Response to Comments from Reviewers: AMT-2017-260

The authors would first of all like to thank the reviewers 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 May 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."

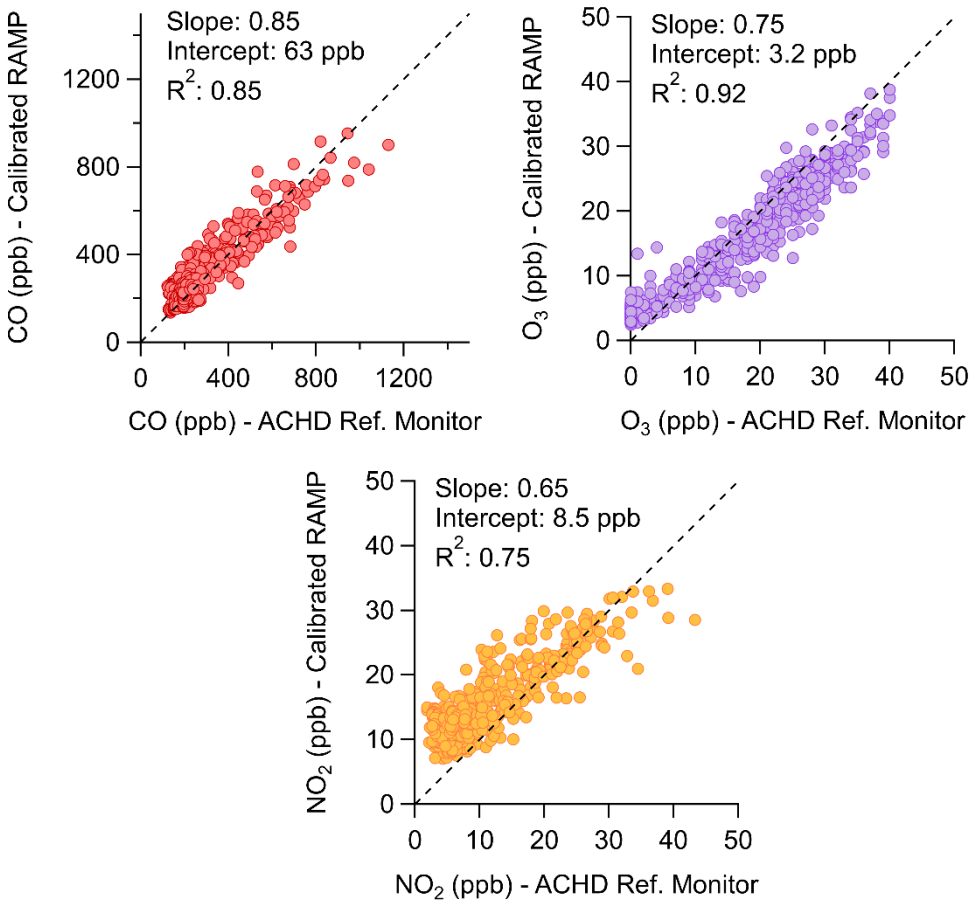

Figure 12: Comparison of CO, NO_2 and O_3 hourly average concentrations measured by a co-located RAMP monitor and the reference monitors at the Allegheny County Health Department (ACHD). The RAMP monitor was first calibrated on the Carnegie Mellon campus prior to deployment.

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.

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 12 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.**"

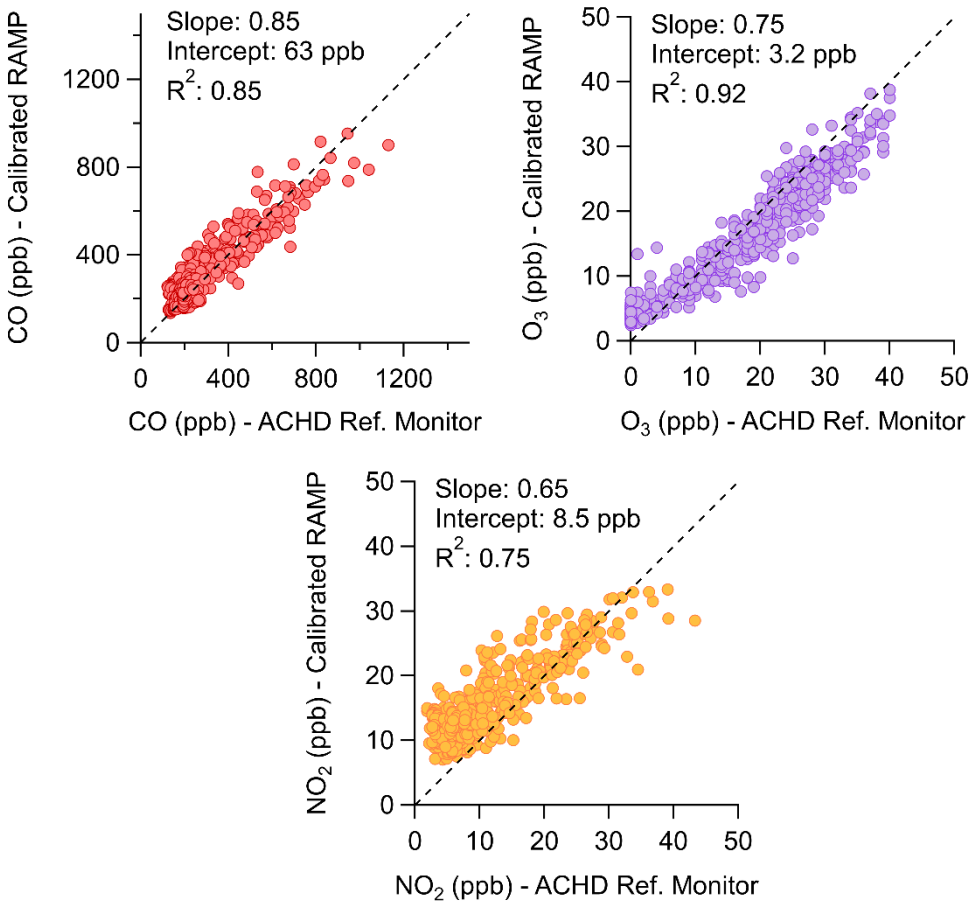

Figure 12: Comparison of CO, NO_2 and O_3 hourly average concentrations measured by a co-located RAMP monitor and the reference monitors at the Allegheny County Health Department (ACHD). The RAMP monitor was first calibrated on the Carnegie Mellon campus prior to deployment.

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:

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

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

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**)."

Response to Comment from Aerodyne Research Inc. (Eben Cross et al.)

1) Scope of work completed:
   The manuscript strongly emphasizes the unprecedented scale/scope of the completed work, stating that 19 RAMP systems were deployed for 6 months. At face value this would constitute a ~ 24wk interval across which to train & test the model. The actual reported tests appear more selective (both in terms of the number of RAMPS and duration of testing interval). As written, this is somewhat misleading. The authors should make an effort to more clearly state the scope of work as it pertains the results presented in the paper.

- Pulling data reported in table 3:
- CO: Test data spanned as few as 10 days, up to 108 days with an average of less than 6 weeks. Figure S7 shows only 16 of 19 RAMPS for evaluation (despite fact that 19 systems were RF-trained)
- NO2: Test data spanned as few as 2 days up 56 days with an average of 3.4 weeks. Figure S8 shows only 10 of 19 RAMPS were evaluated (despite fact that 19 systems were RF-trained)
- O3: Test data spanned 11-103 days (average less than 6 weeks) with 16 out of 19 system evaluated
- CO2 15 out of 19 systems evaluated and the number of days of test data were not tabulated.
- What is the fraction of training-to-test data for each RAMP system for which statistical metrics were reported?
- Data displayed for RAMP #4 in Figure 8 shows 15 weeks of test data. From the average number of test sample days reported in Table 3, is RAMP #4 a significant outlier? Did the majority of other RAMP systems run for shorter periods of time?

Response: We agree that the manuscript could have been more direct in terms of actual training and testing windows – the reason that longer collocations were not possible was due to deploying the RAMP monitors intermittently as part of air quality research campaigns in Pittsburgh, PA. While RAMP #4 did run the longest (was permanently located at the Carnegie Mellon supersite), there were 5 other RAMP monitors for which the testing period was 8-10 weeks. Additionally, we used RAMP #4 to systematically assess the performance of the testing data in Figure 8, and did not observe any significant relationship between weeks of testing data and error metrics up to 15 weeks. We only tested the models if there were at least 48 hrs of collocation data left after training; for 3 of the RAMPs, there was not enough data to properly test the model for all pollutants since it was deployed in the field. For an additional three RAMPs, there was not enough data to test the NO2 model due to the reference monitor being offline and needing repair from the manufacturer. We have added additional language throughout the manuscript to more specifically address the specific training and testing windows.

Modified text in Introduction:

"To ensure calibration model robustness, they were developed for **16-19 RAMP monitors and validated for 10-16 RAMP monitors (depending on pollutant)**, with each monitor containing one sensor per species (CO, CO_2, NO_2, SO_2 and O_3). Furthermore, the study was conducted over a six-month period (August 2016 – February 2017) spanning multiple seasons and a wide range of meteorological conditions. **During this period, RAMP monitors were intermittently deployed for air quality monitoring campaigns, resulting in collocation periods ranging from 5.5 to 16 weeks (median 9 weeks).**"

Additional text in Section 2.2 (new text in bold):

"The experiments involved 95 individual pollutant sensors mounted in 19 unique RAMP monitors. **While the collocation period spanned August 2016-February 2017, many sensors were intermittently deployed for air quality campaigns in Pittsburgh, so the collocation period ranged from 30 days to the study 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 reference monitors.**"

Modified text in Section 4.1 (additional text in bold):

Regression plots for 19 RAMP monitors **for CO, CO_2 and O_3 and 16 RAMP monitors for NO_2** illustrating the goodness of fit of the RF model are provided in the Supplemental Information **(Figures S3-S6). Only 16 of the 19 RAMP monitors had an NO_2 calibration, since the NO_2 monitor malfunctioned during the period when three RAMPs were collocated and so a calibration model could not be built for NO_2 for these three RAMPs. The NO_2 malfunction occurred between late September and early October, which did not significantly impact the range of conditions across the study.**

Modified text in Section 4.2 (additional text in bold):

"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.**"'

2) While the authors point out that the limited NO2 training/test data was due to a malfunction in their reference monitor at the co-location site, that does not explain why only 10 out of the 19 RAMP systems which were trained with the ambient RF model were included in the presented results.
    a. The authors should comment on the impact of the significantly shorter evaluation period on the NO2 results. Specifically, did the loss of the NO2 reference monitor exclude data sampled over the colder or warmer seasons in Pittsburgh and if so, how would this impact the range of conditions across which the RF model was found to be robust?

Response: As noted in the response to the previous comment, only 10 RAMP monitors were tested due to insufficient data available (i.e., as soon as those models had data to train the model, they were deployed for air quality monitoring). We have been more explicit about this in the text

(see response to previous comment). Additionally, the NO_2 monitor experienced issues in late September-early October, which did not affect the range of NO_2 sensors for training.

Modified text in Section 4.1 (additional text in bold):

"Regression plots for 19 RAMP monitors **for CO, CO_2 and O_3 and 16 RAMP monitors for NO_2** illustrating the goodness of fit of the RF model are provided in the Supplemental Information **(Figures S3-S6)**. **Only 16 of the 19 RAMP monitors had an NO_2 calibration, since the NO_2 monitor malfunctioned during the period when three RAMPs were collocated and so a calibration model could not be built for NO_2 for these three RAMPs. The NO_2 malfunction occurred between late September and early October, which did not significantly impact the range of conditions across the study.**"

3) Laboratory calibrations

As the authors' correctly point out, laboratory calibrations have formed the basis for much of the low-cost AQ sensor characterization work completed to-date. The manner in which the laboratory calibration experiments were executed in the current work raises a number of concerns:

- The authors should justify their laboratory calibration approach, specifically, sampling the sensors under 9 LPM of active flow, under air compositions dominated by (presumably) clean air, doped with single species of interest (excluding O3) under RH conditions that are outside of the specified operating range of the electrochemical sensors being trained. Given that these sensors operate under diffusion limited conditions, active vs passive flow can have a significant effect on the rate with which analyte molecules reach the working electrode surface of each electrochemical sensor. From the picture of the RAMP node, it appears that when fully integrated, the sensors are positioned to sample the air passively. This disconnect between the LAB cal. conditions and the ambient sampling configuration should be addressed if the authors are honestly trying to assess the validity of the LAB model on reconciling ambient concentrations from deployed RAMP monitors.

  Response: The design of the sampling manifold was such that the face velocity at the sensor surface would be 1.2 m/s, which is in lower end of wind speed range in Pittsburgh (e.g. average monthly windspeed from Jan-May 2017 was 2.4-3.4 m/s). The gas flow rate for the calibration system was based on the required flow rate for the reference instruments, the need to avoid leaks of ambient air into the system, and to minimize calibration gas consumption. Additionally, each data point was taken after 20 min when gas concentrations had stabilized as seen in the steady gas sensor output voltage. We have added these details to the manuscript, and changed the terminology from flow rate to face velocity for clarity.

Modified text in Section 3.1 (additions in bold):
"The sensors were exposed to each step in the calibration window (Table 1) for 20 minutes and a **face velocity of 1.2 m/s** flowed perpendicular to the sensor surface. **This face velocity is in the lower end of the wind speed range in Pittsburgh, PA (e.g. average monthly windspeed over Jan-May 2017 at 2m height is estimated at 2.4-3.4 m/s).**"

- The lack of any systematic logging or control of temperature and RH under these laboratory conditions limits the overall usefulness (and relevance) of the laboratory calibration to reconciling ambient concentrations. While the LAB model is limited in its sophistication, the execution of the lab experiments themselves also presents environmental conditions that do not overlap with their ambient co-location conditions. This apparent disconnect between the LAB and field needs to be explained further.
  Response: We agree; however, our laboratory calibration was limited by the available infrastructure at the time of the study. The goal of the laboratory calibration was to quantify the correlation between analyte response and calibration concentrations. While the utility of the calibration is limited, it was also useful to know that the CO calibration performed well even with a simple linear model. We have added text to Section 3.1 to emphasize that the laboratory calibration could be improved and better performance is in theory possible.

  Modified text in Section 3.1 (additions in bold):
  "Model performance was evaluated by comparing the calibrated response to reference measurements. We refer to the laboratory univariate linear regression calibration as LAB. Separate LAB calibrations were developed for each sensor (95 individual calibrations). **Due to difficulty controlling temperature and RH over a wide range of known ambient conditions, we focused on the relationship between analyte response and the calibration gas concentration, which any user with access to basic lab infrastructure can do. While beyond the scope of this study, an improved LAB calibration would involve a chamber with variable T and RH to better match ambient conditions.**"

- The absence of any O3 lab calibrations needs to be explained further. Why was this species excluded and given the RF model assessment of the Ox-B431 sensor sensitivities to different parameters, do the authors think this sensor type would provide more reasonable LAB-based calibration models, if such experiments had been conducted?
  Response: We did not conduct a LAB calibration for ozone due to our lack of a controlled low-concentration ozone generator and we did not find suitable ozone

calibration gas. We cannot comment on the outcome of a LAB based calibration for O3 as no such experiments were possible. From the RF model, RH and T seemed to have minimal impact on the ozone calibration, but this would require further investigation.

Modified text in Section 3.1 (additions in bold):
"Laboratory calibrations for O_3 were not performed **due to lack of a suitable ozone calibration gas**."

4) RF Model

With access to 1s reference monitor data it is not clear why the authors chose to use 15 min averages to train and test their RF model. Were shorter or longer time-averages tested and found to be measurably worse than the 15-min averages? What are the implications of using 15-min average data vs 1 or 5-min average data when resolving heterogeneity in local pollution gradients?

Response: The raw RAMP monitor data is reported at 15 second intervals, and down-averaged to 15 minutes for two primary reasons: 1) the goal of the RAMP monitor deployment in Pittsburgh is to quantify long term spatial and temporal variability in air pollution for exposures and 2) to generate a manageable data set for when 50+ monitors are deployed in Pittsburgh.

Modified text in Section 2.2 (additions in bold):
"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 and increase computational efficiency. Regulatory bodies typically make their data available at hourly resolution.**"

The authors should expand on their discussion regarding the lack of any extrapolation in the RF model.

- (related) Figure 5. For RAMPS #9,12,13,18 the authors should explain the straight vertical and horizontal at the ~ (50,50) x,y position on each scatter plot.

Response: This point was made by other reviewers and extensive changes have been made to emphasize this fact (see below). We have also noted that the horizontal features in Figure 5 are a result of the model being unable to extrapolate.

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.**"

Modified text in Section 4.3.3 (additions in bold)

"Systematic underprediction at the highest concentrations was also observed and is likely a consequence of either sensor limitations or the training dataset used to fit the RF model. Unless the range of concentrations in the training data encompasses the range of concentrations during model testing, there will be underpredictions for concentrations in exceedance of the training range **due to the RF model's inability to extrapolate. This is also what causes the horizontal feature for some RAMP monitors at high O_3 concentrations in Figure 5, as the model will not predict beyond its training range."**

It would be informative if the authors could comment on the computational cost of running the model. Does this computational cost place constraints on the time averaging used to train the model in the first place?

Response: We have added a comment on the computational cost of the model. The reviewer is correct that increasing the time resolution would come at a significant computational cost, as each RAMP monitor takes approximately 45 minutes to train at 15 minute resolution, thus when building calibrations for up to 50 RAMP monitors (ultimate goal of the work), increase time resolution could be prohibitive computationally.

Additional text in Section 3.3:

"The computation time to train a complete RAMP monitor with five sensors was approximately 45 minutes. This was another motivating factor for 15 minute resolution data,

as building models at higher time resolutions would have significantly increased computational demand."

5) P13 discussion of explanatory variables

What do the authors mean by permuting? Replace with another dataset that's not related to the current dataset? A more thorough explanation of this process is warranted as this process appears critical to evaluating the importance of various interfering factors on each sensor type.

Response: The term permuting is a mathematical term which means that the signal is randomly shuffled, which is not the same as replacing with another dataset not related to the current data set. Both within the manuscript and within the figure caption for explanatory variable performance we describe permuting by saying "(i.e., randomly shuffled)" and thus we feel that the explanation as offered in the manuscript and the standard nature of this mathematical concept does not warrant an expanded discussion.

Figure 9. Why is CO2 more sensitive to CO than CO2?

