Comparative Performance of Burr Type XII 3P, Dagum Type I 3P and Log-Logistic 3P Distributions in Modeling Ozone (O_3), PM_{10} and $PM_{2.5}$ Concentrations

Saripalli Arun Kumar[1,2], Akiri Sridhar[1*], Sarode Rekha[3], Vasili B. V. Nagarjuna[4] and Ramanaiah M.[5]
1. Department of Mathematics, GSS, GITAM (Deemed to be) University, Visakhapatnam-530045, INDIA
2. Department of Mathematics, Aditya Institute of Technology and Management, Tekkali-532201, INDIA
3. Department of Mathematics, Madanapalle Institute of Technology & Science, Chittoor-517325, INDIA
4. Department of Mathematics, VIT-AP University, Amaravati-522237, INDIA
5. Department of Chemistry, Aditya Institute of Technology and Management, Tekkali-532201, INDIA
*sakiri@gitam.edu

Abstract

This study investigates the suitability of three parameters continuous probability distributions-Burr Type XII 3P, Dagum Type I 3P and Log-Logistic 3P-in modeling secondary air pollutants: ozone (O_3), particulate matters (PM_{10} and $PM_{2.5}$) in Visakhapatnam, an urban region having rapid industrialization. By employing rigorous statistical techniques including maximum likelihood estimation (MLE) and bootstrapping, we estimate distribution parameters and validate model fit through diagnostic plots-skewness vs. kurtosis, P-P and Q-Q plots as well as goodness-of-fit test-statistics, such as Kolmogorov-Smirnov(KS), Anderson-Darling(AD) and Cramér von Mises(CvM) tests. Additional, performance metrics including Akaike information criterion(AIC), Bayesian information criterion(BIC), evaluation metrics like mean absolute error(MAE), mean absolute percentage error(MAPE), mean squared error(MSE), root mean squared error(RMSE) and coefficient of determination(R^2) and cross-validation, were also applied to ensure model robustness.

Results indicate that the Burr Type XII 3P distribution most effectively models the high variability and skewed nature of O_3 concentrations, while the Dagum Type I 3P distribution provides the best fit for PM_{10} and both Burr Type XII 3P and Log-Logistic 3P distributions are suitable for $PM_{2.5}$. These findings offer new insights into the behavior of secondary pollutants, supporting the development of robust air quality monitoring frameworks. R software facilitated all numerical analyses and visualizations of data suited to environmental data modeling.

Keywords: Urbanization, Air Pollution, Statistical Distributions, Secondary Air Pollutants, Model Selection.

Introduction

Air pollution, often called the "silent killer," is a major global health crisis, contributing to an estimated seven million preventable deaths annually, as reported by the World Health Organization (WHO). Among the most harmful pollutants are PM_{10} (particulate matter≤10microns), $PM_{2.5}$ (particulate matter≤2.5microns) and O_3 (ozone) which penetrate deeply into the respiratory system and increasing the risk of severe health issues, such as cardiovascular disease, lung cancer and respiratory infections. While extensive research has focused on extreme pollution levels in megacities, a growing body of evidence suggests that mid-sized, rapidly industrializing cities also face substantial air quality challenges.

Visakhapatnam, a coastal city in India experiencing rapid industrialization, exemplifies this trend. Industrial activities such as port operations, steel production and construction contribute to the release of pollutants, transforming primary emissions into secondary pollutants like O_3, PM_{10} and $PM_{2.5}$ through complex chemical reactions. While air pollution levels in this city are not as extreme as those of major metropolitan areas, they frequently exceed national ambient air quality standards and pose significant public health risks. Given its status as one of India's most polluted mid-sized cities, understanding pollutant trends and variability in Visakhapatnam are essential in developing effective air quality controls and public health strategies.

Existing studies on air pollution modeling have applied various statistical distributions to analyze pollutant concentrations in different urban contexts. For instance, Gavriil et al[10] found the Pearson type VI distribution to best fit PM_{10} and $PM_{2.5}$ data in Athens, with inverse Gaussian, lognormal and Pearson type V also performing well.

Noor et al[15] found the gamma distribution best for PM_{10} in Nilai and log-normal best for Shah Alam, with both cities staying within MAAQG limits despite occasional exceedances in 2007. Benjamin et al[3] demonstrated the superiority of the Dagum distribution over the Generalized Extreme Value (GEV) distribution for modeling tropospheric ozone (O_3) levels. Ahmat et al[2] found the GEV distribution most accurate for predicting PM_{10} concentrations in Malaysia. Thupeng[19] used the Burr-XII distribution to model daily maximum nitrogen dioxide levels in Gaborone, finding it superior to the Dagum and Log-Logistic distributions.

El-Shanshoury[8] identified the Frechet distribution with the Hosking and Wallis plotting position as the best fit for TSP

and PM_{10} concentrations in Ain Sokhna. Jaffar et al[11] reviewed gamma, lognormal and Weibull distributions for modeling air pollution data, highlighting the need for accurate models to predict high pollution events and improve air quality management. Febriantikasari et al[9] applied the Dagum distribution to simulate PM_{10} concentrations in Pekanbaru, Indonesia, concluding that the L-Moments method was the most effective for parameter estimation.

Bhandari[4] used a two-parameter lognormal distribution to fit PM_{10} data in Kathmandu, Nepal and found the method of moments provided the best fit for the observed concentrations. Warsono et al[21] determined that generalized distributions, particularly the Generalized Log-Logistic, provided the best fit for $PM_{2.5}$ concentrations during the COVID-19 pandemic in Jakarta. Omar[16] found that the method of moments was the most effective parameter estimator for the lognormal distribution in predicting PM_{10} concentrations in suburban areas, especially in Jerantut and Sungai Petani. In contrast, the uniformly minimum variance unbiased estimator demonstrated strong accuracy in Muar and Kuantan. Jaffar et al[11] identified Nakagami and Gamma distributions as the best fits for ground-level ozone in Malaysian cities, with Nakagami suitable for Kuala Terengganu and Alor Setar and Gamma for Kota Bharu.

Choopradit et al[6] determined that inverse Gaussian and Pearson type V distributions were most suitable for $PM_{2.5}$ concentrations in Bangkok, highlighting that the 98th percentile of $PM_{2.5}$ levels exceeded Thailand's 24-hour threshold, indicating significant health risks. Kumar et al[13] examined seasonal trends in $PM_{2.5}$ and PM_{10} in Visakhapatnam from January 2020 to December 2022, finding the highest levels in December 2020, with concentrations exceeding NAAQS limits in winter and improved air quality in the summer and monsoon seasons[13].

This study evaluates the fit of three flexible three-parameter distributions-Burr type XII 3P, Dagum type I 3P and Log-Logistic 3P-for modeling daily mean levels of O_3, PM_{10} and $PM_{2.5}$ in Visakhapatnam. Chosen for their ability to capture the heavy-tailed and skewed characteristics of air pollutants, these distributions are tested using data from January 2018 to December 2022 provided by the Andhra Pradesh Pollution Control Board. The model performance is assessed through various methods including skewness-kurtosis plots, PDFs, ECDFs, P-P and Q-Q plots and Maximum Likelihood Estimation (MLE) for parameter estimation.

Bootstrapping generates confidence intervals and model validation is conducted via residual plots and goodness-of-fit tests (K-S, A-D, CvM), alongside AIC and BIC metrics. Additional assessments including error metrics, kernel density estimation and cross-validation, further refine the model comparison. The study aims to identify the most suitable distribution for air quality modeling, supporting targeted policy interventions and public health strategies for mid-sized urban areas.

Material and Methods

Research Area and Data: This research study analyzes daily average ambient air quality data collected from January 2018 to December 2022 by the Andhra Pradesh Pollution Control Board (APPCB) at the Continuous Ambient Air Quality Monitoring Station (https://pcb.ap.gov.in/) in Visakhapatnam (GVMC), an industrial city on India's eastern coast. The city's industrial activities including port operations and steel production, contribute to significant air pollution, making it an important site for air quality monitoring. The dataset includes daily measurements of secondary pollutants-ozone (O_3), PM_{10} and $PM_{2.5}$, along with primary pollutants including CO, NO, NO_2, NOx, NH_3, SO_2 and volatile organic compounds (VOCs) such as xylene, toluene and benzene. Daily data from the five-year period were used to assess pollutants trends. Missing values were imputed via linear interpolation and outliers (values beyond three standard deviations) were handled through Winsorization. This dataset supports advanced statistical modeling (Burr type XII 3P, Log-Logistic 3P and Dagum type I 3P) to analyze secondary pollutant (O_3, PM_{10} and $PM_{2.5}$) variability in the industrial urban setting[5].

Methods: The analysis employed three probability density functions (PDFs)-Burr type XII 3P, Dagum type I 3P and Log-Logistic 3P distributions to model air pollutant concentrations. To evaluate the suitability of these models, several methods were used. Goodness-of-fit tests (GoF) including the Kolmogorov-Smirnov, Anderson-Darling and Cramér-von Mises tests, were conducted using R to assess how well the PDFs fit the data. Additionally, model performance was gauged using criteria such as Log-Likelihood (LL), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Hannan-Quinn Information Criterion (HAIC), Consistent AIC (CAIC) and Adjusted BIC (ABIC). Model selection metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared and Cross-Validation, were further utilized to compare the models. In addition, Kernel Density Estimation, residual analysis and diagnostic plots including PDF, ECDF, Q-Q and P-P plots were employed to provide a comprehensive assessment of the goodness of fit, ensuring a robust evaluation of the statistical models for pollutant concentration distribution.

Statistical Distributions in Air Pollution: Statistical probability distributions are vital for modeling air pollution data, offering a robust framework to capture the inherent variability and trends in environmental quality. This study focuses on modeling the concentrations of ozone (O_3), particulate matter (PM_{10} and $PM_{2.5}$) using continuous three-parameter (3P) distributions: Burr type XII 3P, Dagum type I 3P and Log-Logistic 3P. These distributions were chosen for their adaptability in capturing skewed, heavy-tailed data, which is typical of air pollutant distributions. Table 1 presents the Cumulative Distribution Function (CDF, F(x)) and Probability Density Function (PDF, f(x)) for these distributions[18].

Table 1
The PDF and CDF of Three Parameter Distributions

Distribution	Synopsis	PDF	CDF	Parameters
Burr Type XII 3P	Models data with various shapes and is useful for skewed or heavy-tailed data.	$f(x;\alpha,\omega,\lambda)=\frac{\alpha\cdot\omega\cdot(x-\lambda)^{\alpha-1}}{\left[1+\left(\frac{x-\lambda}{\omega}\right)^{\alpha}\right]^{\alpha+1}}$	$F(x;\alpha,\omega,\lambda)=1-\left[1+\left(\frac{x-\lambda}{\omega}\right)^{\alpha}\right]^{-\infty}$	Where $x>\lambda,\alpha>0$(Shape), $\omega>0$(Scale) and $\lambda>0$ (Location).
Dagum Type I 3P	Versatile in modeling income and wealth data, accounting for skewness and kurtosis[14].	$f(x;\delta,\tau,\vartheta)=\frac{\delta\cdot\tau\cdot x^{\delta-1}}{\vartheta\left[1+\left(\frac{x}{\vartheta}\right)^{\delta}\right]^{\tau+1}}$	$F(x;\delta,\tau,\vartheta)=\left[1+\left(\frac{x}{\vartheta}\right)^{\delta}\right]^{-\tau}$	Where $x>0,\delta>0$ (Shape) $\tau>0$(Shape) and $\vartheta>0$ (Scale)
Log-Logistic 3P	Suitable for modeling survival times and skewed data.	$f(x;\eta,\beta,\gamma)=\frac{\eta\cdot\beta\cdot(x-\gamma)^{\eta-1}}{[\beta+(x-\gamma)^{\eta}]^{\eta+1}}$	$F(x;\eta,\beta,\gamma)=\frac{1}{1+\left(\frac{\beta}{x-\gamma}\right)^{\eta}}$	Where $x>\gamma,\ \eta>0$(Shape), $\beta>0$ (Scale), and $\gamma>0$(Location).

