

SUB PROJECT 2

Developing a self-calibrating, self-diagnosable, and optimum low-cost (LCS) air quality (AQ) sensor network

1. Introduction

The project aims to develop a low-cost sensor (LCS) network for air quality monitoring, providing high spatial-temporal coverage and resolution for comparative analysis. This is crucial as air quality (AQ) is heterogeneous, changing significantly even within small areas. AQ is generally very heterogeneous and often changes significantly even within a small area like an academic campus, classrooms, and office spaces. AQ within a city/urban/non-urban area could be highly heterogeneous and vary significantly between intra-city locations. Thus, policymakers need to consider that before proposing any policy for air pollution mitigation. Traditional AQ monitoring networks are costly, limiting coverage and resolution. LCS offers a cost-effective solution for high-density monitoring, enabling better insights into AQ variations. However, LCS often suffer from accuracy and reliability issues. To address these limitations, the project focuses on developing an LCS network capable of:

- Self-calibration: Adjusting sensor readings based on reference/research grade analyzers/equipment. This ensures the accuracy of data collected by the LCS network.
- Self-diagnosis: Detecting and reporting sensor malfunction, degradation, or replacement needs. This ensures the reliability and continuous, high-quality data collection of the LCS network.

The project utilizes AQ and sensor-related parameters to achieve self-calibration and self-diagnosis. It requires a multidisciplinary team with air quality, networking, and data acquisition expertise.

2. Materials and Methods

2.1. Understanding the Working of PM Sensors

2.1.1. Optical Particle Counters (OPC)

OPCs, which may be handheld or part of a more extensive facility monitoring system, utilize a light scattering method to count and size particles. Ambient air is sampled by the OPC and illuminated by a low-power laser diode. The laser illuminates particulate matter, and a light-sensitive diode counts the particles. OPCs are capable of counting and sizing individual particles. The measuring principle of the OPC is the light scattering of single particles with a semiconductor laser as a light source.



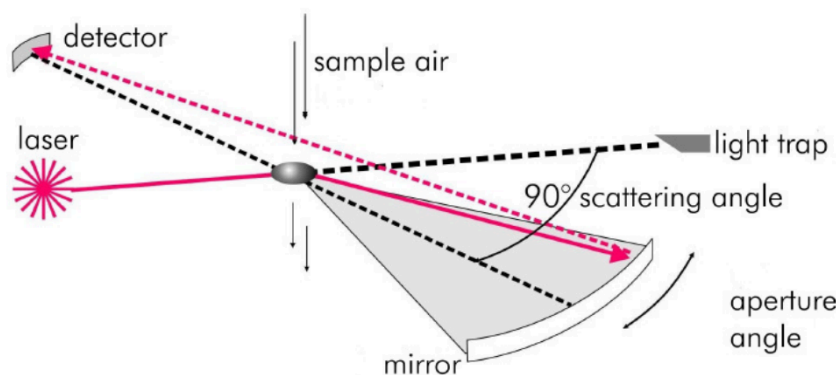


Figure 1. Laser Scattering in Optical Particle Counters

A detector positioned at a 90° scattering angle optimizes the signal-to-noise ratio, allowing detection of particles as small as 10nm. The size distribution of particles is determined by analyzing the light pulses generated as they cross the laser beam. This information calculates dust mass, making OPCs suitable for various applications, including occupational health data compilation, dust analysis, and atmospheric research. The relation between the intensity of scattered light and the size of the particle is discussed in [A Review of Low-Cost Particulate Matter Sensors from the Developers' Perspectives](#)

2.1.2. Nephelometers

Nephelometers, also called photometers, detect particles by measuring the total amount of light scattered by a cloud or batch of particles. The intensity of light scattered by a particle is a function of the particle size, shape, and chemistry, so the response of a nephelometer is a function of particle size for unit mass concentration. Nephelometers provide an integrated output closely related to particle mass concentration. They are typically limited to measuring particle sizes of 0.3 to 10 microns. ([Particles Matter - Met One Instruments](#)).

