air

MDPI

Article

Characterizing the Temporal Variation of Airborne Particulate Matter in an Urban Area Using Variograms

Gokul Balagopal, Lakitha Wijeratne, John Waczak, Prabuddha Hathurusinghe, Mazhar Iqbal, Rittik Patra, Adam Aker, Seth Lee, Vardhan Agnihotri, Christopher Simmons and David J. Lary *

Hanson Center for Space Sciences, University of Texas at Dallas, Richardson, TX 75080, USA; gokul.balagopal@utdallas.edu (G.B.); lhw150030@utdallas.edu (L.W.); john.waczak@utdallas.edu (J.W.); pxh180012@utdallas.edu (P.H.); mazhar.iqbal@utdallas.edu (M.I.); rittik.patra@utdallas.edu (R.P.); adam.aker@utdallas.edu (A.A.); sethlee2024@gmail.com (S.L.); 24agnihotriv@smtexas.org (V.A.); csim@utdallas.edu (C.S.)

* Correspondence: david.lary@utdallas.edu

Abstract: This study aims to determine the optimal frequency for monitoring airborne pollutants in densely populated urban areas to effectively capture their temporal variations. While environmental organizations worldwide typically update air quality data hourly, there is no global consensus on the ideal monitoring frequency to adequately resolve pollutant (particulate matter) time series. By applying temporal variogram analysis to particulate matter (PM) data over time, we identified specific measurement intervals that accurately reflect fluctuations in pollution levels. Using January 2023 air quality data from the Joppa neighborhood of Dallas, Texas, USA, temporal variogram analysis was conducted on three distinct days with varying $PM_{2.5}$ (particulate matter of size $\leq$ 2.5 µm in diameter) pollution levels. For the most polluted day, the optimal sampling interval for $PM_{2.5}$ was determined to be 12.25 s. This analysis shows that highly polluted days are associated with shorter sampling intervals, highlighting the need for highly granular observations to accurately capture variations in PM levels. Using the variogram analysis results from the most polluted day, we trained machine learning models that can predict the sampling time using meteorological parameters. Feature importance analysis revealed that humidity, temperature, and wind speed could significantly impact the measurement time for $PM_{2.5}$. The study also extends to the other size fractions measured by the air quality monitor. Our findings highlight how local conditions influence the frequency required to reliably track changes in air quality.

Keywords: environmental research; particulate matter; air quality; statistics; IOT; variogram; optimization; data science; machine learning; big data; sensors

Academic Editor: Ling Tim Wong

Received: 6 December 2024
Revised: 9 January 2025
Accepted: 14 January 2025
Published: 3 March 2025

Citation: Balagopal, G.; Wijeratne, L.; Waczak, J.; Hathurusinghe, P.; Iqbal, M.; Patra, R.; Aker, A.; Lee, S.; Agnihotri, V.; Simmons, C.; et al. Characterizing the Temporal Variation of Airborne Particulate Matter in an Urban Area Using Variograms. *Air* **2025**, *3*, 7. https://doi.org/10.3390/air3010007

1. Introduction

Air pollution may be defined as the contamination of the air by materials that are produced as a byproduct of either physical, chemical, or biological activity. Pollution can be hazardous for people, animals, plants, and the ecosystem as a whole. Every year, air pollution claims more than 7 million lives [1]. Airborne particulate matter (PM) is widely acknowledged to be one of the most important contributors of air pollution. According to the findings of the United States Environmental Protection Agency (EPA), PM is one of the top six pollutants [2]. Particulate matter encompasses a range of airborne particles with varying sizes, compositions, and sources. Particulate matter is commonly categorized into size fractions such as $PM_{2.5}$ and $PM_{10.0}$. $PM_{2.5}$ (particulate matter with an aerodynamic diameter less than or equal to 2.5 μm) is particularly concerning due to its ability to penetrate deep into the respiratory system, entering the bloodstream. Studies have demonstrated a strong association between $PM_{2.5}$ exposure and adverse health outcomes, including increased mortality rates from cardiovascular and respiratory diseases [3–9].

During the COVID-19 pandemic, studies showed that sudden variations in $PM_{2.5}$ levels induced lung inflammation, which heightened the risk of COVID-19 infections and increased mortality rates [10]. These short-term spikes in PM concentrations, often referred to as "peak exposures", can have disproportionately adverse health effects, highlighting the need for finer temporal resolution in monitoring to better understand and mitigate their health impacts [11,12]. A significant challenge in air quality monitoring is determining the optimal temporal resolution for capturing sudden variations in PM concentrations. While environmental agencies like the EPA typically report PM concentrations at hourly intervals [13], this resolution is sufficient for identifying general trends but often misses rapid fluctuations caused by transient emissions or meteorological phenomena [14,15]. The variability in airborne particulate matter stems from a complex interplay of anthropogenic, natural, and meteorological factors. Major sources of PM include emissions from vehicles, industrial processes, and construction activities, as well as natural events like wildfires, volcanic eruptions, and dust storms [16]. Meteorological conditions, such as wind speed, wind direction, temperature, atmospheric pressure, and relative humidity, play a critical role in the transport, dispersion, and deposition of PM [17]. Additionally, urban structures like street canyons and valleys can influence airflow patterns, creating localized pollution hotspots that further complicate monitoring efforts.

To effectively capture the temporal variability in PM concentrations, variograms can be employed. A variogram [18–20] is a geostatistical tool used to quantify the variability of a parameter. Temporal variograms provide insights into the temporal correlation structures [21] of particulate matter concentrations. The range parameter of the temporal variogram identifies the characteristic time scale beyond which data points are no longer correlated, hence enabling researchers to determine appropriate sampling intervals [15]. This approach is particularly relevant when assessing short-term exposure risks or designing monitoring systems for urban environments. In this study, we aim to determine the optimal sampling frequency for various PM size fractions and assess the influence of meteorological parameters on the required sampling frequency.

By focusing on the temporal fluctuations of airborne particulate content [5–7,9,12], this work addresses a critical gap in the current air quality monitoring practices. Since air quality varies from region to region, we analyze data collected from a specific urban neighborhood (Joppa neighborhood of Dallas, Texas), where moment-to-moment fluctuations in PM concentrations are pronounced due to dynamic emission sources (such as rail yards and shingle factories) and meteorological conditions. In the preliminary phase of this study, we analyze three specific days in January 2023 with varying pollution levels, using variogram analysis to determine the temporal resolution required for accurate PM monitoring. Additionally, we explore the role of meteorological parameters such as wind patterns, temperature, and humidity in modulating PM fluctuations on the most polluted day. By combining high-resolution PM data with advanced statistical methods, this study aims to provide actionable insights into air quality monitoring practices, contributing to improved health risk assessments and policy recommendations.

The findings from this work will be instrumental in guiding the design of air quality monitoring systems with enhanced temporal granularity for the Joppa neighborhood. Such systems will enable the identification of short-term exposure risks, inform mitigation strategies, and ultimately contribute to better public health outcomes. Furthermore, by integrating variogram analysis with meteorological data, this study offers a novel approach to understanding the multifaceted factors influencing PM variability in urban environments.

2. Materials and Methods

2.1. Sensors Overview

The data for this study were collected from the Central Node located in the Joppa neighborhood of the Dallas–Fort Worth Metroplex, a region heavily impacted by industrial encroachment and severe pollution. The Central Node is a powered stationary sensor system. This monitor was designed by Dr. Lakitha Wijeratne of the MINTS Lab at UT Dallas. MINTS stands for the Multiscale Integrated Sensing and Simulation Lab, which focuses on developing intelligent sensing systems. The Central Node is one of these advanced sensing systems. It is a 24/7 distributed sentinel that monitors and posts environmental data frequently and consistently. This node is equipped with an array of sensor systems designed to monitor various environmental parameters, including PM levels recorded every second throughout the day.

This research focuses specifically on the IPS7100, the BME280, and the AirMar sensors within the MINTS Central Node Air Module [22,23]. Using the variogram model, our aim is to determine the optimal sampling interval for each PM size fraction measured by the IPS7100. Meteorological parameters play a critical role in influencing ambient PM concentrations [24–27]. So, we use meteorological data from the BME280 and AirMar sensors to analyze the relationship between PM sampling intervals and meteorological conditions. The sensors used in this study are detailed below (Figure 1):

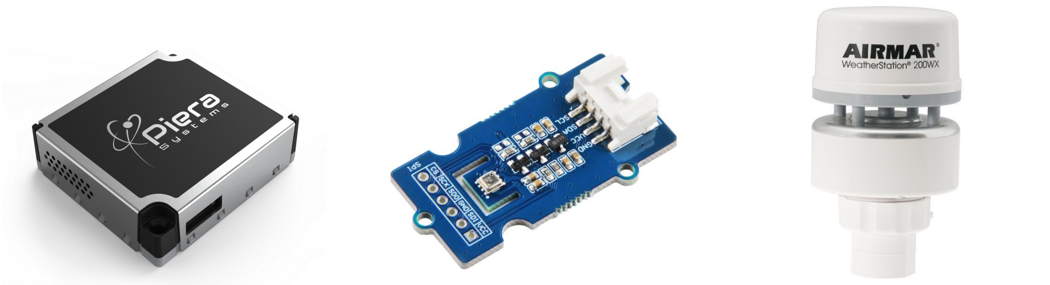

Figure 1. Sensors used in the MINTS Central Node Air Module: (**left**) IPS7100; (**middle**) Bosch BME280; and (**right**) AirMar Weather Station.

- *IPS7100* (manufactured by Piera Systems Inc., located in Ontario, Canada) [28] is an Intelligent Particle Sensor from Piera Systems. It is an optical counter that detects particulate matter size fractions ranging from $PM_{0.1}$ to $PM_{10.0}$. It outputs both the particle count and the concentration for each size fraction, separated into seven bins, one for each size fraction. Therefore, the sensor provides 14 output values. It measures size fractions such as $PM_{0.1}$, $PM_{0.3}$, $PM_{0.5}$, $PM_{1.0}$, $PM_{2.5}$, $PM_{5.0}$, and PM_{10}. The IPS7100 is compatible with UART (Universal Asynchronous Receiver/Transmitter) and I^2C (Inter-Integrated Circuit) protocols. UART is a straightforward, point-to-point communication protocol ideal for sending data asynchronously from a single sensor to a micro-controller. However, it is limited in scalability and is not well suited for systems requiring communication with multiple devices. In contrast, I^2C offers a more versatile solution by supporting multi-device communication through a shared two-wire bus. This makes I^2C an ideal choice for our design, enabling seamless data transmission and reception between the micro-controller and multiple sensors while minimizing wiring complexity and enhancing scalability. It operates at 65 mA and 5 V, making it suitable for low-power applications. The IPS7100 collects PM concentration and counts with a granularity of 1 s, making it ideal for high granularity time-series analysis.

- *Bosch BME280* (manufactured by Bosch Sensortec and sourced from Reutlingen, Germany) [29] measures climate data, including air temperature, atmospheric pressure, and relative humidity. It operates on a voltage range of 1.71 V to 3.6 V and supports the I^2C protocol. Data are collected at a granularity of 10 s. The BME280 provides a quick response rate with high accuracy and resolution while minimizing noise.
- *AirMar Weather Station* (manufactured by AirMar Technology Corporation and sourced from Milford, NH, USA) [30] is a reference sensor designed to measure parameters such as air temperature, barometric pressure, relative humidity, and wind speed. Using ultrasonic technology for wind measurement improves accuracy without requiring moving parts. Wind speed data are collected every 10 s.