Methods of Parameter Estimation: Parameter estimation was conducted using the maximum likelihood estimation (MLE) method, selected for its consistency and efficiency in environmental data modeling. MLE maximizes the likelihood function, allowing us to identify the parameter values that best align the chosen distribution with observed data. For each model, the likelihood function was constructed based on pollutant data and numerical optimization techniques were used to derive the parameters.

To quantify variability and uncertainty, we employed the bootstrap resampling technique with 1,000 resamples. For each resample, MLE was recalculated to generate 95% confidence intervals for the parameters, offering insight into model stability. Narrower intervals indicate a more stable parameter estimate, enhancing the robustness of the fitted models. This approach also provided bias and standard error values, further validating model reliability.

Method of Kernel Density Estimation (KDE): KDE is a non-parametric technique used to estimate the PDF of air pollutant concentrations, (O_3, PM_{10}, $PM_{2.5}$) without assuming a specific distribution. It provides a smooth, continuous estimate of pollutant distributions from observed data, enabling detailed visualization of concentration patterns over time. The bandwidth, or smoothing parameter, is crucial in balancing data pattern capture and noise reduction, preventing over fitting or excessive smoothing. The density estimates at a point x is defined as:

$$\hat{f}(x)=\frac{1}{nh}\sum_{n=1}^{n}K\left(\frac{x-X_i}{h}\right)$$

where K is the kernel function, h is the bandwidth, X_i is the data points and n is the sample size. Unlike parametric models such as Burr type XII 3P, Dagum type I 3P and Log-Logistic 3P, which rely on predefined distributions, KDE adapts to the actual data, making it useful for detecting outliers and assessing model fit.

In this study, KDE was employed to analyze the distribution of pollutant concentrations and to complement the evaluation of parametric models, offering an additional layer of insight into the behavior of air pollutants.

Methods of Model Selection Criteria: After estimating parameters, each model was evaluated for its fit and predictive accuracy using a comprehensive set of selection criteria. These criteria balance model fit with complexity, ensuring that selected models to capture pollutant concentration distributions effectively without overfitting.

Goodness-of-Fit-Tests: To rigorously evaluate the fit of statistical models to observed air pollution data, various goodness-of-fit tests were applied[7]. These tests compare the observed data against the expected values under the fitted model, providing a measure of how well the model captures the underlying distribution of the data.

Kolmogorov-Smirnov (K-S) Test: KS test is sensitive to discrepancies in both the location and shape of the distributions, making it a versatile tool for assessing model adequacy across a broad range of distribution types. The K-S test D is defined as:

$$D=\sup_x\,|\,F_n(x)-F(x)\,|,-\infty<x<\infty$$

where $F_n(x)$ is the ECDF and $F(x)$ is the CDF of the models. A lower KS statistic and higher p-value indicate a close alignment between observed data and model fit.

Anderson-Darling (A-D) Test: A-D test extends the K-S test by placing more weight on the tails of the distribution, making it particularly useful for detecting discrepancies in tail behaviour. The A-D test A^2 is given by:

$$A^2=-n-\frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\left(\ln F(x_i)+\ln\left(1-F(x_{n+1-i})\right)\right)\right]$$

where n is the sample size, X_i are the ordered data points and $F(x_i)$ is the CDF of the fitted model. A larger A^2 value indicates a worse fit, especially in the tails. This test provides a comprehensive assessment of model fit by focusing on deviations in both central and tail areas.

Cramer-von Mises Test: CvM test is the measuring the squared differences between the empirical and theoretical CDFs across the entire data range. Unlike the K-S test, which focuses on deviations at specific points, it provides a global fit assessment. The CvM test W^2 is calculated as:

$$W^2 = \int_{-\infty}^{\infty} \left(F_n(x) - F(x)\right)^2 dF(x)$$

where $F_n(x)$ is the empirical CDF and $F(x)$ is the theoretical CDF. A higher W^2 value indicates a greater deviation from the model. Significance is assessed by comparing the W^2 value to critical values from CvM tables or simulations.

Methods of Graphical Validation: Graphical validation methods provide visual tools to assess the fit of statistical models to data, offering intuitive insights into model performance and potential discrepancies.

Quantile-Quantile (Q-Q) Plots: Q-Q plot is a widely used graphical tool for evaluating how closely a dataset follows a specified distribution. It compares the quantiles of the observed data with those of the fitted distribution. In this plot, observed data values are plotted against the theoretical quantiles derived from the cumulative distribution function (CDF) of the fitted model. If the data align with the assumed distribution, the points will form a straight line, indicating a good fit. The process of constructing a Q-Q plot using n data points is defined as $\{x_i, F^{-1}(p_i)\}$ where i=1, 2,…,n. Q-Q plots compare the quantiles of the observed data against the quantiles of the fitted distribution. Deviations from the 45-degree line in a Q-Q plot reveal discrepancies between the observed and expected distributions, signaling areas where the model may not fit the data accurately.

Probability-Probability (P-P) Plots: P-P plot is a graphical tool used to evaluate the goodness of fit between an empirical dataset and a specified theoretical distribution. It compares the cumulative distribution function (CDF) of the observed data with that of the fitted model by plotting the empirical probabilities against the corresponding theoretical probabilities. If the data align with the assumed distribution, the points will follow the 45-degree reference line, indicating a good fit. P-P plots are especially useful for identifying deviations from the expected distribution, particularly in the tails.

Residual Plots: Residual plots are graphical tools used to assess the goodness of fit of a statistical model by analysing the differences between observed and predicted values (residuals). Typically, residuals are plotted against predicted values or another variable. A well-fitting model shows residuals randomly scattered around zero, indicating no systematic pattern.

Residual plots help to detect issues like model misspecification, heteroscedasticity, or non-linearity. If patterns appear in the plot, it suggests areas for model improvement. These plots complement quantitative goodness-of-fit tests by providing visual insights into model adequacy.

Methods of Model Performance Criteria: Model performance criteria are quantitative measures used to assess the accuracy and efficiency of statistical models in predicting pollution levels[7]. These metrics help to compare different models, ensuring the selection of one that balances complexity with accurate pollutant predictions.

By evaluating these metrics, we can identify which model best captures the variability in pollutants like O_3, PM_{10} and $PM_{2.5}$. Table 2 provides the formulas and descriptions for all performance indicators.

Methods of Model Evaluation Metrics: Model evaluation metrics are critical for assessing the predictive performance and generalizability of statistical models, especially for pollutant concentrations like O_3, PM_{10} and $PM_{2.5}$. Accuracy metrics closer to one indicate better performance while evaluation metrics nearer to zero suggest a better fit. Table 3 lists the formulas and descriptions for all performance indicators.

Results and Discussion

Statistical Characteristics of O_3, PM_{10} and $PM_{2.5}$: Table 4 presents the descriptive statistics for ozone (O_3), PM_{10} and $PM_{2.5}$ concentrations, based on 1,627 observations from January 1, 2018, to December 31, 2022, highlighting the distribution and variability of these pollutants over time[1].

Table 4 shows key statistics for ozone (O_3), PM_{10} and $PM_{2.5}$ concentrations from 2018–2022. Ozone has a low mean concentration (27.75µg/m³), well below the NAAQS standard (180µg/m³), but shows significant variability with occasional high levels (skewness: 2.02, kurtosis: 4.96). PM_{10} exceeds the NAAQS daily limit (70µg/m³) with a mean of 109.75µg/m³ and moderate variability (skewness: 1.14, kurtosis: 1.73). $PM_{2.5}$ also exceeds the NAAQS standard (35µg/m³) with a mean of 45.26µg/m³, showing occasional spikes (skewness: 1.47, kurtosis: 2.92). The coefficient of variation indicates that ozone has the highest relative variability. These findings emphasize the need for stronger air quality controls, especially for ozone and PM_{10}, where safety threshold violations are more likely.

Figures 1 and 2 illustrates daily mean concentrations of ozone (O_3), PM_{10} and $PM_{2.5}$ in Visakhapatnam (2018–2022) revealing significant variability and seasonal patterns, essential for selecting appropriate statistical models. O_3 shows peaks around 2019–2020, likely driven by photochemical reactions while PM_{10} exhibits spikes due to localized pollution events like dust storms. $PM_{2.5}$ shows periodic spikes from vehicular and industrial emissions[16].

These patterns underscore the importance of applying adaptable statistical models to capture the complex characteristics of pollutants in urban regions.

Table 2
Model Information Criteria and their Formulas, Descriptions

Criterion	Description	Formula	Interpretation
LL	Log-Likelihood function measures how well the model explains the observed data, representing the probability of the data given the model's parameters.	$\ln L(\theta)$	Higher Log-Likelihood indicates a greater probability of the model's parameters being consistent with the observed data.
AIC	Akaike Information Criterion balances model fit and complexity by penalizing the number of parameters, thus helping to prevent overfitting	$2.k - 2.\ln(L)$	Lower AIC indicate a better model, as they balance fit and complexity, favoring simpler models with comparable explanatory power.
BIC	Bayesian Information Criterion applies a stronger penalty for complexity than AIC, making it more suitable for larger datasets while promoting parsimony and good fit.	$k.\ln(n) - 2.\ln(L)$	Lower BIC indicates a more parsimonious model, particularly in large datasets, helping to prevent over fitting.
HAIC	Hannan-Quinn Information Criterion balances AIC and BIC, applying a milder penalty for complexity, making it suitable for moderately large samples.	$2.k.ln(\ln(n)) - 2.\ln(L)$	Lower HQIC reflects a balance fit and complexity, with a moderate penalty, making it useful for considering both aspects.
CAIC	Consistent Akaike Information Criterion modifies AIC by applying a stronger penalty for model complexity, offering a more consistent criterion for model selection.	$(k+1).\ln(n) - 2.\ln(L)$	Lower CAIC indicate a model that balances fit and complexity while accounting for sample size, making it robust across varying sample sizes.
ABIC	Adjusted Bayesian Information Criterion modifies BIC to better account for small sample sizes, improving model selection consistency by adjusting the penalty for model complexity	$k.\ln(n) + \frac{2.k^2}{n} - 2.\ln(L)$	Lower ABIC indicate a more optimal model for smaller sample sizes, balancing fit and complexity with an adjusted penalty
Notations: n: Sample size, k: Number of estimated parameters, L: Likelihood function, ln(L): Log-Likelihood and θ: model parameters.			

Table 3
Model Error Metrics and Their Formula, Descriptions

Metric	Description	Formula	Interpretation
MBE	Mean Bias Error quantifies the average bias in predictions, revealing whether the model tends to overestimate or underestimate values.	$\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)$	Measures average bias, with values closer to 0 indicating less bias in the model's predictions.
MAE	Mean Absolute Error quantifies the average magnitude of errors between predicted and actual values, without considering the direction of the errors.	$\frac{1}{n}\sum_{i=1}^{n}\lvert(P_i - O_i)\rvert$	Lower MAE indicates better model performance.
MSE	Mean Squared Error calculates the average of the squared differences between predicted and actual values, giving greater weight to larger errors.	$\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2$	Lower MSE indicates better accuracy.
MdAE	Median Absolute Error is the median of absolute errors, offering robustness to outliers and making it less sensitive to extreme values compared to MAE.	$\text{median}(\lvert(P_i - O_i)\rvert)$	Lower MdAE indicates that the model consistently performs.

