Proceedings of CONF-MPCS 2026 Symposium: Theoretic Physics and Plasma Physics
DOI: 10.54254/2753-8818/2026.DL34440

The Probability of Extreme Pollution Weather Occurrence under Different Meteorological Conditions in Cities Based on Wuhan City

Ziyao Zeng

Shenzhen Senior High School Youwei High School, Shenzhen, China

zhiyuezhi01@gmail.com

Abstract. Understanding the meteorological drivers of extreme pollution events is crucial for formulating effective early warning systems and public health interventions. Based on the air quality and meteorological observation data of Wuhan from 2018 to 2026, the two-stage framework of variance analysis and Bayesian inference is used to explore the impact of meteorological conditions on the probability of extreme pollution events (AQI>200). The analysis of variance is used to screen key meteorological factors, and Bayesian logistic regression is used to estimate the probability of conditions and give the posterior distribution. The results show that extreme pollution accounts for 4.65%, which is characterized by rare events; the probability of extreme pollution is the highest in cold and high-pressure bottom (CHB) weather, which is 14.2% (95% CI: 9.8%~19.6%), which is about 7 times that of the low-pressure groove (LPIT) type (2.1%). This method effectively quantifies the impact of meteorological conditions on the probability of extreme pollution, which can provide a scientific basis for early warning systems

Keywords: Analysis of variance, Bayesian logistic regression, meteorological conditions, posterior probability

1. Introduction

Air pollution has become an important environmental problem affecting the lives of most urban residents [1]. However, the existing literature focuses on the average concentration of pollutants and their health effects [2]. From a Mathematical point of view, extreme pollution events can be regarded as a very few-tailed event in the air quality distribution [3]. This rare event characteristic requires the statistical framework to shift from descriptive analysis to inferential modeling. Its core goal is to estimate the conditional probability of extreme pollution events under specific meteorological conditions [4]. Understanding the probability behavior of extreme pollution is not only an environmental problem, but also a statistical inference problem.

Although the frequency of extreme pollution events is low, their occurrence is not completely random. A large amount of empirical evidence shows that such events show obvious seasonal and meteorological dependence, which means that their probability distribution is influenced by a set of

DOI: 10.54254/2753-8818/2026.DL34440

observable atmospheric variables. Therefore, this problem is very suitable for statistical modeling methods. By treating the concentration of pollutants as random variables whose distribution parameters change with meteorological conditions, the conditional probability of exceeding the risk threshold can be estimated. In addition, extreme value theory provides a strict mathematical framework for estimating the extreme range of extreme pollution levels under given conditions [5].

Based on the above analysis, this study combines Bayesian inference and variance analysis to quantify the impact of different meteorological factors on the probability of extreme pollution. Among them, variance analysis is used to identify key meteorological factors, while the Bayesian method provides a posterior distribution of event probability (not single-point estimation). Specifically, this paper is first based on the air quality and meteorological observation data of Wuhan from 2018 to 2026, and uses variance analysis to screen key meteorological factors; secondly, it uses Bayesian logical regression to estimate the conditional probability of extreme pollution (AQI > 200) under different weather types, and gives its posterior Distribution and confidence interval; finally, by comparing the probability differences of different weather types, the mechanism of the influence of meteorological conditions on the risk of extreme pollution in urban areas is revealed. The research results can provide a statistical basis for early warning and public health decision-making of extreme pollution weather.

2. Methods

2.1. Data source

The air quality data used in this study comes from the national urban air quality real-time release platform of the China Environmental Monitoring Terminal. Monitoring indicators include PM2.5、PM10, SO2, NO2, CO, O3-8h and Air quality index(AQI). Daily monitoring data from 2018 to 2026 were collected from nine national control air quality monitoring stations in Wuhan, covering different functional areas of the city.

Meteorological data comes from the China Meteorological Data Network [6]. The selected meteorological elements include average daily temperature, daily average relative humidity, daily precipitation, average daily wind speed, dominant wind direction, average daily air pressure and sunshine duration.

All raw data underwent quality control procedures: outliers were identified using the boxplot method and excluded. Continuous meteorological variables were standardized using Z-score normalization to eliminate dimensional influences

2.2. Indicators and analysis

Based on meteorological principles and climatic characteristics of Wuhan, three types of indicators:

The thermal factor (average daily temperature) reflects the thermal stability and vertical diffusion conditions of the atmosphere; the dynamic factor (average daily wind speed) characterizes the horizontal conveying capacity and the intensity of turbulence exchange; and the humidity and precipitation factors affect the moisture absorption growth of aerosols and their wet removal process.

Consult and follow the “Technical Regulations on Environmental Air Quality Index (AQI) (Trial Implementation)”. The weather process that defines extreme pollution weather as AQI daily average value is>200 (i.e., severe pollution and above). The two-classified variable Y to indicate whether the day is an extreme pollution day: Yt=1 indicates AQI>200, and Yt=0 indicates AQI≤200.

