Local Reference Free In-Field Calibration of Low Cost Air Pollution Monitoring Sensors

Sushmita Ghosh, Payali Das, Swades De, Shouri Chatterjee, and Marius Portmann

***Abstract*—The real-life deployments of air pollution monitoring systems are sparse, due to large size, high cost, and high-power consumption. Such sparsely deployed sensing stations are unable to provide a fine granular pollution mapping of a given geographical area. By deploying low cost, low power, miniature air pollution monitoring sensor nodes, the air pollution map of the whole area can be accurately measured. However, accuracy of the sensed data of the low cost miniature sensing nodes (MSNs) needs to be addressed. This paper presents an autocalibration method of low cost MSNs, with the help of sparsely deployed high cost sensing stations (HCSSs). The datasets from the HCSSs are collected and used to calibrate the MSN using a suitable learning-based regressor model at the nearby edge node. To this end, this paper proposes a cross-correlation based method of determining the optimum time to re-calibrate the low cost sensors in a multi-sensing node. This method eliminates the requirement of taking the MSNs offline to calibrate/re-calibrate them. To apply the proposed autocalibration method, this paper additionally presents the design of a low cost, low power particulate matter (PM) sensor. To validate the performance of the low cost PM sensor, the calibrated PM data are compared with the data collected from a colocated commercially available PM sensor, which is considered as reference. The low cost PM sensor is 91% more cost efficient and 57% more energy efficient compared to the commercial high cost PM sensor, while maintaining the sensing error within a given threshold.**

***Index Terms*—Autocalibration, automatic recalibration, energy efficiency, learning-based regressor model, low cost PM sensor**

I. Introduction

The air pollution in cities, as well as villages, are increasing day by day due to many reasons, such as increasing vehicles, heavy industries, chemical factories, etc. [1]. To monitor the air quality globally, a large number of air pollution monitoring devices (APMDs) need to be deployed. Conventional good quality sensors are costly, energy consuming, and bulky [2]. The pollution monitoring stations consist of multiple such sensors to monitor multiple air pollutants in the environment. Thus, massive deployment of such monitoring stations is expensive which makes it difficult to scale up the deployment. Large scale field deployed wireless miniature multi-sensing nodes (MSNs) are expected to have limited energy availability [3]. Since powering these field nodes using wired connectivity is not always feasible, these nodes are primarily powered by batteries. Hence, increasing the battery lifetime for uninterrupted sensing operation is very important [4], [5]. High power consuming sensors reduce the battery lifetime [6]. Therefore, design of low cost, energy efficient, and portable pollution monitoring sensors is necessary to deploy them massively for fine-grained pollution localization and mapping [7].

An environmental pollution monitoring board consists of multiple sensors to monitor the environmental parameters, namely, PM_1, $\text{PM}_{2.5}$, PM_{10}, CO, O_3, NO_2, SO_2, temperature, and humidity [8], [9]. Monitoring such parameters using low cost miniature sensors reduces their sensing accuracy. Therefore, the low cost MSN needs to be occasionally calibrated with respect to some accurate reference sensors [10], [11]. However, due to the dynamics of pollution environment and the ambient condition of the sensing module, the accuracy of calibration models decreases with time. Moreover, the low cost sensors may require recalibration more frequently compared to the high cost sensors. Bringing them to the lab for recalibration is expensive and difficult in practice. Deploying a reference sensing node near to the MSN is even more expensive. Thus, local reference free calibration models need to be developed for effectively utilizing the low cost MSNs.

A. Related Works

Although scalable deployment of MSNs is feasible, their sensing quality remains a major issue. To improve the sensing quality, various calibration techniques are proposed in literature. The low cost sensors need to be calibrated with respect to some highly accurate sensor that provides the reference data and validates the calibration model.

Calibration of low cost temperature, humidity, CO, and PM sensor was studied in [12], where four different types of machine learning models, such as, multivariate linear regressor, K-nearest neighbors, random forest, and support vector regressor were used to calibrate the data collected from the sensors. An accurate reference sensor was colocated with the low cost sensor node to collect the true data and develop the calibration model. The calibration model was used to predict the actual data from the uncalibrated low cost sensor data. The work in [13], developed a low cost sensor node with LoRa-based connectivity to monitor CO, NO_2, and PM levels in the air. To increase the accuracy of these sensors, a polynomial regressor-based calibration technique was also used. It has been observed that the non-linear methods provide better accuracy than the linear methods. The works in [14]–[16] proposed sparse Bayesian learning and deep learning models to calibrate the low cost sensors in a densely deployed wireless sensor network.

The above studies are mainly based on the calibration of low cost sensors. However, a few works have been dedicated

S. Ghosh, P. Das, S. De, and S. Chatterjee are with the UQ-IITD Academy and Department of Electrical Engineering, IIT Delhi, New Delhi, India (e-mail: sushmita.ghosh@uqidar.iitd.ac.in, payali.das@ee.iitd.ac.in, swadesd@ee.iitd.ac.in, shouri@ee.iitd.ac.in). M. Portmann is with the School of IT and Electrical Engineering, University of Queensland, Brisbane, Australia (e-mail: marius@itee.uq.edu.au).

to the design and implementation of low cost air pollution monitoring sensors. The design of an indoor air quality monitoring node was proposed in [17] that comprises of multiple communication interfaces such as MODBUS, LoRa, Wi-fi, and NB-IoT, to compare the performance of the communication protocols in terms of packet loss and time delay. The IoT node is able to sense temperature, humidity, dust, CO_2, and formaldehyde periodically. Similarly, a LoRa-based air pollution monitoring system was proposed in [18] to monitor CO, NO_2, and SO_2, where Raspberry Pi was used as a data processor. The work in [19] proposed a portable air quality monitoring system powered by a solar cell to monitor PM levels and concentration of CO in the air. The work in [20] proposed the design of a low power air quality monitoring wearable sensor node to monitor temperature, humidity, CO, NO_2, O_3, etc. All these works focused on the prototype development of pollution monitoring sensors, however, they did not focus on the calibration of these sensors.

B. Research Gap and Motivation

As discussed in Section I-A, the calibration methods proposed in [12], [13] consider an accurate costly reference node colocated with the low cost sensor node. In such methods, reference sensors need to be deployed to collect actual data at the time of calibration and recalibration to retrain the models and validate the sensing accuracy. The studies in [14]–[16] developed calibration models for densely deployed wireless sensor networks, where all the sensors are low cost. In such cases one sensor can be recalibrated from the other spatially distributed sensors. Although the calibration methods are reference free, the sensors need to be replaced when the calibration errors of all the sensors exceed the threshold.

The work in [13] focused on the design of a low cost sensor node. However, in case of field deployment, energy consumption is one of the major concerns for uninterrupted sensing operation, which was not focused. Hence, the above design is not suitable for field deployment. The works presented in [17]–[19], [21], have not focused on the calibration of low cost sensors. The aim of deploying low cost sensor nodes is to replace the need for high cost reference nodes and make the deployment more cost effective without compromising on the sensing quality. Though, a reference sensor can be used at the initial deployment stage, in the long run the sensors accuracy reduces due to aging, temperature, etc., and they need to be re-calibrated after an undetermined period of calibration [10]. The existing works did not address recalibration of the sensors, which is one of the most concerning factors in the deployment scenario.

The conventional recalibration methods involve lab-based calibration and field calibration. In lab-based calibration method, the sensors are brought back to the lab to recompute the calibration coefficients using an elaborate calibration setup and then again deployed the sensors in the field [22]. In the existing field calibration method [12], a reference sensor node is colocated with the low cost sensor node in the field for a certain duration to collect accurate data to retrain the calibration model. The lab calibration method is highly expensive, whereas, field calibration method requires a reference sensor to be *colocated manually* at the time of recalibration. *Thus, local reference free calibration of the low cost miniature sensors and automatic decision on recalibration are of interest.* To this end, *an autocalibration method of MSNs with the help of sparsely deployed high cost sensing stations (HCSSs) is presented in this paper. Since recalibration of the MSNs are performed automatically, without colocating any reference sensor, this method is named as reference free autocalibration method.* A low cost and low power PM sensor is also developed to collect uncalibrated data and apply the proposed calibration method.

It has been observed that the sensing parameters exhibit cross-correlation at the intra-node level and also exhibit spatial correlation at the inter-node level. Considering this as the key point, *this paper presents a method to detect the dynamically varying optimum recalibration instants of the low cost miniature sensors.* The recalibration interval of the sensors depends on the environmental conditions, sensing quality, aging, etc., which is dynamic. Thus, instead of a fixed period, re-calibrating the sensors dynamically based on the requirement can provide more reliable data and also may reduces the recalibration overhead.

C. Contributions

The key features and contributions of this paper are as follows:

1) An autocalibration method of a low cost MSN is proposed in this paper, where the sensors are calibrated using the data collected from the HCSSs that are deployed in the region. The actual data of the MSN deployed location are estimated from the HCSSs data at the base station.
2) To enable automatic in-field recalibration, cross-correlation among the multiple sensed parameter values in the MSN is exploited in estimating the recalibration timing of the multi-sensing node.
3) Various learning-based regression models are explored to find an optimum regressor to calibrate the MSN. Next, a Gaussian process regressor (GPR) based calibration model is also proposed to calibrate the MSNs data, as GPR outperforms the existing calibration models.
4) A low cost, low power PM sensor is developed to collect real-time uncalibrated data and apply the proposed autocalibration method. The developed PM sensor is 91% more cost efficient and 57% more energy efficient compared to the commercially available high cost PM sensor.
5) The accuracy of the calibrated PM data is validated by comparing the calibrated MSNs data with the commercial PM sensor, which shows that the sensing error lies within the acceptable range.

