Responding to Reviewer 1 Comments

General Comments

Overall this a well-written, well-organized contribution to the low-cost sensor literature. The authors demonstrate the importance of bespoke sensor calibration using several techniques, including sensor clustering and various statistical and machine learning techniques. Sensor clustering reduces uncertainty due to inter-sensor differences and overall accuracy is improved using several statistical/machine learning techniques.

Specific Comments

C1. How accurate is 'good enough'?
I think a bit more context regarding this question would be helpful to readers.

C1. Author's Response

Thank you to the reviewer for the advice to be clearer when describing the requirements for sensor performance before they are considered able to perform as instruments in the field. The answer to this question is very application dependant and we cannot therefore provide a definitive statement on how good is good enough. It is for this reason that we included the comparison for NO_2 measured by two identical reference grade instruments in order to provide some reference for our comparisons. In order to try and expand on this point, we have added some text to the manuscript (page 3, lines 2 - 8) describing the standards set for reference grade instrument performance as set by the EU Directive 2008/50/EC, Annex 1(EU, 2008). Conforming to these standards is an obvious target for low cost sensor measurement performance, but that is not to say that reduced accuracy observations do not hold value providing the uncertainties are quantified.

C1. Changes to manuscript

Page 3, lines 2 - 8

For reference monitors in the UK, NO_x, CO and O_3 instruments must produce reproducible measurements for three months that are within 5% of the average for a certain concentration in the field, and results that are linear over a set range (EU, 2008). For NOx this is 0 – 2000 ppb and O_3: 0- 500 ppb and CO: 0 – 50 ppm to ensure that both rural and urban concentration ranges are taken into account. Although the target performance of low-cost sensors is highly application dependent, these standards do provide a benchmark for comparison and highlight the need not only for high accuracy measurements but also reproducibility over long (months) timescales. In order for low cost sensors to be used in atmospheric monitoring or research applications the uncertainty and reproducibility must be quantified across a range of likely environmental conditions.

C2. Overall, using SLR and ML techniques seems to be the largest source of improvement. Is sensor clustering even necessary?

C2. Author's response

Ultimately the clustering and statistical calibration methods are performing different functions in improving sensor performance. In terms of measurement accuracy, the SLR and ML calibration algorithms provide significant improvement over simple linear regression, due to their ability to correct for the multiple cross-sensitivities on sensor signals. In contrast, the function of the clustering approach is not to improve measurement accuracy, but rather sensor reproducibility. As shown in our previous paper (Smith et al., 2017) many sensors show variability in both signal and sensitivity over timescales of days or longer. This variability is very difficult to remove through time averaging, however the lack of correlation of this noise / drift between identical sensors means it can be addressed by instead averaging over multiple sensors. The conclusion of Smith (2017) was that clustering greatly reduces medium-term random noise in the average sensor signal, thus improving confidence in sensor signals and in theory prolonging the time requirements between calibrations.

The time series below illustrates how sensor signals drift apart over time. The plot on the left shows all six sensor signals immediately after calibration to a reference monitor, showing a tight clustering around the median value (red). The plot on the right however shows the drift in individual sensors after a period of 16 days. The use of an average sensor reduces some of this signal variability enabling a more robust calibration to be applied, using algorithms such as SLR or Gaussian Process etc.

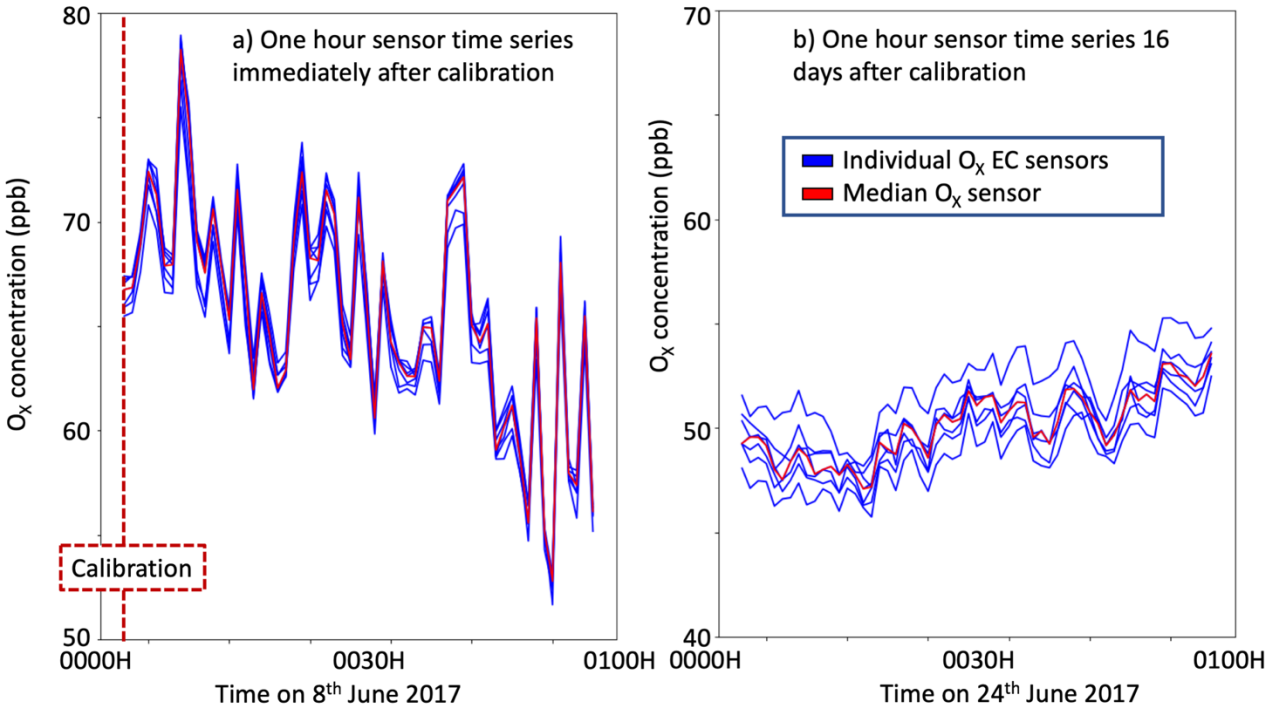


Figure 1. Six individual O_X EC (blue) with the median O_X EC (red), a) immediately after SLR calibration with the reference observations and b) 16 days after the calibration.

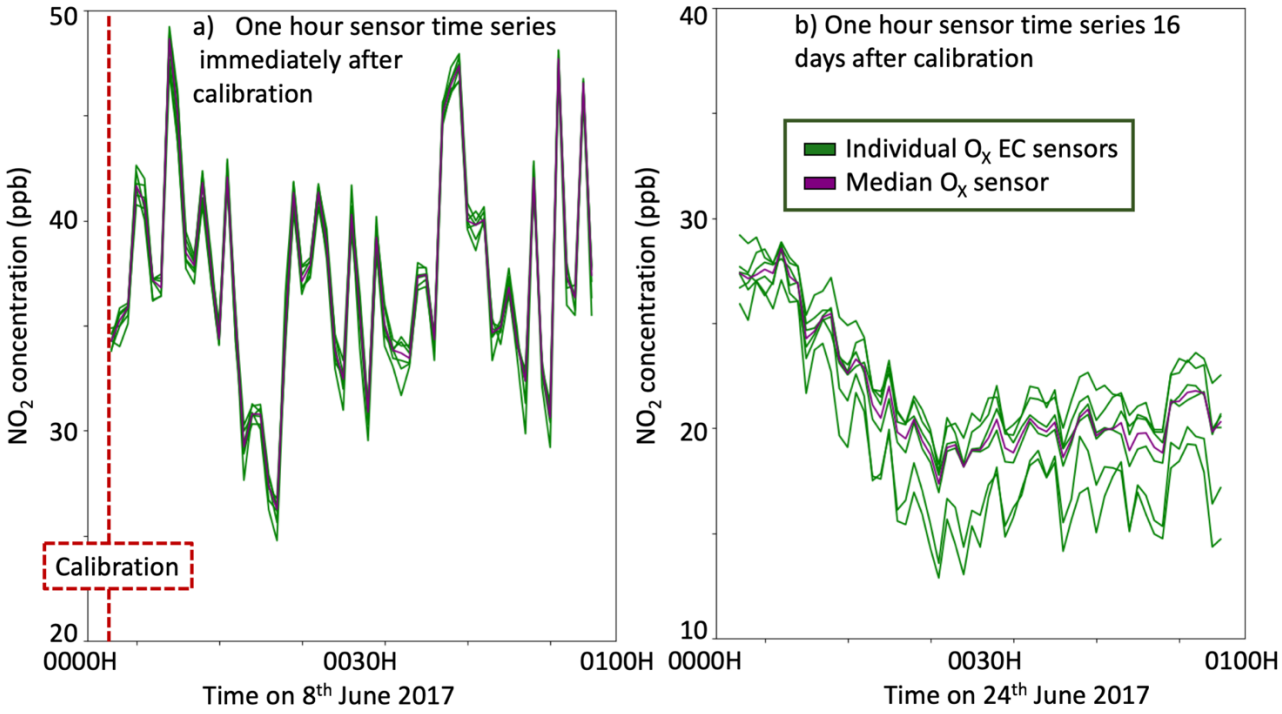


Figure 2. Six individual NO_2 EC (green) with the median NO_2 EC (purple), a) immediately after SLR calibration with the reference observations and b) 16 days after the calibration.

The following text was added to the manuscript to emphasise that clustering and ML are used to target different issues. Page 4, line 26 -27.

C2. Changes to manuscript

The clustering approach was used to improve sensor reproducibility as previously discussed in (Smith et al., 2017), whereas the SLR and ML techniques were applied to improve sensor accuracy by correcting for cross sensitivities.

C3. Does sensor accuracy vary over the observed concentration ranges?

C3. Author's response

This is an excellent question by the reviewer, and we have performed additional analysis below to investigate this.

For each calibration method used in the paper, the data was 25 % of the observed reference concentration range bins. The Root Mean Squared Error (RMSE) and the Normalised Root Mean Squared Error (NRMSE) were calculated for each concentration

bin and the results for NO_2 and O_X are summarised in the tables below. The NRMSE was calculated by dividing the RMSE between the reference observations and the sensor values by the mean reference concentration for the respective bin.

Table 1 and 2 nicely displayed how the different analytical techniques improved the sensor performance at different concentrations. Therefore, we decided to include Table 1 (page16) and a description of the results in the manuscript summarising the NO_2 RMSE and NRMSEs. The O_X summary was very similar so wasn't included, but the authors are happy to include it if the editor wishes.

C3. Changes to manuscript

Table 1 The NRMSE and RMSE between the NO_2 reference and sensor data sets at different concentrations ranges. For each calibration method used in the paper, the data was binned into 25% of the observed reference concentration. The Root Mean Squared Error (RMSE) and the Normalised Root Mean Squared Error (NRMSE) were calculated for each concentration bin and the results for NO_2 and O_X are summarised in the tables below. The NRMSE was calculated by dividing the RMSE between the reference observations and the sensor values by the mean reference concentration for the respective bin.

	NRMSE of Reference vs. NO_2 concentration estimate (RMSE / ppb)				
Concentration range as a % of the max. conc. of reference NO_2	Median	SLR	BLR	BRT	GP
0 - 25 %	1.04 (20.7)	0.59 (11.7)	0.32 (6.3)	0.28 (5.6)	0.29 (5.8)
25 - 50 %	0.69 (47.5)	0.19 (13.3)	0.12 (8.2)	0.22 (15.2)	0.11 (7.9)
50 - 75 %	0.72 (94.9)	0.23 (30.8)	0.26 (34.6)	0.55 (72.5)	0.26 (33.5)
75 - 100 %	0.85 (153.1)	0.10 (17.4)	0.10 (18.8)	0.67 (120.0)	0.10 (18.2)

Page 9, line 18 - 26

The RMSE and NRMSE was calculated after the application of SLR and ML for different reference concentration ranges to indicate where the greatest improvement of the sensor data occurred (see Table 2). The RMSE and NRMSE (calculated by dividing the RMSE by the mean of the concentration bin) were determined between the reference NO_2 observations and the sensor values for four equally spaced reference concentration bins. The ML techniques produced the greatest improvements in the concentration estimates for the lower concentrations of the target measurand where the effect of cross interferences is more significant. The BRT and GP in particular displayed large improvements for the lower NO_2 reference observations. At the higher concentrations of NO_2, the ML algorithms displayed less improvement, where the conditions were outside those of the training data variable space. This was very noticeable for the BRT algorithm due to its inability to extrapolate.