MAPE	Mean Absolute Percentage Error is the average of the absolute percentage errors between predicted and actual values, providing a clear measure of prediction error in percentage terms.	$\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{(P_i-O_i)}{O_i}\right\| \times 100\%$	Lower MAPE indicates better relative accuracy, with predictions closer to the actual values.
RMSE	Root Mean Squared Error is the square root of MSE, providing error magnitude in the same units as the data.	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2}$	Lower values indicate better accuracy.
NRMSE	Normalized RMSE is RMSE normalized by the mean of the observed data.	$\frac{\text{RMSE}}{\text{mean}(O_i)}$	Lower values indicate better accuracy relative to the data range or mean.
CV-RMSE	Coefficient of Variation of RMSE is RMSE normalized by the mean of observed values, expressed as a percentage.	$\frac{\text{RMSE}}{\text{mean}(O_i)} \times 100\%$	Lower values indicate better model performance relative to mean of the observed data.
R^2	Coefficient of Determination measures the proportion of the variance in the dependent variable that is predictable from the independent variables.	$\frac{\sum_{i=1}^{n}(O_i-\bar{O})^2-\sum_{i=1}^{n}(P_i-O_i)^2}{\sum_{i=1}^{n}(O_i-\bar{O})^2}$	Values closer to 1 indicate a better fit.
Cross-Validation (K-Fold)	A method to evaluate model predictive performance and generalizability by partitioning the dataset into k subsets. The model is trained on k−1 subsets and tested on the remaining subset. This process repeats k times, ensuring each subset serves as the test set once[22].	The average performance metrics across all iterations provide a robust measure of the model's ability to generalize to unseen data, helping mitigate overfitting.	The model's generalization capability is assessed to ensure it isn't over fitted to the dataset. Lower variance in performance metrics across folds indicates greater stability.
Notes: P_i: Predicted value, O_i: Observed value, n: Number of observations, $\bar{P}$: Mean of predicted values, $\bar{O}$: Mean of observed values.			

Table 4
Descriptive Statistics Summary for O_3, PM_{10} and $PM_{2.5}$ Concentrations

Statistic	Air Pollutant		
	O_3	**PM_{10}**	**$PM_{2.5}$**
Sample Size (n)	1,627	1627	1627
Minimum (µg/m³)	1.9	12	4
Maximum (µg/m³)	160.2	376	207
1st Quartile	12.2	74	26
Median (µg/m³)	20.7	101	37
Mean (µg/m³)	27.75	109.75	45.26
3rd Quartile	34	133	58
Range (µg/m³)	158.3	364	203
Standard Error of Mean	0.59	1.31	0.7
Lower 95% CI for Mean	26.6	107.19	43.89
Upper 95% CI for Mean	28.9	112.31	46.63
Variance	556.07	2775.9	790.29
Covariance	0.85	0.48	0.62
Standard Deviation (µg/m³)	23.58	52.69	28.11
Skewness	2.02	1.14	1.47
Kurtosis	4.96	1.73	2.92
Trimmed Mean (10%) (µg/m³)	23.41	104.02	41.47
IQR	21.8	59	32
Median Absolute Deviation (µg/m³)	14.53	43	22.24

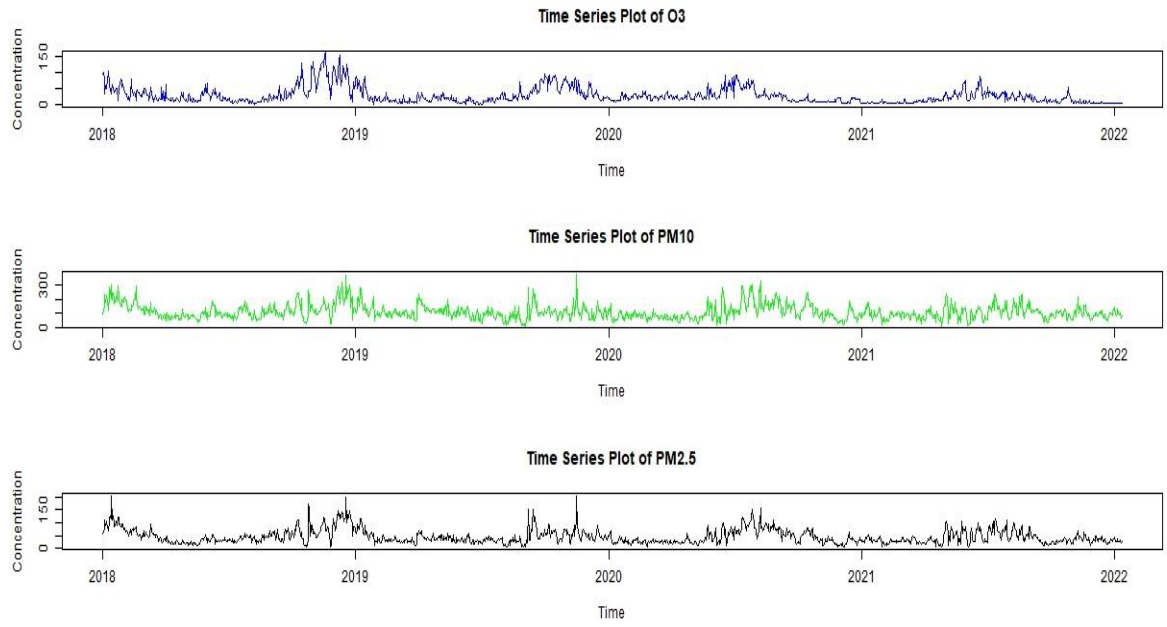

Figure 1: Individual Time Series Plot of Daily Mean Concentrations of O_3, PM_{10} and $PM_{2.5}$.

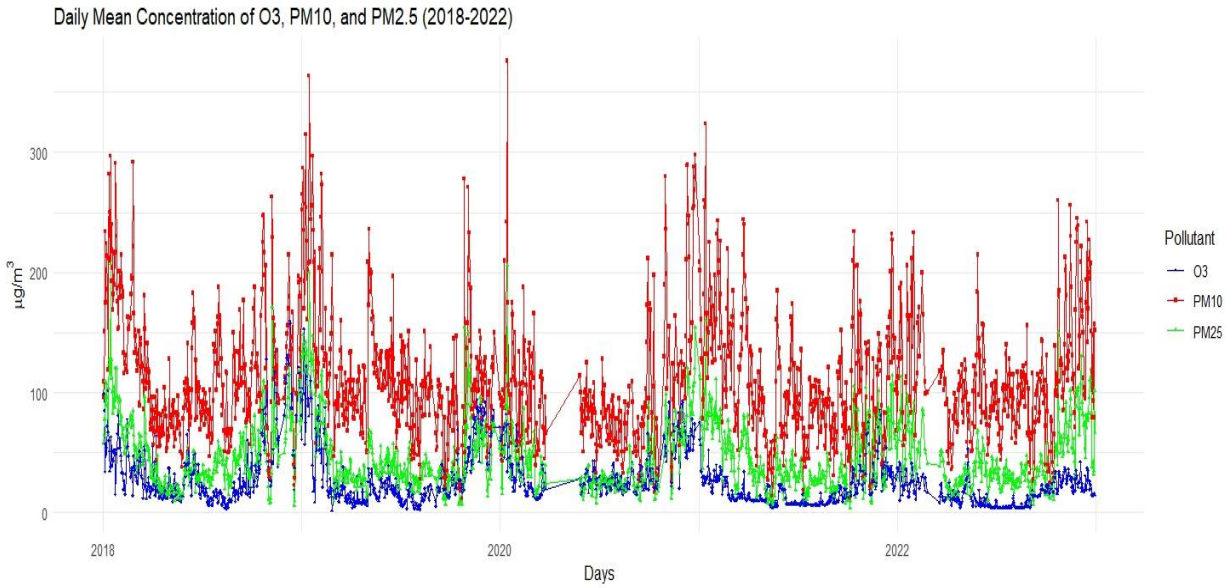

Figure 2: Combined Time Series Plot of Daily Mean Concentrations of O_3, PM_{10} and $PM_{2.5}$.
Figures 1 and 2: Time Series Plot of Daily Mean Concentrations of Ozone (O_3), Particulate Matter (PM_{10} and $PM_{2.5}$) in Visakhapatnam (2018–2022).

Table 5
Descriptive Parameters of Empirical Distribution for Non-Censored Data: O_3, PM_{10} and $PM_{2.5}$ Concentrations.

Pollutant	Min	Max	Median	Mean	Est.Sd	Est.Skewness	Est.Kurtosis
O_3	1.9	160.2	20.7	27.75	23.58	2.03	7.99
PM_{10}	12	376	101	109.75	52.69	1.14	4.74
$PM_{2.5}$	4	207	37	45.26	28.11	1.47	5.94

The Burr type XII 3P, Dagum type I 3P and Log-Logistic 3P distributions were chosen to capture these behaviors. Heavy tails in O_3, $PM_{2.5}$ concentrations align with Burr Type XII 3P and Log-Logistic 3P models, while extreme values in PM_{10} fit the Dagum type I 3P distribution. These insights explain the models' comparative performance based on metrics like AIC, BIC and goodness-of-fit-tests, underscoring the need for advanced distributions to model pollutant variability.

Table 5 presents descriptive parameters for O_3, PM_{10} and $PM_{2.5}$. O_3 displays significant right estimated skewness (2.03) and high kurtosis (7.99), indicating a heavy-tailed

distribution with potential outliers. PM_{10} shows moderate estimated skewness (1.14) and kurtosis (4.74), suggesting a more symmetrical distribution compared to O_3. $PM_{2.5}$ falls between the two, with an estimated skewness of 1.47 and kurtosis of 5.94, indicating a right-skewed distribution but less extreme than O_3. These variations suggest that O_3 may require flexible distributions like Burr type XII 3P or Log-logistic 3P, while PM_{10} could be better modeled by the Dagum type I 3P. $PM_{2.5}$ may benefit from a combination of these approaches.

Figure 3 displays Cullen and Frey graphs for O_3, PM_{10} and $PM_{2.5}$, offering insights for selecting appropriate distributions to model these air pollutants. The graph for O_3 shows high skewness and kurtosis, suggesting that Burr type XII 3P and Log-logistic 3P distributions are suitable due to their capacity to handle heavy tails and significant asymmetry. PM_{10} graph indicates moderate skewness and kurtosis, making Dagum type I 3P distribution a good fit for this pollutant. In contrast, $PM_{2.5}$ graph mirrors O_3, with high skewness and kurtosis, indicating that Burr type XII 3P and Log-logistic 3P distributions are also appropriate for $PM_{2.5}$. Overall, while the Dagum type I 3P is more suitable for PM_{10}, Burr type XII 3P and Log-logistic 3P distributions are better suited for O_3 and $PM_{2.5}$, reflecting their distinct statistical properties.

Parameter Estimates and Confidence Intervals for Each Model Distribution: The Maximum Likelihood Estimation (MLE) method was employed for parameter estimation of each model, with computations performed using R software. Table 6 illustrated the estimated parameters and standard errors for each advanced statistical distribution. The Burr type XII 3P indicates heavy tails and substantial variability for ozone, significant spread for PM_{10} and moderate dispersion for $PM_{2.5}$. The Log-Logistic 3P suggests lighter tails for ozone, reduced variability for PM_{10} and somewhat heavier tails for $PM_{2.5}$. The Dagum type I 3P exhibits significant right skewness and heavy tails for ozone, high variability for PM_{10} and moderate spread with substantial skewness for $PM_{2.5}$. These findings highlight the importance of selecting appropriate distributions for accurate air quality modeling and effective management.

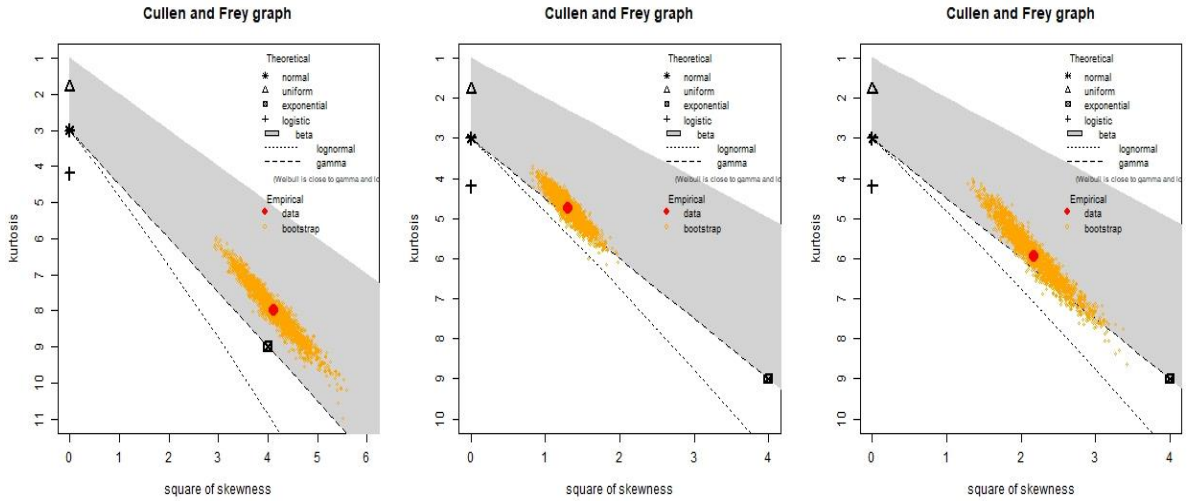

Figure 3: Cullen and Frey Graphs for Assessing Skewness vs. Kurtosis in O_3, PM_{10} and $PM_{2.5}$ Data for Distribution Fitting.

