International Journal of Research Studies on Environment, Earth, and Allied Sciences (IJRSEAS)
| *Volume-3 Issue-4* | *August-2026* | *ISSN: 3066-0637* | *Website:* https://eesrd.us

Manuscript ID:
IJRSEAS-2026-030404

OPEN ACCESS

Quick Response Code:

Website: https://eesrd.us

Creative Commons
(CC BY-NC-SA 4.0)

DOI: 10.5281/zenodo.23391988

DOI Link:
https://doi.org/10.5281/zenodo.23391988

Volume: 3

Issue: 4

Pp. 26-34

Month: August

Year: 2026

E-ISSN: 3066-0637

Submitted: 07 July 2026

Revised: 18 July 2026

Accepted: 10 Aug.2026

Published: 31 Aug. 2026

Address for correspondence:
Pragya Goyal
Centre for Advanced Studies in Social Science and Management, Institution for Development and Communication, Chandigarh, India
Email: pgoyalstats@gmail.com

How to cite this article:
Goyal, P. (2026). Seasonal Dynamics, Extreme Pollution Events, and Hybrid Forecasting of Urban Air Quality in India: A Multi-City Panel Analysis. International Journal of Research Studies on Environment, Earth, and Allied Sciences (IJRSEAS), 3(4), 26–34.
https://doi.org/10.5281/zenodo.23391988

Seasonal Dynamics, Extreme Pollution Events, and Hybrid Forecasting of Urban Air Quality in India: A Multi-City Panel Analysis

Pragya Goyal
Centre for Advanced Studies in Social Science and Management, Institution for Development and Communication, Chandigarh, India

Abstract

Air pollution is a major environmental and public-health concern, particularly in rapidly urbanizing countries such as India. The Air Quality Index (AQI) condenses information on multiple criteria pollutants into a single measure that can be communicated to the public and used to support regulatory action. However, urban AQI dynamics are shaped simultaneously by seasonality, persistent city-specific characteristics, extreme pollution episodes, nonlinear temporal dependence, and changes in the identity of prominent pollutants. This study develops an integrated statistical framework for examining these interconnected dimensions of air quality. Using daily AQI data covering 292 Indian cities from 1 April 2022 to 30 April 2025, the analysis combines exploratory assessment, panel econometric modelling, extreme value theory, hybrid time-series and machine-learning forecasting, and Markov chain modelling of prominent pollutants. Fixed-effects estimates reveal pronounced seasonal variation: relative to the monsoon period, average AQI is substantially higher in summer, post-monsoon, and especially winter. The estimated linear time trend is negative, indicating a decline in average within-city AQI over the study window, although this pattern should not be interpreted as a causal policy effect without further identification. Peaks-over-Threshold modelling of AQI observations above 300 yields an estimated 10-year return level of approximately 453, demonstrating the continuing risk of very severe pollution episodes. The forecasting framework combines an ARIMA model for the linear temporal component with XGBoost learning applied to residual variation using lagged AQI features. Pollutant -transition analysis shows that PM10 and PM2.5 are the most frequent next prominent -pollutant states across a wide range of current states. By integrating econometrics, extreme-value methods, machine learning, and stochastic transition modelling, the study provides a multidimensional assessment of urban air-quality dynamics and a framework for seasonal intervention, extreme-event preparedness, and short-term forecasting.

Keywords: *Air Quality Index; panel data; fixed effects; extreme value theory; Generalized Pareto distribution; ARIMA; XGBoost; Markov chain.*

Introduction

1. Background

Air pollution is among the most persistent environmental and public-health challenges confronting contemporary cities. Exposure to major air pollutants is associated with adverse respiratory and cardiovascular outcomes and constitutes a major global health concern (World Health Organization, 2021). In India, the problem is especially complex because pollution arises from a heterogeneous combination of transport emissions, industrial activity, construction and road dust, solid-fuel use, power generation, biomass burning, and seasonal agricultural practices. Meteorological conditions further modify the accumulation and dispersion of pollutants, producing large differences across places and seasons. The Air Quality Index (AQI) was developed as a communication tool that translates concentrations of major pollutants into a common numerical scale and associated health categories. In the Indian framework, AQI is commonly reported through six categories-Good, Satisfactory, Moderate, Poor, Very Poor, and Severe-providing a standardized basis for communicating air-quality conditions and associated health concerns (CPCB, 2025). A higher AQI indicates poorer air quality and greater potential health concern. Because the index summarizes the prevailing pollution burden in a form that is readily interpretable, it is useful for public advisories, regulatory communication, comparison across locations, and the activation of short -term response measures. An analysis of AQI should therefore go beyond reporting a national or city-level average. Average values may conceal strong seasonal cycles, persistent differences among cities, and short -lived but hazardous episodes in the upper tail of the distribution.

Likewise, forecasting AQI requires methods that can accommodate both regular temporal dependence and nonlinear changes, while pollutant-specific analysis is needed to understand whether the identity of the prominent pollutant persists or shifts over time.

