

## Article

# Modelling NO<sub>2</sub> Emissions at Eskom's Coal-Fired Power Station: Application of Statistical Distributions at Arnot

Mpendulo Wiseman Mamba <sup>1,\*</sup>  and Delson Chikobvu <sup>2</sup> <sup>1</sup> Department of Mathematical and Physical Sciences, Central University of Technology, Bloemfontein 9301, Free State, South Africa<sup>2</sup> Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein 9301, Free State, South Africa; chikobvu@ufs.ac.za

\* Correspondence: mmamba@cut.ac.za

## Abstract

The combustion of coal comes with a heavy price of pollutant emissions. To assist in the planning and management of these emissions and to protect human health, the current study uses the relatively heavy-tailed distributions, namely, the Weibull, Lognormal and Pareto distributions to analyse and characterise the distribution of NO<sub>2</sub> emission (in tons) from Arnot, a coal-fired power station of South Africa's power utility, Eskom. Quantile–quantile (QQ) plots and their corresponding derivative plots for the three distributions are used to characterise the statistical distribution of NO<sub>2</sub> emissions. The strength and advantage of using derivative plots of the three distributions, in particular, for characterising NO<sub>2</sub> emissions from a coal-fuelled power station, is that they are able to better capture and explain the behaviour of the data across different components of this data. Although this method possesses flexible ways of characterisation of data, it is not commonly applied to emissions data, especially NO<sub>2</sub> emissions from a coal-fuelled power station belonging to Eskom, such as Arnot. The choice of the distributions of this study is motivated by their ability to cater to varied tails relative to the exponential distribution. Thus, the tail heaviness ranks of the distributions from lighter to heavier tail, that is, Weibull, Lognormal and Pareto, are taken into consideration in order to arrive at the best-fitting distribution(s). The Weibull distribution with a lighter tail than the Exponential distribution gave the best-fitting distribution over the Lognormal and Pareto distributions for the main body of the data. The Pareto distribution, however, captures the extreme emission tail behaviour much better than the other two distributions. The Kolmogorov–Smirnov and Vasicek–Song (VS) goodness of fit statistics were used to further assess the appropriateness of the fitted distributions. The selection of the Weibull distribution implies that milder high values and less frequent very high NO<sub>2</sub> emission data are expected, showing the weakness of such criteria when extremes are present. For authorities to plan and draw policies for the reduction and management of emissions, these findings may be of interest to them and can assist in better understanding their behaviour and the planning to reduce the impact on humans and the environment. This may also assist practitioners in air quality modelling before other, more sophisticated methods can be explored.



Academic Editors: Valerio Paolini and Francesco Petracchini

Received: 29 December 2025

Revised: 10 February 2026

Accepted: 11 February 2026

Published: 17 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)[Attribution \(CC BY\) license](#).

**Keywords:** heavy-tail; Pareto; light-tail; Weibull; nitrogen dioxide (NO<sub>2</sub>); derivative plot; quantile–quantile (QQ) plot; Arnot power station; Eskom; coal-fired power plant

## 1. Introduction

Since coal is in abundance in South Africa (SA), it is not surprising that this fossil fuel is the main source of electricity generation in the country. Coal is responsible for generating 93% of the SA's electricity [1]. The energy crisis in a number of developing countries persists, and is characterised by the unmet electricity demand due to poor infrastructure, among other reasons [2,3]. Consequently, it is expected that more effort be made to meet this demand to ensure economic growth and development of a country such as SA [3]. However, this cheap method of power generation comes at a price of emissions produced during the combustion of coal. This places coal-fired power stations among the major sources of emission pollutants in the atmosphere [1]. Air pollution, in general, is one of the main role players responsible for diseases and early deaths in the world and continues to cause harm to human health [4,5].

Modelling emissions is one useful instrument for providing early emission-related information behaviour [6]. For authorities to plan and draw policies for the reduction and management of emissions, emissions quantification and prediction can assist in better understanding their behaviour and reduce the impact on humans and the environment. It is important to understand the characteristics and statistical distributions of pollutant emissions for a better characterisation and representation of the data. Probability distributions offer such capabilities to model area-specific air pollutant concentrations. This presents support for the development of methods to combat and control this environmental challenge [7].

Air pollutant concentrations can be considered as statistical random variables and can thus be well described and characterised by means of probability distributions [8]. Although the type of probability distribution used is a special case in every area [9] and is dependent on local characteristics such as proximity to emission source, topology and climate [7], the choice of a proper statistical distribution function to represent the data is very important [8]. It is common to have a right-skewed distribution of air pollutant concentration, such that we have more frequent low concentrations compared to higher (extreme) concentrations that are less frequent or rare in nature [7].

### 1.1. Statement of the Problem

Coal-fired power stations in South Africa are among the main sources of emissions and pose a big threat to human health and the environment. To understand, explain and predict emissions behaviour, probability distributions are a good option. The aim of the current study is to find a model for NO<sub>2</sub> emission, given in tons, from Eskom's Arnot coal-fired power station. This will be done by fitting and comparing three probability distribution functions, namely, the Weibull, Lognormal and Pareto. This will involve analysing the quantile–quantile (QQ) plots and their associated derivative plots for these distributions. The derivative plot will allow the analysis to be done piece wise from the largest values to the lowest values. The derivative plot uses the very highest values and uses the whole data set upon arriving at the lowest values.

### 1.2. Justification of the Study

The three parent distributions provide the benefit of modelling data that may be heavier or lighter-tailed than the Exponential distribution. Although some distributions are common in pollutant modelling, the use of derivative plots is not as common, especially for modelling emission data from a coal-fired power station such as Arnot. Modelling by means of graphical display of the quantile–quantile (QQ) plots and their corresponding derivative plots offers the benefit of exploring different components of the data and finding an appropriate distribution for each component, since different components of a dataset,

including emissions data, may not always have the same distribution throughout the length of the data [10,11]. For example, the central observations of the data may have a different statistical distribution from the tail or the extreme observations [12,13]. The aim is to also capture this tail behaviour accurately, not just the main body.

The Exponential distribution is central to tail modelling, mostly due to its memoryless property, see reference [10] for details. Statistical distributions can be classified into two general categories according to their asymptotic tail behaviour, namely, the sub-exponential and super-exponential (or hyper-exponential) classes. The sub-exponential class has tails that tend to zero less rapidly than the tail of the Exponential distribution. This means that this class has distributions that are heavier-tailed than the Exponential tail. Conversely, the super-exponential class has tails that tend to zero more rapidly than those of the Exponential distribution, meaning that they are lighter-tailed than those of the Exponential distribution [10,14]. The selected distributions cater for all types of tail shapes relative to the Exponential distribution, for example, the Weibull with shape parameter  $\tau > 1$  is lighter tailed than the Exponential distribution, the Weibull with  $\tau = 1$  is the Exponential, the Weibull with  $\tau < 1$  is heavier tailed than the Exponential distribution but lighter than the Lognormal distribution, and the Pareto. The Pareto distribution has the heaviest tail of the three distributions. The appropriateness of a heavy-tailed distribution, compared to a lighter-tailed distribution, implies that more frequent, larger-sized NO<sub>2</sub> emissions are predicted. Thus, fitting a lighter tailed distribution to a heavy tailed NO<sub>2</sub> emissions data may lead to underestimation of risk and this may result in serious or severe implications on human lives and the environment, especially in the upper tail [14]. Consideration of the tail heaviness ranks assists in properly selecting an appropriate distribution to represent the whole length of the data.

### 1.3. Objectives of the Study

To analyse and compare the distribution fit of the Weibull, Lognormal and Pareto distributions, QQ plots and derivative plots of each of the three relatively heavy-tailed distributions are represented graphically. The distributions are listed according to their ranks of tail heaviness in order to determine the best-fitting distribution for modelling NO<sub>2</sub> emission data (in tons) from Eskom's Arnot coal-fired power station.

### 1.4. Contribution of the Study

Characterising the data by considering both the central and upper tail behaviour of the data is very important in the emissions behaviour. This will assist in determining the best models to further explain and predict patterns in the NO<sub>2</sub> emissions data.

For authorities to plan and draw policies for the reduction and management of emissions, statistical modelling of emissions will help to better understand their future behaviour and therefore reduce the impact on humans and the environment. Thus, the modelling technique employed in this study for NO<sub>2</sub> emissions data from Arnot power station will assist in explaining and understanding such behaviours in the data.

## 2. Literature Review

The current section considers some studies where several probability distributions were used to model pollutant concentration data, including NO<sub>2</sub>, from different sources and areas.

One air pollutant that has received attention is particulate matter (PM<sub>10</sub> and PM<sub>2.5</sub>). In one of earlier studies, Karaca et al. [15] used and compared 11 frequency distribution functions, namely, Beta, Erlang, Exponential, Gamma, Weibull, Inverse Gaussian, Lognormal, Pearson V, Pearson VI and Log-logistic, to model 86 daily samples of PM<sub>10</sub>

(particulate matter of size 10 micrometres or less) and PM<sub>2.5</sub> (particulate matter of size 2.5 micrometres or less) in Istanbul Municipality, Turkey, from July 2002 to July 2003. The Kolmogorov–Smirnov and Chi-squared tests were used to assess the goodness of fit of the distribution models. The fitting for PM<sub>10</sub> and PM<sub>2.5</sub> proved the Log-Logistic model as the best-fitting distribution.

