



Article

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

Prabuddha M. H. Dewage ¹, Lakitha O. H. Wijeratne ¹, Xiaohe Yu ², Mazhar Iqbal ¹, Gokul Balagopal ¹, John Waczak ¹, Ashen Fernando ¹, Matthew D. Lary ¹, Shisir Ruwali ¹ and David J. Lary ^{1,*}

- ¹ Department of Physics, The University of Texas at Dallas, Richardson, TX 75080, USA; pxh180012@utdallas.edu (P.M.H.D.); lhw150030@utdallas.edu (L.O.H.W.); mxi190003@utdallas.edu (M.I.); gxb161030@utdallas.edu (G.B.); jxw190004@utdallas.edu (J.W.); aaf170130@utdallas.edu (A.F.); mdl210001@utdallas.edu (M.D.L.); sxr190184@utdallas.edu (S.R.)
- 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 the 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 were 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 datasets 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} analyses. 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 µg/m³, 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 µg/m³ for more than 20% of the time of our analysis period, with the eastern United States and California experiencing concentrations exceeding 9 µg/m³ for over 50% of the time, highlighting the importance of regulatory efforts to maintain annual PM_{2.5} concentrations below 9 µg/m³.

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



Citation: Dewage, P.M.H.; Wijeratne, L.O.H.; Yu, X.; Iqbal, M.; Balagopal, G.; Waczak, J.; Fernando, A.; Lary, M.D.; Ruwali, S.; Lary, D.J. Providing Fine Temporal and Spatial Resolution Analyses of Airborne Particulate Matter Utilizing Complimentary In Situ IoT Sensor Network and Remote Sensing Approaches. *Remote Sens.* **2024**, *16*, 2454. <https://doi.org/10.3390/rs16132454>

Academic Editor: Gregory Giuliani

Received: 25 May 2024

Revised: 26 June 2024

Accepted: 28 June 2024

Published: 3 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

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: PM₁ (particles smaller than 1 µm), PM_{2.5} (particles smaller than 2.5 µm), and PM₁₀ (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}, due to 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 problems and PM, which exhibits significant variations over time and in 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, a focus has been 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 establishing observation stations on the ground. These stations are designed to provide high-temporal-resolution data, specifically in the Dallas area, thus 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 nonlinear 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].

Lary et al. [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, Yu et al. [10] 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.

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 an especially 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, directly collected 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 are 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 were 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 [11] 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 the different micro-environments encountered on very small temporal and spatial scales. Harrison et al. (2015) [17] well demonstrated this point, highlighting the challenges in accurately capturing these variations. However, one major obstacle is the significant 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 Figure 1. These systems, which utilize affordable airborne particle sensors combined with machine learning techniques, have been strategically positioned to effectively monitor environmental conditions. The data collected by these sensors are readily available for real-time analysis via an online dashboard, as detailed by [18].

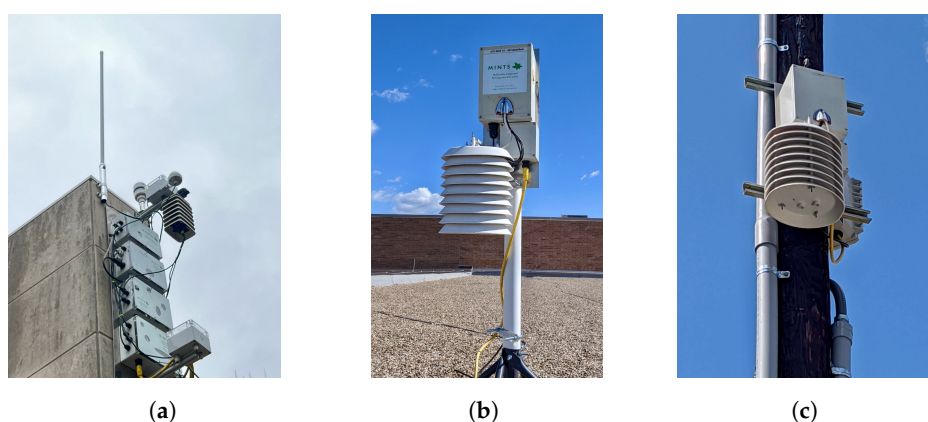


Figure 1. MINTS sensing systems deployment: (a) Central node at Plano, Texas; (b) UTD node at Dallas college, Texas; (c) UTD node at Joppa city, Texas.

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 have a specified accuracy of $\pm 10\%$ for particulate counting (PC), are adept at providing precise and real-time measurements of airborne particulate matter, ranging from PM₁₀ 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 [19].

Additionally, the system incorporates cost-effective gas sensors like the SCD30 for estimating CO₂ levels and the MICS6814 for gauging concentrations of CO, N₂, H₂, NH₃, CH₄, C₃H₄, C₄H₁₀, and C₂H₆OH. 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, sensors are configured to collect data on particulate matter, gases, and climatic conditions at high temporal resolution, capturing readings every 3 s. 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 [20]. 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 PM₁₀, collected by the EPA along with state, local, and tribal air pollution control agencies through hundreds of monitors nationwide [21]. However, negative data values in the AQS can occur due to equipment failures and measurement noise, particularly under very clean atmospheric conditions [11]. For this study, PM_{2.5} data, sampled on an hourly basis, were retrieved using the AQS API Figure 2. These datasets were then employed as ground-truth observations for the purposes of model training and validation.

