BILINGUAL PUBLISHING GROUP

Journal of Atmospheric Science Research

https://journals.bilpubgroup.com/index.php/jasr

ARTICLE

Statistical Modeling of $PM_{2.5}$ Concentrations: Prediction of Extreme Events and Evaluation of Advanced Methods for Air Quality Management

Amaury de Souza [1*] , ***Jose Roberto Zenteno Jimenez*** [2] , ***José Francisco de Oliveira-Júnior*** [3] , ***Kelvy Rosalvo Alencar Cardoso*** [3]

[1] *Institute of Physic, Federal University of Mato Grosso do Sul, Campo Grande 79070-900, Brazil*

[2] *Geophysical Engineering, ESIA-Unidad Ticóman Mayor Gustavo A. Madero. National Polytechnic Institute, México City 07340, México*

[3] *Institute of Atmospheric Sciences (ICAT) Federal University of Alagoas, Maceió 57072-900, Brazil*

ABSTRACT

This study analyzes the statistical behavior of $PM_{2.5}$ concentrations in Brasília using advanced probabilistic and time series modeling to support air quality management and extreme event forecasting. The methods applied include Generalized Extreme Value (GEV) distributions, Bayesian inference with Log-Normal distribution, ARIMA models, and quasi-Gaussian approaches. Model performance was evaluated through statistical metrics such as RMSE, R^2, and the Approximation Index, with parameter estimation improved using the Metropolis-Hastings algorithm. Results show that the GEV 1 model provides a better fit for lower $PM_{2.5}$ concentrations, while GEV 2 performs better at predicting extreme events. The log-logistic and log-normal distributions also demonstrated good fit, capturing asymmetry and long-tail behavior typical of environmental data. The ARIMA model identified seasonal patterns and supported short-term forecasts, though its predictive capacity for extreme values was limited. Bayesian inference allowed robust estimation of parameter uncertainties and revealed the non-negligible likelihood of severe pollution events. The study concludes that model selection should depend on the forecasting objective: GEV for extremes, Log-Normal for general variability, and ARIMA for trends and seasonality. The

*CORRESPONDING AUTHOR:

Amaury de Souza, Institute of Physic, Federal University of Mato Grosso do Sul, Campo Grande 79070- 900, Brazil; Email:amaury.souza@ufms.br

ARTICLE INFO

Received: 21 May 2025 | Revised: 8 July 2025 | Accepted: 16 July 2025 | Published Online: 22 July 2025
DOI: https://doi.org/10.30564/jasr.v8i3.10878

CITATION

Souza, A., Jimenez, J.R.Z., Oliveira Junior, J.F., et al., 2025. Statistical Modeling of PM2.5 Concentrations: Prediction of Extreme Events and Evaluation of Advanced Methods for Air Quality Management. Journal of Atmospheric Science Research. 8(3): 67–92.
DOI: https://doi.org/10.30564/jasr.v8i3.10878

use of MCMC sampling techniques significantly improved model robustness. These findings provide a comprehensive framework for understanding air pollution dynamics and guiding public policy on air quality in urban environments.
Keywords: $PM_{2.5}$; GEV; ARIMA; Bayesian Inference; Metropolis-Hastings

1. Introduction

Air pollution remains a significant environmental and public health concern, particularly in urban areas, where fine particulate matter ($PM_{2.5}$) is a leading contributor to degraded air quality and adverse health outcomes, including respiratory diseases[1]. In Brasília, Brazil's capital, $PM_{2.5}$ concentrations are influenced by various factors such as vehicle emissions, seasonal fires in the Cerrado, and local meteorological conditions[2]. Statistical modeling of $PM_{2.5}$ variability is crucial not only for understanding its temporal and spatial distribution but also for guiding air pollution control and mitigation strategies[3].

Among the various methods for environmental data modeling, probability distribution functions (PDFs) have proven effective in capturing the variability of pollutant concentrations. This study explores the use of the Log-Logistic, Generalized Extreme Value (GEV), and Log-Normal distributions for modeling and predicting $PM_{2.5}$ concentrations in Brasília. These distributions exhibit distinct statistical properties, offering varying levels of accuracy in representing $PM_{2.5}$ concentration data[4,5].

Jimenez et al.[6] conducted a study in Mexico City to model $PM_{2.5}$ concentrations from 2010 to 2018, aiming to identify the best-fitting probability distribution. They compared distributions such as Gamma, Extreme Value, Gumbel, and Weibull, while employing Bayesian inference for daily maximum values. Parameters were estimated using Maximum Likelihood Estimation (MLE) and the Method of Moments, with model performance evaluated through metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Coefficient of Determination (R^2), Approximation Index, and Prediction Accuracy. These metrics help validate the quality of the distribution fits and the reliability of predictions. The study also included a trend analysis of $PM_{2.5}$ concentrations, incorporating Bayesian inference to model daily maximum values and identify potential patterns or temporal changes. The results were compared with official air quality data from Mexico City's environmental authorities to ensure alignment with real-world observations. This validation process is vital for ensuring the accuracy and reliability of predictions.

The comparison of different distributions and estimation techniques allowed for the identification of the best models to represent $PM_{2.5}$ variability. Bayesian inference, applied to Normal and Extreme Value distributions, highlighted the importance of modeling not only the general variability of the data but also rare and extreme high-concentration events. Although the study did not specify which distribution provided the optimal fit for $PM_{2.5}$ data, future research could explore which model—Gamma, Extreme Value, Gumbel, Weibull, or Bayesian approaches—performs best based on the evaluated metrics. Incorporating additional explanatory variables, such as meteorological factors, could further enhance the robustness of $PM_{2.5}$ modeling and forecasting.

This research follows a similar comprehensive approach to Jimenez et al.[6], combining traditional statistical methods (MLE, Method of Moments) with Bayesian inference to model $PM_{2.5}$ concentrations. The robust analysis of various probability distributions, complemented by validation using official data, reinforces the reliability of the results and contributes to a better understanding of air pollution trends in Brasília. To assess the adequacy of the distributions for modeling $PM_{2.5}$ concentrations, statistical metrics such as MSE, RMSE, Absolute Precision (AP), and Concordance Index (AI) (Willmott et al.[7]) were used. These indicators will guide the identification of the most appropriate distribution to represent $PM_{2.5}$ variability in Brasília. This analysis aims to support future studies on air quality dynamics in the region and assist in the development of strategies to mitigate air pollution in the Brazilian capital.

2. Materials and Methods

2.1. Study Area

Brasília is situated at 15.8° S, 47.9° W, with an average elevation of 1,172 meters above sea level. The city

experiences a tropical high-altitude climate, marked by two distinct seasons: a rainy season from October to April and a dry season from May to September. The annual average temperature ranges between 20 °C and 22 °C, with peak temperatures reaching 29 °C to 31 °C during the warmer months and minimum temperatures dropping to 12 °C to 14 °C during the cooler winter period. Relative humidity is typically high during the rainy season but drops significantly during the dry season, often falling below 30%.

The local topography is primarily flat, featuring plateaus and gently undulating terrain. The natural vegetation of the region belongs to the Cerrado biome, one of Brazil's most biodiverse ecosystems, characterized by grasses, shrubs, and small to medium-sized trees that are adapted to nutrient-poor soils and extended dry periods (**Figure 1**).

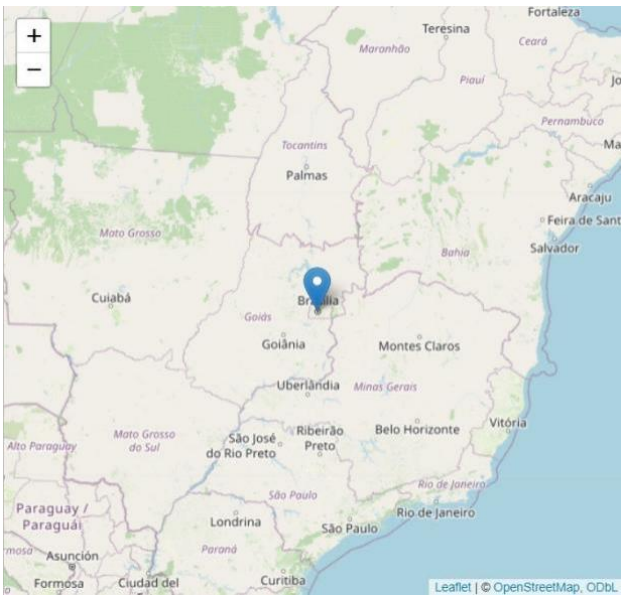


Figure 1. Location of the Federal District and Brasília in Brazil.

2.2. Data

The O_3 concentration data used in this study were provided by the Environmental Information System Integrated with Health (SISAM), managed by the Instituto Nacional de Pesquisas Espaciais (INPE). These data, collected daily by satellite in each municipality of the state, cover a 16-year period from 2000 to 2018.

To assess trends in pollutant concentrations, the Mann-Kendall (MK) test[8,9] was applied to identify any significant increases or decreases. The MK statistic (S) for a time series is computed as follows:

$$S = \sum_{k=1}^{n-1} \sum_{i=k+1}^{n} \operatorname{sgn}(x_i - x_k) \tag{1}$$

Where x_i and x_k represent observations at points i and k, respectively, and n denotes the total number of observations in the series. The sign function (sgn(x)) is defined as:

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{2}$$

To ensure the reliability of the trend analysis in non-random and serially correlated time series, modified MK tests were applied using the variance correction methods proposed by[10,11]. These methods adjust the variance by calculating the effective sample size based on significant serial correlations, thereby ensuring accurate trend detection.

Additionally, the Pettitt test[12], based on the Mann-Whitney test, was used to identify significant change points by splitting the data into two distinct samples and calculating the Pettitt statistic U(t,n). The trend analysis was performed across different temporal scales—daily, monthly, seasonal, and annual—allowing for the capture of variations from multiple perspectives.

After the initial data analysis, several statistical methods were employed for modeling and analyzing $PM_{2.5}$ concentrations, including autoregressive models, Bayesian inference, and extreme value distributions. The methodology followed these steps:

2.2.1. ARIMA Modeling

An Integrated Autoregressive Moving Average (ARIMA) model was used to capture temporal patterns in the data. The model was estimated with various lag orders and smoothing parameters, selecting ARIMA(10,0,5) based on information criteria such as AIC and BIC. The model's fit was evaluated through residual analysis and QQ-plots.

2.2.2. Bayesian Inference with Log-Normal Distribution

To account for the positive skewness in the data, Bayesian inference with a log-normal distribution was applied. The parameters were estimated using the Markov Chain Monte Carlo (MCMC) method via Metropolis-Hastings. This approach provided a more robust representa-

tion of uncertainties in the parameters and facilitated more reliable forecasts for future events.

2.2.3. Extreme Value Distributions - GEV and Log-Logistic

To model extreme events, the Generalized Extreme Value (GEV) and Log-Logistic distributions were fitted to the data.

GEV1 demonstrated a better fit for lower $PM_{2.5}$ concentrations, indicating an asymmetric distribution with a long tail.

GEV2 was more suitable for predicting high pollutant concentrations, reflecting the potential for extreme future events.

Comparative Analysis of Models: The models were compared using:

Residual analysis and QQ-plots to assess the adequacy of the data fitting.

AIC and BIC criteria for model selection.

Cross-validation to test the predictive power of the models.

The results revealed that combining traditional methods (ARIMA), Bayesian inference, and extreme distributions provided more accurate modeling of $PM_{2.5}$ concentrations, capturing both regular patterns and extreme pollution events.

The performance of the adjusted distributions was evaluated using statistical metrics, including Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) to assess model accuracy, Absolute Precision (AP) to measure the mean absolute deviation, and the Concordance Index (IA) to evaluate how well the distributions aligned with the observed data.

Based on this performance information, the distribution with the lowest error values (MSE and RMSE) and the highest precision and concordance values (AP and IA) is considered the most suitable for representing $PM_{2.5}$ concentrations in Brasília.

3. Results

3.1. Descriptive Analysis

In Brazil, fires for land conversion are common during the dry season, particularly in the months leading up to the rainy season. The Midwest region of Brazil, where Brasília is located, is a hotspot for fires[11–15]. These fires are likely to contribute to the increase in pollutant concentrations during this period. **Table 1** summarizes the statistical properties of the $PM_{2.5}$ data, including the mean, standard deviation (SD), and the maximum and minimum values of the Mann-Kendall (MK) test across different time scales. The highest $PM_{2.5}$ concentrations were observed between August and October, corresponding to the months with the greatest SD in concentrations.

Table 1. Descriptive statistics of $PM_{2.5}$ concentrations and *p*-values from the Mann-Kendall (MK) test for temporal trends at various time scales in Brasília.

Months	Average	Place	Maximum	Minimum	*p*-Values
Jan	11.55	15.49	54.90	0.53	0.99
Feb	12.35	6.82	76.30	0.73	0.16
Mar	15.04	9.82	102.28	0.00	0.07
Apr	13.06	5.68	38.10	1.48	0.07
May	14.18	9.53	71.95	1.43	0.89
Jun	11.16	7.50	59.65	1.78	**0.04**
Jul	15.02	23.71	422.40	2.05	**0.01**
Aug	17.67	23.13	308.53	1.85	0.06
Sep	27.69	29.63	315.35	3.00	0.16
Oct	20.57	14.61	125.68	1.70	0.87
Nov	13.61	7.17	107.70	1.78	0.99
Dec	13.19	7.99	153.55	2.08	0.11
				Annual	0.14
				Dry	**0.97**
				daily	0.49

Legend: In bold, significant trends.

Overall, no significant trends were identified for $PM_{2.5}$, although slight variations were observed (**Figure 2**), with an increase of 0.2436 μg/m³/year in June and a decrease of 0.4875 μg/m³/year in July. Breakpoints in the $PM_{2.5}$ series were identified in June 2008 and July 2007. $PM_{2.5}$ can act as a medium for photochemical reactions, facilitating the formation of O_3. $PM_{2.5}$ can act as a medium for photochemical reactions, facilitating the formation of O_3. While SO_2, a primary pollutant from fossil fuel combustion, does not directly contribute to O_3 formation and may even inhibit its production, it can influence the chemical processes that generate O_3, much like NOx. These relationships are context-dependent and can vary based on local conditions, emission sources, and specific atmospheric dynamics. As such, air pollution control policies should prioritize reducing emissions during critical periods, such as fire seasons, to mitigate peak pollutant concentrations. Statistical models are crucial tools for monitoring and predicting these effects, emphasizing the need for models that account for temporal and spatial variations in pollutant concentrations. This approach is essential for effective air pollution mitigation planning[16].

