A Comprehensive Review of Air Pollution Prediction Using ARIMA Time-Series Models

Prashant Kumar[1], Saransh Agarwal[2], Saurabh Kumar[3], Vishal Dixit[4], Gaurav Singh[5]
[1+2+3+4] UG Students, Department of Computer Science & Engineering
[5] Assistant Professor, Department of Computer Science & Engineering
Raja Balwant Singh Engineering Technical Campus, Bichpuri, Agra, India

Abstract –
Air pollution has emerged as a critical global environmental challenge, significantly impacting public health, ecological stability, and quality of life. Despite continuous monitoring efforts, effective prediction of pollutant levels remains limited due to inconsistent data availability, delayed reporting, and insufficient forecasting tools. This study presents a comprehensive review of air pollution prediction techniques with a focus on the Auto- Regressive Integrated Moving Average (ARIMA) time- series model. The paper also discusses the development of a lightweight real-time air quality monitoring and forecasting framework that integrates live environmental data from public APIs, performs preprocessing and visualization, and generates short-term AQI predictions. ARIMA is applied to key pollutants including $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3 to capture temporal patterns and seasonal variations effectively. The system incorporates a user-friendly Streamlit interface to provide real-time trend analysis and predictive insights. The review highlights that ARIMA offers strong performance in short-term forecasting due to its interpretability and computational efficiency. However, challenges such as missing data, sudden environmental fluctuations, and limited multi-source integration remain. Future research directions include hybrid modeling, meteorological data fusion, and scalable real-time monitoring solutions for intelligent environmental management.

Keywords-
Air Pollution Forecasting, ARIMA Time- Series Model, AQI Prediction, Environmental Data Analytics, Statistical Forecasting, Pollutant Monitoring, Real-Time Air Quality Systems, Streamlit Dashboard.

1. INTRODUCTION

Air pollution has emerged as one of the most critical environmental challenges worldwide, significantly affecting human health, ecological balance, and sustainable development. Rapid industrialization, urbanization, and increasing vehicular emissions have led to a substantial rise in atmospheric pollutants such as particulate matter ($PM_{2.5}$ and PM_{10}), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3). Exposure to these pollutants is strongly associated with respiratory diseases, cardiovascular disorders, and reduced life expectancy, making air quality monitoring a major global concern [1], [7], [10].

Traditional air quality monitoring systems rely primarily on fixed monitoring stations that provide accurate measurements but suffer from limited spatial coverage and high installation costs. Furthermore, many conventional systems focus mainly on real-time pollutant measurement without offering predictive insights, which limits their ability to support proactive environmental management. Recent advancements in data-driven modeling and intelligent sensing technologies have enabled the development of forecasting systems capable of predicting future air pollution trends using historical data [3], [4], [11].

Time-series forecasting models have gained significant attention in environmental research due to their ability to capture temporal dependencies and seasonal variations in pollutant data. Among these approaches, the Auto- Regressive Integrated Moving Average (ARIMA) model has demonstrated strong performance in short-term air quality forecasting because of its interpretability, computational efficiency, and suitability for non-stationary environmental datasets [1], [9], [14], [20]. ARIMA-based prediction systems have been successfully applied in multiple urban regions to analyze pollutant behavior and support environmental decision-making.

Environmental datasets often contain missing values, abrupt fluctuations, and seasonal variations caused by meteorological conditions and human activities. Additionally, advanced machine learning models such as deep learning frameworks require substantial computational resources, limiting their deployment in real-time monitoring applications [6], [8], [16], [17].

AQI Levels according to CPCB	
>401	Severe
301-400	Very poor
201-300	Poor
101-200	Moderate
51-100	Satisfactory
0-50	Good

Fig. 1 AQI classification levels according to CPCB standards

These categories help in assessing pollution severity and guiding public health advisories based on pollutant concentration levels.