DOI: 10.54254/2753-8818/2026.DL34440

2.3. Method introduction

To calculate the probability of extreme pollution under different meteorological conditions, continuous meteorological variables need to be converted into discrete classification variables. This study uses two stages to satisfy:

Based on variance analysis:

Discretize the continuous meteorological variables and makes it become categorical variables (Number of intervals K=5~10, ensure that the sample size of each interval is sufficient). Then, single-factor variance analysis is adopted, taking different meteorological factor categories as independent variables, taking the concentration of pollutants as the dependent variable, test the salience of the impact of various meteorological factors on extreme pollution. Through the analysis of variance, identify the key meteorological factors that have a significant impact on the probability of extreme pollution in different seasons.

Conditional probability estimation based on Bayesian inference [7, 8]:

On the basis of determining the key meteorological factors, it adops Bayesian logistic regression and Bayesian generalized linear model to estimate the conditional probability of extreme pollution events.

Suppose Y is the extreme pollution indicator variable, X is the key meteorological factor vector, thus the conditional probability can be expressed as

$$P(Y=1 \mid X)=\pi(X) \tag{1}$$

Under the Bayesian framework, the model parameters Θ give a prior distribution Parking lot(Θ)P(θ), combined with observation data dimension get the posterior distribution

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta) \tag{2}$$

This study adopts a two-stage statistical framework. In the first stage, the continuous meteorological variables are discretized into classification variables (interval number K=5~10), the significance of the impact of each meteorological factor on extreme pollution is tested through single-factor analysis of variance, and the key meteorological factors in different seasons are identified. The advantage of this method is that discrete processing can weaken the interference of measurement error, enhance interpretability, and can capture the threshold effect of factors on pollution. The calculation is efficient, and it is suitable as a preliminary screening tool for high-dimensional factors. In the second stage, based on the key factors selected, Bayesian logistic regression or Bayesian generalized linear model is used to estimate the conditional probability of extreme pollution events, through the Bayesian formula, dimension obtains the posterior distribution of the parameters. The core advantage of the Bayesian method is that the complete posterior distribution and trusted interval of the output probability is not a single-point estimate, which can integrate a priori information to deal with small sample problems (the proportion of extreme pollution events is usually less than 5%), and the posterior distribution can directly serve early warning decision-making. The two-stage method forms a collaborative chain of "screening-assessment", which not only avoids the risk of overfitting of direct modeling of high-dimensional variables, but also overcomes the limitations of the traditional frequency method in the estimation of rare event probability.

DOI: 10.54254/2753-8818/2026.DL34440

3. Results and discussion

3.1. Basic information

Based on the data and statistical methods described in Section 2, this section presents the empirical results of this study. First, report the descriptive statistics of extreme pollution events and meteorological variables. The dataset comprises 2922 daily observations from January 1, 2018, to January 1, 2026, after quality control and outlier removal. Among these, 136 days (4.65%) were identified as extreme pollution days (AQI > 200). Statistical significance assessed at the $\alpha = 0.05$ level.

Table 1. Descriptive statistics and distribution characteristics of major variables

Name	Mean	Standard deviation	Median	Max	Min
PM2.5	65.7903	41.7915	58.5	206	7
PM10	79.4032	45.9291	73.5	221	8
O3	58.6774	19.4706	62.5	94	9
SO2	8.8870	2.0333	9	14	6
NO2	47.9677	20.1363	49.5	86	16
CO	0.9713	0.2745	0.9725	1.611	0.5
AQI	92.5806	50.5937	88.5	256	24

It can be seen from Table 1 that the concentration of PM2.5 shows a significant right-biased distribution. It shows that the pollution concentration is low on most days, and a few extremely high values constitute the tail of the distribution, which confirms the statistical perspective of treating extreme pollution as a "rare event". The precipitation in the first 48 hours showed a highly right-biased distribution, indicating that most precipitation events are low-probability events, which is consistent with meteorological laws.

3.2. Model results

To examine whether extreme pollution probabilities differ significantly across weather types, it conducted a one-way ANOVA. The analysis compares PM2.5 concentrations across four weather types: cold high-pressure bottom (CHB), high-pressure rear (HPR), uniform pressure field (UPF), and low-pressure inverted trough (LPIT).

One-way ANOVA revealed statistically significant differences in PM2.5 concentrations for the four weather types ($F = 98.4$, $p < 0.001$). The effect size ($\eta^2 = 0.18$) indicates that weather type accounts for 18% of the variance in PM2.5 concentrations.

To further identify which specific weather types have significant differences, this study conducted post hoc tests on PM2.5 concentrations under different weather types, the result as Table 2.

DOI: 10.54254/2753-8818/2026.DL34440

Table 2. After-the-fact test results of PM2.5 concentration under different weather types

Weather type	Sample size	PM2.5 mean (μg/m³)	Standard deviation
CHB	142	98.6	42.3
HPR	98	67.2	35.1
UPF	86	58.4	31.6
LPIT	112	42.3	28.4

Based on the key weather types selected by analysis of variance, this study further adopts the Bayesian logistic regression model. Estimate the conditional probability of extreme pollution events (AQI > 200) in different weather types. The model treats the weather type as the independent variable, the extreme pollution indicator variable as the dependent variable, and provides the parameters with no a priori information (Table 3).