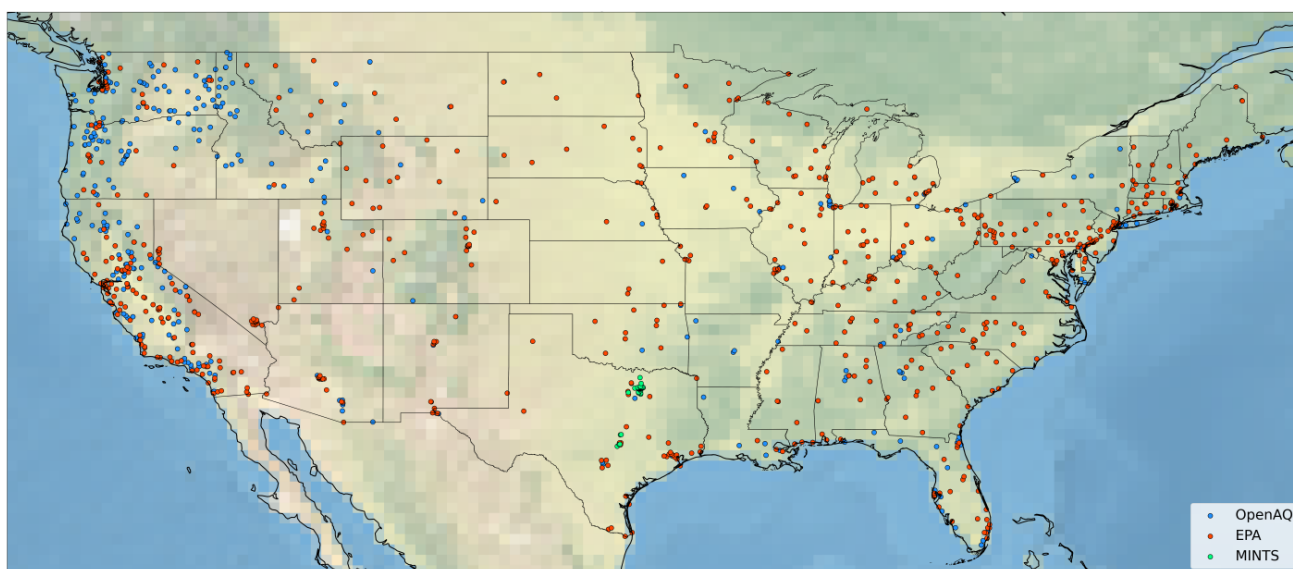


Figure 2. Ground observation sites of EPA (Red), OpenAQ (Blue), and MINTS (Green).

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, NO₂, and CH₄. 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 [22]. For this study, OpenAQ serves as an additional source of hourly sampled PM_{2.5} data, which are utilized for modeling training and validation.

2.2. GOES-16 AOD

In this research, the AOD data from the GOES-16 satellite were 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 [23–27]. AOD, with a spatial resolution as fine as 0.5 km and a temporal resolution reaching up to 30 s, plays a significant role in this study’s analysis.

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, pressure, and temperature. Under elevated relative humidity (RH) conditions, particles experience hygroscopic growth, a process wherein water vapor condenses onto their surfaces, resulting in an increase in particle diameter compared to normal conditions [28,29]. This growth enhances light scattering, significantly impacting aerosol optical depth (AOD) values. Therefore, accounting for RH is crucial when modeling particulate matter (PM) concentrations based on AOD measurements.

For this study, historical weather data were 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 [30]. This service is developed and managed by the European Centre for Medium-Range Weather Forecasts (ECMWF). The ECMWF has created ERA5-Land, a reanalysis dataset 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 dataset 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 m above the ground to a soil depth of 289 cm [31]. The meteorological variables of ERA5-Land that are used for PM_{2.5} modeling are detailed in Table 1.

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

Source	Variables	Temporal Resolution	Spatial Resolution
EPA	PM _{2.5}	1 h	-
OpenAQ	PM _{2.5}	1 h	-
MINTS	PM _{2.5}	3 s	-
ECMWF meteorological	Temperature	1 h	10 km × 10 km
	Pressure	1 h	10 km × 10 km
	Dewpoint temperature	1 h	10 km × 10 km
	Precipitation	1 h	10 km × 10 km
	Skin reservoir	1 h	10 km × 10 km
	Evaporation	1 h	10 km × 10 km
	Specific humidity	1 h	10 km × 10 km
	Relative humidity	1 h	10 km × 10 km
	Wind speed	1 h	10 km × 10 km
	Wind direction	1 h	10 km × 10 km
	Boundary layer height	1 h	10 km × 10 km
	Lake cover	1 h	10 km × 10 km
	Leaf area index, high vegetation	1 h	10 km × 10 km
	Leaf area index, low vegetation	1 h	10 km × 10 km
	Snowfall	1 h	10 km × 10 km
Solar radiation	1 h	10 km × 10 km	
Total cloud cover	1 h	10 km × 10 km	
Specific rain water content	1 h	10 km × 10 km	

Table 1. Cont.

Source	Variables	Temporal Resolution	Spatial Resolution
GOES-16	Aerosol optical depth	5 m	2 km × 2 km
	Data quality flag	5 m	2 km × 2 km
MERRA-2	AOD analysis	3 h	0.312° × 0.25°
	Total column ozone	3 h	0.312° × 0.25°
	Hydrophobic black carbon	3 h	0.312° × 0.25°
	Hydrophilic black carbon	3 h	0.312° × 0.25°
	Hydrophobic organic carbon	3 h	0.312° × 0.25°
	Hydrophilic organic carbon	3 h	0.312° × 0.25°
	SO ₄ sulfate aerosol	3 h	0.312° × 0.25°
	SO ₂ sulfur dioxide	3 h	0.312° × 0.25°
	NH ₃ Ammonia	3 h	0.312° × 0.25°
	NH ₄ Ammonium ion	3 h	0.312° × 0.25°
	NO ₃ Nitrate	3 h	0.312° × 0.25°
	CO Carbon monoxide	3 h	0.312° × 0.25°
	CO ₂ Carbon dioxide	3 h	0.312° × 0.25°
Ancillary Data	Landcover	-	30 m × 30 m
	Population	-	30 arc-second
	Soiltype	-	10 m × 10 m
	Lithology	-	0.5° × 0.5°
	Elevation	-	15 arc-seconds
	Cropland	-	30 m × 30 m
	Building footprint	-	-
	Livestock	-	5 min of arc
	Solar zenith angle	1 h	-
	Solar azimuth angle	1 h	-
	Month	-	-

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 [32].

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 [33]. Studies, including that of Menon et al. (2002) [34], 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 [35].

Sulfate aerosols arise from the oxidation of sulfur dioxide (SO₂) 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 [36], leading to cooling effects. Nitrate aerosols, produced by the 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 [37]. The formation and impact of these pollutants highlight their importance in understanding and modeling climate and air quality dynamics.

2.5. Solar Illumination

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. The geometry of solar illumination is crucial in defining the context of AOD measurements. Solar angles are closely related to the local time and have a huge influence on the AOD quality. The AOD value will not be retrieved due to extreme solar angles [10,38]. 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 that sunlight travels through the atmosphere of Earth to reach the surface.

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 [39]. This study incorporated slowly varying variables such as the population density, elevation, soil type, lithology, land cover, crop type, building footprint, and livestock distribution as important contextual ancillary data. These variables help understand the broader environmental and human factors that can impact the levels of particulate matter.

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) [40], a component of NASA, provides data on population density in the form of raster datasets. These datasets 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 [41]. Geographic variables such as elevation, soil type, lithology, cropland, and land cover offer information on the geological characteristics that could affect the levels of particles.

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 m, covers the entire continental United States.

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 [42].

The National Land Cover Database (NLCD) offers detailed information on land cover and changes over time within the United States. With a 30-m 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 [43,44].

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 dataset presents elevation data on a grid with 15 arc second intervals [45].