Lu [9] compared three distributions, namely the Lognormal, Weibull and type V Pearson distribution in the modelling of PM<sub>10</sub> from three monitoring stations (Hsin-Chu, Sha-Lu and Gian-Jin) in Taiwan over a period of five years. The method of moments and the method of least squares were used for parameter estimation. The Lognormal distribution gave the best fit for the PM<sub>10</sub> data. The method of least squares gave accurate results compared to the method of moments in parameter estimation. The probabilities of exceeding the air quality limits were successfully determined for the three monitoring stations.

For the PM<sub>10</sub> from the Belgrade, Serbia urban area over the period 2003–2013, Perišić et al. [12] fitted the Pearson V, the Lognormal and the Weibull methods for the general distribution, and the two-parameter Exponential and Gumbel probability distributions for extreme values in the determination of a distribution of best fit and estimation of exceedances. To evaluate the goodness of fit, the Kolmogorov–Smirnov and Anderson–Darling tests were used. The Pearson V and Lognormal distributions captured the concentrations observed and the number of exceedances, while at the extremes, the two-parameter Exponential and Gumbel distributions gave satisfactory results for the prediction of critical pollutant loads and for the estimation of required reduction from identified strong sources of emission.

Tharu et al. [16] compared the modelling performance of the Weibull, Gamma and Lognormal distributions in modelling daily average PM<sub>2.5</sub> and total suspended particles (TSP) concentrations in the urban areas of Kathmandu, Nepal, over the period March–October 2024. The Kolmogorov–Smirnov test was used to test how well the distributions fit the data. The Weibull and Lognormal distributions outperformed the Gamma distribution in modelling PM<sub>2.5</sub> and TSP. Another study analysed the statistical distribution of atmospheric PM<sub>10</sub> and PM<sub>2.5</sub> concentrations and found the Gamma and Lognormal distributions as the best-fitting distributions, respectively. The study was conducted in Faial Island in the Azores archipelago during 2024 [17]. Choopradit et al. [18] demonstrated that the inverse Gaussian and Pearson type V distributions outperformed the gamma, Lognormal, Log-Logistic and Weibull in representing daily average PM<sub>2.5</sub> concentrations at quality monitoring locations in Bangkok, Thailand.

Other air pollutants that have been considered in the past include carbon monoxide (CO), NO<sub>2</sub>, ozone (O<sub>3</sub>) and sulphur dioxide (SO<sub>2</sub>), among others. In their study in Lagos, Nigeria, to model CO concentrations, Oguntunde et al. [19] applied three theoretical distributions, namely the Weibull, Gamma and the Lognormal. The Gamma distribution gave the best fit for the data based on the Anderson–Darling and Kolmogorov–Smirnov tests. The characteristics of the pollutant were determined, and the probability of exceeding the set limits was also predicted.

Xue Jiang et al. [8] applied several probability distributions, namely, Lognormal, Gamma, Inverse Gaussian, Log-logistic, Pearson V, Beta, Pearson VI, Weibull, Extreme value distributions, to model the three-year (from 2006 to 2008) daily average concentration data of SO<sub>2</sub>, NO<sub>2</sub> and PM<sub>10</sub> in Xi'an, China. The Kolmogorov–Smirnov statistic, Anderson–Darling statistic, Chi-squared goodness of fit tests and the ML parameter estimation method were used in this study. SO<sub>2</sub>, NO<sub>2</sub> and PM<sub>10</sub> were most represented by the Pearson VI, Extreme Value and Log-Logistic distribution, respectively.

Aydin [20] investigated the statistical characteristics of six-hourly air pollutants data (SO<sub>2</sub>, PM<sub>10</sub>, CO, NO<sub>x</sub>, NO and NO<sub>2</sub>) from two monitoring stations in Sinop, Turkey, col-

lected over the period January–March 2017. They fitted and compared seven distributions, namely, the Gumbel, Weibull, generalised Pareto, Lognormal, Gamma, Rayleigh, and the inverse Weibull distributions to determine the best performing for the pollutants. The Kolmogorov–Smirnov goodness of fit, root mean square error and coefficient of determination criteria were used to determine the best-fitting distributions. The generalised Pareto distribution gave the overall best-fitting distribution, followed by the Lognormal, then the inverse Weibull.

In another study, six probability distribution functions, namely Normal, Gamma, Pearson V, Weibull, Gumbel and Lognormal, were used to characterise the variability of PM<sub>10</sub> and O<sub>3</sub> concentrations by considering the daily maximum moving average concentrations from the Metropolitan Area of São Paulo in Brazil. Goodness of fit metrics used include the Kolmogorov–Sminirnov, Anderson–Darling, and root mean square error. The Lognormal and Gamma distributions gave a good fit for PM<sub>10</sub> and O<sub>3</sub>, respectively [7].

Although statistical modelling using probability distributions on the NO<sub>2</sub> pollutant has received some attention in the past, it is underrepresented in the South African literature. This is especially true for NO<sub>2</sub> emissions from SA's coal-fired power stations. The current study aims to compare the three distributions (Weibull, Lognormal and Pareto distributions) according to their tail heaviness ranks in order to arrive at the best fitting for the NO<sub>2</sub> emissions data. This will be done by employing the QQ and derivative plots of the distributions of NO<sub>2</sub> emissions, given in tons, from one of Eskom's coal-fired power stations, Arnot. Characterisation of data using the derivative plots of the Weibull, Lognormal and Pareto distributions is not common in emission modelling, more so in modelling NO<sub>2</sub> emissions from a power station owned and operated by Eskom's such as Arnot. This is despite their flexibility of being able to capture a dataset's behaviour across all of its components, and can thus indicate if more than one distribution for a single dataset would be appropriate for the components or not. This is particularly important since Osatohanmwun et al. [21] indicated that, if it is possible to partition a dataset into multiple components, the use of parametric distributions, such as those used in this paper, may be of benefit and offer robust and flexible modelling techniques. Common modelling techniques, such as modelling using composite distributions, have fitted a different distribution to each component of a single dataset. For example, Deng et al. [22] fitted a composite model with a Weibull distribution for the bulk of the insurance data and a Pareto distribution for the tail. Akhundjanova [13] used the Lognormal and Pareto distributions for the bulk and tail of the size of national CO<sub>2</sub> emissions data, respectively. Cooray [23], similarly, applied the Lognormal distribution for the bulk and the Pareto distribution for the tail of the insurance data. However, Nadarajah [24] showed that the Burr distribution was better fitted compared to the Pareto for the tail of the loss data, while the Lognormal distribution gave the best fit for the bulk. These studies ([13,22–24]) all employed composite distributions of pairs of parametric distributions. Compared to the assumption of a Pareto type for extreme values and a light-tailed distribution in the main body, Albrecher et al. [25] demonstrated that a dataset can have a distribution that is heavier-tailed, i.e., the Pareto, in the bulk, while the upper tail has a lighter-tailed distribution, such as the Weibull. They used a Weibull-tempered Pareto distribution. The appropriateness of these fitted distributions to multiple components of a single dataset, including emissions data, indicate that indeed a random variable can be partitioned into more than one component and splicing distributions do give a good modelling framework. Thus, incorporating derivative plots can serve as an alternative.

Derivative plots can determine if the full body or a component of the data is likely to be best fitted by a lighter or heavier-tailed distribution than the one being investigated. Thus, as an additional benefit, they can be very useful as graphical diagnosis tools for quantifying

and explaining the body and tail behaviour of the data before more sophisticated modelling techniques like composite and mixture models can be fitted [10,11,25].

### 3. Methodology [10]

The current section firstly presents the extreme value index estimators for determining the tail heaviness of the NO<sub>2</sub> emissions data. The probability distribution functions, QQ plots and corresponding derivative plots of the three distributions, namely, the Weibull, Lognormal and Pareto distributions, ordered in increasing tail heaviness, used in the current study are then presented.

#### 3.1. Extreme Value Index ( $\gamma$ ) Estimates

In this study, the extreme value index (EVI) estimate  $\gamma$  is used to identify the shape of the upper tail [11]. A negative value indicates a lighter-tailed than the extremal Exponential distribution, a positive value indicates a heavier-tailed distribution than the extremal Exponential distribution, while a value of zero indicates that the upper tail is Exponentially distributed. This will assist in identifying potential candidate distribution(s) for the upper tail of NO<sub>2</sub> emissions from Arnot power station.

The Hill and EPD estimators are limited to estimating the extreme value index, EVI ( $\gamma$ ), for cases with  $\gamma > 0$  only. When  $\gamma \leq 0$  is the case, other methods of estimation may be considered. As a result, for this paper, the generalised Hill, moment estimators and the generalised Pareto distribution estimators will be used to estimate the EVI [10].