Literature Review

Existing AQI research may be organized into several broad streams. The first stream examines descriptive and spatio-temporal patterns in pollutant concentrations and AQI. Spatio-temporal approaches are particularly relevant to environmental data because pollution levels vary simultaneously across geographical locations and over time (Cressie & Wikle, 2011). A second stream uses statistical and econometric models to explain temporal or spatial variation. Panel-data methods are particularly useful when repeated observations are available for many cities because they allow the analyst to separate within-city changes from persistent, unobserved city characteristics. A large and rapidly expanding third stream focuses on prediction. Conventional time-series models such as ARIMA are widely used because they provide a transparent representation of autocorrelation, differencing, and moving-average behaviour. Machine-learning methods have also become prominent because AQI may depend on nonlinear interactions and complex lag structures. Gupta et al. (2023), for example, compared support vector regression, random forest regression, and CatBoost regression for multiple Indian cities and emphasized the value of systematic model comparison and clearly reported evaluation metrics. Other studies have combined statistical and machine-learning components in hybrid frameworks so that the first model captures regular linear dependence and the second learns structure remaining in the residuals. M.T. Udristioiu et al. (2023) presents a hybrid machine learning framework for predicting and forecasting particulate matter (PM_1, $PM_{2.5}$, PM_{10}) and air quality index using environmental sensor data. Liu et al. (2019) uses machine learning techniques such as support vector regression and random forest regression to predict air quality indices (AQI) in Beijing and nitrogen oxides (NO) concentrations in an Italian city using real-world environmental datasets. A fourth, less integrated, stream concerns rare and extreme pollution events. Conventional regression and forecasting models are primarily designed to describe the center of a distribution and may not adequately characterize unusually high observations. S.K. Natarajan et al. (2024) proposes an optimized machine learning framework combining Grey Wolf Optimization with Decision Tree Regression to predict AQI in major Indian cities. Extreme Value Theory (EVT), particularly the Peaks-over-Threshold approach and the Generalized Pareto Distribution (GPD), provides a principled framework for modelling exceedances above a high threshold and estimating return levels associated with rare events. Finally, pollutant behaviour can be studied as a sequence of states. A Markov chain provides a simple stochastic framework for estimating the probability that one prominent pollutant is followed by another. Although such a model does not identify causal atmospheric mechanisms, it can reveal persistence, dominant next states, and broad transition patterns in the observed pollutant sequence.

Research gap and contribution

Despite substantial progress, the literature remains methodologically fragmented. Seasonal variation is often examined descriptively; forecasting studies tend to focus mainly on predictive performance; extreme pollution episodes are frequently analyzed separately from longitudinal city-level behaviour; and transitions in the identity of prominent pollutants receive comparatively limited attention. Consequently, there is scope for an integrated framework that examines the center, tail, temporal dependence, and pollutant-state dynamics of AQI within the same multi-city study.

The present study addresses this gap using a large panel of Indian cities. Its contribution is fourfold. First, fixed-effects panel modelling is used to quantify seasonal and longitudinal within-city AQI variation while controlling for time-invariant city heterogeneity. Second, a Peaks-over-Threshold GPD model is used to quantify the upper tail of severe pollution episodes. Third, a hybrid ARIMA–XGBoost framework is used to combine linear temporal structure with nonlinear residual learning. Fourth, a discrete-time Markov chain is used to describe transitions among prominent pollutants. The objective is not merely to compare techniques, but to use complementary statistical tools to answer different questions about the same environmental process.

Objectives of the study

1. To describe the overall and seasonal distribution of AQI across Indian cities.
2. To estimate seasonal and longitudinal within-city AQI effects using panel econometric models.
3. To quantify the tail behaviour and return levels of extreme AQI observations using EVT.
4. To develop a hybrid statistical–machine-learning framework for short-term AQI forecasting.
5. To examine transitions among prominent pollutants using a first-order Markov chain.

Data and Study Design

1. **Data source and coverage**

The study uses daily AQI records compiled from the Central Pollution Control Board (CPCB) monitoring system for the period from 1 April 2022 to 30 April 2025 (CPCB, 2025). After cleaning city identifiers and removing observations with missing AQI values, the data cover 292 Indian cities. Because records containing multiple prominent pollutants were expanded for pollutant-level analysis, the total row count after expansion should be distinguished from the number of unique city-date AQI observations. Each record includes date, state,

area or city, AQI value, air-quality status, the number of monitoring stations, and one or more prominent pollutants. The panel is unbalanced because cities differ in the number and continuity of available observations.

2. **Variables and AQI categories**

The principal outcome variable is the numerical AQI value. Air-quality status is treated as an ordered categorical variable with six levels: Good, Satisfactory, Moderate, Poor, Very Poor, and Severe. The number of monitoring stations is retained as a measure of monitoring intensity. Prominent pollutants include CO, NH_3, NO_2, O_3, PM_{10}, $PM_{2.5}$, SO_2, and SO_3 as recorded in the source data.

3. **Data cleaning and feature engineering**

Dates were converted to a standard date format and used to derive calendar year, month, and weekday. Seasons were classified as winter (December–February), Summer (March–May), Monsoon (June–September), and Post-Monsoon (October–November). Records listing multiple prominent pollutants were separated into long format for pollutant-level analysis. For forecasting, observations were ordered chronologically within city and lagged AQI features were generated at 1-, 7-, and 30-day horizons.

