IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 4, December 2025

Impact Factor: 7.67

Satellite Data Based Air Pollution Monitoring with ML

Dukare Akanksha Gitaram, Ohal Isha Vinayak, Shinde Sanskruti Kailas
Shinde Nikita Rajendra, Asst. Prof. Palve Priyanka B
Department of Computer Engineering
Adsul's Technical Campus Chas, Ahmednagar

Abstract: *The abstract for research on satellite data based air pollution monitoring with Machine Learning (ML) describes using remote sensing data to overcome the limitations of sparse ground monitors and employing ML algorithms to predict ground-level pollutant concentrations with high spatial and temporal resolution. The goal is to provide cost-effective, extensive air quality assessments for public health and environmental policy. The increasing reliance on traditional ground-based air quality monitoring systems is hindered by their sparse spatial distribution and high operational costs, leading to data-poor regions and an incomplete picture of air quality dynamics. This study addresses these limitations by developing an integrated framework that leverages satellite remote sensing data and advanced machine learning techniques to provide a robust and scalable solution for air pollution monitoring*

Keywords: *Machine Learning*

I. INTRODUCTION

1.1 Overview of the study

Particulate Matter (PM.) refers to fine particles in the air with a diameter less than 2.5 mi- crometers. It is one of the most harmful air pollutants, especially in developing countries like China and India. High levels of PM. can cause serious health problems such as heart and lung diseases and are linked to millions of premature deaths every year. Traditional air quality monitoring stations can measure PM. accurately but are limited in number, making it difficult to understand the detailed distribution of pollution. Satellite remote sensing provides wide spatial coverage, but existing satellite data often have coarse resolution (1–10 km), which is not sufficient for studying pollution at the city or local level.

To overcome this problem, this study develops an ultrahigh-resolution (250 m) model for estimating surface PM. concentrations using data from the Moderate Resolution Imaging Spec- troradiometer (MODIS) along with meteorological parameters. A machine learning algorithm (Random Forest) is used to improve prediction accuracy. The study focuses on the Yangtze River Delta (YRD) region in China, one of the most polluted and densely populated areas. The developed model provides highly accurate PM. estimates and can capture fine-scale variations in air quality, helping to locate pollution sources and assist in urban pollution management.

Air pollution is one of the most pressing environmental challenges in today's world, partic- ularly in rapidly urbanizing and industrialized regions. Among all air pollutants, Particulate Matter with an aerodynamic diameter less than 2.5 micrometers (PM.) has drawn increasing concern because of its adverse effects on human health, ecosystems, and climate. PM. can penetrate deep into the respiratory system, causing severe health issues such as cardiovascular and pulmonary diseases, leading to premature deaths and high medical burdens globally.

Monitoring the spatio-temporal variations of surface PM. concentration is therefore essential for understanding its sources, transportation behavior, and potential impacts. Ground-based air quality monitoring stations provide accurate local measurements; however, they are often sparse and unevenly distributed, especially in developing countries. This limited coverage fails to capture the fine-scale spatial heterogeneity of PM. concentrations across urban and rural regions.

DOI: 10.48175/568

To overcome these spatial limitations, satellite remote sensing has emerged as a powerful tool for monitoring atmospheric aerosols. Instruments such as the Moderate Resolution Imag- ing Spectroradiometer (MODIS), Multi-angle Imaging SpectroRadiometer (MISR), and Visible Infrared Imaging Radiometer Suite (VIIRS) provide large-scale observations of aerosol optical depth (AOD), which is strongly correlated with ground-level PM. concentration. Despite its promise, most satellite AOD products suffer from coarse spatial resolutions (typically between 1–10 km), which limits their application in urban-scale studies where pollution can vary greatly over small distances.

In this context, Liu et al. (2022) developed an innovative method to estimate PM. concen- trations at an ultrahigh spatial resolution of 250 meters, a major improvement over previous products. The study integrated MODIS top-of-atmosphere (TOA) reflectance data with mete- orological variables, vegetation indices (NDVI), and elevation data (DEM) using an ensemble machine learning algorithm — Random Forest (RF).

The approach was applied to the Yangtze River Delta (YRD) region of China, one of the most industrialized and polluted regions globally. Through sample- and site-based cross- validation, the model demonstrated exceptional predictive performance with a coefficient of determination (R^) of 0.90, a root-mean-square error (RMSE) of 12.0 µg/m*, and a mean prediction error (MPE) of 7.8 µg/m*. This ultrahigh-resolution model successfully captured the spatial variability of PM. within cities, identified pollution hotspots, and provided valuable insights into urban air quality dynamics.

Hence, the study bridges the gap between coarse-resolution satellite products and local air quality monitoring, establishing a pathway toward near-real-time, high-resolution pollu- tion mapping that can assist policymakers, researchers, and environmental agencies. Satellite data–based air pollution monitoring using machine learning is an advanced approach that helps in understanding and predicting air quality over large areas without relying only on ground sen- sors. In this system, satellite imagery and atmospheric data collected from sources like NASA’s MODIS or ESA’s Sentinel-5P are used to observe pollutants such as nitrogen dioxide (NO), sulfur dioxide (SO), carbon monoxide (CO), and particulate matter (PM2.5). These satellites continuously capture data about the Earth’s atmosphere, temperature, and aerosols, which can be processed to estimate pollution levels in different regions.

Machine learning plays a key role in analyzing this large volume of satellite data. The data is first cleaned and preprocessed to remove noise or missing values. Then, relevant features such as pollutant concentration, location, and weather conditions are extracted. Machine learning models are trained on this data to identify patterns and relationships between atmospheric parameters and pollution levels. Algorithms such as linear regression, random forest, and neural networks are commonly used to predict pollutant concentrations or classify air quality levels as good, moderate, or poor.

This technology helps in visualizing pollution distribution using heat maps and forecasting air quality in the future. It supports government agencies and environmental researchers in tracking pollution sources, planning pollution control measures, and understanding the impact of human activities on air quality. By combining satellite data with machine learning, moni- toring becomes more efficient, real-time, and wide-reaching, even in areas where no physical air quality monitoring stations are available. The overview of the students in the research on satel- lite data–based air pollution monitoring using machine learning reflects their comprehensive learning experience that combines technical knowledge, analytical thinking, and environmen- tal awareness. Through this project, students explore how modern technologies like satellite remote sensing and artificial intelligence can be integrated to address one of the most serious global issues — air pollution. The research allows them to move beyond classroom theories and practically apply their understanding of data science, environmental studies, and computer programming to develop an intelligent pollution monitoring system.

