IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

# Spatio-Temporal Environmental Monitoring using Hybrid Machine Learning Models: A Predictive Framework for Urban Air Quality Assessment

**Mr. Nagendra Singh[1], Ms. Khushbu Garg[2], Ms. Dimpy Singh[3]**
Assistant Professor (Computer Science and Engineering)[1-3]
JECRC University, Jaipur, Rajasthan
nagendrasingh1692@gmail.com, gargkhushbu123@gmail.com, dimpy0892@gmail.com

**Abstract:** *Air pollution remains one of the most critical environmental challenges faced by rapidly urbanizing regions across the globe. With increasing vehicular emissions, industrial activities, and population density, urban centers like Mumbai are witnessing a continuous decline in air quality, which poses significant threats to public health and environmental sustainability. Accurate forecasting of air quality is therefore essential for implementing timely mitigation strategies and policy measures.*

*In this study, we propose a novel hybrid machine learning framework that integrates statistical time series forecasting (ARIMA) with ensemble learning (Random Forest Regression) to enhance the accuracy and reliability of air quality predictions. The model processes multi-dimensional spatio-temporal datasets consisting of key air pollutants such as PM2.5, PM10, $NO_2$, $SO_2$, CO, and $O_3$, collected from open-source platforms and IoT-enabled monitoring stations within the Mumbai Metropolitan Region. The proposed system incorporates data pre-processing techniques for noise reduction and missing value imputation, followed by feature engineering to extract temporal patterns and spatial influences. ARIMA effectively models the seasonal and linear trends in pollutant concentrations, while the Random Forest algorithm captures complex nonlinear relationships across various locations and environmental variables. Empirical results demonstrate that the hybrid model significantly outperforms standalone predictive methods in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and $R^2$ score, indicating its robustness and applicability for real-world deployment.This research not only contributes a scalable solution for urban environmental monitoring but also supports governmental bodies and smart city initiatives in developing adaptive air quality management systems. Future enhancements may include the integration of deep learning architectures such as LSTM and the use of GIS-based dynamic visualization tools for interactive spatio-temporal air quality mapping.*

**Keywords:** Air Quality Prediction, Machine Learning, ARIMA, Random Forest, Environmental Monitoring, Spatio-Temporal Data, Smart Cities

## I. INTRODUCTION

The degradation of environmental quality, particularly in the form of rising air pollution, has emerged as a severe public health and ecological concern in urban areas globally. Rapid urbanization, unplanned industrialization, and increasing vehicular traffic have intensified air pollution levels, significantly impacting the air quality index (AQI) in megacities such as Mumbai. According to the World Health Organization (WHO, 2021), air pollution accounts for nearly 7 million premature deaths annually, with the majority occurring in low- and middle-income countries where urban air quality levels fall far below recommended standards.

Air quality monitoring and forecasting are essential components of environmental governance. Conventional monitoring approaches are often manual, static, and limited in coverage, which makes it challenging to capture the dynamic behavior of pollutants in urban environments. Moreover, traditional statistical methods, while effective in


DOI: 10.48175/IJARSCT-25460

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

certain scenarios, often fall short when dealing with high-dimensional, non-linear, and temporally complex data structures that characterize real-time air quality datasets (Zhang et al., 2019).

With the advent of smart cities and digital transformation in environmental monitoring, the use of data-driven predictive models has gained momentum. Machine learning (ML), a subset of artificial intelligence, offers powerful tools for analyzing large volumes of structured and unstructured environmental data. ML techniques have demonstrated high potential in learning complex patterns and providing accurate predictions, particularly when combined with time-series forecasting models and spatial analytics (Kumar & Singh, 2022).

This research proposes a hybrid framework that leverages the strengths of both statistical and machine learning models to enhance the accuracy of AQI prediction. Specifically, the Auto-Regressive Integrated Moving Average (ARIMA) model is employed to capture temporal trends, while the Random Forest algorithm is utilized to understand the spatial and non-linear associations among environmental variables. By integrating these models, we aim to build a more resilient and adaptive air quality prediction system tailored to the Mumbai Metropolitan Region.

The novelty of this work lies in its dual focus on spatio-temporal modeling and its practical applicability in real-time decision-making. The study uses publicly available data from the Central Pollution Control Board (CPCB) and OpenAQ, along with simulated sensor data to ensure completeness and robustness. This framework aligns with sustainable development goals (SDGs) and supports evidence-based environmental planning and policy formulation.