4. **Analytical framework**

The empirical analysis proceeds in four stages. Exploratory analysis first describes the distribution and seasonal pattern of AQI. Panel econometric models then quantify systematic within-city seasonal and time effects. EVT is subsequently used to focus specifically on the upper tail of the AQI distribution. The forecasting component combines a statistical time-series model with gradient boosting, and the final component uses transition probabilities to summarize changes in prominent pollutants.

Statistical Methodology

1. **Exploratory and seasonal analysis**

Boxplots and category frequencies are used to examine central tendency, dispersion, skewness, outliers, and seasonal differences. The boxplot is especially useful because AQI is right-skewed and contains many high observations; unlike the mean alone, it displays the median, interquartile range, whiskers, and extreme points.

2. **Panel econometric models**

Repeated observations for the same city form a panel structure. Let AQI_{it} denote the average AQI for city i at time t. A pooled regression would treat all city-date observations as independent and would ignore persistent differences in geography, industrial structure, transport intensity, urban form, and other time-invariant characteristics. Panel-data models provide a framework for accounting for such unobserved individual heterogeneity while exploiting repeated observations over time (Croissant & Millo, 2008).

The baseline specification is $AQI_{it} = \alpha_i + \beta_1 PostMonsoon_{it} + \beta_2 Summer_{it} + \beta_3 Winter_{it} + \gamma Year_t + \varepsilon_{it}$,

where α_i represents an unobserved city-specific effect and Monsoon is the omitted reference category. Each seasonal coefficient therefore measures the average within-city difference from Monsoon, conditional on the linear time trend.

3. **Fixed effects, random effects, and the Hausman test**

The fixed-effects estimator removes time-invariant city characteristics through the within transformation and identifies coefficients from changes occurring within the same city over time. The random-effects model instead assumes that the city-specific effect is uncorrelated with the regressors. Because this assumption is restrictive, the Hausman specification test is used to assess the consistency of the random-effects estimator relative to the fixed-effects estimator (Hausman, 1978). Rejection of the null hypothesis supports the fixed-effects specification.

4. **Extended model and robust inference**

An extended fixed-effects model includes the number of monitoring stations. Because repeated observations within a city may exhibit heteroskedasticity and serial dependence, cluster-robust standard errors at the city level are used for this specification. The monitoring-station coefficient is interpreted as an association with observed AQI, not as a causal effect, because monitoring infrastructure may itself be expanded in more polluted locations.

5. **Extreme Value Theory and the POT–GPD approach**

Extreme Value Theory focuses on the probabilistic behaviour of unusually large observations. The Peaks-over-Threshold (POT) method retains observations exceeding a sufficiently high threshold u and models the excess $Y = AQI - u$. Under suitable regularity conditions and for a sufficiently high threshold, the distribution of threshold exceedances can be approximated by the Generalized Pareto Distribution with scale parameter σ and shape parameter ξ (Davison & Smith, 1990).

In this study, u = 300 is used because it marks the upper end of the Very Poor range and isolates severe pollution observations. The scale parameter controls the spread of exceedances, whereas the shape parameter determines tail behaviour. A negative shape parameter implies a finite upper endpoint, which is plausible for an AQI scale capped near 500. Model adequacy is assessed through diagnostic plots comparing empirical and fitted behaviour.

6. **Return-level estimation**

A return level translates the fitted tail model into an interpretable risk measure. A T-year return level is the level expected to be exceeded, on average, once in T years under the fitted stationary model and assumed observation frequency. Return levels should be interpreted probabilistically rather than as deterministic forecasts of the exact year in which an event will occur. Return levels provide an interpretable summary of the magnitude

associated with a specified recurrence horizon in extreme-value analysis (Embrechts et al., 1997).

7. **ARIMA modelling**

ARIMA models describe temporal dependence through autoregressive, differencing, and moving-average components and remain widely used for univariate time-series forecasting (Hyndman & Athanasopoulos, 2021). The selected ARIMA (0,1,1) model uses first differencing to remove nonstationary level behaviour and a moving-average term to represent short-run dependence in forecast errors. The model provides the linear statistical component of the hybrid framework.

8. **XGBoost residual learning**

XGBoost is a scalable gradient-boosting algorithm that sequentially combines decision trees to improve predictive performance while incorporating regularization and computational optimization (Chen & Guestrin, 2016). In the present framework, XGBoost is not fitted directly as a replacement for ARIMA. Instead, it is trained on ARIMA residuals using lag-1, lag-7, and lag-30 AQI features. This design asks whether nonlinear information remains after the linear time-series structure has been modelled.

9. **Hybrid ARIMA-XGBoost framework**

The conceptual decomposition is: observed AQI = linear temporal component + nonlinear residual component. The ARIMA forecast estimates the first component, while XGBoost predicts the residual component. The final hybrid prediction is obtained by adding the two. This architecture is intended to capture complementary structure; however, claims of superior forecasting performance require formal out-of-sample comparison using the same test period and metrics such as RMSE and MAE.

10. **Markov chain modelling**

A discrete-time Markov chain represents transitions among a finite set of states through a transition probability matrix, where the probability of the next state depends on the current state under the first-order Markov assumption (Spedicato, 2017). A first-order discrete-time Markov chain is used to summarise transitions among prominent pollutants. The state space is S = $\{O, NH_3, NO_2, O_3, PM_{10}, PM_{2.5}, SO_2, SO_3\}$. The transition probability p_{ij} represents the estimated probability that state j follows state i in the observed sequence. Rows of the transition matrix sum to one. The first-order Markov assumption implies that the next state is modelled as depending only on the current state.

