Robust statistical calibration and characterization of portable low-cost air quality monitoring sensors to quantify real-time O_3 and NO_2 concentrations in diverse environments

Ravi Sahu[1], Ayush Nagal[2], Kuldeep Kumar Dixit[1], Harshavardhan Unnibhavi[3], Srikanth Mantravadi[4], Srijith Nair[4], Yogesh Simmhan[3], Brijesh Mishra[5], Rajesh Zele[5], Ronak Sutaria[6], Purushottam Kar[2], and Sachchida Nand Tripathi[1]

[1]Department of Civil Engineering, Indian Institute of Technology Kanpur, Kanpur, India
[2]Department of Computer Science and Engineering, Indian Institute of Technology Kanpur, Kanpur, India
[3]Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India
[4]Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India
[5]Department of Electrical Engineering, Indian Institute of Technology, Bombay, India
[6]Centre for Urban Science and Engineering, Indian Institute of Technology, Bombay, India

Correspondence: Sachchida Nand Tripathi (snt@iitk.ac.in)

Abstract. Low-cost sensors offer an attractive solution to the challenge of establishing affordable and dense spatio-temporal air quality monitoring networks with greater mobility and lower maintenance costs. These low-cost sensors offer reasonably consistent measurements, but require in-field calibration to improve agreement with regulatory instruments. In this paper, we report the results of a deployment and calibration study on a network of six air quality monitoring devices built using the Alphasense O_3 (OX-B431) and NO_2 (NO2-B43F) electrochemical gas sensors. The sensors were deployed in two phases over a period of three months at sites situated within two mega-cities with diverse geographical, meteorological and air quality parameters. A unique feature of our deployment is a *swap-out* experiment wherein three of these sensors were relocated to different sites in the two phases. This gives us a unique opportunity to study the effect of seasonal, as well as geographical variations on calibration performance. We report an extensive study of more than a dozen parametric and non-parametric calibration algorithms. We propose a novel local non-parametric calibration algorithm based on metric-learning that offers, across deployment sites and phases, an R^2 coefficient of upto 0.923 with respect to reference values for O_3 calibration and 0.819 for NO_2 calibration. This represents an 4 - 20 percentage point increase in terms of R^2 values offered by classical non-parametric methods. We also offer a critical analysis of the effect of various data preparation and model design choices on calibration performance. The key recommendations emerging out of this study include 1) incorporating ambient relative humidity and temperature into calibration models, 2) assessing the relative importance of various features with respect to the calibration task at hand, by using an appropriate feature weighing or metric learning technique, 3) using local calibration techniques such as KNN, 4) performing temporal smoothing over raw time series data, but being careful to not do so too aggressively, and 5) making all efforts at ensuring that data with enough diversity is demonstrated to the calibration algorithm while training to ensure good generalization. These results offer insights into the strengths and limitations of these sensors, and offer an encouraging opportunity at using them to supplement and densify compliance regulatory monitoring networks.

1 Introduction

Elevated levels of air pollutants have a detrimental impact on human health as well as the economy (Chowdhury et al., 2018; Landrigan et al., 2018). For instance, high levels of ground-level O_3 has been linked to difficulty in breathing, increased frequency of asthma attacks, and chronic obstructive pulmonary disease (COPD). The World Health Organization reported (WHO, 2018) that in 2016, 4.2 million premature deaths worldwide could be attributed to outdoor air pollution, 91% of which occurred in low- and middle-income countries where air pollution levels often did not meet its guidelines. There is a need for accurately real-time monitoring of air pollution levels with dense spatio-temporal coverage.

Existing regulatory techniques for assessing urban air quality (AQ) rely on a small network of Continuous Ambient Air Quality Monitoring Stations (CAAQMS) that are instrumented with accurate air quality monitoring gas analyzers and Beta-Attenuation Monitors and provide highly accurate measurements (Snyder et al., 2013; Malings et al., 2019). However, these networks are established at a commensurately high setup cost and are cumbersome to maintain (Sahu et al., 2020), making dense CAAQMS networks impractical. However, the AQ data offered by these sparse networks, however accurate, limits the ability to formulate effective AQ strategies (Garaga et al., 2018; Fung, 2019).

In recent years, the availability of low-cost AQ (LCAQ) monitoring devices has provided exciting opportunities for finer spatial resolution data (Rai et al., 2017; Baron and Saffell, 2017; Kumar et al., 2015; Schneider et al., 2017; Zheng et al., 2019). The cost of a Federal Reference Method (FRM)-grade CAAQMS system is around USD 200,000, while that of an LCAQ device running commodity sensors is under USD 500 (Jiao et al., 2016; Simmhan et al., 2019). In this manuscript, we use the term "commodity" to refer to sensors or devices that are not custom built and instead sourced from commercially available options. The increasing prevalance of the Internet of Things (IoT) infrastructure allows building large-scale networks of LCAQ devices (Baron and Saffell, 2017; Castell et al., 2017; Arroyo et al., 2019).

Dense LCAQ networks can complement CAAQMS to help regulatory bodies identify sources of pollution and formulate effective policies, allow scientists to model interactions between climate change and pollution (Hagan et al., 2019), allow citizens to make informed decisions, e.g. on their commute (Apte et al., 2017; Rai et al., 2017), and encourage active participation in citizen science initiatives (Gabrys et al., 2016; Commodore et al., 2017; Gillooly et al., 2019; Popoola et al., 2018).

1.1 Challenges in low-cost sensor calibration

Measuring ground-level O_3 and NO_2 is challenging as they occur at parts per billion levels and intermix with other pollutants (Spinelle et al., 2017). LCAQ sensors are not designed to meet rigid performance standards and may generate less accurate data as compared to regulatory-grade CAAQMS (Mueller et al., 2017; Snyder et al., 2013; Miskell et al., 2018). Most LCAQ gas sensors are based either on metal oxide (MOx) or electrochemical (EC) technologies (Pang et al., 2017; Hagan et al., 2019). These present challenges in terms of sensitivity towards environmental conditions and cross-sensitivity (Zimmerman et al., 2018; Lewis and Edwards, 2016). For example, O_3 electrochemical sensors undergo redox reactions in the presence of NO_2. The sensors also exhibit loss of consistency or *drift* over time. For instance, in EC sensors, reagents are spent over time and have a typical lifespan of one to two years (Masson et al., 2015; Jiao et al., 2016). Thus, there is need for reliable

calibration of LCAQ sensors to satisfy performance demands of end-use applications (De Vito et al., 2018; Akasiadis et al., 2019; Williams, 2019).

1.2 Related Works

Recent works have shown that LCAQ sensor calibration can be achieved by co-locating the sensors with regulatory-grade reference monitors and using various calibration models (De Vito et al., 2018; Hagan et al., 2019; Morawska et al., 2018). Zheng et al. (2019) considered the problem of dynamic $\mathrm{PM}_{2.5}$ sensor calibration within a sensor network. For the case of SO_2 sensor calibration, Hagan et al. (2019) observed that parametric models such as linear least squares regression (LS) could extrapolate to wider concentration ranges, at which non-parametric regression model may struggle. However, LS does not correct for dependence on temperature or relative humidity (RH), at which non-parametric models may be more effective.

Since electrochemical sensors are configured to have diffusion-limited responses, and the diffusion coefficients could get affected by ambient temperature, Sharma et al. (2019); Hitchman et al. (1997); Masson et al. (2015) found that at RH exceeding 75% there is substantial error, possibly due to condensation on the potentiostat electronics. Simmhan et al. (2019) used non-parametric approaches such as regression trees along with data aggregated from multiple co-located sensors to demonstrate the effect of training dataset on calibration performance. Esposito et al. (2016) made use of neural networks and demonstrated good calibration performance (with mean absolute error < 2 ppb) for the calibration of NO_2 sensors. However, a similar performance was not observed for O_3 calibration. Notably, existing works mostly use a localized deployment of a small number of sensor, e.g. Cross et al. (2017) who tested two devices, each containing one sensor per pollutant.

1.3 Our Contributions and the SATVAM initiative

The SATVAM initiative (*Streaming Analytics over Temporal Variables from Air quality Monitoring*) has been developing low-cost air quality (LCAQ) sensor networks based on highly portable IoT software platforms. These LCAQ devices include (see Fig. 2) $\mathrm{PM}_{2.5}$ as well as gas sensors. Details on the IoT software platform and SATVAM node cyber infra-structure are available in (Simmhan et al., 2019). The focus of this paper is to build accurate and robust calibration models for the NO_2 and O_3 gas sensors present in SATVAM devices. Our contributions are summarized below:

1. We report the results of a deployment and calibration study involving six sensors deployed at two sites over two phases with vastly different meteorological, geographical and air quality parameters.