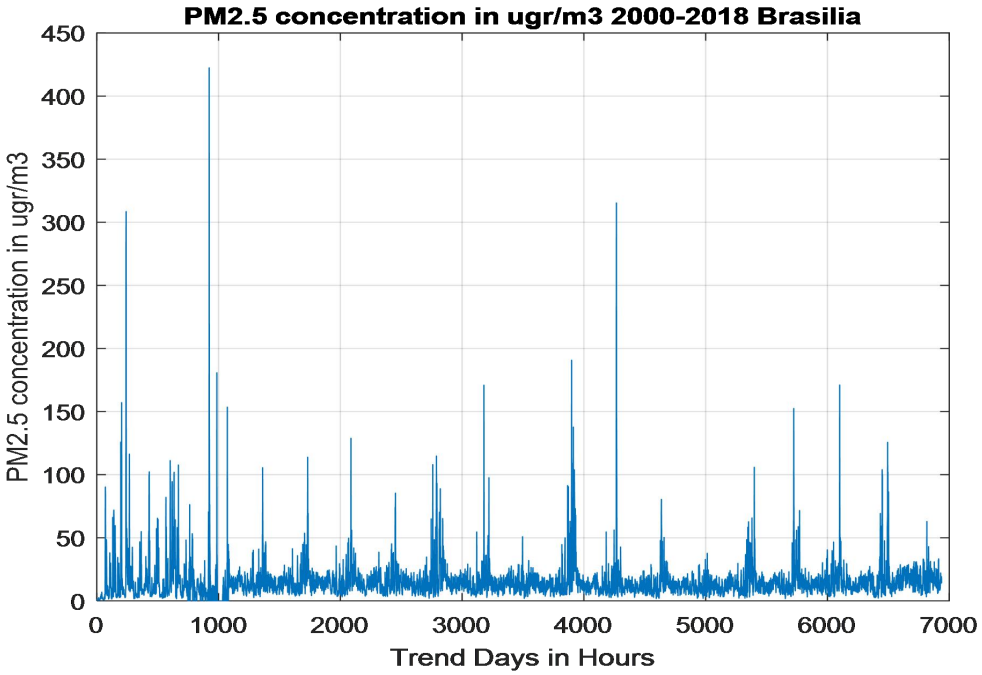
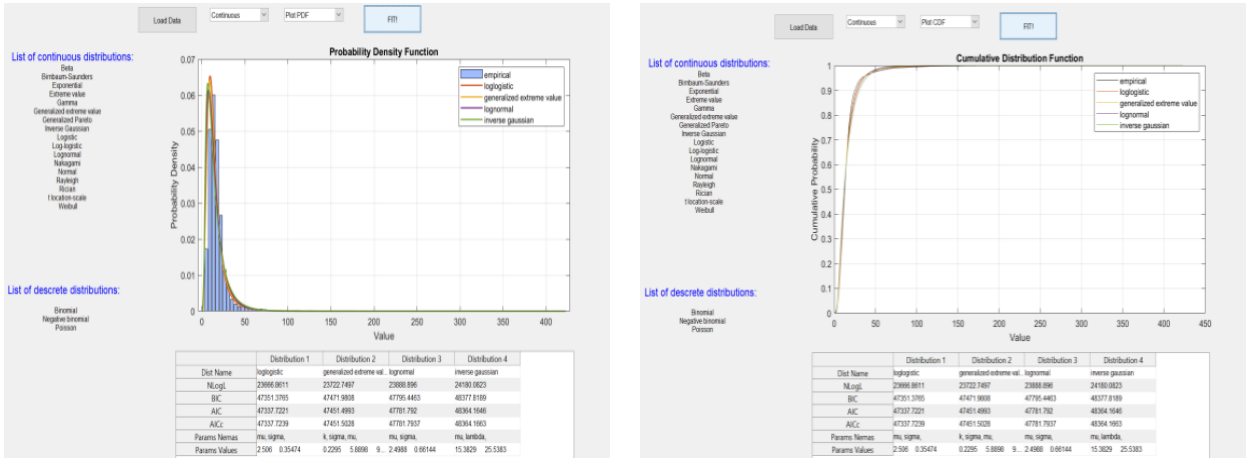


Figure 2. Concentrations of $PM_{2.5}$ Daily Brasilia City 2000–2018.

3.1.1. Results of Probability Distribution Function Fitting Trend of $PM_{2.5}$ Brasília Data (2000–2021)

The Adjusted probability density functions (PDFs), as shown in **Figure 3a**, include distributions such as Generalized Extreme Value (GEV), Weibull, and Exponential, among others. The accompanying table provides statistical metrics that enable an assessment of the fit's quality. The graph indicates that the tested distributions exhibit asymmetric behavior with a right tail, suggesting that the data follows a steeply sloping distribution. This pattern is commonly observed in environmental variables such as wind speed, extreme precipitation, and air pollutants. Also, the overlapping of the curves implies that some distributions provide similar fits. However, selecting the best distribution necessitates evaluating statistical metrics such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Root Mean Square Error (RMSE). The histogram of the empirical distribution displays a sharp peak at low values and a long tail, indicating that high-magnitude events are rare. This pattern may be associated with extreme environmental phenomena. Choosing the most suitable distribution can offer valuable insights into environmental risks, the prediction of extreme events, and the underlying physical processes.



(**a**) (**b**)

Figure 3. (**a**) Adjustment of statistical distributions to the observed data. (**b**) Adjustment of statistical distributions to the cumulative distribution function (CDF) of the observed data.

The adjustment of the cumulative distribution functions (CDFs) is illustrated in **Figure 3b**, which presents the empirical CDF alongside the CDFs modified to fit various statistical distributions. The accompanying table includes statistical metrics that assist in identifying the best fit for the data. The cumulative distribution function increases steeply at low values, indicating that most of the data is concentrated within this range. This observation highlights the strong asymmetry of the distribution and the presence of a long tail on the right, which is a common feature in environmental and hydrological variables. The adjusted distributions closely follow the empirical CDF, suggesting a good statistical fit. However, minor differences can be observed in the tail region, where extreme events occur with low frequency.

Choosing the most appropriate distribution directly affects statistical forecasting and environmental modeling. A proper fit improves the estimation of probabilities of rare events, such as pollution spikes or extreme rainfall, aiding in the formulation of environmental policies and risk management strategies.

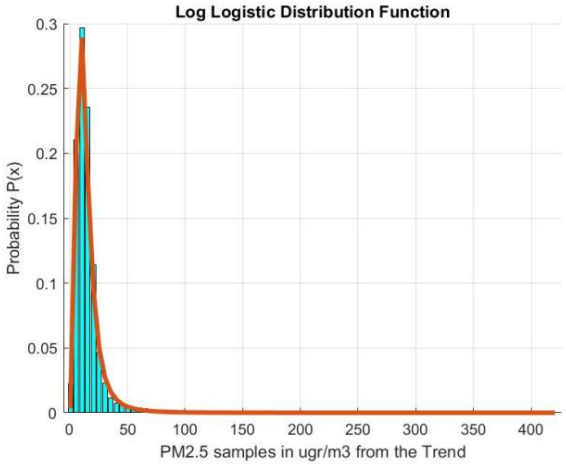
Stationarity analysis refers to the constancy of the statistical characteristics of a time series, such as mean and variance, over time. Tests such as the Dickey-Fuller are commonly used to evaluate this property. If a time series exhibits seasonal trends or significant variations, transformations such as differencing may be necessary to achieve stationarity. The shape of the tails of distributions is critical to understanding the occurrence of extreme events. The asymmetry on the left suggests a greater concentration of low values, indicating a sloping distribution. Depending on the type of adjusted distribution (such as Weibull or Gamma), this factor can impact the modeling of rare events.

The high concentration of values close to zero is relevant in several areas, such as air pollution and disease incidence, where many events have low intensity or frequency. To address this characteristic, logarithmic transformations or adjustments based on specific distributions can be applied that better capture this behavior.

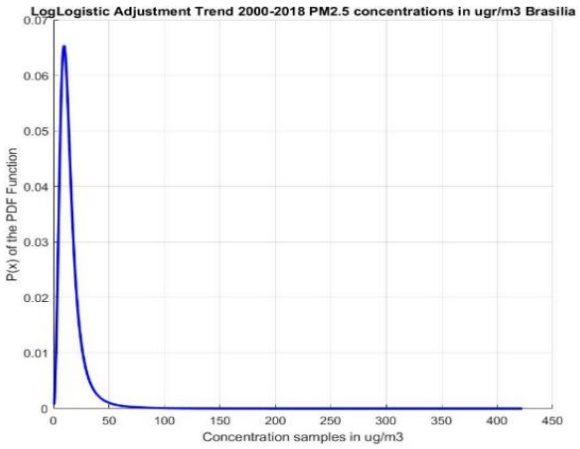
3.1.2. Viewing the settings of 3 Functions: Logistic Log, GEV and Normal Log

Log Logistic

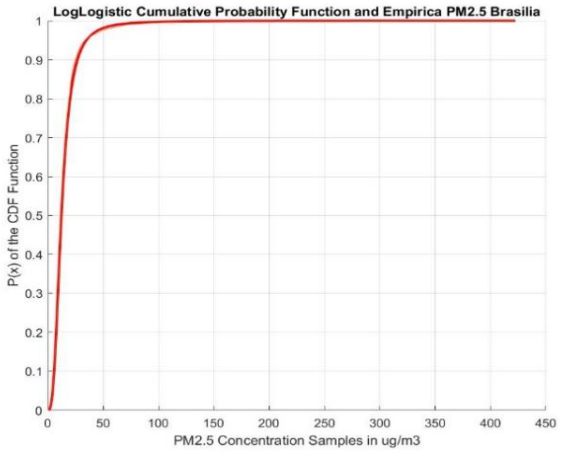
The histogram (**Figure 4a**) and the cumulative distribution function (CDF) (**Figure 4c**) indicate a good fit of the log-logistic distribution to the $PM_{2.5}$ data, effectively capturing the asymmetry and long right tail of the distribution. For this distribution (**Figure 4b**), the model fitting yielded a shape parameter $\sigma = 0.3547$ and a location parameter $\mu = 2.506$. The coefficient of determination ($R^2 = 0.6215$) suggests a moderate fit, while the root mean square error (RMSE = 0.5300) points to some discrepancy between observed and fitted values. The approximation index (0.4569) further indicates that the model may not be ideal for predicting extreme values. The QQ plot (**Figure 4d**) reveals that the empirical and theoretical quantiles of the log-logistic distribution are well aligned in the central portion of the data but exhibit deviations at the extremes—particularly in the upper tail—suggesting that extreme $PM_{2.5}$ events may not be adequately represented by this model.



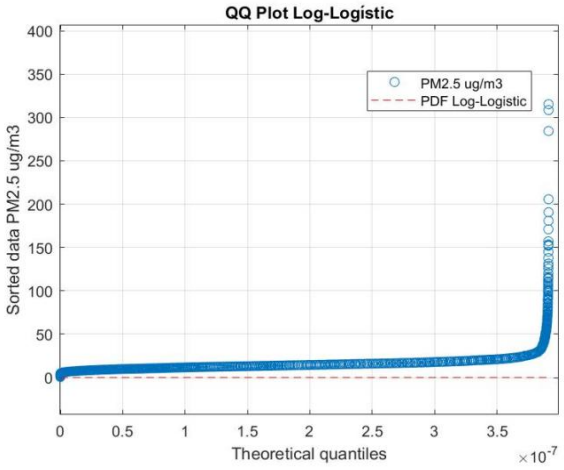
(**a**)



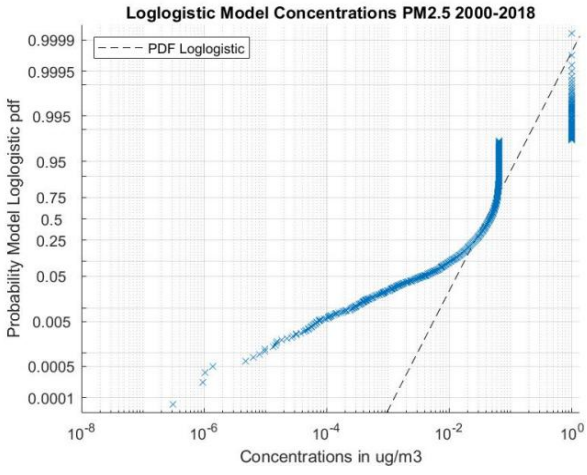
(**b**)

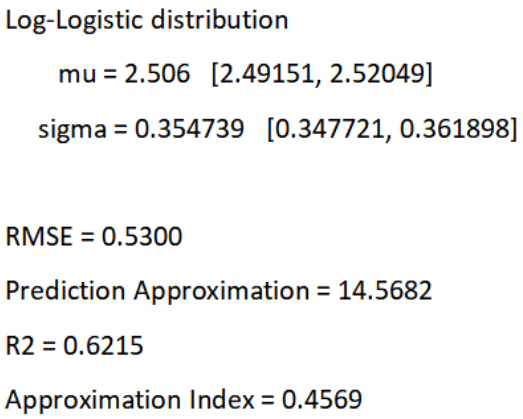


(**c**)



(**d**)





(**e**)

Figure 4. (**a**) (Top left) Log-logistic distribution function fitted to $PM_{2.5}$ concentration data. The histogram represents the observed data, while the red curve displays the adjustment of the distribution. (**b**) (Top right) Adjustment of the log-logistic distribution to the $PM_{2.5}$ time series (2000–2018). The blue curve represents the modeling of the data over time. (**c**) (Center left) Comparison between the empirical cumulative distribution function (CDF) of the $PM_{2.5}$ data and the theoretical CDF of the log-logistic distribution. (**d**) (Center right) QQ plot comparing the empirical quantiles of the $PM_{2.5}$ distribution with the theoretical quantiles of the log-logistic distribution. A good fit is indicated by the proximity of the points to the reference line. (**e**) (Bottom) $PM_{2.5}$ concentrations modeled by the log-logistic distribution between 2000 and 2018, presented on a logarithmic scale.