2.2. Particulate Matter Sensors/Devices

We have evaluated three low-cost pm devices - Atmos (Respirer Living Sciences Private Limited, India), PA Flex 2 (PurpleAir, Inc.), and OPC N3 (Alphasense, Amtech, UK) with two reference-grade devices - GRIMM 11A (GRIMM Aerosol Technik GmbH), and Partector 2 Pro (Naneos Particle Solutions GmbH). The low-cost pm devices were chosen based on their price and portability, i.e., that costs < \$500 and can be placed inside an indoor testing facility with minimum external intervention. Table 1 provides details of the specifications of each device.

	Naneos Partector 2 Pro	GRIMM OPC	Urbansciences ATMOS	PurpleAir Flex 2 Pro	Alphasense OPC N3
PM Sensor	Proprietary	Proprietary	Plantower PM5003	Plantower PM6003	Proprietary
Working Principle	Electron Mobility	Light Scattering - OPC	Light Scattering - Nephelometer	Light Scattering - Nephelometer	Light Scattering - OPC
Approximate Cost	21,00,000	13,00,000	40,000	25,000	40,000
Measurement Range	0.01 m - 0.3 m	0.25 m - 32 m	0.3 m - 10 m	0.3 m - 10 m	0.3 m - 40 m
PM Range	PM0.3	PM1, PM2.5, PM10	PM1, PM2.5, PM10	PM1, PM2.5, PM10	PM1, PM2.5, PM4.25, PM10
Connectivity	Bluetooth, eMMC SD	RS232, eMMC SD	WiFi, eMMC SD	WiFi, eMMC SD	SPI, eMMC SD
Dimensions	8.8 x 14.2 x 3.4 cm	24 x 13 x 7 cm	15 x 10 x 6 cm	8.5 x 8.5 x 12.5 cm	7.5 x 6.4 x 6 cm

Table 1. Key specifications of PM Devices

3. Comparative Study

3.1. Sensor Configuration and Calibration

This study provides a comparative evaluation of various air quality sensors, with a focus on particulate matter (PM) measurement performance. It assesses three low-cost sensors (LCSs)—the PurpleAir Flex II, Alphasense OPC N3, and Urban Sciences Atmos—alongside two reference-grade instruments, the Grimm Aerosol Technik OPC and the Naneos Partector 2 Pro.

3.1.1. Data Resolution and Timestamps

The Naneos Partector 2 Pro records data with the highest frequency at 10-second intervals, followed by the Alphasense OPC N3 at 30-second intervals. The Grimm Aerosol Technik OPC and Urban Sciences Atmos log data every minute, while the PurpleAir Flex II records data every 2 minutes. Timestamp formats also vary across sensors: the Grimm Aerosol Technik OPC and Urban Sciences Atmos use Indian Standard Time (IST), PurpleAir Flex II uses Coordinated Universal Time (UTC), and Naneos Partector 2 Pro uses a hybrid system combining an IST-based start time with incremental intervals. The Alphasense OPC N3, lacking a Real-Time Clock (RTC), presents challenges for timestamp accuracy.

3.1.2. Data Transmission

Each sensor employs different methods for transmitting data to the home server. The Grimm Aerosol Technik OPC uses a Python script to convert data obtained from RS232 Serial Communication to JSON format and transmits it via HTTP POST. PurpleAir Flex II also uses HTTP POST for data transmission, while the Alphasense OPC N3 relies on a microcontroller and RTC unit for WiFi-based JSON packet transmission. Urban Sciences Atmos uses an asynchronous thread to fetch data through the Atmos API.

3.1.3. Accuracy and Calibration

Both the Grimm Aerosol Technik OPC and Naneos Partector 2 Pro, considered highly accurate reference-grade instruments, serve as benchmarks for calibrating the LCSs. Given the known accuracy and reliability limitations of LCSs, machine learning models—specifically linear regression and decision trees—were employed to enhance calibration accuracy by aligning LCS readings with those from the reference instruments.