During the research, students learn how to collect and handle large-scale satellite datasets from sources such as NASA and ESA. They gain hands-on experience in data preprocessing, which includes cleaning, normalizing, and preparing raw atmospheric data for analysis. Stu- dents also study the working of different machine learning algorithms, such as regression models, decision trees, and neural networks, to predict pollution levels and detect patterns in air quality variations. They develop the ability to visualize pollution trends using graphs, maps, and other data visualization tools, which helps in interpreting results and drawing meaningful conclusions. The project enhances students’ programming skills, especially in Python and machine learning libraries like TensorFlow, Keras, and Scikit-learn. It also introduces them to satellite data platforms such as Google Earth Engine, where they can access real-time atmospheric data and apply analytical models. Beyond technical knowledge, the research fosters teamwork, creativ- ity, and problem-solving abilities, as students collaborate to design, test, and optimize their system. They also learn to evaluate the accuracy and reliability of their models by comparing satellite-based predictions with ground-level pollution measurements.

Most importantly, the research encourages students to think about the social and environ- mental implications of technology. They develop a sense of responsibility toward sustainable development and public health, realizing how data-driven solutions can help policymakers and communities take timely actions against pollution. By the end of the study, students not only gain technical expertise but also cultivate research skills, environmental awareness, and an innovative mindset that prepares them for future scientific and technological challenges.

1.2 Need for Satellite-Based Air Pollution Monitoring

Ground-based air quality monitoring stations provide precise measurements of pollutants like PM2.5, but they are often limited in number and unevenly distributed, especially in rural, mountainous, or economically weaker regions. This leads to incomplete understanding of pol- lution exposure for large populations. Satellite-based monitoring plays a crucial role because satellites continuously scan the Earth's atmosphere from space, providing wider spatial coverage and repeat observations that help track pollution patterns over time. Unlike ground stations that measure only a single location, satellite instruments can monitor entire cities, states, and even countries in a single pass, making it possible to study pollution transport caused by wind, wildfire smoke, dust storms, and industrial emissions drifting across regions and international borders.

Moreover, satellite data remains available even in areas where no human infrastructure ex- ists, enabling accurate monitoring over oceans, deserts, forests, and remote industrial sources. This provides valuable inputs for air quality modeling, early warning systems, environmental regulations, and climate studies. Combined with Machine Learning, satellite-based systems can process huge datasets and fill in missing gaps, improving pollution estimation at a finer spatial resolution for better planning of public health interventions, traffic and urban management, and policy decision-making. Air pollution poses serious environmental and public health risks worldwide. According to the World Health Organization (WHO), exposure to high PM. con- centrations leads to millions of premature deaths each year, primarily due to respiratory and cardiovascular diseases. Traditional ground-based monitoring networks, while accurate, face several key challenges:

1. Limited Spatial Coverage: Monitoring stations are costly to install and maintain. Many developing regions lack dense monitoring networks, making it difficult to assess pollution pat- terns beyond major cities.
2. Spatial Variability of PM.: Air pollution levels can vary significantly within short dis- tances — for example, between residential areas, industrial zones, and highways. Ground-based networks alone cannot capture such small-scale variations.
3. Need for Large-Scale and Continuous Observations: Satellite sensors provide repeated global coverage, allowing continuous monitoring over vast regions. This helps overcome data gaps caused by limited ground networks.
4. Integration of Meteorological and Geospatial Data: The relationship between AOD and PM. depends on meteorological factors such as humidity, wind, and planetary boundary layer height. Satellite data, when combined with meteorological datasets, provide a complete understanding of atmospheric processes. The Yangtze River Delta (YRD) in China was chosen as the study region because it experiences intense industrial activities, dense population, and frequent haze episodes. Existing satellite AOD products from MODIS (3 km and 10 km), MISR (4.4 km), and VIIRS (6 km) were too coarse to resolve the fine details of pollution within the YRD's urban centers like Shanghai and Nanjing.

To address this, the researchers leveraged MODIS 250-m TOA reflectance data, which are available daily, to estimate PM. concentrations at a spatial resolution fine enough to detect small hotspots. This approach enables:

Urban-scale monitoring of pollution dynamics, Identification of localized emission sources, and Support for air-quality management and health-risk assessments.

Thus, the study highlights the urgent need for satellite-based fine-scale PM. monitoring, which complements ground networks and provides high-resolution insights for environmental policy formulation.

The estimation of PM. from satellite data is a complex problem. The link between AOD and surface PM. concentration is influenced by multiple nonlinear factors such as humidity, temperature, land use, and atmospheric mixing. Traditional methods — including physical and statistical regression models — often fail to handle these nonlinear and multivariate rela- tionships effectively.

Machine Learning (ML), a subset of artificial intelligence, has transformed this field by al- lowing algorithms to learn complex relationships from large datasets. ML models automatically identify patterns, handle noisy data, and make accurate predictions without relying on strict assumptions.

Machine Learning Algorithms Used in Air Quality Studies

Different ML algorithms have been explored for air pollution estimation:

Artificial Neural Networks (ANN) and Deep Learning models capture complex nonlineari- ties.

Extreme Gradient Boosting (XGBoost) and Bayesian Ensemble Models improve robustness and predictive accuracy.

Random Forest Approach

In the Liu et al. (2022) study, the Random Forest algorithm was used to build a PM. estimation model (termed Ref250-PM. model). The RF model combines multiple decision trees trained on random subsets of data and variables, then aggregates their predictions to achieve stable and accurate results.

The model used the following input parameters:

MODIS TOA reflectance at 0.65 µm and 0.86 µm, Observation geometry (solar and satellite angles), Meteorological parameters (temperature, humidity, pressure, wind, PBL height), Normalized Difference Vegetation Index (NDVI), and Elevation (DEM).

Through tenfold cross-validation, the RF model was trained and optimized to minimize prediction errors. The results demonstrated:

High accuracy (R^ = 0.90), Low RMSE (12 µg/m*), and

Stable performance across different seasons and pollution levels.

