Estimation and Prediction of $PM_{2.5}$ and PM_{10} Concentration in Kathmandu Valley using Remote Sensing and Machine Learning

Indra Paudel [a], Aayush Chand [b], Ashim Poudel [c], Gaurav Baral [d], Aashish Baral [e]

* Corresponding Author: Netra Bahadur Katuwal

[abcde] *Department of Geomatics Engineering, Pashchimanchal Campus, IOE, Tribhuvan University, Nepal*
*The corresponding author is affiliated with the Department of Geomatics Engineering at Tribhuvan University, Pashchimanchal Campus, Pokhara, Nepal.

[a] pas077bge022@wrc.edu.np, [b] pas077bge003@wrc.edu.np, [c] pas077bge005@wrc.edu.np, [d] pas077bge021@wrc.edu.np, [e] pas077bge002@wrc.edu.np,
* Corresponding Author: netra@ioepas.edu.np

Abstract

In recent years public health and the environment have been seriously threatened by air pollution, especially in low and middle-income nations like Nepal, where the ground-level stations are sparse and thus have limited spatial coverage for monitoring air quality at specific hotspot areas. This study uses machine learning algorithms and satellite data to meet the importunate demand for precise $PM_{2.5}$ and PM_{10} prediction. As the ground station, we used data from the monitoring stations of the Department of Environment covering the period January 2022 to December 2023 and satellite-derived parameters, including Aerosol Optical Depth (AOD), Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI) and soil moisture were processed, analyzed and used to train the models. The Gradient Boosting algorithm performed consistently well with average R^2 scores of 0.82 and 0.84 for $PM_{2.5}$ and PM_{10} respectively, proving efficient in catching the non-linear relationships in the air quality data. Likewise, Random Forest also showed reliable accuracy with an average R^2 value of scores 0.80 and 0.82 for $PM_{2.5}$ and PM_{10} respectively. Spatial maps created from model predictions highlight pollution hotspots, giving active insights for targeted interventions. This study demonstrates the potential of integrating satellite data with machine learning techniques to predict air quality indices (AQI) accurately. This finding provides valuable insights for developing targeted intervention strategies. Additionally, this approach aids in air quality monitoring in the regions that lack ground-based monitoring stations, ensuring more comprehensive environmental assessments, supporting data-driven policy and informed decision-making in Nepal and similar areas. Future research should incorporate high-resolution datasets and additional metrological variables to enhance model reliability and scalability.

Keywords

$PM_{2.5}$, PM_{10}, Remote Sensing, Air Quality, Machine Learning

1. INTRODUCTION

Air pollution indicates the presence of toxic substances in the atmosphere, leading to adverse effects on people's health and millions of premature deaths each year as people around the world are exposed to toxic air pollutants that contribute to various respiratory and cardiovascular diseases. Air pollution became the second leading risk factor for death with 8.1 million deaths worldwide in 2021, including children under 5 years of age[1]. It impacts both urban and rural regions, with approximately 91% of the burden falling on low- and middle-income nations, primarily those in Southeast Asia and the Western Pacific. Air pollution is caused by various pollutants, including sulfur dioxide (SO_2), particulate matter (PM), nitrogen dioxide (NO_2), carbon monoxide (CO), and ozone (O_3). As reported by WHO in 2021, the most significant pollutants of public health matter, including particulate matter, carbon monoxide, ozone, nitrogen dioxide, and sulfur dioxide, contribute significantly to breathing and other diseases, making outdoor and indoor air pollution major causes of morbidity and mortality. Addressing this problem of air pollution, which is the second biggest risk factor for noncommunicable diseases, is key to protecting public health.

Particulate matter (PM) indicates the inhalable particles, composed of black carbon, nitrates, sulfate, ammonia, sodium chloride, mineral dust, or water. Particulate matter (PM) varies in size and is typically classified by aerodynamic diameter, with $PM_{2.5}$ and PM_{10} being the most common types in regulatory standards and most significant for health. Particulate matter (PM), also known as atmospheric aerosol, indicates the complex fusion between solid and liquid particles that vary in size and composition and can remain airborne for extended periods[2]. Particulate matter (PM) with aerodynamic diameters smaller than 2.5 μm ($PM_{2.5}$) and 10 μm (PM_{10}) has received significant attention due to its impact on human health and ecosystems, making it a key focus of recent research[3, 4, 5, 6, 7]. Elevated PM_{10} levels are linked to higher hospital admissions for lung and heart diseases, while $PM_{2.5}$ poses a greater health risk as it penetrates deeper into the respiratory system[8]. According to WHO data of 2021, nearly 99% of the global population breathes air that exceeds the organization's guideline limits and contains high pollutant levels, with low- and middle-income countries experiencing the highest exposure.

A 2020 report by Nepal's Ministry of Health and Population attributes 42,100 annual deaths to air pollution, with 19% affecting children under five and 27% adults over 70. It also reduces the average Nepali's life expectancy by 4.1 years. Air pollution significantly contributes to serious causes of death, including COPD, ischemic heart disease, stroke, lower respiratory infections, and neonatal deaths. In many developing countries such as Nepal, air pollution is responsible for an estimated two million premature deaths globally each year[9]. The limited spatial coverage of monitoring stations and the absence of a dense monitoring network, driven by economic and feasibility constraints, may introduce bias in epidemiological studies[9].

The Air Quality Index (AQI) is widely used to estimate air pollution severity, with higher values indicating increased health risks. Ground-based monitors provide accurate surface-level AQI measurements but are often limited to high-pollution areas, and many countries lack sufficient monitoring networks due to high costs[10, 11]. This has led to the exploration of alternative methods like remote sensing, which offers broader spatial coverage but primarily captures pollutants in the upper atmosphere[12]. Despite this limitation, satellite-based observations still help analyze pollution distribution[13, 14, 15].

Efforts have been made to estimate AQI using a combination of remote sensing and ground data[16]. The Sentinel-5P TROPOMI mission is the first Copernicus satellite for monitoring key atmospheric pollutants such as ozone (O_3), methane (CH_4), formaldehyde (HCHO), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and aerosols. However, $PM_{2.5}$ and PM_{10}, which significantly affect air quality, still require ground-based monitoring due to the lack of high-accuracy satellite sensors[11]. Advanced modeling approaches integrating satellite data are enhancing exposure assessments, health studies, and risk evaluations in alignment with WHO air quality guidelines.

Exposure assessments of ambient particles typically rely on three numerical data sources: ground-based monitoring, satellite remote sensing of aerosols, and air quality model simulations. Geostatistical methods like Kriging[17, 18] and land use regression (LUR)[19] have been used to evaluate air pollution from monitoring data though station coverage is often sparse in sub-urban and rural areas[20].

Satellite remote sensing provides atmospheric column measurements of gases and aerosols, aiding ground-level pollution assessment[21]. Given its extensive coverage, aerosol optical depth (AOD) data is widely used to determine the $PM_{2.5}$ and PM_{10} concentrations[22], as these pollutants are key contributors to air pollution. In Nepal, air quality monitoring primarily tracks $PM_{2.5}$ and PM_{10} for AQI calculation. This study aims to predict their concentrations by integrating satellite data with ground-based measurements using multiple machine-learning algorithms.

Air quality monitoring in Nepal faces significant challenges due to the limited spatial extent provided by ground-based sensors. The existing network is sparse, leaving many areas unmonitored and the problem is compounded by inactive sensors, reducing the reliability and comprehensiveness of available data. The insufficient number of stations makes it difficult to record the spatial variability of pollutants like $PM_{2.5}$ and PM_{10}, hindering accurate assessment and effective management of air quality issues. This scarcity of data hampers our understanding of the full extent of the problem and its impact on the public.

