Cover Letter and Author Responses

Point-by-point Responses to Anonymous Referee #1

1. **Query/Comment**: *A revised manuscript needs to be much more focused.*
   Response: We thank the referee for this suggestion. A major revision to the manuscript has been prepared that offers a more focussed discussion.
   Edits to Manuscript: The revision tightens the presentation, and reorganizes portions to retain only those that impact the overall discussion of characterizing calibration models. Other portions e.g. Figure 7, has been moved to the supplementary material. We thank the referee for pointing out the inefficiencies in the structuring of the paper.
2. **Query/Comment**: *Abstract line 15 (performance improvements and missing reference)*
   Response: improvements are indeed in percentage points.
   Edits to Manuscript: In the revision, we have included numerical values in a separate table (a new Table 4) in the main paper which explicitly quantify the improvements. These numerical values are referred to in the abstract and the main text to establish the claimed improvement. The text in section 6.5 (5.2.1 in the revision) has also been modified to similarly refer to Table 4 values to substantiate its claims.
3. **Query/Comment**: *Line 133 - sites D and M are undefined*
   Response: we thank the referee for pointing this missing reference.
   Edits to Manuscript: We have reordered the subsections in the revision so that sites D and M are defined before they are referred to in the discussion.
4. **Query/Comment**: *Section 2.4 does not seem like it needs to be its own section.*
   Response: We agree.
   Edits to Manuscript: The revised manuscript merges this section within section 2.3.
5. **Query/Comment**: *Line 203 - why are reference O3 concentrations <1 ppb scrubbed from the dataset?*
   Response: We thank the referee for pointing this out. We indeed discarded only those reference monitor values that were less than 0ppb and have corrected this typographical error. The reference monitors sometimes offer negative readings when powering up and under some other anomalous operating conditions e.g. condensation at the inlet. However, we note that less than 0.1% of the valid timestamps had reference O3 values between 0 and 1 ppb.
   Edits to Manuscript: The typographical error has been corrected.
6. **Query/Comment**: *Creation of train-test splits in 70-30 ratio in section 3.3.1*
   Response: we chose a 70:30 split since it gave us sizable sets for both training and testing. Machine learning and statistical estimation literature uses various splits such as 70:30, 80:20, etc. Our splits were repeated independently 10 times to allow two-sample tests to be carried out. The k-fold split method as mentioned by the referee, is another alternative. However, with k=10, the resulting 90:10 split offers a rather small test set which we wished to avoid. We verified that the choice of the size of the split (e.g. 70:30 vs 80:20) does not alter the conclusions of the paper.

7. **Query/Comment**: *Creation of subsampled datasets in section 3.3*
   Response: To create the subsampled datasets in section 3.3, we took a split (a split being a 70-30 division among train and test) and randomly subsampled 2500 points from the training portion of the split. The test portion was not altered since the aim of this experiment was to study how lack of training data affects calibration performance. To be more specific, if a dataset contained a total of 10000 valid timestamps, the train-test splits would contain resp. 7000-3000 points. For the subsampled version of this dataset, we would sample 2500 points from the 7000 points, train on those 2500 points and then test on the 3000 test points.
   Edits to Manuscript: We have clarified this in the paper and reordered the placement of section 3.3.1 to better explain this process.
8. **Query/Comment**: *Section 4: enumeration of techniques studied in the paper*
   Response: We agree with the referee and are thankful for the suggestion.
   Edits to Manuscript: We have included a glossary enumerating the names and brief descriptions of all techniques so that references to various algorithms can be readily checked. In order to improve the focus of the paper as suggested by the referee, we have moved details of other algorithms to the supplementary material.
9. **Query/Comment**: *Reason behind choosing k-NN style algorithms*
   Response: k-NN and kernel estimators (kernel ridge regression (KRR) and Nadaraya-Watson (NW)) are well studied non-parametric estimators in literature. These are also known to be asymptotically universal which theoretically guarantees their ability to accurately model complex patterns when given diverse and sufficient data.
   Edits to Manuscript: A brief explanation has been included in section 5.
10. **Query/Comment**: *Figure 5 - replotting*
    Response: we thank the referee for pointing out the improvements to the figures. Figures 5 and 9 in the old version chose to show data for different days as well as different durations which was inefficient. Figure 5 also had a manual labelling error which we have corrected in the revision.
    Edits to Manuscript: We have replotted figures 5 and 9 in the revision to consistently show results across the same two full days of operation (01-02 July and 20-21 Oct) for sake of clarity. However, figure 5 has been moved to the supplementary to make the discussion more focused.
11. **Query/Comment**: *Interpreting Table 3*
    Response: A metric tells us how to compute distance between two points, say 8 dimensional vectors in our case. The Euclidean metric gives equal importance to all 8 dimensions when calculating distances. An alternative interpretation of a Mahalanobis metric is that it tells us how to reorganize dimensions/features so that the resulting distances, when used by the kNN algorithm, give better performance. Table 3 shows us the optimal reorganization found by the metric learning technique. In particular, note that it places heavy emphasis on the Rh and T features. This means that the optimal Mahalanobis metric identifies that a high importance should be placed on Rh and T features when computing distances for use by kNN.
    Edits to Manuscript: A clarification has been added to the revision. However, we felt

that this discussion was not key to the focus of the paper and have moved this discussion to the supplementary material.

12. **Query/Comment**: *Analyzing where various algorithms offer high error in Figures 8, 10*
    Response: We thank the referee for this suggestion. To do this analysis, the (Rh, T) space was divided into various buckets to analyze the performance of each algorithm in each bucket. For data hungry non-parametric algorithms such as RT, NW(ML), KRR, and KNN-D(ML), regions of larger error coincided almost entirely with regions where data was scarce. This is as expected. The least squares method on the other hand demonstrated no such clear trend on regions of high error. We also tracked the errors of various algorithms across the day and found that for O3, whose diurnal levels are more predictable, all algorithms tended to offer relatively larger errors when the (true) O3 levels were higher (i.e. during peak sunlight hours). For NO2, which demonstrates no such predictable diurnal patterns, no patterns in errors were observed either.
    Edits to Manuscript: The revision now contains a new figure 6 and an analysis of situations in which various algorithms offer larger errors.
13. **Query/Comment**: *Section 6.4 typesettting*
    Response: Sections 6.5 through 6.8 were meant to be subsections of section 6.4.
    Edits to Manuscript: We have corrected this formatting error in the revision.
14. **Query/Comment**: *Include a discussion on swapout experiments where training and prediction are done across seasons or sites.*
    Response: Table 5 and section 6.8 in the old version do discuss cases when sensors are trained in one season and tested in another season, which does include cases when the site changes across seasons. We also request the referee to take a look at the comment of Referee #3 on this point and our rebuttal to the comment (please see “General Comments” bullet point 2 in our response to Referee #3).
    Edits to Manuscript: We have changed the title of subsection (section 5.2.4 in the revision) discussing the swapout experiment to highlight this.
15. **Query/Comment**: *Line 118 - Plantower output*
    Response: The Plantower PMS7003 offers readings in microgram per cubic meter, We thank the referee for pointing this out.
    Edits to Manuscript: We have corrected the typographical error in the revision.
16. **Query/Comment**: *Various comments on typography and formatting (e.g. line 133 - defining sites D and M, line 94, adding a glossary, labelling panels in figures 6 and 9, moving figure 7 to supplementary material)*
    Response: We thank the referee for taking pains to point out several improvements in typography and formatting.
    Edits to Manuscript: All suggestions have been incorporated in the revision.

Point-by-point Responses to Anonymous Referee #3

General Comments

1. **Query/Comment**: *Invalid timestamps: were 52% datapoints indeed discarded?*
   Response: Although around half the timestamps were indeed rejected (those that had

even one invalid measurement), it was still the case, especially for summer months, that at least one timestamp (frequently several) were found valid every hour. We note that this does not contradict the rejection of 52% timestamps since site D (resp. site M) offered timestamps at 1 minute (resp. 15 minute) intervals. Thus, the timestamps considered valid could still accurately track diurnal changes in AQ parameters (as indicated by Figure 9). A conservative approach was adopted when rejecting timestamps. We recall that a total of 8 parameters are involved in the training process -- four voltage values, relative humidity and temperature values from the LCAQ sensor, and two reference values (one each for O3 and NO2) from the reference monitors. Timestamps where even one of these parameters had an invalid value were rejected. In future work, data imputation techniques could be adopted to increase the number of valid timestamps.
Edits to Manuscript: We have included a discussion on this in the revision. Table 1 has been revised to include more illustrative examples of rejected timestamps.

2. **Query/Comment**: *Creation of a dataset that is diverse w.r.t. location but not season*
Response: The prospect of investigating the effect of spatial variation alone (without bringing seasonal variations into account) is interesting and we did consider this in our initial experiments but found that cross-sensor calibration is a challenging task in itself. For example, even the relative humidity and temperature sensors present in LCAQ sensors do not present good agreement across sensors. Thus, investigating spatial variation alone would have required us to do some form of “model transfer” of calibration models across LCAQ sensors. This is an encouraging direction for future work.
Edits to Manuscript: We have added a short discussion about this in section 3.2 itself where the derived datasets are discussed.
3. **Query/Comment**: *Out-of-sample generalization of parametric vs non-parametric models*
Response: we thank the referee for making this suggestion. As noted by the referee, performance drops are noticed in both algorithmic paradigms. However, as compared to the non-parametric method KNN-D(ML), the drop for LS is less in some cases, but comparable or worse in others. Of course, when diverse data is provided to both algorithms, KNN-D(ML) is superior at exploiting the additional diversity in data.
Edits to Manuscript: We have updated Table 5 (table 6 in the revision) to include the generalization results for the parametric linear least squares method LS as well.