C3. Author's response

Table 2. The NRMSE (and RMSE) between the O_X reference and sensor data sets at different concentration ranges.

	NRMSE of Reference vs. O_X concentration estimate (RMSE / ppb)				
Concentration range as a % of the max. conc. of reference O_X	Median	SLR	BLR	BRT	GP
0 - 25 %	0.21 (11.0)	0.16 (8.4)	0.10 (5.4)	0.12 (6.0)	0.18 (9.2)
25 - 50 %	0.30 (26.4)	0.12 (10.2)	0.11 (9.4)	0.11 (9.7)	0.14 (12.4)
50 - 75 %	0.36 (50.4)	0.12 (16.3)	0.12 (16.1)	0.10 (14.0)	0.16 (22.4)
75 - 100 %	0.52 (116.1)	0.20 (44.7)	0.26 (58.0)	0.49 (110.9)	0.27 (60.6)

C3. Changes to manuscript

Page 10 line 8 - 11.

The NRMSE was calculated for 4 equally sized reference O_X concentration bins for each analytical method used, in a similar manner to Table 2 for NO_2. The NRMSE improved for SLR and the ML algorithms across all concentration ranges, with BLR and BRT optimal for reducing the error estimate the most. The error was the highest at the higher O_X concentrations for BRT, which was expected due to BRTs inability to extrapolate.

C4. Technical Comments p1 l30. 'site'->situated

C4. Authors response

The wording has been changed on page 1, line 32.

C5 . p2 l8, l19, l21... Check reference parentheses throughout p4 l25.

C5. Author's response

Removed the extra brackets between multiple references and inserted a semi-colon to differentiate two citations for the same reference.

C6. Also Hagan et al. AMT 2018

C6. Author's response

Added the reference into the manuscript on page 5, line 3 as it was relevant to the manuscript.

References

EU: Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner

air for Europe, Eur. Union, 1–62, 2008.

Smith, K., Edwards, P. M., Evans, M. J. J., Lee, J. D., Shaw, M. D., Squires, F., Wilde, S. and Lewis, A. C.: Clustering approaches that improve the reproducibility of low-cost air pollution sensors, Faraday Discuss., 00(0), 1–17, doi:10.1039/C7FD00020K, 2017.

Responding to Reviewer 2 Comments

General Comments

Low cost sensors (LCS) playing an emerging role in the urban environmental monitoring with respect to the possibility of setting up a densely populated gridded network. Nevertheless, the detection limit, the stability and the real-time calibration were in general of question or with difficulty to overcome. In this study, the authors try to use the machine learning (ML) method to enhance of the data quality of LCS which is in general fit the effort of the community to improve the data quality of LCS. The paper is within the scope of AMT and I have the following comments for the authors to consider before publication.

C1. The machine learning method is used to improve the data quality of the LCS. The improvement is clear but still without in-depth explanations. The scientific paper shall not be looks like simply magic. I will be convinced if the authors can provide much more examples as the authors also wrote in their conclusions. Moreover, I did see much better comparison results from LCS (the Cambridge group for the same campaign) with the CAPS instrument on NO2 and other parameters like O_3, CO, etc. So, I wonder if the results presented in this paper can be improved further.

C1. Author's response

We thank the reviewer for this comment, as this is something we really wanted to avoid and have therefore added more detail in order to try and be more explicit about this. During the analysis section of this work the authors made sure that the ML techniques used provided outputs on the decisions they made that could then be compared with laboratory experiments and previous sensor studies, in order to make the methods used not seem like black boxes.

This was underpinned by our choice of ML techniques; BRT was chosen because of the function to extract out the variables gain contributions, GP could produce the uncertainty for each predicted data point and the weights associated with each variable can be extracted from the BLR algorithm (see C4 with Figs. 4 and 5 of this document). The manuscript aimed to compare results from using different techniques on the same dataset and therefore, fully comprehensive explanations of different ML techniques is beyond the scope of this paper. However, we recognise that this was not made clear enough in the manuscript so have added some more detail in the text and some more citations to detailed descriptions of the techniques.

With respect, the authors have not seen any publications from the Cambridge group on this and cannot comment on unpublished work.

Whilst there are many references to studies where LCS have been used successfully in the field, the scope of this manuscript relates to the improvement of low-cost sensor performance for deployment over longer periods of time and possible calibration strategies that would enable this.

C1. Changes to manuscript

Gaussian process reference inserted:

Gaussian process for time series modelling, S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson and S. Aigrain (Roberts et al., 2013)

Page 7, line 27.

XGBoost reference inserted:

Greedy function approximation: a gradient boosting machine (Friedman, 2001).

Page 7, line 17.

C2. Sect. 3.2: during the training period, what kind of regression method is used to calibrate the sensors? According to Cantrell, 2008 (Cantrell, C. A.: Technical Note: Review of methods for linear least-squares fitting of data and application to atmospheric chemistry problems, Atmos. Chem. Phys., 8, 5477–5487, 2008.), bivariate regression algorithm is required to retrieve the robust slope.

C2. Author's response

Thanks for bringing to our attention that the SLR method described in Section 3.2 was unclear. More text has been added to better describe the linear regression process. There were four different types of analytical techniques used in turn to examine the performance of ML versus simple linear regression (SLR). In section 3.2, SLR was used to calibrate the EC sensors against their respective reference instruments.

Using NO_2 as an example, linear parameters in the form of $y = mx + c$ were determined using a linear least squares fit between the NO_2 CAPS reference instrument and the median NO_2 EC sensor. This linear relationship was calculated for the first five days of the deployment – the same five days that were used as the training period for the ML analysis.

Text has been added on Page 6 lines 6 – 10, to provide further detail about SLR.

The training period for the NO_2 EC sensors and NO_2 reference measurements was also re-analysed using bivariate regression (Ordinary Least Squares) and the resulting model was applied to the median NO_2 sensors over the testing period. This produced a bivariate regression NO_2 prediction, shown in green in Figure 1. The regression was performed using the Python statsmodels package.

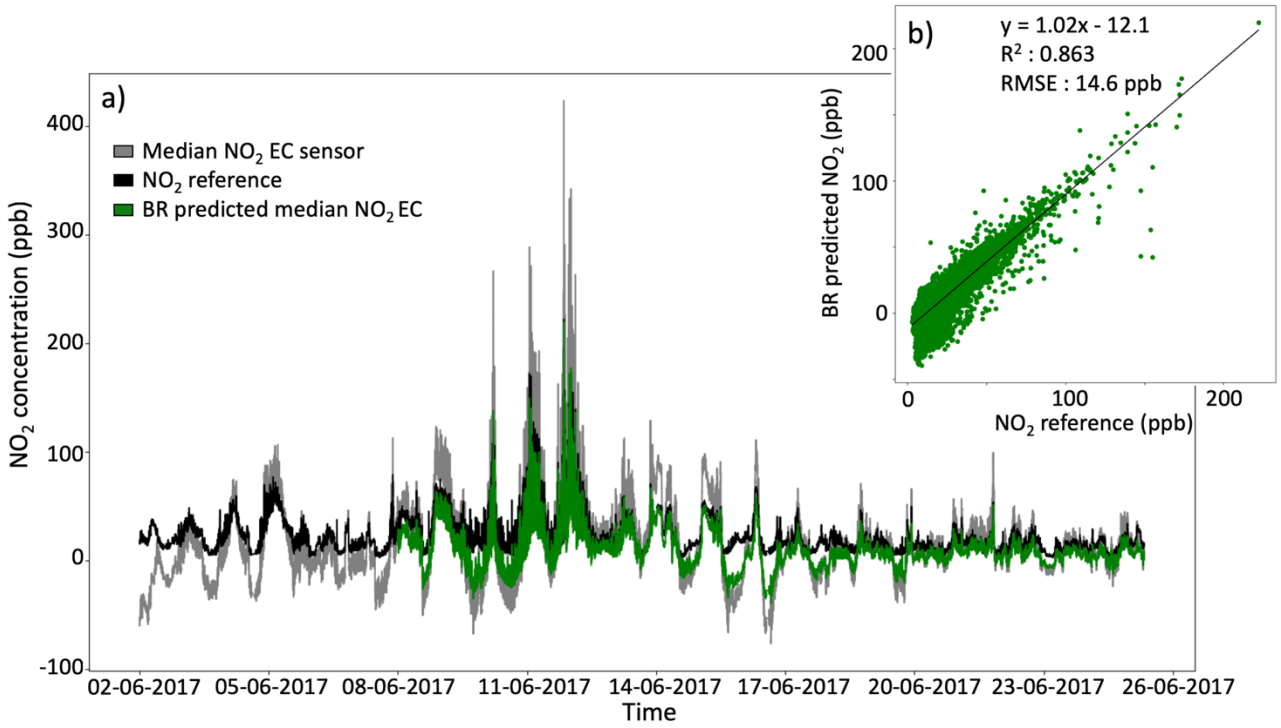


Figure 1. Bivariate regression – Ordinary Least Squares- was performed on the median NO_2 EC sensor (grey) and the NO_2 CAPS measurement during the training data set. and the resulting model was applied to the median NO_2 to produce a bivariate regression (BR) predicted trace (green).

The RMSE was calculated between the BR NO_2 prediction and the NO_2 reference measurements in the testing period and was found to be 14.6 ppb. The bivariate regression prediction therefore contains more error than the simple linear regression in the manuscript (10.42 ppb). This does not change the outcome of the paper which aims to use ML to improve the quality of the NO_2 sensors by correcting for cross interferences and therefore has not been added to the manuscript. The SLR was used to show the calibration of the sensors using linear regression and essentially set a baseline for the improvements.

C2. Changes to manuscript

Using the NO_2 EC as an example, linear parameters in the form of $y = mx + c$ were determined using a linear least squares fit between the NO_2 CAPS reference instrument and the median NO_2 EC sensor for the first five days of the sensor instrument

deployment. Once trained in this manner, these linear calibration factors based on SLR were used to calibrate the median NO_2 sensor and were unchanged for the remainder of the experiment.

C3. Figure 4, Panel A is with linear scale, Panel B-D is with logarithmic scale. Why the authors want to have two different scales?

C3. Author's response

Figure 4a uses a linear scale to compare the uncalibrated median NO_2 EC sensor to the co-located reference NO_2 measurement. The NO_2 sensor signal differed from the reference measurement sufficiently to allow this to be on a linear scale and to show the reader that the NO_2 sensor was able to detect the general trend of the NO_2 concentration patterns, but that there was still a large amount of discrepancy between the two measurements.

However, plotting Fig. 4 b to d) on logarithmic axis shows the fit of the calibrated median NO_2 sensor with the reference measurement. The improvement of the NO_2 measurement across the deployment means that it was difficult to identify times where the concentration estimate contained more error and uncertainty, but the log scale shows this clearly. These higher-error/more uncertain measurements could then be justified by identifying when other variables exhibited measurements that were outside of their training period ranges.

We would therefore like to keep the figure in its current state but are happy to change at the editor's request.

C4. Figure 5 is a nice way to explain the advantage from the ML method. Can the authors do the same for the other ML processing?

C4. Author's response

The Boosted Regression Tree gain contributions for each variable was a major reason for using this as a calibration algorithm, and we are glad the reviewer liked Fig. 5. The gain contributions were also analysed for the O_X EC BRT algorithm and have been added to the manuscript. This function of the BRT was advantageous as it allows the user to identify the key variables that impact the sensor signals which can then be compared with prior knowledge from laboratory experiments and other studies, thus removing some of the "black box" nature of these algorithms.

C4. Changes to manuscript: Gain contributions from O_X BRT added to Figure 5 on page 21.