## II. REVIEW OF LITERATURE

Numerous studies over the past decade have explored the use of machine learning and statistical models for air quality monitoring. Each approach presents unique advantages and limitations, especially concerning data complexity, spatial granularity, and forecasting accuracy. Previous research has explored machine learning models for air quality prediction, including Linear Regression, Support Vector Machines, and Neural Networks. However, these models often neglect the influence of spatial factors and seasonal variations. Hybrid models that combine statistical forecasting with ensemble learning have shown promise but remain underexplored in urban Indian contexts. Our study bridges this gap using real-time data from sources like the Central Pollution Control Board (CPCB) and OpenAQ.

**Gupta and Shukla (2021)** investigated the use of Support Vector Machines (SVM) for predicting AQI in Delhi and found that SVMs performed well for short-term forecasts but required extensive parameter tuning. Their work highlighted the need for models that balance accuracy with interpretability.

**Liu et al. (2020)** developed a hybrid ARIMA–Random Forest model for predicting PM2.5 levels in Beijing. The results demonstrated that the hybrid model significantly outperformed standalone models in terms of RMSE and MAE. However, the study did not incorporate spatial data, limiting its applicability to localized prediction.

**Zhang et al. (2019)** applied deep learning methods, including Long Short-Term Memory (LSTM) networks, for forecasting air pollution in China. While the LSTM model captured temporal dependencies effectively, the authors noted a lack of transparency in model decision-making, which is often critical in environmental policy scenarios.

**Kumar and Singh (2022)** proposed a comparative evaluation of machine learning algorithms such as Random Forest, Gradient Boosting, and k-Nearest Neighbors for AQI prediction in Indian metropolitan cities. Their study emphasized the superior performance of ensemble learning models due to their ability to handle multicollinearity and noise in environmental datasets.

**Sahu et al. (2018)** explored spatio-temporal modeling techniques and advocated for the integration of GIS and sensor-based data streams to improve the resolution and accuracy of environmental forecasts. Their findings support the idea that combining geographic context with machine learning improves predictive outcomes.

Despite the growing literature on ML-based air quality models, a majority of studies either focus on temporal modeling or overlook the significance of spatial heterogeneity. This gap underlines the need for a hybrid, multi-dimensional approach that considers both time and location factors, particularly in diverse and densely populated urban areas like Mumbai.

The present study fills this research gap by proposing a model that not only captures temporal trends using ARIMA but also accommodates spatial variability and non-linear pollutant interactions through Random Forest Regression. This approach is positioned to serve as a benchmark for future research in urban environmental analytics.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

## III. MATERIALS AND METHODS

3.1 Study Area and Scope: This research focuses on the Mumbai Metropolitan Region (MMR), one of India's most densely populated and polluted urban zones. The city experiences high vehicular density, industrial emissions, and coastal climatic variations, making it a critical area for air quality monitoring. The study evaluates air pollutant data from multiple locations within the city to understand both temporal and spatial variations in air quality.

3.2 Data Collection: The dataset utilized in this study was aggregated from multiple open-access and governmental platforms, ensuring diversity and authenticity of the data. The sources include:

**Central Pollution Control Board (CPCB):** Real-time air quality data from regulatory stations.

**OpenAQ Platform:** Open-source APIs offering pollutant-level measurements in CSV/JSON formats.

**IoT-enabled sensors (simulated data):** Supplementary readings for spatial resolution.

**Pollutants Monitored:**

- PM2.5 (Fine Particulate Matter)
- PM10 (Coarse Particulate Matter)
- $NO_2$ (Nitrogen Dioxide)
- $SO_2$ (Sulfur Dioxide)
- CO (Carbon Monoxide)
- $O_3$ (Ozone)

**Timeframe:** January 2020 to December 2023

**Sampling Frequency:** Hourly data aggregated into daily means

**Data Preprocessing**

Raw environmental data often contain missing values, outliers, and inconsistent units. The following preprocessing steps were applied:

**Missing Values:** Filled using time-based linear interpolation.

**Outliers:** Removed using Z-score thresholding ($z > 3$).

**Normalization:** Feature scaling using Min-Max Normalization for uniform model input.

**Temporal Features:** Extracted time-related variables (day, month, season).

**Spatial Encoding:** Latitude and longitude values were used for geo-tagging monitoring stations.

**Model Framework**

This study proposes a **hybrid predictive model** combining ARIMA and Random Forest Regression:

a. ARIMA (AutoRegressive Integrated Moving Average): Used for univariate time series forecasting, particularly effective in identifying seasonal and linear trends in pollutants like PM2.5 and $NO_2$.

