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

General comments

In this study, authors compared different models used to adjust data from low-cost sensors and compared model application in different sites within the Pittsburgh, Pennsylvania region of the USA. These comparisons demonstrate that at the sites different model-based calibration techniques show differing abilities to accurately produce concentration data for the various sensors employed. However, there existing two main problems: 1 for calibration period and deployment, it is unclear which based on the text, the calibration data also used for testing or vice versa, more details about the deployment should be illustrated.

All presented results correspond to "testing" data, which is separate from the "training" data used to calibrate the models. This has been clarified in the text (Page 10, Lines 7-9):

"It should be noted that the metrics presented here are applied only for testing data, i.e., data which were not used to build the calibration models. Model performance on the training data is expected to be higher, and thus less representative of the true capability of the model."

Also, several figures have been added to the supplemental materials illustrating which periods of time are set aside as training and testing data for different sensors. Various other changes have been made throughout the manuscript to clarify what data are being considered, and what models are being applied.

2 Evaluation validity: As listed in Table 4, the average pollutant concentration is relatively small for NO and NO2 both in training and testing period, which about 1.7 and 6.4 ppb level, how did the investigators ensure sensor evaluation validity in such low-level situation, while EC sensor has a lower detection limit at about this level? And for NO, the MAE of all models are even large than average concentration.

Sensor performance was evaluated using only ambient concentrations of pollutants, as their performance in this regime is the most relevant to their performance during actual deployments. This is reflected in the presented results; for example, CvMAE for NO sensors is on the order of 1, indicating absolute errors are of the same magnitude as the readings themselves. We believe this to be an honest way to assess the RAMP sensor and calibration model performance in line with the intended use case of the monitor.

In the specific cases of NO and NO_2, we also present the performance of the calibration models at the "Parkway East" site in Figure 4, which was chosen specifically because it typically experiences higher concentrations of these pollutants than our training (CMU) site. In both cases, the performance of the generalized calibration models at these sites (filled markers) was similar to their performance at the training site (hollow markers); thus, we believe the presented performance results to be robust across most typical ambient concentrations which we would encounter in the city of Pittsburgh or any other city with similar climate and air quality characteristics.

In general, we have focused on assessing the relative calibration model performance in terms of CvMAE (e.g. as presented in Figure 2), since we believe this to be readily understandable and to allow easy performance comparisons between sensors measuring different pollutants. However, we have also included information on absolute performance in terms of MAE (e.g.

1

as presented in Table 5) for interested readers, to allow them to estimate what the expected performance of a similar low-cost sensor package might be in the specific environment they are interested in.

Specific comments

1. Page 1, Line 24: Authors use "These results will help guide future efforts in the calibration and use of low-cost sensor systems worldwide". Perhaps the approaches merit many further trials in the widely differing pollutant and meteorological conditions It is unclear. It is unclear how the study demonstrates protocols that apply worldwide?

We believe that the primary conclusion of the paper (i.e. that generalized models for low-cost gas sensor calibration provide comparable performance to individualized models) is widely applicable to any low-cost sensor package with a similar design to the RAMP monitor (including similar electrochemical sensors and appropriate signal conditioning), regardless of where it is used so long as the generalized calibration model is developed for a network of sensors in the region where the network is deployed. We also believe that the methods used to calibrate models for low-cost gas sensor data presented in the paper are widely applicable to any similar monitoring system that produces clean, reproducible raw data of quality comparable to the RAMP output. Although specific performance metrics will likely differ for different locations, meteorological conditions, and low-cost sensor packages, and thus require future testing to assess he performance of specific monitoring packages in specific environments, we do believe that our qualitative results provide important guidance to future low-cost sensor calibration efforts, and that this guidance is generally applicable worldwide.

2. Page 2, line 16: "These sensors tend to have lower signal-to-noise ratios than regulatory-grade instruments" is a fairly unusual and non-specific way to say they are not as precise or sensitive as conventional air monitors.

This has been corrected (Page 2, Lines 17-18):

"These sensors are less precise and sensitive than regulatory-grade instruments"

3. Page 2, Line 20: Authors mentioned that a nonlinear interaction in the reference (Jiao et al., 2016), however, this paper didn't mention this kind of nonlinear interaction, but cited Gao's paper (Gao et al., 2015) which also didn't mention this, rather it is a nonlinear response for PM sensor.

This citation was incorrect. This has been corrected.

4. Page 2, Line 25: Same problem for reference (Cross et al., 2017) as comment 3.

That paper makes use of high-dimensional model representations involving non-linear component functions (cubic functions are specifically mentioned). We believe this reference is correct.

5. Page 3, line 29: It appears that on-campus field calibrations were conducted during a brief period in summer and early fall. Where winter time calibrations performed as well? If not, how might this impact the application of various results of modelling over winter time conditions?

Field calibrations for some sensors (and for the gRAMP models) included data collected during October, during which conditions were similar to what might be expected during winter (e.g. near-freezing temperatures). Furthermore, time-resolved performance for several sensors across multiple seasons is shown in Figure 7; there is no apparent difference in performance across seasons.

6. At the end Section 2. which describes monitoring methods there is a general reference to the Zimmerman 2018 paper to provide the reader information on protocols. Key factors should be pulled into the current text or figures. Specifically, on Line 30 what were the specific models of regulatory monitors used and how were they located/protected from environmental factors? What methods were employed to measure speciated VOCs mentioned? What is meant by the term "or BTEX"? This does not appear to be covered in the 2018 Zimmerman paper and it does not seem that VOCs are part of this study. Perhaps this needs to be removed from the current text. Further, SO2 is included in the list of pollutants that appear to have been included in this study. However, no data for SO2 appears. Further, several of the models include factors for SO2. Was SO2 measured or employ in this study?

