Environment Protection Engineering

Vol. 50 2024 No. 4

DOI: 10.37190/epe240403

DEMET AYDIN (ORCID: 0000-0002-3491-8392)[1]

A COMPARISON OF THE STATISTICAL DISTRIBUTIONS OF AIR POLLUTION CONCENTRATIONS IN SINOP, TURKEY

The increasing population and industrialization are the reasons for environmental and air pollution around the world. Air pollution is a major threat, especially to human health, both biological and economic. Therefore, determining the properties of air pollutants is very important for researchers and practitioners working in this field. In this study, the statistical distributions of some air pollutants are determined using the Gumbel, Weibull, generalized Pareto, log-normal, gamma, Rayleigh, and inverse Weibull distributions. The data was obtained from stations Boyabat and Merkez stations in Sinop province in 2017. The Kolmogorov–Smirnov test was used to determine the underlying distributions of the air pollution data. Then we use the root mean square error and coefficient of determination criteria to determine which distribution better fits the air pollution data. Finally, numerical results have shown that the generalized Pareto distribution demonstrates the best overall modeling performance, followed by log-normal and inverse Weibull distributions.

1. INTRODUCTION

Air pollution is a problem that threatens human health as well as the ecosystem, accumulating with dense urbanization, inappropriate settlement of cities, the increase in the number of motor vehicles, irregular industrialization, and poor quality fuel use. Mining activities, agriculture, residences, transportation, industrial emissions, and other sources can all be classified as the main sources of air pollution. Much of the air pollution in Turkey comes from the production and use of energy for transport and home heating, among other things [11]. Furthermore, the phenomenon of air pollution frequently occurs during the winter season as a result of inversion conditions that restrict the dispersion of pollutants due to low wind speeds, low mixing heights, and temperature inversions [7]. The scientific community and relevant authorities monitor and analyze their concentrations to determine strategies and solve air pollution problems. The

[1]Department of Statistics, Sinop University, Korucuk Neighborhood, Trafo Street, Sinop 57000, Turkey, email adresses: daydin@sinop.edu.tr

air quality index classification system, widely used all over the world, classifies air quality as good, moderate, bad, or dangerous according to the concentration of pollutants in the air. The methods and criteria used to calculate the index in many countries around the world have been determined according to the air quality standards of the relevant countries. The national air quality index was created as a result of adapting the US-EPA (Environmental Protection Agency) air quality index to our national legislation and limit values. The air quality index is calculated for five main pollutants. These are particulate matter (PM) smaller than about 10 μm (PM10), carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and ozone (O_3).

Mining activities, agriculture, residences, transportation, industrial emissions, and other sources can all be classified as the main sources of air pollution. Much of the air pollution in Turkey comes from the production and use of energy for transport and home heating, among other things [11].

The simplest way to reduce air pollution is to tackle the underlying cause. There are numerous intermediate options to reduce air pollution in the short term. Only a few of the solutions that can be implemented are the transition to electric vehicles, the use of cleaner fuel standards, and the use of renewable energy sources. However, to be effective in any of these measures, governments must acknowledge the effects of air pollution on human health and the economy and take appropriate action [11].

On the other hand, in epidemiological surveillance, the average concentration of air pollutants has been used as an indicator of the degree of atmospheric pollution and its negative impact on human diseases such as chronic bronchitis [13]. For this reason, the statistical distributions of air pollutants have become very important in air pollution studies since it can easily be estimated how many times exceeding air quality standards are exceeded [14]. The distribution often used to model air pollutant concentration is log-normal [8, 10, 14, 15, 18, 19, 21, 22]. However, the level of air pollution varies depending on factors such as the source of pollution and the local meteorology and topography. Therefore, the actual distribution of the atmospheric pollutants does not always correspond to the log-normal one, especially at higher levels of pollution levels [13]. In recent years, various statistical distribution models have been evaluated to meet the objectives of urban air quality management. Therefore, many statistical distributions have been widely used to organize and efficiently describe air pollutant concentrations, including extreme and average concentrations, such as the Weibull distribution [8, 10, 19, 27], the gamma distribution [23], and the log-logistic distribution [6].

Furthermore, in the related literature, there are a considerable number of studies that determine the characteristics of air pollutants that cause air pollution using statistical distributions. Noor et al. [16] used to model PM10 concentration in industrialized areas in Malaysia uLusing the log-normal distribution for 2006 and the gamma distribution for 2007. El-Shanshoury [4] fitted the Frechet distribution known as the inverse Weibull distribution to the model PM10 concentration in Ain Sokhna, Egypt, for 2014. Gavriil et al. [5] conducted the study for two separate years, and the results indicated that Pearson type

VI provided a better fit to PM10. Kan et al. [10] determined the most appropriate distributions for PM10, SO_2, and NO_2 concentrations in Shanghai as log-normal, Pearson V, and extreme values, respectively. Lu [14] found that the theoretical distribution that represents the daily average concentration of PM10 at the Sha-Lu station in Taiwan from 1995 to 1999 is log normal. Lu [15] concluded that the Weibull and log-normal distributions are more suitable for the daily average concentration of the SO_2 distribution at three stations in Taiwan from 1998 to 1999. Leiva et al. [13] showed that the skewed sinh-normal distribution is more suitable for modeling the hourly SO_2 concentration observed at a monitoring station in Santiago in March 2002. Souza et al. [26] claimed that the Rician distribution is more suitable for modeling the average hourly NO concentration from one year to 2015. Gulia et al. [7] indicated that NO_2 concentration is best fitted with log-normal and log-logistic distribution models, respectively, for the winter and summer seasons of 2010 in Delhi. Oguntunde et al. [17] determined that gamma was the best distribution to fit the CO concentration in Lagos state, Nigeria, between 2004 and 2010. Ott et al. [18] demonstrated that the frequency distribution of CO data in US cities can be fitted well with a log-normal distribution. Prieto et al. [20] indicated that the Gumbel distribution is the most appropriate for the annual concentrations of NO_x obtained from 2010 to 2015 in Sao Paulo.

Comparing the modeling performances of the Gumbel, Weibull, log-normal, generalized Pareto, gamma, Rayleigh, and inverse Weibull distributions to determine the characteristics of air pollutants is the aim of this study. These distributions were selected because they are plausible alternatives to the well-known log-normal distribution, which is often used in research on air pollution. For this purpose, hourly data on air pollutants from two monitoring stations from January to March 2017 were analyzed to determine the distributional properties of air pollution. Air pollutants are measured at the following stations: PM10, SO_2, NO, NO_2, NO_x, and CO at the Boyabat station and PM10 and SO_2 at the Merkez station.

The rest of the study is organized as follows. Section 2 of the paper describes the probability density functions that have been used in the specialized literature on modeling air pollution and other natural phenomena. Air pollution data from the study is introduced in Section 3. The results of the statistical analysis are reported in Section 4. Finally, our conclusions obtained from the study are given in Section 5.

2. METHODS

This section of the article briefly describes the probability distributions used in modeling the air pollutants that cause air pollution in this study. Next, the probability density function (pdfs) of all considered distributions for some selected parameter values are displayed in Fig. 1. Almost all of the distributions used in this study are right-skewed, although the Weibull distribution is left-skewed for some shape parameter values.

