Article

Not peer-reviewed version

Providing Fine Temporal and Spatial Resolution Analyses of Airborne Particulate Matter Utilizing Complimentary In-Situ IoT Sensor Network and Remote Sensing Approaches

Prabuddha Madusanka Hathurusinghe Dewage , Lakitha Omal Harindha Wijeratne , Xiaohe Yu , Mazhar Iqbal , Gokul Balagopal , John Waczak , Bharana Ashen Fernando , Matthew Lary , Shisir Ruwali , David J. Lary *

Posted Date: 27 May 2024

doi: 10.20944/preprints202405.1685.v1

Keywords: particulate matter; remote sensing; iot sensor, aerosol optical depth; machine learning

 doi:10.20944/preprints202405.1685.v1

Article

Providing Fine Temporal and Spatial Resolution Analyses of Airborne Particulate Matter Utilizing Complimentary In-Situ IoT Sensor Network and Remote Sensing Approaches

Prabuddha Madusanka Hathurusinghe Dewage [1], Lakitha Omal Harindha Wijeratne [1], Xiaohe Yu [2], Mazhar Iqbal [1], Gokul Balagopal [1], John Waczak [1], Bharana Ashen Fernando [1], Matthew Lary [1], Shisir Ruwali [1] and David J. Lary [1,*]

[1] Department of Physics, The University of Texas at Dallas, Richardson, TX75080, USA
[2] Geospatial Information Science, The University of Texas at Dallas, Richardson, TX 75080, USA
* Correspondence: david.Lary@utdallas.edu

Abstract: This study aims to provide analyses of the levels of airborne particulate matter (PM) using a two-pronged approach that combines data from in situ Internet of Things (IoT) sensor networks with remotely sensed aerosol optical depth (AOD). Our approach involved setting up a network of custom-designed PM sensors that could be powered by the electrical grid or solar panels. These sensors were strategically placed throughout densely populated areas of North Texas to collect data on PM levels, weather conditions, and other gases from September 2021 to June 2023. The collected data was then used to create models that predict PM concentrations in different size categories, demonstrating high accuracy with correlation coefficients greater than 0.9. This highlights the importance of collecting hyperlocal data with precise geographic and temporal alignment for PM analysis. Furthermore, we expanded our analysis to a national scale by developing machine learning models that estimate hourly $PM_{2.5}$ levels throughout the continental United States. These models used high-resolution data from the Geostationary Operational Environmental Satellites (GOES-16) Aerosol Optical Depth (AOD) dataset, along with meteorological data from the European Center for Medium-Range Weather Forecasting (ECMWF), AOD reanalysis, and air pollutant information from the MERRA-2 database, covering the period from January 2020 to June 2023. Our models were refined using ground truth data from our IoT sensor network, the OpenAQ network, and the National Environmental Protection Agency (EPA) network, enhancing the accuracy of our remote sensing PM estimates. The findings demonstrate that the combination of AOD data with meteorological analyses and additional data sets can effectively model $PM_{2.5}$ concentrations, achieving a significant correlation coefficient of 0.849. The reconstructed $PM_{2.5}$ surfaces created in this study are invaluable for monitoring pollution events and performing detailed $PM_{2.5}$ analyzes. These results were further validated through real-world observations from two in situ MINTS sensors located in Joppa (South Dallas) and Austin, confirming the effectiveness of our comprehensive approach to PM analysis. The US Environmental Protection Agency (EPA) recently updated the national standard for $PM_{2.5}$ to 9 $\mu g/m^3$, a move aimed at significantly reducing air pollution and protecting public health by lowering the allowable concentration of harmful fine particles in the air. Using our analysis approach to reconstruct the fine-time resolution $PM_{2.5}$ distribution across the entire United States for our study period, we found that the entire nation encountered $PM_{2.5}$ levels that exceeded 9 $\mu g/m^3$ for more than 20% of the time of our analysis period, with the eastern United States and California experiencing concentrations exceeding 9 $\mu g/m^3$ for over 50% of the time, highlighting the importance of regulatory efforts to maintain annual $PM_{2.5}$ concentrations below 9 $\mu g/m^3$.

Keywords: particulate matter; remote sensing; iot sensor, aerosol optical depth; machine learning

 doi:10.20944/preprints202405.1685.v1

1. Introduction

Airborne particulate matter (PM) consists of tiny solid or liquid particles that float in the air [1]. These particles are typically classified by their aerodynamic diameter into several key sizes: PM_1 (particles smaller than 1 μm), $PM_{2.5}$ (particles smaller than 2.5 μm), and PM_{10} (particles smaller than 10 μm). These particles pose considerable health risks, including lung cancer, stroke, asthma, and cardiovascular disease. Studies have particularly highlighted that $PM_{2.5}$, because of its ability to penetrate deeply into the lungs and enter the bloodstream, poses the most significant health hazard [2–4].

Beyond health implications, PM also plays a critical role in climate dynamics by modifying the atmospheric balance of incoming and outgoing electromagnetic radiation. This modification affects various atmospheric conditions, including temperature, wind patterns, and precipitation. The presence of particulate matter can lead to the formation of fog and acid rain and contributes to the greenhouse effect, as discussed in [5–11].

Given the strong link between various health issues and PM, which exhibits significant variations over time and across different locations, it is crucial to conduct comprehensive studies to better understand the distribution of PM with high temporal and spatial precision [3,11]. Although ground-based monitoring stations are vital, their sparse and uneven distribution across regions makes it difficult to achieve continuous nationwide coverage. To overcome these limitations, numerous studies have explored the use of remote sensing techniques and the expansion of ground observation networks. Consequently, contemporary aerosol detection technologies are mainly categorized into remote sensing and in situ observation systems [12].

A significant hurdle in expanding the reach of precise ground-based monitoring networks is the associated expense. Consequently, there has been a focus on creating calibration techniques for affordable airborne particulate sensors. These methods leverage machine learning to improve the accuracy of sensors in measuring particulate matter [13]. These enhanced sensors offer a way to complement the data collected by the environmental agency monitoring networks [14]. Part of our ongoing research involves the development and implementation of an environmental sensing system. This initiative aims to fill geographical gaps in data collection by setting up observation stations on the ground. These stations are designed to provide high-temporal-resolution data specifically in the Dallas area, thereby augmenting existing environmental monitoring efforts.

Research indicates that useful information on surface-level $PM_{2.5}$ concentrations can be gleaned using satellite-derived Aerosol Optical Depth (AOD) data in conjunction with multivariate non-linear machine learning. This allows us to take into account a variety of contextual factors such as weather conditions and other specific geographical contextual information. As a result, incorporating seasonal information and additional data can uncover temporal patterns and spatial characteristics. These insights enable the identification of changes in the relationship between AOD values and $PM_{2.5}$ concentrations [3,15].

[3] developed a machine learning model to provide daily distributions of $PM_{2.5}$ by utilizing a combination of remote sensing and meteorological datasets, along with ground-based particulate matter measurements spanning from 1997 to 2014. Their research outlines the methodology used and presents global average results for this period, showing that the newly developed $PM_{2.5}$ data product can accurately mirror global $PM_{2.5}$ observations, thus serving as a valuable resource for epidemiological studies.

In a separate study, [10], Yu et al., 2022 enhanced the modeling of $PM_{2.5}$ concentrations with high spatial-temporal resolution. They incorporated data from the Next Generation Weather Radar (NEXRAD), along with information from the European Centre for Medium-Range Weather Forecasts (ECMWF), AOD measurements from the Geostationary Operational Environmental Satellite (GOES-16), and $PM_{2.5}$ concentrations measured by in situ sensors from the Environmental Protection Agency (EPA) across the United States. This approach was designed to improve the accuracy and detail of $PM_{2.5}$ concentration modeling.

 doi:10.20944/preprints202405.1685.v1

1.1. Objectives

This study is driven by two main goals. The first goal is to highlight the importance of collecting high-temporal-resolution data and feature variable observations that are synchronized both spatially and temporally with particulate matter (PM) measurements for accurate PM modeling. We used a specially designed system of IoT sensors, both solar and grid-powered, to detect particulate matter and other environmental parameters, deployed extensively in a densely populated area of North Texas. Our system, named MINTS-AI (Multiscale Multiuse Multimodal Integrated Interactive Intelligent Sensing for Actionable Insights), provides access to a wide range of PM sizes, including $PM_{0.1}$, $PM_{0.3}$, $PM_{0.5}$, $PM_{1.0}$, $PM_{5.0}$, and $PM_{10.0}$. These sizes have been carefully modeled using available feature variables such as weather conditions and light intensity, collected directly at the location of PM data gathering, thus eliminating the need for data interpolation to match specific coordinates. The ability of the system to record data at exceptionally high frequencies (every second) is crucial for understanding the dynamic nature of PM concentrations and their interaction with environmental factors. This approach underscores the potential loss of critical PM distribution characteristics when the spatial and temporal alignment of the feature variables and the PM data is not precise. Moreover, incorporating a comprehensive range of light-intensity measurements, which include over ten distinct levels, significantly enhances the precision of PM modeling alongside other environmental variables.

The second goal broadens the detection capabilities for $PM_{2.5}$ through a blend of on-site and remote sensing techniques, making use of a rich dataset augmented with relevant features. On-site detection involved collecting ground-level $PM_{2.5}$ data from our own IoT sensor network (MINTS-AI), as well as data from the OpenAQ network and the National Environmental Protection Agency (EPA) in the United States. We also compiled Aerosol Optical Depth (AOD) data from the Geostationary Operational Environmental Satellite-16 (GOES-16), meteorological information from the European Centre for Medium-Range Weather Forecasts (ECMWF), aerosol assimilation data with air pollutants from the GrADS Data Server, and additional solar and geographical data from 2020 to the present.

2. Materials

AOD, temperature, pressure, relative humidity, height of the planetary boundary layer, wind speed, and direction are identified as crucial contextual variables for modeling and estimating $PM_{2.5}$ concentrations through satellite-based remote sensing and meteorological data [16]. In addition to these, other specific data types have been recognized as beneficial for accurately modeling $PM_{2.5}$ levels. This includes key meteorological parameters from the European Centre for Medium-Range Weather Forecasts (ECMWF), AOD products from the GOES-16 satellite, relevant air pollutants from the MERRA-2 database, solar variables, and various ancillary variables. The primary data for $PM_{2.5}$, used in this context, were sourced from three platforms: the EPA Air Quality System (AQS), the OpenAQ global air quality data platform, and 30 sensors from the UTD MINTS monitoring network.

Data collection for this study, encompassing $PM_{2.5}$, meteorological variables, AOD, and solar angles, varied in temporal and spatial resolutions and spanned from January 2020 to June 2023. To analyze these data, tree-based machine learning methods [17,18] were used. These methods were chosen for their effectiveness in handling the highly time-sensitive nature of the data, including the target variable $PM_{2.5}$ and other influencing environmental factors.

2.1. $PM_{2.5}$ Ground Observations

2.1.1. MINTS Sensors

Temporal and spatial resolution plays a critical role in air monitoring and modeling systems because air quality can change significantly over microenvironments encountered on very small temporal and spatial scales. Harrison et al. (2015) [19] well demonstrated this point, highlighting the challenges in accurately capturing these variations. However, one major obstacle is the significant

 doi:10.20944/preprints202405.1685.v1

maintenance costs of the sensing devices, coupled with the fact that the existing number of ground-based monitoring sites is too limited to provide comprehensive spatial coverage. To address these challenges, numerous studies, including one by Xiaohoe et al. (2021) [11], have been carried out to improve the precision and coverage of $PM_{2.5}$ data collection efforts.

This study focuses on the development of environmental sensing systems and models to estimate particulate matter, using the foundation provided by the MINTS-AI platform. MINTS-AI, a project spearheaded by the Physics Department at the University of Texas at Dallas, is a collaborative initiative that champions open source and open data principles. The platform has been instrumental in the design and deployment of in situ environmental sensing systems across the Dallas-Fort Worth (DFW) metroplex. These systems, which utilize affordable airborne particle sensors combined with machine learning techniques, have been strategically positioned to monitor environmental conditions effectively. The data collected by these sensors are readily available for real-time analysis via an online dashboard, as detailed by [20].

(a) **(b)** **(c)**

Figure 1. MINTS sensing systems deployment: (a) Central Node at Plano, Texas. (b) UTD Node at Dallas college, Texas. (c) UTD Node at Joppa city, Texas

The central and UTD nodes are integral components of MINTS's advanced stationary sensor systems, playing a key role in environmental data collection via IoT sensors. These systems are equipped with a variety of sensors designed to measure particulate matter, gases, ambient light intensity, and climatic conditions. Particulate matter levels are monitored using the IPS 7100 sensors from Pierra Systems, which are celebrated for their affordability, precision, and high sensitivity. These laser scattering sensors are adept at providing precise and real-time measurements of airborne particulate matter, ranging from PM_{10} to ultrafine $PM_{0.1}$, including particle counts and sizes. In particular, the IPS 7100 boasts low power consumption with the capability to collect and sample rapidly every second [21].