Based on these results, the Log-Logistic distribution stands out as a viable option for modeling environmental variables with asymmetric behavior and a long tail, such as $PM_{2.5}$. However, the fit can be improved by considering alternative distributions, such as Generalized Extreme Value (GEV), which offers greater flexibility to model extreme events. The analysis suggests that, despite the reasonable adjustment, there is room for refinement that improves the prediction and representation of critical $PM_{2.5}$ concentrations.

For the Generalized Extreme Value (GEV) distribution, the histogram (**Figure 5a**) and the cumulative distribution function (CDF) (**Figure 5c**) demonstrate a good fit to the $PM_{2.5}$ data, effectively capturing the asymmetry and the long right tail. The model fitting (**Figure 5b**) produced a shape parameter $\sigma = 6.3956$, a location parameter $\mu = 9.766$, and a form parameter $k = 0.18007$. The positive value of the shape parameter ($k > 0$) indicates a heavy-tailed distribution, suggesting the presence of extreme values in the data. The coefficient of determination ($R^2 = 0.6468$) shows that the GEV model explains approximately 65% of the data variability, which represents a moderate, though not exceptional, fit. The root mean square error (RMSE = 0.5286) and the mean squared error (MSE = 0.2794) indicate relatively low error values, supporting the adequacy of the fit. However, the approximation index (0.4569) and the prediction approximation value (13.7673) suggest that the model may not be optimal for accurately forecasting extreme $PM_{2.5}$ events.

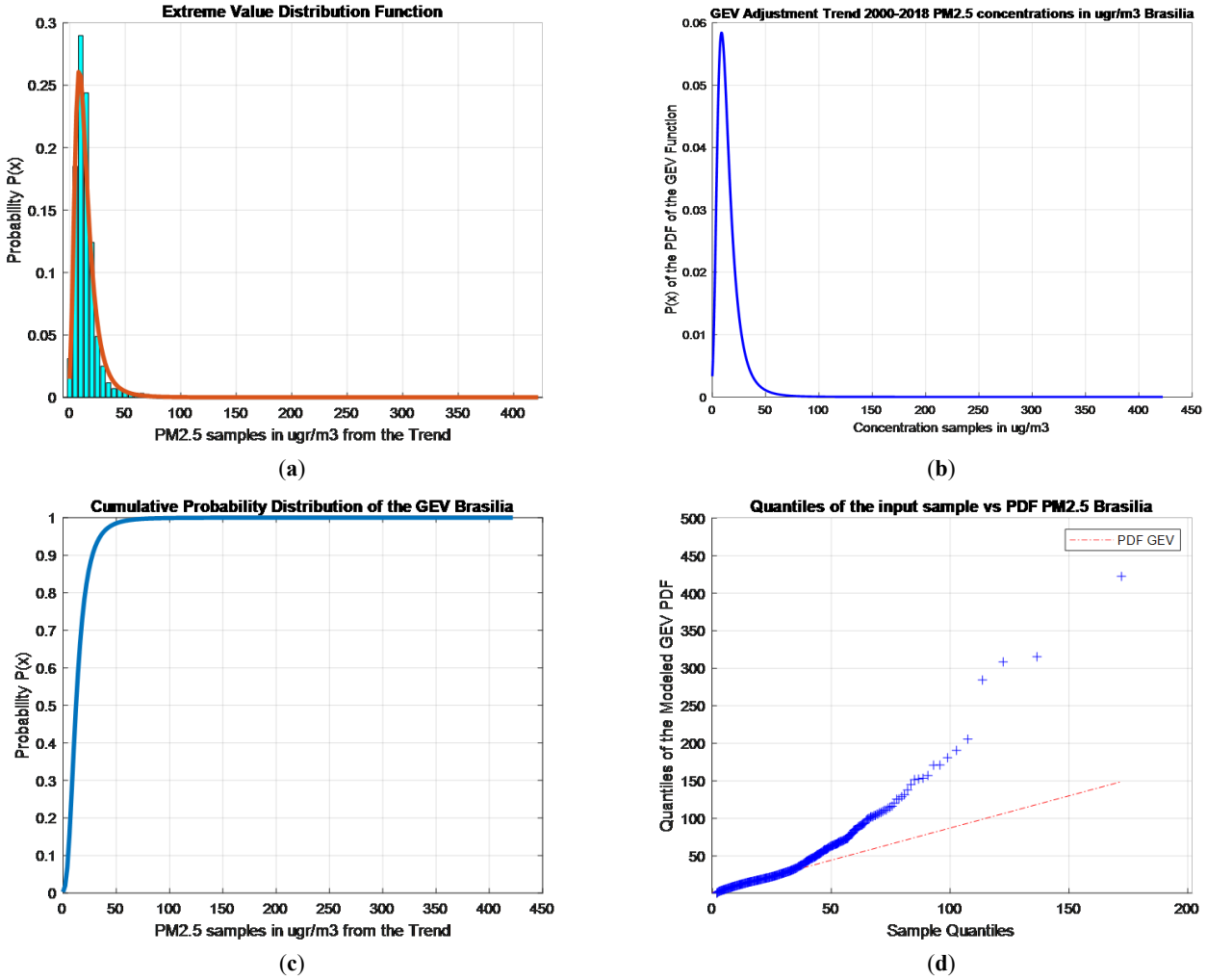
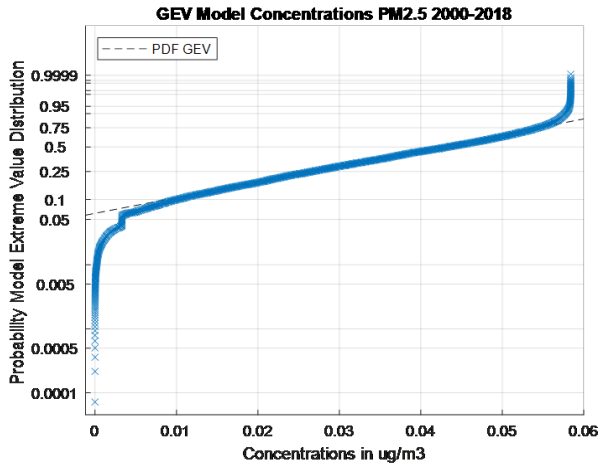


Figure 5. *Cont.*

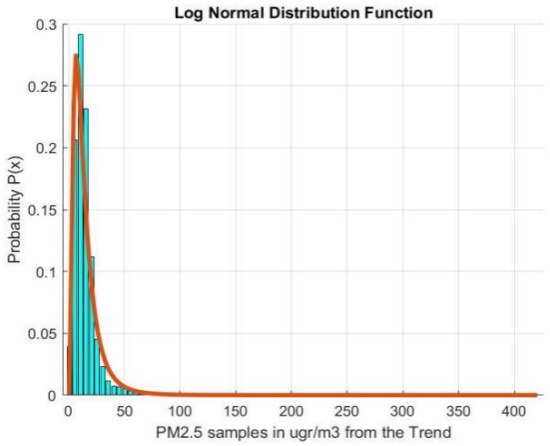


(**e**)

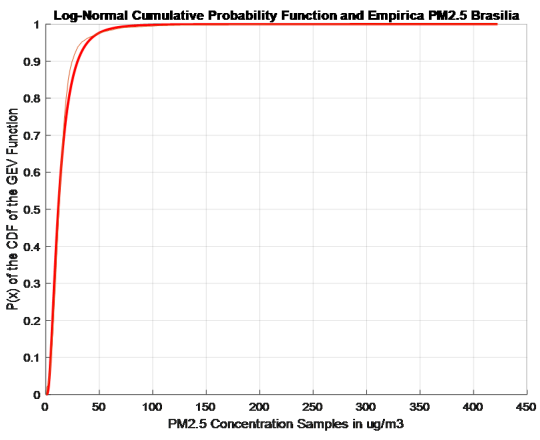
Figure 5. (**a**) Histogram and Adjusted Probability Density Function (PDF): Represents the distribution of the empirical data and the adjusted curve of the GEV distribution. (**b**) Adjusted Cumulative Distribution Function (CDF): Displays the empirical CDF of the data and the fit of the GEV distribution. (**c**) Empirical Probability Plot: Displays the empirical CDF adjusted to the GEV. (**d**) Quantile-Quantile Plot (QQ-Plot): It compares the empirical quantiles with the theoretical quantiles of the GEV. A good fit is indicated by the proximity of the points to the red line. (**e**) Adjusted Cumulative Density Function: Represents the accumulated density of the data adjusted to the GEV distribution.

The QQ plot (**Figure 5d**) confirms the adequacy of the fit, but large deviations in the upper quantiles may indicate that the model does not represent the extreme values well. Confirms the adequacy of the fit, but large deviations in the upper quantiles may indicate that the model does not represent the extreme values well.

Log-Normal distribution is commonly used to model environmental variables, particularly pollutant concentrations, as it effectively captures data asymmetry (right-skewed distribution with a long tail), with a location parameter $\mu = 2.4556$, and the scale parameter $\sigma = 0.7322$. The histogram (**Figure 6a**) and CDF (**Figure 6b**) indicate a good fit, capturing the asymmetry and long tail to the right of the data. The coefficient of determination ($R^2 = 0.6689$) indicates a moderate to good fit, suggesting that the Log-Normal distribution explains approximately 67% of the data variability, surpassing the performance of the Generalized Extreme Value (GEV) distribution ($R^2 = 0.6468$). The RMSE = 0.5118, indicating a reasonable fit. The approximation index (0.4897) points out that the model also may not be ideal for extreme values, with a prediction approximation of 13.6735. Additionally, the QQ-Plot (**Figure 6c**) provides insight into how well the model represents upper and lower quantiles.

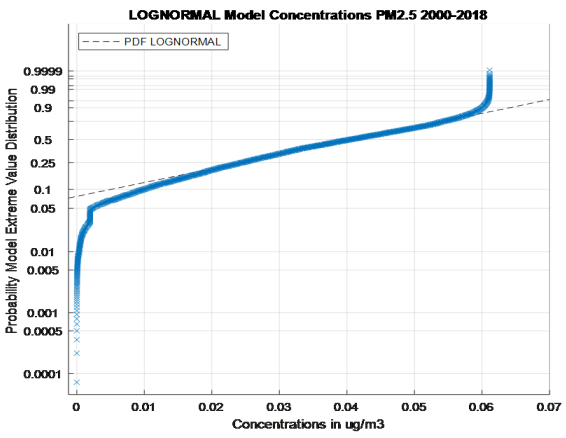


(**a**)

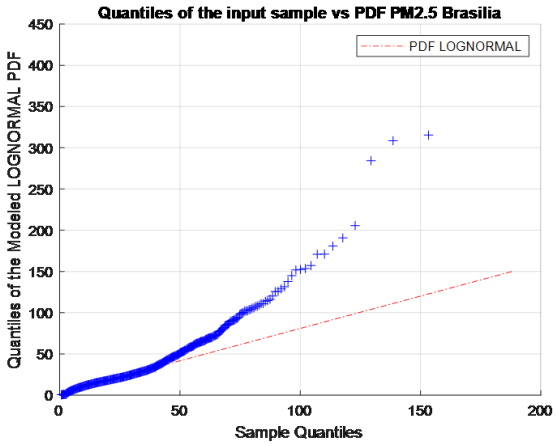


(**b**)

Figure 6. *Cont.*



(**c**)



(**d**)

Log Normal

mu = 2.4556

sigma = 0.7322

RMSE = 0.5118

Prediction Approximation = 13.6735

R2 = 0.6689

Approximation Index = 0.4897

Figure 6. (**a**) Histogram and PDF: Displays the empirical distribution of the data alongside the fitted Log-Normal probability density function. (**b**) Cumulative Distribution Function (CDF): Compares the empirical cumulative distribution with the fitted Log-Normal CDF. (**c**) Quantile-Quantile Plot (QQ-Plot): Evaluates how well the empirical quantiles align with the theoretical quantiles of the Log-Normal distribution. Significant deviations indicate potential model inadequacy. (**d**) Adjusted Cumulative Density Function: Illustrates the cumulative density of the data adjusted to the Log-Normal distribution.

In probability theory, the tail of a distribution represents extreme values, either high or low. A long-tailed distribution indicates that extreme events are more likely to occur than in distributions with shorter tails, such as the Normal distribution. Examples of long-tailed distributions include the Log-Logistic and Generalized Extreme Value (GEV) distributions, which suggest a higher likelihood of observing rare, extreme values (**Table 2**).

Table 2. Distribution Model Fitting Metrics (Log-Logistic, GEV and Log-Normal) for $PM_{2.5}$ Concentrations.

PDF	R^2	RMSE	Approximation Index	KS Test	Chi-Square Test (*p*-Value)
Log-Logistic	0.6215	0.5300	0.4569	0.8831	0.0000
GEV	0.6468	0.5286	0.4696	0.8631	0.0000
Log-Normal	0.6689	0.5118	0.4897	0.8631	0.0000

In the case of air pollutant concentrations, these distributions often exhibit long-tailed behavior, indicating that extreme pollution events are not uncommon. For instance, modeling ozone concentrations in urban areas reveals that peak values occur more frequently than a normal distribution would predict. These peaks may result from factors such as wildfires, industrial emissions, or specific meteorological conditions[13–17].

Moreover, pollutant behavior is influenced by various environmental and atmospheric factors, including wind speed and direction, temperature and humidity, local topography (such as mountains, valleys, and urban structures), and atmospheric chemical reactions. These processes can lead to localized pollution spikes, contributing to long tails in pollutant distributions. Additionally, these pollution spikes can have significant health impacts, particularly involving $PM_{2.5}$, ozone, and NO_2, which are linked to severe respiratory and cardiovascular issues, chronic diseases, and premature mortality[4,18–22].

3.2. Modeling

3.2.1. ARIMA

Applying the ARIMA model and then smoothing the data and removing higher outliers to improve model accuracy. **Figure 7** presents the time series of daily $PM_{2.5}$ concentrations from 2000 to 2018. The original series (blue) displays significant daily variability, marked by abrupt peaks and high-frequency fluctuations. In contrast, the smoothed series (red) highlights long-term trends and seasonal patterns, offering a clearer view of the data's underlying behavior.

