Received Date: Sept. 2024 Revised: Nov. 2024 Accepted: Dec. 2024

Modeling PM2.5 and TSP Concentrations: A Comparison of Weibull, Lognormal, and Gamma Distributions

Nanda Kumar Tharu1*, Subita Baidhya2
1* Central Department of Biotechnology, Tribhuvan University, Kirtipur, Nepal
1 Tri Chandra Multiple Campus, Tribhuvan University, Ghantagahr, Nepal
2 Patan Multiple Campus, Tribhuvan University, Pulchowk, Nepal
Correspondence: nanda.tharu@trc.tu.edu.np

Doi: https://doi.org/10.3126/ppj.v4i2.79165

Abstract

This study evaluates the efficacy of the Weibull, lognormal, and gamma distributions in modeling PM2.5 and TSP concentrations in urban Kathmandu, specifically at Pulchok and Ratnapark, based on data collected from May 1, 2024 to October 31, 2024. The analysis revealed notable disparities in pollution levels, with Pulchok exhibiting higher and more fluctuating concentrations compared to Ratnapark. Both the Weibull and lognormal distributions demonstrated superior fit for the PM2.5 and TSP data, as confirmed by the Kolmogorov-Smirnov test. These decisions highlight the significant influence of urban pollution sources and highly the importance of robust statistical modeling in air quality management and health risk assessment.

Keywords: Air pollution, PM 2.5, TSP concentrations, Weibull distribution, Lognormal distribution, Kolmogorov-Smirnov test

Introduction

Air pollution is a significant environmental health issue worldwide, with particulate matter (PM) such as PM2.5 and total suspended particles (TSP) serving as critical indicators of air quality. Accurate modeling of these pollutants is essential for understanding their impact on public health and for implementing effective control strategies. Among the statistical models commonly used for this purpose, the Weibull, lognormal, and gamma distributions are particularly favored for their ability to accommodate the diverse patterns seen in pollutant concentration data (Cao et al., 2018; Zhang et al., 2020). These models are especially valuable in environmental research, where pollutant concentrations often display skewed distributions and extreme values.

The Weibull distribution is valued for its versatility in modeling data with various shapes, making it suitable for a broad range of pollution datasets, especially those exhibiting skewness and extreme events (Zhou et al., 2019). Similarly, the lognormal distribution is frequently used to model particulate matter concentrations, as it effectively captures data that follows a multiplicative process (Liu et al., 2021). Additionally, the gamma distribution, commonly applied in reliability studies to model event timing, is also useful in air quality modeling because of its ability to account for the variability and skewness often seen in pollutant concentration data (Huang et al., 2017).

In their 2019 study on air pollution in urban India, Ghosh and Chaudhuri utilized Weibull, lognormal, and gamma distributions in conjunction with extreme value theory (EVT) to model PM10 and PM2.5 concentrations. They discovered that the Weibull distribution provided the best fit for the data, particularly for PM2.5, while EVT played a crucial role in forecasting extreme pollution events, which are vital for public health (Cao et al., 2018). Their approach highlights the significance of modeling extreme events to improve air quality management and inform policy decisions (Zhang et al., 2020).

This study seeks to assess the goodness-of-fit of the Weibull, lognormal, and gamma distributions in modeling PM2.5 and TSP concentrations at a urban location Ratnapark. By applying statistical tools such as the Kolmogorov-Smirnov test, we evaluate the suitability of each distribution in capturing the characteristics of the pollutant concentration data. The results offered valuable understandings into the most appropriate distribution models for air quality evaluations in urban settings.

