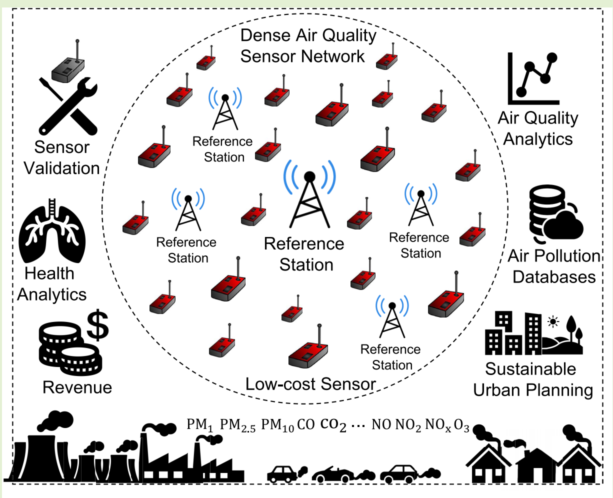
Dense Air Quality Sensor Networks: Validation, Analysis, and Benefits

Martha Arbayani Zaidan, *Member, IEEE*, Yuning Xie, Naser Hossein Motlagh, Bo Wang, Wei Nie, Petteri Nurmi, Sasu Tarkoma, *Senior Member, IEEE*, Tuukka Petäjä, Aijun Ding, and Markku Kulmala

***Abstract*—Air pollution is known to be harmful to human health and the environment. Official air quality monitoring stations have been established across many smart cities around the world. Unfortunately, these monitoring stations are sparsely located and consequently do not provide high-resolution spatio-temporal air quality information. This article demonstrates how a dense sensor network deployment offers significant advantages in providing better and more detailed air quality information. We use data from a dense sensor network consisting of 126 low-cost sensors (LCSs) deployed in a highly populated district in Nanjing, China. Using data obtained from 13 existing reference stations installed in the same district, we propose three LCS validation methods to evaluate the performance of LCSs in the network. The methods assess the reliability, accuracy of tests, and failure and anomaly detection performance. We also demonstrate how the reliable data generated from the sensor network provides deep insights into air pollution information at a higher spatio-temporal resolution. We further discuss potential improvements and applications derived from the dense deployment of LCSs in cities.**



***Index Terms*—Air quality, anomaly detection, low-cost sensors (LCSs), reference stations, sensor network, sensor validation.**

I. Introduction

According to World Health Organization (WHO), air pollution causes approximately 7 million deaths each year. Of this, an estimated 4.2 million deaths are due to outdoor exposure [1]. Air pollution is one of the leading causes of adverse human health effects such as cardiovascular and respiratory illnesses. Exposure to air pollution also has other negative effects, for example, it has been shown to degrade well-being and productivity [2] and it is believed to be linked with an increased risk of COVID-19 infection [3]. Beyond societal effects, air pollution has immense economic consequences as it is associated with increased expenditure in healthcare, including costs of treatment, diagnosis, and medical insurance [4].

Mitigating the adverse effects of pollution requires a detailed understanding of the sources, causes, and consequences of pollutants, which in turn requires comprehensive information about the concentration, distribution, and characteristics of air pollutants. Conventionally, official air quality monitoring stations are installed in cities to study the characteristics of pollutants in urban environments [5]. Unfortunately, due to the high costs of the instruments, their operations, and maintenance, the number of installed official air quality monitoring stations in cities is limited. Thanks to advances in communication and networking technologies, and the Internet-of-Things (IoT) low-cost sensors (LCSs) have emerged as an alternative that can be deployed on a massive scale in cities. This deployment facilitates obtaining hyperlocal air pollution information that can vary by more than eight times within 200 m [6] and offers high resolution of spatio-temporal air quality information [7]. In a massive deployment approach, LCSs can be installed in strategic locations in urban areas or the air quality data can be gathered through crowd-sourced-based sensing [8]. Currently, the

Manuscript received 27 May 2022; accepted 7 October 2022. Date of publication 26 October 2022; date of current version 30 November 2022. This work was supported in part by the MegaSense and Nokia Center for Advanced Research (NCAR); in part by the Atmosphere and Climate Competence Center (ACCC) under Grant 337549; in part by the EMME-CARE Project through the European Union's Horizon 2020 Research and Innovation Programme under Grant 856612; in part by the Academy of Finland Projects under Grant 324576, Grant 345008, Grant 335934, and Grant 339614; in part by the Technology Industries of Finland Centennial Foundation to Urban Air Quality 2.0 Project; and in part by the Helsinki Institute for Information Technology (HIIT). The associate editor coordinating the review of this article and approving it for publication was Prof. Aime Lay-Ekuakille. *(Corresponding author: Martha Arbayani Zaidan.)*

Please see the Acknowledgment section of this article for the author affiliations.

Digital Object Identifier 10.1109/JSEN.2022.3216071

benefits massive-scale LCSs deployments can provide are not sufficiently understood. Indeed, while the benefits of LCSs have been demonstrated in several studies, the existing works have mostly used a limited number of sensors and extrapolated the findings for larger sensor network deployments [9]. The better understand the practical benefits and the usefulness of massive-scale LCS deployments, there thus is a need to study denser deployments, analyzing their practicality and benefits, and the details they can provide. In massive deployments, while sensors operate, they may encounter various challenging issues. For example, sensors may malfunction, degrade over time, and output inaccurate readings [10].

Analyzing the performance of large-scale deployments, however, also poses its own challenges. In practice, LCSs are not placed side by side with the reference instruments, and therefore direct validation of LCSs against the reference instruments is not feasible. The validation methods should evaluate various states of sensor functions, such as sensors' reliability in a sensor network, and individual sensor validations, including sensing accuracy, sensors' failures, and sensors' anomalies. Sensor reliability indicates the general performance of the sensors deployed in a network in providing reliable air quality information on a larger scale (e.g., city district). Accuracy indicates how well are the measurements of an LCS in agreement with the measurement of a reference instrument. Failure indicates to the LCSs when they stop transmitting the data to the edge servers, whereas anomaly refers to the LCSs which generate anomalous data patterns in comparison to the measurements of reference instruments (i.e., drift).

This article demonstrates the benefits of dense LCS deployments by performing a comprehensive sensors validation and data analysis for a dense air quality sensor network. We collect extensive air quality measurement datasets generated from 13 reference stations and 126 units of LCSs which are deployed in Nanjing, China. Ours is among the densest deployments ever to be analyzed. To account for the challenges in validating the benefits of deployments, we propose three methods of sensor validations including: 1) reliability investigation (by means of statistical properties and correlation coefficients) to evaluate all LCSs to observe if they provide reliable measurements as a whole; 2) accuracy tests on few of LCSs which are nearest to the reference stations; and 3) failure and anomaly detection on individual LCSs to evaluate if they generate reliable air quality data. The validation results demonstrate that the sensor network is reliable as a whole and the accuracy tests for pollutant variables PM$_{10}$, PM$_{2.5}$, and O$_3$ explain that the measurements of LCSs are reliable. We further highlight the advantages of having such a dense sensor network in cities and discuss the potential improvements and applications offered by real massive sensor deployments. Among others, our results show that dense deployments can facilitate detecting localized pollution sources or hotspots and capture diurnal variations in pollutant concentrations at different locations within the city.

II. Materials and Methods

This section describes the materials and methods used in this study. The materials provide information about the measurement site, air quality sensors, and air pollution data. The methods describe sensor clustering, air pollution models, and the metrics used to evaluate the performance of sensors.

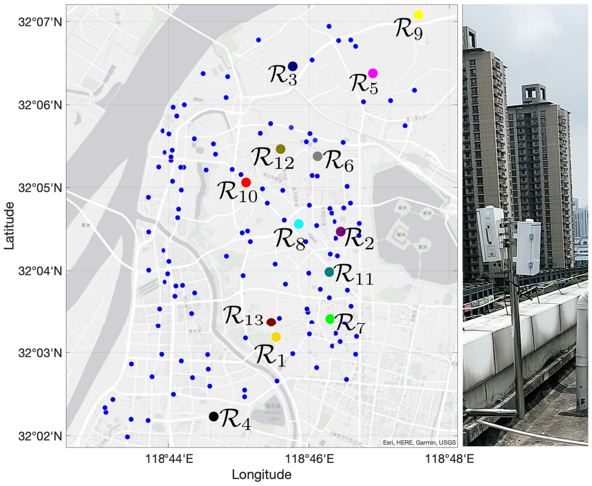


Fig. 1. Sensors network map (left) and an LCS installed on the roof of a building (right), in Gulou, Nanjing. $\boldsymbol{\mathcal{R}}_1$–$\boldsymbol{\mathcal{R}}_{13}$ indicates the 13 reference stations.

A. Measurement Site and Air Quality Monitoring

The measurement site where the sensors (i.e., reference stations and LCSs) are deployed is located at the coordinates of $32^{\circ}3'50''$N $118^{\circ}45'5''$E in Gulou, Nanjing, China (as shown in Fig. 1). The district size is 54.18 km^2 with the population estimated at 1 109 600 in 2018. Given the population density, obtaining reliable and high-resolution air quality information would be beneficial for health exposure analysis as well as air pollution risk and mitigation.

We use air quality data from two types of sensors including: 1) reference stations and 2) a dense LCS network. Fig. 1 depicts the sensor deployment, where the big colored bubbles shown by $\mathcal{R}$ present the reference stations, and the small blue bubbles present the LCSs deployed in the network. This dense network consists of 13 reference stations ($\mathcal{R}$) and 126 units of LCSs. The picture on the right side of the figure also shows an example of an LCS that is installed on the roof of a building. The reference stations and LCSs are installed at different strategic locations in Gulou, Nanjing. These locations cover six different types of environments where we label them with letters **A**–**F**, as presented in Table I.

The technology of LCSs is based on YSRDAQ-07 sensors (Insights Value Technology Company Ltd.) [11]. The LCSs are capable of measuring particulate matter (PM) including PM$_{10}$ (PM with an aerodynamic diameter of 10 μm or less) and PM$_{2.5}$ (PM with an aerodynamic diameter of 2.5 μm or less), ozone (O$_3$), nitrogen dioxide (NO$_2$), carbon monoxide (CO), and sulfur dioxide (SO$_2$). In addition, the reference stations (labeled by $\mathcal{R}$) are operated by the Chinese national and standard monitoring stations. These stations measure the concentrations of air pollutants such as PM$_{10}$, PM$_{2.5}$, CO, NO$_2$, O$_3$, and SO$_2$ [12]. The air quality data from these reference stations were downloaded from the website of the Chinese Environmental Protection Bureau.[1] For our analysis

[1] http://www.cnemc.cn/

TABLE I
ENVIRONMENTAL TYPES WHERE LCSs ARE INSTALLED

Label	Environment
A	The roadsides refer to the monitoring points placed on the side of the city's main roads.
B	The evaluation points monitor air quality for each street in the Gulou district.
C	The construction sites monitor the points located near urban construction areas).
D	The transmission points monitor the points located in the border area of the Gulou District.
E	The sub-station surroundings (Parallel) monitor the points arranged around the Shanxi road national control station.
F	The quality control points are used as a control point, close to the national control station of Shanxi road.

in this article, we gather hourly air quality datasets from the sensor's measurements from March 1 to July 31, 2021.