The resulting matrix is descriptive. A high probability of transition into PM_{10}, for example, indicates that PM_{10} is frequently observed as the next prominent pollutant. It should not be called an absorbing state unless its self-transition probability equals one, and it should not be interpreted as evidence that one pollutant physically causes another.

Results and Discussion

1. Overall AQI distribution and category profile

The overall AQI distribution is strongly right-skewed. The median lies in the lower-to-middle part of the AQI scale, while the upper tail extends to values close to 500. Numerous high observations appear beyond the upper whisker of the boxplot, showing that severe pollution episodes are not adequately represented by the centre of the distribution alone. This distributional pattern motivates the later use of EVT, which is designed specifically for upper-tail behaviour.

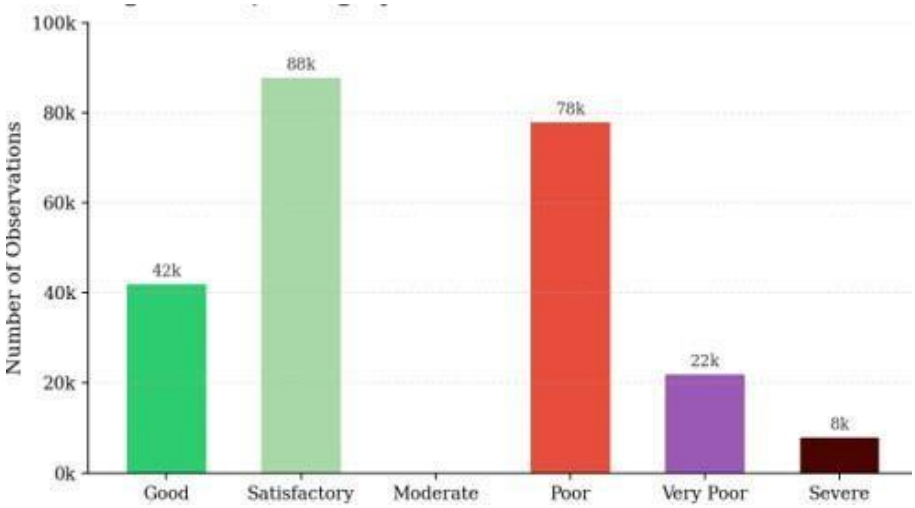

Figure 1. AQI category distribution across all cities and observation years.

Figure 1 presents the frequency distribution across AQI categories. The figure should be read as a description of how often different air-quality conditions occur in the full sample. The concentration of observations in the lower and middle categories indicates that very high pollution is not the modal daily condition across all cities and dates; nevertheless, the visible frequencies in Poor, Very Poor, and Severe categories demonstrate a substantial upper-tail burden. Because the data pool many cities and years, the figure should not be

interpreted as the profile of a typical individual city.

2. Seasonal variation in AQI

Figure 2 shows a clear seasonal ordering. Monsoon has the lowest median and the narrowest central spread, whereas Winter and Post-Monsoon have higher medians and longer upper tails. Summer occupies an intermediate position. The pattern is consistent with the combined role of emissions and meteorology: rainfall can remove airborne particles during the monsoon, while weaker atmospheric dispersion and inversion conditions can allow pollutants to accumulate in winter. Post-monsoon pollution may also coincide with region-specific biomass-burning episodes.

Table 1 summarises the same pattern numerically. The seasonal comparison is important because it shows that deterioration occurs not only through a shift in the median but also through a wider distribution and more extreme upper observations. Thus, the seasonal burden is reflected in both typical conditions and tail risk. Approximate values read from a graph should, however, be replaced in the final manuscript by statistics calculated directly from the data.

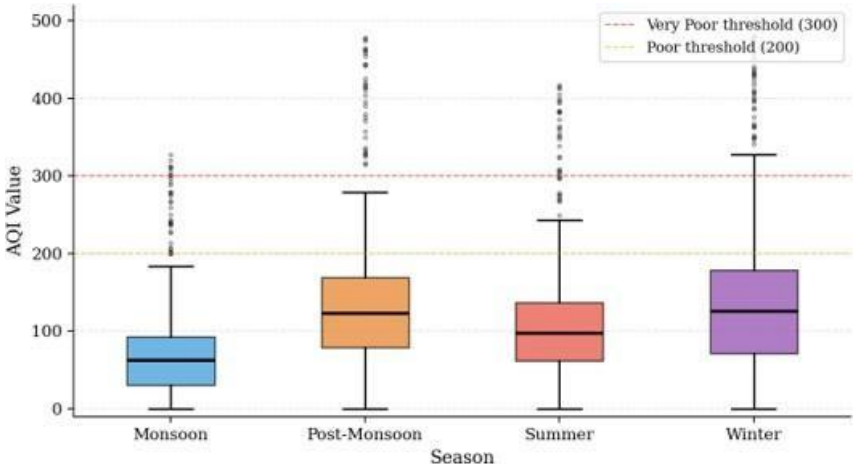

Figure 2. Seasonal AQI variation across Indian cities.

