Asian Journal of Civil Engineering
https://doi.org/10.1007/s42107-025-01572-9

RESEARCH

Machine learning-driven prediction and interpretation of air quality index in industrial environment

R. Muralikrishnan[1] · Suneel Gollapalli[2] · Elayaraja Sellappan[3] · J. Prasanya[4] · V. Murali[5] · A. Vijayakumar[6]

Received: 16 September 2025 / Accepted: 14 October 2025

Abstract

Air pollution in South India's industrial clusters remains poorly studied, despite their growing importance as sources of pollutants in the region. This study presents an in-depth examination of air quality in Gummidipoondi, Tamil Nadu, encompassing statistical analysis, feature importance, and machine learning models. The results showed that $PM_{2.5}$ and PM_{10} are the main features that affect the Air Quality Index (AQI), accounting for more than 99% of its variations. Gaseous pollutants and weather conditions had only a lesser effect. XGBoost was the most accurate machine learning model ($R^2 = 0.9965$, RMSE = 0.0573), compared to regression and neural models. The SHAP summary plot confirmed that $PM_{2.5}$ is the main feature that had a significant impact on AQI. Temporal patterns indicated significant pollution impacts during winter and post-monsoon seasons, attributed to slow boundary layer conditions and daily industrial activities, while monsoon offered natural alleviation through dispersion. These results broaden air quality studies outside urban areas and illustrate the effectiveness of explainable machine learning for real-time surveillance and early warning systems. The study underscores the immediate necessity for focused particulate matter regulation to mitigate health hazards in industrialising corridors.

Keywords Air quality index · Machine learning · Particulate matter · Gaseous pollutants

Introduction

Air pollution is now one of the biggest problems for the environment and public health in the modern world (Manisalidis et al., 2020). The World Health Organisation reports that every year, millions of people die because of fine particulate matter, ground-level ozone, and other pollutants. Air pollution in India has reached dangerous levels in many places. In cities and industrial areas, the amount of particulate matter is sometimes far higher than what is considered safe by national and international norms. The problem has worsened because of rapid industrialisation, urban growth,

✉ R. Muralikrishnan
murali660105@gmail.com

✉ Elayaraja Sellappan
elayaraja86@gmail.com

Suneel Gollapalli
suneel.g@gmrit.edu.in

J. Prasanya
prasanyajayabal@gmail.com

V. Murali
murali_v123@yahoo.co.in

A. Vijayakumar
vijayakumarkct@gmail.com

[1] Jei Mathaajee College of Engineering, Kanchipuram, Tamil Nadu 631552, India

[2] Department of Computer Science and Engineering, GMR Institute of Technology, Rajam, Srikakulam, Andhra Pradesh 532127, India

[3] Department of Civil Engineering, PSG Institute of Technology and Applied Research, Coimbatore 641062, India

[4] Dept of Civil Engineering, Karpaga Vinayaga College of Engineering and Technology, Chengalpattu, Tamil Nadu 603308, India

[5] Department of Civil Engineering, Pandian Saraswathi Yadav Engineering College, Arasanoor, Sivagangai, Tamil Nadu 630 561, India

[6] Department of Civil Engineering, V.S.B. Engineering College, Karur, Tamil Nadu 639 111, India

Published online: 03 November 2025

and rapid transportation (Manisalidis et al., 2020). This is especially true in peri-urban industrial corridors, where environmental monitoring and regulatory enforcement are generally less strict than in big cities.

The Air Quality Index (AQI) is a common term to measure and talk about the state of the air quality. Researchers, policymakers, and the general public can better understand air quality based on AQI, which combines measurements of several pollutants into one value with defined categories (Payus et al., 2022). The Central Pollution Control Board (CPCB) in India calculates AQI by considering pollutants such as $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_3, NH_3, CO, and Pb. Fine particulate matter, especially $PM_{2.5}$, is often considered to be the most important cause of bad air quality. This is mostly because it can enter the respiratory system and cause many heart and lung problems.

