Response to Comments from Reviewer #1 AMT-2017-260

The authors would first of all like to thank reviewer #1 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 #1

The title is a bit ambitious, ambiguous, or both. How much of the performance "gap" is closed by a) improved hardware compared to past studies, b) the algorithm (i.e., Random Forest), c) sensor combinations at each node, and d) range of different sample types collected? Application of machine learning for sensor calibration in the field has been performed before, but the title and abstract seems to give the impression that this reduces the gap. There is much focus given to RF but there is no indication that it has an inherent advantage over other machine learning methods. For instance, it is possible that a MLR model could also handle cross-sensitivities only if it were provided all variables (though RF and other machine learning algorithms are more flexible in that it does not require the assumption regarding global linearity).

The past work of De Vito et al. (2008, 2009) also show encouraging results from a long-term evaluation of field calibrations (for low-cost multi-sensor devices for benzene, CO, and NO2 against government monitoring station instruments using machine learning algorithms).:

De Vito S., Massera E., Piga M., Martinotto L., and Di Francia G.: On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical, 129(2):750–757, doi:10.1016/j.snb.2007.09.060, 2008.

De Vito S., Piga M., Martinotto L., and Di Francia G.: CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization, Sensors and Actuators B: Chemical, 143(1):182–191, doi:10.1016/j.snb.2009.08.041, 2009.

Response: Thank you for suggesting the papers by De Vito et al. and for correctly pointing out that the title is too bold. We have revised the title to: **"A machine learning calibration model to improve low-cost sensor performance".** We have also added references to De Vito et al. 2008 and 2009:

Modified text in Introduction (additions in bold):

"To date, there have been published studies using high-dimensional multi-response models (Cross et al., 2017) and neural networks (Esposito et al., 2016; Spinelle et al., 2015, 2017, **De Vito et al., 2008, 2009**). Spinelle et al. (2015) showed that artificial neural network calibration models could meet European data quality objectives for measuring ozone (uncertainty < 18 ppb); however, meeting these objectives for NO_2 remained a challenge. **In De Vito et al. (2009), the neural network calibration approach was applied to CO, NO_2 and NOx metal oxide sensors in Italy with encouraging results; in general mean relative error was approximately 30%.**"

The manuscript is perhaps too bold in its tone. Accurate predictions are shown for concentration (and T, RH) domains that are present at the location of the reference monitor used for calibration,

even while using different data points. (As stated by the authors, current implementation of RF is limited to the domain of the training set.) Dense network coverage implies monitor placement in different microenvironments (e.g., nearroadway, etc.) which would experience different concentration regimes. Moreover, some of the explanatory variables used for calibration may be surrogates for another variable which may vary differently at another site. There is mention of two RAMPS units deployed in Pittsburgh and their positive evaluation against other reference measurements in a mobile van (p. 17, line 15), but no results are shown.

Response: Due to the currently long length of the manuscript, we have elected to not go into details of the mobile van measurements, and they will be presented in a forthcoming publication. However, we did deploy a RAMP that was calibrated at Carnegie Mellon University at the Allegheny County Health Department (ACHD) in February 2017-May 2017 and observed good agreement between the hourly ACHD concentrations of O_3, NO_2 and CO and the calibrated-RAMP. We have modified the manuscript to include this additional figure.

Below is the complete Section 4.5, which was re-organized to improve narrative flow and now includes the ACHD assessment (additions in bold), followed by the new Figure 12.

"4.5 RF model calibrated RAMP performance in a monitoring context

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, NO_2 and O_3 concentrations were compared (no measurement of CO_2 is made at ACHD). For all pollutants, R^2 was ≥0.75 (CO: 0.85, NO_2: 0.75, O_3: 0.92) and points were clustered around the 1:1 line. NO_2 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 NO_2 (39% CvMAE) and, 3.2 ppb for O_3 (16% CvMAE), in line with the performance metrics in Figure 6.

