JOURNAL of ENERGY in SOUTHERN AFRICA

Volume 28 Number 1

Quantifying South Africa's sulphur dioxide emission efficiency in coal-powered electricity generation by fitting the three-parameter log-logistic distribution

Maseapei Elizabeth Girmay*, Delson Chikobvu

Department of Mathematical Statistics and Actuarial Science, University of the Free State, PO Box 339, Bloemfontein 9300, South Africa

Abstract

This paper fits the three-parameter log-logistic (3LL) distribution to sulphur dioxide (SO_2) monthly emissions in kilograms per gigawatt hour (kg/GWh) and in milligrams per cubic nano metre (mg/Nm^3), at 13 of Eskom's coal fired power-generating stations in South Africa. The aim is to quantify and describe the emission of sulphur dioxide at these stations using a statistical distribution, and to also estimate the probabilities of extreme emissions and exceedances (emissions above a certain threshold). Using the 3LL distribution is proposed as such a distribution. The log-logistic distribution is a special form of a Burr-type distribution. Various goodness-of-fit measures, including the Kolmogorov Smirnov, the Anderson Darling and some graphical tests, are employed to test if the 3LL distribution is a good fit to the data. The maximum likelihood method is used to estimate the parameters. The distribution fit is important as it then becomes possible to quantify and manage the SO_2 emissions effectively. The 3LL distribution, which is compared with three other distributions, gave the best overall fit to most of the power stations.

Keywords: *emission, Eskom, log logistic distribution, goodness of fit, sulphur dioxide, Burr-type distribution*

Highlights

- *Quantification of SO_2 emissions in terms of a statistical distribution*
- *Calculating the probability of SO_2 emissions exceeding certain specified limits*
- *Ranking power stations in terms of SO_2 emissions efficiency*

Journal of Energy in Southern Africa 28(1): 91–103
DOI: http://dx.doi.org/10.17159/2413-3051/2017/v28i1a1530

* *Corresponding author: Tel: +27 (0)51 401 9697*
Email: girmayme@ufs.ac.za

1. Introduction

Eskom is South Africa's electricity public utility, established in 1923 as the Electricity Supply Commission by the government of South Africa. It is the largest producer of electricity in Africa, and among the top seven utilities in the world in terms of generation capacity and among the top nine in terms of sales. Eskom generates approximately 95% of electricity used in South Africa. About 95% of its generating capacity comes from coal. Ash emissions from Eskom's coal-fired power stations have been reduced by more than 90% since the early 1980s due to the installation of efficient pollution abatement technology and the decommissioning of older plants (Eskom Emission Monitoring, 2012). At present, Eskom has 13 coal-fired generating stations giving various emissions, including sulphur dioxide (SO_2). Medupi and Kusile are two new stations and are not included in the analysis.

Coal-fired power stations release harmful chemicals (stack emissions) into the atmosphere, causing environmental problems. Sulphur dioxide, for example, is a precursor to acid deposition (including acid rain) and secondary particulate matter formation, and is also toxic. Eskom must comply with legally prescribed limits on a number of emissions to avoid heavy penalties. Exceeding the emission limits may result in the forced shutdown of generating units. Emission levels must therefore be monitored and alarmed continuously. A stack test is a procedure for sampling a gas stream from a single sampling location at a facility, unit, or pollution control equipment. Each of Eskom's stations could be a unit, and there could also be more than one unit at a station. The stack is used to determine a pollutant emission rate, concentration, or parameter while the facility, unit, or pollution control equipment is operating at conditions that result in the measurement of the highest emission or parameter values (prior to any control device) or at other operating conditions approved by the regulatory authority. Stack emission control programmes are effective only if emissions are controlled at the source, which requires a highly accurate monitoring schedule. Stack emissions are measured for various reasons, including to determine if emission permits requirements are met or exceeded, for emissions permit renewal, or for process control purposes. Reporting on environmental performance has several benefits, including providing management with information to help exploit the cost savings that good environmental performance usually brings; it also gives Eskom the chance to set out what they believe is significant in their environmental performance. Companies that measure (to quantify), manage and communicate their environmental performance are inherently well placed. They understand how to improve their processes, reduce their costs, comply with regulatory requirements and stakeholder expectations, and take advantage of new technologies on the market. This paper describes SO_2 emissions at 13 Eskom coal-fired power stations measured by fitting a three-parameter log-logistic (3LL) distribution. The maximum likelihood method is used to estimate the parameters. The 3LL is compared with three other distributions, namely the normal, log-normal and three-parameter Weibull. A literature review follows in Section 2; methodology is outlined in Section 3; the data and results are given in Section 4 and conclusions in Section 5.

2. Literature review

Little if any modelling of pollutant emissions in South Africa is done. Geogopolous et al. (1982) stated that the answer to the question of which distribution best fits the air quality/emissions data was shown to depend in general on the pollutant, the time period of interest, the averaging time of the data, the location and other factors. Generally there is no priori reason to choose one particular distribution over the other (Seinfeld et al., 1998). Yi Zhang et al. (1994) investigated the statistical distribution of on-road carbon monoxide and hydrocarbon emissions from various locations in the United States, and found that the emissions are statistically gamma−distributed**.** Rumburg et al. (2001) investigated the statistical distribution of particulate matter in Spokane, Washington and concluded that the distribution was best fitted by a three-parameter log-normal distribution and a generalised extreme value distribution. Hadley et al. (2003) investigated the distribution of annual mean daily SO_2 in the United Kingdom and found the log-normal distribution to be a better fit for the data than the normal. One of the most important papers on modelling greenhouse gas emissions is that of Smith (1989), where he applies the extreme value theory to the study of hourly readings of ozone in Houston, Texas, since excessive levels of ozone are taken to indicate high air pollution. Smith concluded that the exponential distribution gave a poor fit in the upper tail whereas the generalised Pareto distribution seemed adequate.

The method of estimation of parameters for the chosen statistical distribution is also important when emission data is available. Ashkar et al. (2003) compared the maximum likelihood (ML) and the generalised probability weighted moments (GPWM) in estimating the parameters of a log-logistic, and found that the ML outperformed the GPWM method over all parameter space and sample sizes. Ashkar et al. (2006) compared the method of generalised moments, ML, the methods of the GPWM and the method of log moments for the estimation of the parameters and quantiles in the two-parameter log-logistic. Their simulation results showed that the GM method outperformed the other competitive methods in the two-parameter log-logistic case when the moment orders are appropriately chosen. Abbas et

al. (2015) proposed the Bayesian method using the metropolis algorithm within the Gibbs sampling under the reference prior to estimate the parameters of the log-logistic. Singh et al. (1993) developed a new competitive method of estimating parameters of the log-logistic based on the principle of maximum entropy (POME) using the Monte Carlo simulated data, and compared it to the methods of moments (MOM), ML, and the probability weighted moments (PWM). They concluded that POME yielded the least parameter bias for all sample sizes. Other parameter estimation problems for the log-logistic distribution are addressed by, among others, Kantam and Srinivasa (2002), Balakrishnan et al. (1987), and Tiku and Suresh (1992). The log-logistic distribution is a special type of a Burr-type XII distribution. Burr (1942) introduced twelve different forms of cumulative distribution functions for modelling data, of which Burr-type X and Burr-type XII received the maximum attention. There is also a thorough analysis of Burr-type XII distribution in Rodriguez (1977), and see also Wingo (1993). Burr-type distributions are very flexible and can be adapted to fit many situations. In this paper the three-parameter log-logistic distribution from the Burr-type XII family is used. The normal is one of the most commonly used distributions where there is a large data set. Log-normal is one of the distributions commonly mentioned in the literature on emissions; the variance of the log-normal increases with the mean. The three-parameter Weibull distribution is also one of the commonly used reported distributions, and it is also heavy-tailed.