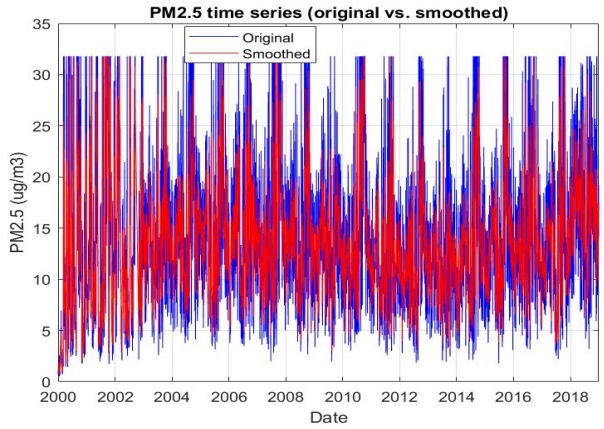


Figure 7. $PM_{2.5}$ Time Series (Original vs. Smoothed): Blue: Original $PM_{2.5}$ Series, Red: Smoothed $PM_{2.5}$ Series.

The data indicates frequent fluctuations, likely due to the influence of intermittent sources such as vehicle and industrial emissions, along with natural events like fires. The null hypothesis has been rejected, suggesting that the time series is likely stationary, as evidenced by the Augmented Dickey-Fuller (ADF) test result of −27.68. Seasonal patterns and trends are also evident, with smoothing techniques revealing recurring cycles that suggest seasonal variations in air pollution. Furthermore, it's important to explore potential long-term trends through statistical tests to determine whether $PM_{2.5}$ concentrations have increased or decreased over time. Hence, smoothing the original series, which contains significant noise, is crucial to help identify structural patterns more easily. Smoothing simplifies trend analysis, supports predictive model development, and aids in detecting anomalous events.

Several adjustments were tested to find the most suitable parameter configuration, as shown in **Table 3**, to better align with the real-time series' behavior. Among the applied parameters, the ARIMA (10,0,5) model emerged as the best fit, incorporating 10 autoregressive (AR) terms, no differencing terms, and 5 moving average (MA) terms. This combination effectively captures both autocorrelation and seasonality in the $PM_{2.5}$ time series.

The results of the ADF test indicate that the time series is stationary, with a *p*-value significantly lower than the common significance level of 0.05. The autoregressive (AR) and moving average (MA) terms show statistically significant values, although there are variations among the lag terms. The AR(1), AR(2), AR(3), AR(4), AR(5), and AR(10) terms exhibit strong significance. In contrast, the AR(8) and MA(4) terms show marginal significance, suggesting that there is room for further refinement. The estimated variance of the model is 152.71, with a *p*-value close to zero, indicating that it effectively captures the variability in the data (**Table 3**).

Table 3. Estimated Parameters and Statistical Significance of the ARIMA Model for $PM_{2.5}$ Concentrations.

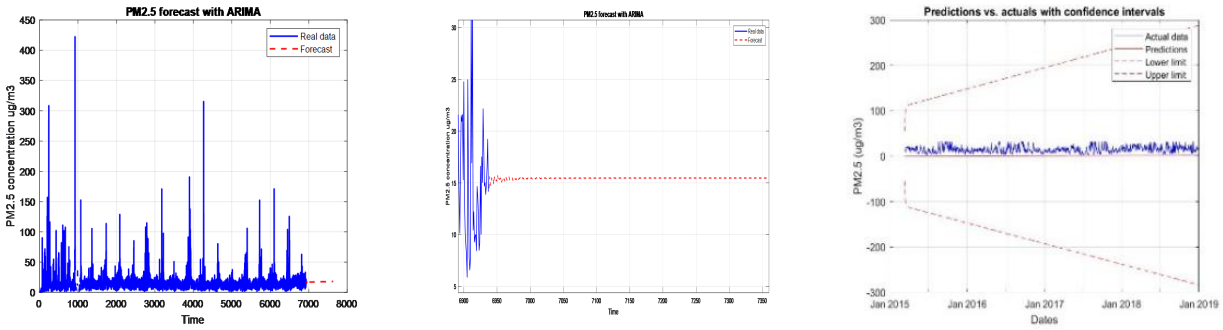
Term	Value	Standard Error	Statistics t	*p*-Value
Constant	1.52E−05	5.12E−05	0.297	0.766
AR(1)	−0.592	14.03	−0.421	0.673
AR(2)	−0.275	0.490	−0.561	0.574
AR(3)	0.2681	0.154	17.388	0.082
AR(4)	0.189	0.385	0.492	0.622
AR(5)	0.028	0.057	0.506	0.612
AR(6)	0.038	0.024	16.071	0.108
AR(7)	0.027	0.039	0.702	0.482
AR(8)	−0.007	0.023	−0.317	0.750
AR(9)	−0.001	0.028	−0.060	0.951
AR(10)	−0.007	0.024	−0.326	0.743
MA(1)	−0.831	14.03	−0.592	0.553
MA(2)	−0.361	18.44	−0.196	0.844
MA(3)	−0.446	0.664	−0.671	0.501
MA(4)	0.387	0.375	10.302	0.302
MA(5)	0.252	0.647	0.389	0.696
Variance	154.43	0.500	308.43	0

The ARIMA(10,0,5) model seems well-suited for analyzing this time series. Further validation, including residual analysis and comparison of model fit statistics (AIC, BIC), will confirm the model's ability to reliably predict future $PM_{2.5}$ concentrations. The autoregressive (AR) coefficients, particularly AR(3), AR(4), AR(5), and AR(6), exhibit strong significance with very low *p*-values, indicating that these variables significantly influence the model's predictions. The moving average (MA) coefficients (1 to 5) are also notable, as they have low *p*-values, suggesting that the moving average plays a crucial role in the model. The model's variance is 151.82, which has a very low *p*-value, confirming the model's effectiveness in capturing data variability.

Figure 8a and **Figure 8b** illustrate a comparison between actual $PM_{2.5}$ concentrations (represented by the blue line) and predictions from the ARIMA model (depicted by the black dotted line). The red dashed lines represent the lower and upper bounds of the confidence interval.

The variations in $PM_{2.5}$ concentrations can be attributed to external factors such as changes in weather conditions, heavy vehicular traffic, industrial activities, or natural events like fires. The notable fluctuations indicate a dynamic environment, likely to be influenced by intermittent sources of pollution. The actual $PM_{2.5}$ time series shows significant variations, especially at the beginning, with values ranging between 12 and 22 $\mu g/m^3$. This behavior suggests that air pollution is subject to abrupt changes due to environmental influences like variable weather, vehicle traffic, or local industrial activities.

The dotted red line in **Figure 8b** represents the model's prediction and suggests that $PM_{2.5}$ concentrations stabilize over time, showing more consistent values than the actual data. In contrast, the actual $PM_{2.5}$ concentrations exhibit fluctuations around a steady average, reflecting the inherent variability in air quality. While the ARIMA forecast captures this stability, there is a noticeable widening of the confidence intervals as time progresses, indicating increasing uncertainty in longer-term predictions.



(**a**) (**b**) (**c**)

Figure 8. (**a**) Time series of $PM_{2.5}$ concentration (particulate matter with a diameter of less than 2.5 micrometers) in $\mu g/m^3$ over time. (**b**) The actual data is represented by the blue line, while the model forecast is indicated by the dotted red line. (**c**) Comparison between actual and forecasted $PM_{2.5}$ values, including confidence intervals.

The model provides reasonably accurate short-term forecasts, with confidence intervals indicating manageable uncertainty. However, as the forecast extends into the future, these intervals widen significantly, suggesting that long-term predictions are less reliable. The model's predictions do not seem to adequately account for the peaks and valleys observed in real data, likely due to the smoothing effect of the modeling process, which may overlook extreme pollution events. At the beginning of the data series, the actual $PM_{2.5}$ concentrations show significant variability. This may reflect the influence of external factors such as climate anomalies or environmental events like wildfires or dust storms. The ARIMA model may not fully capture these influences due to its assumptions of linearity and stationarity in the data. The growing uncertainty in the predictions could indicate the model's limitations in accounting for longer-term variations, especially if external factors, such as climate change or human activities, are influencing the data in ways that the model does not consider.

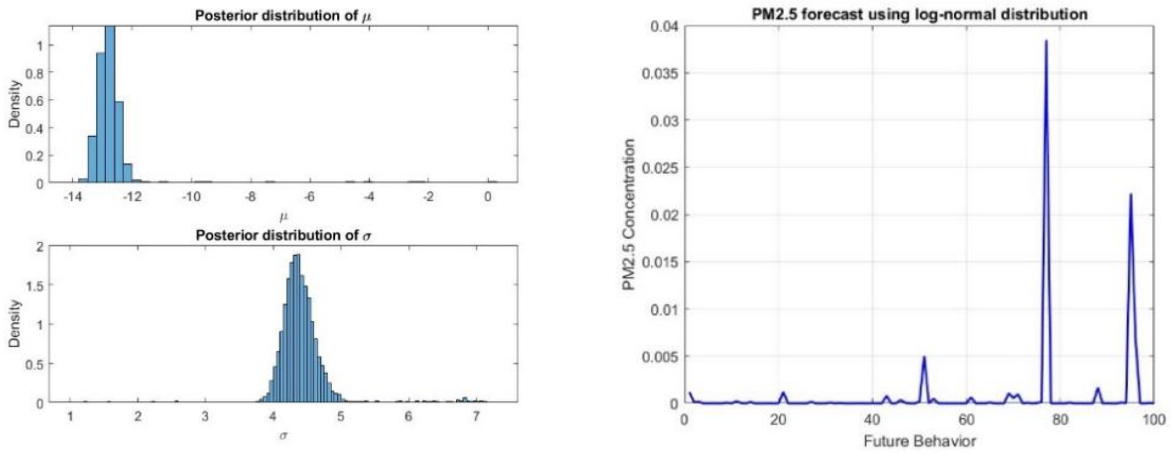
From a specific point in the series, the dotted red line begins to represent the model's prediction, indicating a gradual stabilization of $PM_{2.5}$ concentration around 16 $\mu g/m^3$,

with a slight increase over time. After the initial fluctuations, the ARIMA model's forecast stabilizes, suggesting that the model smooths out early variability. However, this stabilization may not fully capture real-world complexities, particularly when sudden, short-term changes occur. The static nature of the forecast could be due to the model's tendency to over-smooth the data, potentially missing transient events that impact air quality. To enhance the model's ability to capture extreme events and improve prediction accuracy, it is necessary to adjust the model parameters, incorporate exogenous variables, or integrate machine learning techniques. This would account for seasonal or external factors influencing $PM_{2.5}$ concentrations.

The observed stationary behavior of the series implies that it maintains a consistent meaning and variance over time, supporting the assumption of stationarity in ARIMA models. The absence of significant trends or seasonal cycles indicates that the time series behaves predictably within certain limits. If the assessment in 2019 closely mirrors the overall pattern of previous years, it further reinforces the stationary nature of the data, suggesting that no abrupt changes or external shocks significantly affected the series during that period.

3.3. Bayesian inference with Log-Normal Distribution

For the analysis of the positive skewness of the data, as can be observed in **Figure 9a**, most $PM_{2.5}$ values remain low, but significant spikes are observed at specific points (approximately at values 80 and 95 on the X-axis). This suggests that although the concentration of $PM_{2.5}$ is predominantly low, there are sporadic high-intensity events that affect pollution. For this reason, the log-normal distribution is suitable for modeling positive and asymmetric data. The results indicate that while most $PM_{2.5}$ values remain low (**Figure 9b**), there is a significant probability of extreme events occurring, such as sharp spikes in pollution.



(**a**) (**b**)

Figure 9. (**a**) Distribution of daily $PM_{2.5}$ concentrations ($\mu g/m^3$), highlighting the positive asymmetry of the data. (**b**) Frequency distribution of $PM_{2.5}$ concentrations ($\mu g/m^3$), highlighting the predominance of low values and the probability of extreme events occurring, represented by sharp peaks of pollution.

In the log-normal distribution (**Figure 9b**), spikes can represent periods of intense pollution caused by factors such as adverse weather conditions, increased industrial activity, or wildfires. The concentration of $PM_{2.5}$ does not follow a linear pattern. The log-normal distribution suggests that small increases occur more frequently, while large peaks are rare but can occur. Predictions made with this approach can be useful for public policies, as they indicate that, despite long periods with low pollution, it is essential to maintain constant monitoring to prevent environmental crises and respond quickly to high-pollution events.

Figure 10 illustrates the distribution of the original $PM_{2.5}$ concentration values. A high density concentrated around values close to zero is observed, indicating an asymmetric distribution and the predominant presence of very low or even null values. This behavior suggests a large dispersion

in the data, which may indicate the need for transformation, such as normalization, or the use of more appropriate distributions, such as log-normal or Weibull, for more effective modeling.

In addition, the predominance of low values may reflect a censorship effect on the data, possibly due to the detection limits of the sensors. This characteristic should be considered when applying statistical methods or predictive models, and it may be necessary to use techniques such as the removal of outliers or transformations to improve the suitability of the data to parametric methods.

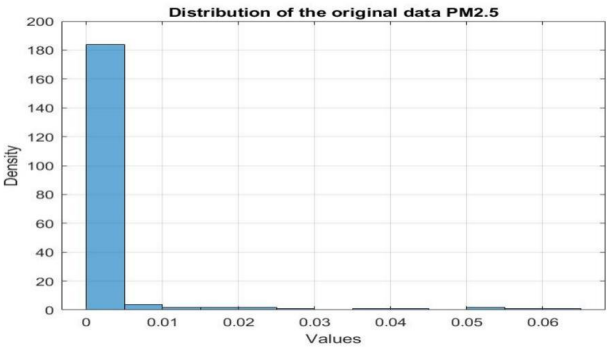


Figure 10. Distribution of the original $PM_{2.5}$ concentration values.

3.4. Comparative Analysis of Distributions

Three distributions were tested to model the observed data: Log-Logistic, Generalized Extreme Value (GEV), and Log-Normal. These distributions are indicated for asymmetric data and are commonly used in environmental and climatic phenomena, especially those with extreme values and long tails.