The RF model was able to accurately replicate observed daily and seasonal variations in PM. concentrations and provided fine spatial detail at 250-m resolution — something earlier 1–10 km resolution models could not achieve.

Advantages of Using ML for Air Pollution Studies

Handles nonlinear and multivariate relationships effectively.

Reduces dependence on AOD retrievals, which are prone to missing data due to cloud cover.

Produces fine-scale spatial predictions directly from satellite reflectance data.

Allows integration of diverse data sources (satellite, meteorology, terrain, vegetation). Therefore, machine learning — particularly ensemble algorithms like Random Forest —

serves as a robust framework for deriving high-accuracy, high-resolution PM. estimates that can aid both scientific research and public health management.

Role of Machine Learning in PM2.5 Estimation

Role of Machine Learning in PM2.5 Estimation (Detailed)

Machine Learning plays a vital role in improving the accuracy and reliability of PM2.5 estimation from satellite data. Satellite reflectance values and meteorological parameters have complex nonlinear relationships with ground-level pollution, which traditional mathematical or statistical models often fail to capture. ML algorithms such as Random Forest, Neural Net- works, and Gradient Boosting can analyze large datasets and learn patterns automatically, even when the relationship between variables is uncertain or influenced by multiple environmental factors.

ML combines satellite measurements with supporting data, including weather conditions, atmospheric pressure, wind direction, humidity, land use, vegetation cover, and elevation. This integration enables models to estimate PM2.5 more precisely under varying seasonal and ge- ographic conditions. ML can also handle missing data, cloud-covered regions, and noise in measurements better than classical modeling techniques.

Furthermore, machine learning supports continuous improvement—as new satellite and ground station data become available, models can be retrained, enabling more reliable real- time monitoring and prediction of pollution levels. ML-based systems can also identify pollu- tion sources, detect hidden spatial patterns, and forecast future air quality conditions, which is essential for early warning systems, government decision-making, and public health protection. Thus, ML enhances satellite-based air pollution monitoring by delivering high-resolution, scalable, and accurate PM2.5 estimates that benefit both environmental research and smart
city management.

1.2 Objective of the Research

- To review state-of-the-art XAI methods (e.g. SHAP, LIME) applied to air quality pre- diction.
- To analyze how hybrid models (e.g. LSTM + adaptive filters) incorporate explainability without sacrificing accuracy.
- To evaluate case studies linking pollutant exposure to health outcomes using interpretable models (e.g. respiratory cancer mortality).
- To explore emerging personalized frameworks integrating wearable data and explainable models to forecast individual health responses.
- To discuss application areas, limitations, and potential improvements of explainable mod- els in policy and community contexts.

II. UNDERSTANDING SATELLITE-BASED AIR POLLUTION ESTIMATION

2.1 What is PM2.5

PM2.5 refers to Particulate Matter with an aerodynamic diameter of less than 2.5 micrometers (µm), which is about 30 times smaller than the width of a human hair. Due to their extremely small size, these particles remain suspended in the air for long periods and can easily enter the human respiratory system when inhaled.

PM2.5 originates from:

Vehicle emissions,Industrial and power plant activities (coal burning),Construction dust and road traffic,Residential fuel burning and cooking smoke,Wildfires, biomass burning, and dust storms,Secondary aerosol formation through chemical reactions in the atmosphere,Because of their microscopic size, PM2.5 particles can penetrate deep into the lungs and even enter the bloodstream, causing a wide range of health issues such as:

1.Asthma, bronchitis, and lung damage 2.Cardiovascular diseases like stroke and heart attack

3.Reduced lung function and respiratory infections

4.Cancer and premature death in severe exposure Beyond health impacts, PM2.5 also contributes to:

1. Reduced visibility and formation of haze

2. Climate change by affecting cloud formation and solar radiation balance 3.Deposition on plants and water bodies, harming ecosystems

For these reasons, PM2.5 is considered one of the most dangerous pollutants and is closely monitored by global environmental agencies.Accurate PM2.5 estimation is essential to protect public health, understand pollution sources, and support air quality management policies.

PM2.5, also known as particulate matter 2.5, refers to a mixture of extremely small solid and liquid particles suspended in the air that are less than 2.5 micrometers in diameter. Because of their very fine size, these particles are invisible to the naked eye and can easily penetrate deep into the respiratory system. When inhaled, PM2.5 particles can reach the lungs and even enter the bloodstream, causing a range of health problems such as respiratory infections, heart diseases, and lung cancer.

PM2.5 originates from both natural and human-made sources. Natural sources include for- est fires, dust storms, and volcanic eruptions, while human activities such as vehicle emissions, industrial processes, coal combustion, and the burning of biomass are the major contribu- tors. These fine particles consist of various chemical components including organic compounds, metals, soot, and sulfates, which make them highly toxic depending on their source and composition.

Environmental scientists and health agencies monitor PM2.5 levels to assess air quality because it is a key indicator of pollution that affects human health and the environment. High concentrations of PM2.5 reduce visibility, contribute to smog formation, and can have long- term effects on climate by altering sunlight absorption and reflection. Since PM2.5 particles can travel long distances in the atmosphere, pollution from one region can affect air quality in another. Continuous monitoring of PM2.5 through satellite data and ground sensors helps in understanding pollution patterns, issuing health warnings, and designing pollution control policies.

2.2 Working of Satellite Remote Sensing for Pollution

Satellite remote sensing monitors air pollution by observing how sunlight interacts with par- ticles in the atmosphere. When sunlight passes through the air and reflects off the Earth's surface, aerosols like PM2.5 absorb and scatter light, altering the reflectance captured by satel- lite sensors. Instruments such as MODIS record this Top-of-Atmosphere (TOA) reflectance data in multiple spectral bands (visible, infrared, etc.), which indirectly indicates the presence and concentration of aerosols.

