Machine Learning Techniques to Improve the Field Performance of Low-Cost Air Quality Sensors

RC comments and AC responses

RC1: 'Comment on amt-2021-282', Anonymous Referee #1, 19 Nov 2021

General Comments:

In this work, the authors developed a machine learning calibration process that combines a 4-stage baseline offset correction and Random Forest Regression Modelling (RF). They adjusted the RF model by identifying readily available training features and optimizing the number of leaf nodes and trees. This work compared the performance of the RF correction model against values from a reference monitor, the raw sensor value, and baseline-corrected sensor values over a time span of ~7 months. This baseline + RF model improved the performance of low-cost NO_2, PM_{10}, and $PM_{2.5}$ sensors relative to the raw and baseline-corrected values. This machine learning technique is a reasonable method to improve data quality from low-cost air sensors and is suitable for publication after minor revisions.

Major:

RC2-1. Alphasense NO2-A43F electrochemical NO2 sensors (and Alphasense NO2-B43F) have a known cross-sensitivity to ozone (Spinelle et al.). Although the Praxis Urban sensor system and the St Ebbe's monitoring site do not appear to measure ozone, the study fails to mention/address this concern. While inclusion of this variable into feature training could restrict the spread of this model to other networks, it could greatly enhance the performance of the NO2 model. Spinelle et al. also found that sensors from the same manufacturer can behave differently in the same environmental conditions. This manuscript would greatly benefit from applying your model to more than one sensor to demonstrate its capability to nullify discrepancies from sensor to sensor. (Spinelle, L.; Gerboles, M.; Kotsev, A.; Signorini, M. Evaluation of Low-Cost Sensors for Air Pollution Monitoring: Effect of Gaseous Interfering Compounds and Meteorlogical Conditions; Publications Office of the European Union:Luxemborg, 2017. https://doi.org/10.2760/548327)

AC. Firstly, our thanks for your time and thought in preparing your very helpful comments.

Thank you for this comment, we think it is very valuable for context and in developing further learnings. We agree that it is worthwhile adding a note on cross sensitivity with ozone and (will) include a reference to Spinelle 2017 in the revised manuscript

We confirm that ozone data is available at the St Ebbes monitoring station and agree that these data would help in evaluating the effectiveness of ozone as an additional training feature for the development of the RF model and improved correction model performance. However, only 6 in 16 sensors deployed across our network have ozone monitoring capability and this was not the focus for application of low-cost sensor data for local air quality management.

Documenting the performance of the models as-is, is valuable as a demonstrator for the performance that is achievable with the constrained approach presented (i.e. without the ozone cross sensitivity training), not least as this is representative of many real-world low-cost sensor applications where (many) NO_2 only electrochemical sensor network in operation.

AC mods: Lines 65-75, reference to Spinelle 2017 & commentary.
Lines 225-235, response to O3-NO2 cross sensitivity.

RC2-2. It is unclear how this model could be applied to sensors throughout a network. Would each sensor need to spend *x* number of months at a reference site to develop the model prior to deployment? How well would a baseline established at the reference site transfer to the deployment site?

AC. For deployment in real world situations I would anticipate that the model, or a variant thereof, would be training for each 'local' network and this model would be directly deployable across a local network e.g. within a town or small city where the influencing variables are likely to be consistent. The correction model itself is constrained by the diversity of data used to train it, both in terms of variability sensor to sensor and in terms of the pollution/environmental conditions to which the sensors are exposed (mainly NO_2 & RH). The more diverse the training data, the greater the applicability of the model. One of the main challenges for most applications, and particularly in a study environment such as Oxford which has generally / relatively good air quality, is the under-representation of higher pollution events in the training datasets which may result in over correction (under prediction) of real-world concentrations. In an ideal situation one could imagine co-location at low, medium, high and very high pollution conditions, but as I am sure you are aware such situations are almost impossible to engineer.

AC mods: No mods required.

RC2-3. Line 163: "The filtering criteria presented in Table 1 were identified empirically from an analysis of typical sensor performance from the sensor network and from similar parameters logged at the St Ebbe's AURN station" It is not fully clear how these criteria were chosen. Was this based on limits set by the sensor manufacturer? Please clarify. It would also be useful to state the sample population percentage that was removed based on these criteria, as you did on line 188.