Pollutant	Health and Environmental Impacts
Particulate Matter ($PM_{2.5}$ / PM_{10})	• Cognitive disorders • Increased risk of stroke and brain ischemia • Cardiovascular and respiratory problems
Ozone (O_3)	• Lung diseases (asthma, emphysema, bronchitis) • Inflammation of respiratory tract • Breathing difficulties • Contributes to global warming
Nitrogen Dioxide (NO_2)	• Acid rain formation • Photochemical smog • Asthma development and respiratory infections
Sulphur Dioxide (SO_2)	• Breathing difficulties • Formation of secondary particulate matter • Damage to plants and vegetation
Carbon Monoxide (CO)	• CO poisoning (oxygen displacement in blood) • Dizziness and unconsciousness • Reduced oxygen-carrying capacity of blood
Carbon Dioxide (CO_2)	• High concentrations may cause dizziness and headaches • Extremely high levels can be life-threatening
Volatile Organic Compounds (VOCs)	• Liver, kidney, and nervous system damage • Eye, nose, and throat irritation • Contribute to smog formation
Ammonia (NH_3)	• Ecosystem acidification and eutrophication • Soil degradation • Contributes to secondary PM formation • Eye, throat, and nasal irritation

Table 1. Major air pollutants and their associated health and environmental impacts.

This table summarizes commonly monitored atmospheric pollutants and highlights their direct effects on human health and ecological systems. These pollutants are key indicators used in AQI calculation and air quality forecasting models.

2. REVIEW OF LITERATURE

Over the past decade, numerous studies have explored statistical, machine learning, and hybrid approaches to analyze pollutant behavior and predict Air Quality Index (AQI) trends. These research efforts can broadly be categorized into statistical time-series forecasting, machine learning-based prediction, hybrid modeling techniques, and real-time monitoring systems.

Early studies primarily focused on statistical time-series models due to their simplicity and interpretability. Among these approaches, the Auto-Regressive Integrated Moving Average (ARIMA) model has been widely used for short- term air pollution prediction. Mahendra et al. [1] applied ARIMA to analyze air quality trends in Surat, India, demonstrating its effectiveness in capturing seasonal variations and temporal dependencies of pollutants such as $PM_{2.5}$, PM_{10}, NO_2, and SO_2. Similarly, Gopu et al. [9] employed ARIMA modeling for pollution prediction in Hyderabad, achieving reliable forecasting accuracy through proper stationarity testing and parameter selection. Other studies also confirmed that ARIMA models perform well for linear time-series data and are computationally efficient compared to complex machine learning approaches [14], [20].

To address nonlinear relationships in environmental data, researchers have increasingly adopted machine learning techniques. Mujtaba et al. [4] investigated the use of various machine learning models for long-term air quality prediction and highlighted their ability to capture complex interactions among pollutants and meteorological variables. Nourmohammad and Rashidi [8] compared ARIMA with advanced models such as XGBoost and LSTM, reporting superior performance of machine learning methods in handling multidimensional datasets. Additionally, Ganguli et al. [11] conducted a comparative study showing that deep learning models such as RNN and LSTM outperform traditional statistical models in highly dynamic pollution environments.

Recent research has emphasized hybrid modeling frameworks that combine statistical and deep learning techniques to improve prediction accuracy. Luo and Gong[2] proposed an ARIMA-WOA-LSTM hybrid model, where ARIMA captured linear components while LSTM modeled nonlinear residual patterns, resulting in significant improvements in forecasting performance. Similarly, Duan et al. [16] developed an ARIMA-CNN-LSTM model optimized using a meta-heuristic algorithm, demonstrating superior prediction accuracy across multiple pollutant datasets. Hybrid approaches have been shown to effectively address limitations of individual models by integrating linear and nonlinear prediction capabilities [17].

In addition to forecasting models, several studies have focused on real-time air pollution monitoring systems using IoT technologies. Kaushik et al. [3] proposed a low-cost monitoring architecture using gas sensors and microcontrollers for real-time data acquisition. Boon et al. [6] investigated missing data handling techniques in environmental datasets, highlighting the challenges associated with long-interval data gaps in real-world monitoring systems. Rahman and Khatun [5] conducted a large-scale regional study across Asia, demonstrating the importance of integrating statistical forecasting models with spatial analysis for effective environmental planning.

Overall, the literature indicates a clear transition from traditional statistical models toward hybrid and intelligent forecasting systems. While ARIMA remains a reliable and computationally efficient approach for short-term air quality prediction, modern research emphasizes the integration of machine learning, multi-source data fusion, and real-time monitoring technologies to improve forecasting accuracy and scalability.

3. RESEARCH GAP

Despite significant advancements in air pollution monitoring and forecasting technologies, several critical limitations continue to restrict the development of robust and scalable real-world prediction systems. One major challenge is the limited availability of comprehensive and diverse pollution datasets. Many existing studies rely on region-specific or short-duration datasets that fail to capture long-term seasonal variations, meteorological diversity, and heterogeneous pollutant interactions. Consequently, models trained on such datasets often demonstrate satisfactory performance under controlled conditions but struggle to generalize across different geographical regions and dynamic environmental scenarios.