2. A unique feature of our deployment is a *swap-out* experiment wherein three of these sensors were relocated to different sites in the two phases (see Sect. 2 for deployment details). This allows us to investigate the efficacy of calibration models when applied to weather and air quality conditions vastly different from those present during calibration. Such an investigation is missing from previous works which mostly consider only localized calibration.

3. We present an extensive study of parametric and non-parametric calibration models, and develop a novel local calibration algorithm based on metric learning that offers stable (across gases, sites and seasons) and accurate calibration.

Figure 1. A map showing the locations of the deployment sites. Fig. 1(b) and (c) on the right show a local-scale map of the vicinity of the deployment sites – namely Site D at MRIU, **D**elhi NCR (Fig. 1(b)) and Site M at MPCB, **M**umbai (Fig. 1(c)), with the sites themselves pointed out using bright green dots. Fig. 1(a) shows the location of the sites on a map of India. **Credit for Map Sources**: Fig. 1(a) is taken from the NASA Earth Observatory with the outlines of the Indian states in red taken from QGIS3.4 Madeira. Fig. 1(b) and (c) were obtained from, and are, © Google Maps. The green markers for the sites in all figures were added separately.

4. We present an analysis of the effect of data preparation techniques such as volume of data, temporal averaging and data diversity, on calibration performance. This yields several take-home messages that can boost calibration performance.

2 Deployment Setup

Our deployment employed a network of LCAQ sensors as well as reference grade monitors for measuring both NO_2 and O_3 concentrations, deployed at two sites across two phases.

2.1 Deployment Sites

SATVAM LCAQ sensor deployment and collocation with reference monitors was carried out at two sites. Fig. 1 presents the geographical locations of these two sites.

1. **Site D**: located within the **D**elhi National Capital Region (NCR) of India at the Manav Rachna International Institute of Research and Studies, Sector 43, Faridabad (28.45°N, 77.28°E, 209 m above mean sea level).
2. **Site M** (in Mumbai): located within the city of **M**umbai at the Maharashtra Pollution Control Board within the university campus of IIT Bombay (19.13°N, 72.91°E, and 50 m above mean sea level).

We refer the reader to the supplementary material for additional details about the two deployment sites. Due to increasing economic and industrial activities, a progressive worsening of ambient air pollution is witnessed at both sites. We considered these two sites to cover a broader range of pollutant concentrations and weather patterns, so as to be able to test the reliability of

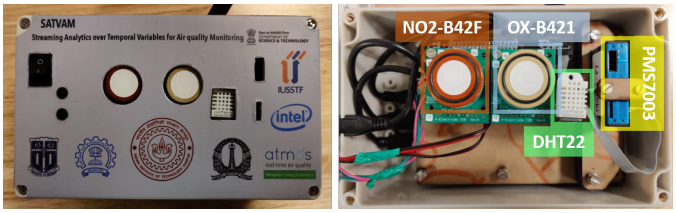

Figure 2. Primary components of the *SATVAM* LCAQ (low-cost air-quality) sensor used in our experiments. The SATVAM device consists of a Plantower PMS7003 $PM_{2.5}$ sensor, Alphasense OX-B431 and NO2-B43F electrochemical sensors, and a DHT22 RH and temperature sensor. Additional components (not shown here) include instrumentation to enable data collection and transmission.

LCAQ networks. It is notable that the two chosen sites present different geographical settings as well as different air pollution levels with site D of particular interest in presenting significantly higher minimum O_3 levels than site M, illustrating the influence of the geographical variability over the selected region.

2.2 Instrumentation

LCAQ Sensor Design: Each *SATVAM* LCAQ device contains two commodity electrochemical gas sensors (Alphasense OX-B421 and NO2-B42F) for measuring O_3 (ppb) and NO_2 (ppb) levels, a PM sensor (Plantower PMS7003) for measuring $PM_{2.5}$ (μg m^{-3}) levels, and a DHT22 sensor for measuring ambient temperature in °C and RH in percentage points. Fig. 2 shows the placement of these components. A notable feature of this device is its focus on frugality with use of the very low-power ContikiOS platform and 6LoWPAN for providing wireless sensor network connectivity.

Detailed information on assembling these different components and the interfacing with an IoT network is described in (Simmhan et al., 2019). These sensors form a highly portable IoT software platform to transmit 6LoWPAN packets at 5 minute intervals containing five time-series data points from individual sensors, namely NO_2, O_3, $PM_{2.5}$ (not presented in this study), temperature and RH. Given the large larger number of devices spread across two cities and seasons in this study, a single border-router edge device was configured at both sites using a Raspberry Pi that acquired data, integrated it, and connected to a cloud facility using a WiFi-link to the respective campus broadband networks. A Microsoft Azure Standard D4s v3 VM was used to host the cloud service with 4 cores, 16 GB RAM and 100 GB SSD storage running an Ubuntu 16.04.1 LTS OS. The Pi edge device was designed to ensure that data acquisition continues even in the event of cloud VM failure.

Reference Monitors: At both the deployment sites, O_3 and NO_2 were measured simultaneously with data available at 1 minute intervals for site D deployments (both Jun and Oct) and 15 minute intervals for site M deployments. O_3 and NO_2 values were measured at site D using an ultraviolet photometric O_3 analyzer (Model 49i O_3 analyzer, Thermo Scientific^{TM}, USA) and a chemiluminescence oxide of nitrogen (NOx) analyzer (Model 42i NOx analyzer, Thermo Scientific^{TM}, USA), respectively. Regular maintenance and multi-point calibration, zero checks, and zero settings of the instruments were carried out following the method described by (Gaur et al., 2014). The lowest detectable limits of reference monitors in measuring O_3 and NO_2 are 0.5 ppb and 0.40 ppb, respectively, and with a precision of $\pm$0.25 ppb and $\pm$0.2 ppb, respectively. Similarly, the deployments at

	Sensors					
	DD1	**DM2**	DD3	MM5	**MD6**	**MD7**
Jun	D	D	D	M	M	M
Oct	D	M	D	M	D	D

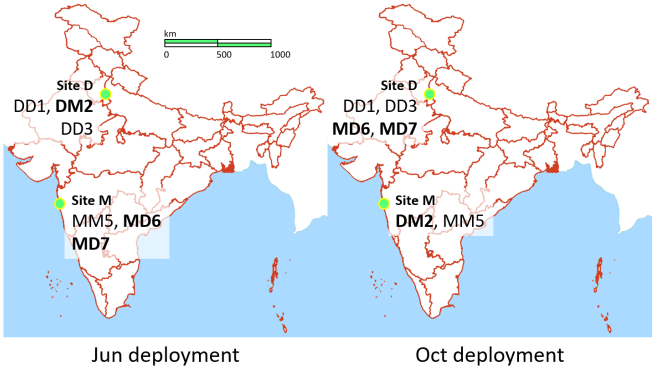

Figure 3. A schematic showing the deployment of the six LCAQ sensors across site D and site M during the two deployments. The sensors subjected to the *swap-out* experiment are presented in bold. The outlines of the Indian states in red was taken from QGIS3.4 Madeira with other highlights (e.g. for oceans) and markers being added separately.

site M had Teledyne T200 and T400 reference-grade monitors installed. These also have a UV photometric analyzer to measure O_3 levels and use chemiluminescence to measure NO_2 concentrations with lowest detectable limits for O_3 and NO_2 of 0.4 ppb and 0.2 ppb respectively and a precision of $\pm$0.2 ppb and $\pm$0.1 ppb respectively. For every deployment, the reference monitors and the AQ sensors were time-synchronized, with the 1 minute interval data averaged across 15 minute intervals for all site M deployments since the site M reference monitors gave data at 15 minute intervals.

2.3 Deployment Details

A total of four field co-location deployments, two each at sites D and M, were evaluated to characterize the calibration of the low-cost sensors during two seasons of 2019. The two field deployments at site D were carried out from 27th Jun–6th Aug 2019 (7 weeks) and 4th Oct–27th Oct 2019 (3 weeks). The two field deployments at site M, on the other hand, were carried out from 22nd Jun–21st Aug 2019 (10 weeks), and 4th Oct–27th Oct 2019 (3 weeks) respectively. For sake of convenience, we will refer to both deployments that commenced in the month of June 2019 (resp. October 2019) as *Jun* (resp. *Oct*) deployments even though the dates of both Jun deployments do not exactly coincide.