3. Research methodology

This section describes the statistical distribution used to describe the SO_2 emissions data and the method of estimating the parameters. The log-logistic distribution is a continuous probability distribution for a non-negative random variable. It is the probability distribution of a random variable whose logarithm has a logistic distribution. It is generated by a transformation of logistic variable, just like the log-normal distribution is obtained from the normal distribution. It is similar in shape to the log-normal distribution but has heavier tails. The log-logistic distribution is a special case of the Burr-type XII distribution (Burr, 1942) and also a special case of the Kappa distribution (Mielke & Johnson, 1973). The Burr distribution is very relevant in the study of atmospheric emissions, it is more flexible and has heavier tails and is then able to model extreme emissions. Its cumulative distribution function can be written in closed form, unlike that of the log-normal. The 3LL distribution has a probability density function given as in Equation 1.

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{-2} \quad \alpha>0, \beta>0, \gamma \le x < +\infty \tag{1}$$

and a cumulative distribution given by Equation 2:

$$F(x) = \left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{-1} \quad \alpha>0, \beta>0, \gamma \le x < +\infty \tag{2}$$

where $\beta > 0$ is a scale parameter, $\alpha > 0$ is a shape parameter and $\gamma \in R$ is a location parameter. The distribution is unimodal when the shape parameter $\alpha > 1$ and its dispersion decreases as the shape parameter α increases.

The mean and variance of the 3LL distribution are given by Equations 3 and 4 respectively:

$$E(X) = \gamma + \beta B\left(1+\frac{1}{\alpha}, 1-\frac{1}{\alpha}\right) \tag{3}$$

$$Var(x) = \beta^2\left[B\left(1+\frac{2}{\alpha}, 1-\frac{2}{\alpha}\right) - B^2\left(1+\frac{1}{\alpha}, 1-\frac{1}{\alpha}\right)\right] \tag{4}$$

3.1 Parameter estimation

The parameters are estimated using the ML method. The ML is asymptotically normal, has the smallest asymptotic variance, and is asymptotically efficient and optimal. With the ML approach, the distributions of the estimators become more and more concentrated near the true value of the parameter being estimated as the sample size (n) increases.

Maximum likelihood method: The probability density function of the 3LL distribution is given in Equation (1) above, with the log likelihood given in Equation 5:

$$L(x;\alpha,\beta,\gamma) = nln\alpha - nln\beta + (\alpha-1)\sum_{i=1}^{n} ln\left(\frac{x_i-\gamma}{\beta}\right) - 2\sum_{i=1}^{n} ln\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right) \tag{5}$$

where n is the sample size. Taking the partial derivatives of each parameter gives Equations 6–8:

$$\frac{\partial L(x;\alpha,\beta,\gamma)}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n} ln\left(\frac{x_i-\gamma}{\beta}\right) - 2\sum_{i=1}^{n}\frac{\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha} ln\left(\frac{x_i-\gamma}{\beta}\right)}{\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)} \tag{6}$$

$$\frac{\partial L(x;\alpha,\beta,\gamma)}{\partial \beta} = -\frac{n}{\beta} - \frac{\alpha-1}{\beta} + \frac{2\alpha}{\beta}\sum_{i=1}^{n}\frac{\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}}{\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)} \tag{7}$$

$$\frac{\partial L(x;\alpha,\beta,\gamma)}{\partial \gamma} = -(\alpha-1)\sum_{i=1}^{n}\frac{1}{x_i-1} + \frac{2\alpha}{\beta}\sum_{i=1}^{n}\frac{\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha-1}}{\left(1+\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha}\right)} \tag{8}$$

The maximum likelihood estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ are obtained by setting each of the above equations to zero and solving them simultaneously. Optimisation computer algorithms such as the Newton-Raphson, Nelder-Mead, and simulated annealing are often used to arrive at the estimates. The normal, log-normal and three-parameter Weibull distribution parameters are estimated in a similar way.

3.2. Goodness of fit tests

The Anderson-Darling, Chi Squared, and Kolmogorov-Smirnov tests are used to test if the 3LL is a good distribution to fit the data. The probability density function (PDF) plots, cumulative distribution function (CDF) plots, together with the probability-probability (P-P) and quantile-quantile (Q-Q) plots are also used as tests.

The Kolmogorov-Smirnov test is a nonparametric test for the equality of continuous, one-dimensional probability or statistical distributions that can be used to compare a sample of observations with a reference probability distribution or to compare two samples. It tries to determine if two data sets differ significantly, and has the advantage of making no assumption about the statistical distribution of data. It quantifies a distance between the empirical distribution function of the sample and the theoretical cumulative distribution function, or between the empirical distribution functions of two samples. The test statistic of the Kolmogorov Smirnov is given by Equation 9:

$$D_n = max\left\{max_{i=1\ldots.n}\left[\frac{1}{n} - \hat{F}(x_i)\right]; max_{i=1\ldots.n}\left[\hat{F}(x_i) - \frac{i-1}{n}\right]\right\} \quad (9)$$

which is the largest vertical difference between the theoretical and empirical CDF for all values of x (Evans et al., 2008), where n is the sample size.

The Anderson Darling test is also a statistical test of whether a given sample of data is drawn from a given probability distribution. It is a modification of the Kolmogorov-Smirnov test and gives more weight to the tails of a statistical distribution than that test. In its basic form, the test does not assume a particular parametric form of the distribution being tested, in which case the test and its set of critical values is distribution free. The test is, however, most often used in context where a family of distributions is being tested, in which case the parameters of that family need to be estimated and account must be taken of this in adjusting either the test-statistic or its critical values. When applied to testing if a normal distribution adequately describes a set of data, it is one of the most powerful statistical tools for detecting most departures from normality.