Raw signals from SO_2 and VOC sensors within the RAMP monitors are used as potential inputs into the calibration models developed to account for possible cross-sensitivities. However, no calibration models were developed for SO_2 or for any VOC species as a part of the work presented in this paper. References to the monitoring of SO_2 and VOCs have therefore been removed from this section. For the other gases, the instrument models have been given, and a general description of the location of the instruments has been provided (Page 4, Lines 3-7):

> "Less than 10 meters from the RAMP monitors, a suite of high-quality regulatory-grade instruments, measuring ambient concentrations of CO (with a Teledyne T300U instrument), CO2 (LICOR 820), O3 (Teledyne T400 Photometric Ozone Analyser), and NO and NO2 (2B Technologies Model 405nm) are stationed to provide true concentration values for these various gases to which the RAMP monitors are exposed. These regulatory-grade instruments are contained within a mobile laboratory van, into which samples are drawn through an inlet 2.5 meters above ground level."

7. Page 4, Line 5: The collocation in ACHD points seems to span the whole year of 2017 in each week at Lawrenceville as shows in Figure 7, while in Table 4 the maximum days of testing period duration only ranged from 75 to 110 days for CO, O3 and NO2. The authors should give more explanation about the deployment? It is unclear and could not be found in the cited reference (Zimmerman et al., 2018). Also, as mentioned in Page 5, Line 9, 80% of collocation is about 28 days less than 1 month, what are other days while RAMP collocation such long time and what is the criteria for period selection?

Table 4 presents information related to the CMU site only. This has been indicated in the caption. Additional details about how data are divided into training and testing periods are provided in the text of Section 2.3, and figures illustrating these periods have been included in the supplemental information as well.

8. Page 4, Line 10: Authors mentioned ACHD Parkway East has high levels of NO and NO2 for maximum at 100ppb and 40ppb, this period also included in testing period, but in Table 4, for whole testing period these two maximum concentrations are 66 and 32.

Table 4 presents information related to the CMU site only. This has been indicated in the caption.

9. Page 4, Line 22: The best model was selected based on performance of individual models, what is the criteria for this?

Pearson r is used to select the best-performing individual model. This has been clarified in the text (Page 4, Line 33 to Page 5, Line 2):

3

"Second, from these individualized models, a best individual calibration model (bRAMP) was chosen, which performed best out of all the individualized models on a testing data set with respect to correlation (Pearson r, see Section 2.4)."

10. Page 9, Line 4: The random forest and hybrid model used all sensor data for training CO, and combined linear model which only used CO, why not introduce other sensors in this procedure?

For the linear models, based on prior work by Zimmerman et al., we used only the signal for the target gas of interest (or, in the case of O_3, signals from both NO_2 and O_3 sensors) along with T and RH since only responses to these factors could be reliably modeled as linear. For the quadratic models, we sought to investigate the impacts of including inputs from the other gas sensors in a non-linear model; for this reason, we investigate both "limited" quadratic models, which use the same limited set of inputs as the linear models, and "complete" quadratic models, which use the inputs of all sensors on the RAMP (as do all of the models discussed later, including the hybrid models).

11. Page 10, Line 8: As mentioned in the text "measurements where the corresponding true value is below an assigned lower limit are removed from the measurement set to be evaluated", while for NO2(10ppb) and O3 (10ppb) which listed in Table 1, the training period or testing period concentration range still start from 0 or 1ppb, does any filter about lower limit come into force in the training or evaluation? The authors should discuss the impact of removal of data below specified minimums as this would appear to skew the actual data and do so in a differential basis depending on ambient conditions.

The removal of values below a lower limit is applied only for the evaluation of the Precision and Bias metrics being discussed in this section, and thus only affects the results presented in Figure 5. This has been clarified in the text (Page 11, Lines 7-9):

"Note that this removal of low values is applied only when computing the precision and bias error metrics, and not when evaluating the other metrics described above."

Due to the construction of these metrics, the inclusion of values below these lower bounds would tend to make the Precision and Bias metrics worse. In other words, in ambient conditions where concentrations of pollutants are higher, the Precision and Bias metrics would be better for the same absolute errors. We have chosen to remove the low values as indicated when evaluating these metrics to allow for consistency with other studies which evaluate these metrics in the same way according to EPA recommendations.

12. Page 16, line 11: "The fact that for most gases a variety of calibration approaches show similar and (for typical uses cases, acceptable) performance may reflect better underlying performance from the RAMP monitor, as similar studies for other low-cost sensor packages showed a wider variability in performance between calibration approaches (see e.g. the summary provided by Zimmerman et al., 2018)." This statement appears to indicate that the RAMP monitors somehow performed better than other similar sensor based monitors. However, there is no rationale for this statement based on the actual monitor package, component details provided or information reported by others. It appears to simply be a diffusion based sensor deployment of Alfasense electrochemical cells and an NDIR CO2 sensor. What other factors might influence system performance?