Specific Comments

1. **Query/Comment**: *Various comments on typography and formatting (e.g. “upto” vs “up to”, typesetting 2.5 as a subscript in PM2.5, typesetting 2 as a superscript in R2, gas labels in figure 9)*
Response: We thank the referee for taking pains to point out several typography and typesetting changes.
Edits to Manuscript: We have incorporated all changes in the revised version.
2. **Query/Comment**: *Abstract: LCAQ are consistent but require calibration for accuracy.*
Response: We thank the referee for suggesting this rewording and agree with the same.
Edits to Manuscript: We have incorporated all changes in the revised version.

3. **Query/Comment**: *Lines 4, 52, 116, 188: reference to the word "commodity"*
   Response: The Alphasense electrochemical sensors used in the SATVAM LCAQ setup were not customized or specifically tailored for our study. Hence we use the term "commodity" to describe them.
   Edits to Manuscript: We have clarified this term at its point of first use in the paper.
4. **Query/Comment**: *Figure 3, lines 182, 184-186: number of sensors getting swapped*
   Response: We thank the referee for pointing this out. It seems we forgot to include a clarificatory remark in the paper. There were indeed 7 sensors deployed in the field of which 4 were swapped across sites. However, one of the sensors DM4 (that was swapped) was experiencing sensor malfunction. Its onboard Rh and T sensors were non-functional for the entire duration of the Jun deployment. For the Oct deployment, its data had extremely large gaps (sometimes spanning several days), which was qualitatively distinct from the other sensors which mostly experienced only intermittent gaps lasting a few minutes. For this reason, this sensor was excluded from our study. Although for sake of full disclosure we still mentioned in our original submission that 7 sensors were used, we forgot to include this clarificatory remark.
   Edits to Manuscript: We have included this clarification in the revision and corrected the number of sensors reported at various places in the paper to be consistent.
4. **Query/Comment**: *Do Rh, T values come from: LCAQ sensors or reference monitors?*
   Response: Rh and T values were obtained from DHT22 sensors located in the individual LCAQ sensors. 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.
   Edits to Manuscript: We have clarified this in the revision in section 3.
5. **Query/Comment**: *Clarify figure 6 labels, add plots showing site variation, and avoid Gaussian fitting for unsigned data*
   Response: we agree with the referee's suggestions and are thankful for the same.
   Edits to Manuscript: We have moved this plot to the supplementary material in the revised version as well as added plots that show differences across sites but in the same season. We have also clarified all aspects of the plot as kindly pointed out in the comments. We have also replaced Gaussian fits (dotted lines) with non-parametric KDE fits which are more appropriate for data that is visibly non-Gaussian.
6. **Query/Comment**: *Line 385: What does "statistically distributed" mean?*
   Response: We thank the referee for pointing out this typographical error. We meant to write "normally distributed". However, we have amended this statement since some of the distributions do not seem normally distributed.
   Edits to Manuscript: We have corrected this typographical error in the revision.
7. **Query/Comment**: *Line 391: The figure does not appear to have 3rd or 4th rows.*
   Response: We regret this formatting error. Our initial submission to the journal was in a two column format (in which Figure 6 did have 4 rows). However, we were requested by the editorial desk to convert to a single column format. We did so but forgot to change this piece of text to reflect the change in formatting. We have corrected this.
   Edits to Manuscript: We have corrected this formatting error in the revision. However, the figure and accompanying discussion has been moved to the supplementary material.

8. **Query/Comment**: *shifting figure 7 to the supplementary and clarifying violin plot details*
   Response: We agree. We used the standard Python-based library seaborn to create the plots. Seaborn seems to calculate medians and interquartile range of the combined left and right data in the case of split violin plots. This can be seen in figure 7 (right) where the median and interquartile ranges correspond to the combined data rather than the left or the right data.
   Edits to Manuscript: We have moved the small tutorial on interpreting violin plots to the supplementary and added this clarification on medians and interquartile ranges.
9. **Query/Comment**: *Section 6.4 appears incomplete*
   Response: The referee is indeed correct in observing that sections 6.5 through 6.8 were meant to be subsections of section 6.4
   Edits to Manuscript: We have corrected this formatting error in the revision.

General Comments on Edits to the Manuscript

Apart from changes to fix typographical or formatting errors (e.g. repeated words "sensor sensor", "inter" vs "intra" in title of section 6.2.2, labelling errors, subscript error in formatting PM2.5, R2), most changes were done to improve the focus of the paper and make the writing more concise. We agree with the comments of both referees that encouraged us to move portions not essential to the core discussion, to the supplementary material.

1. Section 6.1 (Analysis of Raw Data) has been moved to the supplementary material along with detailed descriptions of the deployment sites in section 2.2. It was suggested in the review that Figure 6 etc be moved to the supplementary and we agreed that these portions do not significantly contribute to the core discussion.
2. As suggested in the review, portions of sections 4 and 5 have been moved to the supplementary material. The main text now briefly outlines baseline calibration methods, motivates the proposed method and gives necessary details of the proposed method. We agree with the referee comments on making the presentation of the calibration algorithms tighter.
3. Additional results have been introduced in the main text as suggested in the referee comments, for which we are thankful. For example, a glossary of the acronyms used in the discussion, results of the parametric algorithm LS in the swapout experiments on the aggregated datasets, a discussion on the cases in which various algorithms offer high error, and numerical values of performance improvements offered by the proposed method, in addition to the violin plots.
4. Algorithm 1 in the main paper has been simplified to describe only the proposed KNN-D(ML) algorithm. Earlier the algorithm sought to describe the entire family of KNN-style algorithms which may have been confusing. The general description of the KNN family of algorithms that was earlier present in the main text has been moved to the supplementary material for the interested reader.
5. The discussion around the diagonal entries of the learnt metric and the accompanying Table 3 have been moved to the supplementary material. It seems that the discussion may not be of general interest.

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. ~~Rising awareness of the health risks posed by elevated levels of ground-level and have led to an increased demand for~~ Low-cost sensors offer an attractive solution to the challenge of establishing affordable and dense spatio-temporal air quality monitoring networks ~~. Low-cost sensors used as a part of Internet of Things (IoT) platforms offer an attractive solution~~ with greater mobility and lower maintenance costs~~, and can supplement compliance regulatory monitoring stations. These commodity~~ ~~.~~ . These low-cost sensors ~~have reasonably high accuracy~~ offer reasonably consistent measurements, but require in-field calibration to improve ~~precision~~agreement with regulatory instruments. In this paper, we report the results of a deployment and calibration study on a network of ~~seven~~ 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~~– Faridabad (Delhi National Capital Region) and Mumbai, India. The deployment was done in two phases over a period of three months.~~ . A unique feature of our deployment is a *swap-out* experiment wherein ~~four~~ three of these sensors were relocated to different sites in the two ~~deployment phases. Such a diverse deployment with sensors switching places~~ phases. This gives us a unique opportunity to study the effect of seasonal, as well as geographical variations on calibration performance. We ~~perform~~ report an extensive study of more than a dozen parametric ~~as well as~~ and non-parametric calibration algorithms~~and find local calibration methods to offer the best performance~~. We propose a novel local non-parametric calibration algorithm based on metric-learning that offers, across deployment sites and phases, an ~~average R2 coefficient of 0.873~~ R^2 coefficient of upto 0.923 with respect to reference values for O_3 calibration and ~~0.886~~ 0.819 for NO_2 calibration. This represents an ~~upto 9%~~ 4 - 20 percentage point increase in terms of ~~R2~~ R^2 values offered by classical ~~local calibration methods. In particular, our proposed model far outperforms the default calibration models offered by the gas sensor manufacturer.~~ 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 ~~as free parameters (or features) into all~~ 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) ~~the use of local (or even hyper-local)~~ using local calibration techniques such as ~~k-NN that seem to offer the best performance in high variability conditions such as those encountered in field deployments~~KNN, 4) performing temporal smoothing over raw time series data, ~~say by averaging sensor signals over small windows,~~ 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

~~Adverse effects of air pollution~~ Elevated levels of air pollutants have a detrimental impact on ~~the health of human populations~~ human health as well as the economy (Chowdhury et al., 2018; Landrigan et al., 2018). For instance, high levels of ground-level ~~ozone can cause severe health risks, including but not limited to ,~~ 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. ~~Decision-makers require real-time information on air pollution to formulate effective policies which presents~~ There is a need for ~~monitoring~~ accurately real-time monitoring of air pollution levels ~~accurately~~ with dense spatio-temporal coverage.

