Indonesian Journal of Pure and Applied Mathematics
http://journal.uinjkt.ac.id/index.php/inprime
Vol. 1, No. 1 (2019), pp. 48-56
ISSN 2686-5335
doi: 10.15408/inprime.v1i1.12839

Statistical Modelling of Extreme Data of Air Pollution in Pekanbaru City

Ari Pani Desvina , Elfira Safitri, and Ade Novia Rahma
Mathematics Department, Science and Technology Faculty, UIN Sultan Syarif Kasim Riau
Jl. HR. Soebrantas No. 155 Simpang Baru, Panam, Pekanbaru, 28293
Email: elfira.safitri@uin-suska.ac.id

Abstract

Air pollution is a phenomenon that is often discussed, especially regarding air quality in urban areas. This has become a major contributor to health problems and environmental issues in Asian countries, such as Indonesia, especially Riau Province. The event of forest fires is one of the many events that occurred in Indonesia, especially Riau Province which harmed the population of Indonesia and neighboring countries. The phenomenon of forest forestry generally occurs due to a shift in the season towards drought and can occur in areas prone to forest fires. Therefore, it is necessary to know the model of air pollution distribution by Particulate Matter (PM10) in Pekanbaru City. This study aims to obtain the distribution model for daily air pollution PM10 in Pekanbaru City from 2014 to February 2015. Data were taken from three stations i.e. Sukajadi Station, Tampan Station, and Kulim Station. Four distributions will be tested i.e. Log Pearson III distribution, Gumbel distribution, Generalized Pareto Distribution, and Generalized Extreme Value (GEV) distribution. We test the goodness of fit from these distribution using the Kolmogorov-Smirnov and the Anderson-Darling tests. The result shows that the Generalized Extreme Value (GEV) distribution model was better than the Log Pearson III, Gumbel and Generalized Pareto distribution models for modeling city air pollution data Pekanbaru with three stations namely Sukajadi, Tampan, and Kulim.

Keywords: Anderson-Darling; Generalized Extreme Value (GEV); Kolmogorov-Smirnov.

Abstrak

Pencemaran udara merupakan satu fenomena yang sering dibicarakan, apalagi mengenai kualitas udara di daerah perkotaan. Hal ini menjadi penyumbang utama tentang masalah kesehatan dan isu lingkungan hidup di negara-negara Asia, seperti Negara Indonesia khususnya Provinsi Riau. Peristiwa kebakaran hutan merupakan salah satu peristiwa yang banyak terjadi di Indonesia khususnya Provinsi Riau yang berdampak negatif terhadap penduduk Indonesia dan negara tetangga. Fenomena kebarakan hutan pada umumnya terjadi karena adanya pergeseran musim kearah kemarau dan dapat terjadi di daerah rawan kebakaran hutan. Oleh karena itu, perlu diketahui model distribusi pencemaran udara oleh *Particulate Matter* (PM10) Kota Pekanbaru. Penelitian ini bertujuan untuk mendapatkan model distribusi data harian pencemaran udara oleh *Particulate Matter* (PM10) Kota Pekanbaru Tahun 2014 sampai Februari 2015 dengan tiga stasiun yaitu stasiun sukajadi, stasiun tampan dan stasiun kulim. Adapun distribusi yang digunakan adalah distribusi Log Pearson III, distribusi Gumbel, Distribusi *Generalized Pareto* dan distribusi *Generalized Extreme Value* (GEV). Berdasarkan pembahasan uji kebaikan (*Goodness of Fit*) yaitu uji Kolmogorov-Smirnov dan Anderson-Darling, maka diperoleh bahwa model distribusi *Generalized Extreme Value* (GEV) lebih baik dari pada model distribusi Log Pearson III, Gumbel dan

Generalized Pareto untuk memodelkan data pencemaran udara kota Pekanbaru dengan tiga stasiun yaitu Sukajadi, Tampan dan Kulim.

Kata Kunci: Anderson-Darling; Generalized Extreme Value (GEV); Kolmogorov-Smirnov.