Table 1. Summary Statistics of AQI by Season

Season	Median AQI	IQR	Upper Whisker (approx.)	Dominant Category
Monsoon	~60	~50–100	~150	Good / Satisfactory
Post-Monsoon	~115	~75–165	~320	Moderate / Poor
Summer	~95	~65–145	~265	Satisfactory / Moderate
Winter	~120	~75–185	~325	Moderate / Poor

3. Fixed-effects panel results

Table 2 reports the fixed-effects estimates, with Monsoon as the reference category. The coefficients therefore represent average within-city AQI differences relative to Monsoon after controlling for the common linear year trend. Post-Monsoon AQI is estimated to be 62.91 units higher, Summer AQI 44.30 units higher, and Winter AQI 77.43 units higher than Monsoon. All three effects are statistically significant at conventional levels. The ordering of the coefficients quantitatively confirms the descriptive boxplot. Winter has the largest estimated difference, followed by Post-Monsoon and Summer. The Winter coefficient is about 23% larger than the Post-Monsoon coefficient and roughly 75% larger than the Summer coefficient. This indicates that the seasonal deterioration is not merely a generic non-monsoon effect; its magnitude differs substantially across seasons. The year coefficient is -8.71, indicating an estimated average within-city downward trend of about 8.7 AQI units across successive calendar years during the study period, conditional on season. This is a descriptive longitudinal association. It may reflect changes in emissions, regulation, economic activity, monitoring coverage, sample composition, or other time-varying factors; the coefficient alone does not identify the causal effect of a particular policy. The within R^2 of approximately 0.246 indicates that the included seasonal indicators and linear year trend explain about one quarter of the within-city variation in AQI. For a daily environmental outcome affected by weather, local activity, episodic emissions, and measurement conditions, substantial unexplained variation is expected.

Table 2. Fixed Effects Panel Regression Results (Dependent Variable: Mean AQI)

Variable	Estimate	Std. Error	t-value	p-value
Season: Post-Monsoon	62.909	0.327	192.62	$< 2.2e^{-16}$***
Season: Summer	44.299	0.292	151.82	$< 2.2e^{-16}$***
Season: Winter	77.426	0.300	258.07	$< 2.2e^{-16}$***

Year (linear trend)	−8.714	0.130	−67.12	$< 2.2e^{-16}$***
Within R^2	0.2464	—	—	—
F-statistic	19,252	df = 4, 235,489	—	$< 2.2e^{-16}$

Reference category: Monsoon. *** $p < 0.001$.

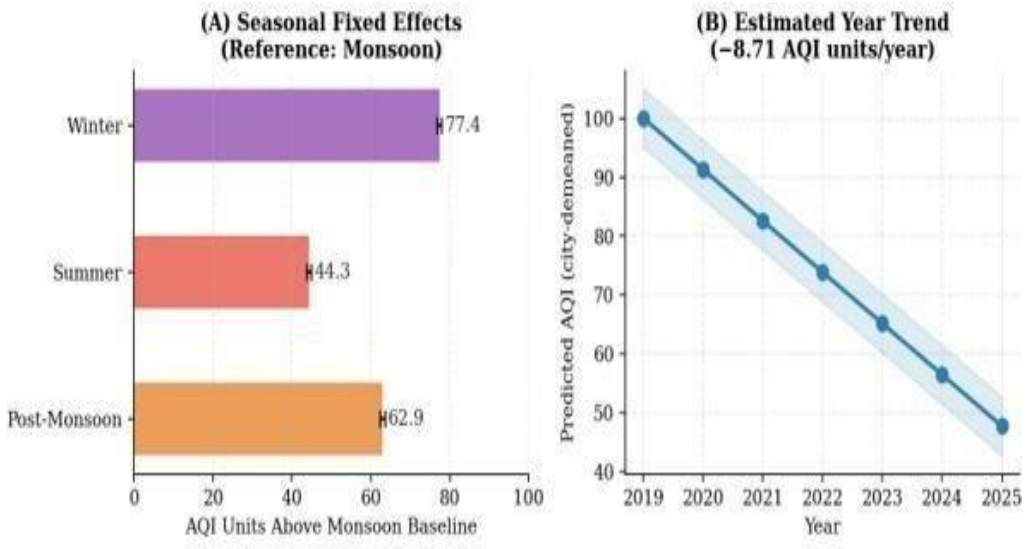

Figure 3. Fixed-effects coefficient estimates and linear time trend.

4. Model selection and monitoring intensity

The Hausman statistic is 76.79 with 4 degrees of freedom and a p-value of 8.3×10^{-16}. The null hypothesis that the random-effects estimator is consistent is therefore rejected. This result indicates that the unobserved city-specific component is correlated with at least some regressors, making the fixed-effects estimator the preferred specification for the reported within-city associations.

In the extended model, the number of monitoring stations has a coefficient of 1.877 with a cluster-robust standard error of 0.788 and p = 0.017. The positive association should not be read as evidence that monitoring stations increase pollution. A more plausible interpretation is that monitoring intensity is endogenous: cities or periods with greater pollution concern may receive more stations, and broader monitoring may detect spatially heterogeneous pollution more completely.