Table 3. After-the-fact test results of PM2.5 concentration under different weather types

Weather type	CHB	LPLT	HPR	UPF
Average probability	0.142	0.021	0.058	0.043

4. Discussion

This study reveals the law of extreme pollution weather in Wuhan and its statistical relationship with meteorological conditions. From the core findings, the frequency of extreme pollution shows a significant downward trend, which is consistent with the results of national air pollution prevention and control. At the same time, studies have shown that the combined pollution of PM2.5 and ozone is becoming more and more prominent in Chinese cities [9]. From a statistical perspective, the two-variable extreme characteristics of compound pollution events (PM2.5 and O3 exceeding the standard at the same time) are more complex than single-pollutant events. Due to the nonlinear interactions among meteorological factors, the traditional single-variable extreme value method struggles to fully capture its risk characteristics. Recent studies have successfully applied extreme value theory to predict future ozone extreme events under different emissions and climate paths [10]. At the practical level, the "screening-assessment" two-stage framework of this study can provide quantitative support for extreme pollution early warning. Future work needs to be extended to a multivariate extreme value framework to explore the conditional probability modeling of PM2.5 extreme events, thereby supporting environmental risk management in the background of air pollution.

5. Conclusion

Based on the air quality and meteorological observation data of Wuhan from 2018 to 2026, this study uses variance analysis and Bayesian inference methods to explore the probability of extreme pollution events (AQI > 200) under different meteorological conditions. Extreme pollution shows a significant downward trend, but the form of pollution has changed. During the study period, the average annual concentration of PM2.5 in Wuhan decreased by 56.3%, while the concentration of ozone increased significantly, indicating that Wuhan is transforming from single particulate pollution to PM2.5-O3 composite pollution. Therefore, the probability structure of extreme pollution events is becoming more and more complex.

The conditional probability estimation based on Bayesian logical regression further quantifies the risk of extreme pollution under different weather types. The posterior probability of extreme pollution in CHB weather is 14.2% (95% CI: 9.8%~19.6%), which is about 7 times that of the low-pressure tank (LPIT) type (2.1%, 95% CI: 0.8%~3.8%). The probability of high-pressure posterior (HPR) and equal pressure field (UPF) is 5.8% and 4.3%, respectively, and there is no significant difference between the two.

This study reveals the statistical relationship between meteorological conditions and urban air pollution from the perspective of the probability of extreme events, and verifies the feasibility of inferring and modeling extreme pollution as a "rare event". The research results provide a quantitative basis for the statistical identification and probability warning of extreme pollution weather, and lay the foundation for the improvement of environmental risk management methods in the context of compound pollution. Future research should incorporate the multivariate extreme value framework to portray the joint probability behavior of PM2.5 and O3 dual-variable extreme events.

References

[1] Wang, Y., Yao, J. L., Zhou, X. L., et al. (2022). Characteristics of air pollution in Suzhou in 2020 and its short-term impact on residents' health. Occupational and Health, 38(15), 2082–2087.
[2] Chen, Y. Y., Zhao, Q. Q., & Li, L. (2025). Related factors of outdoor air pollution affecting respiratory health of the elderly. Journal of Environment and Health, 42(8), 694–698.
[3] Majumder, R., & Richards, J. (2026). Semi-parametric bulk and tail regression using spline-based neural networks. Extremes, Advance online publication.
[4] Yuan, M. X., & Yang, F. (2025). Research on the variation characteristics of atmospheric visibility and its meteorological influencing factors in Wuhan. Journal of Agricultural Catastrophology, 15(10), 194–197.
[5] Ma, Y. L., Ext, C. F., & Liu, C. Y. (2026). Statistical inference of additive tail index model and its application. Journal of Shanxi University (Natural Science Edition), 1–15. Advance online publication.
[6] China Meteorological Data Service Centre. (2026). China meteorological data network. http: //data.cma.cn/
[7] Wan, F., Shen, L., & Li, M. (2025). Projecting future ozone episodes in China by 2060 under diverse emission and climate pathways: Insights from extreme value theory. Geophysical Research Letters, 52(19), Article e2025GL116891.
[8] Ji, X. S., & Fu, K. X. (2026). Covariate conditional diffusion model for multi-resolution time series forecasting. Computer Engineering, 1–12. Advance online publication. (In Chinese)
[9] Song, X. H., Yan, L., Liu, W., et al. (2023). Spatiotemporal characteristics of PM2.5 and ozone composite pollution in Beijing-Tianjin-Hebei and its surrounding areas from 2015 to 2021. Environmental Science, 44(4), 1841–1851.
[10] Luo, Y., Dang, C., Broggi, M., et al. (2025). Stochastic dynamic response analysis via dimension-reduced probability density evolution equation (DR-PDEE) with enhanced tail-accuracy. Probabilistic Engineering Mechanics, 79, Article 103735.