1. INTRODUCTION

Air is the most important substance after water in providing life on earth [1]. Allah the Almighty has created the air between the winds, as explained in the Quran Surah Ar-Rum (48):

اللَّهُ الَّذِي يُرْسِلُ الرِّيَاحَ فَتُثِيرُ سَحَابًا فَيَبْسُطُهُ فِي السَّمَاءِ كَيْفَ يَشَاءُ
وَيَجْعَلُهُ كِسَفًا فَتَرَى الْوَدْقَ يَخْرُجُ مِنْ خِلَالِهِ ۖ فَإِذَا أَصَابَ بِهِ مَنْ يَشَاءُ مِنْ
عِبَادِهِ إِذَا هُمْ يَسْتَبْشِرُونَ

"*Allah sends the winds that stir up clouds and then He spreads them in the sky as He pleases and splits them into different fragments, whereafter you see drops of rain pouring down from them. He then causes the rain to fall on whomsoever of His servants He pleases, and lo, they rejoice at it*".

Air pollution can cause various kinds of diseases to humans such as lung disease, asthma, anemia, and other diseases. The main air pollutant gases are carbon monoxide, carbon dioxide, nitrogen oxides, nitrogen dioxide, particulate matter (PM10) and so on. PM10 is ash or dust less than 10 μm in diameter which can have a more serious effect on human, animal and plant health risks compared to larger particles that are generally formed from immovable sources such as vehicles (vehicle eczos) [2].

BNPB data also states, as many as 49.591 people in Riau suffer from smoke-related diseases such as upper respiratory tract infections (URI/URTI), pneumonia, asthma, eye, and skin irritation. In addition to causing illness, haze in Riau Province, especially Pekanbaru City, has disrupted community activities, such as all educational activities in Riau Province, especially Pekanbaru City, were stopped because of this pollution. One of the universities that stop academic activities was Sultan Syarif Kasim Riau State Islamic University for 4 days. Besides, visibility on the highway is only ± 200 meters, which can disturb the driver's activity.

Other impacts of air pollution are ozone depletion, smoke, acid rain and global warming [3]. Also, ash, smoke, fog, steam or other materials produced by air pollution can obstruct eyesight. Besides the impact on humans, negative impacts also occur on plants and animals. The impact on plants is to cause stunted plant growth, this is caused by obstruction of sunlight to get to the leaves so that the process of photosynthesis is reduced and the level of carbon dioxide uptake is reduced. In animals can cause interference with the respiratory system of animals. Animals that eat fluoridated grass and leaves will cause abnormal bone shape [2].

The problem of air pollution is a phenomenon that is often discussed. Air quality in the urban environment is increasingly tapered and is a major contributor to health problems and environmental issues in Asian. Some mathematical models have been widely used to determine the patterns of movement from air pollution data. In this paper, we will explore some statistical models using some distribution functions i.e. Pearson III log Gumbel generalized Pareto and Generalized Extreme Value (GEV) distribution that appropriate to determine the patterns of movement of PM10 in Pekanbaru City.

2. METHOD

The data used in the study is air pollution data particularly data related to *particulate matter* (PM_{10}). This data is collected daily from January 2014 to February 2015 from Environmental services Pekanbaru city, with three monitoring stations, i.e. Sukajadi, Kulim, and Tampan. We use the distribution of Log Pearson III, Gumbel, Generalized Pareto, and Generalized Extreme Value (GEV) to model the distribution of this air pollution of particulate matter by using easyfit software. The steps in conducting this research are summarized as follows:

I. Select Log Pearson III, Gumbel, Generalized Pareto, and Generalized Extreme Value (GEV) distributions:

Log Pearson III distribution is one of the Pearson distribution family. Log Pearson III distribution refers to Gamma distribution. This distribution is very close to Generalized Extreme Value where it uses three parameters, i.e. scale, location, and shape parameters [4]. The probability density function of Log Pearson III distribution can be expressed as follows [5]:

$$f(x) = \frac{1}{|\beta|\Gamma(\alpha)}\left[\frac{\ln x-\delta}{\beta}\right]^{\alpha-1} exp\left(-\left(\frac{\ln x-\delta}{\beta}\right)\right) \quad \text{where } \alpha > 0, \beta \neq 0\ .$$