2. STUDY AREA

The selected study area for this project is the Kathmandu Valley, which spans an area of 933.73 square kilometers. The geographical boundaries of the valley are defined by its northernmost position at 27.818° N, southernmost position at 27.403° N, easternmost position at 85.5657° E, and westernmost position at 85.189° E.The Kathmandu Valley, in the lesser Himalayas of Central Nepal, is bowl-shaped with a central elevation of 1,425 m. It is surrounded by four mountain ranges: Shivapuri (2,732 m), Phulchowki (2,695 m), Nagarjun (2,095 m), and Chandragiri (2,551 m) with the Bagmati River flowing through it. It's made up of three districts: Kathmandu, Bhaktapur, and Lalitpur with respective population densities of 5169, 3631, and 1433 people per sq.km.. Its subtropical, continental, semi-humid climate has an average temperature of 18.3°C and a mean annual rainfall of 1,439.7 mm. Pollution mainly comes from open fires, vehicular emissions, construction dust, and post-earthquake damage, worsened by low elevation and lack of dispersing winds.

Kathmandu Valley has been chosen as the study area due to its significant concentration of air quality monitoring stations, which is the highest in Nepal. Within this area, there are 7 governmental air quality monitoring stations and 44 purple-air sensors available, providing a substantial amount of ground-based data for model training and validation. This dense network of stations is crucial for developing and validating the models, as it offers a rich dataset for training purposes. While the ultimate goal is to extend the model to cover all of Nepal, starting with the Kathmandu Valley allows for a more manageable and focused approach. The complexity and time-consuming nature of processing remotely sensed data for a large area make it impractical to begin with the entire country.

Additionally, the availability of a relatively large number of ground stations in Kathmandu Valley, though still not as extensive as in some developed cities globally, supports the justification for this study. The data from these stations will enable robust testing and applicability of the project, ensuring the model's accuracy and reliability. Once the model is successfully developed and validated in Kathmandu Valley, it can be extended to cover the entire country with necessary adjustments to the training parameters.

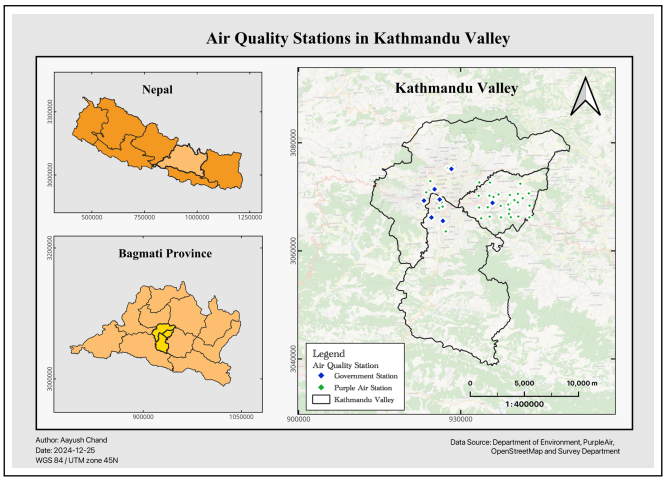

Figure 1: Study Area

3. MATERIALS AND METHODS

3.1 Description of Datasets

To determine the concentration of $PM_{2.5}$ and PM_{10} in the Kathmandu Valley, we utilized data from both remote sensing and ground-based observations (Table 1) spanning two years from January 1, 2022, to December 31, 2023. Our data acquisition strategy focused on obtaining high-quality air quality monitoring data from the Government of Nepal, the Ministry of Population and Environment, Department of Environment, which operates seven air quality monitoring stations across the Kathmandu Valley. These stations provided vital data on various particulate matter, including $PM_{2.5}$ and PM_{10}. Historical data from these stations were accessed through the Department of Environment's Air Quality Monitoring website.

The temporal range for data acquisition extended from January 1, 2022, to December 31, 2023. These two years provided sufficient data to capture seasonal variations, long-term trends, and short-term pollution spikes caused by events such as wildfires, dust storms, industrial accidents, crop burning, and traffic congestion.

The satellite-based air pollution data used in this study were sourced from the MCD19A2.061 MODIS Terra and Aqua MAIAC satellites via the Google Earth Engine Catalog. This dataset enabled the extraction of aerosol optical depth (AOD) information within the study area. Leveraging the MCD19A2.061 product, derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) instruments aboard the Terra and Aqua satellites, the study accessed a daily AOD dataset with a spatial resolution of 1 kilometer. Distributed by NASA's Land Processes Distributed Active Archive Center (LP DAAC), MCD19A2.061 utilized the Multi-angle Implementation of Atmospheric Correction (MAIAC) algorithm, offering daily global coverage for analysis of short-term aerosol variations. This product provided additional parameters such as AOD uncertainty, cloud information, and solar/viewing geometry data, enhancing aerosol characterization and atmospheric correction procedures.

In addition to pollution data from satellite imagery, we incorporated other environmental factors crucial to air quality variation. Environmental factors included the Normalized Difference Vegetation Index (NDVI) from MODIS satellite data, with a spatial resolution of 1 km and a temporal resolution of 1 month. Furthermore, we incorporated land surface temperature (LST) data from MODIS MOD11A1 V6.1, which provided daily LST and emissivity values with 1 KM spatial resolution. These temperature values were derived from the `MOD11_L2` swath product. In regions above 30 degrees latitude, some pixels had multiple observations meeting clear-sky criteria. In such cases, the pixel value represented the average of all qualifying observations. The dataset included both daytime and nighttime surface temperature bands, along with their quality indicator layers (MODIS bands 31 and 32 and six observation layers). Precipitation data were obtained from the CHIRPS dataset of the "UCSB-CHG/CHIRPS/DAILY" image collection, having a spatial resolution of 5.56 kilometers and a daily temporal resolution. Soil moisture data were sourced from the SMAP dataset of the "NASA/SMAP/SPL4SMGP/007" image collection, having a spatial resolution of 9 kilometers and resampled to 1 kilometer and temporal resolution of 3 hours. The soil moisture data was selected with a 3-hour temporal resolution to account for its dynamic nature, as soil moisture levels fluctuate throughout the day based on sunlight and other environmental factors. This resolution was specifically chosen to align the soil moisture data with the exact timestamps of the aerosol optical depth (AOD) data captured by the MODIS sensor over the study area. By synchronizing the temporal resolutions, we aimed to enhance the accuracy of the soil moisture data, ensuring that it corresponded closely to the conditions present during the AOD measurements. This alignment was critical for improving the reliability of the subsequent analysis and modeling.

Table 1: *Description of Datasets*

Datasets	Data Source	Data Type	Spatial Resolution	Temporal Resolution
Shapefile	Survey Department	Vector	-	-
Ground PM2.5 and PM10	Department of Environment	CSV	-	1 Day
AOD	MODIS	Raster	1 KM	1 Day
LST	MODIS	Raster	1 KM	1 Day
NDVI	MODIS	Raster	1 KM	1 Month
Soil Moisture	SMAP	Raster	9 KM (Resampled to 1 KM)	3 Hour

3.2 Methodology

The main aim of our study was to establish a machine learning model to estimate and predict the ground-level air pollution in Kathmandu Valley from 2019 to 2022 through heterogeneous data such as satellite, and meteorological data. In particular, we focused on the determination of ground-level concentrations of $PM_{2.5}$ and PM_{10}, through a machine-learning approach, presented in Figure 2.