Existing regulatory techniques for assessing urban air quality (AQ) rely on a small network of ~~monitoring stations providing highly precise measurements of the pollutants (Snyder et al., 2013; Malings et al., 2019). In developing countries like India, existing city-level air quality monitoring networks are comprised of a proportionally small number of~~ Continuous Ambient Air Quality Monitoring Stations (CAAQMS) ~~. These stations~~ 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 ~~and operating cost . The~~ cost and are cumbersome to maintain (Sahu et al., 2020), making dense CAAQMS networks impractical. However, the AQ data offered by ~~a small number of these monitors across a city~~these sparse networks, however accurate, ~~limit~~ limits the ability to formulate ~~AQ improvement~~ effective AQ strategies (Garaga et al., 2018; Fung, 2019). ~~Moreover, CAAQMS with traditional gas analyzers and filter based monitoring facilities are cumbersome and expensive to install, operate and maintain (Sahu et al., 2020). Consequently, real-time actionable data at the citizen level are currently available at very few locations in a city. There is a need for dense air quality monitoring coverage across cities that can complement the limited spatial resolution of existing air pollution maps that are available for citizens (Kumar et al., 2015; Schneider et al., 2017; Zheng et al., 2019). Adequate information on the real-time spatial and temporal distribution of pollutants would allow citizens to make informed decisions, for instance, on~~

~~their commute (Apte et al., 2017; Rai et al., 2017). However, the overall cost of operating a large number of reference-grade CAAQMS is not practical, especially in developing countries like India.~~

In recent years, the availability of low-cost AQ ~~monitoring sensors for measuring real-time air pollution concentrations~~ (LCAQ) monitoring devices has provided exciting opportunities for finer spatial resolution data ~~(Rai et al., 2017; Baron and Saffell, 2017)~~ (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 ~~monitoring~~ CAAQMS system is around USD 200,000, while that of ~~a low-powered~~ an LCAQ device running commodity ~~AQ~~ sensors is under USD 500 (Jiao et al., 2016; Simmhan et al., 2019). ~~Several low-cost sensors can be installed to complement a few reference monitors for better pollution mapping. In addition,~~ 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 ~~emergence of cloud computing and the~~ Internet of Things (IoT) ~~cyber-infrastructure~~ infrastructure allows building large-scale networks of ~~low-powered AQ monitoring~~ LCAQ devices (Baron and Saffell, 2017; Castell et al., 2017; Arroyo et al., 2019). ~~This paves a way for regulatory bodies to use AQ sensor data to identify patterns and~~

Dense LCAQ networks can complement CAAQMS to help regulatory bodies identify sources of pollution and ~~efficient policy formulation, for~~ formulate effective policies, allow scientists to model ~~the~~ interactions between climate change and pollution~~accurately~~ (Hagan et al., 2019), ~~and to facilitate the participation of the common public~~ 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 ~~more actively~~initiatives (Gabrys et al., 2016; Commodore et al., 2017; Gillooly et al., 2019; Popoola et al., 2018).

~~However, the use of low-cost sensor data at a high temporal resolutions presents challenges as available sensors are not designed to meet rigid performance standards and generate less accurate data than research-grade instruments (Mueller et al., 2017; Snyder~~ ~~. Thus, there is need to evaluate data from real-time sensors and IoT networks made available by manufacturers of AQ devices, for accuracy and precision (Akasiadis et al., 2019; Williams, 2019).~~

1.1 Challenges in low-cost sensor calibration

Measuring ground-level ~~ozone (~~O_3 ~~) and nitrogen dioxide (~~and NO_2 ~~) accurately using sensors~~ is challenging as they occur at parts per billion ~~(ppb) micro-levels~~ levels and intermix with other pollutants (Spinelle et al., 2017). ~~Most commonly available low-cost sensors for these gas-phase compounds~~ 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). ~~Field calibration remains one of the major challenges preventing extensive use of these technologies. Often, sensor calibration is carried out in controlled conditions which differ substantially from real-world conditions.~~

~~In addition, these sensors at times have issues of consistency, stability and~~ 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. ~~Further, the constancy of low-cost sensors is recognized to reduce overtime. Moreover, in electrochemical cells~~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 ~~techniques that meet performance metrics required by~~ of LCAQ sensors to satisfy performance demands of end-use applications~~even at low ambient concentrations (De Vito et al., 2018)~~ (De Vito et al., 2018; Akasiadis et al., 2019; Williams, 2019).

1.2 Related Works

Recent works have ~~demonstrated that valid~~ shown that LCAQ sensor calibration can be achieved by co-locating ~~them with highly accurate~~ the sensors with regulatory-grade reference monitors and using various ~~linear and non-linear~~ calibration models (De Vito et al., 2018; Hagan et al., 2019; Morawska et al., 2018). ~~For the specific~~ 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 ~~simple~~ parametric models such as ~~least squares~~ linear least squares regression (LS) ~~regression~~ could extrapolate to wider concentration ranges, at which non-parametric regression model may ~~fail~~struggle. However, LS does not correct for ~~the~~ dependence on temperature or relative humidity (RH)~~dependence of the signal~~, at which non-parametric models may be more effective.

Since electrochemical sensors are configured ~~such that the responses are~~ to have diffusion-limited responses, and the diffusion ~~coefficient~~ coefficients could get affected by ~~varying temperature, Hitchman et al. (1997); Masson et al. (2015)~~ 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. ~~More recently,~~ 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 ~~similarly impressive~~ similar performance was not observed for O_3 calibration. ~~Moreover~~Notably, existing works ~~have mostly been tested with~~ mostly use a localized deployment of a small number of sensor, ~~for instance~~ e.g. Cross et al. (2017) who tested two ~~sensor~~ devices, each containing one sensor per pollutant~~over 4 months with 35% training data~~.

1.3 Our Contributions and the SATVAM initiative

The SATVAM initiative (*Streaming Analytics over Temporal Variables from Air quality Monitoring*) has ~~focused on the development and calibration of~~ been developing low-cost air quality (LCAQ) sensor networks based on highly portable IoT software platforms. These ~~sensors~~ LCAQ devices include (see Fig. 2) ~~contain PM2.5~~ $\mathrm{PM}_{2.5}$ 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 ~~work of Zheng et al. (2019) considered the related problem of dynamic PM2.5 sensor calibration within a sensor network deployed in Delhi, India. 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 ~~6~~ six sensors deployed at two sites over two phases with vastly different meteorological, geographical and air quality parameters~~, over two phases.~~ .

2. A unique feature of our deployment is a *swap-out* experiment wherein ~~4~~ three of these sensors were relocated to different sites in the two phases (see Sect. 2 for deployment details).

3. ~~The swap-out experiment in particular is crucial in allowing~~ 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. ~~This~~ Such an investigation is missing from previous works which mostly consider only localized calibration~~of a couple of models~~.

4. We present an extensive study of parametric and non-parametric calibration models, ~~both parametric and non-parametric~~ and develop a novel local calibration algorithm based on metric learning that offers ~~both~~ stable (across gases, sites and ~~deployment phases) , as well as accurate calibrationperformance~~seasons) and accurate calibration.

5. We present ~~a critical~~ an analysis of the effect of data preparation techniques such as volume of data, temporal averaging and data diversity, on calibration performance. This ~~study yields several simple yet crucial~~ yields several take-home messages that ~~significantly~~ 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. ~~Here we give details of the deployment setup.~~

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 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.

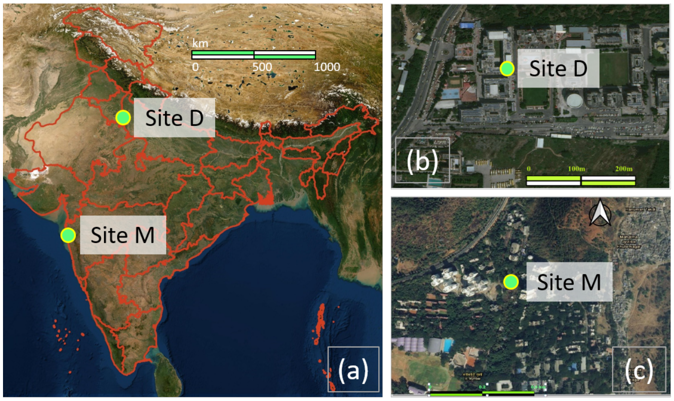

Figure 1. A map showing the locations of the deployment sites. Fig. 1(b) and (c) on the right show a local-scale map of the vicinity of the deployment sites – namely Site D at MRIU, **D**elhi NCR (Fig. 1(b)) and Site M at MPCB, **M**umbai (Fig. 1(c)), with the sites themselves pointed out using bright 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.

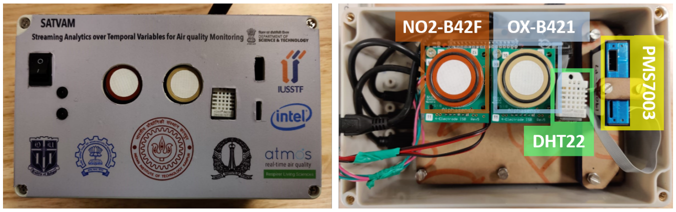
SATVAM
Streaming Analytics over Temporal Variables for Air quality Monitoring
IUSSTF
intel
atmos
NO2-B42F
OX-B421
PMS7003
DHT22

Figure 2. Primary components of the *SATVAM* LCAQ (low-cost air-quality) sensor used in our experiments. The SATVAM ~~ensemble~~ device consists of a Plantower PMS7003 ~~PM2.5~~ $PM_{2.5}$ 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.

2.2 Instrumentation

~~Low-cost~~ 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 ~~PM2.5 (mg~~ $PM_{2.5}$ (μg m^{-3}) levels, and a DHT22 sensor for measuring ambient temperature in °C and RH in ~~percent~~percentage points. Fig. 2 shows the placement of these components. A notable feature of this device is its focus on ~~resource~~ frugality with use of the very low-power ContikiOS platform and 6LoWPAN for providing wireless sensor network ~~communications~~connectivity.