Additionally, the system incorporates cost-effective gas sensors like the SCD30 for estimating CO_2 levels and the MICS6814 for gauging concentrations of CO, N_2, H_2, NH_3, CH_4, C_3H_4, C_4H_{10}, and C_2H_6OH. The BME280 sensor is used to measure temperature, humidity, and pressure, thus aiding in climate analysis. The light intensity is tracked via a sensor capable of detecting peaks across a wavelength range of 300 to 1100 nm. The central node also features an ozone module that employs Optical Absorption Spectroscopy to ascertain ozone levels. This expansive sensor network is actively deployed at various sites in the Dallas-Fort Worth metroplex, dedicated to measuring and reporting particle matter concentrations [12].

For our first study, the primary data on all particulate matter (PM) size fractions and other relevant variables, as well as one of the key sources of ground-truth $PM_{2.5}$ observations for $PM_{2.5}$ modeling, were obtained from the Central and UTD Nodes of the UTD MINTS-AI platform. This platform oversees 32 monitoring locations distributed throughout north Texas in Dallas, Collin, and Tarrant counties. A significant number of these monitoring sites are located in Richardson, near the University of Texas at Dallas, with additional sites in Fort Worth, Carrollton, and Plano. At each site,

 doi:10.20944/preprints202405.1685.v1

sensors are configured to collect data on particulate matter, gases, and climatic conditions at high temporal resolution, capturing readings every 3 seconds. However, the scope for $PM_{2.5}$ reference data is somewhat constrained by the relatively limited number of monitoring locations within a somewhat confined area.

2.1.2. EPA

A primary source of $PM_{2.5}$ data in the United States is the EPA's in situ monitoring network, which includes more than 500 ground-based stations scattered throughout the country [22]. These networks are considered among the most reliable sources for aerosol information. The Air Quality System (AQS) of the EPA is a database that aggregates ambient air pollution data, including $PM_{2.5}$ and PM_{10}, collected by the EPA along with state, local, and tribal air pollution control agencies through hundreds of monitors nationwide. However, AQS does not provide real-time air quality data, making the data available only six months after collection [23]. Additionally, negative data values in the AQS can occur due to equipment failures and measurement noise, particularly under very clean atmospheric conditions [11].

In contrast, the EPA's AirNow program offers real-time air quality information, although these data may not have undergone full verification or validation. For this study, $PM_{2.5}$ data, sampled on an hourly basis, were retrieved using both the AQS API and the AirNow API. These datasets were then employed as ground-truth observations for the purposes of model training and validation.

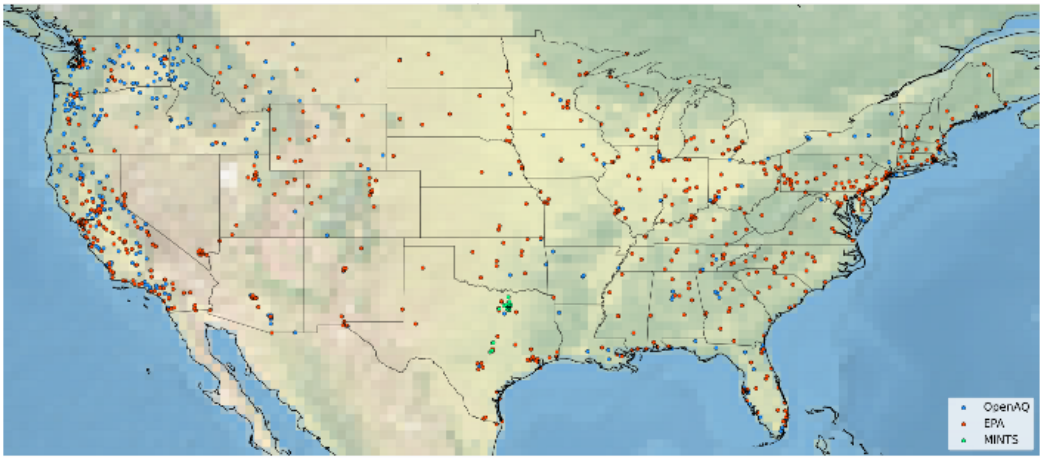


Figure 2. Ground Observation sites of EPA (Red), OpenAQ (Blue) and MINTS (Green).

2.1.3. OpenAQ

In addition to the EPA, OpenAQ, a non-profit organization, facilitates global access to air quality data. It aggregates and standardizes air quality data from all over the world, offering it through a free, open source data platform. Since its launch in 2015, OpenAQ has been collecting historical and real-time data from reference-grade government monitoring stations. The platform covers particulate matter (PM) and various gaseous pollutants, including NO, NO_2, and CH_4. As the largest open source air quality data repository worldwide, OpenAQ provides an API for easy programmatic access to its comprehensive database.

The OpenAQ database incorporates data from approximately 1000 ground-based monitoring stations across the US, including stations from the EPA's in situ monitoring networks [24]. For this study, OpenAQ serves as an additional source of hourly-sampled $PM_{2.5}$ data, which is utilized for modeling training and validation.

 doi:10.20944/preprints202405.1685.v1

2.2. GOES-16 AOD

In this research, the AOD data from the GOES-16 satellite was utilized as one of the key input features. GOES-16, a geostationary weather satellite operated by the National Oceanic and Atmospheric Administration (NOAA) of the United States, is located in a stationary orbit above the Western Hemisphere [25–29]. It offers uninterrupted monitoring of weather conditions throughout the United States, the Caribbean, and South America. The satellite's Advanced Baseline Imager (ABI) is a high-resolution instrument capable of producing images of the Earth's surface and atmosphere. With a spatial resolution as fine as 0.5 km and a temporal resolution reaching up to 30 s, the ABI captures the Earth across 16 spectral bands ranging from visible to infrared wavelengths. This provides vital data on various weather phenomena, including cloud coverage, atmospheric moisture, and temperature [30,31]. AOD, a critical parameter measured by GOES-16, plays a significant role in this study's analysis.

Aerosol Optical Depth (AOD) is a critical parameter for characterizing the role of aerosols on Earth's climate, air quality, and applications in remote sensing. It quantifies the total attenuation of light due to absorption or scattering by aerosol particles in the atmosphere of the Earth [32–37]. Essentially, AOD measures how much sunlight is prevented from reaching the Earth's surface by aerosols in a vertical column of air from the surface to the top of the atmosphere. Measurements of AOD are typically made at specific wavelengths, usually within the visible or near-infrared spectrum, and are reported as a dimensionless value. AOD values can range from 0, suggesting that there are no aerosols present, to values above 1, indicating denser aerosol concentrations in the atmosphere.

The quality and reliability of AOD data are indicated by a Data Quality Flag (DQF), which ranges from 0 to 3. This flag helps users assess the confidence level in the AOD measurements. However, it is important to note that AOD retrieval is challenging in cloudy areas, and the accuracy of AOD data near clouds is less certain. The connection between AOD and $PM_{2.5}$ concentrations is influenced by various factors, including meteorological conditions such as relative humidity and the height of the planetary boundary layer [15,16], which means that this relationship can change over time and at different locations.

2.3. ECMWF Meteorological Data

The levels of airborne particulate matter are significantly influenced by weather conditions, including wind speed, humidity, and temperature. For this study, historical weather data was acquired through the Climate Data Store (CDS) Application Programming Interface (API). The CDS is an extensive digital service that provides a unified web interface to access a wide range of climate and environmental data, including historical, current, and projected future conditions from various sources [38]. This service is developed and managed by the European Centre for Medium-Range Weather Forecasts (ECMWF). Established in 1975, the ECMWF is both a research institute and a round-the-clock operational service, known for producing global numerical weather predictions and maintaining one of the world's largest supercomputing facilities and meteorological data archives [39].

The ECMWF has created ERA5-Land, a reanalysis data set that offers a detailed collection of global atmospheric data spanning from 1979 to the present. ERA5-Land applies the reanalysis technique, which integrates model data with observations from around the world to produce a globally comprehensive and consistent dataset in accordance with physical laws. This data set is structured on a fixed data grid with a spatial resolution of 9 km and provides data updates on an hourly basis. The vertical extent of ERA5-Land ranges from 2 meters above the ground to a soil depth of 289 cm [40]. However, it is important to note that the data from ERA5-Land for these variables are accessible only up to 7 days before the current date. The meteorological variables of ERA5-Land that are used for $PM_{2.5}$ modeling are detailed in Table 1.

 doi:10.20944/preprints202405.1685.v1

2.4. MERRA-2 data

The MERRA-2 dataset, developed by NASA, represents the second iteration of the Modern-Era Retrospective Analysis for Research and Applications. It is an atmospheric reanalysis dataset that combines observational data with sophisticated modeling techniques to create a continuous and high-quality historical account of the Earth's climate system. MERRA-2 utilizes the Goddard Earth Observing System Model, Version 5 (GEOS-5) data assimilation system, which organizes data on a grid with a horizontal resolution of 0.625° by 0.5°. This dataset offers both instantaneous and time-averaged products, available in three-hour intervals [41].

This study incorporates data on air pollutants such as black carbon, sulfate, and nitrate from the MERRA-2 database to improve the precision of its models. Anthropogenic atmospheric aerosols, such as black carbon, are known to adversely affect the global climate [43]. Studies, including Menon et al. (2002) [42], have shown that efforts to reduce black carbon emissions could decelerate the global temperature rise. Additionally, atmospheric aerosols influence atmospheric chemistry; sources such as coal-fired power plants, metal smelting operations, and vehicle emissions release sulfur and nitrogen oxides into the atmosphere. These oxides can react with photochemical products and airborne particles, resulting in the formation of acid aerosols [44].

Sulfate aerosols arise from the oxidation of sulfur dioxide (SO_2) emissions from human activities, such as the burning of fossil fuels, and natural events such as volcanic eruptions. They can significantly affect the climate by reflecting sunlight back into space [45], leading to cooling effects. Nitrate aerosols, produced by oxidation of nitrogen oxides (NO_x) from fossil fuel combustion and biomass burning, contribute to haze and reduced visibility. These aerosols also pose health risks to humans [46]. The formation and impact of these pollutants highlight their importance in understanding and modeling climate and air quality dynamics.

2.5. Solar Illumination

The geometry of solar illumination is crucial in defining the context of Aerosol Optical Depth (AOD) measurements. In $PM_{2.5}$ estimation models, two significant solar-related variables are considered: the solar zenith angle and the solar azimuth angle. These angles influence the distance sunlight travels through the atmosphere of Earth to reach the surface. Specifically, the path length of sunlight through the atmosphere extends with an increase in the sun's zenith angle, which occurs as the sun moves closer to the horizon. This longer journey through the atmosphere allows for more interaction between sunlight and aerosols, leading to increased scattering and absorption of sunlight [47].

As a consequence, AOD values tend to be higher at larger zenith angles because a greater number of aerosols participate in dimming the sunlight. The azimuth angle determines the position of the Sun relative to a specific reference direction, affecting the geometry of light scattering. Variations in azimuth angles can alter the angles at which aerosols scatter sunlight, which in turn influences the observed AOD values, different scattering angles can result in variations in the intensity of light scattered and detected.

2.6. Ancillary Data

In addition to data that change quickly over time, variables that change more slowly can also provide valuable information on environmental, geological, and socioeconomic factors that influence the spatial and temporal distribution of particulate matter concentrations [48]. This study incorporated slowly varying variables such as population density, elevation, soil type, lithology, land cover, crop type, building footprint, and livestock distribution as important contextual ancillary data. These variables help to understand the broader environmental and human factors that can impact the levels of particulate matter.

 doi:10.20944/preprints202405.1685.v1

Population density can significantly influence particulate matter levels due to increased human activities, such as traffic and industrial operations that emit pollutants. The Socioeconomic Data and Applications Center (SEDAC) [49], a component of NASA, provides data on population density in the form of raster data sets. These data sets offer estimates of the population per square kilometer, aligned with figures from national censuses and population registers for the years 2000, 2005, 2010, 2015, and 2020. The available global raster files have a resolution of 30 arc seconds, roughly equivalent to 1 km at the equator.

Topographic features such as mountains and valleys play an important role in the dispersion and accumulation of particulate matter, while trees and other forms of vegetation serve as natural filters, capturing particulate matter and thus mitigating air pollution [50]. Geographic variables such as elevation, soil type, lithology, cropland, and land cover offer information on geological characteristics that could affect the levels of particles.

 doi:10.20944/preprints202405.1685.v1

Table 1. Data source and variables for remote sensing approches.