The test assesses whether a sample comes from a specified distribution. It compares the observed CDF with the expected CDF and is defined in Equation 10:

$$A_n^2 = n\int_{-\infty}^{\infty} \frac{\left(F_n(x) - \hat{F}(x)\right)^2}{\hat{F}(x)[1-\hat{F}(x)]} d\hat{F}(x) \quad (10)$$

with the computational formula given by Equation 11:

$$A_n^2 = -\sum_{i=1}^{n} \frac{2i-1}{n}\left(ln\left(\hat{F}(x_i)\right) + ln\left(1 - \hat{F}(x_{n+1})\right)\right) - n \quad (11)$$

where n is the sample size. According to Abbas et al. (2012), the Anderson-Darling test is superior where there is greater concern for the extreme values in the data. This is very relevant when considering atmospheric emissions data.

The P-P and Q-Q plots are graphical methods used to test the fit of the distributions to the data. The P-P plot assesses whether or not the data set follows the specified distribution. The P-P plot compares the CDF of the distributions by plotting the theoretical values and the points of the empirical distribution against it. The Q-Q plot compares the distributions by plotting their quantiles against each other. With the Q-Q plot, the data are plotted against the theoretical distribution. For both the P-P and Q-Q plots, if the empirical distribution is close to the theoretical distribution, the graph will be a straight line (Beirlant et al., 2004). Departures from the straight line indicate the departures from the theoretical distribution.

4. Results and discussions

The data is from Eskom, for the period January 2005 to January 2012.The efficiency of the power station cannot be determined by absolute emissions, but by the amount of SO_2 emitted in kilograms per gigawatt hours (kg/GWh) of electricity sent out (relative emissions). To accommodate the non-stationarity in the data, the total emissions of SO_2 emitted (in kilograms or milligrams) per month per power station is divided by the total amount of units of power sent out per power station per month or by the volume. The derived data for the power stations, therefore, represents the amount of emission emitted in kilograms to send out one unit of power in gigawatt hours or in milligrams per cubic nano metre (mg/Nm^3). The resultant variables are a measure of efficiency of the station in emitting SO_2 and will be used for the analysis.

Time plots are useful for checking obvious patterns in the data. The plots represent evolving efficiency at the power stations. Time plots of Lethabo and Camden power stations' SO_2 kg/GWh and SO_2 mg/Nm^3 data are given in Figures 1 and 2 respectively. The rest of the time plots are given in supplementary information.

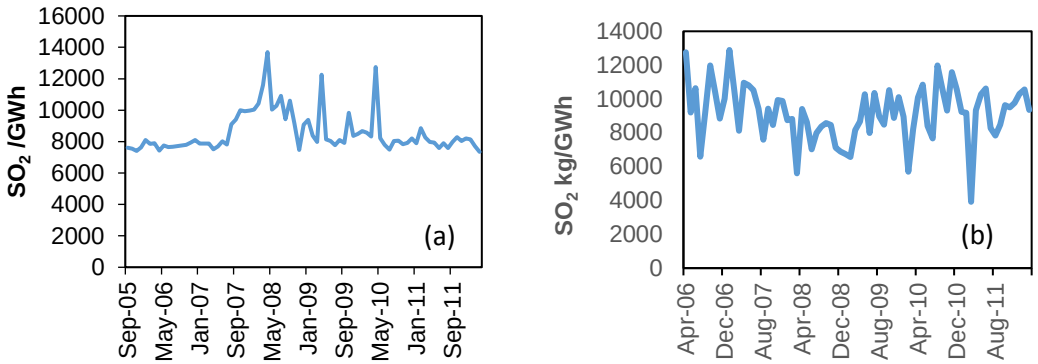


Figure 1: Monthly SO_2 in kg/GWh emissions, where (a) = Lethabo and (b) = Camden.

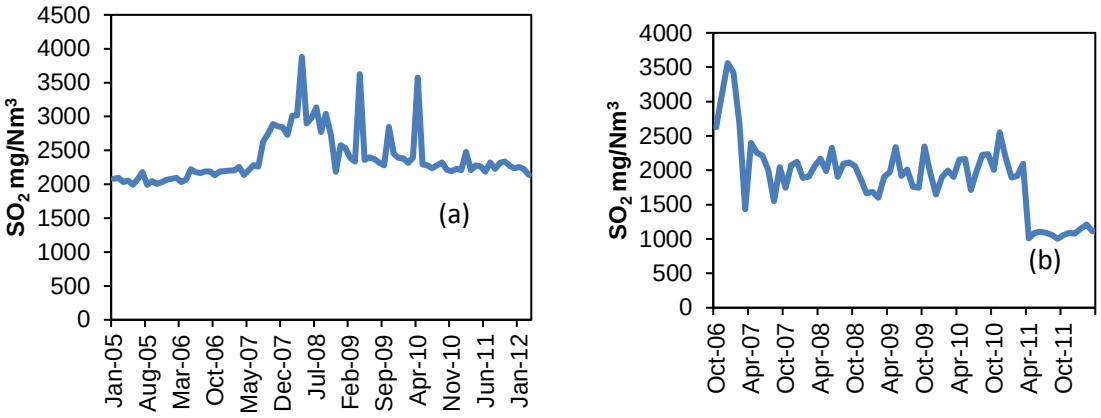


Figure 2: Monthly SO_2 in mg/Nm^3 emissions, where (a) = Lethabo and (b) = Camden.

Table 1: Descriptive statistics of SO_2 in kg/GWh.

Station	*Mean SO_2 (kg/GWh)*	*Standard deviation SO_2 (kg/GWh)*	*Skewness SO_2 (kg/GWh)*	*Kurtosis SO_2 (kg/GWh)*
Arnot	6 270.2	862.76	0.2815	1.3013
Camden	9 162.7	1625.80	-0.4128	0.9573
Duhva	8 572.0	1325.60	0.8272	2.1535
Grootvlei	8 122.3	1791.90	-0.4895	0.9706
Hendrina	8 565.6	1577.10	0.8250	0.0876
Kendal	8 972.5	1388.90	-1.7983	13.7830
Komati	7 031.5	1996.80	-0.2294	2.5996
Kriel	6513.7	746.55	0.6170	0.0591
Lethabo	8468.5	1248.80	2.1048	4.6931
Majuba	8125.6	1247.00	0.6481	0.7672
Matimba	10807	787.27	0.1678	0.3873
Matla	8665.2	10488.00	0.6109	5.4594
Tutuka	8790.1	860.30	-0.0552	0.0927

Arnot, Komati, and Kriel have average monthly SO_2 emissions of around 6000 kg/GWh; Duhva, Grootvlei, Hendrina, Hendrina, Lethabo, Majuba, Matimba, Matla and Tutuka average 8000 kg/GWh; and Camden, Kandal and Matimba average emission around 9000 kg/GWh.

