We would like to again thank the editor and reviewers for their constructive comments. We have tried to address these comments in the attached response document and in the manuscript. Comments are reproduced in black, our responses are in blue.

Associate Editor

Comments to the Author:

One review recommends 'publish as it is', and another recommends 'minor changes' as suggested in below. Please address the recommendations before finalizing the paper for submission and publication.

peer reviews:

This is an improved manuscript – in particular, the additional details about the training and test data sets have substantially clarified the authors' approaches to sensor calibration. Most of my concerns have been addressed; remaining comments are listed below.

If training and test sets are not the same (p. 6, line 1), it's hard to understand the utility (or even the meaning) of Table 4 – combining all test and training data into combined ranges isn't terribly meaningful, since there's substantial overlap between the two. Moreover combining them even risks misleading the reader as to the true ranges of the two sets for individual sensors – the ranges for an individual sensor might be quite different than those shown in the table. I would recommend changing this table substantially to make clear the differences in training/test sets for individual sensors. In addition, these differences should be discussed in the table caption. If there was a way to visualize these differences (maybe some example histograms?) that would be helpful also.

Table 4 lists ranges of upper, lower, and average concentration values across all sensors. This gives an indication of the variability in these concentrations across the RAMPs. For example, for O_3, average concentrations during the training period can vary from 21 to 36 ppb, while the maximum 15-minute-average concentration for the training period can vary between 62 and 128 ppb across RAMPs. Supplemental information figures S11 through S15 provide boxplots depicting the ranges of values for the training and testing set used for each gas and sensor to provide more specific information. The caption for the table has been expanded to better explain this, and make reference to the supplemental information for more details:

> "Table 4: Durations and ranges of testing and training data at CMU in 2017. Durations of the training and testing periods are in days. Ranges indicated are in ppb for all gases, degrees Celsius for temperature (T), and percent for relative humidity (RH). Note that because training and testing periods vary for different RAMPs, as described in Section 2.3, the duration and concentration ranges of the training and testing periods will likewise vary. This table gives an indication of the variability in training and testing period durations across RAMPs, as well as the variability in concentrations, temperature, and relative humidity. The "lower range" indicates the variability in the lowest 15-minute-average concentration experienced by RAMPs during training or testing. Likewise, the "upper range" indicates the variability in highest 15-minute-average concentrations. The "average range" indicates the variability in the average 15-minute-average concentration across RAMPs for either the training or testing period. Further information about these ranges is provided in the supplementary information (see Figures S11-S15)."

Results: in my original review I had suggested that the random forest and hybrid approaches shouldn't be different, since the training and test sets appeared to be identical. But since the ranges given in Table 4 turn out to be combined ranges, and not ranges covered by each individual sensor, this may be incorrect – there may be sensors for which the ranges of the training and test sets differ substantially. (However, whether this is actually the case is hard to evaluate based on the information in the paper and SI.) In such cases the two models may be expected to give different results, so the two could be discussed individually.

Regardless, if hybrid model is to be left in, the authors should still need to provide information on the number of "crossings" between the RF and LR models, and the fraction time evaluated by RF vs time evaluated by LR.

Although ranges vary from RAMP to RAMP, for a given RAMP the same range will be used for the training of all algorithms, including random forest and hybrid models. However, since the performance of the random forest model on calibration of RAMP data has already been discussed in previous work (Zimmerman et al., 2018), we will continue to refrain from including it in the current manuscript.

Information has been included about the active times of the random forest model as part of the hybrid approach, as well as the average interval between "crossings" from the random forest to linear models (P. 12, lines 7-11):

> "For the hybrid models, across gases, their random forest components were typically active from 88% to 93% of the time (or one "crossing" from random forest to linear models every 12 to 17 hours of active sensor time), although for specific RAMPs this ranged from 75% to 99% (one "crossing" every 5 to 83 hours) depending on the ranges of training and testing data. For perspective, the random forest component would be active 90% of the time if the distributions of training and testing concentrations were identical."

P. 16, lines 9-11: it should also be mentioned that the clustering algorithm would likely be improved by use of kNN (rather than k-means-clustering, which is what is used).

This has been added (P. 16, lines 9-11):

> "These results could perhaps be improved further; for instance, our linear and quadratic regression models did not use regularization, nor did we experiment with neural networks involving multiple hidden layers and varying numbers of nodes, nor with the use of different k-means and nearest-neighbor algorithms for clustering."

P. 16, lines 13-15: if the authors are going to continue to make this suggestion based on the current work (even with the new caveat added), they need to be back it up with much more than a citation to another paper. Specifically they need to show some evidence that the differences in performance results from the RAMP circuitry, and not from differences in the training/test set used. I'm not sure how one would do this, but as written the sentence is purely speculative, and not backed up with any substantive evidence.

We will leave the evaluation of the RAMP circuitry performance as a topic for future work. We have therefore removed this statement.

P. 18 line 14: it is stated that a new model should be developed "each year", but this is probably more specific than is warranted from the work. My takeaway from this work is that models stay reasonably robust for timescales of several months, but should be periodically evaluated/updated when used over longer timescales (on the order of every ~6-18 months). I would recommend changing to wording to reflect this. This recommendation is also included in the abstract,

and so should be changed there as well. (As a minor side note, I feel including it in the abstract risks detracting from the more fundamental results of this work, related to generalized models. So I might recommend removing or shortening the sentence in lines 20-22 of the abstract.)

The recommendation has been updated (P. 18, lines 10-13):

> "For long-term deployments, it is recommended that model performance be periodically re-evaluated (using limited co-location campaigns with a subset of the deployed sensors) every 6 months to one year, and that development of new calibration models be contingent on the outcomes of these re-evaluations. This recommendation is due to the noticeable change in performance when models for one year were used for processing data collected in the subsequent year."

The sentence in the abstract has also been changed (P. 1, lines 20-22):

> "For long-term deployments, it is recommended that model performance be re-evaluated and new models developed periodically, due to the noticeable change in performance over periods of a year or more."

SI: in the Response to Reviews the authors state that "the randomized nature of the training approach for some models (such as the random forest models) will lead to slightly different results if these models are re-built." I don't understand this. If the algorithm uses a random seed to generate psuedo-random numbers, the same psuedo-random numbers should be generated each time, so results should be replicable. (More generally, if the results are indeed different when the model is re-run, this represents a potentially major problem, as it calls into question the robustness of the results.)

Pseudo-random number generation is performed within the MATLAB packages related to neural network and random forest model calibration, and so we cannot guarantee that the exact same results will be achieved every time, due to the pseudo-random number generation being performed within these pre-existing packages. However, the resulting differences should not be enough to noticeably affect the results, and are (based on some tests conducted by the authors) several orders of magnitude lower than the inter-RAMP variability presented for the results.