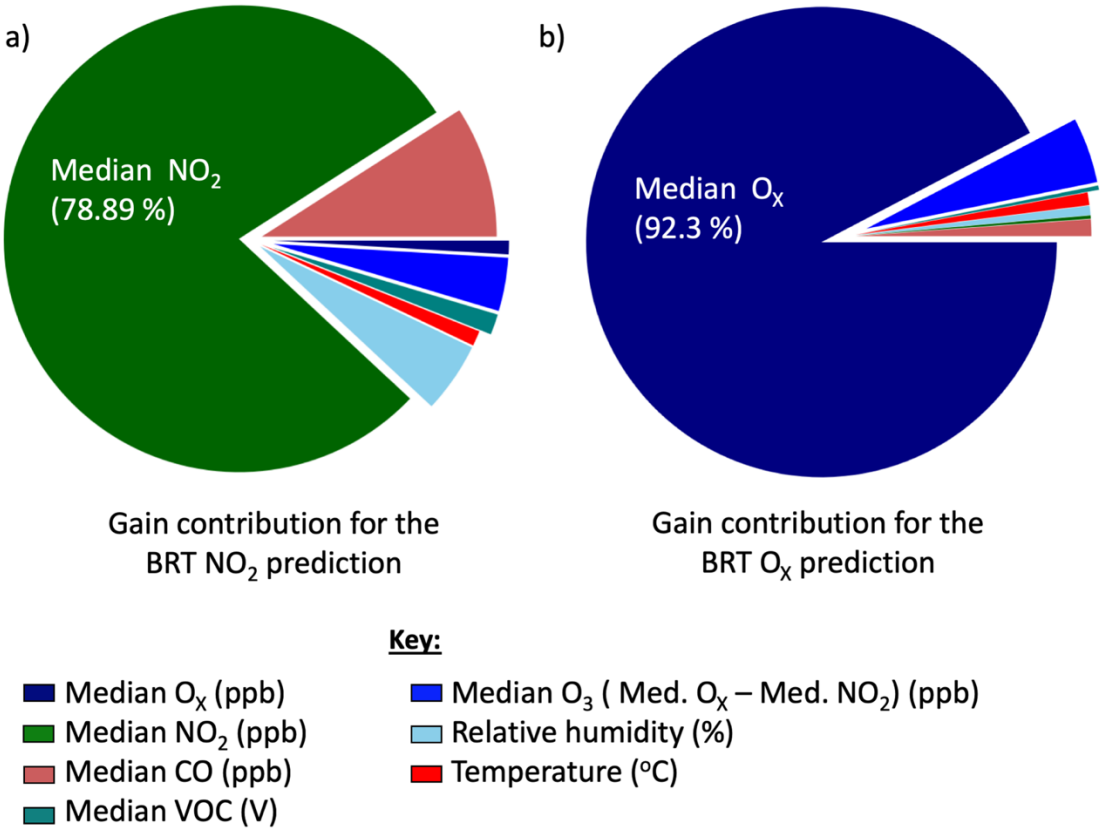


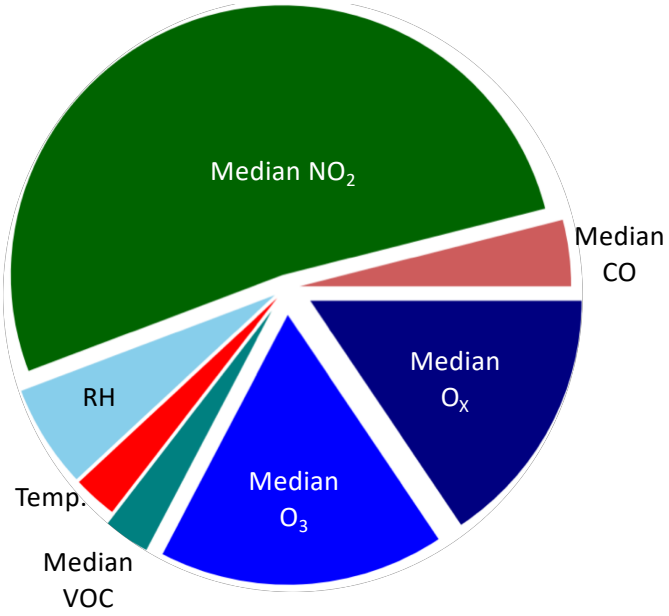
Figure 5: Breakdown of contribution from each variable used by the BRT algorithm to predict the clustered a) NO_2 sensor and b) O_X concentrations.

C4. Author's response

Please note that the values used in the pie chart for the NO_2 concentration estimate gain contributions have been changed slightly. Whilst adding in the O_X pie chart the authors noticed that the previous NO_2 plot was an old version, and this has now been changed to the most up-to-date chart. Where cited in the manuscript, values relating to these plots have been updated accordingly.

It is unfortunately not possible to extract the same information from the Gaussian Process implementation that was used in this work. This approach does however provide a prediction uncertainty, see Fig. 4b, which is very useful when interpreting the predicted concentrations, in particular when they move into variable space outside of that experienced during the model training dataset.

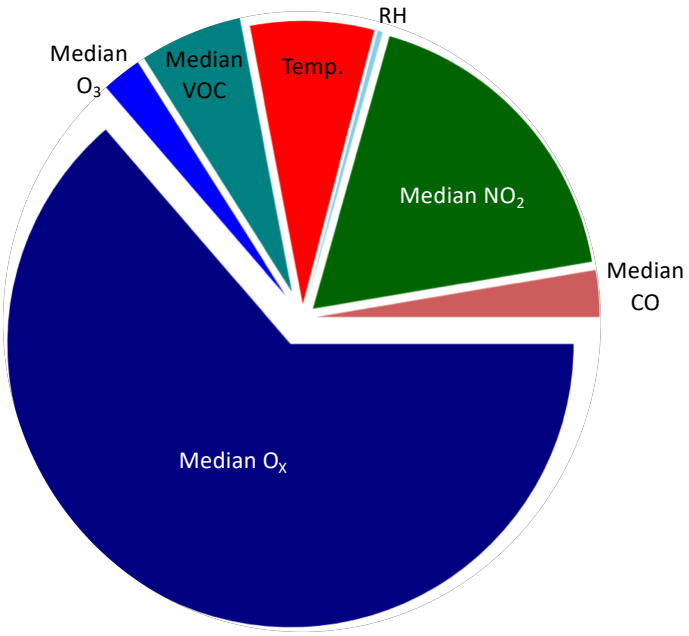
Linear regression weights for variables can be extracted from the BLR algorithm. However, to make assumptions about the relative importance of each of the sensors to the algorithm, all the variables, including the reference observations were normalised to between 0 – 1. The BLR analysis was then repeated with the normalised data. This does not change the algorithm and the concentration estimates were identical to those used in the manuscript after the normalisation process.



Variable	BLR weight
Median VOC sensor	0.039
RH	-0.087
Temperature	0.037
Median NO_2 EC	0.73
Median O_X EC	0.22
Median O_3 EC	-0.24
Median CO EC	0.056

NO_2 bias: -0.489437

Figure 3. Weights for the BLR-predicted NO_2 concentration, with normalised variables prior to analysis.



Variable	BLR weight
Median VOC sensor	0.11
RH	0.0055
Temperature	-0.13
Median NO_2 EC	-0.33
Median O_X EC	1.17
Median O_3 EC	0.043
Median CO EC	0.049

O_X bias: -0.484464

Figure 4. Weights for the BLR-predicted O_X concentration, with normalised variables prior to analysis.

The resultant weights can be used to indicate that, for the NO_2 BLR algorithm, the linear function describing the NO_2 sensor measurement contributed the most to the BLR algorithm. Equally, the linear component of the O_X sensor measurement was the most important variable for determining the BLR algorithm when predicting the O_X concentration estimate. The ability to extract these weights from the BLR analysis is useful for identifying relationships between the sensors, yet this was not included in the manuscript because it overcomplicated the analysis.

These weights output from BLR should not be directly compared to the gain contributions extracted from the BRT, because they are different metrics. The weights from BLR examine the linear relationships between variables whereas the gain contribution from BRT analysed the degree to which each variable contributed to the regression tree decisions – this includes non-linear functions.

If the editor wishes, this can be included in the manuscript.

C5. The ML corrected LCS signal still significantly smaller than those measured by the reference instruments especially for the peak values of Ox? Could the authors provide more discussions on this aspect and what could be the possible improvements on LCS or ML.

C5. Author's response

Machine learning techniques are very powerful at data interpolation, but often fail when it comes to extrapolation beyond the training data variable space. It is for this reason that linear models can often out perform some ML techniques when using small training datasets. The performance of the ML techniques can be greatly improved by randomly distributing the training

data throughout the full timeseries in order to cover more variable space. However, this is not a realistic calibration strategy for low cost sensors and so was not pursued in this work.

In Figure 6, at peak [O_X] the BRT ML corrected O_X concentration estimate does sometimes under-predict the concentration of O_X, compared to the reference measurement. This is due to the median O_X EC sensor reporting values at these times that are slightly higher than the maximum [O_X] observed by the median O_X EC sensor during the training period. The inability of the BRT algorithm to extrapolate caused the BRT predicted [O_X] estimate to be lower than the reference measurements, in a similar manner to the BRT NO_2 prediction. To improve the comparison between the BRT O_X concentration estimate and the O_X reference measurements more training data is required. This will ensure that the concentration range of [O_X] as measured by the EC sensors in the testing period is within the [O_X] range in the training data. This was summarised by a few lines which were added to the text on page 9, lines 32 -34.

C5. Changes to manuscript, page 9, lines 32 -34

The ML technique with the lowest RMSE, BRT, bought the O_X concentration estimate much closer to the reference observations, see Fig. 6, however, during peaks in O_X concentration, the BRT predicted O_X concentration estimate was underpredicted due to BRT's inability to extrapolate.

C6. Technical comments: **In most cases, the multi-citations were not correctly implemented. For example, page 2 line 8, (Caron et al., 2016),(Jiao et al., 2016) should be (Caron et al., 2016; Jiao et al., 2016). This shall be revised throughout the paper.**

C6. Author's response

Thanks for notifying the authors about this error, this issue was addressed above in the Reviewer 1 Technical Comment 2.

C7. Figure 3 is not cited in the main text which I assume should appear somewhere in Sect. 3.2.

C7. Author's response

Figure 3, showing how increasing the number of EC sensors from 1 to 6 within a cluster improves the agreement between the reference measurement and the median sensor signal, was cited within the manuscript in section 3.1 on page 5, lines 34.

References

Friedman, J. H.: Greedy function approximation: a gradient boosting machine, Ann. Statisitcs, 29(5), 1189–1232, 2001.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N. and Aigrain, S.: Gaussian processes for time-series modelling., Philos. Trans. A. Math. Phys. Eng. Sci., 371(1984), 20110550, doi:10.1098/rsta.2011.0550, 2013.

Where changes to the manuscript have occurred, they have been highlighted using green text.

An improved low power measurement of ambient NO_2 and O_3 combining electrochemical sensor clusters and machine learning

Kate R. Smith*[1], Peter M. Edwards[1], Peter D. Ivatt[1], James D. Lee[1,2], Freya Squires[1], Chengliang Dai[1], Richard E. Peltier[3], Mat J. Evans[1,2], Yele Sun[4], Alastair C Lewis[1,2]

[1]Wolfson Atmospheric Chemistry Laboratories, University of York, York, YO10 5DD, United Kingdom
[2]National Centre for Atmospheric Science, University of York, York YO10 5DD, United Kingdom
[3]Environmental Health Science, University of Massachusetts, 686 North Pleasant Street Amherst, MA 01003, USA
[4]State Key Laboratory of Atmospheric Boundary Layer Physics and Atmospheric Chemistry, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

Correspondence to: Peter M. Edwards (pete.edwards@york.ac.uk)

Abstract. Low cost sensors (LCS) are an appealing solution to the problem of spatial resolution in air quality measurement, but they currently do not have the same analytical performance as regulatory reference methods. Individual sensors can be susceptible to analytical cross interferences, have random signal variability and experience drift over short, medium and long timescales. To overcome some of the performance limitations of individual sensors we use a clustering approach using the instantaneous median signal from six identical electrochemical sensors to minimise the randomised drifts and inter-sensor differences. We report here a low power analytical device (< 200 W) that comprises of clusters of sensors for NO_2, O_X, CO and total VOC, and that measures supporting parameters such as water vapour and temperature. This was tested in the field against reference monitors, collecting ambient air pollution data in Beijing, China. Comparisons were made of NO_2 and O_X clustered sensor data against reference methods for calibrations derived from factory settings, in-field simple linear regression (SLR) and then against three machine learning (ML) algorithms. The parametric supervised ML algorithms boosted regression trees (BRT) and boosted linear regression (BLR) and the non-parametric technique Gaussian Process (GP) used all available sensor data to improve the measurement estimate of NO_2 and O_X. In all cases ML produced an observational value that was closer to reference measurements than SLR alone. In combination, sensor clustering and ML generated sensor data of a quality that was close to that of regulatory measurements (using the RSME metric) yet retained a very substantial cost and power advantage.

