Predicting $PM_{2.5}$ Concentration in Under-Monitored Urban Areas Using Satellite-Derived Predictors and Machine Learning

Vincent Nyamollo[1], **Oyedele Adewole**[1], **Yueqing Li**[1]

[1]Department of Industrial and Systems Engineering, Lamar University, Beaumont, TX, USA
vnyamollo@lamar.edu, oadewole1@lamar.edu, yli6@lamar.edu

Abstract

Sub-Saharan Africa (SSA) experiences elevated levels of fine particulate matter ($PM_{2.5}$), often exceeding World Health Organization guidelines, yet monitoring capacity remains severely limited due to sparse ground monitoring networks. While integrating satellite remote sensing with machine learning offers potential solutions to these data gaps, its feasibility in the complex urban environments across SSA remains underexplored. In this study, we evaluated various machine learning approaches to estimate $PM_{2.5}$ concentrations in Nairobi, Kenya, using satellite-derived atmospheric and meteorological predictors. Model choice strongly influenced prediction accuracy and reliability. Support Vector Regressor (SVR) delivered the best overall performance, combining high accuracy with minimal bias ($R^2 = 0.913$, RMSE $= 5.059$, MAE $= 2.987$, MBE ≈ 0). Extreme Gradient Boosting (XGBR) and Random Forest Regressor (RFR) also achieved strong results but showed systematic biases in predicting high $PM_{2.5}$ concentration levels. These results demonstrate that integrating satellite data with machine learning could enhance air quality monitoring in under-monitored urban environments.

Introduction

Rapid economic growth and urbanization across Sub-Saharan Africa (SSA) have intensified anthropogenic activities that not only contribute to regional climate challenges but also degrade air quality (Fisher et al. 2021). Most urban environments in this region experience elevated concentrations of fine particulate matter ($PM_{2.5}$, particles with an aerodynamic diameter $\leq 2.5\ \mu m$ (US EPA 2025)) with daily average levels varying widely between 25–200 $\mu g/m^3$ (Nyaga et al. 2025; Okure et al. 2022). Globally, nine out of ten people are exposed to polluted air, with ambient $PM_{2.5}$ linked to more than 7 million deaths annually (World Health Organization 2018) and an estimated 231.51 million disability-adjusted life years in 2021 (Fang et al. 2025).

Although substantial advances in air quality monitoring have been realized across North America, Europe and parts of Asia, SSA remains critically underrepresented in the literature, despite experiencing some of the fastest rates of urbanization and heightened vulnerability to climate change. This is attributed to limited data availability stemming from sparse and unevenly distributed ground-based monitoring networks (Shaddick et al. 2018). Emerging advances in machine learning and remote sensing hold significant potential to address limitations in air quality monitoring in regions with sparse ground-based observations (Zhou et al. 2024). However, their feasibility and effectiveness in SSA cities with diverse pollution sources and high $PM_{2.5}$ variability remain largely underexplored.

In this study, we assess the use of satellite–machine learning approaches to estimate $PM_{2.5}$ concentrations in Nairobi, Kenya, drawing on satellite-derived atmospheric and meteorological variables. Our findings demonstrate the feasibility of integrating satellite data with machine learning to strengthen air quality monitoring in data-sparse regions. By advancing monitoring capacity in SSA cities, this work could help identify at-risk areas and support evidence-based public health policies and interventions.

Related Work

Machine learning techniques such as ensemble models (i.e., Random Forest and Gradient Boosting) and Support Vector Regression (SVR) have demonstrated robust performance in estimating ground $PM_{2.5}$ concentration from satellite-derived environmental predictors (Makhdoomi, Sarkhosh, and Ziaei 2025; Xiao et al. 2018). For instance, Zhang et al. (2021) showed that Random Forest (RF) models trained on satellite-derived Aerosol Optical Depth (AOD), meteorology, land use and socioeconomic data effectively estimated $PM_{2.5}$, achieving an R-squared of 0.8 and Root Mean Squared Error (RMSE) of 9.40 μg/m^3. Similarly, Amiri and Shahne (2025) applied RF and SVM to estimate $PM_{2.5}$ concentrations in Tehran and found that incorporating meteorological parameters significantly improved prediction accuracy, with RF outperforming other models with accuracies of 94 – 98%.