Gumbel distribution is firstly introduced by Emil Gumber, a mathematician from Germany. Gumbel distribution is a special case of extreme value distribution where the location parameter is equal to zero [7]. The probability density function of Gumbel distribution can be written as [6]:

$$f(x) = \frac{1}{\beta} exp\left[\frac{(\alpha-x)}{\beta}\right] exp\left[-exp\left(\frac{\alpha-x}{\beta}\right)\right], \quad -\infty < x < \infty \text{ where } \beta > 0, \alpha \in \mathbb{R}.$$

Generalized Pareto distribution is one of the continuous distributions that have a probability density function. Muraleedharan and Soares [7] define the density function of Generalized Pareto distribution as follows:

$$f(x;k,\xi,\alpha) = \frac{1}{\alpha}\left(1+\frac{k(x-\xi)}{\alpha}\right)^{-\left(\frac{1}{k}\right)-1}, \text{ for } k \neq 0 \text{ where } x \geq \xi \text{ for } k \geq 0 \text{ and } \ \xi \leq x \leq \xi - \frac{\alpha}{k} \text{ for } \ k < 0.$$

In general, GEV distribution is used to model extreme data that used within the maximum range of specific periods such as on a daily, monthly, and yearly basis. In reality, maximum extreme data is really useful to be used as a reference in preventing the future extreme value [8]. Generalized Extreme Value (GEV) distribution has three parameters, they are scale (σ), location (μ), and shape (ξ) parameters. The probability density function of Generalized Extreme Value (GEV) is expressed as follows:

$$f(x,\mu,\sigma,\xi) = \frac{1}{\sigma}\left\{1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{\frac{1}{\xi}-1} e^{\left\{-1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{\frac{1}{\xi}}}, \xi \neq 0.$$

II. Collect air pollution data in Pekanbaru city
Estimate the parameter values from Log Pearson III, Gumbel, Generalized Pareto, and Generalized Extreme Value (GEV) by using the Maximum Likelihood method. The maximum

likelihood method estimates the unknown parameter based on the observed data. In this research, we estimate the parameters for all distribution models using easyfit software.

III. Develop a statistical model based on the estimated parameters from the data.

IV. Conduct goodness of fit test using Kolmogorov-Smirnov and Anderson Darling tests to determine the best-fitted model among the proposed four distribution model.

The statistic of Kolmogrov-Smirnov test is $D = \max[D^+, D^-]$ where $D^+ = \max_{i=1,\dots,n}\left[\frac{i}{n} - F(x_i)\right]$, $D^- = \max_{i=1,\dots,n}\left[F(x_i) - \frac{(i-1)}{n}\right]$, and $F(x_i)$ is cumulative function. *D* value based on the maximum distance between D^+ and D^-. The statistic of Anderson-Darling test is

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix}(2i-1)\ln F(x_i) + \\ (2n+1-2i)\ln\left(1 - F(x_i)\right)\end{bmatrix}.$$

The model is said to fit the data well if the statistic value of Kolmogorov Smirnov and Anderson-Darling tests are minima [9].

3. RESULTS AND DISCUSSIONS

3.1. Descriptive Statistics for Air Pollution Data in Pekanbaru City

The number of air pollution can increase and decrease at any time. From January 2014 to February 2015, the number of air pollution in Pekanbaru city experiences a variation of decreasing and increasing from month to month. For more details, the following Figure 1 displays the number of daily air pollution in Pekanbaru City. Figure 1 shows that the number of daily air pollution in Pekabaru City from January 2014 to February 2015 for Sukajadi, Tampan, and Kulim stations where the summary of the descriptive statistics presented in Table 1. Table 1 indicates that the lowest number of air pollution occurs in Sukajadi station is 0.01 and reaches the highest to 617.24 in Sukajadi station as well. Furthermore, the means of air pollution in Sukajadi, Tampan, and Kulim stations are 58.55, 28.996, and 86.401, respectively with their corresponding standard deviations are 82.409, 68.977, and 78.42, respectively. Both Figure 1 and Table 1 gives us an overall description of the air pollution data in Pekanbaru city.