Table 6
Estimated Parameter Values of each Distribution for O_3, PM_{10} and $PM_{2.5}$ using MLE

Distributions	Air Pollutants	Parameter 1	Std. Error 1	Parameter 2	Std. Error 2	Parameter 3	Std. Error 3
Burr Type XII 3P	O_3	2.1627	0.0895	0.4190	0.0316	14.7275	0.5073
	PM_{10}	3.1679	0.1177	0.1916	0.0176	73.4368	1.2175
	$PM_{2.5}$	2.6883	0.1212	0.3020	0.0275	27.2696	0.7402
Log-Logistic 3P	O_3	0.4993	0.0138	2.9259	0.0262	1.6352	0.2329
	PM_{10}	0.2764	0.1103	4.5909	0.0325	0.9004	2.8336
	$PM_{2.5}$	0.3673	0.0142	3.5837	0.0339	1.7618	0.9798
Dagum Type I 3P	O_3	2.1326	0.0894	18.6112	1.5456	1.1532	0.1304
	PM_{10}	4.2431	0.1711	114.4079	3.5039	0.6944	0.0558
	$PM_{2.5}$	2.7935	0.1088	36.1097	2.0832	1.0957	0.1121

Table 7 presents the 95% confidence intervals (CIs) for each parameter across the Burr type XII 3P, Log-Logistic 3P and Dagum type I 3P distributions in modeling for O_3, PM_{10} and $PM_{2.5}$. The CIs for O_3 parameters in the Burr type XII 3P distribution are relatively narrow, indicating moderate variability (Shape1: 1.9873–2.3381, Shape2: 0.3571–0.4809, Scale: 13.7333–15.7218). In contrast, the intervals for PM_{10} parameters are broader, suggesting higher uncertainty (Shape1: 2.9371–3.3987, Shape2: 0.1570–0.2261, Scale: 71.0506–75.8230) while $PM_{2.5}$ CIs reflect intermediate variability. The Log-Logistic 3P distribution shows tighter CIs for O_3 (Shape: 0.4723–0.5263, Scale: 2.8745–2.9772) but broader intervals for PM_{10} and $PM_{2.5}$, especially for the threshold parameter (PM_{10}: -4.6535 to 6.4542; $PM_{2.5}$: -0.1585 to 3.6822).

The Dagum type I 3P distribution exhibits moderate variability, with the widest CIs for PM_{10} (Shape1.a: 3.9077–4.5784, Scale: 107.5404–121.2755, Shape2.p: 0.5850–0.8037) and narrower intervals for $PM_{2.5}$. Overall, the Burr type XII 3P distribution provides more precise parameter estimates, while the Log-Logistic 3P shows significant variability in the threshold parameter and the Dagum type I 3P demonstrates moderate uncertainty, particularly for PM_{10}.

Table 7
Confidence Intervals for each Distributions and Pollutants (O_3, PM_{10} and $PM_{2.5}$) using MLE

Distributions	Parameter	O_3 (2.5%CI)	O_3 (97.5%CI)	PM_{10} (2.5%CI)	PM_{10} (97.5%CI)	$PM_{2.5}$ (2.5%CI)	$PM_{2.5}$ (97.5%CI)
Burr Type XII 3P	Shape1	1.9873	2.3381	2.9371	3.3987	2.4507	2.9259
	Shape2	0.3571	0.4809	0.1570	0.2261	0.2480	0.3559
	Scale	13.7333	15.7218	71.0506	75.8230	25.8188	28.7204
Log-Logistic 3P	Shape	0.4723	0.5263	0.2548	0.2980	0.3394	0.3951
	Scale	2.8745	2.9772	4.5272	4.6545	3.5173	3.6501
	Threshold	1.1789	2.0916	-4.6535	6.4542	-0.1585	3.6822
Dagum Type I 3P	Shape1.a	1.9573	2.3078	3.9077	4.5784	2.503	3.0067
	Scale	15.5819	21.6406	107.5404	121.2755	32.0267	40.1928
	Shape2.p	0.8977	1.4088	0.5850	0.8037	0.8760	1.3155

Table 8
Bootstrap Parameter Estimates and Confidence Intervals for Each Pollutant across Model Distributions

Pollutant	Distribution	Parameter	Original Estimate	Bias	Std. Error	95% CI
Ozone(O_3)	Burr Type XII 3P	Shape1	2.1627	0.0007	0.075	(2.026, 2.318)
		Shape2	0.419	-0.0014	0.0243	(0.3733, 0.4655)
		Scale	14.7275	0.0183	0.4821	(13.73, 15.63)
	Log-Logistic 3P	Shape	0.4994	0.0032	0.0163	(0.4722, 0.5365)
		Scale	2.9255	-0.0051	0.031	(2.851, 2.975)
		Threshold	1.6377	0.0707	0.3109	(1.168, 2.409)
	Dagum Type I 3P	Shape1	2.1326	0.0068	0.0793	(1.985, 2.310)
		Shape2	1.1532	0.0007	0.1089	(0.972, 1.409)
		Scale	18.6112	0.0767	1.3612	(15.75, 21.17)
PM_{10}	Burr Type XII 3P	Shape1	3.1679	0.005	0.1141	(2.950, 3.418)
		Shape2	0.1916	-0.0002	0.0141	(0.1633, 0.2176)
		Scale	73.4368	0.0087	1.1019	(13.73, 15.63)
	Log-Logistic 3P	Shape	0.2794	-0.0405	0.0266	(0.2004, 0.2857)
		Scale	4.581	0.1499	0.1011	(4.565, 4.878)
		Threshold	1.8152	-15.743	10.8815	(-30.91, 1.92)
	Dagum Type I 3P	Shape1	4.2431	0.0046	0.1999	(3.965, 4.608)
		Shape2	0.6944	0.0146	0.3558	(0.6028, 0.7965)
		Scale	114.408	-0.2609	4.2554	(108.0, 120.5)
$PM_{2.5}$	Burr Type XII 3P	Shape1	2.6883	0.0099	0.1231	(2.476, 2.948)
		Shape2	0.302	0.0005	0.0251	(0.2519, 0.3505)
		Scale	29.1537	0.001	1.1009	(25.73, 29.04)
	Log-Logistic 3P	Shape	0.3689	0.0053	0.0137	(0.3443, 0.3900)
		Scale	3.5732	-0.0308	0.0652	(3.523, 3.644)
		Threshold	1.835	0.0368	0.5021	(0.195, 3.415)
	Dagum Type I 3P	Shape1	2.846	0.0103	0.1328	(2.630, 2.989)
		Shape2	1.138	-0.0326	0.2698	(0.929, 1.326)
		Scale	36.0377	-0.6462	2.9655	(32.27, 40.02)

Table 8 presents bootstrap analysis results, identifying the Burr type XII 3P, Dagum type I 3P and Log-Logistic 3P distributions as the most reliable models for predicting O_3, PM_{10} and $PM_{2.5}$ concentrations. All models exhibit low bias and tight confidence intervals, making them strong choices for air pollutant modeling. The Burr type XII 3P model showed the highest stability for O_3, while the Dagum type I 3P and Log-Logistic 3P models followed closely. For PM_{10}, the Dagum type I 3P model provided the most reliable estimates, though the Log-Logistic 3P showed instability in its threshold parameter.

For $PM_{2.5}$, the Burr type XII 3P performed well, while the Log-Logistic 3P exhibited significant variability. Overall, the Log-Logistic 3P distribution displayed greater instability, particularly in threshold parameters, indicating a need for refinement in PM_{10} and $PM_{2.5}$ modeling. These findings are valuable for selecting effective models in air quality prediction for Visakhapatnam.

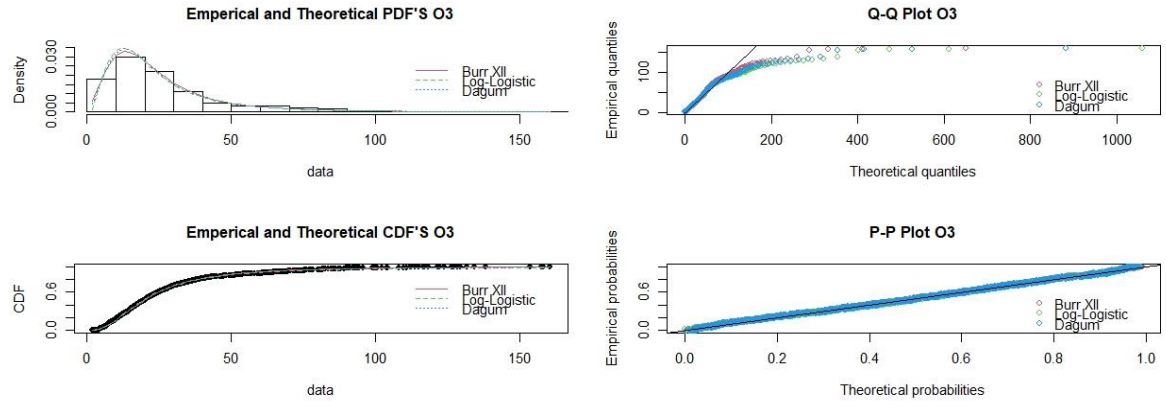

Figure 4: Empirical and Theoretical Comparisons of O_3 Concentrations with Burr Type XII 3P, Log-Logistic 3P and Dagum Type I 3P Distributions

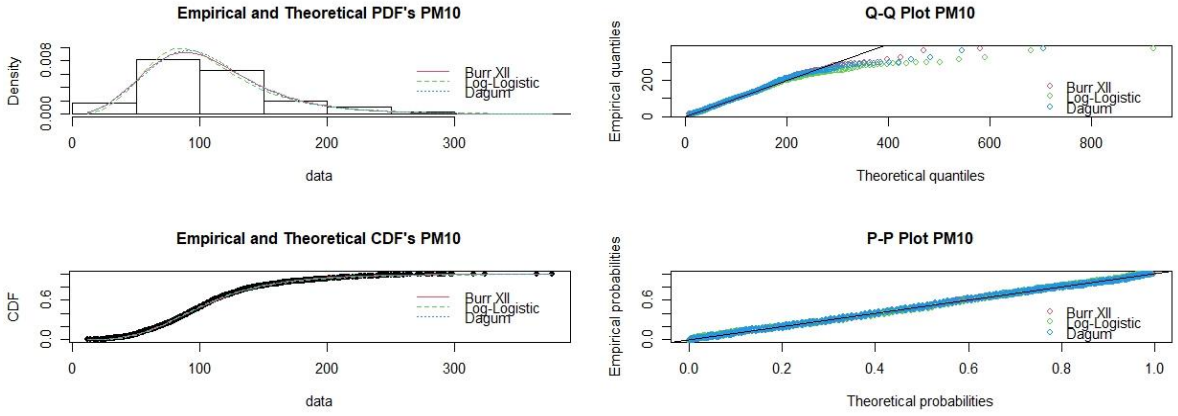

Figure 5: Empirical and Theoretical Comparisons of PM_{10} Concentrations through Burr Type XII 3P, Log-Logistic 3P and Dagum Type I 3P Distributions