5. Extreme pollution episodes

The POT analysis retains 6,226 observations above AQI 300. The fitted GPD has an estimated scale parameter of 54.98 and shape parameter of -0.239. The negative shape estimate implies a bounded upper tail, which is consistent with the finite upper range of the AQI scale. The diagnostic plots should be discussed jointly: agreement between empirical and model quantiles supports the fitted tail over much of the range, while deviations at the most extreme observations indicate where uncertainty is greatest.

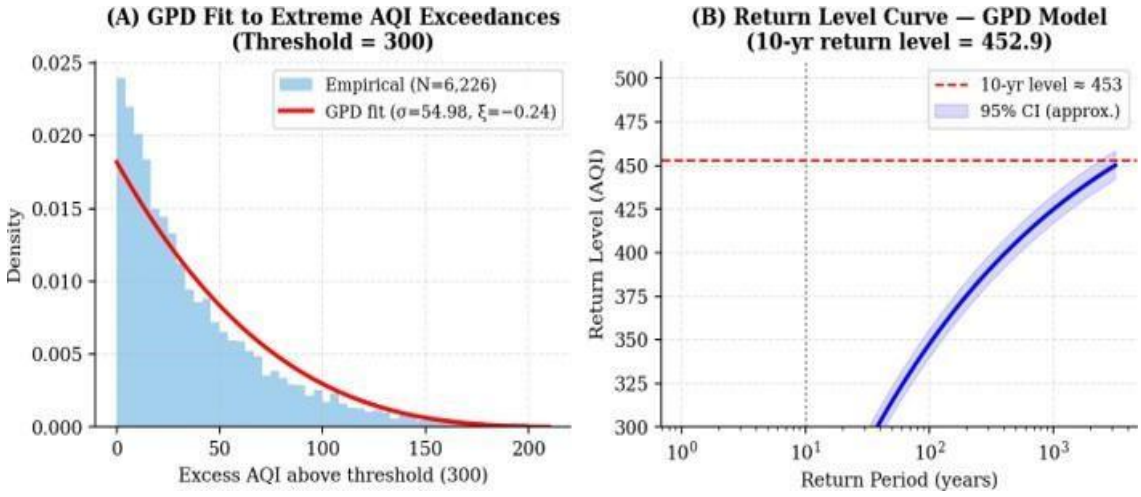

Figure 4. GPD diagnostics and return-level curve for AQI exceedances above 300.

The estimated 10-year return level is 452.88 AQI units. Under the fitted stationary GPD model and the return-period convention used in the estimaion, this value represents an extremely high level associated with a long recurrence horizon. It does not mean that every city will experience AQI 453 exactly once every ten years. The estimate summarizes tail risk for the pooled series used in the model and is conditional on the assumptions of threshold adequacy, stationarity, and the treatment of dependence.

6. Hybrid forecasting results

For the representative city series, automatic model selection identifies ARIMA (0,1,1). The moving-average estimate is -0.888 with standard error 0.068. Training RMSE is 25.37 and MAE is 19.14. These values describe in-sample fit and should not be used alone to establish forecasting superiority. XGBoost is then trained on ARIMA residuals using lagged AQI predictors. The feature-importance plot shows that lag-1 contributes most of the reported gain, indicating that the most recent AQI observation contains the strongest information for modelling the remaining residual structure. Smaller contributions from lag-7 and lag-30 suggest weaker additional information from weekly and monthly lags in this fitted example.

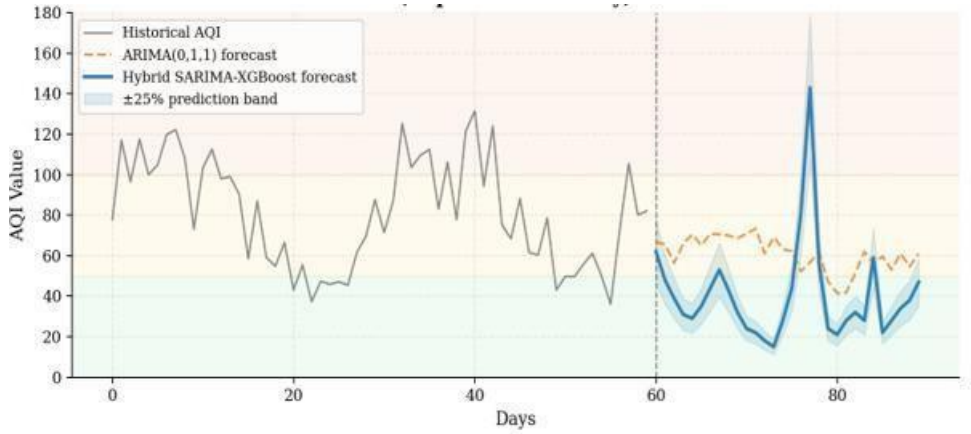

Figure 5. Thirty-step hybrid ARIMA–XGBoost forecast for a representative city.

Figure 5 presents the 30-step hybrid trajectory. The combined series shows short-run variation because the XGBoost residual component modifies the smoother ARIMA path. The figure illustrates how the two modelling components are combined. The present analysis is therefore interpreted as a demonstration of the hybrid framework rather than as evidence of forecasting superiority over standalone ARIMA. A formal out-of-sample comparison may be undertaken in future work.