Table 1 presents the EVI and quantile estimators of the distributions used in this study:

**Table 1.** Estimators of the extreme value index (EVI),  $\gamma$ .

| Method                          | Equation                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|---------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Moment [26]                     | $\hat{\gamma}_{k,n}^M = H_{k,n} + 1 - 0.5 \left( \frac{H_{k,n}^2}{H_{k,n}^{(2)}} \right)^{-1},$ <p>where <math>H_{k,n}^{(2)} = \frac{1}{k} \sum_{j=1}^k (\log X_{n-j+1,n} - \log X_{n-k,n})^2</math>, <math>H_{k,n}^2 = (H_{k,n})^2 = \left( \frac{1}{k} \sum_{j=1}^k \log X_{n-j+1,n} - \log X_{n-k,n} \right)^2</math>, log is the natural logarithm and <math>X_{n-k,n}</math> is the threshold.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| Generalised Hill                | $\hat{\gamma}_{k,n}^{GH} = \frac{1}{k} \sum_{j=1}^k \log UH_{j,n} - \log UH_{k+1,n} = H_{k+1,n} + \frac{1}{k} \sum_{j=1}^k \log H_{j,n} - \log H_{k+1,n},$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| Generalised QQ plot             | $\left( \log \frac{n+1}{k+1}, \log X_{n-k,n} H_{k,n} \right), k = 1, \dots, n-1,$ <p>If the plot is increasing, then <math>\gamma &gt; 0</math>.<br/>         If the plot is decreasing, then <math>\gamma &lt; 0</math>.<br/>         If the plot is horizontal, then <math>\gamma = 0</math>.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| Generalised Pareto Distribution | <p>Suppose <math>X_1, X_2, X_3, \dots, X_n</math> is the random sample from a GPD with distribution function:</p> $F(X) = \begin{cases} 1 - \left( 1 + \gamma \left( \frac{x-\mu}{\sigma} \right) \right)^{\frac{-1}{\gamma}}, & \text{if } \gamma \neq 0 \\ 1 - \exp \left[ - \left( \frac{x-\mu}{\sigma} \right) \right], & \text{if } \gamma = 0 \end{cases}.$ <p><math>\hat{\gamma}</math> is produced by maximising the following log-likelihood function of <math>F(X)</math>:</p> $l(\gamma, \sigma   x_i - u) = \begin{cases} -n \log(\sigma) - \frac{1+\gamma}{\gamma} \sum_{i=1}^n \log \left( 1 + \gamma \left( \frac{x_i - \mu}{\sigma} \right) \right), & \text{if } \gamma \neq 0 \\ -n \log(\sigma) - \frac{1}{\sigma} \sum_{i=1}^n (x_i - \mu), & \text{if } \gamma = 0, \end{cases}$ <p>where <math>\left( 1 + \gamma \left( \frac{x_i - \mu}{\sigma} \right) \right) &gt; 0</math> for <math>i = 1, \dots, n</math>, <math>\mu</math> is the chosen threshold, and <math>\sigma</math> is the scale parameter estimator.</p> <p>Taking the partial derivatives with respect to the parameters <math>\gamma</math>, and equating them to 0, the <math>\hat{\gamma}</math> estimate is obtained [27].</p> |

### 3.2. Weibull Distribution

The Weibull distribution function is

$$F(x) = 1 - \exp(-\lambda x^\tau), x > 0, \quad (1)$$

where  $\lambda$  and  $\tau$  are scale and shape parameters, respectively.  $F(x)$  is sub-Exponential for  $0 < \tau < 1$ . The Weibull distribution is a first Box–Cox transformation of the exponential distribution. That is, if  $\tau = 1$ , the Weibull distribution is the Exponential distribution, and if  $\tau > 1$ , the Weibull distribution is lighter-tailed than the Exponential distribution. Conversely, if  $\tau < 1$ , the Weibull distribution is heavier-tailed than the Exponential distribution.

The Weibull distribution has a QQ plot that is given by

$$\left( \log \left[ -\log \left( 1 - \frac{i}{n+1} \right) \right], \log X_{i,n} \right), i = 1, \dots, n. \quad (2)$$

and the derivative plot is given by

$$\left( \log x_{n-k,n}, \frac{H_{k,n}}{W_{k,n}} \right) \text{ or } \left( k, \frac{H_{k,n}}{W_{k,n}} \right) \quad (3)$$

with  $W_{k,n} = \frac{1}{k} \sum_{j=1}^k \log \log \frac{n+1}{j} - \log \log \frac{n+1}{k+1}$ .

### 3.3. Lognormal Distribution

Transforming the data and fitting the normal distribution to the data yields the log-normal distribution. This distribution is heavier-tailed than any Weibull and has two parameters, given by

$$F(x) = 1 - \frac{1}{\sigma \sqrt{2\pi}} \int_x^\infty \exp \left\{ -\frac{(\log u - \mu)^2}{2\sigma^2} \right\} \frac{du}{u} = \Phi \left( \frac{\log x - \mu}{\sigma} \right), \mu \in \mathbb{R}, \sigma > 0. \quad (4)$$

The Lognormal distribution has a tail that is heavier than that of the Weibull and the Exponential distribution, and is given by

$$\bar{F}(x) \sim \frac{\sigma}{\log x \sqrt{2\pi}} \exp \left\{ -\frac{(\log x - \mu)^2}{2\sigma^2} \right\}, x \rightarrow \infty \quad (5)$$

The QQ plot is given as

$$\left( \Phi^{-1} \left( \frac{i}{n+1} \right), \log X_{i,n} \right), i = 1, \dots, n, \quad (6)$$

where

$\Phi^{-1}$  is the standard normal quantile function. The derivative plot is given by

$$\left( \log x_{n-k,n}, \frac{H_{k,n}}{N_{k,n}} \right) \text{ or } \left( k, \frac{H_{k,n}}{N_{k,n}} \right), \quad (7)$$

with  $N_{k,n} = \frac{n+1}{k+1} \varphi \left( \Phi^{-1} \left( 1 - \frac{k+1}{n+1} \right) \right) - \Phi^{-1} \left( 1 - \frac{k+1}{n+1} \right)$ ,  
since

$$\frac{1}{k} \sum_{j=1}^k \Phi^{-1} \left( 1 - \frac{j}{n+1} \right) - \Phi^{-1} \left( 1 - \frac{k+1}{n+1} \right) = \int_0^1 \Phi^{-1} \left( 1 - u \frac{k+1}{n+1} \right) du - \Phi^{-1} \left( 1 - \frac{k+1}{n+1} \right) = N_{k,n} \quad (8)$$

where  $\varphi$  denotes the standard normal density.

### 3.4. Pareto Distribution

The Pareto distribution is the prime example of a heavy-tailed distribution. The strict Pareto is a sub-exponential for a  $\forall \alpha$  and is given by

$$F(x) = 1 - \left(\frac{x}{\beta}\right)^{-\alpha}, \alpha > 0, 0 < x_0 < x, \quad (9)$$

where  $\beta$  and  $\alpha$  are the scale and shape parameters. Let  $\log X$  be exponentially distributed with  $\lambda = \alpha$  when  $X$  is strict Pareto ( $\alpha$ )-distributed, then the Pareto QQ plot is given as

$$\left(-\log\left(1 - \frac{i}{n+1}\right), \log X_{i,n}\right), i = 1, \dots, n, \quad (10)$$

with a derivative plot

$$(\log x_{n-k,n}, H_{k,n}) \text{ or } (k, H_{k,n}), \quad (11)$$

where  $H_{k,n} = \frac{1}{k} \sum_{j=1}^k \log X_{n-j+1} - \log X_{n-k,n}$  is the estimator of  $\gamma = 1/\alpha$  [28]. A linear QQ plot and horizontal derivative plot (at the level)  $1/\alpha$  indicates that the data follows a Pareto distribution.

### 3.5. Assessing the Goodness of Fit, and Parameter Estimation

Two goodness-of-fit tests, namely, the Kolmogorov–Smirnov test and the Vasicek–Song (VS), are used to assess if a distribution fits the data or not. The drawback with these measures is that they concentrate on evaluating the average fit, unlike the derivative plot, which does a good job of assessing the fit of the extreme emissions. The Akaike Information Criteria (AIC) and Schwarz’s Bayesian Information Criterion (BIC) [29], with formulae, are given as

$$\text{AIC} = -2l(\hat{\theta}; y) + 2p, \quad (12)$$

and

$$\text{BIC} = -2l(\hat{\theta}; y) + 2p \log(n), \quad (13)$$

will be used to select the best-fitting distribution from the Exponential, Weibull, Lognormal and Pareto distributions.

Parameters are estimated by employing the maximum likelihood (ML) method for the three distributions in this study. The method has the advantage of producing a minimum variance estimate of the parameter [9]. Let  $f(x_j|\theta)$  denote a known probability distribution function (i.e., the Weibull, Lognormal and Pareto distributions) with a pair of parameters given as  $\theta = \theta_1, \theta_2$ . If  $L(\theta; x)$  is the likelihood function of the joint density function of the  $n = 108$  independent and identically distributed  $\text{NO}_2$  emissions observations from Arnot power station,  $x_1, x_2, x_3, \dots, x_{108}$ , given as

$$L(\theta_1, \theta_2; x) = \prod_{j=1}^n f(x_j|\theta_1, \theta_2), \quad (14)$$

then the log-likelihood function  $l(\theta_1, \theta_2)$  is given by the equation

$$l(\theta) = \log[L(\theta_1, \theta_2; x)] = \log \left[ \prod_{j=1}^n f(x_j|\theta_1, \theta_2) \right] = \sum_{j=1}^n \log f(x_j|\theta_1, \theta_2), \quad (15)$$

The optimal values of  $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$  are then obtained by maximising  $l(\theta)$  in Equation (15). This is done by solving for  $\theta_1$  and  $\theta_2$  in the equations  $\frac{\partial l(\theta)}{\partial \theta_1} = \frac{\partial l(\theta_1, \theta_2)}{\partial \theta_1} = 0$  and  $\frac{\partial l(\theta)}{\partial \theta_2} = \frac{\partial l(\theta_1, \theta_2)}{\partial \theta_2} = 0$ , respectively.

## 4. Results

The following sections provide the results and the discussion.

### 4.1. Description of the Data

The data used in this paper were obtained from Eskom's Arnot power station located in Middleburg, in the Mpumalanga province in South Africa. It is actual monthly NO<sub>2</sub> emissions data, given in tons, collected from April 2005 to March 2014 (108 months). This data is just one of the emissions datasets collected from the Arnot power station and is selected because NO<sub>2</sub> emissions are one of the statistically understudied dangerous pollutants from coal-fired power stations in South Africa. The monthly NO<sub>2</sub> emissions data used in this study are presented in Table A1 of Appendix A.