2.2. Data Collection

The data collected for this study span from 1 January 2023, 06:00:00 UTC, to 1 February 2023, 05:59:59 UTC, and include measurements from the IPS7100, BME280, and AirMar sensors. The time stamp of the data collected is in UTC, and it will be later converted to the designated local time zone. The data for each sensor are collected as daily CSV files and later processed.

The PM data are captured at a fast sampling rate of 1 s for IPS7100. The 10 s interval for the meteorological sensors (BME 280 and AirMar) is specifically chosen to prevent latency issues that could arise from having multiple sensors communicating on the same bus to guarantee efficient data collection.

The sensors used in the study are prone to measurement errors. So the data collected for this study are checked based on the measurement range for each sensor. Table 1 below provides details of the sensors, including their parameters, units, measurement ranges (range), resolution/frequency of measurements and uncertainty.

Table 1. Specifications and performance of sensors used for air quality and meteorological measurements [28–30].

Sensor	Parameter	Unit	Range	Resolution (Seconds)	Uncertainty
IPS7100	$PM_{0.1}$	$\mu g/m^3$	0–6000	1	±10%
	$PM_{0.3}$	$\mu g/m^3$	0–6000	1	±10%
	$PM_{0.5}$	$\mu g/m^3$	0–6000	1	±10%
	$PM_{1.0}$	$\mu g/m^3$	0–6000	1	±10%
	$PM_{2.5}$	$\mu g/m^3$	0–6000	1	±10%
	$PM_{5.0}$	$\mu g/m^3$	0–6000	1	±10%
	$PM_{10.0}$	$\mu g/m^3$	0–6000	1	±10%
BME280	Air Temperature	°C	−40–85	10	±0.5 °C
	Atmospheric Pressure	hPa	300–1100	10	±1 hPa
	Relative Humidity	%	0–100	10	±3% RH
AirMar	Wind Speed	m/s	0 – 30	10	±2%

2.3. Variogram Analysis

2.3.1. Temporal Variogram Equation

Based on the variogram analysis of the collected PM data, we can characterize the frequency with which airborne particulate observations should be made to adequately resolve the PM time series. We utilize semi-variograms (also known as variograms) [18–20,31–37], a geostatistical model that calculates the autocovariance of the points in a sample. This method is used to identify the temporal fluctuations that occurred over a period of 24 h. We perform the analysis for each of the three distinctly polluted days separately. This allows us to better understand how the sampling period for each of the PM size fractions change over time.

The variogram is calculated as the average squared difference between points separated by a specific distance (time lag). The semi-variogram is half of the variogram, though the terms are often used interchangeably. The equation for a general temporal semi-variogram for the PM data takes the form

$$\gamma(\Delta t) = \frac{1}{2N} \sum_{i=1}^{N} [PM(t_i \pm \Delta\, t) - PM(t_i)]^2$$

where we have the following: $\gamma(\Delta t)$: represents the semi-variogram as a function of time lag.

N: represents the total number of data point pairs for which the variogram is constructed.

$PM(t_i)$: the PM concentration measured for the ith time.

$PM(t_i \pm \Delta t)$: the PM concentration measured before or after a time lag of Δt from the ith time.

2.3.2. Key Elements of a Variogram

The *nugget*, *sill*, and *range* are the three primary elements that make up a variogram. The *nugget* is the y-intercept of the semi-variogram vs. the lag time plot. It can also be defined as the value of the variogram, $\gamma(\Delta t)$, as Δt approaches 0. Ideally, when there are two simultaneous PM observations, they should have the same value. This would result in a nugget value of zero. However, in reality, this is often not the case. Measurement errors inherent to the sensor during sampling or temporal variations of PM at time gaps smaller than the sampling intervals introduce some randomness. This randomness causes slight differences between values that would otherwise be identical. As a result, the nugget is non-zero. This phenomenon is known as the *Nugget Effect* [38].

The *sill* is the point on the y-axis of the semi-variogram vs. the lag time plot, where the variogram starts to level off to a near-constant value. The leveling indicates that beyond that lag distance or time lag, the points become uncorrelated.

The *range* is the distance beyond which the variogram flattens out. Here, it is used to identify the temporal scale beyond which the data are no longer correlated. To obtain it, we find the x coordinate on the semi-variogram vs. the lag time plot that corresponds to 95% of the sill.

Among the attributes described above, range is of particular interest. When determining the range for each time step, sliding windows are an effective tool to use. For each time step and for a sliding window of a specific size, we find the temporal scale beyond which the correlation between measurements diminishes. To determine the range of our variograms, we first fit an empirical variogram to a theoretical model. We then identify the point on this theoretical variogram curve where the correlation effectively reaches zero (or plateaus). The x-coordinate corresponding to this point is defined as the range of the variogram.

Here, the theoretical variogram model used for fitting is an exponential model. Exponential Variograms [39] have been useful in capturing high variability, which is commonly seen in PM data due to localized sources of pollution. It follows the equation

$$\gamma(\Delta t) = C_0 + C(1 - e^{(\frac{-h}{a})})$$

where $\gamma(\Delta t)$ is the semivariance in the lag time Δt, C_0 is the *nugget*, C is the *sill* minus the *nugget*, a is the *range*, and h is the lag time (in our case, it varies from 0 to 5 min).

The model accuracy can be found using the Mean Squared Error (MSE) which is the averaged squared difference between the actual value and the predicted value. The lower the MSE, the better the model will be, indicating smaller errors. The MSE is given by

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

The goodness of fit of the empirical variogram with the theoretical model can also be assessed using the Coefficient of Determination (R^2). This helps in estimating the model performance. The R^2 value can range from 0 to 1, with values closer to 1 indicating a better fit. The R^2 score is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2}$$

where we have the following:

- $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ is the Residual Sum of Squares (RSS);
- $\sum_{i=1}^{N}(y_i - \bar{y})^2$ is the Total Sum of Squares (TSS);
- y_i represents the empirical value;
- $\hat{y}_i$ represents the theoretical value;
- $\bar{y}$ is the mean of all empirical values.

2.3.3. Rolling Temporal Variogram

To determine the variability in the PM measurement time throughout the day at a high granularity, we calculate the *range* for each 1 s time step. To do that, we use sliding windows of 15 min and a time lag of 5 min. Starting from the first second of the day, we create a rolling temporal variogram using the equation

$$\gamma(\Delta\, t) = \frac{1}{2N}\sum_{j=1,i=j}^{M,N+j}[PM(t_i + \Delta\, t) - PM(t_i)]^2$$

The above equation employs a 15 min sliding window that traverses the entire day, week, month, or year of the data at a time step of 1 s. The goal here is to obtain a range value for the shortest resolution possible. Here, i is the variable that defines the width of the window (15 × 60 s = 900 s). j is the variable that decides the duration for which the window must traverse. It helps in generating the moving variogram. It starts off at 1 s and allows our algorithm to generate variograms over each pair of points in time that are 900 s apart (e.g., (1, 901), (2, 902), (3, 903), and so on). N is the total number of pairs of concentrations of PM across which the variogram is constructed. M represents the first value of the last sliding window. This will be equal to the time duration for which PM levels are measured minus the width of one sliding window (15 min). For a day's data, it will be equivalent to

$$(24 \times 60 \times 60)\text{ s} - (15 \times 60)\text{ s} = 85{,}500\text{ s}$$

where Δt is the time lag, and in our study it is 5 min. $PM(t_i)$ is the PM data measured for ith time. $PM(t_i + \Delta t)$ is the PM data measured after a time delay of Δt from the ith time (because we are considering a forward traversing rolling temporal variogram).

2.4. Machine Learning Setup

The temporal variogram is applied to obtain the measurement time for all PM concentrations (from $PM_{0.1}$ to $PM_{10.0}$) for the month of January 2023. After estimating the PM measurement time for each size fraction, we proceed to assess the impact of various meteorological parameters on this measurement time. For this part of the study, we use parameters such as *air temperature* (referred to as temperature), *atmospheric pressure* (referred to as pressure), *relative humidity* (referred to as humidity), and *wind speed*, all of which are recorded by the Central Node. Nonlinear and nonparametric regression models are used to estimate the measurement time for each size fraction [40–42], using meteorological parameters as input. Wind speed, humidity, temperature, and pressure data from the Central Node are recorded at 10 s intervals. The estimated measurement time for each size fraction is considered as the target variable. We develop seven models corresponding to each of the seven size fractions, using the measurement time as the target variable and meteorological parameters as predictors.

For modeling purposes, we use ensemble learning methods like *Random Forest* [43]. Random Forests are often known for their versatility and flexibility. They are widely used in air quality monitoring because they are good at handling high-dimensional data and are robust to noise. They also produce accurate and interpretable results. Random Forests use decision trees, also known as Classification And Regression Trees (CART), as their base learners. Individual decision trees are prone to over-fitting, as they may create individual leaf nodes for each observation. In Random Forest, each tree is fitted using a bootstrapped sample from the original data, and the results are averaged later. This approach is known as bagging. Random Forests also use random subsets of features to split each tree, preventing important features from appearing at the top of each decision tree. This reduces the overfitting and the correlation between trees.

The best Random Forest Regression model is chosen after hyperparameter [44,45] optimization through Random Search CV. Random Search CV combines Random Search and cross validation. Instead of testing all possible values for each hyperparameter, it randomly chooses a subset of combination to evaluate. This approach is useful when multiple hyperparameters are involved. The cross validation part in Random Search CV helps in obtaining a generalized model. In cross validation, we divide the entire data set into multiple folds. In each round, one fold is used for validation, and the other folds are used for training. This process is repeated for each fold, and the results are averaged to obtain a reliable estimate of the model's performance. The model with the best R^2 (or lowest MSE) value is chosen for the test. This helps in model performance when it comes to unseen data.

The feature importance ranking in Random Forest is a key aspect which helps us determine which features contribute the most to the model's goodness of fit. Random Forest ranks its features based on average decrease in variance for regression when a feature is used to split the data. However, this becomes an issue when features are highly correlated, as Random Forest tends to distribute importance across these correlated features. Hence, we choose a model-agnostic feature ranking method such as permutation ranking [46]. Here, the performance of the model (R^2 or MSE) is first calculated on the original data set. Now, for each feature, the values are randomly shuffled among that feature while keeping the other features and target value as is. The model performance is then evaluated using the shuffled feature. If there is a large performance drop, it shows that the feature is very important, while a small drop indicates that the feature is less relevant. If features are correlated to each other, then shuffling them will result in a smaller dip. This can help the model to rely on other correlated features to capture similar information.

3. Results

3.1. Preliminary Analysis

The raw data collected are consolidated into a single data set for each sensor. For each data set, the time stamp is converted to Central Time to align with the local timezone. Duplicate timestamps are identified and consolidated by averaging the values. Rows containing *missing* or *NA* values are removed to ensure data consistency. Invalid numerical values, such as any string-numerical columns in the data set, are converted to numerical format. To maintain temporal consistency, the gaps in the timestamps (1 s intervals for the IPS7100 and 10 s intervals for the meteorological sensors) are imputed using forward and backward imputation methods. The cleaned data sets are then verified to ensure that they contained the correct number of records per day—86,400 records per day for the IPS7100 and 8640 records per day for the BME280 and AirMar sensors. Finally, the cleaned data are stored as CSV files, with timestamps and categorization based on the relevant sensor, ready for analysis.

The next step is to analyze the PM data we cleaned and check how they compare to the various PM standards set by government agencies. The EPA is the United States Environmental Protection Agency. They monitor and report $PM_{2.5}$ and $PM_{10.0}$. Similarly, the World Health Organization (WHO) and the European Union (EU) have defined precise standards for these categories of PM. Their standards are shown in Table 2 below [47–49].