Extreme gradient boosting regressor (XGBR) was also reported to achieve excellent $PM_{2.5}$ predictions in the northern US states with only a smaller number of predictors and RMSE of 3.11 μg/m^3 (Just et al. 2020). In another study focusing on China, Liu et al. (2021) demonstrated that XGBR combined with data calibration techniques outperformed SVR and Ridge Regression in predicting $PM_{2.5}$ in urban areas.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Dataset Description

The study focuses on Nairobi, a metropolitan area of 692 km^2 with a population exceeding 5 million in 2023. The city exemplifies the dual challenge of heightening air pollution and sparse ground monitoring infrastructure. The selection of the satellite-derived predictors was based on their established link to particulate matter. Trace gases such as nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO) influence the secondary formation of aerosols, including those measured by Aerosol Optical Depth (AOD), which is commonly used as a proxy for $PM_{2.5}$ (Shan et al. 2024; Liu et al. 2021; Zhang et al. 2021). Meteorological variables like temperature, precipitation and wind speed were included to account for the transformation, dispersion and removal of particulate matter (Tian et al. 2021).

The dataset covered a five-year period (2020 to 2024). Daily average values for a total of eleven satellite-derived predictors were extracted via Google Earth Engine from different satellite providers. Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), Carbon Monoxide (CO), and Ozone (O_3) were obtained from COPERNICUS Sentinel-5P TROPOMI. AOD was obtained from MODIS and processed using MAIAC algorithm. Environmental variables, including air temperature, land surface temperature, dew point temperature, eastward and northward wind components and precipitation, were all obtained from ECMWF ERA reanalysis.

The target daily ground-observed $PM_{2.5}$ concentration levels from 2020 to 2024 were downloaded from OpenAfrica portal, a public data repository that stores air quality data obtained from low-cost ground-based sensors deployed across Africa. The dataset consisted of hourly air quality variables such as $PM_{2.5}$, PM_{10}, humidity, and temperature with inconsistent reporting frequency. As such, we derived daily averages of $PM_{2.5}$ from the reporting sensors to address the inconsistency in measurement frequency.

Data preprocessing involved dropping rows with missing values in the target column, dropping columns with more than 40% missing values and imputing remaining missing values via time-based interpolation. Rows with $PM_{2.5}$ values exceeding three times the inter-quartile range were treated as outliers and removed. This resulted in seven initial features and one target, whose descriptive statistics are illustrated in Table 1. Advanced feature engineering involved deriving wind speed based on the wind components and introducing smoothing features to capture season variability, local temporal dynamics and correct skewness. The final processed dataset consisted of 12 features, one target variable and a total of 1573 samples.

Methodology

We define the problem as a supervised regression task, where the objective is to learn a mapping $f : X \rightarrow y$ from a set of satellite-derived predictors X to a continuous target ground $PM_{2.5}$ measurements y. Given the training data $\{(X_i, y_i)\}_{i=1}^{N}$, the goal is to minimize a loss function $L(f(X_i), y_i)$, such as the mean squared error and root mean squared error, to enable accurate prediction of y on unseen X. We implemented Random Forest Regressor (RFR), Extreme Gradient Boosting Regressor (XGBR) (Just et al. 2020; Murillo-Escobar et al. 2019) and Support Vector Regressor (SVR) (Stafoggia et al. 2017) for the regression task. We also used Ridge Regression as a regularized linear baseline to evaluate the performance of the complex models (Samad et al. 2023). We split the dataset into 80% training (1258 samples) and 20% testing (315 samples). Hyperparameter optimization was based on 5-fold cross-validation on the training data. Performance was assessed using four evaluation metrics, namely coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE) and mean bias error (MBE) to provide a balance between interpretability, accuracy and prediction bias.