Organization: Section II introduces the system model. Section III explains the proposed autocalibration method, followed by experimental setup for data collection in Section IV. The results are discussed in Section V, followed by concluding remarks in VI.

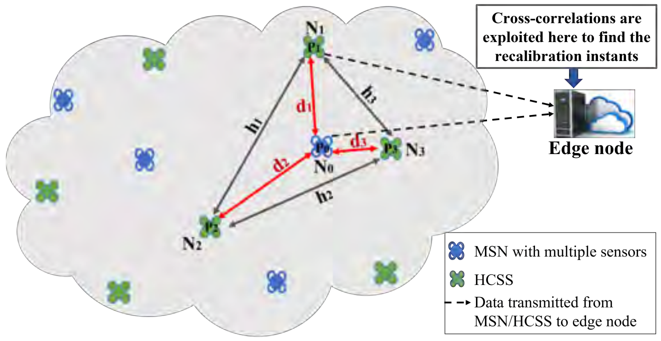

Cross-correlations are exploited here to find the recalibration instants
Edge node
N1
P1
h1
h3
d1
P0
N0
d3
P3
N3
d2
h2
P2
N2
MSN with multiple sensors
HCSS
Data transmitted from MSN/HCSS to edge node

Fig. 1: System model

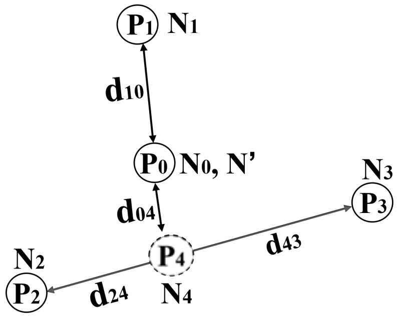


Fig. 2: Estimate the sensing parameter values at point P_0 using section formula.

II. System Model

Consider a sparsely deployed sensor network consisting of high cost and highly accurate air pollution monitoring sensor nodes, as shown in Fig. 1. N_1, N_2, and N_3 are the HCSSs deployed at positions P_1, P_2, and P_3, respectively. The distances between the HCSSs are on the order of kilometers, using which pollution localization and accurate pollution mapping of the whole geographical area are not possible. Deployment of such high cost pollution monitoring stations in a dense manner is difficult to implement in practice. For fine granular pollution mapping and pollution localization low cost MSNs are developed. Such a low cost MSN N_0 consisting of multiple sensors is deployed at P_0 in between the HCSSs. P_1, P_2, and P_3 are at a distance of h_1, h_2, and h_3 from each other. P_0 is at a distance of d_1, d_2, and d_3 from P_1, P_2, and P_3, respectively. All the nodes in the network are connected to a base station to transmit the sensed parameter values. Since, the low cost sensors of the MSNs are less accurate than the HCSSs, the sensing parameters of MSN need to be calibrated/re-calibrated at the edge node on-demand using the data collected from the nearby HCSSs to obtain the accurate values.

III. Proposed Local Reference Free Recalibration

This section describes the proposed local reference free autocalibration method of a MSN deployed in a sparse network of HCSSs, as depicted in the system model in Fig. 1.

A. Local Reference Free Calibration Method

The traditional calibration methods consider that a reference sensor is colocated with the MSN. In such cases the reference sensor node is sufficient to calibrate the MSN. However, in practice, MSNs are deployed to replace the HCSSs. If the MSN is not collinear with the nearby HCSSs, minimum three HCSSs are required to estimated the sensing parameter values at the MSN deployed location. The deviation of the estimated data from the actual data which can be called as estimation error, increases with the increase in distances between the MSN and HCSSs. However, as the pollution generated from a source spread with time, the estimation error in the long term averaged data should be comparatively low. Thus, the estimated long term average data should be similar to the actual data, or there must be a strong correlation between the actual and estimated data, which can be used to predict the actual data from estimated data. As listed in Table I, $Z_p^n(i)$ is the i^{th} instantaneous value of the p^{th} parameter at the n^{th} HCSS. The instantaneous samples of the sensing parameters are collected at a high sampling rate that satisfies the Nyquist criteria, to enable successful reconstruction of the original signal at the edge node [5]. These instantaneous values of the sensing parameters are denoted as short term data, whereas the long term average is computed by averaging the instantaneous values over a long period which is denoted as long term averaged data. The optimum averaging period is decided by minimizing the error between the long term averaged actual data $\bar{Z}_p^{act}$ and the long term averaged estimated data $\bar{Z}_p^{est}$.

To estimate the sensing parameters at P_0, a virtual node N_4 is placed at position P_4, which is the intersecting point of the two lines passing through (N_0, N_1) and (N_2, N_3). Knowing the distance between the nodes, the sensing parameters are first estimated at P_4 using the section formula, as given in (1). These estimated values at P_4 are further used to estimate the sensing parameters at P_0, as given in (2). Thus, the estimated long term averaged data at the point of interest are calculated using (2).

$$\bar{Z}_p^4 = \frac{\bar{Z}_p^2 d_{24} + \bar{Z}_p^3 d_{43}}{d_{24} + d_{43}}. \tag{1}$$

$$\bar{Z}_p^{est} = \frac{\bar{Z}_p^1 d_{10} + \bar{Z}_p^4 d_{04}}{d_{10} + d_{04}}. \tag{2}$$

$\bar{Z}_p^{est}$ has a strong correlation with $\bar{Z}_p^{act}$, which provides the actual data at P_0. Let, $\mathcal{F}_1$ is the underlying estimation function to calculate $\bar{Z}_p^{act}$ from $\bar{Z}_p^{est}$, as expressed in (3), and $\mathbf{A} = \{A_1, A_2, \cdots, A_m\}$ is the vector containing the coefficients of the function. Let $\bar{Z}_p^{pred}$ is calculated from the long term averaged $\bar{Z}_p^{est}$ using (3), which is similar to $\bar{Z}_p^{act}$. The optimum coefficient values of the vector $\mathbf{A}$ should be estimated to minimize the prediction error between $\bar{Z}_p^{pred}$ and $\bar{Z}_p^{act}$.

$$\bar{Z}_p^{act} \approx \bar{Z}_p^{pred} = \mathcal{F}_1(\mathbf{A}, \bar{Z}_p^{est}). \tag{3}$$

The optimum coefficient values of $\mathbf{A} = \{A_1, A_2, \cdots, A_m\}$ are estimated using the least square regressor method by minimizing the deviation between $\bar{Z}_p^{pred}$ and $\bar{Z}_p^{act}$. The optimization function, defined in (4), finds a set of coefficients of the vector $\mathbf{A}$ that minimizes the mean squared error between $\bar{Z}_p^{pred}$ and $\bar{Z}_p^{act}$.

TABLE I: List of Symbols

$Z_p^n(i)$	i^{th} instantaneous value of the p^{th} parameter at the n^{th} HCSS
$\bar{Z}_p^n$	Long term average value of the p^{th} parameter at the n^{th} HCSS
$\bar{Z}_p^{est}$	Long term average value of the p^{th} parameter at position P_0 estimated from the nearby HCSS
$Z_p^{msn}(i)$	i^{th} instantaneous value of the p^{th} parameter collected from uncalibrated MSN N_0
$\bar{Z}_p^{msn}$	Long term average value of the p^{th} parameter collected from uncalibrated MSN N_0
$Z_p^{act}(i)$	i^{th} instantaneous value of the p^{th} parameter collected from high cost reference sensor N' deployed at location P_0, which gives the actual data
$\bar{Z}_p^{act}$	Long term average value of the p^{th} parameter collected from high cost reference sensor N' deployed at location P_0
$\bar{Z}_p^{pred}$	Long term average value of the p^{th} parameter predicted from the estimated data at P_0
$\bar{Z}_p^{cal}$	Long term average value of the p^{th} parameter calibrated from data collected from the MSN at P_0

$$\mathbf{A}^* = \min_{\mathbf{A}} \left\{ \frac{1}{K}\sum_{k=1}^{K} (\bar{Z}_p^{act}(k) - \mathcal{F}_1(\mathbf{A}, \bar{Z}_p^{est}(k)))^2 \right\}. \tag{4}$$

The predicted data can be used to calibrate the data collected from the MSN N_0 using (5). The input to the calibration model is the long term average of uncalibrated MSN data $\bar{Z}_p^{msn}$ and the output is the predicted data $\bar{Z}_p^{pred}$, as given in (5). The calibrated output is defined as $\bar{Z}_p^{cal}$. Once the calibration function $\mathcal{F}_2$ is obtained from the long term average of $\bar{Z}_p^{msn}$ and $\bar{Z}_p^{pred}$, it can be used to calibrate the instantaneous values of the MSN using (6).

$$\bar{Z}_p^{act} \approx \bar{Z}_p^{pred} \approx \bar{Z}_p^{cal} = \mathcal{F}_2(\mathbf{B}, \bar{Z}_p^{msn}). \tag{5}$$

$$Z_p^{act}(i) \approx Z_p^{pred}(i) \approx Z_p^{cal}(i) = \mathcal{F}_2(\mathbf{B}, Z_p^{msn}(i)). \tag{6}$$

Similar to (4), an optimization function is defined in (7) to find the optimum coefficient values of $\mathbf{B} = \{B_1, B_2, \cdots, B_n\}$, by minimizing the mean squared error between $\bar{Z}_p^{cal}$ and $\bar{Z}_p^{pred}$, that subsequently minimizes the error between $\bar{Z}_p^{cal}$ and $\bar{Z}_p^{act}$, which is defined as the calibration error.

$$\mathbf{B}^* = \min_{\mathbf{B}} \left\{ \frac{1}{K}\sum_{k=1}^{K} (\bar{Z}_p^{pred}(k) - \mathcal{F}_2(\mathbf{B}, \bar{Z}_p^{msn}(k)))^2 \right\}. \tag{7}$$

In (4) and (7), K is the total number of samples available in the dataset to estimate the coefficients at the beginning. $\bar{Z}_p^{act}(k)$, $\bar{Z}_p^{est}(k)$, and $\bar{Z}_p^{msn}(k)$ are respectively the k^{th} long term averaged actual, estimated, and uncalibrated MSN data.