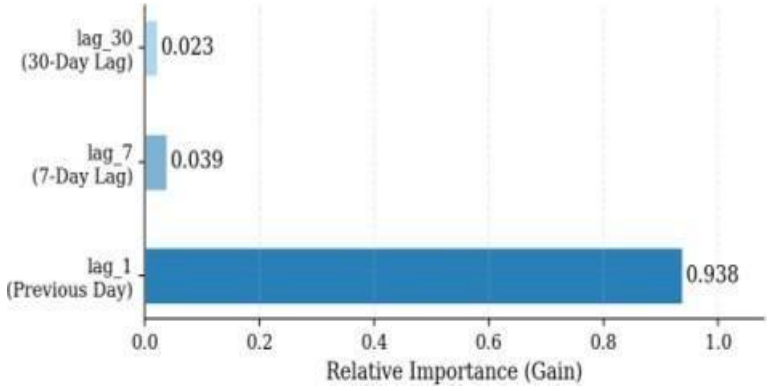

Figure 6. XGBoost feature importance for the residual component.

7. Pollutant transition dynamics

The transition matrix shows that PM_{10} is the most probable next state for most current pollutant states. For example, the probability of transition to PM_{10} is approximately 0.506 from CO, 0.458 from NH_3, 0.493 from PM_{10}, and 0.529 from $PM_{2.5}$. $PM_{2.5}$ is often the second most probable next state. Together, these patterns indicate the strong observed prominence of particulate matter in the sequence. The PM_{10} self-transition probability is approximately 0.493, meaning that when PM_{10} is currently prominent, it remains prominent in the next recorded state in nearly half of observed transitions. However, PM_{10} is not anabsorbing state because the probability of remaining in PM_{10} is less than one. It is more accurate to describe PM_{10} as a dominant or highly recurrent destination state.

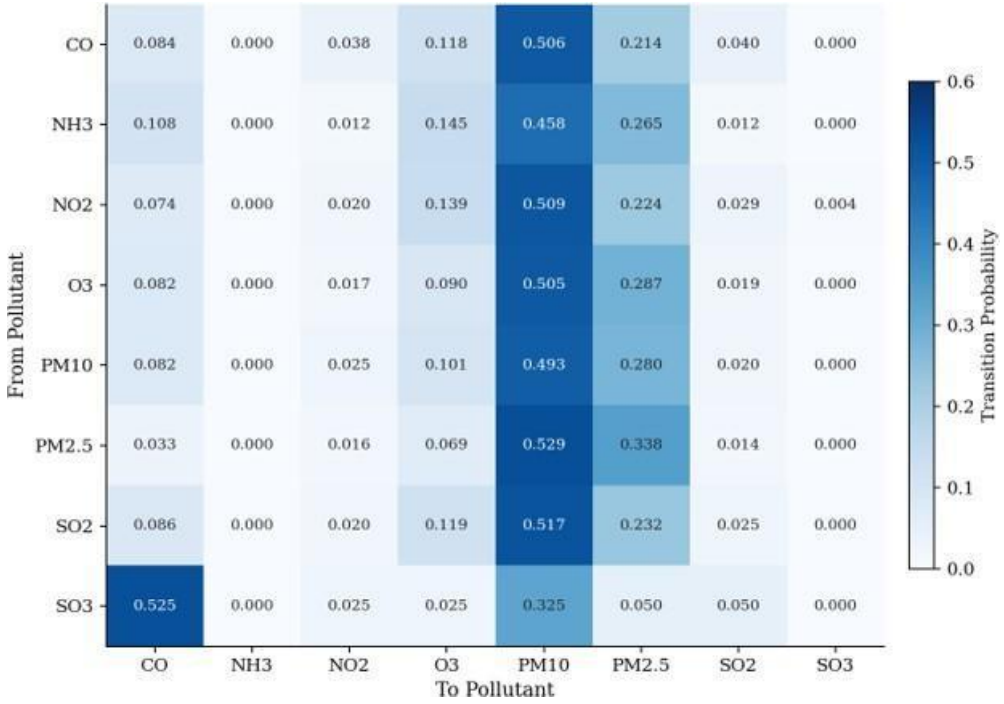

Figure 7. Heatmap of the pollutant transition probability matrix.

The SO_3 row differs from most others, with a relatively high transition probability to CO. Because SO_3 observations appear comparatively uncommon, this pattern should be interpreted cautiously. More generally, the matrix describes sequential association rather than atmospheric causation. The transition analysis is best interpreted as an exploratory description of pollutant-state patterns.

Table 3. Selected rows of the pollutant transition matrix.

From \ To	CO	$NH3$	$NO2$	$O3$	$PM10$	$PM_{2.5}$	$SO2$	$SO3$
CO	0.083	0.000	0.039	0.118	0.506	0.214	0.040	0.000
NH_3	0.108	0.000	0.012	0.145	0.458	0.265	0.012	0.000
PM_{10}	0.081	0.000	0.025	0.101	0.493	0.280	0.020	0.000
$PM_{2.5}$	0.033	0.000	0.016	0.069	0.529	0.338	0.014	0.000
SO_3	0.525	0.000	0.025	0.025	0.325	0.050	0.050	0.000

Policy Implications

The combined results suggest that air-quality management should be both seasonal and risk-sensitive. The large Winter and Post-Monsoon effects support intensified monitoring and mitigation before and during the months in which AQI typically deteriorates. Measures may include dust control, enforcement of industrial and transport standards, management of open burning, and timely public-health communication.