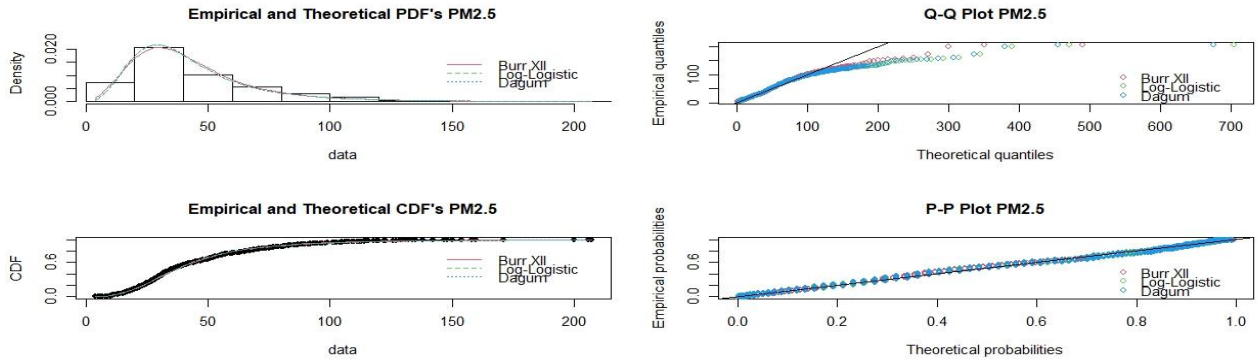

Figure 6: Empirical and Theoretical Comparisons of $PM_{2.5}$ Concentrations Using Burr Type XII 3P, Log-Logistic 3P and Dagum Type I 3P Distributions

Figures 4, 5 and 6 analyze the empirical and theoretical fits of O_3, PM_{10} and $PM_{2.5}$ concentrations using Burr type XII 3P, Log-Logistic 3P and Dagum type I 3P distributions. Across all three figures, the probability density functions (PDFs) indicate that while all models capture the overall data trend, Burr type XII 3P and Dagum type I 3P provide superior accuracy, particularly in the tails. The cumulative distribution functions (CDFs) further confirm this pattern, with these two distributions aligning more closely with empirical data at the extremes. The Q-Q plots show that all models fit central quantiles well but highlight the superior performance of Burr type XII 3P and Dagum type I 3P in capturing extreme values.

Similarly, the P-P plots validate overall fitting accuracy, with these two distributions consistently outperforming Log-Logistic 3P for extreme pollutant concentrations. These findings emphasize the suitability of Burr type XII 3P and Dagum type I 3P distributions for modeling pollutant concentrations, particularly in scenarios requiring precise tail behavior representation, such as environmental risk assessments and air quality modeling. Table 9 analyzes the distributions of air pollutants (O_3, PM_{10} and $PM_{2.5}$) using KDE and various distribution models, revealing distinct distribution characteristics. For O_3, the actual data (bandwidth 3.337) shows a mean of 81.05µg/m³, ranging from -8.11 to 125.63µg/m³. The Burr type XII 3P model (bandwidth 3.154) predicts a higher mean of 309.83µg/m³ and a maximum of 627.81µg/m³, indicating strong right skew.

The Dagum type I 3P model (bandwidth 3.155) offers a lower predicted mean of 438.98µg/m³. In contrast, the Log-Logistic 3P model (bandwidth 3.178) significantly overestimates O_3 concentrations, predicting a median of 1,210.71µg/m³ and a maximum of 2,429.16µg/m³, making it unsuitable. For PM_{10}, empirical data (bandwidth 9.031) reveals a mean of 194µg/m³, ranging from -15.09 to 403.09µg/m³. The Burr type XII 3P model (bandwidth 9.164) predicts a mean of 291.48µg/m³, while the Dagum type I 3P model (bandwidth 8.867) shows a mean of 358.27µg/m³. Both overestimate central tendencies but are more realistic than the Log-Logistic 3P model (bandwidth 9.069), which predicts a median of 732.99 µg/m³ and a maximum of 1,485.41 µg/m³. For $PM_{2.5}$, the actual data (bandwidth 4.898) has a mean of 105.5µg/m³, ranging from -10.69 to 221.69µg/m³. The Burr type XII 3P model (bandwidth 4.468) predicts a mean of 237.50µg/m³ while the Dagum type I 3P model (bandwidth 4.425) yields a higher mean of 337.81µg/m³. The Log-Logistic 3P model (bandwidth 4.436) again inflates the median (647.2µg/m³) and maximum (1,305µg/m³) values, indicating an unsuitable fit. Overall, these results highlight the variability in pollutant distributions across models and the importance of selecting appropriate models for accurate representation.

Figure 7 presents KDE comparisons of observed and predicted concentrations for O_3, PM_{10} and $PM_{2.5}$.

- **O_3 (left panel):** All distributions capture the lower range (0–50µg/m³) well, with Burr type XII 3P and Log-Logistic 3P performing best near the peak. Dagum type I 3P excels in the tail region above 100 µg/m³.
- **PM_{10} (middle panel):** Burr type XII 3P and Dagum type I 3P closely match peak density, while Log-Logistic 3P diverges slightly at mid-range concentrations. Dagum Type I 3P effectively models the right tail above 200µg/m³, indicating its robustness for extremes.
- **$PM_{2.5}$ (right panel):** All distributions align around the peak (20–50µg/m³), with Burr type XII 3P, Dagum type I 3P and Log-Logistic 3P performing better for high concentrations (>100µg/m³).

Figure 8 illustrates residuals for O_3, PM_{10} and $PM_{2.5}$ using the Burr type XII 3P, Log-Logistic 3P and Dagum type I 3P distributions.

- **O_3 (top row):** Burr type XII 3P and Dagum type I 3P show tightly clustered residuals around zero, indicating a good fit, while Log-Logistic 3P shows a slightly wider spread, particularly between indices 500–1000.
- **PM_{10} (middle row):** All distributions perform well, but Burr type XII 3P and Dagum type I 3P exhibit tighter clustering around zero, with Log-Logistic 3P showing larger deviations at higher indices.
- **$PM_{2.5}$ (bottom row):** Burr type XII 3P and Dagum type I 3P maintain close clustering around zero, while Log-Logistic 3P displays more variability and occasional outliers.

Figure 9 displays diagnostic plots for Burr type XII 3P, Log-Logistic 3P and Dagum type I 3P models applied to O_3 concentrations, including residuals vs. predicted values, histograms, Q-Q plots and scale-location plots.

- Residuals are symmetrically scattered around zero, indicating a good model fit.
- Q-Q plots confirm normality, with slight tail deviations.
- Scale-location plots reveal a minor increase in residual spread with higher fitted values, suggesting mild heteroscedasticity.

Figure 10 shows diagnostic plots for PM_{10} concentrations using Burr type XII 3P, Log-Logistic 3P and Dagum type I 3P models.

- **Residuals vs. Predicted Values:** Residuals are randomly distributed around zero, indicating strong model performance.
- **Histograms:** Residuals are tightly centered near zero, suggesting a good fit, though with slight overfitting potential.
- **Q-Q Plots:** Normality is largely confirmed, with minor deviations.
- **Scale-Location Plots:** Slight increases in residual spread indicate mild heteroscedasticity.

These results suggest the models fit PM_{10} concentrations well with minor variability.

Table 9
Summary of KDE for O_3, PM_{10} and $PM_{2.5}$

Air Pollutant	O_3			PM_{10}			$PM_{2.5}$		
Data Type	Actual Data	Burr Type XII 3P	Log-logistic 3P	Dagum Type I 3P	Actual Data	Burr Type XII 3P	Log-logistic 3P	Dagum Type I 3P	Actual Data	Burr Type XII 3P	Log-logistic 3P	Dagum Type I 3P
Band width	3.337	3.154	3.178	3.155	9.031	9.164	9.069	8.867	4.898	4.468	4.436	4.425
Min. x	-8.11	-8.148	-7.748	-8.108	-15.09	-11.58	-19.43	-13.76	-10.69	-9.173	-10.5	-8.871
1st Qu. x	36.47	150.841	601.478	215.435	89.45	139.95	356.78	172.25	47.4	114.162	318.4	164.467
Median x	81.05	309.829	1210.705	438.979	194	291.48	732.99	358.27	105.5	237.498	647.2	337.805
Mean x	81.05	309.829	1210.71	438.979	194	291.48	732.99	358.27	105.5	237.498	647.2	337.805
3rd Qu. x	125.63	468.817	1819.93	662.523	298.55	443.01	1109.2	544.28	163.6	360.833	976.1	511.143
Max. x	170.21	627.805	2429.16	886.067	403.09	594.54	1485.41	730.29	221.69	484.168	1305	684.481
Min. y	$9.74E^{-07}$	$0.00E^{+00}$	$0.00E^{+00}$	$0.00E^{+00}$	$3.12E^{-07}$	$0.00E^{+00}$	$0.00E^{+00}$	$0.00E^{+00}$	$8.77E^{-07}$	$0.00E^{+00}$	$0.00E^{+00}$	$0.00E^{+00}$
1st Qu. y	$3.58E^{-04}$	$0.00E^{+00}$	$0.00E^{+00}$	$0.00E^{+00}$	$6.14E^{-05}$	$2.48E^{-05}$	$0.00E^{+00}$	$0.00E^{+00}$	$1.04E^{-04}$	$3.81E^{-06}$	$0.00E^{+00}$	$0.00E^{+00}$
Median y	$1.80E^{-03}$	$6.32E^{-06}$	$0.00E^{+00}$	$0.00E^{+00}$	$9.45E^{-04}$	$2.03E^{-04}$	$0.00E^{+00}$	$6.66E^{-05}$	$1.21E^{-03}$	$5.41E^{-05}$	$0.00E^{+00}$	$1.50E^{-06}$
Mean y	$5.60E^{-03}$	$1.57E^{-03}$	0.00041	$1.12E^{-03}$	$2.39E^{-03}$	$1.65E^{-03}$	$6.64E^{-04}$	$1.34E^{-03}$	$4.30E^{-03}$	$2.03E^{-03}$	$7.59E^{-04}$	$1.44E^{-03}$
3rd Qu. y	$4.79E^{-03}$	$1.96E^{-04}$	$0.00E^{+00}$	$7.85E^{-05}$	$3.65E^{-03}$	$1.85E^{-03}$	$7.81E^{-05}$	$1.07E^{-03}$	$6.38E^{-03}$	$7.58E^{-04}$	$2.49E^{-05}$	$1.64E^{-04}$
Max. y	$2.99E^{-02}$	$3.03E^{-02}$	0.030415	$3.42E^{-02}$	$9.72E^{-03}$	$9.22E^{-03}$	$9.55E^{-03}$	$1.04E^{-02}$	$2.25E^{-02}$	$1.98E^{-02}$	$2.06E^{-02}$	$2.29E^{-02}$

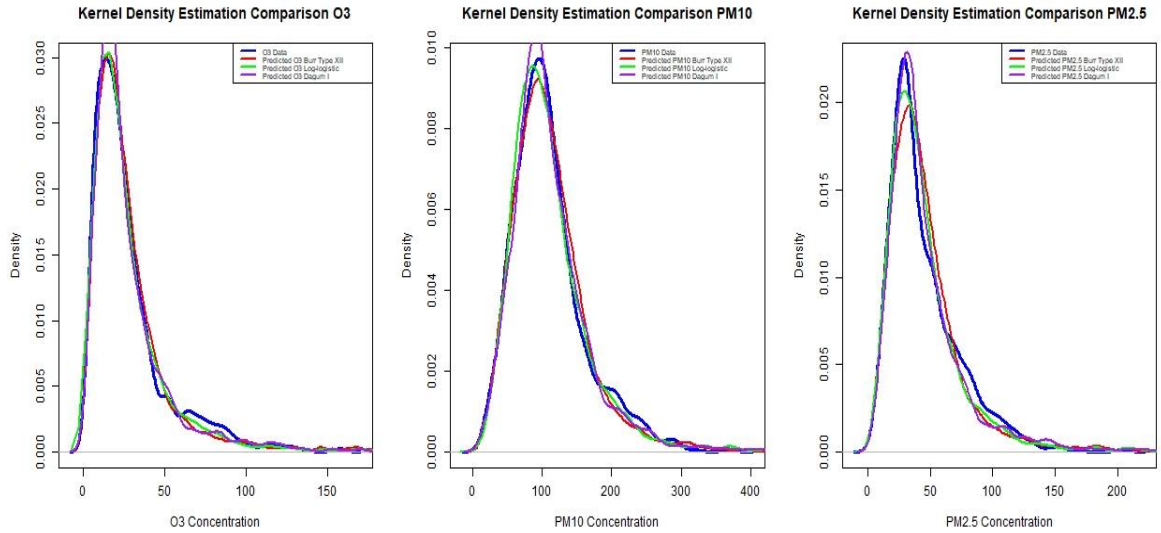

Figure 7: Kernel Density Estimation Comparisons for O_3, PM_{10} and $PM_{2.5}$ vs. Burr Type XII 3P, Log-Logistic 3P and Dagum Type I 3P Distributions

