MODELLING AIRQUALITY OF ILORIN, KWARA STATE WITH FAMILY OF XLINDLEY DISTRIBUTION

Toyin Elijah Faniyi and Bello Ishola Sanni

Department of Mathematics and Statistics, Kwara State University, Malete, Nigeria

elijahfaniyi05@gmail.com

Abstract

Ozone, a major component of smog, forms through chemical reactions involving pollutants like volatile organic compounds and nitrogen oxides, primarily from vehicle emissions and industrial activities. As a key air quality indicator, high ozone levels pose environmental and health risks, causing respiratory issues such as asthma and bronchitis while increasing cardiovascular disease risk. This study proposes the Generalized XLindley Distribution to model ozone levels in Ilorin, Kwara State, and compares it with the Exponential-Lindley, Quasi XLindley, and Inversed XLindley distributions. The Kolmogorov-Smirnov test was used to assess goodness-of-fit, while parameter estimation was performed using the Maximum Likelihood Estimator and the Method of Moments. A simulation study explored model behavior across varying parameters, with survival and hazard functions analyzed for deeper insights. Model selection criteria, including AIC, AICC, BIC, and HQIC, were applied to evaluate efficiency. The Generalized Xlindley Distribution outperformed competing models, providing a more accurate fit. These findings enhance statistical modeling of air pollution data, improving air quality assessment and prediction in urban settings.

Keywords: Weighted Distribution, Xlindley distribution, Maximum Likelihood Estimation, Ozone, Airquality

INTRODUCTION

Air quality refers to the state of the air in our environment, particularly in relation to the presence of pollutants that can affect human health, ecosystems, and climate. Clean air is essential for maintaining public health and environmental balance, while poor air quality is linked to respiratory diseases, environmental degradation, and climate change (WHO, 2021). Airquality is influenced by Meteorological Conditions (Seinfeld & Pandis, 2016), Volcanic Eruptions (Robock, 2000), Wildfires (Jaffe et al., 2020), Industrial Emissions (U.S. EPA, 2022), Vehicle Exhaust (Kim et al., 2018), Burning Fossil Fuels (IPCC, 2021). Airquality significantly impacts environmental health, especially in urban areas with increased human activities. Ozone (O_3), a secondary pollutant, forms through photochemical reactions involving volatile organic compounds (VOCs) and nitrogen oxides (NO_x) in sunlight (Seinfeld & Pandis, 2016). Ground-level ozone harms respiratory health, agriculture, and ecosystems (WHO, 2021).

Air quality is commonly measured using the Air Quality Index (AQI), which categorizes air conditions based on pollutant concentrations. Common Air Pollutants are Particulate Matter (PM2.5 & PM10), Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Dioxide (NO_2) and Ozone (O_3)

Ozone (O_3) is a significant air pollutant and is often considered the second most concerning air pollutant after particulate matter (PM2.5) due to its widespread health and environmental effects (WHO, 2021). Unlike primary pollutants, which are directly emitted into the atmosphere, ozone is a secondary pollutant that forms through chemical reactions in the presence of sunlight (U.S. EPA, 2022).

Ozone, a key component of smog, forms when pollutants from vehicles, industries, and other sources react under sunlight (EPA. 2020). Ground-level ozone contributes to air pollution and public health issues, including respiratory problems and cardiovascular diseases (WHO, 2021).

Anthropogenic sources like transportation, industrial processes, and agriculture significantly drive ozone formation (Seinfeld & Pandis, 2016), along with natural sources such as wildfires (Monks, et al. 2015). Chronic exposure to high ozone levels damages crops, forests, and materials, impacting the environment (Mills, et al. 2011). Vulnerable groups, such as children and the elderly, are particularly at risk (Bell, et al. 2006). Monitoring ozone levels is done through ground-based sensors, satellite data, and atmospheric models (Fishman, Bowman, & Brasseur, 2010). This study analyzed ozone levels in Ilorin, Nigeria, using statistical models derived from the Xlindley distribution family. The airquality data, which focused on ozone levels over a four-month period, were obtained from OpenWeatherMap (OpenWeatherMap, 2024).

The field of distribution theory has advanced with new families of continuous distributions, enhancing the flexibility of traditional probability models. The Lindley Distribution, introduced by Lindley (1958), offers a monotonically decreasing hazard rate, which is an improvement over the exponential distribution (Shanker, 2015). The Xlindley distribution (Chouia, & Zeghdoudi, 2021). Quasi Xlindley distribution (Ibrahim, Shah & Haq, 2023), and Inverse Xlindley distribution (Meriem, et al. 2022). have since been developed. The Generalized Xlindley Distribution (GXLD) extends the Quasi Xlindley model by adding a weighting mechanism, making it more adaptable for various statistical analyses (Merovci, et al. 2014; Beghriche, et al. 2023). This extension broadens the applicability of the model, offering better flexibility for data modeling in different disciplines.

The pdf of Lindley distribution of random variable X, with scale parameter γ is given by:

$$f(x) = \frac{\theta^2}{\theta+1}(1+x)e^{-\theta x}, \quad x > 0, \theta > 0 \tag{1}$$

The idea of this work is to extend Quasi Xlindley Distribution called a Generalized Xlindley Distribution (GXLD) with the hope that it will attract many applications in different disciplines.

On applying the weighted version, the third parameter indexed, to this distribution, it is expected to be more flexible to describe different lifetime data than its sub-models.

In this paper, a New Generalized Quasi Xlindley distribution which includes lindley distribution, exponential - lindley distribution and Quasi Xlindley distribution as particular cases, has been proposed and discussed. The GXLD is developed by incorporating weighting mechanisms into the Quasi Xlindley model, thereby broadening its applicability. Weighted distributions are utilized to modulate the probabilities of the events as observed and transcribed. (Patil, et al, 1987; Saghir, et al., 2017).]. The hazard rate function and Survival function of the distribution have been derived and their shapes have been discussed for varying values of the parameters. The estimation of its parameters has been discussed using maximum likelihood method. Finally, the goodness of fit and the applications of the distribution have been explained through four lifetime data and the fit has been compared with the Quasi XLindley distribution, Power Xlindley distribution and Inverse Xlindley distribution.

METHODS

Generalized XLindley Distribution

In this section, Generalized Xlindley distribution is introduced through the incorporation weighted mechanism into the Quasi Xlindley distribution.

