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.

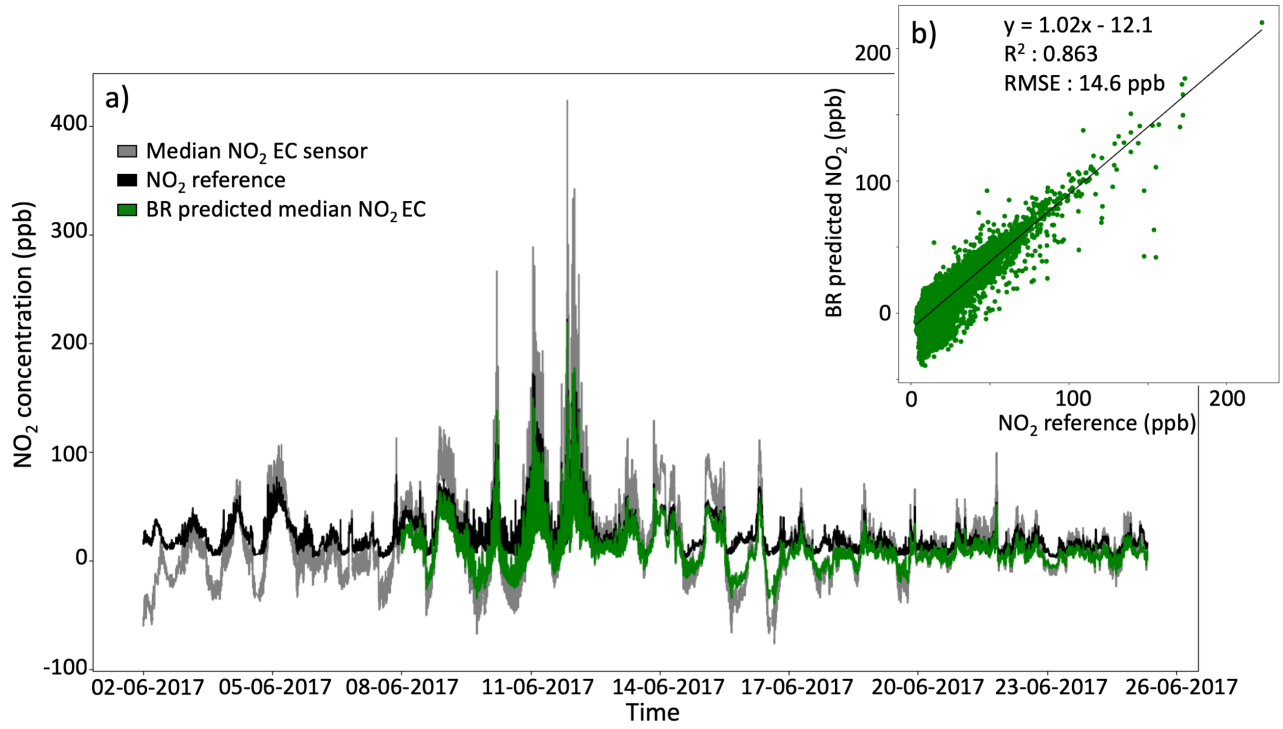


Figure 1. Bivariate regression – Ordinary Least Squares- was performed on the median NO_2 EC sensor (grey) and the NO_2 CAPS measurement during the training data set. and the resulting model was applied to the median NO_2 to produce a bivariate regression (BR) predicted trace (green).

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.