Materials and Methods

Daily average PM2.5 and TSP concentrations in the urban areas of Kathmandu were obtained from the Kathmandu Environment Department's database, covering the period from March 1, 2024, to October 31, 2024. Two strategically selected sites, representing different functional zones of Kathmandu, were chosen for monitoring. These sites were considered to provide a comprehensive overview of the city's air pollution levels. The dataset was complete, with no missing values recorded for the specified variables. The study briefly outlines several fitted distributions used to model the pollution data, as follows:

Weibull Distribution: Let X denote a random variable; the two-parameter Weibull density function is given by

$$f(x;\alpha,\beta)=\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}e^{-\left(x/\beta\right)^{\alpha}}\quad;\qquad \alpha>0,\ \beta>0$$

where α = Shape parameter, β = Scale parameter

Log-normal Distribution: A random variable X follows a log-normal distribution if $\ln(X)$ is normally distributed. Its probability density function is expressed as:

$$f\left(x;\mu,\sigma^2\right)=\frac{1}{x\sqrt{2\pi\sigma^2}}e^{-(\ln x-\mu)^2\big/2\sigma^2}; \qquad x>0,\ -\infty<\mu<\infty,\ \sigma^2>0$$

where, μ = location parameter as well the mean of the distribution and
σ = scale parameter as well the standard deviation of the distribution.

Gamma Distribution: Let X denote a random variable, the two-parameter gamma density function with parameters α and β is specified by

$$f\left(x;\sigma,\beta\right)=\frac{1}{\Gamma(\alpha)\beta^{x}}x^{\alpha-1}e^{-x\!/\!\beta}; \qquad x>0,\ \sigma>0,\ \beta>0$$

where, α = shape parameter and β = scale parameter.

Methods of Parameter Estimation

The parameters of the distributions can be estimated through several techniques, including maximum likelihood estimation (MLE) and method of moments (MOM), among others. In this study, the maximum likelihood estimation method was employed due to its widespread use and its ability to provide estimates with minimum variance.

Overall Fitting distribution and evaluation

In this process, we begin by hypothesizing a family of distributions that might best describe the data, such as Normal, Weibull, Lognormal, or Gamma, based on the data's characteristics. Once a potential distribution is chosen, we estimate the distribution's parameters using methods as maximum likelihood estimation (MLE). After parameter estimation, the quality of fit is evaluated by performing a Kolmogorov-Smirnov (K-S) test. This statistical test is used to determine how well the chosen distribution fits the observed data, with a high p-value indicating that the data fits the distribution well, while a low p-value suggests a poor fit.

RESULTS AND DISCUSSION

This study analyzed hourly PM 2.5 and TSP concentrations at Pulchok and Ratnapark using a combination of advanced statistical tools and distributional modeling techniques. By employing these methods, the research aimed to capture the variability, skewness, and distributional patterns of pollutant concentrations, providing a comprehensive understanding of air quality dynamics in these urban locations.