Results and Discussion

In Nairobi, $PM_{2.5}$ concentrations range from 9.75 to 355.03 $\mu g/m^3$, with a mean of 29.72 $\mu g/m^3$. Levels in some areas exceeded three times the World Health Organization's recommended average (Nyaga et al. 2025). Table 2 summarizes the performance of the models. Time series plots (Figure 1) showed that $PM_{2.5}$ levels are highest during the cool-dry months of July–September and lowest during the rainy season (March–May), consistent with the findings of Nyaga et al. (2025).

SVR achieved the best balance between accuracy and bias, with the highest R^2 of 0.913 and the smallest RMSE of 5.06 $\mu g/m^3$, MAE of 2.99 $\mu g/m^3$ and near-zero overall MBE (-0.097 $\mu g/m^3$ µg/m³), tracking the overall daily (Figure 1) and monthly (Figure 2) trends closely. XGBR also performed well (R^2 = 0.906, RMSE = 5.26 $\mu g/m^3$, MBE $\approx$ 0) compared to RFR and Ridge Regression, which tended to overpredict at high concentrations. Residual analysis was conducted to validate the reliability of these models in predicting $PM_{2.5}$ at low, medium and high concentrations (Figure 3). SVR minimized systematic error well, while XGBR showed slight underprediction and overprediction at high and low $PM_{2.5}$ concentrations, respectively. Random Forest demonstrated a strong medium-range performance but exhibited inconsistent bias at extremes. Ridge Regression showed the weakest overall performance, with the lowest R-squared, larger MAE and modest underprediction at higher concentrations. Scatter plots of the observed vs predicted $PM_{2.5}$ depicted the same trend where kernel and tree-based models aligned tightly with the 1:1 line, while Ridge Regression aligned moderately, signifying its inability to capture nonlinearities (Figure 4). Generally, all feature had substantial contribution to the prediction performance. Residual analysis highlighted that while global metrics such as RMSE and MAE provide an overall assessment of model performance, examining systematic trends and bias in model performance across low, medium and high concentrations is crucial, particularly in air quality prediction, where accurate estimates at extreme pollution levels are essential.

Conclusion and Future Work

We evaluated the performance of four machine learning models in predicting $PM_{2.5}$ concentrations in Nairobi us-

Table 1: Descriptive statistics of satellite-derived $PM_{2.5}$ pollutant predictors and target ground measurements between 2020 to 2024

Variable	Ranges	Mean	Std Dev	Skewness	Kurtosis
AOD	0.048 – 1.585	0.420	0.127	1.024	8.923
CO (μg/m^3)	16.122 – 45.256	26.867	3.384	0.768	2.383
NO_2 (μg/m^3)	0 – 73.7	21.980	9.940	0.711	1.087
Ozone (μg/m^3)	104.488 – 131.14	117.505	4.594	-0.047	-0.450
Air temperature (°C)	14.54 – 22.76	18.855	1.401	-0.114	-0.317
Precipitation (mm/m^2)	0 – 42.938	1.732	3.470	4.163	24.716
Wind speed (m/s)	0.056 – 4.337	1.935	0.770	-0.044	-0.525
$PM_{2.5}$	8.15 – 355.03	29.633	34.955	5.707	37.635

Table 2: Comparison of the performance of the four models on training and testing datasets across the four metrics

Model	Train R^2	Train RMSE	Train MAE	Train MBE	Test R^2	Test RMSE	Test MAE	Test MBE
RR	0.962	7.486	3.625	0.000	0.886	5.793	3.550	0.312
SVR	0.982	5.190	2.571	-0.088	0.913	5.059	2.987	-0.097
RFR	0.994	2.993	1.612	0.004	0.882	5.903	3.355	0.476
XGBR	0.986	4.558	2.655	-0.019	0.906	5.262	3.005	-0.047