AC. Thank you for this comment, we clarify these criteria were developed independently of the manufacturer. Please see sections 2.3.1 to 2.3.4 for an explanation of the derivation of the filter criteria and associated techniques. We will add a footnote to Table 1 to reflect this.

AC mods: Line 629, foot note to Table 1.
Line 172, revised description of method.
Line 181, added proportion of sample population removed.

Minor:

RC2-4. Line 69: "multiple linear regression (MLR) models have been successfully used with variable results" Conflicting statement, please clarify.

AC. We suggest modifying this to "multiple linear regression (MLR) models have been developed with variable results"

AC mods: Line 68, modified text, as above.

RC2-5. Line 136: Please provide more information regarding the location of the sensor relative to the reference instrumentation.

AC. We confirm that sensor and reference instrumentation were co-located at St Ebbes with sensor inlets were within 0.5 metres (gases) and 2 metres (particles). We will add this to the paper

AC mods: Lines 146: dimensions added.

RC2-6. Table 4 & Table 5: Please re-format the column headers as it is currently difficult to differentiate between them.

AC. Thank you for this comment, Tables 4 and 5 have been re-formatted.

AC mods: All tables reformatted.

RC2-7. Line 319: "The performance of each component of the correction method is presented in Table 3" Should read Table 4 I believe. All table references after this point in the manuscript need to be shifted +1 up to Table7.

AC. Thank you for this comment we have corrected the table referencing.

AC mods: Table numbering reviewed & updated throughout

RC2-8. Line 392: “December 2020 saw the occurrence of several pollution events in the particle sensor time series (as also noted above). Although these events were observed throughout Oxford in multiple particle sensor time series, they were not reciprocated in reference measurements, nor in NO2 data” It seems that around 12/25 in Figs 12-14 all corrected sensor values for NO2, PM10, & PM2.5 experience an increase relative to the reference value. Therefore, it does seem like some event affected all three pollutant models. Have you investigated these anomalies further to locate a common factor?

AC. Yes, we confirm this is correct, NO2, PM10 and PM2.5 sensors were all affected by a series of events in Dec 2020 which were not reciprocated in either PM or NO2 reference data and shown in Figs 12-14. We have undertaken some further detailed investigation but have no evidence for associated changes in T & RH local sensor time series nor in independent high resolution weather data. I will modify the text to indicate that no evidence was found in the reference datasets for reciprocal events.

AC mods: Line 388: Commentary on the events provided including to reference to Figs illustrating the effects on PM concentrations

RC2: 'Comment on amt-2021-282', Anonymous Referee #2, 14 Dec 2021

In this work, the capabilities of low-cost sensors for enhancing urban air quality networks is investigated. Statistical and machine learning methods (Random Forest regression) are used for sensor data post-processing and thus for improving the data quality. It is then evaluated whether the achieved corrected sensor data meets European data quality objectives. It is found that the sensors meet the requirements for "indicative" measurements and it is stated that the sensors are "likely to deliver at least comparable data quality to passive sampler methods (for NO2)". These are important findings that might have impact on regulatory air quality measurements. However, I think that the found conclusions are not sufficiently supported in the way this work is presented. I therefore recommend major revisions before this work can be published. My main comments and concerns are the following:

RC2-1. The applied data post-processing approach is in my view not sufficiently explained. The different applied stages are described, that is good, however, some of the stages raise questions: The filters applied in stage 1 are presented in Table 1. If I understand the logic behind the filters as presented, I conclude that all observations at relative humidity > 35% had to be removed. It is unlikely that this is true, please correct (if yes the sensors are useless for most locations).

AC. Firstly, may I extend our thanks for your time in reviewing the paper and the helpful comments. I can clarify that sensor observations of NO2, PM10 and PM2.5 with associated RH values < 35% were excluded from subsequent analyses. Low relative humidity is generally infrequent in Oxford and the UK because of the maritime climate. Looking at Oxford meteorological records in the last 7 years, there has only been 1-day when RH was <35% as a daily mean. However, RH is likely to vary much more at higher time resolutions than this and in preparing our filters we used 15-minute data for Oxford from an independent source http://eodg.atm.ox.ac.uk/eodg/weather/index.html to reality check our assumptions. These data showed that there were ~1,400 15-minute periods (~2.5 weeks) in 2020 when RH values were <35% in Oxford during 2020. On this basis, though the choice of the 35% RH threshold is a precautionary (conservative) measure to screen out sensed values logged during periods when sensor faults may have occurred, we don't believe this inappropriately biases our data. We have updated text and tables accordingly.