A total of six low-cost SATVAM LCAQ sensors were deployed at these two sites. We assign these sensors a unique numerical identifier and a name that describes its deployment pattern. The name of a sensor is of the form `XYn` where `X` (resp `Y`) indicates the site at which the sensor was deployed during the Jun (resp Oct) deployment and `n` denotes its unique numerical identifier. Fig. 3 outlines the deployment patterns for the six sensors DD1, DM2, DD3, MM5, MD6, and MD7.

Table 1. Samples of the raw data collected from the DM2(Jun) and MM5(Oct) datasets. The last column indicates whether data from that time-stamp was used in the analysis or not. Note that DM2(Jun) data, coming from site D, has samples at 1 minute intervals whereas MM5(Oct) data, coming from site M, has samples at 15 minute intervals. The raw voltage values (no2op1, no2op2, oxop1, oxop2) offered by the LCAQ sensor are always integer valued, as indicated in the DM2(Jun) data. However, for site M deployments, due to averaging, the effective voltage values used in the dataset may be fractional, as indicated in the MM5(Oct) data. The symbol × indicates missing values. A bold font indicates invalid values.

	DM2(Jun)										
Time-stamp	O3	NO2	T	RH	no2op1	no2op2	oxop1	oxop2	no2diff	oxdiff	Valid?
29-06 04:21	19.82	20.49	32.7	54.6	212	231	242	209	-19	33	Yes
30-06 08:02	46.363	**-0.359**	36.8	39.6	184	221	234	201	-37	33	No
01-07 04:02	24.38	14.73	32.5	69.7	×	×	×	×	×	×	No
08-07 07:51	**-0.035**	17.147	31.5	97.8	209	238	231	216	-29	15	No
	MM5(Oct)										
Time-stamp	O3	NO2	T	RH	no2op1	no2op2	oxop1	oxop2	no2diff	oxdiff	Valid?
19-10 05:45	×	×	×	×	160.46	188.31	158.31	172.38	-27.85	-14.07	No
19-10 07:15	5.55	11.52	41.47	99.9	170.4	197.2	167.6	181.93	-26.8	-14.33	Yes
20-10 10:45	×	×	28.52	99.9	121.8	154.0	119.3	135.3	-32.2	-16.0	No
22-10 18:30	8.33	10.91	27.87	99.9	143.2	172.3	146.2	155.47	-29.1	-9.27	Yes

Swap-out Experiment. As Fig. 3 indicates, three sensors were swapped with the other site across the two deployments. Specifically, for the Oct deployment, DM2 was shifted from site D to M and MD6 and MD7 were shifted from site M to D.

Sensor Malfunction. Our experiment actually involved a total of seven sensors being deployed. The seventh sensor, named DM4, was supposed to be swapped from site D to site M. However, the onboard RH and temperature sensors for this sensor were non-functional for the entire duration of the Jun deployment and frequently so for the Oct deployment as well. For this reason, this sensor was excluded from our study altogether. To avoid confusion, in the rest of the manuscript (e.g. the abstract, Fig. 3, etc) we report only six sensors of which three were a part of the swapout experiment.

3 Data Analysis Setup

All experiments were conducted on a commodity laptop with an Intel Core i7 CPU with 2.70GHz frequency, 8GB of system memory and running an Ubuntu 18.04.4 LTS operating system. Standard off-the-shelf machine learning and statistical analysis packages such as numpy, sklearn, scipy and metric–learn were used to implement the calibration algorithms.

Raw Datasets and Features. The six sensors across the Jun and Oct deployments, gave us a total of 12 datasets. We refer to each dataset by mentioning the sensor name and the deployment. For example, the dataset DM2(Oct) contains data from the October deployment at site M of the sensor DM2. Each dataset is represented as a collection of eight time series for which each time stamp is represented as an 8-tuple (O3, NO2, RH, T, no2op1, no2op2, oxop1, oxop2) giving us, respectively, the

reference values for O_3 and NO_2 (in ppb), relative humidity (in %) and temperature (in °C) values, and voltage readings (in mV) from the two electrodes present in each of the two gas sensors. These readings represent working (no2op1 and oxop1) and auxiliary (no2op2 and oxop2) electrode potentials for these sensors. We note that RH and T values in all our experiments were obtained from DHT22 sensors in the LCAQ sensors and not from the reference monitors. This was done to ensure that the calibration models, once trained, could perform predictions using data available from the LCAQ sensor alone and not rely on data from a reference monitor. For site D, both the LCAQ sensor as well as the reference monitor data was available at 1 minute intervals. However for site M, since reference monitor data was only available at 15 minute intervals, CAQ sensor data was averaged over 15 minute intervals.

Data Cleanup. Time-stamps from the LCAQ sensors were aligned to those from the reference monitors. For several time-stamps, we found that either the sensor or reference monitors presented with one or more missing or spurious values (see Table 1 for examples). Spurious values included the following cases: a) a reference value for O_3 or NO_2 of > 200 ppb or < 0 ppb (the reference monitors sometimes offered negative readings when powering up and under anomalous operating conditions e.g. condensation at the inlet), b) a sensor temperature reading of > 50 °C or < 1 °C, c) an sensor RH level of > 100 % or < 1 %, and d) a sensor voltage reading (either of no2op1, no2op2, oxop1, oxop2) of > 400 mV or < 1 mV. These errors are possibly due to electronic noise in the devices. All time-stamps with even one spurious or missing value were considered invalid and removed. Across all 12 datasets, an average of 52% of the time-stamps were removed as a result. However, since site D (resp. site M) offered timestamps at 1 minute (resp. 15 minute) intervals i.e. 60 (resp 4) timestamps every hour, at least one timestamp (frequently several) were found still valid every hour in most cases. Thus, the valid timestamps could still accurately track diurnal changes in AQ parameters. The datasets from Jun (resp. Oct) deployments at site D offered an average of 33753 (resp. 9548) valid time-stamps. The datasets from Jun (resp. Oct) deployments in site M offered an average of 2462 (resp. 1062) valid time-stamps. As expected, site D that had data at 1 minute intervals offered more time-stamps than site M that had data at 15 minute intervals. For both sites, more data is available for the Jun deployment (that lasted longer) than the Oct deployment.

3.1 Data Augmentation and Derived Dataset Creation

For each of the 12 datasets, apart from the six data features provided by the LCAQ sensors, we included two augmented features, calculated as shown below

$$\text{no2diff} = \text{no2op1} - \text{no2op2}$$

$$\text{oxdiff} = \text{oxop1} - \text{oxop2}$$

We found that having these augmented features, albeit simple linear combinations of raw features, offered our calibration models a predictive advantage. The *augmented* datasets created this way represented each time-stamp as a vector of 8 feature values (RH, T, no2op1, no2op2, oxop1, oxop2, no2diff, oxdiff), apart from the reference values of O_3 and NO_2.

3.1.1 Train–Test Splits

Each of the 12 datasets was split in a 70:30 ratio to obtain a train-test split. 10 such splits were independently generated for each dataset. All calibration algorithms were offered the same train-test splits. For algorithms that required hyperparameter tuning, a randomly chosen set of 30% of the training data points in that split were used as a held out validation set. All features were normalized to improve the conditioning of the calibration problems. This was done by calculating the mean and standard deviation for each of the 8 features on the training portion of a split, and then mean centering and dividing by the standard deviation all time-stamps in both training and testing portion of that split. An exception was made for the Alphasense calibration models, which required raw voltage values. However, reference values were never normalized in any way.

3.2 Derived Datasets

In order to study the effect of data frequency (how frequently do we record data e.g. 1 minute, 15 minute), data volume (total number of time-stamps used for training), and data diversity (data collected across seasons or sites) on the calibration performance, we created several *derived* datasets as well. All these datasets contained the augmented features.

1. **Temporally Averaged Datasets**: We took the two datasets DD1(Jun) and DM2(Jun) and created four datasets out of each of them by averaging the sensor and reference monitor values at 5 minute, 15 minute, 30 minute and 60 minute intervals. These datasets were named by affixing the averaging interval size to the dataset name, for example DD1(Jun)-AVG5 for the dataset created out of DD1(Jun) with 5 minute averaging, DM2(Jun)-AVG30 for the dataset created out of DM2(Jun) with 30 minute averaging, etc.

2. **Sub-sampled Datasets**: To view the effect of having less training data on calibration performance, we created *sub-sampled* versions of both these datasets by sampling a random set of 2500 time-stamps from the training portion of the DD1(Jun) and DM2(Jun) datasets to get the datasets DD1(Jun)-SMALL and DM2(Jun)-SMALL.

3. **Aggregated Datasets**: Next, we created new datasets by clubbing together data for a sensor across the two deployments. This was done to the data from the sensors DD1, MM5, DM2 and MD6. For example, if we consider the sensor DD1, then the datasets DD1(Jun) and DD1(Oct) were combined to create the dataset DD1(Jun-Oct).