Regulatory agencies must also report compliance with National Ambient Air Quality Standards (NAAQS). In this study, the time resolution and methods used to assess the effectiveness of the RF models (15 min) do not match the metrics used for NAAQS. For example, the NAAQS standard for O_3 is based on the maximum daily maximum 8-hour average, and compliance for NO_2 is based on the 98th percentile of the daily maximum 1-hour averages. While acknowledging that the RAMP monitor collocation period was shorter than typical NAAQS compliance periods (e.g. annually for O_3 and across 3 years for NO_2) it is still worth characterizing

the RAMP performance using the LAB, MLR and RF models (Figure 13). 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**). For NO_2, the 98th percentile of the daily maximum 1-hour averages was 34 ppb for the RF model versus 35 ppb measured using a reference monitor compared to 25 ppb for the MLR model and 51 ppb for the LAB model. The RF model was substantially closer to the reference monitor estimate and the underestimation was only by 1 ppb. Other RF model calibrated RAMP monitors performed similarly, all agreeing within 5 ppb.

Air sensor performance goals by application area are also provided by the US EPA Air Sensor Guidebook (Williams et al., 2014). The performance criteria include maximum precision and bias error rates for applications ranging from education and information (Tier I) to regulatory monitoring (Tier V). The precision estimator is the upper bound of a 90% confidence interval of the coefficient of variation (CV) and the bias estimator is the upper bound of a 95% confidence interval of the mean absolute percent difference between the sensors and the reference (full equations in the Supplemental Information). An overarching goal of RAMP monitor deployments is to use low-cost sensor networks to quantify intra-urban exposure gradients, thus our benchmark performance was Tier IV (Personal Exposure), which recommends that low-cost sensors have precision and bias error rates of less than 30%. For the testing (withheld) periods, we compared the performance of the RF, MLR and LAB models for all the RAMP monitors used in this study to the precision and bias estimators recommended by the US EPA (Figure 1). The performance across the RAMP monitors was summarized using box plots, and only the RF model calibrated RAMPs are suitably precise and accurate for Tier IV (personal exposure) monitoring across CO, NO_2 and O_3. Furthermore, both RF model calibrated CO and O_3 RAMP monitor measurements were below the even more stringent Tier III (Supplemental Monitoring) standards, which recommends precision and bias error rates of <20%. The RF model NO_2 RAMP measurements may reach Tier III in locations with higher NO_2 concentrations.

To demonstrate the improved performance of the RF models in a real-world context, two of the RAMPs used in the evaluation study were deployed for a 6-week period at two nearby sites in Pittsburgh, PA. One RAMP monitor was located on the roof of a building at the Pittsburgh Zoo in a residential urban area, and another was placed approximately 1.5 km away at a near-road site located within 15 m of Highway 28 in Pittsburgh (Figure 15). NO_2 concentrations are known to be elevated up to 200 m away from a major roadway compared to urban backgrounds due to the reaction of fresh NO in vehicle exhaust with ambient O_3 (Zhou and Levy, 2007). Figure 13 shows the diurnal profiles of the RAMPs at the two locations evaluated using the RF and MLR models. The RF model indicates an NO_2 enhancement of approximately 6 ppb at the near-road site (Figure 15, red trace) compared to the nearby urban residential site (Figure 15, blue trace) and there are notable increases in NO_2 during morning and evening rush hour periods, as expected. The concentrations reported by the RF model calibrated RAMPs were further verified with measurements using a mobile van equipped with reference instrumentation at different periods throughout the day. However, applying the MLR model to the RAMP data reveals no significant

difference between the two sites (Figure 15, bottom diurnal). In fact, the MLR model predicts negative concentrations during the day. The results of this preliminary deployment suggest that the RF model calibrated RAMPs could be suitable for quantification of intra-urban pollutant gradients."