For SO_2 in mg/Nm^3 most power stations (Camden, Duhva, Hendrina, Kendal, Lethabo, Majuba, Matla and Tutuka) have an average monthly emission of 2000, and Arnot, Grootvlei, Komati and Kriel 1600. Graphs for Lethabo and Camden are given in Figure 2, the remainder in the supplementary information. Matimba had the highest monthly average of 3000 mg/Nm^3. Matimba received the 2011 National Association of Clean Air award for consistent reduction of point source particulate emission (COPFIT Fact Sheet, Eskom, 2012), but it seems not to be the case for SO_2 emissions. An augmented Dickey-Fuller test for stationarity was carried out for all of the stations for both SO_2 in kg/GWh and SO_2 in mg/Nm^3 at 5% significance level, showing that all stations were stationary, that none of the stations had a trend.

Table 1 gives the descriptive statistics of the power stations for SO_2 in kg/GWh, and Table 2 gives the statistics for SO_2 in mg/Nm^3. Figures 3 and 4 give graphical plots of the monthly means for each station for both SO_2 in kg/GWh and SO_2 in mg/Nm^3 from the worst to the better emitters.

Table 2: Descriptive statistics of SO_2 in mg/Nm^3.

Station	*Mean SO_2 (mg/Nm^3)*	*Standard deviation SO_2 (mg/Nm^3)*	*Skewness SO_2 (mg/Nm^3)*	*Kurtosis SO_2 (mg/Nm^3)*
Arnot	1634.2	181.70	-0.7008	2.0968
Camden	1912.3	535.06	0.4380	1.2521
Duhva	2292.0	347.54	0.4444	2.4256
Grootvlei	1880.6	316.99	-0.1991	-0.1948
Hendrina	2018.1	317.94	1.0537	0.5400
Kendal	2388.5	371.78	1.8954	3.9019
Komati	2335.0	234.96	0.2570	-0.2139
Kriel	1623.3	263.42	1.4775	3.5454
Lethabo	1654.9	210.38	0.4208	-0.5615
Majuba	2033.4	249.95	0.6445	0.3236
Matimba	3165.4	296.94	-0.5155	0.2562
Matla	2302.8	313.61	1.0767	5.7380
Tutuka	2325.0	187.36	0.1694	-0.2861

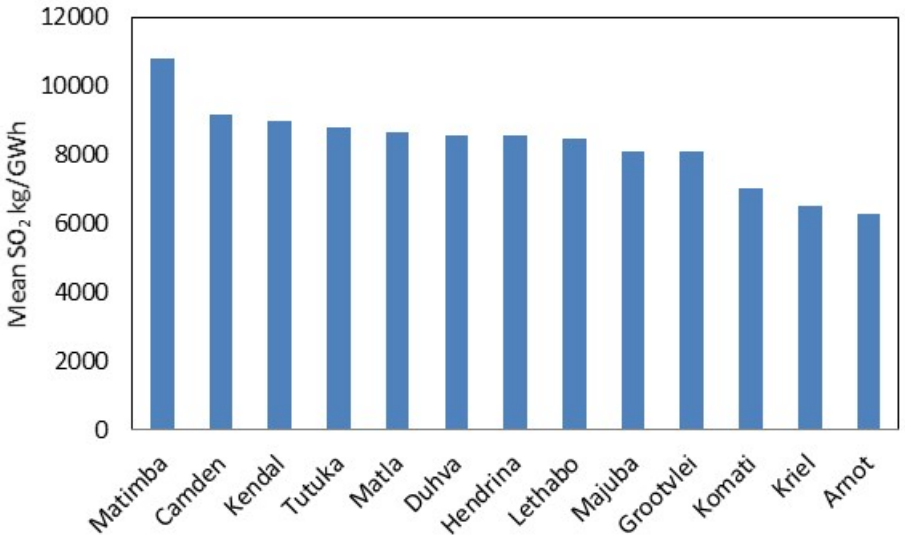


Figure 3: SO_2 in kg/GWh monthly means in an ascending order of efficiency.

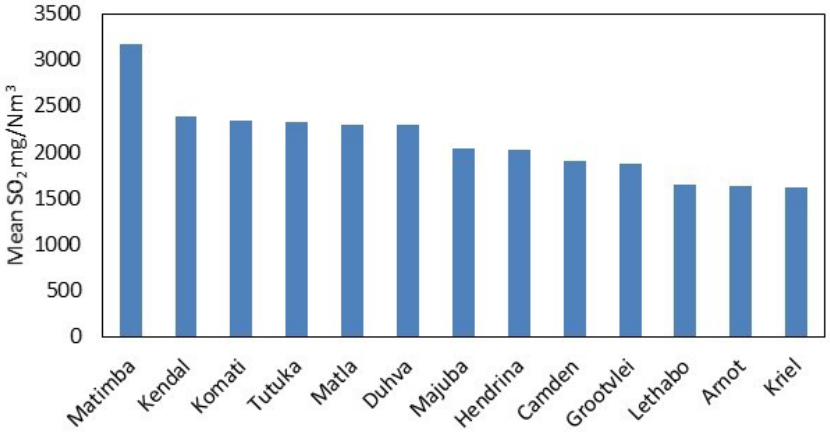


Figure 4: The SO_2 in mg/Nm^3 monthly means in an ascending order of efficiency.

Table 3: Abatement technology used in Eskom power stations (Eskom, 2012).

Power station	*Abatement technology*
Arnot	Fabric filter plants
Camden	Fabric filter plants
Duhva Unit 1 – 3	Fabric filter plants
Duhva Unit 4 – 6	Electrostatic precipitators and flue gas conditioning
Grootvlei Units 1, 5, 6	Fabric filter plants
Grootvlei Units 2, 3, 4	Electrostatic precipitators and flue gas conditioning
Hendrina	Fabric filter plants
Kendal	Electrostatic precipitators and flue gas conditioning
Komati	Electrostatic precipitators and flue gas conditioning
Kriel	Electrostatic precipitators and flue gas conditioning
Lethabo	Electrostatic precipitators and flue gas conditioning
Majubas	Fabric filter plant
Matimba	Electrostatic precipitators and flue gas conditioning
Matla	Electrostatic precipitators and flue gas conditioning
Tutuka	Electrostatic precipitators

From Figures 3 and 4 it can be concluded that Matimba and Kendal are the least efficient stations in emitting SO_2, with Arnot and Kriel the most efficient.

Eskom installed abatement technologies at each power station to reduce the ash emissions. These technologies are said to have an efficiency of at 99%, and over 99.9% in many cases (COPFIT Fact Sheet, Eskom, 2012). These abatement technologies are given in Table 3.

From Figures 3 and 4 in conjunction with Table 3, it can be concluded that the stations that use an electrostatic precipitators and flue gas conditioning technology are less efficient than the ones using fabric filter plants, but other factors also affect emission efficiency, including the age of the plant and the quality of coal used.