Investigating Impact of Diversity in Data. The aggregated datasets are meant to help us study how calibration algorithms perform under seasonally and spatially diverse data. For example, the aggregated datasets DD1(Jun-Oct) and MM5(Jun-Oct) include data that is seasonally diverse but not spatially diverse (since these two sensors were located at the same site for both deployments). On the other hand, the aggregated datasets DM2(Jun-Oct) and MD6(Jun-Oct) include data that is diverse both seasonally as well as spatially (since these two sensors were a part of the swapout experiment). At this point, it is tempting to ask whether aggregated datasets that are diverse spatially but not seasonally diverse can be created as well. Although the prospect of investigating the effect of spatial diversity alone (without bringing seasonal diversity into account) is interesting, this would require aggregating data from two distinct sensors since no sensor was located at both sites during a deployment.

This presents an issue since the various onboard sensors in these LCAQ devices, e.g. RH and temperature sensors, do not present good agreement across devices. Thus, some form of cross-device calibration would have been required which is an interesting but challenging task in itself. This is an encouraging direction for future work but not considered in this study.

3.2.1 Performance Evaluation

The performance of calibration algorithms was assessed using standard error metrics and statistical hypothesis testing.

Error Metrics: calibration performance was measured using four popular metrics: mean averaged error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE), and the coefficient of determination (R^2) (please see the supplementary material for detailed expressions of these metrics).

Statistical Hypothesis Tests: in order to compare the performance of different calibration algorithms on a given dataset (to find out the best performing algorithm), or compare the performance of the same algorithm on different datasets (to find out the effect of data characteristics on calibration performance), we performed paired and unpaired two-sample tests, respectively. Our null hypothesis in all such tests proposed that the absolute errors offered in the two cases considered are distributed identically. The test was applied and if the null hypothesis was rejected with sufficient confidence (an α value of 0.05 was used as the standard to reject the null hypotheses), then a winner was simultaneously identified. Although the Student's t-test is more popular, it assumes that the underlying distributions are normal. However, an application of the Shapiro-Wilk test (Shapiro and Wilk, 1965) to our absolute error values rejected the normal hypothesis with high confidence. Thus, we chose the non-parametric Wilcoxon signed-rank test (Wilcoxon, 1945) when comparing two algorithms on the same dataset, and its unpaired variant, the Mann-Whitney U-test (Mann and Whitney, 1947) for comparing the same algorithm on two different datasets. These tests do not make any assumption on the underlying distribution of the errors and are well-suited for our data.

4 Baseline and Proposed Calibration Models

Our study considered a large number of parametric and non-parametric calibration techniques as baseline algorithms. Table 2 provides a glossary of all the algorithms including their acronyms and brief descriptions. Detailed descriptions of all these algorithms is provided in the supplementary material. Among parametric algorithms, we considered the Alphasense models (AS1-AS4) supplied by the manufacturers of the gas sensors, linear models based on least-squares (LS and LS(MIN)) and sparse recovery (LASSO). Among non-parametric algorithms, we considered regression trees (RT), kernel-ridge regression (KRR), the Nystroem method for accelerating KRR, the Nadaraya Watson estimator (NW), and various local algorithms based on the k-nearest neighbors principle (KNN, KNN-D). In this section we give a self-contained description of our proposed algorithms KNN(ML) and KNN-D(ML).

Notation: For every time-stamp t, the vector $\boldsymbol{x}^t \in \mathbb{R}^8$ denotes the 8-dimensional vector of signals recorded by the LCAQ sensors for that time-stamp, namely (RH, T, no2op1, no2op2, oxop1, oxop2, no2diff, oxdiff), while the vector $\boldsymbol{y}^t \in \mathbb{R}^2$ will denote the 2-tuple of the reference values of O_3 and NO_2 for that time step. However, this notation is unnecessarily cumbersome since we will build separate calibration models for O_3 and NO_2. Thus, to simplify the notation, we will instead use $y^t \in \mathbb{R}$ to

Table 2. Glossary of baseline and proposed calibration algorithms used in our study with their acronyms and brief descriptions. The KNN(ML) and KNN-D(ML) algorithms are proposed in this paper. Please see the supplementary material for details.

Parametric Algorithms		Non-parametric Algorithms		Non-parametric KNN-style Algorithms	
AS1, AS2	Alphasense models (from gas sensor manufacturer)	RT	Regression Tree	KNN	k-nearest Neighors
AS3, AS4		KRR	Kenel Ridge Regression	KNN-D	Distance weighted KNN
LS	Least Squares Regression	NYS	Nystroem Method	KNN(ML)*	KNN (learnt metric)
LS(MIN)	LS with reduced features	NW(ML)	Nadaraya Watson (learnt metric)	KNN-D(ML)*	KNN-D (learnt metric)
LASSO	Sparse Regression				*proposed in this paper

denote the reference value of the gas being considered (either O_3 or NO_2). The goal of calibration will then be to learn a real valued function $f : \mathbb{R}^8 \to \mathbb{R}$ such that $f(\boldsymbol{x}^t) \approx y^t$ for all time-stamps t (the exact error being measured using metrics such as MAE, MAPE, etc described in Sect. 3.2.1). Thus, we will learn two functions, say f_{NO_2} and f_{O_3} to calibrate for NO_2 and O_3 concentrations respectively. Since several of our calibration algorithms will involve the use of some statistical estimation or machine learning algorithm, we will let N (resp. n) denote the number of training (resp. testing) points for a given dataset and split thereof. Thus, we will let $\{(\boldsymbol{x}^t, y^t)\}_{t=1}^N$ denote the training set for that dataset and split with $\boldsymbol{x}^t \in \mathbb{R}^8$ and $y^t \in \mathbb{R}$.

4.1 Proposed Method: Distance-weighed KNN with a Learnt Metric

Our proposed algorithm is a local, non-parametric algorithm that uses a learnt metric. Below we describe the design of this method and reasons behind these design choices.

Non-parametric estimators for Calibration. The simplest example of a non-parametric estimator is the KNN (k nearest neighbors) algorithm that predicts on a test point, the average reference value in the k nearest training points. Other examples (please see the supplementary material for details) include kernel ridge regression (KRR) and the Nadaraya-Watson (NW) estimator. Non-parametric estimators are well-studied and known to be asymptotically universal which guarantees their ability to accurately model complex patterns which motivated their choice. These models can also be brittle Hagan et al. (2019) when used in unseen operating conditions but Sec. 5.2 shows that our proposed algorithm performs comparably to parametric algorithms when generalizing to unseen conditions, but offers far more improvements when given additional data.

Metric Learning for KNN Calibration. As mentioned above, the KNN algorithm uses the closest neighbors to compute its output. To do this, it needs a notion of distance, specifically a *metric*, to compute closeness. The default and most common choice for a metric is the Euclidean distance which gives equal importance to all 8 dimensions when calculating distances between two points say $\boldsymbol{x}^1, \boldsymbol{x}^2 \in \mathbb{R}^8$. However, our experiments in Sect. 5 will show that certain features, e.g. RH and T, seem to have a significant influence on calibration performance. Thus, it is unclear how much emphasis should RH and T receive, as compared to other features such as voltage values e.g. oxop1 while calculating distances between two points. The technique of *metric learning* (Weinberger and Saul, 2009) offers a solution in this respect by learning a customized *Mahalanobis metric* metric that can be used instead of the generic Euclidean metric. A Mahalanobis metric is characterized by a positive semi-

Algorithm 1 The proposed KNN-D(ML) algorithm for distance weighted KNN calibration with a learnt metric.