The analysis was performed using the R statistical programming packages ReIns and fitdistrplus.

### 4.2. Data Exploration

Table 2 shows the descriptive statistics, as well as stationarity and normality tests for the NO<sub>2</sub> emissions data from Arnot power station.

**Table 2.** Descriptive statistics, as well as stationarity and normality tests for the NO<sub>2</sub> emissions, in tons, from the Arnot power station.

|                                                           |                        |
|-----------------------------------------------------------|------------------------|
| <b>N</b>                                                  | 108                    |
| <b>Minimum</b>                                            | 2309.00                |
| <b>Mean</b>                                               | 3963.02                |
| <b>Median</b>                                             | 4014.5                 |
| <b>Standard Deviation</b>                                 | 574.44                 |
| <b>Maximum</b>                                            | 5063.00                |
| <b>Kurtosis</b>                                           | −0.23                  |
| <b>Skewness</b>                                           | −0.33                  |
| <b>Phillips–Perron Unit Root Test</b>                     | <0.01                  |
| <b>Augmented Dickey–Fuller (ADF) test <i>p</i>-value</b>  | 0.03772                |
| <b>Kolmogorov–Smirnov (KS) test for normality of data</b> | $<2.2 \times 10^{-16}$ |
| <b>Anderson–Darling (AD) test for normality of data</b>   | $5.556 \times 10^{-6}$ |

The coefficient of skewness is less than zero for the NO<sub>2</sub> emission data. This indicates that the data is spread more to the left of their means. Kurtosis is a measure of whether the data is heavier or lighter compared to the normal distribution. Since the value of the kurtosis is less than 0, the data has an upper tail that is thinner and shorter than the normal distribution. The skewness may somewhat seem to be closer to 0, suggesting some Normality in the distribution. However, this observation can be dismissed since the kurtosis is less than 0, and the KS and AD tests have *p*-values smaller than 0.05.

In Figure 1, the histogram of the NO<sub>2</sub> emissions data is displayed. One can see that the data is not exactly symmetric. Additionally, the QQ plot shows that the upper-tail data is lighter-tailed than the normal distribution.

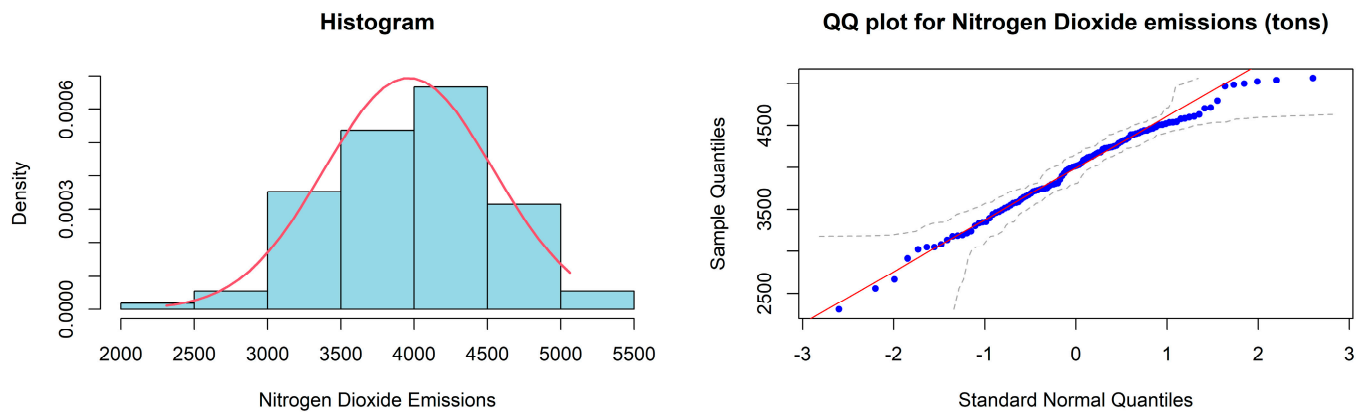


**Figure 1.** Histogram and QQ plot for NO<sub>2</sub> emissions (tons) from Arnot power station. The red line in both plots indicates the fitted normal line (theoretical). The blue dotted line in the QQ plot indicates the data points while the dotted grey line indicates the 95% confidence interval.

Since the Augmented Dickey–Fuller and Phillips–Perron stationarity tests in Table 2 have  $p$ -values that are less than 5%, the NO<sub>2</sub> emissions from Arnot power station are stationary.

#### 4.3. Shape of the Tail

To determine the extreme value index,  $\gamma$ , the generalised Hill (genHill) and moment estimators are used. These estimators are selected since they allow for  $\gamma \leq 0$ . Figure 2 presents the EVI estimates  $\hat{\gamma}$  using the generalised Hill and moment estimators, together with the PoT-based estimates for comparison purposes. Also included in Figure 2 is the generalised QQ plot. These plots will assist in describing and determining the shape of the tail.

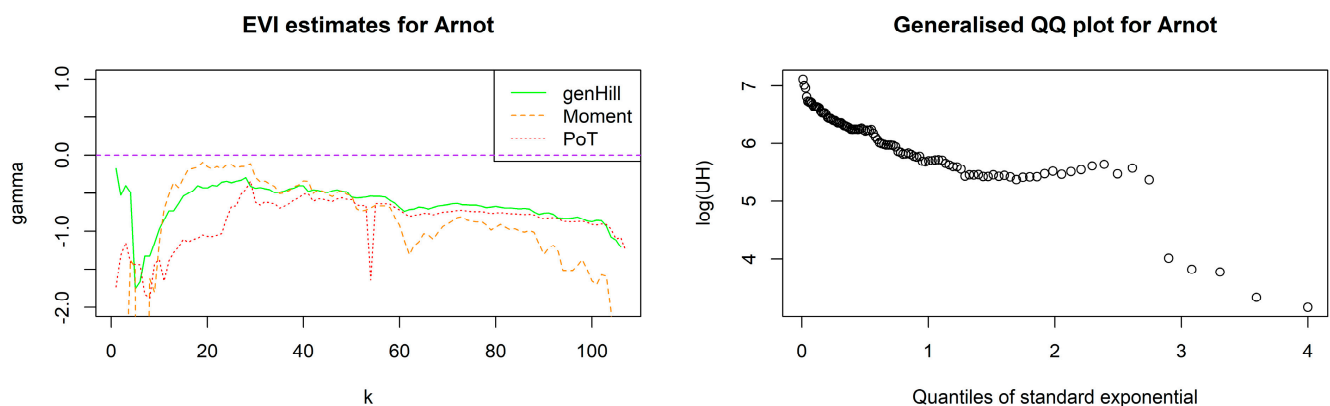


**Figure 2.** Estimates of the EVI (first) and generalised QQ plot (second). The purple dotted line indicates a value of the EVI estimate equal to zero ( $\hat{\gamma} = 0$ ).

Figure 2 shows the EVI estimates  $\hat{\gamma}$  plotted against their corresponding number of exceedances  $k$ . The values of  $\hat{\gamma}$  are either negative or around zero for all values of  $k$ . Additionally, the generalised QQ plot has a somewhat or approximately horizontal tail pattern. This suggests that a distribution from the Gumbel domain may be a good candidate for the data. This thus supports our selection of the Weibull, Exponential and Lognormal distributions for representing the NO<sub>2</sub> emissions data from Arnot power station. The Pareto distribution is included for comparison purposes.

#### 4.4. The Distribution of the Data

Figure 3 below shows the quantile and derivative plots, together with the PP and density plots of the three distributions (Weibull, Lognormal and Pareto) fitted in this study.

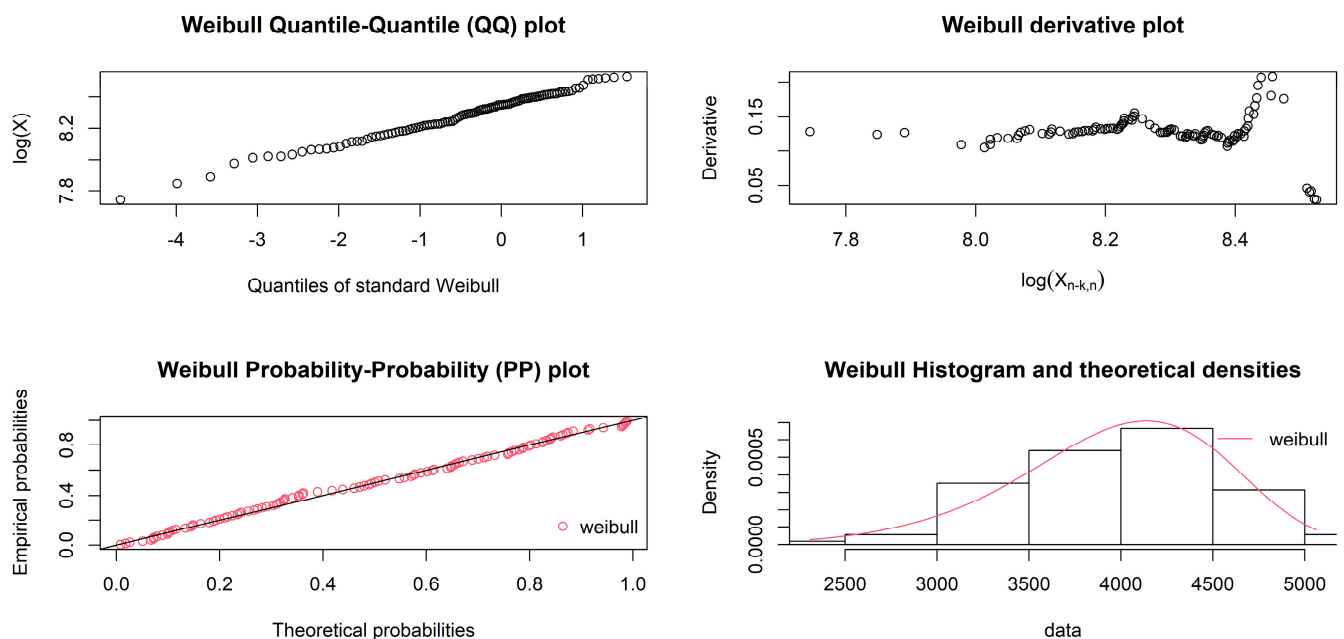


**Figure 3.** QQ plot (top left), derivative plot (top right), and PP plot (bottom left) and density plot (bottom right) for the Weibull distribution.