3.2. Evaluation for Indoors and Outdoors Environment

Indoor environments tend to be more stable with fewer fluctuations in temperature, humidity, and airflow, which can lead to more consistent sensor readings. In contrast, outdoor environments experience greater variability due to changing weather conditions, which can affect sensor accuracy and stability, particularly for low-cost sensors (LCSs). Evaluation would assess each sensor's response to these environmental changes which would help in creating additional calibration or correction factors that are required for PM monitoring.

Field evaluations can provide insights into optimal sensor placement and network configuration to maximize coverage and data reliability for both indoor and outdoor monitoring. For instance, outdoor



sensors may benefit from protective casings or periodic recalibration, while indoor sensors can be placed to minimize airflow disruptions. Linear Regression for Indoor and Outdoor evaluation of the sensors is presented in Figure 2 for all PM size ranges.

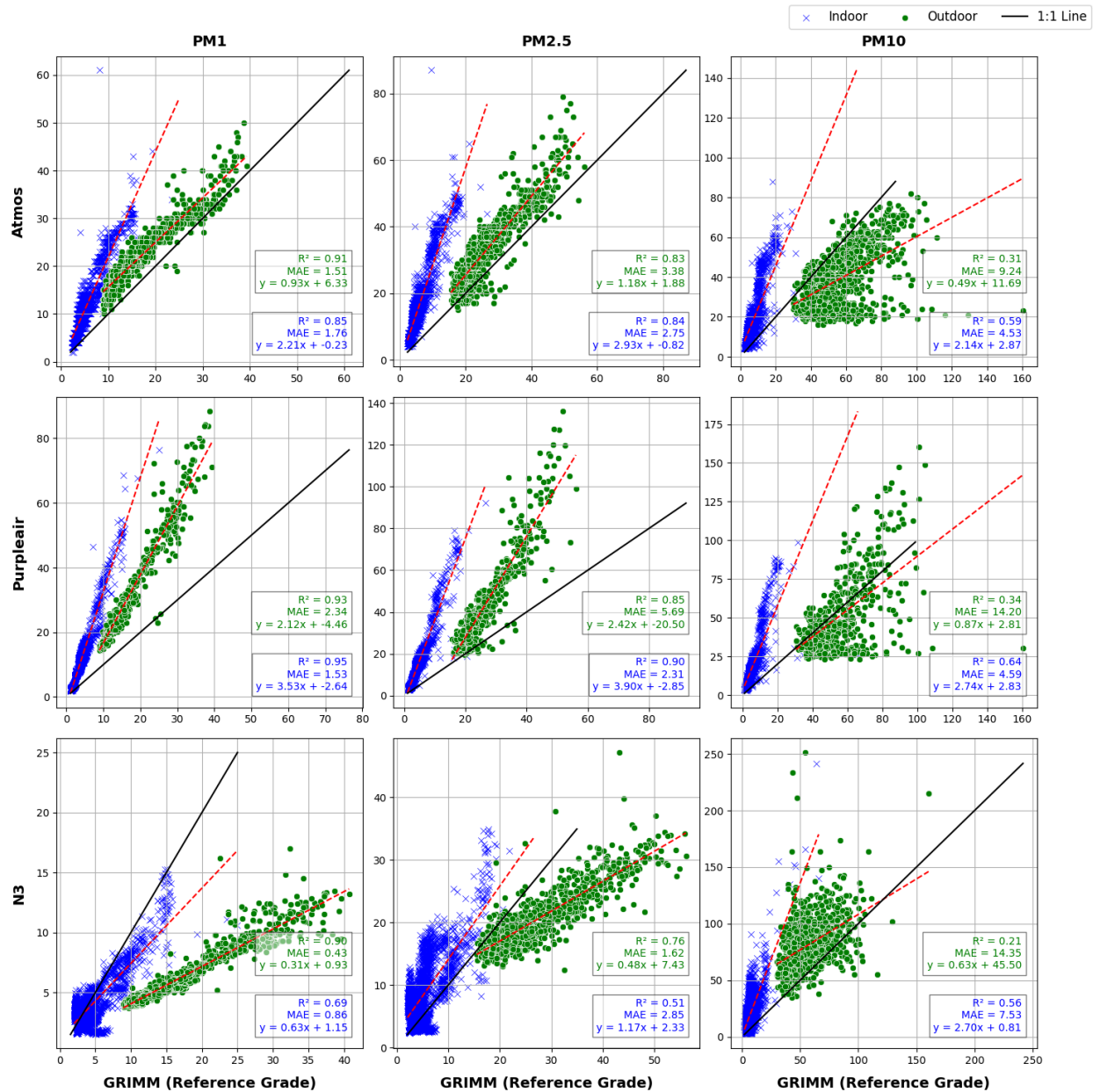


Figure 2. Comparison of LCS with Reference Grade Sensor using Line of best fit (Linear Regression).

Across all sensors, indoor PM concentrations generally show lower mean and median values compared to outdoor concentrations. This difference is notable across all PM sizes (PM1, PM2.5, and PM10). The metrics obtained after regression are highlighted in Table 2.