Detailed information on assembling these different components and the ~~cyber-infrastructure required to make a customized sensor node capable of interfacing within~~ interfacing with an IoT network is ~~available inother works~~described in (Simmhan et al., 2019). These ~~works also describe in detail the formation of~~ sensors form a highly portable IoT software platform to transmit 6LoWPAN packets at 5 minute intervals containing five time-series data points ~~of the~~ from individual sensors, namely NO_2, O_3, ~~PM2.5~~ $PM_{2.5}$ (not presented in this study), temperature and ~~relative humidity (RH). In previous deployments which used only a couple of SATVAM devices, a Raspberry Pi unit was used at each device along with a mesh network to collect and push data to a cloud storage facility. However, for the current deployment that considers a much~~ 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 ~~failures~~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$^{\text{TM}}$, USA) and a chemiluminescence oxide of nitrogen (NOx) analyzer (Model 42i NOx analyzer, Thermo Scientific$^{\text{TM}}$, 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 ±0.25 ppb and ±0.2 ppb, respectively. Similarly, the deployments at 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 ±0.2 ppb and ±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 ~~. The DHT-22 sensor of the SATVAM devices was compared to Vaisala, a reference-grade instrument for temperature and humidity kept alongside the AQ monitors at site D.~~ since the site M reference monitors gave data at 15 minute intervals.

~~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.~~

2.3 ~~Deployment Sites~~

~~SATVAM LCAQ sensor 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 (28.45N, 77.28E, 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 (19.13N, 72.91E, and 50 m above mean sea level).~~

~~**About Site D**: According to Greenpeace India (Times, 2018) and the Niti Aayog, Govt. of India (Aggarwal, 2018), Faridabad was the second most polluted city in India in 2018. Surrounded by the Aravalli Hills, this is a rapidly growing city and a leading industrial center suffering from heavy air pollution that mask the city and its neighborhoods routinely during the fall and winter seasons. The study site is 5 km away from Delhi and near Delhi-Surajkund Highway. It falls in the Indo-Gangetic Plain, which registered critical levels of ambient air pollution attributable to a combination of multiple ambient sources, the use of biomass and coal for household cooking and heating needs, and the stubble or agricultural residue burning (Chowdhury et al., 2019). The deployment site is affected by vehicular traffic which are likely a dominant source of precursors to formation (and volatile organic compounds) and of nitric oxide that reacts with to form the pollutant . The reference monitors were deployed in a laboratory on the first floor of the building with the low-cost AQ monitoring sensors next to its inlets.~~

~~**About Site M**: This site presents relatively lower pollution levels as it is situated within the IIT Bombay campus between the Vihar and Powai lakes, and it is adjacent to the Sanjay Gandhi National park. Less that 5 km to its west side passes the Thane creek (an inlet in the shoreline of the Arabian Sea) that isolates the city of Mumbai from the Indian mainland while the Arabian sea is at around 10 km to its west. The reference monitors were deployed on the rooftop of the building with the low-cost AQ monitoring sensors next to its inlets. All AQ monitoring devices were in a Stevenson box to avoid damage to sensors.~~

~~Due to ever-increasing economic and industrial activities across the city, a progressive worsening of ambient air pollution is nearly inevitable at both study sites. We considered these two polluted sites situated within the Delhi-NCR and Mumbai to cover a broader range of pollutant concentrations and weather patterns, so as to be able to test the reliability of low-cost sensor networks in measuring and levels. 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 levels than site M, illustrating the influence of the geographical variability over the selected region.~~

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 ~~clearly depicts~~ describes its deployment pattern. The name of a sensor is of the form `XYn` where `X`

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

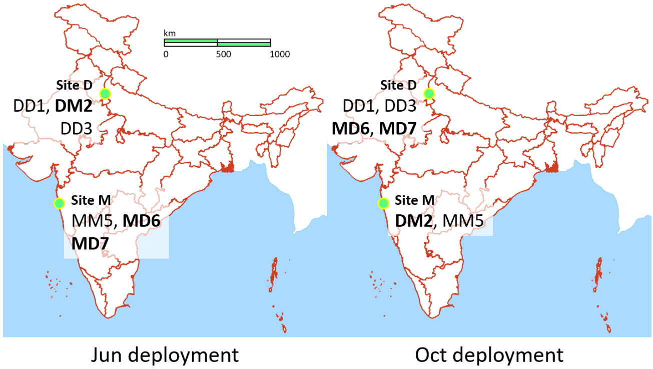

Figure 3. A schematic showing the deployment of the six LCAQ sensors across site D and site M during the two deployments. The sensors subjected to the *swap-out* experiment are presented in bold. The outlines of the Indian states in red was taken from QGIS3.4 Madeira with other highlights (e.g. for oceans)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

(resp Y) indicates the site at which the sensor was deployed during the Jun (resp Oct) deployment and n denotes its unique numerical identifier. ~~The seven sensors are thus named~~ Fig. 3 outlines the deployment patterns for the six sensors DD1, DM2, DD3, MM5, MD6, and MD7. ~~Fig. 3 outlines the deployment patterns.~~

2.4 ~~Swap-out Experiment~~

Swap-out Experiment. As Fig. 3 indicates, ~~two of the sensors from each of the sites were exchanged or swapped out to the other city~~ three sensors were swapped with the other site across the two deployments. Specifically, for the Oct deployment, DM2 was shifted from ~~Delhi to Mumbai~~ site D to M and MD6 and MD7 were shifted from ~~Mumbai to Delhi for the Oct deployment.~~ 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.

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	ox
29-06 04:21	19.82	20.49	32.7	54.6	212	231	242	209	-19	
~~29-06 04:22~~ 30-06 08:02	~~21.89~~ 46.363	~~20.56~~ **-0.359**	~~32.7~~ 36.8	~~54.6~~ 39.6	~~212~~ 184	~~231~~ 221	~~243~~ 234	~~210~~ 201	~~-19~~ -37	
~~29-06~~ 01-07 04:~~23~~ 02	~~22.71~~ 24.38	~~18.17~~ 14.73	~~×~~ 32.5	~~×~~ 69.7	×	×	×	×	×	
~~29-06 04:24~~ 08-07 07:51	~~24.82~~ **-0.035**	~~14.60~~ 17.147	~~32.5~~ 31.5	~~53.3~~ 97.8	~~×~~ 209	~~×~~ 238	~~×~~ 231	~~×~~ 216	~~×~~ -29	~~×~~

	MM5(Oct)							
Time-stamp	O3	NO2	T	RH	no2op1	no2op2	oxop1	oxop2
~~17-10~~ 19-10 05:45	~~7.17~~ ×	~~37.62~~ ×	~~26.12~~ ×	~~99.9~~ ×	~~119.27~~ 160.46	~~152.93~~ 188.31	~~128~~ 158.31	~~133.4~~ 172.
~~17-10 06:00~~ 19-10 07:15	~~8.7~~ 5.55	~~34.11~~ 11.52	~~26.14~~ 41.47	99.9	~~122.93~~ 170.4	~~155.53~~ 197.2	~~131.87~~ 167.6	~~136.47~~ 181
~~17-10 06:15~~ 20-10 10:45	×	×	~~26.25~~ 28.52	99.9	~~121.67~~ 121.8	~~154.13~~ 154.0	~~129.2~~ 119.3	~~134.6~~ 135
~~17-10 06~~22-10 18:30	~~10.86~~ 8.33	~~30.95~~ 10.91	~~26.16~~ 27.87	99.9	~~119.33~~ 143.2	~~151.4~~ 172.3	~~127.27~~ 146.2	~~131.67~~ 155

3 Data Analysis Setup

~~**Testbench:**~~ 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.

3.1 ~~Raw Datasets and Features~~

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~~name~~. 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~~at each time stamp, in addition to~~ , and voltage readings (in mV) from the two electrodes present in each of the two gas sensors. These readings ~~are named no2op1, no2op2, oxop1, and oxop2 and they~~ 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.

3.1 ~~Data Cleanup~~

Data Cleanup. Time-stamps from ~~each of~~ the LCAQ sensors were aligned to those from the reference monitors. ~~We considered only those datapoint that were temporally aligned.~~ 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 ~~a~~ 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, ~~an~~ c) an sensor RH level of > 100 % or < 1 %, ~~a reference value for or~~ and d) a sensor voltage reading (either of no2op1, no2op2, oxop1, oxop2) of > ~~200 ppb or < 1 ppb, or voltage readings from the four sensors at values either >~~ 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.

~~For site D deployments, both the LCAQ sensor as well as the reference monitor data was available at 1 minute intervals. However for site Mdeployments, whereas the LCAQ sensors continued to provide data~~ However, since site D (resp. site M) offered timestamps at 1 minute ~~intervals, the reference monitors at that site were set to provide data at~~ (resp. 15 ~~minute intervals. To align the two time series, LCAQ sensor data was averaged over 15 minute intervals.~~

~~The 3~~ 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. ~~4 from~~ Oct) deployments at site D offered an average of 33753 (resp. 9548) valid time-stamps. The ~~3~~ datasets from Jun (resp. ~~2 from~~ Oct) deployments in site M offered an average of 2462 (resp. 1062) valid time-stamps. As expected, site D ~~deployments offered more valid~~ that had data at 1 minute intervals offered more time-stamps than site M ~~deployments in any season given that the former enjoyed 1 minute interval data whereas the latter deployments~~ that had data at 15 minute intervals. ~~We also note that for~~ 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, ~~namely RH and T values and sensor voltage values (no2op1, no2op2, oxop1, oxop2),~~ we included two ~~derived~~ augmented features, calculated as shown below

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

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

We found that having these ~~derived~~ 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, ~~5~~ 15 minute), data volume (total number of time-stamps used for training), and data diversity (data collected across seasons or ~~cities~~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). ~~This was done in order to study the effect of offering datato the calibration models that is more diverse in terms of location (since DM2 and MD6 moved across sites) and season (Jun vs Oct) .~~

3.2.1 ~~Train–Test Splits~~

~~To create training and test data from each dataset (whether original or derived), we randomly split each dataset 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 samples in that split were used as a held out validation set. All features were normalized to improve the conditioning of~~ **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 ~~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. Also, the reference values were never normalized in any way~~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 ~~Error Metrics and Statistical Hypothesis Testing~~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) (see below). Here n denotes the number of test points for a given dataset and split thereof, the variable t runs over all time-stamps in the testing set, y^t denotes the reference value (either or)at the t-th time-stamp, $\hat{y}^t$ denotes the corresponding value predicted by the calibration model, and $\bar{y}$ denotes the mean reference value i. e. $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y^t$~~R^2) (please see the supplementary material for detailed expressions of these metrics).