Table 1. Descriptive Statistic for the Number of Air Pollution in Pekanbaru City

City	N	Minimum	Maximum	Mean	Std. Deviation
Sukajadi	406	0.01	617.24	58.55	82.409
Tampan	406	1.21	492.56	38.996	68.977
Kulim	424	5.29	596.30	68.401	78.42

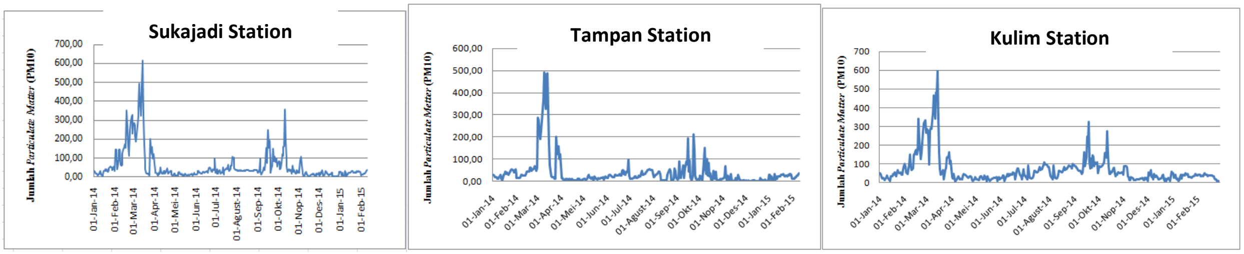


Figure 1. The plot of air pollution data in Sukajadi, Tampan, and Kulim Station

3.2. Parameter Estimation

The parameter estimation using the maximum likelihood method for Log Pearson III, Gumbel, Generalized Pareto, and Generalized Extreme Value (GEV) distributions are shown in Table 2.

Tabel 2. Estimated parameter for Log Pearson III, Gumbel, Generalized Pareto and GEV distributions

Distribution	Parameter		
	Sukajadi Station	**Tampan Station**	**Kulim Station**
Log Pearson III	α = 9,3889	α = 118,27	α = 19,181
	β = -0,34612	β = 0,10529	β = 0,18937
	γ = 6,7479	γ = -9,5011	γ = 0,20681
Gumbel	σ = 64,254	σ = 53,781	σ = 61,144
	μ = 21,461	μ = 7,953	μ = 33,108
Generalized Pareto	k = 0,44718	k = 0,47694	k = 0,31533
	σ = 28,179	σ = 18,904	σ = 37,738
	μ = 7,5775	μ = 2,8548	μ = 13,283
GEV	k = 0,53881	k = 0,56074	k = 0,44493
	σ = 20,343	σ = 13,89	σ = 25,153
	μ = 23,806	μ = 13,813	μ = 34,366

3.3. Modelling Air Pollution in Pekanbaru City

Based on Table 2, the air pollution model for Log Pearson III Distribution for the three stations are

a. Sukajadi station

$$f(x) = \frac{1}{0.34612\Gamma(9.3889)}\left[\frac{\ln x-6.7479}{-0.34612}\right]^{8.3889} e^{\left[-\left(\frac{\ln x-6.7479}{-0.34612}\right)\right]}, \beta > 0. \tag{1}$$

b. Tampan station

$$f(x) = \frac{1}{0.10529x\ \Gamma(118.27)}\left[\frac{\ln x+9.5011}{0.10529}\right]^{117.27} e^{\left[-\left(\frac{\ln x+9.5011}{0.10529}\right)\right]}, \beta > 0. \tag{2}$$

c. Kulim station

$$f(x) = \frac{1}{0.18937x\ \Gamma(19.181)}\left[\frac{\ln x-0.20681}{0.18937}\right]^{18.181} e^{\left[-\left(\frac{\ln x-0.20681}{0.18937}\right)\right]}, \beta > 0. \tag{3}$$

We depict the Log Pearson III distribution for the three stations in Figure 2. Figure 2 shows the histogram of PM10 data and the line expresses the Log Pearson III distribution for three stations based on equation (1) – (3). From these figures, we can see that the line of the Log Pearson III model more than the histogram of PM10 data.