In addition, we use this data to derive air quality index (AQI) data which quantifies overall air quality based on all ambient air pollutants in the monitored area. The two main objectives of AQI are: 1) to inform and alert the public about the risk of exposure to air pollution levels and 2) to enforce required regulatory measures to mitigate the impacts [13]. Indeed, AQI is defined as the maximum of the indexes for six criterion pollutants, including PM_{10}, $PM_{2.5}$, CO, NO_2, O_3, and SO_2 [14] which can be formulated as

$$\mathrm{AQI} = \max\{\mathrm{IAQI}_1,\ \ \mathrm{IAQI}_2,\ \mathrm{IAQI}_3, \ldots,\ \ \mathrm{IAQI}_n\} \tag{1}$$

where IAQI stands for an individual AQI and n is the number of ambient air pollutants.

B. Sensor Clustering

In order to perform LCS validation, the LCS measurement data should be compared to the ground-truth data generated from the nearest reference stations. In our study, there are 13 reference stations ($\mathcal{R}$) and 126 LCSs deployed in Gulou, Nanjing, China. Since there are a considerable number of $\mathcal{R}$s in a densely deployed LCSs network, we use this opportunity to form sensor clusters and validate LCSs against their respective reference stations. Therefore, we form the sensor clusters based on the nearest distances between LCSs and the reference stations. To do this, we use the Haversine equation [15] that calculates the shortest distance between two points over the Earth's surface. The Haversine equation is given by

$$\mathrm{Dist} = 2\,\mathrm{R_E} \arctan\left(\sqrt{a}, \sqrt{(1-a)}\right) \tag{2}$$

where a is calculated by

$$a = \sin^2(\Delta\phi/2) + \cos\phi_1\ \cos\phi_2\ \sin^2(\Delta\lambda/2) \tag{3}$$

where R_E is the Earth's radius (i.e., mean radius = 6 371 km), and ϕ and λ are the latitude and the longitude, respectively. Applying the Haversine equation on the coordinates of the reference stations and LCSs, the distances between each LCS ($\mathcal{L}$) to every $\mathcal{R}$ can be calculated. As the $\mathcal{R}$ coordinates are considered cluster centers, thus, each $\mathcal{L}$ can then be assigned to the nearest cluster center.

Table II presents 13 clusters (shown by $\mathcal{R}_1$–$\mathcal{R}_{13}$) and their respective LCSs $\mathcal{L}$s in those clusters. In the table, the labels (i.e., numbering) of LCSs $\mathcal{L}$ are sorted in ascending order based on their distances to their respective $\mathcal{R}$. Note that the clustering based on the Haversine equation does not list any LCSs ($\mathcal{L}$) in the cluster $\mathcal{R}_9$. The reason is the long distances between the LCSs to $\mathcal{R}_9$ compared to $\mathcal{R}_5$s. This is clearly shown in Fig. 1, by $\mathcal{R}_9$ and $\mathcal{R}_5$, presented with colors yellow and magenta, respectively.

TABLE II
LCSs ($\mathcal{L}$) GROUPED WITHIN CLUSTERS $\mathcal{R}_1$–$\mathcal{R}_{13}$

Cluster $\mathcal{R}$	LCSs ($\mathcal{L}$) with numbering (l)
$\mathcal{R}_1$	3, 31, 39, 41
$\mathcal{R}_2$	81, 21, 80, 86, 79, 85, 30, 84
$\mathcal{R}_3$	33, 14, 27, 13, 71, 32
$\mathcal{R}_4$	118, 120, 123, 122, 126, 92, 36, 1, 35, 43, 107, 76, 89, 8, 106, 52, 113, 45, 40
$\mathcal{R}_5$	16, 15, 37, 9, 48, 57, 38
$\mathcal{R}_6$	119, 56, 2, 53, 121, 90, 125, 87
$\mathcal{R}_7$	19, 99, 69, 100, 51, 70, 61, 60, 74, 54, 98, 29, 49
$\mathcal{R}_8$	64, 82, 63, 83, 26, 6, 50
$\mathcal{R}_9$	–
$\mathcal{R}_{10}$	18, 72, 67, 65, 68, 73, 55, 20, 11, 95, 93, 104, 105, 47, 94, 59, 110, 10, 62, 77, 12, 22, 42, 66, 24, 78, 23, 34, 28, 25, 102, 103, 75, 114
$\mathcal{R}_{11}$	4, 88, 58, 101, 5
$\mathcal{R}_{12}$	116, 17, 124
$\mathcal{R}_{13}$	117, 44, 115, 96, 112, 109, 97, 91, 46, 7, 111, 108

TABLE III
AIC EVALUATED ON FOUR PROBABILITY DISTRIBUTIONS FOR MODELING DIFFERENT AIR POLLUTANTS

Distribution \ **Pollutant**	PM_{10}	$PM_{2.5}$	O_3
Gamma	460,720	387,288	476,220
Log-normal	460,798	389,338	486,897
Rayleigh	481,892	389,644	487,217
Weibull	**457,509**	**385,819**	**475,321**

C. Air Pollutant Model: Weilbull Distribution

Once the clusters have been formed, LCS anomaly detection can be applied by evaluating the LCSs measurements if they lie at the outliers' regime of pollutants' distributions at the respective reference stations. The current scientific understanding suggests that air pollutant concentrations follow a right-skewed distribution, and we consider the most commonly considered families of distributions: Gamma, Log-normal, Rayleigh, and Weibull distributions [16], [17], [18]. We determine the best fitting distribution using the Akaike information criterion (AIC) that is a widely used method for evaluating the fit of different distributions [19], including for air pollutant concentrations [20], [21], [22]. The smaller the AIC value, the better the fit of the distribution. Table III presents the AIC values of the four probability distributions when fit on the data of PM_{10}, $PM_{2.5}$, and O_3 gathered from the 13 reference stations. The best fit in all cases is the Weibull distribution and hence we use the Weibull distribution for modeling air pollutant concentrations at the reference stations ($\mathcal{R}$).

The Weibull probability density function is mathematically defined as

$$f(X;\lambda,k) = \begin{cases} \dfrac{k}{\lambda}\left(\dfrac{X}{\lambda}\right)^{k-1} \exp^{-(X/\lambda)^k}, & X \geq 0 \\ 0, & X\ < 0 \end{cases} \tag{4}$$

where X represents the air pollutants measurement data. The parameter $k > 0$ is the shape parameter and the parameter

$\lambda > 0$ is the scale parameter of the distribution. These parameters can be estimated using the maximum likelihood method. In addition, the quantile (inverse cumulative distribution) function for the Weibull distribution is

$$Q(p;\lambda,k) = \lambda(-\ln(1-p))^{1/k} \tag{5}$$

for $0 \leq p > 1$.

D. Sensor Performance Metrics

We use three metrics to evaluate the sensor performances including biweight midcorrelation, mean absolute error (MAE), and Spearman correlation. While the first two metrics are used to determine accuracy tests (Section III-B), the third metric is used to evaluate the correlation between pollutant variables for reliability investigation (Section III-A). For two vectors of measurement data, X and X', where $X = \{x_1, x_2, \ldots, x_N\}$ and $X' = \{x'_1, x'_2, \ldots, x'_N\}$, the performance metrics can be calculated as follows.

1) Biweight Midcorrelation (R_b): This is a similarity metric that is an alternative to Pearson correlation because it is more robust to outliers. The biweight midcorrelation between X and X', that is, $R_b(X, X')$ can be computed by

$$R_b(X,X') = \frac{\sum_{i=1}^{N}(x_i - \mathrm{med}(x))\, w_i^{(x)}\, \left(x'_i - \mathrm{med}(x')\right) w_i^{(x')}}{\sqrt{\sum_{i=1}^{N}\left[(x_i - \mathrm{med}(x)) w_i^{(x)}\right]^2 \sum_{i=1}^{N}\left[(x'_i - \mathrm{med}(x')) w_i^{(x')}\right]^2}} \tag{6}$$

where weights $w_i^{(x)}$ and $w_i^{(x')}$ are defined as

$$w_i^{(x)} = \left(1 - u_i^2\right)^2 I\left(1 - |u_i|\right) \tag{7}$$

$$w_i^{(x')} = \left(1 - v_i^2\right)^2 I\left(1 - |v_i|\right) \tag{8}$$

where the notation I is the identity function, defined as

$$I(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

and the notations u_i and v_i are defined as

$$u_i = \frac{x_i - \mathrm{med}(x)}{9\,\mathrm{mad}(x)} \tag{10}$$

$$v_i = \frac{x'_i - \mathrm{med}(x')}{9\,\mathrm{mad}(x')}. \tag{11}$$

The notation $\mathrm{med}(x)$ is the median of a vector x, whereas the notation $\mathrm{mad}(x)$ is the median absolute deviation (MAD).

2) Mean Absolute Error: This metric is a measure of errors between paired observations expressing the same measurements. MAE can be calculated by

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{x}_i - x_i|. \tag{12}$$

The notations x and $\hat{x}$ are the data points obtained from the reference instruments and the LCSs, respectively.

3) Spearman's Rank Correlation Coefficient Analysis (R_s): This metric evaluates how well the relationship between two measured variables can be described using a monotonic function. The formula is given by

$$R_s = 1 - \frac{6\sum d_i^2}{N(N^2-1)} \tag{13}$$

where N is the number of measurement data points and $d_i = \mathrm{rank}(X) - \mathrm{rank}(X')$ is the difference between two ranks of each data (i.e., X and X').

III. Sensor Validations

In practice, LCSs that are deployed in the field undergo a laboratory calibration and some might even undergo field calibration while they operate. However, while the LCSs operate in the field, they may malfunction, witness drift, or degrade over time. Thus, it is important to validate their sensing accuracy while they are deployed and operating in the field [23], [24]. We next present sensor validation methods applied to LCSs when they operate as part of a network. First, we perform a reliability investigation to evaluate all of the LCSs in a network to observe if they provide reliable measurements as a whole compared to the measurements of all of the reference stations. Second, we perform accuracy tests on a few of the LCSs which are nearest to the reference stations. These accuracy tests are then generalized to the remaining LCSs in the sensor network as the LCSs are based on the same sensing technology. Third, we perform failure and anomaly detection on the individual sensors to evaluate which sensor functions properly and generates reliable air quality data.

A. Reliability Investigation

The first step for validating sensors is to study the reliability of their measurements. Hence, we compare the overall measurements of LCSs and the reference stations by means of their statistical properties and their correlation analysis between pollutants.

1) Statistical Properties of Pollutant Variables: We first summarize key characteristics of the measurement data generated at the reference stations ($\mathcal{R}$) and the LCSs ($\mathcal{L}$). The median values of $\mathcal{R}$ and $\mathcal{L}$ are closely aligned, and also the MAD values—a common measure of precision in air quality data [25]—are consistent, suggesting that the two types of sensors produce very similar measurements. In contrast, when we compare the mean and standard deviation of the measurements, more variation is observed. This suggests that the measurements also contain some outliers or abnormal events, even if on a whole the measurements are similar. The existence of outliers is also supported by the skewness of the data, which shows the distributions to be consistently right-skewed but that the LCSs tend to have higher tails—a common indicator of outliers. As can be expected, the consistency of the high-quality instruments ($\mathcal{R}$) is higher than that of the LCSs ($\mathcal{L}$) which can be observed from the lower standard deviation.