The signal conditioning employed in the RAMP monitoring package likely contribute to higher signal-to-noise ratios compared to other similar systems, based on discussions with the device manufacturer. However, determination of whether this is the case is beyond the scope of the current paper. The statement has been qualified (Page 16, Lines 11-15):

> "The fact that for most gases a variety of calibration approaches show similar (and for typical uses cases, acceptable) performance may reflect better underlying performance from the RAMP monitor, as similar studies for other low-cost sensor packages showed a wider variability in performance between calibration approaches (see e.g. the summary provided by Zimmerman et al., 2018). This suggests that the primary difference between these monitors, i.e. the internal circuitry which is unique to the RAMP, is the cause for this consistency; however, determination of this is beyond the scope of this paper."

13. Page 25, Figure 4: The performance of CMU site for each sensor is marked in hollow marker, while there are two hollow markers in CO and NO2, which is unclear.

The hollow markers are colored to correspond to the RAMPs deployed to each site. For example, the green filled marker represents the performance of a RAMP at the Lawrenceville site, and the hollow green marker indicates the performance of that same RAMP when it was at the CMU site. This was done so that the variability of performance between sites could be compared with the variability in performance between RAMPs at a common site. This has been clarified in the text (Page 13, Lines 3-5):

> "Filled markers indicate the performance of the models at these sites, while hollow markers indicate the 2017 testing period performance of the corresponding RAMP when it was at the CMU site for comparison."

14. Page 30, Table 4: In the evaluation of testing period, R2 was used, however, in previous figure and text, such as Figure 2-4, Pearson linear correlation coefficient r was used, what is the standard for selection between these two parameters?

In reviewing the literature, there was no clear preference for the r or r^2 metric to represent correlation. When preparing the paper, we had a slight preference for the r metric, as it distinguishes between positive and negative correlations, which we envisioned might occur for some of the worse-performing calibration models and sensors. Therefore, in figures, the r metric is used. In presenting the results in tabular form, the r^2 metric is used, in recognition that some readers may be more familiar with and prefer results presented in that way. For the purposes of these results, where r^2 is given for the best-fit line between the calibrated and "true" data, the information provided by r and r^2 are the same.

15. Page 30, Table 4: The title of the paper is about a general calibration model (gRAMP) and recommended in the conclusion, while it was not compared with iRAMP models in Table 4. This paper focuses on field calibration methods and model application to produce adjusted pollutant concentration data from sensor-based monitors. However, discussions and recommendations are only made regarding field calibration approaches. It would also appear that the data agreement between sensors and regulatory monitoring might be improved by other means. For example, conventional air monitoring methods as applied to regulatory monitoring efforts always include periodic zero and span challenges. Improved circuitry or air conditioning might also improve monitor performance. How might the inclusion of such methods improve the quality of data produced by sensor-based systems?

Comparisons between iRAMP and gRAMP models are provided in Figures 3 through 6. Data corresponding to the results of Table 4 (now Table 5) for the bRAMP and gRAMP models evaluated at the CMU site have been provided as part of the supplementary materials. The main conclusion of the paper is also based on the relative performance of the gRAMP models when tested at a "new" deployment site (i.e. a site distinct from where the models are trained), as depicted in Figures 4 and 5.

5   The possibility of conducting periodic zero- and span-checks on these sensors was considered but was eventually dismissed due to the logistical challenges of performing these checks over a large network of low-cost sensors deployed over a relatively large spatial domain. Improved circuitry within the RAMP monitor to improve signal-to-noise ratios might be a possibility but was not investigated as a part of this paper (discussions with the manufacturer indicated that these monitors already have very good signal conditioning electronics compared to other comparable products). Providing climate control for the RAMP monitor

10 would greatly increase the cost and power consumption of the monitor, and so was not considered. Ongoing work is instead being focused on different automated methods of in-field calibration checks, for example by comparing with nearby regulatory monitoring sites, by comparing with data collected from mobile sampling campaigns, and by using clusters of nearby monitors in an attempt to identify and correct for outliers.

Development of a General Calibration Model and Long-Term Performance Evaluation of Low-Cost Sensors for Air Pollutant Gas Monitoring

Carl Malings[1], Rebecca Tanzer[1,3], Aliaksei Hauryliuk[1], Sriniwasa P.N. Kumar[1], Naomi Zimmerman[2],
Levent B. Kara[3], Albert A. Presto[1,3], and R. Subramanian[1]

[1]Center for Atmospheric Particle Studies, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, USA
[2]Department of Mechanical Engineering, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada
[3]Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, USA

Correspondence to: Carl Malings (cmalings@andrew.cmu.edu)

Abstract. Assessing the intra-city spatial distribution and temporal variability of air quality can be facilitated by a dense network of monitoring stations. However, the cost of implementing such a network can be prohibitive if traditional high-quality, expensive monitoring systems are used. To this end, the Real-time Affordable Multi-Pollutant (RAMP) monitor has been developed, which can measure up to five gases including the criteria pollutant gases carbon monoxide (CO), nitrogen dioxide (NO_2), and ozone (O_3), along with temperature and relative humidity. This study compares various algorithms to calibrate the RAMP measurements including linear and quadratic regression, clustering, neural networks, Gaussian processes, and hybrid random forest/linear regression models. Using data collected by almost seventy RAMP monitors over periods ranging up to eighteen months, we recommend the use of limited quadratic regression calibration models for CO, neural network models for NO, and hybrid models for NO_2 and O_3 for any low-cost monitor using electrochemical sensors similar to those of the RAMP. Furthermore, generalized calibration models may be used instead of individual models with only a small reduction in overall performance. Generalized models also transfer better when the RAMP is deployed to other locations. For long-term deployments, it is recommended that new models be developed each year, due to the noticeable change in performance when models for one year were used for processing data collected in the subsequent year. This makes annually-developed generalized calibration models even more useful since only a subset of deployed monitors are needed to build these models. These results will help guide future efforts in the calibration and use of low-cost sensor systems worldwide.