To further describe the distribution of the data we look at skewness and kurtosis of the data. The skewness is defined by Equation 12:

$$s = \frac{E(x-\mu)^3}{\sigma^3} \qquad (12)$$

where μ and σ are the mean and standard deviation of emission efficiency variable x. The skewness shows how symmetric the data is around the mean. A symmetric data has skewness near zero, while negative skewness indicates that the data is spread more to the left of the mean and positive skewness indicates that the data is spread to the right of the mean. By negative-skewed, the left tail is long relative to the right tail, and by positive-skewed the right tail is long relative to the left. For SO_2 in kg/GWh, Camden, Grootvlei¸ Kendal, Komati and Tutuka

have a negative skewness, indicating that their data is more spread to the left of the mean. For the rest of the stations the data is more spread to the right of the mean. For SO_2 in mg/Nm^3, Arnot, Grootvlei, and Matimba have a negative skew, and all the other stations have a positive skew. Only Grootvlei is confirmed with a negative skewness in both data sets.

The kurtosis is defined by Equation 13:

$$k = \frac{E(x-\mu)^4}{\sigma^4} \quad (13)$$

where μ and σ are the mean and standard deviation of x. Kurtosis is a measure of whether the data sets are heavy- or light-tailed relative to the normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or a lack of outliers. If a distribution has a kurtosis less than three, its tail is shorter and thinner and its peak is flatter and broader than the normal distribution (Brown 2011). For SO_2 in kg/GWh, only Kendal, Lethabo and Matla have a kurtosis higher than three, indicating that they have a central peak higher and sharper and that their tails are longer and fatter than that of normal distribution. All the other stations have a tail shorter and thinner and their peaks are flatter and broader than the normal distribution.

Since most stations are positively skewed for both SO_2 in kg/GWh and SO_2 in mg/Nm^3, a right-skewed distribution needs to be considered, indicating that it could be reasonable to fit a 3LL distribution to the data. Tables 4 and 5 give the parameter estimates of the 3LL distribution for both the SO_2 in kg/GWh and SO_2 in mg/Nm^3. The parameters are estimated using the ML estimation method.

Table 4: Parameter estimates of SO_2 kg/GWh using 3LL distribution.

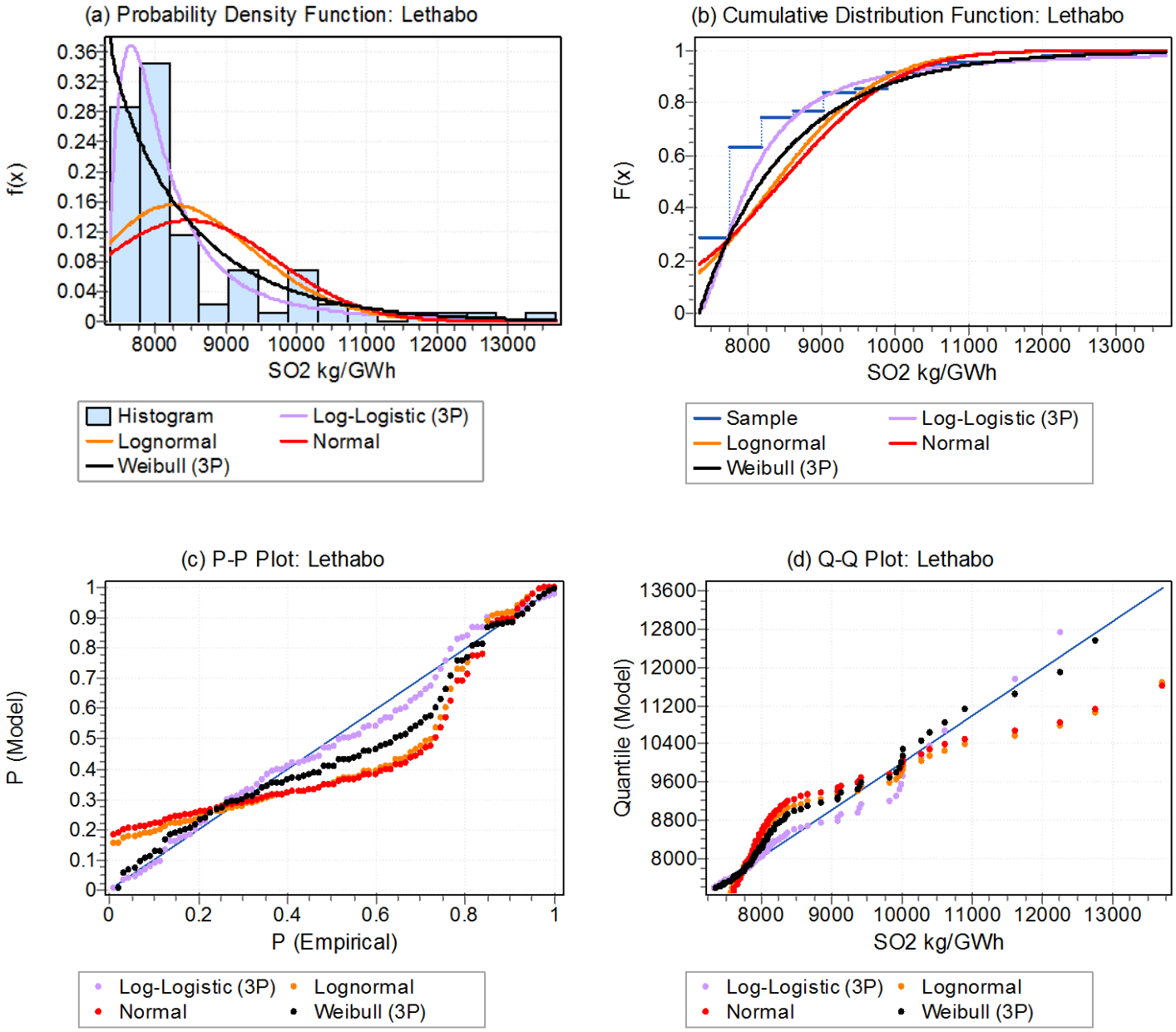
Station	*SO_2 kg/GWh*		
Arnot	$\alpha = 35.53$	$\beta = 18\,237$	$\gamma = -11\,983.2$
Camden	$\alpha = 1.47E+8$	$\beta = 1.31E+11$	$\gamma = -1.31E+11$
Duhva	$\alpha = 9.5008$	$\beta = 6\,768.2$	$\gamma = 1\,688.4$
Grootvlei	$\alpha = 5.54E+7$	$\beta = 5.44E+10$	$\gamma = -5.44E+1$
Hendrina	$\alpha = 4.0839$	$\beta = 3\,381.8$	$\gamma = 4\,873.9$
Kendal	$\alpha = 1.30E+8$	$\beta = 8.56E+10$	$\gamma = -8.56E+10$
Komati	$\alpha = 1.58E+6$	$\beta = 1.63E+9$	$\gamma = -1.63E+9$
Kriel	$\alpha = 1.47E+8$	$\beta = 1.31E+11$	$\gamma = -1.31E+11$
Lethabo	$\alpha = 1.7336$	$\beta = 710.39$	$\gamma = 7\,315.9$
Majuba	$\alpha = 7.8949$	$\beta = 5450.9$	$\gamma = 2555.3$
Matimba	$\alpha = 20.096$	$\beta = 8794.0$	$\gamma = 1972.8$
Matla	$\alpha = 36.565$	$\beta = 1\,911.0$	$\gamma = -11\,273.0$
Tutuka	$\alpha = 6.78E+7$	$\beta = 3.33E+10$	$\gamma = -3.33E+10$

Table 5: Parameter estimates of SO_2 mg/Nm^3 using 3LL distribution.

