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

Anonymous Author(s)
Affiliation
Address
email

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.

1 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 [1]. Most urban environments in this region experience elevated concentrations of fine particulate matter ($PM_{2.5}$, particles with an aerodynamic diameter $\leq 2.5\ \mu m$ [2]) with daily average levels varying widely between 25–200 $\mu g/m^3$ [3][4]. Globally, nine out of ten people are exposed to polluted air, with ambient $PM_{2.5}$ linked to more than 7 million deaths annually [5] and an estimated 231.51 million disability-adjusted life years in 2021 [6].

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 [7]. 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 [8]. However, their feasibility and effectiveness in SSA cities with diverse pollution sources and high $PM_{2.5}$ variability remain largely underexplored.

Submitted to Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

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.

2 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 [9][10]. For instance, Zhang et al.[11] 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 [12] 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 [13]. In another study focusing on China, Liu et al. [14] demonstrated that XGBR combined with data calibration techniques outperformed SVR and Ridge Regression in predicting $PM_{2.5}$ in urban areas.

3 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 $PM_{2.5}$ [15][14][11]. Meteorological variables like temperature, precipitation and wind speed were included to account for the transformation, dispersion and removal of particulate matter[16].

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 2 (see Appendix A). 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. Figure 5 (appendix E) shows the correlation heatmap between the features and the target variable. A detailed description of the data preprocessing is presented in Table 4 (see Appendix F).

4 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) [13][17] and Support Vector Regressor (SVR) [18] for the regression task. We also used Ridge Regression as a regularized linear baseline to evaluate the performance of the complex models [19]. 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. Optimal hyperparameter values and their ranges are listed in Table 3 (see Appendix C). 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 (see Appendix B).

5 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 [3]. Table 1 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. [3].

Table 1: 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

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(a)) and monthly (Figure 1(b)) 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 2). 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 3). Generally, all feature had substantial contribution to the prediction performance (Figure 4). 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.

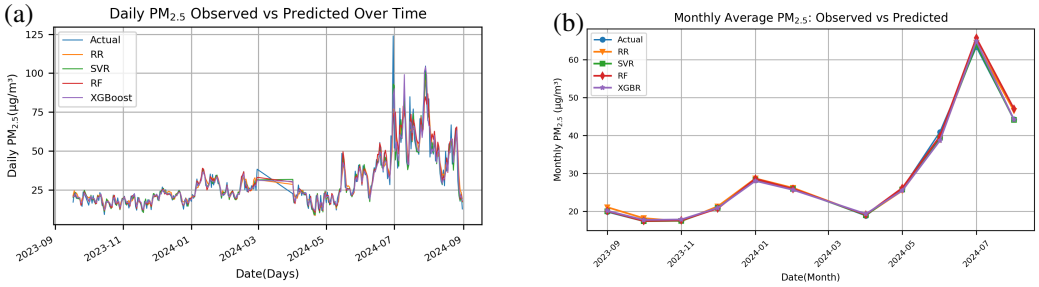

Figure 1: Predicted vs. observed PM2.5 concentrations over time. (a) time series comparison showing daily variability. (b) Monthly average concentration levels. XGBR and SVR closely captured the overall trend, while RR overpredicted at both low and peak concentrations. RFR performed well in predicting the midrange concentrations.

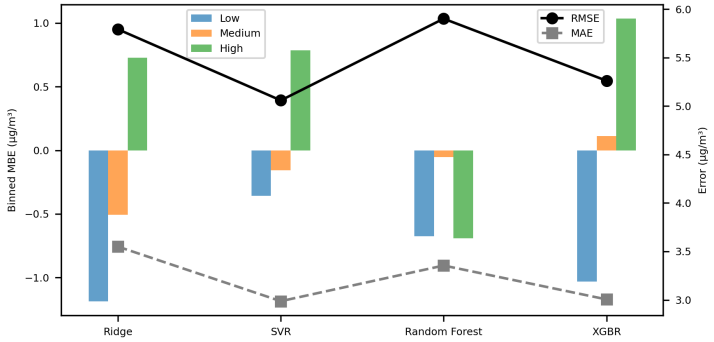

Figure 2: Binned residual analysis evaluating both bias and overall predictive accuracy across low, medium and high pollution levels. The left axis shows mean bias error (MBE, μg/m^3) in predicting low, medium and high $PM_{2.5}$ concentrations, where negative values indicate underprediction and positive values indicate overprediction. The right axis reports the prediction errors (RMSE and MAE, μg/m^3).