Chouia and Zeghdoudi (2012) proposed xlindley distribution as the mixture of exponential and Lindley distribution with proportion given as $p_1 = \frac{\theta}{1+\theta}$ and $p_2 = 1 - \frac{\theta}{1+\theta}$ respectively. The pdf was obtained as:

$$f(x) = \frac{\theta^2(2+\theta+x)e^{-\theta x}}{(1+\theta)^2} \quad x > 0, \theta > 0 \tag{2}$$

Ibrahim, et al (2023) introduced a new generalization of xlindley distribution by incorporating a new parameter into the xlindley distribution. The distribution is named Quasi-XLindley (QXL) distribution with pdf:

$$f(x) = \frac{\theta}{1+\alpha}\left(\alpha + \frac{\theta(1+x)}{1+\theta}\right)e^{-\theta x}, \ \alpha, \theta > 0; \ x > 0 \qquad (3)$$

The aim of this work is to proposed an extension of QXL distribution through introduction of weighted parameter. The weight was quoted by Rather and Ozel (2020) as

$$f_w(x) = \frac{x^c f(x)}{E(x^c)} \qquad (3)$$

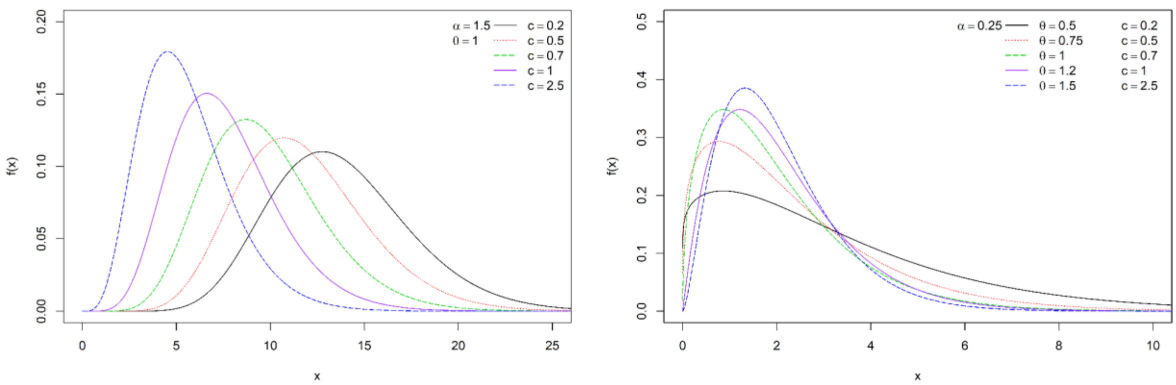
where c > 0 is the weight parameter and the expected value is defined as:

$$E(x^c) = \int_0^\infty x^c \, f(x)dx \qquad (4)$$

Therefore, the Generalized Xlindley distribution is obtained over scale, shape and weight parameter as:

$$f(x;\alpha,\theta,c) = \frac{\theta^{(1+c)}}{1+\alpha+c}\,\frac{x^c(\alpha(1+\theta)+\theta(1+x))}{\Gamma(1+c)}e^{-\theta x}; \qquad x > 0, \alpha, \theta > 0, c > 0 \qquad (5)$$

The figure below is plot of probability density function of GXLD at various parameter values



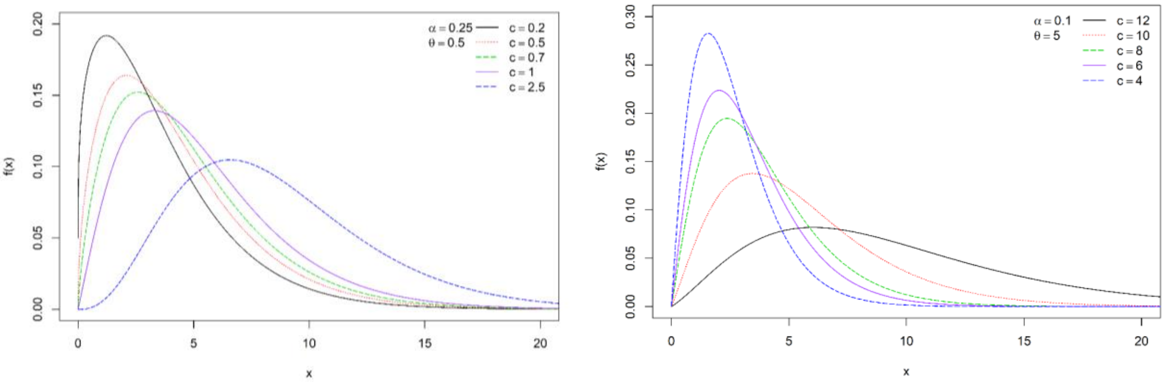


Figure 1: The probability density function of GXLD

The cumulative distribution is obtained as:

$$F(x;\ \alpha,\theta,c) = \frac{\left(x^{c}\theta^{c}(x\theta)^{-c}\left((1+\alpha+c+\theta+\alpha\theta)\Gamma(1+c)-(\alpha+\theta+\alpha\theta)\Gamma(1+c,x\theta)-\Gamma(2+c),x\theta)\right)}{((1+\alpha+c+\theta+\alpha\theta)\Gamma(1+c)} \tag{6}$$

The figure below is plot of cumulative density function of GXLD at various parameter values

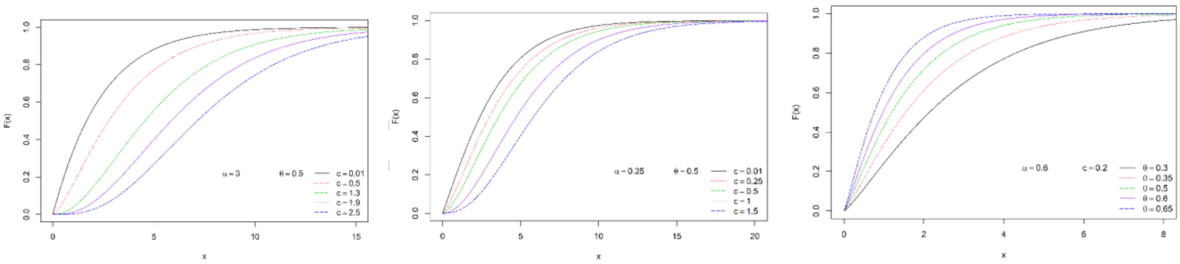


Figure 2: The cumulative density function of GXL distribution

Moments

$$E(x^{r}) = \mu_r' = \frac{\theta^{-r}(1+r+c+\alpha+\theta+\alpha\theta)\Gamma(1+r+c)}{(1+\alpha+c+\theta+\alpha\theta)\Gamma(1+c)} \tag{7}$$