Response: In periods of high humidity, the CO_2 sensor becomes saturated, as the NDIR CO_2 sensor is also sensitive to water. We hypothesize that when the CO_2 sensor saturates, the model must rely on other pollutant signals (e.g., CO) as a predictor of CO_2 concentration. Additionally, short term fluctuations of CO2 are likely from combustion sources which also emit CO. This is currently in the manuscript in Section 4.3.2, but we have also added a few words of clarifying text.

Text from Section 4.3.2 (additions in bold):

"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 **(i.e., becomes saturated in periods of high humidity)**, 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.

The authors state that SO2 concentrations were below detection limits for the duration of the ambient co-location study and therefore not discussed further in the manuscript. While it is true that the SO2 concentrations in Pittsburgh are very low, the extent to which the SO2-B4 sensor output informed the RF model is in fact statistically significant according to the data presented in Figure 9 which indicates that the MSE can change by ~ 20-40% when the SO2 sensor parameter (presumably differential voltage?) is permuted? A more robust assessment of the importance of

the SO2-B4 sensor data to the resulting RF model may be to exclude it altogether from the available input parameters used to train the model.

Response: The SO_2 RAMP sensor may in and of itself have useful cross-sensitivities that may assist model performance. This is likely why the RAMP SO_2 sensor contributes to model performance despite low ambient SO_2 concentrations, thus we elected to include it. We have added some text regarding this hypothesis to Section 4.3.2.

Modified text in Section 4.3.2 (additions in bold):
"**Interestingly, despite low SO_2 concentrations, there was some contribution from the RAMP SO_2 sensor. This may be due to cross-sensitivities within the SO_2 sensor itself, as the SO_2 sensor may respond to more than ambient SO_2, warranting future investigation**. However, in general the SO_2 sensor contributed the least to model performance, thus this sensor could be replaced with a more relevant sensor, such as NO, in future iterations of the RAMP monitor."

6) All goodness of fit discussions relative to Cross et al., 2017 need to be revised according to the results published in the final accepted version of that manuscript.
Response: We have updated our tables and mentions within the manuscript body to be correct and consistent with the final version of the manuscript. We have also removed any mention of the combined testing and training data and updated our comparisons to reflect the new numbers.

7) Additional comments
P11 L15: The figure caption does not indicate this…

   - Figure 4 shows the calibrated RAMP #1 output regressed against the reference monitor concentration for the entire testing period for all three calibration models (LAB, MLR, and RF).

Response: Thank you for pointing this out – the figure caption in the manuscript was from a previous iteration of the figure that did not have the regressions. We missed updating it to reflect the new version of the manuscript. The updated caption is below:

"Figure 4: Example time series and regressions comparing the reference monitor data (black) to statistically average RAMP (RAMP#1) using LAB model (green), multiple linear regression (MLR) model (blue) and random forest (RF) model (pink). The left panel shows only 48 hrs of time series data to illustrate approach; the full evaluations (Table 3) were performed with much larger testing datasets; example regressions from the full data set for RAMP #1 are shown in the right panel."

P12 L20: The text states that the MAE comparison is against the number of points, but Figure 9 displays this data versus the number of weeks, not number of points.

Response: Thank you for pointing this out. This is another example of text that was not fully updated after revising a figure. We have corrected the text to say number of weeks vs number of points. We apologize for the error.

First paragraph of section 2.2 is unnecessarily repetitive

Response: Thank you, upon re-reading the paragraph, we agree and have deleted the redundant sentence describing which sensors are in the RAMP monitor.

Modified text (removed sentence in bold and strikethrough):
"The study uses the Real-time Affordable Multi-Pollutant (RAMP) monitor, which was developed in a collaboration between Carnegie Mellon University and SenSevere. ~~**The RAMP monitor incorporates widely-used Alphasense electrochemical sensors to measure gaseous pollutants (CO, NO2, SO2 O3) and a non-dispersive infrared (NDIR) sensor to measure CO2. The latter sensor also includes modules to measure temperature and relative humidity.**~~ The RAMP uses the following commercially-available electrochemical sensors from Alphasense Ltd: carbon monoxide (CO, Alphasense ID: CO-B41), nitrogen dioxide (NO_2, Alphasense ID: NO2-B43F), sulfur dioxide (SO_2, Alphasense ID: SO2-B4), and total oxidants (O_x, Alphasense ID: Ox-B431). The unit also includes a nondispersive infrared (NDIR) CO_2 sensor (SST CO2S-A) which contains built-in T (method: bandgap) and RH (method: capacitive) measurement."

95 sensor measurements (should be 76).
Response: As part of our response to your first comment, we have revised the text to be more clear on sensor count:

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, many sensors were intermittently deployed for air quality campaigns in Pittsburgh, so the collocation period ranged from 30 days to the full study 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 reference monitors."**

P7 L6 'beta4' should be 'beta3' according to the formula above
Response: Thank you for pointing out this typographical error, it has been corrected to β_3.

P9 L20 missing 'resolution' following 'temporal'

Response: Thank you for pointing out this typographical error, it has been corrected to "a higher temporal resolution"

P11 L13 Figure 2 should read Figure 3

Response: Thank you for pointing out this typographical error, it has been corrected to refer to Figure 3 in the text.

P18 L7 missing 'this' - as written: 'demonstrate that degree'

Response: Thank you for pointing out this typographical error, we have added the word "this" in the sentence in the Conclusions section.

P20 L30 Levy 2014 reference is the same as Moltchanov et al., 2015 reference.

Response: We apologize for the error, it seems as though we had an old version of the reference that was incorrectly imported into our reference software. The references have been merged and we apologize for the mistake.

Figure 2 caption should specify units as 'a.u.' following >255.9

Response: Thank you for suggesting this, we have added units to the CO sensor signal in the Figure 2 caption.

Figure 4 (left) – why do the four different pollutant times series all have unique time periods? If environmental parameters impact the sensors differently (RH, T) then it would be important to keep these parameters self-similar across the evaluation framework presented here (even though it's only 48-hours worth, should be the same 48 hours for all sensors).

Response: The intent was just as a visual snapshot of a period when there was variation in concentration – we also wanted a period when there was uninterrupted testing data. The ozone data is shifted by about two weeks as at that time the ozone reference monitor was offline. Given that we are not including T and RH in our plots, we do not feel it is essential to show the same time period, and it would not change the manuscript in any meaningful way.

Figure 7. It's not clear why there are ~ 10 or fewer data points displayed when data from 19 RAMPS are reportedly presented

Response: In Figure 7, there are 16 data points for CO and O_3, 15 data points for CO_2 and 10 data points for NO_2, as this is the number of RAMP monitor sesnors included as part of the testing data. We have more clearly addressed this within the manuscript and these changes are included earlier in our response to reviewers (your first comment). In Figure 7 it is occasionally difficult to see all 10-16 data points as there was significant overlap in many of the points (model performance was fairly consistent between RAMP monitors).

Figure 8 caption. 'long periods' is relative. Data displayed is for 15 weeks. Lifetime of the sensors is significantly longer than this (~100-150 weeks). Language should be revised accordingly.

- The extent to which the model improves over time should be quantified with 95% confidence intervals on the linear fits. By eye, it looks like this confidence interval would include 0.

Response: Both within the caption and in the manuscript body, the relative term "long periods" has been replaced with "the study period". Given the relatively new sensors we worked with, the focus of the study was not on temporal degradation which may be where the term "long" is more appropriate, as you mention. Thus, we have modified the manuscript to be more specific that the maximum period is 15 weeks. We have also calculated 95% confidence intervals for the slopes and they do include 0.

Modified text in Section 4.3.1 (additions in bold):

"For all the gas species, the MAE was essentially flat across the RAMP monitors **and the 95% confidence interval on the slope included 0**; RAMP monitors with more testing data did not have substantially higher (worse) MAE, suggesting the RF models are robust **over the study period**."

Table 3. Rather than identifying the number of days of sampling/evaluation – it would be more appropriate to identify the total number of data points used in each case study.

- Add an extra column that identifies the time resolution – as this is an important factor that drives signal-to-noise and accuracy and precision metrics as well as various end-use cases of interest.

Response: We agree that adding either time resolution or number of points would be helpful for others (but both would be redundant). As such we have added a column for time resolution, as we think this would be most helpful for others when considering what sort of model performance is achievable at a given time resolution.

Section 4.4. As written, this section oversimplifies the reality of the situation. When analyzing various lower-cost AQ sensor systems it is important to recognize that the combined hardware and software configuration impacts the performance metrics, not the software alone. The authors shouldn't gloss over this fact.

Response: This point was also mentioned by both Reviewer #1 and Reviewer #2 and we have made changes in the manuscript to address this.

Modified text in introduction (additions in bold):

"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)"

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.**"

~~Closing the gap on lower cost air quality monitoring:~~A machine learning calibration model~~s~~ to improve low-cost sensor performance

Naomi Zimmerman[1], Albert A. Presto[1], Sriniwasa P.N. Kumar[1], Jason Gu[2], Aliaksei Hauryliuk[1], Ellis S. Robinson[1], Allen L. Robinson[1], R. Subramanian[1]

[1]Center for Atmospheric Particle Studies, Carnegie Mellon University, Pittsburgh, 15213, USA
[2]Sensevere LLC, Pittsburgh, 15222, USA

Correspondence to: R. Subramanian (subu@cmu.edu)

Abstract. Low-cost sensing strategies hold the promise of denser air quality monitoring networks, which could significantly improve our understanding of personal air pollution exposure. Additionally, low-cost air quality sensors could be deployed to areas where limited monitoring exists. However, low-cost sensors are frequently sensitive to environmental conditions and pollutant cross-sensitivities, which have historically been poorly addressed by laboratory calibrations, limiting their utility for monitoring. In this study, we investigated different calibration models for the Real-time Affordable Multi-Pollutant (RAMP) sensor package, which measures CO, NO_2, O_3, and CO_2. We explored three methods: 1) laboratory univariate linear regression, 2) empirical ~~multivariate linear~~multiple linear regression and 3) machine-learning based calibration models using random forests (RF). Calibration models were developed for 16-19 RAMP monitors (varied by pollutant) using training and testing windows spanning August 2016 through February 2017 in Pittsburgh, PA. The random forest models matched (CO) or significantly outperformed (NO_2, CO_2, O_3) the other calibration models, and their accuracy and precision was robust over time for testing windows of up to 16 weeks. Following calibration, average mean absolute error on the testing dataset from the random forest models was 38 ppb for CO (14% relative error), 10 ppm for CO_2 (2% relative error), 3.5 ppb for NO_2 (29% relative error) and 3.4 ppb for O_3 (15% relative error), and Pearson r versus the reference monitors exceeded 0.8 for most units. Model performance is explored in detail, including a quantification of model variable importance, accuracy across different concentration ranges, and performance in a range of monitoring contexts including the National Ambient Air Quality Standards (NAAQS), and the US EPA Air Sensors Guidebook recommendations of minimum data quality for personal exposure measurement. A key strength of the RF approach is that it accounts for pollutant cross sensitivities. This highlights the importance of developing multipollutant sensor packages (as opposed to single pollutant monitors); we determined this is especially critical for NO_2 and CO_2. The evaluation reveals that only the RF-calibrated sensors meet the US EPA Air Sensors Guidebook recommendations of minimum data quality for personal exposure measurement. We also demonstrate that the RF model calibrated sensors could detect differences in NO_2 concentrations between a near-road site and a suburban site less than 1.5 km away. 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.

1 Introduction

Historically, spatial coverage of air quality monitoring stations has been limited by the high cost of instrumentation; urban areas typically rely on a few reference-grade monitors to assess population scale exposure. However, air pollutant concentrations often exhibit significant spatial variability depending on local sources and features of the built environment (Marshall et al., 2008; Nazelle et al., 2009; Pugh et al., 2012; Tan et al., 2014), which may not be well captured by the existing monitoring networks. In the past several years, there has been a significant increase in the development and applications of low-cost sensor-based air quality monitoring technology (Lewis and Edwards, 2016; McKercher et al., 2017; Moltchanov et al., 2015; Snyder et al., 2013). The use of low-cost air quality sensors for monitoring ambient air pollution could enable much denser air quality monitoring networks at a comparable cost to the existing regime. Increasing the spatial density of air quality monitoring would help quantify and characterize exposure gradients within urban areas and support better epidemiological models. Additionally, more highly resolved air quality information can assist regulators with future policy planning, with identification of hot spots or potential areas of concern (e.g., fracking in rural areas) where more detailed characterization is needed, and with risk mitigation for noncompliant zones. Furthermore, low-cost air quality sensors are generally characterized by their compact size and low power demand. These features enable low-cost sensors to be moved with relative ease to rural areas or developing regions where limited monitoring exists.

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 with~~The challenge with low-cost air quality sensor~~s is~~ calibration ~~at typical ambient pollutant concentrations and environmental conditions. These sensors are~~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). The most common example is for ozone electrochemical sensors, which also undergo redox reactions in the presence of NO_2. Additionally, NO has also been observed to interfere with NO_2, and CO sensors have exhibited some cross-sensitivity to molecular hydrogen in urban environments (Mead et al., 2013). Furthermore, low-cost sensors can be affected by meteorology (Masson et al., 2015b; Moltchanov et al., 2015; Pang et al., 2017; Williams et al., 2013). Most electrochemical sensors are configured such that the reactions are diffusion-limited, and the diffusion coefficient can be affected by temperature (Hitchman et al., 1997); Masson et al. (2015b) have shown that at relative humidity exceeding 75% there is significant error, possibly due to condensation on potentiostat electronics. Lastly, the stability of low-cost sensors is known to degrade over time (Jiao et al., 2016; Masson et al., 2015a). For example, in electrochemical cells, the reagents are consumed over time and have a typical lifetime of 1-2 years.

Deconvolving the effects of cross-sensitivity and stability on sensor performance is complex. Linear calibration models developed in the laboratory perform poorly on ambient data (Castell et al., 2017). Attempts to build calibration models from first principles have shown some success, but the models are difficult to construct and their transferability to new environments

remains unknown (Masson et al., 2015b). Accurate and precise calibration models are particularly critical to the success of dense sensor networks deployed in urban areas of developed countries where concentrations are on the low end of the spectrum of global pollutant concentrations, as poor signal-to-noise ratios and cross-sensitivities 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. 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 NO_x metal oxide sensors in Italy with encouraging results; in general mean relative error was approximately 30%. Cross et al. (2017) built high-dimensional multi-response calibration models for CO, NO, NO_2 and O_3 which had good agreement with reference monitors (slopes 0.476-0.946, R^2 0.3951-0.8896). Esposito et al. (2016) demonstrated excellent performance with dynamic neural network calibrations of NO_2 sensors (mean absolute error < 2 ppb); however, the same performance for O_3 was not observed. Furthermore, these calibrations have only been tested on a small number of sensor packages. For example, Cross et al. (2017) tested two sensor packages, each containing one sensor per pollutant over a four-month period, of which 35% was used as training data. Spinelle et al. (2015) tested a cluster of sensors in a single enclosure, testing 22 individual sensors in total over a period of 5 months, of which 15% was used as training data. Esposito et al. (2016) reported calibration performance on a single sensor package (5 gas sensors per package for measuring NO, NO_2 and O_3) and the model was tested on four weeks of data.

In this study, we aim to improve the calibration strategies of low-cost sensors using a random-forest-based machine learning algorithm, which, to our knowledge, has not been previously applied to low-cost air quality monitor calibrations. To ensure calibration model robustness, they were developed for 16-19 RAMP monitors and validated for ~~19 sensor packages~~10-16 RAMP monitors (depending on pollutant), with each ~~package~~ monitor containing one sensor per species (CO, CO_2, NO_2, SO_2 and O_3) ~~for a total of 95 individual sensors~~. Furthermore, the ~~model training and testing~~study was conducted over a six-month period (August 2016 – February 2017) spanning multiple seasons and a wide range of meteorological conditions. During this period, RAMP monitors were intermittently deployed for air quality monitoring campaigns, resulting in collocation periods ranging from 5.5 to 16 weeks (median 9 weeks)~~, providing one of the most comprehensive low cost air quality sensor calibration investigations to date~~. The fitting of the machine learning algorithms is discussed in detail to determine ideal calibration datasets to maximize performance and minimize overtraining. The performance of the random forest models is compared to traditional laboratory univariate linear models, multiple linear regression models, and EPA performance guidelines. The performance of a given model over time is also discussed.

2 Experimental methods

2.1 Measurement sites

Measurements were made from August 3, 2016 to February 7, 2017 on the Carnegie Mellon University campus in the Oakland neighbourhood of Pittsburgh, PA. ~~The measurement site~~The outdoor ambient testing environment (40°26'31.5"N, 79°56'33"W) is located within a small (< 100 vehicles) limited access, open air parking lot near the center of campus. It consisted of a mobile laboratory equipped with reference-grade instrumentation (Section 2.3) and adjacent lawn space where the RAMP monitors were mounted on tripods (Section 2.2). The dominant local source at the site is vehicle emissions when vehicles enter and exit the parking lot during the morning and evening rush hours. There was also occasional truck traffic and restaurant emissions from nearby on-campus restaurants. The small size of the parking lot (< 100 cars) and few other local sources means that for most of the day the location is essentially an urban background site. During the measurement period, the site mean (range) ambient temperature and relative humidity were 13°C (-15 to 34 °C) and 71% (27 to 98%), respectively.

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.

2.2 Real-time Affordable Multi-Pollutant (RAMP) monitor

The study uses the Real-time Affordable Multi-Pollutant (RAMP) monitor, which was developed in a collaboration between Carnegie Mellon University and SenSevere. ~~The RAMP monitor incorporates widely-used Alphasense electrochemical sensors to measure gaseous pollutants (CO, NO_2, SO_2, O_3) and a non-dispersive infrared (NDIR) sensor to measure CO_2. The latter sensor also includes modules to measure temperature and relative humidity. The RAMP is paired with a Met One Neighborhood PM monitor to measure optical $PM_{2.5}$.~~ The RAMP uses the following commercially-available electrochemical sensors from Alphasense Ltd: carbon monoxide (CO, Alphasense ID: CO-B41), nitrogen dioxide (NO_2, Alphasense ID: NO2-B43F), sulfur dioxide (SO_2, Alphasense ID: SO2-B4), and total oxidants (O_x, Alphasense ID: Ox-B431). The unit also includes a nondispersive infrared (NDIR) CO_2 sensor (SST CO2S-A) which contains built-in T (method: bandgap) and RH (method: capacitive) measurement. The experiments involved 95 individual pollutant sensors mounted in 19 unique RAMP monitors. While the collocation period spanned August 2016-February 2017, many sensors were intermittently deployed for air quality campaigns in Pittsburgh, so the collocation period ranged from 30 days to the full study 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 reference 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 and increase computational efficiency. Regulatory bodies typically make their data available at hourly resolution. The sensors sample passively from the bottom of the unit (Figure 1B), with screens installed to protect the sensors. ~~If operated with the $PM_{2.5}$ monitor, the RAMP monitors require 120-240V AC power; however,~~ Rroughly 3 weeks of measurements of gaseous species, T, and RH are possible on single charge of a built-in 30 amp-hour NiMH battery. The RAMP monitors are either mounted to a steel plate for easy pole mounting or are deployed on tripods approximately 1.5 m above the ground (Figure 1C). In this study, all the RAMP monitors were tripod-mounted at a consistent height.