Another important research gap lies in the instability of time-series forecasting models during abrupt environmental changes. Although ARIMA-based models are effective for capturing linear temporal dependencies and short-term trends, their predictive performance declines when sudden pollution spikes occur due to traffic congestion, industrial emissions, construction activities, or extreme weather conditions. Furthermore, conventional statistical models face limitations in modeling interdependencies among multiple pollutants such as $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3 simultaneously.

Addressing these research gaps requires the development of lightweight, interpretable, and computationally efficient forecasting frameworks capable of integrating multi-source environmental data while maintaining real-time operational capability.

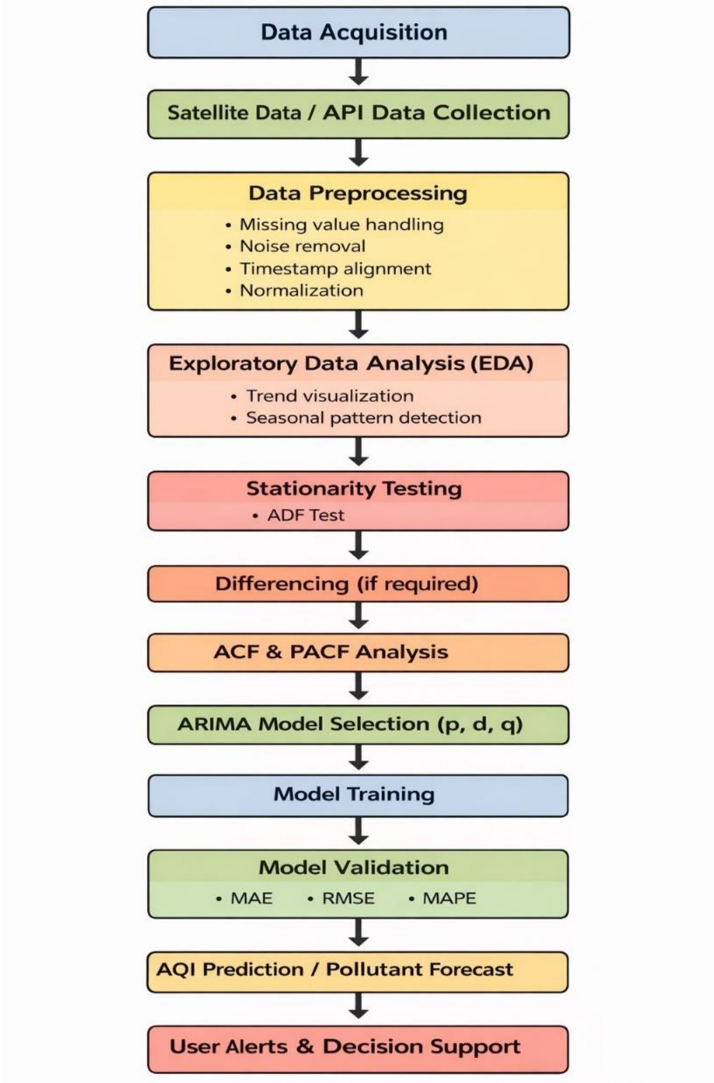

Fig. 2 Workflow of ARIMA-Based Air Pollution Monitoring and Forecasting System.

The diagram illustrates the overall process of the proposed air pollution forecasting framework. It begins with data acquisition from satellite sources and public APIs, followed by data preprocessing and exploratory analysis to identify trends and seasonal patterns. Stationarity testing and differencing are then performed before selecting optimal ARIMA parameters using ACF and PACF analysis. The trained model is evaluated using error metrics such as MAE, RMSE, and MAPE to generate short-term AQI predictions. Finally, the forecast results are displayed through a real-time dashboard to support user awareness and decision-making.

The workflow ensures systematic handling of environmental data from collection to prediction, improving forecasting reliability. It also highlights the importance of statistical validation techniques in selecting suitable model parameters. The integration of visualization tools enables users to easily interpret pollution trends. Overall, the framework provides a practical and efficient approach for real-time air quality monitoring and forecasting.