First moment:

$$E(x) = \mu_1' = \frac{(2+c+\alpha+\theta+\alpha\theta)\Gamma(2+c)}{\theta(1+c+\alpha+\theta+\alpha\theta)\Gamma(1+c)} \tag{8}$$

Second moment:

$$E(x^2) = \mu_2' = \frac{(3+\alpha+c+\theta+\alpha\theta)\Gamma(3+c)}{\theta^2(1+\alpha+c+\theta+\alpha\theta)\Gamma(1+c)} \tag{9}$$

Third Moment:

$$E(x^3) = \mu_3' = \frac{(4+c+\alpha+\theta+\alpha\theta)\Gamma(4+c)}{\theta^3(1+c+\alpha+\theta+\alpha\theta)\Gamma(1+c)} \tag{10}$$

Fourth moment:

$$E(x^4) = \mu_4' = \frac{(5+c+\alpha+\theta+\alpha\theta)\Gamma(5+c)}{\theta^4(1+c+\alpha+\theta+\alpha\theta)\Gamma(1+c)} \tag{11}$$

Statistical summaries are obtained as follows:

$$\text{Variance } (\sigma^2) = \frac{(1+c)^2\left((3+c+\alpha+\theta+\alpha\theta)(2+c)(1+\alpha+c+\theta+\alpha\theta) - (2+c+\alpha+\theta+\alpha\theta)^2\right)}{\theta^2\left(\Gamma(1+c)\right)^2} \tag{12}$$

Coefficient of variation:

$$cv = \frac{\sqrt{2+c^2+4\theta+\theta^2+\alpha^2(1+\theta)^2+2\alpha(1+\theta)(2+c+\theta)+c(3+2\theta)}}{(1+c)(2+c+\alpha+\theta+\alpha\theta)} \tag{13}$$

Skewness:

$$sk = \frac{(4+c+\alpha+\theta+\alpha\theta)\Gamma(4+c)}{\theta^3(1+c+\alpha+\theta+\alpha\theta)\left[\frac{(1+c)(3+2\theta)\left(2+c^2+4\theta+\theta^2+\alpha^2(1+\theta)^2+2\alpha(1+\theta)(2+c+\theta)+c\right)}{\theta^2(1+c+\alpha+\theta+\alpha\theta)^2}\right]^{\frac{3}{2}}\Gamma(1+c)} \tag{14}$$

Kurtosis is obtained as:

$$ks = \frac{(1+c+\alpha+\theta+\alpha\theta)^3(5+c+\alpha+\theta+\alpha\theta)\Gamma(5+\beta)}{(1+c)^2\left(2+c+4\theta+\theta^2+\alpha^2(1+\theta)^2+2\alpha(1+\theta)\right)(2+c+\theta)+c(3+2\theta)^2\Gamma(1+c)} \tag{15}$$

Index of Dispersion is obtained as:

$$ID = \frac{\left(2+c^2+4\theta+\theta^2+\alpha^2(1+\theta)(2+c+\theta)+c(3+2\theta)\right)}{\theta(1+c+\alpha+\theta+\alpha\theta)(2+c+\alpha+\theta+\alpha\theta)} \qquad (16)$$

Harmonic mean is obtained as:

$$HM = \left(\frac{c}{1+\alpha+\theta}\left[\frac{c+\alpha}{\Gamma(1+\theta)} + \frac{\theta c(1+\alpha)}{\Gamma(1+\theta)}\right]\right)^{-1} \qquad (17)$$

Table 1a: Statistical Properties of GXLD for α = 0.3 and 0.8

α	θ	c	μ	Var	Sk	Ks	ID	HM
0.3	0.3	0.3	7.1859	6.7935	6.0160	11.12028	2.9610	0.7538
		0.8	7.0817	39.8099	0.4159	0.39326	2.9863	0.3012
		1.2	7.0242	69.7530	0.1915	0.14933	3.0072	0.2076
		2.0	6.9467	139.6385	0.1033	0.05457	3.0453	0.1301
	0.8	0.3	2.3390	2.4924	1.4282	1.94411	1.0743	2.5253
		0.8	2.3646	6.9482	0.3476	0.37259	1.0824	0.9554
		1.2	2.3799	11.1155	0.2088	0.19547	1.0895	0.6403
		2.0	2.4021	21.0158	0.1589	0.09615	1.1036	0.3871
	1.2	0.3	1.4399	1.4093	0.9954	1.29031	0.6991	4.1772
		0.8	1.4709	3.3830	0.3255	0.35933	0.7047	1.5574
		1.2	1.4901	5.2453	0.2167	0.21014	0.7095	1.0345
		2.0	1.5192	9.6948	0.1862	0.11446	0.7191	0.6173
	2.0	0.3	0.7738	0.6293	0.7209	0.90760	0.4032	7.7778
		0.8	0.7979	1.3472	0.3055	0.34356	0.4069	2.8723
		1.2	0.8137	2.0296	0.2266	0.22305	0.4100	1.8954
		2.0	0.8390	3.6690	0.2197	0.13480	0.4158	1.1186
0.8	0.3	0.3	6.2374	17.7240	1.4282	1.94411	2.8649	0.9470
		0.8	6.3057	49.4097	0.3476	0.37259	2.8864	0.3583
		1.2	6.3465	79.0438	0.2088	0.19547	2.9054	0.2401
		2.0	6.4055	149.4454	0.1589	0.09615	2.9429	0.1452
	0.8	0.3	2.0621	3.0061	1.2476	2.09181	1.1204	2.6365
		0.8	2.1163	7.2167	0.4492	0.62610	1.1211	0.9901
		1.2	2.1509	11.2295	0.3185	0.37818	1.1231	0.6607
		2.0	2.2042	20.9049	0.2933	0.21148	1.1288	0.3969
	1.2	0.3	1.2833	1.4240	1.2113	2.19277	0.7583	4.0376
		0.8	1.3235	3.2483	0.5038	0.76145	0.7577	1.5126
		1.2	1.3501	4.9989	0.3789	0.48262	0.7581	1.0078
		2.0	1.3926	9.2463	0.3664	0.28212	0.7602	0.6040
	2.0	0.3	0.7018	0.5435	1.1984	2.35796	0.4606	6.9006
		0.8	0.7258	1.1803	0.5785	0.94772	0.4602	2.5806
		1.2	0.7424	1.7955	0.4619	0.63177	0.4602	1.7172
		2.0	0.7703	3.2986	0.4655	0.38800	0.4607	1.0270