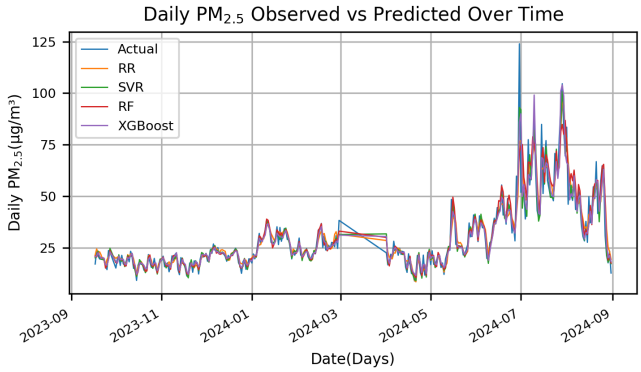

Figure 1: Predicted vs. observed monthly average concentration levels. XGBR and SVR closely captured the overall trend, while RR overpredicted at both low and peak concentrations. RFR performed well in predicting the midrange concentrations.

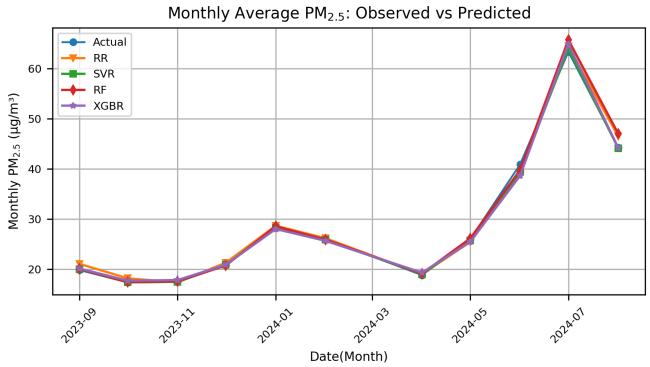

Figure 2: Time series comparison showing daily predicted vs. observed PM2.5 concentrations.XGBR and SVR performed well in capturing the trend, while RR over-predicted the daily levels at low and peak concentrations.

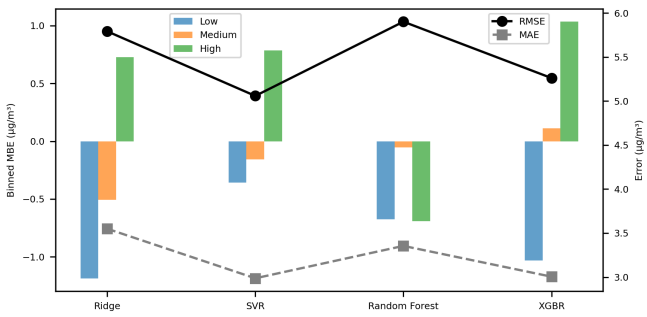

Figure 3: Binned residual analysis evaluating both bias and overall predictive accuracy across low, medium and high pollution levels. The left axis shows mean bias error (MBE, μg/m^3) in predicting low, medium and high $PM_{2.5}$ concentrations, where negative values indicate underprediction and positive values indicate overprediction. The right axis reports the prediction errors (RMSE and MAE, μg/m^3).

ing satellite-derived variables. SVR achieved the best overall balance of robustness and generalizability with near-negligible bias, followed by XGBR, which was biased at extremes despite the high accuracy score. RFR was particularly stable for mid-range $PM_{2.5}$ levels but less reliable at the extremes. Ridge Regression consistently underperformed due to its simplicity and limited ability to model nonlinear relationships. Residual analysis revealed that most models tended to underpredict at low $PM_{2.5}$ concentrations and overpredict at high concentrations, with RFR and Ridge Regression showing less consistency at the extremes. These findings demonstrate the feasibility of ensemble boosting and kernel methods to capture complex interactions among meteorological variables, trace gases and particulate matter, underscoring their potential to enhance air quality monitoring in urban environments with limited ground-based observation networks.