Require: training data points $\{(\boldsymbol{x}^t, y^t)\}_{t=1}^N$, neighborhood size k

Ensure: a prediction from the KNN-D(ML) model

$\boldsymbol{\Sigma} \leftarrow$ use training data points to learn a Mahalanobis metric using the technique from (Weinberger and Tesauro, 2007)

Receive feature vector $\tilde{\boldsymbol{x}} \in \mathbb{R}^8$ for a test data point

Find the k training data points (say $i_1, \ldots, i_k$) that are closest to $\tilde{\boldsymbol{x}}$ in terms of the learnt Mahalanobis distance $d^{\text{Maha}}(\cdot, \cdot; \boldsymbol{\Sigma})$

For all $l = 1 \ldots k$, let $\alpha^l = (d^{\text{Maha}}(\tilde{\boldsymbol{x}}, \boldsymbol{x}^{i_l}; \boldsymbol{\Sigma}))^{-1}$

$\hat{y} = \dfrac{\sum_{l=1}^k \alpha^l \cdot y^{t_l}}{\sum_{l=1}^k \alpha^l}$

return Calibrated value $\hat{y}$ for the test data point

definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{8\times 8}$ and calculates the distance between any two points as follows

$$d^{\text{Maha}}(\boldsymbol{x}^1, \boldsymbol{x}^2; \boldsymbol{\Sigma}) = \sqrt{(\boldsymbol{x}^1 - \boldsymbol{x}^2)^\top \boldsymbol{\Sigma} (\boldsymbol{x}^1 - \boldsymbol{x}^2)}$$

Note that the Mahalanobis metric recovers the Euclidean metric when $\boldsymbol{\Sigma} = I_8$ is the identity matrix. Now, whereas metric learning for KNN is popular for classification problems, it is uncommon for calibration and regression problems. This is due to regression problems lacking of a small number of "classes". To overcome this problem, we note that other non-parametric calibration algorithms such as NW and KRR also utilize a metric indirectly (please see the supplementary material) and there does exist a technique to learn a Mahalanobis metric to be used alongwith the NW algorithm (Weinberger and Tesauro, 2007). This allows us to adopt a *two-stage* algorithm that first learns a Mahalanobis metric suited for the NW algorithm and then uses it to perform KNN-style calibration. Algorithm 1 describes the resulting KNN-D(ML) algorithm.

5 Results and Discussion

The goals of using low-cost AQ monitoring sensors vary widely. This section critically assesses a wide variety of calibration models. First we look at the performance of the algorithms on individual datasets i.e. when looking at data within a site and within a season. Next, we look at derived datasets (Sec 3.2) which look at the effect of data volume, data averaging and data diversity on calibration performance.

5.1 Effect of Calibration Model on Calibration Performance

We compare the performance of calibration algorithms introduced in Sect. 4. Given the vast number of algorithms, we execute a sort of tournament where divide algorithms into small families, decide the winner within a family and then compare the winners across families. The detailed per-family comparisons are available in the supplementary material and only summarized here. We use the Wilcoxon paired two sample test (see Sect. 3.2.1) to compare two calibration algorithms on the same dataset. However, for visual inspection, we also provide *violin plots* of the absolute errors offered by the algorithms. We refer the reader to the supplementary material for some pointers on how to interpret violin plots.

5.1.1 Interpreting the Two-sample Tests

We refer the reader to Table 2 for a glossary of algorithm names and abbreviations. As mentioned earlier, we used the paired Wilcoxon signed ranked test to compare two algorithms on the same dataset. Given that there are 12 datasets and 10 splits for each dataset, for ease of comprehension, we provide globally averaged statistics of wins scored by an algorithm over another. For example, say we wish to compare RT and KRR as done in Tab 3. We perform the test for each individual dataset and split. For each test, we either get a win for RT (in which case RT gets a +1 score and KRR gets 0), or a win for KRR (in which case KRR gets a +1 score and RT gets 0) or else the null hypothesis is not refuted (in which case both get 0). The average of these scores is then shown. For example, in Tab 3 (left), row 3 column 2 records a value of 0.63 implying that in 63% of these tests, KRR won over RT in case of O_3 calibration, whereas row 2 column 3 records a value of 0.22 implying that in 22% of the tests, RT won over KRR. In the balance (1 - 0.63 - 0.22 = 0.15) i.e. 15% of the tests, neither algorithm could be declared a winner.

5.1.2 Intra-family Comparison of Calibration Models

We divide the calibration algorithms (see Table 2 for a glossary) into four families: 1) the Alphasense family (AS1, AS2, AS3, AS4), 2) linear parametric models (LS, LS(MIN) and LASSO), 3) kernel regression models (KRR, NYS), and 4) KNN-style algorithms (KNN, KNN-D, NW(ML), KNN(ML), KNN-D(ML)). We included the Nadaraya-Watson (NW) algorithm in the fourth family since it was used alongwith metric learning, as well as because as explained in the supplementary material, the NW algorithm behaves like a "smoothed" version of KNN algorithm. The winners within these families are described below.

1. **Alphasense**: All four Alphasense algorithms exhibit extremely poor performance across all metrics on all datasets, offering extremely high MAE and low R^2 values. This is corroborated by previous studies (Lewis and Edwards, 2016; Jiao et al., 2016; Simmhan et al., 2019).

2. **Linear Parametric**: Among the linear parametric algorithms, LS was found to offer the best performance.

3. **Kernel Regression**: The Nystroem method NYS was confirmed to be an accurate but accelerated approximation for KRR with the acceleration being higher for larger datasets.

4. **KNN and Metric Learning Models**: Among the KNN family of algorithms, the distance weighted KNN algorithm that uses a learnt metric i.e. KNN-D(ML) was found to offer the best accuracies across all datasets and splits.

5.1.3 Global Comparison of Comparison Models

We took the best algorithms from all the families (except Alphasense models that gave extremely poor performance) and regression trees (RT) and performed a head-to-head comparison to assess the winner. The two-sample tests (Table 3) as well as violin plots (Fig. 4) indicate that KNN-D(ML) algorithm continues to emerge as the overall winner. Table 4 additionally establishes that KNN-D(ML) can be upto 8 - 20 percentage points better than classical non-parametric algorithms such as KRR in terms of R^2 coefficient. The improvement is much more prominent for NO_2 calibration which seems to be more challenging

Table 3. Results of the pairwise Wilcoxon signed rank tests across all model types (see Sect. 5.1.1 for a key). KNN-D(ML) beats every other algorithm comprehensively and is scarcely ever beaten. (with the exception of NW(ML) which it still beats 58% of the time on NO_2 and 62% on O_3) The overall ranking of the algorithms is indicated to be KNN-D(ML) > NW(ML) > KRR > RT > LS.

NO_2					
	LS	RT	KRR	NW(ML)	KNN-D(ML)
LS	0	0	0	0	0
RT	0.97	0	0.38	0.16	0
KRR	1	0.4	0	0	0
NW(ML)	1	0.75	1	0	0.07
KNN-D(ML)	1	1	1	0.58	0

O_3					
	LS	RT	KRR	NW(ML)	KNN-D(ML)
LS	0	0.01	0	0	0
RT	0.83	0	0.22	0	0
KRR	1	0.63	0	0.01	0
NW(ML)	1	0.97	0.96	0	0.02
KNN-D(ML)	1	1	0.97	0.62	0

Table 4. A comparison of algorithms across families on the DD1 and MM5 datasets across seasons with respect to the R^2 metric. All values are averaged across 10 splits. Bold values indicate the best performing algorithm in terms of mean statistics.

O_3					NO_2				
	DD1		MM5			DD1		MM5	
	Jun	Oct	Jun	Oct		Jun	Oct	Jun	Oct
LS	0.843±0.006	0.969±0.002	0.334±0.035	0.846±0.019	LS	0.341±0.013	0.623±0.005	0.375±0.049	0.321±0.026
RT	0.852±0.005	0.971±0.003	0.488±0.071	0.393±0.224	RT	0.674±0.015	0.913±0.014	0.487±0.064	0.358±0.087
KRR	0.885±0.005	0.987±0.002	0.719±0.037	0.935±0.02	KRR	0.608±0.019	0.957±0.003	0.728±0.034	0.673±0.059
NW(ML)	0.895±0.004	0.988±0.001	**0.74 ±0.038**	**0.943±0.026**	NW(ML)	0.717±0.017	**0.97 ±0.003**	**0.771±0.026**	**0.751±0.039**
KNN-D(ML)	**0.923±0.003**	**0.99 ±0.001**	**0.744±0.043**	**0.943±0.025**	KNN-D(ML)	**0.819±0.015**	**0.977±0.002**	0.759±0.022	**0.751±0.043**

as compared to O_3 calibration. Fig. 5 presents two cases where the KNN-D(ML) models offer excellent agreement with the reference monitors across significant spans of time.

Analyzing High Error Patterns. Having analyzed the calibration performance of various algorithms including KNN-D(ML), it is interesting to note under what conditions do these algorithms incur high error. Non-parametric algorithms such as RT and KNN-D(ML) are expected to do well in the presence of good amounts of diverse data. Fig 6 confirms this by classifying timestamps into various bins according to weather conditions. KNN-D(ML) and RT do offer high average error mostly in bins where there were less training points. Fig 6 also confirms a positive correlation between high concentrations and higher error although this effect is more pronounced for LS than KNN-D(ML).

5.2 Effect of Data Preparation on Calibration Performance

We now critically assess the robustness of these calibration models, as well as identify the effect of other factors, such as temporal averaging of raw data, total amount of data available for training, and diversity in training data. We note that some of these studies were made possible only because the experimental setup enabled us to have access to sensors that did not change their deployment sites, as well as those that did change their deployment site during the swap-out experiment.