The air pollution model for Gumbel Distribution for the three stations are

a. Sukajadi station

$$f(x) = \frac{1}{64.254}\exp\left(\frac{21.461-x}{64.254}\right)\exp\left(\exp\left(\frac{21.461-x}{64.254}\right)\right), \quad -\infty < x < \infty. \tag{4}$$

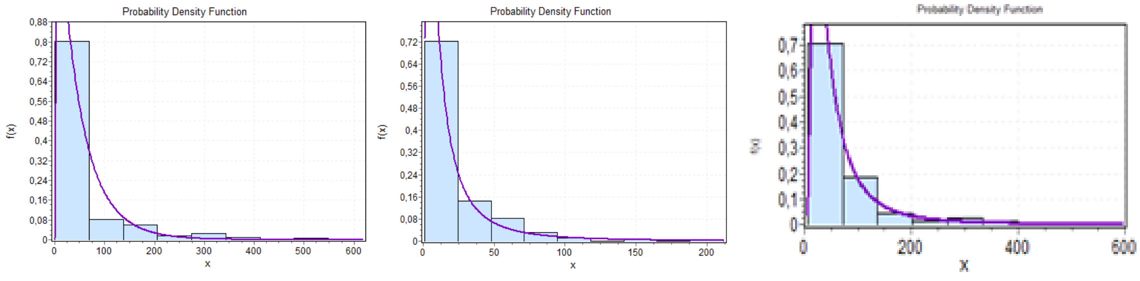
b. Tampan station

$$f(x) = \frac{1}{53.781}\exp\left(\frac{7.953-x}{53.781}\right)\exp\left(\exp\left(\frac{7.953-x}{64.254}\right)\right), \quad -\infty < x < \infty. \tag{5}$$

c. Kulim station

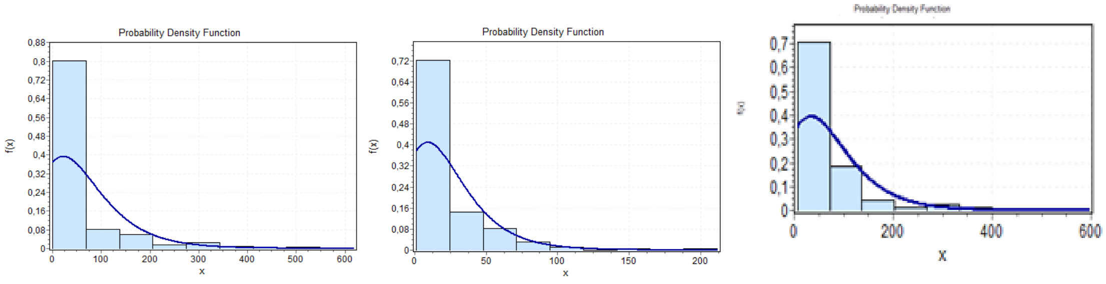
$$f(x) = \frac{1}{61.144}\exp\left(\frac{33.108-x}{61.144}\right)\exp\left(\exp\left(\frac{33.108-x}{64.254}\right)\right), \quad -\infty < x < \infty. \tag{6}$$

We depict the Gumbel distribution for the three stations in Figure 3. We depict the Gumbel distribution for the three stations in Figure 3. Figure 3 shows the histogram of PM10 data and the line expresses the Gumbel distribution for three stations based on equation (4) – (6). From these figures, we can see that the line of the Gumbel model lower than the histogram of PM10 data.



(a) (b) (c)

Figure 2. The plot of Log Pearson III model for air pollution in Sukajadi (a), Tampan (b), and, Kulim Station (c)



(a) (b) (c)

Figure 3. The plot of Gumbel distribution for air pollution in Sukajadi (a), Tampan(b), and, Kulim Station (c)

The air pollution model for Generalized Pareto Distribution for the three stations are

a. Sukajadi Station

$$f(x;k,\xi,\alpha) = \frac{1}{28.179}\left(1+\frac{0.44718(x-7.5775)}{28.179}\right)^{-\left(\frac{1}{0.44718}\right)-1}. \tag{7}$$

b. Tampan Station

$$f(x;k,\xi,\alpha) = \frac{1}{18.904}\left(1+\frac{0.47694(x-2.8548)}{18.904}\right)^{-\left(\frac{1}{0.47694}\right)-1}. \tag{8}$$

c. Kulim Station

$$f(x;k,\xi,\alpha) = \frac{1}{37.738}\left(1+\frac{0.31533(x-13.283)}{37.738}\right)^{-\left(\frac{1}{00.31533}\right)-1}. \tag{9}$$

We can depict the distribution for the three stations as shown in Figure 4. We depict the Generalized Pareto distribution for the three stations in Figure 4. Figure 4 shows the histogram of PM10 data and the line expresses the Generalized Pareto distribution for three stations based on equation (7) – (9). From these figures, we can see that the line of the Generalized Pareto model exceeds the histogram of PM10 data.