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) [46] 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 concen-

tration levels. Microsoft Maps offers a comprehensive open dataset of building footprints for the United States. This dataset 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 [47].

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 min of arc, which is roughly equivalent to 10 km at the equator. The data detail the total number of each species per pixel (5 min 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 emissions from enteric fermentation and manure management, together with the disruption of nitrogen and phosphorus cycles [48].

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 [49], 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 [50].

3.1. All PM Size Fractions Modeling—MINTS Observation

In this phase of the study, data were exclusively acquired 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 CO₂ 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 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.

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

Table 2. *Cont.*

Sensor	Variables
AS7262	Violet Blue Green Yellow Orange Red
TSL2591	Luminosity Infrared Full spectrum Visible light Lux
VEML6075	Ultraviolet A Ultraviolet B

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 solely based on the high-temporal-resolution data from MINTS, limiting our feature variables to those available within the MINTS dataset.

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, CO₂ concentrations from SCD30, and light intensities from AS7262, TSL2591, and VEML6075. Meanwhile, Group 3 is tailored to assess the impact of light intensities specifically on different PM size fractions. The data sets for each group include approximately 617,000 entries, split into two parts: 80% of the data are used for training purposes, and the remaining 20% are reserved for testing.

Table 3. MINTS observation PM groups.

Group	Weather	CO ₂	Light
1	✓		
2	✓	✓	✓
3			✓

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 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 dataset. 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.

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 s, 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, were 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 these data usable for geographical analyses, it is necessary to transform the AOD data into a geographic coordinate system. This transformation relies on metadata that include 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, provide an approximate spatial resolution of 50 km × 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 using the locations of ground-based PM observation sites from the EPA, OpenAQ, and MINTS as the reference coordinate system. A multilinear interpolation method was used to ensure that the datasets 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 dataset had missing AOD values, which were then filled with the corresponding AOD data from the MERRA-2 dataset to complete the dataset.

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	January 2020–June 2023			✓
2	US	January 2020–June 2023	✓		✓
3	US	January 2020–June 2023	✓	✓	✓
4	US	January 2020–June 2023	✓	✓	
5	TX	September 2021–June 2023	✓	✓	✓
6	TX	September 2021–June 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 datasets 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.

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 dataset with numerous variables, and machine learning algorithms excel at managing complex datasets 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 was shown to be effective

for PM_{2.5} modeling using AOD and meteorological variables in previous research [11,51]. 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.

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 (µg/m ³)	Test R	Test RMSE (µg/m ³)
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

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 a 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.

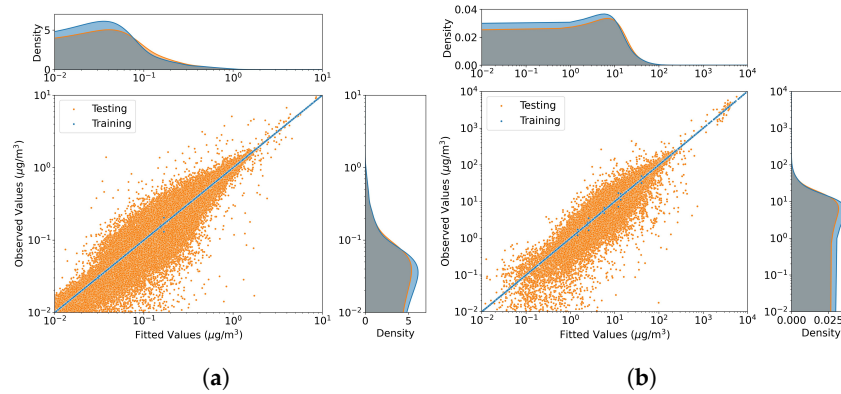


Figure 3. Scatter diagrams depicting the training and testing datasets for Group-2 (incorporate all the feature variables within MINTS system): (a) $PM_{0.1}$; and (b) $PM_{10.0}$.

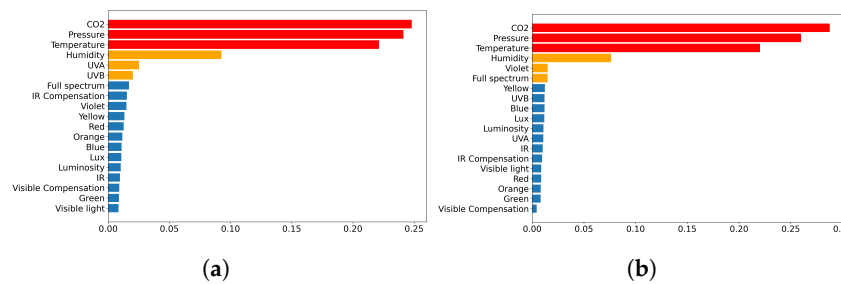


Figure 4. Importance of features for Group-2 (incorporate all the feature variables within MINTS system): (a) $PM_{0.1}$; and (b) $PM_{10.0}$.

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 the particles.

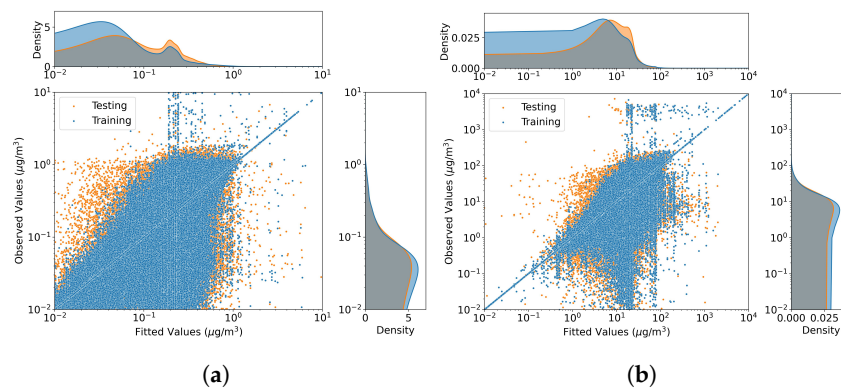


Figure 5. Scatter plots depicting the training and testing data for Group-3 (incorporate only light sensing variables within MINTS system): (a) $PM_{0.1}$; and (b) $PM_{10.0}$.

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.