In their simplest configuration, electrochemical sensors function based on a redox reaction within an electrochemical cell in which the target analyte oxidizes the anode and the cathode is proportionally reduced (or vice versa, depending on target analyte). The subsequent movement of charge between the electrodes produces a current which is proportional to the analyte reaction rate, which can be used to determine the analyte concentration. The Alphasense electrochemical sensors utilize a more complex configuration by using four electrodes (working, reference, counter and auxiliary) to account for zero current changes. Essentially, the auxiliary electrode, which is not exposed to the target analyte, accounts for ~~baseline~~ changes in the sensor baseline signal under different meteorological conditions. Additional details on the theory of operation for electrochemical sensors can be found in Mead et al. (2013).

The RAMP monitors log two output signals from each of the Alphasense sensors: one from the auxiliary electrode and the other from the working electrode. The net sensor response is determined by subtracting the auxiliary electrode signal from that of the working electrode. In theory, for a target analyte a linear relationship should exist between the net sensor signal for that analyte and ambient analyte concentrations, and this expectation forms the basis of univariate linear regression models built from laboratory calibrations. However, as noted in the introduction, even with an auxiliary electrode, electrochemical sensors may insufficiently account for the impacts of temperature (which affects the rate of diffusion) and relative humidity under high humidity conditions where condensation is possible. This has motivated researchers to construct ~~multivariate linear~~multiple linear regression models (MLR) to account for these temperature and humidity effects (Jiao et al., 2016). While these

calibration models typically improve performance relative to univariate linear models (Spinelle et al., 2015, 2017), they typically do not incorporate any cross-sensitivities to other pollutants or any non-linearities in the response. In this study, we attempt to build a calibration model for each analyte with no underlying assumptions regarding the calibration model structure and allow the models to consider directly the full suite of data being reported by the RAMP monitors using a machine learning approach.

2.3 Reference instrumentation

Reference measurements were made on ambient air continuously drawn through an inlet on the roof of the ~~supersite~~ mobile laboratory located approximately 2.5 m above ground. Gaseous pollutants were drawn through approximately 4 m of 0.953 cm outer diameter Teflon fluorinated ethylene propylene (FEP) tubing with a six-port stainless steel manifold for flow distribution to the gas analyzers. Measurements were made using direct absorbance at 405 nm for NO_2 (2B Technologies Model 405 nm), a gas filter correlation infrared analyzer for CO (Teledyne T300U), a non-dispersive infrared analyzer for CO_2 (LICOR 820), UV absorption for O_3 (Teledyne T400 Photometric Ozone Analyzer) and by UV fluorescence for SO_2 (Teledyne T100A UV Fluorescence SO_2 Analyzer). The time resolution for all reference measurements was 1 s.

The reference gas analyzers were checked and calibrated weekly using calibration gas mixtures, except for O_3 which is calibrated biannually at a nearby regulatory monitoring site. 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. The gas analyzers at the regulatory monitoring site are checked daily and thus this normalization helped correct for any baseline drift during the days between calibration. No significant drift was observed for CO_2 or O_3.

3 Calibration methods

Three calibration methods were evaluated: (1) a laboratory-based univariate linear regression based on net sensor response when exposed to calibration gases, (2): an empirical ~~multivariate linear~~multiple linear regression of net sensor response, T and RH regressed against reference monitor concentrations, and (3): a random forest machine learning model using net responses from all sensors, T, and RH to predict reference monitor concentrations. Calibration models were constructed for the CO, NO_2, CO_2 and O_3 sensors in each RAMP monitor. In this study, no calibration models were built for SO_2 due to a combination of reference instrument malfunction and SO_2 concentrations measured with the reference instrumentation being below the instrument detection limit (<0.4 ppbv) for most of the campaign (no nearby sources of SO_2). While lab calibrations were

conducted for the SO_2 sensors, this data will be the subject of a future publication on air quality in industrial areas where SO_2 is more commonly detected.

3.1 Laboratory-based univariate linear regression (LAB)

Prior to outdoor collocation, the sensors inside the RAMP monitors were calibrated in a laboratory environment using a custom manufactured sensor bed and calibration gas mixtures. The sensors were exposed to each step in the calibration window (Table 1) for 20 minutes and a ~~flow rate of 9 LPM~~face velocity of 1.2 m/s flowed perpendicular to the sensor surface. This face velocity is in the lower end of the wind speed range in Pittsburgh, PA (e.g. average monthly windspeed over Jan-May 2017 at 2m height is estimated at 2.4-3.4 m/s). ~~-~~The sensor response at each calibration step was averaged once the signal had stabilized (steady sensor output voltage). Temperature and relative humidity were not controlled during the calibration due to lack of available infrastructure at the time of the study. The temperature was at levels typical of indoor laboratory environments (approx. 20 °C), and the dry calibration gas provided very little humidity (RH <5%). Calibrations were built for CO, NO_2 and CO_2. Laboratory calibrations for O_3 were not performed due to a lack of suitable O_3 calibration gas.~~.~~

The laboratory calibration follows a standard univariate linear regression model of regression net (CO, NO_2) or raw (CO_2) signal against the reference gas concentration (Eq. 1)

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

Model performance was evaluated by comparing the calibrated response to reference measurements. We refer to the laboratory univariate linear regression calibration as LAB. Separate LAB calibrations were developed for each sensor (~~95~~37 individual calibrations, 9-14 per pollutant). Due to difficulty controlling temperature and RH over a wide range of known ambient conditions, we focused on the relationship between analyte response and the calibration gas concentration, which any user with access to basic lab infrastructure can do. While beyond the scope of this study, an improved LAB calibration would involve a chamber with variable T and RH to better match ambient conditions.

3.2 Empirical ~~multivariate linear~~multiple linear regression (MLR)

Following laboratory calibration, the individual sensors were mounted in the RAMP monitors and deployed outdoors adjacent to the Carnegie Mellon University supersite. The collocation period varied by RAMP, with a minimum collocation period of 6 weeks and a maximum collocation period of the entire 6-month study period. The collocation window varied due to intermittent deployment of some RAMP monitors for ongoing air quality monitoring campaigns in the Pittsburgh area. To build calibration models, the collocation period was separated into a training and testing period identical to that used for the random forest calibration (see Section 3.3). Due to the previously established influence of T and RH on sensor response (Jiao

et al., 2016; Masson et al., 2015b; Spinelle et al., 2015, 2017), a multiple linear regression (MLR) model was used to calibrate the output from each sensor using net sensor response to the target analyte (e.g. CO for the CO sensor), T and RH as explanatory variables (Eq. 2), similar to the approach described in a recent a European Union report on protocols for evaluating and calibrating low-cost sensors (Spinelle et al., 2013).

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

The training data was used to calculate the model coefficients (β_0 through β_{43}) and the model performance was evaluated on withheld testing data. Separate ~~multivariate linear~~multiple linear regression models were developed for each sensor (~~95~~73 individual models). We refer to these models as MLR.

3.3 Random forest model (RF)

A random forest (RF) model is a machine learning algorithm for solving regression or classification problems (Breiman, 2001). It works by constructing an ensemble of decision trees using a training data set; the mean value from that ensemble of decision trees is then used to predict the value for new input data. Briefly, to develop a random forest model, the user specifies the maximum number of trees that make up the forest, and each tree is constructed using a bootstrapped random sample from the training data set. The origin node of the decision tree is split into sub-nodes by considering a random subset of the possible explanatory variables (m_{try}). The training algorithm splits the tree based on which of the ~~random subsets of~~ explanatory variables in each random subset is the strongest predictor of the response. The number of random explanatory variables considered at each node (denoted m_{try}) is tuned by the user. This process of node splitting is repeated until a terminal node is reached; the user can specify the maximum number of sub-nodes or the minimum number of data points in the node as the indication to terminate the tree. For our random forest models, the terminal node was specified using a minimum node size of 5 data points per node.

To illustrate the method, consider building a random forest model for one RAMP monitor using a single decision tree and a subset of 100 training data points to build a CO calibration model (Figure 2). In this highly simplified example, at the first node, the net CO sensor signal is the strongest predictor of the CO reference monitor concentration, with a natural split in the data at a net CO sensor voltage of 255.9 a.u~~.~~. If sensor voltage exceeds 255.9 a.u., a cluster of 7 data points from the training data predicts an average CO concentration of 357 ppb, if CO net sensor voltage is ≤255.9 a.u. then the data goes to the next decision node, in which net CO sensor signal is again the strongest predictor of the CO reference monitor concentration, with a natural break in the data at a net CO sensor voltage of 167.3 a.u. The splitting proceeds until all the training data are assigned to a terminal node. The prediction value for each terminal node is the average reference monitor concentration of training points assigned to that node. To apply the algorithm (i.e., predict the CO concentration from a set of measured inputs), the user

takes the measured T and the net CO, NO_2 and O_3 signals and follows the path through the tree to the appropriate terminal node. The predicted CO concentration for that tree is then the average training value associated with that terminal node. This process is then repeated through multiple trees (Figure 2 shows only one simple tree) and the predictions from each tree are averaged to determine the final output from the entire random forest model. In this simple example, there are only six possible CO concentrations the random forest model will output. In practice, each tree has hundreds of terminal nodes and the forest typically comprises hundreds of trees, which means that there are thousands of possible answers. The model prediction for a given set of inputs is the average prediction across all the hundreds of trees that comprise the forest.

The random forest model's critical~~main~~ 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. To maximize utilization of the training data set to avoid missing any spikes during the training window, a k-fold cross validation approach was used. A k-fold cross-validation divides the data into k equal sized groups (where k is specified by the user) and k repeats are used to tune the model. Consider an example where k is equal to 5 (a 5-fold cross-validated random forest model). With a 5-fold validation, five unique random forest models are constructed, one for each fold. In building the first random forest, the first 20% (1/k) of the data will be the testing data, and the remaining 80% [(1-k)/k] of the data will be used as training. In building the second random forest, the next 20% of the data will be used as test data, and the first 20% and remaining 60% will be used to train. This is repeated until the data are fully covered, at which point the random forest model is created by combining the five (k) individual models into one large random forest model. This helps to minimize bias in training data selection when predicting new data, and ensures that every point in the training window is used to build the model.

In this study, reference gas data, RAMP net sensor data for CO, NO_2, SO_2, O_3, and RAMP raw sensor data for CO_2, T, and RH were collected at 15 second resolution, time-matched, and down-averaged to 15 min intervals (IGOR Pro v6.34), which is a higher temporal resolution than the 1 h intervals at which typical regulatory monitoring information are reported and minimized computational cost. The down-sampled data were then imported into R (ver. 3.3.3, "Another Canoe") for random forest model building. R is an open-source package for tuning and cross-validating many classes of statistical models, including random forest models. The cross-validated random forest models were compiled using the open-source "caret" package (Kuhn et al., 2017). The model considered all RAMP data (net voltage outputs from the five gas sensors plus T and RH, 7 possible variables total) as potential explanatory variables to predict the reference monitor gas concentration. 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. The computation time to train a complete RAMP monitor with five sensors was approximately 45 minutes. This was another motivating factor for 15 minute resolution data, as building models at higher time resolutions would have significantly increased computational demand.

When fitting the random forest models with the training data, the main tuning parameter is the number of explanatory variables to consider at each decision node (m_{try}). To determine the optimal m_{try}, the root mean square error (RMSE, equation in Supplemental Information) and the coefficient of determination (R^2) were calculated on the withheld folds of the training data (Figure 3, step 2) for m_{try} equal to 2, 4 or 7 to span the complete variable range. The random subset of explanatory variables considered at each node was chosen based on which value of m_{try} minimized RMSE. The cross-validation and the subset of explanatory variables randomly considered at each node (m_{try}) was tuned using the caret package in R (Kuhn et al., 2017). Following random forest model generation and tuning, the five 100 tree models were combined to create a final model with 500 trees. This process was repeated for each sensor to create ~~95~~ 73 separate random forest models. The final models convert the RAMP output signals into calibrated concentrations. The model conversion was done within R, where it exists as a standalone object compatible with the standard R configuration.

Data from three RAMP monitors (15 individual gas sensors) were used to investigate the optimal training period, which was determined by comparing the training data size to mean absolute error (MAE, the average of the absolute value of the residuals). The optimal training period was the period beyond which increases in the length of the training window (and therefore size of the training dateset) no longer resulted in significant reductions in the MAE. The initial training window evaluated was 1 week, and 1 week increments in training period duration were considered until MAE was minimized. The optimal collocation window was determined to be 4 weeks (or 2688 data points at 15-minute resolution). 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 ~~intermittently distributed~~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 ~~intermittent~~ 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 ~~degree of collocation~~ training data utilized in this study is equivalent to collocating the RAMP monitors with reference monitors for 3-4 days every 1-2 months. ~~However, if the MAE using the initial consecutive collocation is satisfactory for the application, this calibration strategy was not substantially less accurate than the distributed collocations.~~ 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.

3.4 Metrics for performance evaluation

The evaluation of the different models was conducted on 15-minute averaged testing data (i.e., data withheld entirely from model building). Metrics to quantitatively compare the LAB, MLR and RF model output to the reference monitor concentrations included Pearson r, which is a measure of the strength and direction of a linear relationship, and the coefficient of variation of the mean absolute error (CvMAE, Eq. 3). For comparing the RF model performance to other published studies, we also evaluated mean bias error, mean absolute error, slope of the linear regression of RF model calibrated RAMP data and reference data, and coefficient of determination (R^2).

$$\mathrm{CvMAE} = \frac{\mathrm{MAE}}{\mathrm{Avg.\ Reference\ Conc.}} = \frac{1}{\mathrm{Avg.\ Reference\ Conc.}} \times \left[\frac{1}{n}\sum_{i=1}^{n}|\mathrm{Model}_i - \mathrm{Reference}_i|\right], \quad (3)$$

Another useful tool for visually comparing competing models is a target diagram (Jolliff et al., 2009). A target diagram illustrates the contributions of the centered root mean square error (CRMSE, which is RMSE corrected for bias) and the mean bias error (MBE) towards total RMSE. In a target diagram, the x-axis is the CRMSE, the y-axis is the MBE and the vector distance to the origin is the RMSE. 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/$\sigma_{reference}$ (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, (3) whether the variance of the residuals is smaller than the variance of the reference measurements (inside circle of radius 1) or larger than the variance of the reference measurements (outside circle), and (4) the error (nRMSE), the vector distance between the coordinate and the origin. Details of equations required to build a target diagram are provided in the Supplemental Information. Model performance metrics were calculated in R (ver. 3.3.3, "Another Canoe") using the "tdr" package (Perpinan Lamigueiro, 2015).

4 Results and Discussion

4.1 Calibration model goodness of fit: comparing model predictions to training data

Following model building, the goodness of fit between the model output concentrations and the reference monitor concentrations during the training window (i.e. the data used to build the model) were evaluated for all three calibration model approaches (laboratory univariate linear regression "LAB", field-based multiple linear regression "MLR" and field-based random forest "RF"). For the training period, the calibrated CO and O_3 concentrations were all highly correlated (Pearson r > 0.8) with the reference monitor concentrations for all the calibration model approaches (Table 2). 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). Furthermore, CvMAE for each species was ≤5% during the training window for the RF models, substantially outperforming the other models.

Regression plots for ~~all~~ 19 RAMP monitors~~s~~ and for CO, CO_2 and O_3 and 16 RAMP monitors for NO_2 ~~and all four gas species~~ illustrating the goodness of fit of the RF model are provided in the Supplemental Information (Figures S3-S6). Only 16 of the 19 RAMP monitors had an NO_2 calibration, since the NO_2 monitor malfunctioned during the period when three RAMPs were collocated and so a calibration model could not be built for NO_2 for these three RAMPs. For the RF models, Table 2 also provides the random subset of explanatory variables sampled for splitting at each decision node (m_{try}) to achieve the lowest model RMSE. ~~In general, the larger the m_{try}, the simpler the underlying structure of the model. The advantage of a lower m_{try} is that subtle relationships between explanatory variables and the response can be probed. For example, if there is one dominant variable but the model is permitted to consider all 7 explanatory variables at each decision node, then the model will most frequently split the data based on the dominant variable, potentially masking the effect of other variables on the response. If the goodness of fit of the calibration model is improved by decreasing m_{try}, this suggests more complex variable interactions (Strobl et al., 2008).~~ In general, the larger the m_{try}, 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., m_{try}=7), then the model will most frequently split the data based on the dominant variable. By contrast, the advantage of a lower m_{try} is that subtle relationships between explanatory variables and the response can be probed. When randomly selecting fewer explanatory variables (m_{try}=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 m_{try}, this suggests more complex variable interactions with the response (Strobl et al., 2008).

Using the m_{try} metric, we observed that the underlying RF model structure is the simplest for CO, that some model explanatory variable complexities exist for the O_3 and NO_2 models, and that the CO_2 model is the most complex and relies on subtle

relationships between the explanatory variables to best fit the data (lowest m_{try} had the best results). This finding matches our expectations based on the LAB and MLR models; these simpler models performed best for CO and worst for CO_2. The trends in the m_{try} metric highlights the value of the RF model approach which directly accounts for multiple pollutants. This appears to be critical for O_3, NO_2 and CO_2 sensors because they are cross-sensitive to other pollutants. Cross-sensitivities have been shown to have a minimal impact on CO sensors, with the only notable cross-sensitivity being to molecular hydrogen (Mead et al., 2013). The poor performance of linear models at predicting CO_2 concentration is not surprising, as the sensor was observed to measure high concentrations under periods of high relative humidity (e.g., during rain) and in some cases during heavy rain will be saturated at 2000 ppm, the upper limit of the sensor, and then is reset to 400 ppm daily, as per manufacturer recommendations. The increase in CO_2 under high humidity conditions is likely due to the interference of water with CO_2 in the NDIR signal. Linear models are poorly suited to describe this behaviour.

