17t	39.95	−0.06	0.40	0.20	196	281.60	310.30	331.00	350.88
26t	30.81	0.16	0.56	0.27	294	323.91	383.03	434.16	491.59
27t	57.46	−0.02	0.88	0.68	315	377.90	426.30	462.38	498.00
30t	46.84	0.15	0.41	0.53	32	267.27	331.14	385.81	446.64
32t	28.22	0.33	0.97	0.75	91	283.46	360.14	435.45	529.88
35t	62.15	0.01	0.18	0.12	381	415.23	472.96	516.63	560.30
40t	55.48	0.04	0.38	0.20	509	427.25	489.78	538.67	588.99
43t	-	-	-	-	-	-	-	-	-
44t	39.49	0.31	0.49	0.41	30	240.02	310.78	379.08	463.44
46t	42.06	−0.11	0.32	0.23	185	275.12	300.3	317.79	334.05
47t	41.54	0.12	0.33	0.22	268	372.48	437.84	492.16	551.07
59t	32.74	−0.01	0.39	0.32	103	246.44	275.92	298.13	320.26
61t	34.78	−0.10	0.76	0.63	71	231.99	255.37	271.72	287.00
63t	-	-	-	-	-	-	-	-	-
67t	55.28	−0.09	0.22	0.25	290	340.14	374.25	398.31	420.96
69t	53.91	−0.01	0.10	0.11	264	366.69	413.94	449.41	484.63

Table 7 Parameter estimates, number of exceedances and return levels of daily maximum data using threshold 180

Code	Parameters		p-value		Number of	Return levels			
	Scale	Shape	W^2	A^2	exceedances	2-year	5-year	10-year	20-year
24t	116.1	−0.14	0.78	0.38	2,393	651.56	693.91	722.49	748.39
36t	65.06	−0.05	0.93	0.92	195	389.82	438.09	473.09	506.82
73t	132.7	−0.24	0.39	0.20	185	519.50	562.79	589.87	612.88