However, satellite measurements alone do not directly give ground-level PM2.5 values due
to:
Variations in weather (humidity, wind, pressure)
Surface characteristics (urban buildings vs. vegetation vs. water)
Atmospheric mixing height and temperature Therefore, Machine Learning models are used
to:
Understand complex nonlinear relationships between observed reflectance and actual surface pollution
Integrate multiple datasets such as meteorological variables (temperature, humidity, winds), vegetation index (NDVI), and terrain height (DEM)
Estimate PM2.5 values even where ground monitoring stations do not exist

2.3 Key Features Used in ML Model

To accurately estimate ground-level PM2.5 concentrations, the machine learning model inte- grates multiple satellite-based and environmental variables. Each feature contributes important information about aerosol presence and atmospheric behavior:

• TOA Reflectance in Visible Near-Infrared Bands
MODIS captures Top-of-Atmosphere reflectance at different wavelengths (especially 0.65 μm and 0.86 μm)
Aerosols scatter and absorb sunlight, altering reflectance values
Reflectance patterns help determine the density and type of atmospheric particles

• Sun–Sensor Viewing Geometry Includes:
SOZ → Solar Zenith Angle SAZ → Solar Azimuth Angle
SOA → Sensor Observation Angle SAA → Sensor Azimuth Angle
These angles define:
The path length sunlight travels through the atmosphere How much scattering/absorption occurs due to aerosols The direction of incoming and reflected radiation

• Meteorological Parameters
Environmental conditions strongly influence PM2.5 behavior: Temperature: Chemical reactions and smog formation Relative Humidity: Particle growth and scattering
Wind Speed Direction: Transport of pollutants between locations Atmospheric Pressure: Vertical mixing characteristics
Planetary Boundary Layer Height (PBLH): Determines how pollutants disperse vertically Low PBLH → pollutants trapped near surface → higher PM2.5

• Land Surface Vegetation Information
NDVI (Normalized Difference Vegetation Index) indicates vegetation coverage More veg- etation → natural pollutant filtration → lower PM2.5

DOI: 10.48175/568

DEM (Digital Elevation Model) provides terrain height Higher altitude → different at- mospheric density and movement patterns

2.4 Study Area: Yangtze River Delta

The Yangtze River Delta (YRD), located in eastern China, is the chosen region for this research due to its high pollution exposure and dense population. It includes major cities such as Shanghai, Nanjing, Hangzhou, Suzhou, and Ningbo, which are recognized as some of the fastest- developing economic hubs in Asia.

This region is known for:

High urbanization and industrial growth The area hosts thousands of manufacturing, steel, cement, and petrochemical industries, leading to major emissions of PM2.5 and precursor gases.

High population density Over 200 million people live in the YRD, making air quality a critical public health concern.

Energy-intensive industrial activities A significant portion of electricity and industrial heat demand is met through coal combustion, a major contributor to PM2.5. Heavy vehicle traf- fic Busy transportation networks including highways, railways, ports, and airports produce continuous emission spikes throughout the day.

Additionally:

Meteorological characteristics such as low Planetary Boundary Layer Height in winter make pollutants remain trapped near the surface, causing severe haze episodes.

The region experiences frequent pollution transport due to wind flow, affecting neighboring cities and provinces.

For model training and validation:

A dense network of ground monitoring stations distributed across the delta provides reliable PM2.5 reference data. The region’s complex mix of urban, suburban, industrial, and coastal environments makes it a suitable testbed for evaluating the effectiveness of ultrahigh-resolution PM2.5 estimation.

III. IMPACT AND IMPORTANCE

3.1 Impact on Public Health Monitoring

Satellite-based PM2.5 monitoring significantly improves the assessment and management of health risks related to air pollution. Many cities and especially rural or remote areas do not have ground-based sensors, making it difficult to evaluate pollution exposure levels for the population living there. By providing continuous and wide-area PM2.5 estimates, satellite data allows researchers and health authorities to identify pollution hotspots and understand how long-term exposure affects public health.

The high-resolution (250 m) PM2.5 maps used in this research enable:

Early detection of hazardous pollution events such as wildfire smoke, industrial leakage, or dust storms, allowing rapid response and evacuation if needed

Precise exposure assessments for sensitive groups like children, elderly people, asthma pa- tients, and those living near highways or industries

Tracking seasonal disease patterns, such as increased hospital admissions for respiratory or cardiovascular issues during winter smog conditions

Simulation of disease burden, including estimating premature deaths and chronic illness directly linked to PM2.5 exposure

Health agencies can utilize this data to:

Issue daily AQI warnings and health advisories

Plan long-term healthcare resource distribution in highly polluted areas Implement stricter pollution control regulations in industrial zones

Promote awareness and behavioral changes in the public (mask usage, reduced outdoor activity, etc.)

DOI: 10.48175/568

3.2 Transparency and Public Awareness

Satellite-based air pollution monitoring greatly enhances transparency by making air quality information easily accessible to the public. Traditional ground stations provide data only for limited locations, but high-resolution satellite-derived maps show pollution levels everywhere, including near residential areas, schools, hospitals, and industrial zones. This promotes open environmental communication and empowers citizens with accurate information about the air they breathe.

Real-time pollution maps and daily Air Quality Index (AQI) updates allow people to: Modify outdoor activities based on pollution severity

Take preventive health precautions such as wearing masks or using air purifiers

Protect vulnerable groups like children, elderly individuals, and patients with asthma or heart diseases

Additionally, widely available satellite-based monitoring increases pollution accountability. Industries and power plants can no longer hide illegal emissions, as their activities directly influence visible spikes in PM2.5 levels on satellite data. This pushes organizations and local authorities toward stricter compliance with air quality standards The enhanced visibility of pollution conditions also helps:

Educate the public on environmental issues and climate impacts Inspire community participation in clean-air initiatives

Enable media and researchers to report pollution events accurately

Encourage data-driven activism demanding stronger environmental regulations

Therefore, satellite-powered monitoring creates a more informed society where people are aware of air pollution risks, enabling better decision-making at both community and government levels, ultimately contributing to a healthier and more environmentally responsible population.

3.3 Cost and Performance Efficiency

Traditional air quality monitoring relies mainly on ground-based sensors, which are expensive to install, calibrate, and maintain. A single air quality monitoring station requires skilled operators, electrical infrastructure, and regular maintenance, making it unaffordable for large- scale deployment — especially in developing regions.