Table 2. Global air quality standards for $PM_{2.5}$ and $PM_{10.0}$.

Pollutant	Standard	Daily Limit (μg/m^3)	Annual Limit (μg/m^3)
$PM_{2.5}$	EPA	35	9
	EU	25	10
	WHO	15 (for 3–4 days)	5
$PM_{10.0}$	EPA	150	None
	EU	45	20
	WHO	45	15

To compare our data with PM standards, we select three distinctly polluted days from Joppa in January 2023. This selection process utilizes a calendar plot, which provides a clear and intuitive visualization of Joppa's PM data by displaying daily average concentrations in a calendar format. This approach facilitates the identification of patterns and trends throughout the month. $PM_{2.5}$ is chosen for this study, as it is one of the most critical pollutants reported by the EPA and will also be used in the variogram analysis results. For generating the calendar plot, we use the *openair* package [50] in R, a powerful toolkit designed for air quality and atmospheric data analysis. Figure 2 illustrates the calendar plot of daily average $PM_{2.5}$ concentrations for January 2023.

Lighter colors on the above plot such as yellow indicate lower pollution levels, whereas darker colors such as orange and red signify higher pollution levels.

The following three days are selected from the calendar plot and ranked according to their daily mean $PM_{2.5}$ concentrations:

- Most polluted—2 January (31.87 μg/m^3);
- Intermediately polluted—29 January (17.97 μg/m^3);
- Least polluted—24 January (4.8 μg/m^3).

To effectively illustrate the comparison of PM data from our selected days with the global PM standards, we create time series plots for the selected days, overlaying the global PM standards for reference. The methodology and computational framework required for this study, as well as subsequent studies, will be implemented in Europa [51], the High Performance Computing resource provided by University of Texas at Dallas's Office of

Information Technology. For the selected days, the $PM_{2.5}$ concentration (the most commonly reported PM size fraction globally) is plotted at both the hourly and second-wise resolutions. Table 3 shows the statistical summary of the $PM_{2.5}$ concentration for the selected days.

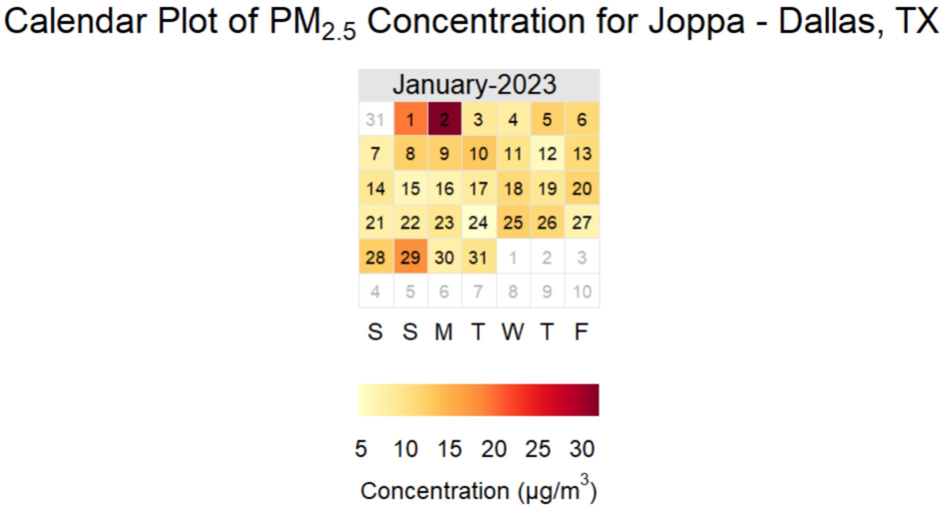

Figure 2. Variation in the daily average $PM_{2.5}$ concentration recorded in the Joppa neighborhood of Dallas, Texas, during January 2023. The calendar format displays the daily average concentrations using a gradient of color tones, where darker shades indicate higher pollution levels, and lighter shades represent lower levels. The color bar below the calendar provides a reference scale for the concentrations in µg/m^3.

Table 3. Statistical summary of $PM_{2.5}$ concentration for the selected days.

Date	2 January	29 January	24 January
Count	86,400	86,400	86,400
Mean (µg/m^3)	31.87	17.97	4.80
Std Dev (µg/m^3)	21.91	8.93	3.35
Min (µg/m^3)	0.00	3.29	0.07
25% (µg/m^3)	25.40	10.62	1.34
50% (µg/m^3)	31.23	17.23	5.18
75% (µg/m^3)	36.90	22.04	7.19
Max (µg/m^3)	1,265.41	135.04	28.06

The $PM_{2.5}$ time series plots are given below (Figure 3):

The plots include legends denoting the daily average $PM_{2.5}$ concentration limits set by key regulatory organizations. Each organization's standard is represented with distinct color-coded dashes:

- **EPA Standard**: 35 µg/m^3 (red dash);
- **EU Standard**: 25 µg/m^3 (blue dash);
- **WHO Standard**: 15 µg/m^3 (green dash).

We now analyze the peak exposure in each second-wise plot compared to the hourly plots. The table below shows the comparison.

The time series plot in Figure 3 and the standard deviation of second-wise $PM_{2.5}$ data (21.91 µg/m^3) reported in Table 4 for 2 January highlight the significant variability, with the maximum value (1265.41 µg/m^3) far exceeding the EPA threshold. This contrasts with the much lower standard deviation (7.43 µg/m^3) in the hourly data, reflecting the dampening effect of averaging. Similarly, for 29 January, the second-wise data have a standard deviation of 15.32 µg/m^3, again showcasing the higher variability captured at finer temporal scales compared to hourly analysis (which is 8.36 µg/m^3).

The second-wise data, as shown for 2 January and 29 January, capture short-lived but intense peaks in $PM_{2.5}$ levels that are otherwise masked in the hourly averages. Around 27,949 peaks on 2 January exceeded 35 µg/m^3 (388 rose above 100 µg/m^3), with the highest peak recorded at 1265.41 µg/m^3. In contrast, the hourly data for the same day identified only 9 instances exceeding 35 µg/m^3 (0 values above 100 µg/m^3), with a maximum hourly value of just 40.33 µg/m^3.

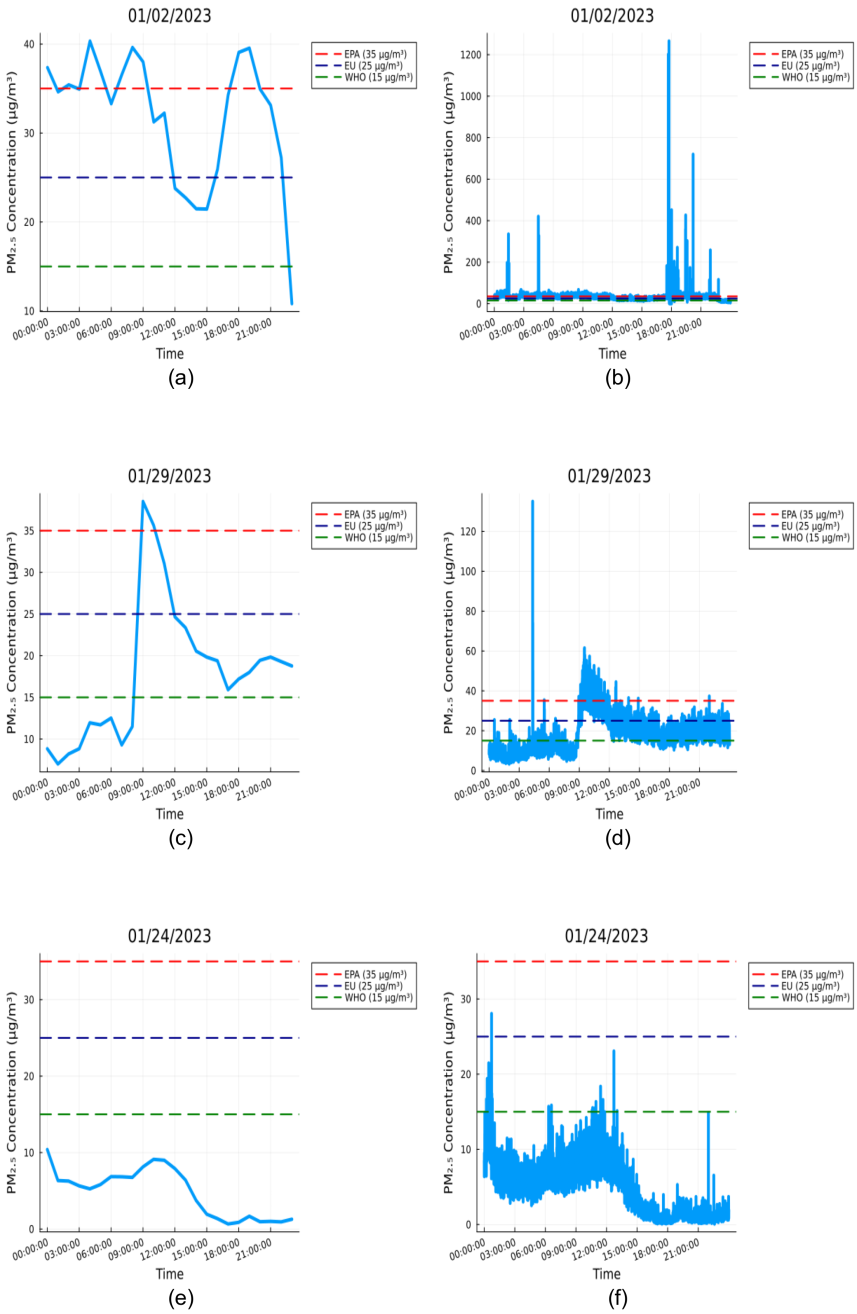

Figure 3. Time series plots of $PM_{2.5}$ concentrations for three days in January 2023: the most polluted day (2 January 2023), the intermediate polluted day (29 January 2023), and the least polluted day (24 January 2023) in Joppa, Dallas, Texas. (**a**,**c**,**e**) Hourly $PM_{2.5}$ concentrations for these days. (**b**,**d**,**f**) The $PM_{2.5}$ concentrations at one-second intervals. The horizontal dashed lines indicate air quality standards set by the EPA (35 µg/m^3, indicated by a red dashed line), the EU (25 µg/m^3, indicated by a blue dashed line), and the WHO (15 µg/m^3, indicated by a green dashed line).

Table 4. Summary of $PM_{2.5}$ peak exposure across selected days.

Date	2 January	29 January	24 January
Count of Peaks in Second-wise $PM_{2.5}$ Data Above EPA Limit	27,949	5039	0
Maximum $PM_{2.5}$ Concentration in Second-wise Data (μg/m^3)	1265.41	135.04	28.06
Standard Deviation of Second-wise $PM_{2.5}$ Data (μg/m^3)	21.91	15.32	5.45
Count of Peaks in Hourly $PM_{2.5}$ Data Above EPA Limit	9	2	0
Maximum $PM_{2.5}$ Concentration in Hourly Data (μg/m^3)	40.33	38.51	10.40
Standard Deviation of Hourly $PM_{2.5}$ Data (μg/m^3)	7.43	8.36	3.16

Table 4 shows that for 29 January, the second-wise data recorded 5039 peaks above 35 μg/m^3 (6 peaks exceed 100 μg/m^3), with a maximum peak of 135.04 μg/m^3. However, the hourly data for this day identified only two instances above 35 μg/m^3 (0 values exceeding 100 μg/m^3) and a maximum hourly value of 38.51 μg/m^3. Although the magnitude of the peaks on 29 January was less extreme than that for 2 January, the second-wise data still reveal notable pollution events that remain undetected in the hourly averages.