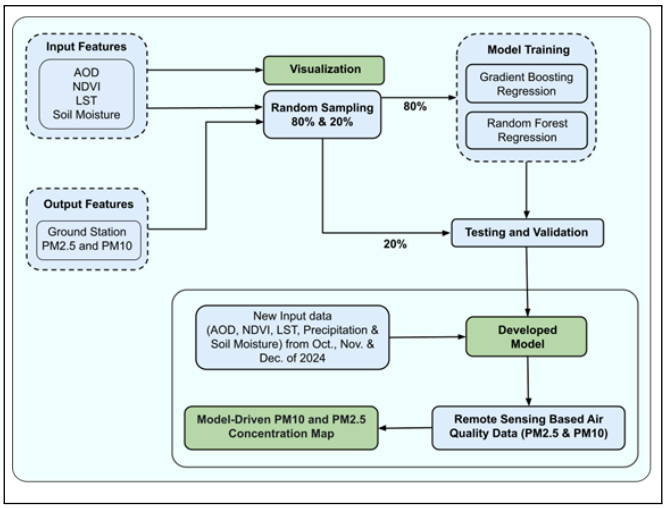

Figure 2: Methodological flowchart used for estimation of $PM_{2.5}$ and PM_{10}

3.2.1 Data Extraction, Cleaning, Processing, and Analysis

In this work, the ground-based $PM_{2.5}$ andPM_{10} data were collected as daily averages, ensuring consistency and reliability in temporal resolution for analysis. The AOD data used in this study was sourced from MODIS, Terra's satellite, and similarly had a daily temporal resolution. This alignment of temporal resolution between ground-based and satellite-derived AOD data minimized discrepancies during analysis. Land Surface Temperature (LST) data, another critical variable, was also obtained with a daily temporal resolution, enabling consistent comparisons across datasets.

The NDVI dataset, however, had a temporal resolution of one month, captured precisely at the start of each month. To address this temporal disparity, NDVI values were interpolated by assigning the value of the first day of the month to all days within 15 days before and after this date. This method ensured that NDVI values represented a consistent vegetation index across the month while maintaining computational simplicity. Soil moisture data presented a unique challenge, as it had a higher temporal resolution of three hours, resulting in eight data points per day. To ensure that the soil moisture data aligned with the temporal resolution of the other variables, the value closest to the time when AOD data was extracted was selected. The remaining seven values for each day were excluded from the analysis. For spatial alignment, point values for each remote sensing dataset were extracted for the exact pixel in which the ground station was located, ensuring spatial correspondence between the datasets. Once all datasets were successfully extracted, they were combined by matching ground station locations and observation times. During preprocessing, missing values and duplicates were identified and removed using Python's Pandas library to enhance data quality and consistency. We detect and remove outliers, which are data points that vary outside of the range of three standard deviations from the mean. To ensure uniformity across different variables, data scaling and normalization techniques were applied. These preprocessing steps collectively enhanced the dataset's suitability for model development. To better understand the dataset and identify patterns, various visualization techniques were employed. Scatter plots, boxplots, and Kernel Density Estimation (KDE) plots were created to examine data distribution and detect potential issues such as skewness or anomalies. For instance, an important insight emerged from the visualization of precipitation data, which revealed that its values were zero more than 75% of the time. This lack of variability indicated that precipitation would likely not contribute significantly to the model and was, therefore, excluded from further analysis.

Furthermore, a correlation heatmap was plotted to examine the relationships between predictor and target variables. This heatmap revealed valuable insights into variable dependencies. The correlation among predictor variables was found to be weak, meaning multicollinearity was not a concern. As a result, feature scaling was deemed unnecessary. After careful evaluation, all variables except precipitation were retained for further analysis, as they showed potential relevance to the target variables. When analyzing the relationships between independent and dependent variables, AOD emerged as a particularly important factor, exhibiting the strongest positive correlation with $PM_{2.5}$ and PM_{10} concentrations. This strong association underscored the importance of AOD as a predictive variable in understanding air quality.

Through careful data extraction, cleaning, and preprocessing, the study established a solid foundation for the subsequent modeling and analysis of $PM_{2.5}$ andPM_{10} levels. Each step, from aligning temporal resolutions to identifying and addressing data anomalies, was carefully designed to verify that the datasets were accurate, consistent, and insightful, paving the way for efficient predictive analysis.

3.2.2 Model Development

This study's primary motivation is to estimate the ground $PM_{2.5}$ and PM_{10} values from the remote sensing observations. To achieve this, we used 2 algorithms: Random Forest and Gradient Boosting models.

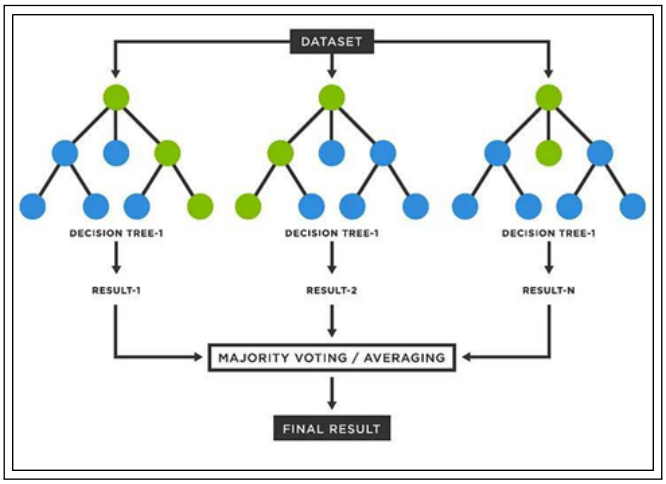

Figure 3: Decision Tree for Multiple Random Forest Regression Model

Simple Random Forest Regression (SRFR) assumes that the error term (ϵ) has a normal distribution with a constant variance and zero mean (homoscedasticity). It's often used as a theoretical concept or for specific statistical tests. Multiple

Random Forest Regression (MRFR) is a powerful ensemble machine learning technique. It combines the predictions of multiple decision trees to create a more robust and accurate model, especially for complex relationships between variables. No single equation represents a random forest model. It's built from a collection of decision trees, each making predictions based on splitting rules learned from the data. The final prediction is an average or weighted average of the individual tree predictions.

Gradient Boosting is a powerful machine-learning technique widely used for regression and classification problems. It is an ensemble method that builds models sequentially, where each new model corrects the errors of the previous one. By combining the strengths of multiple weak learners, usually decision trees, Gradient Boosting creates a robust model capable of capturing complex patterns in the data. This method is particularly effective when dealing with structured data, as it focuses on reducing bias and variance to improve prediction accuracy.

After completing the preprocessing steps, a total of 987 rows of clean and consistent data remained, ready for model development. This dataset was split into training and testing subsets to analyze the model performance. Specifically, 80% of the data was allocated for training the models, ensuring they could learn patterns and relationships within the dataset, while the remaining 20% was reserved for testing purposes, allowing us to validate the models' predictive capabilities.

Two models were developed for this study: Random Forest, and Gradient Boosting. Separate models were created to predict $PM_{2.5}$ and PM_{10} concentrations. Each model was carefully designed to capture the relationships between the predictor variables and the target variables, leveraging their unique strengths.

To enhance the models' performance, hyperparameters were fine-tuned for each model using a trial-and-error technique. This involved systematically adjusting key parameters, such as the learning rate and the number of estimators for Gradient Boosting, to find the optimal combination that delivered the best results. Fine-tuning ensured that the models were neither underfitted nor overfitted, striking a balance between generalization and precision.

Finally, the fine-tuned models were used to predict $PM_{2.5}$ and PM_{10} concentrations using the test datasets. This step allowed us to evaluate the models' accuracy and reliability in predicting air quality parameters. By comparing the predicted values with the actual test data, we could determine how well each model performed and assess their potential for real-world applications.

3.2.3 Testing and Validation

Out of total cleaned and processed data, 80% data were used for training the models and 20% data were used for testing and validation of the models. We applied random sampling to split our data into training and testing sets. This technique ensures that each individual has an equal chance of being chosen, creating representative samples and minimizing bias. Random splitting involves dividing the dataset into a training set, used for model training, and a testing set, used for performance evaluation. This method ensured both sets represent the overall data distribution, providing reliable inferences about the population. Root-mean square error (RMSE), Mean Square Error (MSE), R-squared (R^2), Adjusted R-squared, methods were used for accuracy assessment. If y is the predicted value, $\hat{y}$ is the observed value (AQI-S), n is the total of data, and v is the residual, then:

$$RMSE = \sqrt{\frac{\sum(y-\hat{y})^2}{n}} = \sqrt{\frac{\sum v^2}{n}} \tag{1}$$

The MSE measured the average of the squared differences between predicted and observed values, providing a more severe penalty for larger errors. It was expressed as:

$$MSE = \frac{\sum(y-\hat{y})^2}{n} \tag{2}$$

One of the key accuracy metrics, R-squared (R^2) was calculated to measure prediction accuracy.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3}$$

where:

$\bar{y}$ = Mean of observed values.