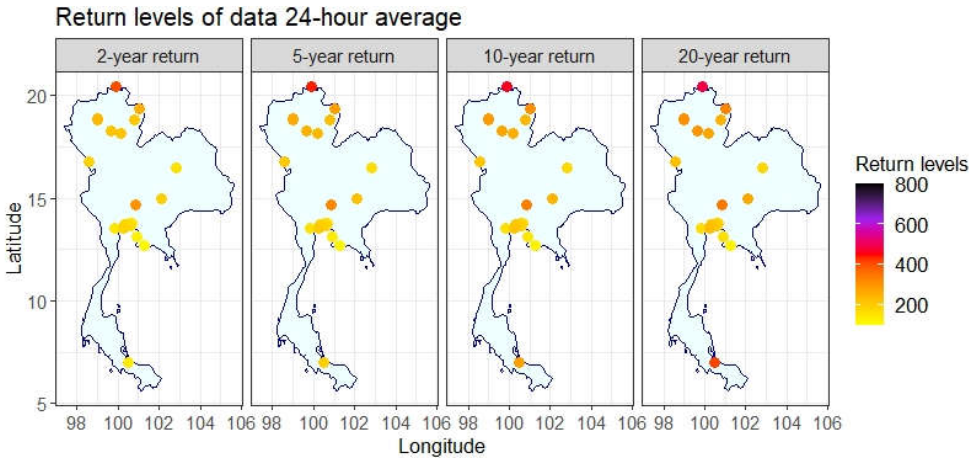

Figure 6 Return levels of PM_{10}: 24-average data

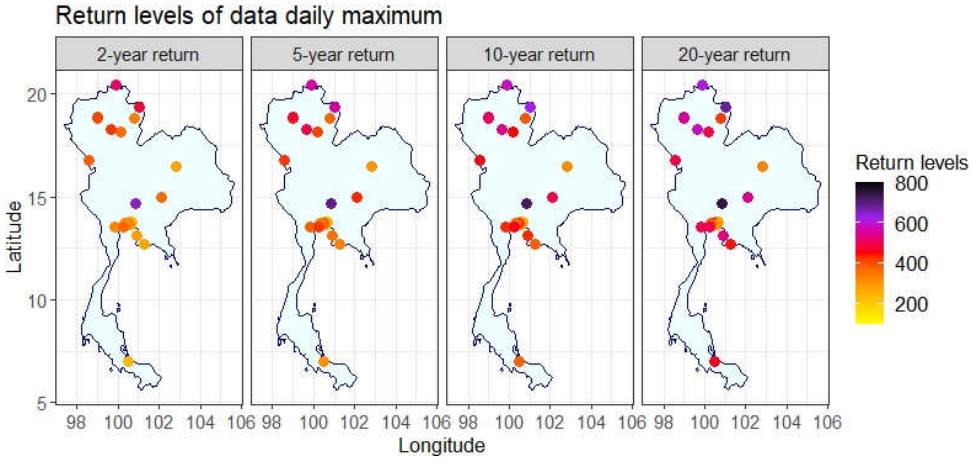

Figure 7 Return levels of PM_{10}: daily maximum data

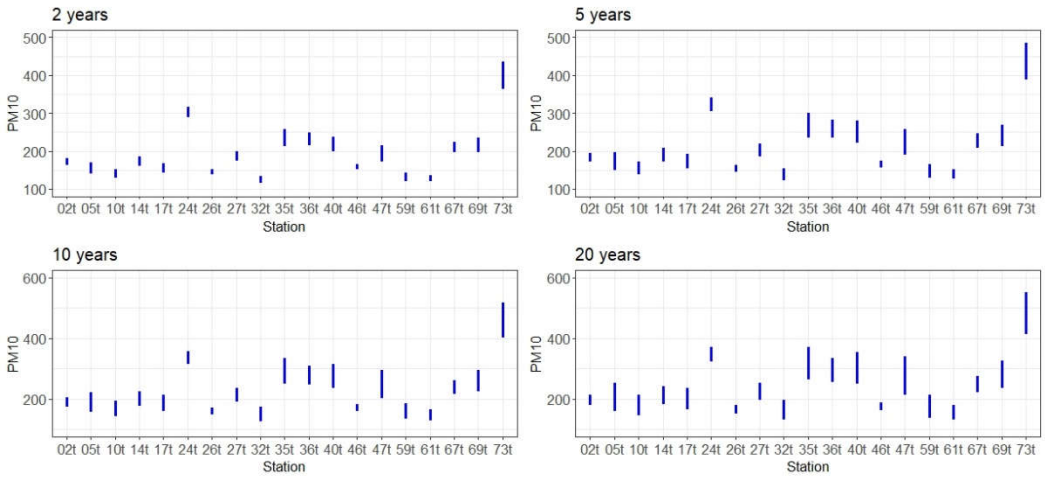

Figure 8 95% confidence intervals of return levels of PM_{10}: 24-average data

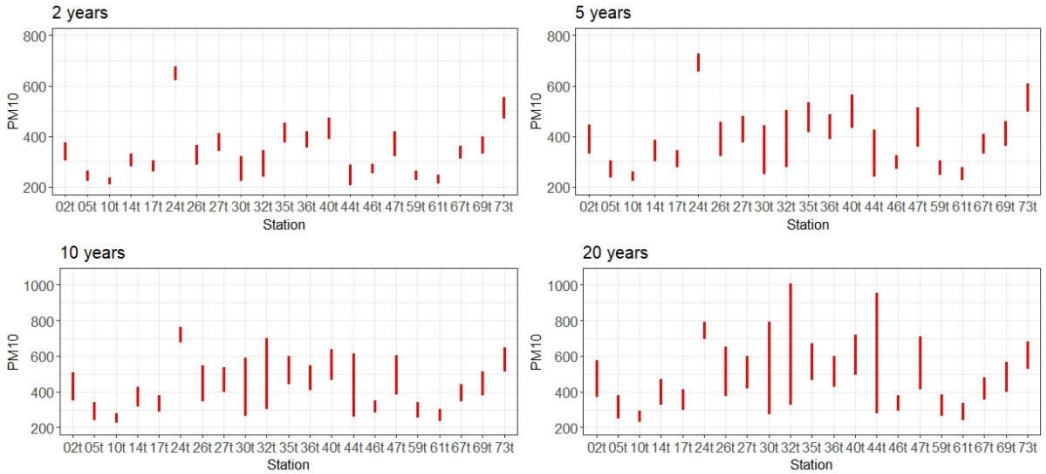

Figure 9 95% confidence intervals of return levels of PM_{10}: daily maximum data

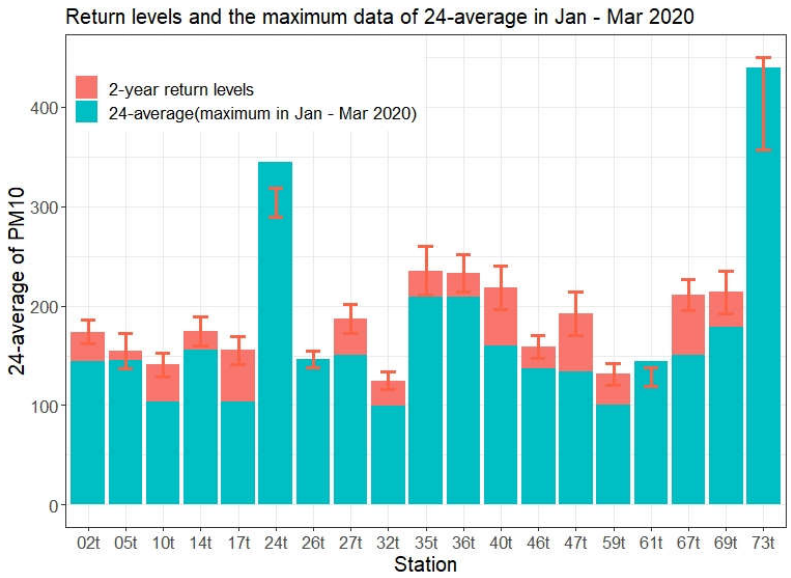

Figure 10 Comparison between 2-year return levels of the maximum of 24-average data in January-May 2020

4. Discussion

The POT approach and the GPD modeling are applied to provide the extreme events of PM_{10} in the future. We study the extreme events of PM_{10} concentrations in two representations, 24-average, and the daily maximum. It is important to note that the time series plots in Figures 2-3 suggesting trends and seasonal variation in the data. However, removing trends and seasonality leads to a small number of data and even smaller when considering the data exceeds the selected threshold. It leads to a limited number of stations to study. As we aim to investigate the extremes of PM_{10} across the country so that we relax the assumption of independence.

One of the challenging issues in extreme value analysis is to determine a threshold. There are no promising methods to assess an optimal threshold. We apply the graphical method, the MRL plot

together with the goodness of fit with Cramer-van Miss and Anderson-Darling statistics, to test chosen thresholds based on parameter estimates obtained from MLE. We found that the thresholds of 90 and 150 are plausible in most stations for 24-average data and daily maximum data, respectively, except stations 24t, 73t, and 36t, where 120 and 180 thresholds are used. Using these thresholds, we can still have a good number of exceedances to model without worrying about the bias and variance in parameter estimation, which can lead to poor return level estimation. Besides, we choose to use the same threshold for as many stations as long as the GoF tests are valid for convenience and simplicity. The parameter estimates are then used to estimate return levels of 2, 5, 10, and 20 years and their 95% confidence intervals to show the predictive extreme values and uncertainty in the future. In this work, confidence intervals are determined by the bootstrap distribution of the exceedance probability. The profile likelihood (Glotzer et al. 2017), as well as the Bayesian approach (Renard et al. 2006), can be alternative methods to govern confidence intervals.