Conversely, the data for 24 January, a day with relatively clean air, highlight the absence of such transient peaks. Both second-wise and hourly analyses show no values exceeding 35 μg/m^3, and the maximum hourly value is only 10.40 μg/m^3.

The substantial difference in data across various temporal resolutions highlights the smoothing effect of hourly averaging, which significantly diminishes the visibility of transient pollution events. These events can have critical implications for public health. The significant variance observed underscores the need for advanced methods, such as variograms, to assess PM concentration variability across different days and determine the optimal sampling frequency for each day. Rolling temporal variograms are particularly useful, as they can effectively track abrupt variations in the highest temporal resolution, providing a more accurate understanding of pollution dynamics.

3.2. Case Study 1: Estimating the PM Measurement Time Using Rolling Temporal Variograms

3.2.1. Analyzing Temporal Variograms for Different Windows of $PM_{2.5}$ Concentration Across Three Distinctly Polluted Days

Rolling temporal variograms are plotted for each window to capture the variability in the data at each time step. The rolling aspect of the temporal variogram allows us to analyze these variations dynamically across the data set. The chosen time step is 1 s, as it corresponds to the highest resolution available from the Central Nodes PM data. The window duration is set to 15 min, with a maximum time delay of 5 min. The first window begins at 00:00:00 on the selected day and spans 15 min, ending at 00:15:00. The second window starts at 00:00:01 and ends at 00:15:01. This process is repeated iteratively, with each subsequent window shifting by one second, until the final window, which ends at 23:59:59 for the selected day. This rolling window approach ensures that temporal variograms provide a continuous and detailed representation of variability in $PM_{2.5}$ concentration throughout the day.

Figure 4 shows the variograms for two different windows (the 10th and 50,000th) of $PM_{2.5}$ pollution on three distinctly polluted days in January 2023. These windows are chosen at random from all the 85,500 (for each day).

For a selected window, the cross defines the point where the correlations stops. For the various windows, we can see that the correlation stops at different values of time lag. If we consider the 50,000th window in Figure 4b,d,f, we can see that the range of the variogram is 40.9, 76.3, and 193.8 s, respectively. These are the time scales at which the $PM_{2.5}$ concentration must be measured during the 50,000th window for the highest, intermediate, and least polluted days to capture all the variations. Similarly, we can find the measurement time of the $PM_{2.5}$ concentration across these days for all windows (85,500 windows for each day). Apart from the range and sill, the plots also show the nugget. We notice that the nuggets are very close to zero, indicating a low tolerance or error in the measured $PM_{2.5}$ concentration. We also note the the R^2 value (Coefficient of Determination) for the various windows across the day. Across the different windows shown in Figure 4, the lower bound of R^2 is 0.6, and the upper bound is 0.9. This indicates that the theoretical model (Exponential Variogram) fits decently with most of the Empirical Variogram data.

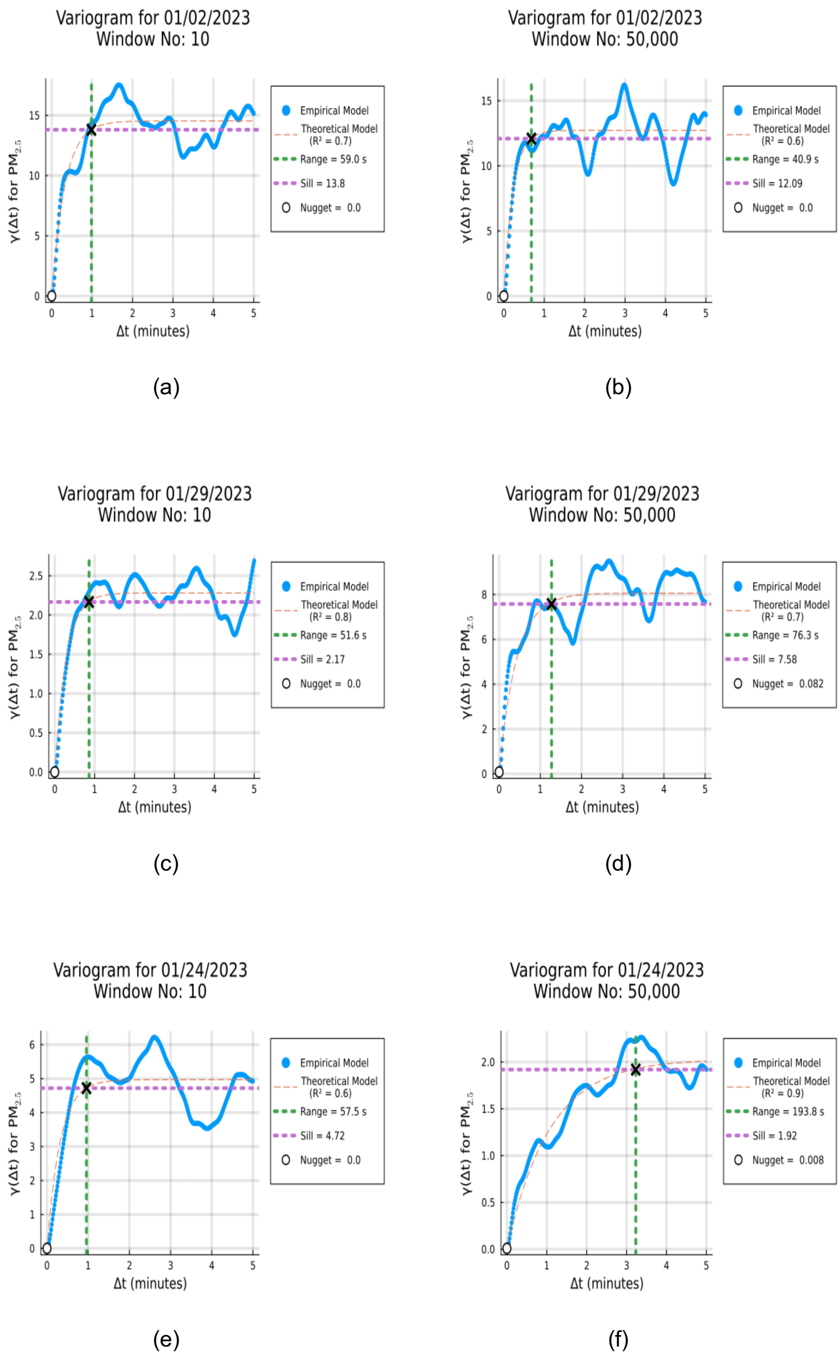

Figure 4. Variogram plots of $PM_{2.5}$ concentration in decreasing order of pollution (2 January 2023, 29 January 2023, and 24 January 2023). (**a**,**c**,**e**) Variograms for Window Number 10 on the respective days. (**b**,**d**,**f**) Window Number 50,000 on the same days. Window Numbers 10 and 50,000 were randomly selected from a total of 85,500 windows for each day. The theoretical variogram model is represented by yellow dashed lines, while the empirical model is depicted in blue. The cross indicates the coordinates of the sill and range, as shown in the legend, and the white dot represents the nugget. The R^2 value for each plot is also included in the legend.

3.2.2. Identifying the Optimal Measurement Intervals for $PM_{2.5}$ Concentration on Days with Varying Pollution Levels

Based on the Rolling Variogram analysis across 85,500 windows for each day selected in the study, a Probability Density Function (PDF) is plotted for the $PM_{2.5}$ measurement time as shown in Figure 5. This analysis helps identify the most frequent measurement time for $PM_{2.5}$, which corresponds to the mode (x-coordinate) of the peak y-value in the PDF. The mode provides a critical insight into how quickly the sensor should measure/sample the $PM_{2.5}$ concentration to effectively capture all variations in a given neighborhood.

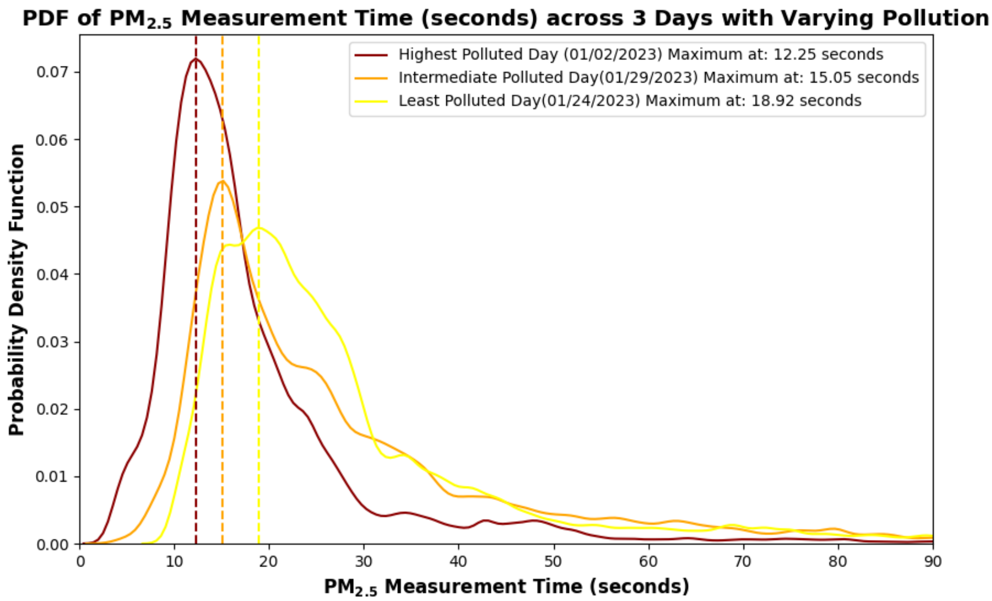

Figure 5. Probability Density Plot of $PM_{2.5}$ measurement time across 3 distinctly polluted days of January 2023 for Joppa, Dallas, Texas. The peak measurement times are marked at 12.25 s, 15.05 s, and 18.92 s for the highest, intermediate, and least polluted days, respectively.

Considering days with varying pollution levels, the most polluted day is expected to exhibit the shortest peak measurement time due to its larger variance, necessitating a faster sampling rate. Conversely, the least polluted day is expected to show the longest peak measurement time, with moderately polluted days falling in between based on their variance levels. This relationship is derived from the computation of the variogram, which is fundamentally based on variance. On 2 January, the most polluted day in the study, the $PM_{2.5}$ concentration would require a significantly faster measurement rate compared to other days.

The optimal sampling time of $PM_{2.5}$ follows the order below:

$$2 \text{ January} < 29 \text{ January} < 24 \text{ January}.$$

Table 5 presents the estimated peak measurement times for each day in increasing order along with their corresponding 95% confidence intervals (peak measurement time CI) [52,53]. The confidence intervals for the peak measurement times are calculated using nonparametric bootstrapping. In this method, the estimated $PM_{2.5}$ sampling times are repeatedly resampled with replacement, and the mode is calculated for each resample. These intervals indicate that there is a 95% certainty that the optimal sampling times for these days lie within the specified ranges. The confidence intervals highlight the precision of the estimated optimal sampling times. Narrow intervals indicate higher precision in the estimation of $PM_{2.5}$ sampling times. Since the confidence intervals for the sampling times across the days do not overlap, we can conclude that there is a significant difference in the optimal $PM_{2.5}$ sampling times among the days.

Table 5. Estimated $PM_{2.5}$ sampling time for 3 distinctly polluted days in increasing order.