We also analyze the uncertainty of the measurements by considering the standard error of the mean (CI in the table).

TABLE IV
STATISTICAL PROPERTIES OF POLLUTANT VARIABLES

Pollutant	$\mathcal{R}$					$\mathcal{L}$				
Variables	Median	Mean ± CI	Std	MAD	Skewness	Median	Mean ± CI	Std	MAD	Skewness
AQI	51	54.04 ± 4.46	33.70	16.96	4.46	48.89	52.60 ± 19.04	33.54	20.00	4.20
PM_{10} [μg/m^3]	48	33.03 ± 5.23	48.53	19.97	4.51	46	55.18 ± 21.11	79.87	21.17	72.83
$PM_{2.5}$ [μg/m^3]	25	19.19 ± 9.95	112.52	11.18	85.30	21.65	26.65 ± 19.10	83.55	11.15	76.90
CO [ppb]	0.81	0.85 ± 0.30	0.38	0.19	9.87	0.4	1.41 ± 18.64	5.88	1.00	7.45
NO_2 [ppb]	24	29.26 ± 9.23	19.48	12.26	1.57	20	22.55 ± 8.85	13.82	9.25	1.5
O_3 [ppb]	56	68.09 ± 13.82	49.16	32.08	1.09	65.21	78.00 ± 30.42	54.37	35.00	2.98
SO_2 [ppb]	8	8.56 ± 1.54	6.79	1.56	43.66	4	4.92 ± 1.48	6.25	1.91	20.98

As can be expected, the uncertainty of the LCSs $\mathcal{L}$ is higher than with the reference stations $\mathcal{R}$. There are two reasons for this. First, naturally, a part of the differences is explained by the reference stations using components with higher precision. By comparing the MAD in Table IV, we can see that this effect is relatively small overall and explains only partially the higher spread of the measurements. The most important factor stems from the lower-cost sensors covering a broader geographic area. Air quality measurements tend to have significant variation even within a small geographic distance and the higher number of LCSs results in a larger spread in the measurements. For example, as seen in Fig. 1, the west part of the Gulou has a sparse concentration of reference stations $\mathcal{R}$, whereas the number of LCSs is abundant. Thus, the low-cost measurements contain pollutant concentrations from areas that are not covered by the reference stations which results in higher variation and uncertainty in the overall data.

In terms of individual pollutants, the largest variation occurs for CO and SO_2. As with the other variables, the precision (i.e., MAD) for these pollutants is consistent and thus the main reason for the differences is the variation in spatial coverage. The best indication of consistency, however, is the AQI column. The AQI combines all six pollutants into a single value and is sensitive to fluctuations in any single pollutant. As the values of the AQI are consistent—as indicated by the similar mean, median, MAD, and skewness—this indicates that both $\mathcal{R}$ and $\mathcal{L}$ generate similar readings for all pollutants. The uncertainty in the low-cost measurements $\mathcal{L}$ is higher, which again mostly reflects the difference in geographic coverage of the sensor instruments. Taken together, the results suggest that the deployed LCS network provides reliable data that is similar to the measurements provided by the reference stations and complements the air quality information that can be acquired from the city of Nanjing.

2) Correlation Analysis Between Pollutant Variables: The correlation analysis between pollutant variables demonstrates the relationship between the pollutant variables measured at $\mathcal{R}$ and $\mathcal{L}$. Fig. 2 shows the heatmap (matrix plot) of absolute number of Spearman correlation coefficients (R_s) for different pollutants measured by $\mathcal{R}$ and $\mathcal{L}$. The cells on the right side of the diagonal (i.e., yellow cells) show the correlation coefficients between pollutants measured at $\mathcal{R}$. The cells on the left side of the diagonal show the correlation coefficients between pollutants measured at $\mathcal{L}$. It can be seen that both sides of the diagonal present almost similar coefficient values. For example, the correlation between AQI and PM_{10} present a high value (shown by dark blue) in $\mathcal{R}$ ($R_s \approx 0.94$) and in $\mathcal{L}$ ($R_s \approx 0.75$). Likewise, the correlation between $PM_{2.5}$ and PM_{10} are also high (dark blue) in $\mathcal{R}$ ($R_s \approx 0.84$) and in $\mathcal{L}$ ($R_s \approx 0.79$). Another example is the variable NO_2 which is correlated well with the $PM_{2.5}$ by presenting $R_s \approx 0.41$ (shown by light blue) for $\mathcal{R}$ and $R_s \approx 0.49$ for $\mathcal{L}$. Indeed, the comparison between correlation coefficients is not always similar between $\mathcal{R}$ and $\mathcal{L}$ as some sensors might contain missing data. Nevertheless, the relationship similarity (i.e., correlation coefficients) between the majority of the pollutant variables measured by $\mathcal{R}$ and $\mathcal{L}$ are similar indicating the reliability of the sensor network.

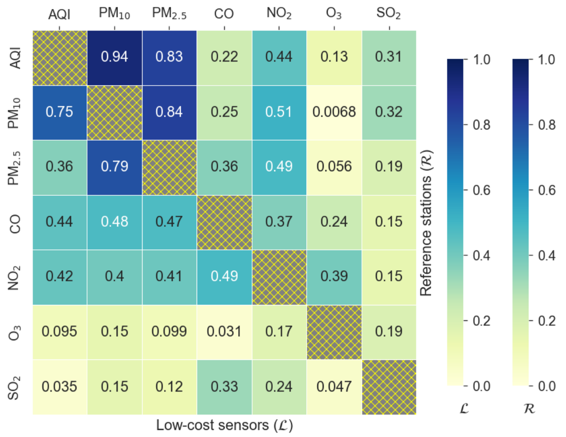


Fig. 2. Correlation coefficients of air pollutants measured via $\mathcal{R}$ (upper triangle) and $\mathcal{L}$ (lower triangle).

B. Accuracy Tests

After the field deployment of the LCSs ($\mathcal{L}$) and while they operate, they usually are not placed side by side at reference stations ($\mathcal{R}$). Therefore, one method to evaluate the performance accuracy of $\mathcal{L}$ is by performing accuracy tests on some nearest sensors to $\mathcal{R}$. Note that even this approach is approximate as in practice it is difficult to ensure that the air intake of the LCS is perfectly aligned with that of the reference station. Hence, comparing individual measurements is not meaningful, and instead aggregates calculated over a longer time interval should be used instead. In this article, accuracy tests provide an indication of how similar the measurements of $\mathcal{L}$ and $\mathcal{R}$ [26] are. Since the LCSs ($\mathcal{L}$) used in our study are identical (i.e., they have the same hardware and software), the accuracy results largely transfer to the other sensors.

Table V presents the summary of accuracy tests for $\mathcal{L}$ (e.g., $\mathcal{L}_3$ and $\mathcal{L}_{21}$) which have the nearest distances (*Dist* in *km*) to

TABLE V

SUMMARY OF ACCURACY TESTS FOR THE NEAREST LCSS ($\mathcal{L}$) TO THE REFERENCE STATIONS ($\mathcal{R}$)

$\mathcal{R}$	$\mathcal{L}$	*Env*	*Dist*	Biweight Midcorrelation (R_b)							Mean Absolute Error (MAE)						
				AQI	PM_{10}	$PM_{2.5}$	CO	NO_2	O_3	SO_2	AQI	PM_{10}	$PM_{2.5}$	CO	NO_2	O_3	SO_2
$\mathcal{R}_1$	$\mathcal{L}_3$	B	0.53	0.63	0.86	0.77	–	–	0.83	–	13.13	11.58	7.1	–	–	19.34	–
$\mathcal{R}_2$	$\mathcal{L}_{21}$	F	0.27	0.61	0.83	0.82	–	–	0.91	–	17.24	14.4	24.75	–	–	15.76	–
$\mathcal{R}_3$	$\mathcal{L}_{33}$	C	0.45	0.71	0.8	–	–	–	–	–	17.18	13.73	–	–	–	–	–
$\mathcal{R}_4$	$\mathcal{L}_{118}$	B	0.68	–	–	0.62	–	0.61	0.58	–	–	–	6.61	–	6.13	21.27	–
$\mathcal{R}_5$	$\mathcal{L}_{16}$	B	0.67	0.7	0.81	0.71	–	0.35	0.79	–	15.62	14.14	7.85	–	11.48	22.95	–
$\mathcal{R}_6$	$\mathcal{L}_{56}$	C	0.41	–	0.7	0.66	–	–	–	–	19.46	11.31	8.56	–	–	–	–
$\mathcal{R}_7$	$\mathcal{L}_{19}$	B	0.39	0.72	0.73	0.55	–	0.55	0.83	–	14.13	16.81	10.17	–	8.37	20.01	–
$\mathcal{R}_8$	$\mathcal{L}_{64}$	C	0.34	0.71	0.79	–	–	–	–	–	15.48	13.95	–	–	–	–	–
$\mathcal{R}_9$	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
$\mathcal{R}_{10}$	$\mathcal{L}_{18}$	B	0.21	0.72	0.84	0.79	–	–	0.79	–	16.68	36.54	7.54	–	–	24.08	–
$\mathcal{R}_{11}$	$\mathcal{L}_4$	B	0.40	0.7	0.82	0.75	–	–	0.71	–	14.59	14.16	7.60	–	–	25.19	–
$\mathcal{R}_{12}$	$\mathcal{L}_{17}$	B	0.58	0.7	0.81	0.68	–	–	0.75	–	15.48	14.76	7.81	–	–	23.93	–
$\mathcal{R}_{13}$	$\mathcal{L}_{117}$	A	0.22	–	0.85	0.65	–	0.5	0.62	–	–	5.99	6.35	–	6.87	17.82	–
Mean of metrics				0.68	0.80	0.70	–	0.51	0.79	–	15.90	15.21	9.43	–	8.21	21.15	–

$\mathcal{R}$ (i.e., $\mathcal{R}_1 - \mathcal{R}_{13}$). The table also shows the environment (*Env*) that refers to the different strategic locations [e.g., roadsides (**A**) and construction sites (**C**)] in Gulou, Nanjing. As described in Section II-D, we use performance metrics of biweight midcorrelation (R_b) and MAE for evaluating the nearest LCSs $\mathcal{L}$ and their respective $\mathcal{R}$ for the pollutant variables (e.g., AQI, PM_{10}, and others.)