A return level can be useful in indicating how often extreme events are likely to occur. In other words, the return levels in 2, 5, 10, and 20 years shown in Tables 4-7 are expected to be exceeded on average once in those years. The return levels of 24-average at station 73t in Chiang Rai are the highest, and next is station 24t in Saraburi. For the daily maximum data, the return levels of these two stations are the other way around. The 2-year return periods of 24-average can reach over 400 and can be almost 600 in 20 years. These levels are considered hazardous levels that can cause serious health effects.

We validate our results by comparing the 2-year return levels of 24-average data to the latest real data available online in http://air4thai.pcd.go.th/webV2/index.php. The comparison is shown in Figure 10. We can see that the estimates of returns are higher than the recent data in most stations. Interestingly, the recent data at stations 24t, 26t, and 73t already surpass the estimates. Precisely, the current extreme value in February 2020 at station 24t is 345 while our estimated return level 303.91, and in March 2020, the recent extreme value at station 73t is 439 while our estimated return is 403.43. It is important to note that the 2-year return levels are the expected extreme values in the 2-year period, so the 24-average of PM_{10} might be higher. From this result, we could say that the extreme value analysis presented in this work is reliable on account of the recent extreme PM_{10} concentration in Thailand.

5. Conclusions

In this work, we present the extreme value analysis of daily PM_{10} concentration using the POT approach modeled by GPD. The daily concentration is expressed in two forms; the 24-average and a maximum of 24 hours, called the daily maximum data. Determining a threshold for GPD is essential as the estimated parameters from a small threshold can be biased due to the deviation of the distribution from the GPD. In contrast, a high threshold leads to a small sample size, hence high variance. Therefore, we balance between bias and deviation to obtain an optimal threshold. We use different thresholds for different data sets and stations based on MRL plots and GoF tests to obtain the feasible estimated return levels. The return levels and their uncertainties are quantified to assess the extreme events in the future. Hence, they can be used to indicate the risk of extreme events. The stations that have high return levels should be cautiously monitored.

Acknowledgments

This research is supported by the Data Science Research Center, Department of Statistics, Faculty of Science, Chiang Mai University. We would like to thank Pollution Control Department,

Air Quality and Noise Management Bureau, The Ministry of Natural Resources and Environment of Thailand for valuable data.

References

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

Brabson BB, Palutikof JP. Tests of the generalized Pareto distribution for predicting extreme wind speeds. J Appl Meteorol Climatol. 2000; 39(9): 1627-1640.

Brook RD, Rajagopalan S, Pope III CA, et al. Particulate matter air pollution and cardiovascular disease: an update to the scientific statement from the American Heart Association. Circulation. 2010; 121(21): 2331-2378.

Chen G, Balakrishnan N. A general purpose approximate goodness-of-fit test. J Qual Tech. 1995; 27(2): 154-161.

Choulakian V, Stephens MA. Goodness-of-fit tests for the generalized Pareto distribution. Technometrics. 2001; 43(4): 478-484.

Coles S. An introduction to statistical modeling of extreme values. London: Springer Science & Business Media; 2001.

Dupuis DJ. Exceedances over high thresholds: a guide to threshold selection. Extremes. 1999; 1(3): 251-261.

Engeland K, Hisdal H, Frigessi A. Practical extreme value modelling of hydrological floods and droughts: a case study. Extremes. 2004; 7(1): 5-30.

Gilleland E, Katz RW. extRemes 2.0: An extreme value analysis package in R. J Stat Softw. 2016; 72(8): 1-39.

Gilli M, Këllezi E. An application of extreme value theory for measuring financial risk. Comput Econ. 2006; 27(2): 207-228.

Glotzer D, Pipiras V, Belenky V, Campbell B, Smith T. Confidence intervals for exceedance probabilities with application to extreme ship motions. Revstat Stat J. 2017; 15(4): 537-563.