Table 1b: Statistical Properties of GXLD for α = 1.2

α	θ	c	μ	Var	Sk	Ks	ID	HM
1.2	0.3	0.3	5.7595	22.5490	0.9954	1.29031	2.7962	1.0443
		0.8	5.8834	54.1285	0.3255	0.35933	2.8188	0.3893
		1.2	5.9606	83.9240	0.2167	0.21014	2.8382	0.2586
		2.0	6.0768	155.1162	0.1862	0.11446	2.8763	0.1543
	0.8	0.3	1.9249	3.2039	1.2113	2.19277	1.1374	2.6917
		0.8	1.9853	7.3086	0.5038	0.76145	1.1366	1.0084
		1.2	2.0252	11.2474	0.3789	0.48262	1.1372	0.6718
		2.0	2.0889	20.8042	0.3664	0.28212	1.1402	0.4027
	1.2	0.3	1.2062	1.4132	1.3422	2.75824	0.7876	3.9689
		0.8	1.2470	3.1612	0.6111	1.03231	0.7851	1.4894
		1.2	1.2748	4.8501	0.4774	0.67128	0.7840	0.9934
		2.0	1.3207	8.9771	0.4708	0.40132	0.7831	0.5965
	2.0	0.3	0.6667	0.4981	1.5367	3.64186	0.4941	6.4734
		0.8	0.6892	1.0915	0.7698	1.47332	0.4919	2.4324
		1.2	0.7051	1.6689	0.6240	0.98505	0.4905	1.6239
		2.0	0.7326	3.0913	0.6214	0.60471	0.4885	0.9767

From table 1, the follows can be deduced

- as α, θ, and c increase (i.e. tends to ∞), the μ becomes smaller, variance, skewness, and kurtosis tends to increase, showing a wider, more extreme distribution with lower central values.
- Also, as α, θ, and c increase, skewness values are mostly positive but decrease, suggesting a shift towards a more symmetric distribution with larger parameters.
- The kurtosis also tends to increase with increasing α, θ, and c, indicating that higher values of these parameters are associated with more extreme values (heavier tails).
- Both ID and HM tend to decrease as α, θ, and c increase, indicating diminishing returns or a saturation effect with larger values of these parameters.
- *c* modifies the central tendency and tail behavior of the distribution

In general, α determines the overall scale of the distribution, with higher values compressing it and reducing both the mean and skewness. The parameter θ influences the dispersion.

Survival Function

The survival function is a function that gives the probability that a patient, device, or other object of interest will survive past certain time (Jiang, & Guterman, 2024). Let the lifetime *x* be a continuous random variable with cumulative density function *F*(*x*) and probability density function *f*(*x*) on the interval [0, ∞]. survival function of GXLD is given as:

$$s(x;\alpha,\theta,c) = 1 - F(x;\alpha;\theta,c) \tag{18}$$

$$s(x;\alpha,\theta,c) = 1 + \frac{x^c\theta^c(x\theta)\big(-(1+c+\alpha+\theta+\alpha\theta)\Gamma(1+c)+(\alpha+\theta+\alpha\theta)\Gamma(1+c,x\theta)+(2+c,x\theta)\big)}{c(1+c+\alpha+\theta+\alpha\theta)\Gamma(1+c)} \tag{19}$$

Figure 3 present the shapes of the survival function of the proposed GXLD at selected values of α, θ and *c*.

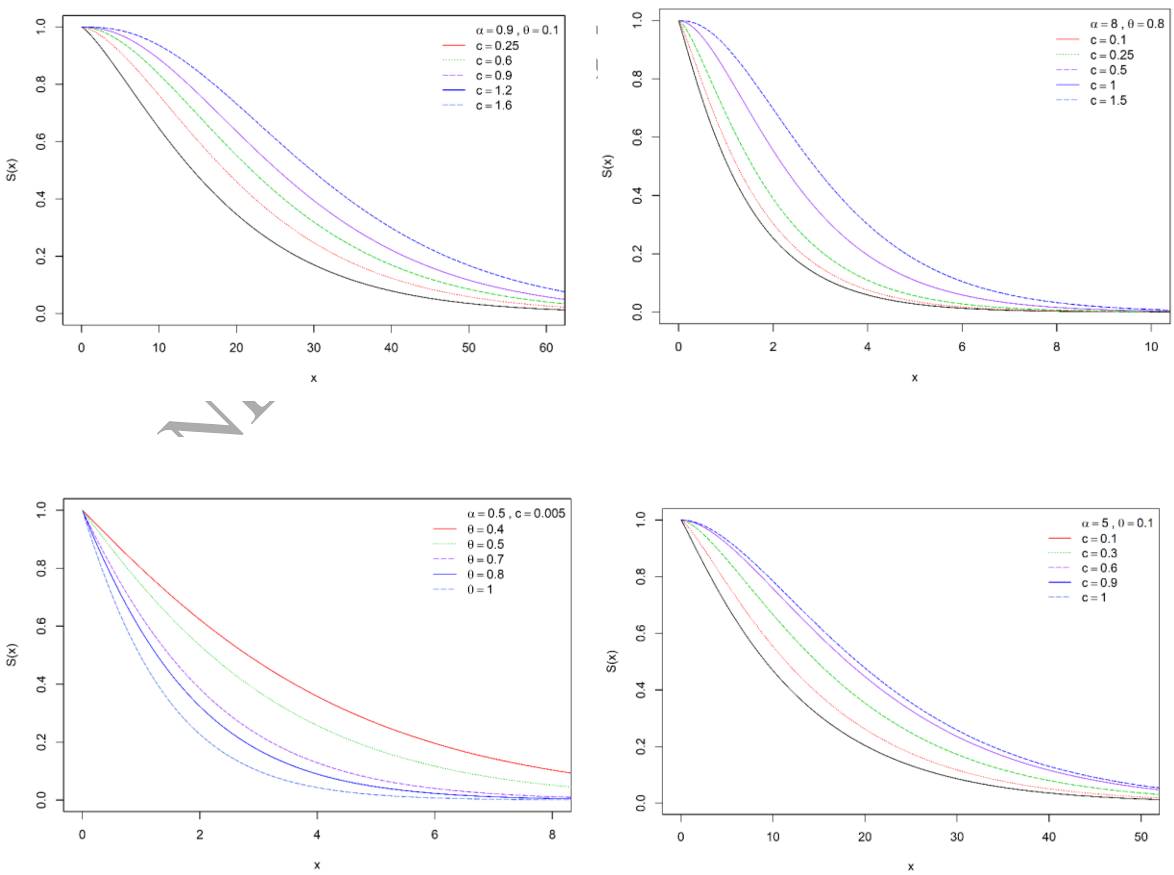


Figure 3: The survival function of GXLD for various choices of parameter

Hazard Function

The hazard function *h*(*x*) of an event is the probability of the failure of the event at time *x*. It is the probability that an event will fail at a given time *x*. The hazard function *h*(*x*) is given as:

$$h(x) = \frac{f(x)}{1-F(x)} = \frac{f(x)}{s(x)} = \frac{f(x;\alpha,\theta,c)}{s(x;\alpha,\theta,c)} \tag{20}$$

the hazard function of GXLD is given as

$$h(x;\alpha,\theta,c) = \frac{\left(\alpha+\frac{(1+x)}{1+\theta}\right)\theta e^{-x\theta}}{(1+\alpha)\left(1+\frac{x^c\theta^c(x\theta)(1+c+\alpha+\theta+\alpha\theta)\Gamma(1+c)+(\alpha+\theta+\alpha\theta)\Gamma(2+c,x\theta)}{c(1+c+\alpha+\theta+\alpha\theta)\Gamma(1+c)}\right)} \tag{21}$$

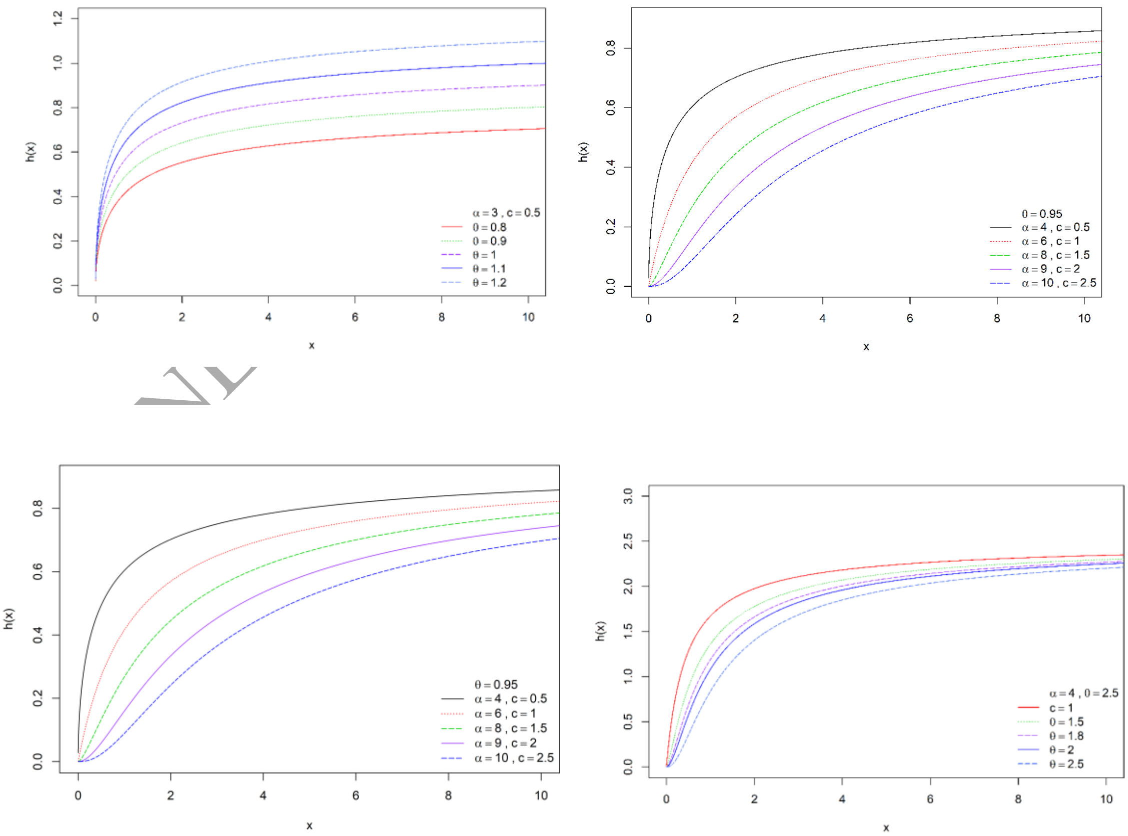


Figure 4: The hazard function of GXLD for various choices of parameter

Figure 4 presented the shapes of the hazard function of the proposed GXLD at selected values of α, θ and c. The plots show an increasing pattern which suggests that items are more likely to fail at the beginning after which they maintain a constant failure rate. This means that the failure may likely happen during the useful life of the item and failure occurs at random.

Maximum Likelihood Estimation (MLEs)

The Maximum Likelihood Estimate (MLE) is a widely used method for estimating the parameters of an assumed probability distribution. This is because of MLE estimators have desirable properties such as consistency, asymptotic efficiency, and invariance. To obtain the maximum likelihood estimators of the parameters of the GXLD, let $x_1, x_2, ..., x_n$ be a random sample of size n from the GXLD with the log-likelihood function is defined as

$$L(\alpha,\theta,c;x) = \prod_{i=1}^{n} f(x;\alpha,\theta,c) \quad (22)$$

To find the MLE $\hat{\alpha},\hat{\theta},\hat{c}$ take the partial derivatives of the log-likelihood function with respect to α,θ and c set them to zero as: to differentiate with respect to α

$$\frac{\partial L(\alpha,\theta,c;x)}{\partial\alpha} = \frac{n}{1+c+\alpha}\sum_{i=1}^{n}\frac{1}{\alpha+\theta+x_i\theta+\alpha\theta} \quad (23)$$

$$\frac{\partial L(\alpha,\theta,c;x)}{\partial\theta} = \sum_{i=1}^{n}\left(\frac{1+c}{\theta}+\frac{1+x_i+\alpha}{\alpha+\theta+x_i\theta+\alpha\theta}-n\right) \quad (24)$$

$$\frac{\partial L(\alpha,\theta,c;x)}{\partial c} = \frac{\theta^{1+c}x^{c}\left(\alpha(1+\theta)+\theta(1+x)\right)e^{-\theta x}}{(1+\alpha+c)\Gamma(1+c)}\left[ln(\theta)-\frac{1}{(1+\alpha+c)}+\ln(x)-\psi(1+c)\right] \quad (25)$$