6 Conclusion and Future Work

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

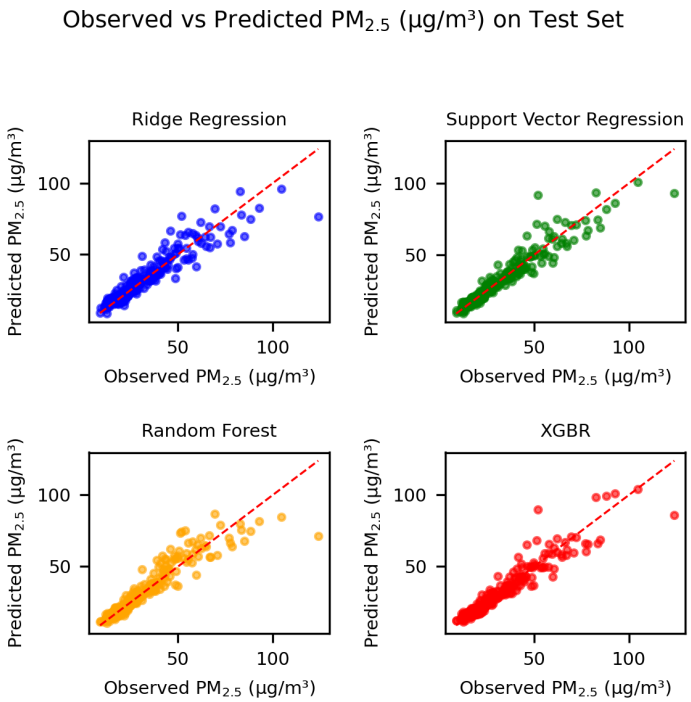

Figure 3: Scatter plots of 5-fold cross-validation of the daily $PM_{2.5}$. SVR and tree-based models aligned well with the 1:1 line, indicating satisfactory performance compared to baseline.

References

[1] Samantha Fisher, David C Bellinger, Maureen L Cropper, Pushpam Kumar, Agnes Binagwaho, Juliette Biao Koudenoukpo, Yongjoon Park, Gabriella Taghian, and Philip J Landrigan. Air pollution and development in africa: impacts on health, the economy, and human capital. *The Lancet Planetary Health*, 5(10):e681–e688, 2021.

[2] US EPA. Particulate matter (pm) basics. `https://www.epa.gov/pm-pollution/particulate-matter-pm-basics`, 2025. Accessed: 2025-08-14.

[3] Ezekiel W Nyaga, Michael R Giordano, Matthias Beekmann, Daniel M Westervelt, Michael Gatari, John Mungai, Godwin Opinde, Albert A Presto, Emilia Tjernström, V Faye McNeill, et al. 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, 2025.

[4] Deo Okure, Joel Ssematimba, Richard Sserunjogi, Nancy Lozano Gracia, Maria Edisa Soppelsa, and Engineer Bainomugisha. 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, 2022.

[5] World Health Organization. 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-` 2018. Accessed: 2025-08-14.

[6] Tao Fang, Yanbo Di, Yang Xu, Na Shen, Haojun Fan, Shike Hou, and Xiaoxue Li. 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, 2025.

[7] Gavin Shaddick, Matthew L Thomas, Heresh Amini, David Broday, Aaron Cohen, Joseph Frostad, Amelia Green, Sophie Gumy, Yang Liu, Randall V Martin, et al. 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, 2018.

[8] Shiyun Zhou, Wei Wang, Long Zhu, Qi Qiao, and Yulin Kang. Deep-learning architecture for pm2. 5 concentration prediction: A review. *Environmental Science and Ecotechnology*, 21:100400, 2024.

[9] Ahmad Makhdoomi, Maryam Sarkhosh, and Somayyeh Ziaei. Pm2. 5 concentration prediction using machine learning algorithms: an approach to virtual monitoring stations. *Scientific Reports*, 15(1):8076, 2025.

[10] Qingyang Xiao, Howard H Chang, Guannan Geng, and Yang Liu. An ensemble machine-learning model to predict historical pm2. 5 concentrations in china from satellite data. *Environmental science & technology*, 52(22):13260–13269, 2018.

[11] Danlu Zhang, Linlin Du, Wenhao Wang, Qingyang Zhu, Jianzhao Bi, Noah Scovronick, Mogesh Naidoo, Rebecca M Garland, and Yang Liu. A machine learning model to estimate ambient pm2. 5 concentrations in industrialized highveld region of south africa. *Remote sensing of environment*, 266:112713, 2021.

[12] Zahra Amiri and Maryam Zare Shahne. 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, 2025.

[13] Allan C. Just, Kodi B. Arfer, Johnathan Rush, Michael Dorman, Alexandra Shtein, Alexei Lyapustin, and Itai Kloog. 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, 2020.

[14] Bing Liu, Xianghua Tan, Yueqiang Jin, Wangwang Yu, and Chaoyang Li. Application of rr-xgboost combined model in data calibration of micro air quality detector. *Scientific Reports*, 11(1):15662, 2021.