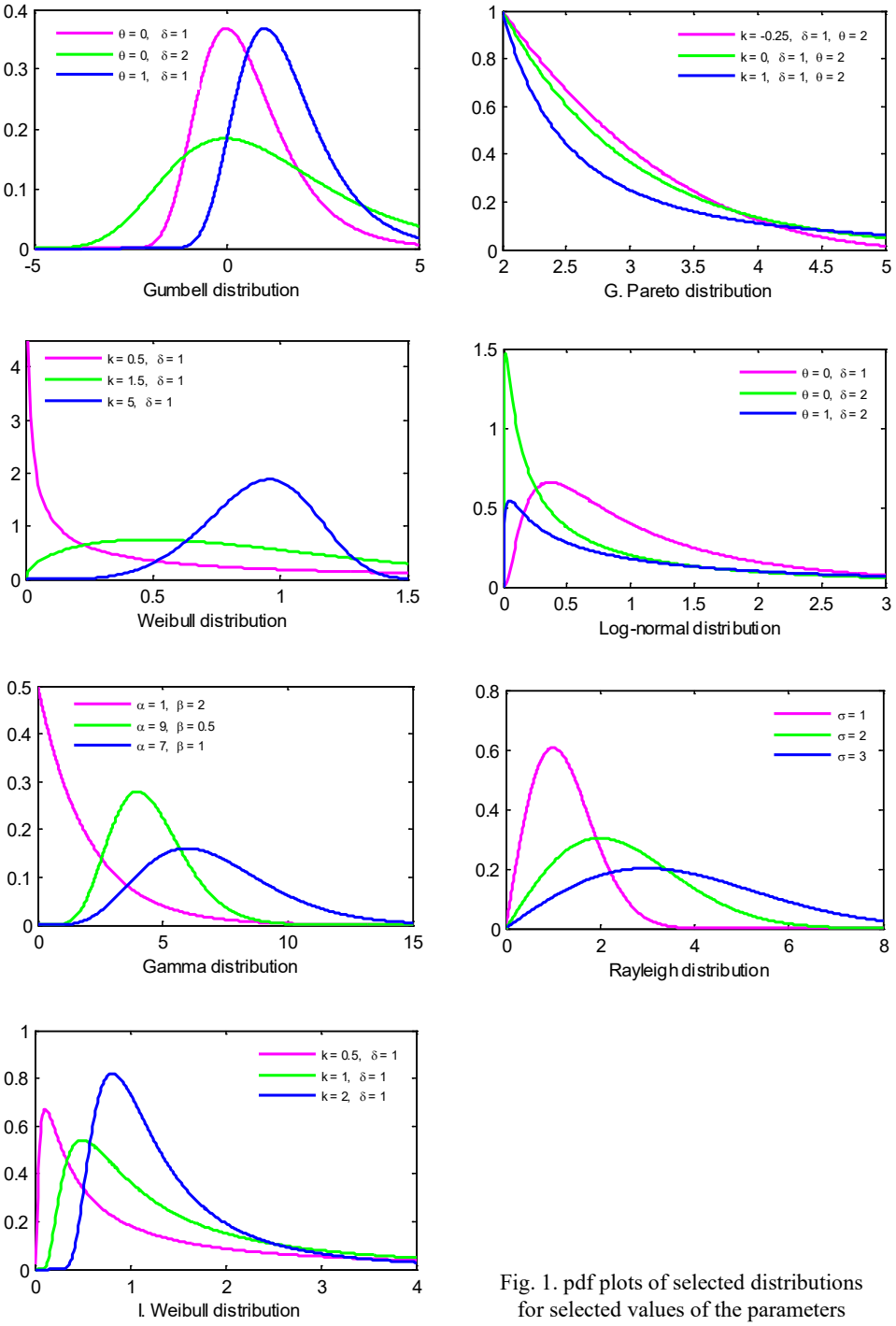

Fig. 1. pdf plots of selected distributions for selected values of the parameters

2.1. GUMBEL DISTRIBUTION

Let X be a random variable; the two-parameter Gumbel distribution probability density function (pdf) and cumulative density function (cdf) are given as

$$f(x)=\frac{1}{\delta}\exp\left(-\frac{x-\theta}{\delta+\exp\left(-\frac{x-\theta}{\delta}\right)}\right),\quad x\in\mathbb{R}$$

and

$$F(x)=\exp\left(-\exp\left(-\frac{x-\theta}{\delta}\right)\right)$$

$\theta\in\mathbb{R}$ is the location parameter, and $\delta>0$ the scale parameter. The expected value $E(X)$, variance $\mathrm{Var}(X)$, skewness β_1, and kurtosis of β_1 X are:

$$E(X)=\theta+\delta\gamma$$

$$\mathrm{Var}(X)=\frac{\pi^2}{6}\delta^2$$

$$\beta_1=1.14,\quad \beta_2=5.40$$

γ is the Euler's constant, which has an approximate value of 0.5772.

2.2. WEIBULL DISTRIBUTION

Let X denote a random variable having a Weibull distribution with the shape parameter k and scale parameter δ. pdf and cdf of X are given by

$$f(x)=\frac{k}{\delta}\left(\frac{x}{\delta}\right)^{k-1}\exp\left(-\frac{k}{\delta}\right)^{k},\quad x>0,\ \delta>0$$

and

$$F(x)=1-\exp\left(-\frac{x}{\delta}\right)^{k}$$

Basic characteristics of the Weibull distribution, such as the mean, variance, skewness and kurtosis are as follows:

$$E(X)=\delta\,\Gamma\left(1+\frac{1}{k}\right)$$

$$\mathrm{Var}(X)=\delta^2\left(\Gamma\left(1+\frac{2}{k}\right)-\Gamma^2\left(1+\frac{1}{k}\right)\right)$$

$$\beta_1=\frac{E(X^3)-3\mu E(X^2)+2\mu^3}{\sigma^3}$$

$$\beta_2=\frac{E(X^4)-4\mu E(X^3)+6\mu^2E(X^2)-3\mu^4}{\sigma^4}-3$$

where

$$\mu=E(X),\quad \sigma=\left(\mathrm{Var}(X)\right)^{1/2}$$

The expected values used in the above-mentioned skewness and kurtosis values can be easily calculated using the following general formula:

$$E(X^r)=\delta^r\Gamma\left(1+\frac{r}{k}\right),\quad r=1,2,\ldots$$

2.3. GENERALIZED PARETO DISTRIBUTION

Let X be a random variable that has a generalized Pareto distribution with location parameter θ, scale parameter δ, and shape parameter k. The pdf and cdf of X are given by

$$f(x)=\begin{cases}\dfrac{1}{\delta}\left(1+k\dfrac{x-\theta}{\delta}\right)^{-1-1/k}, & k\neq 0,\\[2ex] \dfrac{1}{\delta}\exp\left(-\dfrac{x-\theta}{\delta}\right), & k=0,\end{cases}\qquad k\in\mathbb{R},\ \theta\in\mathbb{R},\ \delta>0$$

and

$$F(X)=\begin{cases}1-\left(1+k\dfrac{x-\theta}{\delta}\right)^{-1-1/k}, & k\neq 0\\[2ex] 1-\exp\left(-\dfrac{x-\theta}{\delta}\right), & k=0\end{cases}$$

The domains of X depend on the shape parameter k and are given as

$$\begin{cases} \theta \le x \le \infty, & k \ne 0 \\ \theta \le x \le \theta - \dfrac{\delta}{k}, & k = 0 \end{cases}$$

The mean, variance, skewness and kurtosis of X are:

$$E(X) = \theta + \frac{\delta}{1-k}, \quad k < 1$$

$$\mathrm{Var}(X) = \frac{\delta^2}{(1-2k)(1-k)^2}, \quad k < \frac{1}{2}$$

$$\beta_1 = \frac{2(1+k)(1-2k)^{1/2}}{1-3k}, \quad k < \frac{1}{3}, \qquad \beta_2 = \frac{6\left(1+k-6k^2-2k^3\right)}{(1-3k)(1-4k)}, \quad k < \frac{1}{4}$$

2.4. LOG-NORMAL DISTRIBUTION

Let X denote a random variable; the pdf and cdf of log-normal distribution with location parameter $\theta \in \mathbb{R}$ and scale parameter $\delta > 0$ are given by

$$f(x) = \frac{1}{x(2\pi)^{1/2}\,\delta} \exp\left(-\frac{(\ln x - \theta)^2}{2\delta^2} \right), \quad x > 0,\ \theta \in \mathbb{R},\ \delta > 0$$

and

$$F(x) = \Phi\left(\frac{\ln x - \theta}{\delta} \right)$$

Basic characteristics of the log-normal distribution are:

$$E(X) = \exp\left(\theta + \frac{1}{2}\delta^2 \right)$$

$$\mathrm{Var}(X) = \exp\left(2\left(\theta + \delta^2\right)\right) - \exp\left(2\theta + \delta^2\right)$$

$$\beta_1 = \left(2 + \exp\left(\delta^2\right)\right)\left(\exp\left(\delta^2\right) - 1\right)^{1/2}$$

$$\beta_2 = \exp\left(4\delta^2\right) + 2\exp\left(3\delta^2\right) + 3\exp\left(2\delta^2\right) - 6$$

2.5. GAMMA DISTRIBUTION

Let X be a random variable having gamma distribution and the pdf and cdf of X are