The Adjusted R^2 improved upon R^2 by accounting for the number of predictors in the model. It prevented overfitting by adjusting for model complexity. Its formula was:

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(n-1)}{(n-k-1)} \tag{4}$$

where:

R^2 = Adjusted coefficient of determination

N = Total number of observations

K = Number of predictors

The MAE calculated the average of absolute errors between the observed and predicted values. It was less sensitive to outliers compared to RMSE and was defined as:

$$MAE = \frac{\sum|y-\hat{y}|}{n} \tag{5}$$

where:

$|y-\hat{y}|$ = Absolute error of each prediction

3.2.4 Visualization of Prediction Results

As part of this study, our developed models were applied to predict $PM_{2.5}$ and PM_{10} concentrations across Kathmandu Valley. For the visualization phase, recent input data, including Aerosol Optical Depth (AOD), Soil Moisture, Normalized Difference Vegetation Index (NDVI), and Land Surface Temperature (LST), were collected at 10-day intervals during October, November, and December 2024. Predictions were generated at 935 evenly spaced points, each 1 km apart, using both trained models, and the final $PM_{2.5}$ and PM_{10} values were obtained by averaging the outputs from both

approaches. The predicted values were then mapped using a color scale ranging from white to red, where white represented lower concentrations and red indicated higher concentrations, effectively capturing spatial pollution variations. These visualizations help to highlight pollution hotspots and temporal trends, offering valuable insights into air quality distribution. This approach demonstrates the capability of the model in identifying high-risk zones, aiding policymakers in designing targeted air quality management strategies.

4. RESULTS

4.1 Data Visualization

To represent the spatial and temporal variations of air quality and associated environmental parameters, we developed comprehensive maps for each parameter: **Aerosol Optical Depth (AOD)**, **Land Surface Temperature (LST)**, **Normalized Difference Vegetation Index (NDVI)**, and **Soil Moisture**. These maps were created for four distinct four-month intervals spanning the years 2022 and 2023, providing a seasonal perspective of air quality dynamics in the Kathmandu Valley.

4.1.1 Aerosol Optical Depth (AOD)

Figure 4 presents a time-series analysis of the Average Aerosol Optical Depth (AOD) in Kathmandu Valley for the years 2022 and 2023, subdivision into seasonal periods. The average AOD was extracted from the `MCD19A2.061` MODIS Terra and Aqua MAIAC satellite data using the Google Earth Engine Catalog. These datasets, with a 1 km spatial resolution, were averaged over the designed intervals to visualize the concentration of aerosols across the study area. AOD is a key indicator of air pollution, determining how much sunlight is blocked or scattered by airborne particles like dust, smoke, and pollutants. Higher AOD values indicate poorer air quality. The aerosol optical depth (AOD) values observed in Kathmandu Valley range from 0.12 to 0.74. During the winter and pre-monsoon period (January–April), the AOD value is at 0.58 (2022) and 0.69 (2023), recommended very high pollution consistent with hazardous air quality conditions. However, the AOD value is declining during the monsoon season (May-August), at 0.12 (2022), indicating the fresh air due to pollutants carried by rainfall. But, the exceptionally high AOD of 0.68 (May-August 2023) suggests potential anomalies such as wildfires, dust storms, vehicles, etc. Moderate AOD (0.12–0.33) was observed during the post-monsoon periods (September and December), which was above optimal level but in a comparable range with urban areas.

According to WHO and NASA the standard AOD range for healthy air (is < 0.2), Kathmandu valley mostly AOD levels above 0.3, and during winter goes up above 0.6, indicating consistently poor air quality. To more precisely identify health risks, these results highlight the necessity of linking AOD data with ground-level PM2.5 data and the importance of targeted pollution control actions, particularly during high-AOD seasons. Effective mitigation strategies should be informed by additional study that examines both local and worldwide sources of pollution.

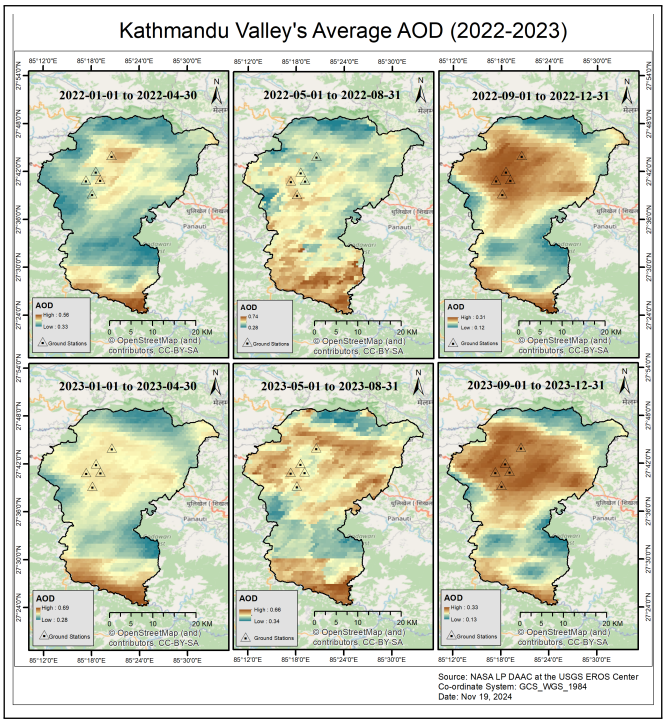

Figure 4: Kathmandu Valley AOD (2022-2023)

4.1.2 Land Surface Temperature (LST)

LST data were derived from the MODIS MOD11A1 V6.1 dataset, which provides daily temperature and emissivity values with a 1 km spatial resolution. The data were aggregated to represent average LST for each four-month interval. These maps help correlate temperature variations with aerosol distributions, emphasizing seasonal effects on air quality.

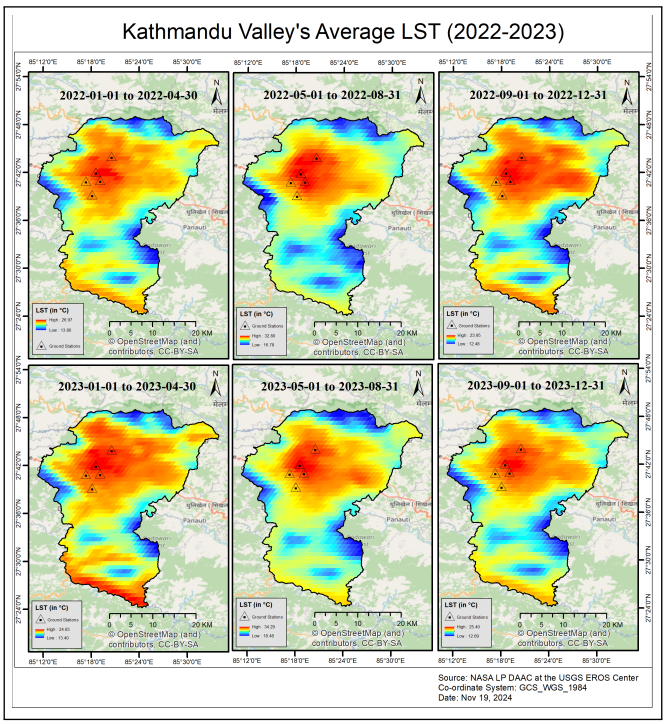

Figure 5: Kathmandu Valley LST (2022-2023)

4.1.3 Normalized Difference Vegetation Index (NDVI)

NDVI data were sourced from MODIS satellite data at a 1 km spatial resolution and a monthly temporal resolution. The interval-averaged NDVI maps provide insights into vegetation coverage and health, factors that influence air quality by regulating particulate deposition and secondary aerosol formation.