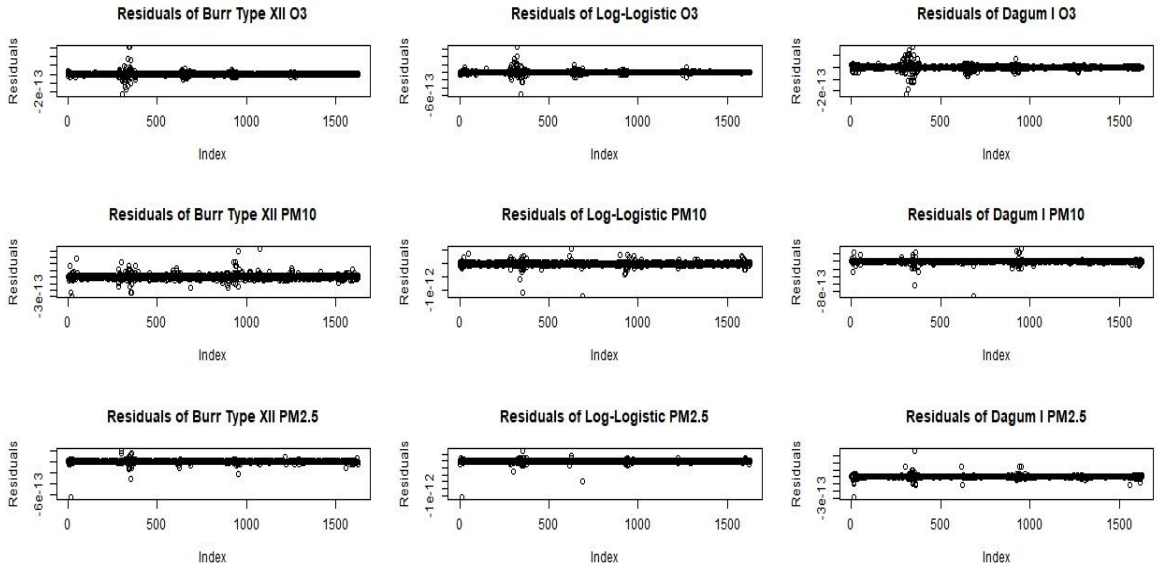

Figure 8: Residual Analysis for O_3, PM_{10} and $PM_{2.5}$ Using the Burr Type XII 3P, Log-Logistic 3P and Dagum Type I 3P Distributions

Figure 11 presents diagnostic plots for $PM_{2.5}$ concentrations using Burr type XII 3P, Log-Logistic 3P and Dagum type I 3P models.

- **Residuals vs. Predicted Values:** Residuals are randomly distributed around zero, showing effective model performance.
- **Histograms:** Residuals are centered near zero, indicating a good fit with minor precision concerns.
- **Q-Q Plots:** Residuals align with normality.
- **Scale-Location Plots:** Slight residual spread increase suggests mild heteroscedasticity.

Goodness-of-Fit-tests and their p-values: Selecting an accurate distribution model for pollutants like O_3, PM_{10} and $PM_{2.5}$ is crucial for reliable predictions. This study compares the three-parameter Burr type XII, Log-Logistic and Dagum type I distributions using K-S, C-VM and A-D tests. The p-values from these tests indicate how well the observed data fits each model, helping to identify the best distribution for each pollutant. Table 10 summarizes these findings.

The analysis presented in table 10 evaluates the suitability of different distributions for modeling air pollutant concentrations. For O_3, the Burr type XII 3P distribution is the best fit, with lower KS (0.0245), CVM (0.1198) and AD (1.7585) statistics and higher p-values (0.2819, 0.4964, 0.1252), indicating a close match with the observed data. For PM_{10}, the Dagum type I 3P distribution shows the strongest fit, with the lowest test statistics (KS=0.0185, CVM=0.0943, AD=0.8719) and the highest p-values (0.6326, 0.6136, 0.4319).

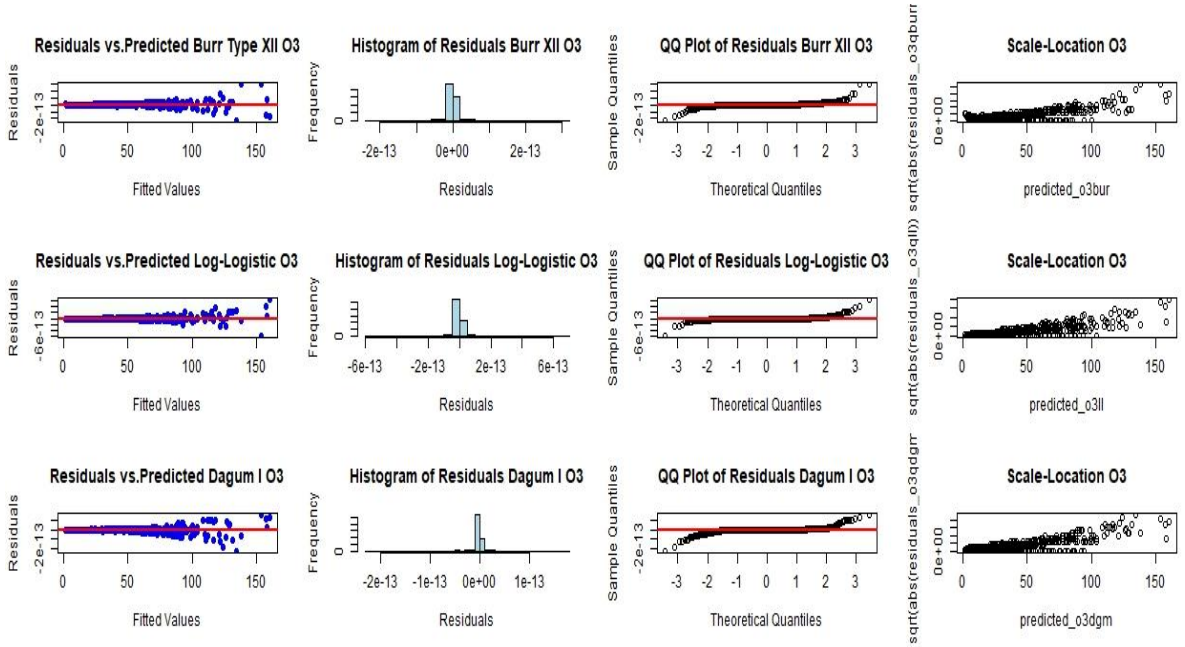

Figure 9: Diagnostic Plots for Burr Type XII 3P, Log-Logistic 3P and Dagum Type I 3P Models for O_3 Concentrations

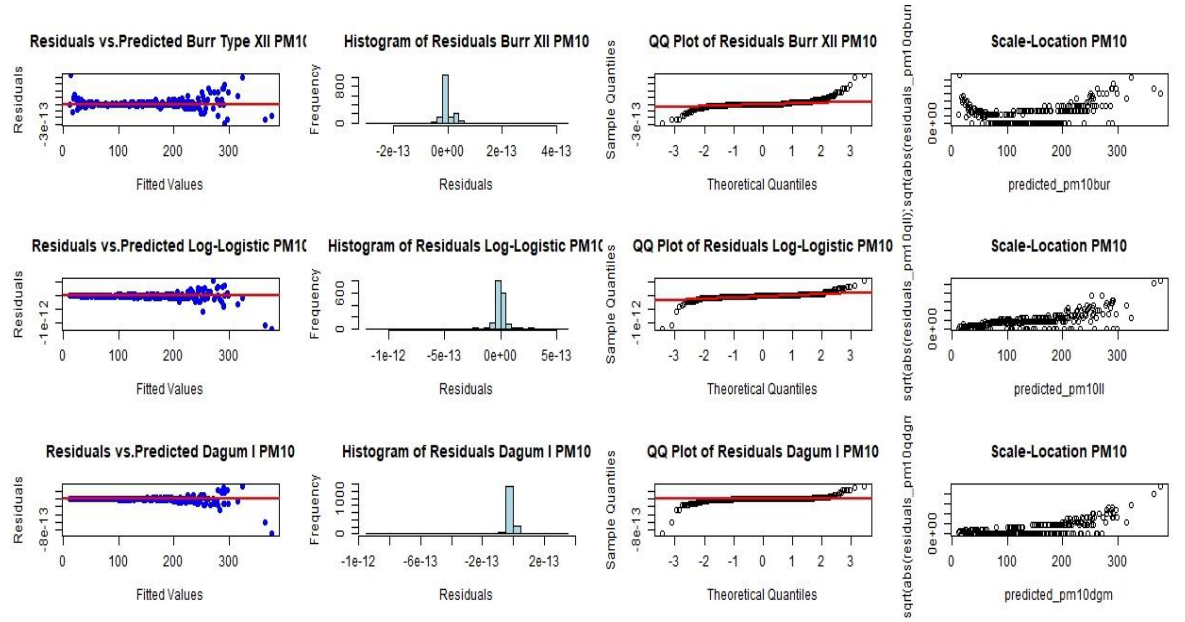

Figure 10: Diagnostic Plots for Burr Type XII 3P, Log-Logistic 3P and Dagum Type I 3P Models for PM_{10} Concentrations

Table 10
Goodness of Fit Statistics and their p-values for O_3, PM_{10} and $PM_{2.5}$ over each distributions

Air Pollutant	Distributions	KS (D)	CVM (W^2)	AD (A^2)	p-value (KS)	p-value (CVM)	p-value (AD)
O_3	Burr Type XII 3P	0.0245	0.1198	1.7585	0.2819	0.4964	0.1252
	Log-Logistic 3P	0.0279	0.1560	1.9490	0.1582	0.3717	0.0980
	Dagum Type I 3P	0.0255	0.1448	1.8230	0.2390	0.4059	0.1152
PM_{10}	Burr Type XII 3P	0.0216	0.1613	1.1536	0.4332	0.3570	0.2859
	Log-Logistic 3P	0.0244	0.1872	1.8744	0.2861	0.2936	0.1078
	Dagum Type I 3P	0.0185	0.0943	0.8719	0.6326	0.6136	0.4319
$PM_{2.5}$	Burr Type XII 3P	0.0491	0.6635	3.8820	0.0008	0.0156	0.0099
	Log-Logistic 3P	0.0399	0.4657	3.0548	0.0113	0.0487	0.0257
	Dagum Type I 3P	0.0402	0.4901	3.1655	0.0103	0.0422	0.0226