Source	Variables
EPA	$PM_{2.5}$
OpenAQ	$PM_{2.5}$
MINTS	$PM_{2.5}$
ECMWF meteorological	Temperature Pressure Dewpoint Temperature Precipitation Skin Reservoir Evaporation Specific Humidity Relative Humidity Wind Speed Wind Direction Boundary Layer Height Lake Cover Leaf Area Index, High Vegetation Leaf Area Index, Low Vegetation Snowfall Solar Radiation Total cloud cover Specific Rain Water Content
GOES-16	Aerosol Optical Depth Data Quality Flag
MERRA-2	AOD Analysis Total Column Ozone Hydrophobic Black Carbon Hydrophilic Black Carbon Hydrophobic Organic Carbon Hydrophilic Organic Carbon SO_4 Sulphate Aerosol SO_2 Sulphur Dioxide NH_3 Ammonia NH_4 Ammonium Ion NO_3 Nitrate CO Carbon monoxide CO_2 Carbon dioxide
Ancillary Data	Landcover Population Soil Type Lithology Elevation Cropland Building Footprint Livestock Solar Zenith Angle Solar Azimuth Angle Month

The Cropland Data Layer (CDL) is a geospatial product generated by the United States Department of Agriculture (USDA) using moderate-resolution satellite imagery combined with extensive agricultural ground truth, identifying around 250 different crop types. This dataset, with a spatial resolution of 30 meters, covers the entire continental United States.

 doi:10.20944/preprints202405.1685.v1

Soil data are provided by the National Cooperative Soil Survey through the Web Soil Survey (WSS), an initiative of the USDA Natural Resources Conservation Service (NRCS), which details approximately 100 soil suborder categories [51].

The National Land Cover Database (NLCD) offers detailed information on land cover and changes over time within the United States. With a 30-meter resolution, the NLCD categorizes land into 16 classes, including various types such as water bodies, urban areas, barren lands, forests, shrublands, grasslands, agricultural areas, and wetlands [52,53].

Bathymetric data, crucial for mapping ocean floors and land elevations, are provided by the General Bathymetric Chart of the Oceans (GEBCO), an international consortium of ocean mapping experts. This data set presents elevation data on a grid with 15-arc second intervals [54].

Lithology, which encompasses the geochemical, mineralogical, and physical properties of rocks, influences numerous Earth surface processes, including the transport of materials to ecosystems, soils, rivers, and oceans. The Global Lithological Map (GLiM) was developed by Hartmann and Moosdorf (2012) [55] by synthesizing regional geological maps and literature, offering a representation of global rock types at a spatial resolution of 0.5°. This classification includes 16 lithological classes, providing a comprehensive view of the Earth's surface composition.

Building footprint data are crucial for identifying the number of buildings around a specific location, which can influence wind dynamics and consequently affect PM concentration levels. Microsoft Maps offers a comprehensive open data set of building footprints for the United States. This data set is created through the application of computer vision algorithms in satellite imagery, resulting in 129,591,852 polygonal representations of building footprints in all 50 states of the United States and the District of Columbia [56].

Gridded Livestock Data (GLD) provides a comprehensive overview of the global distribution of various species of livestock in 2015, including cattle, sheep, goats, buffaloes, horses, pigs, chickens, and ducks. This dataset is accessible for free through the Harvard Dataverse repository. It features a spatial resolution of 5 minutes of arc, which is roughly equivalent to 10 km at the equator. The data detail the total number of each species per pixel (5 minutes of arc). It is available in two formats: a dasymetric product and an areal-weighted product, both derived using redistribution methods. For this study, we chose to use the dasymetric product in the TIFF file format. This decision was influenced by the significant environmental impact of livestock farming, especially in terms of greenhouse gas emission from enteric fermentation and manure management, together with the disruption of nitrogen and phosphorus cycles [57].

3. Methodology

This project uses Europa High-Performance Computing (HPC) resources, overseen by the Cyberinfrastructure Research Computing (CIRC) team at the University of Texas at Dallas. Europa is a computing cluster that includes nodes from the decommissioned Stampede supercomputer [58], originally developed by the Texas Advanced Computing Center (TACC) at the University of Texas at Austin. Stampede stood out as a significant and robust supercomputer within the United States, widely utilized for open science research efforts [59].

3.1. All PM size fractions modeling - MINTS Observation

In this phase of the study, data were acquired exclusively through the MINTS sensing system, encompassing 31 sensors positioned in various locations across Texas. PM measurements were obtained using the IPS7100 sensor, which was then utilized as target variables for the machine learning models. The analysis framework integrated a variety of variables from different sensors within the MINTS sensor unit as feature variables (Table 2). These variables encompassed CO_2 concentration measured by the SCD30 sensor and environmental parameters such as temperature, humidity, and atmospheric pressure, all monitored by the BME280 sensor. The study also included data on visible light intensities across different color bands of the AS7262 sensor, as well as ambient light intensities

 doi:10.20944/preprints202405.1685.v1

detected by the TSL2591 sensor. To further enhance the feature set, data related to the infrared (IR) and ultraviolet (UV) light intensities of the VEML6075 sensor were incorporated.

3.1.1. Data Matching

Since all sensors are integrated within a single unit in the MINTS sensing system, there is no need to align the data based on spatial coordinates. Data sampling occurs every 10 s, but it is important to note that the recording times across the different sensors are not synchronized. To effectively align the various sensor data with the PM measurements, we implemented a one-minute time aggregation approach. This method addresses the challenge of matching the high temporal resolution of our data with that of other sensing systems, which generally have lower temporal resolutions. As a result, our analysis is based solely on the high-temporal-resolution data from MINTS, limiting our feature variables to those available within the MINTS dataset.

Table 2. MINTS embedded sensors and variables.

Sensor	Variables
IPS7100	$PM_{0.1}$ $PM_{0.3}$ $PM_{0.5}$ $PM_{1.0}$ $PM_{2.5}$ $PM_{5.0}$ $PM_{10.0}$
BME280	Temperature Pressure Humidity
SCD30	CO_2
AS7262	Violet Blue Green Yellow Orange Red
TSL2591	Luminosity Infrared Full Spectrum Visible Light Lux
VEML6075	Ultraviolet A Ultraviolet B

3.1.2. Experiment Design

To explore the effectiveness of different variables from various sensors across different PM size fractions, we organized the variables into three unique group configurations (Table 3). Each group contains seven specialized models, each addressing different PM size categories. Group 1 models are built using only meteorological data from BME280 sensor. Group 2 models use a wider range of variables, including meteorological data from BME280, CO_2 concentrations from SCD30, and light intensities from AS7262, TSL2591 and VEML6075. Meanwhile, Group 3 is tailored to assess the impact of light intensities on different PM size fractions specifically. The data sets for each group include around 617, 000 entries, split into two parts: 80% of the data is used for training purposes, and the remaining 20% is reserved for testing.

The model's training involved selecting a range of potentially optimized hyperparameters with an understanding that the training performance heavily depends on various factors. One such critical

 doi:10.20944/preprints202405.1685.v1

factor is the number of trees in tree-based models, which represents a key hyperparameter. Achieving an optimal balance is crucial because increasing the number of trees not only influences the model's performance but also raises the demand on computer memory resources. Therefore, a careful decision was made regarding the number of trees to fit within the constraints of the available computational infrastructure. After training, the model underwent a validation process using the test data set. This step includes assessing performance metrics like the Root Mean Square Error (RMSE) and the Correlation Coefficient (R) to gauge the model's accuracy and predictive ability.

Table 3. MINTS Observation PM Groups.

Group	Weather	CO_2	Light
1	✓		
2	✓	✓	✓
3			✓

3.2. $PM_{2.5}$ modeling - In-situ and Remote Sensing

3.2.1. Data Matching

This study on estimating ground-level $PM_{2.5}$ concentrations analyzed three and a half years of historical data, covering the period from January 2020 to June 2023. The variables used in this study were sourced from various databases, each with its own temporal and spatial resolutions. Ground-level $PM_{2.5}$ data from the EPA Air Quality System (AQS) and OpenAQ, along with ECMWF meteorological data, are available at a temporal resolution of one hour and were used as is, without the need for aggregation. Conversely, $PM_{2.5}$ data collected by the MINTS platform have a native temporal resolution of 3 seconds, necessitating aggregation to align with the one-hour temporal resolution of other data sources. Aerosol Optical Depth (AOD) data from the GOES-16 satellite, which are recorded every five minutes, was selected based on the timestamp closest to the $PM_{2.5}$ observation timestamps for consistency. Atmospheric gas data, obtained from the MERRA-2 GEOS-5 model, have a temporal resolution of three hours. Linear temporal interpolation was used to fill in the gaps between data points, ensuring that all variables match the $PM_{2.5}$ observation timestamps accurately.

Following the harmonization of all highly dynamic data to a consistent one-hour temporal resolution, feature variables such as the AOD data from GOES-16, meteorological data from ECMWF, and solar angles were aligned with ground-based $PM_{2.5}$ measurements. These $PM_{2.5}$ measurements were sourced from three distinct platforms: the EPA Air Quality System (AQS), OpenAQ, and the MINTS platform, and were used as the target variable in the analysis.

Data from various sources come with different spatial resolutions and utilize distinct grid coordinate systems. The AOD data from GOES-16 have a fine spatial resolution of 2 km by 2 km. However, the original AOD data, stored in NetCDF format on Amazon S3, adhere to the GOES-R Advanced Baseline Imager (ABI) fixed-grid projection coordinate system. To make this data usable for geographical analyses, it is necessary to transform the AOD data into a geographic coordinate system. This transformation relies on metadata that includes details about the perspective point height and the sweep angle axis. After conversion, the AOD data are ready for further analysis.

The European Centre for Medium-Range Weather Forecasts (ECMWF) Climate Data Store presents its meteorological variables from the ERA5 land reanalysis in GRIB grid files, featuring a horizontal resolution of 0.1°. Meanwhile, data from the MERRA-2 GEOS-5 model, available in netCDF-4 format, provides an approximate spatial resolution of 50 km x 50 km, offering a broader spatial coverage for analysis.

To effectively train a machine learning model, it is crucial to synchronize all datasets, which contain various variables, in terms of both time and spatial coordinates. The alignment of the coordinates of the dataset was achieved by using the locations of ground-based PM observation sites from the EPA,

 doi:10.20944/preprints202405.1685.v1

OpenAQ, and MINTS as the reference coordinate system. A multilinear interpolation method was used to ensure that the data sets were accurately aligned.

After the matching process was completed, a data table was assembled. This table includes synchronized time and coordinates for each entry, alongside $PM_{2.5}$ observation values, meteorological factors, AOD, air pollutant gases, and solar illumination geometry. In addition, ancillary data from various sources were integrated into the table by aligning their spatial coordinates with the reference coordinate system. This integration included relevant data values but did not consider the temporal aspect of the data.

It is important to note that GOES-16 AOD data are available only during daylight hours and in cloud-free locations. The Data Quality Flag (DQF) included with the AOD data provides insight into the quality of the AOD measurements. To maintain high data integrity, only AOD values classified as high quality, based on DQF information, were selected for use. As a consequence, many entries in the data set had missing AOD values, which were then filled with the corresponding AOD data from the MERRA-2 dataset to complete the data set.

3.2.2. Experiment Design

To explore the effects of incorporating data from MINTS $PM_{2.5}$, MERRA-2, and other sources on $PM_{2.5}$ modeling, six unique model configurations were developed (Table 4). The first model, Model-1, is the basic model that includes the MINTS data but excludes the Ancillary and MERRA-2 data. Model-2 is designed to examine the impact of ancillary data on $PM_{2.5}$ modeling. Model-3 aims to assess the contribution of MERRA-2 data and incorporates all available features, being used for reconstructing national ground-level $PM_{2.5}$ concentrations. Model-4, which excludes MINTS data, investigates the influence of additional in situ observations. Models 5 and 6 focus specifically on the effects of including MINTS $PM_{2.5}$ data, reflecting the limited duration of MINTS data availability and the geographical limitation of MINTS observation sites to Texas. All models use ECMWF meteorological variables, GOES-16 AOD data as basic features, and target $PM_{2.5}$ values from EPA and OpenAQ, with variations in the inclusion of features between different models.

Table 4. $PM_{2.5}$ model categories. The first four models are designed for $PM_{2.5}$ modeling across the entire United States, while the last two models specifically target the Texas region. The distinction among these models lies in the incorporation of ancillary data, MERRA-2 data, and MINTS $PM_{2.5}$ data.