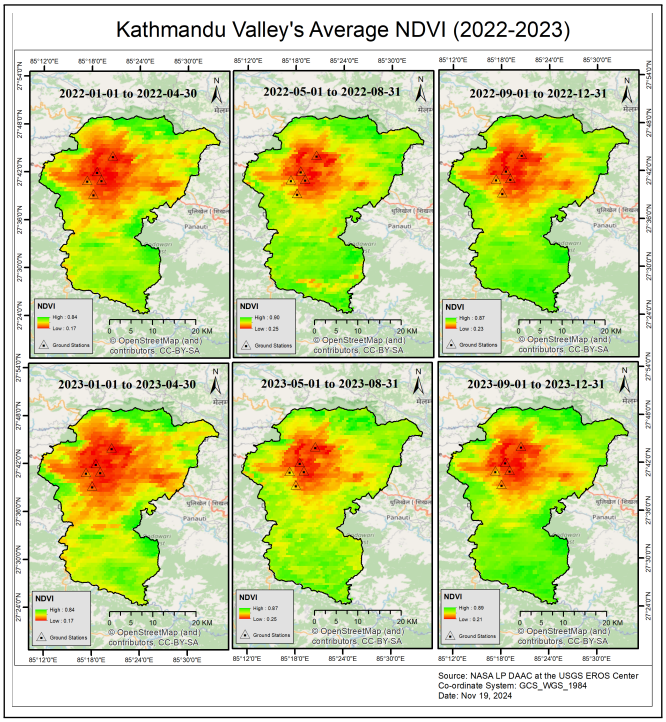

Figure 6: Kathmandu Valley NDVI (2022-2023)

4.1.4 Soil Moisture

Soil moisture data were obtained from the SMAP dataset ("NASA/SMAP/SPL4SMGP/007"), originally available at an 11 km resolution. These data were resampled to a 1 km spatial resolution and aggregated for four-month intervals. The resulting maps highlight the influence of soil moisture on surface-level air quality by affecting aerosol resuspension and deposition dynamics.

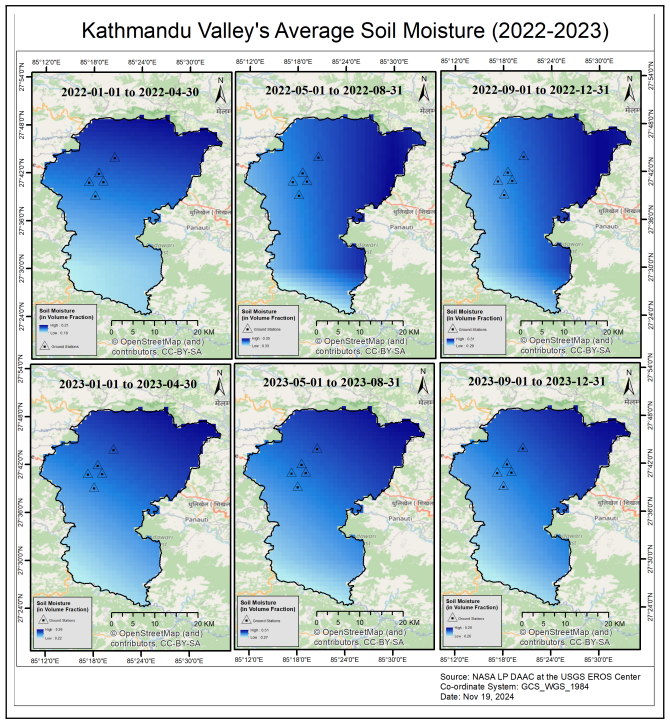

Figure 7: Kathmandu Valley Soil Moisture (2022-2023)

By integrating these environmental variables, the study provides a holistic understanding of air quality variation in the Kathmandu Valley, aiding in the identification of high-risk zones and contributing factors.

4.2 Statistical Evaluation

Figure 8 shows the data distribution of predictor and target variables through the KDE plots. The data distribution of the AOD is positively skewed, with most values concentrated between 0.2 and 0.7, with a long tail expanding up to nearly 2. This indicates that lower AOD values dominate, representing clearer atmospheric conditions most of the time, while higher AOD values are relatively rare. The distribution of LST is close to a normal distribution, with values predominantly between 15°C and 35°C, peaking around 27°C. The distribution of LST shows moderate variability in temperature, with most regions having mild to warm surface temperatures most of the time. The NDVI values are moderately skewed towards lower values, concentrating between 0.2 and 0.6, with fewer values closer to 0.7. This represents sparse variation in most areas, indicating dense vegetation is less common in the study area. The precipitation data values are strongly right skewed, with more than 75% of values being zero, due to which precipitation data were not utilized for further analysis, considering its less significance in model development. The distribution of soil moisture is close to a normal distribution, with most values concentrated between 0.2 and 0.4, peaking around 0.3, indicating fairly uniform soil moisture levels across the regions with moderate variations. The distribution of $PM_{2.5}$ is close to a normal distribution with a slight right skew. Most of the values lie between 50 and 150 $\mu g/m^3$, reflecting moderately high levels of $PM_{2.5}$, which could indicate significant air pollution in the regions analyzed. The distribution of PM_{10} is right-skewed, with values concentrated between 50 and 100 $\mu g/m^3$, with fewer values exceeding 200 $\mu g/m^3$. This indicates

a higher concentration of coarse particulate matter in some areas, but most values suggest moderately polluted air.

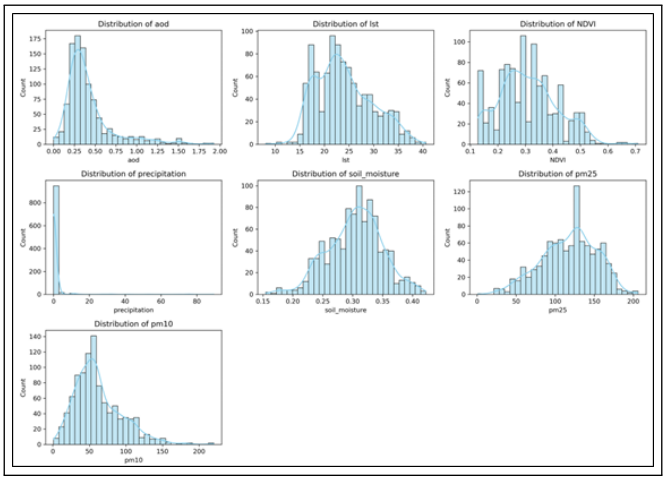

Figure 8: Data distribution of predictor and target variables

The boxplots in figure 9 also show the distribution of the environmental variables. The analysis of these boxplots provides insights into the central tendencies, variabilities, and presence of outliers for each parameter. The median AOD value lies close to 0.3, and the interquartile range is between 0.2 and 0.5, suggesting variability in aerosol concentration within this range. A significant number of outliers are present above 0.5, representing occasional high aerosol loads. The distribution of LST shows the median temperature is approximately 25°C, and the IQR spans from 20°C to 30°C. A few outliers exist beyond 40°C, indicating rare occurrences of extreme surface temperatures. The IQR range of NDVI values ranges from approximately 0.2 to 0.4, with a median NDVI around 0.3, pointing to sparse vegetation coverage in the study area. A small number of outliers are above 0.6, indicating small areas with higher vegetation density. The median value of precipitation is 0, and the IQR is tightly clustered at low values, showing consistent precipitation levels. Numerous outliers extending beyond 80 mm suggest occasional heavy rainfall events. The median soil moisture value is close to 0.3, reflecting moderate moisture levels across the study area. The IQR spans from 0.25 to 0.35, indicating consistent soil moisture conditions. A few outliers below 0.15 represent drier regions. The IQR of PM_{10} ranges between 50 and 100 µg/m³, with a median value of approximately 75 µg/m³, indicating moderate air pollution levels. Outliers above 150 µg/m³ reflect occasional severe air pollution. The median $PM_{2.5}$ value is around 100 µg/m³, highlighting significant particulate matter pollution. The IQR ranges from 75 to 150 µg/m³. Outliers below 50 µg/m³ represent occasional lower air pollution levels.