A reference sensor node N' at P_0 is placed at the beginning to collect actual data to find $\mathcal{F}_1$. It is assumed that the function $\mathcal{F}_1$ is valid for position P_0 and does not change with time. Thus, the reference node at P_0 is not required further to calibrate and recalibrate the MSN node at P_0. Hence this method is reference free. However, to maintain the accuracy of the proposed autocalibration model, the function $\mathcal{F}_1$ needs to be redefined and its coefficients needs to be recomputed by placing a reference sensor node at P_0 during the recalibration of HCSSs.

The proposed autocalibration method is applicable to any WSN consisting of sparsely deployed HCSSs and low cost MSNs, where the deployment location of a MSN should be within the area covered by three or more nearby HCSSs.

B. Machine Learning-based Calibration Model

As discussed in Section III-A, suitable regressors have to be chosen to find $\mathcal{F}_1$ and $\mathcal{F}_2$ that minimizes the prediction error and the calibration error. Initially $\mathcal{F}_1$ is derived to predict the actual dataset. Further, the predicted dataset is used to calibrate the MSN using $\mathcal{F}_2$. However, the accuracy of the calibration model decreases with time, which needs to be recalibrated after a certain period by collecting data from the HCSSs.

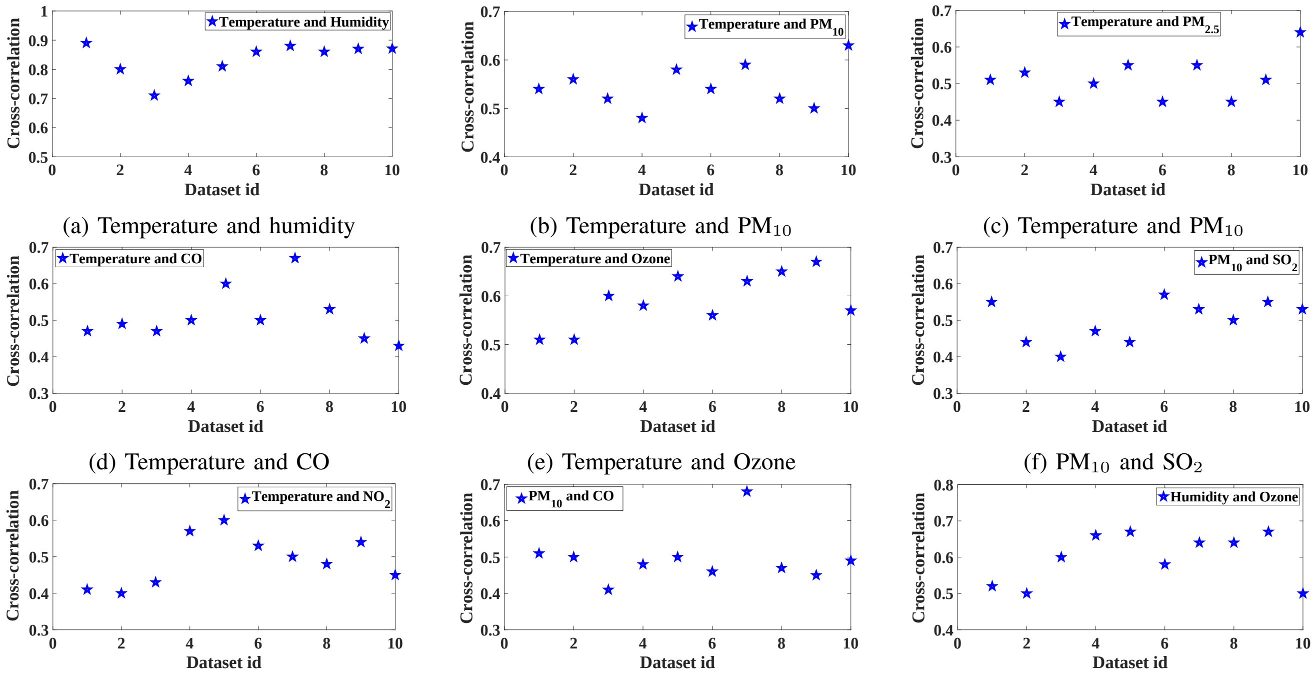
Various machine learning models are available in literature to calibrate the MSNs [23]. In order to find appropriate functions $\mathcal{F}_1$ and $\mathcal{F}_2$, four different regressor models, namely, multivariate linear regressor (MLR), Polynomial regressor, Support Vector regressor (SVR), and Gaussian process regressor (GPR) are used in this work. To find the estimation function, the input and the target vectors of the regressors are respectively $\bar{\mathbf{Z}}_p^{est} = \{\bar{Z}_p^{est}(1), \cdots, \bar{Z}_p^{est}(J)\}^T$ and $\bar{\mathbf{Z}}_p^{act} = \{\bar{Z}_p^{act}(1), \cdots, \bar{Z}_p^{act}(J)\}^T$, where J is the number of training samples. Similarly, to find the calibration function, the input and the target vectors of all the regressors are respectively $\bar{\mathbf{Z}}_p^{msn} = \{\bar{Z}_p^{msn}(1), \cdots, \bar{Z}_p^{msn}(J)\}^T$ and $\bar{\mathbf{Z}}_p^{pred} = \{\bar{Z}_p^{pred}(1), \cdots, \bar{Z}_p^{pred}(J)\}^T$.

The linear regressor model involves a linear combination of the input variables. Since, the input is one-dimensional, the output of the calibration model is expressed as, $\bar{Z}_p^{cal} = a_0 + a_1\bar{Z}_p^{msn}$, where $\mathbf{A} = [a_0, a_1] \in \mathbb{R}^{1\times 2}$ [24]. While training the model, the optimum values of a_0 and a_1 are estimated for different combinations of $\bar{Z}_p^{pred}$ and $\bar{Z}_p^{msn}$.

If the relation between $\bar{Z}_p^{msn}$ and $\bar{Z}_p^{pred}$ is non-linear, the linear regressor can not provide the best fit. In such cases, polynomial regressor model can be used to find the underlying calibration function. The output is expressed as, $\bar{Z}_p^{cal} = a_0 + a_1\bar{Z}_p^{msn} + a_2(\bar{Z}_p^{msn})^2 + \cdots + a_M(\bar{Z}_p^{msn})^M$, where $\mathbf{A} = [a_0, a_1, a_2, \cdots, a_M] \in \mathbb{R}^{1\times M}$ [24].

The linear and polynomial regressors try to find the underlying function in its original domain, however, the data may not be strongly correlated in its original domain. SVR transforms the data to high-dimensional space by using kernel functions and performs linear regression. Thus, $\bar{Z}_p^{msn}$ is transformed to $\phi(\bar{Z}_p^{msn})$ using the kernel functions, such as, linear, polynomial, radial basis function, etc. and the output of the calibration model is expressed as, $\bar{Z}_p^{cal} = a_0 + a_1\phi(\bar{Z}_p^{msn})$, where $\mathbf{A} = [a_0, a_1] \in \mathbb{R}^{1\times 2}$ [24].

SVR considers a fixed parametric model of the data, which may not be valid in case of non-stationary sensing signals. In such cases GPR performs better than SVR, polynomial, and linear regressor models [25]. To predict



(a) Temperature and humidity

(b) Temperature and PM$_{10}$

(c) Temperature and PM$_{10}$

(d) Temperature and CO

(e) Temperature and Ozone

(f) PM$_{10}$ and SO$_2$

(g) Temperature and NO$_2$

(h) PM$_{10}$ and CO

(i) Humidity and Ozone

Fig. 3: Cross-correlation among the sensing parameters collected from high cost reference sensors.

$\bar{Z}_p^{pred}$ from $\bar{Z}_p^{msn}$, GPR finds the underlying function $\mathcal{F}_2$ as, $\bar{Z}_p^{pred}(*) \approx \bar{Z}_p^{cal}(*) = F_2(\bar{Z}_p^{msn}(*)) = F_2(*)$, such that, $F_2 \sim N(0, K_{J\times J})$, where $K_{J\times J}$ is the covariance matrix [25]. According to [26], [27], the mean and the covariance functions of the GPR model are respectively

$$\overline{F_2(*)} = K(\bar{\mathbf{Z}}_p^{msn}, \bar{Z}_p^{msn}(*))[K(\bar{\mathbf{Z}}_p^{msn}, \bar{\mathbf{Z}}_p^{msn}) + \sigma^2 I]^{-1}\bar{\mathbf{Z}}_p^{pred}. \tag{8}$$

$$\begin{aligned}\mathrm{Cov}(F_2(*)) = k(\bar{Z}_p^{msn}(*), \bar{Z}_p^{msn}(*)) + K(\bar{Z}_p^{msn}(*), \bar{\mathbf{Z}}_p^{msn})\\ [K(\bar{\mathbf{Z}}_p^{msn}, \bar{\mathbf{Z}}_p^{msn}) + \sigma^2 I]^{-1} K(\bar{\mathbf{Z}}_p^{msn}, \bar{Z}_p^{msn}(*)).\end{aligned} \tag{9}$$

Using the proposed method, the MSN can be calibrated/recalibrated without the need of actual data collected from the reference sensor at P_0, as the data collected from the HCSSs can predict the reference data at P_0 using the estimation function.