Particulate matter is the most important factor in AQI calculation, and it also shows how different human-made sources affect the air quality. In India, some of the main sources of $PM_{2.5}$ and PM_{10} are industrial combustion processes, vehicle emissions, biomass burning, and construction activities (Joshi et al., 2025). Thermal power stations, steel processing units, and chemical businesses in peri-urban industrial clusters emit more particulate pollution (Zalakeviciute et al., 2020). Meteorological factors, namely Temperature, humidity, wind speed, and boundary layer dynamics, play a major role in the dispersion of the pollutants (Andújar-Maqueda et al., 2025). High winds and rain during monsoon seasons make it easier for pollutants to spread, while stagnant conditions in winter months keep contaminants close to the ground. Because of this complexity, it is hard to forecast air quality using typical statistical methods that presume linear connections (Mujtaba et al., 2025).

In recent years, the world has seen remarkable progress in technology and computer science. Among these developments, artificial intelligence (AI) has emerged as a leading field, drawing global attention. Unlike many other disciplines, AI focuses on creating machines and systems capable of demonstrating intelligence, learning independently, and making decisions (Kaveh, 2024). Machine learning is used to model the non-linear and high-dimensional relationship between contaminants and weather conditions. Machine learning determines the hidden patterns, adjusts to complicated interactions, and makes more accurate predictions than conventional regression or time-series methods (Tao et al., 2023). Ensemble-based models like Random Forest and Extreme Gradient Boosting (XGBoost) are great for estimating air quality because they can mix many decision trees and lower the problem of overfitting (Meena et al., 2024). Support Vector Regression (SVR) and artificial neural networks, such as multi-layer perceptrons, are becoming more common in this field because they can find complicated, non-linear relationships between environmental data (Sananmuang et al., 2024). Importantly, combining machine learning with explainable AI frameworks like SHAP (SHapley Additive exPlanations) not only makes predictions more accurate but also makes them easier to understand, which is important for making decisions.

Several studies have shown that machine learning can be used to predict air quality on a worldwide scale. Research in China, South Korea, and other East Asian areas has shown that tree-based ensemble models can always do better than linear models at predicting $PM_{2.5}$ levels. For example, researchers in Beijing used XGBoost on large-scale monitoring datasets and got far better at making accurate predictions than with traditional approaches (Su et al., 2023). In India, similar trends have been observed, with Random Forest and gradient boosting techniques utilised in Delhi and Mumbai to analyse complicated pollutant-meteorology relationships (de Bont et al., 2024). Recent studies have emphasised the growing air pollution challenge across South Asian cities, linking particulate matter with severe health outcomes (Pant et al., 2024a, b). Risk-based assessments further highlight the importance of integrating predictive models with public health perspectives (Pant et al., 2023). In the industrial domain Pant et al., 2024a, b applied a Naïve Bayes classifier to predict $PM_{2.5}$ concentrations in an industrial cluster, underscoring the relevance of tailored machine learning approaches in non-urban contexts. Nonetheless, the majority of current research is predominantly focused on megacities, with relatively insufficient emphasis on peri-urban industrial clusters. These places are important to regional emission profiles because they have a high concentration of factories and energy production facilities, despite being less densely populated. They not only make the air quality worse in their own area, but they also add to pollution in nearby cities.

Most of the research done in India has been on big cities like Delhi, Bengaluru, Hyderabad, Mumbai and Chennai (K et al., 2024; Ravindiran et al., 2023, 2025). Unlike earlier studies that predominantly focused on urban megacities, this research provides new insights into peri-urban industrial environments such as Gummidipoondi, Tamil Nadu. The novelty lies in the integration of machine learning models with SHAP and LIME explanations to not only achieve highly accurate AQI predictions but also to provide transparent interpretability of pollutant contributions in an underexplored industrial context. This signifies a critical research gap, as these areas are substantial sources of particulate emissions yet do not receive equivalent levels of ongoing monitoring, public awareness, or research-informed policy measures. The current study rectifies these limitations by executing an extensive investigation of AQI