1 Introduction

Current regulatory methods for assessing urban air quality rely on a small network of monitoring stations providing highly precise measurements (at a commensurately high setup and operating cost) of specific air pollutants (e.g. Snyder et al., 2013). The United States Environmental Protection Agency (EPA) determines compliance with national air quality standards at the county level using data collected by local monitoring stations. Many rural counties have at most a single monitoring site; urban counties may be more densely instrumented, though not at the neighborhood scale. For instance, the Allegheny County Health

Department (ACHD) maintains a network of ten monitoring stations which collect continuous and/or 24-hour data for the two-thousand-square-kilometer Allegheny County (with a population of 1.2 million) in Pennsylvania, USA, with only one of these stations providing continuous data for all EPA criteria pollutants listed in the National Ambient Air Quality Standards (NAAQS) (Hacker, 2017). However, air pollutant concentrations can vary greatly even within urban areas due to the large

5   number and variety of sources (Marshall et al., 2008; Karner et al., 2010; Tan et al., 2014). This variability could lead to inaccurate estimates of air quality based on these sparse monitoring data (Jerrett et al., 2005).

One approach to increasing the spatial resolution of air quality data is the use of dense networks of low-cost sensor packages. Low-cost monitors are instruments which combine one or more comparatively inexpensive sensors (typically electrochemical or metal oxide sensors) with independent power sources and wireless communication systems. This allows larger numbers of

10   monitors to be employed at a similar cost to a more traditional monitoring network as described above. The general goals of low-cost sensing include supplementing existing regulatory networks, monitoring air quality in areas that have lacked this in the past (for example in developing countries), and increasing community involvement in air quality monitoring through the provision of sensors and the resulting data to community volunteers to support more informed public decision-making and engagement in air quality issues (Snyder et al., 2013; Loh et al., 2017; Turner et al., 2017). Several pilot programs of low-cost

15   sensor network deployment have been attempted, in Cambridge, UK (Mead et al., 2013), Imperial Valley, California (Sadighi et al., 2018; English et al., 2017), and Pittsburgh, Pennsylvania (Zimmerman et al., 2018).

There are several trade-offs resulting from the use of low-cost sensors. These sensors are less precise and sensitive than regulatory-grade instruments at typical ambient concentrations due to cross-sensitivities to other pollutants and dependence of the sensor response to ambient temperature and humidity (Popoola et al., 2016). These interactions are often nonlinear,

20   meaning that linear regression models developed under controlled laboratory conditions are often insufficient to accurately translate the raw sensor responses into concentration measures (Castell et al., 2017). Due to the variety of interactions and atmospheric conditions which can affect sensor performance, covering the range of conditions to which the sensor will be exposed using laboratory calibrations is difficult. Field calibrations of the sensors are thus necessary, with the sensors being collocated with highly accurate regulatory-grade instruments. Various calibration methods that have been explored include the

25   determination of sensor calibrations from physical and chemical principles (Masson et al., 2015), higher-dimensional models to capture nonlinear interactions (Cross et al., 2017), and nonparametric approaches including artificial neural networks (Spinelle et al., 2015) and k-nearest-neighbors (Hagan et al., 2018). Recent work by our group compared lab-based linear calibration models with multiple linear regression and non-parametric random forest algorithms based on ambient collocations (Zimmerman et al. 2018). The machine learning algorithm using random forests on ambient collocation data enabled low-cost

30   electrochemical sensor measurements to meet EPA data quality guidelines for hot spot detection and personal exposure for NO_2 and supplemental monitoring for CO and ozone (Zimmerman et al., 2018).

There remain several unanswered questions with respect to the calibration of data collected by low-cost sensors which we seek to answer in this work by examining **data collected by almost seventy Real-time Affordable Multi-Pollutant (RAMP) monitors over periods ranging up to eighteen months** in the city of Pittsburgh, PA, USA. First, although various models

have been applied to perform calibrations in different contexts, a thorough comparison on a common set of data of several different forms of calibration models applied to multi-pollutant measurements has yet to be performed. We seek to provide such a comparison and thereby draw robust conclusions about which calibration approaches work best overall and in specific contexts. Second, in previous work with the RAMP monitors and in work with other sensors, unique models have been developed for each sensor. This requires that extensive collocation data be collected for each low-cost sensor, which may not be feasible if large sensor networks are to be deployed. Therefore, it is important to investigate how well a single generalized calibration model can perform when applied across different individual sensors. Third, it is important to quantify the generalizability of models calibrated using data collected at a specific location to other locations across the same city where the sensors might be deployed, which may not share the same ratios of pollutants. This question is examined with several RAMPs that are co-located with regulatory monitors in the city of Pittsburgh, PA, USA. Finally, we seek to address the stability of calibration models over time by tracking changes in performance over the course of a year, and from one year to the next. Overall, we find support for using a generalized model for a network of RAMPs, developed based on local collocation of a subset of RAMPs. This reduces the need to collocate each node of a network, which otherwise can significantly increase network operating costs. These results will help guide future deployment efforts for RAMP or similar lower-cost air quality monitors.