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |y^t - \hat{y}^t|$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \frac{|y^t - \hat{y}^t|}{y^t} \times 100\%$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (y^t - \hat{y}^t)^2}$$

$$\text{R}^2 = 1 - \frac{\sum_{t=1}^{n} (y^t - \hat{y}^t)^2}{\sum_{t=1}^{n} (y^t - \bar{y})^2}$$

Statistical Hypothesis Tests: in order to compare the performance of different calibration algorithms on a given dataset (~~e.g.,~~ to find out the best performing algorithm), or compare the performance of the same algorithm on different datasets (~~e.g.,~~

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 ~~by the two algorithms on the same dataset (in case of a paired test) or the same algorithm across different datasets (in case of an unpaired test) were sampled from the same distribution~~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 ~~most popularly used in such situations, it essentially~~ more popular, it assumes that the underlying distributions are normal. However, an application of the Shapiro-Wilk test (Shapiro and Wilk, 1965) ~~rejected the null hypotheses of the errors being normally distributed~~ to our absolute error values rejected the normal hypothesis with high confidence. ~~As a result~~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 ~~used~~ considered a large number of ~~both parametric ,~~ parametric and non-parametric calibration techniques ~~. Since several of these techniques are standard, we describe them in the *Supporting Information* document supplied with this paper. In particular, the Supporting Information document describes several parametric calibration algorithmsincluding~~ 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 ~~and sparse recovery, as well as several~~ (LS and LS(MIN)) and sparse recovery (LASSO). Among non-parametric ~~calibration algorithmssuch as regression trees~~ algorithms, we considered regression trees (RT), kernel-ridge regression ~~, and~~ (KRR), the Nystroem method ~~. We describe here in the main paper, only those baseline calibration models upon which our proposed technique is developed.~~ 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 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

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

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 ~~$k-$NN Regression Variants~~Proposed Method: Distance-weighed KNN with a Learnt Metric

~~The k-nearest neighbor~~ Our proposed algorithm is a ~~*local* proximity-based learning algorithm that makes predictions on test samples based on which are the training samples that most resemble the test sample. Resemblance is usually calculated using a metric such as the Euclidean metric. We implement several k-nearest neighbor variants. Algorithm 1 gives pseudo code for these variants.~~ local, non-parametric algorithm that uses a learnt metric. Below we describe the design of this method and reasons behind these design choices.

~~**k-NN with Euclidean Distance (KNN):**~~**Non-parametric estimators for Calibration.** The ~~vanilla~~ simplest example of a non-parametric estimator is the KNN (k ~~-nearest algorithm (KNN)~~ nearest neighbors) algorithm that predicts on a test ~~sample~~point, the average reference value in the k nearest training ~~samples. The neighborhood size k was tuned over the fine grid $[2, 4, 6, 8, 10, 15, 20]$ using held-out validation. Standard implementation of kd-trees were used to accelerate the process of discovering the nearest neighbors for a test sample.~~

~~**Distance weighted k-NN (KNN-D):** We also implemented a *distance-weighted* version of this algorithm wherein closest neighbors for particular test sample are weighted according to their Euclidean distance to the test point with closer pointsgetting more weightage. We found this to favorably improve our accuracy.~~

4.2 ~~Kernel Regression Variants~~

~~In statistics and machine learning, the notion of a *kernel* refers to a function that assigns a similarity value to two vectors (Murphy, 2012) . Thus, a kernel is of the form $K : \mathbb{R}^8 \times \mathbb{R}^8 \rightarrow \mathbb{R}$ which, when given two vectors $\boldsymbol{x}^1, \boldsymbol{x}^2 \in \mathbb{R}^8$, assigns a value $K(\boldsymbol{x}^1, \boldsymbol{x}^2) \in \mathbb{R}$ denoting how similar are these vectors. A popularly used kernel is the *Gaussian* kernel (aka the *RBF* kernel) that calculates this similarity as $K(\boldsymbol{x}^1, \boldsymbol{x}^2) = \exp(-\gamma \cdot \|\boldsymbol{x}^1 - \boldsymbol{x}^2\|_2^2)$ where $\|\cdot\|_2$ denotes the Euclidean norm and γ is a *bandwidth* parameter that controls the scale at which similarity values go down. The Nadaraya-Watson estimator and~~ points. Other examples (please see the supplementary material for details) include kernel ridge regression ~~are two popular forms of kernel regression algorithms. A closely related cousin is Gaussian-process regression.~~

~~Below we describe~~ (KRR) and the Nadaraya-Watson ~~estimator as it is useful in the developement of our proposed technique. Kernel ridge regression is described in the Supporting Information document.~~(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.

~~**Nadaraya-Watson (NW):** Given a training set $\{(\boldsymbol{x}^t, y^t)\}_{t=1}^N$, the NW estimator (Nadaraya, 1964; Watson, 1964) makes a prediction on a new (testing) data point $\boldsymbol{x} \in \mathbb{R}^8$ as follows~~

$$f^{\text{NW}}(\boldsymbol{x}) = \frac{\sum_{t=1}^N y^t \cdot K(\boldsymbol{x}^t, \boldsymbol{x})}{\sum_{t=1}^N K(\boldsymbol{x}^t, \boldsymbol{x})}$$

~~The intent of this estimatoris clear – the final prediction is a weighted sum of reference values y^t in the training set with the weight of a training sample $t \in [N]$ being proportional to $K(\boldsymbol{x}^t, \boldsymbol{x})$ i. e. how similar is that training sample to the test sample. Notice also the similarity between NW and KNN-D in the way they make predictions.NW almost behaves like a "smoothed" version of KNN-D by performing weighing using kernel values instead of inverse Euclidean distances and considering all training samples instead of just the neighbors. This observation will be useful later.~~

5 ~~Proposed Calibration Model: *k*-NN variants with a learnt metric~~

~~Below we propose a novel application the *metric learning* technique to build non-parametric calibration models that offer superior performance compared to other models.~~

~~feature vector for a test sample $\tilde{\boldsymbol{x}}$, training samples $\{(\boldsymbol{x}^t, y^t)\}_{t=1}^N$, neighborhood size k, weighing rule, metrica prediction $\hat{y}$ for the test sample using one of the KNN, KNN-D, KNN(ML) or KNN-D(ML) calibration model depending on the weighing rule and metric arguments $\boldsymbol{\Sigma} \leftarrow I_8 \boldsymbol{\Sigma} \leftarrow$ use training samples to learn a Mahalanobis metric using the technique from (Weinberger and Tesau Find the k training samples (say $i^1, \ldots, i^k$) that are closest to $\tilde{\boldsymbol{x}}$ in terms of the learnt Mahalanobis distance $d^{\text{Maha}}(\cdot, \cdot; \boldsymbol{\Sigma})$ $\hat{y} = \frac{1}{k}\sum_{l=1}^k y^{t_l}$ For all $l = 1 \ldots k$, let $\alpha^l = (d^{\text{Maha}}(\boldsymbol{x}, \boldsymbol{x}^{t_l}; \boldsymbol{\Sigma}))^{-1}$ $\hat{y} = \frac{\sum_{l=1}^k \alpha^l \cdot y^{t_l}}{\sum_{l=1}^k \alpha^l}$ $\hat{y}$ Variants of k-NN based calibration~~

4.1 ~~Metric Learning~~

~~As~~ **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 ~~, the inclusion of~~ that certain features, e.g. RH and T~~as additional features benefits~~ , seem to have a significant influence on calibration performance. ~~However~~Thus, it is unclear how much ~~importance should these features receiveas opposed to the other features that are based on voltage readings (~~emphasis should RH and T receive, as compared to other features such as voltage values e.g. ~~no2op1, oxop2, no2diff etc). This is particularly true of k-NN and kernel regression, both of which find neighbors or calculate kernel values by relying on the Euclidean distance which assigns equal importance to all 8 features.~~

~~It is well established that k-NN style algorithms stand to gain if used with a customized metric~~ 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~~(Weinberger and Saul, 2009). It is most popular to replace the Euclidean metric with a learnt *Mahalanobis metric*. This metric~~. A Mahalanobis metric is characterized by a positive semi-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 ~~k-NN is very~~ KNN is popular for classification problems, it is uncommon for calibration and regression problems. This is ~~partly because of technical problems posed by regression problems which lack~~ due to regression problems lacking of a small number of “classes”.

To overcome this problem, we ~~recall our earlier observation that the NW algorithm almost behaves like a smoothed version of the KNN-D algorithm. Given that~~ 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 ~~(Weinberger and Tesauro, 2007)~~ to learn a Mahalanobis metric ~~for use with~~ to be used alongwith the NW algorithm ~~, we~~ (Weinberger and Tesauro, 2007). This allows us to adopt a *two-stage* algorithm that first learns a Mahalanobis metric suited for the NW ~~estimator and then using it with the KNN and~~ algorithm and then uses it to perform KNN-style calibration. Algorithm 1 describes the resulting KNN-D~~algorithms. The method proposed by Weinberger and Tesauro (2007) learns the metric by attempting to minimize the leave-one-out RMSE over the training samples.~~

~~**Metric learning with Nadaraya-Watson (NW(ML)) and k-NN algorithms (KNN(ML), KNN-D(ML)):** We call the variants of NW, KNN and KNN-D when used with a learnt metric, respectively NW~~(ML) ~~, KNN(ML) and~~ algorithm.