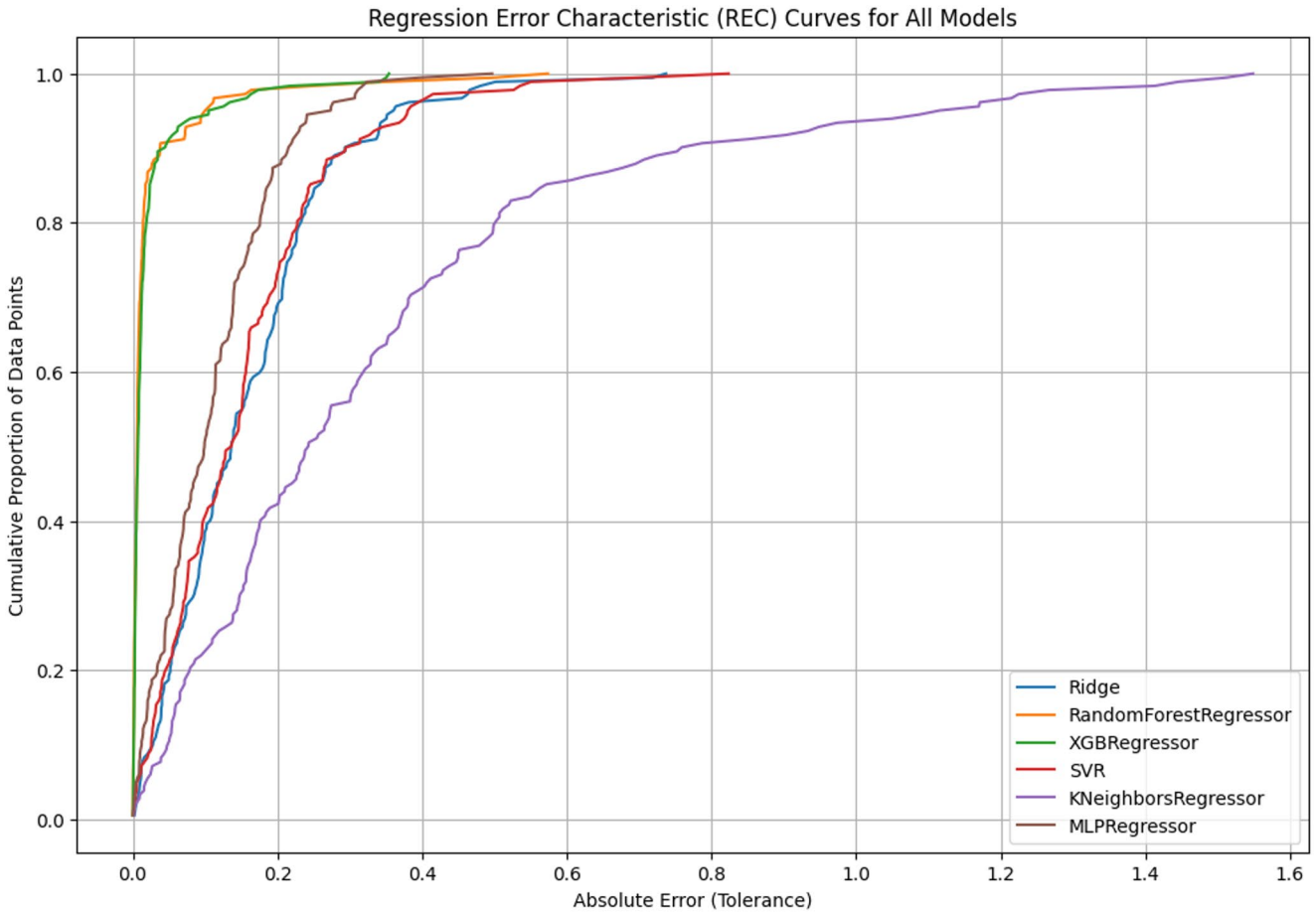

Fig. 12 Regression Error Characteristic (REC) curves for all machine learning models

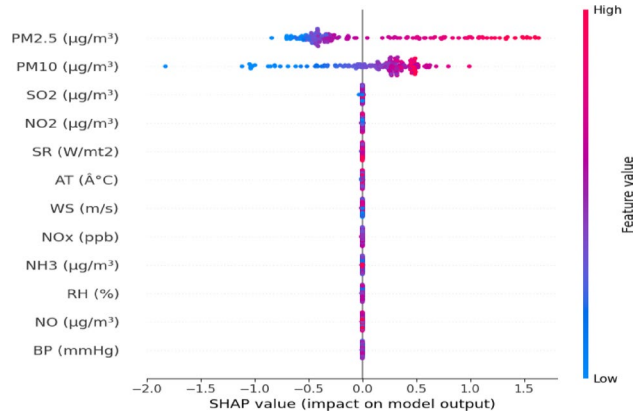

Fig. 13 SHAP summary plot AQI prediction

pollutants are present. This provides a clear and understandable reason for the model outputs, which is important for targeted interventions and clear policy communication.