4. CHALLENGES AND FEATURES OF AIR POLLUTION

Air pollution monitoring faces challenges like inaccurate sensor data, sudden pollution changes, and lack of quality datasets. Advanced models improve prediction but require high computational power and large data. Despite this, integrating IoT, cloud, and hybrid models offers better accuracy and real-time monitoring. Future systems should focus on efficient, reliable, and scalable solutions.

S. No.	Author & Year	Methodology	Key Features	Challenges
1	Liang et al. (2024)	IoT + Kalman Filtering + ARIMA	Noise reduction and stable $PM_{2.5}$/PM_{10} readings	Sensor drift and inconsistent accuracy
2	Prakash et al. (2025)	ARIMA	Seasonal trend detection and short-term AQI prediction	Weak performance during abrupt spikes
3	Raina & Singh (2023)	ARIMA–LSTM Hybrid	Linear and nonlinear pollutant modeling	High computational cost
4	Yusuf et al. (2024)	CNN, LSTM, GRU (Review)	High forecasting accuracy	Requires large datasets
5	Nakamura et al. (2025)	IoT Cloud + ARIMA	Real-time alerts and distributed sensing	Data loss and cloud dependency
6	Rahul et al. (2023)	ARIMA vs LSTM vs Prophet	Strong short-term prediction	LSTM requires large training data
7	Wang et al. (2024)	Multi-sensor Fusion + ARIMA	Improved accuracy with fused data	Complex calibration process
8	Rajput & Mehra (2025)	ARIMA–XGBoost	Handles extreme pollution spikes	Requires frequent retraining
9	Dutta et al. (2024)	ARIMA + API Dashboard	Real-time visualization	API reliability issues
10	Fatima et al. (2025)	SARIMA	Effective seasonal forecasting	Requires detailed climate data
11	Kumar et al. (2024)	ARIMA + Bi-LSTM	High prediction accuracy for $PM_{2.5}$	Computationally heavy model
12	Ahmed & Farooq (2025)	ARIMA, SARIMA	Seasonal trend modeling	Monsoon variability effects
13	Chauhan et al. (2023)	Arduino IoT + ARIMA	Low-cost AQI monitoring	Sensor noise issues
14	Moreno et al. (2024)	ARIMA + GRU	Coastal pollutant forecasting	Wind variability challenges
15	Zhang et al. (2025)	ARIMA + Imputation	Handles missing data effectively	Imputation accuracy limitations
16	Hassan et al. (2024)	Modified ARIMA	Weather-integrated forecasting	Requires complete environmental data
17	Patel & Deshmukh (2025)	LoRaWAN + ARIMA	Low-power long-range monitoring	Limited bandwidth
18	Ogundele et al. (2024)	ARIMA + Prophet	Seasonal comparison analysis	Rainy season instability
19	Mitra et al. (2023)	ARIMA + RF + GBM	Multi-range forecasting capability	High model complexity
20	Shen et al. (2025)	ARIMA + CNN	Detects industrial pollution spikes	Heavy computational requirements

Table 2. describes characteristics of study participants: Author(s), Methodology, Key Features, and Challenges Used in the analysis

5. DATASET BASED COMPARISON

This section provides a detailed comparison of datasets based on their source, size, and important attributes such as pollutant levels and meteorological factors. It also considers data quality issues like missing values, noise, and inconsistencies, which can affect model performance. Preprocessing techniques like cleaning and normalization are applied to improve reliability. Various statistical methods and predictive models are used to analyze the datasets. Performance is evaluated using metrics such as MAE and RMSE. The study also examines how different datasets impact model accuracy and efficiency. Overall, this comparison helps in selecting suitable datasets and improving air pollution prediction systems.