These three natural log-likelihood do not seem to be solved directly. However, the Fisher's scoring method was used to solve these equations and an Hessian Matrix of the log-likelihood function *lnL*, which consists of second-order partial derivatives of *lnL* with respect to the parameters α, θ and c was obtained

$$I(\alpha,\theta,c) = \begin{bmatrix} \frac{\partial^2 InL}{\partial\alpha^2} & \frac{\partial^2 InL}{\partial\alpha\partial\theta} & \frac{\partial^2 InL}{\partial\alpha\partial c} \\ \frac{\partial^2 InL}{\partial\theta\partial\alpha} & \frac{\partial^2 InL}{\partial\theta^2} & \frac{\partial^2 InL}{\partial\theta\partial c} \\ \frac{\partial^2 InL}{\partial c\partial\theta} & \frac{\partial^2 InL}{\partial c\partial\theta} & \frac{\partial^2 InL}{\partial c^2} \end{bmatrix}$$

where α_0, θ_0 and c_0 are the initial values of α, θ and c respectively. These equations were solved iteratively till sufficiently close values of $\hat{\alpha}, \hat{\theta}$ and $\hat{c}$ are obtained.

Simulation Study

In this section, we evaluate $\hat{\alpha}_{MLE}, \hat{\theta}_{MLE}$ and $\hat{c}_{MLE}$ through a brief simulation study. The simulation study of the GXLD is carried out by choosing random samples, say n = 50, 100, ..., 5000. These samples are obtained using the inverse cdf. The simulation study is conducted for the combination values α, θ and c respectively. The judgment about the performances of the $\hat{\alpha}_{MLE}, \hat{\theta}_{MLE}$ and $\hat{c}_{MLE}$ are made by considering two evaluation criteria. These criteria are given by

Mean Square Error (MSE)

MSE $(\hat{\alpha}_{MLE}) = \frac{1}{N}\sum_{i=1}^{n}(\hat{\alpha}_i - \alpha)^2$

Bias

Bias $(\hat{\alpha}_{MLE}) = \frac{1}{N}\sum_{i=1}^{n}(\hat{\alpha}_i - \alpha)$

The above evaluation criteria are also computed for $(\hat{\theta}_{MLE})$and $(\hat{c}_{MLE})$. The simulation study is performed for the combination of parameter values by choosing random samples say n = 50, 100, 300, 500, 700, 1000, 5000. The results are presented numerically in Tables 3.

Table 2: The numerical illustration of the Simulation Study of the GXLD for $\hat{\alpha}_{MLE}$, $\hat{\alpha}_{MLE}$and $\hat{c}_{MLE}$

n	Parameter	MLE	Biases	RMSE
50	α	208.09	206.59	206.59
	θ	33.47	30.97	30.96
	c	15.11	13.61	13.61
100	α	125.29	123.79	123.79
	θ	28.83	26.33	26.33
	c	12.29	10.79	10.79
300	α	565.21	563.71	563.71
	θ	36.18	33.68	33.68
	c	15.89	14.39	14.39
500	α	372.53	371.03	371.03
	θ	26.67	24.17	24.17
	c	11.21	9.71	9.71
700	α	642.54	641.04	641.04
	θ	30.28	27.78	27.78
	c	12.90	11.40	11.40
1000	α	467.13	465.63	465.63
	θ	29.80	27.30	27.29
	c	12.81	11.31	11.30
5000	α	1103.33	1101.83	1101.83
	θ	30.52	28.02	28.01
	c	13.17	11.67	11.47

Interpretation

From the results of the simulation of GXLD in Table 3, the followings were observed:

- Smaller sample sizes lead to high bias and RMSE, particularly for α, suggesting that α is more sensitive to sample size.
- As n -> ∞, θ and c parameters show relatively better performance compared to α,
- The RMSE values being equal to Bias indicates that the variance component is minimal or negligible, meaning most of the error comes from bias, not variability.

- As n -> ∞, especially from n = 500 onwards, θ and c show significant improvements, while α still struggles with high bias, suggesting that further modifications or reparameterization might be needed for α.

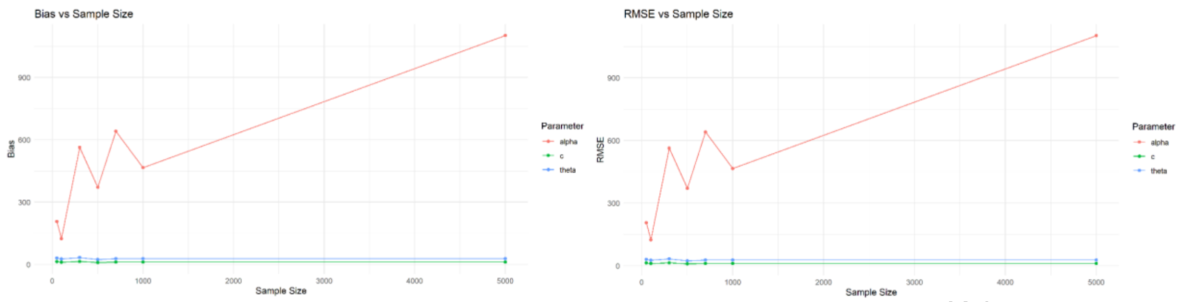


Figure 5 presented the plots of bias and RMSE against different sample size

Models Selection Method

The model selection criteria considered in this thesis are the AIC (Akaike Information Criterion) by Akaike (1974), AICC (Corrected Akaike Information Criterion) by Kletting and Glatting (2009), HQIC (Hannan-Quinn Information Criterion) by Maïnassara and Kokonendji (2016)] and BIC (Bayesian Information Criterion) by Weakliem (1999)and Kolmogorov Smirnov by Massey (1951), Anderson - Darling, (1952) and Pearson (1900). Where the AIC, AICC, BIC, HQIC and KS are obtained as follows:

$$AIC = 2k - 2\ln(L) \quad (26)$$

$$AICC = AIC + \frac{k(k-1)}{n-k-1} \quad (27)$$

$$BIC = kln(n) - 2\ln(L) \quad (28)$$

$$HQIC = -2\ln(L) + 6\ln(\ln(n)) \quad (29)$$

$$KS = sup_x|F_n(x) - F_o(x)| \quad (30)$$

$$AD = -n - \sum_{i=1}^{n} \frac{2i-1}{n}\left[log(F(x_i)) + log\left(1 - F(x_{n-i+1})\right)\right] \quad (31)$$

$$\chi_i^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \quad (32)$$

Where k is the number of parameters in the distribution and n is the number of observations.