1 Introduction

Low cost sensors (LCS) are an attractive prospect for use in complex urban environments where more atmospheric measurements are required to build up a better resolved map of highly heterogeneous pollution patterns. There are numerous reports of low-cost, low-powered sensors commercially available for most of the criteria pollutants. Air pollution measurement has been historically a heavily regulated analytical environment. Many countries have extensive programmes of air quality measurement, and measurements often situated within a legal framework with prescribed methods of measurement. Air quality monitoring stations use relatively power intensive equipment, have a high start-up cost and require skilled personnel for calibration and maintenance. A consequence is that, even in wealthy countries, observations are sparse with sites often located

1–10 km^2 apart (McKercher et al., 2017). Pollutants often exhibit steep spatial concentration gradients over short distances (Broday et al., 2017) and limited measurement locations mean hotspots are often missed (Mead et al., 2013).

LCS provide an opportunity to increase the density of atmospheric measurements and reduce the uncertainty that arises when interpolating between current reference monitors. This has many uses, most notably allowing better validation of atmospheric models (Broday et al., 2017). The lower power and size associated with LCS, along with high frequency measurements, makes them an attractive prospect for mobile use and for personal exposure assessment (Williams et al., 2013). Many low-cost sensors are commercially available, either as stand-alone sensors or as multisensory platforms (Caron et al., 2016; Jiao et al., 2016) (for example, AQMesh (Broday et al., 2017)). There has been a rapid expansion in the number of publications evaluating such devices recently. Single devices containing sensors for the measurement of criteria pollutants such as CO, NO_2, total VOC and O_3 cost a fraction of the price (sensor box approx. cost: £5k) of establishing an equivalent measurement site with reference instruments (Mead et al., 2013) (£200k). Perhaps more importantly sensors can be placed in locations where power is limited or can only be generated through solar resources. The operating costs of low power devices also are a very attractive feature.

However, the literature contains many examples of where LCS approaches can suffer from relatively poor analytical performance, when compared against reference instruments. Whilst such a comparison is perhaps not always appropriate to make in such a highly regulated field of measurement, the benchmark test of any new analytical device will be against the regulatory reference. Significant uncertainty in measurements is introduced because individual sensors each have a unique response to simple environmental conditions such as humidity and temperature (Smith et al., 2017; Moltchanov et al., 2015). This can lead to a relatively high degree of inter-sensor variability and response drift (Lewis et al., 2016; Spinelle et al., 2017) over durations as short as a few hours (Jiao et al., 2016; Masson et al., 2015), rendering in-laboratory calibrations (where the interfering variables are controlled or non-existent) ineffective (Smith et al., 2017). Electrochemical (EC) sensors can display some chemical cross-interferences with other pollutants that are likely to be present (Mead et al., 2013; Lewis et al., 2016; Masson et al., 2015), and accounting for these can be difficult when the relative concentration ratios of the target measurand and interferences change. Metal oxide sensors often lack selectivity and provide only a rough bulk measure of a particular pollutant class such as VOCs, and the responses generated can depend on the chemical of the mixture presented to the sensor.

Although some LCS vendors supply a factory calibration with their sensors, these are not always applicable in the real-world, where ambient conditions are substantially different to the calibration conditions in the factory. Previous studies have shown that sensors co-located with reference instruments can be used to reproduce typical pollution patterns (Jiao et al., 2016; Mead et al., 2013) but there is a significant challenge when attempting to calculate absolute pollutant concentrations with a single deployed sensor device. Recent efforts using multivariate regression models (Zampolli et al., 2004) and pattern recognition analysis (Jiao et al., 2016) have characterised these responses to the environmental conditions and provided insight into processes that generate the sensor signal (Zampolli et al., 2004; Hong et al., 1996). Thus far, there are no agreed standard calibration or correction procedures for sensor data, or indeed what data standards low cost sensor data should work towards.

For reference monitors in the UK, NO_X, CO and O_3 instruments must produce reproducible measurements for three months that are within 5% of the average for a certain concentration in the field, and results that are linear over a set range (EU, 2008). For NO_X this is 0 – 2000 ppb, for O_3: 0- 500 pbb and CO: 0 – 50 ppm to ensure that both rural and urban concentration ranges are taken into account. Although the target performance of low-cost sensors is highly application dependent, these standards do provide a guide for comparison and highlight the need not only for high accuracy measurements but also reproducibility over long (months) timescales. In order for low cost sensors to be used in atmospheric monitoring or research applications the uncertainty and reproducibility must be quantified across a range of likely environmental conditions.

If regulatory reference methods are taken as the benchmark, the implication with current single sensors would be very frequent calibration, possibly hourly or daily. Previous work shows that clustering sensors and using the median sensor signal of the cluster can help minimise some of the effect of medium-term noise and limit the effects of inter-sensor variability (Smith et al., 2017). This practice was adopted here during the building and development of a multi-sensor instrument deployed alongside reference instruments.

2 Experimental

2.1 Analytical description of the instrument

A range of different sensors were mounted into sealed flow-cells such that the sensing element of each was exposed to a continually flowing sample of air. The flow cells were in turn installed inside in a 4U aluminium box (177 mm H x 460 mm D x 483 mm W), which had a metal partition to keep the sensors shielded from electrical interference from the pumps and power supplies (Fig. 1). The number of sensors and their type are shown in Table 1.

Two microcontrollers (Arduino Uno) were used to collect the data from the sensors. Each Arduino recorded 3 Hz data from 25 sensors, and this was then averaged to 2 seconds and sent to a Latte Panda mini-computer for formatting and storage. Two KNF pumps drew ambient air through a sample line at atmospheric pressure over the sensors at a constant rate (c.a. 4 L min^{-1}). Two fans were installed on the box panels to pull air through the box in an attempt to reduce instrument overheating. The power supplies were selected for their low electrical noise, and Adafruit ADS1115 16-Bit ADC boards further minimised this issue. A schematic of the instrument is shown in Fig. 1. The overall power budget of the device when operating was approximately 52 W, with a breakdown of components as follows: 18 x EC sensors: 9 W, 32 x MOS sensors and internal heaters: 9.4 W, 2 x RH/temp sensors: 0.01 W, 2 x diaphragm pumps: 16.8 W, 2 x fans: 2.8 W, 2 x Arduino Uno's: 0.58 W, Latte Panda micro-computer: 10 W, 3 x power supplies: 3 W.

We note that this type of approach differs from the majority of LCS air quality instruments described in the literature and that are commercially available. In most cases the emphasis in LCS design has been minimising cost and size. Clearly an instrument that contains >40 individual sensors is not optimised with cost or size as its main design goals. Instead, we have focused on data reliability as well as the advantages associated with electrical power consumption compared against a suite of traditional reference instruments.

Figure 2 summarises in simple terms how device costs and power consumption compare between a single sensor device, a six-sensor clustered approach and a reference instrument, using the example of ozone. The clustered approach, whilst more expensive than a single sensor, retains a very substantial power advantage over the reference creating potential for deployment in remote or off-grid locations, or in developing countries where electrical supplies can be both costly and unreliable. The next key question therefore is whether a more complex and expensive clustered sensor instrument can meet similar data quality as reference instruments, and therefore offer a direct alternative, but with lower power and operational costs.

2.2 Sensor test deployment in Beijing

The multi-sensor instrument described in section 2.1 was deployed alongside research-grade reference instruments in Beijing, China during a large air quality experiment between 29th May and 26th June 2017. Beijing has well documented issues around air quality (Zhang et al., 2016) meaning concentrations of pollutants were anticipated to be elevated and to show a large dynamic range. Beijing also experiences warm, humid summers (Chan and Yao, 2008); during the deployment reported here air temperature fluctuated between 15.6 – 41.2 °C and absolute humidity ranged between 3.82 – 17.83 g m^{-3}. In combination these conditions provide a robust and wide-ranging test of instrument performance.

Both sensors and reference instruments were located at the Institute of Atmospheric Physics (IAP) site (latitude 39.978, longitude 116.387), which is situated to the north of central Beijing. All instruments were housed in converted sea container laboratories for this study. Reference instruments for NO_2 and O_X EC were co-located and sampled from the same 3 m high inlet, with sample bypass flow provided by a common diaphragm pump. The NO_2 reference measurement was by cavity attenuated phase shift (CAPS) spectroscopy (Teledyne T500U, Teledyne, California), with a 100 ppb NO_2 in N_2 calibration source. The NO_2 reference measurements had 5% uncertainty and 0.1 ppbv precision. O_3 reference was measured at 1-minute averages by a Thermo Environmental UV absorption photometer (TEI49i), traceable for calibration to the UK National Physical Laboratory primary ozone standard with an uncertainty of 2 %, and a precision of 1 ppb.

2.3 Data analysis approaches

The clustering approach was used to improve sensor reproducibility as previously discussed in (Smith et al., 2017), whereas the SLR and ML techniques were applied to improve sensor accuracy by correcting for cross sensitivities.

The median voltage signal from of each of the sensor clusters was calculated automatically by the built-in computing device and software, and then that value converted to concentration units using four different numerical techniques: i) simple linear regression (SLR), ii) boosted regression trees (BRT), iii) boosted linear regression (BLR) and iv) Gaussian Process (GP).

Machine learning techniques (methods ii-iv) are powerful tools for identifying relationships between variables and have been shown to support improved concentration estimates that correct interferences in low cost sensors(Geron, 2017; Zimmerman et al. 2017; Lin et al. 2018; Esposito et al. 2016; Hagan et al., 2018).

The full dataset from all sensors (chemical and environmental) was used in the ML algorithms with a subset of the time-series (2nd June – 8th June 2017) treated as training data. Following training, the ML algorithms were then applied to the testing data set (8th June – 26th June 2017), outputting a corrected concentration value. The median of each sensor cluster of CO, NO_2, O_3, VOC, plus humidity and temperature were used by the three different ML algorithms to determine the viability and relative performance of supervised, self-optimisation techniques as a method for correcting for cross interferences. Examples of both parametric (boosted linear regression, BRL and boosted regression trees, BRT) and non-parametric (Gaussian Process, GP) techniques were assessed. BRT was chosen as a numerical method since it provides diagnostics about how the decision trees are constructed, essentially identifying which sensor signals are used in the calculation (Chen and Guestrin, 2016; Geron, 2017). The results can then be compared to known relationships from previous laboratory studies and ensuring that the prediction is in large part a measurement rather than a model value. Gaussian Process (GP) was used because of its proven ability to handle noisy data and it ability to provide the estimations of uncertainty for each data point in the testing data (Geron, 2017; Rasmussen and Williams, 2006).

3. Results and Discussion

3.1 How clustering improves performance

Previous laboratory studies (Smith et al., 2017) have shown that clustering sensors was one potential technical approach to reducing effects of hour to day drift in individual sensor response and limited the effects of inter-sensor manufacturing variability. The median sensor signal was shown to be a more reliable predictor of the true pollutant value (versus the mean) and the effect of deteriorating or highly variable sensors was minimised. This approach has been extended here to field observations and to a wider range of different chemical species. The EC sensors output two voltages; one from the working electrode (WE) and one from the auxiliary electrode (AE). The standard calibration procedure subtracts the effect of the auxiliary electrode from the working electrode (the electrode exposed to the ambient air and oxidising compounds) effectively helping to correct for some of the temperature and humidity effects. The manufacturer supplies individual conversion factors and equations for each sensor and these were applied to each sensor prior to use within the cluster. Each sensor within a cluster was initially normalised to give a common voltage output.