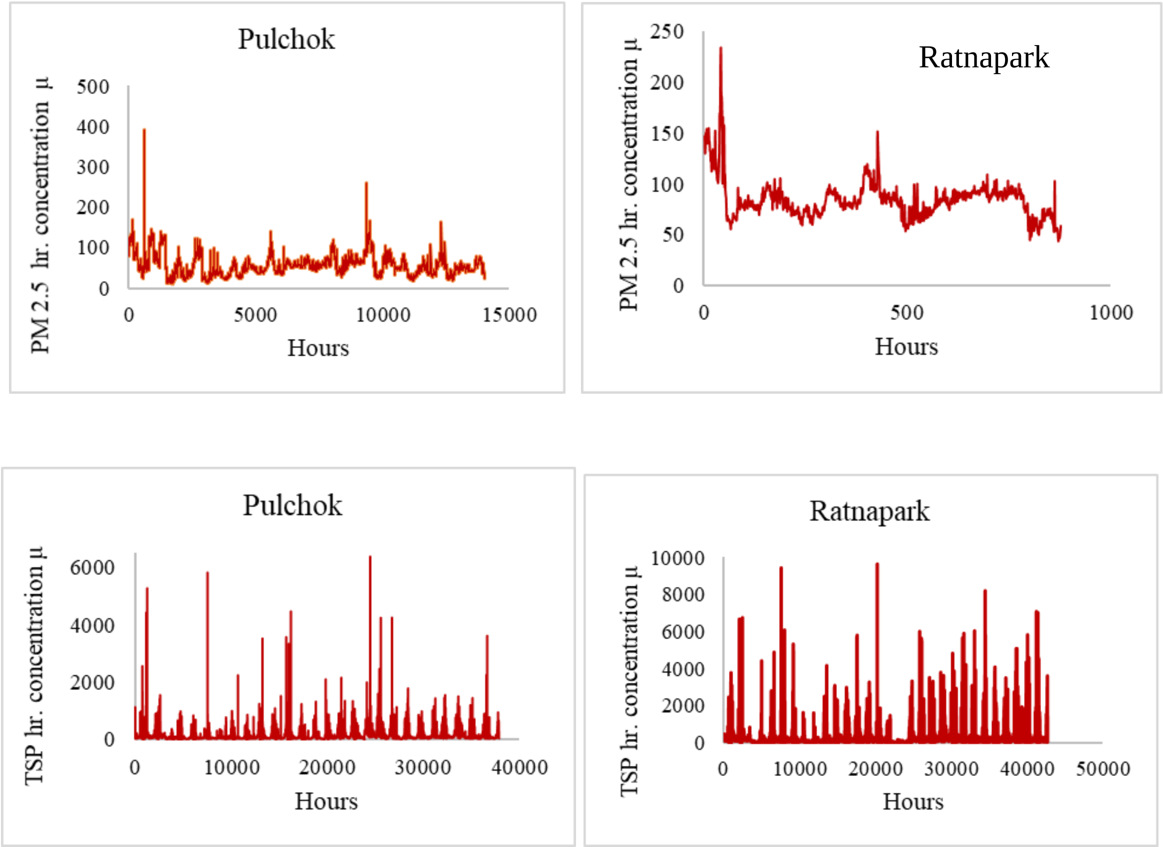


Figure 1. Ambient air concentration PM2.5 and TSP

The graphs compare hourly PM2.5 concentrations at Pulchok and Ratnapark, highlighting key differences in pollution levels and trends. Pulchok shows a wider range of concentrations, peaking up to 400-500 µg/m³, with frequent and pronounced spikes over a longer time frame (15,000 hours). In contrast, Ratnapark exhibits lower peaks (200-250 µg/m³) and more stable trends over a shorter period (1,000 hours).

Pulchok's higher baseline levels (50-100 µg/m³) and greater variability suggest significant pollution sources like traffic or industry. Ratnapark, with lower baseline levels (around 50 µg/m³) and fewer extreme spikes, indicates comparatively better air quality. Overall, Pulchok experiences more severe and variable pollution episodes than Ratnapark.

Table 1. Descriptive statistics of air concentration PM 2.5 and TSP

Station		N	Mean	S.D.	Minimum	P $_{25}$	Median	P $_{75}$	Maximum
Pulchok	PM 2.5	75564	51.803	25.184	4.30	35.60	47.30	63.60	653.20
Ratnapark	PM 2.5	75564	90.057	43.551	4.60	59.20	84.50	115.40	391.40
Pulchok	TSP	37,782	133.931	150.147	4.30	57.60	89.80	160.00	6389.70
Ratnapark	TSP	37,782	168.951	280.029	4.60	75.80	112.00	167.20	9648.30

The table 1 presents the descriptive statistics for PM2.5 and TSP concentrations at Pulchok and Ratnapark. For PM2.5, Pulchok shows a mean concentration of 51.803 µg/m³ with a standard deviation of 25.184 µg/m³, indicating moderate pollution with notable variability. The minimum concentration is 4.30 µg/m³, and the maximum is 653.20 µg/m³, reflecting occasional extreme pollution spikes. In comparison, Ratnapark has a higher mean concentration of 90.057 µg/m³ and a standard deviation of 43.551 µg/m³, indicating more severe pollution with greater fluctuation. Its minimum value is 4.60 µg/m³, and the maximum is 391.40 µg/m³, suggesting less extreme but more persistent pollution levels compared to Pulchok.