Station	*SO_2 kg/GWh*		
Arnot	$\alpha = 1.59E+8$	$\beta = 1.54E+8$	$\gamma = -1.54E+10$
Camden	$\alpha = 56.437$	$\beta = 16128$	$\gamma = 14216.0$
Duhva	$\alpha = 22.921$	$\beta = 4283.4$	$\gamma = -2006.4$
Grootvlei	$\alpha = 2.08E+8$	$\beta = 3.72E+10$	$\gamma = -3.72E+10$
Hendrina	$\alpha = 3.8$	$\beta = 592.72$	$\gamma = 1354.7$
Kendal	$\alpha = 12.997$	$\beta = 1755.2$	$\gamma = 565.36$
Komati	$\alpha = 4.0893$	$\beta = 525.86$	$\gamma = 1050.1$
Kriel	$\alpha = 5.215$	$\beta = 632.8$	$\gamma = 993.04$
Lethabo	$\alpha = 2.193$	$\beta = 317.12$	$\gamma = 1\,956.7$
Majuba	$\alpha = 6.054$	$\beta = 835.69$	$\gamma = 1166.6$
Matimba	$\alpha = 3.9E+8$	$\beta = 6.47E+10$	$\gamma = -6.47E+10$
Matla	$\alpha = 16.875$	$\beta = 2689.4$	$\gamma = -404.36$
Tutuka	$\alpha = 14.396$	$\beta = 154.2$	$\gamma = 770.18$

The normal, log-normal and three-parameter Weibull distributions were also fitted to the data for comparisons. Figure 4 shows the fit to the Lethabo station, with the comparisons for the distributions.

Probability-probability and quantile-quantile plots: The P-P and Q-Q plots are graphical methods used to test the fit of the distributions to the data. Departures from the straight line indicates departures from the theoretical distribution. Figure 5 (split over this page and the next) gives the PDF, CDF, P-P and Q-Q plots of Lethabo station for both SO_2 kg/GWh and SO_2 mg/Nm^3.



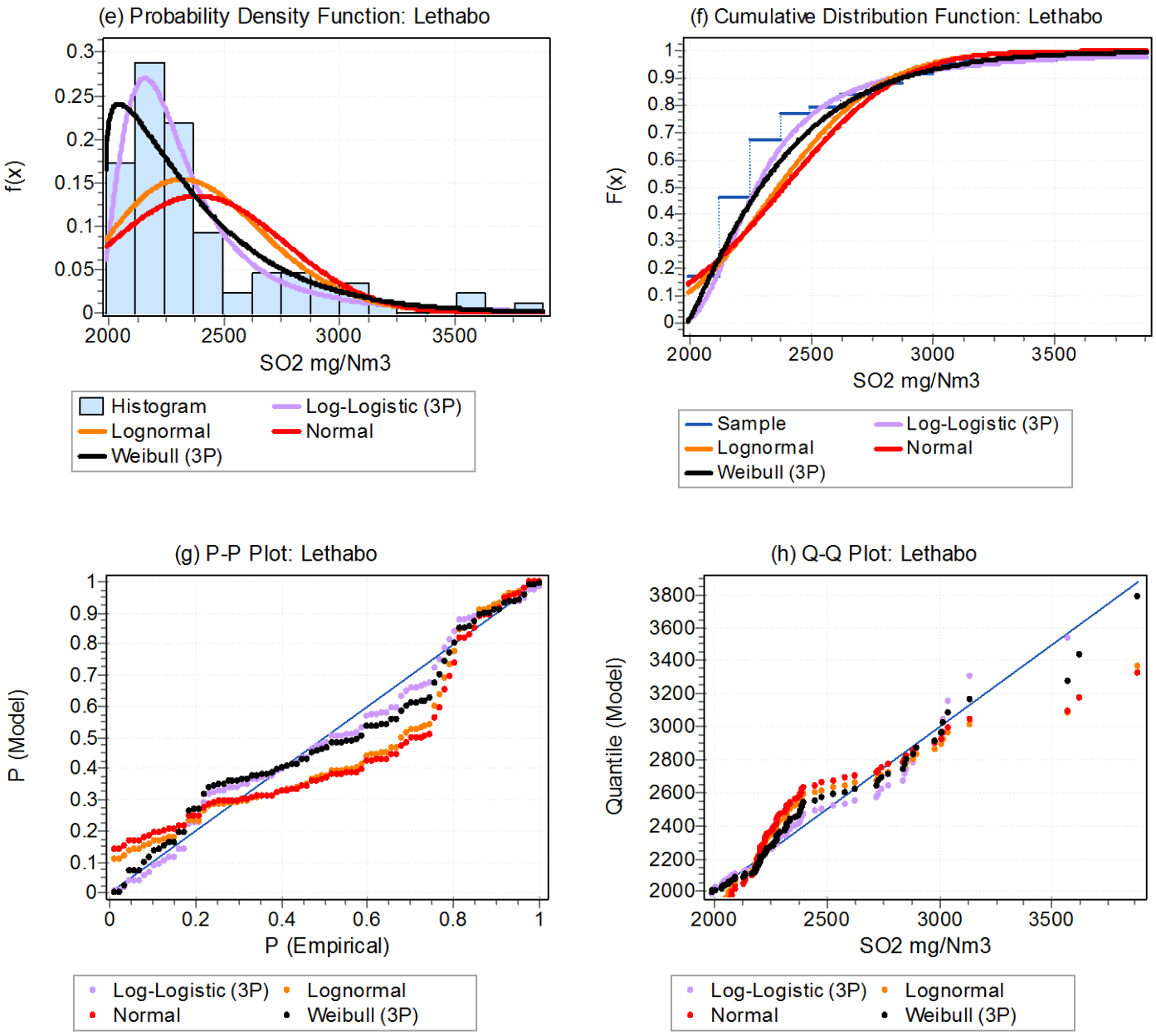


Figure 5: Distribution fit to the SO_2 kg/GWh and SO_2 Nm^3 for Lethabo station.