Model	Spatial Coverage	Time Span	Ancillary	MERRA-2	MINTS
1	US	Jan 2020 - Jun 2023			✓
2	US	Jan 2020 - Jun 2023	✓		✓
3	US	Jan 2020 - Jun 2023	✓	✓	✓
4	US	Jan 2020 - Jun 2023	✓	✓	
5	TX	Sep 2021 - Jun 2023	✓	✓	✓
6	TX	Sep 2021 - Jun 2023	✓	✓	

The datasets for Models 1, 2, and 3 contain 1,521,790 entries, while Model-4 has 1,512,889 entries. Models 5 and 6 have significantly fewer entries, with 61,889 and 52,988 entries, respectively, due to the restricted geographic scope to Texas and the shorter data period. These data sets are divided into training and testing sets with a ratio of 90% to 10%, a common practice for training and evaluating machine learning models. The models are trained using a tree-based machine learning approach, optimized with selected hyperparameters. The performance of these models is then validated in the testing set, using metrics such as the root mean square error (RMSE) and the Correlation Coefficient (R) to evaluate accuracy.

 doi:10.20944/preprints202405.1685.v1

3.3. Machine Learning Approaches

The machine learning approach is particularly well suited for studies like this for several reasons. First, PM concentrations are affected by a wide array of factors, including those beyond the scope of this study. Secondly, there is a notable absence of theoretical models capable of accurately depicting the relationships between various variables and PM concentrations. Lastly, this study relies on a substantial data set with numerous variables, and machine learning algorithms excel at managing complex data sets that traditional data analysis methods might find challenging.

Although different machine learning models, including neural networks and XGBoost, can be applied to PM modeling, tree-based methods like random forests offer distinct advantages. For example, tree-based models tend to perform more efficiently with large datasets. Furthermore, ensemble machine learning techniques, which combine multiple weak learners into a robust model, are particularly effective in minimizing bias and variance, offering a clear understanding of how each variable contributes to the prediction of the model [11].

In this study, the Extra Tree (ET) regression algorithm, an enhancement of the Random Forest algorithm, was chosen for modeling $PM_{2.5}$. The ET model has been shown to be effective for $PM_{2.5}$ modeling using AOD and meteorological variables in previous research [11,60]. It constructs numerous decision trees, each trained on a randomly selected subset of features and data samples, introducing additional randomness into the model. This not only speeds up the training process but also makes the model less prone to overfitting from noisy data.

4. Results

4.1. MINTS all PM size fraction modeling

In this section, we specifically focus on the use of data only from the MINTS sensing system. The modeling efforts are categorized into three main groups, each defined by a unique set of feature variables. Additionally, each main group is further divided into seven subcategories, targeting different PM size fractions.

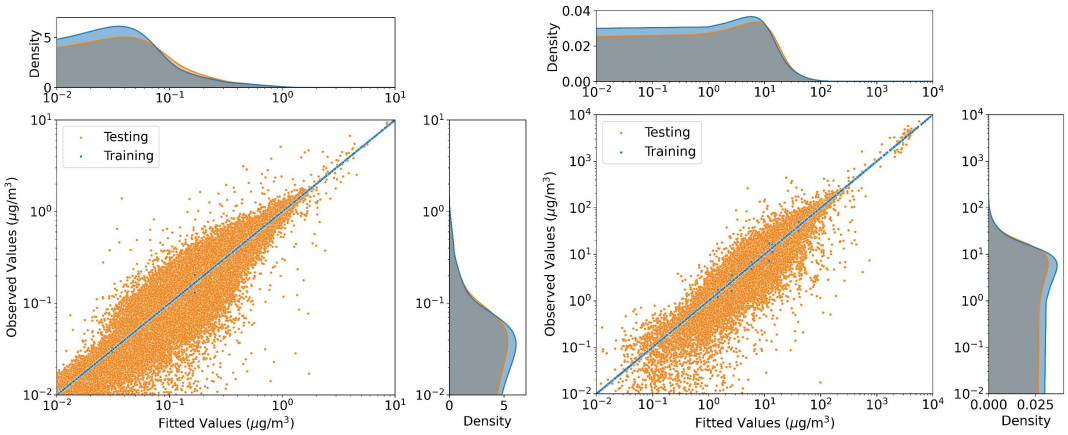
Of these main groups, Group-2, which utilizes all the features available from the MINTS system, shows the highest correlation coefficients (R values) in the test data compared to the other groups (Table 5). Within Group-2, the variation in R values between subcategories is relatively minor. In particular, when using just three meteorological variables (temperature, pressure, and humidity) in Group-1, the models show impressively high performance on the test data, with R values reaching around 0.92. Group-3, designed to explore the effect of light intensity from various frequency channels on different PM size fractions, found that models for $PM_{0.1}$, relying solely on light intensity data, produced higher R values on the test data than those for other PM size fractions within the same group.

Scatter plots were created to illustrate the correlation between predicted and actual PM levels for all specified groups and across different PM size categories. This paper selectively features the most illustrative scatter plots for visual analysis. Figure 3 shows the scatter plots for the smallest ($PM_{0.1}$) and largest ($PM_{10.0}$) PM size fractions within Group-2, which showed superior performance compared to the other groups. Additionally, Figure 4 shows plots depicting the relative importance of various features in the models analyzed. These graphs clearly demonstrate that carbon dioxide, pressure, temperature, and humidity are crucial factors for both $PM_{0.1}$ and $PM_{10.0}$ sizes. Furthermore, for the smallest particles ($PM_{0.1}$), light intensities in the ultraviolet A and B spectrum play a vital role. In contrast, for the larger particles ($PM_{10.0}$), light intensities in the violet and full spectrum ranges make significant contributions to the predictive accuracy of the models.

 doi:10.20944/preprints202405.1685.v1

Table 5. Three main groups are sub-categorized on PM size fractions. The respective evaluation results for all the sub-categories are presented.

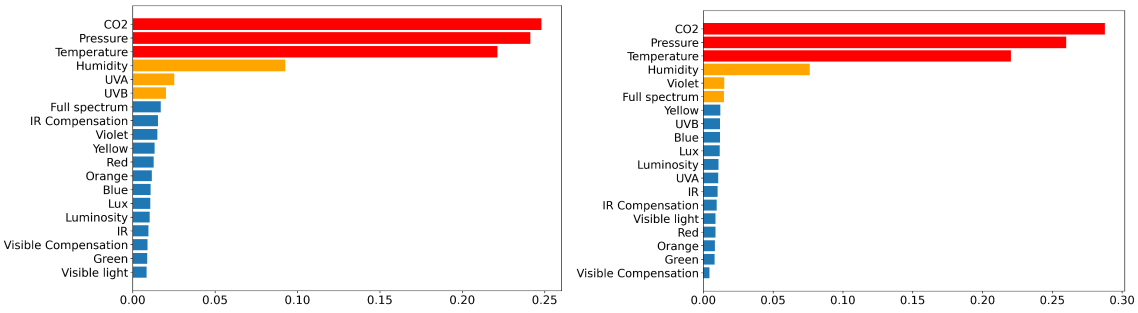
Group	PM	Sample size	Train R	Train RMSE	Test R	Test RMSE
1	$PM_{0.1}$	616,301	0.999	0.016	0.914	0.152
	$PM_{0.3}$	616,866	1.0	0.923	0.923	18.953
	$PM_{0.5}$	617,760	1.0	1.138	0.911	22.277
	$PM_{1.0}$	617,765	1.0	1.202	0.937	19.151
	$PM_{2.5}$	617,767	1.0	1.976	0.923	26.273
	$PM_{5.0}$	617,771	1.0	2.276	0.932	30.352
	$PM_{10.0}$	617,771	1.0	2.304	0.933	31.165
2	$PM_{0.1}$	616,301	1.0	0.0	0.978	0.077
	$PM_{0.3}$	616,866	1.0	0.003	0.978	10.545
	$PM_{0.5}$	617,760	1.0	0.006	0.977	11.576
	$PM_{1.0}$	617,765	1.0	0.003	0.978	11.376
	$PM_{2.5}$	617,767	1.0	0.019	0.973	15.747
	$PM_{5.0}$	617,771	1.0	0.021	0.979	17.528
	$PM_{10.0}$	617,771	1.0	0.021	0.978	18.273
3	$PM_{0.1}$	616,301	0.707	0.274	0.312	0.36
	$PM_{0.3}$	616,866	0.571	40.509	0.044	50.633
	$PM_{0.5}$	617,760	0.597	42.575	0.053	55.74
	$PM_{1.0}$	617,765	0.609	44.09	0.063	56.271
	$PM_{2.5}$	617,767	0.648	54.793	0.11	69.386
	$PM_{5.0}$	617,771	0.617	69.17	0.095	84.653
	$PM_{10.0}$	617,771	0.608	72.213	0.091	87.307



(a) **(b)**

Figure 3. Scatter diagrams depicting the training and testing datasets for Group-2 (incorporate all the feature variables within MINTS system): (a) $PM_{0.1}$. (b) $PM_{10.0}$.

 doi:10.20944/preprints202405.1685.v1

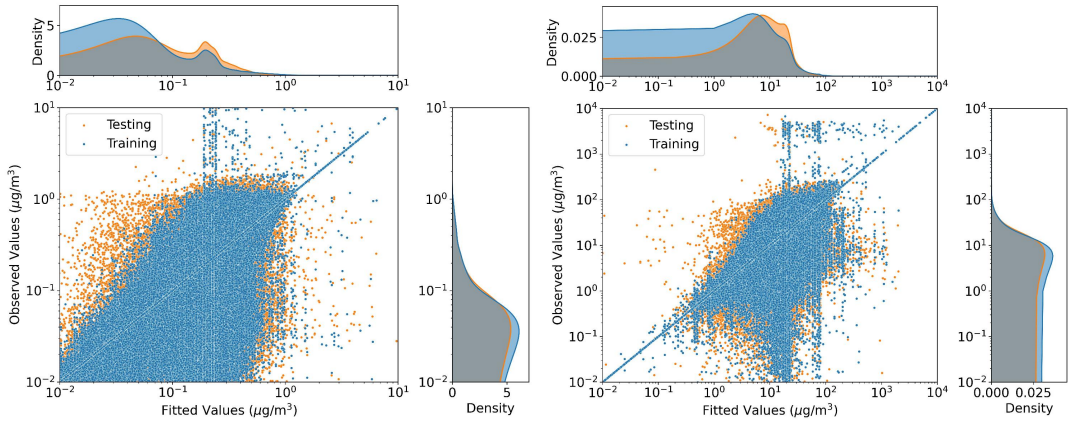


(a) **(b)**

Figure 4. Importance of features for Group-2 (incorporate all the feature variables within MINTS system): (a) $PM_{0.1}$. (b) $PM_{10.0}$.

Figures 5 and 6 illustrate the scatter and feature importance plots for $PM_{0.1}$ and $PM_{10.0}$, focusing on Group-3 (incorporate only light sensing variables within MINTS system). These plots are instrumental in highlighting the light intensity frequency ranges that significantly impact model development, clearly differentiating between the sizes of particles.

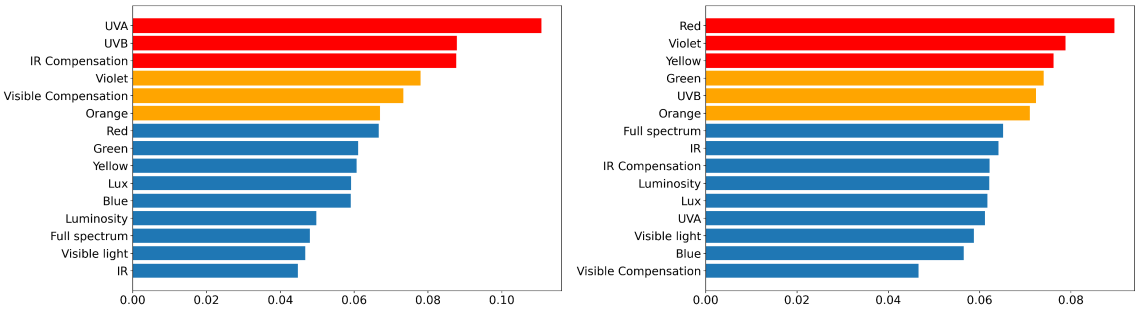
Consistent with the size-dependent light scattering properties of aerosols, our analysis reveals that for fine particle modeling ($PM_{0.1}$), light intensities in the ultraviolet A and B frequency ranges contain valuable information. On the other hand, for the larger particle size ($PM_{10.0}$), light intensities in the red and violet frequency ranges play a more critical role in the construction of predictive models. This clarification of the importance of the features provides insight into the unique characteristics and variables useful for modeling each PM size fraction.



(a) **(b)**

Figure 5. Scatter plots depicting the training and testing data for Group-3 (incorporate only light sensing variables within MINTS system): (a) $PM_{0.1}$ and (b) $PM_{10.0}$.

 doi:10.20944/preprints202405.1685.v1



(a) **(b)**

Figure 6. Importance of features for Group-3 (incorporate only light sensing variables within MINTS system): (a) $PM_{0.1}$. (b) $PM_{10.0}$.