5 Results and Discussion

The goals of using low-cost AQ monitoring sensors vary widely. This ~~study focuses on critically assessing~~ section critically assesses a wide variety of calibration models~~and assessing the suitability of low-cost sensors for spatially dense AQ monitoring networks.~~

5.1 ~~Analysis of Raw Data~~

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) ~~. The modification required to execute NW(ML) with a learnt metric is straightforward – we simply start using an alternate kernel given by~~
    ~~$K^{\text{Maha}}(\boldsymbol{x}^1, \boldsymbol{x}^2; \boldsymbol{\Sigma}) = \exp(-(d^{\text{Maha}}(\boldsymbol{x}^1, \boldsymbol{x}^2; \boldsymbol{\Sigma}))^2)$~~
    ~~We note that this alternate kernel does not require an explicit bandwidth parameter since any such parameter can be absorbed into the matrix $\boldsymbol{\Sigma}$ itself. Algorithm 1 details pseudo-code for KNN(ML) and 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 $\alpha^l = (d^{\text{Maha}}(\tilde{\boldsymbol{x}}, \boldsymbol{x}^{i_l}; \boldsymbol{\Sigma}))^{-1}$
    $\hat{y} = \frac{\sum_{l=1}^k \alpha^l \cdot y^{t_l}}{\sum_{l=1}^k \alpha^l}$
    return Calibrated value $\hat{y}$ for the test data point
```

~~A scatter plot showing variations in RH and T at the two sites across the two deployments. The sites offer substantially diverse weather conditions. Site D exhibits wide variations in RH and T levels during both deployments. Site M exhibits almost uniformly high RH levels during the Oct deployment which coincided with the retreating monsoons.~~

~~Time series showing the variation in the raw parameters measured using the reference monitors (and concentrations) as well as those measured using the SATVAM LCAQ sensors (RH, T, no2op1, no2op2, oxop1, oxop2). The left figure considers a 24 hour periods during the Jun deployment (28 June 2019) at site D whereas the right figure considers the Oct deployment (12 October 2019) at site M. Values for site D are available at 1 minute intervals while those for site M are averaged over 15-min intervals.~~

~~Normalized frequency distributions for various data series. Data from Jun deployments (resp Oct deployments) is shown in red (black) in all plots. The plots in row 1 show, from left to right, variations in the reference values at site D by considering data from the DD1 sensor for (Fig. **??**(a)) and (Fig. **??**(b)). Fig. **??**(c) and (d) show the same for site M by considering data from the MM5 sensor. Recall that both the DD1 and MM5 sensors did not participate in the swap-out experiment and remained at the same site for both deployments. The figures in row 2 plot explore cross site variations in no2diff (Fig. **??**(e) and (g)) and oxdiff (Fig. **??**(f) and (h)) values by considering data from the DM2 and MD6 sensors both of which participated in the swap-out experiment.~~

~~Our deployment strategy, consisting of two sites at geographically diverse locations and experiencing varying air pollution levels, two extended deployments during months experiencing significant variations in RH and T, as well as the swap-out experiment, were aimed at covering a wide range of real-world ambient working conditions (Cross et al., 2017). As we shall see, data from such diverse operating conditions is crucial for proper calibration of these sensors in order to not expect drastic extrapolations from the models during actual deployment.~~

~~To illustrate this, refer to Fig. ?? which shows the RH and T ranges observed during the two deployments across the two sites. It is clear that both sites offer extremely diverse meteorological conditions, with only site M offering somewhat uniformly high RH values during the Oct deployment. We also present in Fig. ??, time series over 24 hour periods from two deployments at the two sites.~~

~~The reference data for the site D Jun deployment indicates that levels exhibit a diurnal trend with a midday peak mainly at around 1500 hrs, while levels tend to peak usually in the morning and in the evening to midnight, suggesting nearby roadways could be a predominant source of pollution. Site M on the other hand presents far lower levels. Ambient RH and T values were observed to vary inversely to each other at site D in both deployments and site M during the Jun deployment. However, site M experienced a near continuous 100% RH level during the Oct deployment. The sensor voltages (no2op1, no2op2, oxop1, oxop2) can be seen to have good correlation in the plots.~~

~~The two sites and deployments also exhibit significant diversity with respect to absolute concentrations. The reference levels from site M (available at 15 minute intervals) ranged from 0.01-44.13 ppb in the Jun deployment and from 0.01-58.44 ppb in the Oct deployment, respectively. At the same time, the reference levels from site D ranged from 0.70-65.49 ppb and from 0.86-159.55 ppb during the Jun and Oct deployments, respectively. Similarly, reference levels also differ significantly across the sites with site M levels ranging from 0.70-65.49 ppb and 0.86-160.41 ppb during the Jun and Oct deployments respectively and those for site D ranging from 0.70-141.47 ppb and from 0.80-180.00 ppb for the same deployments. In general, Site D experienced higher concentration levels, as well as peaks, than site M. Furthermore, concentration levels were found to go up for both sites during the Oct deployment as compared to the Jun deployment. Such diversity in concentration levels are expected to empower calibration models to offer accurate predictions across wide ranges of operating conditions.~~

~~As deployments experienced several cloudy days, peaks of observed levels are not consistent throughout the deployments. Such influence of meteorological parameters on pollutant levels is well recognized in past literature (Gaur et al., 2014; Tiwari et al., 2015; S with effects such as scavenging of PM and gaseous pollutants that occur due to rain that may result in lower concentration peaks of PM2.5 levels (not considered in this study) and lower mixing ratio of , or higher range of concentrations of same pollutants during winter, being observed.~~

~~In order to better understand global trends in cross-site~~ . First we look at the performance of the algorithms on individual datasets i.e. when looking at data within a site and ~~cross-deployment variations, Fig. ?? plots histograms indicating the statistical distribution of reference values as well as sensor voltage readings for various sites and deployments. It is notable that both the reference values, as well as the sensor readings, seem to be statistically distributed across both sites and deployments, with the possible exception of levels at site D during the Oct deployment (see Fig. ?? Row 1 left) which seems to have a bimodal distribution.~~ 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.

~~These plots demonstrate that site D experiences appreciably greater levels for both and . This can be verified by comparing rows 1 and 2 of Fig. ??. This is understandable since site M is located in a coastal city whereas site D is situated at a more arid location. For both sites, in general the Oct deployment offers larger concentration levels as compared to the Jun deployment.~~

~~This is reflected in the plots in rows 3 and 4 of Fig. ?? which show that the distribution of the voltage differentials differs significantly when the same sensor is relocated to a different site during a different season.~~

5.1 Effect of Calibration Model on Calibration Performance

We compare the ~~calibration algorithms discussed~~ performance of calibration algorithms introduced in Sect. 4~~and also those in the Supporting Information document~~. Given the vast ~~set of models that we consider, we first compare within a family of algorithms (these comparisons are presented in detail in the Supporting Information document but summarized below as well) and present here, only comparisons across the winners of those families~~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. ~~See Fig. ?? for a brief description~~ We refer the reader to the supplementary material for some pointers on how to interpret ~~a violin plot.~~ violin plots.