All distributions have important characteristics for the study case. Log-Normal Distributions assume that the logarithms of data follow a normal distribution, often used for environmental variables with exponential growth. Log-Logistic Distributions is Similar to Log-Normal but with heavier tails, ideal for modeling frequent extreme events in hydrology and precipitation. While, GEV (Generalized Extreme Value) Distributions focus more on extreme events in the upper tail, capable of Gumbel, Fréchet, or Weibull behaviors for flexible modeling.

When adjusting the data to the tested distributions, it was observed that the probability values obtained were similar between the models, suggesting that they all provide a good fit to the data. However, some important differences deserve to be highlighted: The Log-Normal distribution captured the asymmetry of the data well, but may underestimate the frequency of extreme values due to the less heavy tail when compared to the Log-Logistic and GEV distributions; The GEV, as it specializes in modeling extreme events, was the most effective in capturing the occurrence of high values. For forecasts involving extreme events, such as pollution peaks or heavy rainfall, the GEV would be the most suitable distribution.

The high value of $\sigma = 4.42$ in Log-Normal indicates a large dispersion of the data. Compared to Log-Logistic, which also has the flexibility to model dispersion, the ideal choice depends on the adherence of the distribution to the tail of the data. For overall data modeling, Log-Normal offers a good fit, capturing well the asymmetry and central distribution of the observed values. For extreme event modeling, the GEV is the best option, being specifically designed to capture very high values. Log-Logistic can be an intermediate alternative, since it combines characteristics of the previous distributions, being useful when the data has heavy tails, but without necessarily involving rare extreme events.

3.4.1. Gaussian PDF adjustment

Figure 11a illustrates the application of the Gaussian distribution function (normal) to the concentration of $PM_{2.5}$ (fine particulate matter) between 2000 and 2018. The distribution of $PM_{2.5}$ data suggests that, although most measurements focus on low values, there are rare episodes of severe pollution. This pattern is useful for predicting trends and guiding environmental policies to mitigate extreme pollution events. The highest probability density is associated with $PM_{2.5}$ concentrations below 50 $\mu g/m^3$, which indicates that most of the data was collected at relatively low concentrations. There are concentrations above 100 $\mu g/m^3$, although rare, with some values reaching up to 400 $\mu g/m^3$, suggesting exceptional episodes of high pollution. The distribution is right-skewed, indicating rare pollution spikes. Low concentrations suggest improved air quality, but extreme events like wildfires need attention.

The histogram of $PM_{2.5}$ concentrations between 2000 and 2018 (**Figure 11b**), with an adjusted Gaussian probability density function (PDF) superimposed. Although the Gaussian fit represents the core values well, it does not adapt adequately to the tails, implying that alternative distributions must be considered for more accurate modeling and fore-

casting. Most samples have $PM_{2.5}$ concentrations below 50 $\mu g/m^3$, indicating that the air quality was within acceptable standards for most of the period. There are some values higher than 100 $\mu g/m^3$, but rare, possibly related to specific high-pollution events, such as fires or industrial emissions. The histogram shows a right tail, suggesting positive skewness and a poor fit with the Gaussian distribution. Alternative distributions, like log-normal, may fit better.

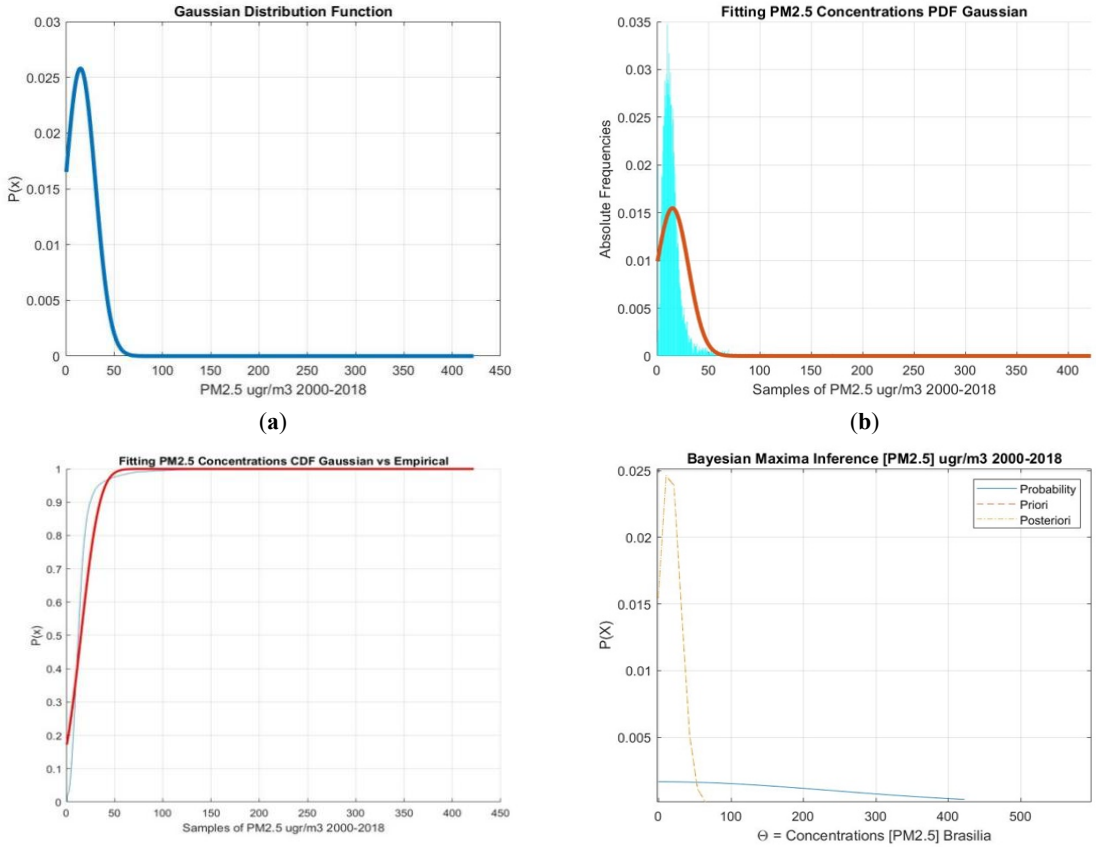


Figure 11. (**a**) "Gaussian Distribution of $PM_{2.5}$ Concentration between 2000 and 2018". (**b**) Adjustment of Gaussian Distribution to $PM_{2.5}$ Concentrations (2000–2018)". (**c**) Adjustment of the Cumulative Distribution Function (CDF) of $PM_{2.5}$ Concentrations: Gaussian vs. Empirical (2000–2018). (**d**) Prior (A Priori Distribution) – Represents the initial assumption about the distribution of maximum PM_2 values, before the data is incorporated. -Posterior (Posterior Distribution) – Represents the adjusted distribution after the incorporation of the observed data.

Empirical CDF (probably in light blue) based on observed data and Adjusted Gaussian CDF (probably in red) fitted to the data via a normal distribution can be observed in **Figure 11c**. The presence of extreme values may indicate sporadic episodes of high pollution, possibly associated with fires or thermal inversions. Environmental policies should focus not only on averting concentrations, but also on mitigating these critical events, which can have severe impacts on public health. The form of the CDF shows that most concentrations of $PM_{2.5}$ are concentrated at low values, with a rapid growth of the accumulated function up to about 100 $\mu g/m^3$. Above this threshold, the probabilities stabilize, indicating that very high concentrations are rare events. This suggests that the Gaussian distribution may be adequate to represent the central variability of the data but may underestimate extreme pollution events.

Figure 11d shows the Bayesian inference for the maximum values of PM_2 concentration. in Brasília in the period from 2000 to 2018, highlighting the a priori and a posteriori distribution of extreme values. Where the priori distribution (dashed line) presents an initial assumption about the maximum possible values and, the posteriori distribution (continuous line) reflects the updating of this assumption based on the observed data.

The a priori distribution represents the initial assumption about the maximum values of $PM_{2.5}$ before incorporation of observational data. The posteriori distribution, obtained after the Bayesian update, reflects the influence of real data on the modeling of extremes. It is observed that the posterior distribution presents a more accentuated displacement and adjustment, indicating that the collected data significantly modified the initial predictions about the maximum concentrations.

Another relevant aspect is the presence of a long tail in the posterior distribution, which suggests the existence of sporadic extreme events, with concentrations higher than 400 μg/m^3. These events can be associated with seasonal factors, such as fires, thermal inversions, and periods of atmospheric stability that favor the accumulation of pollutants. This pattern reinforces the need for continuous monitoring and air pollution control policies, especially at critical times of the year.

The histogram (**Figure 12a**) shows a higher frequency of values between 100 and 300 μg/m^3, with a sharp peak around 250 μg/m^3, indicating that most $PM_{2.5}$ observations are concentrated in this range. The presence of a tail on the right suggests the occurrence of extreme events, with elevated $PM_{2.5}$ values, although less frequent. The brown line, which represents the probability density function (PDF) adjusted by the distribution of extreme values, shows that statistical modeling adequately captures the asymmetry of the distribution and the elevated $PM_{2.5}$ values. High concentrations of $PM_{2.5}$ may be related to fire events, thermal inversion, and emissions from industrial and vehicular sources.

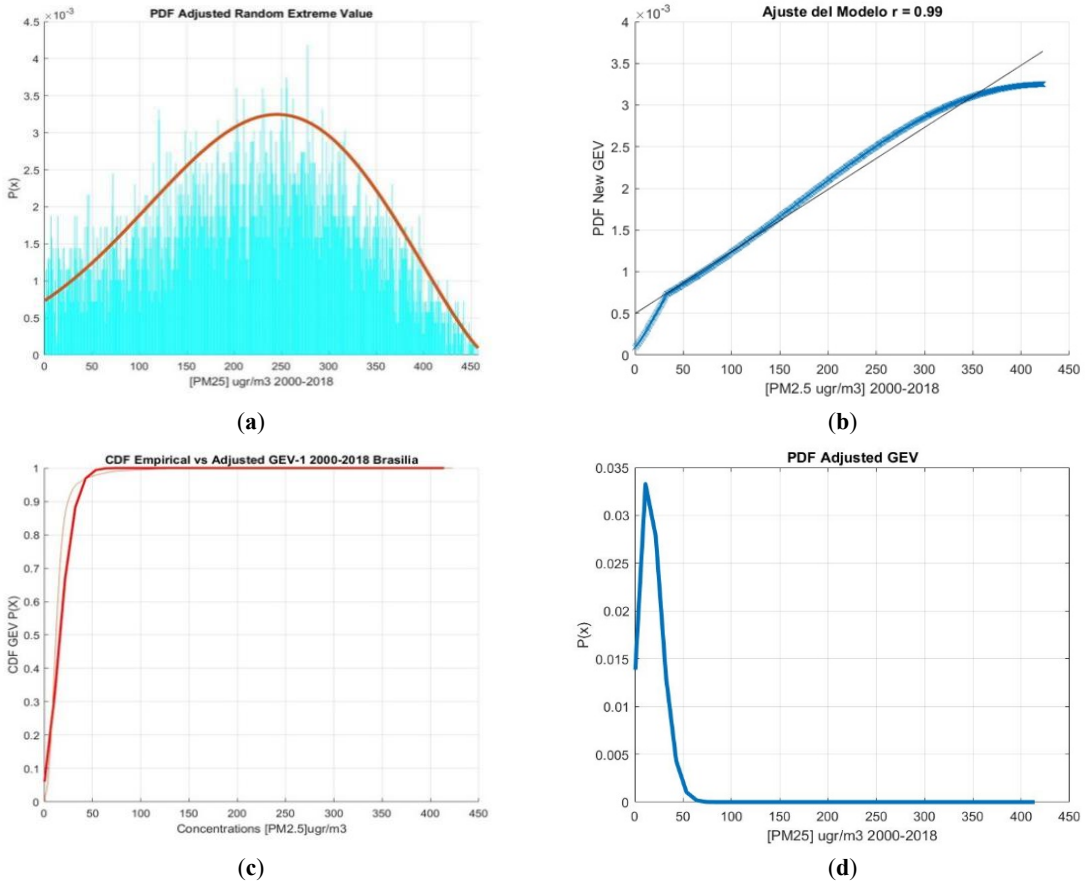


Figure 12. (**a**) Histogram of $PM_{2.5}$ concentration (μg/m^3) in the period from 2000 to 2018, with the adjustment of a probability density function (PDF) based on the distribution of extreme values (Extreme Value Distribution). (**b**) Statistical model adjustment to a distribution of $PM_{2.5}$ (μg/m^3) in the period 2000–2018. (**c**) Comparison between the empirical cumulative distribution function (CDF) and the adjusted CDF of the Generalized Distribution of Extreme Values (GEV) for $PM_{2.5}$ concentrations (μg/m^3) in Brasília in the period 2000–2018. The adjusted curve (red line) shows good adherence to the empirical data. (**d**) Adjusted Probability Distribution (PDF) of $PM_{2.5}$ Concentration in the period 2000–2018 using the Generalized Extreme Value Distribution (GEV). The adjustment suggests that the data follows asymmetric behavior, with a longer tail on the right, indicating events of high $PM_{2.5}$ concentration.

The fit of a statistical model for $PM_{2.5}$ concentrations in the period 2000–2018 (**Figure 12b**), with the blue line representing the observed data and the grey line the adjusted GEV model. The value of r = 0.99, r = 0.99, r = 0.99 suggests that the fit of the GEV model is excellent, indicating that the adjusted distribution accurately describes the data, especially at the highest concentrations. The high value reveals that the GEV model very accurately describes the highest concentrations of $PM_{2.5}$, which is essential for the prediction of extreme events. The excellent fit suggests that the GEV distribution can be used to predict extreme air pollution events, which have major implications for environmental policymaking and pollution mitigation actions.

The empirical cumulative distribution function (CDF) (in black) with the CDF adjusted by the GEV distribution (in red) (**Figure 12c**). The excellent grip between the two corners indicates a high-quality fit. The curve shows a sharp growth at first, indicating that most $PM_{2.5}$ concentrations are in lower ranges. Extreme concentrations, on the other hand, are less frequent, which is in line with the observed distribution. The almost complete overlap between the empirical CDF and the CDF adjusted by the GEV distribution suggests that the model describes the distribution of the data very well.