In Figure 3, the Weibull distribution produced QQ and PP plots that do not show significant deviation from the 45° line, a density plot that shows minimal deviation from the fitted line, and a derivative plot that is approximately constant except at the tail end, where it increases before decreasing. This indicates that the Weibull distribution is a good fit for the NO<sub>2</sub> emissions data for the greater part of the body of the data, while towards the tail end, a heavier-tailed distribution like the Lognormal or Pareto distribution may be more appropriate.

Considering the plots in Figure 4, an overall good fit is obtained from the Lognormal distribution since minimal deviation from the Lognormal is observed from the expected on the QQ, PP and density plots. However, taking a closer look at the derivative plot, an overall somewhat decreasing pattern is observed, suggesting that a lighter-tailed distribution than the Lognormal distribution may be a better fit for the NO<sub>2</sub> emissions data. The slightly upward kink in the derivative suggests very extreme emissions near the tail end of the data. This implies that a heavier-tailed distribution than the Lognormal distribution, like the Pareto, may be more appropriate in this region of the data. The QQ plot shows some deviations with values below the fitted line in the tails, thus supporting the decision of a lighter-tailed distribution for the data in some sections of the data.

The rapidly decreasing Pareto distribution derivative plot and the downward curving QQ plot in Figure 5 indicate that the Pareto distribution is not a good fit and that a lighter-tailed distribution than the Pareto distribution may be a good fit. However, it seems to capture the very extreme observations at the very tail end. In this region, the Pareto derivative plot is approximately constant.

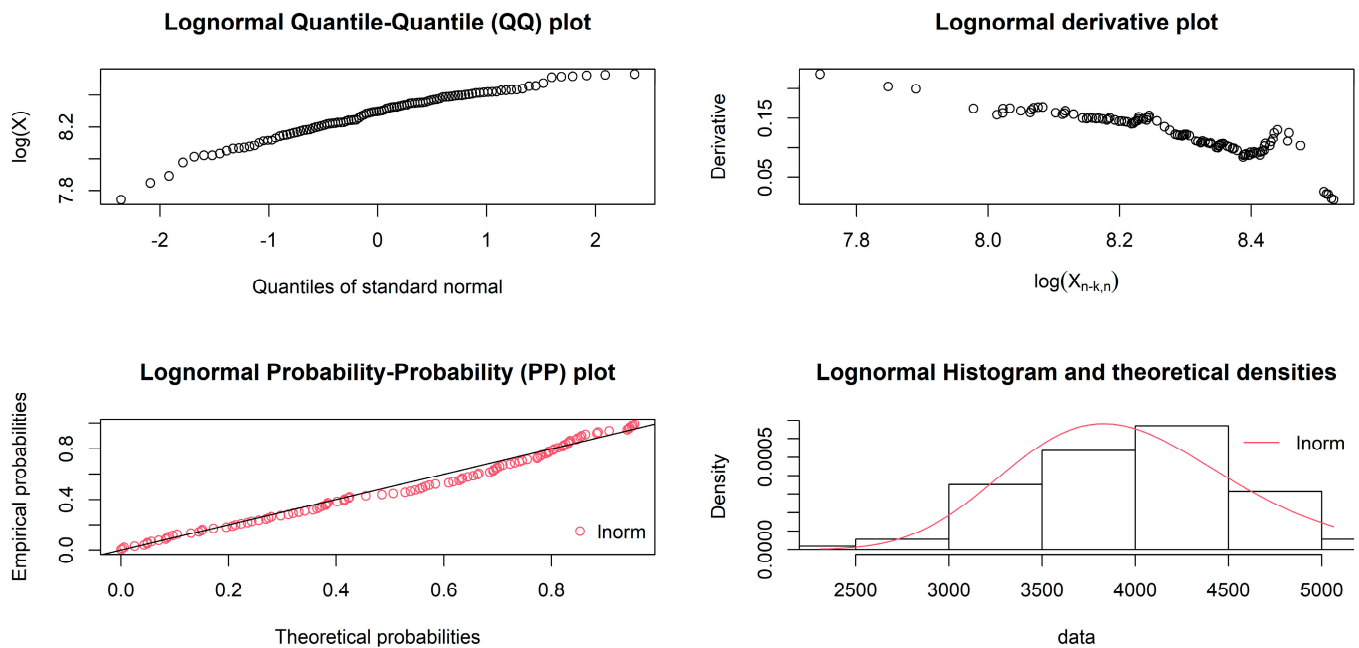


**Figure 4.** QQ plot (top left), derivative plot (top right), and PP plot (bottom left) and density plot (bottom right) for the Lognormal distribution.

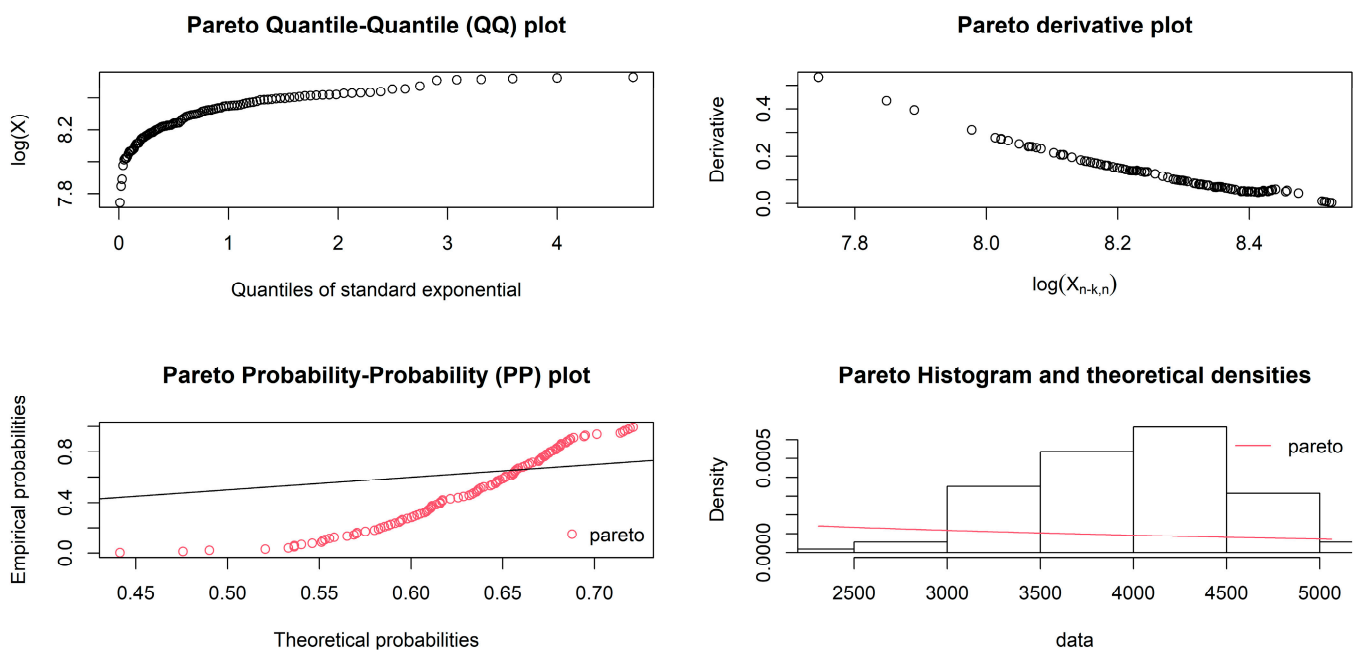


**Figure 5.** QQ plot (top left), derivative plot (top right), and PP plot (bottom left) and density plot (bottom right) for the Pareto distribution.

#### 4.5. Goodness of Fit Tests

The KS goodness of fit test for each of the three distributions (Weibull, Lognormal and Pareto distributions) is performed. It tests the null hypothesis  $H_0$ : The data follows the specified distribution. The Vasicek–Song (VS) goodness-of-fit test is also included to assess how well the selected distributions fit the data. This test has more power compared to other goodness-of-fit tests like the KS test [30]. Table 3 presents the results of the test and the maximum likelihood (ML) parameter estimates for the three distributions.