	Indoor			Outdoor		
	PM1	PM2.5	PM10	PM1	PM2.5	PM10
Key Metrics GRIMM 11A						
Range	1.1 – 25 $\mu\text{g}/\text{m}^3$	1.1 – 26.5 $\mu\text{g}/\text{m}^3$	1.1 – 65.9 $\mu\text{g}/\text{m}^3$	8.9 – 40.7 $\mu\text{g}/\text{m}^3$	15.5 – 56.1 $\mu\text{g}/\text{m}^3$	29.9 – 160 $\mu\text{g}/\text{m}^3$
Mean	4.56 $\mu\text{g}/\text{m}^3$	5.27 $\mu\text{g}/\text{m}^3$	6.19 $\mu\text{g}/\text{m}^3$	17.77 $\mu\text{g}/\text{m}^3$	27.30 $\mu\text{g}/\text{m}^3$	55.76 $\mu\text{g}/\text{m}^3$
Median	3.9 $\mu\text{g}/\text{m}^3$	4.5 $\mu\text{g}/\text{m}^3$	5.1 $\mu\text{g}/\text{m}^3$	16.1 $\mu\text{g}/\text{m}^3$	24.8 $\mu\text{g}/\text{m}^3$	51.7 $\mu\text{g}/\text{m}^3$
Co-efficient of Determination (r^2)						
ATMOS	0.85	0.84	0.58	0.9	0.82	0.3
PurpleAir Flex 2	0.94	0.9	0.64	0.93	0.85	0.33
Alphasense OPC N3	0.68	0.51	0.56	0.91	0.7	0.2
Mean Absolute Error (MAE)						
ATMOS	1.75	2.74	4.53	1.5	3.38	9.23
PurpleAir Flex 2	1.53	2.3	4.59	2.34	5.69	14.1
Alphasense OPC N3	0.86	2.84	7.52	0.42	1.61	14.3
Root Mean Square Error (RMSE)						
ATMOS	2.45	3.92	7.63	2	4.49	11.72
PurpleAir Flex 2	2.21	3.96	8.97	3.9	8.41	19.91
Alphasense OPC N3	1.11	3.55	10.6	0.71	2.26	19.82

Table 2. Performance Metrics for LCS for Indoor and Outdoor Datasets.

