Response to Comments from Reviewer #2 AMT-2017-260

The authors would first of all like to thank reviewer #2 for the insightful comments on the work we have submitted for publication, and the editor for the opportunity to improve the manuscript. Under each comment there is a summary of the response (red text), in addition to the text from the paper that was modified, if applicable.

Reviewer #2

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.

Response: We thank the reviewer for the comments on the manuscript. Respectfully, our calibration represents a real-world application tested under real-world conditions. The development and testing of the calibration occurred outdoors in an urban background environment with variable local sources such as passenger vehicles, trucks and restaurant emissions from nearby restaurants on the Carnegie Mellon campus. As such, there were real-world variants between the training and testing data. We do agree that the manuscript as written in the original submission only focused on testing data from RAMP monitors also at the Carnegie Mellon campus. To further demonstrate the suitability of the calibrated RAMP monitors in other real-world environments where traditional reference monitors were deployed, we have modified the manuscript to include a comparison of a RAMP monitor calibrated at Carnegie Mellon and then moved to the Allegheny County Health Department (ACHD), where there is an independent set of reference monitors for

CO, NO_2 and O_3. The ACHD site has more nearby sources than the Carnegie Mellon site (more traffic and restaurants) and different land use classifications. Comparing the CMU calibrated RAMP to the ACHD data, we found good agreement for the pollutants at similar performance levels (based on CvMAE, Pearson r) to the testing data originally presented in the manuscript. These results are included in a new Figure 14 with additional text. We have also added additional wording to Section 2.2 to indicate the nature of the real-world environments tested as part of this study.

Additional text in Section 4.5 (new text in bold), followed by the new Figure 12 (old Figures 12-14 now shifted by one):

"We further assess the RAMP monitor performance against **three** metrics: **1) comparison of a RAMP monitor calibrated at Carnegie Mellon against an independent set of regulatory reference monitors at the Allegheny County Health Department**, 2) for NAAQS compliance, and 3) for suitability for exposure measurements as per the US EPA Air Sensor Guidebook (Williams et al., 2014). We also demonstrate the benefit of improved performance of the RF models in a real-world deployment at two nearby sites in Pittsburgh, PA.

From February through April 2017, a RAMP calibrated at the Carnegie Mellon Campus was deployed at the Allegheny County Health Department (ACHD) to test the performance of the RAMP relative to an independent reference monitor (Figure 12). The ACHD reports data hourly, so RAMP data were down-sampled to hourly averages and the CO, NO2 and O3 concentrations were compared (no measurement of CO2 is made at ACHD). For all pollutants, R2 was ≥0.75 (CO: 0.85, NO2: 0.75, O3: 0.92) and points were clustered around the 1:1 line. NO2 performed the most poorly, with a large cluster of points in the 5-10 ppb range where the model is known to underperform. The MAE was 49 ppb (17% CvMAE) for CO, 4.7 ppb for NO2 (39% CvMAE) and, 3.2 ppb for O3 (16% CvMAE), in line with the performance metrics in Figure 6."

Additional description of the ACHD site in Section 2.1:

"**The RAMP monitors have also been intermittently deployed across the Pittsburgh region as part of ongoing air quality monitoring research. To demonstrate the accuracy of the calibrated RAMP, we also show data from a RAMP monitor which was first calibrated at Carnegie Mellon University and then moved to the Allegheny County Health Department (ACHD, 40°27'55.6"N, 79°57'38.9"W) from February – May 2017. The ACHD site has independent reference monitors for CO, NO_2 and O_3 and thus comparing data from these two sites enables an independent real-world assessment of model performance. The ACHD site is characterized by increased traffic volume, restaurant density and industry relative to the Carnegie Mellon site.**"