**Table 3.** Kolmogorov–Smirnov (KS) goodness of fit test and maximum likelihood (ML) parameter estimates for the Weibull, Lognormal and Pareto distributions.

| Component                 | Sample Size (n) | Test and Parameter Estimates | Distribution |                 |             |                        |                |                       |
|---------------------------|-----------------|------------------------------|--------------|-----------------|-------------|------------------------|----------------|-----------------------|
|                           |                 |                              | Weibull      |                 | Lognormal   |                        | Pareto         |                       |
|                           |                 |                              | Statistic    | <i>p</i> -value | Statistic   | <i>p</i> -value        | Statistic      | <i>p</i> -value       |
| Bulk<br>( $x \leq 4506$ ) | n = 89          | KS                           | 0.0700       | 0.7489 ***      | 0.0947      | 0.3785                 | 0.3781         | <0.0001               |
|                           |                 | VS                           | 0.1693       | 0.4053 ***      | 0.2532      | $1.043 \times 10^{-7}$ | 1.1150         | $2.2 \times 10^{-16}$ |
|                           |                 | Parameter estimates          | $\hat{\tau}$ | $\hat{\lambda}$ | $\hat{\mu}$ | $\hat{\sigma}$         | $\hat{\alpha}$ | $\hat{\beta}$         |
|                           |                 |                              | 9.7914       | 4003.8568       | 8.2336      | 0.1359                 | 2.0447         | 2309                  |
| Tail<br>( $x > 4506$ )    | n = 19          | KS                           | 0.2311       | 0.2245          | 0.2200      | 0.2742                 | 0.1857         | 0.4734 ***            |
|                           |                 | VS                           | 0.8995       | 0.0056          | 0.7438      | 0.0014                 | 0.4340         | 0.105 ***             |
|                           |                 | Parameter estimates          | $\hat{\tau}$ | $\hat{\lambda}$ | $\hat{\mu}$ | $\hat{\sigma}$         | $\hat{\alpha}$ | $\hat{\beta}$         |
|                           |                 |                              | 24.0977      | 4833.2211       | 8.4612      | 0.0424                 | 21.1226        | 4509                  |

$x$  is NO<sub>2</sub> emissions from the Arnot power station. \*\*\* The best-fitting distribution.

The data in Table 3 are divided into two components as suggested by the derivative plots above, with the first component representing the bulk of the data ( $x \leq 4506$ ) and the second component representing the upper tail ( $x > 4506$ ), that is, the extremely high NO<sub>2</sub> emissions. This is to verify the suitability of the distributions, as a single dataset may be represented by multiple distributions [10].

Table 3 shows that both the Weibull and Lognormal distributions produced *p*-values > 0.05 for the KS goodness of fit test in the bulk of the NO<sub>2</sub> emissions data. This indicates that we fail to reject the null hypothesis and thus conclude that there is sufficient evidence to support that the data belongs to the Weibull and Lognormal distributions on average. The Pareto distribution, however, produced *p*-values < 0.05, indicating rejection of the null hypothesis for this component of the NO<sub>2</sub> emissions data. This goodness-of-fit test shows that the Weibull distribution is an overall best fit for the bulk of the data since it produced the smallest test statistic values with the KS statistic = 0.07. Parameter estimates are obtained by the maximum likelihood method of estimation as  $\hat{\tau} = 9.7914$  and  $\hat{\lambda} = 4003.8568$  for the shape and scale parameters. Since the Weibull distribution produced a shape parameter estimate that is greater than one, that is,  $\hat{\tau} = 9.7914 > 1$ , the tail is lighter than that of an Exponential distribution [10]. In the upper tail, the Pareto distribution with the smallest KS test statistic (=0.1857) estimates the very extreme emissions better than the Weibull and the Lognormal distributions. The ML parameters of the Pareto distribution were estimated and given as  $\hat{\alpha} = 21.1226$  and  $\hat{\beta} = 4509$  for the shape and scale, parameters.

In Table 3, the more powerful VS test explicitly demonstrates the Weibull with  $\tau > 1$  and Pareto distributions as the only and best-fitting distributions for the bulk of the data ( $x \leq 4506$ ) and the upper tail ( $x > 4506$ ), respectively. These distributions have *p*-values greater than 0.05 and the smallest test statistic for the VS test.

Table 4 presents the Akaike Information Criteria (AIC) and Schwarz Bayesian Information Criterion (BIC) for the three distributions within each component. A lower value of the AIC and BIC is preferred, indicating the best fit for each component. The Weibull distribution provides the best fit for the bulk of the NO<sub>2</sub> emissions data, whereas the Pareto distribution produces the best fit in the upper tail, as indicated by the lowest AIC and BIC values in their respective analyses. This confirms the findings of the diagnostic plots (QQ, derivative, PP, and density plots) and the goodness of fit statistics (the KS and VS tests) mentioned above.

**Table 4.** Akaike Information Criteria (AIC) and Schwarz’s Bayesian Information Criterion (BIC) for the Weibull, Lognormal and Pareto distributions.

| Component                 | Sample Size (n) | Model Adequacy | Distribution  |           |              |
|---------------------------|-----------------|----------------|---------------|-----------|--------------|
|                           |                 |                | Weibull       | Lognormal | Pareto       |
| Bulk<br>( $x \leq 4506$ ) | n = 89          | AIC            | 1348.3970 *** | 1366.8680 | 1520.2720    |
|                           |                 | BIC            | 1353.3740 *** | 1371.8450 | 1525.2490    |
| Tail<br>( $x > 4506$ )    | n = 19          | AIC            | 262.8421      | 259.3847  | 247.6116 *** |
|                           |                 | BIC            | 264.7310      | 261.2736  | 249.5005 *** |

$x$  is NO<sub>2</sub> emissions from the Arnot power station. \*\*\* The best-fitting distribution.

## 5. Discussion

The study’s aim was to compare and determine the best-fitting parent probability distributions by comparing the QQ and derivative plots of the Weibull, Lognormal and Pareto distributions, while also considering their tail heaviness rank. These probability distributions cater for varied tail heaviness relative to the Exponential distribution, with the Weibull distribution with  $\tau > 1$  being the lightest-tailed distribution and the Pareto the heaviest. The QQ and derivative plots possess the power of demonstrating the appropriateness of a distribution not just in the bulk of the data but also in the tail, and thus give information of a possible candidate distribution for the data in extremities; see Albrecher et al. [10].

The extreme value index (EVI) was estimated by utilising the generalised Hill, moment, and generalised QQ plot. These estimators showed that the NO<sub>2</sub> emissions data were likely to follow a distribution in the Gumbel domain with either a lighter-tailed distribution than the Exponential distribution, like the Weibull with  $\tau > 1$ , or a heavier-tailed distribution than the Exponential distribution, like the Weibull with  $\tau < 1$  or the Lognormal distribution and Pareto. The generalised QQ plot clearly demonstrated this with a constant horizontal pattern at the largest values, the upper tail, indicating the appropriateness of a distribution from the Gumbel domain. These results demonstrate the justification of the selected distribution for the NO<sub>2</sub> emissions data. However, often the worry is in the extreme upper tail. The Pareto seems to capture some heavy-tail behaviour in this dataset.

The Weibull and Lognormal distributions are common in pollutant emission modelling studies [31–38]. However, their use in modelling pollutant emissions does not usually make a distinction between their tail heaviness ranking to arrive at the best-fitting probability distribution for the upper tail as well. Thus, for the Weibull distribution fitting in particular, it is seldom, if ever, highlighted whether the distribution of the data is lighter (with  $\tau > 1$ ) or heavier-tailed (with  $\tau < 1$ ) than the Exponential distribution. Additionally, there is little evidence, if any at all, of the application of the very powerful joint use of the QQ and derivative plots [10] in the modelling of pollutant emissions, especially the application in NO<sub>2</sub> emissions from coal-fired power stations. The current study then employs the three distributions that cater for varied tails’ heaviness and shapes by jointly utilising the distributions’ QQ and derivative plots to close the gap and explain NO<sub>2</sub> emissions data behaviour more fully with results showing that the Weibull distribution is the model of best fit for the main body of data and the Pareto distribution being a better model for the very extreme emissions.

The energy crisis in several emerging nations continues to be marked by unmet electrical demand resulting from inadequate infrastructure, among other factors. Therefore, increased efforts are anticipated to meet this demand and ensure economic growth and development in a country like South Africa. This coal-fired type of electricity generates emissions resulting from coal combustion. Authorities can utilise statistical models, such

as the Weibull, Lognormal and Pareto distribution models, to formulate policies aimed at the reduction and management of emissions, thereby enhancing understanding of their behaviour and mitigating their impact on humans and the environment. Understanding the tail heaviness of the data is very important, particularly the upper tail, that is, the very large values of NO<sub>2</sub> emissions, since they are associated with exacerbated health risks and even death within a short period. This region of a dataset controls both the frequency and size of the extremely large NO<sub>2</sub> emissions [14]. If the data has a Weibull distribution with  $\tau > 1$ , the tail of the data is lighter than if it were to follow an Exponential distribution, suggesting that lesser extremely large NO<sub>2</sub> emissions observations produced from Arnot power station can be expected in the future and/or future NO<sub>2</sub> emissions extremes may be small in size. Conversely, if the data follows a Weibull with  $\tau < 1$  or Lognormal distribution, the tail of the data is heavier than if it were to follow an Exponential distribution, suggesting that moderately large NO<sub>2</sub> emissions observations can be expected in the future and/or the chances of experiencing very large NO<sub>2</sub> observations are moderate. If the Pareto fits the data well, then the chances of observing very large NO<sub>2</sub> emissions are higher and/or future extremes may be larger in size [14]. Therefore, modelling a heavy upper tail of NO<sub>2</sub> emissions using a lighter-tailed distribution, like the Weibull or Lognormal, may lead to underestimation of risk with potential implications on human lives and the environment [14]. The results demonstrate that NO<sub>2</sub> emissions can be characterised and explained by the Weibull and Pareto distributions. The Weibull distribution has a lighter tail than the Exponential distribution for the main body of data, but the Pareto distribution suggests a heavier tail or the possibility of very extreme emissions. These results are consistent with those by [22,39], where the data were best represented by the Weibull and Pareto, for the bulk and tail components, respectively. The modelling technique employed herein should guide authorities and power utilities, such as Eskom, in the planning and policymaking for air quality management.