C. Automatic Recalibration of Low Cost Air Pollution Monitoring Sensors in MSN

The sensors are initially calibrated in the lab environment before being deployed in the field. However, the sensing accuracy decreases with time. In such cases, bringing them to the lab for recalibration is not practically feasible always. The low cost sensors need to be recalibrated more frequently than the HCSSs. In such cases, recalibrating the low cost sensors unnecessarily increases the overhead cost.

Consider that a MSN consists of multiple sensors for monitoring various parameters in the environment. In this work it is proposed to exploit the cross-correlation among the parameters to find the optimum recalibration instants. Under normal circumstances, cross-correlation between two parameters could be strong or medium. Since the calibration error is random and increases with time, the cross-correlation decreases, which indicates the recalibration instants of the corresponding sensors.

To explore the cross-correlation among the air pollution monitoring parameters, datasets of ten monitoring stations were collected from the website of [28]. Each dataset contains eight parameters, namely, PM$_{2.5}$, PM$_{10}$, CO, O$_3$, NO$_2$, SO$_2$, temperature, humidity. Fig. 3 shows the cross-correlation among the parameters for ten monitoring stations. Dataset id in Fig. 3 denotes the index of each dataset, collected from one monitoring station. In most cases, the cross-correlation is above 0.4, however, in some cases it is above 0.5 and 0.7. Thus, a suitable correlation threshold can be chosen from this observation. If cross-correlation between any two parameters, mentioned in Fig. 3 falls below the threshold, the corresponding sensors need to be recalibrated. The recalibration instants can be verified by collecting the data from the HCSSs. For the system model in Fig. 1, the correlation between the estimated and calibrated data is much higher than the correlation between the estimated data and uncalibrated low cost sensor data. As the sensing accuracy reduces with time, the correlation coefficient also decreases, which indicates the recalibration instants.

The multi-sensing nodes send data to edge node, where the MSNs data are calibrated to find Z_p^{cal}. Simultaneously, the cross-correlation among the parameters are exploited at the edge node. If the correlation coefficient between any two parameters falls below a threshold, the edge node collects data from the three HCSSs and compute the estimated data Z_p^{est} of those sensing parameters. If the correlation between the Z_p^{cal} and Z_p^{est} also falls below a threshold, the edge node retrains the calibration models by collecting recent samples

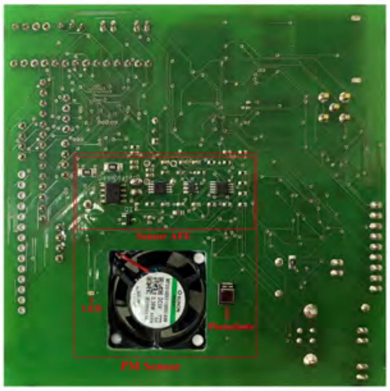


Fig. 4: Prototype implementation of low cost PM sensor.

from the MSN and the HCSSs. The estimation of optimum cross-correlation threshold is discussed in Section V-E.

It is notable that in the proposed reference free calibration method the recalibration decision at the edge node is purely based on the multiple sensed parameters at the MSN. The remotely located HCSSs data are then collected by the nearby edge node to which the MSN is connected and this dataset is used to recompute the calibration coefficients of the MSN.

IV. Experimental Setup For Data Collection

The proposed autocalibration method is applicable to any low cost air pollution parameter, such as, PM, CO, O_3, NO_2, SO_2, etc. In this exercise, a low cost PM sensor is designed to validate the efficacy of this method. Section IV-A and IV-B explains the design and deployment of low cost PM sensor.

A. Design of Low Cost PM sensor

To validate the proposed autocalibration method, discussed in Section III, a MSN consisting of a low cost PM sensor of dimension 4.5 cm $\times$ 5.5 cm has been developed. According to the design reference in [29], an optical particle counter consists of a low cost light emitting source, such as LED, a photodetector, a series of transimpedance amplifiers, and a fan. The designed PM sensor is shown in Fig. 4, which is able to monitor $PM_{2.5}$ and PM_{10} in the air. The sensor is completely covered by a black box with one side of the fan exposed to the air such that no external light can enter inside the box. The fan is placed in between the LED and photodetector in the board to draw air from outside environment. The emitted light from the LED passes through the PM contaminated air and falls on the photodetector. Based on the incident intensity of the light, the amplitude of the current through the photodetector changes. The analog front end (AFE) of the sensor, shown in Fig. 4, consists of a series of transimpedance amplifiers and filters to amplify the current of the photodetector, convert it to a voltage signal, and eliminate the high frequency noise of the signal. The amplified and filtered voltage is finally measured at the output. Based on the voltage level, the concentration of the PM particles is calculated. An Arduino UNO board consists of ATMEGA328P microcontroller is used to read the output voltage of the PM sensor. The microcontroller is programmed using Arduino integrated development environment (IDE) to compute the concentration levels of $PM_{2.5}$ and PM_{10} in μg/m^3 from the voltage detected at the output of the PM sensor [29].

B. Deployment of Low Cost PM Sensor

The developed low cost miniature PM sensor has been deployed in the campus of IIT Delhi. Along with the MSN, a high cost OPC N3 PM sensor is also deployed to collect accurate data initially to compare with the calibrated data and validate the autocalibration method discussed in Section III. Although OPC N3 does not perform like the HCSSs, it is calibrated using a Beta attenuation monitor (BAM) to provide accurate data [30]. $PM_{2.5}$ and PM_{10} data are collected from both the MSN and the OPC N3 PM sensor in October 2021 at the campus of IIT Delhi. As shown in Fig. 5(a), the low cost PM data are collected using Arduino board. The OPC N3 PM sensor is also placed along with it. The MSN is surrounded by three nearest air pollution monitoring stations deployed by the central Pollution Control Board (CPCB), depicted in Fig. 5(b) [28]. Let N_1, N_2, and N_3 be respectively the HCSSs deployed at locations P_1 (Sri Aurobindo Marg), P_2 (R. K. Puram), and P_3 (Sirifort). They are 2 km away from the MSN N_0, deployed at location P_0 (inside IIT Delhi).

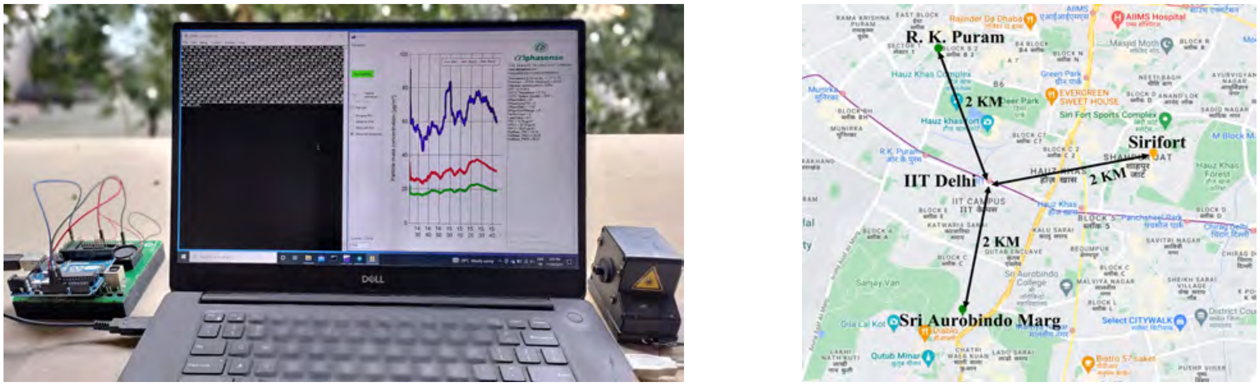
The uncalibrated data collected by low cost MSN as well as the calibrated data from the HCSSs are transmitted to the edge node, where the calibration algorithm is implemented to calibrate the MSN data. Since the calibration coefficients vary with the dynamics of the environment, the calibration models are retrained at the edge node after an optimal recalibration interval (discussed in Section III-C) using the previously stored data.

V. Results and Discussions

As discussed in Section IV-A, a low cost PM sensor has been developed based on the design circuitry provided in [29]. This sensor module has been used for studying our proposed autocalibration method. It is intended to validate the accuracy of low cost PM sensor such that the sensing error remains with in an acceptable range.

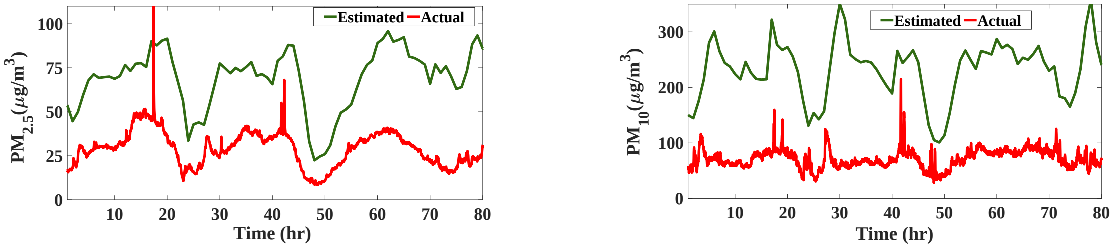
A. Estimation of Actual Data $\bar{Z}_p^{act}$ from Long Term Averaged Estimated Data $\bar{Z}_p^{est}$

$PM_{2.5}$ and PM_{10} data were collected inside the IIT Delhi campus using both the low cost PM sensor and the OPC N3 PM sensor, which are denoted respectively as N_0 and N'. The HCSS datasets N_1, N_2, and N_3 were collected from the website of CPCB [28]. Before determining the Nyquist criteria, the data are collected from the sensor at a very high rate such as $f_s > 1$ Hz, where f_s is the rate of oversampling. From the temporal samples, the maximum frequency F_m of each parameter was computed such that 99% of the total energy of the signal is concentrated within that frequency range. From the power spectral density (PSD), the observed maximum frequency of $PM_{2.5}$ and PM_{10} were noted to be respectively 0.0062 Hz and 0.008 Hz. Thus, the Nyquist sampling rate for $PM_{2.5}$ and PM_{10} are respectively



(a) (b)

Fig. 5: (a) Experimental setup for data collection; (b) high cost air pollution monitoring nodes deployed by CPCB [28].