APPLICATION

Dataset: The dataset is 2902 data points measure of ozone in airquality per hour of September to December 2024 Ilorin, Kwara State. (OpenWeatherMap, 2025).

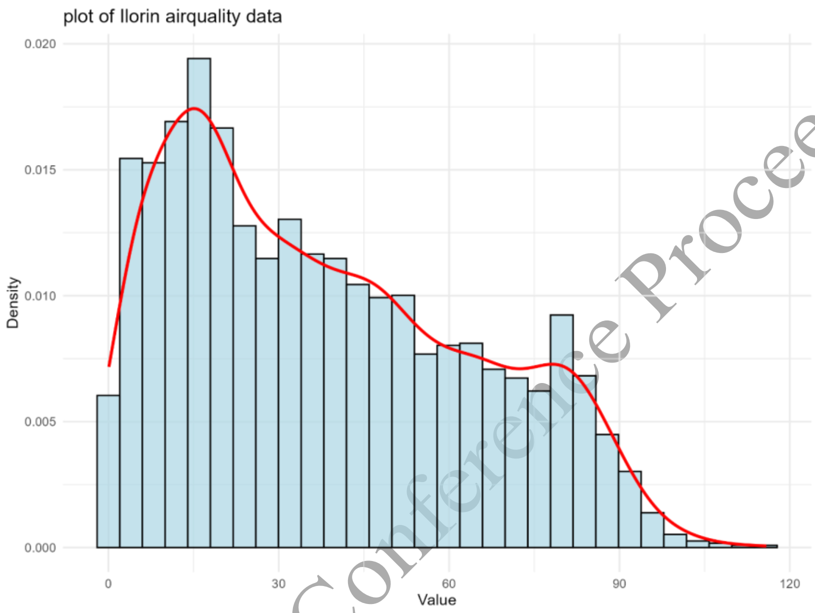


Figure 6: Airquality dataset of Ilorin, Nigeria

Table 3: Parameter estimates and goodness of fit test statistics for dataset

Model	MLE	-ll	AIC	AICC	HQIC	BIC	K-S	KS p-value	AD	Chi-square p-value
GXLD	$\hat{\alpha} = 0.39284$ $\hat{\theta} = 0.04904$ $\hat{c} = 0.14807$	13270.98	26547.19	26547.20	26554.20	26557	0.1088	0.0000	0.0471	0.1129
QXLD	$\hat{\alpha} = 0.1973$ $\hat{\theta} = 0.04736$	13272.76	26550.11	26550.01	26554.20	26561	0.1271	0.0000	0.0322	0.1485
IXLD	$\hat{\alpha} = 29.8916$	17217.2	34436.4	34436.40	34438.55	34442	0.7643	0.0000	0.0081	0.0000
XLD	$\hat{\theta} = 0.05025$	13290	26581	26581	26575	26587	0.2161	0.0000	0.0011	0.0026

Table 3 shows goodness of fit tested with the airquality dataset, its shows that GXLD performed better with the lowest AIC (26547.19), kolmogorov smirnov value 0.1088 indicate that the empirical and theoretical distributions are closer, suggesting a better fit. The KS p-value (0.0000), shows that the difference observed is statistically significant suggesting.

Anderson-Darling p-value it is the best model for this dataset among the model consider. Other models like QXLD, PXLD and IXLD have higher AIC and negative log-likelihood, making them less preferred. The second and the other datasets performed the same, as GXLD shows lowest value of *-ll*, *AIC*, *AICC*, *HQIC*, *BIC*, *K-S*.

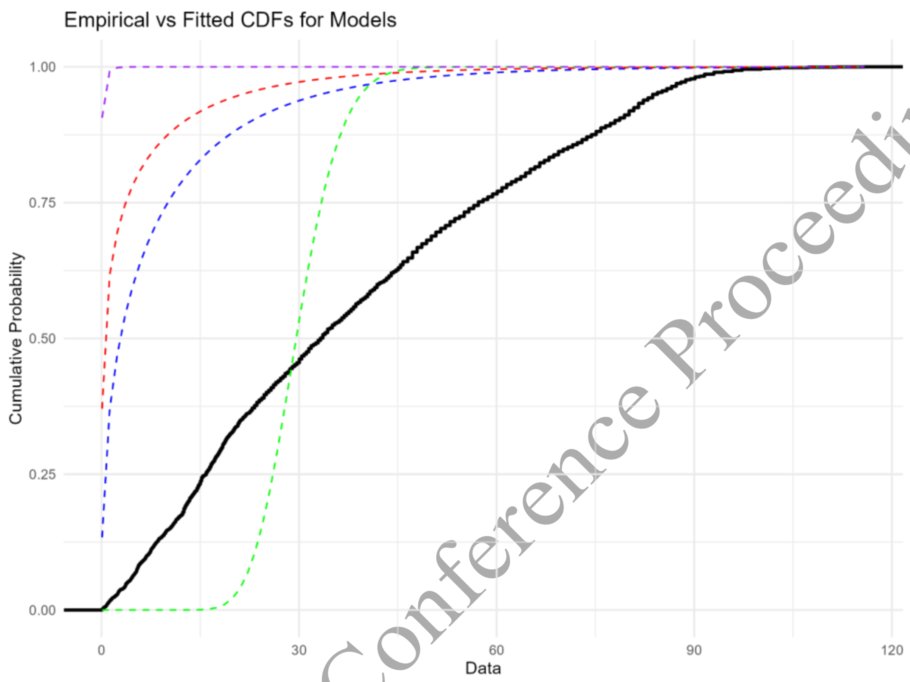


Figure 7: plot of Empirical versus Fitted CDFs

CONCLUSION

This paper aims to model airquality of Ilorin, Kwara State, Nigeria, using the XLindley distribution family. The ozone component of the air quality data was obtained from the OpenWeatherMap website, covering four months from September 1 to December 31, 2024. Four members of the XLindley distribution family, including the baseline distribution, were considered: the XLindley distribution (XLD), Inverse XLindley distribution (IXLD), Quasi XLindley distribution (QXLD), and Generalized XLindley distribution (GXLD). The goal was to identify the best-fitting model for the dataset using model selection criteria such as log-likelihood (ll), Akaike Information Criterion (AIC), Corrected Akaike Information Criterion

(AICC), Bayesian Information Criterion (BIC), and Hannan-Quinn Information Criterion (HQIC).