For TSP, Pulchok has a mean concentration of 133.931 µg/m³ with a standard deviation of 150.147 µg/m³, reflecting significant variability and occasional very high TSP levels (maximum of 6389.70 µg/m³). Ratnapark shows a higher mean concentration of 168.951 µg/m³ and a much larger standard deviation of 280.029 µg/m³, with extreme spikes (maximum of 9648.30 µg/m³), indicating higher and more erratic pollution episodes compared to Pulchok. Overall, Ratnapark experiences more severe and variable air pollution, especially in terms of TSP concentrations, than Pulchok, suggesting different sources or patterns of pollution at the two locations.

Distributional Models

In this study, the Weibull, lognormal, and gamma distributions were fitted to the PM 2.5 and TSP concentration data to assess the best-fit models for the pollution levels at Ratnapark. The Weibull distribution showed a good fit for both PM 2.5 and TSP concentrations, capturing the skewness and heavy tails of the pollution data. The lognormal and gamma distributions also provided reasonable fits, with the lognormal distribution being particularly suitable for modeling the right-skewed nature of the PM 2.5 data, while the Weibull distribution was more appropriate for TSP concentrations, reflecting the variability and extreme pollution spikes observed.

Table 2. Fitted Distribution and Goodness-of-fit Statistics for PM 2.5

	Estimated Parameter		Kolmogorov Smirnov Test	p-value
Distribution	Shape/Mean log	Scale/Sd log		
Weibull	1.872	80.247	0.10	0.23
Lognormal	4.101	0.586	0.121	0.13
Gamma	3.276	0.0462	0.0147	0.08

The table 2 presents the fitted distribution types (Weibull, Lognormal, and Gamma) along with their estimated parameters and goodness-of-fit statistics, specifically the Kolmogorov-Smirnov (KS) test p-values.

The shape parameter is 1.872, and the scale parameter is 80.247 in Weibull Model. The Kolmogorov-Smirnov test p-value is 0.23 > 0.05, which is greater than the significance level of 0.05, indicating that the Weibull distribution fits the data well and does not significantly deviate from the observed data. Likewise, the shape parameter is 4.101, and the scale parameter is 0.586 in lognormal distribution. The KS test p-value is 0.13, which also exceeds 0.05, suggesting that the lognormal distribution provides a good fit to the data, with no significant deviation from the observed distribution. And the shape parameter is 3.276, and the scale parameter is 0.0462 in Gamma distributional, with a p-value of 0.08. While the p-value is slightly below 0.10, it still suggests a reasonable fit, but it might not be as strong as the Weibull and Lognormal distributions in terms of adherence to the observed data.

Inclusively, all three distributions (Weibull, Lognormal, and Gamma) provide reasonable fits, as indicated by their p-values being greater than 0.05, suggesting that these distributions do not significantly differ from the observed pollution data. However, the Weibull and Lognormal distributions show slightly better fits, as their p-values are higher compared to the Gamma distribution.