The PurpleAir Flex II sensor typically records higher mean and median values across all PM sizes compared to the other low-cost sensors. This suggests it may have a higher sensitivity or a tendency to overestimate concentrations in both environments. The Grimm reference sensor shows lower mean and median values in the indoor data compared to the low-cost sensors, suggesting that LCSs may overestimate PM levels indoors. The PurpleAir Flex II has the highest R^2 values across PM1 (0.95), PM2.5 (0.90), and PM10 (0.64), suggesting a relatively strong correlation with the Grimm reference sensor indoors. The R^2 values are generally lower outdoors, indicating that the LCSs perform better at capturing indoor PM levels relative to the Grimm sensor but struggle with outdoor measurements. Notably, for PM10, R^2 values are particularly low across all LCSs, with Atmos and Alphasense N3 showing very low correlation, which points to difficulties in accurately capturing PM10 levels in outdoor environments.

PurpleAir Flex II has the lowest MAE for PM1 in indoor conditions (1.53), indicating strong accuracy relative to Grimm. However, for PM10, both Atmos and Alphasense N3 show high MAE values in outdoor data, suggesting difficulties in accurately measuring larger particle sizes outdoors.

Overall, PurpleAir Flex II shows the highest accuracy for indoor (live) data across all PM sizes, particularly for PM1 and PM2.5. However, Atmos also performs well in indoor conditions for PM1, with relatively strong correlation and lower error metrics.



4. Data Collection Portal - Graphic User Interface

4.1. Data Analysis

A dedicated AWS-based data collection portal was developed to streamline the subsequent analysis. This web interface facilitated cleaning and standardizing raw data obtained from each LCS, ensuring a consistent and reliable dataset for further investigation. A Simplified Flow of Operation for DCP is described in Figure 4.

The DCP collocates and co-times the data received from various low-cost sensors (LCSs), ensuring that data from different sources is synchronized and can be easily compared. This is achieved by creating a testbed where all the sensors, including the reference-grade instruments, are placed within a controlled environment. This arrangement allows for simultaneous data collection from each device, resulting in a synchronized dataset for analysis. The DCP's visualization capabilities are still under development, and future improvements include adding more user-friendly features and a more professional-looking dashboard.