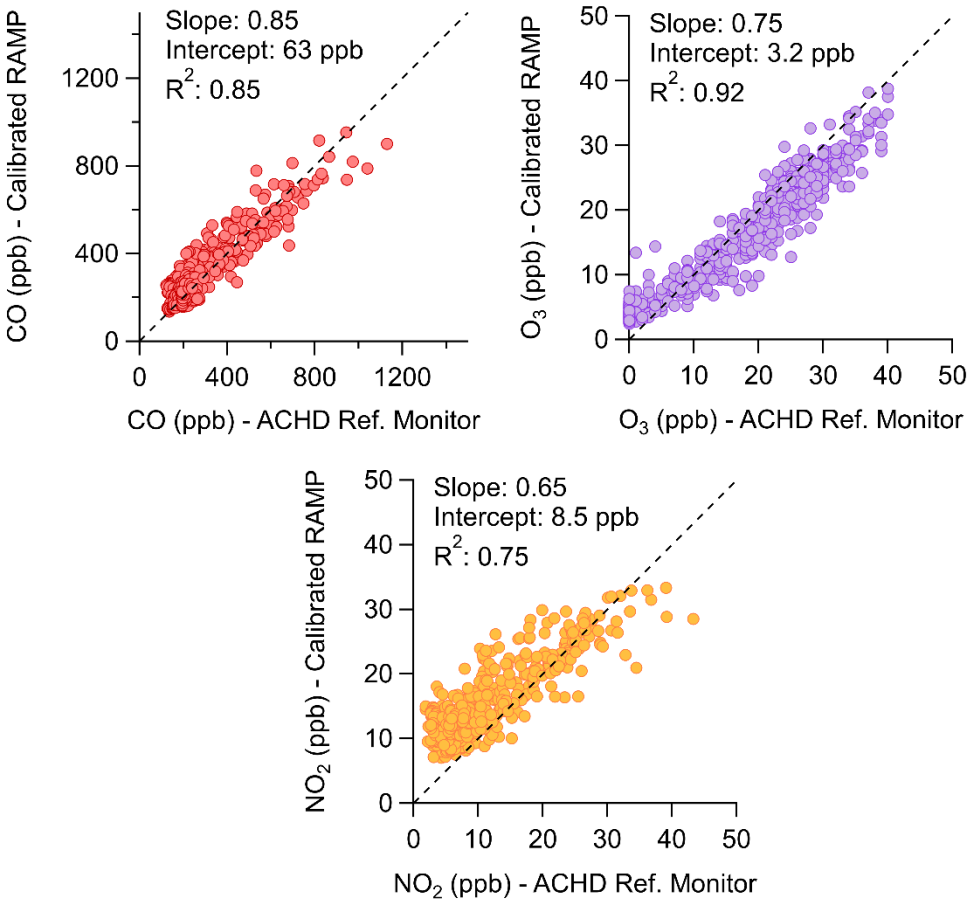

Figure 12: Comparison of CO, NO_2 and O_3 hourly average concentrations measured by a co-located RAMP monitor and the reference monitors at the Allegheny County Health Department (ACHD). The RAMP monitor was first calibrated on the Carnegie Mellon campus prior to deployment.

Since corrections of the supersite reference monitors against the Allegheny County Health Department instruments are necessary, why not make this Allegheny County Health Department site the reference site? Given the local contributions of vehicle emissions to CO and NO2 that are present in the parking lot site, how were the corrections for baseline drift determined?

Response: We have added two sentences to section 2.3 to describe the baseline correction approach. We would like to emphasize that the baseline corrections were modest and did not substantially affect the dataset from our reference monitors. The incentive for using the Carnegie Mellon site as the reference monitoring station is due to the higher time resolution of the data (we report at 1 Hz), the availability of the data in near-real time, and the ability to explore calibrations for pollutants not measured at the Allegheny County Health Department (ACHD) (e.g., CO_2). Given the large numbers of RAMPs and availability of reference-grade instruments at CMU, the

CMU Supersite was much easier to access and hence used as the reference site. Other users who do not have the facilities we do could use their local regulatory monitors as a reference site if accessible.

Modified text in Section 2.3 (additions in bold)

"The CO and NO_2 analyzers experience modest baseline drift between weekly calibrations, on the order of approximately 40 ppb for CO and 2 ppb for NO_2. Hence, baseline pollutant concentrations were normalized to a nearby regulatory monitoring site (Allegheny County Health Department, Air Quality Division, Pittsburgh, PA). **The baseline correction was done using a linear regression between the beginning and end of the week on the baseline signals (local source spikes removed). The regression was based on daytime differences, as night time inversions may cause real differences in the baseline signals between the two sites.**"

While the authors describe the use of 5-fold CV to selection the explanatory variables to use, the choice of 5 data points per terminal node / 100 trees per fold does not seem to be explained. This was also selected in the CV process?