4.2 Evaluation of models using testing data

To test the performance of the three different calibration models, the models were applied to the testing data that were not used for model fitting. The RAMP monitor concentrations after correction using the calibration models were compared to the actual measured reference concentrations (Figure 3~~2~~, step 5). To illustrate the approach, in Figure 4, we show a very short time-series of the testing data (~48-hour window) for RAMP #1. This RAMP monitor's performance is representative of the average model performance across the ~~19~~ RAMP monitors and therefore illustrates the quality of an average model. Figure 4 also shows the calibrated RAMP #1 output regressed against the reference monitor concentration for the entire testing period for all three calibration models (LAB, MLR, and RF). For this period, the RF model ~~clearly~~ outperformed the LAB and MLR models for all pollutants except for CO. Differences between the different models were smallest for CO and O_3 and largest for CO_2 and NO_2; the LAB models essentially did not reproduce the reference concentrations for CO_2 and NO_2. To illustrate the consistency of the RF model calibrated RAMP monitors across the entire suite of monitors, regressions for all the RAMP monitors for O_3 are shown in Figure 5. Regression plots for all RAMP monitors across the other gases are provided in the Supplemental Information (Figures S7-S10).

To assess the overall model performance, two performance metrics (Pearson r and CvMAE) were calcul~~a~~ted 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. This aggregate assessment shows that the MLR and RF models are interchangable for CO, as both models achieved Pearson r >0.9 and CvMAE <15%. The LAB model achieved a similar Pearson r, but CvMAE doubled to ~30%. For CO_2, NO_2, and O_3, the RF model substantially outperforms the LAB and MLR calibration models on the testing data. On average, Pearson r exceeded 0.8 for the RF model for CO_2 and NO_2 versus < 0.6 for the LAB and MLR calibration models.

Furthermore, the RF model performance was more consistent across the RAMP monitors than the MLR and LAB models. For example, the Pearson r for ~~N~~O_{3}~~2~~ ranged from 0.92 to 0.95 for the RF models versus 0.74 to 0.89 for the MLR models. This means that essentially all the RF models for O_3 performed well versus only a subset of the MLR models. The consistency of the different models is indicated by the smaller range in the box plots of Figure 6.

To compare the LAB, MLR and RF models, target diagrams were constructed for the four gases using all three calibration models for each RAMP monitor (Figure 7). The target diagrams show that, on average, across the RAMP monitors the random sensor error (distance to origin) was smaller for RF models and the RF models showed the least RAMP-to-RAMP variability (less disperse). This contrasts with the MLR models, whose bias and extent of model standard deviation varied much more widely between RAMP monitors, especially for CO_2. For the LAB models, the error for CO_2 and NO_2 was approximately an order of magnitude larger than for the RF and MLR models and had to be plotted on a separate inset due to their poor performance. Across all gases, the RF models on average were biased towards predicting concentrations slightly lower than the reference (i.e., slight tendency to under-predict, $MBE/\sigma_{reference} < 0$). Thus, we conclude that the low CvMAE, high Pearson r correlations, lowest bias and lowest absolute error characteristics of the RF models for all four gases are significant improvements compared to conventional calibration approaches (LAB and MLR).

4.3 Detailed assessment of RF model performance

To investigate the performance of the RF models in greater detail, we assessed the effect of amount of testing data on model performance, the relative importance of the seven explanatory variables, the performance of the models across the different concentration ranges, and the number of data points needed in each concentration range to optimize the fit.

4.3.1 Drift over amount of testing data

The first assessment was of amount of testing data. 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). Again, all the data have 15 min temporal resolution. The ~~number of points~~amount of data used to test the model performance varied by RAMP monitor and by pollutant, as reference monitors were occasionally offline for maintenance and calibration, and some RAMP monitors were intermittently deployed for concurrent air quality monitoring campaigns in Pittsburgh. To assess the effect of ~~number of testing points~~testing window size on conclusions regarding RF model performance, we compared the MAE to the number of ~~points~~ weeks in the testing window (Figure 8). For all the gas species, the MAE was essentially flat across the RAMP monitors and the 95% confidence interval on the slope included 0; RAMP monitors with more testing data did not have substantially higher (worse) MAE, suggesting the RF models are robust ~~over time~~over the study period. For NO_2, the most data available for testing was approximately 8 weeks due to instrument maintenance and repair taking the NO_2 reference monitor offline for 6 weeks of the study. Figure 8 also shows MAE over time from one RAMP, RAMP #4, which remained at the Carnegie Mellon supersite for

the entirety of the six-month study. For RAMP #4, MAE was calculated for an increasing cumulative number of weeks forward in time, and again, MAE was consistent (and in some weeks improved) over time.

4.3.2 RF model explanatory variable importance

While RF models are non-parametric, some sense of the model structure can be gained by examining the relative importance of the explanatory variables. The importance of each variable was quantified by comparing the percent increase in mean square error (MSE) if the explanatory variable signal is permuted (i.e., randomly shuffled). If an explanatory variable strongly affects the model performance, permuting that variable results in a large increase in MSE. Conversely, if a variable is not a strong predictor of the response, then permuting the variable does not significantly increase the MSE. Figure 9 shows for each of the gases (CO, CO_2, NO_2 and O_3) the increase in MSE when the explanatory variables were permuted. For both CO and O_3, the signal from the sensor measuring the target analyte (CO or O_3) is the most important explanatory variable, as expected. For the O_3, the second most important variable was the NO_2 signal, an expected cross-sensitivity, as the ozone sensor measures total oxidants (O_3 + NO_2) (Spinelle et al., 2015).

The explanatory variable importance is more complex for CO_2 and NO_2. For CO_2, all variables are 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 the CO_2 signal when the measured CO_2 has reached full-scale (i.e. becomes saturated in periods of high humidity), 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. For the NO_2 model, RH was the most important explanatory variable followed by the NO_2 sensor signal, highlighting again the importance of including meteorological data within sensor packages. The NO_2 model was also more strongly affected by temperature than the other pollutants. We hypothesize that the sensitivity of the NO_2 sensor to ambient NO_2 is suppressed in Pittsburgh, which has low ambient NO_2 concentrations compared to other cities where these sensors have been evaluated (see Table 3). NO_2 is lowest when O_3 ~~ozone~~ is highest in the summer, and thus the NO_2 RF model effectively uses T and RH as indicators for seasonality when NO_2 is low and the sensor response is supressed. Furthermore, the relatively equal variable importance of several of the explanatory variables within a model suggests that a cluster of sensors measuring many different species is critically important to build robust calibration models. Interestingly, despite low SO_2 concentrations, there was some contribution from the RAMP SO_2 sensor. This may be due to cross-sensitivities within the SO_2 sensor itself, as the SO_2 sensor may respond to more than ambient SO_2, warranting future investigation. However, in general the SO_2 sensor contributed the least ~~The only sensor channel~~

~~that did not contribute significantly~~ to ~~any~~ model performance ~~was the SO_2 sensor~~, thus this sensor could be replaced with a more relevant sensor, such as NO, in future iterations of the RAMP monitor. These findings highlight the value of bundling sensors for measuring a suite of pollutants together, as the different sensors can capture (at least to some extent) cross-sensitivities to other pollutants and improve the model performance for other sensors.

4.3.~~3~~2 RF model performance as a function of ambient concentration

In Section 4.2, predicted concentrations were normalized to average reference monitor concentration to ~~compare~~ quantitatively compare differences between the ~~different~~ calibration models (CvMAE). To evaluate the RF model performance at different reference concentrations, the testing data were divided into deciles for which the median reference monitor concentration, the absolute residual, and the residual normalized to the reference monitor concentration were calculated (Figure 10). For all species, the RF models tended to overestimate at lower concentrations, and underestimate at the highest concentrations. For the CO RF model, the normalized residual is within 10% of the reference monitor concentration by the 20th percentile of the data (>100 ppb), and continues to improve until the 50th percentile when it plateaus at a normalized residual of about 5%. The US EPA requires a limit of detection of 100 ppb for CO instruments used for regulatory monitoring (United States Environmental Protection Agency, 2014), thus our performance meets that goal. In the top decile, the average absolute CO residual for the RF models approximately doubles but the relative error is still around 5%. However, the top decile spans the broadest concentration range due to the lognormal shape of the CO concentration distribution, and these points are difficult to capture in training data sets.

For the CO_2 RF model, agreement with the reference monitor data are within a few percent up to the 90th percentile, when agreement drops to within 5%. This is possibly due to the RF model actively supressing high CO_2 sensor signals, as the sensor is prone to reading erroneously high concentrations during rain events. Additionally, the top decile of the data spans a wide range of CO_2 concentrations due to the lognormal shape of the CO_2 distribution. As with CO, the NO_2 RF model agreement with the reference monitor plateaus around the 50th percentile mark; however, the NO_2 RF-model error exceeds 100% for the lowest decile (<5 ppb), suggesting an effective sensitivity of the sensor of 5 ppb. For the O_3 RF model, the effective sensitivity is also around 5 ppb; when the average reference monitor concentration increased from 5 ppb to 10 ppb (from first to second decile), the normalized residual decreased from over 100% to about ~~than~~ 20%. The US EPA limit of detection for federal regulatory monitors is 10 ppb for both NO_2 and O_3, suggesting that as with CO, the RF model performance is within 20% of regulatory standards (United States Environmental Protection Agency, 2014).

Systematic underprediction at the highest concentrations was also observed and is likely a consequence of ~~either sensor limitations or~~ the training dataset used to fit the RF model. Unless the range of concentrations in the training data encompasses the range of concentrations during model testing, there will be underpredictions for concentrations in exceedance of the training

range due to the RF model's inability to extrapolate. This is also what causes the horizontal feature for some RAMP monitors at high O_3 concentrations in Figure 5, as the model will not predict beyond its training range. Additionally, the performance of the RF model is sensitive to the number of data points at a given concentration and the model performance. 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.

To illustrate the impact of number of training data points on the RF model, we binned the data for the representative RAMP (RAMP #1) by concentration and the average concentration measured by the reference monitors was plotted against the average concentration from the calibrated RAMP (Figure 11). The uncertainty in the ~~random forest~~RF model was plotted as the standard deviation of the model solutions from the 500 trees and the bins were colour coded by the number of data points within each bin. Figure 11 illustrates that for every pollutant, agreement with the reference monitor and uncertainty in the model prediction was larger for concentration bins containing fewer than 10 data points. This disproportionately impacted the upper end of the pollutant distribution where fewer data points were collected due to the intermittent and variable nature of high pollutant episodes. This suggests that a minimum of 10 data points at a given concentration are needed to adequately train the RF model, which may inform future RF model building. At NO_2 concentrations below 5 ppb, deviations from the 1:1 line were also observed despite the training dataset containing more than 100 data points at these concentrations. As was concluded from Figure 10, 5 ppbv appears to be the sensitivity limit of these low-cost sensors for NO_2.

4.4 Comparison of results to other published studies

In this section, we compare the performance of our RF models to results from other recent studies including the EuNetAir project in Italy (Borrego et al., 2016) and EPA Community Air Sensor Network (CAIRSENSE) project (Jiao et al., 2016). Additionally, a handful of studies have tested the field performance of low-cost sensors both 'out of the box' with factory calibrations (Castell et al., 2017; Duvall et al., 2016), and after a machine-learning-based calibration (Cross et al., 2017; Esposito et al., 2016; Spinelle et al., 2015, 2017). ~~The number of sensors and length of deployment used here is generally greater than those previous studies.~~ We compare the performance of our RF models to these studies in Table 3. While several low-cost sensor calibration studies have investigated calibration models within laboratory environments (Masson et al., 2015a; Mead et al., 2013; Piedrahita et al., 2014; Williams et al., 2013), we have elected to limit our comparison to field data.

There was not a substantial difference in performance of the RF model calibrated vs. LAB calibrated RAMP for CO, and performance was best for this pollutant on the 'out-of-the-box' factory calibrated performance assessments in EuNetAir and CAIRSENSE, suggesting that rigorous calibration models may not be critical for CO. However, the RAMP CO RF model did provide improved performance (smallest MAE, 38 ppb) at lower average concentrations compared to the EuNetAir study. Similarly, the 'out-of-the-box' performance of the CO sensors tested as part of CAIRSENSE and by the 24 AQMesh sensors tested in Castell et al. (2017) was poorer than the RF model calibrated RAMP. Of those studies that used an advanced algorithm to calibrate the sensors (Cross et al., 2017; Spinelle et al., 2017), the CO RF model resulted in ~~greater than or equivalent~~the highest R^2 values and slightly lower slopes; the slope closest to 1 was reported by Cross et al. (2017).~~s. While the R^2 of the CO HDMR model of Cross et al. (2017) is highest, it is difficult to estimate its true predictive performance due to its statistical metrics being calculated over the whole collocation period of which 35% of the data were used for training. Therefore, it blends goodness of fit and predictions.~~

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 ~~Therefore,~~ advanced calibration models, such as those using machine learning, ~~are~~ may be critical ~~to~~ for accurate measurements of ambient NO_2. Furthermore, sensor performance was correlated with average ambient concentration; studies in areas with higher NO_2 concentrations had the best performance, consistent with our observations (Figure 10). For studies using advanced NO_2 sensor calibration models (Cross et al., 2017; Esposito et al., 2016; Spinelle et al., 2015), Esposito et al. (2016) had the best performance, with a MAE of < 2 ppb; however, this evaluation was done in a location with high NO_2 concentrations, 45 ppbv (Air Quality England, 2015), more than three times higher than the 12 ppbv in Pittsburgh. In addition, they only evaluated one sensor array so the robustness of the approach is unknown. In our study, the MAEs across the NO_2 RF model RAMPs ranged from 2.6-3.8 ppb, which is almost as good as Esposito et al. (2016), but at less than one third the ambient concentrations. The slope ~~and R^2~~ of the HDMR model for NO_2 of Cross et al. (2017) does exceed that of the RAMP RF ~~,~~model, but the R^2 and MAE values are similar between both studies. ~~but again their performance metrics appear to be calculated over the entire collocation period, which includes 35% training data.~~ Similarly, the annual average NO_2 concentrations in 2015 were 15 ppb at the Massachusetts regulatory site used as a reference in Cross et al. (2017) (Massachusetts Department of Environmental Protection, 2016), 3 ppb higher than the average concentration observed in our study. As shown in Figure 10, an increase of a few ppb of NO_2 can result in almost 100% reductions in relative residuals in our model, potentially explaining discrepancies in the slope between our study and Cross et al. (2017). ~~thus this effect is not surprising.~~ Furthermore, for identical factory calibrated sensors out of the box, such as the Cairclip and AQMesh, a 5 ppb increase in average NO_2 concentration results in R^2 values more than doubling. As such, the excellent performance of the RF model for NO_2 at average ambient concentrations of 12 ppbv shows promise.

For O_3, the RF model, the calibrated data from Spinelle et al., (2015), and the measurements from the Aeroqual SM50 (Jiao et al., 2016) performed the best. Good performance from the Aeroqual when measuring NO_2 has also been previously observed (Delgado-Saborit, 2012). However, the results were the most consistent across the RAMP monitors calibrated with RF models, with relative standard deviations of <20% across the ~~19~~ 16 RAMPs for all markers of statistical performance. This performance consistency also holds for the CO and NO_2 RF models. The O_3 RF models were built in Pittsburgh, PA, which has historically had issues with NAAQS ozone compliance, thus while our model was seemingly one of the most accurate and robust, some of this performance may be attributed to the higher ambient O_3 concentrations. From this comparison, we conclude that the RAMP monitor calibrated with a RF model is unique in that it is more accurate when considering the combined suite of pollutants (i.e., all pollutants were accurately measured), it is consistent between many units (<20% relative standard deviation in performance metrics across 10-16~~9~~ monitors), and is precise even at lower ambient concentrations.

4.5 RF model calibrated RAMP performance in a monitoring context

We further assess the RAMP monitor performance against ~~two~~ 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~~1~~) for NAAQS compliance, and 3~~2~~) 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 May 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 site 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 ~~by regulators when considering compliance to National Ambient Air Quality Standards (NAAQS)~~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~~2~~). 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 MLR, 0.82 for RF). 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.

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 13). 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 13, red trace) compared to the nearby urban residential site (Figure 13, 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 13, 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.

Air sensor performance goals by application area are also provided by t~~T~~he US EPA Air Sensor Guidebook (Williams et al., 2014) ~~provides air sensor performance goals by application area~~. 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 14). 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 153). 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 153, red trace) compared to the nearby urban residential site (Figure 153, 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 153, 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.

5 Conclusions

This study demonstrates that the RF model applied to the RAMP low-cost sensor package can accurately characterize air pollution concentrations at the low levels typical of many urban areas in the United States and Europe. The fractional error of the models at a 15-minute time resolution was <5% for CO_2, approximately 10-15% for CO and O_3 and approximately 30% for NO_2, corresponding to mean absolute errors of 10 ppm, 38 ppb, 3.4 ppb and 3.5 ppb, respectively. This performance meets the recommended precision and accuracy error metrics from the US EPA Air Sensor Guidebook for Personal Exposure (Tier IV) monitoring. We demonstrate that this degree of sensitivity allows quantification of intra-urban gradients. Furthermore, the calibration models were well-constrained across 10-169 RAMP units (all performance metrics <20% relative standard deviation), and showed minimal degradation over the duration of the collocation study (August 2016 – February 2017),

While the iteration of the RAMP used in this study was equipped with an SO_2 sensor, no calibration model was possible due to SO_2 concentrations at our supersite being below reference instrument detection limits. One feature of the RAMP monitor is that the sensors are modular and can be readily replaced. The assessment of explanatory variable importance combined with the sub-detection limit levels SO_2 during the study suggests that the RAMP monitor did not substantially benefit from the

presence of the SO_2 sensor in this urban background environment. Future iterations of the RAMP will be equipped with NO sensors, which may be more relevant in an urban context.

The RF-models described here were built on four weeks of training data equally distributed in 3.5 day periods throughout the entire collocation. This is nominally equivalent to 3-4 days of calibration every 2 months. As previously mentioned, the low-cost sensor modules within the RAMP monitors can be readily replaced, and as such, we recommend for a large urban deployment to prepare a set of sensors at a regulatory monitoring site and to exchange sensors as they malfunction or as calibration models drift. Since the completion of this study, the sensors have been deployed in Pittsburgh for over 4 months, and changes in the calibration models over longer periods of deployment (1 year or more) will be discussed in a future work. Additionally, the sensors were first opened in July 2016, and characterized over the first 7 months of exposure to ambient environments. During this period, no significant temporal drift or sensor degradation was observed, but longer observational studies are likely needed to characterize sensor decay and end-of-life.