4.2. Complimentary In-Situ and Remote Sensing $PM_{2.5}$ modeling

This section looks at the creation of four national $PM_{2.5}$ estimation models, each notable for its high temporal resolution and distinguished by different target variables and $PM_{2.5}$ observation sources. Additionally, two regional $PM_{2.5}$ models were developed, categorized based on the observation sources used. The purpose of classifying these regional models is to demonstrate the benefits of improving PM estimation models with additional ground-based observations and to evaluate the effectiveness of incorporating MINTS data.

The national data set includes a comprehensive collection of approximately 1,521,790 observations and 53 predictor variables. The regional dataset contains about 61,889 observations with the same set of feature variables, all employed in the model training and testing phases. The data was split into training and testing segments in a 90:10 ratio. Training data were used for model fitting, with the performance of the models evaluated in both data sets. Table 6 offers a detailed examination of essential evaluation metrics, such as the correlation coefficients between actual observations and the predictions made by machine learning, model R scores, and root mean square error (RMSE) figures, all based on test data. These metrics collectively facilitate an evaluation of the models' accuracy and predictive capability.

Table 6. Model categories as well as their corresponding evaluation result are listed.

Model	Sample size	Train R	Train RMSE	Test R	Test RMSE
1	1,521,790	0.998	0.388	0.793	3.673
2	1,521,790	0.998	0.388	0.816	3.501
3	1,521,790	0.998	0.388	0.849	3.201
4	1,512,889	0.998	0.392	0.834	3.364
5	61,889	0.998	0.527	0.872	4.474
6	52,988	0.997	0.565	0.816	4.253

The base model, referred to as Model-1, utilizes $PM_{2.5}$ data collected from a variety of sources, including the Environmental Protection Agency (EPA), OpenAQ, and the MINTS-AI environmental sensing system. This initial model relies exclusively on ECMWF meteorological data and Aerosol Optical Depth (AOD) feature variables from the GOES-16 satellite, achieving a correlation coefficient (R) of 0.793. The introduction of additional data to the base model leads to an improvement in the R-value, which climbs from 0.793 to 0.816. Following this, Model-3, which integrates both supplementary data and MERRA-2 data, reaches an R value of 0.849, indicating a further improvement in model performance. In contrast, removing the MINTS-AI environmental sensing data from Model-3 results in a decrease in the R value to 0.834. Importantly, incorporating MINTS data into the regional model, identified as Model-5, significantly improves the model performance, demonstrating the valuable impact of the MINTS data on the accuracy of $PM_{2.5}$ estimations.

 doi:10.20944/preprints202405.1685.v1

The scatter diagram comparing measured versus estimated values for Model-3 (seen in Figure 7) visually demonstrates the correlation between actual (measured) and predicted (estimated) values for a specific target variable. This plot is instrumental in pinpointing the strengths of the model and areas that need refinement, thus serving as a crucial tool for assessing model performance and identifying potential enhancements. To aid in the analysis of overlapping data points, marginal histograms are incorporated into the figure. Furthermore, the importance ranking of the predictors (shown in Figure 8) is designed to highlight the contribution of each variable to Model-3's predictive capability. Variables ranked with higher importance scores exert a more substantial influence on the model predictions. In particular, the most critical variables, according to the feature importance chart, include Aerosol Optical Depth (AOD) analysis (utilizing AOD data from MERRA-2), specific humidity, AOD from GOES-16, dew point temperature, carbon monoxide and carbon dioxide.

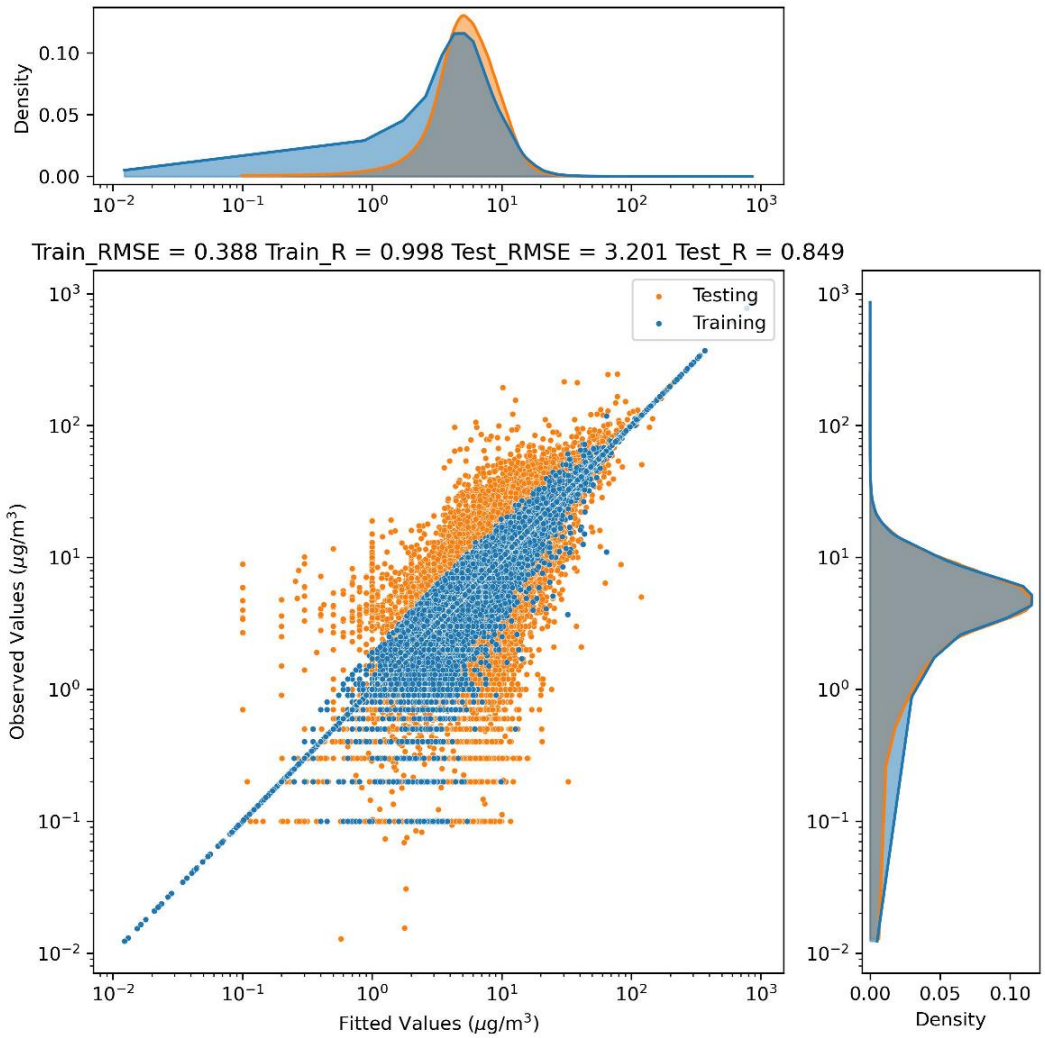


Figure 7. Scatter plots depicting training and testing data in log scale, accompanied by marginal probability density functions, illustrating the analysis conducted for Model-3.

 doi:10.20944/preprints202405.1685.v1

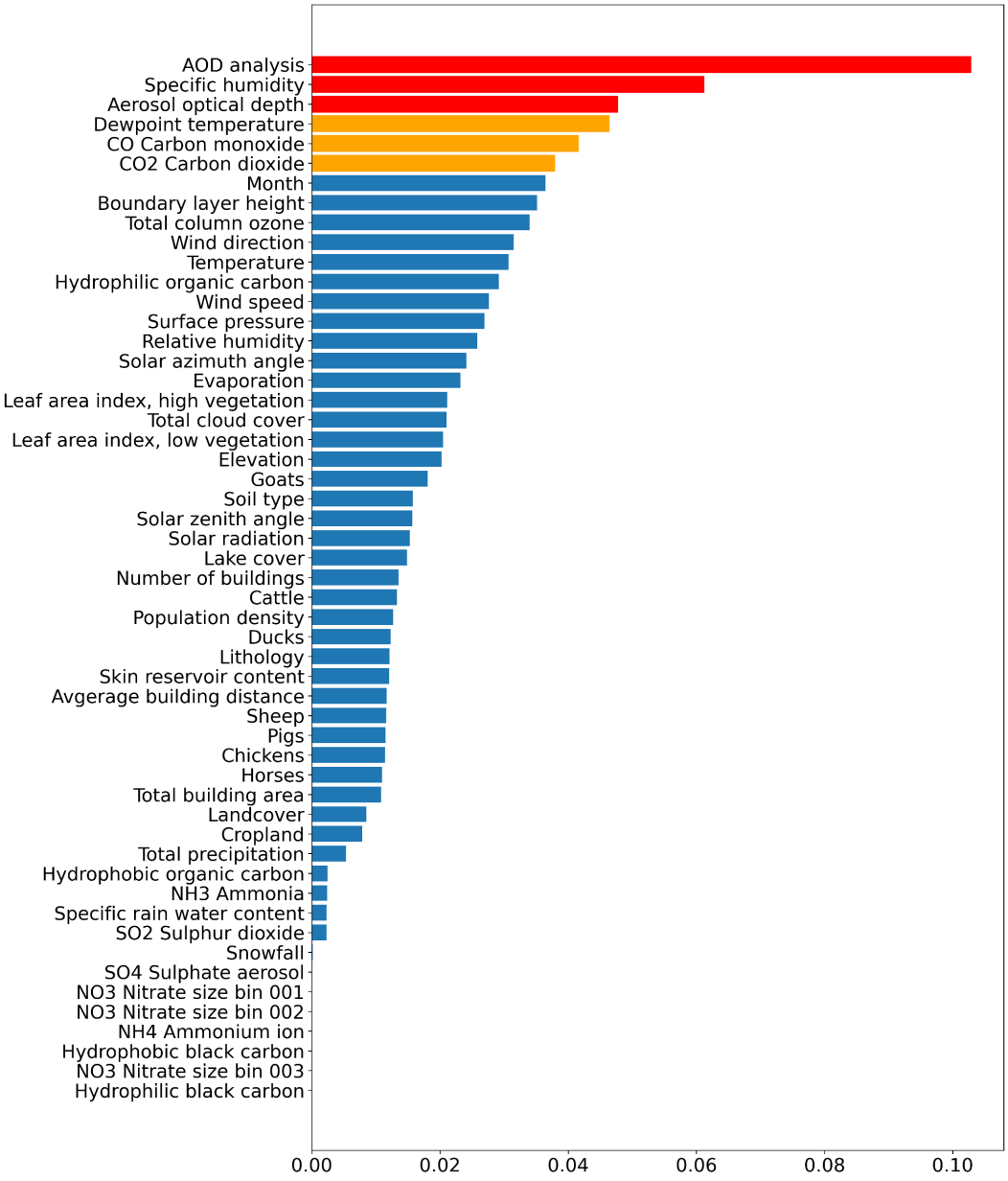


Figure 8. Feature importance score for Model-3.

4.3. Nationwide $PM_{2.5}$ model validation

Model-3, which incorporates all available features and $PM_{2.5}$ data sources, stands out for its exceptional performance in mapping ground-level $PM_{2.5}$ concentrations throughout the United States. The detail and precision of this $PM_{2.5}$ mapping are influenced by the resolution of the remote sensing data employed. To ensure uniformity in all ground-level $PM_{2.5}$ concentration maps, the ECMWF meteorological data grid, which measures approximately 10 km x 10 km, is used as the standard coordinate framework. However, when using data from different sources, which may follow various coordinate systems, it becomes necessary to align them with the standard grid using linear interpolation to ensure consistency.

Wildfires significantly contribute to the increase and change in the composition of airborne particulate matter, including both primary and secondary pollutants, which can affect human health and the environment. Large wildfire events in the United States have been linked to specific weather conditions, such as droughts, high temperatures, low humidity, and strong winds, which are conducive to the ignition and propagation of wildfires. Figure 9 illustrates the $PM_{2.5}$ concentrations on the ground

 doi:10.20944/preprints202405.1685.v1

as estimated by Model-3 during one of the most significant wildfire events in the US, the Santa Clara Unit (SCU) Lightning Complex fire in California in 2020. This fire, sparked by dry lightning on August 16, was eventually contained in early October.