Figure 13 shows the QQ plot for the GEV 1 distribution fitted to $PM_{2.5}$ concentration data from 2000 to 2018. The plot serves as a graphical diagnostic tool to assess the adequacy of the statistical model in representing the empirical data. In the central portion of the distribution (approximately quantiles 10 to 50), the points lie relatively close to the 1:1 line, suggesting that the GEV 1 distribution provides a reasonable fit for typical $PM_{2.5}$ concentrations.

However, significant deviations are observed in the upper quantiles (above quantile 60), where the points rise steeply above the reference line. This behavior indicates that the GEV 1 model underestimates extreme values, failing to adequately capture high $PM_{2.5}$ concentration events—an essential concern for public health risk assessments and air quality management strategies.

When compared to other distributions tested in the study, the Log-Logistic distribution also showed a good fit for the central values but displayed similar limitations in the upper tail. The Log-Normal distribution, while achieving the highest coefficient of determination ($R^2 = 0.6689$), likewise showed deviations in the extremes, as evidenced in its respective QQ plot, indicating it may not be ideal for modeling rare pollution events.

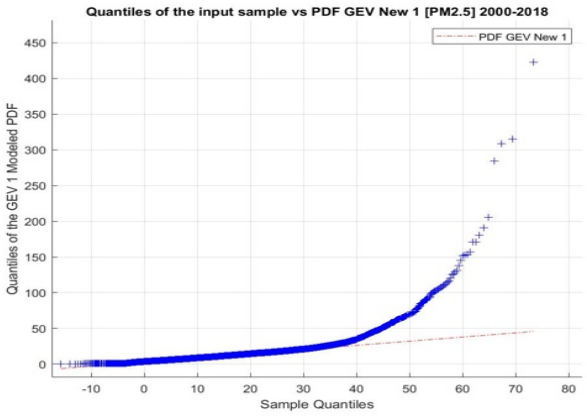


Figure 13. Quantile-Quantile (QQ) plot comparing the empirical quantiles of daily $PM_{2.5}$ concentrations in Brasília (2000–2018) with the theoretical quantiles of the Generalized Extreme Value distribution (GEV 1). The red dashed line represents the 1:1 reference line, indicating perfect agreement between the observed and theoretical quantiles.

In contrast, the GEV 2 distribution, which was fitted specifically to capture the behavior of extreme values, showed superior performance in the upper tail of the distribution. Its heavier tail provided a better representation of high-concentration events, making it a more robust alternative for extreme value modeling.

Therefore, the QQ plot in **Figure 13** highlights the importance of selecting statistical models based on the intended purpose of the analysis. While GEV 1 may be suitable for describing the general behavior of $PM_{2.5}$ data, distributions like GEV 2 are more appropriate when the objective is to predict or understand extreme pollution events.

GEV 2

The probability density function adjusted by the Generalized Extreme Value (GEV-2) distribution (**Figure 14a**) for $PM_{2.5}$ concentration data between 2000 and 2018. The distribution shows a positive asymmetric shape, typical of GEV, with a peak around 50–60 $\mu g/m^3$, indicating the most frequent concentration of $PM_{2.5}$. After this point, the density decays rapidly, but with a long tail, suggesting the presence of extreme events with high concentrations of $PM_{2.5}$. This behavior is indicative that, although most of the data is concentrated in moderate concentrations, extreme pollution events still occur with some frequency. The GEV model

seems to be adequate to represent the distribution of $PM_{2.5}$ concentrations, but its validity must be confirmed by statistical goodness-of-fit tests. The long tail suggests the need for specific strategies to mitigate extreme pollution events, such as fires and weather conditions that favor the concentration of pollutants.

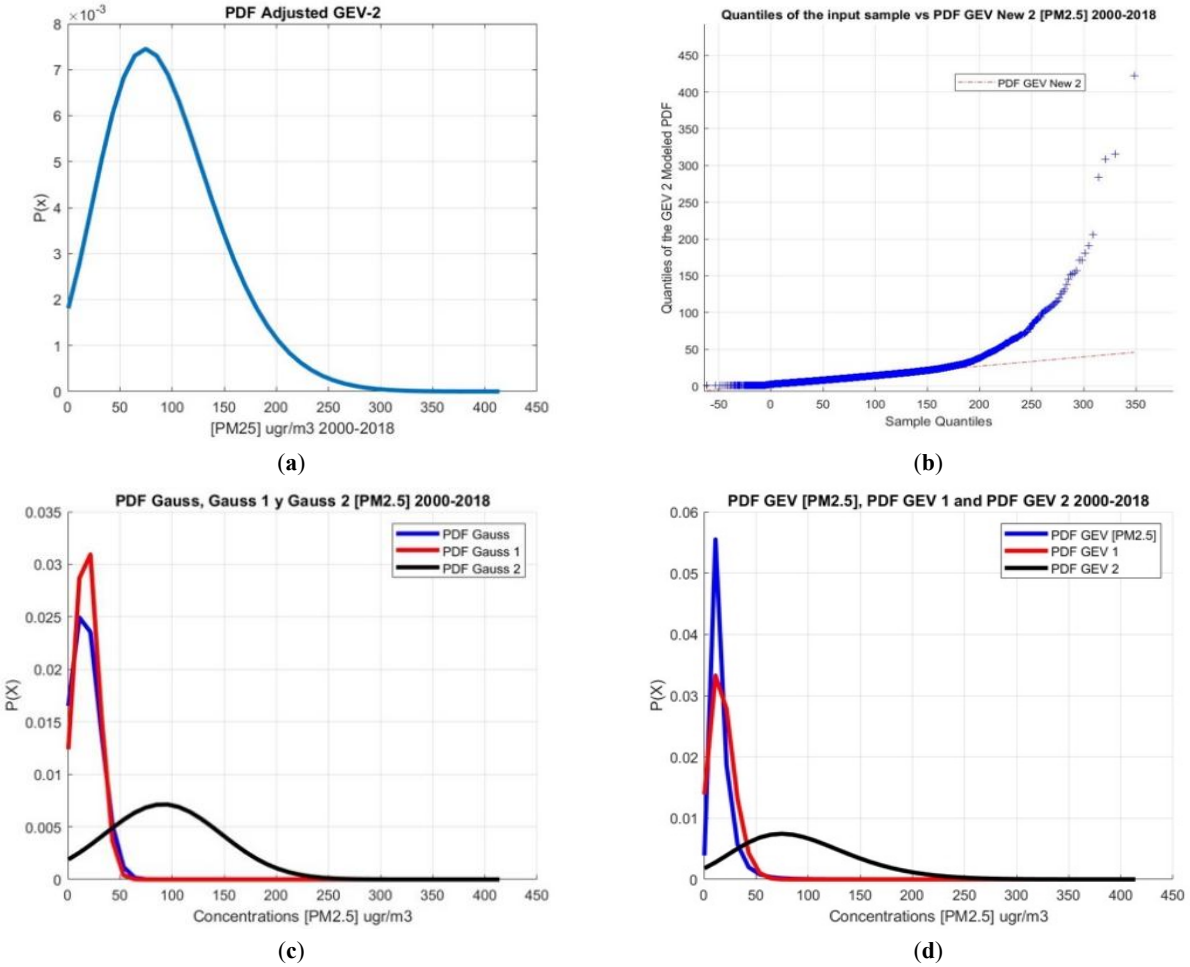


Figure 14. (**a**) Adjusted Probability Density Function (PDF) of the Generalized Extreme Value (GEV-2) distribution for $PM_{2.5}$ concentration in the period 2000–2018. (**b**) Quantile-Quantile (QQ) graph comparing the quantiles of the $PM_{2.5}$ sample (2000–2018) with the Generalized Extreme Value (GEV-2) distribution. (**c**) probability distribution (PDF) of $PM_{2.5}$ concentrations for the period 2000–2018, adjusted for three different Gaussian distributions (Gauss 1, Gauss 2 and Gauss 3). (**d**) Probability density function (PDF) adjusted for $PM_{2.5}$ concentrations for the period 2000–2018, using Generalized Extreme Value (GEV) distributions. The curves represent different fits: GEV ($PM_{2.5}$) in blue, GEV 1 in red and GEV 2 in black.

The predominance of concentrations around 50–60 $\mu g/m^3$ may indicate a chronically high level of pollution, with implications for public health, since values above 25 $\mu g/m^3$ are considered harmful to health, according to the WHO. In addition, the presence of extreme events, represented by the long tail of the distribution, highlights the need for public policies focused on air pollution control.

The QQ graph (**Figure 14b**) evaluates the adequacy of the adjustment of the GEV-2 distribution to the $PM_{2.5}$ concentration data. Although GEV-2 provides a reasonable fit for most $PM_{2.5}$ data, its limitation in modeling extreme events suggests the need for refinement in the model or consideration of other distributions. Modeling rare and extreme events is essential to account for the variability of $PM_{2.5}$ concentrations, and strategies to deal with these episodes should be implemented. From quantile 50 onwards, a significant deviation of the points from the reference line is observed, indicating that the GEV-2 underestimates the extreme values of $PM_{2.5}$. This deviation is most evident in the upper tail, where the sample quantiles grow sharply in relation to the values predicted by the adjusted distribution. In the lower and intermediate quantiles, the dots closely follow the refer-

ence line (dashed), suggesting that the GEV-2 model has a good fit for the data from the central part of the distribution, i.e., for the most common concentrations of $PM_{2.5}$.

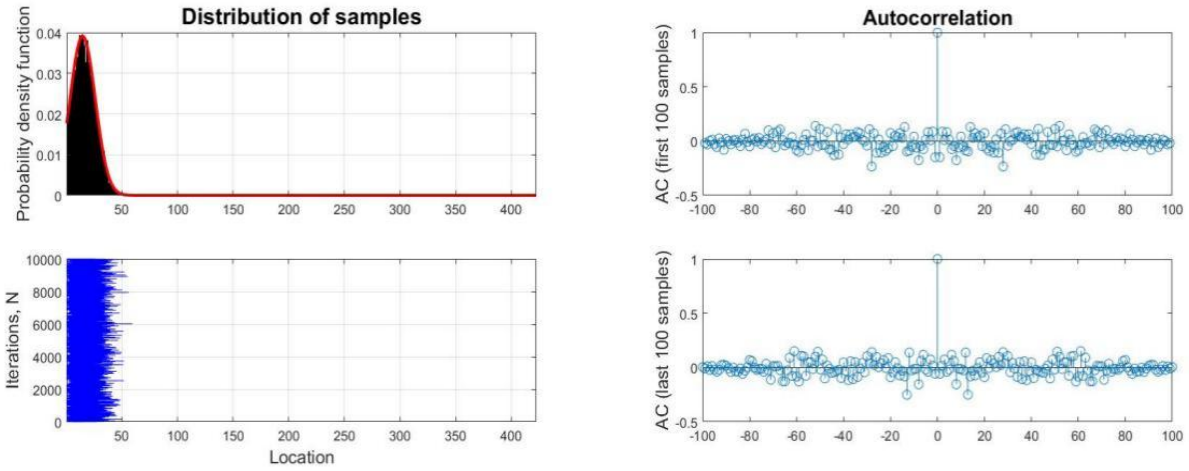
The graph in **Figure 14c** compares three Gaussian distributions to model the $PM_{2.5}$ data. Gauss 1 and Gauss 2 (red and blue), both have a sharp peak around 50 $\mu g/m^3$, which indicates that this range concentrates most of the $PM_{2.5}$ observations. Gauss 3 (black) presents a more pronounced asymmetrical to the right, indicating the presence of extreme events of high $PM_{2.5}$ concentration, although with a lower probability. The overlap between the Gauss 1 and Gauss 2 distributions suggests that both adequately model the main part of the data distribution, especially for the most frequent values. Gauss 3, with a long tail, represents a distribution that can capture severe pollution events. The predominant concentration of $PM_{2.5}$ around 50 $\mu g/m^3$ may indicate chronic pollution, with possible impacts on public health. The long tail of Gauss 3 reinforces the need to investigate and monitor high-concentration events, such as fires or adverse weather episodes.

The graph in **Figure 14d** compares the GEV distribution with Gaussian distributions for the $PM_{2.5}$ data. The blue distribution (GEV $PM_{2.5}$) shows a very sharp peak around 50 $\mu g/m^3$, indicating that most of the data is concentrated in this value. The red distribution (GEV 1) follows a similar pattern, but with a smoothing in the peak region. The black distribution (GEV 2) has a more pronounced long tail, which suggests that this model attempts to capture extreme events of high $PM_{2.5}$ concentration. This suggests that episodes of intense pollution can be sporadic and are linked to specific atmospheric conditions, such as fires or thermal inversions. A high concentration of $PM_{2.5}$ around 50 $\mu g/m^3$ may indicate a chronic air pollution problem.

The comparison of Bayesian Inference models using Log-Normal and Generalized Extreme Value (GEV) distributions revealed key differences in modeling $PM_{2.5}$ concentrations. GEV 1 captured lower concentrations well but was variable at higher levels. In contrast, GEV 2 provided a better overall fit, particularly at higher concentrations, making it more effective in modeling extreme pollution events and indicating a greater likelihood of severe pollution in the future. The discrepancy between these models becomes evident in the QQ plots, where GEV 2 stands out in its ability to model high-concentration events. However, it's essential to note that both GEV models act as predictors and don't necessarily offer an exact fit to the input data, underlining the importance of complementary validation methods. Furthermore, these observations align with results from the ARIMA model and traditional Bayesian Inference, reinforcing the robustness of the methods employed.

The findings from this analysis emphasize the significance of considering a variety of statistical approaches for modeling air pollution. While GEV 2 serves as a robust predictor for extreme events, Log-Normal Bayesian Inference remains effective for modeling the general data structure. The incorporation of the Metropolis-Hastings algorithm through MCMC represents a crucial step forward, providing a more nuanced probabilistic approach to predict future $PM_{2.5}$ patterns and extreme air pollution events. Since GEV models are primarily used for predicting extreme events, the next step in the analysis involves using the Metropolis-Hastings algorithm, a key technique within Bayesian Inference and for sampling from complex distributions. This algorithm is a variant of the Markov Chain Monte Carlo (MCMC) method, commonly used when the target distribution cannot be directly sampled, such as with GEV distributions and other density functions lacking an explicit analytical form.