The GXLD exhibited the lowest values across the model selection criteria, with ll = 13270.98, AIC = 26547.19, AICC = 26547.20, HQIC = 26554.20, and BIC = 26557.01. This indicates that it provides the best fit by striking an optimal balance between goodness-of-fit and model complexity.

Additionally, three goodness-of-fit tests were conducted to evaluate the performance of the empirical and theoretical distributions: the Kolmogorov-Smirnov (KS) test, the Anderson-Darling (AD) test, and the Chi-square test. The GXLD achieved a KS statistic of 0.1088 with a p-value of 0.0000, suggesting that it may not perfectly capture the distribution but performs better than the other models. The AD test yielded a p-value of 0.0471, while the Chi-square test produced a p-value of 0.1129, further indicating the superior performance of the GXLD compared to the alternative models considered (Wilks, 2011).

FUTURE RESEARCH

Having consider GXLD as best model for airquality dataset of Ilorin Kwara State, it is therefore suggested that future research should extend the analysis to other pollutants, such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and carbon monoxide (CO), to provide a more comprehensive assessment of air quality of the city. Extending the study to assess the potential health impacts of varying ozone levels, especially during peak pollution periods, could also provide valuable information for policymakers and health authorities. The performance of the GXLD suggests its suitability for the dataset; however, more refined estimation techniques, such as Bayesian inference or bootstrapping, could be employed to improve parameter accuracy and reduce estimation variability. Predictive models such as Machine learning approaches i.e. time series models or ensemble learning, could be integrated with the distribution-based modelling for enhanced predictive.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19* (6), 716–723.

Anderson, T. W., & Darling, D. A. (1952). *Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes*. The Annals of Mathematical Statistics, 23(2), 193-212. https://doi.org/10.1214/aoms/1177729337

Beghriche, A., Tashkandy, Y. A., Bakr, M. E., Halim, Z., Gemeay, A. M., Hossain, M. M., et al. (2023). The inverse xlindley distribution: Properties and application. *IEEE Access*.

Bell, M. L., et al. (2006). Ozone and short-term mortality in 95 US urban communities, 1987–2000. *JAMA*, 295(5), 1127-1135.

Chouia, S., & Zeghdoudi, H. (2021). The xlindley distribution: Properties and application. *Journal of Statistical Theory and Applications*, 20 (2), 318–327.

EPA. (2020). *Ground-Level Ozone: Health and Environmental Impacts*. U.S. Environmental Protection Agency.

Fishman, J., Bowman, K. W., & Brasseur, G. P. (2010). *Satellite Observations of Tropospheric Ozone*. Springer.

Ibrahim, M., Shah, M. K. A., & Haq, M. Ahsan-ul. (2023). New two-parameter xlindley distribution with statistical properties, simulation and applications on lifetime data. International *Journal of Modelling and Simulation*, 1–14.

Intergovernmental Panel on Climate Change (IPCC). (2021). *Climate Change 2021: The Physical Science Basis*. Cambridge University Press.

Jaffe, D. A., O'Neill, S. M., Larkin, N. K., Holder, A. L., Peterson, D. L., Halofsky, J. E., & Rappold, A. G. (2020). Wildfire and air quality: Perspectives on a changing world. *Atmospheric Chemistry and Physics*, 20(5), 3473-3481.

Jiang, F., & Guterman, E. (2024). Survival analysis. *Statistical Methods in Epilepsy*, 124–142.

Kim, K. H., Kabir, E., & Kabir, S. (2018). A review on the human health impact of airborne particulate matter. *Environmental International*, 74, 136-143.

Kletting, P., & Glatting, G. (2009). Model selection for time-activity curves: the corrected akaike information criterion and the f-test. *Zeitschrift für medizinische Physik*, *19*(3), 200–206.

Lindley, D. V. (1958). Fiducial distributions and bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, 102–107.

Maïnassara, Y. B., & Kokonendji, C. C. (2016). Modified schwarz and hannan–quinn information criteria for weak varma models. *Statistical Inference for Stochastic Processes*, *19*, 199–217.

Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78. https://doi.org/10.1080/01621459.1951.10500769

Meriem, B., Gemeay, A. M., Almetwally, E. M., Halim, Z., Alshawarbeh, E., Abdulrahman, A. T., et al. (2022). [retracted] the power xlindley distribution: Statistical inference, fuzzy reliability, and covid-19 application. *Journal of Function Spaces, 2022* (1), 9094078.

Merovci, F., Sharma, V. K., et al. (2014). The delta-lindley distribution: properties and applications. *Journal of applied mathematics*, 2014 .

Mills, G., et al. (2011). Impacts of tropospheric ozone on vegetation. *Critical Reviews in Environmental Science and Technology*, 41(6), 1012-1042.

Monks, P. S., et al. (2015). Tropospheric ozone and its precursors from global to regional scales. *Atmospheric Chemistry and Physics*, 15(5), 8881-8960.

OpenWeatherMap. (2025). *OpenWeatherMap API documentation*. OpenWeatherMap. Retrieved from https://openweathermap.org

Patil, G. P., Rao, C. R., Zelen, M., & Patil, G. P. (1987). *Weighted distributions*. Citeseer.

Pearson, K. (1900). *X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*. Philosophical Magazine, 50(302), 157-175. https://doi.org/10.1080/14786440009463897

R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Rather, A., & Ozel, G. (2020). The weighted power Lindley distribution with applications on the life time data. *Pakistan Journal of Statistics and operation research, 16*(2).

Robock, A. (2000). Volcanic eruptions and climate. *Reviews of Geophysics*, 38(2), 191-219.

Saghir, A., Hamedani, G. G., Tazeem, S., & Khadim, A. (2017). Weighted distributions: A brief review, perspective and characterizations. *International journal of Statistics and Probability*

Seinfeld, J. H., & Pandis, S. N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. Wiley.

Shanker, R. (2015). Shanker distribution and its applications. *International journal of statistics and Applications, 5* (6), 338–348.

U.S. Environmental Protection Agency (EPA). (2022). *Air Quality Index: A Guide to Air Quality and Your Health*. Retrieved from www.epa.gov

Weakliem, D. L. (1999). A critique of the bayesian information criterion for model selection. *Sociological Methods & Research, 27* (3), 359–397.

Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences*. Academic Press.

Wolfram, S. (1991). Mathematica: a system for doing mathematics by computer. Addison Wesley Longman Publishing Co., Inc.

World Health Organization. (2021). *Air Quality Guidelines: Global Update 2021*. Retrieved from https://www.who.int