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

and

$$F(x)=\frac{\gamma\left(\alpha,\frac{x}{\beta}\right)}{\Gamma(\alpha)}$$

where α is the shape parameter, β is the scale parameter, and γ is

$$\gamma(a,x)=\int_0^x u^{a-1}\mathrm{e}^{-u}du$$

The following formulas are used to obtain the distributional characteristics of gamma distribution:

$$E(X)=\alpha\beta,\quad \mathrm{Var}(X)=\alpha\beta^2$$

$$\beta_1=\frac{2}{\alpha^{1/2}},\quad \beta_2=3+\frac{6}{\alpha}$$

2.6. RAYLEIGH DISTRIBUTION

Let X be a random variable with the pdf

$$f(x)=\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right),\quad x>0$$

and cdf

$$F(x)=1-\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

where $\sigma>0$ is the scale parameter. The mean, variance, skewness, and kurtosis of X having Rayleigh distribution are given by

$$E(X) = \sigma\left(\frac{\pi}{2}\right)^{1/2}$$

$$\mathrm{Var}(X) = \frac{4-\pi}{2}\sigma^2$$

$$\beta_1 = \frac{2\pi^{1/2}(\pi-3)}{(4-\pi)^2}, \quad \beta_2 = -\frac{6\pi^2 - 24\pi + 16}{(4-\pi)^2}$$

2.7. INVERSE WEIBULL DISTRIBUTION

Let X denote a random variable that has an inverse Weibull distribution with pdf and cdf given by

$$f(x) = \delta k x^{-(k+1)} \exp\left(-\delta x^{-k}\right), \quad x > 0,\ \delta > 0,\ \kappa > 0$$

and

$$F(x) = \exp\left(-\delta x^{-k}\right)$$

where k is the shape parameter and δ the scale parameter. To obtain the basic characteristics of the inverse Weibull distribution, the mean, variance, skewness, and kurtosis of X are:

$$E(X) = \delta^{1/k}\Gamma\left(1-\frac{1}{k}\right)$$

$$\mathrm{Var}(X) = \delta^{2/k}\left(\Gamma\left(1-\frac{2}{k}\right) - \Gamma^2\left(1-\frac{1}{k}\right)\right)$$

$$\beta_1 = \frac{E(X^3) - 3\mu E(X^2) + 2\mu^3}{\sigma^3}, \quad \beta_2 = \frac{E(X^4) - 4\mu E(X^3) + 6\mu^2 E(X^2) - 3\mu^4}{\sigma^4} - 3$$

where $\mu = E(X)$ and $\sigma = (\mathrm{Var}(X))^{1/2}$. The expected values used in the skewness and kurtosis can be easily calculated using the following general equation:

$$E(X^r) = \delta^{r/k}\Gamma\left(1-\frac{r}{k}\right), \quad k > r$$

2.8. LEAST SQUARES ESTIMATION

The least squares (LS) method [25] is widely used in the statistical estimation process due to its easy applicability. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample of size n from a distribution having a pdf $f(x_i|\theta)$ where θ is an unknown parameter. Let $X_{(i)}$, $i = 1, 2, \ldots, n$ be the ith order statistics obtained by arranging. Let $X_{(i)}$, $i = 1, 2, \ldots, n$ in ascending order of magnitude (i.e., $X_{(1)} < X_{(2)} < \ldots X_{(n)}$ and $x_{(i)}$, $i = 1, 2, \ldots, n$ ordered observations corresponding to $X_{(i)}$. The LS estimator of the parameter θ is obtained by minimizing the following function:

$$G(\theta) = \sum_{i=1}^{n} \left(F\left(x_{(i)}\right) - p_i \right)^2$$

where p_i is the estimate of $F\left(x_{(i)}\right)$, which is generally taken to be $p_i = i/(n + 1)$ [9, 12].

2.9. KOLMOGOROV–SMIRNOV TEST

Goodness-of-fit tests are used to assess how well a proposed model fits a particular data set. Furthermore, test statistics are usually based on the deviations between the data and the observed predictions of the model. In this context, many tests, such as the chi-square test and the Kolmogorov–Smirnov test, have been suggested in the literature for goodness of fit. In this study, the Kolmogorov–Smirnov test (K-S) was applied to verify that the data were taken from a particular distribution. Let $F_n(x)$ be the cdf of a random sample $\{X_1, X_2, \ldots, X_n\}$ of size n and $S(x)$ be the empirical cdf of the $\{X_1, X_2, \ldots, X_n\}$. Accordingly, the K-S test statistic is defined as:

$$KS = \max_x \left| S(x) - F_n(x) \right|$$

2.10. MODEL EVALUATING TEST

Many criteria are used to compare the data modeling performance of a proposed distribution or to determine the best fit to the data from among the assumed probability distributions. Among these, the most commonly used criteria are Akaike information, Bayes information, the root mean square error (RMSE), and the coefficient of determination (R^2) [1, 3, 24]. In this study, RMSE and R^2, which are widely preferred, are used to determine the best modeling distribution of the air pollution data among the supposed distributions. For calculations, the following formulas are used:

$$\mathrm{RMSE} = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{F}\left(x_{(i)} - p_i\right)^2 \right)^{1/2}$$

and

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(\hat{F}\left(x_{(i)}\right) - p_i\right)^2}{\sum_{i=1}^{n}\left(\hat{F}\left(x_{(i)}\right) - \tilde{F}(x)_{(i)}\right)^2}$$

where $\hat{F}\left(x_{(i)}\right)$ is the estimate of $F\left(x_{(i)}\right)$, $\tilde{F}\left(x_{(i)}\right)$ is the mean of $\hat{F}\left(x_{(i)}\right)$ and

$$\tilde{F} = \frac{1}{n}\sum_{i=1}^{n}\hat{F}$$

3. AIR POLLUTANT DATA

Sinop is situated at the northern edge of the Turkish side of the Black Sea coast. It is located at 42°01′44″ north latitude and 35°09′19″ east longitude. Its surface area is 5.862 km^2, and its borders are 475 km in total, 300 km of which are land and 175 km are seaside. From the national air quality and monitoring network, we selected two stations at Merkez (city center) and Boyabat, the only stations in the Sinop region for which measurement data are available for 2017. Figure 2 shows where the selected stations are located on a map of Turkey. The monitoring data sets in the two stations were recorded from the official website of the Ministry of Environment and Urbanization of the Republic of Turkey (https://www.csb.gov.tr).

Fig. 2. Geographic location of Merkez and the Boyabat stations in Sinop, Turkey

We have characterized the distributions of two pollutants recorded in Merkez, PM10 and SO_2, and the distributions of six pollutants recorded in Boyabat, PM10, SO_2, NO, NO_2, NO_x, and CO. We provided the study with all hourly concentrations of the pollutants from 1 January 2017 to 31 March 2017. Some descriptive statistics have been calculated to summarize the data and provide preliminary information about the data. The descriptive statistics used in this study are the mean, standard deviation (*SD*), maximum value (Max), range, skewness (β_1), kurtosis (β_2), and the percentage of available data (Data [%]). The calculated values for the data set are given in Table 1 for each month and each station.

Table 1

Some descriptive statistics on air pollutants at Merkez and Boyabat

Station	Month	Pollutant	Mean	*SD*	Max	Range	β_1	β_2	*n*	Data [%]
Merkez	Jan	PM10	53	42.5	472	463	3.75	25.01	712	99
		SO_2	18	21.9	165	165	2.66	12.06	669	93
	Feb	PM10	66	55	588	577	3.79	25.31	487	75
		SO_2	22	24	186	183	2.27	10.20	649	100
	Mar	PM10	52	32.2	259	248	2.33	10.80	716	99
		SO_2	15	14.2	139	135	3.79	24.11	717	99
Boyabat	Jan	PM10	97	66.1	668	660	2.10	12.13	717	99
		SO_2	48	47.7	319	313	2.91	13.00	718	100
		NO	32	44.3	330	330	3.12	15.40	718	100
		NO_2	56	29.3	203	193	0.80	3.74	718	100
		NO_x	88	67.1	410	400	1.89	7.13	718	100
		CO	1325	1111.6	7431	7427	1.71	6.89	715	99
	Feb	PM10	92.72	56.5	297	288	1.00	3.49	648	100
		SO_2	48.64	49.8	384	381	2.45	10.72	648	100
		NO	28.78	44.2	335	335	2.96	14.36	648	100
		NO_2	57	29	139	130	0.59	2.52	648	100
		NO_x	85	65.1	442	432	1.56	5.89	648	100
		CO	1269	1117.6	5625	5618	1.20	3.96	645	99
	Mar	PM10	76.12	48.1	227	219	0.91	3.18	616	85
		SO_2	34.84	50.4	533	532	4.68	35.09	716	99
		NO	13.58	25.3	241	241	4.46	30.70	685	95
		NO_2	43.25	25.4	137	132	0.76	2.90	685	95
		NO_x	56.86	45.2	311	306	1.85	7.96	685	95
		CO	848.70	735.3	4122	4122	1.22	4.54	704	98

For the Merkez station, the average concentrations of both pollutants, PM10 and SO_2, are the highest in February (see Table 2). PM10 in February has the highest range value (which is the difference between the highest and lowest values for a given set of data). The maximum, that is, the highest observation value, is PM10 recorded in February at 588 mg/m^3. Furthermore, the largest standard deviations of PM10 and SO_2 concentrations (55 and 24 mg/m^3, respectively) are observed in February. The smallest

range values for PM10 and SO_2 concentrations (248 and 135 mg/m^3, respectively) in Merkez and almost all pollutant concentrations are obtained in March.