Figure 15a shows the distribution of the samples in two distinct panels. The histogram (black bars) displays the frequency of samples at different locations, and the red line represents the probability density function (PDF). The asymmetry of the distribution may indicate that the variable analyzed presents extreme values or outliers in small regions. The behavior of the PDF suggests that statistical models such as the Weibull distribution, log-normal, or gamma may be appropriate for fitting the data. The uniformity of the number of samples per location in the second panel confirms that data collection was homogeneous, avoiding significant sampling bias. If the context is environmental, this distribution may be associated with measurements of pollutants or climatic variables with high spatial and temporal variability.



(**a**) (**b**)

Figure 15. a-Sample distribution: (**a**) Probability density function (PDF) of the samples (red line) superimposed on the histogram (black bars); (**b**) Autocorrelation function (AC) for different segments of the time series: (**a**) Autocorrelation of the first 100 samples; (**b**) Autocorrelation of the last 100 samples.

The bottom panel shows that the distribution of the samples seems uniform in terms of quantity at each location. This pattern indicates that data collection was done consistently across the study domain, ensuring a significant number of sampling points. The uniformity of the number of samples per location confirms that data collection was homogeneous, avoiding significant sampling bias. If the context is environmental, this distribution may be associated with measurements of pollutants or climatic variables with high spatial and temporal variability.

The autocorrelation function (AC) (**Figure 15b**) for two subsets of data in a time series, highlighting the correlation between different lags. The absence of significant correlation suggests that the data may be approximately independent in time, which may indicate a stochastic process with no memory structure. For the top panel most autocorrelation values are close to zero, indicating a significant absence of correlation for different lags. For the bottom panel the pattern is similar, with no clear temporal dependency structure. CA values continue to be randomly distributed along lags, with no indication of seasonal patterns or trends. This lack of correlation may suggest that the phenomenon studied does not have strong temporal dependence or that it was correctly pre-processed to remove trends and seasonality. Also, the lack of meaningful autocorrelation may indicate that methods based on autoregressive models (such as ARIMA) may not be ideal, and the use of non-autoregressive approaches, such as neural networks or machine learning-based models, is preferable.

The results indicate that the GEV 1 model exhibits a behavior similar to Bayesian inference with the Log-Normal distribution, providing values close to zero and displaying long tails in the PDF. This model fits better with $PM_{2.5}$ data at low concentrations but has high variability for high concentrations. This characteristic can be seen in the QQ chart.

On the other hand, the GEV 2 model performs better when considering the full set of data, especially for higher concentration values. This suggests that there is a higher probability of future events having elevated $PM_{2.5}$ concentrations. However, GEV 2 does not fit the initial sample values as well, as both GEV models function as predictors. These findings corroborate the results obtained with both ARIMA and standard Bayesian Inference.

The next step will be the implementation of the Metropolis-Hastings algorithm, one of the most widely used methods in Bayesian inference and in the sampling of probability distributions. The method makes use of Markov Chain Monte Carlo (MCMC), allowing the generation of samples of complex distributions, especially when direct sampling is not feasible.

3.5. Sampling Methodology: Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm, part of the Monte Carlo methods via Markov Chain Monte Carlo (MCMC),

was used in this study to estimate the parameters of complex probability distributions, especially the Generalized Extreme Value (GEV) distribution. The choice of this algorithm is justified by the fact that, in many cases, such as the GEV distribution, it is not possible to directly sample from its posterior distribution, as it lacks a simple analytical form.

The main advantage of the Metropolis-Hastings algorithm is its ability to generate samples from complex distributions through a Markov chain, allowing parameter estimation in a Bayesian context even when the likelihood function is intractable. In this study, the method was applied after formulating the Bayesian model with Log-Normal and GEV distributions, enabling robust sampling to predict extreme $PM_{2.5}$ concentrations.

The application of the algorithm aimed to:

- Increase the robustness of parameter estimates for the GEV models, especially in the tails of the distribution where extreme events occur;
- Incorporate uncertainty into the forecasts of critical pollutant concentrations;
- Allow comparison between classically fitted distributions (MLE) and those estimated by Bayesian inference with MCMC.

The results obtained with Metropolis-Hastings were consistent with observed values and allowed a more faithful representation of data asymmetry and dispersion. This approach is especially relevant in environmental modeling, where extreme events, although rare, have a significant impact on public health and air quality control strategies.

The application of the Metropolis-Hastings algorithm enabled the generation of samples from the posterior distributions of the parameters of the GEV and Log-Normal models, allowing for more precise estimation of $PM_{2.5}$ extremes in Brasília between 2000 and 2018. **Figure 15a** shows the distribution of the generated samples, highlighting the positive skewness typical of extreme environmental events.

Figure 15b presents the autocorrelation function (ACF) for the first and last 100 values of the sample chain. The low autocorrelation observed indicates good efficiency of the Markov chain, signaling convergence and relative independence among the samples. This confirms the quality of the simulations carried out by the algorithm and the robustness of the Bayesian inference achieved.

The comparison between the prior distribution (representing the initial knowledge about the parameters) and the posterior distribution (after incorporating the observed data) is illustrated in **Figure 16a**. The posterior curve shows a shift relative to the prior, with greater probability concentration in values associated with extreme episodes, such as wildfires and thermal inversions, demonstrating the impact of real data on the final estimate.

Figure 16b reinforces the good fit of the probability density function (PDF) generated by the GEV model adjusted via Metropolis-Hastings, which adequately captured the long tail of the distribution, associated with elevated $PM_{2.5}$ concentrations. This feature is essential for predicting rare and potentially critical public health events.

Moreover, the use of the Bayesian approach with MCMC sampling enabled the estimation of credible intervals for the parameters, providing a more comprehensive assessment of the uncertainty associated with the forecasts. This represents an advance over classical maximum likelihood approaches, which often underestimate variability in heavy-tailed distributions.

When a distribution cannot be sampled directly—as in the case of GEV, which does not have a simple analytical form—the Metropolis-Hastings algorithm makes it possible to generate samples that follow the desired distribution.

Figure 16a shows the Bayesian inference applied to the modeling of extreme values of $PM_{2.5}$ concentration in Brasília between 2000 and 2018. The Bayesian approach uses prior information (a priori distribution) to analyze maximum $PM_{2.5}$ concentrations, leading to a posterior distribution that aligns with observed data. The a priori distribution indicates that high $PM_{2.5}$ values are rare, while the posterior distribution adjusts slightly to fit extreme values better. The blue curve illustrates the adjusted probability function with a long tail, indicating that extremely high $PM_{2.5}$ events, though infrequent, are possible. This behavior is characteristic of extreme value distributions like the Generalized Extreme Value (GEV) or Pareto distribution. This long tail underscores the importance of ongoing air quality monitoring in Brasília, especially during dry seasons or wildfires. Further research can explore how meteorological and human factors affect maximum $PM_{2.5}$ concentrations.

Probability density function (PDF) shown in **Figure 16b**, adjusted using the GEV distribution for $PM_{2.5}$ concen-

trations (μg/m³) in the period 2000–2018. The PDF shows a positive asymmetry (right tail), indicating that most $PM_{2.5}$ concentrations are at low values, while extreme events of high concentration are less frequent, but possible. The modal value of the distribution (peak of the curve) occurs around 20–30 μg/m³, suggesting that this concentration range was the most common during the analyzed period. The long tail of the data shows that most $PM_{2.5}$ values are low, but there are occasional high spikes, likely from wildfires or weather events. The Generalized Extreme Value (GEV) distribution effectively models these extremes. While low to moderate levels are generally good for air quality, the extreme spikes are concerning, as high $PM_{2.5}$ exposure can harm public health.

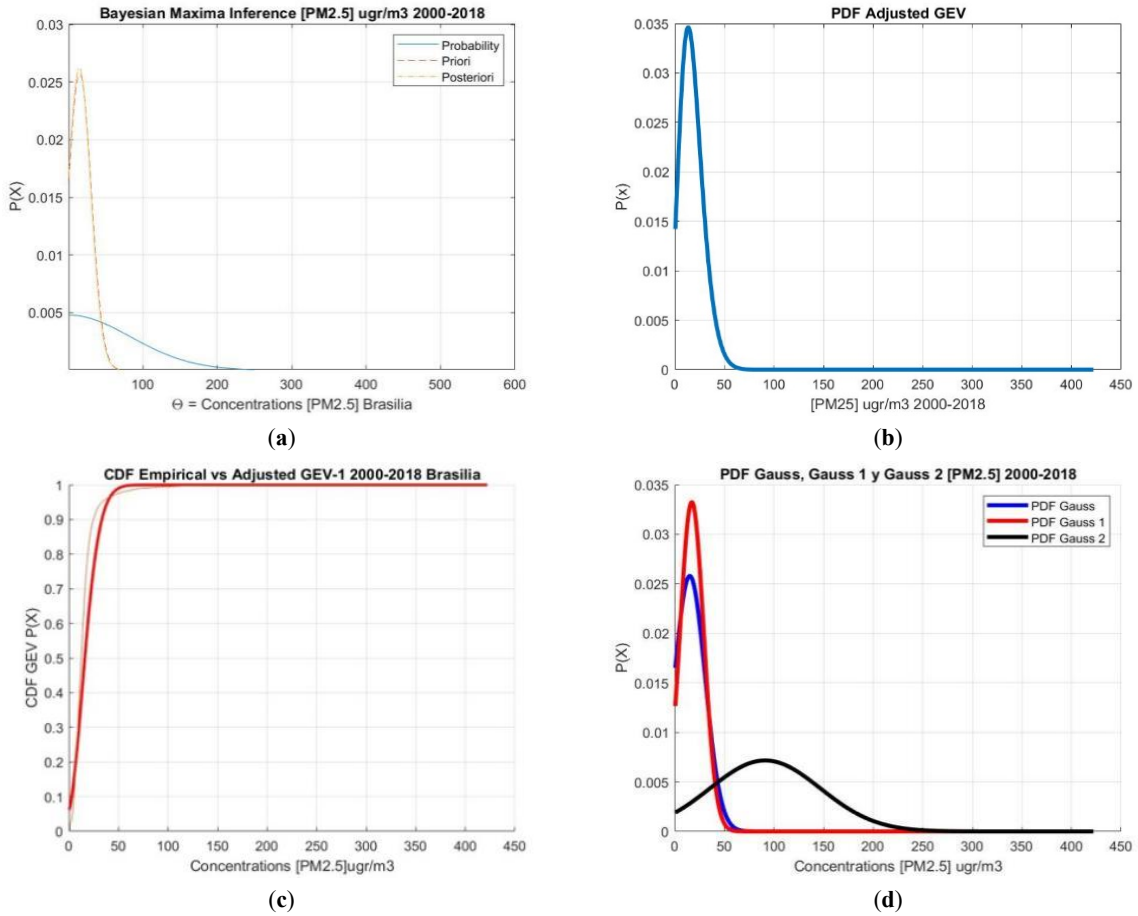


Figure 16. (**a**) Bayesian inference for maximum $PM_{2.5}$ concentration values in Brasília in the period 2000–2018. The figure shows the priori (brown dashed line) and a posteriori (green dashed line) distributions, in addition to the adjusted probability distribution (blue line). (**b**) Probability density function (PDF) adjusted using the GEV distribution for $PM_{2.5}$ concentrations (μg/m^3) in the period 2000–2018. (**c**) Comparison between the empirical Cumulative Distribution Function (CDF) and the CDF adjusted by the GEV distribution for $PM_{2.5}$ concentrations (μg/m^3) in the period 2000–2018 in Brasília. (**d**) Probability Density Function (PDF) adjusted for $PM_{2.5}$ concentrations in Brasília between 2000 and 2018, comparing different Gaussian distributions.

The CDF curve (**Figure 16c**) exhibits rapid growth at low concentrations of $PM_{2.5}$, indicating that most values are concentrated in this range. The curve stabilizes for higher concentrations, which reflects the low frequency of extreme events. The CDF shows the cumulative probability of finding $PM_{2.5}$ values below a given threshold. For example, if the curve reaches 0.9 around 50 μg/m³, this indicates that 90% of the observations were below this value. Asymmetry in the tail suggests that there are some unusually high concentrations, which, while rare, can have significant impacts on air quality and public health. The proximity between empirical and adjusted CDF reinforces the adequacy of the GEV distribution to model the variability of $PM_{2.5}$ concentrations.

In **Figure 16d** 3 different representation lines can be observed, PDF Gauss as a blue line, PDF Gauss 1 as a red line, and Gauss 2 PDF as black line. The graph shows that the dis-

tribution of $PM_{2.5}$ is asymmetric on the right, with a long tail that extends to high concentration values (above 300 μg/m³). This indicates the presence of extreme air pollution events, possibly related to fires or industrial and vehicular emissions. The adjustment of the distributions suggests that the concentration of $PM_{2.5}$ can be well represented by a combination of Gaussian distributions, possibly reflecting different emission regimes. The predominance of lower concentrations, but with the presence of extreme events, indicates that air pollution in Brasília may be associated with both continuous sources (traffic, industry) and sporadic and seasonal events (fires).

The Gauss 1 PDF curve (red) shows a steeper peak and closely follows the empirical distribution for lower concentrations (between 0 and 100 μg/m³). The Gauss PDF curve (blue) also follows the trend of the empirical distribution, but with a less precise fit at the lower concentrations. The Gauss 2 PDF curve (black) represents a second Gaussian distribution, which better captures the tail of the distribution, i.e., the extreme values. The occurrence of extreme events suggests the need for continuous monitoring of air quality and preventive measures to mitigate episodes of high pollution, especially during the dry season.

Figure 17 shows the probability density functions (PDF) adjusted with the Generalized Extreme Value (GEV) distribution for $PM_{2.5}$ concentrations in the period from 2000 to 2018. The curves indicate an asymmetric distribution to the right, with a long tail, suggesting the occurrence of extreme events, i.e., high concentrations of $PM_{2.5}$ that appear with low frequency.