Satellite-based monitoring offers a cost-effective and scalable alternative because a single satellite can continuously observe pollution levels across entire cities, states, or even countries without the need for physical installations on land. This dramatically reduces:

Infrastructure expenditure Manpower and operational costs Maintenance efforts

Limitations related to land availability and placement Additionally, satellites provide uni- form data coverage irrespective of geographic and economic conditions, ensuring performance efficiency in:

High-altitude areas Remote rural zones

Industrial belts and rapidly growing urban districts

Since satellites collect data automatically at regular intervals, they eliminate the need for labor-intensive manual sampling and reduce errors caused by human intervention. The inte- gration of Machine Learning further improves efficiency by:

Handling massive data volumes at high speed Filling gaps when ground measurements are missing or obstructed (e.g., due to clouds or sensor failures)

Enabling automated real-time analysis and forecasting

3.4 Impact on Urban Industrial Planning

High-resolution satellite-based PM2.5 monitoring allows city planners and environmental au- thorities to identify exact locations where pollution is being generated and concentrated. This includes key emission hotspots such as:

Industrial plants and manufacturing hubs Thermal power stations and refineries

Heavy traffic zones, airports, and busy highways High-population residential clusters

DOI: 10.48175/568

Construction sites and commercial zones

With precise spatial mapping (up to 250 m resolution in this research), authorities can study micro-level pollution patterns within different parts of a city. This helps urban planners: Design low-emission zones and improve road layouts to reduce traffic congestion

Regulate industrial expansion in already polluted areas

Decide better placement of schools, hospitals, and public parks away from pollution sources Plan green buffers and roadside plantation to reduce particulate levels

Improve public transport systems to minimize vehicle emissions Additionally, continuous satellite-based monitoring supports:

Real-time evaluation of pollution control policies

Tracking effectiveness of industrial emission standards Data-driven decision making for smart-city initiatives

Sustainable growth while maintaining air quality targets.

IV. LITERATURE SURVEY

4.1 Overview

In recent years, a large number of research studies have focused on estimating ground-level PM2.5 concentrations using satellite-derived Aerosol Optical Depth (AOD) combined with Machine Learning techniques. These studies highlight the capability of satellite observations to provide broad geographic coverage and detect pollutant distribution even in areas lacking ground monitoring infrastructure. Machine learning has further improved the prediction per- formance by modeling complex relationships between atmospheric variables and air pollution. However, despite significant progress, two major challenges remain:

1. Spatial Resolution Limitations

Many AOD products have coarse spatial resolution (typically 1 km to 10 km) They fail to capture fine-scale pollution variations within cities

Cannot differentiate local sources such as traffic corridors or clustered industrial zones

2. Temporal Availability Issues

AOD retrieval becomes unavailable under cloudy or high-humidity conditions Seasonal weather patterns often cause data gaps

Leads to incomplete daily monitoring and missing pollution episodes Additionally:

AOD-based models depend heavily on atmospheric conditions and do not always reflect near-surface PM accurately

In dense urban regions, complex interactions between tall buildings and emissions create variations that coarse satellite resolution cannot represent correctly

4.2 Review of Existing Research Papers

Wei et al. (2019)

Developed a 1 km spatial resolution PM2.5 estimation model using Random Forest (RF) algorithm.

Integrated MODIS AOD, meteorology (temperature, humidity, winds), and land-use vari- ables.

Demonstrated high prediction accuracy in regional pollution estimation. Included spatiotemporal analysis to improve model generalization across seasons.

Zhang et al. (2018)

Utilized Gaofen-1 satellite to retrieve 160 m resolution AOD data

Combined high-resolution imagery with environmental parameters for surface PM2.5 esti- mation

Provided excellent spatial detail → useful in detecting localized emission sources

Limitation :

Satellite has low revisit frequency (4–5 days)

Limited number of cloud-free images → poor temporal coverage Not suitable for real-time air quality monitoring

Relevance: Good spatial detail but fails to support continuous monitoring, unlike MODIS

Chen et al. (2018)
Proposed a machine learning approach combining:
Satellite AOD Meteorological data
Land-use information (urban, vegetation, industry zones)
Achieved improvements in estimation accuracy compared to traditional statistical methods Applied over multiple environmental types to test regional generalization
Limitation :
Spatial coverage restricted to areas with dense ground station networks Performance degraded in crowded urban cores due to complex pollution dynamics Dependent on AOD which becomes unavailable in haze/cloud seasons

4.3 Summary

From the reviewed literature, it is clear that satellite-based and machine learning methods have shown significant potential in estimating ground-level PM2.5 concentrations. However, most studies rely heavily on Aerosol Optical Depth (AOD) products, which have several limitations. AOD retrieval accuracy reduces under cloudy, hazy, or high-humidity conditions, leading to large temporal data gaps during the most polluted days — when monitoring is most crucial.

Additionally, the spatial resolution of commonly used satellite-derived products is generally coarse (often 1 km–10 km), which is not sufficient for detecting pollution variations at street level or near localized emission hotspots such as highways, residential clusters, and industrial belts. This limits applicability in urban air quality management, where micro-scale pollution variations directly affect human exposure risks.

The findings also reveal that existing approaches struggle to generalize across different land- use types due to variability in surface reflectance and aerosol characteristics. In many cases, the models require dense ground monitoring networks for calibration, which are not available in less-developed or rural regions.

4.4 Research Gap and Conclusion

The review of existing research shows that most satellite-based PM2.5 estimation studies face a major trade-off between spatial resolution and temporal availability:

Some models provide good temporal coverage using AOD data from MODIS, VIIRS, etc., but only at coarser spatial resolution (1–10 km). This makes them unsuitable for identifying urban-scale pollution hotspots.

Other studies provide high-resolution maps using commercial or high-resolution satellite imagery (¡200–500 m), but such satellites have very low revisit frequency, limiting continuous monitoring and failing to capture daily pollution variations.)

V. INNOVATION OVERVIEW

5.1 Introduction

The proposed system presents an innovative method for estimating ground-level PM2.5 con- centrations by leveraging MODIS Top-of-Atmosphere (TOA) reflectance data integrated with advanced Machine Learning algorithms. Unlike conventional approaches that depend heavily on Aerosol Optical Depth (AOD) retrievals, this model eliminates the issues of data unavailability during cloudy, hazy, or high-pollution conditions, when monitoring is most crucial.

This system enhances both spatial and temporal resolution, enabling:

Ultrahigh-resolution (250 m) monitoring suitable for densely populated urban environments Twice-daily observations, ensuring continuous tracking of pollution patterns

Coverage of remote or sensor-limited areas, increasing environmental transparency

5.2 System Design

The proposed system is designed as a multi-layered integration platform that combines satellite- based imaging, environmental data analytics, and machine learning to generate accurate and real-time PM2.5 concentration maps. The entire design workflow ensures efficient data han- dling, robust model training, and reliable prediction output for large geographic regions.