The calibration models were developed in Pittsburgh, which had higher O_3 and lower NO_2 compared to several published field-based calibrations and measurements with low-cost sensors. Our results and those of other studies demonstrate that low-cost sensor performance generally increases with increasing ambient concentration, but despite this, the RF models for NO_2 had the second lowest mean absolute error (<4 ppbv) even at low NO_2 concentrations. The good performance of the RF models across all pollutants can likely be attributed to the ability of the RF models to account for pollutant and meteorological cross-sensitivities, highlighting the importance of building multipollutant sensor packages.

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. Developing highly resolved air pollutant maps will assist researchers, policymakers and communities in developing new policies or mitigation strategies to enhance human health. Going forward, a random forest calibrated RAMP network of up to 50 nodes will be deployed in Pittsburgh, PA. This robustly calibrated network will help support better epidemiological models, aid in policy planning, and identify areas where more assessment is needed.

Competing interests

Author J. Gu is the CEO of SenSevere, the developer and manufacturer of the RAMP hardware. The extent of J. Gu's involvement was solely in development management, and improvement of the hardware in the RAMP monitors, and not in data analysis. Authors N. Zimmerman and R. Subramanian may in the future act as consultants for SenSevere on low-cost sensor calibration. The data output from the SenSevere hardware in conjunction with the calibration algorithms presented in

this paper yields significantly more accurate measurements than previously reported, and are the subject of provisional patent application. The authors declare no other competing interests.

Acknowledgements

Funding for this study was provided by the Environmental Protection Agency (Assistance Agreement Nos. RD83587301 and 83628601), and the Heinz Endowment Fund (Grants E2375 and E3145). N. Zimmerman's funding was provided by the NSERC Postdoctoral Fellowship (PDF-487660-2016). The authors also wish to thank A. Ellis and J.S. Apte for helpful conversations, and M. Schurman Boehm for her assistance with the laboratory calibrations.

References

Air Quality England: Air Pollution Report, 1st January to 31st December 2016, Cambridge Parker Street (Site ID: CAM 1), , 1–4 [online] Available from: http://www.airqualityengland.co.uk/assets/downloads/airqualityengland-statistics-report-CAM1-2015.pdf, 2015.

Bart, M., Williams, D. E., Ainslie, B., McKendry, I., Salmond, J., Grange, S. K., Alavi-Shoshtari, M., Steyn, D. and Henshaw, G. S.: High density ozone monitoring using gas sensitive semi-conductor sensors in the lower Fraser valley, British Columbia, Environ. Sci. Technol., 48(7), 3970–3977, doi:10.1021/es404610t, 2014.

Borrego, C., Costa, A. M., Ginja, J., Amorim, M., Coutinho, M., Karatzas, K., Sioumis, T., Katsifarakis, N., Konstantinidis, K., De Vito, S., Esposito, E., Smith, P., Andre, N., Gerard, P., Francis, L. A., Castell, N., Schneider, P., Viana, M., Minguillon, M. C., Reimringer, W., Otjes, R. P., von Sicard, O., Pohle, R., Elen, B., Suriano, D., Pfister, V., Prato, M., Dipinto, S. and Penza, M.: Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise, Atmos. Environ., 147(2), 246–263, doi:10.1016/j.atmosenv.2016.09.050, 2016.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, 2001.

Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D. and Bartonova, A.: Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?, Environ. Int., 99, 293–302, doi:10.1016/j.envint.2016.12.007, 2017.

Cross, E. S., Lewis, D. K., Williams, L. R., Magoon, G. R., Kaminsky, M. L., Worsnop, D. R. and Jayne, J. T.: Use of electrochemical sensors for measurement of air pollution : correcting interference response and validating measurements, Atmos. Meas. Tech. Discuss., 2017–138, 1–17, doi:10.5194/amt-2017-138, 2017.

Delgado-Saborit, J. M.: Use of real-time sensors to characterise human exposures to combustion related pollutants, J. Environ. Monit., 14(7), 1824, doi:10.1039/c2em10996d, 2012.

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 Actuators, B Chem., 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 Actuators, B Chem., 143(1), 182–191, doi:10.1016/j.snb.2009.08.041, 2009.

Duvall, R. M., Long, R. W., Beaver, M. R., Kronmiller, K. G., Wheeler, M. L. and Szykman, J. J.: Performance evaluation and community application of low-cost sensors for ozone and nitrogen dioxide, Sensors (Switzerland), 16(10), 1–14, doi:10.3390/s16101698, 2016.

Esposito, E., De Vito, S., Salvato, M., Bright, V., Jones, R. L. and Popoola, O.: Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems, Sensors Actuators, B Chem., 231, 701–713, doi:10.1016/j.snb.2016.03.038, 2016.

Hitchman, M. L., Cade, N. J., Kim Gibbs, T. and Hedley, N. J. M.: Study of the Factors Affecting Mass Transport in Electrochemical Gas Sensors, Analyst, 122(November), 1411–1417, doi:10.1039/a703644b, 1997.

Jiao, W., Hagler, G., Williams, R., Sharpe, R., Brown, R., Garver, D., Judge, R., Caudill, M., Rickard, J., Davis, M., Weinstock, L., Zimmer-Dauphinee, S. and Buckley, K.: Community Air Sensor Network (CAIRSENSE) project: Evaluation of low-cost sensor performance in a suburban environment in the southeastern United States, Atmos. Meas. Tech., 9(11), 5281–5292, doi:10.5194/amt-9-5281-2016, 2016.

Jolliff, J. K., Kindle, J. C., Shulman, I., Penta, B., Friedrichs, M. A. M., Helber, R. and Arnone, R. A.: Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, J. Mar. Syst., 76(1–2), 64–82, doi:10.1016/j.jmarsys.2008.05.014, 2009.

Kuhn, M., 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, 2017.

Lewis, A. and Edwards, P.: Validate personal air-pollution sensors, Nature, 535, 29–31, 2016.

Marshall, J. D., Nethery, E. and Brauer, M.: Within-urban variability in ambient air pollution : Comparison of estimation methods, , 42, 1359–1369, doi:10.1016/j.atmosenv.2007.08.012, 2008.

Massachusetts Department of Environmental Protection: Massachusetts 2015 Air Quality Report. [online] Available from: http://www.mass.gov/eea/docs/dep/air/priorities/15aqrpt.pdf, 2016.

Masson, N., Piedrahita, R. and Hannigan, M.: Approach for quantification of metal oxide type semiconductor gas sensors used for ambient air quality monitoring, Sensors Actuators, B Chem., 208, 339–345, doi:10.1016/j.snb.2014.11.032, 2015a.

Masson, N., Piedrahita, R. and Hannigan, M.: Quantification method for electrolytic sensors in long-term monitoring of ambient air quality, Sensors (Switzerland), 15(10), 27283–27302, doi:10.3390/s151027283, 2015b.

McKercher, G. R., Salmond, J. A. and Vanos, J. K.: Characteristics and applications of small, portable gaseous air pollution monitors, Environ. Pollut., 223, 102–110, doi:10.1016/j.envpol.2016.12.045, 2017.

Mead, M. I., Popoola, O. A. M., Stewart, G. B., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J. J., McLeod, M. W., Hodgson,

T. F., Dicks, J., Lewis, A., Cohen, J., Baron, R., Saffell, J. R. and Jones, R. L.: The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks, Atmos. Environ., 70, 186–203, doi:10.1016/j.atmosenv.2012.11.060, 2013.

Moltchanov, S., Levy, I., Etzion, Y., Lerner, U., Broday, D. M. and Fishbain, B.: On the feasibility of measuring urban air pollution by wireless distributed sensor networks, Sci. Total Environ., 502, 537–547, doi:10.1016/j.scitotenv.2014.09.059, 2015.

Nazelle, A. De, Rodríguez, D. A. and Crawford-brown, D.: Science of the Total Environment The built environment and health : Impacts of pedestrian-friendly designs on air pollution exposure, Sci. Total Environ., 407(8), 2525–2535, doi:10.1016/j.scitotenv.2009.01.006, 2009.

Oshiro, T. M., Perez, P. S. and Baranauskas, J. A.: How Many Trees in a Random Forest?, in Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings, edited by P. Perner, pp. 154–168, Springer Berlin Heidelberg, Berlin, Heidelberg., 2012.

Pang, X., Shaw, M. D., Lewis, A. C., Carpenter, L. J. and Batchellier, T.: Electrochemical ozone sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air-quality monitoring, Sensors Actuators, B Chem., 240, 829–837, doi:10.1016/j.snb.2016.09.020, 2017.

Perpinan Lamigueiro, O.: tdr: Target Diagram, [online] Available from: https://cran.r-project.org/package=tdr, 2015.

Piedrahita, R., Xiang, Y., Masson, N., Ortega, J., Collier, A., Jiang, Y., Li, K., Dick, R. P., Lv, Q., Hannigan, M. and Shang, L.: The next generation of low-cost personal air quality sensors for quantitative exposure monitoring, Atmos. Meas. Tech., 7(10), 3325–3336, doi:10.5194/amt-7-3325-2014, 2014.

Pugh, T. A. M., Mackenzie, A. R., Whyatt, J. D. and Hewitt, C. N.: Effectiveness of Green Infrastructure for Improvement of Air Quality in Urban Street Canyons, Environ. Sci. Technol., 46(14), 7692–7699, 2012.

Snyder, E. G., Watkins, T. H., Solomon, P. A., Thoma, E. D., Williams, R. W., Hagler, G. S. W., Shelow, D., Hindin, D. A., Kilaru, V. J. and Preuss, P. W.: The changing paradigm of air pollution monitoring, Environ. Sci. Technol., 47(20), 11369–77, doi:10.1021/es4022602, 2013.

Spinelle, L., Aleixandre, M. and Gerboles, M.: Protocol of evaluation and calibration of low-cost gas sensors for the monitoring of air pollution, Eur. Comm. JRC Techical Reports, EUR 26112, doi:10.2788/9916, 2013.

Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M. and Bonavitacola, F.: Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide, Sensors Actuators, B Chem., 215, 249–257, doi:10.1016/j.snb.2015.03.031, 2015.

Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M. and Bonavitacola, F.: Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2, Sensors Actuators, B Chem., 238, 706–715, doi:10.1016/j.snb.2016.07.036, 2017.

Tan, Y., Lipsky, E. M., Saleh, R., Robinson, A. L. and Presto, A. A.: Characterizing the Spatial Variation of Air Pollutants and the Contributions of High Emitting Vehicles in Pittsburgh, PA, 2014.

United States Environmental Protection Agency: Appendix D, Measurement Quality Objectives and Validation Templates, in QA Handbook Volume II, pp. 5–12., 2014.

Williams, D. E., Henshaw, G. S., Bart, M., Laing, G., Wagner, J., Naisbitt, S. and Salmond, J. a: Validation of low-cost ozone measurement instruments suitable for use in an air-quality monitoring network, Meas. Sci. Technol., 24(6), 65803, doi:10.1088/0957-0233/24/6/065803, 2013.

Williams, R., Kilaru, V. J., Snyder, E. G., Kaufman, A., Dye, T., Ruttler, A., Russell, A. and Hafner, H.: Air Sensor Guidebook, EPA/600/R-14/159., 2014.

Zhou, Y. and Levy, J. I.: Factors influencing the spatial extent of mobile source air pollution impacts: a meta-analysis, BMC Public Health, 7(1), 89, doi:10.1186/1471-2458-7-89, 2007.

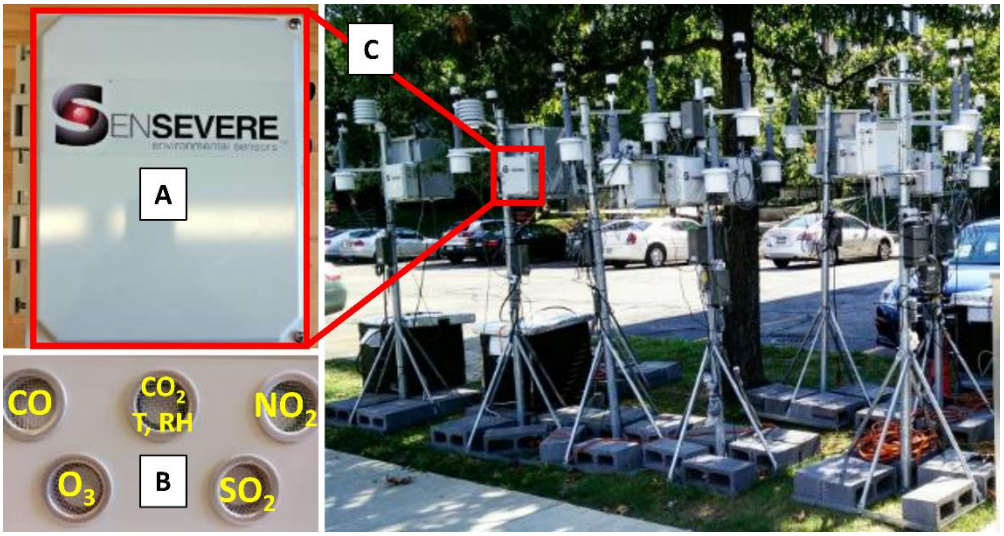

Figure 1: Photographs of the RAMP monitors and the sampling set up. (A) Front view of the RAMP unit in the NEMA-rated enclosure. (B) Bottom view of the RAMPs with sensor layout labelled in yellow. (C) Example of collocation set-up using tripod mounting (not pictured: supersite containing the reference monitors, immediately beside the tripods).

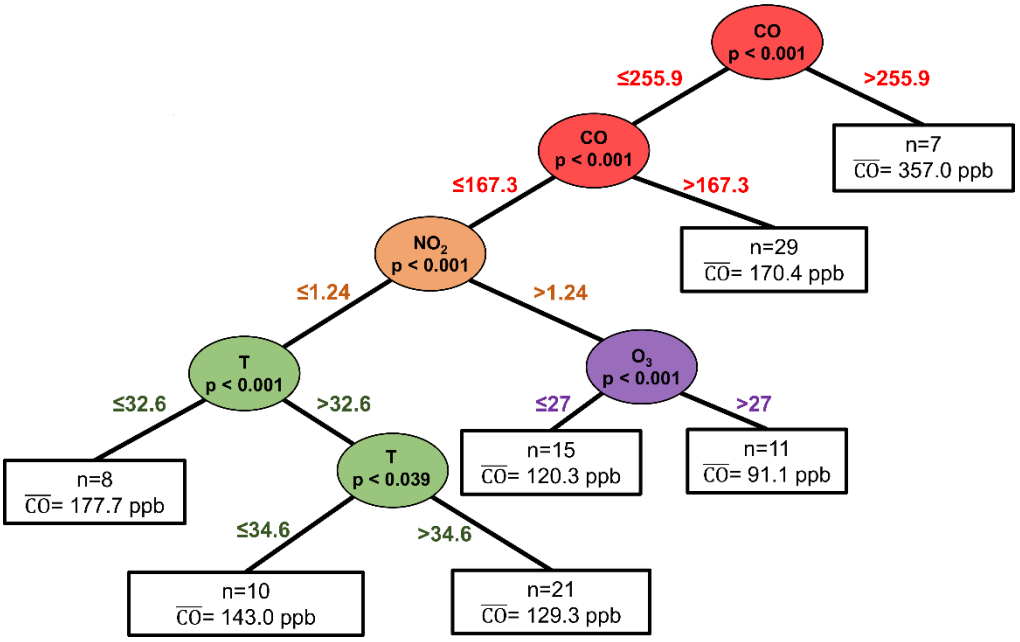

Figure 2: Simplified illustration of one potential CO random forest tree for one RAMP using 100 data points (the trees within the actual models are significantly more complex and 500 such trees are included in the final models). Tree nodes are coloured by splitting variable and split point is overlaid on the branch (e.g., at first split, points with CO sensor signal >255.9 a.u. are sent to a terminal node, the remaining points go to the next splitting node). $\overline{CO}$ is the average CO reference monitor concentration (ppb) in each terminal node; n = number of data points in each terminal node.

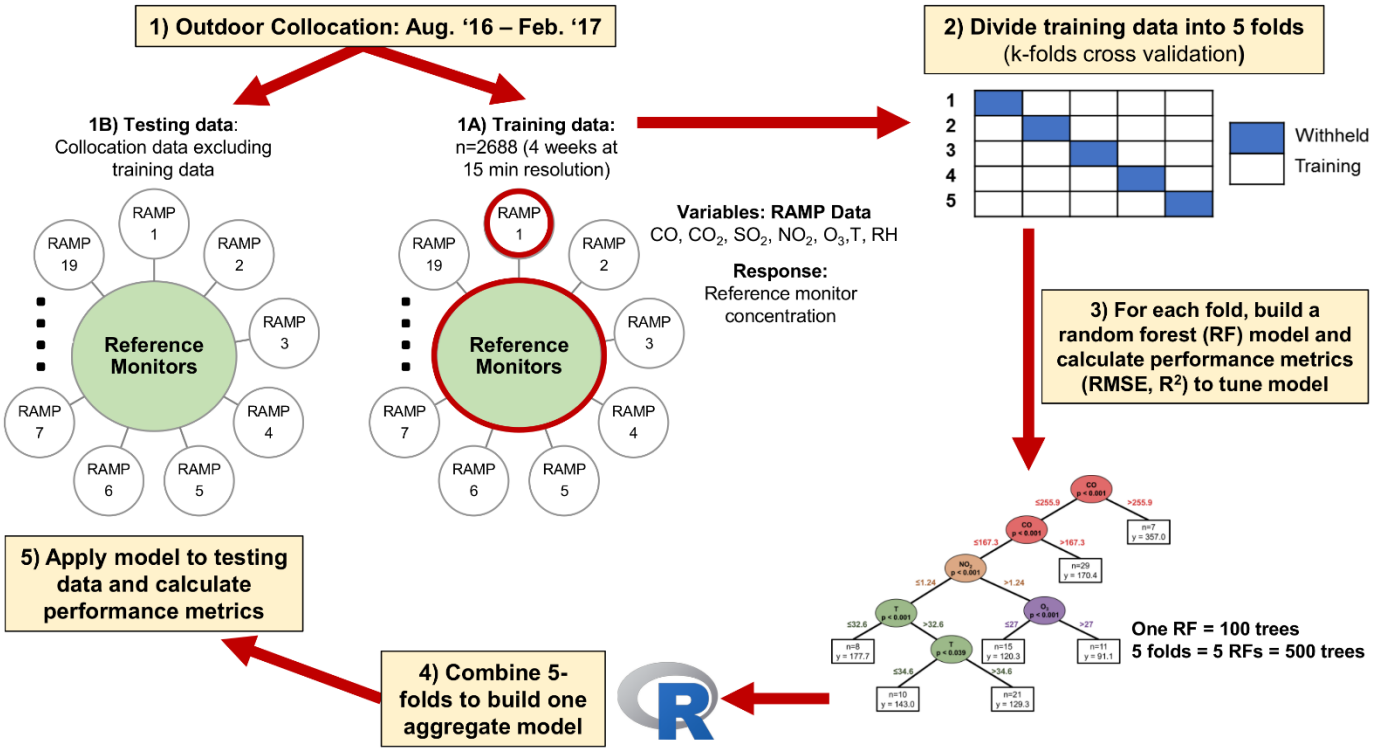

Figure 3: Flow path for data collection and RF model fitting and testing. From collocation period, 2688 points were sub-selected as training (1A) data while the remaining data were used for model testing (1B). The training data were further divided into 5 cross-validation folds and each fold was used to tune and build an RF model. All five models were then combined in R to build one cumulative model and the predictive power of the model was assessed for the withheld testing data.