Date	Peak Measurement Time (Seconds)	Peak Measurement Time CI (Seconds)
2 January	12.25	[12.13, 12.44]
29 January	15.05	[14.99, 15.18]
24 January	18.92	[18.65, 19.29]

3.2.3. Assessing Measurement Errors and $PM_{2.5}$ Variability Across Days with Varying Pollution Levels

Table 6 represents the estimated peak nugget and peak nugget confidence intervals for each day in increasing order of estimated $PM_{2.5}$ sampling time.

Table 6. Estimated $PM_{2.5}$ Measurement Error for 3 distinctly polluted days

Date	Peak Nugget $(\mu g/m^3)^2$	Peak Nugget CI $(\mu g/m^3)^2$
2 January	0	[0, 0]
29 January	0	[0, 0]
24 January	0	[0, 0]

The peak nugget represents the mode of all estimated nugget values from the 85,500 windows. It corresponds to the most recurring measurement error or noise associated with the sensor. The unit of the nugget is the same as that of the sill, measured in $(\mu g/m^3)^2$. The peak nugget confidence interval (CI) indicates that, with 95% confidence, the true mode of the nugget distribution lies within the given range of values. In this case, for all three days, the CI is [0, 0], indicating that across days with varying pollution levels, the measurement error consistently approaches 0.

In the context of our study, the zero peak nugget value and its 95% confidence interval across all three days suggest low measurement error and high consistency in the sensor's readings. This indicates that the temporal structure of the data is well-defined, enabling the reliable determination of the optimal $PM_{2.5}$ measurement time. Consequently, this highlights the IPS7100 sensor's reliability in capturing accurate and consistent air quality data.

3.2.4. Analyzing $PM_{2.5}$ Variability Across Days with Distinct Pollution Levels using Different Sampling Intervals

Having determined the optimal sampling times and measurement errors for $PM_{2.5}$ for each of the selected days, we use these peak measurement times (shown in Table 5) to average the data and compare the variance retained relative to the highly granular 1 s data. The table below summarizes the standard deviation captured for each day using 1 s sampling, 1 h sampling, and the optimal sampling intervals. This analysis will ensure that the relevant variance is preserved while reducing the temporal resolution.

In Table 7, the columns *SD (1s)* and *SD (1hr)* represent the standard deviations of $PM_{2.5}$ concentrations for each day when the data are sampled at 1 s and 1 h resolutions, respectively. *SD (Peak Time)* represents the standard deviations of $PM_{2.5}$ data calculated for the optimal sampling intervals (peak measurement times) of 12.25 s, 15.05 s, and 18.92 s for the highest, intermediate, and least polluted days, respectively.

Table 7. $PM_{2.5}$ standard deviations (in $\mu g/m^3$) and standard deviation ratios for different temporal resolutions.

Date	SD (1 s)	SD (1 h)	SD (Peak Time)	SD (Peak Time)/SD (1 s)	SD (1 h)/SD (1 s)
2 January	21.91	7.44	19.11	0.87	0.34
29 January	8.93	8.36	8.83	0.99	0.94
24 January	3.35	3.16	3.31	0.99	0.94

The ratio *SD (Peak Time)/SD (1 s)* quantifies the proportion of variance retained from the original 1-second granularity when the data are averaged at the optimal sampling intervals. Similarly, the ratio *SD (1 h)/SD (1 s)* quantifies the proportion of variance retained from the original 1 s granularity when the data are averaged at 1 h intervals. A standard deviation ratio close to 1 indicates how well the variance is retained at a coarser interval, providing a measure to evaluate how effectively the variability in $PM_{2.5}$ concentrations is preserved at coarser temporal resolutions [54].

For 2 January, the standard deviation for the optimal sampling interval is 19.11 $\mu g/m^3$, which is very close to the standard deviation of 21.91 $\mu g/m^3$ observed at the highly granular 1 s resolution. In contrast, the standard deviation for the 1 h resolution is only 7.44 $\mu g/m^3$. This indicates that the optimal sampling interval retains nearly three times more variability than the 1 h resolution. This is further supported by comparing the ratio of SD (Peak Time)/SD (1 s) to SD (1 h)/SD (1 s) for 2 January. These findings demonstrate that using a 12.25 s granularity effectively preserves a significant portion of the variance captured at the 1 s resolution. For 29 January and 24 January, the standard deviations for both the hourly resolution and the optimal sampling intervals are relatively similar. However, the standard deviations and the ratios of standard deviations for the optimal sampling intervals are slightly higher than those for the 1 h resolution. This indicates that while the optimal sampling intervals provide an improvement over hourly resolution in capturing variability, the differences are less pronounced compared to 2 January, likely due to the lower overall variability on these days.

We have established an efficient methodology to capture the variations in $PM_{2.5}$ concentrations for distinctly polluted days using this approach. Similarly, this method is applied to other size fractions measured by the IPS7100 (such as $PM_{0.1}$, $PM_{0.3}$, $PM_{0.5}$, $PM_{1.0}$, $PM_{5.0}$, and PM_{10}) using the Joppa data set. Based on the data obtained from this analysis, we aim to investigate how these optimal sampling times are influenced by meteorological parameters and estimate them using climate data from the BME280 and AIRMAR sensors. To achieve this, we employ machine learning models. This study focuses on predicting the sampling times for all PM size fractions on the highly polluted day (2 January 2023) of January 2023.

3.3. Case Study 2: Estimating PM Measurement Time for January 2023's Most Polluted Day Using Ensemble Machine Learning Models

The input features for this study comprise the meteorological measurements, such as temperature, humidity, pressure, and wind speed, recorded by the Central Node in Joppa on 2 January 2023. The target variables denote the PM measurement intervals in minutes for various size fractions, spanning from 0.1 μm to 10.0 μm, which are estimated using the temporal variogram model. The measurement intervals are in minutes instead of seconds to reduce variance and overfitting. The best performance is attained through the utilization of an ensemble machine learning (ML) model called the Random Forest Regressor (RFR). Since the resolution of the meteorological parameters is 10 s, the number of data points that are chosen is under 8640 (86,400 s/10). The train test split used is 80:20. We use Python's *scikit-learn* learn [55] library to implement the Random Forest Regression. To optimize the model performance and identify the best performing model, hyperparameters are tuned by Random Search with 5-fold cross validation.

The following hyperparameters from *scikit-learn*'s Random Forest Regressor are considered for tuning:

- *n_estimators*: The number estimators are chosen to range between *100*, and *500*.
- *max_features*: This is the maximum number of features to be considered at each split. We select between *sqrt* and *log2*. This is performed to reduce overfitting.

- *max_depth*: The maximum depth of each tree could be *10*, *20*, *30*, *40*, or *None*. Here, *None* indicates unrestricted depth, which means all nodes are expanded until all leaves are pure or until they contain fewer than the minimum number of samples required for a split.
- *min_samples_split*: This is the minimum number of samples required to split a node, and it has a default value of *2*. Here, the optimization method could choose from *2*, *5*, or *10*.
- *min_samples_leaf*: This is the minimum number of samples required to be at the leaf node. By default, it is *1*; here, for tuning purposes, it is chosen from *1*, *2*, or *4*.

After tuning, we find that the same set of hyperparameters yields the best results across all the RFR models. The optimal hyperparameters are as follows:

- n_estimtors: 300;
- max_features: log2;
- max_depth: None;
- min_samples_split: 2;
- min_samples_leaf: 1.

The plots below depict the performance of the RFR models for these hyperparameters:

Figure 6 shows the results of the hyperparameter optimized RFR models. The x axis for each plot shows the actual PM measurement time estimated by the variogram model, while the y axis represents the predicted values from the RFR model. For a perfect fit, the scatter plot will be a straight line with a slope of 1 as represented by the 1:1 line. From Figure 6 panels (a) to (g), we can see that our model performs very well. The blue dots are the data (80% of values) we used to train the RFR model and the orange dots are the independent validation data (20% of values) we used to test the RFR model.

The distributions on either axis of the scatter plot from Figure 6 represent the marginal distributions of the predicted PM measurement time and the actual PM measurement time. These distributions are independent of each other and provide an insight into the spread and frequency of values for both the training and test sets. As seen from the scatter plots for all PM size fractions measurement time, the actual test values (y) and the predicted test sets ($\hat{y}$) show similar distributions and spread. This similarity indicates that the model fits the data well. The count of each size fractions train test split is provided in the legend, along with the goodness of fit (R^2 value).

Figure 7 shows the feature importance ranking using permutation importance for each of the size fractions (panels (a) to (g)). The bar plots for the feature importance metric are sorted in decreasing order of importance, indicating that the most important feature for the model is at the top, followed by the second most important feature, and so on. From the panels for each size fraction, it can be observed that for all the size fractions less than 2.5 µm, the top features (Figure 7 panels (a) to (d)) are temperature, humidity, and pressure (atmospheric pressure). However, as the particle size increases, wind speed begins to gain importance (Figure 7 panels (e) to (g)). This may be associated with processes like saltation [56,57], where the higher wind speeds may generate sufficient kinetic energy to force larger particles to lift off from the ground, travel a short distance, and then fall back down.

Table 8 shows the results of the evaluation metrics (R^2 and MSE in seconds^2) for each model. We can observe that the testing data perform really well, with the R^2 values being close to those of the training data, though slightly less. The slight difference suggests a good fit, and it shows that the models generalize well and capture all the variations effectively for each size fraction. The lower MSE of the models for both training and testing suggests that the models have minimal errors.