For future work, we aim to improve model robustness to capture hourly and neighborhood-level variability by employing hybrid deep-learning and ensembled approaches and expanding the dataset to incorporate local predictors like traffic data, land use and population activity. Furthermore, applying transfer learning could help test the generalizability in multiple cities across Sub-Saharan Africa and provide actionable insights for policymakers. This could help support scalable air quality monitoring frameworks across the region.

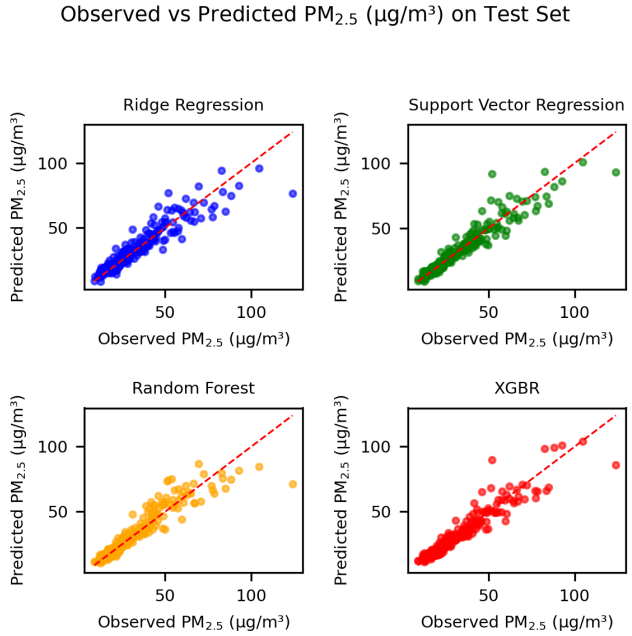

Figure 4: Scatter plots of 5-fold cross-validation of the daily $PM_{2.5}$. SVR and tree-based models aligned well with the 1:1 line, indicating satisfactory performance compared to baseline.

References

Amiri, Z.; and Shahne, M. Z. 2025. Modeling PM2. 5 concentration in tehran using satellite-based Aerosol optical depth (AOD) and machine learning: Assessing input contributions and prediction accuracy. *Remote Sensing Applications: Society and Environment*, 38: 101549.

Fang, T.; Di, Y.; Xu, Y.; Shen, N.; Fan, H.; Hou, S.; and Li, X. 2025. Temporal trends of particulate matter pollution and its health burden, 1990–2021, with projections to 2036: a systematic analysis for the global burden of disease study 2021. *Frontiers in Public Health*, 13: 1579716.

Fisher, S.; Bellinger, D. C.; Cropper, M. L.; Kumar, P.; Binagwaho, A.; Koudenoukpo, J. B.; Park, Y.; Taghian, G.; and Landrigan, P. J. 2021. Air pollution and development in Africa: impacts on health, the economy, and human capital. *The Lancet Planetary Health*, 5(10): e681–e688.

Just, A. C.; Arfer, K. B.; Rush, J.; Dorman, M.; Shtein, A.; Lyapustin, A.; and Kloog, I. 2020. Advancing methodologies for applying machine learning and evaluating spatiotemporal models of fine particulate matter (PM2.5) using satellite data over large regions. *Atmospheric Environment*, 239: 117649.

Liu, B.; Tan, X.; Jin, Y.; Yu, W.; and Li, C. 2021. Application of RR-XGBoost combined model in data calibration of micro air quality detector. *Scientific Reports*, 11(1): 15662.

Makhdoomi, A.; Sarkhosh, M.; and Ziaei, S. 2025. PM2. 5 concentration prediction using machine learning algorithms: an approach to virtual monitoring stations. *Scientific Reports*, 15(1): 8076.

Murillo-Escobar, J.; Sepulveda-Suescun, J.; Correa, M.; and Orrego-Metaute, D. 2019. Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: Case study in Aburrá Valley, Colombia. *Urban climate*, 29: 100473.