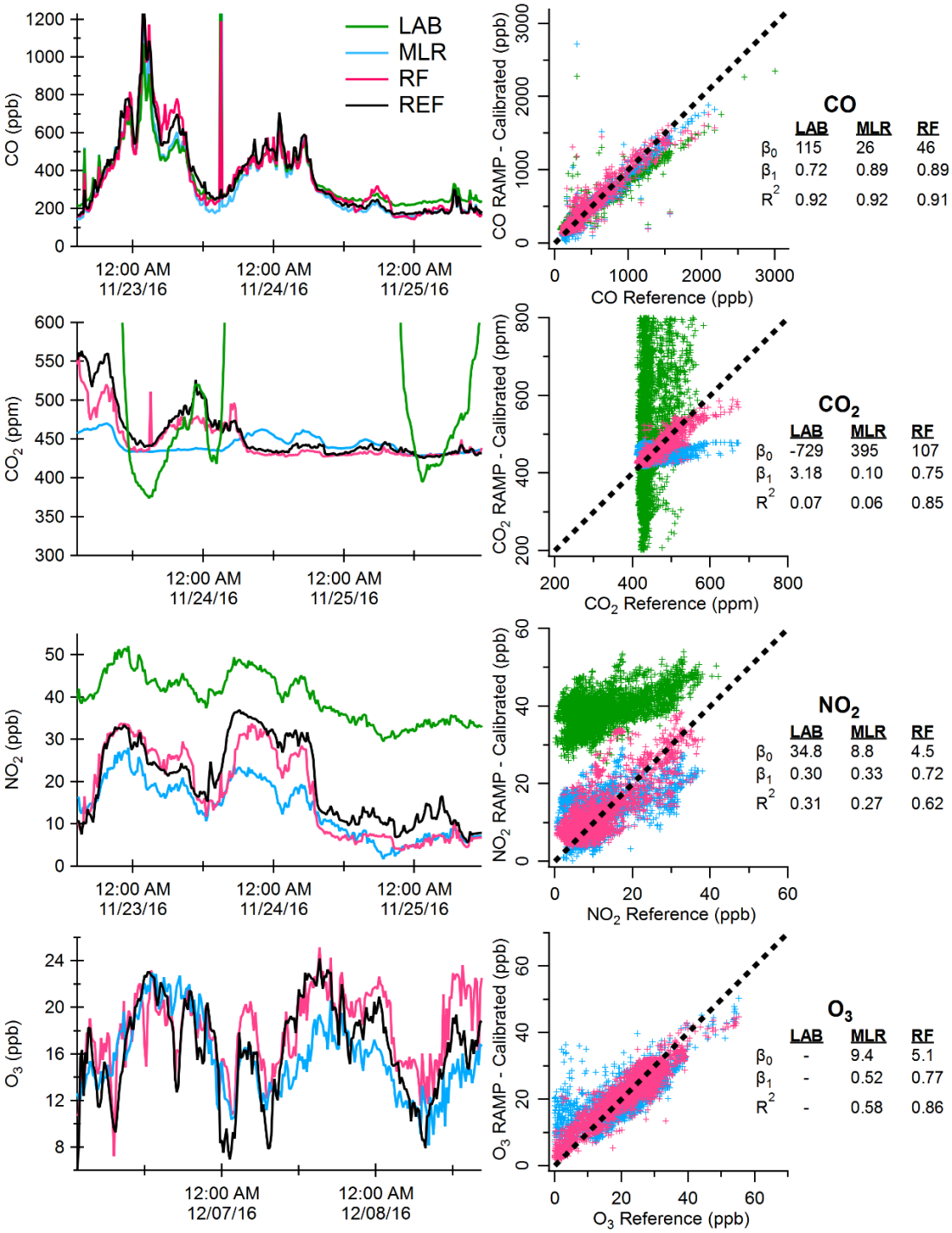

Figure 4: Example time series and regressions comparing the reference monitor data (black) to statistically average RAMP (RAMP#1) using LAB model (green), multiple linear regression (MLR) model (blue) and random forest (RF) model (pink). The left panel~~his example~~ shows only 48 hrs of time series data to illustrate approach; the full evaluations (Table 3) were performed with much larger testing datasets; example regressions from the full data set for RAMP #1 are shown in the right panel.~~.~~

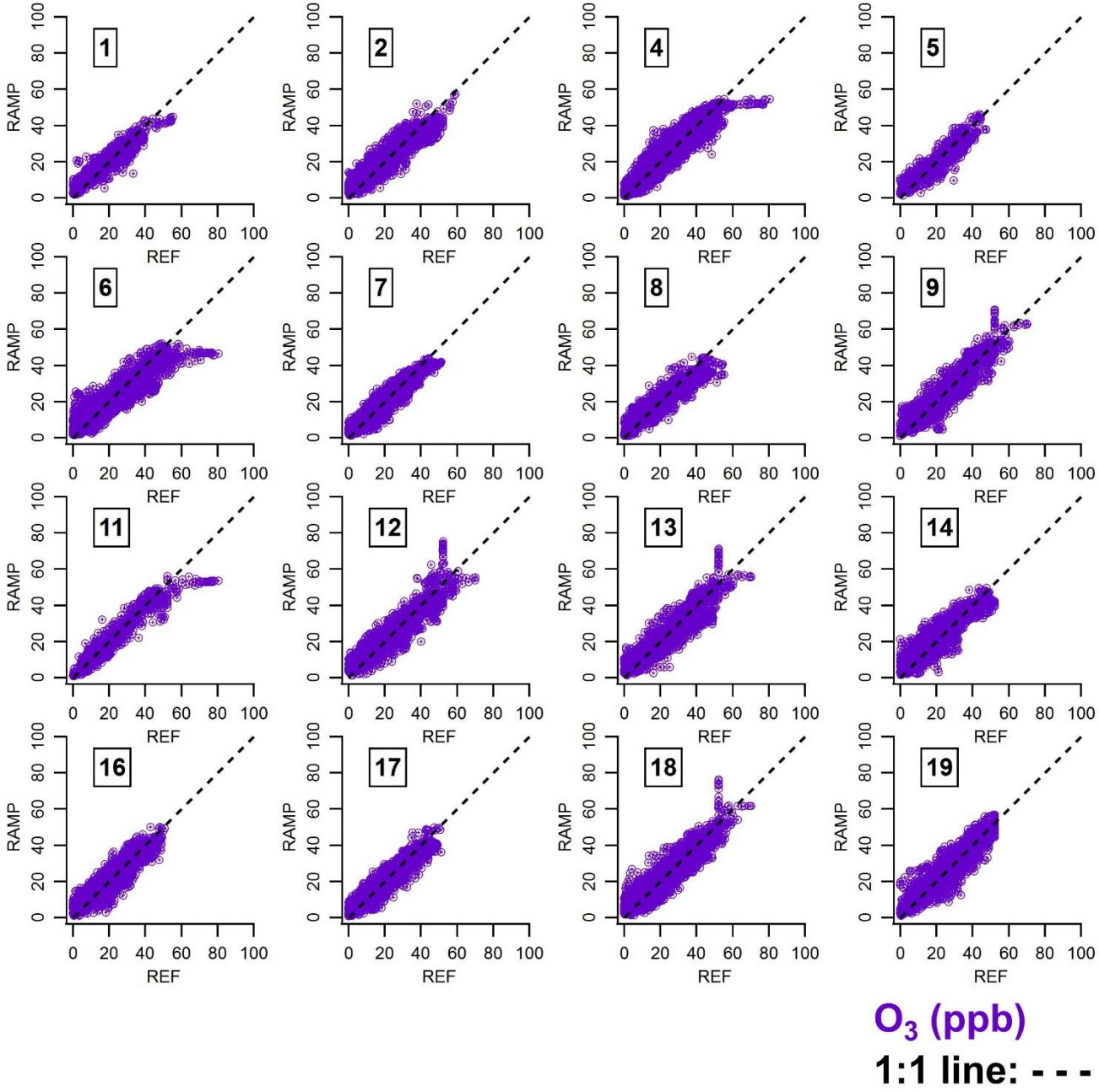

Figure 5: RF model performance for ozone evaluated using the testing data (data withheld from building model). Correlation plots show predicted ozone concentration ("RAMP") versus the reference monitor concentration ("REF") for 16 RAMP units. All values are in ppb, and the 1:1 line is drawn as a black dashed line.

5

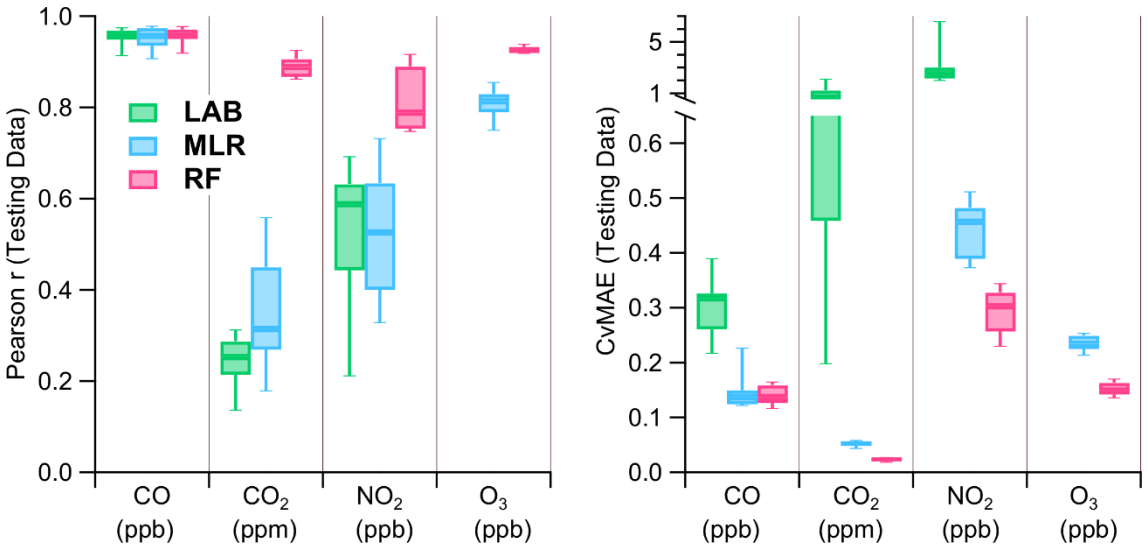

Figure 6: Performance of different calibration models against reference monitor testing data (data not included in model fitting). *Left:* Pearson r correlation coefficient (higher = better, maximum of 1) of different calibration models '(LAB', green; 'MLR', blue; 'RF', pink) versus reference monitor. *Right:* The CvMAE (coefficient of variation of the MAE; MAE normalized by average reference concentration, lower = better) for the three calibration methods. The box plots show the range across the 10-~~16~~9 RAMPs (whiskers: 10th and 90th percentile, box edges: 25th and 75th percentile).

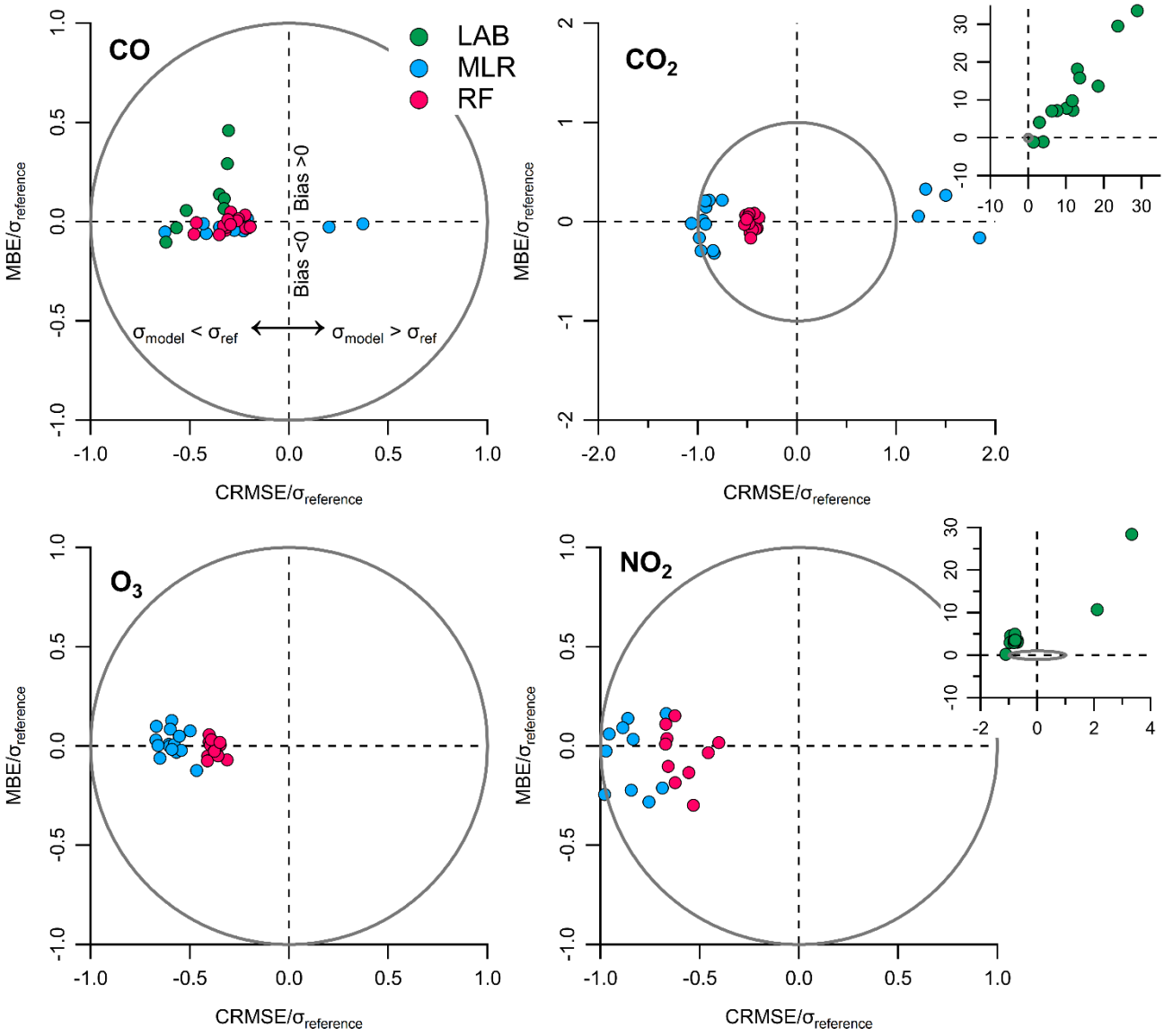

Figure 7: Target diagrams for CO, CO_2, NO_2 and O_3 to compare the LAB, MLR and RF model performance. The y-axis is the bias relative to the reference and the x-axis is the bias-adjusted RMSE (CRMSE) normalized by reference monitor standard deviation; the vector distance between any given point and the origin is the RMSE normalized by the standard deviation of the reference measurements. The CRMSE is in the left plane if model standard deviation is smaller than the standard deviation of the reference observations, and vice versa. If data falls within the circle, then the variance of the residuals is smaller than the variance of the reference measurements. The target diagram for the LAB model for CO_2 and NO_2 is shown in the inset figure because of the order of magnitude difference in MBE and CRMSE compared to the MLR and RF models.