## 6. Conclusions

Three distributions, namely, the Weibull, Lognormal and the Pareto distributions, are fitted and compared by making use of the QQ and derivative plots of the distributions. These distributions have varied tails compared to the Exponential tail. The results of the study showed that NO<sub>2</sub> emission data, in tons, from Arnot power station show that the main body of data is best represented by the light-tailed Weibull distribution, and the Pareto is more applicable at the very extreme emissions, as evidenced by the QQ and derivative plots, KS and VS tests, AIC and BIC values, and EVI estimates. The appropriateness of the Pareto in the upper tail indicates that very high and more frequent NO<sub>2</sub> emissions are expected, compared to those that would have been generated by the Weibull and Lognormal distributions.

The findings demonstrate the significance of employing the QQ together with the derivative plots in emissions monitoring, in particular NO<sub>2</sub> emissions from Arnot, a coal-fired power station operated by Eskom. The derivative plots' flexibility of capturing the NO<sub>2</sub> emissions behaviour using the two distributions for the bulk and upper tail of the data is demonstrated in this study. They can also identify if the full emission dataset or its components can be best represented by a lighter or heavier-tailed distribution compared to the one investigated. This facilitates the statistical and air quality analysis of NO<sub>2</sub> emissions by enabling the practitioner to depend on the graphical plotting of the QQ plot and its derivatives, and parametric methods (Weibull and Pareto distributions). Using this modelling approach for the NO<sub>2</sub> emissions, the emissions can be explained and predicted.

This study contributes to the methodology that can be used to analyse and model NO<sub>2</sub> emissions from coal-fuelled power stations, a practical atmospheric global challenge with an impact on human lives and the environment.

Pollutant emissions data is rarely modelled by jointly using the QQ and derivative plots while taking into account the tail heaviness ranking of the distributions used. The current study uses these plots to close the gap and better explain NO<sub>2</sub> emission data patterns. It also provides references for researchers, policy makers, regulatory bodies, and power utilities, such as Eskom, looking to understand or reproduce the methods applied in this study for similar emissions to the environment. The utilisation of these modelling methods across various pollutants, geographic regions, and other coal-fuelled power stations operated by Eskom or otherwise can yield significant insights into the results' applicability and generalisability.

## 7. Further Studies and Limitations

Further studies may explore modelling the NO<sub>2</sub> emissions from Arnot data using extreme value models such as the generalised extreme value distribution (GEVD) and the generalised Pareto distribution (GPD) in the modelling of very high values. These distributions, compared to parent distributions, only focus on extremely high (or low) values of the data and discard the bulk of the data. Thus, providing a more sophisticated approach to understanding the behaviour of the data in the tails compared to parent distributions. The methods applied in this study may also be applied to other Eskom coal-fired power stations.

One major drawback of the derivative plots and the current version of software available for their implementation is that they are limited to only four distributions, namely, the Exponential, Weibull, Lognormal and the Pareto. Thus, developments in that regard would enable further comparisons of other statistical distributions such as the Gamma, Burr and similar distributions frequently used to model emissions.

**Author Contributions:** Writing the original draft of this manuscript, M.W.M.; review, editing, and supervision, D.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

|                 |                                          |
|-----------------|------------------------------------------|
| EVI             | extreme value index                      |
| SA              | South Africa                             |
| NO <sub>2</sub> | nitrogen dioxide                         |
| MLE             | maximum likelihood estimation            |
| BIC             | Schwarz's Bayesian Information Criterion |
| AIC             | Akaike Information Criteria              |
| CO              | carbon monoxide                          |
| O <sub>3</sub>  | ozone                                    |
| SO <sub>2</sub> | sulphur dioxide                          |

|                   |                                                    |
|-------------------|----------------------------------------------------|
| GEVD              | generalised extreme value distribution             |
| GPD               | generalised Pareto distribution                    |
| PM <sub>10</sub>  | particulate matter of size 10 micrometres or less  |
| PM <sub>2.5</sub> | particulate matter of size 2.5 micrometres or less |
| PP                | probability-probability                            |
| EPD               | extended Pareto distribution                       |
| PoT               | peaks over threshold                               |
| QQ                | quantile–quantile                                  |
| N                 | number of observations                             |
| genHill           | generalised Hill                                   |

## Appendix A

**Table A1.** Monthly nitrogen dioxide (NO<sub>2</sub>) emissions data, given in tons.

| Nitrogen Dioxide (NO <sub>2</sub> ) Emission (in Tons) | Month of Emission |
|--------------------------------------------------------|-------------------|
| 3462                                                   | April 2005        |
| 3021                                                   | May 2005          |
| 2309                                                   | June 2005         |
| 2561                                                   | July 2005         |
| 3718                                                   | August 2005       |
| 4238                                                   | September 2005    |
| 4003                                                   | October 2005      |
| 3303                                                   | November 2005     |
| 3048                                                   | December 2005     |
| 4288                                                   | January 2006      |
| 3707                                                   | February 2006     |
| 3667                                                   | March 2006        |
| 3742                                                   | April 2006        |
| 4605                                                   | May 2006          |
| 4119                                                   | June 2006         |
| 3492                                                   | July 2006         |
| 3473                                                   | August 2006       |
| 3184                                                   | September 2006    |
| 4439                                                   | October 2006      |
| 3732                                                   | November 2006     |
| 3526                                                   | December 2006     |
| 4966                                                   | January 2007      |
| 3441                                                   | February 2007     |
| 4332                                                   | March 2007        |
| 3216                                                   | April 2007        |
| 3513                                                   | May 2007          |
| 4597                                                   | June 2007         |
| 4509                                                   | July 2007         |

Table A1. Cont.

| Nitrogen Dioxide (NO <sub>2</sub> ) Emission (in Tons) | Month of Emission |
|--------------------------------------------------------|-------------------|
| 4506                                                   | August 2007       |
| 3337                                                   | September 2007    |
| 3133                                                   | October 2007      |
| 3578                                                   | November 2007     |
| 3808                                                   | December 2007     |
| 3352                                                   | January 2008      |
| 3748                                                   | February 2008     |
| 3928                                                   | March 2008        |
| 3194.496565                                            | April 2008        |
| 4230.084644                                            | May 2008          |
| 4708.410787                                            | June 2008         |
| 4466.214804                                            | July 2008         |
| 4543.381095                                            | August 2008       |
| 3081.554301                                            | September 2008    |
| 3724.211074                                            | October 2008      |
| 3745.350429                                            | November 2008     |
| 4010                                                   | December 2008     |
| 4250.626883                                            | January 2009      |
| 3570.002285                                            | February 2009     |
| 3049.098874                                            | March 2009        |
| 3896.844898                                            | April 2009        |
| 3773.618573                                            | May 2009          |
| 4483.672838                                            | June 2009         |
| 4100.335240                                            | July 2009         |
| 4145.246646                                            | August 2009       |
| 4627.092120                                            | September 2009    |
| 3683.758421                                            | October 2009      |
| 3805.282085                                            | November 2009     |
| 4394.895518                                            | December 2009     |
| 4411.757832                                            | January 2010      |
| 4391.796378                                            | February 2010     |
| 5022.998042                                            | March 2010        |
| 4126.136522                                            | April 2010        |
| 4176.738399                                            | May 2010          |
| 4534.364329                                            | June 2010         |
| 5039.382416                                            | July 2010         |
| 3614.121097                                            | August 2010       |
| 3852.732547                                            | September 2010    |

**Table A1.** *Cont.*

| <b>Nitrogen Dioxide (NO<sub>2</sub>) Emission (in Tons)</b> | <b>Month of Emission</b> |
|-------------------------------------------------------------|--------------------------|
| 4163.229131                                                 | October 2010             |
| 4181.204149                                                 | November 2010            |
| 3394.769128                                                 | December 2010            |
| 4086.758762                                                 | January 2011             |
| 3586.296039                                                 | February 2011            |
| 3976.728056                                                 | March 2011               |
| 3650.302634                                                 | April 2011               |
| 4791.368994                                                 | May 2011                 |
| 4261.699668                                                 | June 2011                |
| 4998                                                        | July 2011                |
| 4537                                                        | August 2011              |
| 4322                                                        | September 2011           |
| 4441.629070                                                 | October 2011             |
| 4458.775161                                                 | November 2011            |
| 4984.964488                                                 | December 2011            |
| 4242.134007                                                 | January 2012             |
| 3547.665060                                                 | February 2012            |
| 4304.482749                                                 | March 2012               |
| 4129                                                        | April 2012               |
| 4700                                                        | May 2012                 |
| 4583                                                        | June 2012                |
| 4404                                                        | July 2012                |
| 4429                                                        | August 2012              |
| 4575                                                        | September 2012           |
| 5063                                                        | October 2012             |
| 3989                                                        | November 2012            |
| 3177                                                        | December 2012            |
| 3239                                                        | January 2013             |
| 3794                                                        | February 2013            |
| 4228                                                        | March 2013               |
| 3348                                                        | April 2013               |
| 3637                                                        | May 2013                 |
| 2671                                                        | June 2013                |
| 4048                                                        | July 2013                |
| 4352                                                        | August 2013              |
| 4521                                                        | September 2013           |
| 4217                                                        | October 2013             |

Table A1. Cont.

| Nitrogen Dioxide (NO <sub>2</sub> ) Emission (in Tons) | Month of Emission |
|--------------------------------------------------------|-------------------|
| 3793                                                   | November 2013     |
| 2915                                                   | December 2013     |
| 3961                                                   | January 2014      |
| 4030                                                   | February 2014     |
| 4019                                                   | March 2014        |