(a) PM$_{2.5}$ (b) PM$_{10}$

Fig. 6: Short term variation of (a) Estimated PM$_{2.5}$ and actual PM$_{2.5}$; (b) Estimated PM$_{10}$ and actual PM$_{10}$.

$F_s = 2F_m = 0.0124$ and $F_s = 2F_m = 0.016$, which gives a sampling interval of $T_s \approx 80$ sec for PM$_{2.5}$ and $T_s \approx 62$ sec for PM$_{10}$. As a little conservative measure, in our experiment the sampling interval of the PM sensor was set as 60 sec for collecting the instantaneous samples or the short term data from both the low cost and the reference sensor.

After collecting the initial dataset, the long term averaged values of PM$_{2.5}$ and PM$_{10}$ at P_0 are estimated from the long term averaged HCSS data using (1) and (2). Fig. 6 presents the short term variation of estimated and actual data with time. It can be observed that the estimated data are not able to provide the local variations at P_0, hence a sensor needs to be deployed to monitor the local variations of the pollutants accurately to provide the real-time data. However, the estimation error reduces for the long term averaged data.

To find the estimation error, mean absolute error (MAE) is considered as the performance metric. If $\bar{Z}_p^{est}(j)$ and $\bar{Z}_p^{act}(j)$ are respectively the long term averaged j^{th} samples of the estimated dataset and the reference dataset for the p^{th} parameter, the MAE is given by

$$e_e = \frac{1}{J}\sum_{j=1}^{J} |(\bar{Z}_p^{est}(j) - \bar{Z}_p^{act}(j))|. \tag{10}$$

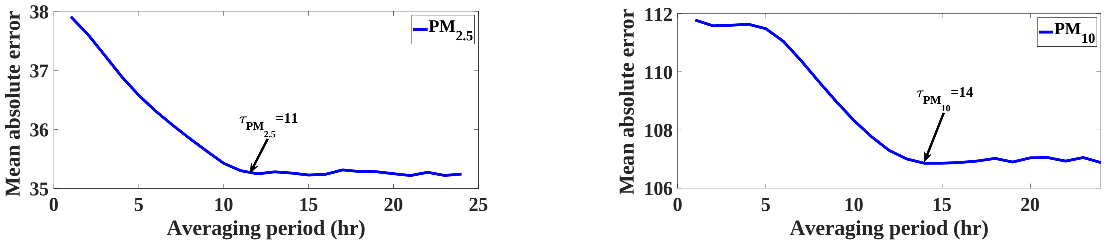
To find the optimum averaging periods $\tau_{PM_{2.5}}$ and $\tau_{PM_{10}}$ for PM$_{2.5}$ and PM$_{10}$, respectively, e_e is computed using (10) with varying τ. Fig. 7, presents the variation of e_e with τ. The optimum values are set as $\tau_{PM_{2.5}} = 11$ hrs and $\tau_{PM_{10}} = 14$ hrs.

The variations of the estimated and actual values of PM$_{2.5}$ with 11 hrs moving average and PM$_{10}$ with 14 hrs moving average are respectively shown in Figs. 8(a) and 8(c). Although Figs. 8(b) and 8(d) show that a linear relation exits between the averaged estimated and actual data, four different types of regressors, such as, linear regressor, polynomial regressor, SVR, and GPR were used to predict the actual data from the estimated data.

To analyze the accuracy of the predicted PM data, root mean squared error (RMSE) is considered as the performance metric. If $\bar{Z}_p^{pred}(j)$ and $\bar{Z}_p^{act}(j)$ are respectively the long term averaged j^{th} samples of the predicted dataset and the reference dataset, the RMSE is given by

$$e_p = \sqrt{\frac{1}{J}\sum_{j=1}^{J} (\bar{Z}_p^{pred}(j) - \bar{Z}_p^{act}(j))^2}. \tag{11}$$

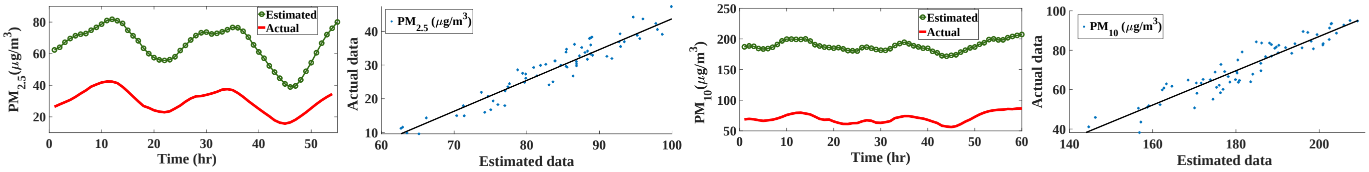
A comparison of prediction error using different regressors is listed in Table II. It can be observed that the RMSE values are similar. According to [31], RMSE < 7 μg/m^3 is acceptable for PM parameters. Thus, all the regressor models are meeting the accuracy of the data. However, linear regressor is the simplest model with least computational overhead. Hence, a linear function can be adopted to predict the actual data from the estimated data such that $\bar{Z}_p^{pred} = \alpha\bar{Z}_p^{est} + \beta$. By solving (4), the optimum values of α and β are obtained as, $\alpha^* = 0.598$, $\beta^* = -9.236$ for PM$_{2.5}$ and $\alpha^* = 0.116$, $\beta^* = 47.38$ for PM$_{10}$. Variations of averaged estimated, actual, and predicted values for PM$_{2.5}$ and PM$_{10}$ are shown in Fig. 9.



(a) $\mathrm{PM}_{2.5}$

(b) PM_{10}

Fig. 7: Variation of mean absolute error of (a) $\mathrm{PM}_{2.5}$, and (c) PM_{10} with averaging period.



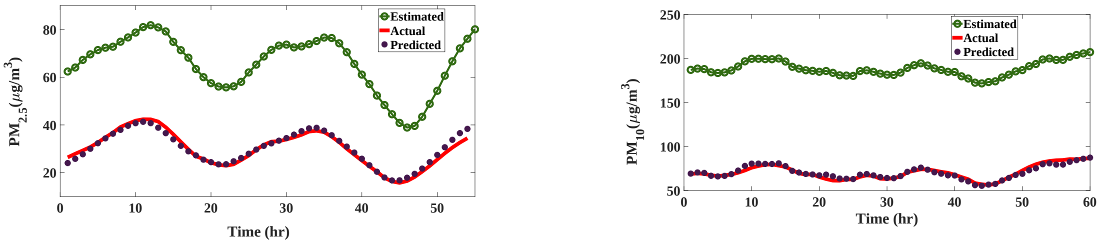
(a) $\mathrm{PM}_{2.5}$

(b) $\mathrm{PM}_{2.5}$

(c) PM_{10}

(d) PM_{10}

Fig. 8: Variation of long term averaged (a) $\mathrm{PM}_{2.5}$ and (c) PM_{10} with time, (b) Estimated $PM_{2.5}$ versus actual $PM_{2.5}$, and (d) Estimated PM_{10} versus actual PM_{10}.



(a) $\mathrm{PM}_{2.5}$

(b) PM_{10}

Fig. 9: Comparison of the estimated, actual, and predicted long term average values of (a) $\mathrm{PM}_{2.5}$, and (b) PM_{10}.

TABLE II: Comparison of error between predicted and actual data for different regressors

Regressor	Linear	Polynomial	Support vector regressor	Gaussian process regressor
Prediction error of $\mathrm{PM}_{2.5}$ (RMSE in μg/m^3)	0.21	0.20	0.21	0.21
Prediction error of PM_{10} (RMSE in μg/m^3)	0.44	0.43	0.44	0.42

TABLE III: Variation of prediction error with the area of triangle

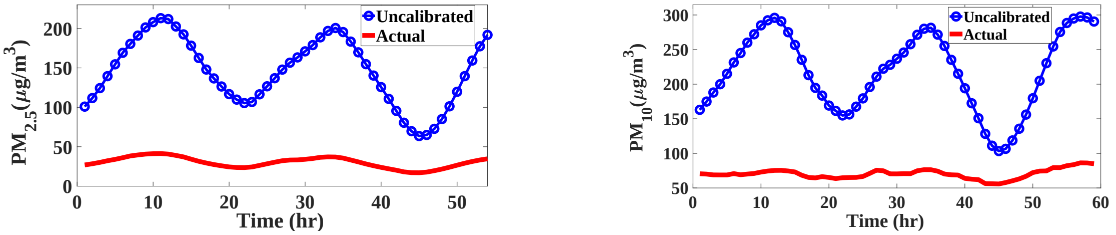
Area (km^2)	3.675	18.253	36.412	44.159
Prediction error of PM_{10} (RMSE in μg/m^3)	0.44	1.1	1.6	1.7
Prediction error of $\mathrm{PM}_{2.5}$ (RMSE in μg/m^3)	0.21	0.51	1.2	1.4

B. Variation of Prediction Error with Distance of MSN from HCSSs

The prediction error between the actual and predicted signals depends on the distance between the HCSSs and the MSN. Considering three HCSSs at a time, the signal estimation method is applied on multiple HCSSs deployed far from the MSN to find an optimum area of the triangle to estimate the actual signal. Table III shows that the error increases with the increase in area of the triangle, i.e., when the chosen HCSSs are more and more away from the MSN position. As the predicted data is further used to calibrate the low cost MSNs, the calibration error is higher than the prediction error. Thus, considering the prediction error threshold in terms of RMSE as 1.6 μg/m^3 [32], and the calibration error threshold as 7 μg/m^3 [31], the reference HCSS locations chosen such that the area of the triangle is below 36 km^2. This observation of increased prediction error as a function of HCSS distance highlights the limit of the proposed local reference free calibration approach.