AC mods: Line 176, clarification provided

RC2-2. For stage 2, the authors refer to the original publication (and source code) for information about the applied baseline and drift correction method. Without consulting the original paper, the reader has no information how baseline and drift correction technically has been done. Some brief technical description about the applied method would be helpful and should be provided, maybe also in the form

of supplementary information. If I understand correctly, then stage 2 forces the baseline to be zero and by doing so, sensor drift is also corrected.

AC. Thank you for this observation, we will provide a brief description of the airPLS algorithm for clarity. I can also confirm that your understanding is correct. Stage 2 uses the airPLS technique to correct or normalise sensor offset. Offsets of course do vary from sensor to sensor and over time and the airPLS technique offers significant utility in providing a flexible, fast and programmatically easy way of handling them.

AC mods: Lines 185-196. Add commentary on airPLS technique

RC2-3. Stage 3 then compensates for this zeroing and adds an urban scale background concentration. Based on the measurements from an urban background reference site, constant background concentrations have then been determined and added. Firstly, there is no information given how the values for the average uplift have been determined. It is necessary that the authors describe how the given values have been obtained.

AC. The uplift is calculated at the same resolution as the urban background reference i.e. 15-minute resolution. Raw sensor data (at 10s resolution) are aggregated to 15-minute average resolution to align with the same temporal datum as the reference dataset. The requisite baseline uplift is then calculated by difference for each 15-minute observation. We will add this additional information into the revised manuscript.

AC mods: Line 202, 339, 361, 369, 375, clarification on the time resolution at which the uplift/compensation is calculated.

RC2-4. Secondly, an urban background concentration that is constant over time appears to be an oversimplification. This assumption should be explained and justified. If this approach is in a real world application applied to a sensor network across a city, then this would also mean that the urban background is assumed to be constant in time and across the entire city. This is ways too simple. The authors themselves state on page 10 that "the availability of a reliable and high-quality city background ... is essential". Please discuss the consequences for bias and error and potential limitations of this oversimplified approach for background determination.

AC. Using the compensation method described above, the uplift is time varying. We agree that its application is limited to the spatial representativeness of the urban background field characterised by the reference location. The reference location used is in this research is part of the UK compliance monitoring network and conforms to stringent siting criteria set by European air quality Directives to promote local respresentivity. In addition, the study area, Oxford, is a relatively small city with uncomplicated local and surrounding topography, and well understood emissions and emission sources. We do not feel, therefore, that the method is over simplified. However, in larger cities and

places with complex terrain, topography or emissions, we agree that over-simplification of the real-world may occur. In such cases it may be prudent to use multiple reference stations to characterise baseline conditions. We will add such caveats to the paper.

AC mods: Line 169, 487, clarification of the time resolution of compensation method & likely representiveness of the urban background within this study.

RC2-5. In Figure 5 an example of the processing of raw sensor data from stage 1 to stage 4 is presented for NO2 from the sensor system that was co-located at the reference station. For the final data as shown in Figure 5e, the agreement between corrected sensor data and reference NO2 must be considered as very poor. The sensor data is biased high by about 20ppb and shows a very different temporal variability. The data quality as expressed by the MAE and presented in the result section are certainly not achieved during the shown time period. The authors should explain the shortcoming of their data correction method here.

AC. Thank you very much for this observation. To clarify, Figure 5 only presents, in an illustrative way, the handling of sensor offset (and its possible drift over time). This is a preparatory step prior to correcting sensor interference effects using the RF regression model also described in the paper. We agree that the agreement between corrected sensor data and reference NO2 in Fig 5e overall is poor. However, agreement in the baseline of the corrected sensor and reference method datasets is good. These part-corrected (baseline corrected) data are passed to the RF model to correct for environmental interferences. We include a paragraph at the end of section 1.2 and start of section 2.3 to this effect. We will add an explanatory footnote to Figure 5 for clarity.