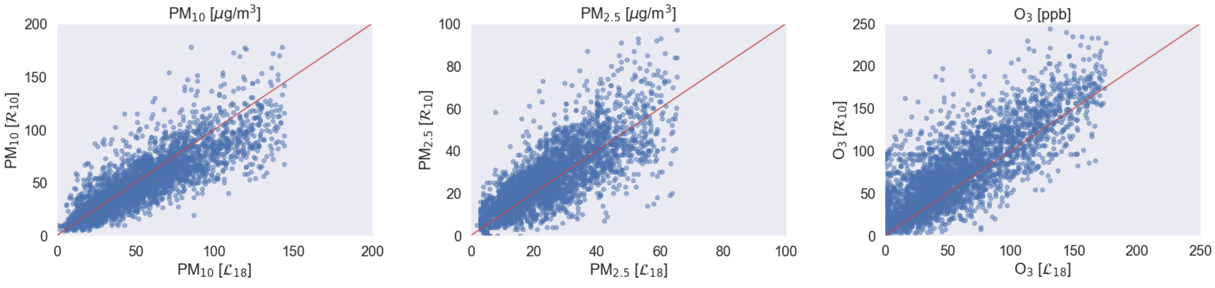
In Table V, all $\mathcal{R}$s present closest distances with some $\mathcal{L}$s (e.g., $\mathcal{R}_1$ has the nearest distance to $\mathcal{L}_3$), except $\mathcal{R}_9$ as there are no $\mathcal{L}$ installed nearby this reference station. In addition, if there is a nearest $\mathcal{L}$ to $\mathcal{R}$ but it does not generate enough data, in this case, we perform accuracy tests on the next nearest $\mathcal{L}$ to $\mathcal{R}$. For measurements of PM_{10}, the values of R_b range between 0.7 and 0.86. These results explain that PM_{10} sensors function well without drifts. Similarly, the results by the metric MAE approve the measurement performance of PM_{10} sensors by presenting values in the range between 5 and 16 μg/m^3, except for $\mathcal{L}_{18}$. The sensor $\mathcal{L}_{18}$ might measure high pollutant concentrations emitted from dust particles, which are often found in the evaluation points (i.e., *Env* **B**). In addition, $\mathcal{L}_{118}$ located near to $\mathcal{R}_4$ do not present any R_b values because data in $\mathcal{L}_{118}$ are missing for PM_{10} measurements, indicating this sensor is being faulty. The mean of the MAE values for PM_{10} is 15.21 μg/m^3 which is within the precision (MAD) of the sensor, as shown in Table IV and indicates a reasonable accuracy. As PM_{10} particles are larger in size, they tend to fall to the ground faster and thus the pollutant concentrations are more heavily localized than for other particles. The main bulk of PM_{10} results from street dust and traffic. LCSs capture this more effectively than reference stations as they sample the air directly within the streets and other urban structures instead of requiring specialized instruments that are located close to the pollution source.

In all LCSs ($\mathcal{L}$) for the measurements of $PM_{2.5}$, the values of R_b range between 0.55 and 0.82, while MAE values range between 6 and 11 μg/m^3. The mean of MAE values for $PM_{2.5}$ is 9.43 μg/m^3 which again is within the precision of the $PM_{2.5}$ measurements (column MAD in Table IV). These results confirm that $PM_{2.5}$ measurements for all $\mathcal{L}$ provide similar readings to $\mathcal{R}$, indicating that $PM_{2.5}$ sensors are accurate. However, $PM_{2.5}$ sensors for $\mathcal{L}_{33}$ and $\mathcal{L}_{64}$ do not exhibit any values indicating that they might be in a failure state as no data is transmitted.

In addition, there are no values for R_b and MAE for the gas sensors of CO and SO_2, as there are no sufficient data collected by the LCSs ($\mathcal{L}$) listed in the table. In contrast, O_3 sensors (for those $\mathcal{L}$ that provide data) demonstrate promising performance with R_b values range between 0.58 and 0.88, and MAE values range between 17 and 23 ppb. For the LCSs ($\mathcal{L}$) that provide NO_2 data, the results for R_b values range between 0.5 and 0.61, except for the sensor near $\mathcal{R}_5$ with $R_b \approx 0.35$ ($\mathcal{L}_{16}$). Their MAE values also range between 6.13 and 11.48 ppb. These results show good performance by metrics R_b and MAE, except $\mathcal{L}_{16}$ that might measure different NO_2 concentration as the distance to $\mathcal{R}_5$ is about 0.67 km. In NO_2 measurements, however, there are more than half of LCSs ($\mathcal{L}$) that do not exhibit any values for R_b and MAE, indicating that those $\mathcal{L}$ might be in a faulty state that does not generate NO_2 data. As with the other pollutants, the mean of MAE values for both O_3 and NO_2 sensors is within the precision of the sensors.

In our study, in order to further analyze the results presented in Table V, we perform visualization of scatter plot between $\mathcal{L}_{18}$ and $\mathcal{R}_{10}$. We select these two sensing units as they have the shortest distance (i.e., 0.21 km) to each other among all of the sensor units in the network. These scatter plots for pollutant variables PM_{10}, $PM_{2.5}$, and O_3 are depicted in Fig. 3. In the plots, the x-axis is the LCS measurements, whereas the y-axis presents the measurements at reference stations. The red line is the reference line. The results in the figure confirm that $\mathcal{L}_{18}$ has accurate measurement because the majority of data points from LCS measurements for the three pollutant variables lie around the reference line.

The discrepancies between LCSs and reference instruments in the scatter plots (as shown in Fig. 3) result from a combined effect of multiple factors. These include meteorological factors, such as wind speed and direction, anthropogenic factors, such as vehicles' emissions, and spatial factors related to the deployment locations. The meteorological effects also are observable from dispersion, for example, when wind speed is low, the air pollutants remain close to the emission source. In contrast, when the wind speed is high, the air pollutants



(a) (b) (c)

Fig. 3. Scatter plots between the reference station 10 ($\mathcal{R}_{10}$) to the nearest LCS ($\mathcal{L}_{18}$). (a) PM_{10}. (b) $PM_{2.5}$. (c) O_3.

disperse into the environment [27]. The measurement differences between LCSs against the reference instruments also depend on wind direction [28]. As an example, from Fig. 3, we see that when the wind is in the direction of $\mathcal{R}_{10}$ and $\mathcal{L}_{18}$, the pollutant concentrations tend to be closer than when the wind direction is reversed. Similarly, sensor locations affect the measurements, especially when combined with meteorological factors. For example, exhausts for vehicles in close proximity to an LCS are observable by the LCS but are not necessarily observable at the reference instrument, particularly if the wind direction is away from the reference station. Reference stations often also sample the air at a higher altitude than the LCS and this also factors into the overall dispersion. Isolating these effects in a real-world deployment, such as ours, is difficult, and thus there necessarily is more variation in the measurements as a result of such factors. The main motivation for deploying LCS is to capture this variation at a finer resolution and to help explain and understand it. Note that wind speed does not impact the quality of the measurements as both the reference instrument and the LCSs regulate the air volume that is used for measurements [29] and hence the only effects resulting from wind speed and wind direction are due to differences in the pollutant concentrations and dispersion patterns. We will further discuss this concern later in Section IV-A.

In conclusion, our accuracy tests confirm that the nearest LCSs which provide data of PM_{10}, $PM_{2.5}$, and O_3 are correlated with their respective reference stations. These correlations state that those LCSs function properly without drifts and provide accurate measurements. This type of accuracy testing is considered to be effective as the LCSs ($\mathcal{L}$) are identical, and as they are deployed in one city district which presents a unique air quality profile. Therefore, other deployed LCSs within the same city district which are located far from reference stations $\mathcal{R}$ can also be considered to be accurate.

C. Sensor Failure and Anomaly Detections

In addition to the reliability investigation and accuracy tests, in our study, we perform sensor failure and anomaly detections for all LCSs ($\mathcal{L}$) in the sensor network.

1) Sensor Failure: Sensor failure can be detected when the sensors stop transmitting the data to the server (i.e., a computer edge server or a cloud server, where all sensor data are collected, processed, and analyzed). Sensor failure usually occurs due to major faults in the power unit, sensing, and communication modules [30], [31]. In practice, identifying sensor failure is a relatively simple task. Continuous inspections and maintenance can be performed indeed whenever a server stops receiving the sensor data. In our study, we assume sensor failure happens when a sensor consecutively stops transmitting data for $\mathcal{T}_1$ hours. We then exclude the sensor from the analysis (i.e., failure) when the ratio of available data of sensor number l of pollutant variable p_v ($N_l^{p_v}$) to total period of analysis (N) is less than or equal to a threshold $P\%$: $(N_l^{p_v}/N) \times 100\% \leq P\%$. It is worth noting that the missing data at the server is not always because of sensor failure; instead, it might occur due to the maintenance activities in a sensor network. Due to this reason, in our analysis, we exclude the sensors' data that had: 1) more than five days of consecutive missing data (i.e., $\mathcal{T}_1 = 120$ h) and 2) the ratio of available data is not sufficient ($P \leq 50\%$).

We analyze the reliability of the sensor network using the mean time between failures (MTBFs). The MTBF is a common measure for describing the expected time between two failures of a repairable system, such as a sensor [32]. We use MTBF to analyze how likely a sensor is to fail within a certain time period, and how often a certain type of sensor failure may occur. We measure the failure time of each sensor and we then estimate the MTBF for each sensor type. The MTBF for the three pollutants PM_{10}, $PM_{2.5}$, and O_3 are 4984.57, 2801.67, and 2661.76 h, respectively. Thus, the reliability ranges from 3 to 5 months and can be used to determine sensor maintenance schedules. These values are generally in line with expectations as the performance of air quality sensors degrades over time through the accumulation of dirt at the air inlet which degrades the sensor performance [29]. Note that this corresponds to the failure of the sensing unit and does not incorporate temporary errors resulting in anomalous sensor readings. We discuss this next.

2) Sensor Anomaly Detection: Sensor anomaly detection is used to identify sensors that drift and generate anomalous data patterns in comparison to the measurements of reference instruments. However, a sensor that drifts still can transmit data to a computer server but the data might be inaccurate or poor quality [33]. To evaluate the sensor's drift, we propose an anomaly detection method based on outlier analysis. The method comprises two steps. First, we model each pollutant variable p_v at every reference station $\mathcal{R}$ using a Weibull distribution $f_{\mathcal{R}}^{p_v}(X; \lambda, k)$ as stated in (4). Second, we evaluate the measurements for each p_v at each LCS l (numbered in Table II) that belongs to a cluster $\mathcal{R}$ on the Weibull distribution $f_{\mathcal{R}}^{p_v}(X; \lambda, k)$. Thus, we use the notation $\mathcal{L}_{\{\mathcal{R},l\}}^{p_v}$

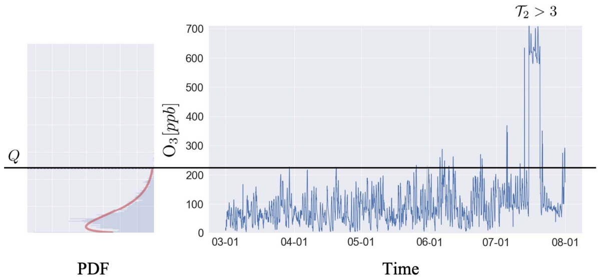


Fig. 4. Anomaly detection for LCS measurements.