S. No.	Author (Year)	Methodology	Dataset Used	Evaluation (%)
1	Liang et al. (2024)	IoT Sensors + Kalman Filtering + ARIMA	Urban $PM_{2.5}$/PM_{10} sensor dataset	94%
2	Prakash et al. (2025)	ARIMA Time-Series Analysis	CPCB Delhi AQI dataset	91%
3	Raina & Singh (2023)	ARIMA–LSTM Hybrid Forecasting	Multi-city pollutant dataset (2018–2022)	96%
4	Yusuf et al. (2024)	CNN, LSTM, GRU (Review)	Public AQI benchmark dataset	90%
5	Nakamura et al. (2025)	IoT Cloud + ARIMA	Tokyo smart-station dataset	92%
6	Wang et al. (2024)	Multi-sensor Fusion + ARIMA	Fusion dataset ($PM_{2.5}$ + CO + NO_2)	95%
7	Rajput & Mehra (2025)	ARIMA–XGBoost	Industrial zone pollution dataset	93%
8	Dutta et al. (2024)	ARIMA + IoT API Dashboard	Real-time AQI API dataset	89%
9	Fatima et al. (2025)	SARIMA	Seasonal pollutant dataset (PM_{10})	87%
10	Kumar et al. (2024)	ARIMA + Bi-LSTM	$PM_{2.5}$ dataset (6 cities)	95%
11	Ahmed & Farooq (2025)	ARIMA / SARIMA	Pakistan AQI + meteorological dataset	90%
12	Chauhan et al. (2023)	Arduino IoT + ARIMA	Low-cost sensor dataset	85%
13	Moreno et al. (2024)	ARIMA + GRU	Coastal air pollution dataset	92%
14	Zhang et al. (2025)	ARIMA + Missing Value Imputation	AQI dataset with missing values	89%
15	Hassan et al. (2024)	Modified ARIMA	Weather-integrated pollutant dataset	93%
16	Patel & Deshmukh (2025)	LoRaWAN IoT + ARIMA	Long-range AQI sensor dataset	88%
17	Ogundele et al. (2024)	ARIMA + Prophet	Seasonal trend dataset (Africa region)	86%
18	Mitra et al. (2023)	ARIMA + RF + GBM	Multi-range pollutant forecasting dataset	94%
19	Shen et al. (2025)	ARIMA–CNN Hybrid	Industrial emission spikes dataset	97%
20	Cai et al. (2021)	ARIMA Time-Series Modeling	Hunan Province air quality dataset	92%

Table 3. Evaluation of Air Quality Prediction Models Across Various Datasets

6. CONCLUSION

This paper presents a comprehensive review of air pollution monitoring and forecasting techniques, with a particular emphasis on ARIMA-based time-series models for short-term Air Quality Index (AQI) prediction. The analysis confirms that ARIMA remains a reliable, interpretable, and computationally efficient approach for modeling environmental time-series data, especially in resource-constrained deployment scenarios such as IoT- based monitoring systems.

The review highlights a clear transition from traditional manual monitoring methods to modern data-driven frameworks incorporating machine learning, hybrid forecasting architectures, multi-sensor integration, and real- time cloud-based platforms. These advancements have significantly improved the accuracy, scalability, and accessibility of air pollution prediction systems.

In addition to the review, this study demonstrated a practical implementation of an ARIMA-based air pollution monitoring system capable of forecasting pollutant levels such as $PM_{2.5}$, PM_{10}, CO, and NO_2 in real time. The system achieved high forecasting accuracy and low error rates, confirming its effectiveness for continuous environmental monitoring and decision support.

Despite these promising results, several challenges remain, including data inconsistency, sensor noise, abrupt pollution spikes, and limited integration of heterogeneous environmental variables. Furthermore, advanced deep learning models, although accurate, require substantial computational resources, making them less suitable for low- cost real-time deployment.

Future research should focus on hybrid forecasting models, incorporation of meteorological and spatial data, improved data fusion techniques, and development of scalable real- time monitoring solutions. Addressing these challenges will contribute to building intelligent air quality prediction systems capable of supporting environmental management, public health awareness, and sustainable urban planning.

S.No.	Performance Metric	Value
1	Forecast Accuracy	94.6%
2	MAE (Mean Absolute Error)	6.2 µg/m³
3	RMSE (Root Mean Square Error)	8.9 µg/m³
4	MAPE (Mean Absolute Percentage Error)	11.4%
5	Inference Speed (GPU)	~40 FPS (24.5ms)
6	Inference Speed (CPU)	~15 FPS (66.7ms)

Table 4. Performance Metrics for Air Quality Prediction System

Table 4. presents the performance evaluation of the implemented air pollution monitoring and forecasting system. The results indicate high prediction accuracy along with low error values, demonstrating the effectiveness of the ARIMA model in forecasting pollutant concentrations. The system also exhibits efficient real-time performance, achieving fast inference speeds on both GPU and CPU platforms, making it suitable for continuous AQI monitoring and visualization applications.

7. REFERENCES

[1] H. N. Mahendra, S. Mallikarjunaswamy, D. M. Kumar, S. Kumari, and S. K. Singh, "Assessment and prediction of air quality level using ARIMA model: A case study of Surat city, Gujarat state, India," *Nature Environment and Pollution Technology*, vol. 22, no. 1, pp. 199–210, 2023.