2 Methods

2.1 The RAMP Monitor

The RAMP monitor (Fig. 1) was jointly developed by the Center for Atmospheric Particle Studies at Carnegie Mellon University (CMU) and a private company, SenSevere (Pittsburgh, PA). The RAMP package combines a power supply, control circuitry, cellular network communications capability, a memory card for data storage, and up to five gas sensors in a weatherproof enclosure. All RAMPs incorporate a nondispersive infrared (NDIR) CO_2 sensor produced by SST Sensing (UK), which also measures temperature and relative humidity. All RAMPs have one sensor that measures CO and one sensor that measures NO_2. Of the remaining sensors, one is either an SO_2 or NO sensor, and the other measures either a combination of oxidants (referred to hereafter as an Ozone or O_3 sensor, since this is its primary function in the RAMP) or Volatile Organic Compounds (VOCs). The VOC sensor is an Alphasense (UK) PID and all other unspecified sensors are Alphasense B4 electrochemical units. Specially designed signal processing circuitry ensures relatively low noise from the electrochemical sensors. Further details of the RAMP are provided elsewhere (Zimmerman et al., 2018). **Data collected from a total of 68 RAMP monitors are considered in this work.**

2.2 Calibration Data Collection

Following Zimmerman et al. (2018), RAMP monitors are deployed outdoors on a parking lot located on the CMU campus for a calibration based on collocated monitoring with regulatory-grade instruments. The parking lot (40°26'31"N by 79°56'33"W)

3

is a narrow strip between a low-rise academic building to the south and several tennis courts to the north. RAMP monitors are deployed for one month or more to allow for exposure to a wide range of environmental conditions; in 2017, these deployments took place in the summer and fall. Less than 10 meters from the RAMP monitors, a suite of high-quality regulatory-grade instruments, measuring ambient concentrations of CO (with a Teledyne T300U instrument), CO_2 (LICOR 820), O_3 (Teledyne

5   T400 Photometric Ozone Analyser), and NO and NO_2 (2B Technologies Model 405nm) are stationed to provide true concentration values for these various gases to which the RAMP monitors are exposed. These regulatory-grade instruments are contained within a mobile laboratory van, into which samples are drawn through an inlet 2.5 meters above ground level. Using sensor signal data collected by the RAMPs during this collocation period together with data collected by these regulatory-grade instruments, calibration models are created for each RAMP monitor prior to its deployment, as described in

10   Sect. 2.3. Further details on the regulatory-grade instrumentation and the collocation process are provided in previous work (Zimmerman et al., 2018).

In addition to collocation at the CMU campus, additional special collocation deployments of RAMP monitors were performed, in order to allow independent comparisons between the RAMP monitor data and regulatory monitors at different locations. One RAMP monitor was collocated with ACHD regulatory monitors at their Lawrenceville site (40°27'56"N by 79°57'39"W),

15   an urban background site where all NAAQS criteria pollutant concentrations are measured. The ACHD Parkway East site, located alongside the I-376 highway (40°26'15"N by 79°51'49"W), was chosen as an additional collocation site for observing higher levels of NO and NO_2: up to ~100 ppb for NO and ~40 ppb for NO_2. For reference, the NAAQS limit for one-hour maximum NO_2 is 100 ppb (https://www.epa.gov/criteria-air-pollutants/naaqs-table).

2.3 Gas Sensor Calibration Models

20   Various computational models were applied to the sensor readings of the RAMPs (i.e. the net signal, or raw response minus reference signal, from each electrochemical gas sensor, together with the outputs of the CO_2, temperature, and humidity sensor) to estimate gas concentrations, based entirely on ambient collocations of the RAMPs with regulatory-grade monitors. These models, outlined in the following subsections, include parametric models such as linear and quadratic regression models, a semi-parametric Gaussian process regression model, and non-parametric nearest-neighbor clustering, artificial neural network,

25   and hybrid random forest/linear regression models. A common difficulty of non-parametric methods is generalizing beyond the training data set. For example, if no high concentrations are observed during the collocation period, then the resulting trained nonparametric model will be unable to estimate such high concentrations if it is exposed to these during deployment. This is of potential concern for air quality applications, as the detection of high concentrations is an important consideration. Parametric models avoid this difficulty, but at the cost of lower flexibility in the types of input-output relationships they can

30   capture.

Models using each of these algorithms were calibrated in three separate categories. First, **individualized RAMP calibration models (iRAMP)** were created for each RAMP, using only the data collected by gas sensors in that RAMP and the regulatory monitors. Individualized models are applied only to data from the RAMP on which they were trained. Second, from these

individualized models, a **best individual calibration model (bRAMP)** was chosen, which performed best out of all the individualized models on a testing data set with respect to correlation (Pearson r, see Section 2.4). This model was then used to correct data from all other RAMPs which shared the same mix of gas sensors (to ensure that the inputs to the model would be consistent). Third, **general calibration models (gRAMP)** were developed by taking the median of the data from a subset

5   of the RAMP monitors deployed at the same place and time and treating this as a virtual "typical RAMP", for which models were calibrated for each gas sensor (the median is used rather than the mean to reduce the effects of any erroneous measurements by a few gas sensors in some RAMP monitors on the "typical" signal). The motivation for the use of gRAMP models is similar to that of Smith el at. (2017); however, while in that work it is recommended that the median from a set of duplicate low-cost sensors be used to improve performance, in this work we use that method to develop the gRAMP calibration