Algorithm 1 Sensor Analysis: Failure and Anomaly Detection

```
1: Select $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$
2: Define $\mathcal{T}_1$, $\mathcal{T}_2$ and $P$
3: Define probability range $p$ for $Q^{p_\upsilon}_{\mathcal{R}}(p;\lambda,k)$
4: Estimate $\lambda$ and $k$ for $f^{p_\upsilon}_{\mathcal{R}}(X;\lambda,k)$
5: Compute $Q^{p_\upsilon}_{\mathcal{R}}(p;\lambda,k)$
6: while $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ operates do
7:     Count $\mathcal{C}_1$
8:     Count $\mathcal{C}_2$
9:     if $\mathcal{C}_1 \geq \mathcal{T}_1$ then
10:        $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$: failure mode
11:    else if $\mathcal{C}_1 < \mathcal{T}_1$ then
12:        $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$: function and anomaly detection is activated
13:        if $\mathcal{C}_2 \geq \mathcal{T}_2$ then
14:            $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ is considered to be anomalous
15:        else
16:            $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ functions properly
17:        end if
18:    end if
19: end while
20: if $\frac{N^{p_\upsilon}_l}{N} \times 100\% < P\%$ then
21:    $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ is excluded from analysis
22: else
23:    $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ is included for analysis
24: end if
```

when referring to the measurements of a p_υ. For instance, the LCS number 103 of pollutant variable O_3, which belong to the cluster $\mathcal{R}_{10}$ (see Table II), where can then use the notation $\mathcal{L}^{O_3}_{\mathcal{R}_{10},103}$. If measurements of $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ crosses the outlier line consecutively for more than a threshold time $\mathcal{T}_2$ (in hours), the sensor $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ is assumed to be anomalous. In our study, the outlier line refers to the quantile 0.99 ($q99$) of the Weibull distribution $Q^{p_\upsilon}_{\mathcal{R}}(p;\lambda,k)$, as stated in (5).

For example, Fig. 4 illustrates the proposed anomaly detection applied for $\mathcal{L}^{O_3}_{\mathcal{R}_{10},103}$. In this figure, the Weibull distribution is modeled at the reference station $\mathcal{R}_{10}$, labeled as $f^{O_3}_{10}(X;\lambda,k)$ (left subfigure). In this example, the $\mathcal{L}^{O_3}_{\mathcal{R}_{10},103}$ measurement crosses the outlier line Q and continues for more than the threshold time 3 h (i.e., $\mathcal{T}_2 > 3$, as shown on the right subfigure). This situation is considered to be a sensor anomaly. Note that in our analysis, we have chosen $\mathcal{T}_2$ to be 3 h as the threshold time. The selection of $\mathcal{T}_2$ is a user choice and needs further field engineering and research work to agree on a specific threshold.

Algorithm 1 presents our proposed method for identifying a sensor failure and detecting an anomaly. In line 1, we select $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$, that is, a pollutant variable p_υ from an LCS l that belongs to a cluster $\mathcal{R}$. In line 2, we define $\mathcal{T}_1$, $\mathcal{T}_2$, and P. In this line, the parameter $\mathcal{T}_1$ indicates sensor failure and explains the maximum accepted hours for an LCS that does not transmit data. $\mathcal{T}_2$ refers to the maximum accepted hours for measurement $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ crossing $Q^{p_\upsilon}_{\mathcal{R}}$. The parameter P in % refers to the maximum accepted percentage of missing data from the total measurements. In line 3, we define p that is the probability range for quantile function $Q^{p_\upsilon}_{\mathcal{R}}(p;\lambda,k)$. In line 4, the algorithm uses maximum likelihood to estimates parameters λ and k of Weibull distribution, $f^{p_\upsilon}_{\mathcal{R}}(X;\lambda,k)$ [see (4)]. Line 5 uses the parameters of p, λ, and k (obtained from the previous steps) and computes the quantile function $Q^{p_\upsilon}_{\mathcal{R}}(p;\lambda,k)$.

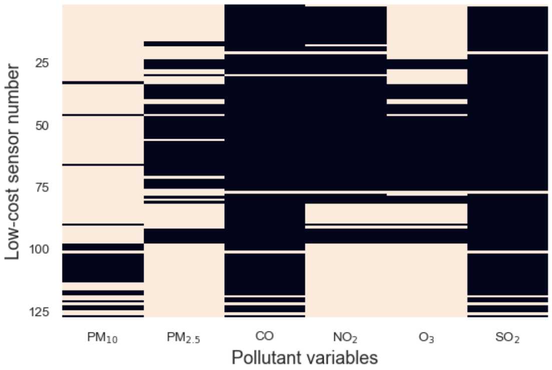
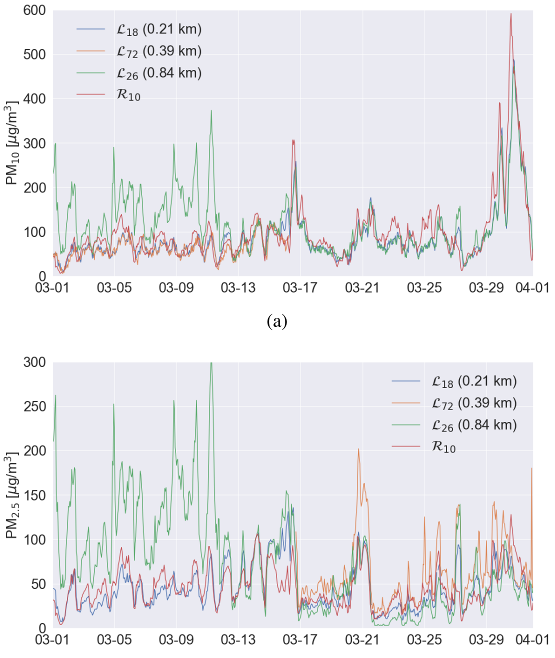


Fig. 5. Individual analysis: sensor failure and anomaly.

From lines 6 to 19, $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ operates. In line 7, the parameter $\mathcal{C}_1$ counts the occurrence of consecutive missing data. In line 8, the parameter $\mathcal{C}_2$ counts the occurrence number of $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ measurement ($x^{p_\upsilon}_l$) crosses the Q line, that is, $x^{p_\upsilon}_l > Q^{p_\upsilon}_{\mathcal{R}}(p;\lambda,k)$. The notation $x^{p_\upsilon}_l$ is the measurement data point of pollutant variable p_υ in a sensor l. In line 9, if $\mathcal{C}_1$, that is, the occurrence of consecutive missing data is bigger of equal to the threshold $\mathcal{T}_1$, then in line 10, $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ is considered to be in a failure mode. If $\mathcal{C}_1$ is smaller than $\mathcal{T}_1$ (line 11), then the $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ functions and anomaly detection method is activated (line 12). In lines 13 and 14, $\mathcal{C}_2$ is bigger or equal to $\mathcal{T}_2$, then $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ is considered to be anomalous. Otherwise, in lines 15 and 16, $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ is considered to function well. We exclude the sensor data from analysis, if the ratio of the available $\mathcal{L}^{p_\upsilon}_{\{\mathcal{R},l\}}$ data ($N^{p_\upsilon}_l/N$) is less than the accepted percentage of missing data from the total period of analysis ($(N^{p_\upsilon}_l/N) \times 100\% < P$) as stated in lines 20–24. Note that some sensor data are not missing because the measurements of some pollutant variables are not available in the LCSs. For example, some LCSs have been equipped with the sensing units of PM$_{10}$, PM$_{2.5}$, and O_3, but sensing units for measuring other pollutants, such as CO and SO_2 are not attached in the LCS unit. Therefore, Algorithm 1 also functions in providing information about which sensor data are available adequately for then being used in data analytics.

The results generated from Algorithm 1 for all LCS failures and anomalies are depicted in Fig. 5. The x-axis shows the pollutant sensor type and the y-axis shows the number of LCSs. The color cream indicates that the sensors function well. The black color indicates that the sensors are in failure or in anomaly modes, where Algorithm 1 detects and filters them out. The sensors which are in anomaly or failure modes are



(a)

(b)

Fig. 6. Time-series plots for PM$_{10}$ (top plot) and PM$_{2.5}$ (bottom plot) for the three nearest LCSs to $\boldsymbol{\mathcal{R}}_{10}$. (a) PM$_{10}$. (b) PM$_{2.5}$.

therefore recommended to be inspected manually. In addition, Fig. 5 shows that almost all sensors for CO and SO$_2$ are in anomaly or failure modes (presented with black color), and also more than half of sensors measuring NO$_2$ have similar situations. Due to this reason, in our work, we do not perform deep analysis for these pollutant variables in Section IV.

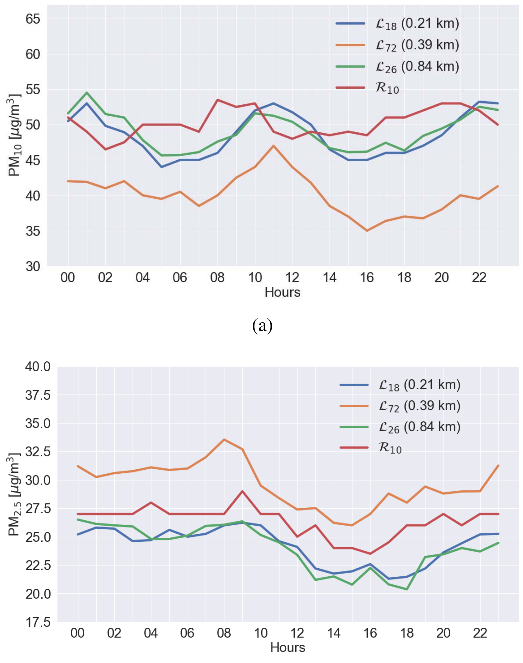
IV. Advantages of Dense Sensor Deployment

This section explains the advantages of deploying dense air quality sensors in city districts. The advantages include local pollution monitoring, hotspots, and environmental analyses explained in Sections IV-A–IV-C.

A. Local Pollution Monitoring

One of the key motivations for deploying dense LCSs in a city district is to provide local pollution monitoring in high resolution. For example, Fig. 6 depicts time-series plots for the measurements of PM$_{10}$ (top subfigure) and PM$_{2.5}$ (bottom subfigure) from the three nearest LCSs $\mathcal{L}$ to $\mathcal{R}_{10}$, between March 1 and April 1, 2021. As shown in the figure, the nearest two LCSs which are $\mathcal{L}_{18}$ (blue) and $\mathcal{L}_{72}$ (orange) follow the PM$_{10}$ and PM$_{2.5}$ readings measured at $\mathcal{R}_{10}$ (red). Indeed, due to the close distances of $\mathcal{L}_{18}$ (0.21 km) and $\mathcal{L}_{72}$ (0.39 km), their PM$_{10}$ and PM$_{2.5}$ measurements are similar with the PM concentrations measured at $\mathcal{R}_{10}$.

However, the measurements of the third nearest LCS $\mathcal{L}_{26}$ (green) which is located about 0.84 km from $\mathcal{R}_{10}$ do not follow the PM readings measured at $\mathcal{R}_{10}$. This is particularly evident from the period between March 1 and 13. During this period, the measurements of both sensing units are different after March 13, the readings of $\mathcal{L}_{26}$ start to follow the measurements of $\mathcal{R}_{10}$ again. This is due to the fact that between March 1 and 13, the PM concentrations' discrepancy between $\mathcal{L}_{26}$ and $\mathcal{R}_{10}$ occurs due to local emissions emitted next to $\mathcal{L}_{26}$. As measurements are similar at both sensing units after March 13, we can safely conclude that the measurements of $\mathcal{L}_{26}$ between March 1 and 13 are not a sensor drift. Note that a sensor drift takes place when the readings of an LCS continuously generates data that does not follow the measurement patterns of a reference station $\mathcal{R}$.