Conclusion

This study shows that fine particulate matter is the main cause of poor air quality in Gummidipoondi. Controlling $PM_{2.5}$ is the best strategy to improve the AQI. Temporal assessments highlighted the concerns present during winter and post-monsoon months when air stagnation exacerbates pollutant accumulation. The use of advanced ensemble models showed that machine learning can make very accurate and clear predictions that can be used in decision-support systems. Policy action should put the strict regulation of industrial emissions, the good management of vehicle traffic along freight routes, and the building of more infrastructure for continuous monitoring at the top of the list. Integrating AI-driven prediction frameworks into regional governance can facilitate early warning systems, mitigate exposure risks, and inform long-term strategic planning. Future research should encompass multi-site industrial clusters, incorporate satellite-based monitoring, and assess health burden correlations to enhance the evidentiary base for sustainable air quality management in South India. While our findings fit well with large-city research, a limitation is

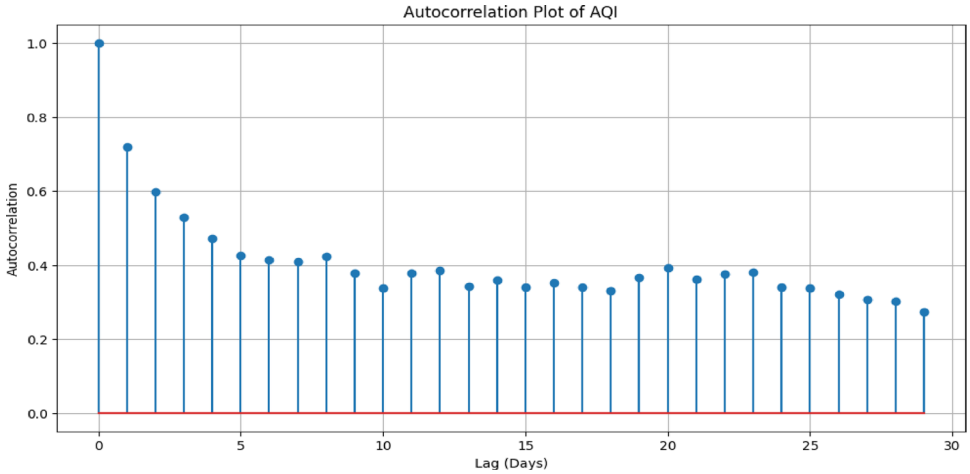

Fig. 14 Autocorrelation plot of AQI over a 30-day lag

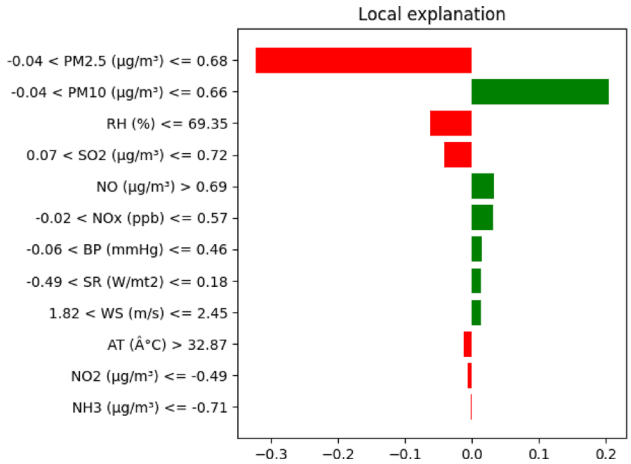

Fig. 15 LIME-based local explanation of AQI prediction

the relatively short monitoring duration and single-site context, which might not capture spatial heterogeneity seen in multi-cluster studies. These limitations suggest directions for future work, including expanding to multi-site, multi-year datasets and integrating satellite-based observations, as recommended by recent reviews.

Author contributions R. Muralikrishnan: Conceptualization, Methodology, Supervision, Writing – Original Draft, Project Administration. Suneel Gollapalli: Data Curation, Software Implementation, Formal Analysis, Visualization, Writing – Review & Editing.Elayaraja Sellappan: Methodology, Model Development, Supervision, Writing – Review & Editing.Prasanya J: Data Collection, Preprocessing, Investigation, Validation.V. Murali: Literature Review, Statistical Analysis, Resources, Data Interpretation.A. Vijayakumar: Model Evaluation, Technical Review, Writing – Final Review & Proofreading.