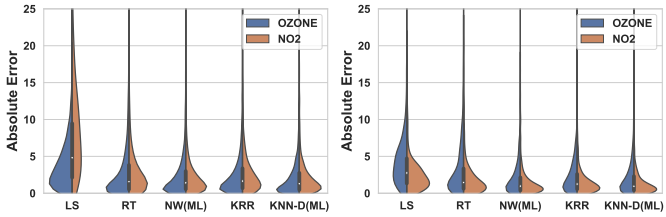

Figure 4. The violin plots on the left and right depict the distribution of absolute errors incurred by various models on respectively, the DD1(Oct) and MM5(Jun) datasets. KNN-D(ML) offers visibly superior performance than several other algorithms such as LS and RT.

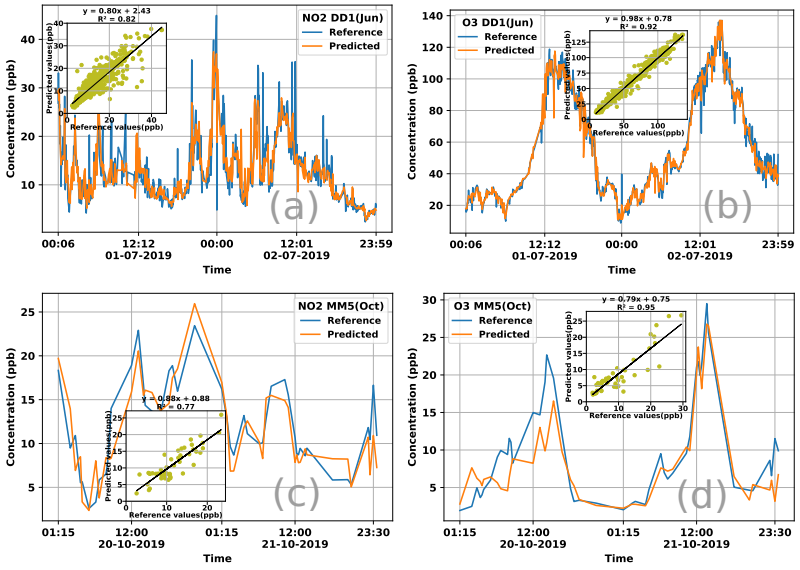

Figure 5. Time series for a duration of 24 hours of the reference values and those predicted by the KNN-D(ML) algorithm for NO_2 and O_3 concentration using data from the DD1 and MM5 sensors. The legend of each plot notes the gas for which calibration is being reported, the deployment season, as well as the sensor from which data was used to perform the calibration. Each plot also contains a scatter plot as an inset showing the correlation between the reference and predicted values of the concentrations. For both deployments and both gases, KNN-D(ML) can be seen to offer excellent calibration and agreement with the FRM-grade monitor.

5.2.1 Some Observations on Original Datasets

The performance of KNN-D(ML) on the original datasets itself gives us indications on how various data preparation methods can affect calibration performance. Table 4 shows us that in most cases, the calibration performance is better (with higher R^2) for O_3 than NO_2. This is another indication that NO_2 calibration is more challenging that O_3 calibration. Moreover, for both gases and in both seasons, we see site D offering a better performance than site M. This difference is more prominent for NO_2

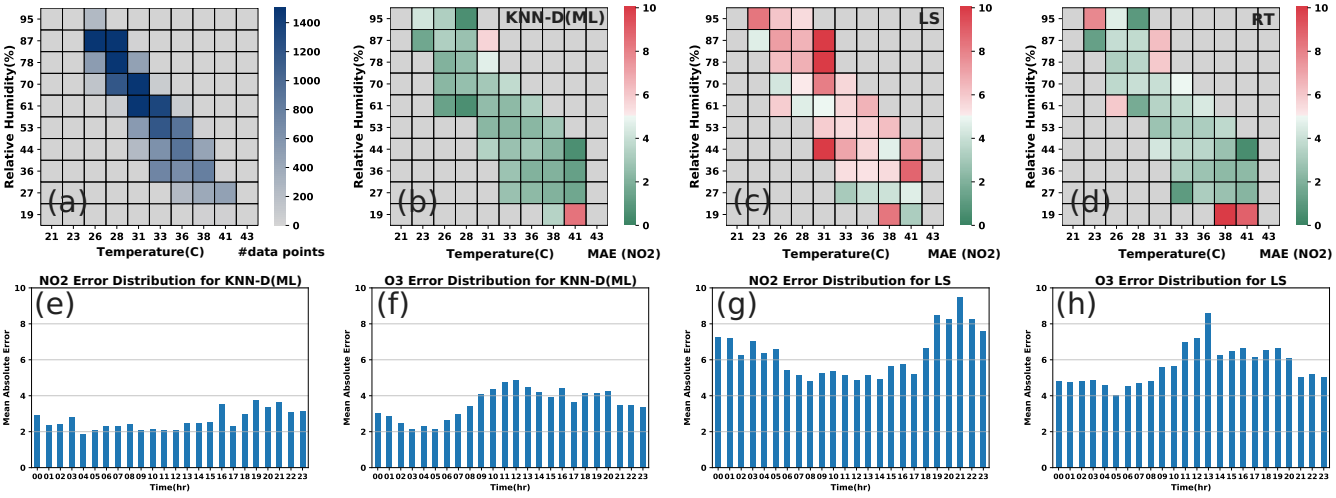

Figure 6. Analyzing error distributions of LS, KNN-D(ML, RT.. Fig 6(a) shows the number of training data points in various weather condition bins. Figs 6(b,c,d) show the MAE for NO_2 calibration offered by the algorithms in those same bins. Non-parametric algorithms such as KNN-D(ML) and RT offer poor performance (high MAE) mostly in bins that had less training data. No such pattern is observable for LS. Figs 6(e,f,g,h) show the diurnal variation of MAE for KNN-D(ML) and LS at various times of day. O_3 errors exhibit a diurnal trend of being higher (more so for LS than KNN-D(ML)) during daylight hours when O_3 levels are high. No such trend is visible for NO_2.

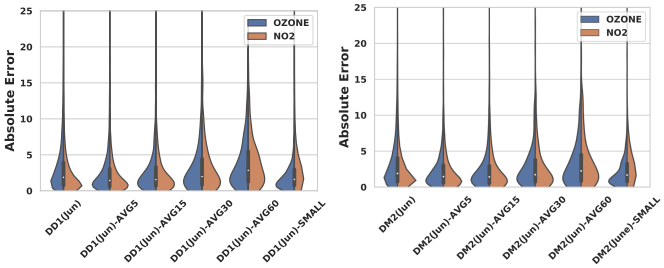

Figure 7. Effect of temporal data averaging, and lack of data on the calibration performance of the KNN-D(ML) algorithm on temporally averaged and sub-sampled versions of the DD1(Jun) and DM2(Jun) datasets. Notice the visible deterioration in the performance of the algorithm when aggressive temporal averaging, e.g. across 30 minute windows, is performed. NO_2 calibration performance seems to be impacted more adversely by lack of enough training data or aggressive averaging than O_3 calibration.

than for O_3. This indicates that paucity of data and temporal averaging may be affecting calibration performance negatively, as well as that O_3 calibration might be less sensitive to these factors than NO_2 calibration.

5.2.2 Effect of Temporal Data Averaging

Recall that data from sensors deployed at site M had to be averaged over 15 minute intervals to align them with the reference monitor timestamps. To see what effect such averaging has on calibration performance, we use the temporally averaged datasets

Table 5. Results of the pairwise Mann-Whitney U tests on the performance of KNN-D(ML) across temporally averaged versions of the DD1 dataset (see Sect. 5.1.1 for a key). The dataset names are abbreviated, e.g. DD1(Jun)-AVG5 is referred to as simply AVG5. Results are reported over a single split. AVG5 wins over any other level of averaging and clarifies that mild temporal averaging (e.g. over 5 minute windows) boosts calibration performance, whereas aggressive averaging e.g. 60 minute averaging in AVG60, degrades performance.