(a)

(b)

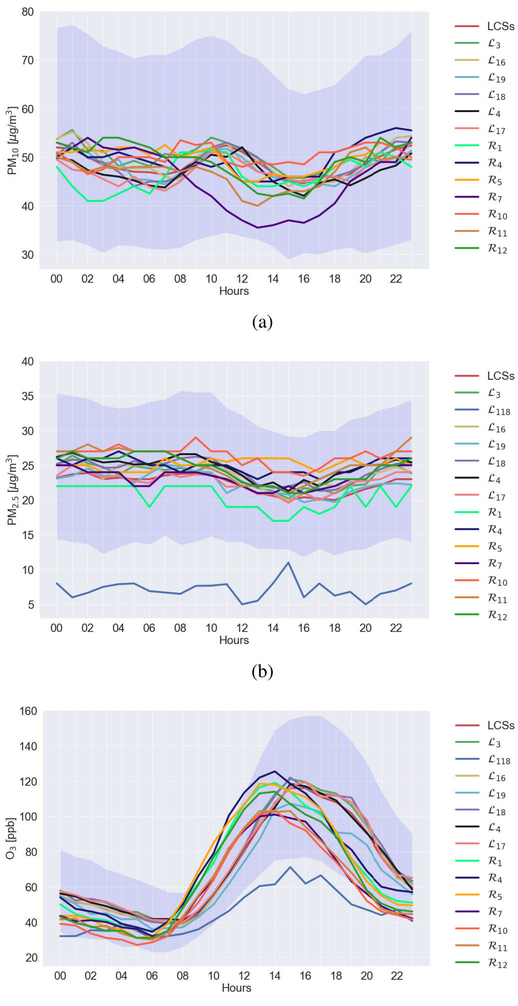
Fig. 7. Median diurnal cycles for PM$_{10}$ (top plot) and PM$_{2.5}$ (bottom plot) for the three nearest LCSs to $\boldsymbol{\mathcal{R}}_{10}$. (a) PM$_{10}$. (b) PM$_{2.5}$.

Fig. 7 demonstrates the median diurnal cycles of PM$_{10}$ (top subfigure) and PM$_{2.5}$ (bottom subfigure) for measurements at $\mathcal{R}_{10}$ and its three nearest LCSs, that is, $\mathcal{L}_{18}$, $\mathcal{L}_{26}$, and $\mathcal{L}_{72}$. These subfigures show that the median diurnal cycles of PM$_{10}$ and PM$_{2.5}$ at $\mathcal{R}_{10}$ differ from the three LCSs. While the median diurnal cycles of $\mathcal{L}_{18}$ (blue) and $\mathcal{L}_{26}$ (green) are similar, the median diurnal cycle of $\mathcal{L}_{72}$ (orange) shows a different pattern, that is, the lowest for PM$_{10}$ and the highest for PM$_{2.5}$.

Considering that the LCSs ($\mathcal{L}$) are accurate as presented in the validation in Section III, the results show that the diurnal cycles' discrepancies between $\mathcal{R}_{10}$ and its three nearest $\mathcal{L}$ may occur due to the existence of local pollution sources, for example, from nearby vehicles emissions. The reason is because $\mathcal{R}_{10}$ and the LCSs ($\mathcal{L}$) around it are located in the evaluation point environment (**B**) which monitor air quality for each street in the Gulou. As described in Section II-A, there are different environments where the LCSs are deployed and each environment has its own air pollution profile, such as environment (**B**) which presents variations in pollution concentration caused by traffic. The results of time-series and diurnal cycles (shown in Figs. 6 and 7) emphasize the significance of dense sensors deployment as they are capable of measuring local pollutant concentrations at high resolutions.

B. Pollution Patterns and Environmental Analysis

A dense air quality sensor deployment provides significant data and enables the analysis of pollution patterns at



(a)

(b)

(c)

Fig. 8. Median of diurnal cycles for sensing units deployed in the evaluation environment (**B**). (a) PM$_{10}$. (b) PM$_{2.5}$. (c) O$_3$.

local levels in city districts. For example, Fig. 8 illustrates the median diurnal cycles for pollutants PM$_{10}$, PM$_{2.5}$, and O$_3$ for the evaluation environment (**B**). Note that the different environments where the LCSs are deployed are explained in Section II-A.

In Fig. 8, the lines (shown with different colors) represent the median diurnal cycles of the measurements obtained from $\mathcal{R}$ and different LCSs ($\mathcal{L}$) deployed in environment (**B**). While the red line (LCSs shown in all subfigures) shows the median of diurnal cycles, the shaded area presents the 25%–75% percentile of diurnal cycles for all LCSs ($\mathcal{L}$) in the environment (**B**). In the measurements in Fig. 8, however, the patterns of median diurnal cycles are almost similar for the majority of LCSs ($\mathcal{L}$) and reference stations $\mathcal{R}$, the median diurnal cycle discrepancies still exist. The discrepancies would provide valuable information by performing a proper investigation that helps understand the air pollution sources and causes.

Fig. 8(a) shows that the median diurnal cycles of LCSs ($\mathcal{L}$) and $\mathcal{R}$ lie within the shaded area. Their concentrations and patterns are also similar, in the range between 40 and 60 μg/m^3. In this figure, the highest diurnal cycle of PM$_{10}$ concentration is obtained from $\mathcal{L}_5$ (black), especially before 14:00. The reason would be the location of the sensor device where it is installed with a high construction works in its surroundings. In contrast, $\mathcal{L}_7$ (purple) shows that the lowest level of the median diurnal cycle of PM$_{10}$ presents the pollution level of the location and environment where it is installed.

Fig. 8(b) and (c) shows that the median diurnal cycles obtained from all LCSs $\mathcal{L}$ and $\mathcal{R}$ are almost similar and lie on the shaded area, except for $\mathcal{L}_{118}$ (blue). These figures show that the PM$_{2.5}$ and O$_3$ concentrations are very low at $\mathcal{L}_{118}$ because this sensor device is located at a school and surrounded by trees. Based on these results, the authorities who are in charge of maintaining the deployed sensors can investigate and learn from the location where $\mathcal{L}_{118}$ is installed, the reasons why the pollution level (i.e., PM$_{2.5}$ and O$_3$ concentration) is low there in this case.

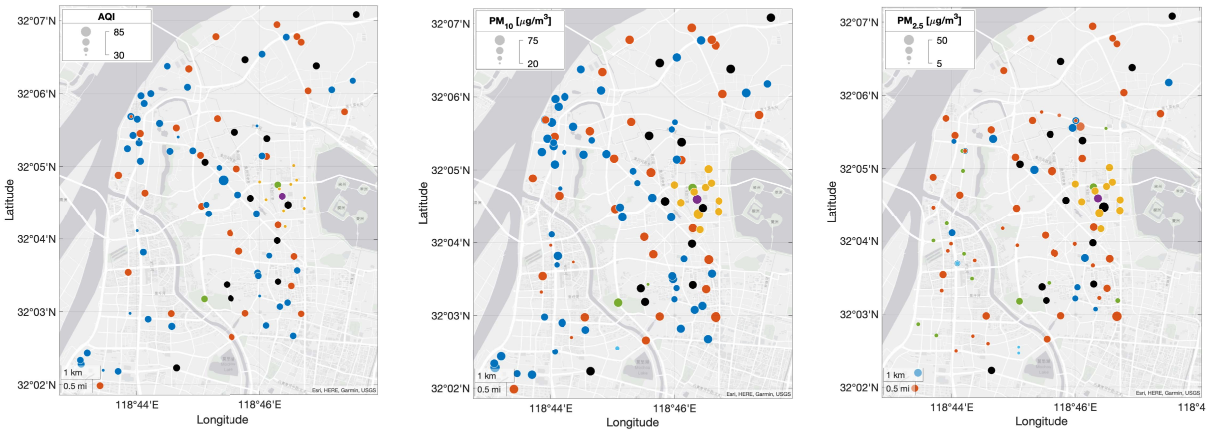
In conclusion, having a dense air quality sensor network in urban areas helps authorities: 1) to understand the pollutants' variations in each area where the sensor is installed and 2) to influence policymakers and designers to improve the environments in cities by learning from the least and highest polluted environments.

C. Pollution Hotspot Analysis

An air pollution hotspot refers to an area where the pollution concentration level crosses a threshold level (defined by authorities) for consecutive days [24]. One way to detect air pollution hotspots in cities is to deploy dense air pollution sensors. For example, Fig. 9 illustrates the hotspots of aggregated air pollutants obtained from LCSs ($\mathcal{L}$) and reference stations $\mathcal{R}$ deployed in Gulou, Nanjing, China. Fig. 9(a) shows the pollution hotspots for AQI, and Fig. 9(b) and (c) illustrates hotspots for PM$_{10}$ and PM$_{2.5}$, respectively. The colored circles in the subfigures in Fig. 9 indicate to different environmental types as explained in Section II-A. The sizes of colored circles also represent the pollution concentration levels.

Fig. 9(a) shows that majority of worst AQI hotspots are presented by blue and red circles, while the highest hotspot is shown by the blue-colored circle (in the middle of the map). The color blue represents the areas with construction work (**C**) and the color red refers to the evaluation points (**B**) (i.e. the area where the sensors monitor air quality for each street in the district). Fig. 9(b) illustrates the hotspots for PM$_{10}$ concentrations, where the highest PM$_{10}$ concentrations are dominated by construction sites (blue). These results are expected because PM$_{10}$ coarse particles are mainly sourced from dust particles emitted through construction activities [34]. Fig. 9(c) depicts the hotspots for PM$_{2.5}$ concentrations. It can be seen that high PM$_{2.5}$ concentrations mainly take place at evaluation (red) and parallel (yellow) points. The parallel points represent air pollution concentrations around the Shanxi Road National Control Station. These results are also expected because the environments presented by these colors are the main roads in Gulou. Generally, the concentrations of PM$_{2.5}$ fine particles are high on the roads, which are mainly sourced from gasoline or diesel combustion emitted by motor vehicle emissions [34].

For the pollution hotspots in the other types of environments in Gulou, the results show that interestingly the sensors which area placed on roadsides (green) do not capture high



(a) (b) (c)

Fig. 9. Different hotspot levels of AQI, PM$_{2.5}$, and PM$_{10}$ measured via sensors network. The different colored circles represent different environments: roadside (**A**-●), evaluation (**B**-●), construction (**C**-●), transport (**D**-●), parallel (**E**-●), quality control (**F**-●), reference station ($\mathcal{R}$-●). (a) AQI. (b) PM$_{10}$. (c) PM$_{2.5}$.

pollutant concentrations. Another conclusion would be that most vehicles that move in the evaluation (red) and parallel (yellow) points might experience traffic jams. In addition, the vehicles which move on the city's main roads (red) might move smoothly which may result in less emission generation. Indeed, these results show that air pollution hotspots analysis based on mapping the pollution concentration to the environment types enables identifying pollution sources. The information achieved from the results would also allow the authorities to take proper actions to minimize the impact of air pollution in those environments [35].

V. Discussion

This section discusses potential improvements and some applications obtained from the deployment of dense air quality sensors network.

A. Potential Improvements

1) In-Field Sensor Calibrations: In-field sensor calibrations are important procedures to ensure LCSs perform accurate and reliable measurements. Even though LCSs undergo laboratory calibration before deploying in the field, the fluctuations in meteorological conditions and changes in anthropogenic sources would lead to drifts in the calibrated sensors. Thanks to ground-truth data obtained from reference air quality monitoring stations that can be used to develop sensor calibration models to improve the measurements of LCSs in a sensor network [36].