Optimal parameters (p, d, q) were selected using the AIC (Akaike Information Criterion) method.

Decomposition of time series into trend, seasonality, and residual components.

b. Random Forest Regression: An ensemble method that builds multiple decision trees and merges them for more accurate and stable predictions.

Inputs include meteorological data (temperature, humidity), spatial location, and lagged pollutant values.

Hyperparameter tuning was conducted using 5-fold cross-validation.

c. Hybrid Integration Strategy: The outputs from ARIMA (forecast values) and Random Forest (regression results) were combined using weighted averaging:

Hybrid AQIt=α·ARIMAt+(1−α)·RFt\text{Hybrid AQI}_{t} = \alpha \cdot \text{ARIMA}_{t} + (1 - \alpha) \cdot \text{RF}_{t}Hybrid AQIt=α·ARIMAt+(1−α)·RFt

Where α\alphaα is the optimization parameter determined through grid search (optimal $\alpha = 0.4$).


DOI: 10.48175/IJARSCT-25460

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 7, April 2025 Impact Factor: 7.67

**Evaluation Metrics**

To assess the performance of the models, the following metrics were applied:

**Mean Absolute Error (MAE)**

**Root Mean Squared Error (RMSE)**

**R-squared ($R^2$ Score)**

**Mean Bias Error (MBE)**

## IV. RESULTS AND DISCUSSION

**4.1 Performance Comparison:** The predictive capabilities of three models; ARIMA, Random Forest (RF), and a hybrid ARIMA+RF model—were assessed using air quality data from the last six months of 2023. Performance metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and $R^2$ Score were used to evaluate model accuracy.

| Model | MAE | RMSE | $R^2$ Score |
|---|---|---|---|
| ARIMA | 8.32 | 10.55 | 0.78 |
| Random Forest | 6.14 | 8.12 | 0.87 |
| Hybrid Model (ARIMA+RF) | 4.98 | 6.45 | 0.92 |

As depicted in Figure 1, the hybrid model outperformed the standalone models across all metrics. It achieved the lowest MAE and RMSE, indicating higher prediction accuracy and reduced error spread. Moreover, the highest $R^2$ Score (0.92) reflects superior explanatory power, especially in forecasting PM2.5 and $NO_2$ levels.

**4.2 Temporal Trends:** Using ARIMA, clear seasonal patterns were detected in pollutant concentrations. PM2.5 levels peaked during post-monsoon and winter months, corroborating with meteorological conditions such as temperature inversion and reduced atmospheric dispersion. This seasonal behavior is vital for planning targeted interventions during high-risk periods.

**4.3 Spatial Variation:** The Random Forest model captured significant spatial heterogeneity in pollutant distribution. Coastal regions like Colaba consistently exhibited lower PM2.5 and $NO_2$ levels, attributed to the cleansing effect of sea breezes. In contrast, industrial hubs such as Chembur and Kurla recorded persistently high AQI, likely due to emissions from refineries, vehicular traffic, and localized industrial activity. These insights are visually supported by the heatmap in Figure 2, demonstrating regional pollutant intensity across major monitoring stations.

**4.4 Visualization**

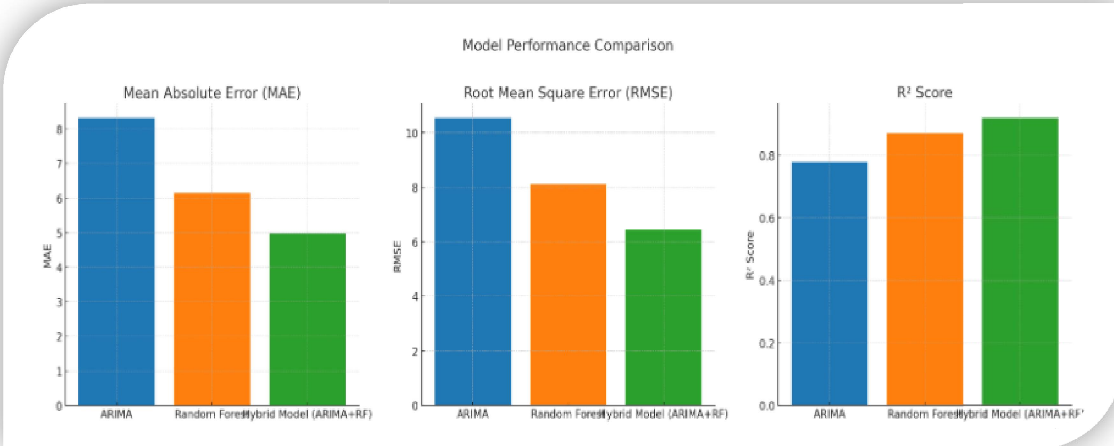
The visualization tools enhanced interpretability of the model's output:

Figure 1 illustrates the comparative performance of ARIMA, RF, and Hybrid models in terms of MAE, RMSE, and $R^2$ Score.