	O_3						NO_2				
	DD1(Jun)	AVG5	AVG15	AVG30	AVG60		DD1(Jun)	AVG5	AVG15	AVG30	AVG60
DD1(Jun)	0	0	0	0	0	DD1(Jun)	0	0	0	1	1
AVG5	1	0	1	1	1	AVG5	1	0	1	1	1
AVG15	1	0	0	1	1	AVG15	0	0	0	1	1
AVG30	1	0	0	0	1	AVG30	0	0	0	0	1
AVG60	0	0	0	0	0	AVG60	0	0	0	0	0

(see Sect. 3.1). Fig. 7 presents the results of applying the KNN-D(ML) algorithm on data that is not averaged at all (i.e. 1 minute interval timestamps), as well as data that is averaged at 5, 15, 30 and 60 minute intervals. The performance for 30 and 60 minute averaged datasets is visibly inferior that that for the non-averaged dataset as indicated by the violin plots. This leads us to conclude that excessive averaging can erode the diversity of data and hamper effective calibration. To distinguish among the other temporally averaged datasets for which visual inspection is not satisfactory, we also performed the unpaired Mann-Whitney U test, the results for which are shown in Tab 5. The results are striking in that they reveal that moderate averaging, for example at 5 minute intervals, seems to benefit calibration performance. However, this benefit is quickly lost if the averaging window is increased much further at which point, performance almost always suffers.

5.2.3 Effect of Data Paucity

Since temporal averaging also decreases the amount of data as a side-effect, in order to tease these two effects apart, we also considered the sub-sampled versions of these datasets (see Sect. 3.1). Fig. 7 also shows that reducing the amount of training data has an appreciable negative impact on calibration performance. However, NO_2 calibration performance seems to be impacted more adversely by lack of enough training data or aggressive averaging than O_3 calibration.

5.2.4 The Swapout Experiment: Effect of Data Diversity

Tab 6 describes an experiment wherein we took the KNN-D(ML) model trained on one dataset and used it to make predictions on another dataset. To avoid bringing in too many variables such as cross-device calibration (see Sec 3.2), this was done only in cases where both datasets belonged to the same sensor but for different deployments. Without exception, such *transfers* led to a drop in performance. We confirmed that this was true not just for non-parametric methods such as KNN-D(ML) but also parametric models like LS. This is to be expected since the sites D and M experience largely non-overlapping ranges of RH and T across the two deployments. We refer the reader to the supplementary material for a plot of RH and T values experienced at both sites in both deployments. Thus, it is not surprising that the models performed poorly when faced with unseen RH and

Table 6. A demonstration of the impact of data diversity and data volume on calibration performance. All values are averaged across 10 splits. The results for LS diverged on some of the datasets on a few splits and those splits were removed while averaging to give LS an added advantage. Bold values indicate the better performing algorithm. The first two rows present the performance of the KNN-D(ML) and LS calibration models when tested on data for a different season (deployment) but in the same site. This was done for the DD1 and MM5 sensors that did not participate in the swap-out experiment. The next two rows present the same, but for sensors DM2 and MD6 that did participate in the swap-out experiment and thus, their performance is being tested not only for a different season, but also a different city. The next four rows present the dramatic improvement in calibration performance once datasets are aggregated for these four sensors. NO_2 calibration is worse affected by these variations (average R^2 in first four rows being -3.69) than O_3 calibration (average R^2 in first four rows being -0.97).

	KNN-D(ML)				LS			
	O_3		NO_2		O_3		NO_2	
Train → Test	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2
DD1(Jun) → (Oct)	21.82	0.19	21.86	-0.64	**12.88**	**0.73**	**12.73**	**0.22**
MM5(Oct) → (Jun)	**8.33**	**-3.75**	**15.79**	**-12.28**	10.39	-4.83	17.06	-21.67
DM2(Jun) → (Oct)	13.04	0.41	9.05	-0.99	**9.36**	**0.68**	**5.95**	**0.1**
MD6(Jun) → (Oct)	**16.71**	**-0.72**	30.9	-0.85	21.12	-1.29	**25.67**	**-0.23**
DD1(Jun-Oct)	**3.3**	**0.956**	**2.6**	**0.924**	11.7	0.29	13.0	0.38
MM5(Jun-Oct)	**2.5**	**0.902**	**1.8**	**0.814**	4.28	0.32	5.51	0.67
DM2(Jun-Oct)	**3.7**	**0.916**	**2.8**	**0.800**	6.13	0.79	6.72	0.26
MD6(Jun-Oct)	**1.9**	**0.989**	**1.8**	**0.975**	7.01	0.71	6.36	0.91

T ranges. To verify that this is indeed the case, we ran the KNN-D(ML) algorithm on the aggregated datasets (see Sect. 3.1) which combine training sets from the two deployments of these sensors. Tab 6 confirms that once trained on these more diverse datasets, the algorithms resume offering good calibration performance on the entire (broadened) range of RH and T values. However, KNN-D(ML) is superior at exploiting the additional diversity in data than LS. We note that parametric models are expected to generalize better on unseen conditions than non-parametric models and indeed we observe this in some cases in Tab 6 where on DD1 and DM2 datasets, LS generalized better than KNN-D(ML). However, we also observe some cases such as MM5 and MD6 where KNN-D(ML) generalizes comparable to or better than LS.

6 Conclusions and Future Work

In this study we presented results of field deployments across two seasons of LCAQ sensors across two sites having diverse geographical, meteorological, and air pollution parameters. A unique feature of our deployment was the *swap-out* experiment wherein three of the six sensors were transported across sites in the two deployments. To perform highly accurate calibration of these sensors, we experimented with a wide variety of standard algorithms but found a novel method based on *metric learning* to offer the strongest results. A few key takeaways from our statistical analyses are:

1. Incorporating ambient RH and T, as well as the emphaugmented features oxdiff and noxdiff (see Sect. 3), into the calibration model improves calibration performance.

2. Non-parametric methods such as KNN offer the best performance on these calibration tasks. However, they stand to gain significantly through the use of metric learning techniques, which automatically learn the relative importance of each feature, as well as *hyper-local* variations such as distance-weighted KNN. These indicate that these calibration tasks operate in high variability conditions where local methods offer the best chance at capturing subtle trends.

3. Performing smoothing over raw time series data obtained from the sensors may help improve calibration performance but only if done over short windows. Very aggressive smoothing done over long windows is detrimental to performance.

4. Calibration models are data-hungry as well as diversity hungry. This is especially true of local methods like KNN variants. Offering these techniques limited amounts of data or even data that is limited in diversity of RH, T or concentration levels, may result in calibration models that generalize very poorly.

5. Although all calibration models see a decline in performance when tested in unseen operating conditions, calibration models for O_3 seem to be less sensitive than those for NO_2 calibration.

Our results offer encouraging options for using LCAQ sensors to complement CAAQMS in creating dense and portable monitoring networks. Avenues for future work include the study of long-term stability of electrochemical sensors and characterizing drift or deterioration patterns in these sensors and correcting for the same, and rapid calibration of these sensors that requires minimal collocation with a reference monitor.

Code and data availability. The code and data used in this study are available upon request to author Purushottam Kar (purushot@cse.iitk.ac.in).

Competing interests. Author Ronak Sutaria is the CEO of Respirer Living Sciences Pvt. Ltd. which builds and deploys low-cost sensor based air quality monitors with trade-name 'Atmos - Realtime Air Quality' monitoring sensor networks. Ronak Sutaria's involvement was primarily in the development of the air quality sensor monitors and the big data enabled application programming interfaces to access the temporal data but not in the data analysis. Author Brijesh Mishra, subsequent to the work presented in this paper, has joined the Respirer Living Sciences team. The authors declare no other competing interests.

Acknowledgements. This research has been supported under the Research Initiative for Real-time River Water and Air Quality Monitoring program funded by the Department of Science and Technology, Government of India, and Intel® and administered by the Indo-United States Science and Technology Forum (IUSSTF).

References

Akasiadis, C., Pitsilis, V., and Spyropoulos, C. D.: A multi-protocol IoT platform based on open-source frameworks, Sensors, 19, 4217, 2019.

Apte, J. S., Messier, K. P., Gani, S., Brauer, M., Kirchstetter, T. W., Lunden, M. M., Marshall, J. D., Portier, C. J., Vermeulen, R. C., and Hamburg, S. P.: High-resolution air pollution mapping with Google street view cars: exploiting big data, Environmental Science & Technology, 51, 6999–7008, 2017.

Arroyo, P., Herrero, J. L., Suárez, J. I., and Lozano, J.: Wireless sensor network combined with cloud computing for air quality monitoring, Sensors, 19, 691, 2019.

Baron, R. and Saffell, J.: Amperometric gas sensors as a low cost emerging technology platform for air quality monitoring applications: A review, ACS Sensors, 2, 1553–1566, 2017.

Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., and Bartonova, A.: Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?, Environment International, 99, 293–302, 2017.

Chowdhury, S., Dey, S., and Smith, K. R.: Ambient $\mathrm{PM}_{2.5}$ exposure and expected premature mortality to 2100 in India under climate change scenarios, Nature Communications, 9, 1–10, 2018.