AC mods: Line 650, footnote added to Figure 5.
Lines 163, 210, clarifications added to clearly demarcate sensor offset correction model & environmental interference models

RC2-6. The authors write in the methods and materials section (section 2.2) that 16 sensor units were deployed across the city of Oxford. One of the sensor units was co-located at the St. Ebbe's reference station. Most results of this research has been obtained from the co-located sensor unit (albeit sometimes not explicitly stated), only data from two of the remaining 15 sensors has been used for this study (for Figure 4). I find mentioning the sensor network somewhat misleading, when in fact most of the data is not used. But more importantly, there is no information provided about how the sensor units have been calibrated before deployment. The only information about calibration is given in section 3.1, however, it remains unclear if the sensor units were deployed after factory calibration or the authors performed a lab calibration. This should be explained in more detail. Then, I wonder about the huge (up to 80ppb) and different offsets of the different sensor units as shown in Figure 4. How can this be explained when presumable all sensors were

calibrated in the same way? The authors mention these huge and different offsets but do not question them. I think the authors should discuss these offsets and provide an explanation. As an user, I would be alerted when seeing such a behaviour of calibrated measurement systems.

AC. Thank you for this observation we have deleted reference to the 16 sensors in section 2.2 to avoid confusion.

You are correct also, we do not mention sensor calibration extensively. For information the sensor systems were calibrated by the manufacturers. No other calibration, other than acceptance tests upon receipt of the sensor systems was conducted. We agree the offsets observed are unexplained, for the reasons you allude to. However, in our experience this is not atypical sensor behaviour. It is in our view consistent with real-word sensor data uncertainty that needs to be handled and can be done so with the methods we present. The evidence we present indicates that the methods perform well under the conditions set out. We will add a comment to this effect into the paper.

AC mods: Line 149 reference to 16 sensors removed.
Line 134, now confirms the calibration status of sensors

RC2-7. The main result of sensor performance is the MAE from the unseen data relative to the reference. The numbers in the abstract do not agree with the numbers in Table 5, please correct. The time resolution of the data used for calculating the MAE's should be given.

AC. Thank you very much for this comment, you are correct, we do have a consistency. I have updated throughout.

AC mods: Updated throughout

RC2-8. In section 3.2.3 the performance of the sensors is compared against European data quality objectives and used the approach as defined for demonstrating equivalence to reference methods. The authors do this for the validation data set and the so-called unseen data. I think the validation data set cannot be used for this purpose. Although the validation data has not been used for model training, it is a random sample of the training data and must be considered as being part of the training data. The uncertainty estimated using the validation dataset (Table 6) are too optimistic. For the unseen data set it can be seen that the performance of the PM sensor is much lower compared to the validation data. The author argue for some very special environmental conditions during the considered time period (December 2020). However, this is probably more a realistic scenario for a real world application and when sensors are used at conditions that deviate from conditions during the model training period. In Table 7 the R2 values for PM10 and

PM2.5 are 0.27 and 0.45 respectively, it is hard for me to believe that this is sufficient for fulfilling the expanded uncertainty objective.

AC. You raise several very important and interesting points here. We feel it is valid to present performance and uncertainty estimates for both the validation set and the out-of-sample (unseen) set. Not least because the differences in the two are not well documented in a peer reviewed setting and we see transparency benefits in doing so and they are in concentration units at least relatively small. Also, I believe the validity of the validation set results depends upon how the methods presented in the paper are applied in an operational setting. A 'traditional' view on the type of correction methods presented might be as a tool that is developed / configured once (or irregularly) and is valid for application many times on (multiple) sensor datasets of the same type. The assumption here being that it is relatively easy to train the model to a sufficiently steady state to deliver satisfactory performance. In such a case, I agree the unseen dataset performance is more relevant. An alternative view is that it is not at all easy to train the model to a sufficiently steady state to deliver satisfactory performance - likely linked to RFs inability extrapolate outside of its training range. Hence, to get to the steady state some very diverse AQ data are needed for training, which of course takes time and resource to acquire. However, until such a time as the data is acquired to achieve steady state, if the model is regularly retrained as new data become available the validation set performance is more applicable. By presenting both, we believe we can allow the reader to make a judgement on which is the most useful for their application and, therefore, the likely uncertainty .