Nyaga, E. W.; Giordano, M. R.; Beekmann, M.; Westervelt, D. M.; Gatari, M.; Mungai, J.; Opinde, G.; Presto, A. A.; Tjernström, E.; McNeill, V. F.; et al. 2025. Seasonal

multisite low-cost sensor measurements to estimate spatial and temporal variability of particulate matter pollution in Nairobi, Kenya. *Atmospheric Pollution Research*, 16(10): 102630.

Okure, D.; Ssematimba, J.; Sserunjogi, R.; Gracia, N. L.; Soppelsa, M. E.; and Bainomugisha, E. 2022. Characterization of ambient air quality in selected urban areas in Uganda using low-cost sensing and measurement technologies. *Environmental Science & Technology*, 56(6): 3324–3339.

Samad, A.; Garuda, S.; Vogt, U.; and Yang, B. 2023. Air pollution prediction using machine learning techniques–An approach to replace existing monitoring stations with virtual monitoring stations. *Atmospheric Environment*, 310: 119987.

Shaddick, G.; Thomas, M. L.; Amini, H.; Broday, D.; Cohen, A.; Frostad, J.; Green, A.; Gumy, S.; Liu, Y.; Martin, R. V.; et al. 2018. Data integration for the assessment of population exposure to ambient air pollution for global burden of disease assessment. *Environmental science & technology*, 52(16): 9069–9078.

Shan, Y.; Zhu, Y.; Qi, Y.; Yang, Y.; Mu, J.; Liu, M.; Li, H.; Zhang, J.; Nie, Y.; Liu, Y.; Zhao, M.; Zhang, X.; Zhang, L.; Wang, Y.; Li, H.; Shen, H.; Zhang, Y.; Wang, X.; Huang, L.; Wang, W.; and Xue, L. 2024. Insights into atmospheric trace gases, aerosols, and transport processes at a high-altitude station (2623 m a.s.l.) in Northeast Asia. *Atmospheric Environment*, 326: 120482.

Stafoggia, M.; Schwartz, J.; Badaloni, C.; Bellander, T.; Alessandrini, E.; Cattani, G.; de' Donato, F.; Gaeta, A.; Leone, G.; Lyapustin, A.; Sorek-Hamer, M.; de Hoogh, K.; Di, Q.; Forastiere, F.; and Kloog, I. 2017. Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. *Environment International*, 99: 234–244.

Tian, X.; Cui, K.; Sheu, H.; Hsieh, Y.; and Yu, F. 2021. Effects of rain and snow on the air quality index, PM2. 5 levels, and dry deposition flux of PCDD/Fs. Aerosol Air Qual. Res. 21, 210158.

US EPA. 2025. Particulate Matter (PM) Basics. https://www.epa.gov/pm-pollution/particulate-matter-pm-basics. Accessed: 2025-08-14.

World Health Organization. 2018. 9 out of 10 people worldwide breathe polluted air, but more countries are taking action. https://www.who.int/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action. Accessed: 2025-08-14.

Xiao, Q.; Chang, H. H.; Geng, G.; and Liu, Y. 2018. An ensemble machine-learning model to predict historical PM2. 5 concentrations in China from satellite data. *Environmental science & technology*, 52(22): 13260–13269.

Zhang, D.; Du, L.; Wang, W.; Zhu, Q.; Bi, J.; Scovronick, N.; Naidoo, M.; Garland, R. M.; and Liu, Y. 2021. A machine learning model to estimate ambient PM2. 5 concentrations in industrialized highveld region of South Africa. *Remote sensing of environment*, 266: 112713.

Zhou, S.; Wang, W.; Zhu, L.; Qiao, Q.; and Kang, Y. 2024. Deep-learning architecture for PM2. 5 concentration prediction: A review. *Environmental Science and Ecotechnology*, 21: 100400.