Table 2

Parameter estimates of the statistical distributions of PM10 and SO_2 for the Merkez station

Month	Gumbel	Weibull	Generalized Pareto	Log--normal	Gamma	Rayleigh	Inverse Weibull
PM10							
Jan	$\hat{\theta} = 35.55$ $\hat{\delta} = 18.96$	$\hat{k} = 1.53$ $\hat{\delta} = 56.83$	$\hat{\alpha} = 0.01$ $\hat{\beta} = 29.64$ $\hat{\theta} = 20.43$	$\hat{\theta} = 3.73$ $\hat{\delta} = 0.53$	$\hat{\alpha} = 3.76$ $\hat{\beta} = 12.39$	$\hat{\sigma} = 36.76$	$\hat{k} = 2017.40$ $\hat{\delta} = 2.15$
Feb	$\hat{\theta} = 43.19$ $\hat{\delta} = 23.46$	$\hat{k} = 1.49$ $\hat{\delta} = 69.73$	$\hat{\alpha} = -0.02$ $\hat{\beta} = 38.53$ $\hat{\theta} = 23.72$	$\hat{\theta} = 3.93$ $\hat{\delta} = 0.54$	$\hat{\alpha} = 3.67$ $\hat{\beta} = 15.50$	$\hat{\sigma} = 44.73$	$\hat{k} = 2413.40$ $\hat{\delta} = 2.08$
Mar	$\hat{\theta} = 37.80$ $\hat{\delta} = 18.17$	$\hat{k} = 1.88$ $\hat{\delta} = 56.54$	$\hat{\alpha} = -0.14$ $\hat{\beta} = 33.02$ $\hat{\theta} = 21.80$	$\hat{\theta} = 3.78$ $\hat{\delta} = 0.49$	$\hat{\alpha} = 4.51$ $\hat{\beta} = 10.66$	$\hat{\sigma} = 38.48$	$\hat{k} = 4708.50$ $\hat{\delta} = 2.34$
SO_2							
Jan	$\hat{\theta} = 7.56$ $\hat{\delta} = 8.92$	$\hat{k} = 7.56$ $\hat{\delta} = 8.92$	$\hat{\alpha} = 0.69$ $\hat{\beta} = 8.06$ $\hat{\theta} = 2.09$	$\hat{\theta} = 2.27$ $\hat{\delta} = 1.11$	$\hat{\alpha} = 1.04$ $\hat{\beta} = 14.10$	$\hat{\sigma} = 8.39$	$\hat{k} = 7.17$ $\hat{\delta} = 1.05$
Feb	$\hat{\theta} = 9.99$ $\hat{\delta} = 13.04$	$\hat{k} = 1.00$ $\hat{\delta} = 20.33$	$\hat{\alpha} = 0.71$ $\hat{\beta} = 10.77$ $\hat{\theta} = 2.19$	$\hat{\theta} = 2.52$ $\hat{\delta} = 1.17$	$\hat{\alpha} = 0.97$ $\hat{\beta} = 20.53$	$\hat{\sigma} = 11.52$	$\hat{k} = 7.59$ $\hat{\delta} = 0.98$
Mar	$\hat{\theta} = 9.34$ $\hat{\delta} = 5.25$	$\hat{k} = 1.42$ $\hat{\delta} = 15.78$	$\hat{\alpha} = 0.44$ $\hat{\beta} = 5.97$ $\hat{\theta} = 5.69$	$\hat{\theta} = 2.41$ $\hat{\delta} = 0.57$	$\hat{\alpha} = 3.37$ $\hat{\beta} = 3.70$	$\hat{\sigma} = 9.76$	$\hat{k} = 84.44$ $\hat{\delta} = 2.02$

In the Boyabat station, the highest average concentrations of PM10, NO, NO_x, and CO (97, 32, 88, and 1325 mg/m^3, respectively) were observed in January. The highest average concentrations of SO_2 and NO_2 (48.64 and 57 mg/m^3, respectively) were observed at the same station in February. Boyabat has the smallest *SD* values for PM10, NO, NO_2, NO_x, and CO concentrations (48.1, 25.3, 25.4, 45.2, and 735.3 mg/m^3, respectively) in March. Except for NO_2 in three months and PM10 in March at the station, the skewness values for all air pollutants are greater than one. In terms of kurtosis, all but NO_2 in February and March are larger than three. Therefore, the distribution of air pollutants is leptokurtic with heavy tails, as observations of air pollutants with a kurtosis higher than three have extreme values on their tails. On the contrary, air pollutants with

a kurtosis coefficient lowerthan three, that is, NO_2 in February and March, have a platykurtic distribution with lighter and shorter tails with fewer outliers.

To compare the modeling performances of the distributions, we first obtain the LS estimates of the parameters of interest for the air pollution data. The estimated values of the parameters that are used in subsequent analyzes are presented in Tables 2 and 3.

Table 3

Parameter estimates of statistical distributions
of PM10, SO_2, NO, NO_2, NO_x and CO at the Boyabat station

Month	Gumbel	Weibull	Generalized Pareto	Log--normal	Gamma	Rayleigh	Inverse Weibull
PM10							
Jan	$\hat{\theta} = 66.28$ $\hat{\delta} = 45.52$	$\hat{k} = 1.64$ $\hat{\delta} = 106.09$	$\hat{\alpha} = -0.19$ $\hat{\beta} = 82.79$ $\hat{\theta} = 26.84$	$\hat{\theta} = 4.38$ $\hat{\delta} = 0.66$	$\hat{\alpha} = 2.53$ $\hat{\beta} = 7.271.48$	$\hat{\sigma} = 71.48$	$\hat{k} = 1.73$ $\hat{\delta} = 1294.30$
Feb	$\hat{\theta} = 66.40$ $\hat{\delta} = 44.03$	$\hat{k} = 1.76$ $\hat{\delta} = 101.64$	$\hat{\alpha} = -0.16$ $\hat{\beta} = 77.76$ $\hat{\theta} = 27.23$	$\hat{\theta} = 4.35$ $\hat{\delta} = 0.65$	$\hat{\alpha} = 2.54$ $\hat{\beta} = 36.02$	$\hat{\sigma} = 69.25$	$\hat{k} = 1.78$ $\hat{\delta} = 1530.90$
Mar	$\hat{\theta} = 51.98$ $\hat{\delta} = 38.78$	$\hat{k} = 1.73$ $\hat{\delta} = 82.98$	$\hat{\alpha} = -0.25$ $\hat{\beta} = 74.24$ $\hat{\theta} = 17.44$	$\hat{\theta} = 4.15$ $\hat{\delta} = 0.70$	$\hat{\alpha} = 2.22$ $\hat{\beta} = 34.36$	$\hat{\sigma} = 1.64$	$\hat{k} = 607.55$ $\hat{\delta} = 1.64$
SO_2							
Jan	$\hat{\theta} = 26.96$ $\hat{\delta} = 18.06$	$\hat{k} = 1.25$ $\hat{\delta} = 47.89$	$\hat{\alpha} = 0.17$ $\hat{\beta} = 26.53$ $\hat{\theta} = 12.52$	$\hat{\theta} = 3.49$ $\hat{\delta} = 0.68$	$\hat{\alpha} = 2.53$ $\hat{\beta} = 15.15$	$\hat{\sigma} = 28.70$	$\hat{k} = 43.34$ $\hat{\delta} = 1.20$
Feb	$\hat{\theta} = 25.82$ $\hat{\delta} = 25.16$	$\hat{k} = 1.16$ $\hat{\delta} = 47.84$	$\hat{\alpha} = 0.02$ $\hat{\beta} = 40.45$ $\hat{\theta} = 4.76$	$\hat{\theta} = 3.47$ $\hat{\delta} = 0.94$	$\hat{\alpha} = 1.39$ $\hat{\beta} = 31.28$	$\hat{\sigma} = 29.65$	$\hat{k} = 43.34$ $\hat{\delta} = 1.20$
Mar	$\hat{\theta} = 15.47$ $\hat{\delta} = 15.73$	$\hat{k} = 0.99$ $\hat{\delta} = 30.89$	$\hat{\alpha} = 0.26$ $\hat{\beta} = 21.21$ $\hat{\theta} = 3.33$	$\hat{\theta} = 2.97$ $\hat{\delta} = 1.00$	$\hat{\alpha} = 1.30$ $\hat{\beta} = 20.71$	$\hat{\sigma} = 17.97$	$\hat{k} = 18.41$ $\hat{\delta} = 1.12$
NO							
Jan	$\hat{\theta} = 12.06$ $\hat{\delta} = 18.42$	$\hat{k} = 12.00$ $\hat{\delta} = 18.42$	$\hat{\alpha} = 0.63$ $\hat{\beta} = 16.90$ $\hat{\theta} = 0.47$	$\hat{\theta} = 2.70$ $\hat{\delta} = 1.41$	$\hat{\alpha} = 0.74$ $\hat{\beta} = 37.18$	$\hat{\sigma} = 14.51$	$\hat{k} = 5.68$ $\hat{\delta} = 0.80$
Feb	$\hat{\theta} = 8.32$ $\hat{\delta} = 14.93$	$\hat{k} = 0.77$ $\hat{\delta} = 4.42$	$\hat{\alpha} = 1.06$ $\hat{\beta} = 9.73$ $\hat{\theta} = 0.30$	$\hat{\theta} = 0.77$ $\hat{\delta} = 4.42$	$\hat{\alpha} = 1.06$ $\hat{\beta} = 9.73$	$\hat{\sigma} = 8.83$	$\hat{k} = 3.48$ $\hat{\delta} = 0.71$