We use the raw sensor voltages and the manufacturers calibration values to gain an initial concentration. One method of determining the improvement in the concentration estimated by the sensors is to compare the range of slopes obtained against reference instrument for a range of different numbers of sensors. This is shown for the first time for an electrochemical NO_2 sensor in Fig. 3. As the number of sensors in a cluster is increased, the observed range of values for the unique permutations of the groups narrows considerably, greatly improving measurement precision. The slope does not however converge on 1:1 since there is a difference in the factory calibration of the sensors compared to the reference instrument. The cluster versus reference comparison using simple factory calibration can be seen in Fig. 4a.

3.2 Simple Linear Regression

The first data calibration approach used was simple linear regression (SLR), applied to calibrate the median sensor signal using the reference instrument concentration from the first five days of the experiment (the training period). The sensor concentrations were corrected using linear parameters from training period calibration and subsequent sensor performance was assessed by comparing against the co-located reference instrument. Using the NO_2 EC as an example, linear parameters in the form of $y = mx + c$ were determined using a linear least squares fit between the NO_2 CAPS reference instrument and the median NO_2 EC sensor for the first five days of the sensor instrument deployment. Once trained in this manner, these linear calibration factors based on SLR were used to calibrate the median NO_2 sensor and were unchanged for the remainder of the experiment.

The different pollutant clusters showed variable performance against their respective reference over the 21 days. We use here root mean squared error (RMSE) as a metric to evaluate the performance of various clusters and different data calibration approaches. We also calculate the RMSE between two approximately co-located NO_2 reference grade instruments (4.3 ppb) during the same field deployment to quantify what might be considered the 'optimum comparison' that could be expected between the sensors and the reference approach. During the campaign a localised source of NO/NO_2 was emitted into the vicinity downwind of the second NO_2 CAPS instrument, and hence not observed by it. For a fair comparison of the two NO_2 reference measurements the data between the 10th and 14th June, when the NO/NO_2 emission occurred, was removed. Unfortunately, there was not a co-located CO reference instrument or multiple co-located reference observations of O_3 available for this study. The CO sensor median was still included with the total VOC median, RH and temperature in the sensor variables for training and testing the ML algorithms, but we were unable to make a comparison.

Applying SLR, the NO_2 sensor cluster gave a root mean squared error (RMSE) of 10.42 ppb and RMSE = 10.44 ppb for the O_X cluster median signal with the sum of the $NO_2 + O_3$ reference measurements. The ambient NO_2 concentrations varied over a wide range from below 2 ppb to in excess of 200 ppb and the clustered NO_2 package performed well at capturing this range of observed concentrations, but with substantial discrepancies between the median NO_2 EC sensor and the NO_2 CAPS reference instrument when the reference NO_2 concentrations were below 10 ppb. This finding fits well with previous work that shows the impact of cross-sensitivities on EC sensors is most important at low target compound concentrations (Lewis et al., 2016). The Alphasense OX-B431 sensors detected both O_3 and NO_2. They respond proportionately, but independently to concentrations of O_3 and NO_2, hence the O_X EC were calibrated with and compared to the sum of the O_3 and NO_2 reference measurements. The median value from the O_X cluster showed the best correlation with the respective reference measurements (O_X $R^2 = 0.95$, NO_2 $R^2 = 0.86$).

3.3 Using machine learning (ML) algorithms to calibrate the median sensor cluster

Each ML algorithm was trained and then tested using the same 1-minute average sensor data as the SLR in section 3.2, split into the same training and testing sets each time. The training data was the first 8490 data points of the measurement period, and the testing set the remaining 25956 data points. For BRT and BLR the python XGBoost implementation was used to train, cross validate and test the models. This scalable learning system is open source, computationally efficient, and has performed well on other platforms (Rasmussen and Williams, 2006). Both BRT and BLR have different hyperparameters that allow the

ML algorithm to be tuned so that the algorithm can detect trends within the data, without overfitting. Hyperparameters, such as the learning step can be increased or decreased to allow a good fit to the training data, and to optimise the performance of the algorithm (Geron, 2017). To tune the ML algorithm hyperparameters a five-fold cross validation of the training set was used to build the classification models, with a randomisation seed of 42 each time. The seed randomises the data, so it does not matter the value of the seed, just that it is consistent for the cross validation. During the cross validation process, the algorithm trained on one-fold of the training data set and made a prediction based on these learnt relationships over the other four folds to test out the associated rules it has found. The hyperparameters were decided by minimising the mean absolute error (MAE) between the predicted folds and the training label (Shi et al., 2017). Once decided, these hyperparameters were fixed and the algorithm then tested on data that it has not yet seen, i.e. the testing data set.

BRT uses gradient boosted regression trees to integrate large numbers of decision trees, and this improves the overall performance of the trees (Rasmussen and Williams, 2006; Friedman, 2001). Through a process where many decision trees are working on the training data set the algorithm generates a set of rules by which the training data is linked to the training label (Shi et al., 2017). By discarding trees that do not have much impact on the MAE, the algorithm is more efficient at determining the relationships between variables. The nature of decision trees means BRT is not limited to identifying linear functions, unlike BLR. During the same cross validation process as described for BRT, BLR identifies the linear relationships between the sensor variables and uses these correlations to predict the compound response during the testing period. BLR is simpler than BRT but works well when there are multiple linear trends between variables. Gaussian Process (GP) uses the Gaussian distribution over functions and can be a powerful tool for regression and prediction purposes (Rasmussen and Williams, 2006). It is a flexible model which generalises the Gaussian distribution of the functions that make up the properties of each variables function (Rasmussen and Williams, 2006). GP can be used as a supervised learning technique once suitable properties for the covariance functions (kernels) are found, then a GP model can be created and interpreted (Roberts et al., 2013). For this study there were two kernels used to train and predict the sensor data. These were Matern32 (k1) and Linear (k2) functions. They were added together (k1 + k2) to enable both linear (k2) and non-linear (k1) relationships between the variables to be detected, as it was observed in the laboratory that the relationships between the variables could be either (Lewis et al., 2016; Smith et al., 2017). The hyperparameters were then self-optimised using the training data by the open-sourced python package running the algorithm, GPy. The GP, BRT and BLR predicted responses were then compared to the reference data over the testing period, and a RMSE calculated to investigate how well the ML algorithm performed.

3.4 Sensor cluster data with ML processing – NO_2 cluster

Figure 4 shows the predicted NO_2 time series using the median cluster value and the three ML calibrations compared with the reference measurement. The median sensor with individual factory corrections (Fig. 4a) clearly detects the major trend in NO_2

concentration, but often under predicts at times when the NO_2 concentration is low. At higher concentrations the median sensor overpredicts the NO_2 signal, leading to a RMSE of 86.7 ppb.

3.4.1 Gaussian process (GP)

The GP ML algorithm predicted the NO_2 concentration with a RMSE of 5.2 ppb compared to the reference measurement, the lowest for all the different ML techniques. The Matern32 kernel is adept at capturing the more typical (sub 50 ppb) NO_2 concentrations, due to its ability to model cross-sensitivities on the sensor signals but struggled to extrapolate to highest concentrations. One advantage of using GP to predict compound concentrations is that an uncertainty on the predicted values is also calculated. This uncertainty is shown in Fig. 4b (light yellow shading), as ± 2 standard deviations on the predicted data points. It is clear that there are periods when there is more uncertainty in the prediction. There are four main periods where the GP prediction appeared low, and the uncertainty was high: 1500H 8th June, 1700H 9th June, 1400H 15th June and 1400H 16th June. These over-extrapolated data points all occurred when the temperature reached +40 °C and exceeded the maximum temperature recorded during the training period (35.8 °C), coinciding with the NO_2 concentration and RH were low (Fig. 4e). Machine learning techniques all have difficulty making predictions when the testing and training data sets cover different variable space, but the calculation of a prediction uncertainty which takes this into account highlights when this could potentially be an issue and could be used to inform calibration strategies.

3.4.2 Boosted Regression Trees (BRT)

The BRT prediction (Fig. 4c) was very good during periods when the test data did not exceed concentrations of NO_2 seen in the training data (~79 ppb). However, the classification nature of the BRT algorithm means it is incapable of extrapolation, so the prediction cannot capture the high concentrations of NO_2 that were observed between the 10th – 14th June (the NO_2 CAPS instrument recorded a maximum NO_2 concentration of 222.2 ppb during the testing period). Between this time a localised source of NO/NO_2 was emitted. Overall, the RMSE between the BRT NO_2 prediction and the NO_2 CAPS reference measurement was 7.2 ppb, an improvement on SLR (10.4 ppb) of ~30% despite its inability to capture NO_2 concentrations outside of those experienced during the training data period. This improvement for the lower concentrations of NO_2, is due to the BRT model's ability to better correct for some cross sensitivities on the sensor signals, such as the effect of humidity. With the dates omitted for the localised source of NO/NO_2 (described in section 3.2) the RMSE for BRT prediction was 6.1 ppb, showing that the BRT prediction does well at capturing the trends in NO_2 when extrapolation is not required.

The BRT algorithm outputs a gain feature called gain, which can be used to identify how much each variable contributes to the predicted sensor response and these are shown in Fig. 5a. The median NO_2 sensor signal was (encouragingly) the largest contributor to the NO_2 concentration prediction, followed by data from the CO cluster and the relative humidity sensor. This is consistent with previous laboratory results, where it was observed that the NO_2 sensor signal had a CO interference and was affected by changing humidity (Lewis et al., 2016).

3.4.3 Boosted Linear Regression (BLR)

The BLR predicted NO_2 concentration was comparable to the GP prediction, with a RMSE of 6.6 ppb. When the NO/NO_2 localised source was removed the RMSE did not change substantially (6.3 ppb) suggesting that this technique was good at

extrapolating to the NO_2 concentrations outside the range of the training data. BLR assumes purely linear trends between variables, meaning it does not represent non-linear relationships, but the linear nature of the relationships allows BLR to extrapolate trends beyond the ranges seen in the training data. Figure 5d shows the predicted BLR NO_2 signal fully capturing the maximum NO_2 concentrations between the 10^{th} – 14^{th} June. Overall, the RMSE between the BLR prediction and NO_2 reference measurement were slightly better than the BRT suggesting that the inter-sensor relationships were often approximately linear over the variable space observed. The similarity between the GP and BLR predictions are not surprising given the use of the linear kernel in the GP algorithm. The BLR also over-extrapolated the predicted NO_2 concentration during the same periods as the GP prediction, suggesting that the linear kernel contributed substantially to the GP prediction but that the training data was not adequate to capture deviations from this linearity.

Figure 7a summarises how a progressively improved RMSE can be achieved as NO_2 sensors are first used in a cluster, and then the various different numerical methods applied to calibration, ultimately producing performance that is close to the reference vs reference RMSE. Figure 7a also highlights the evidence that the uncertainty in the sensor concentrations is greatly reduced if the sensors are calibrated in field (using SLR) or if ML procedures are applied. The GP prediction was the ML calibration technique that was closest to the RMSE between the two reference instruments. The RMSE and NRMSE was calculated after the application of SLR and ML for different reference concentration ranges to indicate where the greatest improvement of the sensor data occurred (see, Table 2). The RMSE and NRMSE (calculated by dividing the RMSE by the mean of the concentration bin) were determined between the reference NO_2 observations and the sensor values for four equally spaced reference concentration bins. The ML techniques produced the greatest improvements in the concentration estimates for the lower concentrations of the target measurand where the effect of cross interreferences is more significant. The BRT and GP in particular displayed large improvements for the lower NO_2 reference observations. At the higher concentrations of NO_2, the ML algorithms displayed less improvement, where the conditions were outside those of the training data variable space. This was very noticeable for the BRT algorithm due to its inability to extrapolate.

3.5 Sensor cluster data with ML processing – O_x cluster.

The data from the median O_X sensor versus the $NO_2 + O_3$ reference measurements is shown in Fig. 6, along with the best performing ML data processing method. During peaks in O_X concentration the factory calibrated sensor values tend to produce over estimates of the O_X concentrations (e.g. maximum O_X concentration observed by reference was 253 ppb, the median O_X sensor 426 ppb). The ML technique with the lowest RMSE, BRT, bought the O_X concentration estimate much closer to the reference observations, see Fig. 6, however, during peaks in O_X concentration, the BRT predicted O_X concentration estimate was underpredicted due to BRT's inability to extrapolate.