Response: The typical range of cross-validations that are explored is from 3-20 folds. We observed that by 5 folds, the model performance had roughly stabilized, thus to optimize computational power we chose the minimum number of folds such that an increase in folds produced a <5% increase in model RMSE and R^2. Similarly, random forests are typically constructed with 64-128 trees, so we chose a number in the middle of this range (100 trees). We agree that these details should be included in the manuscript, and have been added to Section 3.3.

Modified text in Section 3.3 (additions in bold):

"The number of trees was capped at 100 per fold, and a five-fold cross-validation was used for a total of 500 trees. Therefore, the predicted value for a given set of measured inputs is the average value from this set of 500 trees (each tree provides one prediction). **The k-value was chosen by identifying the minimum number of folds for which an increase in the fold size increased model performance less than 5% on the held-out data. The number of trees was chosen based on the work of Oshiro et al. (2012), who suggested that the number of trees range from 64-128**."

p. 14 Line 18 paragraph: Is this not possibly a limitation of the hardware?

Response: In this instance, we do not believe it is a limitation of the hardware. In our laboratory calibrations, we have exposed the sensors to several ppm of NO_2 and have not observed a flat response (i.e., sensors are sensitive at high concentrations).

Minor comments:

Section 2.2: Data coverage (i.e., missing data) and the time resolution should be stated here rather than (or in addition to) later in the manuscript.

Response: Thank you for this comment that has also been pointed out by other reviewers. We have been more upfront with missing data and time resolution earlier in the manuscript to make the scope of the work clear.

Modified text in Section 2.2 (additions in bold)

"The experiments involved 95 individual pollutant sensors mounted in 19 unique RAMP monitors. **While the collocation period spanned August 2016-February 2017, some sensors were intermittently deployed for air quality campaigns in Pittsburgh, thus the range of collocation available ranged from 30 days to the full collocation period, depending on the unit. Additionally, calibrations were not built for sensors for which reference data was below detection limits or if reference monitoring units were malfunctioning, reducing the total number of sensors in this experiment to 73, due to issues with the SO_2 and NO_2 monitors.**

The electrochemical sensor outputs were measured using electronic circuitry custom designed by SenSevere optimized for signal stability. The circuitry includes custom electronics to drive the device, multiple stages of filtering circuitry for specific noise signatures, and an analog-to-digital converter for measurement of the conditioned signal. The RAMP monitors are housed in a NEMA-rated weather proof enclosure (Figure 1A) and equipped with GSM cards to transmit data using cellular networks to an online server. The RAMP monitors also log data to an SD card as a fail-safe in case of wireless data transfer issues. **The data is logged to the server at ~15 second resolution and down-sampled to 15-minute averages, which was deemed to be an appropriate time resolution for assessing spatial variability in air pollution exposure and to reduce the size of the dataset. Regulatory bodies typically make their data available at hourly resolution.**"

P. 9 Line 15 to end of paragraph. The authors switch from describing "intermittent" collocation to "distributed" collocation. Given the discussion of multiple RAMP monitors, "distributed" can be confusing. Also, "degree of collocation" is referring to frequency or effective duration?

Response: Thank you for pointing this out, we agree that it is confusing. We have switched the terminology to "consecutive" and "non-consecutive" collocations.

Modified Text in Section 3.3 (additions/changes in bold)

"This was evaluated for a consecutive collocation window and for 8 **non-consecutive** collocation windows equally distributed throughout the whole collocation period (August 2016 – February 2017) in half week increments. Details of this evaluation are provided in the Supplemental Information, but the **non-consecutive** collocations generally performed slightly better, with reductions in MAE of 12 ppb (4% relative error) for CO, 2 ppm for CO_2 (0.4% relative error), 0.4 ppb for NO_2 (4% relative error), and 1.6 ppb for O_3 (7% relative error) compared to the consecutive four-week collocation. The motivation for exploring **non-consecutive** collocation windows dispersed throughout the study period was to ensure that the training period covered a complete range of gas species concentrations, temperatures and relative humidity. In practice, **the training data** utilized in this study is equivalent to collocating the RAMP monitors with reference monitors for 3-4 days every 1-2 months. **If non-consecutive collocation is inconvenient or not possible,**

consecutive collocation may be satisfactory as determined by MAE and other accuracy parameters needed for the application at hand."

p. 10 Line 19: value of correlation for NO2 and CO2 with reference monitors is missing.