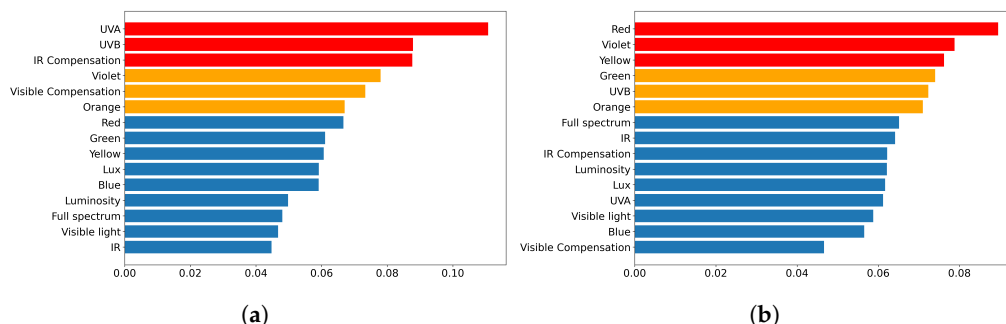


Figure 6. Importance of features for Group-3 (incorporate only light sensing variables within MINTS system): (a) $PM_{0.1}$; and (b) $PM_{10.0}$.

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 dataset 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 were 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 datasets. 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 ($\mu g/m^3$)	Test R	Test RMSE ($\mu g/m^3$)
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 exclusively relies 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.

The scatter diagram comparing the 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.

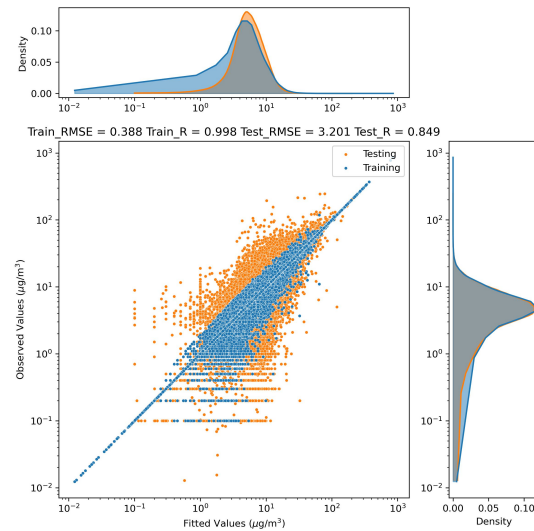


Figure 7. Scatter plots depicting training and testing data in log scale, accompanied by marginal probability density functions, illustrating the analysis conducted for Model-3.

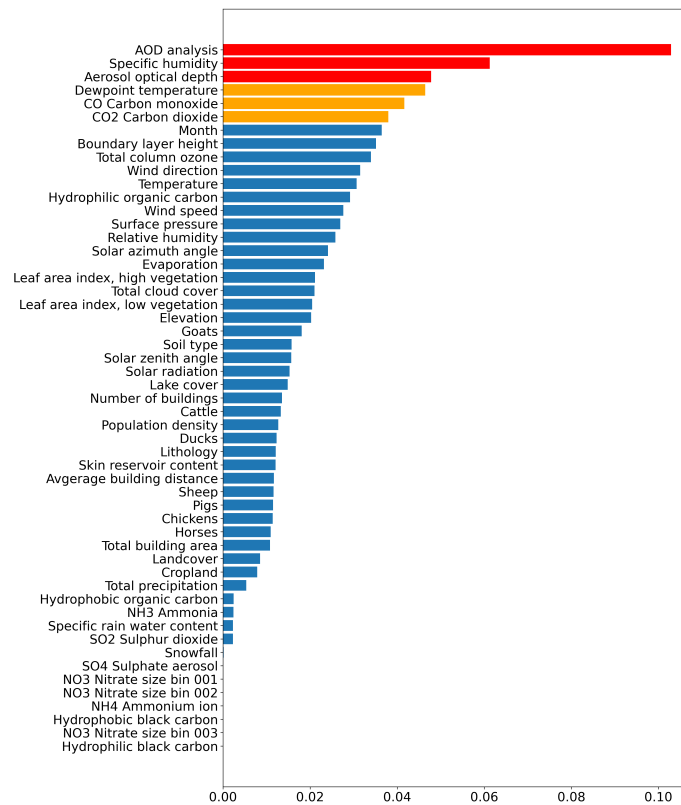


Figure 8. Feature importance score for Model-3.

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. A comprehensive input dataset for the machine learning model was prepared through several preprocessing steps. To ensure uniformity in all ground-level PM_{2.5} concentration maps, the ECMWF meteorological data grid, which measures approximately 10 km × 10 km and covers the whole US region, is used as the standard coordinate framework. This grid array was transformed from a two-dimensional shape into a one-dimensional format and then combined into a tabular structure, such as a dataframe. This coordinate dataframe, containing latitude and longitude, was used as the reference coordinate dataframe. 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. The low dynamic ancillary data were augmented to this reference coordinate dataframe by matching the locations. Since all other feature variables vary with timestamps, this reference coordinate dataframe with matched ancillary data was duplicated for hourly timestamps. ECMWF meteorological data were incorporated into the corresponding hourly reference dataframe by matching the spatial coordinates. Time-interpolated MERRA-2 data were also integrated into the respective timestamp dataframes by matching location coordinates. AOD data were then aligned with the respective timestamp dataframes by matching location coordinates. Solar angles for specific datetime dataframes were generated using the spatial coordinates. The resulting dataframes for each timestamp contained spatially matched feature variables data. These enriched dataframes were sequentially inputted into the machine learning model to generate the hourly dataframes of estimated PM_{2.5} concentrations at all location coordinates. These output dataframes were transformed into two-dimensional arrays of latitude, longitude, and estimated PM_{2.5} to visualize the PM_{2.5} reconstruction maps.

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 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 9a,b offer visual insights into the ground-level PM_{2.5} concentrations recorded at two different times: 9 PM and midnight on 2 October 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, Figure 9c,d 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 µg/m³ 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 exclusively focuses on results within the state of Texas. Specifically, we scrutinize data from three distinct timestamps on 1 January 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 h. 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).

