Response to Comments from Aerodyne Research Inc: AMT-2017-260

The authors would first of all like to thank Aerodyne (and specifically Dr. Eben Cross) 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.

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 SO2 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.**"