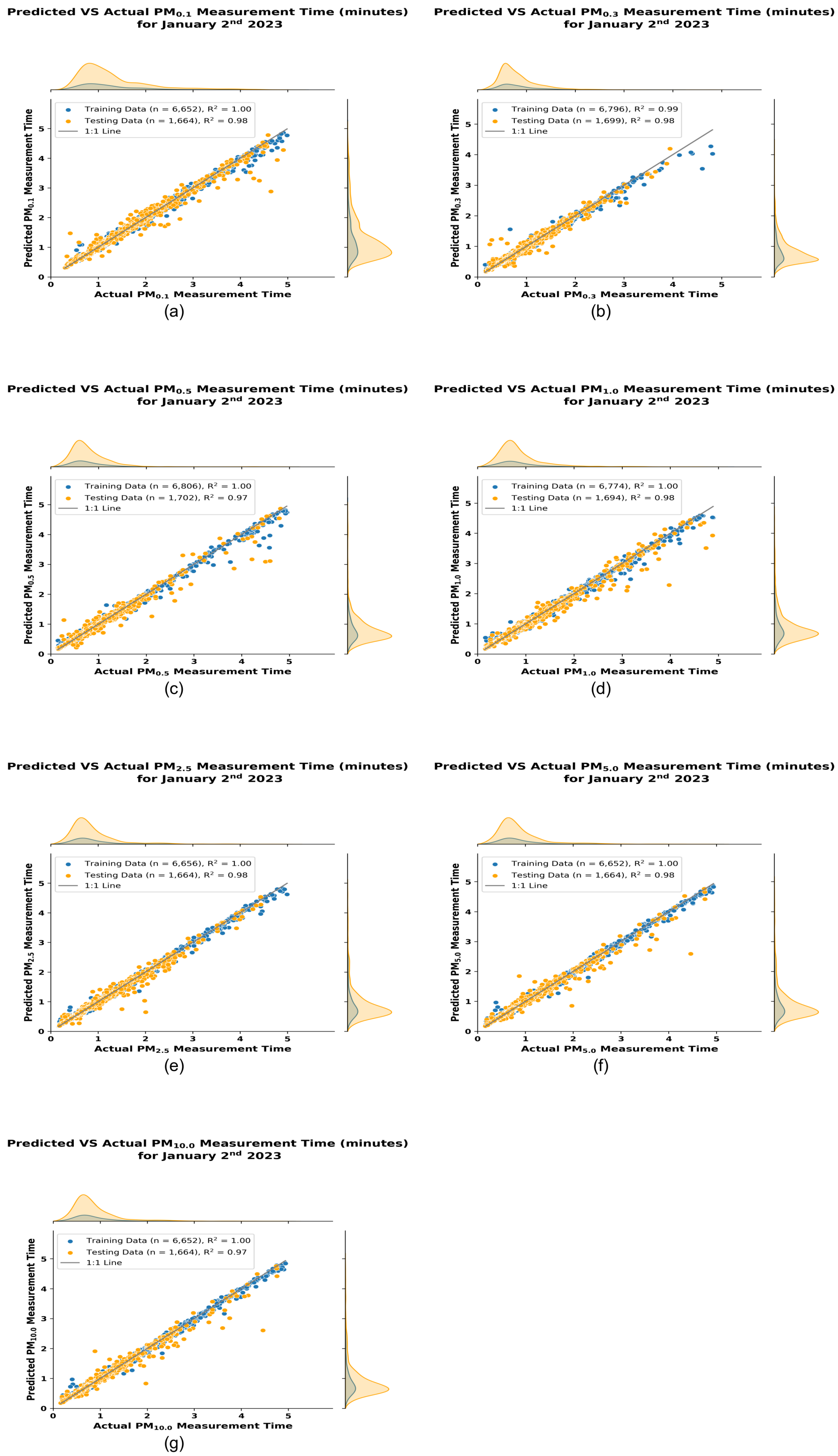

Figure 6. Scatter plots of the hyperparameter-optimized Random Forest Regression models for PM measurement time (minutes) on 2 January 2023, in Joppa, Dallas, Texas. Panels (**a**–**g**) correspond to the measurement times of $PM_{0.1}$, $PM_{0.3}$, $PM_{0.5}$, $PM_{1.0}$, $PM_{2.5}$, $PM_{5.0}$, and $PM_{10.0}$, respectively. The blue dots represent the training data, while the orange dots represent the testing data. The R^2 values and the marginal distributions for both training and testing data are provided, along with the 1:1 reference line for evaluating prediction accuracy.

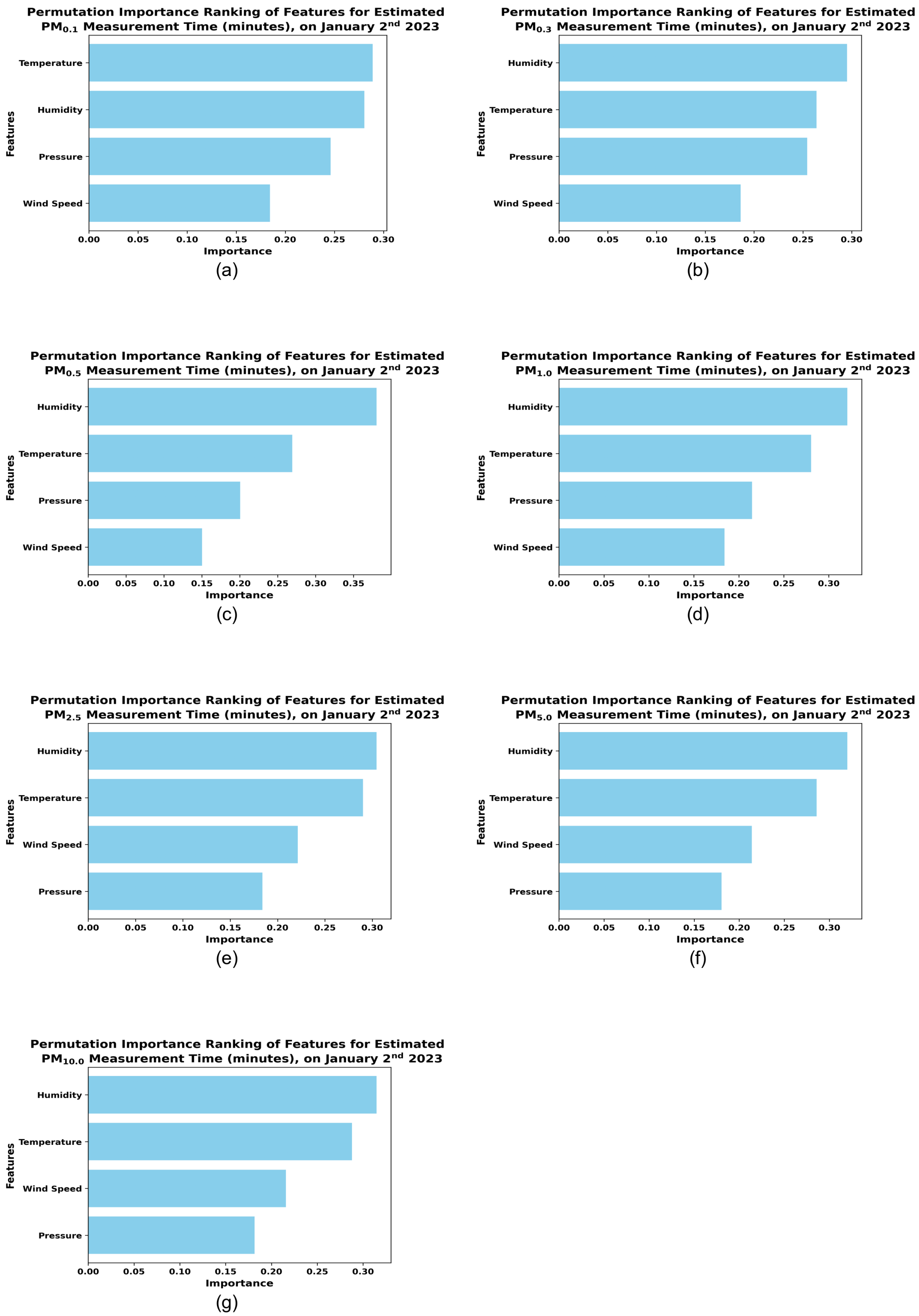

Figure 7. Permutation importance rankings of features influencing estimated PM measurement time (in minutes) for various particulate matter (PM) size fractions on 2 January 2023, in Joppa, Texas. Panels (**a**–**g**) correspond to the measurement times of $PM_{0.1}$, $PM_{0.3}$, $PM_{0.5}$, $PM_{1.0}$, $PM_{2.5}$, $PM_{5.0}$, and $PM_{10.0}$, respectively. Each bar chart displays the relative importance of meteorological parameters—humidity, temperature, pressure, and wind speed—in influencing the measurement time for each PM size fraction.

Table 8. Coefficient of Determination (R^2) and Mean Squared Error (MSE) for the machine learning models.

PM Measurement Time	Training R^2	MSE Training (s^2)	Testing R^2	MSE Testing (s^2)
$PM_{0.1}$	1.00	0.0015	0.98	0.0118
$PM_{0.3}$	0.99	0.0011	0.98	0.0051
$PM_{0.5}$	1.00	0.0012	0.97	0.0091
$PM_{1.0}$	1.00	0.0011	0.98	0.0102
$PM_{2.5}$	1.00	0.0009	0.98	0.0061
$PM_{5.0}$	1.00	0.0009	0.98	0.0085
$PM_{10.0}$	1.00	0.0009	0.97	0.0087

4. Discussion

The variability in particulate matter (PM) concentrations at the neighborhood scale highlights critical challenges in environmental monitoring and public health protection. Traditional air quality monitoring methods, which often rely on fixed hourly observation windows, can miss critical short-term variations in air quality as evidenced by the high standard deviation values and peaks in PM concentration data for 2 January and 29 January (refer to Figure 3 and Tables 3 and 4 from Section 3.1). This finding aligns with studies by Harrison et al. [36,37] from MINTS Lab, which emphasize the need for finer spatial and temporal resolutions in neighborhood-scale air quality monitoring. Additionally, the work of Liu [58] highlights the importance of high-resolution temporal data in improving our understanding of the variability and dynamics of airborne particulates. However, he also warns that excessively high resolution can present notable challenges, such as increased data storage requirements and higher bandwidth demands for data transfer.

This implies that a one-second resolution from the Central Node may not be the most efficient approach. Instead, the temporal variogram results indicate a more practical sampling interval of 12.25 s, which falls within the 95% confidence interval of [12.13 s, 12.44 s]. This interval is particularly suitable for highly polluted days, such as 2 January. Furthermore, the peak nugget and its 95% confidence interval for all three selected days, including the highly polluted day, are 0 and [0, 0], respectively. These estimates indicate a high degree of temporal correlation in the data and minimal measurement errors, affirming the reliability of the IPS7100 sensor in accurately capturing temporal variations in PM measurements. Notably, the standard deviation retained by the 12.25 s sampling interval is 0.87 times that of the 1 s resolution, demonstrating that it achieves comparable data variability while significantly improving efficiency. Adopting this optimal sampling interval can further reduce the Central Node's power consumption by approximately 2 h, making it an ideal choice for monitoring on such days.

The public health significance of this issue is further supported by Wijeratne [59], who advocates for low-latency air quality data to empower individuals to make informed decisions. For instance, consider a scenario where a neighborhood experiences a 10 min fire outbreak. If the air quality is monitored only once per hour, this short-lived but potentially hazardous event might go unrecorded in government reports. Consequently, an asthma patient relying on such hourly data might mistakenly believe it is safe to go outdoors, unaware of lingering harmful particulates in the air. In addition to this, Dr. Wijeratne also mentions the need to measure the meteorological parameters to calibrate these high-resolution, low-cost particulate sensors. A similar approach is used in our study to identify the meteorological parameters associated with PM variability. In our approach, we use Random Forest Regression models to estimate the sampling time for PM using temperature, atmospheric pressure, humidity, and wind conditions. The RFR models indicate that for particles 1.0 μm and smaller, temperature, atmospheric pressure, and humidity significantly influence sampling times, while wind speed becomes more critical for larger particle sizes, hence requiring particle-specific monitoring needs. These findings indicate the complex interaction between meteorological conditions and PM variability, consistent with other studies that use ML to estimate PM levels from meteorological data [22,23]. An interesting observation is the influence of meteorological events, such as rainfall, on daily PM variability. For instance, 24 January, one of the least polluted days in our data set, experienced morning rainfall, which likely reduced airborne particulates by washing them out of the atmosphere. This aligns with studies showing that precipitation can effectively remove pollutants and influence daily PM concentrations [60,61].

Despite the promising findings, our study has certain limitations and identifies avenues for future research. The fluctuating environmental conditions and inherent noise associated with the sensor reading pose challenges to data accuracy. For example, the IPS7100 can only measure up to 6000 $\mu g/m^3$ for particles $\leq$2.5 μm, with a sampling interval of 1 s. In addition to that, the nugget calculated by the temporal variogram is not always zero, indicating potential measurement errors or temporal variability for a resolution smaller than the IPS7100 measurement interval of 1 s. Similarly, for the BME280 climate, data may become noisy if the sampling rate is below 100 ms. Additionally, the AIRMAR sensor measures wind speed with an uncertainty of ±2% of the measured wind speed, with the error increasing proportionally as the wind speed rises. The rolling temporal variograms and ML models developed in this study are specific to the Joppa neighborhood of Dallas, Texas, reflecting its unique weather conditions, industrial activities, and traffic patterns. Consequently, these findings cannot be generalized to other neighborhoods or regions of the world without further investigation [62,63].