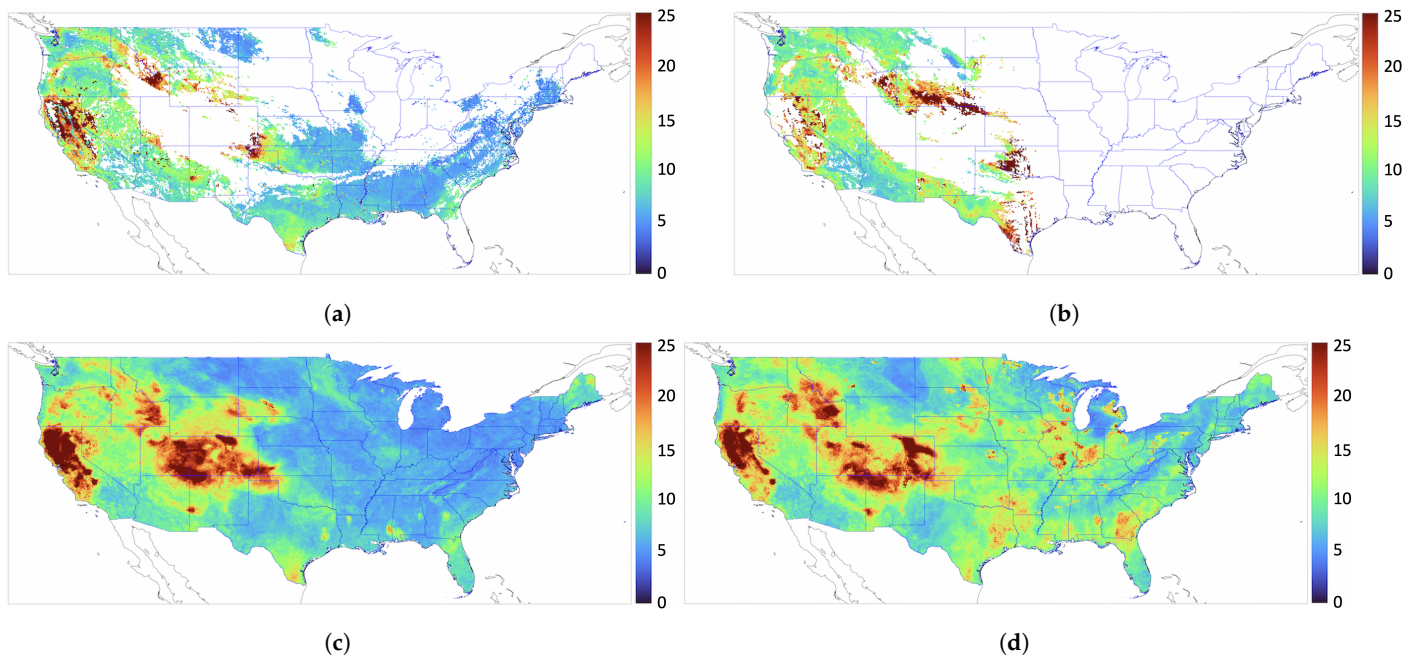


Figure 9. $PM_{2.5}$ reconstruction during the Santa Clara Unit (SCU) Lightning Complex fire in 2020. Panels (a,b) are for 9 PM UTC and midnight on October 2, respectively, using a specialized version of Model-2 that exclusively incorporates AOD data from GOES-16. Panels (c,d) are for the same times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

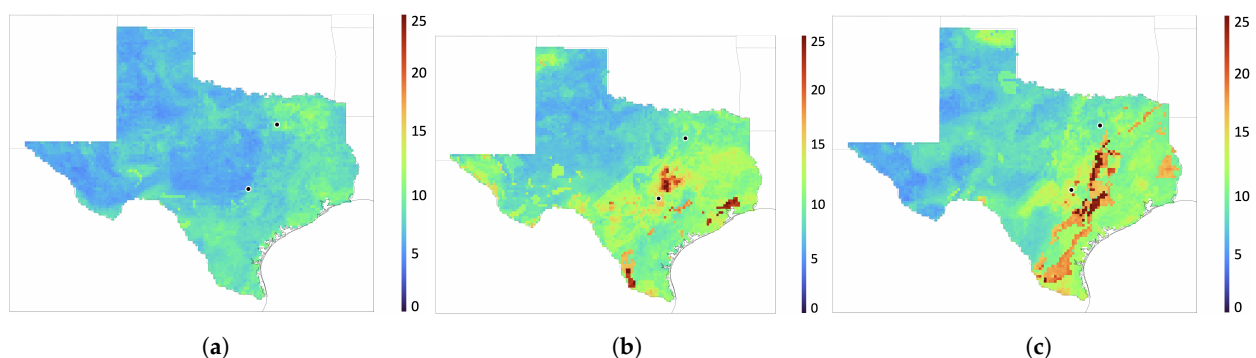


Figure 10. Reconstructed $PM_{2.5}$ concentrations across the Texas region at three distinct timestamps on 1 January 2023, in UTC. The black solid circle in the north corresponds to the MINTS ground sensor located in Joppa (south Dallas), while the black solid circle in the south represents the MINTS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

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 31 December 2022 (equivalent to 1 January 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 h 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 h 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.

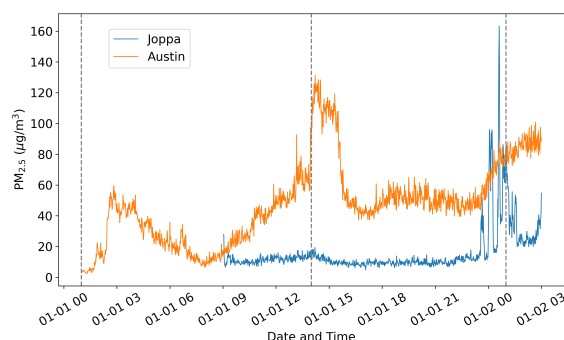


Figure 11. PM_{2.5} measurements obtained from two MINTS in situ sensors located in Joppa (depicted in blue) and Austin (shown in orange). The timestamps indicated by the gray dashed lines align with those presented in Figure 10.

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, concerns remain 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 µg/m³ to 9 µg/m³, aiming to mitigate the adverse health impacts and associated costs. The EPA estimates that adhering to this new standard could lead to potential savings of up to USD 46 billion in avoided healthcare and hospitalization costs by 2032 [52–60].

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 dataset allows us to calculate the fraction of time during which PM_{2.5} concentrations exceeded five distinct threshold levels (8 µg/m³, 9 µg/m³, 10 µg/m³, 11 µg/m³, and 12 µg/m³) 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 µg/m³ 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 µg/m³ 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 µg/m³ for more than 50% of the time during the same period. These estimates underscore the importance of regulatory measures aiming to maintain annual PM_{2.5} concentrations below 9 µg/m³.