The air pollution model for GEV Distribution for the three stations are

a. Sukajadi station

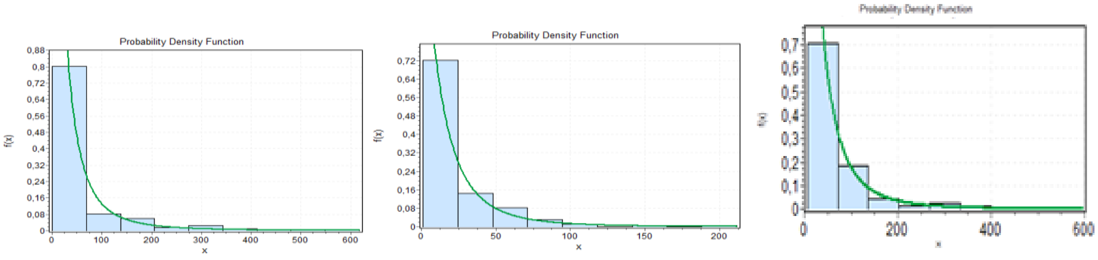
$$f(x,\mu,\sigma,\xi) = \frac{1}{20.343}\left\{1+0.53881\left(\frac{x-23.806}{20.343}\right)\right\}^{0.8559} e^{\left\{-1+0.53881\left(\frac{x-23.806}{20.343}\right)\right\}^{\frac{1}{0.53881}}}. \tag{10}$$

b. Tampan station

$$f(x,\mu,\sigma,\xi) = \frac{1}{13.89}\left\{1+0.56074\left(\frac{x-13.813}{13.89}\right)\right\}^{0.7834} e^{\left\{-1+0.56074\left(\frac{x-13.813}{13.89}\right)\right\}^{\frac{1}{0.56074}}}. \tag{11}$$

c. Kulim station

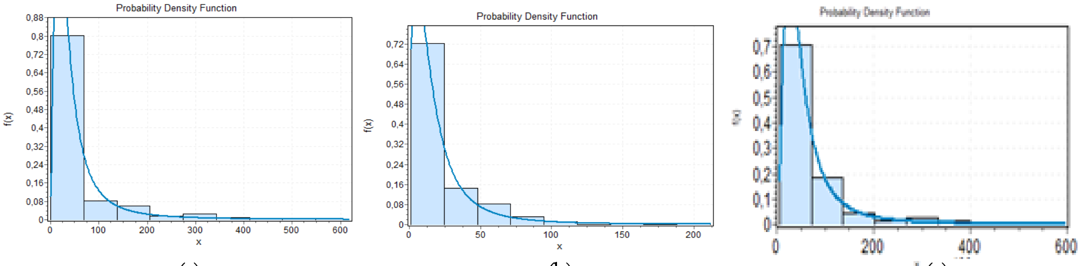
$$f(x,\mu,\sigma,\xi) = \frac{1}{25.153}\left\{1+0.44493\left(\frac{x-34.366}{25.153}\right)\right\}^{1.2475} e^{\left\{-1+\xi\left(\frac{x-34.366}{25.153}\right)\right\}^{\frac{1}{0.44493}}}. \tag{12}$$



(a) (b) (c)

Figure 4. The plot of Generalized Pareto Distribution for air pollution in Sukajadi (a), Tampan(b), and, Kulim Station (c).

We can depict the distribution for the three stations as shown in Figure 5. We depict the GEV distribution for the three stations in Figure 5. Figure 5 shows the histogram of PM10 data and the line expresses the GEV distribution for three stations based on equation (10) – (12). From these figures, we can see that the line of the GEV model more than the histogram of PM10 data.