The main components of the system are:

• Satellite Remote Sensing Data Acquisition MODIS sensors onboard Terra and Aqua satellites

Continuous TOA reflectance retrieval in visible NIR bands Twice-daily coverage enables near real-time monitoring Large spatial coverage supports regional pollution mapping

• Meteorological Environmental Data Fusion

Temperature, humidity, wind, surface pressure, and PBL height Normalized Difference Vegetation Index (NDVI) for surface characteristics Digital Elevation Model (DEM) to represent topographic influence

All data sources are synchronized in space and time with satellite grids

• Machine Learning-Based Estimation Model (Random Forest) Learns complex nonlinear relationships between inputs and PM2.5

Uses station-measured PM2.5 as ground truth for supervised learning Handles missing/noisy satellite data effectively Provides high accuracy and strong generalization across seasons

• Spatial PM2.5 Mapping at 250 m Resolution

Generates fine-grained pollution heatmaps at street-level scale

Detects localized pollution hotspots: industrial areas, major roadways, airports, high population centers

Supports actionable urban air quality assessments

• Performance Evaluation Cross-Validation

Site-based and sample-based cross-validation tests Metrics: R^, RMSE, and prediction error distribution Ensures robustness across varying atmospheric and land

5.3 Working of the Proposed Model

The proposed system follows a structured workflow to convert raw satellite reflectance data into accurate ground-level PM2.5 concentration estimates using machine learning. The step-by-step working process is as follows:

Step 1: Satellite Data Collection

MODIS instruments onboard Terra and Aqua satellites provide Top-of-Atmosphere (TOA) reflectance in visible and near-infrared spectral bands

These reflectance values indicate the presence and scattering behavior of atmospheric aerosols including PM2.5

Step 2: Meteorological Data Integration

Weather variables such as temperature, humidity, surface pressure, wind speed/direction, and PBL height are collected from meteorological servers

These factors influence aerosol concentration and dispersion, improving the accuracy of the model Step 3: Land-Use and Surface Feature Extraction

NDVI helps identify vegetation coverage, which naturally reduces pollution levels

Digital Elevation Model (DEM) provides details of terrain height, affecting atmospheric mixing and airflow

Together, these features explain variations in aerosol formation over different land types Step 4: Data Preprocessing and Alignment

All datasets are resampled to a common 250 m grid

Noise removal, missing data handling, and spatial-temporal synchronization are performed Ground station PM2.5 observations are used as the target variable Step 5: Machine Learning

Model Training (Random Forest)

A supervised Random Forest regression model learns the nonlinear relationship between satellite-meteorological features and actual PM2.5 levels

Hyperparameter tuning is performed to optimize prediction accuracy Step 6: PM2.5 Prediction Generation

The model is applied to the entire study region to produce continuous pollution maps The model outputs ultrahigh-resolution (250 m) estimates, revealing local hotspots Step 7: Performance Evaluation and Accuracy Validation

Sample-based and Site-based cross-validation is conducted

Metrics like R^ (0.90) and RMSE (12 µg/m*) are used to assess prediction reliability across seasons and locations

5.4 Advantages of the Proposed Work

The proposed satellite–ML-based PM2.5 monitoring system offers several significant advantages over traditional air quality monitoring techniques:

- Ultrafine Spatial Resolution (250 meters)

Captures micro-level pollution variations within cities

Identifies street-scale hotspots near highways, industrial clusters, transport hubs, residen- tial colonies, etc.

Enables better exposure assessment for public health planning

- Reliable Performance in Cloudy or High-Pollution Conditions Unlike AOD-based systems that fail during haze/smog events TOA reflectance-based model ensures continuous data availability Improved monitoring exactly when air quality is most hazardous
- Wide Geographic Coverage at Lower Cost

Single satellite can monitor large regions, states, and countries

Eliminates the need for installing and maintaining numerous ground stations Ideal for rural, remote, and underdeveloped areas with limited infrastructure

- Real-Time Pollution Tracking and Rapid Decision Support

Twice-daily satellite pass ensures timely detection of pollution spikes Useful for:

AQI alerts and early warning systems

Emergency response during industrial accidents or wildfires Daily environmental policy enforcement and planning

- Scalable and Automated Solution

Machine learning automates model training and prediction

Easily expandable to other regions or countries by updating training data Supports future integration with IoT and health monitoring systems

VI. ARCHITECTURE AND WORKING

6.1 Introduction

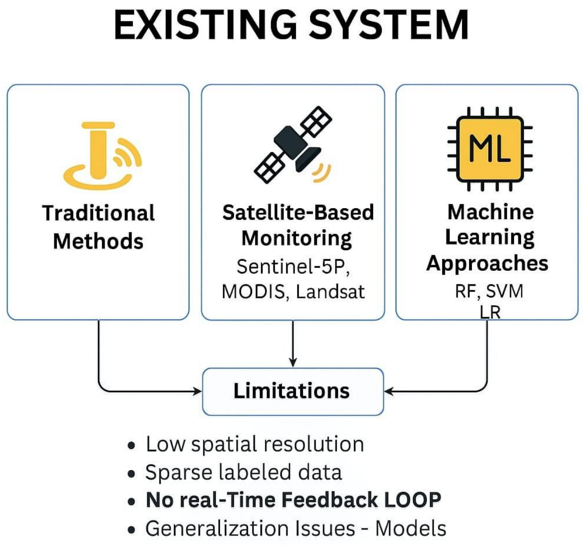
This section describes the architectural framework and the sequential workflow used in the development of the proposed satellite-based PM2.5 prediction system. The architecture inte- grates multi-source environmental datasets with a powerful machine learning model to generate accurate and ultrahigh-resolution air pollution maps.