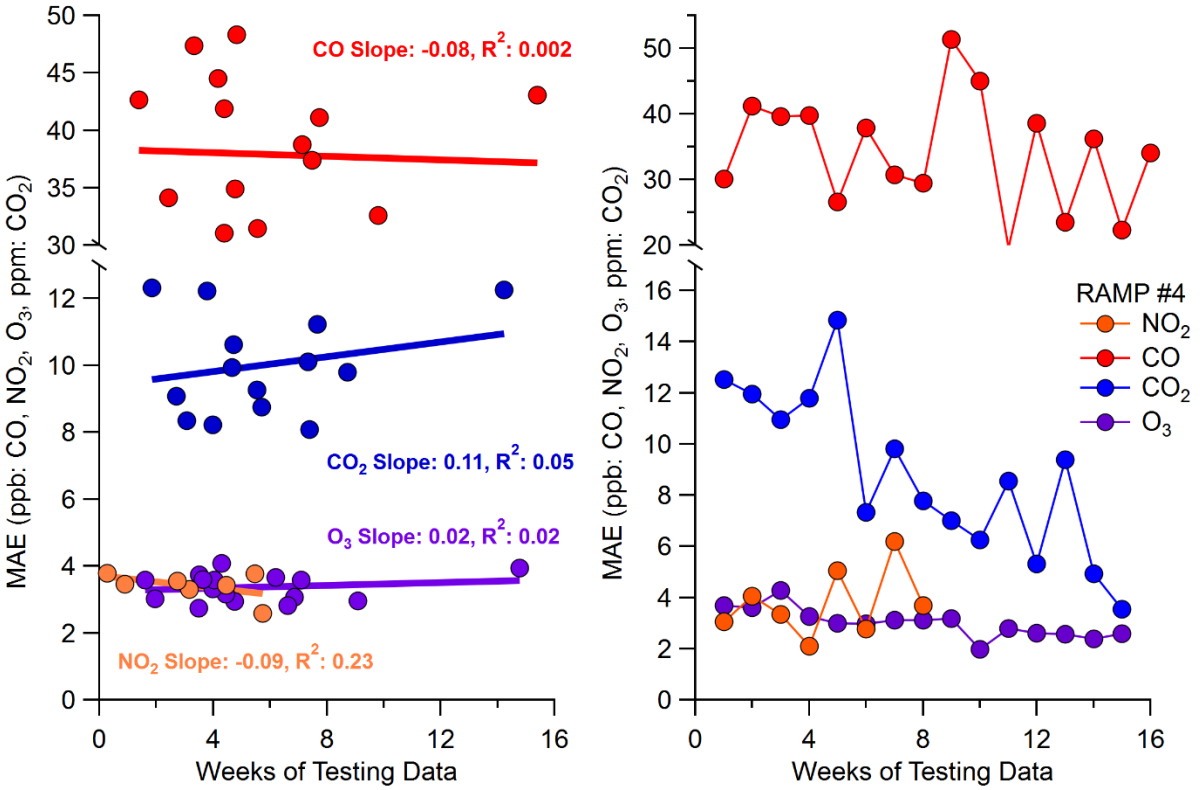

Figure 8: ***Left:*** **Mean absolute error (MAE) versus the length of the testing period for CO (red), CO_2 (blue), NO_2 (orange) and O_3 (purple) for all the RAMPs.** ***Right:*** **Changes in MAE over time for the RAMP with the longest testing window (RAMP #4). The figure shows that the MAE is generally unchanged (or in some cases improves) as the amount of testing data increases, suggesting the RF models are stable over ~~long~~ the study period~~s~~.**

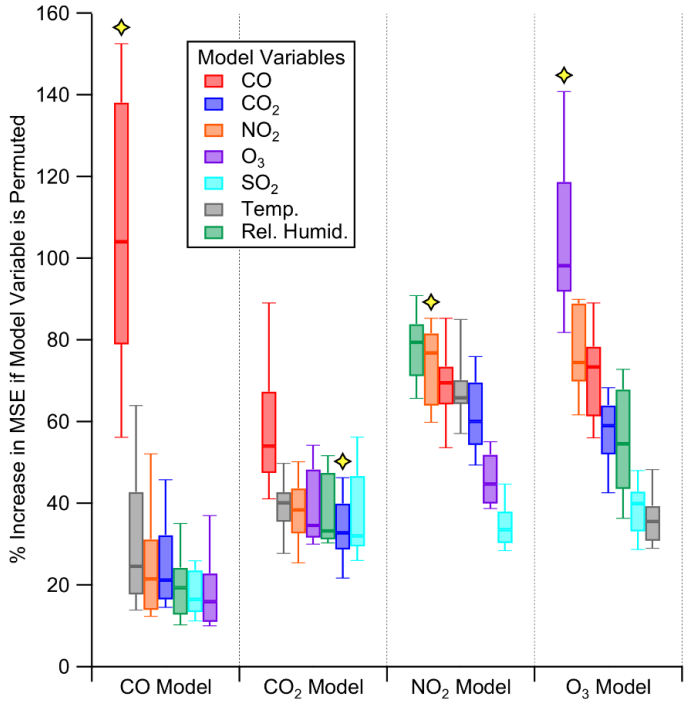

Figure 9: Importance of the explanatory variables to each of the RF models. For each model, the explanatory variables are rank ordered from most to least important, and the sensor response corresponding to the target analyte is marked with a yellow star. The box plots represent the range of importance across the 10-16~~19~~ RAMPs (whiskers: 10th and 90th percentile, box edges: 25th and 75th percentile). The relative importance is determined by calculating the increase in mean square error if the explanatory variable is permuted (i.e., randomly shuffled).

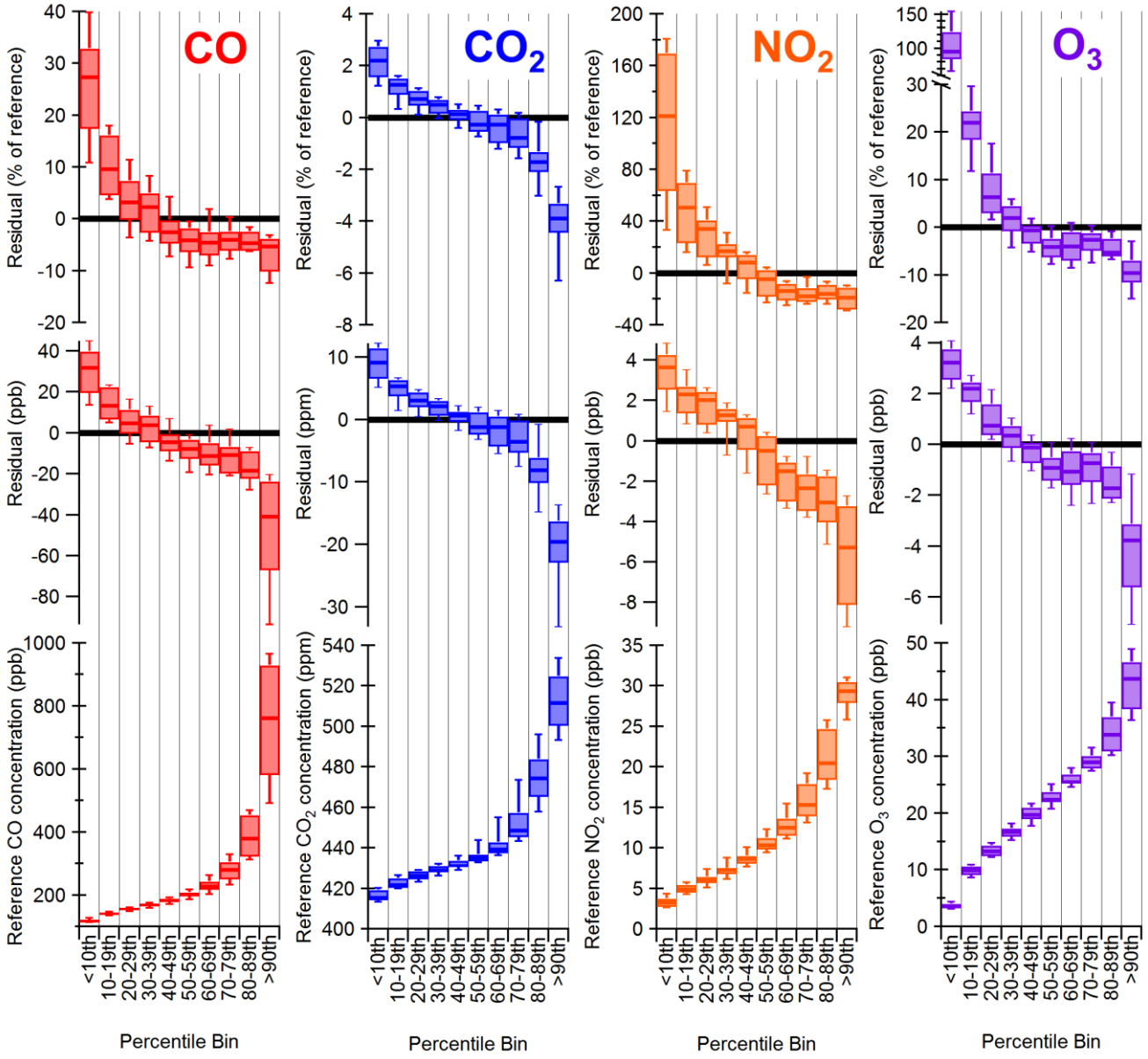

Figure 10: Box plots from the 10-169 RAMP monitors of median concentrations measured by monitors (bottom) and median model residuals (middle) and model residuals normalized to the reference concentration (top) for each pollutant, divided into deciles. The box plots provide the range of medians by the different RAMP monitors.

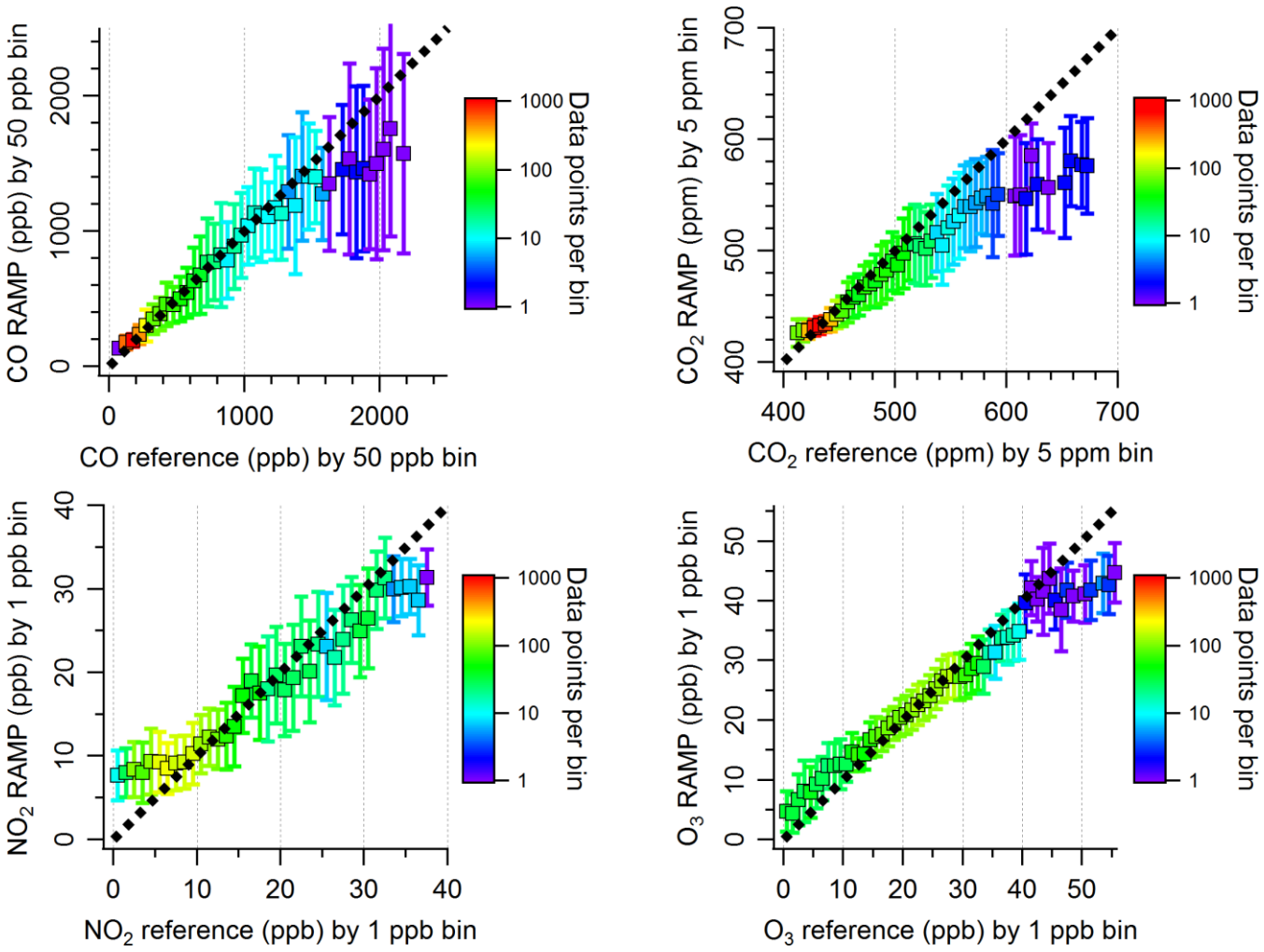

Figure 11: Illustrating the range of predictions from the 500 trees for RAMP #1. The testing data were binned and averaged. The concentration measured by the ~~reference~~ calibrated RAMP monitors is then plotted against the average concentration from the ~~model~~reference monitor. The error bars represent 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.

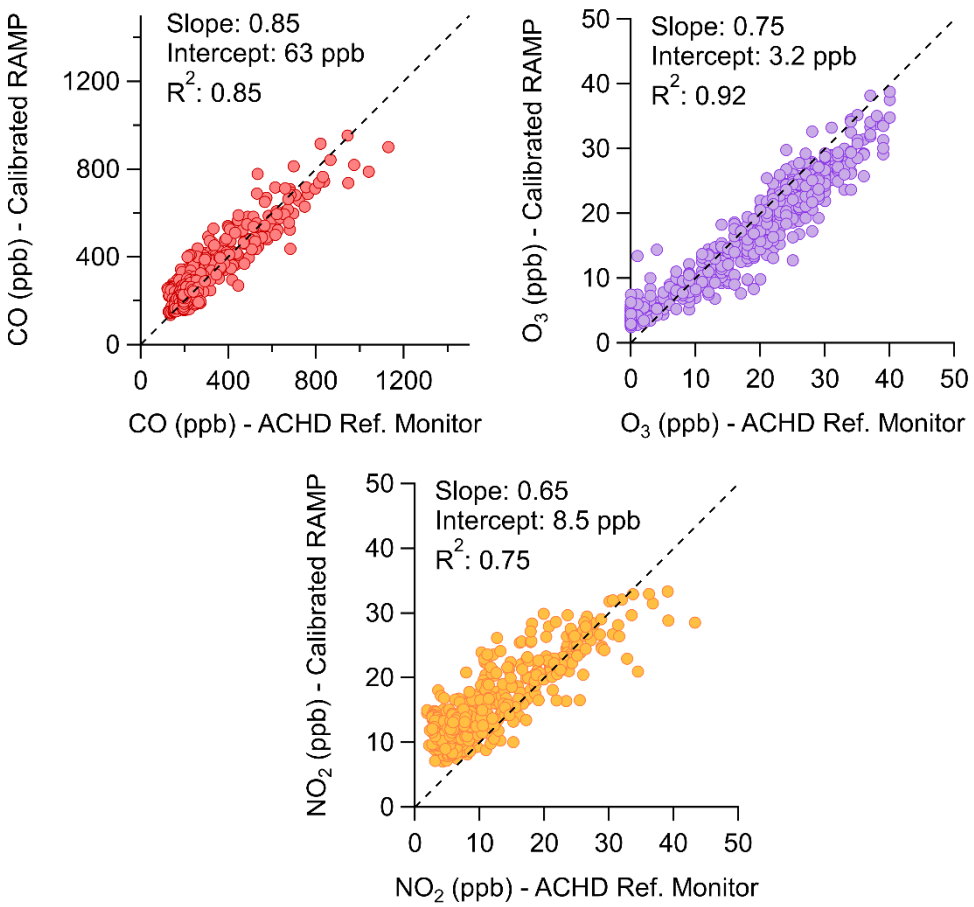

Figure 12: Comparison of CO, NO_2 and O_3 hourly average concentrations measured by a co-located RAMP monitor and the reference monitors at the Allegheny County Health Department (ACHD). The RAMP monitor was first calibrated on the Carnegie Mellon campus prior to deployment.

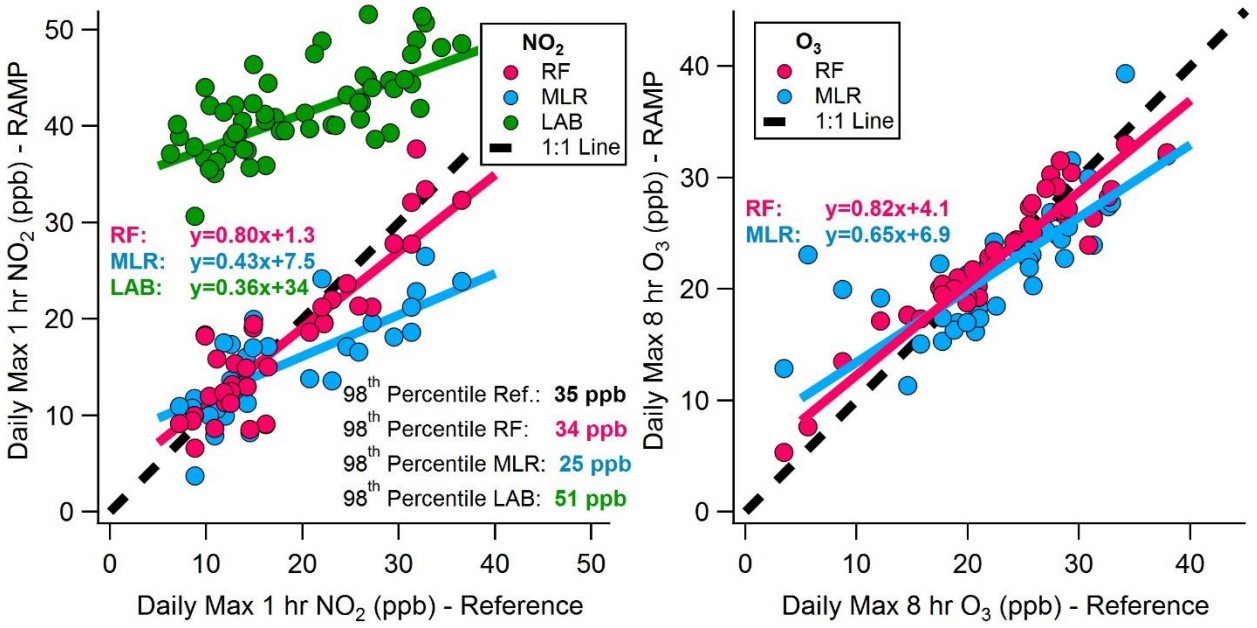

Figure 1~~3~~2: Performance of one representative RAMP (RAMP#1) for NAAQS compliance metrics (O_3: Daily Max 8 h, NO_2: 98th percentile of Daily Max 1 h averages) Right: comparison of daily 8 hr maximum reference monitor ozone concentrations (x-axis) to MLR and RF models. Left: comparison of daily 1 h maximum reference monitor concentrations versus the LAB, MLR and RF models. The NO_2 standard is the 98th percentile of the daily 1 h maximums.

