Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC2
https://doi.org/10.5194/amt-2021-282-RC2, 2021

Comment on amt-2021-282

Anonymous Referee #2

Referee comment on "Machine learning techniques to improve the field performance of low-cost air quality sensors" by Tony Bush et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-282-RC2, 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:

- 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).
- 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. 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. 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.

- 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.
- 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 behavior of calibrated measurement systems.
- 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.
- 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.

Other comments:

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).

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.

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.

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.