The extreme-value results provide a complementary perspective to average AQI. Emergency planning should account for rare but very high pollution levels, including protocols for health advisories, protection of vulnerable populations, and temporary restrictions where legally and operationally appropriate. Return levels can support preparedness, but they should be updated as longer data series become available and preferably estimated with models that allow seasonality and temporal dependence.

The transition analysis reinforces the importance of particulate matter. Because PM10 and PM2.5 frequently appear as next prominent states, interventions directed at road and construction dust, combustion sources, transport, and industrial emissions remain central. At the same time, source-specific policy conclusions require pollutant concentration data and source-apportionment evidence beyond the transition matrix alone.

Limitations and Future Research

This study has several limitations. First, the panel is unbalanced and monitoring coverage differs across cities and years. Second, the linear year trend is descriptive and does not identify the causal effect of a specific policy. Third, the pooled EVT model assumes stationarity and may be affected by serial and cross-city dependence; declustering and nonstationary GPD models would be useful extensions. Fourth, the forecasting illustration is based on a single city and is presented primarily to demonstrate the hybrid modelling framework; formal out-of-sample evaluation across multiple cities would strengthen future work. Fifth, the pollutant-transition analysis is exploratory and can be extended in future research through city-specific transition construction, state-frequency reporting, uncertainty measures, and sensitivity checks.

Future work can extend the framework through spatial panel models, city-specific or hierarchical EVT, nonstationary tail parameters, rolling-origin forecast evaluation, multi-city forecasting, meteorological covariates, and quasi-experimental designs for policy evaluation. Source-apportionment data could also be integrated to connect observed pollutant transitions with plausible emission mechanisms.

Conclusion

This study presents an integrated assessment of urban AQI dynamics across 292 Indian cities during the period from 1 April 2022 to 30 April 2025. The descriptive and fixed-effects results reveal pronounced seasonality, with the largest within-city AQI elevation in Winter and the lowest levels during Monsoon. The negative time coefficient indicates an overall downward trend within the study window, although causal attribution requires additional research. The EVT component demonstrates that the upper tail contains substantial environmental risk and yields a 10-year return-level estimate near 453. The hybrid forecasting framework illustrates how a statistical time-series model and machine learning can be combined to represent linear and residual nonlinear structure. The Markov chain analysis further shows the strong recurrence of particulate matter as a prominent pollutant state. Taken together, the methods provide complementary information: panel econometrics describes systematic within-city variation, EVT quantifies tail risk, hybrid forecasting addresses short-term temporal prediction, and Markov chains summarize pollutant-state dynamics. This multidimensional approach offers a more complete statistical description of urban air quality than any single method used in isolation.

Acknowledgement

The author gratefully acknowledges the Central Pollution Control Board (CPCB) for providing the air-quality data used in this study, covering 292 Indian cities from April 2022 to April 2025. The author also expresses sincere gratitude to the academic and research community whose scholarly contributions in air-quality assessment, panel-data econometrics, extreme value theory, time-series forecasting, machine learning, and Markov-chain modelling provided valuable methodological guidance. Special appreciation is extended to the institutions and individuals who supported the compilation, analysis, and interpretation of the data. Finally, the author acknowledges all those whose encouragement and assistance contributed to the successful completion of this research on urban air-quality dynamics in India.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Central Pollution Control Board (CPCB). (2025). National Air Quality Index. Ministry of Environment, Forest and Climate Change, Government of India.
2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
3. Cressie, N., & Wikle, C. K. (2011). Statistics for Spatio-Temporal Data. Wiley.
4. Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. Journal of Statistical Software, 27(2), 1–43.
5. Davison, A. C., & Smith, R. L. (1990). Models for exceedances over high thresholds. Journal of the Royal Statistical Society: Series B, 52(3), 393–425.
6. Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance. Springer.
7. Gupta, N. S., Mohta, Y., Heda, K., Armaan, R., Valarmathi, B., & Arulkumaran, G. (2023). Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. Journal of Environmental and Public Health, 2023, Article 4916267.
8. Hausman, J. A. (1978). Specification tests in econometrics. Econometrica, 46(6), 1251–1271. Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and Practice (3rd ed.). OTexts.
9. Liu, H., Li, Q., Yu, D., & Gu, Y. (2019). Air quality index and air pollutant concentration prediction based on machine learning algorithms. *Applied sciences*, *9*(19), 4069.
10. Natarajan, S. K., Shanmurthy, P., Arockiam, D., Balusamy, B., & Selvarajan, S. (2024). Optimized machine learning model for air quality index prediction in major cities in India. *Scientific reports*, *14*(1), 6795.
11. Spedicato, G. A. (2017). Discrete time Markov chains with R. The R Journal, 9(2), 84–104.
12. Udristioiu, M. T., Mghouchi, Y. E., & Yildizhan, H. (2023). Prediction, modelling, and forecasting of PM and AQI using hybrid machine learning. *Journal of Cleaner Production*, *421*, 138496.
13. Wang, S., Ren, Y., & Xia, B. (2023). Estimation of urban AQI based on interpretable machine learning. *Environmental Science and Pollution Research*, *30*(42), 96562-96574.
14. World Health Organization. (2021). WHO Global Air Quality Guidelines: Particulate Matter, Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. WHO.