Future research will extend this analysis by examining multiple months of PM data to establish a robust understanding of the optimal measurement intervals. Additionally, we aim to backtrack the polluted air in the study area to identify pollution origins and assess their influence on the PM sampling intervals. Further investigations will also analyze counts of various PM size fractions (particle counts) to provide deeper insights into variability across particle sizes.

5. Conclusions

In this research, we focus on identifying the optimal temporal resolution for days with varying pollution levels in a neighborhood. The results of our first case study indicate that, with 95% confidence, $PM_{2.5}$ (particulate matter with an aerodynamic diameter $\leq$2.5 μm) concentrations on highly polluted days in the Joppa neighborhood should be observed at intervals between 12.13 s and 12.44 s. We select 12.25 s as the optimal interval, as it lies within this confidence interval and corresponds to the mode derived from the PDF (Probability Density Function) of $PM_{2.5}$ measurement times. This analysis underscores the importance of high-frequency observations, with the 12.25 s interval retaining nearly the same level of variability as the 1 s interval on highly polluted days. Such rapid observations allow for capturing short-term variations in PM concentrations, providing more accurate air quality assessments compared to 1-hour averaging. The error computation using the peak nugget and peak nugget CI across the three distinctly polluted days affirms the suitability of the IPS7100 sensor for capturing precise temporal variations in PM measurements without much random noise.

The second case study highlights how this measurement time/sampling time can be dynamically estimated for each PM (particulate matter) size fraction using the meteorological parameters such as temperature, air pressure, humidity, and wind speed. For the most polluted day in the data set, the measurement interval for PM size fractions from 0.1 μm to 1.0 μm is primarily influenced by variations in relative humidity, air temperature, and atmospheric pressure. The influence of humidity can be attributed to the hygroscopic nature of smaller particles, which absorb moisture and increase in size under high humidity [64,65]. Additionally, higher temperatures can intensify chemical reactions, increasing the concentration of smaller particles. Warmer air tends to trap pollutants more effectively, worsening pollution levels. Furthermore, smaller particles often accumulate more in high-pressure environments because the air is more stagnant, giving pollutants less of a chance to spread out [64–66]. For $PM_{2.5}$ and above, wind plays a significant role by transporting these particles through processes like saltation, leading to fluctuations in pollution levels and affecting the required measurement intervals. Meteorological events such as rainfall also significantly impact PM concentrations by efficiently removing particulates from the atmosphere. This is evident on 24 January, a day with notably low pollution levels, likely due to the morning rainfall which reduced airborne particulates and resulted in longer optimal measurement intervals [60,61,67]. These findings from the second case study underscore the complex interactions between pollution levels, meteorological conditions, and sampling frequency, emphasizing the importance of adaptive monitoring strategies.

Together, these findings provide a scalable approach for low-cost air quality systems to dynamically adapt to pollution levels, enhancing urban air quality monitoring while conserving power through more precise and granular PM data collection. This approach aligns with the principles of community-based air quality monitoring [68,69], which emphasize the importance of high temporal resolution to identify neighborhoods disproportionately affected by air pollution from transportation and industrial sources. To address these challenges, low-cost, high-frequency air quality monitoring systems and dashboards are recommended to track air quality metrics in pollution hot spots [70,71].

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

Funding: The following grants were helpful in this work: The US Army (Dense Urban Environment Dosimetry for Actionable Information and Recording Exposure, US Army Medical Research Acquisition Activity, BAA CDMRP Grant Log #BA170483. The Texas National Security Network Excellence Fund Award for Environmental Sensing Security Sentinels. EPA 16th annual P3 Awards Grant Number 83996501, entitled Machine Learning-Calibrated Low-Cost Sensing. SOFTWERX award for Machine Learning for Robotic Teams. 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 original data used for this study will be made available by the authors upon request.

Acknowledgments: The authors highly acknowledge the support received from the University of Texas at Dallas Office of Sponsored Programs, the Natural Science and Mathematics Department of the University, and the Physics Department.

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

Abbreviations

The following abbreviations are used in this manuscript:

CI	Confidence Interval
EPA	Environmental Protection Agency

EU	European Union
IoT	Internet of Things
I^2C	Inter-Integrated Circuit
IPS	Intelligent Particle Sensor
ML	Machine Learning
MSE	Mean Squared Error
PDF	Probability Density Function
PM	Particulate Matter
$PM_{0.1}$	Particulate Matter with an aerodynamic diameter $\leq$ 0.1 µm
$PM_{0.3}$	Particulate Matter with an aerodynamic diameter $\leq$ 0.3 µm
$PM_{0.5}$	Particulate Matter with an aerodynamic diameter $\leq$ 0.5 µm
$PM_{1.0}$	Particulate Matter with an aerodynamic diameter $\leq$ 1.0 µm
$PM_{2.5}$	Particulate Matter with an aerodynamic diameter $\leq$ 2.5 µm
$PM_{5.0}$	Particulate Matter with an aerodynamic diameter $\leq$ 5.0 µm
PM_{10}	Particulate Matter with an aerodynamic diameter $\leq$ 10 µm
RFR	Random Forest Regression
s	Second
SD	Standard Deviation
UART	Universal Asynchronous Receiver/Transmitter
WHO	World Health Organization

References

1. World Health Organization. Air Pollution: The Invisible Health Threat. 2021. Available online: https://www.who.int/news-room/feature-stories/detail/air-pollution--the-invisible-health-threat (accessed on 28 December 2024).
2. United States Environmental Protection Agency. Criteria Air Pollutants. 2024. Available online: https://www.epa.gov/criteria-air-pollutants (accessed on 28 December 2024).
3. Boldo, E.; Medina, S.; Le Tertre, A.; Hurley, F.; Mücke, H.G.; Ballester, F.; Aguilera, I.; Daniel Eilstein on behalf of the Apheis Group. Apheis: Health impact assessment of long-term exposure to $PM_{2.5}$ in 23 European cities. *Eur. J. Epidemiol.* **2006**, *21*, 449–458. . [CrossRef] [PubMed]
4. Boldo, E.; Linares, C.; Lumbreras, J.; Borge, R.; Narros, A.; García-Pérez, J.; Fernández-Navarro, P.; Pérez-Gómez, B.; Aragonés, N.; Ramis, R.; et al. Health impact assessment of a reduction in ambient $PM_{2.5}$ levels in Spain. *Environ. Int.* **2011**, *37*, 342–348. [CrossRef] [PubMed]
5. Brook, R.D.; Rajagopalan, S.; Pope, C.A., III; Brook, J.R.; Bhatnagar, A.; Diez-Roux, A.V.; Holguin, F.; Hong, Y.; Luepker, R.V.; Mittleman, M.A.; et al. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. *Circulation* **2010**, *121*, 2331–2378. [CrossRef] [PubMed]
6. Brook, R.D.; Bard, R.L.; Kaplan, M.J.; Yalavarthi, S.; Morishita, M.; Dvonch, J.T.; Wang, L.; Yang, H.y.; Spino, C.; Mukherjee, B.; et al. The effect of acute exposure to coarse particulate matter air pollution in a rural location on circulating endothelial progenitor cells: Results from a randomized controlled study. *Inhal. Toxicol.* **2013**, *25*, 587–592. . [CrossRef] [PubMed]
7. Pope, C.A.; Brook, R.D.; Burnett, R.T.; Dockery, D.W. How is cardiovascular disease mortality risk affected by duration and intensity of fine particulate matter exposure? An integration of the epidemiologic evidence. *Air Qual. Atmos. Health* **2011**, *4*, 5–14. [CrossRef]
8. Pope, C.A., III; Dockery, D.W. Health effects of fine particulate air pollution: Lines that connect. *J. Air Waste Manag. Assoc.* **2006**, *56*, 709–742. [CrossRef]
9. Ballester, F.; Medina, S.; Boldo, E.; Goodman, P.; Neuberger, M.; Iñiguez, C.; Künzli, N. Reducing ambient levels of fine particulates could substantially improve health: A mortality impact assessment for 26 European cities. *J. Epidemiol. Community Health* **2008**, *62*, 98–105. [CrossRef]
10. Hermanns, S.; von Schneidemesser, E.; Caseiro, A.; Koch, S. The association between different timeframes of air pollution exposure and COVID-19 incidence, morbidity and mortality in German counties in 2020. *Environ. Health* **2024**, *23*, 112. [CrossRef]
11. Lung, S.C.C.; Tsou, M.C.M.; Cheng, C.H.C.; Setyawati, W. Peaks, sources, and immediate health impacts of $PM_{2.5}$ and PM_1 exposure in Indonesia and Taiwan with microsensors. *J. Expo. Sci. Environ. Epidemiol.* **2024**, 1–14. [CrossRef]
12. Brook, R.D.; Xu, X.; Bard, R.L.; Dvonch, J.T.; Morishita, M.; Kaciroti, N.; Sun, Q.; Harkema, J.; Rajagopalan, S. Reduced metabolic insulin sensitivity following sub-acute exposures to low levels of ambient fine particulate matter air pollution. *Sci. Total Environ.* **2013**, *448*, 66–71. [CrossRef]