10   model, but then apply this calibration to the outputs of individual sensors rather than to the median of a group of sensors. RAMPs were divided into training and testing sets for the gRAMP models randomly, with the caveat that the two RAMPs deployed to the Lawrenceville and Parkway East sites were required to be part of the testing set. Data from about three quarters of the RAMP monitors (53 out of 68) were used for developing the general calibration models (although not all of these monitors were active at the same time). Data from the remaining 15 RAMP monitors were used for testing, ensuring that the

15   testing data are completely distinct from the training data. For the gRAMP models, the set of possible model inputs was restricted to ensure that, for each gas, all necessary model inputs would be provided by every RAMP (e.g. for NO models, only CO, NO, NO_2, T, and RH could be used as inputs since all RAMP monitors measuring NO would also measure these, but not necessarily any of the other gases). Thus, each of the calibration model algorithms were applied in three categories, yielding iRAMP, bRAMP, and gRAMP variants of each model. Finally, note that, for brevity, we will refer to iRAMP, bRAMP, or

20   gRAMP model variants when discussing specific results; however, when drawing general conclusions about low-cost electrochemical gas sensor calibration methods, we will use less RAMP-specific terms (such as "generalized models").

In all cases, models were calibrated using training data, which consists of the RAMP monitor data collected during the collocation period (which are measurements of the input variables, i.e. the signals from the various gas sensors) together with the readings of the regulatory-grade instruments with which the RAMP monitor was collocated (which are the targets for the

25   output variables). These collocation data are down-averaged from their original sampling rates to 15-minute averages, to ensure stability of the trained models and minimize the effects of noise on the training process. From the collocation data, eight equally sized, equally spaced time intervals are selected to serve as training data for the calibration models. The amount of training data is selected to be either 80% of the collocation data or four weeks of data (corresponding to 2688 15-minute-averaged data points), whichever is smaller. The minimum amount of training data is 21 days; if less than this is available, no iRAMP model

30   is trained for this RAMP, and thus no iRAMP model performance can be assessed for it (although bRAMP and gRAMP models trained on other RAMPs are still applied to this RAMP for testing). Training data for gRAMP models are obtained in the same way, although in that case it is the data for the virtual "typical RAMP" which are divided, rather than data for individual RAMPs. Any remaining data from the collocation period are left aside as a separate testing set, on which the performance of the trained models is evaluated. Note that due to differences in which RAMPs and/or regulatory-grade instruments were

operating at a given time, training and testing periods are not necessarily the same for all RAMPs and gases; for example, a certain time may be part of the training period for the CO model for one RAMP, and be part of the testing period for the O₃ model of another RAMP. However, the training and testing periods for a given RAMP and gas are always distinct. The division of data collected at the CMU site in 2017 into training and testing periods is illustrated in the supplemental information (Figures S6-S10). The division of data collected at the CMU site in 2016 is carried out in a similar manner. The choice of averaging period, of minimum and maximum training times, and the method for dividing between training and testing periods are motivated by previous work with the RAMP monitors (Zimmerman et al., 2018). All data collected at sites other than the CMU site (i.e. the Lawrenceville or Parkway East sites) are reserved for testing; no training of calibration models is done using data collected at these other sites, and so they represent a true test of the performance of the models at an "unseen" location.

2.3.1 Linear and Quadratic Regression Models

Linear regression models represent perhaps the simplest and most common method for gas sensor calibration, and have been used extensively in prior work (Spinelle et al., 2013, 2015; Zimmerman et al., 2018). A linear regression model (sometimes called a multi-linear regression model in the case that there are multiple inputs) describes the output as an affine function of the inputs. Here, linear functions are used where the sets of inputs are restricted to the signal of the sensor for the gas in question along with temperature and relative humidity. For example, the calibrated measurement of CO from the RAMP, c_{CO}, is an affine function of the signal of the CO sensor, s_{CO}, and the temperature T and relative humidity RH measured by the RAMP:

$$c_{\text{CO}} = \alpha_{\text{CO}} s_{\text{CO}} + \alpha_{\text{T}} T + \alpha_{\text{RH}} RH + \beta_{CO}, \tag{1}$$

Coefficients α_{CO}, α_{T}, and α_{RH} and offset term β_{CO} are calibrated from training data to minimize the root-mean-square difference of c_{CO} and the measured CO concentration from the regulatory-grade instrument. The one exception to this general formulation is for evaluation of c_{O_3}, where both s_{O_3} and s_{NO_2} are used as inputs (along with T and RH); this is done to account for the fact that the sensor for O₃ also responds to NO₂ concentrations (Afshar-Mohajer et al., 2018).

In addition to linear regressions, quadratic regressions were also applied. These are the same as linear regressions but can involve second-order interactions of the input variables. For example, for CO, a quadratic regression function would be of the following form:

$$c_{\text{CO}} = \alpha_{\text{CO}} s_{\text{CO}} + \alpha_{\text{CO}^2} s_{\text{CO}}{}^2 + \alpha_{\text{T}} T + \alpha_{\text{T}^2} T^2 + \alpha_{\text{RH}} RH + \alpha_{\text{RH}^2} RH^2 + \alpha_{\text{CO,T}} s_{\text{CO}} T + \alpha_{\text{CO,RH}} s_{\text{CO}} RH + \alpha_{\text{T,RH}} T\, RH + \beta_{CO}, \tag{2}$$

Note that, as above, a reduced set of inputs is used here. Quadratic regression models using such reduced sets (the same sets used for linear regression) are hereafter referred to as "limited" quadratic regression models; in contrast, models making full use of all available gas, temperature, and humidity sensor inputs from a given RAMP are referred to as "complete" quadratic regression models.