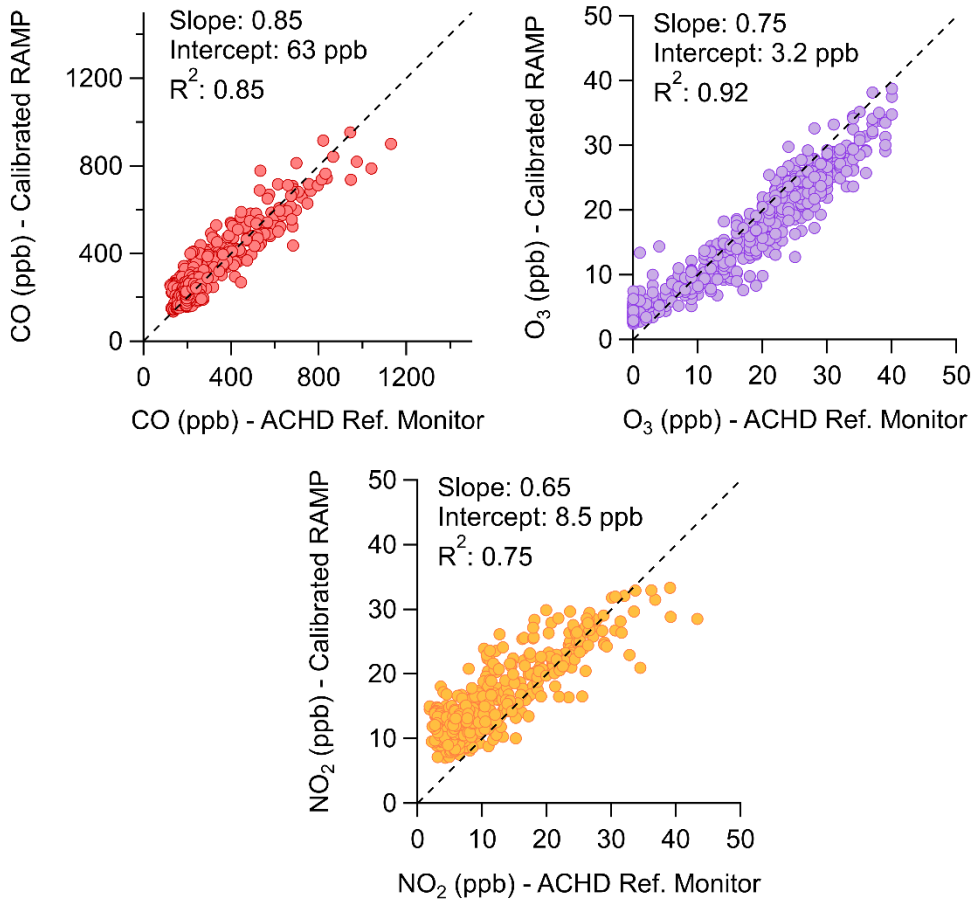

Figure 12: Comparison of CO, NO_2 and O_3 hourly average concentrations measured by a co-located RAMP monitor and the reference monitors at the Allegheny County Health Department (ACHD). The RAMP monitor was first calibrated on the Carnegie Mellon campus prior to deployment.

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 concentrations. 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.

Response: We agree that there are some gas sensing units that will never be suitable for air quality measurement applications and we have modified the text to more directly to address the numerous limiting factors for low-cost sensors. In the abstract, we only make this claim regarding our specific

unit (which is suited to measurement of ambient air pollutants), and not all gas sensing units, thus we have left the abstract unchanged.

Modified text in introduction:

"The two primary requirements of low cost sensors for ambient measurement are 1) hardware that is sensitive to ambient pollutant concentrations, and 2) calibration of the sensors. The latter is the focus of this study. A primary challenge of low-cost sensor calibration is that the sensors are prone to cross-sensitivities with other ambient pollutants (Bart et al., 2014; Cross et al., 2017; Masson et al., 2015b; Mead et al., 2013)"

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?

Response: We agree that this is a critical point to further emphasize. We have changed some of the language and added additional text on a possible solution for extrapolating, such as a hybrid RF and MLR model, where the MLR model is used for concentrations beyond the 95th percentile of the training data.

Modified text in Section 3.3 (additions in bold)

"The random forest model's **critical** limitation is that its ability to predict new outcomes is limited to the range of the training data set; in other words, it will not predict data with variable parameters outside the training range (no extrapolation). Therefore, a larger and more variable training data set should create a better final model. **In this study, our collocation window covered a broad range of concentrations and meteorological conditions; however, in situations where shorter collocation windows with less diverse training ranges are desired, the RF model may not be suitable as a standalone model. This is discussed further in Section 4.3.2.**"

Modified text in Section 4.3.2 (additions in bold)

"To build a robust model, many data points are required at a given concentration to probe the extent of the ambient air pollutant matrix. In this study, the training windows were dispersed throughout the collocation period to ensure good agreement of gas species and meteorological conditions during both the training and testing windows (see Supplemental Information). **The RF model may not work well in cases where such a diverse collocation window is not possible or where concentrations are routinely expected to exceed the training window. In such**

situations, hybrid calibration models such as combined RF-MLR where MLR is used for concentrations above the training window range may be suitable, as MLR tends to perform better when concentrations are higher."

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 pollutants. On page 11 (line 5) the authors mention "The poor performance of linear models at predicting 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.

Response: The numbers in Table 2 correspond to the goodness of fit of the model (i.e., performance of the withheld folds in the training data). As such, the scatter plots in Figure 5 are not related to Table 2 – but to the scatter plots in the Supporting Information (SI Figures S3-S6). The data shown in Figure 5 is for testing data (Section 4.2) We have added references to the SI figures in the text directly to minimize confusion, and modified the caption of Table 2 to direct the reader to Section 4.1 (discussion of goodness of fit). Additionally, as pointed out by reviewer number 1, there was a typo in the manuscript, the statement that the Pearson r varied from 0.92-0.95 is for O_3, not for NO_2 and was a simple typographical error that has been corrected.

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

Response: Thank you for noticing this – we have corrected all instances of "multivariate linear regression" to "multiple linear regression", these were typographical errors.

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.

