Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2017-260-RC2, 2017

Atmospheric Measurement Techniques
Discussions

Interactive comment on "Closing the gap on lower cost air quality monitoring: machine learning calibration models to improve low-cost sensor performance" *by* Naomi Zimmerman et al.

Anonymous Referee #2

Received and published: 17 September 2017

This paper explores the performance of a low-cost sensor unit for measurement of urban air quality. A total of nineteen multi-pollutant sensor packages (called RAMP) which measure CO, NO2, O3 and CO2 as well as rH and Temp have been used from August 2016 through February 2017 in Pittsburgh PA next to an air quality monitoring site where reference instruments have been operated. Measurements from the reference instruments have been used as independent variables for investigation of different models for the calibration of the low-cost sensor units. The responses from all sensors in the RAMP units have been used for prediction of the pollutant concentration. It was found that calibration models based on a machine learning technique (Random Forests, RF) performed much better than (multiple) linear regression models. The authors find that the combination of the RF calibration approach and the multi-pollutant sensor package accounts for pollutant cross sensitivities and is a promising approach for the use of low-cost air quality sensors. The manuscript covers a relevant and emerging topic and adds to a growing number of studies on sensor calibration and performance of low-cost sensors. However, there are some technical flaws in the manuscript that should be corrected. The manuscript can be published in AMT after consideration of the following comments.

The overall message of the manuscript is in my view too optimistic and can for readers be misleading. The authors should make clear that the good performance of the sensors found in this calibration study does not imply that the sensor unit is capable of providing similarly accurate air quality measurements in a real-world application. A good performance of sensor units in a calibration exercise like the study at hand is certainly necessary but not sufficient for the suitability of the sensors for real world air quality measurements. It should be clear that the manuscript is targeting on the good data quality obtained when combining the multi-pollutant sensor unit and RF and that a full assessment of the performance of the RAMPs within a sensor network for air quality measurements under real world conditions requires future research (and solutions for the quality assurance and quality control of the deployed sensors). The authors touch this point briefly in the conclusions section, however, for readers the impression remains that the RAMPS sensor units are ready for being used for urban air quality assessments. For example, in the conclusions section, last paragraph, it is stated that "Overall, we conclude that with careful data management and calibration using advanced machine learning models, that low-cost sensing with the RAMP monitors may significantly improve our ability to resolve spatial heterogeneity in air pollutant concentrations.". This conclusion is not justified by the available study and should be kept for the future work when results on the data quality as obtained in real world applications are available. As another example, the authors write on page 14, lines 14-16 "The US EPA limit of detection for federal regulatory monitors is 10 ppb for both NO2 and O3, suggesting that as with CO, the RF model performance is within 20% of regulatory standards (United States Environmental Protection Agency, 2014)". This is again misleading: It can be concluded from this calibration study that the performance of sensors with an updated calibration meet those requirements, the data quality that can be achieved with the sensor under real world conditions is something different and currently not known. Please revise the text carefully.

Another point that I find irritating and that should be rephrased is the last sentence in the abstract ("From this study, we conclude that combining RF models with the RAMP monitors appears to be a very promising approach to address the poor performance that has plagued low cost air quality sensors.") and again on page 3 lines 1-3 ("as poor signal-to-noise ratios may hamper their ability to distinguish between intra-urban sites. As such, there has been increasing interest in more sophisticated algorithms (e.g., machine learning) for low cost sensor calibration."). These two statements are misleading as they imply that the limiting factor of sensor based data is data processing and not the gas sensing unit itself. It is well known that there are sensors available that are not sensitive and selective enough for the measurement of air pollutants at ambient concen-trations. Sophisticated algorithms will not be able to help here. The text should be changed so that the message of the paper is that sophisticated algorithms can improve the performance of those sensors that are generally suited for the measurement of ambient air pollutants.

On page 8, second paragraph it is stated that "The random forest model's main limitation is that its ability to predict new outcomes is limited to the range of the training dataset; in other words, it will not predict data with variable parameters outside the training range.". This is a relevant and important point and should further be discussed, i.e. the authors should elaborate on the practical consequences for using sensors. For example, the calibration model for O3 might not be applicable for peak summer concentrations when the training data has been measured during the cold season (how is the situation here, training data has been measured form August to February, is it applicable for peak ozone as typically observed in June/July?). This issue is even more

important for a multipollutant unit like the RAMP as pollutants like ozone have highest concentrations during summer and other primary pollutants often show highest concentrations during the cold season. Does this mean that calibration measurements need to cover a whole year, or what are the strategies for dealing with this situation?

The average Pearson correlation coefficients (e.g. the 0.99 for LAB and RF – even for CO2) are hardly to believe, given e.g. the scatter plots in Figure 4. There is a lot of scattering for all pollu-tants. On page 11 (line 5) the authors mention "The poor performance of linear models at predict-ing CO2 concentration is not surprising . . .". why then r=0.99 in Table 2? This needs to be checked or requires a convincing explanation. In addition, on page 11 line 31 it is said that "the Pearson r for NO2 ranged from 0.92 to 0.95". Again, this is very hard to believe, looking at Figure 5 there are a few RAMPs where I expect that r is smaller than 0.92 (e.g. #4, #6 #19). Please correct, or add the r values to the plots in Figure 5.

Other comments: The authors use alternately the terms "multivariate linear regression" and "multiple linear regres-sion". The method applied here is multiple linear regression and not multivariate linear regression which is something different. Use solely the term multiple linear regression.

On page 4, lines 20-21. The RAMP version with PM2.5 sensor does not need to be mentioned here since PM2.5 measurements are not used in the study. The notation of equations 1 and 2 is poor and should be improved. The measurements with the reference instruments are used in the models as independent variables, this should be clear. So use something like y_reference (t) = . . . instead of Corrected_MLR etc.

Page 8, line 22. The software package R should be correctly cited, see citation() in R.

Page 10, first paragraph. What is "the standard deviation of the model"? Is this the standard devi-ation of the model predictions? Please be clear and correct.

Page 12, line 8: "Smaller bias of RF models than the reference method?" Do you really mean that the RF corrected sensor data have a smaller bias than the reference? How can this be, the reference measurements have been used as independent variable for training the RF models.

Page 14, line 9, it was found that the CO signal was the most important variable in the RF model for CO2. This likely poses strong limitations for using calibrated CO2 sensors in another environment than the location where the training data was obtained. The sensor calibration can likely not be transferred to rural environments, i.e. away from combustion sources, were CO and CO2 might not be strongly in-terlinked. What about measurements during the vegetation period, when CO2 uptake by plants can changes the relationship between CO2 and CO2 in urban environments? The authors should address this issue.

Legend of Figure 11 is wrong, should be RAMP vs. Reference, not the other way around.

Page 15, lines 24-26: “For NO2, the performance of ‘out-of-the-box’ low-cost sensors varied widely and half the sensors in the EuNetAir study (Borrego et al., 2016) reported errors larger than the average ambient concentrations. Therefore, advanced calibration models, such as those using machine learning, are critical to accurate measurements of ambient NO2.”. As mentioned earlier, this is too simple and is neglecting the requirements for the gas sensing unit. If the sensor strongly responds to other factors than covered by the available predictors, not even advanced calibration models can be successful. So the quality of the sensing unit itself is key. The text should be revised.

Figure 12: More relevant than the slope and intercept of the regression of the RAMP against the reference would be the uncertainty of the daily 8h max as measured with the RAMP. This could be expressed by corresponding confidence intervals.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2017-260, 2017.