Grimshaw SD. Computing maximum likelihood estimates for the generalized Pareto distribution. Technometrics. 1993; 35(2): 185-191.

Gyarmati-Szabó J, Bogache LV, Chen H. Nonstationary POT modelling of air pollution concentrations: Statistical analysis of the traffic and meteorological impact. Environmetrics. 2017; 28(5): e2449.

Hosking JRM, Wallis JR. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics. 1987; 29(3): 339-349.

Leadbetter MR. Extremes and local dependence in stationary sequences. Z Wahrscheinlichkeit. 1983; 65: 291-306.

Lin H, Liu T, Xiao J, et al. Hourly peak $PM_{2.5}$ concentration associated with increased cardiovascular mortality in Guangzhou, China. J Expo Sci Environ Epidemiol. 2017; 27(3): 333-338.

Martins LD, Wikuats CF, Capucim MN, et al. Extreme value analysis of air pollution data and their comparison between two large urban regions of South America. Weather Clim Extrem. 2017; 18: 44-54.

Masseran N, Razali AM, Ibrahim K, Latif MT. Modeling air quality in main cities of Peninsular Malaysia by using a generalized Pareto model. Environ Monit Assess. 2016; 188(1): 65.

Osman YZ, Fealy R, Sweeney JC. Modelling extreme temperatures in Ireland under global warming using a hybrid peak-over-threshold and a generalised Pareto distribution approach. Int J Global Warm. 2015; 7(1): 21-47.

Pickands J 3rd. Statistical inference using extreme order statistics. Ann Stat. 1975; 3(1): 119-131.

Pollution Control Department. Booklet on Thailand State of Pollution 2018. Pollution Control Department Ministry of Natural Resources and Environment. Bangkok: S.Mongkon Press Limited Partnership; 2019.

Pope CA 3rd, Burnett RT, Thun MJ, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Jama. 2002; 287(9): 1132-1141.

Pothirat C, Chaiwong W, Liwsrisakun C, et al. The short-term associations of particular matters on non-accidental mortality and causes of death in Chiang Mai, Thailand: a time series analysis study between 2016-2018. Int J Environ Health Res. 2019; 31(5): 538-547.

R Core Team R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria; 2019 [cited 2019 Dec 15]. Available from: http://www.R-project.org/.

Ragulina G, Reitan T. Generalized extreme value shape parameter and its nature for extreme precipitation using long time series and the Bayesian approach. Hydrolog Sci J. 2017; 62(6): 863-879.

Renard B, Lang M, Bois P. Statistical analysis of extreme events in a non-stationary context via a Bayesian framework. Case study with peak-over-threshold data. Stoch Environ Res Risk Assess. 2006; 21(2): 97-112.

Scarrott C, MacDonald A. A review of extreme value threshold estimation and uncertainty quantification. Revstat Stat J. 2012; 10(1): 33-60.

Thompson PA, Cai Y, Reeve DE, Stander J. Automated threshold selection methods for extreme wave analysis. Coast Eng. 2009; 56(10): 1013-1021.

Vichit-Vadakan N, Ostro BD, Chestnut LG, et al. Air pollution and respiratory symptoms: results from three panel studies in Bangkok, Thailand. Environ Health Perspect. 2001; 109 (suppl 3): 381-387.

Viroj W. PM_{10} in the atmosphere and incidence of respiratory illness in Chiang Mai during the smoggy pollution. Stoch Environ Res Risk Assess. 2008; 22(3): 437- 440.

World Health Organization. Air quality guidelines. Global Update 2005. Particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Germany: Druckpartner Moser; 2006.

Worms J, Worms R. Estimation of second order parameters using probability weighted moments. ESAIM P S. 2012; 16: 97-113.

Zikova N, Masiol M, Chalupa DC, Rich DQ, Ferro AR, Hopke PK. Estimating hourly concentrations of $PM_{2.5}$ across a metropolitan area using low-cost particle monitors. Sensors. 2017; 17(8): 1922.