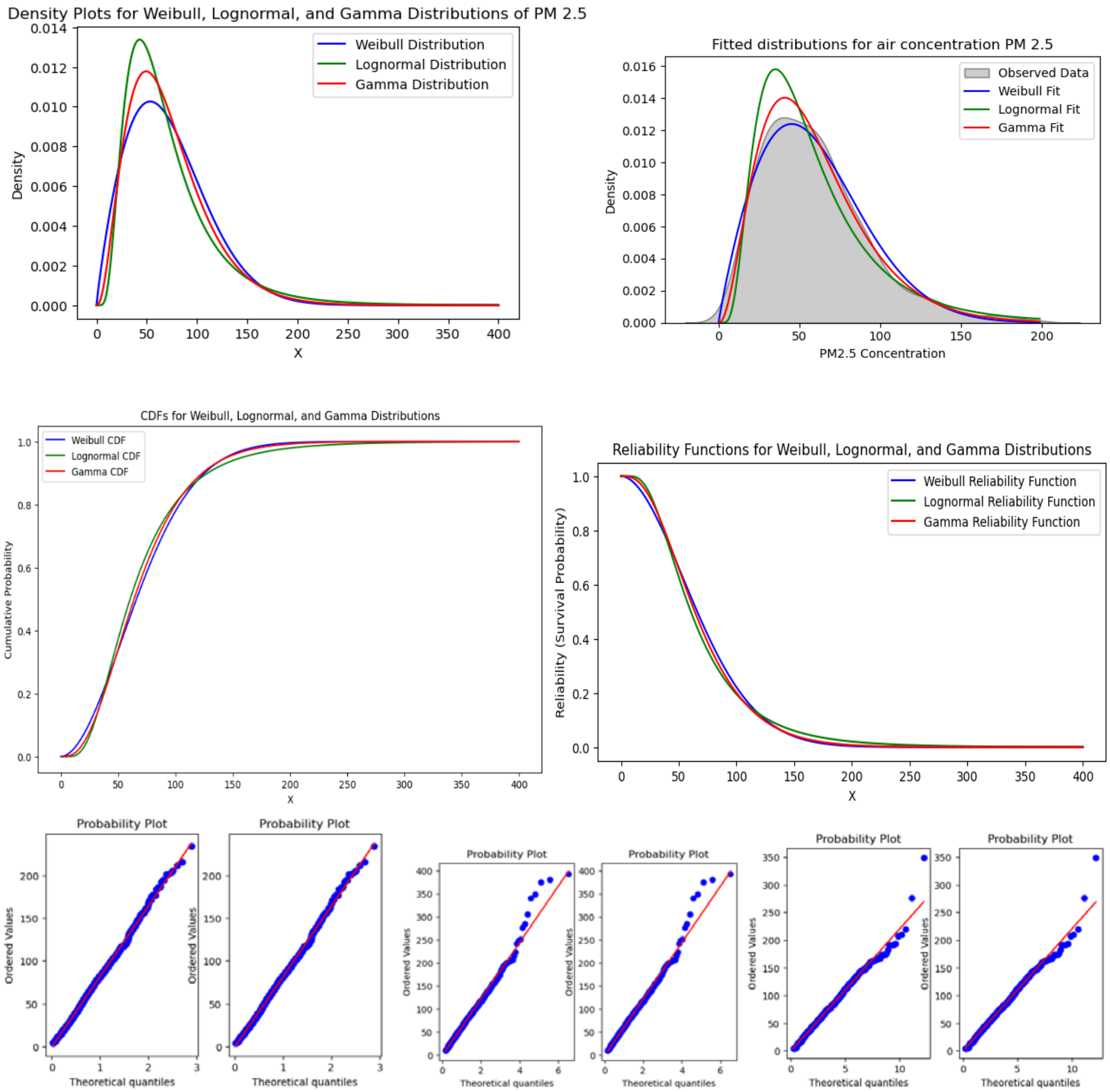


Figure 2. Distributional characteristics of air concentration PM 2.5

Table 3. Fitted Distribution and Goodness-of-fit Statistics using Kolmogorov Test for TSP

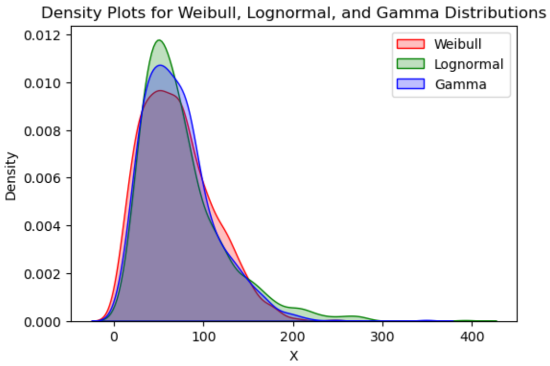
	Estimated parameter		Kolmogorov Smirnov Test	p-value
Distribution	Shape/Mean log	Scale/Sd log		
Weibull	1.105	158.80	0.157	0.17
Lognormal	4.677	0.7485	0.0207	0.06