13. United States Environmental Protection Agency. Air Data Basic Information. 2024. Available online: https://www.epa.gov/outdoor-air-quality-data/air-data-basic-information (accessed on 27 December 2024).
14. Jacob, D.J.; Winner, D.A. Effect of climate change on air quality. *Atmos. Environ.* **2009**, *43*, 51–63. [CrossRef]
15. Xu, W.; Sun, J.; Liu, Y.; Xiao, Y.; Tian, Y.; Zhao, B.; Zhang, X. Spatiotemporal variation and socioeconomic drivers of air pollution in China during 2005–2016. *J. Environ. Manag.* **2019**, *245*, 66–75. [CrossRef] [PubMed]
16. Odubo, T.C.; Kosoe, E.A. Sources of Air Pollutants: Impacts and Solutions. In *Air Pollutants in the Context of One Health: Fundamentals, Sources, and Impacts*; Springer: Cham, Switzerland, 2024; p. 75. [CrossRef]
17. Li, X.; Ma, Y.; Wang, Y.; Liu, N.; Hong, Y. Temporal and spatial analyses of particulate matter (PM_{10} and $PM_{2.5}$) and its relationship with meteorological parameters over an urban city in northeast China. *Atmos. Res.* **2017**, *198*, 185–193. [CrossRef]
18. Matheron, G. Principles of geostatistics. *Econ. Geol.* **1963**, *58*, 1246–1266. [CrossRef]
19. Bohling, G. Introduction to geostatistics and variogram analysis. *Kans. Geol. Surv.* **2005**, *1*, 1–20.
20. Oliver, M.; Webster, R. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena* **2014**, *113*, 56–69. [CrossRef]
21. Chen, C.C.; Wu, C.F.; Yu, H.L.; Chan, C.C.; Cheng, T.J. Spatiotemporal modeling with temporal-invariant variogram subgroups to estimate fine particulate matter $PM_{2.5}$ concentrations. *Atmos. Environ.* **2012**, *54*, 1–8. [CrossRef]
22. Dewage, P.M.; Wijeratne, L.O.; 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. [CrossRef]
23. Yu, X.; Lary, D.; Simmons, C.; Wijeratne, L. High Spatial-Temporal $PM_{2.5}$ Modeling Utilizing Next Generation Weather Radar (NEXRAD) as a Supplementary Weather Source. *Remote Sens.* **2022**, *14*, 495. [CrossRef]
24. Trivedi, D.K.; Ali, K.; Beig, G. Impact of meteorological parameters on the development of fine and coarse particles over Delhi. *Sci. Total Environ.* **2014**, *478*, 175–183. [CrossRef]
25. Hien, P.; Bac, V.; Tham, H.; Nhan, D.; Vinh, L. Influence of meteorological conditions on $PM_{2.5}$ and $PM_{2.5–10}$ concentrations during the monsoon season in Hanoi, Vietnam. *Atmos. Environ.* **2002**, *36*, 3473–3484. [CrossRef]
26. Galindo, N.; Varea, M.; Gil-Moltó, J.; Yubero, E.; Nicolás, J. The influence of meteorology on particulate matter concentrations at an urban Mediterranean location. *Water Air Soil Pollut.* **2011**, *215*, 365–372. [CrossRef]
27. Zha, Y.; Gao, J.; Jiang, J.; Lu, H.; Huang, J. Monitoring of urban air pollution from MODIS aerosol data: Effect of meteorological parameters. *Tellus B Chem. Phys. Meteorol.* **2010**, *62*, 109–116. [CrossRef]
28. Pieras Systems. IPS Datasheet V1.0.2. 2021. Available online: https://pierasystems.com/wp-content/uploads/2021/03/IPS-Datasheet-V1.0.2.pdf (accessed on 11 November 2024).
29. Bosch Sensortec. BME280 Datasheet. 2018. Available online: https://www.mouser.com/datasheet/2/783/BST-BME280-DS002-1509607.pdf (accessed on 11 November 2024).
30. Airmar Technology Corporation. 110WX WeatherStation® Instrument for Land-Based Applications. 2021. Available online: https://www.airmarweb.com/uploads/brochures/110WX-LAND.pdf (accessed on 11 November 2024).
31. Chiles, J.P.; Delfiner, P. *Geostatistics: Modeling Spatial Uncertainty*; John Wiley & Sons: Hoboken, NJ, USA, 2012; Volume 713. [CrossRef]
32. Song, W.; Jia, H.; Li, Z.; Tang, D. Using geographical semi-variogram method to quantify the difference between NO_2 and $PM_{2.5}$ spatial distribution characteristics in urban areas. *Sci. Total Environ.* **2018**, *631*, 688–694. . [CrossRef]
33. Cressie, N.; Hawkins, D.M. Robust estimation of the variogram: I. *J. Int. Assoc. Math. Geol.* **1980**, *12*, 115–125. [CrossRef]
34. Genton, M.G. Highly robust variogram estimation. *Math. Geol.* **1998**, *30*, 213–221. [CrossRef]
35. Abzalov, M. Introduction to Geostatistics. In *Applied Mining Geology*; Springer: Cham, Switzerland, 2016; pp. 233–237. [CrossRef]
36. Harrison, W.A.; Lary, D.; Nathan, B.; Moore, A.G. The neighborhood scale variability of airborne particulates. *J. Environ. Prot.* **2015**, *6*, 464. [CrossRef]
37. Harrison, W.A.; Lary, D.J.; Nathan, B.J.; Moore, A.G. Using remote control aerial vehicles to study variability of airborne particulates. *Air Soil Water Res.* **2015**, *8*, ASWR-S30774. [CrossRef]
38. Shi, Y. Evaluation method of continuous compaction quality of highway water stable base. *J. Phys. Conf. Ser.* **2021**, *1983*, 012020. [CrossRef]
39. Benavides-Bravo, F.G.; Soto-Villalobos, R.; Cantú-González, J.R.; Aguirre-López, M.A.; Benavides-Ríos, Á.G. A Quadratic–Exponential Model of Variogram Based on Knowing the Maximal Variability: Application to a Rainfall Time Series. *Mathematics* **2021**, *9*, 2466. [CrossRef]
40. 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] [PubMed]
41. Ruwali, S.; Fernando, B.; Talebi, S.; Wijeratne, L.; Waczak, J.; Madusanka, P.M.; Lary, D.J.; Sadler, J.; Lary, T.; Lary, M.; et al. Estimating Inhaled Nitrogen Dioxide from the Human Biometric Response. *Adv. Environ. Eng. Res.* **2024**, *5*, 011. [CrossRef]

42. Ruwali, S.; Talebi, S.; Fernando, A.; Wijeratne, L.O.; Waczak, J.; Dewage, P.M.; Lary, D.J.; Sadler, J.; Lary, T.; Lary, M.; et al. Quantifying Inhaled Concentrations of Particulate Matter, Carbon Dioxide, Nitrogen Dioxide, and Nitric Oxide Using Observed Biometric Responses with Machine Learning. *BioMedInformatics* **2024**, *4*, 1019–1046. [CrossRef]
43. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
44. Feurer, M.; Hutter, F. Hyperparameter optimization. In *Automated Machine Learning: Methods, Systems, Challenges*; Springer: Cham, Switzerland, 2019; pp. 3–33. [CrossRef]
45. Probst, P.; Wright, M.N.; Boulesteix, A.L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1301. [CrossRef]
46. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [CrossRef]
47. United States Environmental Protection Agency. Reconsideration of the National Ambient Air Quality Standards for Particulate Matter. 2024. Available online: https://www.federalregister.gov/documents/2024/03/06/2024-02637/reconsideration-of-the-national-ambient-air-quality-standards-for-particulate-matter (accessed on 8 November 2024).
48. World Health Organization. What are the WHO Air Quality Guidelines? 2021. Available online: https://www.who.int/news-room/feature-stories/detail/what-are-the-who-air-quality-guidelines (accessed on 8 November 2024).
49. Smart Air Filters. Global Air Quality Standards for $PM_{2.5}$ & PM_{10}. 2024. Available online: https://smartairfilters.com/en/blog/global-air-quality-standards-pm2-5-pm10/ (accessed on 8 November 2024).
50. Carslaw, D.C.; Ropkins, K. Openair—An R package for air quality data analysis. *Environ. Model. Softw.* **2012**, *27*, 52–61. [CrossRef]
51. Center for Institutional Research Computing (CIRC). User Guide for CIRC Systems. 2024. Available online: https://docs.circ.utdallas.edu/user-guide/systems/index.html (accessed on 11 November 2024).
52. O'Brien, S.F.; Yi, Q.L. How do I interpret a confidence interval? *Transfusion* **2016**, *56*, 1680–1683. [CrossRef]
53. Hazra, A. Using the confidence interval confidently. *J. Thorac. Dis.* **2017**, *9*, 4125–4130. [CrossRef]
54. Topics, S. Variance Ratio. Available online: https://www.sciencedirect.com/topics/computer-science/variance-ratio (accessed on 11 November 2024).
55. Jolly, K. *Machine Learning with Scikit—Learn Quick Start Guide: Classification, Regression, and Clustering Techniques in Python*; Packt Publishing Ltd.: Birmingham, UK, 2018. Available online: https://books.google.com/books?hl=en&lr=&id=28t1DwAAQBAJ&oi=fnd&pg=PP1&dq=Jolly,+K.+Machine+learning+with+scikit-learn+quick+start+guide:+classification,+regression,+and+clustering+541+techniques+in+Python3B+Packt+Publishing+Ltd,+2018.&ots=GbEhT9qK_Z&sig=guYzkX6iKEC3UA9mrWQjiwx1Rqk#v=onepage&q&f=false (accessed on 11 November 2024).
56. Shao, Y.; Wyrwoll, K.H.; Chappell, A.; Huang, J.; Lin, Z.; McTainsh, G.H.; Mikami, M.; Tanaka, T.Y.; Wang, X.; Yoon, S. Dust cycle: An emerging core theme in Earth system science. *Aeolian Res.* **2011**, *2*, 181–204. [CrossRef]
57. Yang, X.; Zhou, C.; Yang, F.; Meng, L.; Huo, W.; Mamtimin, A.; He, Q. Saltation Activity on Non-Dust Days in the Taklimakan Desert, China. *Remote Sens.* **2022**, *14*, 2099. [CrossRef]
58. Liu, X. Physical Studies of Airborne Pollen and Particulates Utilizing Machine Learning. Ph.D. Thesis, The University of Texas at Dallas, Dallas, TX, USA, 2019. Available online: https://utd-ir.tdl.org/items/270f88c6-f502-438c-8089-ecceb4f6fe97 (accessed on 11 November 2024).
59. Wijeratne, L. Coupling Physical Measurement With Machine Learning for Holistic Environmental Sensing. Ph.D. Thesis, The University of Texas at Dallas, Dallas, TX, USA, 2021. Available online: https://utd-ir.tdl.org/items/5550059c-4af4-4da8-97a3-c5050baec902 (accessed on 24 November 2024).
60. Guo, L.C.; Zhang, Y.; Lin, H.; Zeng, W.; Liu, T.; Xiao, J.; Rutherford, S.; You, J.; Ma, W. The washout effects of rainfall on atmospheric particulate pollution in two Chinese cities. *Environ. Pollut.* **2016**, *215*, 195–202. [CrossRef]
61. Zhao, X.; Sun, Y.; Zhao, C.; Jiang, H. Impact of precipitation with different intensity on PM2. 5 over typical regions of China. *Atmosphere* **2020**, *11*, 906. [CrossRef]
62. ScienceDirect. Variogram: Definition, Components, and Use in Geostatistics. In *ScienceDirect Topics: Earth and Planetary Sciences*; ScienceDirect: Amsterdam, The Netherlands, 2024. Available online: https://www.sciencedirect.com/topics/earth-and-planetary-sciences/variogram#:~:text=II.&text=There%20are%20two%20reasons%20why,experimental%20points%20in%20different%20directions (accessed on 11 November 2024).
63. Spatial Heterogeneity of Ecological Environment Quality in China's North-South Transitional Zone. *Int. J. Environ. Res. Public Health* **2024**, *19*, 2236. [CrossRef]
64. Zhang, L.; Cheng, Y.; Zhang, Y.; He, Y.; Gu, Z.; Yu, C. Impact of air humidity fluctuation on the rise of PM mass concentration based on the high-resolution monitoring data. *Aerosol Air Qual. Res.* **2017**, *17*, 543–552. [CrossRef]
65. Emekwuru, N.; Ejohwomu, O. Temperature, humidity and air pollution relationships during a period of rainy and dry seasons in Lagos, West Africa. *Climate* **2023**, *11*, 113. [CrossRef]
66. Sindosi, O.; Katsoulis, B.; Bartzokas, A. An objective definition of air mass types affecting Athens, Greece; the corresponding atmospheric pressure patterns and air pollution levels. *Environ. Technol.* **2003**, *24*, 947–962. . [CrossRef]

67. World Weather Online. Weather Forecast for Dallas, USA on 24 January 2024. 2024. Available online: https://world-weather.info/forecast/usa/dallas_1/24-january/ (accessed on 8 November 2024).
68. MINTS Lab, University of Texas at Dallas. SharedAirDFW Project. 2021. Available online: https://mints.utdallas.edu/2021/01/01/sharedairdfw/ (accessed on 24 November 2024).
69. Georgetown Climate Center. Community-Based Air Quality Monitoring: A Tool for Equitable Climate Policy. 2024. Available online: https://www.georgetownclimate.org/articles/community-based-air-quality-monitoring-equitable-climate-policy.html (accessed on 27 November 2024).
70. MINTS Lab, University of Texas at Dallas. Central Node Demo Dashboard. 2024. Available online: http://mdash.circ.utdallas.edu:3000 (accessed on 24 November 2024).
71. SharedAirDFW. SharedAirDFW: Air Quality Monitoring in Dallas—Fort Worth. 2024. Available online: https://www.sharedairdfw.com/ (accessed on 24 November 2024).