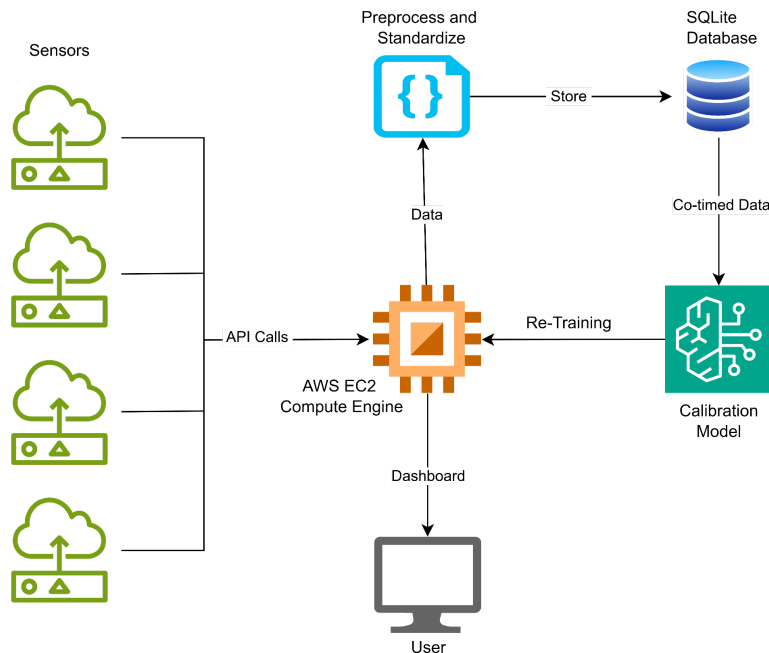


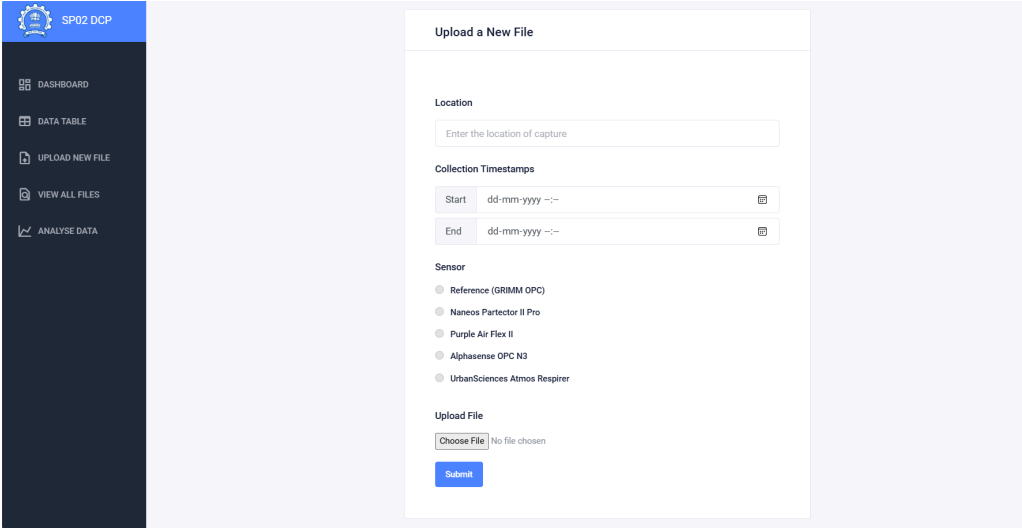
Figure 4. Working of Data Collection of Portal.

4.1.1. Static Data Analysis

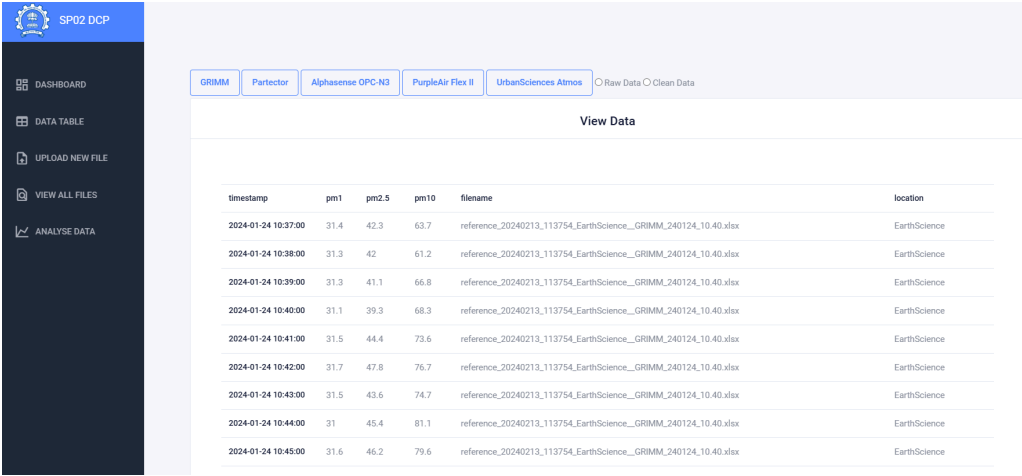
For Static Data Analysis, Data can be uploaded manually to the DCP via “Upload New Files” tab presented in Figure 5a. Here Optional Parameters such as Location can be added. This helps in future analysis of the data. This data is then cleaned and standardized following a specific syntax to store the data in the database. After standardization it can be viewed in “Data Table” tab. Similarly, all the

uploaded files and cleaned files can be accessed or modified by GUI inside the “View All Files” tab described in Figure 5b and Figure 5c respectively. All the uploaded data can be viewed and analyzed using the auto-charting feature in “Analyze Data” tab. This feature is demonstrated in Figure 5d.