Response: Thank you, this has been added.

Modified text in Section 4.1 (additions in bold):

"However, only the RF model achieved strong correlations between the reference monitor and the RAMPs for NO_2 and CO_2 **(Pearson r: 0.99)**."

p. 10 Line 22: insert figure numbers (SI Fig S3-S6).

Response: Thank you, this has been added.

Modified text in Section 4.1 (additions in bold):

"Regression plots for all 19 RAMPs and all four gas species illustrating the goodness of fit of the RF model are provided in the Supplemental Information **(Figures S3-S6)**."

p. 10 Line 30: The relationship between m_try and model complexity is not very clear.

Response: We have edited Section 4.1 to add additional details to help make this connection clearer. In general, by having a larger m_try, there is a higher probability that one dominant variable will be what the split is decided on. In other words, there is a lower probability that all the variables will participate in the model structure. If the model performance improves by diversifying the variables it splits based on, it is generally considered to have a more complex underlying structure. We have modified the text to better convey this point.

Modified Section 4.1 below (additions in bold)

"In general, the larger the mtry, the simpler the underlying structure of the model. For example, if there is one dominant variable but the model is permitted to consider all 7 explanatory variables at each decision node (i.e., mtry=7), then the model will most frequently split the data based on the dominant variable. **By contrast,** the advantage of a lower mtry is that subtle relationships between explanatory variables and the response can be probed. **When randomly selecting fewer explanatory variables (mtry=2 or 4) at each decision node, the probability of selecting a dominant variable decreases and the model is forced to split the data into sub-nodes based on variables which may have a smaller (but real) effect on the response**. If the goodness of fit of the calibration model is improved by decreasing mtry, this suggests more complex variable interactions **with the response** (Strobl et al., 2008)."

p. 11 Line 13: "clearly outperformed" -> not for CO

Response: As a general theme, we have toned down the language. We agree that for CO, any calibration seems to perform well and have modified the manuscript to reflect this.

Modified text in Section 4.2 (additions in bold, also removed the word "clearly":

"For this period, the RF model ~~clearly~~ outperformed the LAB and MLR models **for all pollutants except for CO**."

p. 11 Line 21: insert figure numbers (SI Figs S7-S10). Slopes, correlations, or some of the metrics listed in Table S2 included in the panels would be informative. Why are some RAMPS not included?

Response: Thank you, we acknowledge that why some RAMPs were not included was not totally clear, so we have made several revisions throughout the manuscript to be more descriptive of the calibration and collocation process. The total study domain was from August 2016 – February 2017, but RAMP monitors were intermittently deployed for air quality campaigns, so the average collocation period ranged from 5.5-15 weeks (median 9 weeks). After determining that 4 weeks of data was needed for proper calibration, some RAMP monitors did not have sufficient data to build a complete model (only 16 of the 19 RAMPs for NO_2) and some did not have enough data for a meaningful testing period (minimum threshold 48 hrs, actual test window: 1.4-15.5 weeks). Thus for testing the model, the total number of RAMP monitors was reduced to 16 for CO and O_3, 15 for CO_2 and to 10 for NO_2. We have modified the text in several sections to indicate this more clearly, with one example shown below. We have also added references to the Figure numbers in the text, and added the MAE and Pearson r metrics to the panels in Figures S7-S10, as requested (not showing here due to size of Figures, but is in Revised Manuscript).

Modified text (additions in bold) in Section 4.2:

"To assess the overall model performance, two performance metrics (Pearson r and CvMAE) were calculated for each RAMP monitor using the entire testing dataset (Figure 6). **In this study, any data remaining after training were used to test model performance, provided there were at least 48 hours of testing data (192 data points). This reduced the number of RAMP monitors included for testing the model to 16 for CO and O_3, 15 for CO_2 and 10 for NO_2.** The size of the testing dataset varied from 1.4 to 15 weeks, with a median value of 5 weeks.

p. 11 Line 31: "NO2" -> "O3" here?

Response: Yes, thank you, that was a typographical error and has now been corrected.