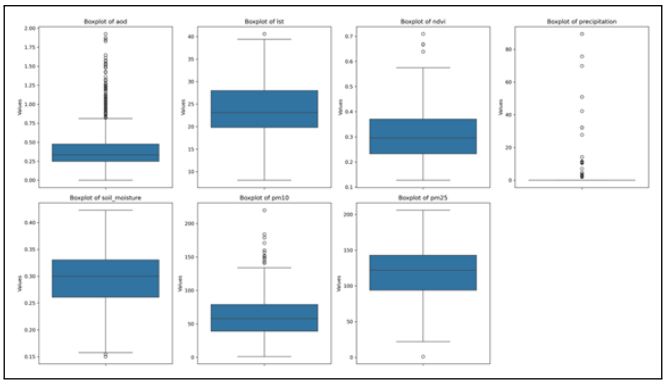

Figure 9: Boxplots of predictor and target variables

The correlation heatmaps in Figure 10 provide a visual representation of the relationships between variables, measured using Pearson's correlation coefficient. The values range from -1 to 1, where 1 represents a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation between variables. The AOD shows a maximum positive correlation with PM_{10} (0.53) and $PM_{2.5}$ (0.49). This suggests that, as AOD increases, PM concentrations also tend to increase. AOD has a slight positive correlation with LST (0.32), indicating that areas with higher aerosol levels might experience slightly higher surface temperatures. LST shows a moderate positive correlation with NDVI (0.39), indicating that vegetation density is somewhat related to surface temperature. LST shows a weak negative correlation with $PM_{2.5}$ (-0.08). NDVI shows a weak negative correlation with $PM_{2.5}$ (-0.22) and PM_{10} (-0.05), suggesting that areas with denser vegetation may have slightly lower PM levels, possibly due to vegetation acting as a filter to pollutants. Soil moisture exhibits a negative correlation with PM_{10} (-0.43) and $PM_{2.5}$ (-0.43), suggesting that higher soil moisture may be associated with lower particulate matter levels, likely due to reduced dust generation. It is weakly negatively correlated with AOD (-0.30), indicating that aerosols might decrease with increasing soil moisture. PM_{10} and $PM_{2.5}$ have a strong positive correlation (0.77), indicating that these two variable metrics are closely related and often change together.

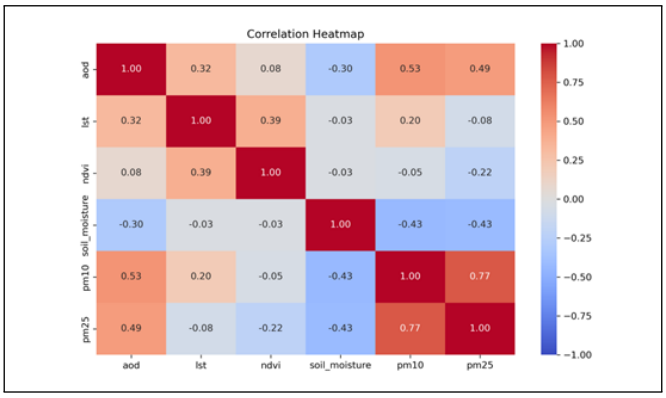

Figure 10: Correlation Heatmap

4.3 Random Forest Model

The Random Forest model demonstrates strong performance in predicting $PM_{2.5}$ concentrations. The R^2 score of 0.8 indicates that the model explains 80% of the variance in $PM_{2.5}$

data, showcasing its ability to capture the complex relationships in the dataset. The error metrics further highlight the model's accuracy, with a Mean Absolute Error (MAE) of 10.98 and a Root Mean Squared Error (RMSE) of 15.16. These metrics suggest that the predictions are both precise and reliable, minimizing deviations from the observed values. For PM_{10} concentrations, the Random Forest model exhibits similar results. The R^2 score of 0.82 signifies that the model explains 82% of the variance in PM_{10} data, demonstrating a strong correlation between predicted and actual values. The error metrics are also favorable, with an MAE of 9.31 and an RMSE of 15.05, indicating a high degree of accuracy and consistency in the predictions.

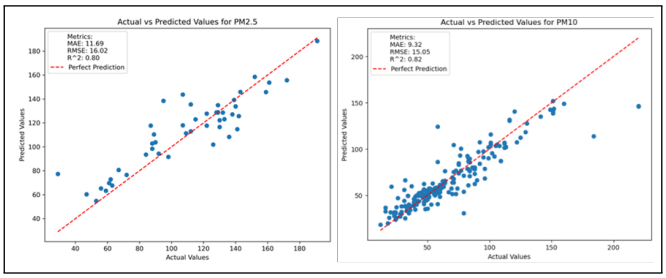

Figure 11: Actual vs Predicted Values (Random Forest)

The cross-validation results provide further evidence of the model's reliability. For $PM_{2.5}$, the metrics are stable across folds, as shown in the left panel of the figure 12. The R^2 values remain high, while the MAE and RMSE stay low, underscoring the model's generalization capability. Similarly, for PM_{10}, the right panel of the figure 12 reveals consistent performance across different folds. The stability in R^2, MAE, and RMSE values indicates that the model maintains its accuracy and robustness on unseen data.

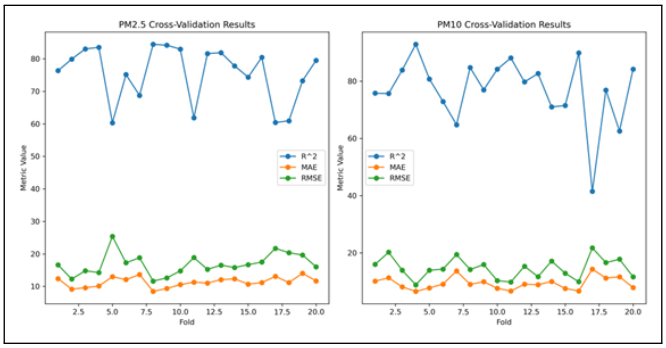

Figure 12: Cross-Validation Results (Random Forest)

In summary, the Random Forest model effectively predicts $PM_{2.5}$ and PM_{10} concentrations, as evidenced by its high R^2 scores, low error metrics, and consistent cross-validation performance. These results highlight the model's capability to handle the complexities of the dataset and provide reliable predictions, making it a valuable tool for air quality analysis and forecasting.

4.4 Gradient Boosting Model

The Gradient Boosting model exhibits strong predictive performance for both $PM_{2.5}$ and PM_{10}, achieving metrics that underscore its accuracy and effectiveness. For $PM_{2.5}$, the model achieves an R^2 of 0.82, signifying that 82% of the variance in the data is explained by the model. Additionally, it records a Mean Absolute Error (MAE) of 10.42, indicating the average absolute difference between predicted and actual values, and a Root Mean Square Error (RMSE) of 14.12, which provides insight into the magnitude of prediction errors. Similarly, for PM_{10}, the model demonstrates even higher accuracy with an R^2 of 0.84, capturing 84% of the variance in the data. The model also achieves a low MAE of 8.16 and an RMSE of 14.08, reflecting its ability to make precise predictions with minimal error. These metrics highlight the robustness of the Gradient Boosting model for air quality analysis.