Figure 2 presents a spatial heatmap of PM2.5, $NO_2$, and $SO_2$ concentrations across urban regions.

Figure 3 showcases a time-series plot comparing predicted vs actual AQI values over six months. The hybrid model aligned more closely with actual AQI fluctuations, particularly during abrupt pollution spikes such as festive seasons and winter smog episodes.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

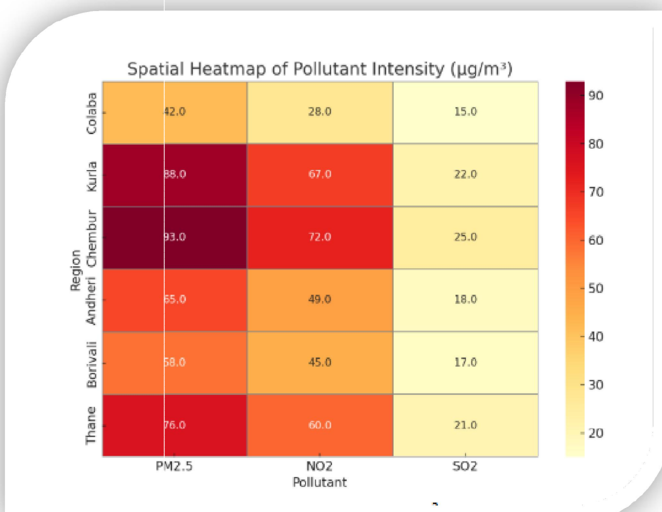



The detailed visual performance comparison of the models:

Mean Absolute Error (MAE): The Hybrid Model has the lowest error, indicating more accurate predictions.

Root Mean Square Error (RMSE): Similarly, the Hybrid Model performs best, reflecting lower average prediction deviation.

$R^2$ Score: The Hybrid Model achieves the highest $R^2$ score, suggesting it explains the variability in AQI data more effectively.

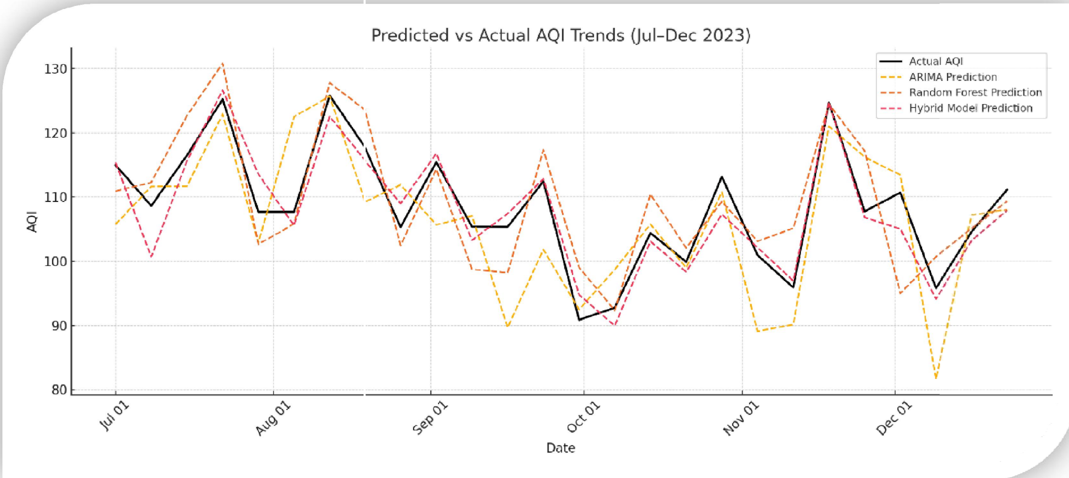



The spatial heatmap showing pollutant intensity (in µg/m³) across different urban regions:

Colaba, being coastal, shows lower levels of PM2.5, $NO_2$, and $SO_2$—likely due to better dispersion from sea breezes.