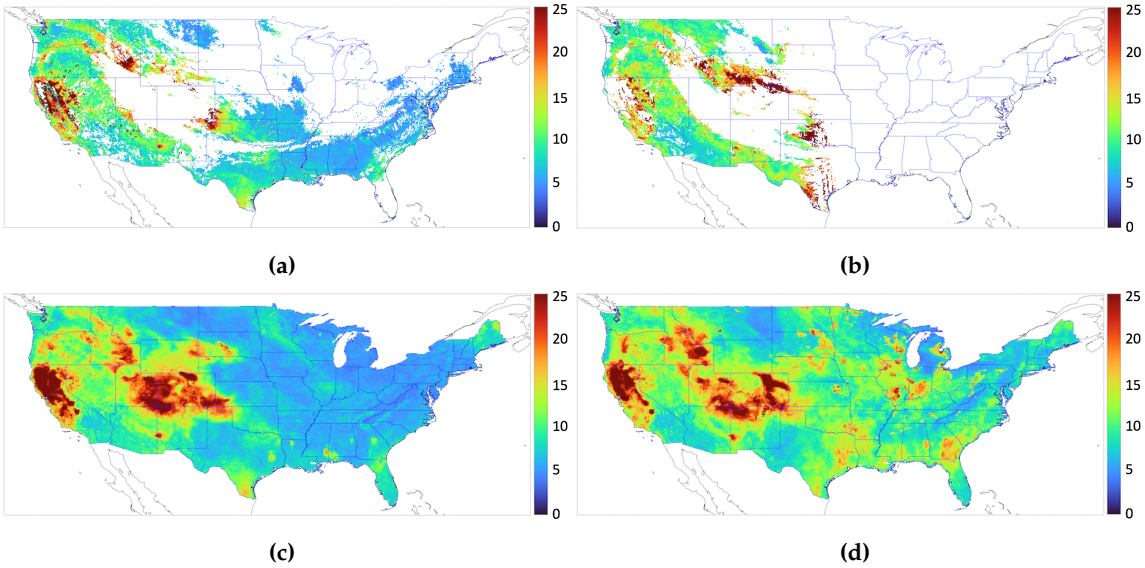
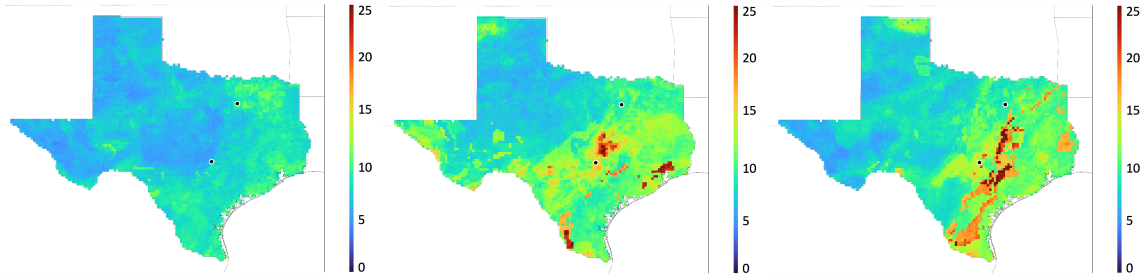


Figure 9. $PM_{2.5}$ reconstruction during the Santa Clara Unit (SCU) Lightning Complex fire in 2020. Panels (a) and (b) are for 9 PM UTC and midnight on October 2, respectively, using a specialized version of Model-2 that exclusively incorporates AOD data from GOES-16. Panels (c) and (d) are for the same times but using the original Model-3. Areas with $PM_{2.5}$ concentrations exceeding the 25 $\mu g/m^3$ threshold are highlighted in red.

Figures 9a and 9b offer visual insights into the ground-level $PM_{2.5}$ concentrations recorded at two different times: 9 PM and midnight on October 2, 2023. These visualizations were produced using a modified version of Model-3, specifically trained without incorporating MERRA-2 Aerosol Optical Depth (AOD) data. On the other hand, Figures 9c and 9d depict the $PM_{2.5}$ concentrations at the same times, but were generated using the original version of Model-3, which includes a comprehensive set of feature parameters. Both variations of the model successfully identified areas of high $PM_{2.5}$ concentrations in California, with the pollution spreading to the northeast over the three-hour interval. In particular, the specialized version of Model-3 encounters limitations due to the absence of GOES-16 AOD data in areas covered by clouds, resulting in gaps in the $PM_{2.5}$ concentration estimates. To overcome these limitations, the original Model-3 supplements missing GOES-16 AOD observations with MERRA-2 AOD data, ensuring a more detailed portrayal of $PM_{2.5}$ concentrations throughout the region. The chosen color scale adheres to the guidelines of the World Health Organization (WHO), setting the threshold at 25 $\mu g/m^3$ for the annual mean concentration of $PM_{2.5}$, beyond which there is a significant risk to health. This threshold is used as the upper limit to visualize the map data, in accordance with global health standards.

The coverage of the MINTS sensing system is limited to the north Texas region. To comprehensively evaluate the performance of the model in $PM_{2.5}$ reconstruction, our analysis focuses exclusively on results within the state of Texas. Specifically, we scrutinize data from three distinct timestamps on January 1, 2023, comparing them with $PM_{2.5}$ observations collected by two MINTS in situ sites located in Joppa and Austin, represented by solid black circles on the maps in Figure 10. This figure visually presents the $PM_{2.5}$ reconstruction results generated by Model-3 at these three timestamps, each separated by a minimum interval of 11 hours. Similarly, Figure 11 provides a time series illustrating $PM_{2.5}$ observations recorded by the ground sensors of the two MINTS in the cities of Joppa (blue) and Austin (orange).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2024 doi:10.20944/preprints202405.1685.v1



(a) (b) (c)

Figure 10. Reconstructed $PM_{2.5}$ concentrations across the Texas region at three distinct timestamps on January 1, 2023, in UTC. The black solid circle in the north corresponds to the MINTS ground sensor located in Joppa (south Dallas), while the black solid circle in the south represents the MINTS ground sensor located in Austin. The subfigures depict the following timestamps: (a) 2023 January 1 at 01:00 AM UTC, (b) 2023 January 1 at 02:00 PM UTC, and (c) 2023 January 2 at 01:00 AM UTC.

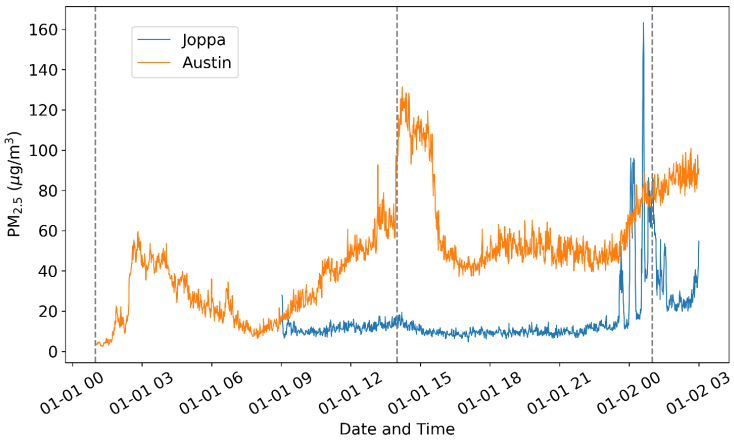


Figure 11. $PM_{2.5}$ measurements obtained from two MINTS in-situ sensors located in Joppa (depicted in blue) and Austin (shown in orange). The timestamps indicated by the gray dashed lines align with those presented in Figure 10.

In particular, the three gray dashed lines in Figure 11 correspond to the timestamps of the $PM_{2.5}$ reconstruction maps shown in Figure 10. Specifically, Figure 10a depicts a relatively less polluted environment at both locations around 7 PM Central Time on December 31, 2022 (equivalent to January 1, 2023, at 01:00 UTC). This finding aligns with similar observations of lower pollution concentrations made by the Austin MINTS ground sensor at the same time (corresponding to the first gray dashed line). Approximately 13 hours later, the model captures elevated $PM_{2.5}$ concentrations near Austin, while concentrations in the Joppa area remain lower (Figure 10b). This pattern closely mirrors the observations recorded by the two MINTS ground sensors, with high $PM_{2.5}$ concentrations observed in Austin and lower levels in Joppa. In a subsequent timeframe, approximately 24 hours after the initial observation, the model indicates an expansion of higher $PM_{2.5}$ concentrations, particularly in the Joppa area (Figure 10c). This trend is aligned with the simultaneous observation of higher concentrations by both MINTS ground sensors at both locations.

4.4. Time fraction of $PM_{2.5}$ concentration exceed thresholds in 2022

Since 2000, there has been a notable 42% decrease in overall $PM_{2.5}$ levels in the United States, attributed to the implementation of clean air regulations. Despite this progress, there remains concern about the need for further reductions. In February 2024, responding to these concerns, the Environmental Protection Agency (EPA) revised the national standards of ambient air quality for PM. Specifically, the annual primary $PM_{2.5}$ standard was revised downward from 12 $\mu g/m^3$ to 9 $\mu g/m^3$, aiming to mitigate the adverse health impacts and associated costs. The EPA estimates that

 doi:10.20944/preprints202405.1685.v1

adhering to this new standard could lead to potential savings of up to $46 billion in avoided healthcare and hospitalization costs by 2032 [67,68].

In this section, we used our Model-3 machine learning to estimate hourly $PM_{2.5}$ concentrations across the entire United States for the year 2022. The resulting data set allows us to calculate the fraction of time during which $PM_{2.5}$ concentrations exceeded five distinct threshold levels (8 $\mu g/m^3$, 9 $\mu g/m^3$, 10 $\mu g/m^3$, 11 $\mu g/m^3$, and 12 $\mu g/m^3$) throughout the entirety of 2022. The accompanying figure illustrates maps showing the percentage of time that $PM_{2.5}$ concentrations exceeded the specified threshold levels, with color-coded representations corresponding to the percentage values.

As shown in Figure 12a, certain areas in the eastern United States and California exhibit elevated percentage values, indicating that these regions experienced $PM_{2.5}$ concentrations exceeding the threshold of 12 $\mu g/m^3$ for more than 20% of the time throughout the year 2022. However, Figure 12d illustrates that the entire United States shows elevated percentage values, suggesting that the entire nation encountered $PM_{2.5}$ concentrations exceeding the threshold of 9 $\mu g/m^3$ for more than 20% of the time in 2022. In particular, the eastern United States and California regions sustained $PM_{2.5}$ concentrations that exceeded the threshold of 9 $\mu g/m^3$ for more than 50% of the time during the same period. These estimates underscore the importance of regulatory measures aimed at maintaining annual $PM_{2.5}$ concentrations below 9 $\mu g/m^3$.

 doi:10.20944/preprints202405.1685.v1

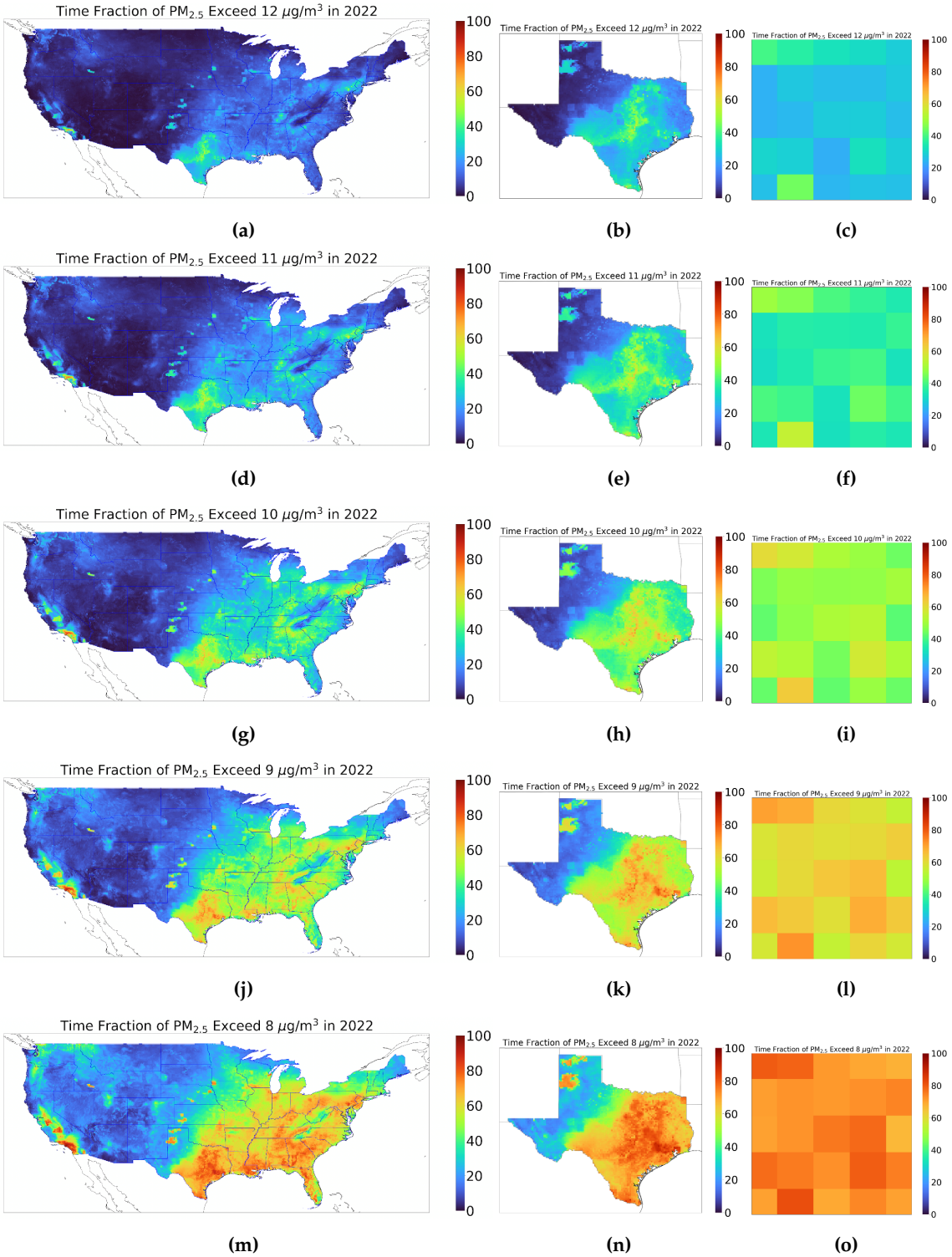


Figure 12. Percentage of time exceeding $PM_{2.5}$ concentration thresholds throughout the entirety of 2022, as estimated by Model-3. The subfigures (a), (d), (g), (j) and (m) illustrate the duration exceeding thresholds 12 $\mu g/m^3$, 11 $\mu g/m^3$, 10 $\mu g/m^3$, 9 $\mu g/m^3$, and 8 $\mu g/m^3$ over the US, respectively. The subfigures (b), (e), (h), (k) and (n) illustrate the corresponding $PM_{2.5}$ exceeding in Texas regions and the subfigures (c), (f), (i), (l) and (o) illustrate the corresponding $PM_{2.5}$ exceeding in Dallas regions, respectively