Funding No funding was received.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

Andújar-Maqueda, J., Ortiz-Amezcua, P., Cariñanos, P., Abril-Gago, J., De Linares, C., de Arruda Moreira, G., Bravo-Aranda, J. A., Granados-Muñoz, M. J., Alados-Arboledas, L., & Guerrero-Rascado, J. L. (2025). The role of atmospheric boundary layer wind and turbulence on surface pollen levels. *Agricultural and Forest Meteorology*, *371*, 110584. https://doi.org/10.1016/J.AGRFORMET.2025.110584

de Bont, J., Krishna, B., Stafoggia, M., Banerjee, T., Dholakia, H., Garg, A., Ingole, V., Jaganathan, S., Kloog, I., Lane, K., Mall, R. K., Mandal, S., Nori-Sarma, A., Prabhakaran, D., Rajiva, A., Tiwari, A. S., Wei, Y., Wellenius, G. A., Schwartz, J., & Ljungman, P. (2024). Ambient air pollution and daily mortality in ten cities of india: A causal modelling study. *The Lancet Planetary Health*, *8*(7), e433–e440. https://doi.org/10.1016/S2542-5196(24)00114-1

Joshi, D. C., Negi, P., Devi, S., Lohani, H., Kumar, R., Gupta, M., & Ming, L. C. (2025). Fine particulate matter (PM2.5, PM10): A silent catalyst for chronic lung diseases in India; a comprehensive review. *Environmental Challenges*, *20*, 101215. https://doi.org/10.1016/J.ENVC.2025.101215

K, K., S.K, A., R, D., & Ravindiran, G. (2024). Integrating machine learning techniques for air quality index forecasting and insights from pollutant-meteorological dynamics in sustainable urban environments. *Earth Science Informatics*, *17*(4), 3733–3748. https://doi.org/10.1007/S12145-024-01382-8/METRICS

Kalbarczyk, R., & Kalbarczyk, E. (2020). Meteorological conditions of the winter-time distribution of nitrogen oxides in Poznań: A proposal for a catalog of the pollutants variation. *Urban Climate*, *33*, 100649. https://doi.org/10.1016/J.UCLIM.2020.100649

Kaveh, A. (2024). Applications of artificial neural networks and machine learning in civil engineering. *Studies in Computational Intelligence*, *1168*, 1–474. https://doi.org/10.1007/978-3-031-66051-1/COVER

Kaveh, A., & Khavaninzadeh, N. (2024). Shear strength of cellular steel beams predicted by hybrid ANFIS-ECBO model. *Arabian Journal for Science and Engineering*, 1–22. https://doi.org/10.1007/S13369-024-09802-Z/TABLES/5

Kaveh, A., Bakhshpoori, T., & Hamze-Ziabari, S. M. (2018). M5' and Mars based prediction models for properties of Self-Compacting

concrete containing fly Ash. *Periodica Polytechnica Civil Engineering*, *62*(2), 281–294. https://doi.org/10.3311/PPCI.10799

Liu, B. H., Zhang, L. W., Wei, Y. Q., & Chen, C. (2024). Dual power transformation and Yeo–Johnson techniques for static and dynamic reliability assessments. *Buildings 2024*, *14*(11), 3625. https://doi.org/10.3390/BUILDINGS14113625. *14*.

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, *8*, 505570. https://doi.org/10.3389/FPUBH.2020.00014/BIBTEX

Meena, K. K., Bairwa, D., & Agarwal, A. (2024). A machine learning approach for unraveling the influence of air quality awareness on travel behavior. *Decision Analytics Journal*, *11*, 100459. https://doi.org/10.1016/J.DAJOUR.2024.100459

Mujtaba, M. A., Munir, M. A., Ali, S., Petrů, J., Ansar, T., Akhlaq, W., Ahmad, M., Iqbal, H., Ali, F., Bashir, M. N., & Alexander, T. (2025). Using machine learning for air quality prediction and sustainable urban planning. *Sustainable Futures*, *10*, 100981. https://doi.org/10.1016/J.SFTR.2025.100981

Pant, A., Sharma, S., & Pant, K. (2023). Evaluation of machine learning algorithms for air quality index (AQI) prediction. *Journal of Reliability and Statistical Studies*, *16*(2), 229–242. https://doi.org/10.13052/JRSS0974-8024.1621