## References

- Riekert, J.W.; Koch, S.F. Projecting the external health costs of a coal-fired power plant: The case of Kusile. *J. Energy S. Afr.* **2012**, *23*, 52–66. [\[CrossRef\]](#)
- Ateba, B.B.; Prinsloo, J.J. Strategic management for electricity supply sustainability in South Africa. *Util. Policy* **2019**, *56*, 92–103. [\[CrossRef\]](#)
- Pollet, B.G.; Staffell, I.; Adamson, K.-A. Current energy landscape in the Republic of South Africa. *Int. J. Hydrogen Energy* **2015**, *40*, 16685–16701. [\[CrossRef\]](#)
- He, H.; Schäfer, B.; Beck, C. Spatial heterogeneity of air pollution statistics in Europe. *Sci. Rep.* **2022**, *12*, 12215. [\[CrossRef\]](#)
- Shah, A.S.V.; Lee, K.K.; McAllister, D.; Hunter, A.; Nair, H.; Whiteley, W.; Langrish, J.P.; E Newby, D.; Mills, N.L. Short term exposure to air pollution and stroke: Systematic review and meta-analysis. *BMJ* **2015**, *350*, h1295. [\[CrossRef\]](#)
- Jaffar, M.I.; Hamid, H.A.; Yunus, R.; Raffee, A.F. Fitting Statistical Distribution Functions of Air Pollutant Concentration in Different Urban Locations in Malaysia. *J. Eng. Sci. Res.* **2023**, *7*, 7–11. [\[CrossRef\]](#)
- Dário, M.S.; Novais, D.G.; Pauliquevis, T.; Rizzo, L.V. Long-term trends and probability distribution functions of air pollutant concentrations in the megacity of São Paulo. *Derbyana* **2024**, *45*. [\[CrossRef\]](#)
- Jiang, X.; Deng, S.; Liu, N.; Shen, B. The statistical distributions of SO<sub>2</sub>, NO<sub>2</sub> and PM<sub>10</sub> concentrations in Xi'an, China. In Proceedings of the 2011 International Symposium on Water Resource and Environmental Protection, Xi'an, China, 20–22 May 2011; pp. 2206–2212. [\[CrossRef\]](#)
- Lu, H.-C. The statistical characters of PM10 concentration in Taiwan area. *Atmos. Environ.* **2002**, *36*, 491–502. [\[CrossRef\]](#)
- Albrecher, H.; Beirlant, J.; Teugels, J.L. *Reinsurance: Actuarial and Statistical Aspects*; John Wiley & Sons: Hoboken, NJ, USA, 2017.
- Beirlant, J.; Bladt, M. Tail classification using non-linear regression on model plots. *Extremes* **2025**, *28*, 345–369. [\[CrossRef\]](#)
- Perišić, M.; Stojić, A.; Stojić, S.S.; Šoštarić, A.; Mijić, Z.; Rajšić, S. Estimation of required PM10 emission source reduction on the basis of a 10-year period data. *Air Qual. Atmos. Health* **2015**, *8*, 379–389. [\[CrossRef\]](#)
- Akhundjanov, S.B.; Devadoss, S.; Luckstead, J. Size distribution of national CO<sub>2</sub> emissions. *Energy Econ.* **2017**, *66*, 182–193. [\[CrossRef\]](#)
- Papalexiou, S.M.; Koutsoyiannis, D.; Makropoulos, C. How extreme is extreme? An assessment of daily rainfall distribution tails. *Hydrol. Earth Syst. Sci.* **2013**, *17*, 851–862. [\[CrossRef\]](#)
- Karaca, F.; Alagha, O.; Ertürk, F. Statistical characterization of atmospheric PM10 and PM2.5 concentrations at a non-impacted suburban site of Istanbul, Turkey. *Chemosphere* **2005**, *59*, 1183–1190. [\[CrossRef\]](#) [\[PubMed\]](#)
- Tharu, N.K.; Baidhya, S. Modeling PM2.5 and TSP Concentrations: A Comparison of Weibull, Lognormal, and Gamma Distributions. *Patan Prospect. J.* **2024**, *4*, 69–78. [\[CrossRef\]](#)
- Meirelles, M.G.; Vasconcelos, H.C. Physical–Statistical Characterization of PM10 and PM2.5 Concentrations and Atmospheric Transport Events in the Azores During 2024. *Earth* **2025**, *6*, 54. [\[CrossRef\]](#)
- Choopradit, B.; Paitoon, R.; Srinuan, N.; Kwankaew, S. Application of the Parametric Bootstrap Method for Confidence Interval Estimation and Statistical Analysis of PM2.5 in Bangkok. *WSEAS Trans. Environ. Dev.* **2024**, *20*, 215–225. [\[CrossRef\]](#)
- Oguntunde, P.E.; Odetunmbi, O.A.; Adejumo, A.O. A Study of Probability Models in Monitoring Environmental Pollution in Nigeria. *J. Probab. Stat.* **2014**, *2014*, 864965. [\[CrossRef\]](#)
- Aydin, D. A comparison of the statistical distributions of air pollution concentrations in Sinop, Turkey. *Environ. Prot. Eng.* **2024**, *50*, 47. [\[CrossRef\]](#)
- Osatohanmwen, P.; Oyegue, F.O.; Ogbonmwan, S.M.; Muhwava, W. A General Framework for Generating Three-Components Heavy-Tailed Distributions with Application. *J. Stat. Theory Appl.* **2024**, *23*, 290–314. [\[CrossRef\]](#)
- Deng, M.; Aminzadeh, M.S. Bayesian predictive analysis for Weibull-Pareto composite model with an application to insurance data. *Commun. Stat.-Simul. Comput.* **2022**, *51*, 2683–2709. [\[CrossRef\]](#)
- Cooray, K.; Cheng, C.-I. Bayesian estimators of the lognormal–Pareto composite distribution. *Scand. Actuar. J.* **2015**, *2015*, 500–515. [\[CrossRef\]](#)

24. Nadarajah, S.; Bakar, S.A.A. New composite models for the Danish fire insurance data. *Scand. Actuar. J.* **2014**, *2014*, 180–187. [[CrossRef](#)]
25. Albrecher, H.; Araujo-Acuna, J.C.; Beirlant, J. TEMPERED PARETO-TYPE MODELLING USING WEIBULL DISTRIBUTIONS. *ASTIN Bull.* **2021**, *51*, 509–538. [[CrossRef](#)]
26. Dekkers, A.L.M.; Einmahl, J.H.J.; De Haan, L. A Moment Estimator for the Index of an Extreme-Value Distribution. *Ann. Stat.* **1989**, *17*, 1833–1855. [[CrossRef](#)]
27. Coles, S. *An Introduction to Statistical Modeling of Extreme Values*; Springer: London, UK, 2001.
28. Hill, B.M. A Simple General Approach to Inference About the Tail of a Distribution. *Ann. Stat.* **1975**, *3*, 1163–1174. [[CrossRef](#)]
29. Dobson, A.J.; Barnett, A.G. *An Introduction to Generalized Linear Models*, 3rd ed.; Texts in Statistical Science Series; Chapman & Hall/CRC: Boca Raton, FL, USA, 2008.
30. Lequesne, J.; Regnault, P. vsgoftest: An R Package for Goodness-of-Fit Testing Based on Kullback-Leibler Divergence. *Code Snippet 1. J. Stat. Softw.* **2020**, *96*, 1–26. [[CrossRef](#)]
31. Aleksandropoulou, V.; Eleftheriadis, K.; Diapouli, E.; Torseth, K.; Lazaridis, M. Assessing PM 10 source reduction in urban agglomerations for air quality compliance. *J. Environ. Monit.* **2012**, *14*, 266–278. [[CrossRef](#)]
32. Sharma, P.; Sharma, P.; Jain, S.; Kumar, P. An integrated statistical approach for evaluating the exceedence of criteria pollutants in the ambient air of megacity Delhi. *Atmos. Environ.* **2013**, *70*, 7–17. [[CrossRef](#)]
33. Mishra, G.; Ghosh, K.; Dwivedi, A.K.; Kumar, M.; Kumar, S.; Chintalapati, S.; Tripathi, S. An application of probability density function for the analysis of PM2.5 concentration during the COVID-19 lockdown period. *Sci. Total Environ.* **2021**, *782*, 146681. [[CrossRef](#)]
34. Nwaigwe, C.C.; Ogbonna, C.J.; Achem, O. On the Modeling of Carbon Monoxide Flaring in Nigeria. *Int. J. Stat. Probab.* **2018**, *7*, 94. [[CrossRef](#)]
35. Gao, C.; Deng, S.; Jiang, X.; Guo, Y. Analysis for the Relationship Between Concentrations of Air Pollutants and Meteorological Parameters in Xi'an, China. *J. Test. Eval.* **2016**, *44*, 1064–1076. [[CrossRef](#)]
36. De Souza, A.; Olaofe, Z.O.; 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*, 569. [[CrossRef](#)]
37. Gulia, S.; Nagendra, S.M.S.; Khare, M. Extreme Events of Reactive Ambient Air Pollutants and their Distribution Pattern at Urban Hotspots. *Aerosol Air Qual. Res.* **2017**, *17*, 394–405. [[CrossRef](#)]
38. Hrishikesh, C.G.; Nagendra, S.M.S. Study of meteorological impact on air quality in a humid tropical urban area. *J. Earth Syst. Sci.* **2019**, *128*, 118. [[CrossRef](#)]
39. Scollnik, D.P.M.; Sun, C. Modeling with Weibull-Pareto Models. *N. Am. Actuar. J.* **2012**, *16*, 260–272. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.