 doi:10.20944/preprints202405.1685.v1

5. Conclusions

Environmental agencies often depend on a small set of airborne particulate monitoring stations, which are often unevenly spread out, leading to low temporal resolution in PM observations. These inherent constraints limit the precision of PM modeling due to the significant variability in PM concentrations at fine scales and over time. To address these issues, the UTD MINTS-AI platform has implemented a specialized environmental monitoring network tailored for use in local communities in Texas. This network is specifically designed to gather PM data, along with relevant environmental variables, with high temporal resolution and fine spatial detail.

In this paper, we have concentrated on two distinct studies related to PM modeling. In the first study, we underscored the significance of raw data collection within a synchronized temporal and spatial coordinate system for effective PM modeling. In the second study, we enhanced $PM_{2.5}$ modeling by employing an asynchronized temporal and spatial coordinate system, leveraging pertinent remote sensing data.

In the first study, to achieved underscore the significance of a synchronized temporal and spatial coordinate system, we exclusively utilized data only from the MINTS sensing system recorded between September 2021 and June 2023. This restricted data collection to the MINTS sensing system was intentional, as it allows access to both PM data and other pertinent environmental data at precisely the same location with synchronized time stamps. The decision to utilize the extra tree regression model, based on its strong performance in prior research and efficient computational processing, proved successful in tackling these challenges. Modeling activities were categorized based on environmental factors, incorporating all available feature variables (all available variables from the embedded sensors within MINTS system) that exhibited superior performance across different PM size fractions. Specifically, variables such as carbon dioxide, pressure, temperature, and humidity emerged as the most influential during the modeling phase. Moreover, it was discovered that high-frequency band light intensities played a secondary role in modeling fine PM sizes, whereas low-frequency band light intensities had a more significant impact on modeling larger PM sizes. It is noteworthy that the modeling of the fine PM size fraction ($PM_{0.1}$) resulted in higher correlation coefficient (R) values compared to coarser PM size fractions in Group-3, which relied solely on the light intensity variables. This result indicates that, for smaller particle sizes, Mie scattering can be beneficial in accurately capturing specific particle characteristics. This can be attributed to the fact that the diameter of $PM_{0.1}$ particles falls within the ultraviolet wavelength range, which improves the model's capability to capture finer details of PM concentrations. Importantly, when a model is built solely on light intensity data from different frequency bands, it becomes clear that variations in the fine PM size fraction can be effectively captured by high-frequency band intensities.

It is important to highlight that using only three environmental factors, namely temperature, pressure, and humidity, has been proven to be effective in modeling various PM size fractions with high performance, as evidenced by high R values, as long as the data were collected in a synchronized temporal and spatial coordinate system. This effectiveness can be attributed to the advantage of having data collected at the exact geographical location where PM observations are made. This means that all data are gathered at the same coordinates with synchronized timestamps, eliminating the need for data alignment or interpolation, which are crucial in PM modeling. Additionally, the data is captured at a high temporal resolution, allowing for a comprehensive representation of PM variations and related changes in feature variables. Importantly, the timestamps for different variables are closely synchronized, reducing the introduction of noise that often occurs during data alignment processes. This synchronization enhances the model's capability to detect subtle nuances in PM fluctuations. However, it is crucial to recognize that such ideal circumstances are often unattainable in real-world situations. When modeling PM that involves integrating environmental data from different sources, requiring spatial and temporal data alignment, a more extensive set of environmental factors is typically needed to achieve satisfactory model performance. This was demonstrated in the second study, where PM2.5 modeling incorporated complementary in-situ and remote sensing approaches.

 doi:10.20944/preprints202405.1685.v1

The development of nationwide $PM_{2.5}$ models in the second study, a diverse array of predictor variables was harnessed. This included high-temporal AOD data derived from the GOES-16 geostationary satellite, meteorological variables sourced from the ECMWF, ancillary data gathered from various external sources, location-specific solar angles, and reanalysis data related to AOD and air pollutant gases, obtained from the MERRA-2 database. The model training process was stratified into categories based on the inclusion of feature variables and the sources of ground observations of PM. As noted above, these variables originate from disparate sources, each characterized by distinct coordinate systems and temporal resolutions. To align these datasets, a linear interpolation method was applied, albeit with noticeable consequences on model performance. Interestingly, the model that incorporated all available feature parameters and utilized data from all sources of PM observation exhibited the most favorable performance, particularly in terms of R values, in the context of the nationwide $PM_{2.5}$ modeling. In particular, among the most influential variables that contributed to this performance were AOD, specific humidity, dew point temperature, carbon monoxide, and carbon dioxide.

Based on the comparative analysis of models, it becomes evident that the inclusion of auxiliary and MERRA-2 data as supplementary feature variables improves the accuracy of the model, as reflected in higher R values. This augmentation helps to better discern variations in $PM_{2.5}$ concentrations with respect to both temporal and spatial dimensions. Furthermore, the integration of environmental sensing data from the MINTS-AI platform, although limited to a small number of sites within the Texas region, has a positive impact on the precision of nationwide $PM_{2.5}$ models. These findings underscore the potential advantages of incorporating additional ground-based observations and their associated data into PM modeling, as they contribute to improved model accuracy.

Although the increase in the R value for the national model resulting from the integration of MINTS environmental sensing data may not be substantial, due to the limited number of MINTS sites located primarily in Texas, there is a discernible enhancement in regional models with the inclusion of MINTS data. This observation suggests that $PM_{2.5}$ exhibits intricate variations on a very fine spatial scale. To capture more nuanced features or to achieve highly accurate $PM_{2.5}$ estimates, it is imperative to expand the network of ground sensing systems, ensuring an even distribution in a broader geographical area.

Using our analysis approach to reconstruct the fine-time resolution $PM_{2.5}$ distribution across the entire united states for our study period, we found that the entire nation encountered $PM_{2.5}$ levels that exceeded 9 $\mu g/m^3$ for more than 20% of the time of our analysis period, with the eastern United States and California experiencing concentrations exceeding 9 $\mu g/m^3$ for over 50% of the time, highlighting the importance of regulatory efforts to maintain annual $PM_{2.5}$ concentrations below 9 $\mu g/m^3$.

Funding: This research was funded by the following grants: Support from the University of Texas at Dallas Office of Sponsored Programs, Dean of Natural Sciences and Mathematics, and Chair of the Physics Department is gratefully acknowledged. TRECIS CC* Cyberteam (NSF 2019135), NSF OAC-2115094 Award, and EPA P3 grant number 84057001-0.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors acknowledge the OIT-Cyberinfrastructure Research Computing group at the University of Texas at Dallas and the TRECIS CC* Cyberteam (NSF 2019135) for providing HPC resources that contributed to this research, NSF OAC-2115094 Award, and EPA P3 grant number 84057001-0.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Boucher, O. Atmospheric Aerosols: Properties and Climate Impacts; Springer Netherlands: 2015. ISBN 978-9401796484.

 doi:10.20944/preprints202405.1685.v1

2. Chen, R., Li, Y., Ma, Y., Pan, G., Zeng, G., Xu, X., Chen, B., Kan, H. Coarse particles and mortality in three Chinese cities: the China Air Pollution and Health Effects Study (CAPES). *Science of the Total Environment* **2011**, *409* (23), 4934–4938.
3. D. J. Lary, F. S. Faruque, N. Malakar, A. Moore, B. Roscoe, Z. L. Adams, Y. Eggelston. Estimating the global abundance of ground level presence of particulate matter ($PM_{2.5}$). *Geospatial Health* **2014**, *8*(3), 611–630.
4. C. Arden Pope III, Richard T. Burnett, Michael J. Thun, Eugenia E. Calle, Daniel Krewski, Kazuhiko Ito, George D. Thurston. *Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution*. *Jama*, 2002, 287 (9), 1132–1141.
5. Dubovik, O., Holben, B., Eck, T. F., Smirnov, A., Kaufman, Y. J., King, M. D., Tanré, D., Slutsker, I. Variability of absorption and optical properties of key aerosol types observed in worldwide locations. *Journal of the Atmospheric Sciences* **2002**, *59* (3), 590–608.
6. Charlson, R. J., Schwartz, S. E., Hales, J. M., Cess, R. D., Coakley, Jr. J. A., Hansen, J. E., Hofmann, D. J. Climate forcing by anthropogenic aerosols. *Science* **1992**, *255* (5043), 423–430.
7. Pöschl, U. Atmospheric aerosols: composition, transformation, climate and health effects. *Angewandte Chemie International Edition* **2005**, *44* (46), 7520–7540.
8. National Research Council and others. *A Plan for a Research Program on Aerosol Radiative Forcing and Climate Change*. National Academies Press, 1996.
9. Chin, M. *Atmospheric Aerosol Properties and Climate Impacts*. DIANE Publishing Company, 2009. ISBN 978-1437912616. https://books.google.com/books?id=IgJZXXgtHmQC
10. X. Yu, D. J. Lary, C. S. Simmons, L. O. H. Wijeratne. High Spatial-Temporal $PM_{2.5}$ Modeling Utilizing Next Generation Weather Radar (NEXRAD) as a Supplementary Weather Source. *Remote Sensing* **2022**, *14* (3), 495.
11. X. Yu, D. J. Lary, C. S. Simmons. $PM_{2.5}$ Modeling and Historical Reconstruction over the Continental USA Utilizing GOES-16 AOD. *MDPI Remote Sensing* **2021**, *13*(23), 4788.
12. L. O. H. Wijeratne. *Coupling Physical Measurement With Machine Learning for Holistic Environmental Sensing*. PhD thesis, The University of Texas at Dallas, 2021.
13. D. J. Lary, T. Lary, B. Sattler. *Using Machine Learning to Estimate Global $PM_{2.5}$ for Environmental Health Studies*. *Environmental Health Insights*, 2015, 1, doi: 10.4137/EHI.S15664, 41–52.
14. Lakitha O.H. Wijeratne, Daniel R. Kiv, Adam R. Aker, Shawhin Talebi, David J. Lary. *Using Machine Learning for the Calibration of Airborne Particulate Sensors*. *Sensors*, 2019, 20 (99).
15. H. Zhang, R. M. Hoff, J. A. Engel-Cox. The Relation between Moderate Resolution Imaging Spectroradiometer (MODIS) Aerosol Optical Depth and $PM_{2.5}$ over the United States: A Geographical Comparison by U.S. Environmental Protection Agency Regions. *The Air & Waste Management Association* **2009**, *59* (11), 1358–1369.
16. C. Zheng, C. Zhao, Y. Zhu, Y. Wang, X. Shi, X. Wu, T. Chen, F. Wu, Y. Qiu. *Analysis of influential factors for the relationship between $PM_{2.5}$ and AOD in Beijing*. *Atmos. Chem. Phys.*, 2017, 17, 13473–13489.
17. Z. Zang, D. Li, Y. Guo, W. Shi, X. Yan. *Superior $PM_{2.5}$ Estimation by Integrating Aerosol Fine Mode Data from the Himawari-8 Satellite in Deep and Classical Machine Learning Models*. *Remote Sens.*, 2021, 13 (2779).
18. Xiaohe Yu, David J. Lary, Christopher S. Simmons. *$PM_{2.5}$ Modeling and Historical Reconstruction over the Continental USA Utilizing GOES-16 AOD*. *Remote Sensing*, 2021, 13 (23), MDPI.
19. Harrison, W. A. *In-situ Observation of Atmospheric Particulates*. The University of Texas at Dallas, 2015.
20. Shawhin Talebi. *Physical Quantification of the Interactions Between Environment, Physiology, and Human Performance*. PhD thesis, The University of Texas at Dallas, 2022.
21. Piera Systems. *IPS Series Sensor*. Piera Systems Inc, 2022.
22. United States Environment Protection Agency EPA. *Air Quality System (AQS) API*. 2020. https://aqs.epa.gov/aqsweb/documents/data_api.html
23. *Air Quality System (AQS) Data API*. U.S. Environmental Protection Agency (EPA). n.d. [Accessed: March 26, 2023]. https://aqs.epa.gov/aqsweb/documents/data_api.html
24. *OpenAQ - About*. OpenAQ. n.d. [Accessed: March 26, 2023]. https://openaq.org/about/
25. FE Volz, TH Lee, TF LaPenta, JD Spinhirne, GB Hulley, JJ O'Brien. *Geostationary operational environmental satellite system–R (GOES-R)*. *Bulletin of the American Meteorological Society*, 2000, 81 (10), 2345–2363.
26. Timothy J. Schmit, Paul Griffith, Matthew M. Gunshor, Jay Daniels, Steve Goodman, William Lebair, Doug Lindholm, Steve Miller, Scott Rudlosky, Dan Miller. *Introducing the Next-Generation Advanced Baseline Imager on GOES-R*. *Bulletin of the American Meteorological Society*, 2017, 98 (4), 681–698.