AC mods: Line 390, further justification / confirmation provided.

Other comments:

RC2-9. The mean absolute error (MAE) is used in the paper for quantification of the sensor performance. Would be nice to have the formula available to see how exactly this quantity was calculated (could be given as a supplementary information).

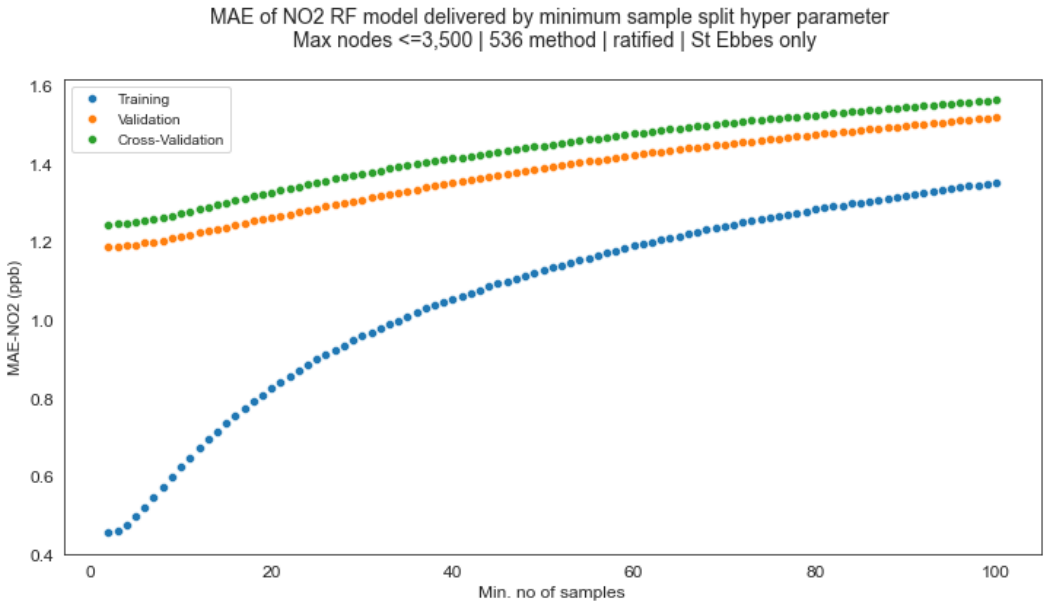
AC. We can provide a reference (also below), for MAE (RF modelling in general) in the revised manuscript. https://scikit-learn.org/stable/modules/model_evaluation.html#mean-absolute-error

AC mods: Line 248, 280, reference(s) provided.

RC2-10. Random forest regression: My impression is that the hyperparameter settings for training the models allowed very and probably too large trees. In particular the minimum number of samples per node (set to a min of 2 samples per node) appears to be very small and might be prone to overfitting. Please comment this.

AC. The RFR models do not appear overly sensitive to tree size and we have found them to be resistant to overfit. With regard to your min sample split query, our tests showed that this default setting worked well; see the fig below for info. The relatively shallow and uniform gradient of the cross validation &

validation curves suggest over-fit at low min sample split values is not a particular issue. I have added comment to the text to indicate that we have performed a reality-check on the hyperparameters chosen to assess the impact of deviations from the parameters identified.



AC mods: Line 279, clarification on reanalysis work to check assumptions used, added

RC2-11. In section 3.2.2. it is referred to Table 3 but this should be Table 4. The different correction steps are difficult to interpret. Please improve formatting. The wrong numbering of tables also continues for the next tables 5, 6 and 7.

AC. Thank you for this observation, also spotted by another reviewer and amended in the revised manuscript.

AC mods: Table numbering reviewed & updated throughout

RC2-12. Section 3.2.2 the MAE values for corrected NO2, PM10 and PM2.5 are given. The temporal resolution of the data used for calculating the given numbers should be mentioned.

AC. Thank you, for highlighting this oversight. We (will) provide the relevant temporal resolution in the revised manuscript.

AC mods: Reviewed & updated throughout