Table 3

Parameter estimates of statistical distributions
of PM10, SO_2, NO, NO_2, NO_x and CO at the Boyabat station

Mar	$\hat{\theta}=3.34$ $\hat{\delta}=6.44$	$\hat{k}=0.76$ $\hat{\delta}=2.31$	$\hat{\alpha}=-26.86$ $\hat{\beta}=25.19$ $\hat{\theta}=3.06$	$\hat{\theta}=0.76$ $\hat{\delta}=2.31$	$\hat{\alpha}=0.46$ $\hat{\beta}=24.03$	$\hat{\sigma}=4.05$	$\hat{k}=1.66$ $\hat{\delta}=0.65$
				NO_2			
Jan	$\hat{\theta}=66.28$ $\hat{\delta}=45.52$	$\hat{k}=1.64$ $\hat{\delta}=106.09$	$\hat{\alpha}=-0.19$ $\hat{\beta}=82.79$ $\hat{\theta}=26.84$	$\hat{\theta}=4.38$ $\hat{\delta}=0.66$	$\hat{\alpha}=2.53$ $\hat{\beta}=37.21$	$\hat{\sigma}=71.48$	$\hat{k}=12.9430$ $\hat{\delta}=1.73$
Feb	$\hat{\theta}=42.35$ $\hat{\delta}=26.15$	$\hat{k}=2.14$ $\hat{\delta}=63.00$	$\hat{\alpha}=-0.50$ $\hat{\beta}=60.91$ $\hat{\theta}=16.17$	$\hat{\theta}=3.92$ $\hat{\delta}=0.59$	$\hat{\alpha}=3.05$ $\hat{\beta}=18.90$	$\hat{\sigma}=44.81$	$\hat{k}=1130.00$ $\hat{\delta}=1.90$
Mar	$\hat{\theta}=30.43$ $\hat{\delta}=22.18$	$\hat{k}=1.81$ $\hat{\delta}=47.79$	$\hat{\alpha}=-0.37$ $\hat{\beta}=46.62$ $\hat{\theta}=9.54$	$\hat{\theta}=3.60$ $\hat{\delta}=0.69$	$\hat{\alpha}=2.32$ $\hat{\beta}=18.97$	$\hat{\sigma}=33.37$	$\hat{k}=222.69$ $\hat{\delta}=1.62$
				NO_x			
Jan	$\hat{\theta}=55.65$ $\hat{\delta}=45.53$	$\hat{k}=1.48$ $\hat{\delta}=93.25$	$\hat{\alpha}=-0.15$ $\hat{\beta}=80.93$ $\hat{\theta}=16.31$	$\hat{\theta}=4.22$ $\hat{\delta}=0.78$	$\hat{\alpha}=1.91$ $\hat{\beta}=44.43$	$\hat{\sigma}=62.29$	$\hat{k}=295.43$ $\hat{\delta}=1.45$
Feb	$\hat{\theta}=52.54$ $\hat{\delta}=43.77$	$\hat{k}=1.45$ $\hat{\delta}=90.22$	$\hat{\alpha}=0.08$ $\hat{\beta}=64.58$ $\hat{\theta}=18.41$	$\hat{\theta}=4.17$ $\hat{\delta}=0.79$	$\hat{\alpha}=1.83$ $\hat{\beta}=44.64$	$\hat{\sigma}=57.75$	$\hat{k}=281.38$ $\hat{\delta}=1.45$
Mar	$\hat{\theta}=34.72$ $\hat{\delta}=30.12$	$\hat{k}=1.40$ $\hat{\delta}=60.02$	$\hat{\alpha}=-0.01$ $\hat{\beta}=47.50$ $\hat{\theta}=10.33$	$\hat{\theta}=3.75$ $\hat{\delta}=0.82$	$\hat{\alpha}=1.73$ $\hat{\beta}=31.63$	$\hat{\sigma}=38.75$	$\hat{k}=119.36$ $\hat{\delta}=1.38$
				CO			
Jan	$\hat{\theta}=785.82$ $\hat{\delta}=734.42$	$\hat{k}=1.26$ $\hat{\delta}=1398.30$	$\hat{\alpha}=-0.10$ $\hat{\beta}=1273.70$ $\hat{\theta}=158.87$	$\hat{\theta}=6.89$ $\hat{\delta}=0.88$	$\hat{\alpha}=1.53$ $\hat{\beta}=841.40$	$\hat{\sigma}=898.39$	$\hat{k}=5710.30$ $\hat{\delta}=1.31$
Feb	$\hat{\theta}=694.39$ $\hat{\delta}=829.75$	$\hat{k}=1.12$ $\hat{\delta}=1300.70$	$\hat{\alpha}=0.00$ $\hat{\beta}=1288.90$ $\hat{\theta}=27.07$	$\hat{\theta}=6.76$ $\hat{\delta}=1.11$	$\hat{\alpha}=1.05$ $\hat{\beta}=1243.40$	$\hat{\sigma}=847.61$	$\hat{k}=677.66$ $\hat{\delta}=1.02$
Mar	$\hat{\theta}=483.93$ $\hat{\delta}=569.04$	$\hat{k}=483.93$ $\hat{\delta}=569.04$	$\hat{\alpha}=-0.25$ $\hat{\beta}=1100.30$ $\hat{\theta}=-33.81$	$\hat{\theta}=6.41$ $\hat{\delta}=1.08$	$\hat{\alpha}=1.06$ $\hat{\beta}=854.05$	$\hat{\sigma}=608.82$	$\hat{k}=692.91$ $\hat{\delta}=1.08$

The Kolmogorov–Smirnov test is used to determine whether the data are suitable for the assumed distributions and RMSE and R^2 criteria to determine which distribution fits the data better. The results are presented in Tables 4 and 5. In the evaluation of the results of the K-S test, hypothesis H_0 is established as follows: H_0: The air pollutants are the assumed distribution and cannot be rejected if the calculated K-S value is less than the table K-S value (K-S_t). In other words, the distribution of air pollution data will be the assumed distribution in hypothesis H_0. It should be noted that the K-S values are compared with K-$S_t = 1.36/(n)^{1/2}$ values. Here, both the smallest RMSE and the highest R^2 in Tables 4 and 5 are in bold to show which distribution provides a better fit to the data.