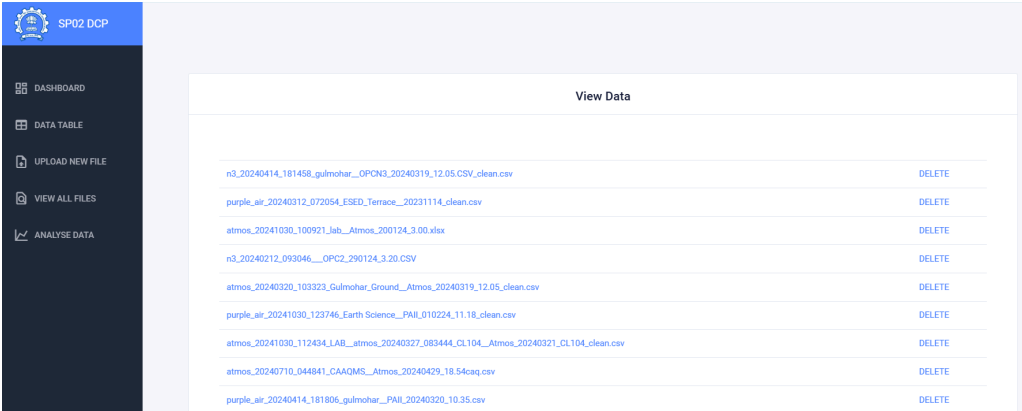
a.



b.



c.



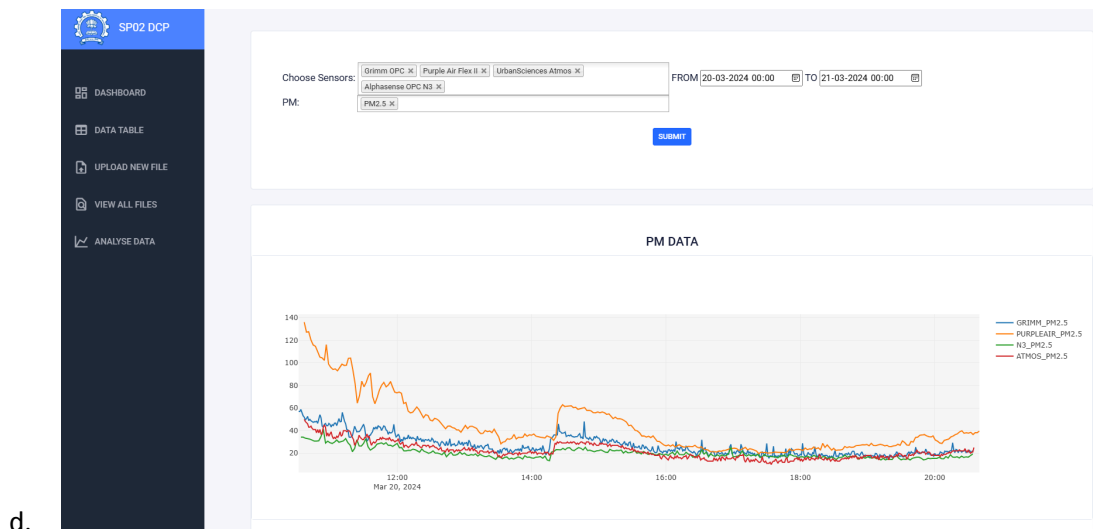


Figure 5. a. Manual Data Upload on Data Collection Portal. b. View Table for Raw/ Clean Data. c. View or Modify all files in the database. d. Visualize Uploaded data on DCP.

4.1.2. Real-time Data Analysis

In Atmos, an asynchronous thread fetches the latest data from the Atmos sensor and sends it to the Data Collection Portal using the Atmos API. The PurpleAir device uses an HTTP POST communication protocol to send sensor data to the DCP endpoint. Alphasense OPC N3 lacks the hardware to send information wirelessly, so we interfaced it with a microcontroller and an RTC unit to establish a WiFi connection and transmit data to the DCP as a JSON packet. A simplified block diagram for interfacing Alphasense OPC N3 is presented in Figure 6.