 doi:10.20944/preprints202405.1685.v1

27. Anthony J. Mannucci, Philip W. Stephens, Liam M. Kilcommons, Chung-Huei Wang, James M. McTiernan, C. Ho, W. Schreiner. *Early results from GOES-16 and GOES-17 magnetometer and magnetometer inversion algorithm*. *Space Weather*, 2019, 17 (11), 1452–1462.
28. DallaSantina Timothy, Murphy Martin, Anthony Reale, James Martin, Doug Lindholm, Mark Smith, Emily Berndt, David Biscan, Bradley Zavodsky, Kyle Burke, and others. *The GOES-R Proving Ground: Accelerating User Readiness for the Next-Generation Geostationary Environmental Satellites*. *Bulletin of the American Meteorological Society*, 2018, 99 (4), 631–651.
29. David C. Wooten, Raymond D. Blevins. *Geostationary operational environmental satellite R-series: The next generation of geostationary weather satellites*. *Journal of Applied Meteorology and Climatology*, 2016, 55 (7), 1493–1512.
30. NASA's Earth Observing System. *Geostationary Operational Environmental Satellite-16*. 2022. https://eospso.nasa.gov/missions/geostationary-operational-environmental-satellite-16
31. Timothy J. Schmit, Mathew M. Gunshor, W. Paul Menzel, James J. Gurka, Jun Li, A. Scott Bachmeier. *Introducing the Next-generation Advanced Baseline Imager on GOES-R*. *American Meteorological Society*, 2005, 86 (8), 1079–1096.
32. Yoram J. Kaufman, Didier Tanré, Olivier Boucher. *Aerosol optical thickness and atmospheric path radiance*. *Journal of Geophysical Research: Atmospheres*, 1998, 103 (D22), 25867–25880.
33. T. F. Eck, B. N. Holben, J. S. Reid, O. Dubovik, A. Smirnov, N. T. O'Neill, I. Slutsker, S. Kinne. *Aerosol optical depth measurements by airborne sun photometry during SAFARI 2000*. *Journal of Geophysical Research: Atmospheres*, 2003, 108 (D13), 8494.
34. Ralph A. Kahn, Brian J. Gaitley, John V. Martonchik, David J. Diner, Kathleen A. Crean. *Aerosol optical thickness from satellite and sun photometer measurements*. *Journal of Geophysical Research: Atmospheres*, 1997, 102 (D14), 16815–16830.
35. Lorraine A. Remer, Yoram J. Kaufman, Didier Tanré, Shana Mattoo, Donald A. Chu, J. Vanderlei Martins, Rong-Rong Li, Charles Ichoku, Robert C. Levy, Richard G. Kleidman, and others. *Aerosol remote sensing over oceans using the MODIS-EOS spectral radiances*. *Journal of Geophysical Research: Atmospheres*, 2005, 110 (D10).
36. Didier Tanré, Yoram J. Kaufman, Jay R. Herman, Shana Mattoo, Omar Torres, Brent N. Holben, Ilya Slutsker. *Aerosol remote sensing from POLDER/ADEOS over the ocean: improved retrieval using a nonspherical particle model*. *Journal of Geophysical Research: Atmospheres*, 1997, 102 (D14), 16989–17013.
37. Brent N. Holben, TF Eck, I. Slutsker, Didier Tanré, JP Buis, A. Setzer, E. Vermote, JA Reagan, YJ Kaufman, T. Nakajima, and others. *AERONET—A Federated Instrument Network and Data Archive for Aerosol Characterization*. *Remote Sensing of Environment*, 1998, 66 (1), 1–16.
38. Baudouin Raoult, Cédric Bergeron, Angel López Alós, Jean-Noël Thépaut, Dick Dee. *Climate service develops user-friendly data store*. 2017. https://www.ecmwf.int/en/newsletter/151/meteorology/climate-service-develops-user-friendly-data-store
39. ECMWF. *About us*. n.d. https://www.ecmwf.int/en/about
40. Climate Data Store (CDS). *ERA5-Land hourly data from 1950 to present*. 2019. https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview
41. M. G. Bosilovich, R. Lucchesi, and M. Suarez, *MERRA-2: File Specification*. GMAO Office Note No. 9 (Version 1.1), 73 pp, available from http://gmao.gsfc.nasa.gov/pubs/office_notes.
42. Menon, S., Hansen, J., Nazarenko, L., Luo, Y. Climate effects of black carbon aerosols in China and India. *Science* **2002**, *297* (5590), 2250–2253.
43. Fuzzi, S., Baltensperger, U., Carslaw, K., Decesari, S., Denier Van Der Gon, H., Facchini, M. C., Fowler, D., Koren, I., Langford, B., Lohmann, U., and others. Particulate matter, air quality and climate: lessons learned and future needs. *Atmospheric Chemistry and Physics* **2015**, *15* (14), 8217–8299.
44. M. T. Kleinman, R. F. Phalen, W. J. Mautz, R. C. Mannix, T. R. McClure, T. T. Crocker. Health effects of acid aerosols formed by atmospheric mixtures. *Environmental Health Perspectives* **1989**, *79*, 137.
45. IPCC. *Climate Change 2013: The Physical Science Basis*. In: *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P.M. Midgley (Eds.). Cambridge University Press, 2013, Chapter 7, pp. 571–658. https://www.ipcc.ch/report/ar5/wg1/.

46. EPA. *Air Quality Guide for Nitrogen Dioxide*. U.S. Environmental Protection Agency, 2010. https://www.epa.gov/sites/production/files/2015-08/documents/no2_aqg_summary.pdf.
47. Ming Zhang, Yingying Ma, Yifan Shi, Wei Gong, Shihua Chen, Shikuan Jin, Jun Wang. *Controlling factors analysis for the Himawari-8 aerosol optical depth accuracy from the standpoint of size distribution, solar zenith angles and scattering angles Atmospheric Environment*, 2020, 233, 1352-2310
48. Xiaohe Yu. *CLOUD DETECTION AND $PM_{2.5}$ ESTIMATION USING MACHINE LEARNING*. PhD thesis, The University of Texas at Dallas, 2021.
49. *SEDAC GPW-v4 Population Density, Rev11*. Socioeconomic Data and Applications Center. Accessed: March 26, 2023. https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev11
50. David J. Nowak, Satoshi Hirabayashi, Allison Bodine, Eric Greenfield. *Air pollution removal by urban forests in Canada and its effect on air quality and human health*. *Urban Forestry & Urban Greening*, 2018, 29, 40–48.
51. NRCS. *Web Soil Survey (WSS)*. 2019. https://websoilsurvey.sc.egov.usda.gov/app/
52. *Multi-Resolution Land Characteristics (MRLC) - National Land Cover Database (NLCD)*. Multi-Resolution Land Characteristics (MRLC). 2019. Accessed: March 26, 2023. https://www.mrlc.gov/data?f%5B0%5D=year%3A2019
53. *National Land Cover Database Class Legend and Description*. Multi-Resolution Land Characteristics Consortium. Unknown. Accessed: March 26, 2023. https://www.mrlc.gov/data/legends/national-land-cover-database-class-legend-and-description
54. *Gridded Bathymetry Data*. General Bathymetric Chart of the Oceans. n.d. https://www.gebco.net/data_and_products/gridded_bathymetry_data/
55. Jens Hartmann, Nils Moosdorf. *The new global lithological map database GLiM: A representation of rock properties at the Earth surface*. *Geochemistry, Geophysics, Geosystems*, 2012, 13(12), Q12004.
56. Pedro Camargo. *USBuildingFootprints*. 2022. [Accessed: March 26, 2023]. https://github.com/microsoft/USBuildingFootprints
57. Marius Gilbert, Gaëlle Nicolas, Giusepina Cinardi, Thomas P. Van Boeckel, Sophie O. Vanwambeke, G. R. William Wint, and Timothy P. Robinson. *Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010*. *Scientific Data*, volume 5, Article number: 180227, 2018.
58. *CIRC Systems*. CIRC team at UT Dallas. [Accessed: Current Year]. https://docs.circ.utdallas.edu/user-guide/systems/index.html
59. *"Stampede's" comprehensive capabilities to bolster U.S. open science computational resources*. Texas Advanced Computing Center. 2011. https://www.tacc.utexas.edu/-/-stampede-s-comprehensive-capabilities-to-bolster-u-s-open-science-computational-resources
60. Bin Chen, Zhihao Song, Feng Pan, and Yue Huang. *Obtaining vertical distribution of $PM_{2.5}$ from CALIOP data and machine learning algorithms*. *Science of The Total Environment*. Elsevier, 2021. **Volume 805**.
61. Sangeeta Sharma, Xiaolu Zhang, Cenlin Lin, and Jiangfeng Li. *Wildfire emissions, detection, and impacts on air quality*. *Environment International*, volume 92, pages 1–3, 2016. Publisher: Elsevier.
62. Yueyang Jiang, Yang Sun, and Michelle L. Bell. *Wildfires and their impacts on air quality in the western US*. *Current Pollution Reports*, volume 5, number 3, pages 229–239, 2019. Publisher: Springer.
63. Katelyn Westrick, Philip E Higuera, Maryellen Barnes, Paul A Duffy, Feng Sheng Hu, James A Lutz, Alexander L Metcalf, T Scott Rupp, and Cathy Whitlock. *Increased heat, drought, and insect outbreaks have contributed to severe wildfires in the western United States*. *Global Change Biology*, volume 26, number 11, pages 6106–6121, 2020. Publisher: Wiley Online Library.
64. Shelton Johnson, Arjan J Meddens, and Jeffrey A Hicke. *Effects of drought and insect outbreaks on epigaeic beetle communities in western USA deciduous forests*. *Agricultural and Forest Entomology*, volume 17, number 2, pages 160–171, 2015. Publisher: Wiley Online Library.
65. Jeremy L Weiss, Phillip J van Mantgem, and Simon C Brewer. *US wildfires, 1984–2012: A spatial temporal analysis of trends, drivers, and climatic associations*. *Annals of the American Association of Geographers*, volume 107, number 1, pages 1–12, 2017. Publisher: Taylor & Francis.
66. *WHO Air Quality Guidelines*. *Howpublished:* https://www.c40knowledgehub.org/s/article/WHO-Air-Quality-Guidelines?language=en_US
67. *Particulate Matter (PM) Pollution*. U.S. Environmental Protection Agency (EPA). n.d. [Accessed: February 10, 2024]. https://www.epa.gov/pm-pollution/final-reconsideration-national-ambient-air-quality-

 doi:10.20944/preprints202405.1685.v1

standards-particulate-matter-pm?emci=8c4af901-18c2-ee11-b660-002248223197&emdi=06d4332d-11c6-ee11-b660-002248223848&ceid=5660439

68. V. Gewin, *Air Pollution Threatens Millions of Lives. Now the Sources Are Shifting*, *Scientific American*, Feb. 8, 2024, https://www.scientificamerican.com/article/air-pollution-threatens-millions-of-lives-now-the-sources-are-shifting/.