The main advantages of linear and quadratic regression models are their ease of implementation and calibration, as well as their ability to be readily interpreted, e.g., the relative magnitudes of the regression coefficients correspond to the relative

importance of the different inputs in producing the output. The main disadvantage of these models is their inability to compute complicated relationships between input and output which are beyond that of a second-order polynomial. The training and application of linear and quadratic regression models are implemented using custom-written routines for the MATLAB programming language (version R2016b).

2.3.2 Gaussian Process Models

Gaussian processes are a form of regression which generalizes the multivariate Gaussian distribution to infinite dimensionality (Rasmussen and Williams, 2006). For the purposes of calibration, we make use of a simplified variant of a Gaussian process model. From the training data, both the signals of the RAMP monitors and the readings of the regulatory-grade instruments are transformed such that their distributions during the training period can be approximately modelled as standard normal distributions. This transformation is accomplished by means of a piecewise linear transformation, where the domain is segmented and for each segment different linear mappings are applied. After this transformation, an empirical mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is computed for the regulatory-grade and RAMP measurements. The transformed measurements can then be described using a multivariate Gaussian distribution. For example, for a RAMP measuring CO, SO_2, NO_2, O_3, and CO_2, this distribution would be:

$$\left\{c'_{CO}, c'_{SO_2}, c'_{NO_2}, c'_{O_3}, c'_{CO_2}, s'_{CO}, s'_{SO_2}, s'_{NO_2}, s'_{O_3}, s'_{CO_2}, T', RH'\right\}^{\mathrm{T}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

where, for example, c'_{CO} represents the concentration measurement for CO following the transformation. The mean vector and covariance matrix are divided as follows:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{conc} \\ \boldsymbol{\mu}_{RAMP} \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{conc,conc} & \boldsymbol{\Sigma}_{conc,RAMP} \\ \boldsymbol{\Sigma}^{\mathrm{T}}_{conc,RAMP} & \boldsymbol{\Sigma}_{RAMP,RAMP} \end{bmatrix}, \tag{4}$$

where $\boldsymbol{\mu}_{conc}$ represents the mean of the (transformed) concentration measurements of the regulatory-grade instrument, $\boldsymbol{\mu}_{RAMP}$ represents the mean of the (transformed) signal measurements from the RAMP, $\boldsymbol{\Sigma}_{conc,conc}$ represents the covariance of the (transformed) concentrations, $\boldsymbol{\Sigma}_{RAMP,RAMP}$ represents and covariance of the (transformed) RAMP signals, and $\boldsymbol{\Sigma}_{conc,RAMP}$ represents the covariance between the (transformed) concentrations and RAMP signals ($\boldsymbol{\Sigma}^{\mathrm{T}}_{conc,RAMP}$ is the transpose of $\boldsymbol{\Sigma}_{conc,RAMP}$). Once these vectors and matrices have been defined, the model is calibrated.

Given a new set of signal measurements from a RAMP, denoted as $\mathbf{y}_{RAMP} = \left\{s_{CO}, s_{SO_2}, s_{NO_2}, s_{O_3}, s_{CO_2}, T, RH\right\}^{\mathrm{T}}$, these are transformed using the piecewise linear transformation defined above to give the set of transformed signal measures $\mathbf{y}'_{RAMP}$. These are then used to estimate the concentrations measured by the RAMP with the standard conditional updating formula of the multivariate Gaussian as follows:

$$\left\{c'_{CO}, c'_{SO_2}, c'_{NO_2}, c'_{O_3}, c'_{CO_2}\right\}^{\mathrm{T}} = \boldsymbol{\mu}_{conc} + \boldsymbol{\Sigma}_{conc,RAMP}\boldsymbol{\Sigma}^{-1}_{RAMP,RAMP}(\mathbf{y}'_{RAMP} - \boldsymbol{\mu}_{RAMP}), \tag{5}$$

The inverse of the original piecewise linear transformation is then applied to these transformed concentration estimates to yield the appropriate concentration estimates in their original units.

The main advantage of a Gaussian process calibration model of this form is its robustness to incomplete or inaccurate information; for example, if a signal from one gas sensor were missing or corrupted by a large voltage spike, in the former case the missing input could be "filled in" by the correlated measurements of other sensors, while in the latter case estimates would be "reigned in" by the more reasonable measures of the other sensors. A major disadvantage of this calibration model

5   is its continued use of what is basically a linear regression formula; the only difference being in the non-linear transformation from the original measurement space to the standard normal variable space used by the model. Furthermore, during the calibration process, the ratios of concentration for the pollutants of the collocation site may be "learned" by the model, making it less likely to predict differing ratios during field deployment. The training and application of Gaussian process calibration models are accomplished using custom-written routines in the MATLAB programming language.

10   ### 2.3.3 Clustering Model

The clustering model presented here seeks to estimate the outputs corresponding to new inputs by searching for input-output pairs in the training data for which the distance (by a predefined distance metric in a potentially high-dimensional space) between the new input and the training inputs is minimized, and using the average of several outputs corresponding to these nearby inputs (the "nearest neighbors"). In a traditional k-nearest-neighbors approach, such as that used in previous work

15   (Hagan et al., 2018), every input-output pair from the training data is stored for comparison to new inputs. Although this provides the best possible estimation performance via this approach, storing these data and performing these comparisons are computation- and memory-intensive. Therefore, in this work, the input data are first clustered, i.e., grouped by proximity of the input data. These clusters are then represented by their centroid, with the corresponding output being the mean of the outputs from the clustered inputs. In this work, training data are grouped into one thousand clusters using the 'kmeans' function

20   in MATLAB. Euclidian distance in the multidimensional space of the sensor signals from each RAMP is used. For estimation, the outputs of the five nearest neighbors to a new input are averaged.