C. GPR based Calibration Model

A comparison of data collected from uncalibrated MSN and the actual data for $\mathrm{PM}_{2.5}$ and PM_{10} is shown in Fig. 10. It can be observed that the MSN over-estimates the data and hence it needs to be calibrated to find the accurate values. As discussed in Section I-A and III-B, various machine learning models, such as, linear regressor [12], polynomial regressor



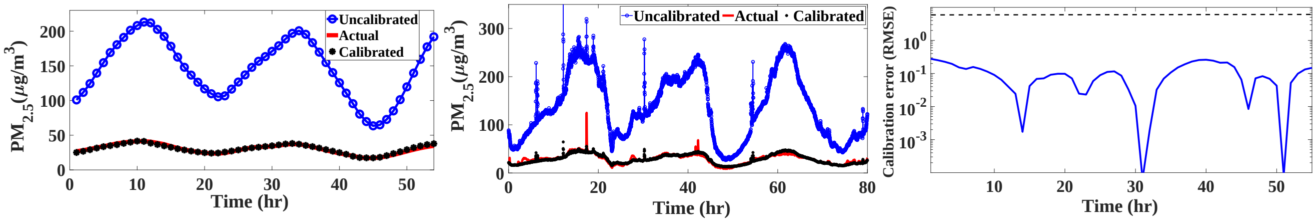
(a) PM$_{2.5}$

(b) PM$_{10}$

Fig. 10: Long term average of the particulate matter (PM) data collected from the uncalibrated MSN and the high cost reference sensing station installed in IIT Delhi campus.

TABLE IV: Comparison of sensing error between calibrated MSNs data and actual data

Regressor	Linear [12]	Polynomial [13]	Support vector regressor [12]	GPR (proposed)
Calibration error of PM$_{2.5}$ (RMSE in μg/m^3)	0.98	0.8	0.94	0.76
Calibration error of PM$_{10}$ (RMSE in μg/m^3)	1.14	1.10	1.18	1.04



(a)

(b)

(c)

Fig. 11: Comparison of uncalibrated, actual, and calibrated sensor data: (a) Long term average values of PM$_{2.5}$; (b) Instantaneous values of PM$_{2.5}$; (c) Variation of calibration error of PM$_{2.5}$.

[13], SVR [12], are used to calibrate the low cost sensors data. In this work, GPR based calibration model is proposed due to its higher prediction accuracy compared to other regressors [27].

The model is trained and tested using the long term averaged values for both PM$_{2.5}$ and PM$_{10}$, where $\bar{Z}_p^{msn}$ and $\bar{Z}_p^{act}$ are respectively the input and output of the model for the p^{th} parameter. Based on the training, once the expected accuracy is achieved, the same calibration model is used to predict the actual instantaneous values of PM data from the instantaneous values of the uncalibrated low cost PM sensor.

To analyze the accuracy of the calibrated PM data, RMSE is considered as the performance metric. If $Z_p^{cal}(j)$ and $Z_p^{act}(j)$ are respectively the j^{th} instantaneous values of the calibrated dataset and the reference dataset, the RMSE is given by

$$e_c = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(Z_p^{cal}(j) - Z_p^{act}(j))^2}. \qquad (12)$$

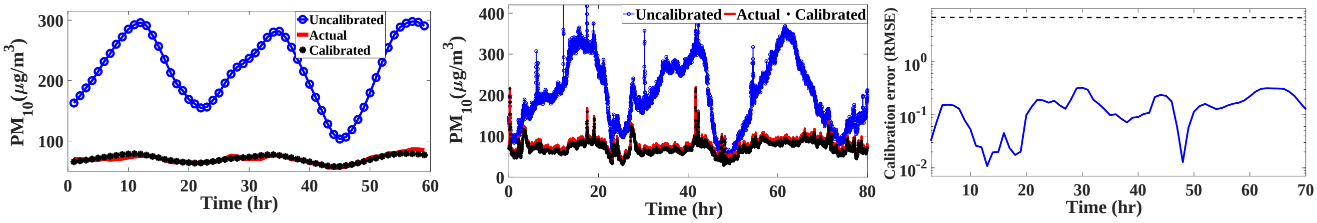
The error threshold is set as $e_c^{th} = 7\mu$g/m^3 [31]. The e_c values for the calibration models with different regressors, computed using (12), are listed in Table IV. The calibration error is calculated from the instantaneous values of the actual and calibrated dataset of PM$_{2.5}$ and PM$_{10}$. It can be observed that, using all the regressors the error achieved is lower than the threshold. Since GPR gives the minimum error, it is chosen for calibrating the designed low cost PM sensor in this work.

Comparison of averaged actual PM data (collected from OPC N3), uncalibrated low cost PM data, and the calibrated PM data are shown in Fig. 11(a) and Fig. 12(a) for PM$_{2.5}$ and PM$_{10}$, respectively. Once the calibration coefficients are estimated, the same model is used to calibrate the instantaneous values of PM$_{2.5}$ and PM$_{10}$. Fig. 11(b) and Fig. 12(b) presents the short term variation of PM$_{2.5}$ and PM$_{10}$, respectively. It can be observed that, the GPR based calibration model is able to find the actual data from the uncalibrated MSNs data. Fig. 11(c) and Fig. 12(c) show the variation of calibration error with time. The figures show that the error lies within the threshold after calibration.

From Fig. 13 it is observed that the training and cross validation errors of the calibration model are minimum if the training length is around $45 - 50$ samples for both PM$_{2.5}$ and PM$_{10}$. Thus, the prediction error is also minimum for this range of training samples. Hence, the regressors are trained using 45 samples of the long term averaged dataset.

D. Automatic Recalibration Performance

To validate the cross-correlation method for finding the optimum recalibration instants, as discussed in Section III-C, the following approach in taken. A calibrated multisensing node is used to capture the variation of environmental parameters, and their cross-correlation values are computed. Along with this, the cross-correlation of the sensed data from an uncalibrated multisensing node is also computed. The two cross-correlation

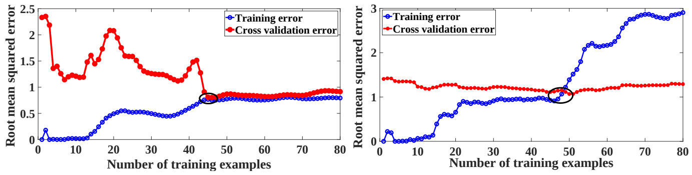


(a)

(b)

(c)

Fig. 12: Comparison of uncalibrated, actual, and calibrated sensor data: (a) Long term average values of PM$_{10}$; (b) Instantaneous values of PM$_{10}$; (c) Variation of calibration error of PM$_{10}$.



(a) PM$_{2.5}$

(b) PM$_{10}$

Fig. 13: Variation of training and cross validation error of (a) PM$_{2.5}$, and (b) PM$_{10}$ with number of training examples.

matrices are compared to make a decision on recalibration of the uncalibrated node. The sensors include DHT11 to monitor temperature and humidity, Alphasense OPC N3 PM sensor to monitor PM$_{2.5}$ and PM$_{10}$, and Alphasense AFE A4 Gas sensors to monitor CO, O$_3$, NO$_2$, SO$_2$. Thus, six sensor modules monitor eight parameters. Although 28 combinations are created from eight parameters considering two at a time to find the cross-correlation, it has been observed that only few combinations exhibit average correlation coefficient above 0.5 with calibrated data. Hence, these are considered as correlated, whereas the other combinations of parameters are not correlated [33]. For the correlated parameters, Fig. 14 presents the variation of cross-correlation with time for both the calibrated and uncalibrated data. It can be clearly observed that the cross-correlation of uncalibrated data is lower than that of the calibrated data. Thus, a cross-correlation threshold can be set to find optimum time for recalibration.