[15] Ye Shan, Yujiao Zhu, Yanbi Qi, Yu Yang, Jiangshan Mu, Mingxuan Liu, Hongyong Li, Ji Zhang, Yanqiu Nie, Yuhong Liu, Min Zhao, Xin Zhang, Lingli Zhang, Yufei Wang, Hong Li, Hengqing Shen, Yuqiang Zhang, Xinfeng Wang, Liubin Huang, Wenxing Wang, and Likun Xue. 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, 2024.

[16] X Tian, K Cui, HL Sheu, YK Hsieh, and F Yu. 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, 2021.

[17] J Murillo-Escobar, JP Sepulveda-Suescun, MA Correa, and D Orrego-Metaute. Forecasting concentrations of air pollutants using support vector regression improved with particle swarm optimization: Case study in aburrá valley, colombia. *Urban climate*, 29:100473, 2019.

[18] Massimo Stafoggia, Joel Schwartz, Chiara Badaloni, Tom Bellander, Ester Alessandrini, Giorgio Cattani, Francesca de' Donato, Alessandra Gaeta, Gianluca Leone, Alexei Lyapustin, Meytar Sorek-Hamer, Kees de Hoogh, Qian Di, Francesco Forastiere, and Itai Kloog. Estimation of daily pm10 concentrations in italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. *Environment International*, 99:234–244, 2017.

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

Appendix

A Descriptive Statistics

Table 2: 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 ($^\circ$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

B Metrics

The following metrics were used to evaluate model performance: Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Bias Error (MBE). Here, y_i represents the observed value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the mean of observed values, and n is the total number of observations.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{4}$$

C Hyperparameters

Table 3: Hyperparameter values considered during model tuning for Ridge Regression, SVR, Random Forest, and XGBR. The optimal values selected through cross-validation are indicated in **bold**.

Algorithm	Hyperparameters
RR	$\alpha = [\mathbf{100}, 2.78 \times 10^2, 7.74 \times 10^2, 2.15 \times 10^3, 5.99 \times 10^3,$ $1.67 \times 10^4, 4.64 \times 10^4, 1.29 \times 10^5, 3.59 \times 10^5, 1 \times 10^6]$
SVR	Kernel: [**'linear'**, 'rbf'], C = [1, 10, **50**, 100], Epsilon: [0.1, **0.5**, 1, 2], gamma = [**'scale'**, 'auto', 0.01, 0.1, 1]
RFR	Estimators = [100, 200, 300, 400, **600**, 800], max_depth = [5, **10**,15], max_features = [**'sqrt'**, 0.3, 0.5, 0.7]
XGBR	Estimators = [100, 200, 300, 400, **600**, 800], max_depth = [1, **2**, 5, 10], learning_rate = [0.001, **0.01**, 0.1], subsample = [0.4, **0.6**, 1.0]

D Feature Importance

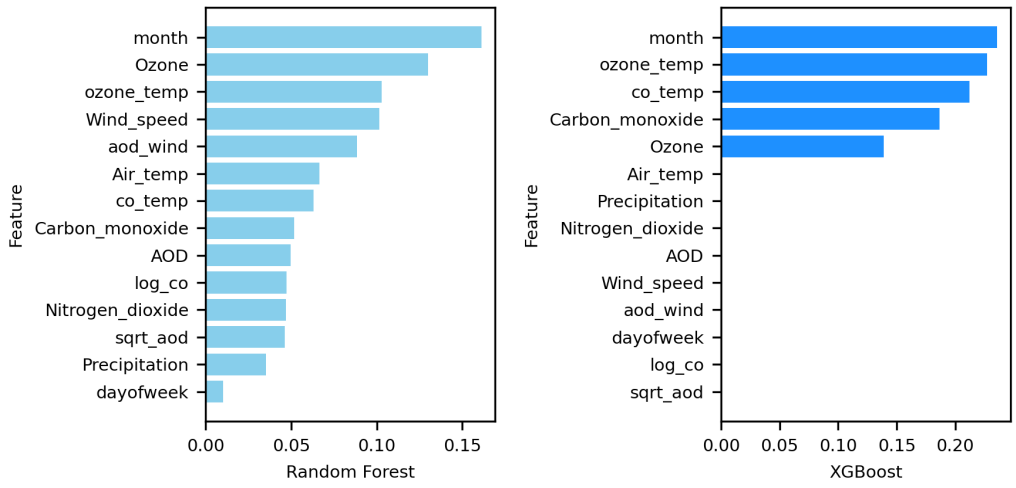

Figure 4: Feature importance for RFR and XGBR. All the features relatively have strong influence on the performance of the models