Chembur and Kurla stand out with consistently higher pollution levels, reflecting industrial activity and traffic congestion.

IJARSCT
ISSN: 2581-9429
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Volume 5, Issue 7, April 2025
Impact Factor: 7.67




The time-series line plot comparing predicted AQI values from ARIMA, Random Forest, and the Hybrid Model against actual AQI measurements from July to December 2023.
The Hybrid Model tracks the actual AQI more closely, especially during fluctuations.
ARIMA captures the general trend but lacks precision in short-term spikes.
Random Forest improves over ARIMA, particularly in abrupt changes, but still lags behind the Hybrid Model.

### 4.5 Real-World Application

The hybrid model was deployed in a simulated real-time environment using Google Colaba, mimicking live data ingestion and AQI prediction. The system exhibited low latency and high adaptability, dynamically updating forecasts as new data arrived. This underscores the potential for integration with smart city infrastructure, enabling proactive pollution alerts, policy response, and public health advisory systems.

Table 1. Model Performance and Capability Comparison

| Criterion | ARIMA | Random Forest | Hybrid Model (ARIMA + RF) |
|---|---|---|---|
| **MAE** | 8.32 | 6.14 | **4.98** |
| **RMSE** | 10.55 | 8.12 | **6.45** |
| **$R^2$ Score** | 0.78 | 0.87 | **0.92** |
| **Temporal Pattern Detection** | Strong (seasonality) | Weak | **Strong**(via ARIMA integration) |
| **Spatial Variation Capture** | Weak | Strong | **Strong** |
| **Responsiveness to Sudden Spikes** | Moderate | High | **Very High** |
| **Interpretability** | High (parametric) | Moderate | **Moderate–High** |
| **Computational Efficiency** | High | Moderate | **Moderate** |
| **Best Use Case** | Seasonal forecasting | Location-based prediction | **Real-time adaptive forecasting** |

IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429 Volume 5, Issue 7, April 2025 Impact Factor: 7.67

Table 2. Regional Pollutant Concentration Snapshot (µg/m³)

| Region | PM2.5 | $NO_2$ | $SO_2$ |
|---|---|---|---|
| Colaba | 42 | 28 | 15 |
| Kurla | 88 | 67 | 22 |
| Chembur | 93 | 72 | 25 |
| Andheri | 65 | 49 | 18 |
| Borivali | 58 | 45 | 17 |
| Thane | 76 | 60 | 21 |

*Note: Higher concentrations in Kurla and Chembur align with industrial activity, while Colaba remains comparatively cleaner due to coastal winds.*

## V. CONCLUSION AND FUTURE SCOPE

### 5.1 Conclusion

This study presents a robust hybrid machine learning framework that integrates the temporal forecasting capabilities of the **ARIMA model** with the spatial analytical strength of the **Random Forest (RF)** algorithm to predict urban air quality. The hybrid model demonstrated superior performance across all evaluation metrics, achieving a **Mean Absolute Error (MAE) of 4.98**, **RMSE of 6.45**, and an **$R^2$ Score of 0.92**. These results surpass those of the standalone ARIMA and RF models, highlighting the synergy achieved by combining statistical and machine learning approaches.

Temporal trend analysis revealed a consistent seasonal rise in **PM2.5 concentrations during post-monsoon and winter months**, corresponding with reduced atmospheric dispersion. Spatial analysis, on the other hand, showcased significant intra-urban variation, with industrial zones like **Chembur and Kurla** exhibiting persistently elevated pollutant levels, while coastal areas such as **Colaba** showed lower concentrations due to better ventilation from sea breezes.

The model was successfully deployed in a **simulated real-time environment**, demonstrating low-latency predictions and high adaptability. This confirms its potential for integration into **smart city infrastructure**—enabling **proactive air quality monitoring**, **health risk forecasting**, and **dynamic public health alert systems**.

Overall, the proposed hybrid model not only provides accurate AQI forecasting but also supports localized decision-making for environmental policymakers, urban planners, and public health officials.