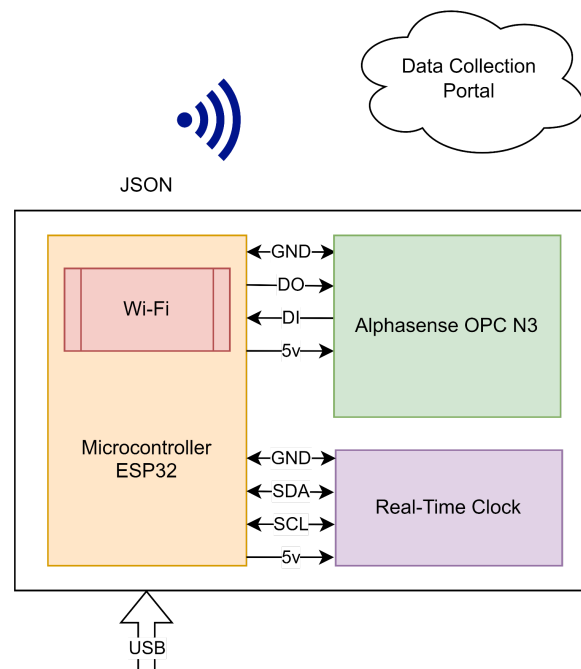


Figure 5. Block Diagram for Interfacing of Alphasense OPC N3 and ESP32 Board.

Real-time data is observed on the Dashboard of DCP in the form of Gauge Plots which denote the PM values for each sensor in the form of a Gauge ranging from good to extremely severe PM readings. Dashboard also has real-time regression based scatter plot to showcase raw and calibrated readings along with metrics such as r^2 , slope, intercept, Mean Absolute Error (MAE) and Root Mean Square Error (RSME).

4.2. Model Creation

We are currently working with Linear Regression and SGD Regressor for real-time data analysis. Further comparison with Multi-parameter Neural Network such ANN is planned for the project. After finalization of the working framework, an independent model will be trained to self-calibrate the sensors with minimal need of reference grade sensors.

5. Future Work

In future research, we aim to address several critical areas to enhance the reliability and autonomy of low-cost sensor (LCS) networks for air quality monitoring. A network-based degradation prediction algorithm that can detect malfunctioning sensors without manual intervention. By comparing each sensor's readings to the average corrected readings of the other sensors in the network, this algorithm will identify potential malfunctions if any sensor's measurements deviate beyond a set threshold, $\bar{\delta}$, over a predefined period, such as one week. This predictive model will help maintain network accuracy and reduce maintenance requirements. Selecting the most suitable machine learning algorithms is also a priority. Given the need for models that balance computational efficiency with high predictive accuracy, we will evaluate various algorithms with batch learning capabilities. Such models are ideal for real-time processing in LCS networks, where computational resources are often limited. The focus will be on those algorithms that provide optimal accuracy with minimal computational demands, ensuring that the system remains efficient and scalable.

Additionally, we will extend our models to account for environmental factors, specifically temperature, humidity, and dew point, which can influence particulate matter readings. By incorporating these parameters, we aim to create multi-parameter models that increase sensor accuracy in diverse environmental conditions. This approach will allow for a more robust analysis of LCS data, ensuring reliability across varying atmospheric contexts. Real-time calibration of LCS devices without the need for reference-grade instruments represents another pivotal goal. Through machine learning, we will pursue methods for independent, data-driven calibration, allowing LCS devices to self-correct in real time. This would eliminate reliance on costly reference instruments, making widespread LCS deployment more feasible and affordable.

Finally, we will evaluate the calibration model's effectiveness by conducting parallel performance tests. Using metrics such as r^2 , slope, and RMSE, we will assess the model's ability to improve accuracy and reliability, both with and without corrective interventions. This comprehensive evaluation will contribute valuable insights into model performance, validating its application for autonomous air quality monitoring networks.