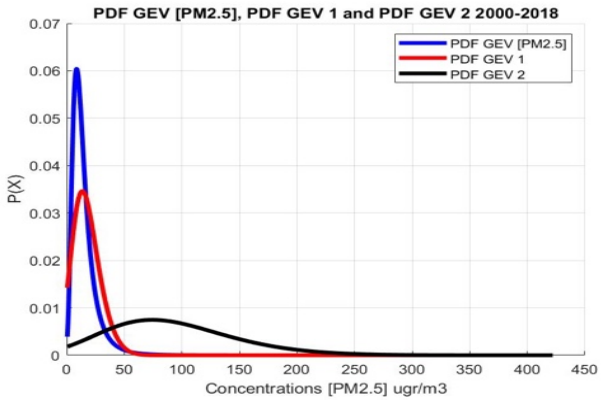


Figure 17. Probability density functions (PDF) adjusted using the Generalized Extreme Value (GEV) distribution for $PM_{2.5}$ concentration in the period 2000–2018. The curves represent different fits: PDF GEV for the $PM_{2.5}$ data (blue), PDF GEV 1 (red) and PDF GEV 2 (black).

The blue curve, representing the GEV fit for $PM_{2.5}$ data, shows a sharp peak at lower concentrations, which is consistent with the prevalence of reduced values observed in the original distribution. On the other hand, the red curve (PDF GEV 1) shows a similar behavior, but with a less pronounced peak. The black curve (PDF GEV 2) demonstrates a more pronounced elongation in the tail, indicating that this adjustment can more accurately represent extreme values.

The differences observed between the adjustments may reflect variations in the adequacy of the GEV distribution, depending on the parameterization used. The choice of the most appropriate model should be supported by statistical metrics, such as the Kolmogorov-Smirnov test, the AIC (Akaike Information Criterion) or the BIC (Bayesian Information Criterion), which allow the evaluation of which PDF best fits the $PM_{2.5}$ data.

In addition, the presence of extreme values reinforces the importance of investigating possible meteorological or anthropogenic factors associated with episodes of high PM2 concentration 5. This includes phenomena such as fires, thermal inversions, and seasonal emission patterns, which can play a crucial role in the occurrence of these extreme events.

The application of the improved methodology generated results consistent with previous analyses. The GEV 1 model presented the best fit for the data at low concentrations of $PM_{2.5}$, with well-distributed probabilities close to zero, a behavior that is reinforced by the Gaussian fit, also suggesting a good fit of the model for this concentration range.

On the other hand, the GEV 2 model, although it did not present the best global fit, demonstrated a higher probability for high $PM_{2.5}$ concentrations in future events. The presence of long tails in both GEV distributions indicates a higher probability of extreme events, which reinforces the need for models that take this characteristic into account in air quality forecasts.

To validate these results, three complementary statistical approaches were used:

ARIMA models, which analyze the time dependence of the series and suggest cyclical and trend patterns.

Bayesian inference with Log-Normal distribution, which captures the positive asymmetry of the data and allows for robust probabilistic estimates.

Methodology based on almost all Gaussian behavior, applied as a reference to evaluating the distribution of data.

The results indicate that the distribution of $PM_{2.5}$ concentrations does not follow purely Gaussian behavior, justifying the need for more advanced statistical models to capture its variability and predict air pollution episodes with greater accuracy.

These findings provide a solid scientific basis for the development of sustainable urban and rural policies, especially as it relates to air quality monitoring and public health interventions. By identifying the probability distributions that best model $PM_{2.5}$ concentrations, it is possible to improve environmental monitoring systems, allowing for more accurate detection of air pollution variations and trends. This information is crucial for the formulation of effective emission control strategies and for the implementation of preventive measures that protect the health of the population. In addition, the validation of models with official data reinforces the reliability of forecasts, helping public managers to make informed decisions and prioritize actions aimed at improving air quality and reducing the risks associated with exposure to fine particulate matter.

The adjusted statistical models—particularly the Generalized Extreme Value (GEV-2)—proved to be powerful tools not only for analyzing the distribution of $PM_{2.5}$ concentrations but also for predicting extreme pollution events. The robustness of the Bayesian inference, supported by the Metropolis-Hastings algorithm, improved the accuracy of estimates and the representation of uncertainties.

These models provide important support for public policy formulation by allowing the identification of critical periods with a higher risk of pollution peaks, such as the dry season [23,24]. The analysis showed that extreme values of $PM_{2.5}$ are not rare and tend to occur under specific atmospheric and seasonal conditions [25,26]. Thus, the implementation of targeted mitigation strategies can be guided by the outputs of these models.

We emphasize the public health impacts associated with elevated $PM_{2,5}$ levels, including respiratory and cardiovascular risks [1,27]. The study also highlights the importance of continuous air quality monitoring, especially during the dry season, when fire incidence and atmospheric stagnation increase pollutant concentrations [28].

Therefore, the integration of extreme value theory, time series models, and Bayesian inference creates a robust framework for understanding, forecasting, and managing air pollution risks in urban areas. Future studies may expand this approach by incorporating exogenous variables and machine learning techniques to further refine environmental forecasting.

4. Conclusions

Statistical analysis of $PM_{2.5}$ concentrations demonstrated that the distribution of the data is asymmetric and highly variable, justifying the use of advanced statistical models for more accurate predictions. The results indicate that the GEV 1 model is more suitable for low concentrations, while GEV 2 is more effective for predicting extreme events, suggesting that pollution peaks may occur more frequently than predicted by conventional distributions.

Validation of the models using ARIMA, Bayesian Inference with Log-Normal distribution and quasi-Gaussian modeling reinforced the robustness of the approach and the need for long-tailed distributions to adequately capture the behavior of air pollution. Furthermore, the findings suggest that extreme air pollution events are more frequent than a Gaussian distribution would indicate, highlighting the importance of specialized models for environmental predictions.

The choice of the statistical model should be guided by the specific objective of the prediction: Bayesian Inference with Log-Normal is useful for characterizing the general behavior of concentrations; ARIMA allows identifying seasonal patterns and trends; and GEV models are essential for predicting extreme events. The inclusion of sampling techniques, such as the Metropolis-Hastings algorithm, can improve the accuracy of estimates and contribute to more effective air quality planning.

For future studies, it is recommended to incorporate meteorological variables and emission sources to improve predictions. In addition, the combination of statistical methods and machine learning can offer a more comprehensive approach to modeling air pollution and support public policies aimed at mitigating environmental and public health impacts.

Author Contributions

Conceptualization, A.S., J.R., J.F. and K.R.; methodology, A.S., J.R., J.F. and K.R.; software, A.S., J.R., J.F. and K.R.; validation, A.S., J.R., J.F. and K.R.; formal analysis,

A.S., J.R., J.F. and K.R.; investigation, A.S., J.R., J.F. and K.R.; resources, A.S., J.R., J.F. and K.R. A.S., J.R., J.F. and K.R.; data curation, A.S., J.R., J.F. and K.R.; writing—original draft preparation, A.S., J.R., J.F. and K.R.; writing—review and editing, A.S., J.R., J.F. and K.R.; visualization, A.S., J.R., J.F. and K.R.; supervision, A.S., J.R., J.F. and K.R.; funding acquisition, J.F. All authors have read and agreed to the published version of the manuscript.

Funding

This study did not receive external funding.

Institutional Review Board Statement

This study does not involve human or animal subjects.

Informed Consent Statement

All authors have given their consent to participate in this study.

Data Availability Statement

All authors have approved the publication of this work.

Acknowledgments

The authors would like to express their sincere gratitude to the universities for their institutional support in the development of this research. We are also thankful to the administrative and academic teams of both institutions for providing the necessary infrastructure, technical assistance, and encouragement that significantly contributed to the successful completion of this work.

Conflicts of Interest

The authors declare no competing interests.

References

[1] World Health Organization (WHO), 2021. Air quality guidelines: Global update 2021. WHO: Geneva, Switzerland. Retrieved August 7, 2025, from https://www.who.int/publications/i/item/9789240034228

[2] Souza, A.D., Oliveira-Júnior, J.F.D., Cardoso, K.R.A., et al., 2025. The Impact of Meteorological Variables on Particulate Matter Concentrations. Atmosphere, 16(7), 875. DOI: https://doi.org/10.3390/atmos16070875

[3] de Moura, F.R., da Silva Júnior, F.M.R., 2023. 2030 Agenda: discussion on Brazilian priorities facing air pollution and climate change challenges. Environ Sci Pollut Res Int, 30(3), 8376–8390. DOI: https://doi.org/10.1007/s11356-022-24601-5

[4] Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer-Verlag. DOI: https://doi.org/10.1007/978-1-4471-3675-0

[5] Hoinaski, L., Will, R., Ribeiro, C.B., 2024. Brazilian Atmospheric Inventories – BRAIN: a comprehensive database of air quality in Brazil. Earth System Science Data, 16, 2385–2405. DOI: https://doi.org/10.5194/essd-16-2385-2024

[6] Jimenez, J.R.Z., 2019. Prediction of concentrations of suspended particle levels of 2.5 micrometers ($PM_{2.5}$) in Mexico City with probability distribution functions and its trend. International Journal of Latest Research in Engineering and Technology, 5(4), 1–17.

[7] Willmott, C.J., Matsuura, K., Robeson, S.M., 2009. Ambiguities inherent in sums-of-squares-based error statistics. Atmospheric Environment, 43(3), 749–752. DOI: https://doi.org/10.1016/j.atmosenv.2008.10.005

[8] Mann, H.B., 1945. Nonparametric tests against trend. Econometrica, 13(3), 245–259. DOI: https://doi.org/10.2307/1907187

[9] Kendall, M.G., 1975. Rank Correlation Methods, 4th ed. Charles Griffin: London, UK.

[10] Hamed, K.H., Rao, A.R., 1998. A modified Mann-Kendall trend test for autocorrelated data. Journal of Hydrology. 204(1–4), 182–196. DOI: https://doi.org/10.1016/S0022-1694(97)00125-X

[11] Yue, S., Wang, C.Y., 2004. The Mann-Kendall test modified by effective sample size to detect trend in serially correlated hydrological series. Water Resources Management, 18(3), 201–218. DOI: https://doi.org/10.1023/B:WARM.0000043140.61082.60

[12] Pettitt, A.N., 1979. A non-parametric approach to the change-point problem. Applied Statistics, 28(2), 126–135. DOI: https://doi.org/10.2307/2346729

[13] Mota, P.H.S., Rocha, S.J.S.S., Castro, N.L.M., et al., 2019. Forest fire hazard zoning in Mato Grosso State, Brazil. Land Use Policy, 88, 104206. DOI: https://doi.org/10.1016/j.landusepol.2019.104206

[14] Tomei, J., Lyrio de Oliveira, L., de Oliveira Ribeiro, C., et al., 2020. Assessing the relationship between sugarcane expansion and human development at the municipal level: A case study of Mato Grosso do Sul, Brazil. Biomass and Bioenergy. 141, 105700. DOI: https://doi.org/10.1016/j.biombioe.2020.105700

[15] Abreu, M.C., Lyra, G.B., de Oliveira Júnior, J.F., et al., 2022. Temporal and spatial patterns of fire activity in

three biomes of Brazil. Science of the Total Environment. 844, 157138. DOI: https://doi.org/10.1016/j.scitotenv.2022.157138

[16] Nunes, R.S.C., Souza, A., Villar Hernández, B.J., et al., 2023. Fires in Brazilian biomes. Mercator (Fortaleza). 22, e22023. DOI: https://doi.org/10.4215/rm2023.e22023

[17] Correia Filho, W.L.F., Da Costa, R.R., Tavella, R.A., et al., 2024. Evaluation of the $PM_{2.5}$ concentrations in South America: Climatological patterns and trend analysis. Atmospheric Environment, 338, 120800. DOI: https://doi.org/10.1016/j.atmosenv.2024.120800

[18] Silveira, V.R., Oliveira Júnior, J.F., Silva, M.S., et al., 2021. Analysis of urban industrial expansion and increasing level of ozone concentration as subsiding an environmental management plan for the east of Rio de Janeiro metropolitan area – Brazil. Land Use Policy, 101, 105148. DOI: https://doi.org/10.1016/j.landusepol.2020.105148

[19] Clauset, A., Shalizi, C.R., Newman, M.E.J., 2009. Power law distributions in empirical data. SIAM Review, 51(4), 661–703. DOI: https://doi.org/10.1137/070710111

[20] Pope, C.A.,III, Burnett, R.T., Thun, M.J., et al., 2002. Lung cancer, cardiopulmonary mortality, and long term exposure to fine particulate air pollution. JAMA, 287(9), 1132–1141. DOI: 10.1001/jama.287.9.1132

[21] Brook, R.D., Franklin, B., Cascio, W., et al., 2004. Air pollution and cardiovascular disease: A statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association. Circulation, 109(21), 2655–2671. DOI: 10.1161/01.CIR.0000128587.30041.C8

[22] Bell, M. L., Dominici, F., Samet, J. M., 2005. A meta analysis of time series studies of ozone and mortality with comparison to the National Morbidity, Mortality, and Air Pollution Study. Epidemiology, 16(4), 436–445. DOI: 10.1097/01.ede.0000165817.40152.85

[23] Pope, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: lines that connect. Journal of the Air & Waste Management Association, 56(6), 709–742. DOI: 10.1080/10473289.2006.10464485

[24] Pereira, G.M., Kamigauti, L.Y., Pereira, R.F., et al., 2025. Source apportionment and ecotoxicity of $PM_{2.5}$ pollution events in a major Southern Hemisphere megacity: influence of a biofuel-impacted fleet and biomass burning. Atmospheric Chemistry and Physics, 25, 4587–4606. DOI: https://doi.org/10.5194/acp-25-4587-2025

[25] Jacob, D.J., Winner, D.A., 2009. Effect of climate change on air quality. Atmospheric Environment, 43(1), 51–63. DOI: 10.1016/j.atmosenv.2008.09.051

[26] Silveira, G.d.O., Azevedo, G.M.G.V.d., Tavella, R.A., et al., 2025. A Pilot Study with Low-Cost Sensors: Seasonal Variation of Particulate Matter Ratios and Their Relationship with Meteorological Conditions in Rio Grande, Brazil. Climate, 13(4), 71. DOI: https://doi.org/10.3390/cli13040071

[27] Lelieveld, J., Evans, J.S., Fnais, M., et al., 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature, 525(7569), 367–371. DOI: https://doi.org/10.1038/nature15371

[28] Moreira, G.A., Carbone, S., Guerrero-Rascado, J.L., et al., 2025. Evidence of the consequences of the prolonged fire season on air quality and public health from 2024 São Paulo (Brazil) data. Scientific Reports, 15, Article 28337. DOI: https://doi.org/10.1038/s41598-025-08542-w