2) Virtual Sensors: Virtual sensors enable the estimation of air pollutant concentrations that are not measured by physical sensors. For example, black carbon (BC) is known to be a vital variable in air quality assessments. Unfortunately, the instruments needed for measuring these pollutants are expensive (e.g., a proper BC measurement setup is on the order of approximately \$50 000) [37]. Using machine-learning models indeed enables the developing BC virtual sensors by training the models on ground-truth data obtained from reference instruments. Then, using the inputs from the measurements of LCSs in a sensor network, the BC virtual sensors can be activated and scaled up.

3) Sensors Network Faults Detection and Identification: Sensors network faults detection and identification become an important procedure to ensure sensors' continuous operation. As described in Section III-C, we propose failure and anomaly detection methods. However, our methods have not been verified yet by evaluating them against the ground-truth data. To achieve this, controlled experiments can be done by installing LCSs and reference instruments side by side. The experiments are important to collect more ground-truth sensor data that can be used to develop fault detection and identification methods [10]. Implementing these methods ensures the continuous operation of LCSs in a network.

4) Data Fusion Methods: Data fusion methods based on geostatistics can be used to merge air quality measurements in a sensor network with spatial information obtained from an urban-scale air quality model. The missing air quality information (i.e., the spatial data gaps) usually occurs due to sensor failure and uneven sensor deployment in a network. Data fusion methods enable data integration between sensor network data and modeling data obtained from urban-scale air quality models, such as the Episode dispersion model [38] and Enfuser model [39], which helps filling the air quality data gaps in both space and time.

5) Data Communication and Computing Technologies: Data communication and computing technologies are important concerns in the deployment of LCSs in a network. Thanks to the advancements in wireless and communication technologies such as 5G networks which support the massive deployment of sensors and provide rapid computation through their edge computers [40]. Indeed, using 5G for a sensor network deployed in city areas enables data collection and processing of sensors' continuous measurements [41]. For example, developed sensor calibration and virtual sensor models can be deployed at large scales at the edge computing platforms offered by 5G.

B. Current and Future Applications

In Section IV, we present several advantages in deploying a dense air quality sensor network including local pollution monitoring, understanding pollution patterns and environmental analysis, and analyzing pollution hotspots. In addition to these applications, here we discuss other potential applications that can be derived from a dense air quality sensor network deployment. One potential application of such a dense sensor network is to develop green-path route maps using fine-grained air quality information. These maps, for example, can be used by pedestrians and cyclists to plan their journeys with the cleanest air [42]. Sensor network also benefits the public with additional information on particle pollution levels ($PM_{2.5}$) in the air, particularly during wildfires [43]. Furthermore, air quality sensor network has been used as alert systems to detect ambient odor concentrations within ports due to illegal release from the degasification of liquid-carrying vessels or unintended leaks [44].

It is expected that in the near future, air quality sensor network deployment would generate profits and revenues for the investors of air quality infrastructures and service providers. The information obtained from such a sensor network is expected to move toward sophisticated business models, as described in [6]. The total size of the global air quality monitoring market which includes both reference stations, LCS hardware, and services is estimated to reach $6.4 billion in the next three years [45]. Therefore, any application that is developed based on data obtained from a dense air quality sensor network would offer benefits for the authorities and organizations operating and managing the network.

VI. Conclusion

In this article, we present sensor validation methods and data analysis for a dense air quality sensor network. We show solutions to challenges in a large-scale sensor network deployment. We use data from a dense air quality sensor network deployment, located in Nanjing downtown, China, that comprises 126 LCSs and 13 reference stations. Since the majority of sensors deployed in the network are based on LCSs, they are prone to have low-quality data.

Therefore, we propose three methods of sensor validation. First, we perform a reliability investigation to evaluate all LCSs in the network to observe if they provide reliable measurements as a whole in comparison to the measurements of all reference stations. Thus, we compare the measurements between all LCSs and the reference stations by means of statistical properties and correlation coefficients between pollutant variables measured at both sensing units. Second, we perform accuracy tests on a few of the LCSs which are nearest to the reference stations. The accuracy tests are generalized to the remaining LCSs in the sensor network as the LCSs are based on the same sensing technology, as they are identical units. Third, we perform failure and anomaly detection on individual LCSs to evaluate which sensor functions properly and generates reliable air quality data. From the validation results, we conclude that the sensor network is reliable as a whole and the accuracy tests indicate that the sensors for PM_{10}, $PM_{2.5}$, and O_3 are accurate to be used in the analysis.

Due to sensor failure and anomaly, we propose an algorithm to filter sensor data for our data analysis. Based on data analysis, we demonstrate that the dense air quality sensor network generates high-resolution air quality data which benefit: 1 local air pollution monitoring; 2) pollution patterns and environmental analysis; and 3) pollution hotspot analysis. Finally, we discuss potential approaches for improving the dense sensor network in terms of technologies and applications. We also discuss how the dense sensor network can generate profits for investors of a dense sensor network. Naturally, there is room for further improvements. For example, the effects of meteorological factors, and especially those that affect dispersion such as wind speed and direction, and anthropogenic factors, such as emissions from vehicles or heavy industry, need further study and evaluation to better understand the benefits and limitations in LCSs in a wide range of situations. Nevertheless, our work has demonstrated that data captured by LCS is generally accurate, precise, and consistent, closely aligning with reference stations the vast majority of the time. Our results pave the way toward broader adoption of LCS networks to increase the coverage of air quality information and to support new types of hybrid deployments for collecting air quality data.

Acknowledgment

This project was part of the Pan-Eurasian Experiment (PEEX) Program. This work was part of a close research collaboration between the Joint International Research Laboratory of Atmospheric and Earth System Sciences (JirLATEST) at Nanjing University, China, Institute for Atmospheric and Earth System Research (INAR) and the Department of Computer Science at University of Helsinki, Finland.

Martha Arbayani Zaidan is with the Joint International Research Laboratory of Atmospheric and Earth System Sciences, Nanjing University, Nanjing 210023, China, also with the Department of Computer Science and the Institute for Atmospheric and Earth System Research (INAR), University of Helsinki, 00560 Helsinki, Finland, and also with the Nanjing Atmospheric Environment and Green Development Research Institute (NAGR), Nanjing 210000, China (e-mail: martha.zaidan@helsinki.fi).

Yuning Xie is with the Nanjing Atmospheric Environment and Green Development Research Institute (NAGR), Nanjing 210000, China (e-mail: ynxie@nagr.com.cn).

Naser Hossein Motlagh, Petteri Nurmi, and Sasu Tarkoma are with the Department of Computer Science, University of Helsinki, 00560 Helsinki, Finland (e-mail: naser.motlagh@helsinki.fi; petteri.nurmi@helsinki.fi; sasu.tarkoma@helsinki.fi).

Bo Wang is with Gulou Environment Protection Department, Nanjing 210008, China (e-mail: wangwu-5@163.com).

Wei Nie and Aijun Ding are with the Joint International Research Laboratory of Atmospheric and Earth System Sciences, Nanjing University, Nanjing 210023, China (e-mail: niewei@nju.edu.cn; dingaj@nju.edu.cn).

Tuukka Petäjä and Markku Kulmala are with the Institute for Atmospheric and Earth System Research (INAR), University of Helsinki, 00560 Helsinki, Finland, and also with the Joint International Research Laboratory of Atmospheric and Earth System Sciences, Nanjing University, Nanjing 210023, China (e-mail: tuukka.petaja@helsinki.fi; markku.kulmala@helsinki.fi).

References

[1] *World Health Statistics 2019: Monitoring Health for the SDGs, Sustainable Development Goals*, World Health Organization, Geneva, Switzerland, 2021.

[2] A. A. Almetwally, M. Bin-Jumah, and A. A. Allam, “Ambient air pollution and its influence on human health and welfare: An overview,” *Environ. Sci. Pollut. Res.*, vol. 27, no. 20, pp. 24815–24830, Jul. 2020.

[3] Y. Zhu, J. Xie, F. Huang, and L. Cao, “Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China,” *Sci. Total Environ.*, vol. 727, Jul. 2020, Art. no. 138704.

[4] F. Chen and Z. Chen, "Cost of economic growth: Air pollution and health expenditure," *Sci. Total Environ.*, vol. 755, Feb. 2021, Art. no. 142543.

[5] Z. Zhou, Z. Ye, Y. Liu, F. Liu, Y. Tao, and W. Su, "Visual analytics for spatial clusters of air-quality data," *IEEE Comput. Graph. Appl.*, vol. 37, no. 5, pp. 98–105, May 2017.

[6] K. Schäfer et al., "High-resolution assessment of air quality in urban areas—A business model perspective," *Atmosphere*, vol. 12, no. 5, p. 595, 2021.

[7] N. H. Motlagh et al., "Toward massive scale air quality monitoring," *IEEE Commun. Mag.*, vol. 58, no. 2, pp. 54–59, Feb. 2020.

[8] J. Huang et al., "A crowdsource-based sensing system for monitoring fine-grained air quality in urban environments," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3240–3247, Apr. 2019.

[9] M. A. Fekih et al., "Participatory air quality and urban heat islands monitoring system," *IEEE Trans. Instrum. Meas.*, vol. 70, 2020, Art. no. 9503914.

[10] M. A. Zaidan et al., "Intelligent air pollution sensors calibration for extreme events and drifts monitoring," *IEEE Trans. Ind. Informat.*, early access, Feb. 15, 2022, doi: 10.1109/TII.2022.3151782.

[11] T. V. Kokkonen et al., "The effect of urban morphological characteristics on the spatial variation of $PM_{2.5}$ air quality in downtown Nanjing," *Environ. Sci., Atmos.*, vol. 1, no. 7, pp. 481–497, 2021.

[12] C. Zhou, G. Wei, J. Xiang, K. Zhang, C. Li, and J. Zhang, "Effects of synoptic circulation patterns on air quality in Nanjing and its surrounding areas during 2013–2015," *Atmos. Pollut. Res.*, vol. 9, no. 4, pp. 723–734, 2018.

[13] B. R. Gurjar, T. M. Butler, M. G. Lawrence, and J. Lelieveld, "Evaluation of emissions and air quality in megacities," *Atmos. Environ.*, vol. 42, no. 7, pp. 1593–1606, Mar. 2008.

[14] L. Xu et al., "Spatiotemporal pattern of air quality index and its associated factors in 31 Chinese provincial capital cities," *Air Qual., Atmos. Health*, vol. 10, no. 5, pp. 601–609, 2017.

[15] C. C. Robusto, "The cosine-haversine formula," *Amer. Math. Monthly*, vol. 64, no. 1, pp. 38–40, 1957.

[16] J. Taylor, A. Jakeman, and R. Simpson, "Modeling distributions of air pollutant concentrations—I. Identification of statistical models," *Atmos. Environ. (1967)*, vol. 20, no. 9, pp. 1781–1789, 1986.

[17] B. Rumburg, R. Alldredge, and C. Claiborn, "Statistical distributions of particulate matter and the error associated with sampling frequency," *Atmos. Environ.*, vol. 35, no. 16, pp. 2907–2920, Jun. 2001.

[18] J. Wang, X. Zhang, Z. Guo, and H. Lu, "Developing an early-warning system for air quality prediction and assessment of cities in China," *Expert Syst. Appl.*, vol. 84, pp. 102–116, Oct. 2017.