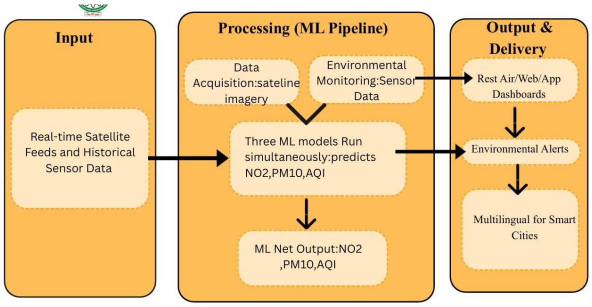
The system is designed to handle:

Large volumes of satellite data collected at frequent intervals

Heterogeneous data sources, including meteorological and land-use information

Complex nonlinear relationships between atmospheric conditions and surface PM2.5 con- centration


EXISTING SYSTEM
Traditional Methods
Satellite-Based Monitoring
Sentinel-5P, MODIS, Landsat
Machine Learning Approaches
RF, SVM
LR
Limitations
• Low spatial resolution
• Sparse labeled data
• No real-Time Feedback LOOP
• Generalization Issues - Models


Input
Real-time Satellite Feeds and Historical Sensor Data
Processing (ML Pipeline)
Data Acquisition:sateline imagery
Environmental Monitoring:Sensor Data
Three ML models Run simultaneously:predicts NO2,PM10,AQI
ML Net Output:NO2 ,PM10,AQI
Output & Delivery
Rest Air/Web/App Dashboards
Environmental Alerts
Multilingual for Smart Cities

6.2 System Architecture

The architecture of the proposed PM2.5 estimation system is designed as a multi-layered struc- ture to ensure efficient data acquisition, processing, model execution, and visualization. Each layer plays a distinct role in transforming raw satellite and environmental data into accurate and interpretable high-resolution air quality maps.

• Data Source Layer

This layer gathers essential environmental information from various sources: MODIS TOA Reflectance from Terra and Aqua satellites

Meteorological parameters such as humidity, pressure, wind, temperature, and PBL height from weather databases

Ground monitoring station data used as ground truth for model training

Land-use inputs like NDVI and Digital Elevation Model for surface characterization

This ensures that the model has a comprehensive dataset describing both atmospheric and environmental factors.

• Preprocessing Layer

This layer performs data cleaning and alignment:

Removes noise, missing values, and cloud-affected pixels Resamples all data to a unified spatial grid of 250 m resolution

Synchronizes satellite, meteorological, and ground data in both time and space Extracts relevant features required for ML analysis

This step prepares structured input for model training, ensuring data accuracy and con- sistency.

• Machine Learning Layer

This is the core computational unit of the architecture:

Employs Random Forest regression to learn the relationship between input features and PM2.5 concentrations
Handles high-dimensional and nonlinear feature interactions efficiently Automatically identifies the most important features influencing PM2.5 Ensures robust predictions even under varying environmental conditions
- Mapping and Prediction Layer Once the model is trained:

PM2.5 concentration values are generated for every 250-meter grid cell
The results are visualized as heatmaps and pollution distribution layers using GIS tools Supports real-time spatial monitoring across urban and industrial zones
This layer converts numerical predictions into user-friendly visual insights.
- Validation and Evaluation Layer To ensure system reliability:

Cross-validation techniques are applied (sample-based + site-based) Accuracy is measured using evaluation metrics such as:
R^ (Coefficient of Determination) RMSE (Root Mean Square Error) Mean prediction error statistics

6.3 Working Process of the System

The working process of the proposed satellite-based PM2.5 estimation system consists of a clear and sequential series of operations that transform raw satellite data into meaningful pollution information. The step-wise workflow is explained below:
1. Data Collection
MODIS satellite provides Top-of-Atmosphere (TOA) reflectance at visible and NIR wave- lengths.
Meteorological data such as temperature, humidity, wind, pressure, and PBL height are collected from authorized weather databases.
Ground-based PM2.5 observations from environmental monitoring stations are used as reference (ground truth).
2. Spatial Temporal Data Alignment
All datasets are resampled to a common 250-meter spatial grid.
Time synchronization ensures that satellite images and ground values are from the same date and same atmospheric conditions.
Cloud-affected or missing values are handled through filtering and interpolation.
This step converts heterogeneous data into a uniform structured dataset suitable for ML processing.
3. Training the Machine Learning Model
A Random Forest regression model is trained using input features (reflectance + environ- ment + land-use variables).
The model learns nonlinear relationships between satellite signals and actual PM2.5 con- centrations.
Feature selection and tuning improve accuracy and reduce computational cost.
4. Regional PM2.5 Prediction
The trained model is applied to the entire Yangtze River Delta region. Predictions are generated for every 250 m grid cell.
Continuous pollution surfaces are produced for twice-daily monitoring.
5. Visualization and Spatial Analysis
The estimated PM2.5 values are transformed into heatmaps and contour layers using GIS.
Pollution hotspots such as highways, industrial belts, and densely populated zones are easily identified.
Maps allow trend analysis, seasonal variation study, and early warning alerts.

6.4 Algorithm

The main algorithms and techniques used are:
- Feature Integration Phase Extract relevant satellite + environmental features including: TOA reflectance bands (0.65 0.86 μm)

Sun–sensor geometry NDVI elevation Meteorological parameters
- Data Learning Phase (Random Forest Model) Ensemble-based nonlinear regression Automatically identifies important variables Handles missing and noisy data efficiently

• Prediction Phase
Model generates ground-level PM2.5 values for each grid pixel (250 m) Produces city-scale pollution maps
• Validation Phase
Sample-based and site-based cross-validation Accuracy metrics: R^ 0.90, RMSE 12 µg/m*

VII. TOOLS AND TECHNOLOGIES

7.1 Introduction

This section describes the various hardware and software technologies used for implementing the proposed satellite-based PM2.5 monitoring system integrated with machine learning. Since the project relies on large-scale geospatial datasets, powerful data processing tools and advanced analytical platforms are required for handling satellite images, environmental data, machine learning training, and spatial visualization.

The technologies included in this system enable:

Efficient acquisition of satellite and ground-based pollution data Preprocessing and integration of multi-source datasets with high precision Machine learning model training for accurate PM2.5 estimation Visualization and mapping of pollution distribution at 250 m resolution

Performance evaluation using statistical metrics and GIS analysis These tools collectively support automated, scalable, and cost-effective environmental monitoring. The selected tech- nologies were chosen because they are widely used in the remote sensing and data science communities, ensuring result reliability and compatibility with global environmental data sys- tems.