~~**(Interpreting violin plots)** Two violin plots based on synthetic error data (i.e. the data does not correspond to any actual model) are shown above. Violin plots display numeric data by showing quartile/percentile information, as well as a rotated kernel density plot to show the distribution of the data. The left figure offers a *symmetric* violin plot on a single data source (calibration in this synthetic example). The thin vertical line in the middle represents the inter-percentile range between the 0.05 and 0.95 percentiles. The thicker and shorter vertical line represents the inter-quartile range between the 0.25 and 0.75 quartiles. The white dot in middle represents the median. The right figure offers a *split* violin plot that considers two data sources together for ease of comparison.~~

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 (~~top~~left), row 3 column 2 records a value of ~~0.46~~ 0.63 implying that in ~~46~~63% of these tests, KRR won over RT in case of O$_3$ calibration, whereas row 2 column 3 records a value of ~~0.24~~ 0.22 implying that in ~~24~~22% of the tests, RT won over KRR. In the balance (1 - ~~0.46~~ 0.63 - ~~0.24~~ 0.22 = ~~0.3~~0.15) i.e. ~~30~~15% of the tests, neither algorithm could be declared a winner.

5.1.2 ~~Inter-family~~ Intra-family Comparison of Calibration Models

~~The calibration models described in Sect. 4 and in the Supporting Information document can be classified in to four broad~~ We divide the calibration algorithms (see Table 2 for a glossary) into four families: 1) the Alphasense family (~~containing the four models A1 to A4~~AS1, AS2, AS3, AS4), 2) linear parametric models (LS, LS(MIN) and LASSO), 3) kernel regression models (KRR~~and the Nystroem method~~, NYS), and 4) ~~the *k*NN family including algorithms that use metric learning~~ KNN-style algorithms (KNN, KNN-D, NW(ML), ~~etc)– please see the Supporting Information document for details of algorithms not described here in the main paper such as LS(MIN) etc.~~

~~A summary of the results of comparing models and algorithms *within* these families is given below. The next section will compare the winners across these families to determine a *global* winner~~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 ~~R2~~ 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**: ~~We confirmed the utility of the Nystroem method as~~ The Nystroem method NYS was confirmed to be an accurate but accelerated approximation for KRR ~~kernel ridge regression (KRR) and that the acceleration is generally~~ with the acceleration being higher for larger datasets.

4. ~~***k*-NN**~~ **KNN and Metric Learning Models**: Among the ~~k-NN~~ KNN family of algorithms, the distance weighted ~~k-NN~~ 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

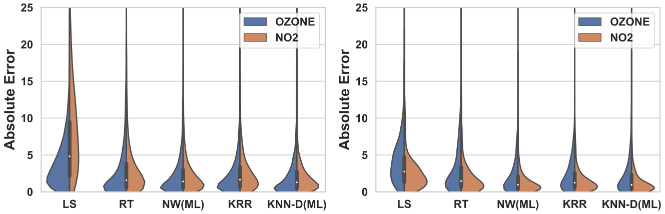

Figure 4. The violin plots on the left and right depict the distribution of absolute errors incurred by various models on respectively, the DD1(Oct) and MM5(Jun) datasets. KNN-D(ML) offers visibly superior performance than several other algorithms such as LS and RT.

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. (~~mostly 100% of the time~~ with the exception of NW(ML) which it still beats 58% of the time ~~)~~ on NO_2 and ~~is scarcely ever beaten.~~ 62% on O_3) The overall ranking of the algorithms is indicated to be KNN-D(ML) > NW(ML) > KRR > RT > LS.

	NO_2					O_3					
	LS	RT	KRR	NW(ML)	KNN-D(ML)		LS	RT	KRR	NW(ML)	KNN-D(ML)
LS	0	0	0	0	0	LS	0	0.01	0	0	0
RT	0.97	0	0.38	0.16	0	RT	0.83	0	0.22	0	0
KRR	1	0.4	0	0	0	KRR	1	0.63	0	0.01	0
NW(ML)	1	0.75	1	0	0.07	NW(ML)	1	0.97	0.96	0	0.02
KNN-D(ML)	1	1	1	0.58	0	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**

We took the best algorithms from all ~~families (parametric, kernel regression, k-NN) as well as regression trees~~ the families (except Alphasense models that gave extremely poor performance) and regression trees (RT) and performed a head-to-head comparison ~~for these~~ to assess the winner~~(Alphasense models were not considered given their extremely poor performance). The KNN-D(ML) algorithm continued to emerge as the winner as indicated by the~~ . The two-sample tests (Table 3) as well as ~~the~~ violin plots (Fig. 4) ~~.~~

~~The linear transformation $\Sigma^{\frac{1}{2}}$ learnt for calibration on the dataset DD1(Jun). Note the large emphasis the transformation places on RH and T, increasing their importance while calculating the Mahalanobis distance while placing comparatively less importance on the oxop1, oxop2 and oxdiff features which is understandable since this metric was learnt for calibration. TRHno2op1no2op2oxop1oxop2no2diffoxdiffT **10.19** 3.29 -1.95 -2.12 3.73 4.29 -0.66 -1.44 RH 3.52 **13.22** 1.43 1.46 -2.32 -2.60 -0.25 0.49 no2op1 -0.17 -0.69 **6.92** 6.20 -3.65 -3.93 0.27 -0.12 no2op2 -0.27 -0.81 5.66 **6.96** -2.94 -3.20 0.51 0.11 oxop1 1.27 -0.19 1.94 2.11 **1.51** 0.50 0.74 0.27 oxop2 0.89 -0.81 3.34 3.58 -0.86 **0.03** 0.86 0.24 no2diff -0.74 -0.68 -4.01 -3.94 6.89 7.12 **2.82** 1.88 oxdiff 2.71 3.45 -7.03 -7.36 7.95 8.54 -0.32 **1.32**~~

~~On the left hand side, the top (resp. bottom) figure exhibits a time series of the reference values and those predicted by the KNN-D(ML) algorithm for (resp.) concentrations at site D (resp. site M) during the Jun (resp. Oct) deployment. On the right~~

~~hand side are scatter plots showing the correlation between the reference and predicted values of the concentrations. For both deployments, KNN-D(ML) can be seen to offer excellent calibration and agreement with the FRM-grade monitor.~~

5.2 ~~The Effect of Metric Learning~~

~~Recall that in Sect. ??, we discussed the need for metric learning in order to place appropriate emphasis on various features, such as RH and T that are known to hugely influence calibration. To assess whether metric learning is indeed discovering such emphasis, Tab ?? shows the linear transformation corresponding to the Mahalanobis metric learnt by the NW~~indicate that KNN-D(ML) algorithm continues to emerge as the overall winner. Table 4 additionally establishes that KNN-D(ML) ~~technique for~~ 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 on the DD1(Jun) dataset. This is essentially the matrix $\boldsymbol{\Sigma}^{\frac{1}{2}}$ where $\boldsymbol{\Sigma}$ is the matrix corresponding to the Mahalanobis metric. We point out the following aspects of the matrix by concentrating on the diagonal entries. The diagonal entries corresponding to no2op1, no2op2 and no2diff have much higher values that those for oxop1, oxop2 and oxdiff. This makes sense since this metric was being learnt for~~ calibration ~~. The diagonal entries corresponding to RH and T are by far the largest. This implies that the method did find it crucial to put more emphasis on these two features while calculating distances.~~ which seems to be more challenging as compared to O_3 calibration. Fig. 5 presents two cases where the ~~models offered by metric learning~~ 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 ~~present studies~~ 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 ~~due to~~ during the swap-out experiment.

5.3 ~~Some Observations on Original Datasets~~

~~Before we proceed to perform studies with the temporally averaged, sub-sampled and aggregated datasets (see!3.1), first we look at the~~

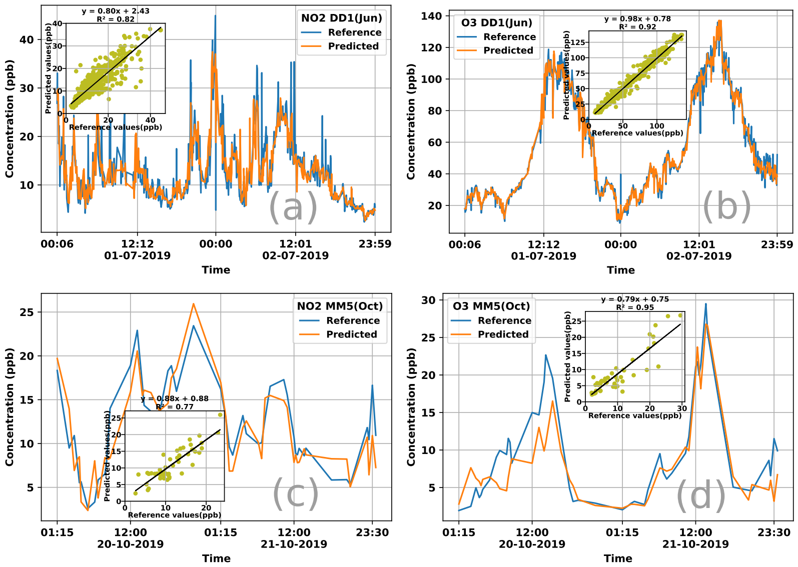

Figure 5. Time series for a duration of 24 hours of the reference values and those predicted by the KNN-D(ML) algorithm for NO_2 and O_3 concentration using data from the DD1 and MM5 sensors. The legend of each plot notes the gas for which calibration is being reported, the deployment season, as well 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.

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 ~~have been~~ are abbreviated~~here. For example~~, e.g. DD1(Jun)-AVG5 is referred to as simply AVG5. ~~These results~~ Results are reported over a single split. ~~The performance of KNN-D(ML) on~~ AVG5 wins over ~~its performance with~~ any other level of averaging ~~. It is clear~~ and clarifies that mild temporal averaging (e.g. over 5 minute windows) ~~positively impacts~~ boosts calibration performance~~. On the other hand~~, ~~the performance with extremely~~ whereas aggressive averaging e.g. ~~on AVG60 is almost always inferior than any other level of~~ 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

5.2.1 Some Observations on Original Datasets

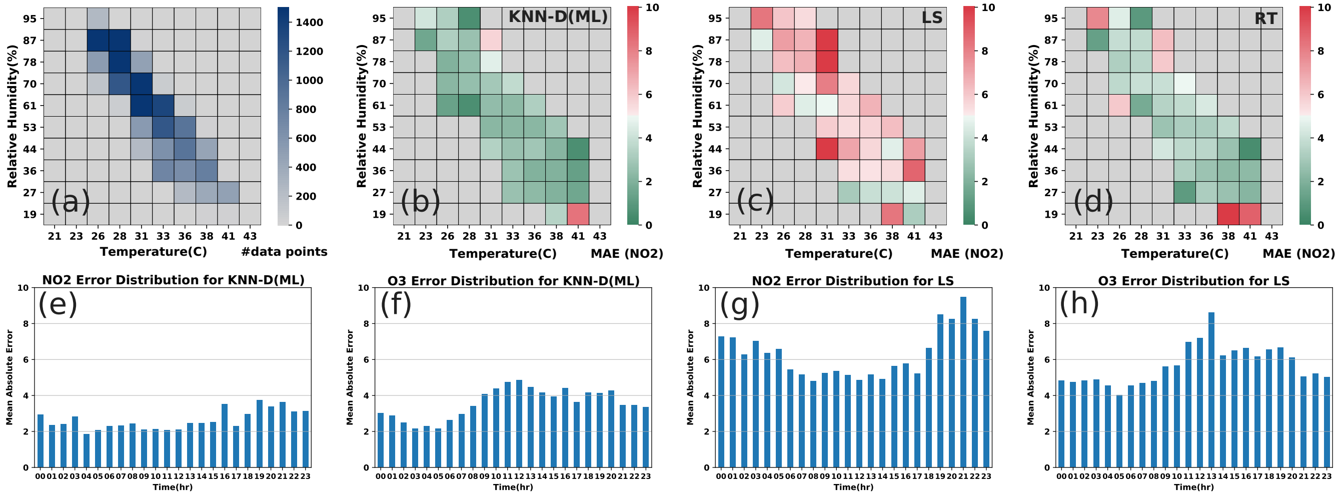

Figure 6. Analyzing error distributions of LS, KNN-D(ML, RT.. Fig 6(a) shows the number of training data points in various weather condition bins. Figs 6(b,c,d) show the MAE for NO_2 calibration offered by the algorithms in those same bins. Non-parametric algorithms such as KNN-D(ML) and RT offer poor 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.