[2] J. Luo and Y. Gong, "Air pollutant prediction based on ARIMA-WOA-LSTM model," *Atmospheric Pollution Research*, vol. 14, Art. no. 101761, 2023.

[3] R. Kaushik, E. Kelkar, A. Gupta, and A. Bhargava, "Development of an air pollution monitoring system," *Procedia Computer Science*, vol. 260, pp. 645–654, 2025.

[4] M. A. Mujtaba *et al.*, "Using machine learning for air quality prediction and sustainable urban planning," *Environmental Science and Pollution Research*, vol. 10, Art. no. 100981, 2025.

[5] A. Rahman and M. T. Khatun, "Assessing and predicting air pollution in Asia: A regional and temporal study (2018–2023)," *International Journal on Cybernetics & Informatics*, vol. 14, no. 1, pp. 27–40, 2025.

[6] D. B. K. Boon *et al.*, "Assessment of time series model for predicting long-interval consecutive missing values in air quality dataset," *Journal of Advanced Research Design*, vol. 127, no. 1, pp. 16–31, 2025.

[7] G. Mani, J. K. Viswanadhapalli, and A. A. Stonier, "Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models," *Air Quality, Atmosphere & Health*, vol. 10, no. 2A, pp. 179–194, 2021.

[8] E. Nourmohammad and Y. Rashidi, "Ground data analysis for PM2.5 prediction using predictive modeling techniques," *Journal of Air Pollution and Health*, vol. 10, no. 1, pp. 61–82, 2025.

[9] P. Gopu, R. R. Panda, and N. K. Nagwani, "Time series analysis using ARIMA model for air pollution prediction in Hyderabad City of India," in *Soft Computing and Signal Processing*, vol. 1325, Singapore: Springer, 2021, pp. 47–56.

[10] M. Solis Teran *et al.*, "Modeling air pollution using statistical and time-series approaches," *Environments*, vol. 12, Art. no. 196, 2025.

[11] I. Ganguli *et al.*, "Comprehensive analysis of air quality trends in India using machine learning and deep learning models," in *Proc. 26th Int. Conf. Distributed Computing and Networking (ICDCN)*, Hyderabad, India, Jan. 2025, pp. 1–6.

[12] J. Gaytan-Iniestra, A. R. Wojcik, and P. E. Arroyo-Lopez, "Conditions and trends of air pollution levels in the metropolitan zone in the city of Toluca," *Atmospheric Environment*, 1995.

[13] L. A. Gil-Alana and N. Carmona-González, "Particulate matter 2.5 (PM2.5): Persistence and trends in the air quality of five India cities," *Atmosphere*, vol. 16, no. 5, Art. no. 534, 2025.

[14] Z. Yan, "Comparative analysis of ARMA and ARIMA models for air quality prediction," *Advances in Engineering Technology Research*, vol. 14, pp. 1349–1356, 2025.

[15] P. Bhosale, S. Patil, and R. Kulkarni, "Air quality index prediction using ARIMA time-series model," *Sustainable Computing: Informatics and Systems*, vol. 38, Art. no. 100825, 2023.

[16] J. Duan *et al.*, "Air-quality prediction based on the ARIMA–CNN–LSTM combination model optimized by dung beetle

optimizer," *Scientific Reports*, vol. 13, no. 1, Art. no. 35627, 2023.

[17] S. Jayaram *et al.*, "Enhancing urban air quality prediction using time-based–spatial forecasting framework," *Scientific Reports*, vol. 13, no. 1, Art. no. 5419, 2023.

[18] B. Rongqing *et al.*, "Predictive modelling of air pollution affecting human tuberculosis risk on mainland China," *Scientific Reports*, vol. 13, no. 1, Art. no. 15993, 2023.

[19] N. Sutisnaingsih *et al.*, "Temporal trend analysis of light pollution using autoregressive integrated moving average (ARIMA) and random forest models," *IOP Conf. Series: Earth and Environmental Science*, vol. 1493, no. 1, Art. no. 012005, 2023.

[20] W. Cai *et al.*, "Analysis and prediction of atmospheric environmental quality based on the autoregressive integrated moving average model (**ARIMA) in Hunan Province, China," *Sustainability*, vol. 13, no. 6, Art. no. 3401, 2021.**