Commodore, A., Wilson, S., Muhammad, O., Svendsen, E., and Pearce, J.: Community-based participatory research for the study of air pollution: a review of motivations, approaches, and outcomes, Environmental Monitoring and Assessment, 189, 378, 2017.

Cross, E. S., Williams, L. R., Lewis, D. K., Magoon, G. R., Onasch, T. B., Kaminsky, M. L., Worsnop, D. R., and Jayne, J. T.: Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements, Atmospheric Measurement Techniques, 10, 3575, 2017.

De Vito, S., Esposito, E., Salvato, M., Popoola, O., Formisano, F., Jones, R., and Di Francia, G.: Calibrating chemical multisensory devices for real world applications: An in-depth comparison of quantitative machine learning approaches, Sensors and Actuators B: Chemical, 255, 1191–1210, 2018.

Esposito, E., De Vito, S., Salvato, M., Bright, V., Jones, R., and Popoola, O.: Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems, Sensors and Actuators B: Chemical, 231, 701–713, 2016.

Fung, P. L.: Calibration of Atmospheric Measurements in Low-cost Sensors, in: Data Science for Natural Sciences Seminar (DSNS 2019), 2019.

Gabrys, J., Pritchard, H., and Barratt, B.: Just good enough data: Figuring data citizenships through air pollution sensing and data stories, Big Data & Society, 3, 1–14, 2016.

Garaga, R., Sahu, S. K., and Kota, S. H.: A review of air quality modeling studies in India: local and regional scale, Current Pollution Reports, 4, 59–73, 2018.

Gaur, A., Tripathi, S., Kanawade, V., Tare, V., and Shukla, S.: Four-year measurements of trace gases (SO 2, NO x, CO, and O 3) at an urban location, Kanpur, in northern India, Journal of Atmospheric Chemistry, 71, 283–301, 2014.

Gillooly, S. E., Zhou, Y., Vallarino, J., Chu, M. T., Michanowicz, D. R., Levy, J. I., and Adamkiewicz, G.: Development of an in-home, real-time air pollutant sensor platform and implications for community use, Environmental Pollution, 244, 440–450, 2019.

Hagan, D. H., Gani, S., Bhandari, S., Patel, K., Habib, G., Apte, J. S., Hildebrandt Ruiz, L., and Kroll, J. H.: Inferring Aerosol Sources from Low-Cost Air Quality Sensor Measurements: A Case Study in Delhi, India, Environmental Science & Technology Letters, 6, 467–472, 2019.

Hitchman, M., Cade, N., áKim Gibbs, T., and Hedley, N. M.: Study of the factors affecting mass transport in electrochemical gas sensors, Analyst, 122, 1411–1418, 1997.

Jiao, W., Hagler, G., Williams, R., Sharpe, R., Brown, R., Garver, D., Judge, R., Caudill, M., Rickard, J., Davis, M., et al.: Community Air Sensor Network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern United States, Atmospheric Measurement Techniques, 9, 5281–5292, 2016.

Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L., and Britter, R.: The rise of low-cost sensing for managing air pollution in cities, Environment International, 75, 199–205, 2015.

Landrigan, P. J., Fuller, R., Acosta, N. J., Adeyi, O., Arnold, R., Baldé, A. B., Bertollini, R., Bose-O'Reilly, S., Boufford, J. I., Breysse, P. N., et al.: The Lancet Commission on pollution and health, The Lancet, 391, 462–512, 2018.

Lewis, A. and Edwards, P.: Validate personal air-pollution sensors, Nature, 535, 29–31, 2016.

Malings, C., Tanzer, R., Hauryliuk, A., Kumar, S. P., Zimmerman, N., Kara, L. B., Presto, A. A., and Subramanian, R.: Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring, Atmospheric Measurement Techniques, 12, 903–920, 2019.

Mann, H. B. and Whitney, D. R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, Annals of Mathematical Statistics, 18, 50–60, 1947.

Masson, N., Piedrahita, R., and Hannigan, M.: Quantification method for electrolytic sensors in long-term monitoring of ambient air quality, Sensors, 15, 27 283–27 302, 2015.

Miskell, G., Salmond, J. A., and Williams, D. E.: Solution to the problem of calibration of low-cost air quality measurement sensors in networks, ACS Sensors, 3, 832–843, 2018.

Morawska, L., Thai, P. K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., Bedini, A., Chai, F., Christensen, B., Dunbabin, M., et al.: Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone?, Environment International, 116, 286–299, 2018.

Mueller, M., Meyer, J., and Hueglin, C.: Design of an ozone and nitrogen dioxide sensor unit and its long-term operation within a sensor network in the city of Zurich, Atmospheric Measurement Techniques, 10, 3783, 2017.

Pang, X., Shaw, M. D., Lewis, A. C., Carpenter, L. J., and Batchellier, T.: Electrochemical ozone sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air-quality monitoring, Sensors and Actuators B: Chemical, 240, 829–837, 2017.

Popoola, O. A., Carruthers, D., Lad, C., Bright, V. B., Mead, M. I., Stettler, M. E., Saffell, J. R., and Jones, R. L.: Use of networks of low cost air quality sensors to quantify air quality in urban settings, Atmospheric Environment, 194, 58–70, 2018.

Rai, A. C., Kumar, P., Pilla, F., Skouloudis, A. N., Di Sabatino, S., Ratti, C., Yasar, A., and Rickerby, D.: End-user perspective of low-cost sensors for outdoor air pollution monitoring, Science of The Total Environment, 607, 691–705, 2017.

Sahu, R., Dixit, K. K., Mishra, S., Kumar, P., Shukla, A. K., Sutaria, R., Tiwari, S., and Tripathi, S. N.: Validation of Low-Cost Sensors in Measuring Real-Time PM_{10} Concentrations at Two Sites in Delhi National Capital Region, Sensors, 20, 1347, 2020.

Schneider, P., Castell, N., Vogt, M., Dauge, F. R., Lahoz, W. A., and Bartonova, A.: Mapping urban air quality in near real-time using observations from low-cost sensors and model information, Environment International, 106, 234–247, 2017.

Shapiro, S. S. and Wilk, M.: An analysis of variance test for normality (complete samples), Biometrika, 52, 591–611, 1965.

Sharma, A., Mishra, B., Sutaria, R., and Zele, R.: Design and Development of Low-cost Wireless Sensor Device for Air Quality Networks, in: IEEE Region 10 Conference (TENCON), 2019.

Simmhan, Y., Nair, S., Monga, S., Sahu, R., Dixit, K., Sutaria, R., Mishra, B., Sharma, A., SVR, A., Hegde, M., Zele, R., and Tripathi, S. N.: SATVAM: Toward an IoT Cyber-infrastructure for Low-cost Urban Air Quality Monitoring, in: 15th IEEE International Conference on e-Science (eScience 2019), 2019.

Snyder, E. G., Watkins, T. H., Solomon, P. A., Thoma, E. D., Williams, R. W., Hagler, G. S., Shelow, D., Hindin, D. A., Kilaru, V. J., and Preuss, P. W.: The Changing Paradigm of Air Pollution Monitoring, Environ. Sci. Technol., 47, 11 369–11 377, 2013.

Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M., and Bonavitacola, F.: Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2, Sensors and Actuators B: Chemical, 238, 706–715, 2017.

Weinberger, K. Q. and Saul, L. K.: Distance Metric Learning for Large Margin Nearest Neighbor Classification, Journal of Machine Learning Research, 10, 207–244, 2009.

Weinberger, K. Q. and Tesauro, G.: Metric Learning for Kernel Regression, in: 11th International Conference on Artificial Intelligence and Statistics (AISTATS), 2007.

WHO: Ambient (outdoor) air pollution, WHO Fact Sheet, https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health. Accessed on 11 November 2019., 2018.

Wilcoxon, F.: Individual Comparisons by Ranking Methods, Biometrics Bulletin, 1, 80–83, 1945.

Williams, D. E.: Low Cost Sensor Networks: How Do We Know the Data Are Reliable?, ACS Sensors, 4, 2558–2565, 2019.

Zheng, T., Bergin, M. H., Sutaria, R., Tripathi, S. N., Caldow, R., and Carlson, D. E.: Gaussian process regression model for dynamically calibrating and surveilling a wireless low-cost particulate matter sensor network in Delhi., Atmospheric Measurement Techniques, 12, 2019.

Zimmerman, N., Presto, A. A., Kumar, S. P., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., and Subramanian, R.: A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring., Atmospheric Measurement Techniques, 11, 2018.