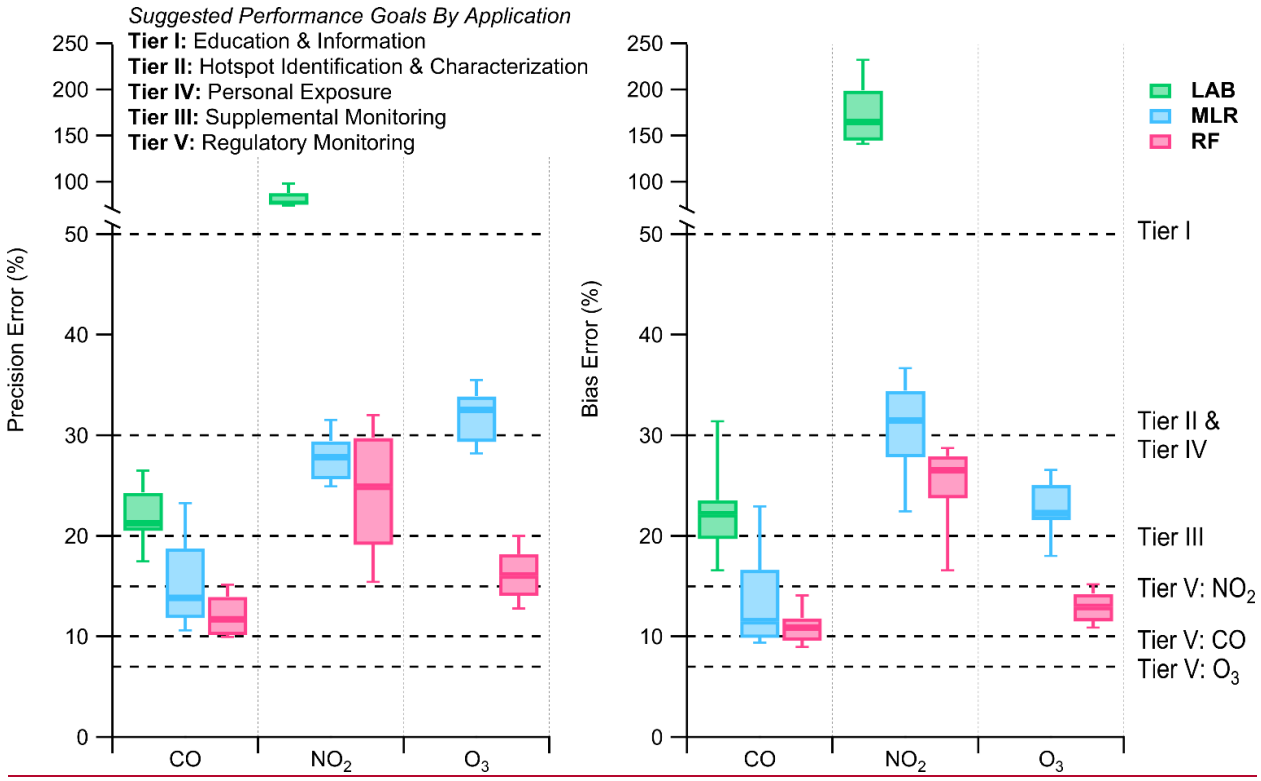

Figure 14: Precision (left) and bias (right) estimates of RAMP monitors calibrated using LAB, MLR, and RF models compared to the suggested performance goals by application as recommended in the EPA Air Sensor Guidebook. The precision estimator is the upper bound of the coefficient of variation (upper bound of the relative standard deviation, RSD). The box plots are the range of performance across the calibrated RAMP monitors (testing data only). The calibrated RAMP monitors meet the recommended error limits for exposure (Tier IV).

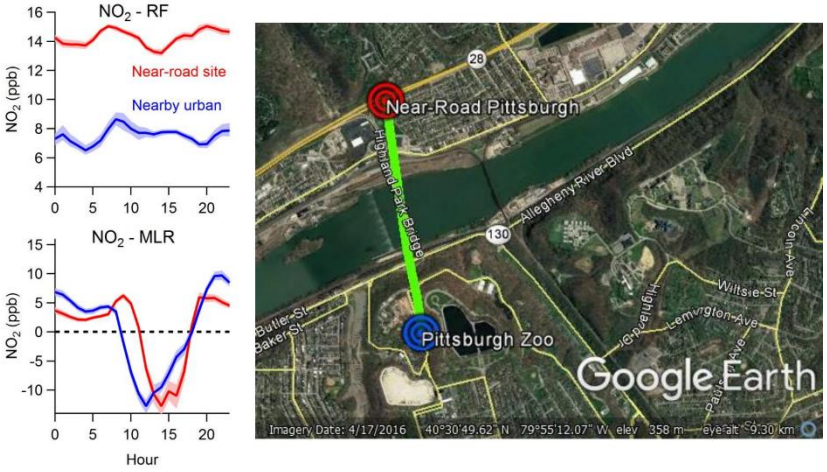

Figure 153: Left: Diurnal NO_2 patterns at two nearby sites (one urban, one near-road) measured by RAMP monitors calibrated using RF models (top) or MLR models (bottom), Right: Satellite view of the two sites, which were ~1.5 km apart. The urban site was at the Pittsburgh Zoo and the near-road site was within 15 m of Highway 28.

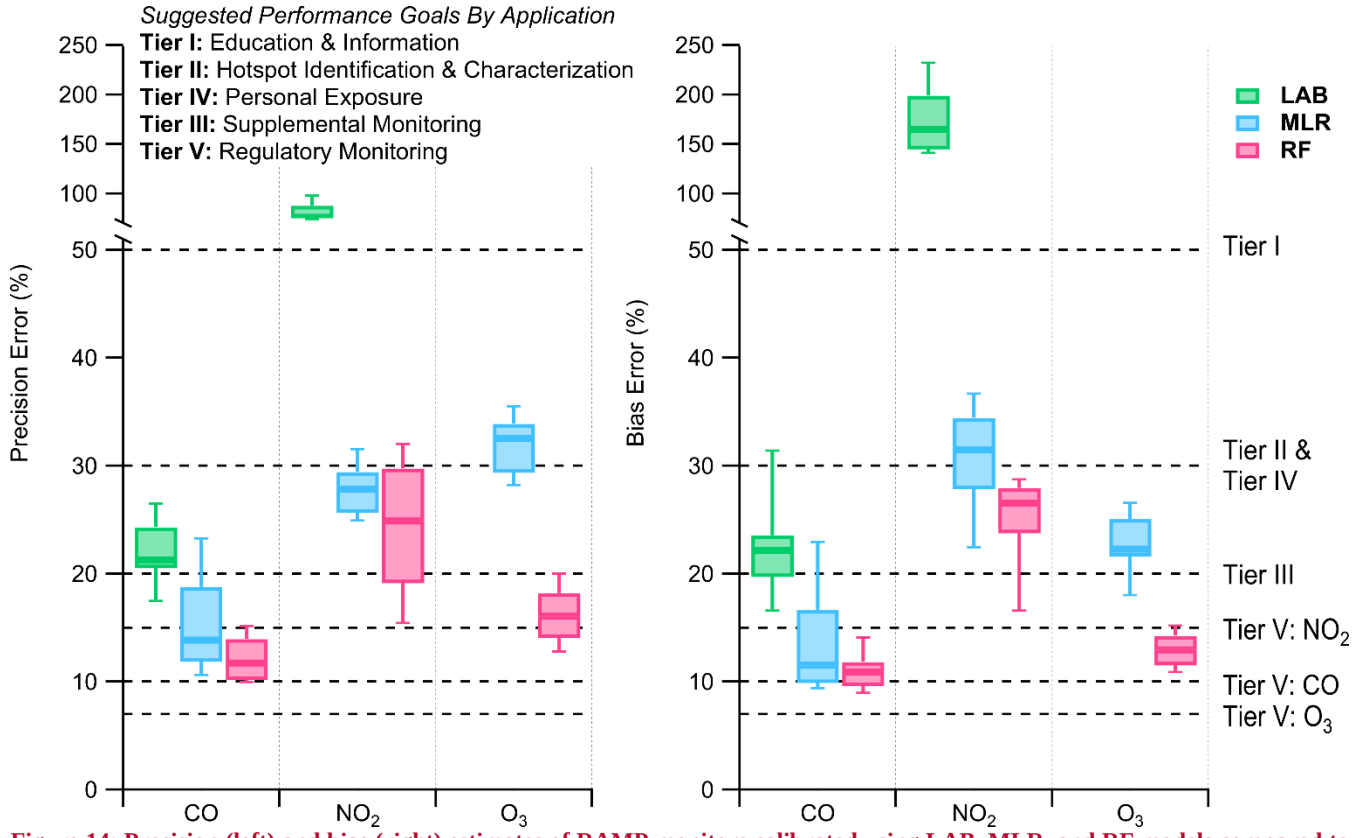

Figure 14: Precision (left) and bias (right) estimates of RAMP monitors calibrated using LAB, MLR, and RF models compared to the suggested performance goals by application as recommended in the EPA Air Sensor Guidebook. The precision estimator is the upper bound of the coefficient of variation (upper bound of the relative standard deviation, RSD). The box plots are the range of performance across the calibrated RAMP monitors (testing data only). The calibrated RAMP monitors meet the recommended error limits for exposure (Tier IV).

Table 1: Calibration ranges for laboratory-based calibration (LAB)

Pollutant	Calibration Range	Points per Calibration
CO	0 – 1600 ppb	3-4
NO_2	0 – 50 ppb	3-4
CO_2	0 – 500 ppm	3-4

Table 2: Performance metrics for fits to training data (i.e., goodness of fit) discussed in Section 4.1.

Type	Species	# RAMPs	Avg. Pearson r (±SD)	Avg. MAE (±SD)	Avg. CvMAE (±SD)	β_0 (±SD)	β_1 (±SD)	β_2 (±SD)	β_3 (±SD)
LAB	CO	9	0.99 (±0.01)	132 (±32 ppb)	38% (±17%)	-119 (±53)	0.82 (±0.69)	-	-
	CO_2	14	0.99 (±0.01)	28 (±24 ppm)	24% (±12%)	20 (±36)	0.98 (±0.13)	-	-
	NO_2	14	0.99 (±0.01)	35 (±8 ppb)	188% (±48%)	-14 (±4.9)	0.62 (±0.15)	-	-
Type	**Species**	**# RAMPs**	**Avg. Pearson r (±SD)**	**Avg. MAE (±SD)**	**Avg. CvMAE (±SD)**	**β_0 (±SD)**	**β_1 (±SD)**	**β_2 (±SD)**	**β_3 (±SD)**
MLR	CO	19	0.94 (±0.06)	39 (±13 ppb)	15% (±5%)	32 (±50)	1.3 (±0.2)	-1.1 (±2.8)	-0.1 (±0.6)
	NO_2	16~~9~~	0.59 (±0.17)	4.6 (±0.7 ppb)	42% (±5%)	3.9 (±16)	1.2 (±0.5)	0.1 (±0.3)	-0.1 (±0.2)
	O_3	19	0.81 (±0.06)	5.1 (±0.6 ppb)	24% (±2%)	9.4 (±14)	0.92 (±0.2)	0.1 (±0.2)	-0.2 (±0.2)
	CO_2	19	0.49 (±0.13)	19 (±3 ppm)	4% (±1%)	390 (±72)	0.1 (±0.1)	-0.8 (±0.7)	0.1 (±1.0)
Type	**Species**	**# RAMPs**	**Avg. Pearson r (±SD)**	**Avg. MAE (±SD)**	**Avg. CvMAE (±SD)**	**Median m_{try}**	**m_{try} = 2**	**m_{try} = 4**	**m_{try} = 7**
RF	CO	19	0.99 (±0.00)	7.9 (±1.5 ppb)	3% (±0.5%)	7	11%	21%	68%
	NO_2	16~~9~~	0.99 (±0.01)	0.5 (±0.1 ppb)	5% (±1%)	4	21%	74%	5%
	O_3	19	0.99 (±0.00)	0.7 (±0.1 ppb)	3% (±0.4%)	4	0%	84%	16%
	CO_2	19	0.99 (±0.00)	1.7 (±0.3 ppm)	0.4% (±0.1%)	2	74%	21%	5%

LAB: Laboratory calibration (Eq. 1), MLR: multiple linear regression (Eq. 2), RF: random forest model.
For the LAB and MLR models, the fit coefficients are provided.
For the RF models, the median mtry value across the 16-19 RAMPs and the breakdown of the mtry tuning results (m_{try} which minimized RMSE) across the 16-19 RAMPs results are provided.

Table 3: Comparison to other published studies.

	Project	Location	Sensor Node	Type	N (days)	Time Res. (min)	AvgConc (ppb)	Slope	R^2	MAE (ppb)	MBE (ppb)
CO	EuNetAir[1]	Aveiro, PT	AirSensorBox	EC	6	60	330	NR	0.76	90	0
	EuNetAir[1]	Aveiro, PT	NanoEnvi	EC	9	60	330	NR	0.53	100	100
	EuNetAir[1]	Aveiro, PT	Cambridge CAM11	EC	14	60	330	NR	0.87	180	-200
	EuNetAir[1]	Aveiro, PT	AQMesh	EC	15	60	330	NR	0.86	50	0
	CAIRSENSE[2]	Decatur, GA, US	AQMesh	EC	110-111	60	330	NR	0.77-0.87	NR	NR
	CAIRSENSE[2]	Decatur, GA, US	Air Quality Egg	MOS	115-196	60	310	NR	<0.25	NR	NR
	Castell et al.[3]	Kirkeveien, NO	AQMesh	EC	72	15	NR	0.88*	0.36	150	-150
	Spinelle et al.[4]	Ispra, IT	Figaro, e2V	EC, MOS	85	60	230	1.01-1.38	0.29-0.37	NR	NR
	Cross et al.[5]	Boston, MA, US	ARISense	EC	~~120~~75 **	5	--	0.9~~46**~~	0.88~~96**~~	~~NR~~24.8	-10.4~~NR~~
	This Study	**Pittsburgh, PA, US**	**RAMP**	**EC**	**41 [10-108]**	**15**	**270 (±30)**	**0.86 (±0.09)**	**0.91 (±0.05)**	**38 (±6.5)**	**0.1 (±0.2)**
NO_2	EuNetAir[1]	Aveiro, PT	Cambridge CAM11	EC	14	60	16	NR	0.84	5.61	-2.3
	EuNetAir[1]	Aveiro, PT	AirSensorBox	EC	7	60	16	NR	0.06	20.2	17.7
	EuNetAir[1]	Aveiro, PT	NanoEnvi	EC	7	60	16	NR	0.57	14.9	13.1
	EuNetAir[1]	Aveiro, PT	ECN_Box_10	EC	11	60	16	NR	0.89	4.95	-1
	EuNetAir[1]	Aveiro, PT	AQMesh	EC	6	60	16	NR	0.89	1.46	0
	EuNetAir[1]	Aveiro, PT	ISAG	MOS	13	60	16	NR	0.02	16.2	349.5
	CAIRSENSE[2]	Decatur, GA, US	Cairclip	EC	194-285	60	11	0.96	<0.25-0.57	NR	NR
	CAIRSENSE[2]	Decatur, GA, US	AQMesh	EC	110-111	60	10	NR	<0.25	NR	NR
	CAIRSENSE[2]	Decatur, GA, US	Air Quality Egg	MOS	115-196	60	11	NR	<0.25	NR	NR
	Duvall et al.[6]	Houston, TX, US	Cairclip	EC	24	60	5.5	0.25	0.01	NR	NR
	Duvall et al.[6]	Denver, CO, US	Cairclip	EC	30	60	5.1	0.04	<0.01	NR	NR
	Castell et al.[3]	Kirkeveien, NO	AQMesh	EC	72	15	NR	0.2-0.38*	0.24	26.2	13.3
	Esposito et al.[7]	Cambridge, UK	SnaQ	EC	28	1	NR	NR	0.83	1.27	NR
	Spinelle et al.[8]	Ispra, IT	αSense, Citytech	EC	86	60	9	0.64-0.79	0.55-0.59	NR	NR
	Cross et al.[5]	Boston, MA, US	ARISense	EC	~~89~~120 **	5	NR	0.8~~3**~~	0.6~~980**~~	3.45~~NR~~	~~NR~~1.20
	This Study	**Pittsburgh, PA, US**	**RAMP**	**EC**	**24 [2-56]**	**15**	**12 (±1.4)**	**0.64 (±0.11)**	**0.67 (±0.12)**	**3.48 (±0.36)**	**-0.4 (±1.13)**
O_3	EuNetAir[1]	Aviero, PT	AirSensorBox	EC	6	60	17	NR	0.13	22.12	19.2
	EuNetAir[1]	Aviero, PT	NanoEnvi	MOS	9	60	17	NR	0.77	7.66	6.5
	EuNetAir[1]	Aviero, PT	Cambridge CAM11	EC	11	60	17	NR	0.14	21.5	15.7
	EuNetAir[1]	Aviero, PT	AQMesh	EC	6	60	17	NR	0.7	2.4	0
	EuNetAir[1]	Aviero, PT	ISAG	MOS	13	60	17	NR	0.12	360.12	356.1
	CAIRSENSE[2]	Decatur, GA, US	Aeroqual SM50	GSS	168-281	60	18	0.81-0.96	0.82-0.94	NR	NR
	CAIRSENSE[2]	Decatur, GA, US	Cairclip	EC	194-285	60	17	0.68-0.85	0.68-0.88	NR	NR
	CAIRSENSE[2]	Decatur, GA, US	AQMesh	EC	110-111	60	15	NR	<0.25	NR	NR
	Duvall et al.[6]	Houston, TX, US	Cairclip	EC	24	60	32	0.93	0.80	NR	NR
	Duvall et al.[6]	Denver, CO, US	Cairclip	EC	30	60	46	1.19	0.77	NR	NR
	Castell et al.[3]	Kirkevein, NO	AQMesh	EC	72	15	NR	0.11-0.26*	0.29	19.9	6.8

Esposito et al.[7]	Cambridge, UK	SnaQ	EC	28	1	NR	NR	0.69	7.45	--
Spinelle et al.[8]	Ispara, IT	αSense, Citytech	EC	82-84	60	30	1.02-1.12	0.86-0.91	NR	NR
Cross et al.[5]	Boston, MA, US	ARISense	EC	~~~87120~~ **	5	NR	0.4762**	0.3951**	NR7.34	NR0.78
This Study	**Pittsburgh, PA, US**	**RAMP**	**EC**	**38 [11-103]**	**15**	**22 (±1.4)**	**0.82 (±0.05)**	**0.86 (±0.02)**	**3.36 (±0.41)**	**-0.14 (±0.46)**

[1](Borrego et al., 2016), [2](Jiao et al., 2016), [3](Castell et al., 2017), [4](Spinelle et al., 2017), [5](Cross et al., 2017), [6](Duvall et al., 2016), [7](Esposito et al., 2016), [8](Spinelle et al., 2015)

EC=electrochemical, MOS=metal oxide sensor, GSS=gas sensitive semiconductor. NR= not reported in manuscript.
For RAMP data, bracketed data is range (for N days) or standard deviation (all other metrics) across all the RAMP units.
*values for slopes only provided for a subset of 2 of 24 sensors
~~**Cross et al. performance metrics reported for full collocation dataset which includes both testing (65% of data) and training (35%) of data. Performance metrics for other studies are only based on testing data not used for model fitting/training.~~