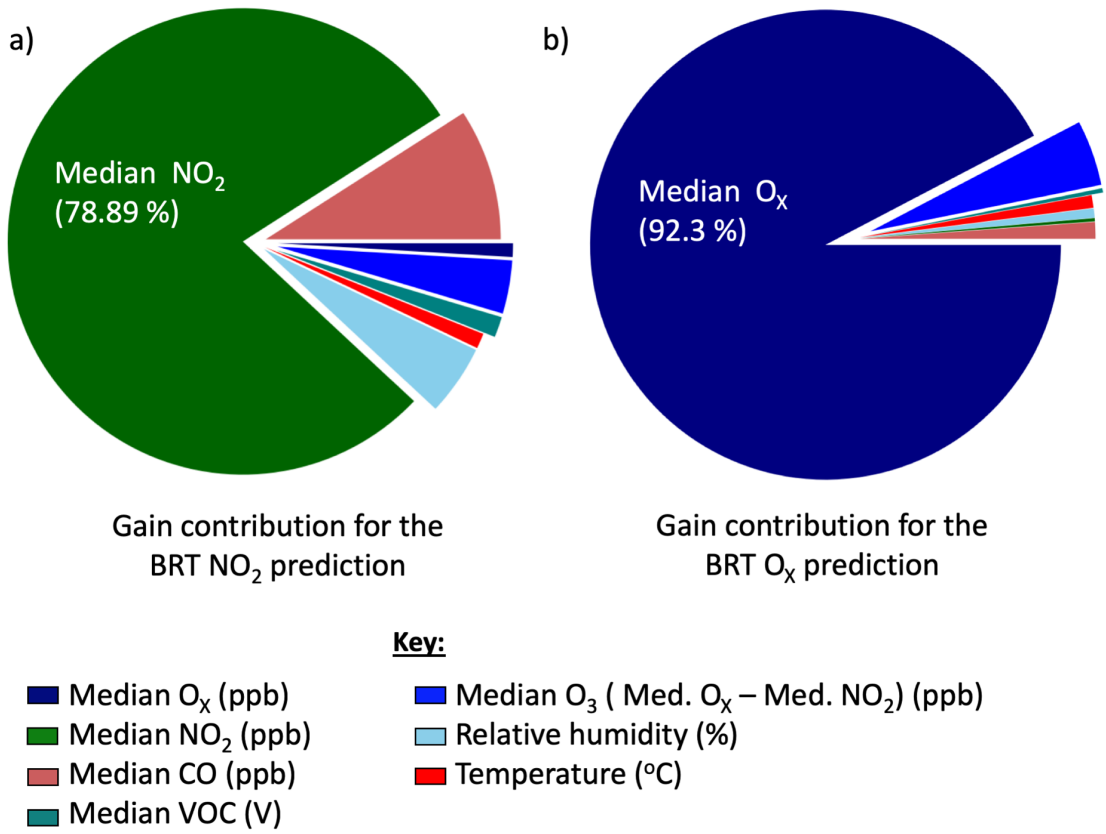


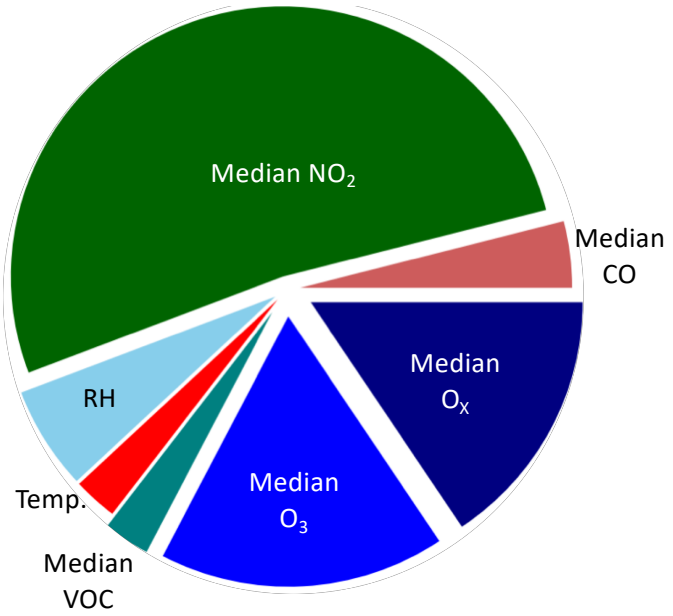
Figure 5: Breakdown of contribution from each variable used by the BRT algorithm to predict the clustered a) NO_2 sensor and b) O_X concentrations.

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.

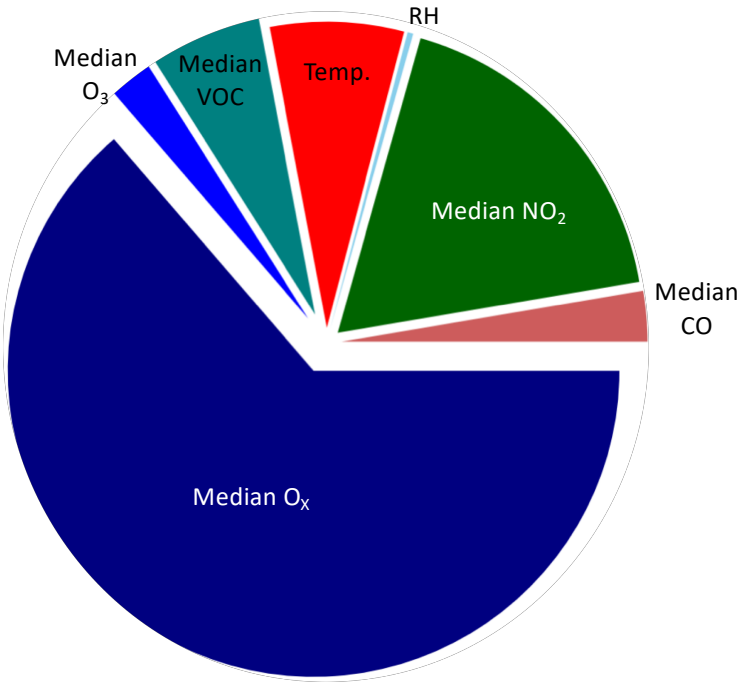
Linear regression weights for variables can be extracted from the BLR algorithm. However, to make assumptions about the relative importance of each of the sensors to the algorithm, all the variables, including the reference observations were normalised to between 0 – 1. The BLR analysis was then repeated 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.