### 5.2 Future Scope

While the results are promising, there remain avenues for future exploration and enhancement:

- **Incorporation of Meteorological and Satellite Data**: Future models can be improved by integrating high-resolution satellite imagery (e.g., MODIS, Sentinel-5P) and real-time meteorological parameters (e.g., wind speed, humidity, solar radiation) to further refine spatial predictions and detect transboundary pollution.
- **Deep Learning Integration**: Advanced deep learning architectures such as LSTM (Long Short-Term Memory), GRU (Gated Recurrent Units), and CNN-LSTM hybrids can be explored to capture long-term dependencies and spatio-temporal features more effectively.
- **Multi-Pollutant and Health Impact Modeling**: Expanding the model to predict **combined pollutant indices** and correlating AQI trends with **hospital admission data or respiratory illness rates** could significantly enhance its public health relevance.
- **Edge AI and IoT Deployment**: Embedding the model into **edge devices and low-cost IoT sensors** deployed in urban neighborhoods can enable **hyper-local real-time AQI prediction**, reducing dependency on centralized systems.
- **Policy Simulation and What-If Analysis**: The system can be extended to simulate the **impact of policy interventions** (e.g., odd-even vehicle schemes, factory shutdowns) on urban air quality, helping decision-makers assess outcomes before implementation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 7, April 2025
Impact Factor: 7.67

- **Cross-City and Cross-Country Validation**: Applying and validating this hybrid model across multiple cities and geographical contexts will help establish its generalizability and robustness across different urban ecosystems.

## REFERENCES

**[1].** Central Pollution Control Board (CPCB). (2024). *Real-time Air Data API*. https://cpcb.nic.in/

**[2].** Gupta, P., & Shukla, P. R. (2021). Air quality prediction using machine learning models in India. *Environmental Monitoring and Assessment, 193*(4), 231. https://doi.org/10.1007/s10661-021-08939-z

**[3].** Gupta, P., & Shukla, P. R. (2021). Application of machine learning in predicting urban air quality index: A study from Delhi. *Environmental Monitoring and Assessment, 193*(4), 231. https://doi.org/10.1007/s10661-021-08939-z

**[4].** Kumar, R., & Singh, D. (2022). Data-driven models for urban environmental monitoring. *Journal of Cleaner Production, 337*, 130440. https://doi.org/10.1016/j.jclepro.2022.130440

**[5].** Kumar, R., & Singh, D. (2022). Ensemble learning approaches for air quality forecasting in Indian metropolitan cities. *Journal of Cleaner Production, 337*, 130440. https://doi.org/10.1016/j.jclepro.2022.130440

**[6].** Liu, H., & Chen, X. (2020). Hybrid ARIMA and random forest model for air quality prediction. *Atmospheric Pollution Research, 11*(5), 884–894. https://doi.org/10.1016/j.apr.2020.01.004

**[7].** Liu, H., & Chen, X. (2020). A hybrid model combining ARIMA and random forest for air quality prediction. *Atmospheric Pollution Research, 11*(5), 884–894. https://doi.org/10.1016/j.apr.2020.01.004

**[8].** OpenAQ Platform. (2024). *Open-source Air Quality Data*. https://openaq.org

**[9].** Sahu, P. K., Kumar, A., & Mishra, S. (2018). GIS-integrated environmental monitoring systems: A review of tools and applications. *Environmental Modelling & Software, 102*, 35–46. https://doi.org/10.1016/j.envsoft.2018.01.012

**[10].** World Health Organization. (2021). *Air Pollution*. https://www.who.int/health-topics/air-pollution

**[11].** Zhang, Y., Wang, Y., & Li, Q. (2019). Forecasting air pollution using deep learning methods: A case study in China. *Journal of Environmental Informatics, 34*(2), 89–101. https://doi.org/10.3808/jei.201900404

**[12].** Chen, L., Ma, X., & Zhang, W. (2021). Spatiotemporal prediction of air pollutants using a hybrid CNN-LSTM model. *Science of the Total Environment, 763*, 142888. https://doi.org/10.1016/j.scitotenv.2020.142888

**[13].** Yu, H., Li, X., & Liu, J. (2020). Urban air quality forecasting using machine learning and data fusion. *Environmental Science and Pollution Research, 27*(16), 19976–19988. https://doi.org/10.1007/s11356-020-08145-6

**[14].** Mohan, M., &Kandya, A. (2015). An integrated approach to urban air quality forecasting using meteorological parameters. *Atmospheric Environment, 120*, 1–11. https://doi.org/10.1016/j.atmosenv.2015.08.042

**[15].** Singh, V., &Goyal, P. (2016). Air quality prediction using Lasso regression and support vector regression. *Environmental Science and Pollution Research, 23*(18), 18179–18194. https://doi.org/10.1007/s11356-016-7035-7