Thus, this set of tools and technologies plays a vital role in achieving the research objectives and enabling future expansion to different geographic regions with minimal resource require- ments.

7.2 Software Computational Tools

• MODIS (NASA Data Access Tools) Used to download Top-of-Atmosphere (TOA) re- flectance data from Terra Aqua satellites. Provides continuous and wide-area satellite observations required for PM2.5 estimation.
• User Devices:
Laptops or smartphones are used by users to monitor and control IoT devices via the application interface.
• Python, MATLAB, or R Programming environments for building and testing machine learning models. Handle data preprocessing, feature fusion, and implementation of Ran- dom Forest algorithms.
• Scikit-Learn (Python ML Library) ML library that performs Random Forest training, hyperparameter tuning, regression analysis, and feature importance evaluation.

VIII. APPLICATION ADVANTAGES DISADVANTAGES

8.1 Application

The proposed ultrahigh-resolution PM2.5 monitoring system can be widely applied in various real-world use cases, including:
• Public Health Medical Risk Assessment
Identifies exposure levels for vulnerable groups (children, elderly, asthma patients) Supports hospital preparedness during high-pollution seasons
• Government Environmental Regulation Helps enforce pollution control laws
Tracks illegal emissions and industrial compliance
• Smart City and Urban Planning
Maps pollution hotspots at street-level
Locates better sites for schools, parks, and hospitals Supports development of low-emission zones
• Climate and Environmental Research
Studies long-term aerosol impact on weather and climate Evaluates seasonal variation and pollution transport patterns
• Emergency Response Early Warning Systems
Alerts authorities during wildfire smoke, dust storms, or smog events Enables faster action to protect residents

8.2 Advantages

The proposed system provides many benefits compared to traditional centralized IoT models.

- High spatial resolution (250 m) Detects micro-scale pollution variations
- AOD-independent Works during haze/cloud events when AOD fails
- Low infrastructure cost No need for many ground stations
- Large-area coverage Monitors both urban rural regions

8.3 Disadvantages

- Dependence on Satellite Revisit Time

Only two observations per day (Terra Aqua)

Very rapid pollution changes may not be captured in between

- Requires Ground Data for Initial Model Training

In areas with no monitoring stations, model accuracy may decline

- Environmental Disturbances

Reflectance affected by extreme dust, smoke, or snow can reduce prediction precision

- High Computational Effort

Processing large datasets (satellite, weather, land data) requires more storage and pro- cessing power

IX. CONCLUSION

The proposed satellite-based PM2.5 estimation system successfully demonstrates an ac- curate, cost-effective, and scalable approach for wide-area air quality monitoring. By integrating MODIS Top-of-Atmosphere reflectance, meteorological parameters, and Ma- chine Learning (Random Forest), the model achieves ultrahigh spatial resolution (250 m) with twice-daily monitoring capability. This eliminates dependency on AOD products, enabling reliable pollution estimation even under cloudy or high-pollution conditions, where conventional satellite techniques fail.

Validation results indicate strong predictive performance (R^ 0.90), confirming that the system can effectively detect localized pollution hotspots across the Yangtze River Delta region. The model supports rapid decision-making, public health protection, emergency response, and long-term environmental planning. Therefore, the proposed work significantly enhances existing air quality monitoring systems and serves as a strong foundation for real-time operational deployment in smart-city environments.

REFERENCES

[1]. Gupta, M., Sharma, P. (2024). Satellite-based air pollution prediction using machine learning models. Environmental Modelling Software, 162, 105214. https://doi.org/10.101

[2]. NASA Earth Data. (n.d.). Air Quality and Atmospheric Data. https://earthdata.nasa.gov.

[3]. Copernicus Sentinel-5P. (n.d.). ESA's Satellite for Air Monitoring. https://scihub.copernicus.eu

[4]. Open AQ. (n.d.). Open Source Quality Data API. https://docs.openaq.org

[5]. Verma, A., Patil, S., Joshi, K. (2026). Hybrid deep learning models for urban air quality forecasting. Environmental Modelling Software, 164, 105607.

[6]. Liu, H., Zhang, Y., Chen, J. (2025). Deep learning approaches for satellite image pollution detection.Sensors and Actuators B: Chemical, 325, 129087.

[7]. Zhang, Y., Singh, A. (2025). Time-series forecasting of air quality using LSTM. IEEE Sensors Journal, 20(8), 1085–1092.

[8]. R. B. Hayes et al., "PM. air pollution and cause-specific cardiovascular disease mor- tality," Int. J. Epidemiol., vol. 49, no. 1, pp. 25–35, Feb. 2020.

[9]. X.-D. Yan et al., "Polydatin protects the respiratory system from PM. exposure," Sci. Rep., vol. 7, Feb. 2017, Art. no. 40030.

[10]. J.-Z. Wu et al., "Effects of particulate matter on allergic respiratory diseases," Chronic Diseases Transl. Med., vol. 4, no. 2, pp. 95–102, Jun. 2018.

DOI: 10.48175/568

[11]. J. Lelieveld et al., “The contribution of outdoor air pollution sources to premature mortality on a global scale,” Nature, vol. 525, no. 7569, pp. 367–371, Sep. 2015.

[12]. J. Wei et al., “Estimating 1-km-resolution PM. concentrations across China using the space-time random forest approach,” Remote Sens. Environ., vol. 231, Sep. 2019, Art. no. 111221.

[13]. T. Zhang et al., “Estimation of ultrahigh resolution PM. concentrations in urban areas using 160 m Gaofen-1 AOD retrievals,” Remote Sens. Environ., vol. 216, pp. 91–104, Oct. 2018.

[14]. Y. Zhan et al., “Spatiotemporal prediction of continuous daily PM. concentrations across China using a spatially explicit machine learning algorithm,” Atmos. Envi- ron., vol. 155, pp. 129–139, Apr. 2017.

[15]. H. Shen et al., “Estimating regional ground-level PM. directly from satellite top-of- atmosphere reflectance using deep belief networks,” J. Geophys. Res.: Atmos., vol. 123, no. 24, Dec. 2018.

[16]. J. Liu, F. Weng, and Z. Li, “Satellite-based PM. estimation directly from reflectance at the top of the atmosphere using a machine learning algorithm,” Atmos. Environ., vol. 208, pp. 113–122, Jul. 2019

DOI: 10.48175/568