The PDF, CDF together with the P-P and Q-Q plots shows a better fit for the log-logistic compared to the other distributions. Looking at the Q-Q plots it can be observed that the quantiles of 3LL and the three-parameter Weibull are closer to the straight line compared to the normal and the log-normal distributions. A goodness of fit is also done on the stations. To test for the goodness of fit, the Kolmogorov-Smirnov and Anderson-Darling tests are considered, and the results for these are given here. The Kolmogorov-Smirnov critical value at 0.05 and 0.01 level of significance for Arnot, Duhva, Hendrina, Kendal, Kriel, Lethabo, Majuba, Matimba, Matla and Tutuka are 0.14355 and 0.17223 respectively, and for Camden its 0.15755 and 0.18903. For Grootvlei and Komati, since their emissions were recorded from April 2009, the critical values are 0.22119 and 0.2632 respectively. The Anderson-Darling critical values at 0.05 and 0.01 significance level are 2.5018 and 3.9074 respectively. The null and alternative hypotheses are given as:

- H_0: The data follows the three-parameter log-logistic distribution.
- H_a: The data do not follow the three-parameter log-logistic distribution.

Table 6 gives the Kolmogorov-Smirnov and Anderson-Darling test results for each station.

Table 6: Anderson-Darling and Kolmogorov-Smirnov test results (3LL).

		SO_2 kg/GWh			SO_2 mg/Nm^3		
		Statistic	*p-value*	*Reject?*	*Statistic*	*p-value*	*Reject?*
Arnot	AD	0.2861		No	0.2336		No
	KS	0.0592	0.9000	No	0.0486	0.9801	No
Camden	AD	3.1772		No	0.2885		No
	KS	0.0743	0.7938	No	0.0599	0.9449	No
Duhva	AD	0.2163		No	0.2019		No
	KS	0.0584	0.9146	No	0.0507	0.9707	No
Grootvlei	AD	0.2223		No	0.4312		No
	KS	0.0782	0.9678	No	0.1172	0.6631	No
Hendrina	AD	0.2709		No	0.6082		No
	KS	0.0679	0.7924	No	0.0840	0.5436	No
Kendal	AD	0.82056		No	0.5138		No
	KS	0.0782	0.6330	No	0.0792	0.6178	No
Komati	AD	0.2556		No	0.2169		No
	KS	0.0753	0.9773	No	0.0827	0.9494	No
Kriel	AD	0.3206		No	0.6006		No
	KS	0.0637	0.8492	No	0.0725	0.7228	No
Lethabo	AD	0.4868		No	0.7281		No
	KS	0.0641	0.8443	No	0.0962	0.3733	No
Majuba	AD	0.4052		No	0.2672		No
	KS	0.0777	0.6413	No	0.0535	0.9751	No
Matimba	AD	0.2354		No	1.0090		No
	KS	0.0458	0.9894	No	0.0901	0.4543	No
Matla	AD	0.7477		No	0.2802		No
	KS	0.0724	0.7245	No	0.0552	0.9401	No
Tutuka	AD	0.3725		No	0.2674		No
	KS	0.0616	0.8754	No	0.0596	0.8983	No

Both the Anderson-Darling and Kolmogorov-Smirnov tests do not reject the null hypothesis for any of the stations for both SO_2 kg/GWh and SO_2 mg/Nm^3 at 5% and 1% significance level. Supplementary information gives the QQ plots for all the other stations. Tables 7 and 8 give the probabilities of exceedances above a given threshold, where t is the threshold and $P(X > t)$ is the probability of the exceedances.

Tables 7 and 8 show that Grootvlei, Kendal and Komati have a probability of almost zero for exceeding 5000 kg/GWh SO_2 emission level per month, with Arnot, Grootvlei Majuba having a probability of almost zero for exceeding 1000 mg/Nm^3 SO_2 emmission level per month. Matimba, Lethabo and Matla have a probability of 1 for exceeding 7000 kg/GWh SO_2 emmission level per month with Matla, Tutuka and Duhva have a probability of 1 for exceeding 2000 mg/Nm^3 SO_2 emmissions per month. This confirms that the least efficient stations with regard to emission of SO_2 are Matimba, Lethabo and Matla.

Conclusions

Monthly SO_2 emissions in kg/GWh and in mg/Nm^3 have been considered for Eskom's 13 coal-fired power-generating stations. The 3LL fits the data of these stations best, and makes it is possible to quantify (in terms of a statistical distribution). This quantification helps to monitor and manage the SO_2 emissions effectively.

The parameters of the log-logistic distribution are estimated by the ML method. Kolmogorov-Smirnov and Anderson-Darling tests are used to test for the

goodness of fit of the 3LL distribution for the 13 stations. The PDF, CDF, P-P and Q-Q plots are used to show how well the distribution fits the data.

Considering Figures 3 and 4 together with Table 4, it can be concluded that the stations that use electrostatic precipitators and flue gas conditioning technology are less efficient than those using fabric filter plants technology, although other factors, such as the age of the plant and the quality of coal used affect emission efficiency.

Table 7: The SO_2 in kg/GWh probabilities of exceedances for each station (3LL).

		$t = 5000$	$t = 6000$	$t = 7000$	$t = 8000$	$t = 10000$
Arnot	$P(X > t)$	0.9396	0.6319	0.1759	0.0287	0.0007
Camden	$P(X > t)$	0.0059	0.0021	0.0008	0.0003	0.00004
Duhva	$P(X > t)$	0.9999	0.9864	0.9091	0.6601	0.1244
Grootvlei	$P(X > t)$	0.0061	0.0022	0.0008	0.0003	≈ 0
Hendrina	$P(X > t)$	1.0000	0.9889	0.8694	0.5796	0.1546
Kendal	$P(X > t)$	0.0005	0.0001	0.0000	≈ 0	≈ 0
Komati	$P(X > t)$	0.0078	0.0030	0.0011	0.0004	≈ 0
Kriel	$P(X > t)$	0.9942	0.7477	0.2144	0.0478	0.0046
Lethabo	$P(X > t)$	≈ 1	≈ 1	≈ 1	0.5163	0.0908
Majuba	$P(X > t)$	0.9983	0.9740	0.8336	0.5023	0.0786
Matimba	$P(X > t)$	1.0000	1.0000	1.0000	0.9995	0.8622
Matla	$P(X > t)$	0.9994	0.9945	0.9585	0.7669	0.0817
Tutuka	$P(X > t)$	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0

Table 8: SO_2 in mg/Nm^3 probabilities of exceedances for each station (3LL).