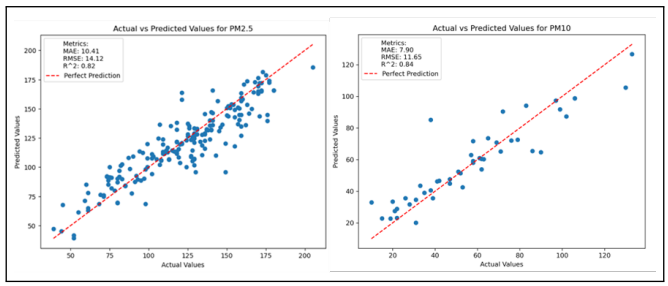

Figure 13: Actual vs Predicted values (Gradient Boosting)

The cross-validation results further establish the reliability of the Gradient Boosting model, showing consistent performance across all 20 folds. For $PM_{2.5}$, the cross-validation results indicate high R^2 values, with minimal variability, ensuring that the model generalizes well to unseen data. Both MAE and RMSE remain stable across the folds, indicating that the model avoids overfitting and maintains a balanced tradeoff between bias and variance. For PM_{10}, the cross-validation results are equally impressive, with consistently high R^2 values across folds and minimal fluctuations in MAE and RMSE, emphasizing the robustness and adaptability of the model to varying data distributions. The visual representation of cross-validation metrics supports these observations, with smooth trends and negligible outliers.

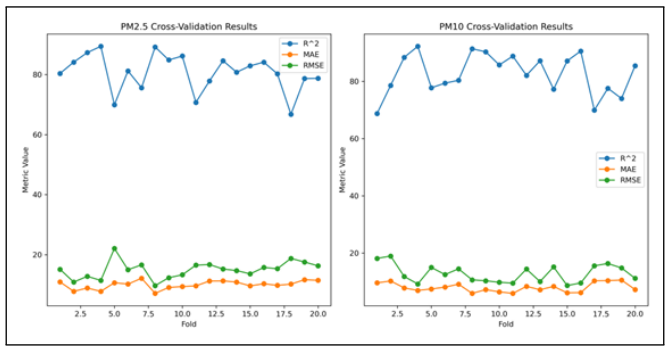

Figure 14: Cross-Validation Results (Gradient Boosting)

In conclusion, the Gradient Boosting model demonstrates exceptional predictive accuracy and consistency for both $PM_{2.5}$ and PM_{10} concentrations. Its ability to achieve high R^2 values and maintain low errors across folds underscores its suitability for air quality monitoring and prediction tasks. This

makes it a reliable and efficient choice for environmental applications, supporting informed decision-making for air pollution management and public health initiatives.

4.5 Model-Driven $PM_{2.5}$ and PM_{10} Concentration Map

To visualize the predicted air quality across the Kathmandu Valley, two maps (figure 15) were developed for average $PM_{2.5}$ and PM_{10} concentrations during October, November, and December 2024, at 10-day intervals. The maps are based on the predicted values generated by our model, using Aerosol Optical Depth (AOD), Soil Moisture, Normalized Difference Vegetation Index (NDVI), and Land Surface Temperature (LST) as input parameters. These inputs were collected at 10-day intervals for the three months. The predictions were made for 935 equally spaced points at a 1 km spacing across the Kathmandu Valley. The model output was then visualized to create average concentration maps of $PM_{2.5}$ and PM_{10}. Key insights from these maps include the spatial variation of particulate matter, highlighting areas of higher pollution levels during the study period.

The $PM_{2.5}$ concentration map (Figure 15 left) displays average values for the study period, with the spatial variation showing hotspots of higher concentrations toward the central regions of the valley. The predicted $PM_{2.5}$ values range from 119.1 µg/m³ to 158.2 µg/m³, with the highest concentrations observed near urbanized and densely populated areas. This pattern indicates significant contributions from vehicular emissions, industrial activities, and urban heating systems in these regions. The PM_{10} concentration map (Figure 15 right) highlights a similar distribution, with concentrations ranging from 64.2 µg/m³ to 108.5 µg/m³. Higher values are evident in the central parts of the valley, closely aligning with $PM_{2.5}$ trends. The elevated PM_{10} levels in these areas suggest additional contributions from road dust and construction activities, which are prominent in urban and semi-urban zones.

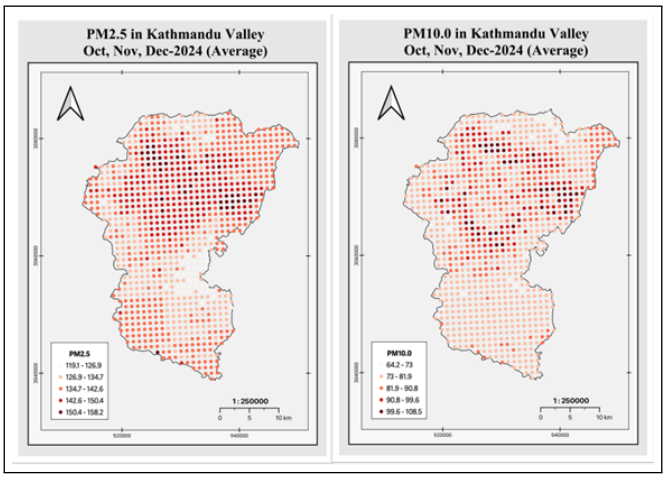

Figure 15: $PM_{2.5}$ and PM_{10} Concentration Map

These maps clearly demonstrate spatial and temporal trends, with certain regions within the valley exhibiting consistently higher pollutant concentrations. The visualizations emphasize the capability of our model to identify pollution hotspots and temporal variations. These maps offer critical insights into spatial pollution trends and can aid policymakers in targeting interventions to reduce air pollution in high-risk areas.

5. Discussion

Air pollution poses severe challenges to public health and the environment, particularly in low- and middle-income countries like Nepal, where limited spatial coverage of air quality monitoring stations hinders effective assessment. This study integrates remote sensing data with ground-based observations to estimate $PM_{2.5}$ and PM_{10} concentrations in the Kathmandu Valley using machine learning models: Random Forest, and Gradient Boosting. Data spanning 2022–2023 from ground stations and satellite-derived parameters, including Aerosol Optical Depth (AOD), Land Surface Temperature (LST), Normalized Difference Vegetation Index (NDVI), and soil moisture, were processed and analyzed.

The results reveal that Gradient Boosting outperformed other models, achieving R^2 values of 0.82 and 0.84 for $PM_{2.5}$ and PM_{10}, respectively, demonstrating its robustness in capturing the non-linear relationships in air quality data. Random Forest also showed high accuracy, with R^2 scores exceeding 0.80. Spatial maps generated from model predictions highlight pollution hotspots, offering actionable insights for targeted interventions.

While the study underscores the potential of remote sensing and machine learning to address gaps in air quality monitoring, it acknowledges limitations, including coarse data resolution and the exclusion of key meteorological factors. Future research should incorporate higher-resolution datasets and additional variables to improve model reliability and scalability. This work demonstrates a viable approach to enhancing air quality assessments, supporting data-driven policy decisions in Nepal and similar regions globally.

Table 2: Metrics for LR, RF, and GBR models

Metrics	Random Forest		Gradient Boosting	
	$PM_{2.5}$	PM_{10}	$PM_{2.5}$	PM_{10}
MAE	10.99	9.32	10.42	8.16
MSE	229.97	226.64	199.55	198.40
RMSE	15.17	15.05	14.13	14.09
R2 Score	0.80	0.82	0.82	0.84
Adjusted R2 Score	0.79	0.82	0.82	0.84

6. Conclusion

There is a very low availability of air quality stations in Nepal, making it almost impossible to accurately monitor the air quality in most regions with ground-based sensors. Installation of air quality measurement stations in all the areas demands high cost and is almost impossible to cover the entire region of the country. One possible solution to this problem of limited spatial coverage of air quality stations is to use remotely sensed raster data to monitor air quality. This study shows the ability of remotely sensed data to predict air quality, particularly the concentration of $PM_{2.5}$ and PM_{10} to solve the above-stated problem with ground-based

monitoring stations. Like the multiple studies in the global context show the ability of machine learning and remote sensing data to determine air quality, our study also supports this in the context of Nepal.