A summary of RMSE improvements, implemented for all methods can be found in Fig. 7b. BLR and BRT performance was near identical indicating the O_X sensors have largely linear relationships governing their performance, at least over the variable

space observed. The 30% of the data used to train the ML algorithms included a range of O_X concentrations much more representative of the total observation period than was the case for NO_2, and so only limited extrapolation beyond the training dataset was needed. The BRT algorithm gain was again used to determine the largest contributing variables to the BRT O_X prediction, Fig. 5b. The median O_X sensor value made the largest contribution to the BRT O_X prediction (92%). The median CO sensor contributed 1.5% to the prediction. The NRMSE was calculated for 4 equally sized reference O_X concentration bins for each analytical method used, in a similar manner to Table 2 for NO_2. The NRMSE improved for SLR and the ML algorithms across all concentration ranges, with BLR and BRT optimal for reducing the error estimate the most. The error was the highest at the higher O_X concentrations for BRT, which was expected due to BRTs inability to extrapolate.

3.6. A measurement vs a sensor model

ML algorithms are skilful at detecting patterns within a dataset and the work shown in this study is evidence that they can improve the performance of LCS, as measured by a reported concentration value compared to a reference. Each of the sensor predictions made by the ML algorithms could be justified by previous experience with working with similar EC sensors in the laboratory and from reported studies. For example, the predicted NO_2 sensor response was formed based upon decision trees that were primarily influenced by the median NO_2 sensor value, then small adjustments were made to the prediction using the median CO EC and humidity data. This is reasonable based on previous laboratory experiments showing NO_2 sensors responding to CO and changing humidity. When using the sensors to correct cross interferences and changing meteorological conditions, the prediction is an optimised version of the sensor response that essentially calibrates for identified cross-sensitivities.

However, ML algorithms can also be used to make predictions of compounds, for example nitric oxide (NO), that are simply correlated to other air pollution variables, but that are not physically measured by a specific sensor. As an example, in this study a reference grade NO measurement was made from the same sampling line as the sensor instrument and this was used to make a NO-prediction using BRT, based on information gathered by the other chemical sensors. From previous laboratory studies it is known that NO is a cross interference on the NO_2 and O_X EC sensors (Lewis et al., 2016), and therefore we could expect that an NO prediction would use these two variables. However, ambient NO concentrations are closely linked to the concentrations of NO_2 and O_3 via steady state inter-conversion, and this underlying chemistry might also be identified by the algorithm and used to predict NO.

Using a BRT model and sensor cluster median values from the sensor instrument deployment, it was possible to correctly identify when the major NO peaks would occur and predict NO concentrations with a RMSE of 10.5 ppb, even though our instrument did not actually contain a NO sensor. This corresponds to a Normalised Root Mean Squared Error (NRMSE) of 0.37. For comparison, the NRMSE for the BRT NO_2 and O_X predictions were 0.11 and 0.08 respectively, and the two NO_2 reference instruments gave a NRMSE of 0.06, so the NO prediction contains a high degree of uncertainty although appears to be quite good initially. When we interrogate the decision tree model however, we find that the prediction is largely based on the chemical relationship between NO_2 and O_X, and not on any cross-sensitivities on sensor signals. In this rather extreme

example it could be claimed that this NO prediction is not a measurement but a model (Hagler et al., 2018), and highlights the challenge of interpreting low cost sensor measurements that exist in something of an analytical grey area due to their reliance on complex calibration algorithms.

4. Conclusions

Using a combination of clustering sensors and machine learning data processing, a lower cost and relatively low power air quality instrument has made measurements of NO_2 and O_X that were close to the RMSE of reference instruments (over the period of study). Clustering of sensors adds little to the overall power budget of an instrument but is a very easy way to overcome individual sensor drift and irreproducibility. Further data treatments such as in-field calibration with SLR or supervised ML techniques can further optimise the sensor data. SLR was seen to improve median sensor concentrations to some degree but struggled to accurately calibrate the sensor data at the lower concentrations. ML techniques were able to further improve the sensor performance because they could correct multiple trends between the sensor variables eliminating some cross-interferences. BLR and BRT were seen to be most powerful at predicting the compound response and used information content from other variables that was reasonable based on previous lab studies. The GP approach was advantageous in that a standard error could be calculated for each predicted data point. Therefore, this identified regions within the data where the prediction was more uncertain, for example, if the testing data significantly deviated from the variable space observed during training. BLR was the simplest technique and worked well when the functions between the sensor variables were linear, for example during the O_X sensor prediction. The time required to train and run the model was reduced when using BLR and BRT over GP. A longer period of data collection, of at least a few months to a year of sensor data, is needed to establish how long such algorithms accurately predict the reference observations. It appears that as a minimum the use of ML calibration techniques would increase the time required between physical calibrations and allow the use of sensor instruments as part of a network or to run in isolated environments, after the instrument was calibrated over as large a range of conditions it is likely to experience as possible. Data that occurs outside the training data ranges can then be flagged and treated with a higher level of uncertainty.

Author contributions

KS, PE, designed and developed the sensor instrument. KS, PE, PI and CD contributed to analysis of sensor data.
FS, JL and YS provided reference data. All authors contributed to the writing of the manuscript.

Acknowledgements

AIRPRO grant NE/N007115/1, AIRPOLL grant NE/N006917/1, NCAS/NERC ACREW, Peter M. Edwards acknowledges a Marie Skłodowska-Curie individual fellowship, Kate R. Smith and Freya A. Squires acknowledge NERC SPHERES DTP PhDs. Peter D. Ivatt acknowledges an NCAS Studentship PhD.

The authors declare that they have no competing interests.

References

Broday, D. M., Arpaci, A., Bartonova, A., Castell-Balaguer, N., Cole-Hunter, T., Dauge, F. R., Fishbain, B., Jones, R. L., Galea, K., Jovasevic-Stojanovic, M., Kocman, D., Martinez-Iñiguez, T., Nieuwenhuijsen, M., Robinson, J., Svecova, V. and Thai, P.: Wireless distributed environmental sensor networks for air pollution measurement-the promise and the current reality, Sensors (Switzerland), 17(10), doi:10.3390/s17102263, 2017.

Caron, A., Redon, N., Hanoune, B. and Coddeville, P.: Performances and limitations of electronic gas sensors to investigate an indoor air quality event, Build. Environ., 107, 19–28, doi:10.1016/j.buildenv.2016.07.006, 2016.

Chan, C. K. and Yao, X.: Air pollution in mega cities in China, Atmos. Environ., 42, 1–42, doi:10.1016/j.atmosenv.2007.09.003, 2008.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, Knowl. Discov. Data Min., doi:10.1145/2939672.2939785, 2016.

Esposito, E., De Vito, S., Salvato, M., Bright, V., Jones, R. L. and Popoola, O.: Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems, Sensors Actuators, B Chem., 231, 701–713, doi:10.1016/j.snb.2016.03.038, 2016.

Friedman, J. H.: Greedy function approximation: a gradient boosting machine, Ann. Statisitcs, 29(5), 1189–1232, 2001.

Geron, A.: Hands-On Machine Learning with Scikit-Learn and TensorFlow, First Edit., edited by N. Tache, N. Adams, and R. Monaghan, O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472., 2017.

Hagan, D. H., Isaacman-Vanwertz, G., Franklin, J. P., Wallace, L. M. M., Kocar, B. D., Heald, C. L. and Kroll, J. H.: Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments, Atmos. Meas. Tech., 11(1), 315–328, doi:10.5194/amt-11-315-2018, 2018.

Hagler, G. S. W., Williams, R., Papapostolou, V. and Polidori, A.: Air Quality Sensors and Data Adjustment Algorithms: When Is It No Longer a Measurement?, Environ. Sci. Technol., 52(10), 5530–5531, doi:10.1021/acs.est.8b01826, 2018.

Hong, H.-K., Shin, H. W., Park, H. S., Yun, D. H., Kwon, C. H., Lee, K., Kim, S.-T. and Moriizumi, T.: Gas identification using micro gas sensor array and neural-network pattern recognition, Sensors and Actuators, 4005(96), 68–71, 1996.

Jiao, W., Hagler, G., Williams, R., Sharpe, R., Brown, R., Garver, D., Judge, R., Caudill, M., Rickard, J., Davis, M., Weinstock, L., Zimmer-Dauphinee, S. and Buckley, K.: Community Air Sensor Network (CAIRSENSE) project: Evaluation of low-cost sensor performance in a suburban environment in the southeastern United States, Atmos. Meas. Tech. Discuss., (June), 1–24, doi:10.5194/amt-2016-131, 2016.

Lewis, A. C., Lee, J., Edwards, P. M., Shaw, M. D., Evans, M. J., Moller, S. J., Smith, K., Ellis, M., Gillott, S., White, A. and Buckley, J. W.: Evaluating the performance of low cost chemical sensors for air pollution research., Faraday Discuss., 189, 85–103, doi:10.1039/C5FD00201J, 2016.

Lin, Y., Dong, W. and Chen, Y.: Calibrating Low-Cost Sensors by a Two-Phase Learning Approach for Urban Air Quality Measurement, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. Artic., 2(18), doi:10.1145/3191750, 2018.

Masson, N., Piedrahita, R. and Hannigan, M.: Quantification method for electrolytic sensors in long-term monitoring of

ambient air quality, Sensors (Switzerland), 15(10), 27283–27302, doi:10.3390/s151027283, 2015.

McKercher, G. R., Salmond, J. A. and Vanos, J. K.: Characteristics and applications of small, portable gaseous air pollution monitors, Environ. Pollut., doi:10.1016/j.envpol.2016.12.045, 2017.

Mead, M. I., Popoola, O. A. M., Stewart, G. B., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J. J., Mcleod, M. W., Hodgson, T. F., Dicks, J., Lewis, A., Cohen, J., Baron, R., Saffell, J. R. and Jones, R. L.: The use of electrochemical sensors for monitoring urban air quality in low-cost , high-density networks, Atmos. Environ., 70, 186–203, doi:10.1016/j.atmosenv.2012.11.060, 2013.

Moltchanov, S., Levy, I., Etzion, Y., Lerner, U., Broday, D. M. and Fishbain, B.: On the feasibility of measuring urban air pollution by wireless distributed sensor networks, Sci. Total Environ., 502, 537–547, doi:10.1016/j.scitotenv.2014.09.059, 2015.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning, 2nd Editio., The MIT Press, Cambridge, Massachusetts., 2006.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N. and Aigrain, S.: Gaussian processes for time-series modelling., Philos. Trans. A. Math. Phys. Eng. Sci., 371(1984), 20110550, doi:10.1098/rsta.2011.0550, 2013.

Shi, X., Li, Q., Qi, Y., Huang, T. and Li, J.: An accident prediction approach based on XGBoost, 2017 12th Int. Conf. Intell. Syst. Knowl. Eng., 1–7, doi:10.1109/ISKE.2017.8258806, 2017.

Smith, K., Edwards, P. M., Evans, M. J. J., Lee, J. D., Shaw, M. D., Squires, F., Wilde, S. and Lewis, A. C.: Clustering approaches that improve the reproducibility of low-cost air pollution sensors, Faraday Discuss., 00(0), 1–17, doi:10.1039/C7FD00020K, 2017.

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 Actuators, B Chem., 238, 706–715, doi:10.1016/j.snb.2016.07.036, 2017.

Williams, D. E., Henshaw, G. S., Bart, M., Laing, G., Wagner, J., Naisbitt, S. and Salmond, J. A.: Validation of low-cost ozone measurement instruments suitable for use in an air-quality monitoring network, Meas. Sci. Technol., 24(6), 065803, doi:10.1088/0957-0233/24/6/065803, 2013.

Zampolli, S., Elmi, I., Ahmed, F., Passini, M., Cardinali, G. C., Nicoletti, S. and Dori, L.: An electronic nose based on solid state sensor arrays for low-cost indoor air quality monitoring applications, Sensors Actuators, B Chem., 101(1–2), 39–46, doi:10.1016/j.snb.2004.02.024, 2004.