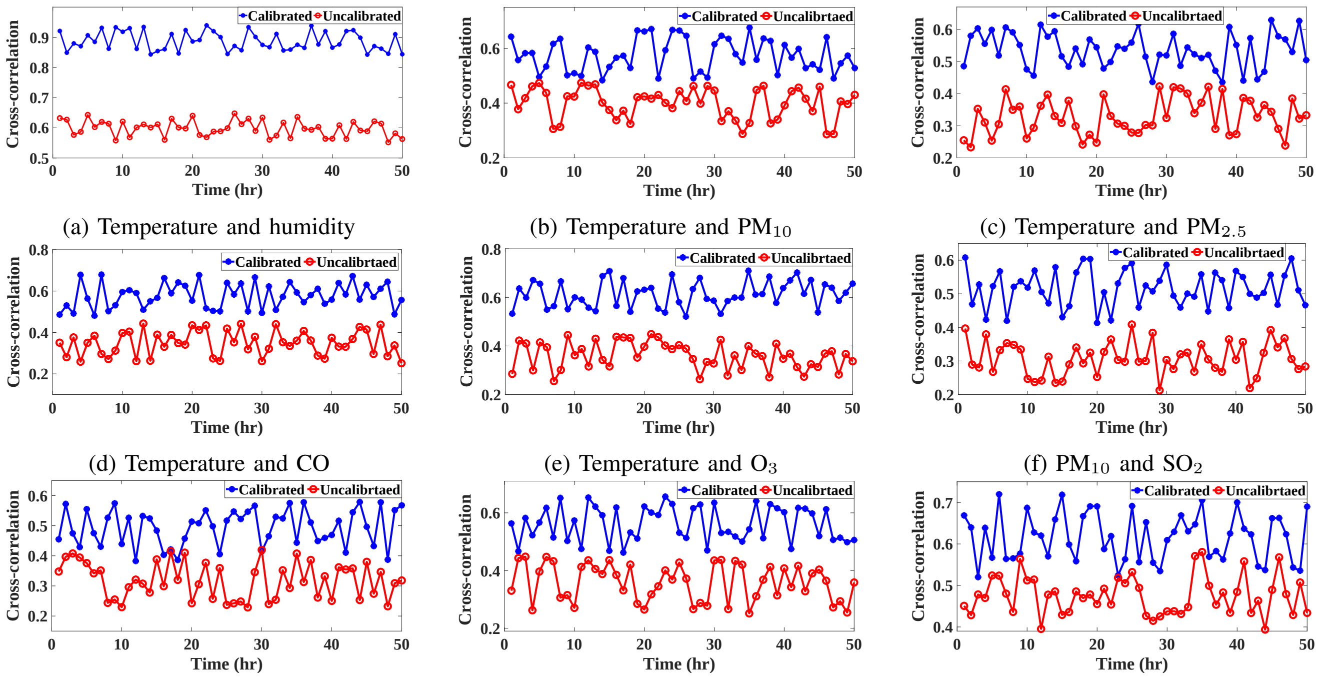
The data collected from the old sensors over-estimates the actual parameter values, as shown in Fig. 15. Hence the data are calibrated using GPR models, where the input to the regressor is the uncalibrated data and the output is the actual data collected from the calibrated new sensors. Fig. 15 shows that the old sensor data follows the actual data after calibration, which validates the efficiency of the GPR based calibration model.

E. Recalibration Overhead

In the presence of HCSSs, the MSN can be recalibrated using the data collected from the cloud. In this case, the calibration overhead is negligible as the sensors need not be calibrated in the lab or no reference sensor need to be colocated to collect actual data for recalibration. However, in the absence of HCSSs, the MSN has to be recalibrated by colocating a high cost sensor node consists of reference sensors. The reference sensor node visits the deployed location, collects sufficient samples to recalibrate the models, and transmits them to the edge node, where the models are recalibrated. In this case the calibration overhead in terms of cost or energy is very large. Thus, estimating optimum recalibration instants can reduce the overhead.

To validate the efficiency of the cross-correlation based detection of optimum recalibration instants, three months data have been collected from the uncalibrated MSN and the calibrated high cost sensors. The PM and temperature data collected from the low cost sensors have been calibrated using the PM and temperature data collected from the high cost sensors. The calibration error of the low cost sensors increases with time. When the sensors give erroneous reading, it leads to reduction in cross-correlation among the parameters. Consider the example of temperature and PM$_{2.5}$. They exhibit a good correlation in the range of $0.5-0.6$, when the sensors give accurate measurements. However, Fig. 16a shows that the correlation decreases with time, as the measurement error of the sensors increases. Fig. 16a presents the variation of cross-correlation between the calibrated absolute temperature (AT) and PM$_{2.5}$ data along with the calibration error of PM$_{2.5}$ with time. It can be observed that the calibration error increases and cross-correlation decreases with time. Since the calibration error is unknown in the absence of reference data, a suitable cross-correlation threshold between AT and PM$_{2.5}$ is set to find the optimum recalibration instants. In case of periodic recalibration, the calibration error in the data is exploited and a suitable time is estimated beforehand to calibrate the MSNs periodically. If the error threshold is set as 0.09, the recalibration period is 15 days, as shown in Fig. 16a. Let t be the overhead for one-time recalibration of the sensor node. Thus, the total recalibration overhead in three months for a fixed recalibration period is $6t$. However, the sensing signals vary dynamically, and hence the recalibration period also varies dynamically. In Fig. 16a, it has been observed that the cross-correlation is 0.57 when the calibration error of 0.09. Thus, a cross-correlation threshold c_{th} is set as 0.57 to find the optimum recalibration instants, such that the calibration error remains within the threshold.

Considering $c_{th} = 0.57$, the calibration models are retrained with the recent samples when the cross-correlation between the calibrated temperature and PM$_{2.5}$ falls below the threshold. Fig. 16b shows that the cross-correlation varies dynamically. The models are retrained adaptively, which gives



(a) Temperature and humidity

(b) Temperature and PM_{10}

(c) Temperature and $PM_{2.5}$

(d) Temperature and CO

(e) Temperature and O_3

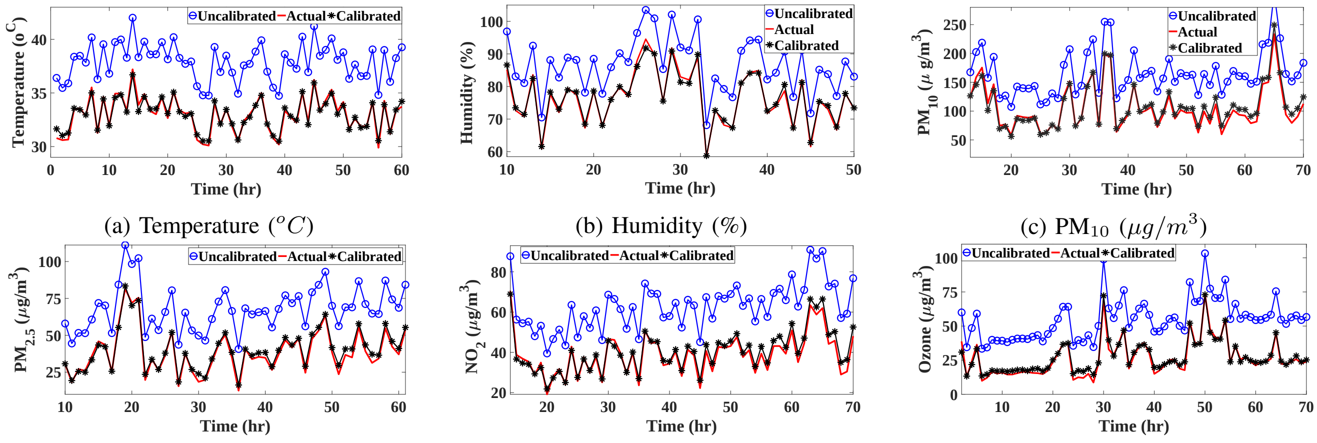
(f) PM_{10} and SO_2

(g) Temperature and NO_2

(h) PM_{10} and CO

(i) Humidity and O_3

Fig. 14: Comparison of cross-correlation among the sensing parameters collected from calibrated and uncalibrated low cost sensors.



(a) Temperature (^oC)

(b) Humidity (%)

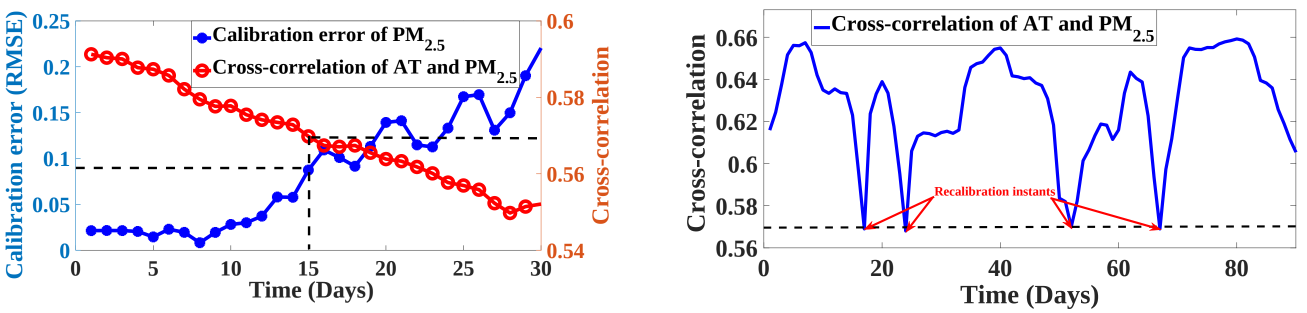
(c) PM_{10} $(\mu g/m^3)$

(d) $PM_{2.5}$ $(\mu g/m^3)$

(e) NO_2 $(\mu g/m^3)$

(f) Ozone $(\mu g/m^3)$

Fig. 15: GPR based calibration of the uncalibrated sensing parameters.



(a)

(b)

Fig. 16: (a) Variation of cross-correlation between temperature and $PM_{2.5}$, calibration error of $PM_{2.5}$ with time, and (b) Cross-correlation based adaptive retaining of the calibration models.

TABLE V: Performance comparison

Performance parameters	OPC N3 PM sensor [30]	low cost PM sensor
Energy consumption of PM sensor per hour (J)	1773	749
Sensing error (RMSE in μg/m^3)	reference	0.76 for PM$_{2.5}$ and 1.04 for PM$_{10}$
Cost of PM sensor (USD)	427.25	36.50

a total recalibration overhead of $4t$ in three months. Thus, the recalibration overhead is reduced by $(\frac{6t-4t}{6t}) \times 100\% = 33.33\%$. In this analysis, the low calibration error threshold is chosen to estimate the overhead from the limited number of samples. However, the recalibration interval increases if the error threshold is increased.