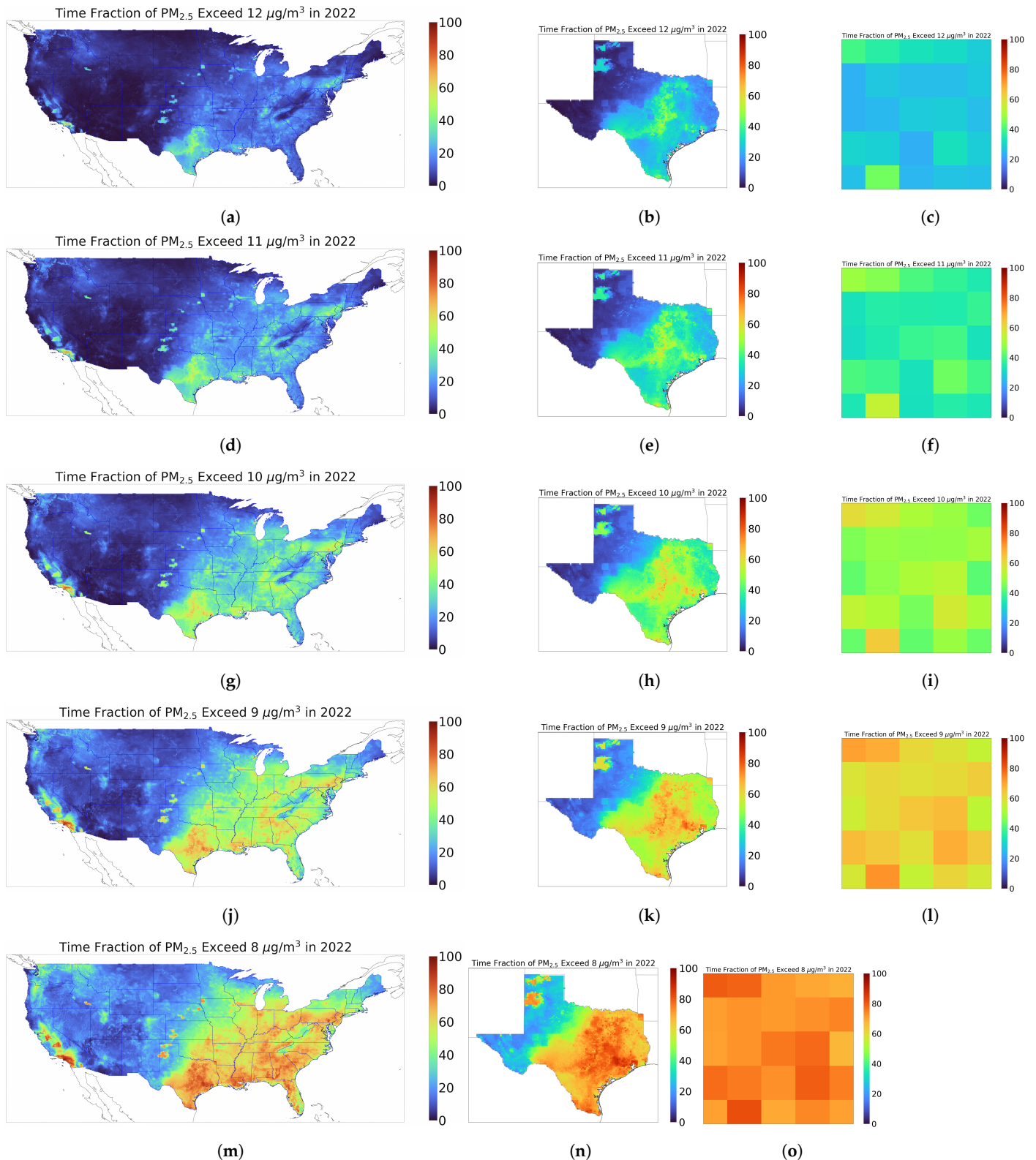


Figure 12. Percentage of time exceeding PM_{2.5} concentration thresholds throughout the entirety of 2022, as estimated by Model-3. The subfigures (a,d,g,j,m) illustrate the duration exceeding the thresholds 12 µg/m³, 11 µg/m³, 10 µg/m³, 9 µg/m³, and 8 µg/m³ over the US, respectively. The subfigures (b,e,h,k,n) illustrate the corresponding PM_{2.5} exceeding in Texas regions and the subfigures (c,f,i,l,o) illustrate the corresponding PM_{2.5} exceeding in Dallas regions, respectively.

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 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 asynchronous temporal and spatial coordinate system, leveraging pertinent remote sensing data.

In the first study, in order to 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 restriction of 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 a 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 solely relied 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 are 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, which 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 PM_{2.5} modeling incorporated complementary in situ and remote sensing approaches.

With 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 re-analysis 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 µg/m³ for more than 20% of the time of our analysis period, with the eastern United States and California experiencing concentrations exceeding 9 µg/m³ for over 50% of the time, highlighting the importance of regulatory efforts to maintain annual PM_{2.5} concentrations below 9 µg/m³.

Author Contributions: Conceptualization, D.J.L.; Methodology, P.M.H.D., L.O.H.W., X.Y., M.I., G.B., J.W., A.F., M.D.L., S.R. and D.J.L.; Supervision, D.J.L.; Project administration, D.J.L.; Funding acquisition, D.J.L. All authors have read and agreed to the published version of the manuscript.

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 upon 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