Pant, A., Joshi, R. C., Sharma, S., & Pant, K. (2024a). Air quality and public health risk assessment: A case of an industrial area in Haridwar City, Uttarakhand (India). *Indian Journal of Public Health*, *68*(2), 222–226. https://doi.org/10.4103/IJPH.IJPH_279_23

Pant, A., Pant, K., Pathak, N., & Ram, M. (2024b). Prediction of particulate matter (PM2.5) for industrial area based on Naive Bayes classifier. *Lecture Notes in Networks and Systems*, *786*, 189–195. https://doi.org/10.1007/978-981-99-6547-2_15

Payus, C. M., Syazni, N., M. S., & Sentian, J. (2022). Extended air pollution index (API) as tool of sustainable indicator in the air quality assessment: El-Nino events with climate change driven. *Heliyon*, *8*(3), e09157. https://doi.org/10.1016/J.HELIYON.2022.E09157

Raiaan, M. A. K., Sakib, S., Fahad, N. M., Mamun, A., Al, Rahman, M. A., Shatabda, S., & Mukta, M. S. H. (2024). A systematic review of hyperparameter optimization techniques in convolutional neural networks. *Decision Analytics Journal*, *11*, 100470. https://doi.org/10.1016/J.DAJOUR.2024.100470

Ravindiran, G., Rajamanickam, S., Kanagarathinam, K., Hayder, G., Janardhan, G., Arunkumar, P., Arunachalam, S., AlObaid, A. A., Warad, I., & Muniasamy, S. K. (2023). Impact of air pollutants on climate change and prediction of air quality index using machine learning models. *Environmental Research*, *239*, 117354. https://doi.org/10.1016/J.ENVRES.2023.117354

Ravindiran, G., Karthick, K., Rajamanickam, S., Datta, D., Das, B., Shyamala, G., Hayder, G., & Maria, A. (2025). Ensemble stacking of machine learning models for air quality prediction for Hyderabad City in India. *IScience*, *28*(2), 111894. https://doi.org/10.1016/j.isci.2025.111894

Sananmuang, T., Mankong, K., & Chokeshaiusaha, K. (2024). Multilayer perceptron and support vector regression models for feline parturition date prediction. *Heliyon*, *10*(6), e27992. https://doi.org/10.1016/J.HELIYON.2024.E27992

Sasmita, S., Kumar, D. B., & Priyadharshini, B. (2022). Assessment of sources and health impacts of PM10 in an urban environment over Eastern coastal plain of India. *Environmental Challenges*, *7*, 100457. https://doi.org/10.1016/J.ENVC.2022.100457

Singh, V., Singh, S., Biswal, A., Kesarkar, A. P., Mor, S., & Ravindra, K. (2020). Diurnal and Temporal changes in air pollution during COVID-19 strict lockdown over different regions of India. *Environmental Pollution*, *266*, 115368. https://doi.org/10.1016/J.ENVPOL.2020.115368

Su, M., Liu, H., Yu, C., & Duan, Z. (2023). A novel AQI forecasting method based on fusing Temporal correlation forecasting with Spatial correlation forecasting. *Atmospheric Pollution Research*, *14*(4), 101717. https://doi.org/10.1016/J.APR.2023.101717

Tao, H., Jawad, A. H., Shather, A. H., Al-Khafaji, Z., Rashid, T. A., Ali, M., Al-Ansari, N., Marhoon, H. A., Shahid, S., & Yaseen, Z. M. (2023). Machine learning algorithms for high-resolution prediction of Spatiotemporal distribution of air pollution from meteorological and soil parameters. *Environment International*, *175*, 107931. https://doi.org/10.1016/J.ENVINT.2023.107931

Yadav, R., Sahu, L. K., Beig, G., Tripathi, N., Maji, S., & Jaaffrey, S. N. A. (2019). The role of local meteorology on ambient particulate and gaseous species at an urban site of Western India. *Urban Climate*, *28*, 100449. https://doi.org/10.1016/J.UCLIM.2019.01.003

Zalakeviciute, R., Rybarczyk, Y., Granda-Albuja, M. G., Diaz Suarez, M. V., & Alexandrino, K. (2020). Chemical characterization of urban PM10 in the tropical Andes. *Atmospheric Pollution Research*, *11*(2), 343–356. https://doi.org/10.1016/J.APR.2019.11.007