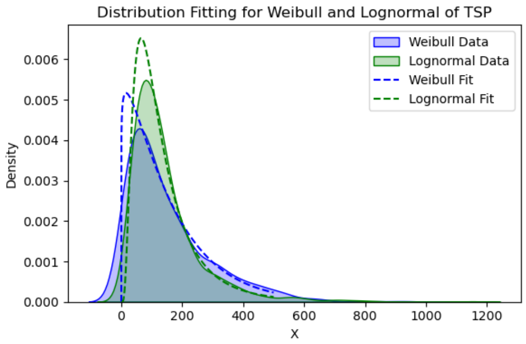
The table 3 presents the fitted distribution types (Weibull and Lognormal) for Total Suspended Particles (TSP) concentrations, alongside their estimated parameters and goodness-of-fit statistics based on the Chi-square test.

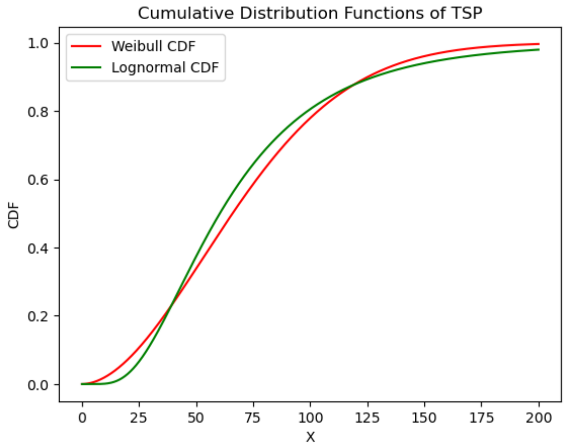
The shape parameter is 1.105, and the scale parameter is 158.80. The Chi-square test yielded a p-value of 0.17, which exceeds the commonly accepted threshold of 0.05, indicating that the Weibull distribution provides an excellent fit to the TSP data. This suggests that the observed distribution of TSP concentrations aligns well with the Weibull model, with no significant deviations.

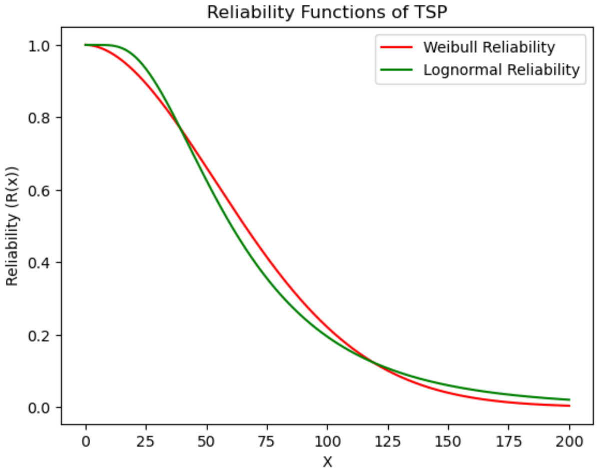
The shape parameter is 4.677, and the scale parameter is 0.7485. The Chi-square test resulted in a p-value of 0.06, which, while slightly lower than 0.17, remains above 0.05, indicating a satisfactory fit. This implies that the lognormal distribution is also an appropriate model for the TSP data, with no substantial evidence of misfit.

Both the Weibull and Lognormal distributions exhibit strong goodness-of-fit, with p-values greater than 0.05, indicating that the models accurately represent the TSP concentration data. The Weibull distribution, however, provides a slightly better fit, as reflected by its higher p-value.