- Boucher, O. *Atmospheric Aerosols: Properties and Climate Impacts*; Springer: The Netherlands, 2015. ISBN 978-9401796484.
- 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). *Sci. Total Environ.* **2011**, *409*, 4934–4938. [[CrossRef](#)] [[PubMed](#)]
- Lary, D.J.; Faruque, F.S.; Malakar, N.; Moore, A.; Roscoe, B.; Adams, Z.L.; Eggelston, Y. Estimating the global abundance of ground level presence of particulate matter (PM_{2.5}). *Geospat. Health* **2014**, *8*, 611–630. [[CrossRef](#)] [[PubMed](#)]
- Pope, C.A., III; Burnett, R.T.; Thun, M.J.; Calle, E.E.; Krewski, D.; Ito, K.; Thurston, G.D. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *Jama* **2002**, *287*, 1132–1141. [[CrossRef](#)] [[PubMed](#)]
- 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. *J. Atmos. Sci.* **2002**, *59*, 590–608. [[CrossRef](#)]
- Charlson, R.J.; Schwartz, S.E.; Hales, J.M.; Cess, R.D.; Coakley, J.A., Jr.; Hansen, J.E.; Hofmann, D.J. Climate forcing by anthropogenic aerosols. *Science* **1992**, *255*, 423–430. [[CrossRef](#)] [[PubMed](#)]
- Pöschl, U. Atmospheric aerosols: Composition, transformation, climate and health effects. *Angew. Chem. Int. Ed.* **2005**, *44*, 7520–7540. [[CrossRef](#)] [[PubMed](#)]
- National Research Council. *A Plan for a Research Program on Aerosol Radiative Forcing and Climate Change*; National Academies Press: Cambridge, MA, USA, 1996.
- Chin, M. *Atmospheric Aerosol Properties and Climate Impacts*; DIANE Publishing Company: 2009. ISBN 978-1437912616. Available online: <https://books.google.com/books?id=IgJZXXgtHmQC> (accessed on 26 March 2023).
- Yu, X.; Lary, D.J.; Simmons, C.S.; Wijeratne, L.O.H. High Spatial-Temporal PM_{2.5} Modeling Utilizing Next Generation Weather Radar (NEXRAD) as a Supplementary Weather Source. *Remote Sens.* **2022**, *14*, 495. [[CrossRef](#)]
- Yu, X.; Lary, D.J.; Simmons, C.S. PM_{2.5} Modeling and Historical Reconstruction over the Continental USA Utilizing GOES-16 AOD. *Remote Sens.* **2021**, *13*, 4788. [[CrossRef](#)]
- Wijeratne, L.O.H. Coupling Physical Measurement with Machine Learning for Holistic Environmental Sensing. Ph.D. Thesis, The University of Texas at Dallas, Richardson, TX, USA, 2021.
- Lary, D.J.; Lary, T.; Sattler, B. Using Machine Learning to Estimate Global PM_{2.5} for Environmental Health Studies. *Environ. Health Insights* **2015**, *1*, 41–52. [[CrossRef](#)]
- Wijeratne, L.O.; Kiv, D.R.; Aker, A.R.; Talebi, S.; Lary, D.J. Using Machine Learning for the Calibration of Airborne Particulate Sensors. *Sensors* **2019**, *20*, 99. [[CrossRef](#)]
- Zhang, H.; Hoff, R.M.; Engel-Cox, J.A. 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. *Air Waste Manag. Assoc.* **2009**, *59*, 1358–1369. [[CrossRef](#)]
- Zheng, C.; Zhao, C.; Zhu, Y.; Wang, Y.; Shi, X.; Wu, X.; Chen, T.; Wu, F.; Qiu, Y. Analysis of influential factors for the relationship between PM_{2.5} and AOD in Beijing. *Atmos. Chem. Phys.* **2017**, *17*, 13473–13489. [[CrossRef](#)]
- Harrison, W.A. *In-Situ Observation of Atmospheric Particulates*; The University of Texas at Dallas: Richardson, TX, USA, 2015.
- Talebi, S. Physical Quantification of the Interactions Between Environment, Physiology, and Human Performance. Ph.D. Thesis, The University of Texas at Dallas, Richardson, TX, USA, 2022.
- Piera Systems. *IPS Series Sensor*; Piera Systems Inc.: Mississauga, ON, Canada, 2022.
- United States Environment Protection Agency EPA. *Air Quality System (AQS) API*; United States Environment Protection Agency EPA: Washington, DC, USA, 2020. Available online: https://aqsweb/documents/data_api.html (accessed on 26 March 2023).
- Air Quality System (AQS) Data API*; U.S. Environmental Protection Agency (EPA): Washington, DC, USA, 2023. Available online: https://aqsweb/documents/data_api.html (accessed on 26 March 2023).
- OpenAQ-About. OpenAQ. n.d. Available online: <https://openaq.org/about/> (accessed on 26 March 2023).
- Volz, F.E.; Lee, T.H.; LaPenta, T.F.; Spinhirne, J.D.; Hulley, G.B.; O'Brien, J.J. Geostationary operational environmental satellite system—R (GOES-R). *Bull. Am. Meteorol. Soc.* **2000**, *81*, 2345–2363.
- Schmit, T.J.; Gunshor, M.M.; Menzel, W.P.; Gurka, J.J.; Li, J.; Bachmeier, A.S. Introducing the Next-Generation Advanced Baseline Imager on GOES-R. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 681–698. [[CrossRef](#)]
- Mannucci, A.J.; Stephens, P.W.; Kilcommons, L.M.; Wang, S.-H.; McTiernan, J.M.; Ho, C.; Schreiner, W. Early results from GOES-16 and GOES-17 magnetometer and magnetometer inversion algorithm. *Space Weather* **2019**, *17*, 1452–1462.
- Timothy, D.S.; Martin, M.; Reale, A.; Martin, J.; Lindholm, D.; Smith, M.; Berndt, E.; Biscan, D.; Zavodsky, B.; Burke, K. The GOES-R Proving Ground: Accelerating User Readiness for the Next-Generation Geostationary Environmental Satellites. *Bull. Am. Meteorol. Soc.* **2018**, *99*, 631–651.
- Wooten, D.C.; Blevins, R.D. Geostationary operational environmental satellite R-series: The next generation of geostationary weather satellites. *J. Appl. Meteorol. Climatol.* **2016**, *55*, 1493–1512.