Since the data distribution of influencing factors is most likely to be nonlinear. This study suggests choosing nonlinear models like bagging and boosting-based regression models e.g. Random Forest and Gradient Boosting as in this study which are supposed to be highly efficient in predicting the concentrations of pollutants despite complex distribution of data. In our case, the gradient boosting model is the most reliable and the random forest model is also efficient for analysis with accuracy over 80%. This Study shows that the Aerosol Optical Depth (AOD) and soil moisture show the strongest influence on predicting $PM_{2.5}$ and PM_{10} concentrations compared to LST and NDVI. In conclusion, Integration of remotely sensed data with machine learning models can solve the problem of limited spatial coverage of air quality stations in Nepal.

References

[1] UNICEF. Air pollution accounted for 8.1 million deaths globally in 2021, becoming the second leading risk factor for death, including for children under five years. https://www.unicef.org/rosa/press-releases/air-pollution-accounted-81-million-deaths-globally-2021#:~:text=Air%20pollution%20accounted%20for%208.1,for%20children%20under%20five%20years, June 2024.

[2] Barbara Arvani, R. Bradley Pierce, Alexei I. Lyapustin, Yansen Wang, Grazia Ghermandi, and Sergio Teggi. High spatial resolution aerosol retrievals used for daily particulate matter monitoring over po valley, northern italy. *Atmospheric Chemistry and Physics*, 15:123–155, 2015.

[3] Katalin Bodor, Róbert Szép, and Zsolt Bodor. The human health risk assessment of particulate air pollution (pm2.5 and pm10) in romania. *Toxicology Reports*, 9:556–562, 2022.

[4] Fatemeh Faraji Ghasemi, Sina Dobaradaran, Reza Saeedi, Iraj Nabipour, Shahrokh Nazmara, Dariush Ranjbar Vakil Abadi, Hossein Arfaeinia, Bahman Ramavandi, Jörg Spitz, Mohammad javad Mohammadi, and Mozhgan Keshtkar. Levels and ecological and health risk assessment of pm2.5-bound heavy metals in the northern part of the persian gulf. *Environmental Science and Pollution Research*, 27:5305 – 5313, 2019.

[5] Leão MLP Zhang L da Silva Júnior FMR. - effect of particulate matter (pm(2.5) and pm(10)) on health indicators: climate. - *Environmental geochemistry and health*, - 45(- 5):– 2229–2240.

[6] Masud Yunesian, Roohollah Rostami, Ahmad Zarei, Mehdi Fazlzadeh, and Hosna Janjani. Exposure to high levels of pm2.5 and pm10 in the metropolis of tehran and the associated health risks during 2016–2017. *Microchemical Journal*, 150:104174, 2019.

[7] Maria A. Zoran, Roxana S. Savastru, Dan M. Savastru, and Marina N. Tautan. Assessing the relationship between surface levels of pm2.5 and pm10 particulate matter impact on covid-19 in milan, italy. *Science of The Total Environment*, 738:139825, 2020.

[8] Luka Mamić, Mateo Gaparović, and Gordana Kaplan. Developing pm2.5 and pm10 prediction models on a national and regional scale using open-source remote sensing data. *Environmental Monitoring and Assessment*, 195, 2023.

[9] Saurav Timilsina, Pawan Gautam, and Kundan Lal Shrestha. Relation between modis-based aerosol optical depth and particulate matter in kathmandu using regression model. *Journal of Environment Sciences*, 2023.

[10] Matthew J. Bechle, Dylan B. Millet, and Julian D. Marshall. Remote sensing of exposure to no2: Satellite versus ground-based measurement in a large urban area. *Atmospheric Environment*, 69:345–353, April 2013. Funding Information: This work was supported by NSF-Division of Chemical, Bioengineering, Environmental, and Transport Systems (CBET) (grant 0853467). We acknowledge the NASA GES DISC for the dissemination of OMI data, and the US EPA AQS Data Mart for the dissemination of EPA monitor data.

[11] Giuseppe Lo Re, Daniele Peri, and Salvatore Davide Vassallo. *Urban Air Quality Monitoring Using Vehicular Sensor Networks*, pages 311–323. Springer International Publishing, Cham, 2014.

[12] Randall V. Martin. Satellite remote sensing of surface air quality. *Atmospheric Environment*, 42(34):7823–7843, 2008.

[13] Mikalai Filonchyk, Michael P. Peterson, and Volha Hurynovich. Air pollution in the gobi desert region: Analysis of dust-storm events. *Quarterly Journal of the Royal Meteorological Society*, 147:1097 – 111, 2021.

[14] Anjar Dimara Sakti, Tania Septi Anggraini, Kalingga Titon Nur Ihsan, Prakhar Misra, Nguyen Thi Quynh Trang, Biswajeet Pradhan, I. Gede Wenten, Pradita Octoviandiningrum Hadi, and Ketut Wikantika. Multi-air pollution risk assessment in southeast asia region using integrated remote sensing and socio-economic data products. *Science of The Total Environment*, 854:158825, 2023.

[15] Zihao Zheng, Zhiwei Yang, Zhifeng Wu, and Francesco Marinello. Spatial variation of no2 and its impact factors in china: An application of sentinel-5p products. *Remote Sensing*, 11(16), 2019.

[16] Chitrini Mozumder, K. Venkata Reddy, and Deva Pratap. Air pollution modeling from remotely sensed data using regression techniques. *Journal of the Indian Society of Remote Sensing*, 41:269–277, 2013.

[17] Michael Jerrett, Richard T. Burnett, Renjun Ma, 3rd Pope, C. Arden, Daniel Krewski, K. Bruce Newbold, George Thurston, Yang Shi, Neal Finkelstein, Eugenia E. Calle, and Michael J. Thun. Spatial analysis of air pollution and mortality in los angeles. *Epidemiology (Cambridge, Mass.)*, 16(6):727–736, 2005.

[18] Laina D. Mercer, Adam A. Szpiro, Lianne Sheppard, Johan Lindström, Sara D. Adar, Ryan W. Allen, Edward L. Avol, Assaf P. Oron, Timothy Larson, L.-J. Sally Liu, and Joel D. Kaufman. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (nox) for the multi-ethnic study of atherosclerosis and air pollution (mesa air). *Atmospheric Environment*, 45(26):4412–4420, 2011.

[19] Marloes Eeftens, Rob Beelen, Kees de Hoogh, Tom Bellander, Giulia Cesaroni, Marta Cirach, Christophe Declercq, Audrius Dėdelė, Evi Dons, Audrey de Nazelle, Konstantina Dimakopoulou, Karine Eriksen, Gustaf Falq, Paul Fischer, Cristina Galassi, Regina Gražulevičienė, Joachim Heinrich, Barbara Hoffmann, Michael Jerrett,

Dirk Keidel, and Gerard Hoek. Development of land use regression models for pm(2.5), pm(2.5) absorbance, pm(10) and pm(coarse) in 20 european study areas; results of the escape project. *Environmental Science Technology*, 46(20):11195–11205, 2012.

[20] Tao Xue, Yixuan Zheng, Dan Tong, Bo Zheng, Xin Li, Tong Zhu, and Qiang Zhang. Spatiotemporal continuous estimates of pm2.5 concentrations in china, 2000–2016: A machine learning method with inputs from satellites, chemical transport model, and ground observations. *Environment International*, 123:345–357, 2019.

[21] Randall V. Martin. Satellite remote sensing of surface air quality. *Atmospheric Environment*, 42:7823–7843, 2008.

[22] Shuhui Wu, Yuxin Sun, Rui Bai, Xingxing Jiang, Chunlin Jin, and Yong Xue. Estimation of pm2.5 and pm10 mass concentrations in beijing using gaofen-1 data at 100 m resolution. *Remote Sensing*, 16(4), 2024.