Response: The two instances where PM2.5 are mentioned in Section 2.2 have been removed, as Reviewer #2 is correct that they are not used in the study. We have also revised the notation of Equations 1 and 2

Modified Equations 1 and 2 below:

$$y_{\text{reference}}(t) = \beta_0 + \beta_1 \times [\text{Net Sensor Response } (CO, NO_2) \text{ or Raw Sensor Response } (CO_2)], \quad (1)$$

$$y_{\text{reference}}(t) = \beta_0 + \beta_1 \times [\text{Net Sensor Resp. } (CO, NO_2, O_3) \text{ or Raw Sensor Resp. } (CO_2)] + \beta_2 \times T + \beta_3 \times RH, \quad (2)$$

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

Response: We agree that the correct way to cite an R package is using citation() in R, and this is how the citation was generated. There appeared to be an issue translating the BibTeX file into the document, which we have now resolved.

Modified citation:

"Kuhn, M., Contributions from: Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., The R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. and Hunt., T.: caret: Classification and Regression Training, [online] Available from: https://cran.r-project.org/package=caret, R package version 6.0-76, 2017."

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

Response: The reviewer is correct, we mean the standard deviation of the model predictions. We have modified the text accordingly.

Modified text (additions in bold)

"Since CRMSE is always positive, a further dimension is added: if the standard deviation of the model **predictions (calibrated sensor data)** exceeds the standard deviation of the reference measurements, the CRMSE is plotted in the right quadrants and vice versa. To match previously constructed target diagrams (Borrego et al., 2016; Spinelle et al., 2015, 2017), the CRMSE and MBE were normalized by the standard deviation of the reference measurements, and thus the vector distance in our diagrams is RMSE/σreference (nRMSE). The resulting diagram enables visualization of four diagnostic measures: (1) whether the model tends to overestimate (MBE > 0) or underestimate (MBE < 0), (2) whether the standard deviation of the model **predictions (calibrated sensor data)** is larger (right plane) or smaller (left plane) than the standard deviation of the reference measurements, …"

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.

Response: As written, the manuscript states the RF model responses were "biased slightly lower" than the reference measurements, which we mean as "tend to underpredict" (negative MBE). This is not the same as saying the RF calibrated sensor data has less bias than the reference monitors, which we agree is not possible. We have rephrased to make this clearer.

Modified text in Section 4.2 (additions in bold):

"Across all gases, the RF models on average were **biased towards predicting** concentrations slightly lower than the reference **(i.e., slight tendency to underpredict, MBE/$\sigma_{reference}$ <0)**."

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 interlinked. 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.

Response: We agree that given the dependence of the CO_2 calibration on the CO signal that the sensors would likely not be suited for rural environments. We have added additional text to the manuscript to emphasize that these models would likely only perform best in urban environments unless a custom calibration was built in a rural environment.

Modified text in Section 4.3.1 (additions in bold):

"The explanatory variable importance is more complex for CO_2 and NO_2. For CO_2, all variables are important roughly equally important, with CO being the most important. This is likely due to the strong meteorological effect of humidity on the measured CO_2 concentration; the model must rely on other primary pollutants to predict CO2 signal when the measured CO_2 has reached full-scale, and short-term fluctuations of CO_2 are likely from combustion sources (e.g., vehicular traffic in urban areas) which also emit CO. This highlights the value of having sensors for multiple pollutants in the same monitor. Including measurements of additional pollutants helps the RF model correct for cross-sensitivities. **However, the drawback of this cross-sensitivity in the model is that the RF model may not perform well in areas where the characteristic source ratios of CO and CO_2 have changed. For example, this model was calibrated in an urban environment with many traffic and combustion-related sources nearby. Such a model would be expected to perform poorly for CO_2 in a heavily vegetated rural environment where CO and CO_2 are not strongly linked**."

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

Response: Thank you for noticing this, the caption of Figure 11 has been corrected.

Modified caption (changes in bold):

Figure 11: Illustrating the range of predictions from the 500 trees for RAMP #1. The testing data were binned and averaged. The concentration measured by the ~~reference~~ **calibrated RAMP monitors** is then plotted against the average concentration from the ~~model~~ **reference monitor**. The error bars represent the standard deviation of the answers from the 500 trees and the bins are colour coded by the number of data points within each bin. The dashed black line is the 1:1 line.

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.

Response: We agree with the Reviewer's assessment that calibration alone cannot correct all sensors and we have modified the text to emphasize the complexity of the problem

Modified text in Section 4.4 is below (additions in bold):

For NO_2, 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. **While the quality of the baseline gas sensing unit remains critical (in which case no calibration should work), we suggest that advanced calibration models, such as those using machine learning, may be critical for accurate measurements of ambient NO_2.**"

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.

Response: We have added 95% confidence intervals to the MLR and RF-calibrated RAMP monitor daily 8h max as compared to the reference monitors to incorporate this uncertainty.

Modified Text in Section 4.5 (additions in bold):

"For the representative RAMP monitor used previously (RAMP #1), daily maximum 8-hour O_3 was in good agreement between the RF calibrated RAMP and the reference monitor, with all data points falling roughly along the 1:1 line (slope: 0.82, **95% CI: 0.81-0.83**), while for the MLR model, concentrations were skewed slightly low (slope of 0.65, **95% CI: 0.63-0.67**)."