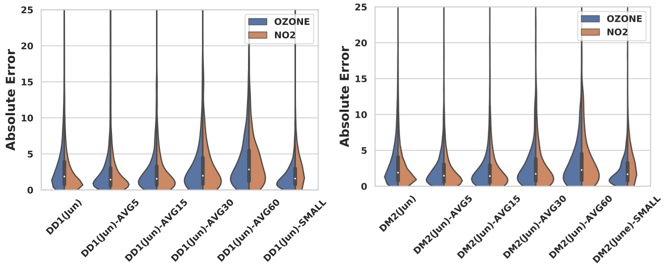

Figure 7. Effect of temporal data averaging, and lack of data on the calibration performance of the KNN-D(ML) algorithm on temporally averaged and sub-sampled versions of the DD1(Jun) and DM2(Jun) datasets. Notice the visible deterioration in the performance of the algorithm when aggressive temporal 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.

The performance of KNN-D(ML) on the original datasets ~~to gain some indications on the effects of these~~ itself gives us indications on how various data preparation methods ~~on~~ can affect calibration performance. ~~We will then confirm these indications using the new datasets. If we consider only datasets obtained from site D deployments, then we find that on these datasets (irrespective of whether in Jun or Oct), KNN-D(ML) offers an extremely high average R2 of 0.952~~ Table 4 shows us that in most cases, the calibration performance is better (with higher R^2) for O_3 ~~calibration~~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. ~~However, the same value for site M deployments (yet again across Jun and Oct deployments) is much lower at 0.762. We observe a similar but less stark difference~~ This difference is more prominent for NO_2 ~~calibration with site D deployments enjoying an average R2 of 0.915 with KNN-D(ML) whereas site M having only 0.846~~than for O_3. This indicates that paucity of data and ~~aggressive~~ temporal averaging may be affecting calibration performance negatively, ~~and more directly than seasonal variations. The above observations also indicate~~ as well as that O_3 calibration might be ~~more~~ less sensitive to these factors than NO_2 calibration.

5.3 ~~Effect of Temporal Data Averaging~~

5.2.1 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 (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 ~~is invariably hurt~~almost always suffers.

5.3 ~~Effect of Data Paucity~~

5.2.1 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.

5.3 ~~Effect of Data Volume and Diversity~~

5.2.1 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

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 ~~learnt~~ 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. ~~Also notable is the fact that~~ NO_2 calibration ~~seems~~ is worse affected by these variations (average ~~R2~~ R^2 in first four rows being ~~-3.68~~ -3.69) than O_3 calibration (average ~~R2~~ R^2 in first four rows being ~~-2.92~~ -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)	~~28.9±1.7~~ 21.82	~~-1.63±0.39~~ 0.19	~~33.1±0.94~~ 21.86	~~-0.87±0.08~~ -0.64	**12.88**	**0.73**	**12.73**	**0.22**
MM5(Oct) → (Jun)	~~8.9±1.69~~ **8.33**	~~-4.14±2.4~~ **-3.75**	~~15.9±2.9~~ **15.79**	~~-9.6±2.9~~ **-12.28**	10.39	-4.83	17.06	-21.67
DM2(Jun) → (Oct)	~~19.0±1.3~~ 13.04	~~-8.12±1.5~~ 0.41	~~17.1±0.97~~ 9.05	~~-0.45±0.12~~ -0.99	**9.36**	**0.68**	**5.95**	**0.1**
MD6(Jun) → (Oct)	~~18.8±0.57~~ **16.71**	~~-0.83±0.08~~ **-0.72**	~~29.6±0.85~~ 30.9	~~-0.77±0.09~~ -0.85	21.12	-1.29	**25.67**	**-0.23**
DD1(Jun-Oct)	~~3.3±0.14~~ **3.3**	~~0.939±0.006~~ **0.956**	~~2.7±0.06~~ **2.6**	~~0.958±0.003~~ **0.924**	11.7	0.29	13.0	0.38
MM5(Jun-Oct)	~~1.8±0.13~~ **2.5**	~~0.814±0.05~~ **0.902**	~~2.5±0.19~~ **1.8**	~~0.902±0.04~~ **0.814**	4.28	0.32	5.51	0.67
DM2(Jun-Oct)	~~3.7±0.13~~ **3.7**	~~0.909±0.009~~ **0.916**	~~3.0±0.02~~ **2.8**	~~0.762±0.008~~ **0.800**	6.13	0.79	6.72	0.26
MD6(Jun-Oct)	~~1.8±0.007~~ **1.9**	~~0.975±0.002~~ **0.989**	~~1.9±0.02~~ **1.8**	~~0.989±0.0006~~ **0.975**	7.01	0.71	6.36	0.91

to ~~poor~~ a drop in performance. We confirmed that this was true not just for ~~calibration models learnt using~~ non-parametric methods such as KNN-D(ML) but also parametric models like LS~~or LASSO or RT.~~

~~This finding, although concerning at first, seems reasonable when we observe Fig. ??. Not only do the sites and deployments individually span wide~~ . This is to be expected since the sites D and M experience largely non-overlapping ranges of RH and T ~~, but these ranges are not entirely overlapping either. Given our earlier confirmation of the importance these parameters have in calibration~~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 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 ~~excellent~~ 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 ~~a diverse field deployment of low-cost AQ monitoring~~ field deployments across two seasons of LCAQ sensors across two sites having ~~vastly different~~ diverse geographical, meteorological, and air pollution parameters~~, as well as two deployments set in seasons offering diverse RH and temperature conditions~~. A unique feature of our deployment was the *swap-out* experiment wherein ~~four of the seven~~ 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 ~~algorithms based on standard statistical estimation techniques~~ standard algorithms but found a novel method based on *metric learning* to offer the strongest results~~(as verified by statistical two-sample tests) across sites and deployment conditions at predicting both and concentrations.~~

. A few key takeaways from our statistical analyses are:

1. Incorporating ambient RH and T~~into the calibration model offers a definite advantage in achieving superior calibration performance. The inclusion of the *augmented*~~ , as well as the emphaugmented features oxdiff and noxdiff ~~we describe in~~ (see Sect. 3~~also positively impact the~~), into the calibration model improves calibration performance.

2. ~~Local~~ Non-parametric methods such as ~~*k*-NN~~ 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 ~~*k*-NN~~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 ~~this smoothing is non-aggressive e.g.~~ done over short windows. Very aggressive smoothing done over long windows is detrimental to ~~calibration~~ performance.

4. Calibration models are data-hungry as well as diversity hungry. This is especially true of local methods like ~~*k*-NN~~ 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. ~~calibration seems to be more sensitive to unseen variations in operating conditions than~~ 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 ~~low-cost AQ~~ LCAQ sensors to complement CAAQMS in creating dense and portable monitoring networks~~which can enable a range of studies in AQ, source apportionment, human health impacts and~~

~~atmospheric chemistry studies. Among avenues~~. Avenues for future work ~~, an especially interesting one is~~ include the study of long-term stability of electrochemical sensors and characterizing drift or deterioration patterns in these sensors and correcting for the same~~. Another interesting challenge is ultra~~, 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

Aggarwal, M.: NITI Aayog's action plan for air pollution: Old wine in a new bottle?, News article at Mongabay-India, https://india.mongabay.com/2018/07/niti-aayogs-action-plan-for-air-pollution-old-wine-in-a-new-bottle/. Accessed 01 April 2020., 2018.

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.

Chowdhury, S., Dey, S., Guttikunda, S., Pillarisetti, A., Smith, K. R., and Di Girolamo, L.: Indian annual ambient air quality standard is achievable by completely mitigating emissions from household sources, Proceedings of the National Academy of Sciences, 116, 10 711–10 716, 2019.

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.

Murphy, K. P.: Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.

Nadaraya, E. A.: On Estimating Regression, Theory of Probability and Its Applications, 9, 141–142, 1964.

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.

Times, H.: Faridabad second most polluted city in country, Gurgaon 12th, Newspaper article, https://www.hindustantimes.com/gurgaon/faridabad-second-most-polluted-city-in-country-gurgaon-12th/story-MJoViOnWDehKKw1QEEeLlO.html. Accessed on 01 April 2020., 2018.

Tiwari, S., Bisht, D., Srivastava, A., and Gustafsson, Ö.: Simultaneous measurements of black carbon and $\mathrm{PM}_{2.5}$, CO, and NO x variability at a locally polluted urban location in India, Natural Hazards, 75, 813–829, 2015.

Watson, G. S.: Smooth regression analysis, Sankhyā: The Indian Journal of Statistics, Series A, 26, 359–372, 1964.

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.