F. Energy Saving Performance

The energy consumption of the designed low cost PM sensor is compared with that of the reference OPC N3 PM sensor. This section presents the performance comparison of the low cost PM sensor and the proposed autocalibration model with the reference PM sensor.

As discussed in Section V-C, RMSE $= 7$ μg/m^3 is taken as the error threshold [31]. It has been observed that the error between the sensed low cost PM data and reference PM data is higher than the threshold. However, after calibration, the error between the calibrated PM data and the reference data reduces and lies below the tolerance threshold.

From the experimental results it has been observed that the OPC N3 PM sensor draws 187 mA of current during turn ON period and 55 mA of current while sensing. The turn on period is 28 sec. The operating voltage of the sensor is 5 V. On the other hand, the proposed on-board PM sensor consumes only 84 mA current during both turn ON and sensing period, and the turn ON period is only 5 sec. The energy efficiency of the low cost PM sensor is calculated as:

$$\text{Energy saved} = \left[\frac{E_h - E_l}{E_h}\right] \times 100\%. \tag{13}$$

In (13), E_h and E_l are the total energy consumed in one hour by the OPC N3 and low cost PM sensors, respectively. As listed in Table V, the cost of on-board PM sensor is 91% less compared with the reference high cost PM sensor. Moreover, the proposed low cost PM sensor saves up to 57% energy compared to the OPC N3 PM sensor, while maintaining sensing error of 0.76 and 1.04 for PM$_{2.5}$ and PM$_{10}$, respectively, after applying the proposed local reference free calibration model.

VI. Concluding Remarks

The proposed local reference free in-field calibration method can be used in a sparsely deployed sensing stations for accurate pollution mapping. The MSNs can be calibrated and recalibrated without colocating any high cost reference sensor. The proposed cross-correlation based method of estimating the optimum recalibration instants performs well in reducing the recalibration overhead significantly. As an implementation-based low cost sensor calibration verification exercise, prototype design of a low cost, low power PM sensor and its implementation has also been presented. The low cost PM sensor is 91% more cost efficient and 57% more energy efficient compared to the OPC N3 PM sensor while maintaining sensing error within RMSE $= 0.76$ for PM$_{2.5}$ and RMSE $= 1.04$ for PM$_{10}$. The GPR based calibration models perform well in calibrating the PM data. Although the sensed PM data over-estimates the actual PM data, the accuracy increases after calibration. Acceptable range of calibration error validates the efficiency of the proposed autocalibration method and the accuracy of the design.

References

[1] P. Das, S. Ghosh, S. Chatterjee, and S. De, "Energy harvesting-enabled 5G advanced air pollution monitoring device," in *proc. IEEE 5GWF, Bangalore, India, India*, 2020, pp. 218–223.

[2] D. Santi, E. Magnani, M. Michelangeli, R. Grassi, B. Vecchi, G. Pedroni, L. Roli, M. C. De Santis, E. Baraldi, M. Setti *et al.*, "Seasonal variation of semen parameters correlates with environmental temperature and air pollution: A big data analysis over 6 years," *Environmental Pollution*, vol. 235, pp. 806–813, 2018.

[3] T. He, S. Krishnamurthy, J. A. Stankovic, T. Abdelzaher, L. Luo, R. Stoleru, T. Yan, L. Gu, J. Hui, and B. Krogh, "Energy-efficient surveillance system using wireless sensor networks," in *proc. of the 2nd international conference on Mobile systems, applications, and services, Boston, MA, USA*, 2004, pp. 270–283.

[4] V. Gupta, S. Tripathi, and S. De, "Green sensing and communication: A step towards sustainable IoT systems," *J. Indian Inst. Sc.*, pp. 1–16, 2020.

[5] H. Harb and A. Makhoul, "Energy-efficient sensor data collection approach for industrial process monitoring," *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 661–672, 2017.

[6] S. De and R. Singhal, "Toward uninterrupted operation of wireless sensor networks," *Computer*, vol. 45, no. 9, pp. 24–30, 2012.

[7] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet Things j.*, vol. 1, no. 1, pp. 22–32, 2014.

[8] M. V. Ramesh and V. P. Rangan, "Data reduction and energy sustenance in multisensor networks for landslide monitoring," *IEEE Sensors J.*, vol. 14, no. 5, pp. 1555–1563, 2014.

[9] W. Qiao, W. Tian, Y. Tian, Q. Yang, Y. Wang, and J. Zhang, "The forecasting of PM2. 5 using a hybrid model based on wavelet transform and an improved deep learning algorithm," *IEEE Access*, vol. 7, pp. 142 814–142 825, 2019.

[10] "Determination of particle size distribution single light interactions methods—part 4: Light scattering airborne particle counter for clean spaces," *ISO 21501-4*, 2018.

[11] S. Komarizadehasl, B. Mobaraki, H. Ma, J.-A. Lozano-Galant, and J. Turmo, "Low-cost sensors accuracy study and enhancement strategy," *Applied Sciences*, vol. 12, no. 6, p. 3186, 2022.

[12] M. A. Zaidan, N. H. Motlagh, P. L. Fung, D. Lu, H. Timonen, J. Kuula, J. V. Niemi, S. Tarkoma, T. Petäjä, M. Kulmala *et al.*, "Intelligent calibration and virtual sensing for integrated low-cost air quality sensors," *IEEE Sensors J.*, vol. 20, no. 22, pp. 13 638–13 652, 2020.

[13] S. Ali, T. Glass, B. Parr, J. Potgieter, and F. Alam, "Low cost sensor with IoT LoRaWAN connectivity and machine learning-based calibration for air pollution monitoring," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2020.

[14] Y. Wang, A. Yang, X. Chen, P. Wang, Y. Wang, and H. Yang, "A deep learning approach for blind drift calibration of sensor networks," *IEEE Sensors J.*, vol. 17, no. 13, pp. 4158–4171, 2017.

[15] Y. Wang, A. Yang, Z. Li, X. Chen, P. Wang, and H. Yang, "Blind drift calibration of sensor networks using sparse bayesian learning," *IEEE Sensors Journal*, vol. 16, no. 16, pp. 6249–6260, 2016.

[16] R. T. Rajan, A. Das, J. Romme, F. Pasveer *et al.*, "Reference-free calibration in sensor networks," *IEEE sensors letters*, vol. 2, no. 3, pp. 1–4, 2018.

[17] L. Zhao, W. Wu, and S. Li, "Design and implementation of an IoT-based indoor air quality detector with multiple communication interfaces," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9621–9632, 2019.

[18] M. Rosmiati, M. F. Rizal, F. Susanti, and G. F. Alfisyahrin, “Air pollution monitoring system using lora module as transceiver system,” *Telkomnika*, vol. 17, no. 2, pp. 586–592, 2019.

[19] M. Rossi and P. Tosato, “Energy neutral design of an iot system for pollution monitoring,” in *2017 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS)*, 2017, pp. 1–6.

[20] P. Zappi, E. Bales, J. H. Park, W. Griswold, and T. Š. Rosing, “The citisense air quality monitoring mobile sensor node,” in *Proc. 11th ACM/IEEE conference on information processing in sensor networks, Beijing, China*, 2012, pp. 16–19.

[21] S. Dhingra, R. B. Madda, A. H. Gandomi, R. Patan, and M. Daneshmand, “Internet of things mobile–air pollution monitoring system (iot-mobair),” *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5577–5584, 2019.

[22] Oizom polludrone. [Online]. Available: https://oizom.com/product/polludrone-air-pollution-monitoring/

[23] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[24] C. M. Bishop, “Pattern recognition,” *Machine learning*, vol. 128, no. 9, 2006.

[25] C. Rasmussen and C. Williams, *Gaussian processes for machine learning*. MIT Press: Cambridge, MA, 2006.

[26] S. Ghosh, S. De, S. Chatterjee, and M. Portmann, “Edge intelligence framework for data-driven dynamic priority sensing and transmission,” *IEEE Transactions on Green Communications and Networking*, pp. 1–1, 2021.

[27] ——, “Learning-based adaptive sensor selection framework for multi-sensing wsn,” *IEEE Sensors J.*, vol. 21, no. 12, pp. 13 551–13 563, 2021.

[28] CPCB. Central Control Room for Air Quality Management - All India. [Online]. Available: https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data

[29] T. Instruments. PM2.5/PM10 particle sensor analog front-end for air quality monitoring design. [Online]. Available: https://www.ti.com/lit/ug/tidub65c/tidub65c.pdf?ts=1627496934960

[30] Alphasense. OPC-N3 particle monitor. [Online]. Available: https://www.alphasense.com/wp-content/uploads/2019/03/OPC-N3.pdf

[31] R. Duvall *et al.* Performance testing protocols, metrics, and target values for fine particulate matter air sensors: Use in ambient, outdoor, fixed sites, non-regulatory supplemental and informational monitoring applications. US EPA Office of Research and Development. [Online]. Available: https://cfpub.epa.gov/si/si_public_record_Report.cfm?dirEntryId=350785&Lab=CEMM

[32] S. Chae, J. Shin, S. Kwon, S. Lee, S. Kang, and D. Lee, “PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network,” *Scientific Reports*, vol. 11, no. 1, pp. 1–9, 2021.

[33] V. Gupta and S. De, “Collaborative multi-sensing in energy harvesting wireless sensor networks,” *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 426–441, 2020.