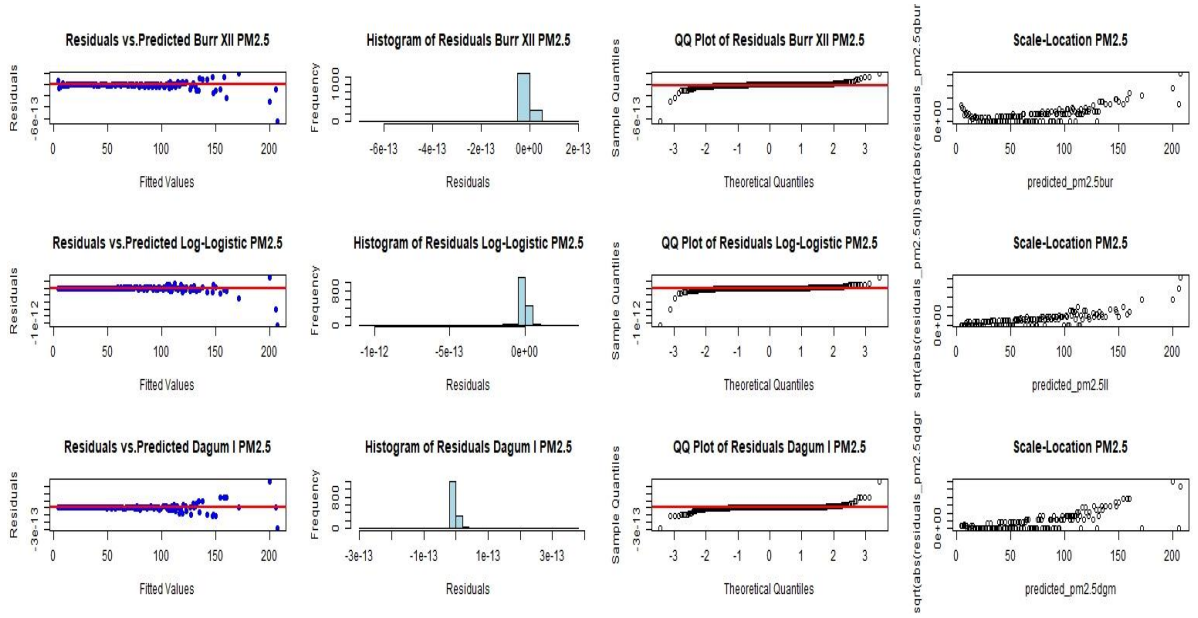

Figure 11: Diagnostic Plots for Burr Type XII 3P, Log-Logistic 3P and Dagum Type I 3P Models for $PM_{2.5}$ Concentrations

Table 11
Model Selection Results for O_3, PM_{10} and $PM_{2.5}$.

Pollutants	Distributions	LL	AIC	BIC	HQIC	CAIC	ABIC
O_3	Burr Type XII 3P	-6843.409	13692.82	13709	13698.82	13712	13709.00
	Log-Logistic 3P	-6835.94	13677.88	13694.06	13683.88	13697.06	13694.07
	Dagum Type I 3P	-6842.995	13691.99	13708.17	13697.99	13711.17	13708.18
PM_{10}	Burr Type XII 3P	-8591.42	17188.84	17205.02	17194.84	17208.02	17205.03
	Log-Logistic 3P	-8603.301	17212.60	17228.79	17218.61	17231.79	17228.79
	Dagum Type I 3P	-8592.983	17191.97	17208.15	17197.97	17211.15	17208.15
$PM_{2.5}$	Burr Type XII 3P	-7422.437	14850.87	14867.06	14856.88	14870.06	14867.06
	Log-Logistic 3P	-7422.652	14851.3	14867.49	14857.31	14870.49	14867.49
	Dagum Type I 3P	-7423.451	14852.90	14869.08	14858.91	14872.08	14869.09

In contrast, for $PM_{2.5}$, the Log-Logistic 3P distribution, despite showing significant deviations (KS (0.0399) p-value: 0.0113, CVM (0.4657) p-value: 0.0487, AD (3.0548) p-value: 0.0257), exhibited relatively lowest test statistics and higher p-values compared to the Burr type XII 3P model, indicating a better fit overall. These results indicate that alternative customizing approaches may be necessary for $PM_{2.5}$. Overall, the findings emphasize the importance of selecting appropriate distributions based on rigorous statistical tests to achieve accurate air quality pattern.

Model Selection Criteria: The optimal distribution for O_3, PM_{10} and $PM_{2.5}$ concentrations is identified using model selection criteria assessing both fit accuracy and model complexity. This study compares three-parameter distributions-Burr type XII 3P, Log-Logistic 3P and Dagum type I 3P based on metrics such as Log-Likelihood (LL), AIC and BIC and other metrics as shown in table 11. These criteria guide the selection of the most balanced model for each pollutant[17].

Table 11 highlights the best-fitting distributions for air pollutant concentrations based on selection criteria:

- **O_3:** Log-Logistic 3P, with the lowest AIC (13677.88), BIC (13694.06) and other metrics, offers the best balance of fit and complexity.
- **PM_{10}:** Burr type XII 3P is optimal, showing the lowest AIC (17188.84), BIC (17205.02) and related values, indicating the best fit.
- **$PM_{2.5}$:** Burr type XII 3P again proves most suitable, with minimum AIC (14850.87), BIC (14867.06) and others, accurately capturing $PM_{2.5}$ data.

These results underscore the importance of model selection criteria in reliable modeling pollutant concentrations.

Model Error Metrics: Table 12 provides comprehensive selection metrics for evaluating the fit of Burr Type XII 3P, Log-Logistic 3P and Dagum type I 3P distributions to model O_3, PM_{10} and $PM_{2.5}$ concentrations. Metrics include MBE, MAE, MSE, MdAE, MAPE, RMSE, NRMSE, CV-RMSE and R^2, all of which assess model accuracy and predictive capability. These metrics collectively offer detailed insights into each model's performance, facilitating precise distribution selection for pollutant concentration data[20].

Table 12 highlights distinct distribution preferences for modeling air pollutants:

- **O_3:** The Burr type XII 3P distribution shows the lowest errors (MBE: -1.3e-15, MAE: 5.6e-15, MSE: 3.4e-28) and high consistency, despite a low R^2, indicating an optimal fit for O_3 data.
- **PM_{10}:** The Dagum type I 3P distribution performs best with minimal errors (MBE: -2.7e-15, MAE: 1.6e-14, MSE: 1.2e-27) and reliable metrics, proving suitable for PM_{10}.
- **$PM_{2.5}$:** Dagum type I 3P provides the most accurate fit overall for $PM_{2.5}$, despite the Log-Logistic 3P models lower RMSE and NRMSE in some aspects.

These findings underscore the importance of using multiple metrics to determine the best-fitting model for accurate air quality analysis: Burr type XII 3P for O_3, Dagum type I 3P for PM_{10} and Dagum type I 3P for $PM_{2.5}$.

Model Selection with Cross-Validation: The cross-validation process evaluated the performance of the Burr type XII 3P, Dagum type I 3P and Log-Logistic 3P distributions for O_3, PM_{10} and $PM_{2.5}$. Each dataset was split into 70% for training and 30% for testing to assess model fit and predictive accuracy. Performance was measured using metrics such as Log-Likelihood (L-L), AIC, BIC, MSE, MAE, RMSE, MedAE, MAPE, R^2 and goodness-of-fit tests (KS, CVM and AD Tests), as displayed in table 13.

Table 12
Model Error Metrics for O_3, PM_{10} and $PM_{2.5}$ vs. Each Model Distribution

Model Metrics	O_3			PM_{10}			$PM_{2.5}$		
	Burr Type XII 3P	Log-Logistic 3P	Dagum Type I 3P	Burr Type XII 3P	Log-Logistic 3P	Dagum Type I 3P	Burr Type XII 3P	Log-Logistic 3P	Dagum Type I 3P
MBE	$-1.3e^{-15}$	$-8.7e^{-16}$	$6.6e^{-16}$	$-2.7e^{-15}$	$-2.2e^{-18}$	$2.6e^{-15}$	$2.4e^{-15}$	$1.4e^{-15}$	$-7.9e^{-16}$
MAE	$5.6e^{-15}$	$9.0e^{-15}$	$5.3e^{-15}$	$1.6e^{-14}$	$3.2e^{-14}$	$9.4e^{-15}$	$7.1e^{-15}$	$9.3e^{-15}$	$5.1e^{-15}$
MSE	$3.4e^{-28}$	$1.2e^{-27}$	$2.7e^{-28}$	$1.2e^{-27}$	$4.5e^{-27}$	$1.5e^{-27}$	$5.9e^{-28}$	$1.3e^{-27}$	$3.2e^{-28}$
MdAE	$1.8e^{-15}$	$1.8e^{-15}$	$8.9e^{-16}$	$7.1e^{-15}$	$2.8e^{-14}$	$0.0e^{+00}$	$3.6e^{-15}$	$3.6e^{-15}$	$0.0e^{+00}$
MAPE	$1.8e^{-14}$	$1.7e^{-14}$	$1.1e^{-14}$	$2.2e^{-14}$	$2.5e^{-14}$	$5.6e^{-15}$	$1.7e^{-14}$	$1.4e^{-14}$	$6.6e^{-15}$
RMSE	$1.8e^{-14}$	$3.5e^{-14}$	$1.7e^{-14}$	$3.5e^{-14}$	$6.7e^{-14}$	$3.9e^{-14}$	$2.4e^{-14}$	$3.6e^{-14}$	$1.8e^{-14}$
NRMSE	$1.2e^{-16}$	$2.2e^{-16}$	$1.0e^{-16}$	$9.5e^{-17}$	$1.8e^{-16}$	$1.1e^{-16}$	$1.2e^{-16}$	$1.8e^{-16}$	$8.8e^{-17}$
CV-RMSE	$6.6e^{-16}$	$1.3e^{-15}$	$6.0e^{-16}$	$3.1e^{-16}$	$6.1e^{-16}$	$3.6e^{-16}$	$5.3e^{-16}$	$7.9e^{-16}$	$4.0e^{-16}$
R^2	$-1.6e^{+30}$	$-4.5e^{+29}$	$-2.0e^{+30}$	$-2.3e^{+30}$	$-6.2e^{+29}$	$-1.8e^{+30}$	$-1.3e^{+30}$	$-6.1e^{+29}$	$-2.5e^{+30}$

Table 13
Cross-Validation Summary for O_3, PM_{10}, $PM_{2.5}$ across Each Distribution

Model Metric	O_3			PM_{10}			$PM_{2.5}$		
	Burr Type XII 3P	Dagum Type I 3P	Log-Logistic 3P	Burr Type XII 3P	Dagum Type I 3P	Log-Logistic 3P	Burr Type XII 3P	Dagum Type I 3P	Log-Logistic 3P
L-L	-4768.32	-4769.32	-4764.12	-5992.97	-5994.64	-5997.70	-5145.85	-5147.52	-5146.99
AIC	9542.65	9544.64	9534.25	11991.94	11995.28	12001.40	10297.71	10301.04	10299.99
BIC	9557.76	9559.75	9549.36	12007.05	12010.39	12016.52	10312.82	10316.15	10315.10
MSE	$1.94e^{-06}$	$2.91e^{-06}$	$4.25e^{-06}$	$1.64e^{-07}$	$1.21e^{-07}$	$1.41e^{-07}$	$2.17\times e^{-6}$	$2.43\times e^{-6}$	$2.41\times e^{-6}$
MAE	0.0009	0.0011	0.0012	0.0003	0.0003	0.0003	0.0009	0.0010	0.0011
RMSE	0.0014	0.0017	0.0021	0.0004	0.0004	0.0004	0.0015	0.0016	0.0016
MedAE	0.00054	0.00053	0.00051	0.0003	0.0002	0.0002	0.0006	0.0008	0.0008
MAPE (%)	64.600	73.270	80.660	108.462	128.254	140.802	37.406	42.177	43.476
R^2	0.9755	0.9632	0.9463	0.9823	0.9869	0.9848	0.9340	0.9260	0.9260
KS Test (p-value)	0.0457 (0.2590)	0.0549 (0.1049)	0.0608 (0.0535)	0.0477 (0.2166)	0.0438 (0.3058)	0.0479 (0.2109)	0.0844 (0.0019)	0.0825 (0.0026)	0.0804 (0.0036)
CVM Test (p-value)	0.3067 (0.1295)	0.3761 (0.0836)	0.4231 (0.0628)	0.2763 (0.1578)	0.2489 (0.1898)	0.2643 (0.1711)	0.7303 (0.0107)	0.6953 (0.0130)	0.6585 (0.0160)
AD Test (p-value)	2.4952 (0.0498)	2.3502 (0.0595)	2.2969 (0.0635)	2.0885 (0.0822)	1.8695 (0.1085)	1.7259 (0.1307)	4.9419 (0.0030)	4.3918 (0.0056)	4.1107 (0.0077)