Table 4

K-S, RMSE, and R^2 for PM10 and SO_2 at the Merkez station

Distribution	January			February			March		
	K-S	RMSE	R^2	K-S	RMSE	R^2	K-S	RMSE	R^2
PM10									
Gumbel	0.0560	0.0318	0.9882	0.0565	0.0302	0.9894	**0.0505**	0.0244	0.9931
Weibull	0.1181	0.0318	0.9202	0.1202	0.0691	0.9176	0.0890	0.0536	0.9569
Generalized Pareto	0.0702	0.0183	0.9961	0.0607	0.0175	0.9964	0.0627	0.0203	0.9952
Log-normal	**0.0447**	0.0210	0.9948	**0.0432**	0.0190	0.9957	**0.0322**	0.0163	0.9969
Gamma	0.0555	0.0316	0.9882	0.0581	0.0296	0.9897	0.0540	0.0265	0.9917
Rayleigh	0.0747	0.0430	0.9764	0.0715	0.0405	0.9795	0.0840	0.0461	0.9702
Inverse Weibull	**0.0323**	**0.0110**	**0.9986**	**0.0249**	**0.0113**	**0.9985**	**0.0387**	**0.0137**	**0.9978**
SO_2									
Gumbel	0.1037	0.0660	0.9513	0.1808	0.0686	0.9448	0.0973	0.0577	0.9622
Weibull	0.4108	0.2326	0.7467	0.1357	0.0481	0.9720	0.1480	0.0833	0.8845
Generalized Pareto	0.0628	0.0207	0.9949	0.0701	0.0300	0.9888	0.0722	0.0259	0.9919
Log-normal	**0.0522**	**0.0261**	**0.9917**	0.1111	0.0348	0.9847	0.0895	0.0472	0.9733
Gamma	0.0700	0.0438	0.9771	0.1446	0.0472	0.9724	0.0911	0.0568	0.9623
Rayleigh	0.2142	0.1155	0.9059	0.2085	0.1351	0.8808	0.0952	0.0649	0.9502
Inverse Weibull	0.0585	0.0216	0.9943	0.0754	0.0348	0.9845	0.0879	0.0317	0.9878

Table 4 indicates that the distribution of PM10 in January is log-normal and inverse Weibull according to the test rule. However, the calculated values of the RMSE and R^2 criteria in Table 4 are examined to determine which distribution provides a better fit to the data than the others, that is, log-normal or inverse Weibull. The RMSE and R^2 are the most useful fit comparison measures. Low RMSE values and high R^2 values indicate a better fit of the assumed distribution. Table 4 presents the results of the goodness of fit test conducted on the pollution data obtained at the Merkez station. The inverse Weibull distribution is more suitable than the others for PM10 concentration data in January–March, and the distribution log-normal for SO_2 data in January. However, since hypothesis H_0 is rejected, the distribution of SO_2 data in February and March could not be modeled using one of the distributions discussed in this study.

Table 5

K-S, RMSE, and R^2 for pollutants PM10, SO_2, NO, NO_2, NO_x, and CO at the Boyabat station

Distribution	Jan			Feb			Mar		
	K-S	RMSE	R^2	K-S	RMSE	R^2	K-S	RMSE	R^2
PM10									
Gumbel	**0.0501**	0.0257	0.9920	**0.0426**	0.0256	0.9922	**0.0447**	0.0220	0.9942
Weibull	0.0596	0.0285	0.9894	**0.0523**	0.0286	0.9902	**0.0528**	0.0251	0.9930
Generalized Pareto	**0.0484**	0.0125	0.9981	0.0648	0.0154	0.9972	0.0584	0.0150	0.9973
Log-normal	**0.0278**	**0.0117**	**0.9983**	**0.0319**	**0.0113**	**0.9985**	**0.0380**	**0.0130**	**0.9979**
Gamma	**0.0371**	0.0188	0.9957	**0.0344**	0.0191	0.9956	**0.0322**	0.0131	0.9979
Rayleigh	0.0579	0.0380	0.9849	0.0780	0.0388	0.9843	0.0822	0.0443	0.9809
Inverse Weibull	0.0600	0.0262	0.9916	0.0713	0.0254	0.9922	0.0784	0.0326	0.9873
SO_2									
Gumbel	0.0905	0.0391	0.9829	0.0840	0.0357	0.9853	0.0865	0.0442	0.9778
Weibull	0.1048	0.0649	0.9331	**0.0387**	0.0230	0.9933	0.0693	0.0404	0.9770
Generalized Pareto	0.0585	0.0199	0.9954	**0.0344**	0.0116	0.9984	**0.0286**	0.0123	0.9982
Log-normal	0.0601	0.0213	0.9947	**0.0269**	**0.0095**	**0.9989**	**0.0253**	**0.0121**	**0.9982**
Gamma	0.0862	0.0344	0.9864	**0.0369**	0.0168	0.9967	0.0603	0.0264	0.9918
Rayleigh	0.1204	0.0482	0.9763	0.1493	0.0820	0.9470	0.1463	0.0919	0.9357
Inv. Weibull	**0.0309**	**0.0128**	**0.9980**	0.0666	0.0317	0.9879	**0.0490**	0.0241	0.9929
NO									
Gumbel	0.1459	0.0639	0.9525	0.1745	0.0863	0.9181	0.1862	0.0809	0.9288
Weibull	0.4657	0.2700	0.6913	0.4759	0.2694	0.6773	0.3824	0.2206	0.7603
Generalized Pareto	**0.0374**	0.0171	0.9964	**0.0464**	0.0205	0.9949	0.4569	0.2676	0.6977
Log-normal	**0.0368**	**0.0165**	**0.9967**	**0.0384**	**0.0195**	**0.9954**	0.1591	0.0420	0.9811
Gamma	**0.0498**	0.0266	0.9913	0.0660	0.0410	0.9797	0.1591	0.0455	0.9783
Rayleigh	0.2393	0.1506	0.8637	0.2966	0.1593	0.8516	0.2754	0.1711	0.8371
Inv. Weibull	0.0714	0.0331	0.9864	0.0821	0.0274	0.9909	0.1591	0.0490	0.9734
NO_2									
Gumbel	**0.0354**	0.0144	0.9974	**0.0364**	0.0163	0.9967	**0.0503**	0.0218	0.9941
Weibull	**0.0325**	0.0135	0.9978	**0.0461**	0.0264	0.9924	**0.0502**	0.0259	0.9926
Generalized Pareto	**0.0348**	**0.0110**	**0.9986**	**0.0313**	**0.0089**	**0.9991**	**0.0223**	**0.0073**	**0.9994**
Log-normal	0.0523	0.0220	0.9941	0.0562	0.0189	0.9956	0.0552	0.0197	0.9952
Gamma	**0.0288**	0.0123	0.9981	**0.0332**	0.0126	0.9980	0.0285	0.0146	0.9974
Rayleigh	**0.0315**	0.0130	0.9980	**0.0341**	0.0184	0.9960	0.0697	0.0402	0.9838
Inv. Weibull	0.0953	0.0436	0.9768	0.0993	0.0400	0.9804	0.0993	0.0400	0.9801
NO_x									
Gumbel	0.0754	0.0299	0.9892	0.0823	0.0417	0.9796	0.0780	0.0372	0.9836
Weibull	**0.0507**	0.0237	0.9931	0.0604	0.0364	0.9840	**0.0517**	0.0289	0.9898
Generalized Pareto	**0.0251**	**0.0102**	**0.9987**	**0.0401**	**0.0116**	**0.9984**	**0.0277**	**0.0090**	**0.9990**
Log-normal	**0.0295**	0.0144	0.9975	**0.0268**	0.0141	0.9976	**0.0286**	0.0117	0.9983
Gamma	**0.0414**	0.0181	0.9960	**0.0514**	0.0287	0.9899	**0.0449**	0.0225	0.9938
Rayleigh	0.0890	0.0597	0.9678	0.1382	0.0693	0.9579	0.1292	0.0715	0.9564
Inv. Weibull	0.0673	0.0315	0.9877	0.0636	0.0212	0.9945	0.0693	0.0272	0.9909

Table 5

K-S, RMSE, and R^2 for pollutants PM10, SO_2, NO, NO_2, NO_x, and CO at the Boyabat station