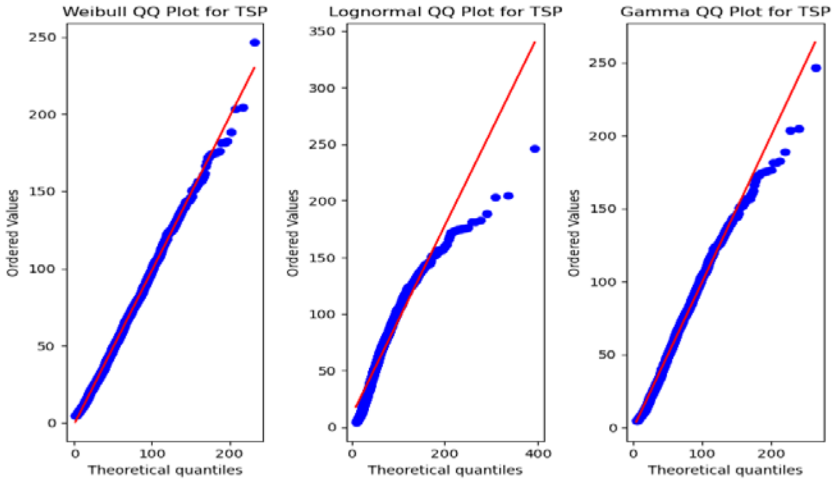


Figure 3. Distributional Characteristics of TSP

This study highlights significant differences in air pollution levels between Pulchok and Ratnapark in Kathmandu. Pulchok, influenced by river side dust, traffic congestion and industrial activities, exhibited higher and more erratic pollution spikes, while Ratnapark had more consistent but still elevated pollution levels. Descriptive statistics revealed variability in both locations, with Pulchok showing more extreme fluctuations.

The study's statistical modeling using Weibull, lognormal, and gamma distributions revealed that Weibull and lognormal models best fit the data for PM2.5 concentrations. The gamma distribution, although suitable for TSP, was less accurate. These results underline the importance of selecting appropriate statistical models for better air quality management by prediction and public health risk assessment.

Conflict of Interest

The authors affirm that there are no conflicts of interest to disclose.

Acknowledgement

We extend our gratitude to the Kathmandu Environment Department for providing access to the database.

References

Cao, J., Xu, B., Zhang, Q., & Li, G. (2018). Assessment of air pollution in urban areas using statistical models. *Journal of Environmental Management*, *224*, 80-89. https://doi.org/10.1016/j.jenvman.2018.07.001

Ghosh, A., & Chaudhuri, R. (2019). Air pollution modeling using probability distributions and extreme value theory: A case study in urban India. *Atmospheric Pollution Research*, *10*(2), 367-373. https://doi.org/10.1016/j.apr.2018.12.002

Huang, R., Hu, B., & Wang, M. (2017). Modeling the temporal and spatial distribution of particulate matter using statistical distributions. *Atmospheric Environment, 161*, 132-141. https://doi.org/10.1016/j.atmosenv.2017.04.035

Huang, Y., Li, Y., & Chen, H. (2017). *Statistical distribution models for air pollutant concentration: A comparative analysis*. Environmental Pollution, 226, 135-142. https://doi.org/10.1016/j.envpol.2017.03.052

Liu, H., Zhang, X., & Xu, B. (2021). Application of lognormal distribution in modeling particulate matter concentrations in urban environments. *Environmental Pollution, 275*, 116539. https://doi.org/10.1016/j.envpol.2021.116539

Liu, X., Zhang, J., & Wang, Y. (2021). *Application of lognormal distribution in modeling particulate matter concentrations*. Environmental Science & Technology, 55(6), 3523-3531. https://doi.org/10.1021/acs.est.0c06473

Sharma, R., & Chauhan, A. (2016). Evaluation of air quality data and statistical fitting for PM concentrations in North India: A statistical approach. *Journal of Environmental Science and Health, Part A, 51*(5), 388-396. https://doi.org/10.1080/10934529.2016.1166967

Zhang, Z., Chen, L., & Wang, F. (2020). Weibull distribution fitting for air quality data: A case study on urban air pollution levels. *Atmospheric Environment, 240*, 117724. https://doi.org/10.1016/j.atmosenv.2020.117724

Zhou, W., Liu, Y., & Li, F. (2019). Statistical modeling of particulate matter concentration using Weibull distribution in a mega city. *Environmental Monitoring and Assessment, 191*(6), 382. https://doi.org/10.1007/s10661-019-7429-x

Zhou, Y., Wang, X., & Li, H. (2019). *Application of Weibull distribution in environmental air quality analysis*. Journal of Environmental Protection, 10(5), 775-784. https://doi.org/10.4236/jep.2019.105049