(a) (b) (c)

Figure 5. The plot of GEV Distribution for air pollution in Sukajadi (a), Tampan(b), and, Kulim Station (c)

3.4. The goodness of Fit Test

To examine the distribution models that fit the observed data among the four distribution models, we use Kolmogorov-Smirnov and Anderson-Darling tests. By using easyfit software, we obtain the results of the goodness of fit test as presented in Table 3, Table 4, and Table 5. According to Table 3, Table 4, and Table 5, it can be seen the statistic values from the Kolmogorov-Smirnov test and Anderson-Darling test. Based on these statistics, the GEV has the smallest statistic than the others. Therefore, we conclude that the best-fitted model distribution for air pollution in Pekanbaru city is GEV distribution.

Table 3. The statistic of Kolmogorov-Smirnov and Anderson-Darling tests on air pollution in Sukajadi Station

Distribusi	Kolmogorov-Smirnov	Anderson-Darling
Log Pearson III	0.16055	13.379
Gumbel	0.26028	44.445
Generalized Pareto	0.08656	79.103
GEV	0.09249	3.3805

Table 4. The statistic of Kolmogorov-Smirnov and Anderson-Darling tests on air pollution in Tampan Station

Distribusi	Kolmogorov-Smirnov	Anderson-Darling
Log Pearson III	0.05615	1.9913
Gumbel	0.32188	54.532
Generalized Pareto	0.0582	79.241
GEV	0.04544	1.3519

Table 5. The statistic of Kolmogorov-Smirnov and Anderson-Darling test on air pollution in Kulim Station

Distribusi	Kolmogorov-Smirnov	Anderson-Darling
Log Pearson III	0.03453	0.59652
Gumbel	0.21934	31.967
Generalized Pareto	0.053	80.243
GEV	0.03178	0.53426

4. CONCLUSION

This research determines the appropriate distribution for air pollution data in Pekanbaru city particularly for the daily data of PM10 measured from 2014 to February 2015 in stations Sukajadi, Tampan, and Kulim using distribution Log Pearson III, Gumbel, Generalized Pareto, and GEV. Visually, based on the graphic of these distributions and histograms the PM10 data, there is no distribution close to the histogram of PM10 data. However, these results are subjective, so we use Kolmogorov-Smirnov and Anderson-Darling tests to test the goodness of fit for these distributions. The result shows that the GEV distribution model is superior to the Log-Pearson III distribution model, Gumbel, and Generalized Pareto. It can be concluded that the GEV model fits the data well to model air pollution of PM10 in Pekanbaru city.

REFERENCES

[1] N. Millington and et al., "The Comparison of GEV, Log-Pearson Type 3 and Gumbel Distributions in the Upper Thames River Watershed under Global Climate Models," Water Resources Research Report Department of Civil and Environmental Engineering the University, 2011.

[2] U. Zaini, Pengenalan Pencemaran Udara. Cetakan kedua. Kuala Lumpur., Dewan Bahasa dan Pustaka, 2000.

[3] T. Godish, Air Quality, 3rd edition, New York: Lewis Publisher, 1997.

[4] M. Jannah, Prediksi Nilai Risiko Severitas Klaim Distribusi Frèchet dan Gumbel, Thesis Mahasiswa Institut Teknologi Bandung, 2015.

[5] K. Arora and V. Singh, "A Comparative Evaluation of the Estimators of the Log Pearson Type (LP) 3 Distribution," *Journal of Hidrology,* vol. 105, pp. 19-37, 1989.

[6] M. Evans and et al., Statistical Distributions, 2nd Edition, New York: Wiley, 1975.

[7] G. Muraleedharan and C. Soares, 2011.

[8] M. Gilli and E. Këllezi, "An Application of Extreme Value Theory for Measuring Financial Risk," *Computational Economics,* vol. 27, no. 2, pp. 207-228, 2006.

[9] H. Thode, "Characteristic and Moment Generating Functions of Generalized Pareto (GP3) and Weibull Distribution," *Journal of Scientific Research and reports,* vol. 14, pp. 1861-1874, 2014.