CO									
Gumbel	0.0550	0.0284	0.9905	0.1013	0.0388	0.9820	0.0963	0.0247	1.0000
Weibull	**0.0324**	0.0163	0.9967	**0.0354**	0.0187	0.9961	0.5469	0.2923	0.7000
Generalized Pareto	**0.0462**	0.0124	0.9982	**0.0245**	**0.0081**	**0.9992**	**0.0304**	**0.0080**	**1.0000**
Log-normal	**0.0342**	0.0124	0.9982	0.0621	0.0230	0.9936	0.0877	0.0401	0.9815
Gamma	**0.0255**	**0.0110**	**0.9986**	**0.0225**	0.0095	0.9989	**0.0417**	0.0181	1.0000
Rayleigh	0.1332	0.0743	0.9550	0.1607	0.1124	0.9125	0.1550	0.1028	0.9000
Inverse Weibull	0.0699	0.0344	0.9859	0.1046	0.0452	0.9753	0.1318	0.0603	1.0000

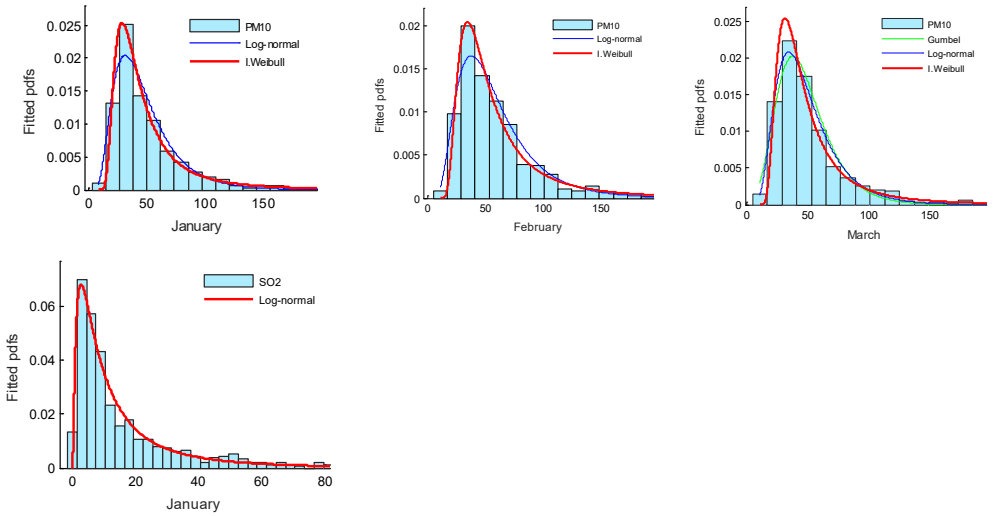

Fig. 3. Fitting specific distributions to data on air pollutant concentrations at the Merkez station for each month

According to Table 5, the distribution of PM10 is Gumbel, generalized Pareto, log-normal, and gamma in January–March, respectively. Furthermore, both model evaluation tests show that the log-normal distribution is well matched with the measured PM10 data at the Boyabat station for all months, and the log-normal distribution is more suitable to represent the SO_2 distribution in February and March. Otherwise, the inverse Weibull distribution is appropriate for representing the SO_2 concentration in January. The distribution of NO data is also log-normal and generalized Pareto in January and February, respectively. However, the NO data in January does not match any of the distributions discussed in this study. The generalized Pareto distribution is seen to be more suitable to represent both NO_2 and NO_x concentrations for all months. For CO concentration, the gamma distribution provides a better fit with measured data in January, and the generalized Pareto distribution is more suitable in February and March.

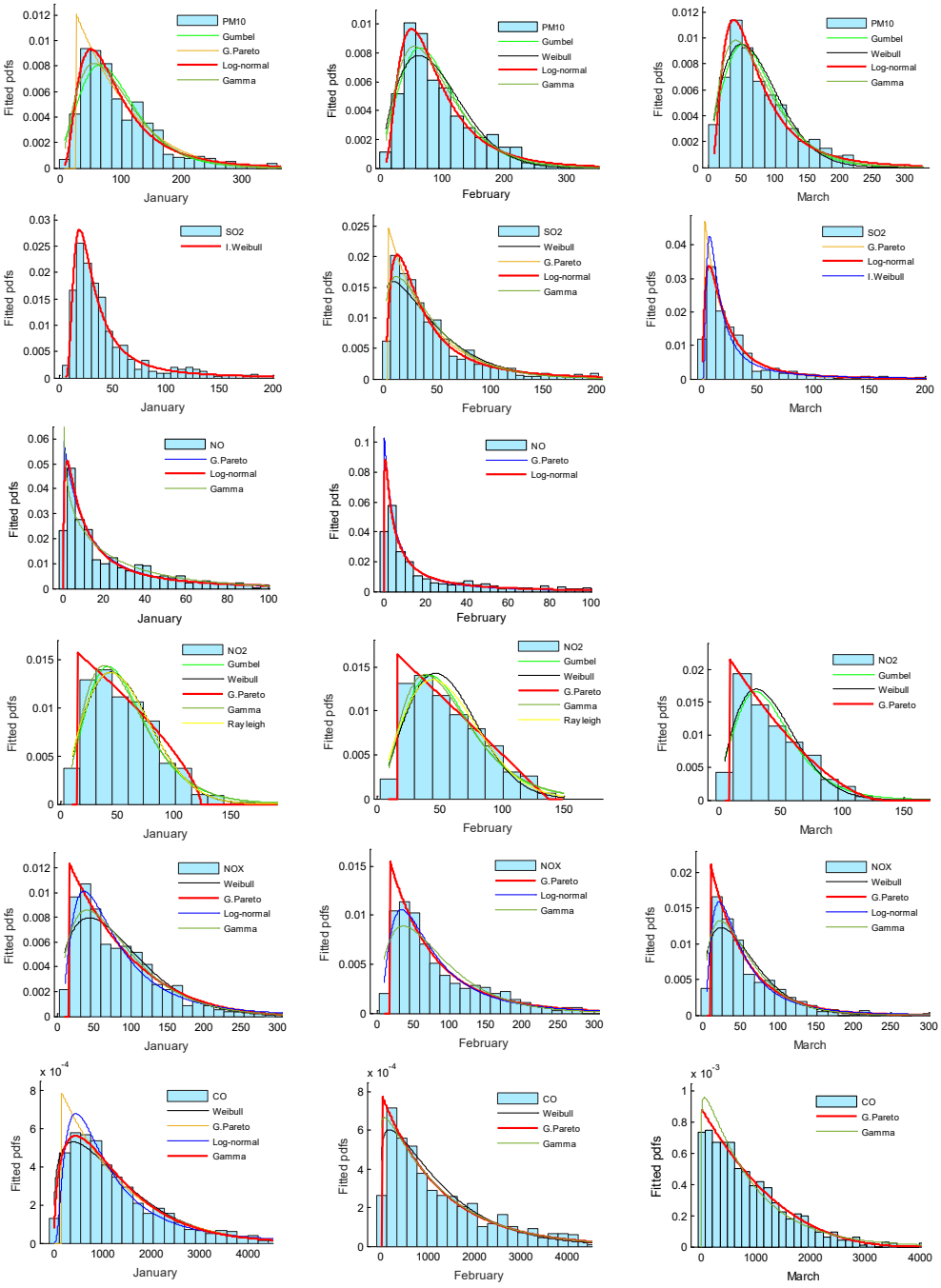

Fig. 4. Fitting specific distributions to data on air pollutant concentrations at the Boyabat station for each month

Furthermore, to illustrate how the pdfs match the data, the histograms and fitted pdf plots for the air pollution data at each station have been shown in Figs. 3 and 4. These results agree with the findings given in Tables 4 and 5. It is also clear from this figures that the inverse Weibull and log-normal distributions show the best performance for modeling the peak of PM10 in the January–March and SO_2 in January for the Merkez station, among others, respectively. However, all distributions are insufficient to model SO_2 data sets collected in February and March. As far as the Boyabat station is concerned, the log-normal distribution for PM10 and the generalized Pareto distribution for NO_2 and NO_x demonstrate better fitting performance than the others for each month.

On the other hand, the log-normal and generalized Pareto distributions perform better than other distributions in air pollutant observations, especially for February and March. However, all considered distributions remain insufficient to model NO data sets recorded in February and March.

4. SUMMARY AND CONCLUSIONS

The study aims to statistically model the concentrations of air pollutants harming both the environment and human health. It attempts to determine which air pollution distribution models are better suited for air pollution in January, February, and March. Two stations in Sinop province, Merkez and Boyabat, provided the air pollution data used in the investigation. Seven statistical distributions were used, namely Gumbel, Weibull, generalized Pareto, log-normal, gamma, Rayleigh, and inverse Weibull.