Zhang, H., Wang, S., Hao, J., Wang, X., Wang, S., Chai, F. and Li, M.: Air pollution and control action in Beijing, J. Clean. Prod., 112, 1519–1527, doi:10.1016/j.jclepro.2015.04.092, 2016.

Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L. and Subramanian, R.: Closing the gap on lower cost air quality monitoring: machine learning calibration models to improve low-cost sensor performance, Atmos. Meas. Tech. Discuss., (2), 1–36, doi:10.5194/amt-2017-260, 2017.

Sensor data for this research has been submitted to the PURE repository and has received the D.O.I:

10.15124/1a0c64b0-433b-4eec-b5c7-64d3de0a0351

The reference data can be found on the CEDA website under the Atmospheric Pollution and Human Health in a Developing Megacity (APHH) project.

Table 3: Summary of sensors used within the instrument.

Measurand	**Sensor type**	**Manufacturer**	**Number of sensors in each cluster**	**Number of clusters**
Carbon monoxide (CO)	Electrochemical CO-B4	Alphasense	6	1
Oxidising gases (O_X)	Electrochemical OX-B431	Alphasense	6	1
Nitrogen dioxide (NO_2)	Electrochemical NO2-B43F	Alphasense	6	1
Total VOC	Metal oxide TGS2602	Figaro	8	4
Temperature and humidity	Transducer (HPP809A031)	TE Connectivity	1	2

Table 2: The NRMSE and RMSE between the NO_2 reference and sensor data sets at different concentrations ranges. For each calibration method used in the paper, the data was binned into 25% of the observed reference concentration. The Root Mean Squared Error (RMSE) and the Normalised Root Mean Squared Error (NRMSE) were calculated for each concentration bin and the results for NO_2 and O_X are summarised in the tables below. The NRMSE was calculated by dividing the RMSE between the reference observations and the sensor values by the mean reference concentration for the respective bin.

	NRMSE of Reference vs. NO_2 concentration estimate (RMSE / ppb)				
Concentration range as a % of the max. conc. of reference NO_2	Median	SLR	BLR	BRT	GP
0 - 25 %	1.04 (20.7)	0.59 (11.7)	0.32 (6.3)	0.28 (5.6)	0.29 (5.8)
25 - 50 %	0.69 (47.5)	0.19 (13.3)	0.12 (8.2)	0.22 (15.2)	0.11 (7.9)
50 - 75 %	0.72 (94.9)	0.23 (30.8)	0.26 (34.6)	0.55 (72.5)	0.26 (33.5)
75 - 100 %	0.85 (153.1)	0.10 (17.4)	0.10 (18.8)	0.67 (120.0)	0.10 (18.2)

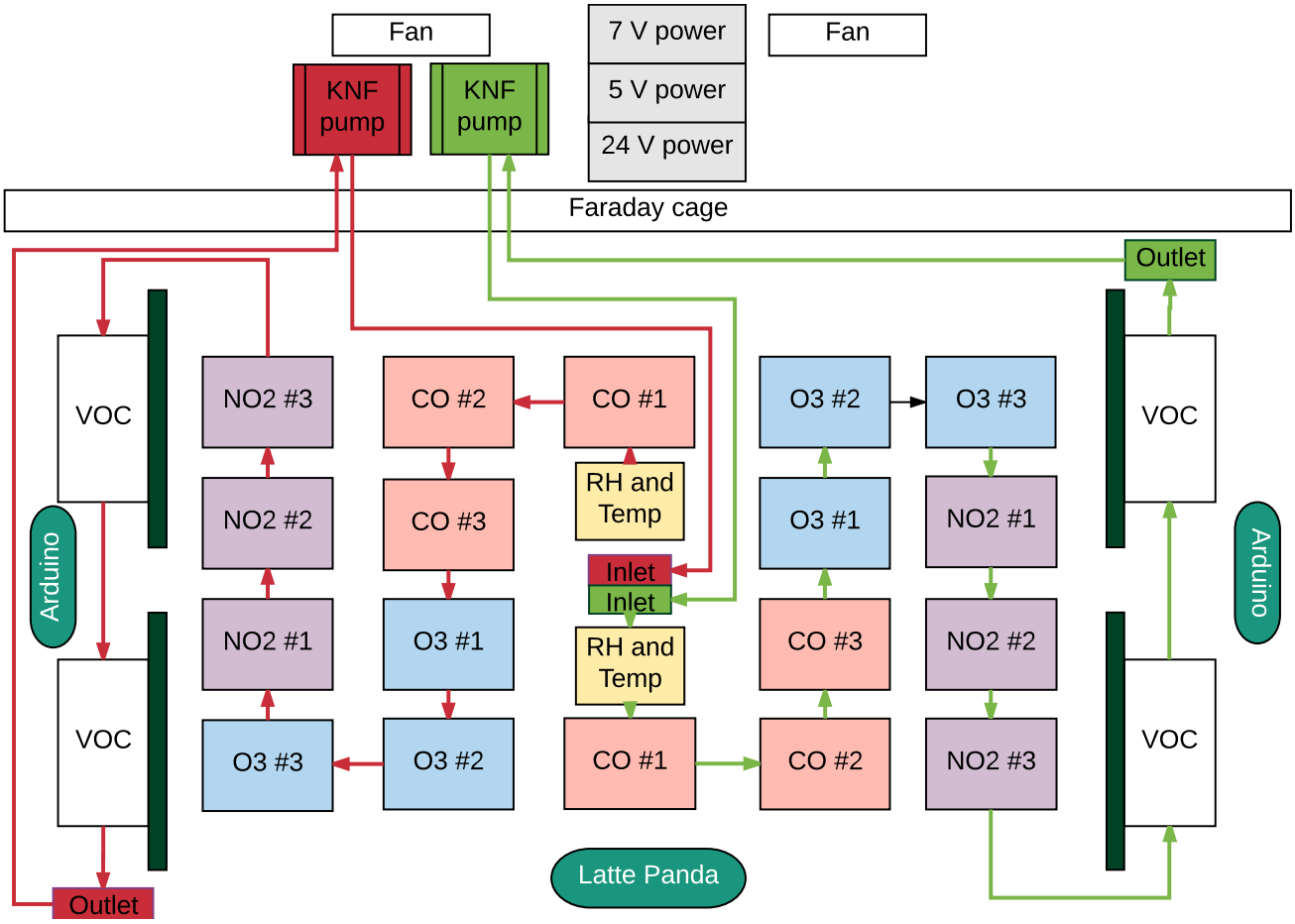


Figure 1: Schematic representation of the gas flow-paths and basic layout of the sensors and components within the device.

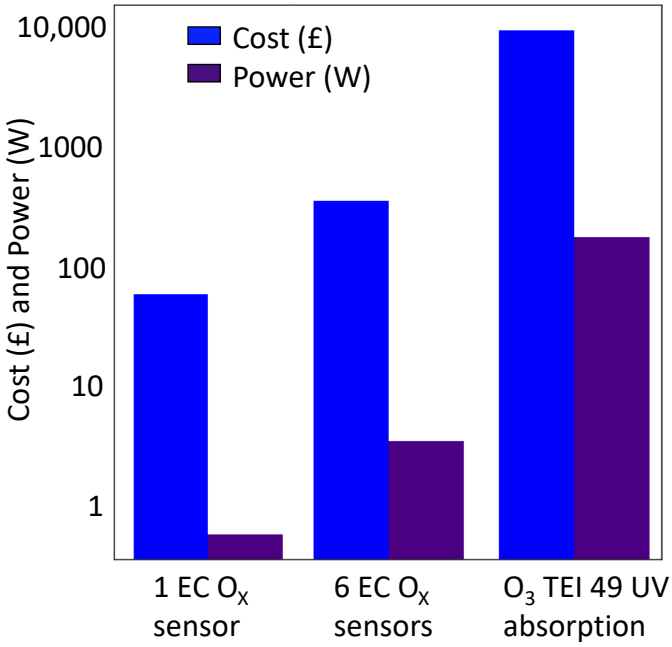


Figure 2: Cost (blue) and power (purple) competitiveness for a single O_X EC sensor device, a clustered six-sensor device and a reference UV ozone monitor.

5

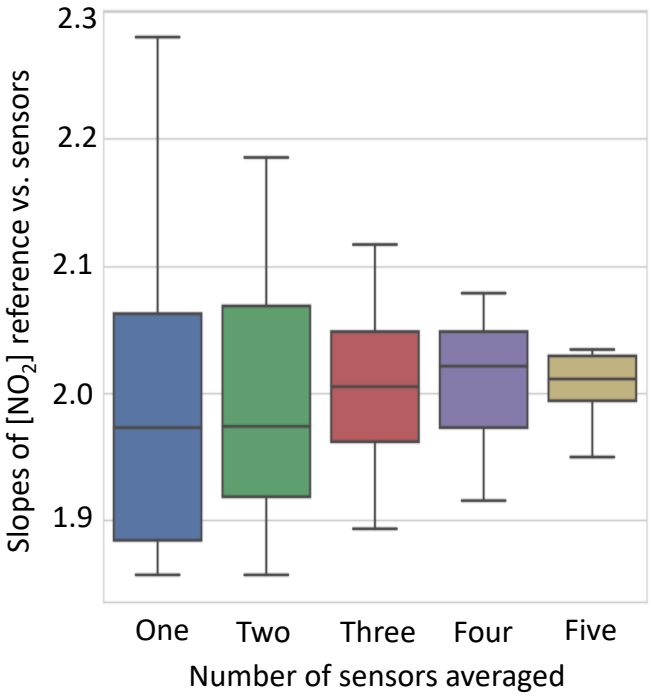


Figure 3: Comparison of slopes of concentrations derived from clusters of NO_2 EC sensors against a reference instrument for ambient Beijing air. As the number of sensors used increases, the spread in data, as seen through the difference in slope, narrows. If data from 3 out of 6 sensors is used there are 20 possible permutations of sensors. The average signal of each was calculated, then plotted against the reference NO_2 CAPS measurements and the gradient extracted. The 20 gradients of these correlation plots (sensitivities) are then plotted in the boxplots above, with the median, 25th percentile, 75th percentile in the box and the 5th and 95th percentile on the whiskers.

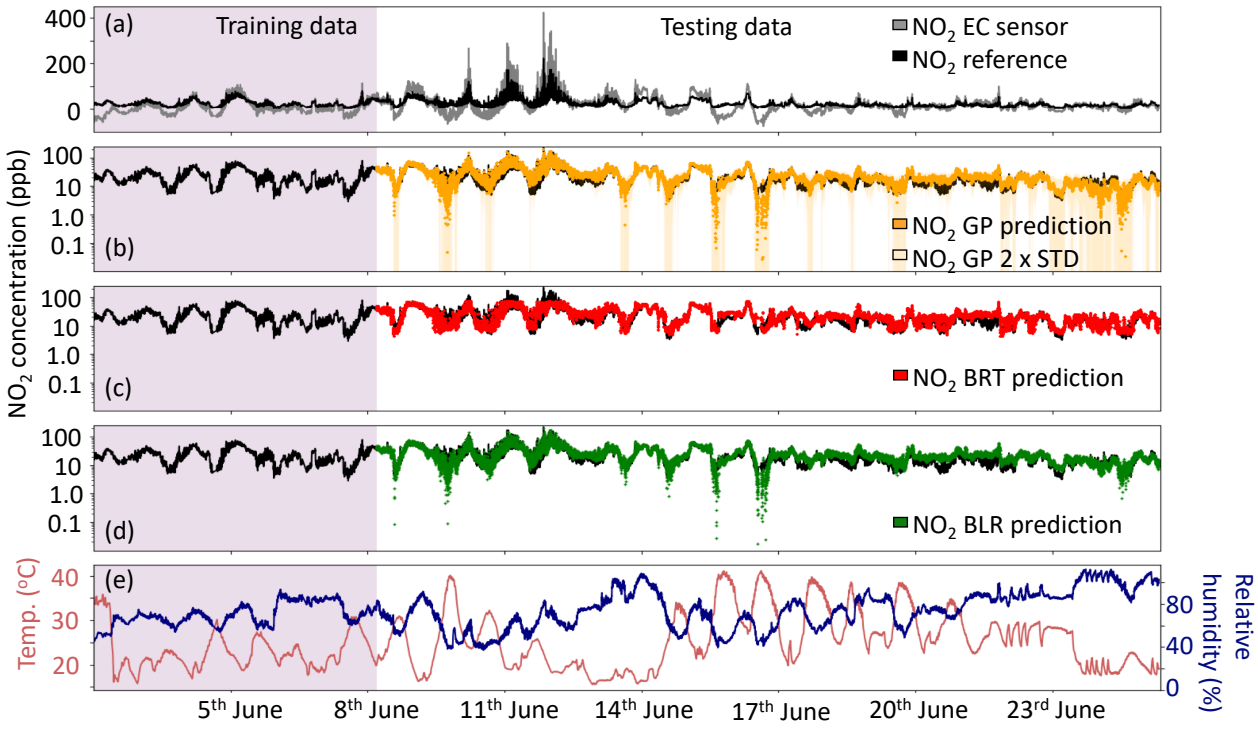


Figure 4: a) Comparison of the median NO_2 sensor using individual factory calibrations, (b) the NO_2 GP prediction ± 2 σ, (c) NO_2 BRT prediction and (d) NO_2 BLR prediction ML techniques. The purple shaded area shows the data used to train the ML algorithms. The black line in all subplots is the York NO_2 CAPS measurement, which was used as a reference. Panel (e) shows the relative humidity (%) and temperature (°C) during the sensor instrument deployment. N.B. Panels (b), (c) and (d) are plotted with a logarithmic y-axis.