		$t = 1000$	$t = 1500$	$t = 2000$	$t = 2500$	$t = 3000$
Arnot	$P(X > t)$	0.0000	0.0000	0.0000	0.0000	0.0000
Camden	$P(X > t)$	0.2459	0.1570	0.0961	0.0753	0.0335
Duhva	$P(X > t)$	1.0000	1.0000	1.0000	≈ 1	0.0273
Grootvlei	$P(X > t)$	0.0037	0.0003	≈ 0	≈ 0	≈ 0
Hendrina	$P(X > t)$	≈ 1	0.9952	0.4200	0.0756	0.0202
Kendal	$P(X > t)$	1.0000	≈ 1	0.9322	0.2201	0.0140
Komati	$P(X > t)$	1.0000	0.6543	0.0818	0.0156	0.0047
Kriel	$P(X > t)$	0.0000	0.0000	0.0000	0.0000	0.0000
Lethabo	$P(X > t)$	≈ 1	1.0000	0.9875	0.2349	0.0684
Majuba	$P(X > t)$	0.0000	0.0000	0.0000	0.0000	0.0085
Matimba	$P(X > t)$	0.0024	0.0001	≈ 0	≈ 0	≈ 0
Matla	$P(X > t)$	1.0000	0.9971	0.8688	0.2146	0.0184
Tutuka	$P(X > t)$	1.0000	≈ 1	0.9633	0.1620	0.0050

Tables 7 and 8 show Arnot, Grootvlei and Kriel as the most efficient stations and Matimba, Lethabo and Matla as the least efficient. Looking at the results of the goodness fit, at both 5% and 1%, the null hypothesis for all stations for both the SO_2 in kg/GWh and in mg/Nm^3 cannot be rejected and it is therefore concluded that the data follows the 3LL distribution. The calculated probabilities can be used to estimate costs of exceeding the given limits.

The goodness of fit tests considered show that the 3LL fits the data of the Duhva, Hendrina, Kendal, Komati, Kriel, Lethabo and Matimba better than other stations. For SO_2 in mg/Nm^3, the three-parameter log-logistic fits the data of Camden, Duhva, Hendrina, Komati, Kriel, and Lethabo best. From the results it shows that three-parameter log-logistic fits the positively skewed data better than the nega-

tively skewed data. For the negatively skewed stations, it is only the three-parameter Weibull distribution that does better than the 3LL distribution. The three-parameter Weibull distribution is very close to the 3LL distribution in terms of goodness of fit. In emissions monitoring, however, concerns are more with high emissions (positively skewed), which give rise to undesirable consequences.

Reporting on environmental performance has several benefits, including providing management information to help exploit the cost savings that good environmental performance usually brings and, giving Eskom the opportunity to set out what they believe is significant in their environmental performance.

Further research will be done on the other Burr-type distributions to see if they will fit the data of most stations similar or better than the 3LL. The impact of the age of the plant and the quality of the coal used on atmospheric emission efficiency also requires research.

Acknowledgement

The authors are grateful to Eskom for providing the necessary data that made this research possible. Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to Eskom.

References

Abbas, K.; Alamgir.; Khan, S.A.; Ali, A.; Khan, D.M.; Khalil, U. 2012. Statistical analysis of wind speed data in Pakistan. *World Applied Sciences Journal* 18(11): 1533–1539.

Ashkar, F. and Mahdi, S. 2003. Comparison of two fitting methods for log–logistic. *Water Resources Research* 39(8): SWC 7.1–SWC 7.8

Balakrishnan, N. and Malik, H.J. 1987. Best linear unbiased estimation of the location and scale parameters of the log–logistic distribution. *Communication in statistics A: Theory and Methods* 16: 3477–95.

Beirlant, J., Goegebeur, Y., Teugels J., De Waal, D. and Ferro C. 2004. *Statistics of extremes.* West Sussex: John Willey & Sons.

Brown, S. 2011. Measures of shape: Skewness and kurtosis. Available online at: http://web.ipac.caltec.edu/staff/fmasci/home/statistics_refs/SkewStatSignif.pdf.

Burr, I. 1942. Cumulative frequency functions. *Annals of Mathematical Statistics*. 13(2): 215–232.

Eskom. 2011. Eskom emission monitoring. Available online at: http://www.wonderware.co.za/content/Eskom%20Emission%20Monitoring.pdf.

Eskom. 2012. Kusile and Medupi coal–fired power stations under construction. Available online at: http://www.eskom.co.za/OurCompany/SustainableDevelopment/ClimateChangeCOP17/Documents/Kusile_and_Medupi_coal–fired_power_stations_under_construction.pdf.

Eskom. 2012. Climate change COP 17 fact sheet. Available online at: http://www.eskom.co.za/OurCompany/SustainableDevelopment/ClimateChangeCOP17/Documents/Air_quality_and_climate_change.pdf.

Evans, D.L., Drew, J.H. and Leemis, L.M. 2008. The distribution of the Kolmogorov–Smirnov, Cramer–von Mises and Anderson–Darling tests statistics for exceptional populations with estimated parameters. *Communication in Statistics–Simulation and Computation* 37: 1396–1421.

Georgopoulos, G.P. and Seinfeld H.J. 1982. Statistical distributions of air pollutant concentrations. *Environmental Science and Technology* 16(7): 401A–416A.

Hadley, A. and Toumi, R. 2003. Assessing changes to the probability distribution of sulphur dioxide in the UK using lognormal model. *Atmospheric Environment* 37: 1461–1474.

Mielke, P.W., & Johnson, E.S. 1973. Three parameter Kappa distribution maximum likelihood estimates and likelihood ratio tests. *Monthly Weather Review* 101, 701–709.

Mitchell, B. 1971. A comparison of Chi Square and Kolmogorov–Smirnov tests. *Area* 3(4): 237–241. Available online at: https://www.jstor.org/stable/20000590?seq=1#fndtn–page_scan_tab_contents

Rumburg, B., Alldredge, R. and Claiborn, C. 2001. Statistical distributions of particulate matter and error associated with sampling frequency. *Atmospheric Environment* 35: 2907–20.

Seifeld, J.H. and Pandis, S.N., 1998. *Atmospheric chemistry and physics: From air pollution to climate change.* New York: Wiley.

Singh, V.P., Gou, H. and Yu, F.X., 1993. Parameter estimation for 3–parameter log–logistic distribution (LLD3) by Pome. *Stochastic Hydrology and Hydraulics* 7: 163–177.

Smith, L.R. 1989. Extreme value analysis of environmental time series: An application to trend detection in ground–level ozone. *Statistical Science* 44: 367–393.

Tiku, M.L. and Suresh, R.P. 1992. A new method of estimation for location and scale parameters. *Journal of Statistical Planning and Inference* 30: 281–292.

Wingo, D.R. (Metrica). 1993. Maximum likelihood methods for fitting the Burr Type XII distribution to multiply (progressively) censored Life Data. 40: 203–210. doi:10.1007/BF02613681.

Zaharim, A., Najid, S.K., Razali, A.M. and Sopian, K. 2009. Analysing Malaysian wind speed data using statistical distribution. Proceedings of the 4th IASME/WSEAS International Conference on Energy and Environment, Cambridge, UK.