It can be seen from the statistical analysis that although the Weibull distribution is frequently used to model air pollution data, other distributions that are more suited for modeling the data include the generalized Pareto, log-normal, and inverse Weibull distributions. Several findings on the statistical distributions employed in the modeling of air pollutants differ from those reported in the literature. For example, the distribution of PM10 at the Merkez station follows the inverse Weibull distribution in January, February, and March. However, while the SO_2 distribution is log-normal in January, the distribution of the data for the other months could not be determined among the assumed distributions. For this reason, curve-fitting plots could not be drawn on the histogram of the data.

In the Boyabat station, the distribution of PM10 is log-normal for each of the three months. The SO_2 distribution is the inverse Weibull in January and the log-normal in other months. Although the distribution of NO items is determined to be log-normal in January, the distribution of the data in February and March could not be determined. The distributions of NO_2 and NO_x are generalized Pareto for each of the three months. The distribution of CO is determined as gamma in January and generalized Pareto in February and March. Additionally, the fit graphs are generally consistent with these results (see Figs. 3 and 4). It should be clear that some of the compatibility graphs examined, for example, some of the histograms of the graphs of SO_2 in Fig. 3 and NO in

Fig. 4, do not seem to depict the data adequately. Because curve fittings on the histograms are drawn using only assumed distributions, graphs that are more suitable for other distributions that are not addressed in this study may be obtained, which can be considered another study subject.

Finally, the findings of this study demonstrated that different times of the year may have different distributions of certain contaminants. Statistical analyses also show that the commonly used log-normal distribution is often inadequate to model air pollution data. It can be observed that the generalized Pareto distribution demonstrates the best overall modeling performance, followed by log-normal and inverse Weibull distributions. The inverse Weibull distribution also provides flexibility for modeling environmental data sets. Furthermore, suitable theoretical distributions allow for effective prediction of days in the upcoming year that surpass city air quality regulations, provided the most relevant distribution that statistically models the pollutant is identified.

ACKNOWLEDGEMENTS

This work was supported by the Sinop University Scientific Research Project No. FEF-1901-16-11.

REFERENCES

[1] AKAIKE H., *Information theory and an extension of the maximum likelihood principle*, [In:] B.N. Petrov, F. Caski (Eds.), Proc. Second International Symposium on Information Theory, Akademiai Kiado, Budapest 1973, 267–281.

[2] AKGÜL F.G., ŞENOĞLU B., ARSLAN T., *An alternative distribution to Weibull for modeling the wind speed data. Inverse Weibull distribution*, En. Conv. Manage., 2016, 114, 234–240. DOI: 10.1016/j.enconman.2016.02.026.

[3] AKGÜL F.G., ŞENOĞLU B., *Comparison of wind speed distributions. A case study for Aegean coast of Turkey*, En. Sourc., A. Rec., Util. Environ. Eff., 2023, 45 (1), 2453–2470. DOI: 10.1080/15567036.2019.1663309.

[4] EL-SHANSHOURY G.I., *Fitting the probability distribution functions to model particulate matter concentrations*, Arab J. Nucl. Sci. Appl., 2017, 50, 108–122.

[5] GAVRIIL I., GRIVAS G., KASSOMENOS P., CHALOULAKOU A., SPYRELLIS N., *An application of theoretical probability distributions, to the study of PM10 and PM2.5 time series in Athens, Greece*, Glob. NEST J., 2006, 8 (3), 241–251.

[6] GOKHALE S., KHARE M., *A review of deterministic, stochastic and hybrid vehicular exhaust emission models*, Int. J. Trans. Manage., 2004, 2 (2), 59–74. DOI: 10.1016/j.ijtm.2004.09.001.

[7] GULIA S., NAGENDRA S.S., KHARE M., *Extreme events of reactive ambient air pollutants and their distribution pattern at urban hotspots*, Aerosol Air Qual. Res., 2017, 17 (2), 394–405. DOI: 10.4209/aaqr.2016.06.0273.

[8] HADLEY A., TOUMI R., *Assessing changes to the probability distribution of sulphur dioxide in the UK using lognormal model*, Atmos. Environ., 2002, 37 (24), 455–467. DOI: 10.1016/S1352-2310 (02)01003-8.

[9] HERD G.R., *Estimation of reliability from incomplete data*, Proc. 6th National Symposium on Reliability and Quality Control, IEEE, New York 1960.

[10] KAN H.D., CHEN B.H., *Statistical distributions of ambient air pollutants in Shanghai, China*, Biomed. Environ. Sci., 2004, 17 (3), 366–372.

[11] KARA N.O., *Air pollution in Istanbul*, Netherlands Enterprise Agency, The Hague 2018.
[12] JOHNSON L.G., *The statistical treatment of fatigue experiments*, Elsevier, New York 1964.
[13] LEIVA V., VILCA F., BALAKRISHNAN N., SANHUEZA A., *A skewed sinh-normal distribution and its properties and application to air pollution*, Commun. Stat.-Theor. M., 2010, 39 (3), 426–443. DOI: 10.1080/03610920903140171.
[14] LU H.C., *The statistical characters of PM10 concentration in Taiwan area*, Atmos. Environ., 2002, 36 (3), 491–502. DOI: 10.1016/S1352-2310 (01)00245-X.
[15] LU H.C., *Comparisons of statistical characteristic of air pollutants in Taiwan by frequency distribution*, Air, Waste Manage. Assoc., 2003, 53 (5), 608–616. DOI: 10.1080/10473289.2003.10466194.
[16] NOOR N.M., TAN C.Y., ABDULLAH M.M.A.B., RAMLI N.A., YAHAYA A.S., *Modelling of PM10 concentration in industrialized area in Malaysia. A case study in Nilai*, [In:] International Conference on Environment and Industrial Innovation IPCBEE, 2011, 12, 48–58.
[17] OGUNTUNDE P.E., ODETUNMIBI O.A., ADEJUMO A.O., *A study of probability models in monitoring environmental pollution in Nigeria*, J. Prob. Stat., 2014, 1–6. DOI: 10.1155/2014/864965.
[18] OTT W.R., MAGE D.T., RANDERCKER V.W., *Testing the validity of the lognormal probability model: computer analysis of carbon monoxide data from US cities*, EPA 600/4-79-040, US Environmental Protection Agency, Research Triangle Park, NC, 1979.
[19] PAPANASTASIOU D.K., MELAS D., *Application of PM10's statistical distribution to air quality management. A case study in Central Greece*, Water Air Soil Poll., 2010, 207, 115–122. DOI: 10.1007/s11270-009-0123-8.
[20] PRIETO W., CREMASCO M., *Application of probability density functions in modelling annual data of atmospheric NO_x temporal concentration*, Chem. Eng. Trans., 2017, 57, 487–492. DOI: 10.3303/CET1757082.
[21] RUMBERG B., ALLDREDGE R., CLAIBORN C., *Statistical distributions of particulate matter and the error associated with sampling frequency*, Atmos. Environ., 2001, 35, 2907–2920. DOI: 10.1016/S1352-2310 (00)00554-9.
[22] SIMPSON R.W., DALY N.J., JAKEMAN A.J., *The prediction of maximum air pollution concentrations for TSP and CO Larsen's model and the ATDL model*, Atmos. Environ., 1983, 17, 2497–2503. DOI: 10.1016/0004-6981 (83)90075-6.
[23] SINGH P., *Simultaneous confidence intervals for the successive ratios of scale parameters*, J. Stat. Plan. Inf., 2004, 36 (3), 1007–1019. DOI: 10.1016/j.jspi.2004.08.006.
[24] SCHWARZ G., *Estimating the dimension of a model*, Ann. Stat., 1978, 6 (2), 461–464.
[25] SWAIN J., VENKATRAMAN S., WILSON J., *Least-squares estimation of distribution function in Johnson's translation system*, J. Stat. Comp. Simul., 1988, 29, 271–297. DOI: 10.1080/00949658808811068.
[26] SOUZA A., OLAOFE Z., KODICHERLA S.P.K., IKEFUTI P., NOBREGA L., SABBAH I., *Probability distributions assessment for modeling gas concentration in Campo Grande, MS, Brazil*, Eur. Chem. Bull., 2018, 6 (12), 569–578. DOI: 10.17628/ecb.2017.6.569–578.
[27] WANG X., MAUZERALL D.L., *Characterizing distributions of surface ozone and its impact on grain production in China, Japan and South Korea: 1990 and 2020*, Atmos. Environ., 2004, 38 (74), 4383–4402. DOI: 10.1016/j.atmosenv.2004.03.067.