Table 14
Performance Metrics and Goodness-of-Fit Tests for O_3, PM_{10} and $PM_{2.5}$ using 10-Fold Cross-Validation

Average Metrics	O_3			PM_{10}			$PM_{2.5}$		
	Burr Type XII 3P	Dagum Type I 3P	Log-Logistic 3P	Burr Type XII 3P	Dagum Type I 3P	Log-Logistic 3P	Burr Type XII 3P	Dagum Type I 3P	Log-Logistic 3P
MSE	$8.23e^{-05}$	$8.38e^{-05}$	$7.98e^{-05}$	$1.47e^{-06}$	$1.45e^{-06}$	$1.56e^{-06}$	$6.40\times e^{-6}$	$6.51\times e^{-6}$	$6.45\times e^{-6}$
MAE	0.0051	0.0052	0.0051	0.0009	0.0009	0.0009	0.0017	0.0017	0.0017
RMSE	0.0068	0.0070	0.0069	0.0012	0.0012	0.0012	0.0024	0.0024	0.0024
MAPE (%)	56.4207	55.2689	54.0315	8618	9864	1064	6.38×10^{12}	7.53×10^{12}	7.72×10^{12}
R^2	0.6272	0.5957	0.5953	0.8117	0.8058	0.7870	0.7788	0.7661	0.7681
MedAE	0.0037	0.0037	0.0036	0.0007	0.0007	0.0007	0.0012	0.0012	0.0012
L-L	-6156.37	-6155.80	-6149.89	-7731.36	-7732.77	-7737.26	-6679.44	-6680.42	-6679.81
AIC	12318.75	12317.6	12305.78	15468.71	15471.54	15480.52	13364.88	13366.84	13365.62
BIC	12334.62	12333.47	12321.65	15484.58	15487.41	15496.38	13380.75	13382.71	13381.49
KS	0.2369	0.2372	0.2345	0.1407	0.1401	0.1398	0.1529	0.1484	0.1481
CVM	4.7967	4.8325	4.7802	1.1815	1.1744	1.1777	1.1390	1.1018	1.1017
AD	25.7533	25.9662	25.6969	7.0657	6.9741	7.0152	6.7643	6.5188	6.5155

Table 13 evaluates the performance of Burr Type XII 3P, Dagum Type I 3P and Log-Logistic 3P distributions for O_3, PM_{10} and $PM_{2.5}$:

- **O_3:** The Log-Logistic 3P model achieves the highest Log-Likelihood (LL: -4764.12) and the lowest AIC (9534.25) and BIC (9549.36), indicating a good fit. However, the Burr type XII 3P model excels in prediction accuracy with the lowest MSE (1.94e-06), MAE (0.0009), RMSE (0.0014) and the highest R^2 (0.9755), making it the most suitable distribution for O_3.
- **PM_{10}:** The Dagum type I 3P model offers the best fit, with the lowest AIC (11995.28) and BIC (12010.39) and the highest R^2 (0.9869). It also shows the lowest MSE (1.21e-07), MAE (0.0003) and RMSE (0.0004), confirming its optimal choice for PM_{10}.
- **$PM_{2.5}$:** The Burr type XII 3P distribution provides the best overall performance, with the lowest AIC (10297.71), BIC (10312.82), MSE (2.17e−6), MAE (0.0009) and RMSE (0.0015). Despite similar R^2 values for Dagum type I 3P and Log-Logistic 3P (0.9260), Burr type XII 3P outperforms in goodness-of-fit tests. Overall, Burr type XII 3P is preferred for O_3 and $PM_{2.5}$; while Dagum type I 3P is optimal for PM_{10}.

Table 14 presents the 10-fold cross-validation results for O_3, PM_{10} and $PM_{2.5}$, revealing distinct model performance patterns:

- **O_3:** The Log-Logistic 3P distribution emerges as the best model for predictive accuracy, with the lowest MSE (7.98e-05), MAE (0.0051) and RMSE (0.0069). Although Log-Likelihood (-6149.89) is higher and AIC (12305.78) and BIC (12321.65) values are lower than those of the Burr type XII 3P distribution, its superior predictive accuracy makes it the preferred choice.
- **PM_{10}:** The Burr type XII 3P distribution is most effective, achieving the lowest MSE (1.47e-06), MAE (0.0009) and RMSE (0.0012), along with the highest R^2 (0.8117). It also excels in Log-Likelihood (-7731.36), AIC (15468.71) and BIC (15484.58) metrics, supported by strong goodness-of-fit results in KS (0.1407) and CVM (1.1815) tests.
- **$PM_{2.5}$:** The Burr type XII 3P distribution is again the most suitable model, with the lowest MSE (6.40e-6), MAE (0.0017) and RMSE (0.0024), alongside the highest R^2 (0.7788). It performs well in Log-Likelihood (-6679.44), AIC (13364.88) and BIC (13380.75), further confirmed by other statistical measures.

Overall, while the Log-Logistic 3P distribution is optimal for O_3 due to its predictive accuracy, the Burr type XII 3P distribution consistently outperforms other models for PM_{10} and $PM_{2.5}$ across multiple metrics.

Conclusion

This research evaluated the suitability of three flexible probability distributions-Burr type XII 3P, Dagum type I 3P and Log-Logistic 3P-for modeling the concentrations of secondary pollutants O_3, PM_{10} and $PM_{2.5}$ in Visakhapatnam, a rapidly industrializing city in India. Using robust statistical techniques, the analysis highlights that the Burr type XII 3P distribution is the most effective in capturing the heavy-tailed distribution of O_3 concentrations.

The Dagum type I 3P proves optimal for PM_{10} concentrations due to its ability to model moderate variability and skewness, while both Burr type XII 3P and Log-Logistic 3P distributions provided reliable models for $PM_{2.5}$. These models offer a valuable basis for improving air quality assessments and targeted pollution control in similar mid-sized cities. Future studies could apply these models to other urban areas and to explore seasonal variations to further enhance air quality predictions. This work supports environmental management efforts to address the public health challenges posed by pollution in growing urban regions.

References

1. Abdullah A.M., Ismail M., Yuen F.S., Abdullah S. and Elhadi R.E., The Relationship between Daily Maximum Temperature and Daily Maximum Ground Level Ozone Concentration, *Polish Journal of Environmental Studies*, **26(3)**, 517-523 **(2017)**

2. Ahmat H., Yahaya A.S. and Ramli N.A., Prediction of PM_{10} Extreme Concentrations in Urban Monitoring Stations in Selangor, Malaysia Using Three Parameters Extreme Value Distributions (EVD), *Jurnal Teknologi*, **77(32)**, 37-46 **(2015)**

3. Benjamin S.M., Humberto V.H. and Barry C.A., Use of the Dagum Distribution for Modeling Tropospheric Ozone Levels, *Journal of Environmental Statistics*, **5(5)**, 1-11 **(2013)**

4. Bhandari P.S., Use of Lognormal Distribution Model to Fit the Observed Data for Air Pollutants Concentration, *International Journal of Multidisciplinary and Current Research*, **8**, 932-936 **(2020)**

5. Binbusayyis A., Khan M.A. and Emmanuel W.R., A Deep Learning Approach for Prediction of Air Quality Index in Smart City, *Discover Sustainability*, **5(1)**, 1-23 **(2024)**

6. Choopradit B., Paitoon R., Srinuan N. and Kwankaew S., Application of the Parametric Bootstrap Method for Confidence Interval Estimation and Statistical Analysis of $PM_{2.5}$ in Bangkok, *WSEAS Transactions on Environment and Development*, **20**, 215-225 **(2024)**

7. Chaudhary A.K., Telee L.B.S. and Kumar V., Extended Kumaraswamy Exponential Distribution with Application to COVID-19 Dataset, *Journal of Nepal Mathematical Society*, **6(1)**, 1-10 **(2023)**

8. El-Shanshoury G.I., Fitting the Probability Distribution Functions to Model Particulate Matter Concentrations, *Arab Journal of Nuclear Sciences and Applications*, **50**, 108-122 **(2017)**

9. Febriantikasari E., Adnan A., Yendra R. and Muhaijir M.N., Using Dagum Distribution to Simulate Concentration of PM_{10} in Pekanbaru City, Indonesia, *Applied Mathematical Sciences*, **13(9)**, 439-448 **(2019)**

10. Gavriil I., Grivas G., Kassomenos P., Chaloulakou A. and Spyrellis N., An Application of Theoretical Probability Distributions to the Study of PM_{10} and $PM_{2.5}$ Time Series in Athens, Greece, *Global NEST Journal*, **8(3)**, 241-251 **(2006)**

11. Jaffar M.I., Hamid H.A., Yunus R. and Raffee A.F., Fitting Statistical Distribution Functions of Air Pollutant Concentration in Different Urban Locations in Malaysia, *International Journal of Environmental Research*, **7(2)**, 07-11 **(2023)**

12. Jaffar M.I., Hamid H.A., Yunus R. and Raffee A.F., Fitting Statistical Distribution on Air Pollution: An Overview, *International Journal of Engineering & Technology*, **7(3.23)**, 40-44 **(2018)**

13. Kumar SR., Rao C.R., Rao C.P. and Ramani V.S., Seasonal Trend Analysis of Major Air Pollutant ($PM_{2.5}$ and PM_{10}) Concentration in Visakhapatnam During 2020–2022: A Case Study, *Indian Journal of Science and Technology*, **17(3)**, 258-269 **(2024)**

14. López-Rodríguez F., García-Sanz-Calcedo J., Moral-García F.J. and García-Conde A.J., Statistical Study of Rainfall Control: The Dagum Distribution and Applicability to the Southwest of Spain, *Water*, **11(3)**, 453 **(2019)**

15. Noor N.M., Tan C.Y., Ramli N.A., Yahaya A.S. and Yusof N.F.F.M., Assessment of Various Probability Distributions to Model PM_{10} Concentration for Industrialized Area in Peninsula Malaysia: A Case Study in Shah Alam and Nilai, *Australian Journal of Basic and Applied Sciences*, **5(12)**, 2796-2811 **(2011)**

16. Omar M.U., Assessment of Parameter Estimator of Lognormal Distribution for a Better Long-Term Prediction of PM_{10} Concentrations in Suburban Areas, *EnvironmentAsia*, **15(2)**, 23-33 **(2022)**

17. Sherwani R.A.K., Ashraf S., Abbas S. and Aslam M., Marshall Olkin Exponentiated Dagum Distribution: Properties and Applications, *Journal of Statistical Theory and Applications*, **22(1)**, 70-97 **(2023)**

18. Souza A. et al, Determination of the Best Probability Distribution of Fit for Ozone Concentration Data in Campo Grande-MS-Brazil, *Eur Chem Bull*, **8**, 291-300 **(2019)**

19. Thupeng W.M., Use of the Three-Parameter Burr XII Distribution for Modeling Ambient Daily Maximum Nitrogen Dioxide Concentrations in the Gaborone Fire Brigade, *American Scientific Research Journal for Engineering, Technology and Sciences (ASRJETS)*, **26(2)**, 18-32 **(2016)**

20. Tu T., Su Y. and Ren S., FC-MIDTR-WCCA: A Machine Learning Framework for PM2.5 Predictions, *IAENG International Journal of Computer Science*, **51(5)**, 544-552 **(2024)**

21. Warsono W., Antonio Y., Yuwono S., Kurniasari D., Suroso E., Yushananta P. and Hadi S., Modeling Generalized Statistical Distributions of PM2.5 Concentrations During the COVID-19 Pandemic in Jakarta, Indonesia, *Decision Science Letters*, **10(3)**, 393-400 **(2021)**

22. Wang Y., Huang L., Huang C., Hu J. and Wang M., High-Resolution Modeling for Criteria Air Pollutants and the Associated Air Quality Index in a Metropolitan City, *Environment International*, **172**, 107752 **(2023)**.

(Received 19th December 2024, accepted 21st January 2025)