28. Crilley, L.R.; Shaw, M.; Pound, R.; Kramer, L.J.; Price, R.; Young, S.; Lewis, A.C.; Pope, F.D. Evaluation of a Low-cost Optical Particle Counter (Alphasense OPC-N2) for Ambient Air Monitoring. *Atmos. Meas. Tech.* **2018**, *11*, 709–720. [[CrossRef](#)]
29. Di Antonio, A.; Popoola, O.A.; Ouyang, B.; Saffell, J.; Jones, R.L. Developing a Relative Humidity Correction for Low-Cost Sensors Measuring Ambient Particulate Matter. *Sensors* **2018**, *18*, 2790. [[CrossRef](#)]
30. Raoult, B.; Bergeron, C.; Alós, A.L.; Thépaut, J.-N.; Dee, D. Climate Service Develops User-Friendly Data Store. ECMWF Newsl. **2017**, *151*, 22–27. Available online: <https://www.ecmwf.int/en/newsletter/151/meteorology/climate-service-develops-user-friendly-data-store> (accessed on 26 March 2023).
31. Climate Data Store (CDS). *ERA5-Land Hourly Data from 1950 to Present*; Climate Data Store (CDS): 2019. Available online: <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-land?tab=overview> (accessed on 26 March 2023).
32. Bosilovich, M.G.; Luchesi, R.; Suarez, M. MERRA-2: File Specification. GMAO Office Note No. 9 (Version 1.1). Available online: http://gmao.gsfc.nasa.gov/pubs/office_notes (accessed on 26 March 2023).
33. Fuzzi, S.; Baltensperger, U.; Carslaw, K.; Decesari, S.; Denier van der Gon, H.; Facchini, M.C.; Fowler, D.; Koren, I.; Langford, B.; Lohmann, U.; et al. Particulate matter, air quality and climate: lessons learned and future needs. *Atmos. Chem. Phys.* **2015**, *15*, 8217–8299. [[CrossRef](#)]
34. Menon, S.; Hansen, J.; Nazarenko, L.; Luo, Y. Climate effects of black carbon aerosols in China and India. *Science* **2002**, *297*, 2250–2253. [[CrossRef](#)] [[PubMed](#)]
35. Kleinman, M.T.; Phalen, R.F.; Mautz, W.J.; Mannix, R.C.; McClure, T.R.; Crocker, T.T. Health effects of acid aerosols formed by atmospheric mixtures. *Environ. Health Perspect.* **1989**, *79*, 137. [[CrossRef](#)] [[PubMed](#)]
36. 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*; Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M., Eds.; Cambridge University Press: Cambridge, UK, 2013; Chapter 7, pp. 571–658. Available online: <https://www.ipcc.ch/report/ar5/wg1/> (accessed on 26 March 2023).
37. EPA. *Air Quality Guide for Nitrogen Dioxide*; U.S. Environmental Protection Agency: Washington, DC, USA, 2010. Available online: https://www.epa.gov/sites/production/files/2015-08/documents/no2_aqg_summary.pdf (accessed on 26 March 2023).
38. Zhang, H.; Kondragunta, S.; Laszlo, I.; Zhou, M. Improving GOES Advanced Baseline Imager (ABI) aerosol optical depth (AOD) retrievals using an empirical bias correction algorithm. *Atmos. Meas. Tech.* **2020**, *13*, 5955–5967. [[CrossRef](#)]
39. Yu, X. Cloud Detection and PM_{2.5} Estimation Using Machine Learning. Ph.D. Thesis, The University of Texas at Dallas, Richardson, TX, USA, 2021.
40. *SEDAC GPW-v4 Population Density, Rev11*; Socioeconomic Data and Applications Center. Available online: <https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev11> (accessed on 26 March 2023).
41. Nowak, D.J.; Hirabayashi, S.; Bodine, A.; Greenfield, E. Air pollution removal by urban forests in Canada and its effect on air quality and human health. *Urban For. Urban Green.* **2018**, *29*, 40–48. [[CrossRef](#)]
42. NRCS. *Web Soil Survey (WSS)*; NRCS: Washington, DC, USA, 2019. Available online: <https://websoilsurvey.sc.egov.usda.gov/app/> (accessed on 26 March 2023).
43. Multi-Resolution Land Characteristics (MRLC)—National Land Cover Database (NLCD). Multi-Resolution Land Characteristics (MRLC). 2019. Available online: <https://www.mrlc.gov/data?f%5B0%5D=year%3A2019> (accessed on 26 March 2023).
44. National Land Cover Database Class Legend and Description. Multi-Resolution Land Characteristics Consortium. Available online: <https://www.mrlc.gov/data/legends/national-land-cover-database-class-legend-and-description> (accessed on 26 March 2023).
45. Gridded Bathymetry Data. General Bathymetric Chart of the Oceans. n.d. Available online: https://www.gebco.net/data_and_products/gridded_bathymetry_data/ (accessed on 26 March 2023).
46. Hartmann, J.; Moosdorf, N. The new global lithological map database GLiM: A representation of rock properties at the Earth surface. *Geochem. Geophys. Geosyst.* **2012**, *13*, Q12004. [[CrossRef](#)]
47. Pedro Camargo. *USBuildingFootprints*. 2022. Available online: <https://github.com/microsoft/USBuildingFootprints> (accessed on 26 March 2023).
48. Gilbert, M.; Nicolas, G.; Cinardi, G.; Van Boeckel, T.P.; Vanwambeke, S.O.; Wint, G.R.; Robinson, T.P. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Sci. Data* **2018**, *5*, 180227. [[CrossRef](#)] [[PubMed](#)]
49. CIRC Systems. CIRC Team at UT Dallas. Available online: <https://docs.circ.utdallas.edu/user-guide/systems/index.html> (accessed on 26 March 2023).
50. "Stampede's" *Comprehensive Capabilities to Bolster U.S. Open Science Computational Resources*; Texas Advanced Computing Center: Austin, TX, USA, 2011. Available online: <https://www.tacc.utexas.edu/-/-stampede-s-comprehensive-capabilities-to-bolster-u-s-open-science-computational-resources> (accessed on 26 March 2023).
51. Chen, B.; Song, Z.; Pan, F.; Huang, Y. Obtaining vertical distribution of PM_{2.5} from CALIOP data and machine learning algorithms. *Sci. Total Environ.* **2021**, *805*, 150338. [[CrossRef](#)]
52. *Particulate Matter (PM) Pollution*; U.S. Environmental Protection Agency (EPA): Washington, DC, USA. Available online: <https://www.epa.gov/pm-pollution/final-reconsideration-national-ambient-air-quality-standards-particulate-matter-pm?emci=8c4af901-18c2-ee11-b660-002248223197&emdi=06d4332d-11c6-ee11-b660-002248223848&ceid=5660439> (accessed on 10 February 2024).

53. Gewin, V. Air Pollution Threatens Millions of Lives. Now the Sources Are Shifting, Scientific American. Available online: <https://www.scientificamerican.com/article/air-pollution-threatens-millions-of-lives-now-the-sources-are-shifting/> (accessed on 8 February 2024).
54. Zhang, M.; Ma, Y.; Shi, Y.; Gong, W.; Chen, S.; Jin, S.; Wang, J. Controlling factors analysis for the Himawari-8 aerosol optical depth accuracy from the standpoint of size distribution, solar zenith angles and scattering angles. *Atmos. Environ.* **2020**, *233*, 1352–2310 [[CrossRef](#)]
55. Sharma, S.; Zhang, X.; Lin, C.; Li, J. Wildfire emissions, detection, and impacts on air quality. *Environ. Int.* **2016**, *92*, 1–3.
56. Jiang, Y.; Sun, Y.; Bell, M.L. Wildfires and their impacts on air quality in the western US. *Curr. Pollut. Rep.* **2019**, *5*, 229–239.
57. Westrick, K.; EHiguera, P.; Barnes, M.; ADuffy, P.; Sheng, P.; James, H.; Alex, A.; LMetcalf, E.; Rupp, T.; Whitlock, C. Increased heat, drought, and insect outbreaks have contributed to severe wildfires in the western United States. *Glob. Chang. Biol.* **2020**, *26*, 6106–6121.
58. Johnson, S.; Meddens, A.J.; Hicke, J.A. Effects of drought and insect outbreaks on epigeic beetle communities in western USA deciduous forests. *Agric. For. Entomol.* **2015**, *17*, 160–171.
59. Weiss, J.L.; van Mantgem, P.J.; Brewer, S.C. US wildfires, 1984–2012: A spatial temporal analysis of trends, drivers, and climatic associations. *Ann. Am. Assoc. Geogr.* **2017**, *107*, 1–12.
60. WHO Air Quality Guidelines. Howpublished. Available online: https://www.c40knowledgehub.org/s/article/WHO-Air-Quality-Guidelines?language=en_US (accessed on 26 March 2023).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.