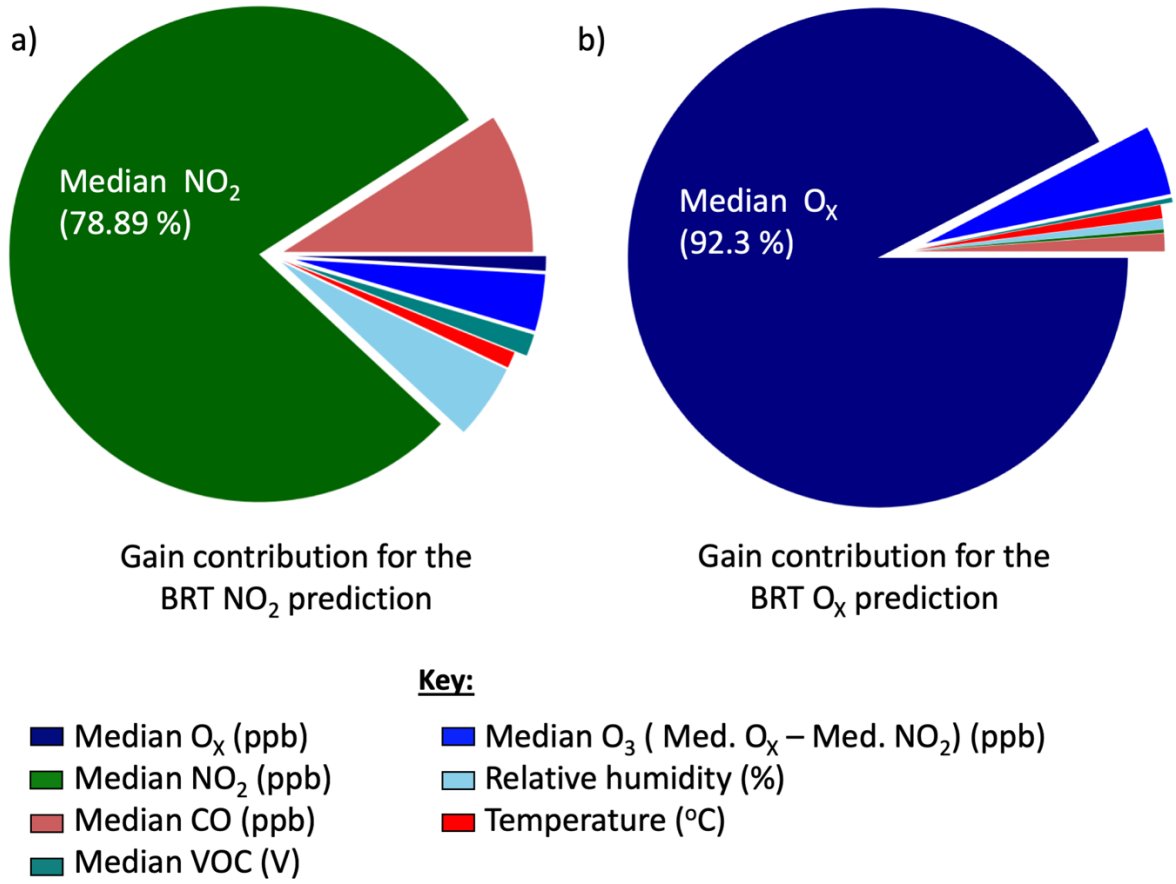


Figure 5: Breakdown of contribution from each variable used by the BRT algorithm to predict the clustered a) NO_2 sensor and b) O_X concentrations.

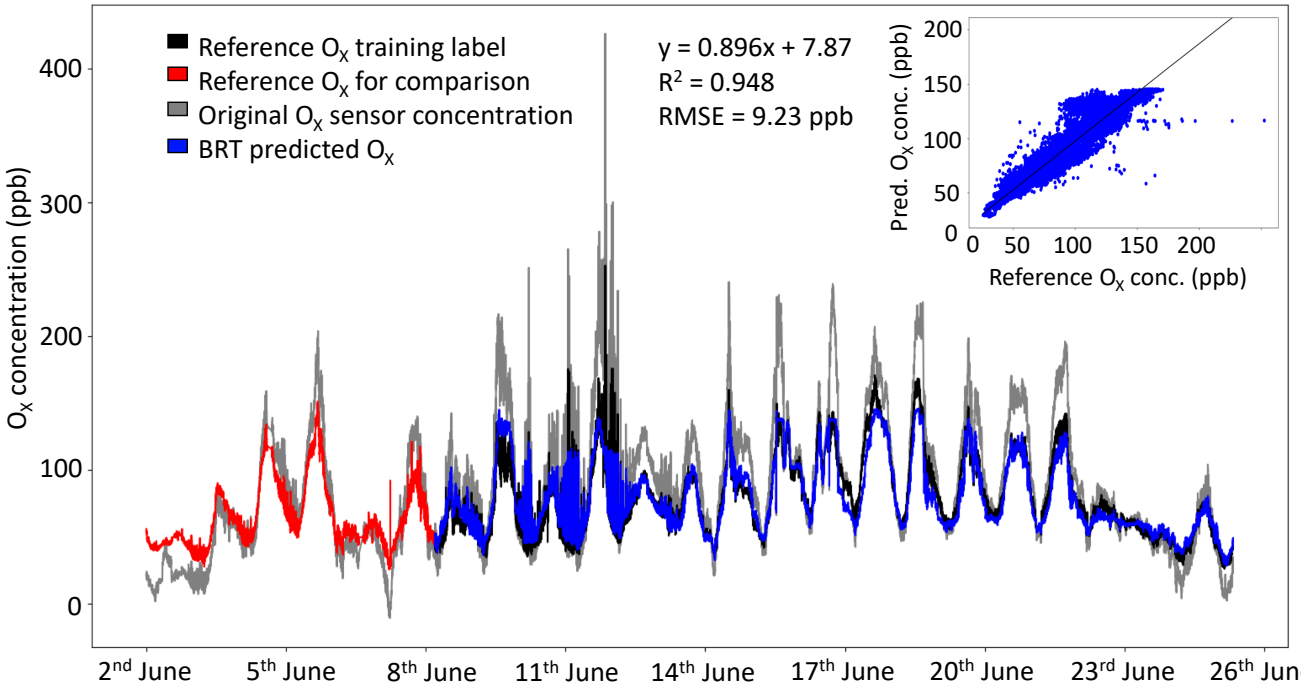


Figure 6: Factory calibrated median sensor concentration (grey), reference $O_3 + NO_2$ data (black) and BRT O_X prediction (blue) for cluster of Ox sensors. The reference measurements that were used as the training label are displayed in red. Inset: The correlation plot for the testing dataset, comparing the reference data and the BRT predicted O_X sensor signal.

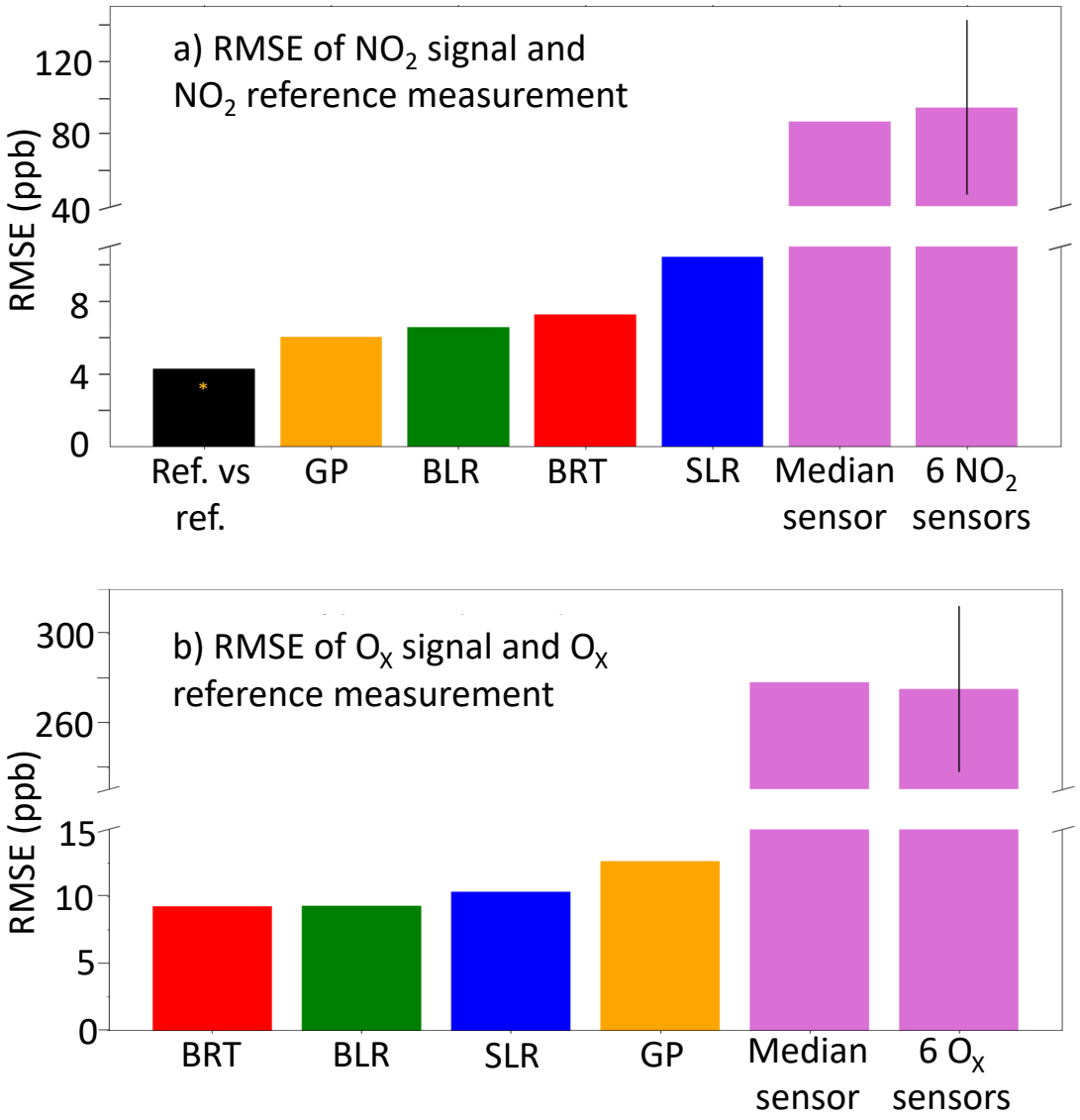


Figure 7: Comparison of the RMSE calculated for electrochemical sensor signal data treatment including individual sensors and a cluster of six using factory calibration, SLR and three ML techniques; when available, a reference versus reference RMSE is also included. a) NO_2, b) O_X.

5