E Correlation Heatmap

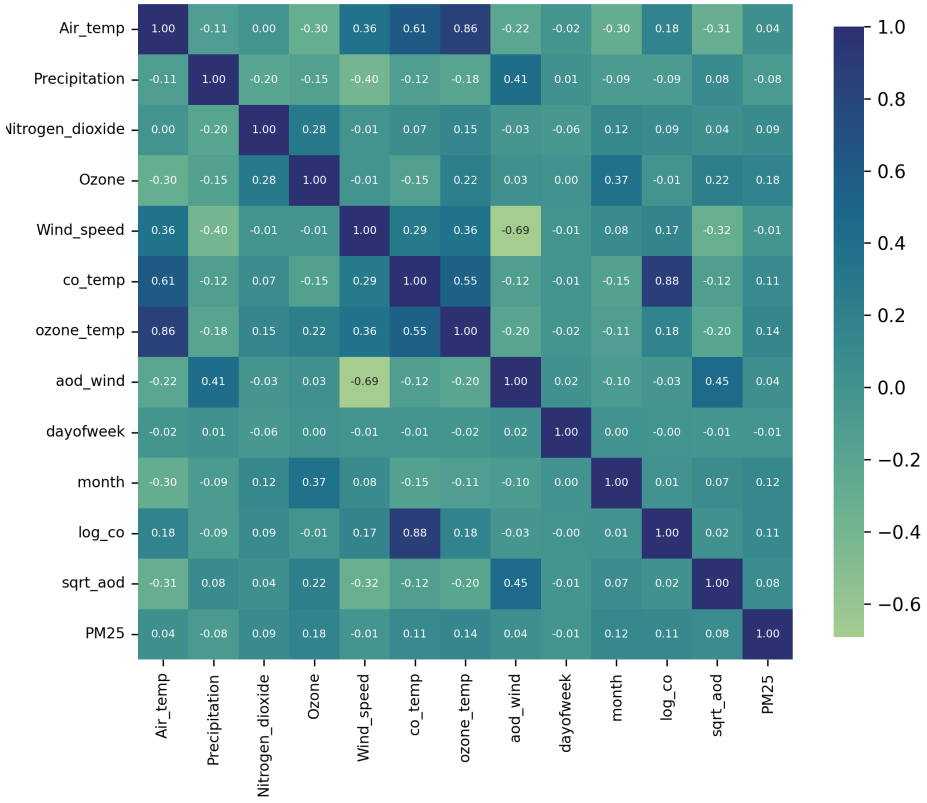

Figure 5: Correlation heatmap between predictors and the target variable

F Data Preprocessing

Table 4: Summary of data preprocessing and feature engineering.

Step 1: Raw Data

- **Predictors** (Extracted from satellite via Google Earth Engine): Date, nitrogen dioxide, AOD, Air temp, northward wind component, eastward wind component, sulfur dioxide, carbon monoxide, ozone, precipitation, surface pressure, land surface temperature, dew point temperature
- **Target** (Downloaded from Open Africa): $PM_{2.5}$
- **Initial processing**: unit conversion, dropping rows whose target values are missing, setting date as index
- **Results**: 13 features and 1 target variable

Step 2: Handling Missing Values

- Nitrogen dioxide (25.4% missing) - Time-based interpolation to preserve trend
- AOD (23.6% missing) - Time-based interpolation to preserve trend
- Sulfur dioxide (60% missing) - Drop column as it exceeds 40%
- Carbon monoxide (10.6% missing) - Time-based interpolation to preserve trend
- Ozone (2.4% missing) - Time-based interpolation to preserve trend
- Precipitation (0.6% missing) - Forward/backward fill as rainfall does not smoothly vary
- Surface pressure (100% missing) - Drop column as it exceeds 40%
- Land surface temperature (43.8% missing) - Drop column as it exceeds 40%
- Dew_point_temp (56.7% missing) - Drop column as it exceeds 40%
- **Results**: 8 features and 1 target variable

Step 3: Feature Engineering

- **Temporal features** - capture predictable weekly and seasonal variations
  - dayofweek (Integer) - Day of week (0 = Monday, 6 = Sunday)
  - month (Integer) - Month of year (1–12)
- **Weather-Pollutant Interactions** - capture complex relationships between meteorological conditions and pollution levels
  - co_temp - CO * Air temp
  - ozone_temp - Ozone * Air temp
  - Wind_speed - sqrt(wind_v² + wind_u²)
  - aod_wind - AOD / (Wind_speed + 0.1)
- **Skewed corrections through monotonic transformations** - Transform skewed variables to stabilize variance
  - Carbon Monoxide - log(CO)
  - AOD - sqrt(AOD)
- Dropped unnecessary columns: AOD, Carbon_monoxide, wind_u, wind_v
- Drop remaining rows with null values

Step 4: Final Dataset

- **Features**: Air_temp, Precipitation, Nitrogen_dioxide, Ozone, Wind_speed, co_temp, ozone_temp, aod_wind, dayofweek, month, log_co, sqrt_aod
- **Target**: $PM_{2.5}$
- **Total rows**: 1573 samples