[19] J. E. Cavanaugh and A. A. Neath, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *WIREs Comput. Statist.*, vol. 11, no. 3, p. e1460, May 2019.

[20] V. Leiva, F. Vilca, N. Balakrishnan, and A. Sanhueza, "A skewed sinh-normal distribution and its properties and application to air pollution," *Commun. Statist.-Theory Methods*, vol. 39, no. 3, pp. 426–443, Jan. 2010.

[21] G. Martínez-Flórez, H. Bolfarine, and H. W. Gómez, "The log-power-normal distribution with application to air pollution," *Environmetrics*, vol. 25, no. 1, pp. 44–56, Feb. 2014.

[22] N. A. Al-Dhurafi, N. Masseran, Z. H. Zamzuri, and M. A. M. Safari, "Modeling the air pollution index based on its structure and descriptive status," *Air Qual., Atmos. Health*, vol. 11, no. 2, pp. 171–179, Mar. 2018.

[23] S. Marco, A. Ortega, A. Pardo, and J. Samitier, "Gas identification with tin oxide sensor array and self-organizing maps: Adaptive correction of sensor drifts," *IEEE Trans. Instrum. Meas.*, vol. 47, no. 1, pp. 316–321, Feb. 1998.

[24] E. Lagerspetz et al., "MegaSense: Feasibility of low-cost sensors for pollution hot-spot detection," in *Proc. IEEE 17th Int. Conf. Ind. Inform. (INDIN)*, vol. 1, Jul. 2019, pp. 1083–1090.

[25] A. R. Whitehill, M. Lunden, S. Kaushik, and P. Solomon, "Uncertainty in collocated mobile measurements of air quality," *Atmos. Environment: X*, vol. 7, Oct. 2020, Art. no. 100080.

[26] N. H. Motlagh et al., "Low-cost air quality sensing process: Validation by indoor-outdoor measurements," in *Proc. 15th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, Nov. 2020, pp. 223–228.

[27] S. Munir, M. Mayfield, D. Coca, S. A. Jubb, and O. Osammor, "Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities—A case study in Sheffield," *Environ. Monit. Assessment*, vol. 191, no. 2, pp. 1–22, 2019.

[28] S. Feinberg et al., "Long-term evaluation of air sensor technology under ambient conditions in denver, Colorado," *Atmos. Meas. Techn.*, vol. 11, no. 8, pp. 4605–4615, 2018.

[29] F. Concas et al., "Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis," *ACM Trans. Sensor Netw.*, vol. 17, no. 2, pp. 1–44, 2021.

[30] R. N. Duche and N. P. Sarwade, "Sensor node failure detection based on round trip delay and paths in WSNs," *IEEE Sensors J.*, vol. 14, no. 2, pp. 455–464, Feb. 2014.

[31] P. J. Basford, F. M. J. Bulot, M. Apetroaie-Cristea, S. J. Cox, and S. J. Ossont, "LoRaWAN for smart city IoT deployments: A long term evaluation," *Sensors*, vol. 20, no. 3, p. 648, Jan. 2020.

[32] T. Jin, *Reliability Engineering and Services*. Hoboken, NJ, USA: Wiley, 2019.

[33] S. De Vito, G. Fattoruso, M. Pardo, F. Tortorella, and G. Di Francia, "Semi-supervised learning techniques in artificial olfaction: A novel approach to classification problems and drift counteraction," *IEEE Sensors J.*, vol. 12, no. 11, pp. 3215–3224, Nov. 2012.

[34] C. K. Chan and X. Yao, "Air pollution in mega cities in China," *Atmos. Environ.*, vol. 42, no. 1, pp. 1–42, Jan. 2008.

[35] P. Kortoçi et al., "Air pollution exposure monitoring using portable low-cost air quality sensors," *Smart health*, vol. 23, Mar. 2022, Art. no. 100241.

[36] M. A. Zaidan et al., "Intelligent calibration and virtual sensing for integrated low-cost air quality sensors," *IEEE Sensors J.*, vol. 20, no. 22, pp. 13638–13652, Nov. 2020.

[37] M. A. Zaidan, D. Wraith, B. E. Boor, and T. Hussein, "Bayesian proxy modelling for estimating black carbon concentrations using white-box and black-box models," *Appl. Sci.*, vol. 9, no. 22, p. 4976, Nov. 2019.

[38] P. Schneider, N. Castell, M. Vogt, F. R. Dauge, W. A. Lahoz, and A. Bartonova, "Mapping urban air quality in near real-time using observations from low-cost sensors and model information," *Environ. Int.*, vol. 106, pp. 234–247, Sep. 2017.

[39] L. Johansson, A. Karppinen, and K. Loven, "Evaluation of air quality using dynamic land-use regression and fusion of environmental information," in *Proc. 2nd Int. Workshop Environ. Multimedia Retr.*, Jun. 2015, pp. 33–38.

[40] X. Su et al., "Intelligent and scalable air quality monitoring with 5G edge," *IEEE Internet Comput.*, vol. 25, no. 2, pp. 35–44, Mar./Apr. 2021.

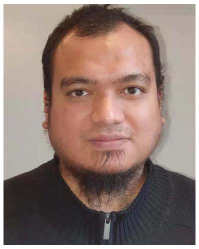
[41] N. H. Motlagh et al., "mMTC deployment over sliceable infrastructure: The Megasense scenario," *IEEE Netw.*, vol. 35, no. 6, pp. 247–254, Nov./Dec. 2021.

[42] A. Nurminen, A. Malhi, L. Johansson, and K. Främling, "A clean air journey planner for pedestrians using high resolution near real time air quality data," in *Proc. 16th Int. Conf. Intell. Environ. (IE)*, Jul. 2020, pp. 44–51.

[43] R. Williams et al., "Deliberating performance targets workshop: Potential paths for emerging $PM_{2.5}$ and O_3 air sensor progress," *Atmos. Environ., X*, vol. 2, Apr. 2019, Art. no. 100031.

[44] B. Milan, S. Bootsma, and I. Bilsen, "Advances in odour monitoring with E-noses in the port of rotterdam," *Chem. Eng. Trans.*, vol. 30, pp. 145–s150, Sep. 2012.

[45] *Air Quality Monitoring System Market Size & Forecast 2019–2025*. Accessed: Mar. 19, 2022. [Online]. Available: https://www.kbvresearch.com/air-quality-monitoring-system-market/



Martha Arbayani Zaidan (Member, IEEE) received the Ph.D. degree in automatic control and systems engineering from The University of Sheffield, Sheffield, U.K., in 2014.

He was a Postdoctoral Research Associate at Maryland University, USA, and a Fellow at Aalto University, Espoo, Finland. He was a Research Associate Professor at Nanjing University, Nanjing, China. He is a Senior Researcher at the Department of Computer Science and the Institute for Atmospheric and Earth System Research (INAR), University of Helsinki, Helsinki, Finland. His research interests include artificial intelligence and machine learning for intelligent control systems, health monitoring technologies, applied physics, atmospheric, and environmental sciences.

Yuning Xie received the Ph.D. degree in atmospheric sciences from the School of Atmospheric Sciences, Nanjing University, Nanjing, China, in 2017.

He was a Postdoctoral Research Associate with East China Normal University, Shanghai, China. His research interests include atmospheric composition measurement, environmental big data mining, smart city technology, and atmospheric/environmental sciences.

Naser Hossein Motlagh received the D.Sc. degree in networking technology from the School of Electrical Engineering, Aalto University, Espoo, Finland, in 2018.

He was a Postdoctoral Fellow with the Helsinki Institute for Information Technology (HIIT)–Helsinki Center for Data Science (HiDATA) Program, Helsinki, Finland. He is a Docent and a Researcher at the Department of Computer Science, University of Helsinki, Helsinki, within the Nokia Center for Advanced Research (NCAR), Finland. His research interests include the Internet of Things, wireless sensor networks, environmental sensing, smart buildings, and unmanned aerial and underwater vehicles.

Bo Wang is a Senior Engineer from Gulou Environment Protection Department, Nanjing, China. He is an expert in urban environment monitoring and protection. His research interests include low-cost sensor networks for outdoors and buildings, data-driven policy making, urban carbon monitoring, and enactment of environment protection laws.

Wei Nie received the Ph.D. degree in environmental sciences from Shandong University, Jinan, China, in 2012.

He was a Postdoctoral Researcher with the University of Helsinki, Helsinki, Finland. He is an Associate Professor of Atmospheric Chemistry with the School of Atmospheric Sciences, Nanjing University, Nanjing, China. His research interests include secondary organic aerosol formation, atmospheric new particle formation, and heterogeneous chemistry.

Petteri Nurmi received the Ph.D. degree in computer science from the Department of Computer Science, University of Helsinki, Helsinki, Finland, in 2009.

He is an Associate Professor of Distributed Systems and the Internet of Things at the Department of Computer Science, University of Helsinki. His research interests include distributed systems, pervasive data science, and sensing systems.

Sasu Tarkoma (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Helsinki, Helsinki, Finland, in 2006.

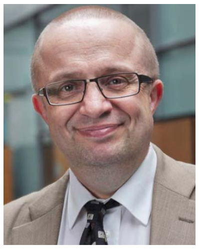
He is a Professor of Computer Science with the University of Helsinki. He is a Visiting Professor with the 6G Flagship, University of Oulu, Oulu, Finland. He has authored four textbooks and has published over 250 scientific articles. He holds ten granted U.S. patents. His research interests include Internet technology, distributed systems, data analytics, and mobile and ubiquitous computing.

Tuukka Petäjä received the Ph.D. degree and Docent title in physics from the University of Helsinki, Helsinki, Finland, in 2006 and 2011, respectively.

He was a Postdoctoral Researcher with the U.S. National Center for Atmospheric Research (NCAR), Boulder, CO, USA. He is a Professor of Experimental Atmospheric Sciences with the Institute for Atmospheric and Earth System Research (INAR), University of Helsinki. He and his team are currently in charge of the development of aerosol particle measuring equipment for continuous measurements. His research interests include atmospheric aerosol particles and their role in climate change and air quality.

Aijun Ding received the Ph.D. degree in meteorology from Nanjing University, Nanjing, China, in 2004.

He is a Professor of Atmospheric Environment and Atmospheric Physics and the Dean of the School of Atmospheric Sciences at Nanjing University. He had about eight-year research experience with the Hong Kong Polytechnic University, Hong Kong, from 2011 to 2009, as a Research Assistant, a postdoctoral, and a Research Fellow, and has been a Full Professor at Nanjing University since 2009. His research interests include air pollution-weather/climate interactions, tropospheric ozone, chemical transport modeling, and Lagrangian dispersion modeling.



Markku Kulmala received the Ph.D. degree in theoretical physics from the University of Helsinki, Helsinki, Finland, in 1988.

He is an Academy Professor and the Head of the Institute for Atmospheric and Earth System Research (INAR), University of Helsinki. He is the Founder of the International Station for Measuring Ecosystem-Atmosphere Relations (SMEAR) Observation Networks. He has published more than 800 SCI articles, including more than 40 articles in Science and Nature. His research interests include atmospheric aerosol nucleation and growth mechanisms, the kinetics of atmospheric aerosols and clusters, and biosphere–aerosol–cloud–climate interactions.

Dr. Kulmala is currently a member of the Academy of Europe and a Foreign Academician of the Chinese Academy of Sciences.