A major advantage of this approach is its simplicity and flexibility, allowing it to capture complicated nonlinear input-output relationships by referring to past records of these relationships, rather than attempting to determine the actual pattern which these relationships follow. Such a method can perform very well when the relationships are stable, and when any new input

25   with which the model is presented is similar to at least one of the inputs from the training period. However, as with all nonparametric models, generalizing beyond the training period is difficult, and the model will tend to perform poorly if the "nearest neighbors" of a new input are in fact quite far away, in terms of the distance metric used, from this input.

2.3.4 Artificial Neural Network Model

The artificial neural network model, or simply neural network, is a machine learning paradigm which seeks to replicate, in a

30   simplified manner, the functioning of an animal brain in order to perform tasks in pattern recognition and classification (Aleksander and Morton, 1995). A basic neural network consists of several successive layers of "neurons". These neurons each receive a weighted combination of inputs from a higher layer (or the signal inputs, if they are in the top layer) and apply

a simple but nonlinear function to them, producing a single output which is then fed on into the next layer. By including a variety of possible functions performed by the neurons and appropriately tuning the weights applied to inputs fed from one layer to the next, highly complicated nonlinear transformations can be performed in successive small steps.

Neural networks have been applied to a large number of problems, including the calibration of low-cost gas sensors (Spinelle et al., 2015). Neural networks represent an extremely versatile framework, and are able to capture nearly any nonlinear input-output relationship (Hornik, 1991). Unfortunately, to do so may require vast amounts of training data, which it is not always practical to obtain. Calibration of these models is also a time-consuming process, requiring many iterations to tune the weightings applied to values passed from one layer to the next. In this work, neural networks were trained and applied using the 'Netlab' toolbox for MATLAB (Nabney, 2002). The network has a single hidden layer with twenty nodes. To limit the computation time needed for model training, the number of allowable iterations of the training algorithm was capped at ten thousand; this cap was typically reached during the training.

2.3.5 Hybrid Random Forest and Linear Regression Models

A random forest model is a machine learning method which makes use of a large number of decision "trees". These trees are hierarchical sets of rules which group input variables based on thresholding (e.g. "the third input variable is above or below a given value"). The thresholds used for these rules as well as the inputs they are applied to and the order in which they are applied are calibrated during training. The final groupings of input variables from the training data, located at the end or "leaves" of the branching decision tree, are then associated with the mean values of the output variables for this group (similar to a clustering model). For estimating an output given a new set of inputs, each decision tree within the random forest applies its sequence of rules to assign the new data to a specific "leaf", and outputs the value associated with that leaf. The output of the random forest is the average of the outputs of each of its trees.

A primary shortcoming of the random forest model (which it shares with other nonparametric methods) is its inability to generalize beyond the range of the training data set, i.e., outputs of a random forest model for new data can only be within the range of the values included as part of the training data. For this reason, the standard random forest model was expanded into a hybrid random forest/linear regression model. The use of this approach for RAMP data was suggested by Zimmerman et al. (2018). Furthermore, it is similar to the approach of Hagan et al. (2018), who hybridize nearest neighbor and linear regression models. In this modified model, a random forest is applied to new data to estimate the concentrations of various measured pollutants. For example, the concentration of CO measured by a RAMP including sensors for CO, SO_2, NO_2, O_3, and CO_2 is estimated using a random forest as:

$$c_{CO} = \mathrm{RF}_{CO}\big(s_{CO}, s_{SO_2}, s_{NO_2}, s_{O_3}, s_{CO_2}, T, RH\big), \tag{6}$$

If this estimated concentration exceeds a given value (in this case, 90% of the maximum concentration value observed during the training, corresponding to about 1ppm in the case of CO), a linear model of the form of Eq. (1) is instead used to estimate the concentration. This linear model is calibrated using a 15% subset of the training data with the highest concentrations of the target gas and is therefore better able to extrapolate beyond the upper concentration value observed during the training period.

This hybrid model therefore is designed to combine the strengths of the random forest model, i.e. its ability to capture complicated nonlinear relationships between various inputs and the target output, with the ability of a simple linear model to extrapolate beyond the set of data on which the model is trained. Random forests are implemented using the 'TreeBagger' function in MATLAB, and custom routines are used to implement hybrid models.

2.4 Assessment Metrics

In the following section, the performance of the calibration models in translating sensor signals to concentration estimates is assessed in several ways. It should be noted that the metrics presented here are applied only for testing data, i.e., data which were not used to build the calibration models. Model performance on the training data is expected to be higher, and thus less representative of the true capability of the model. The estimation bias is assessed as the mean normalized bias (MNB), the average difference between the estimated and actual values, divided by the mean of the actual values. That is, for n measurements:

$$\text{MNB} = \frac{\sum_{i=1}^{n}(c_{\text{estimated},i} - c_{\text{true},i})}{\sum_{i=1}^{n}(c_{\text{true},i})}, \tag{7}$$

where $c_{\text{estimated},i}$ is the measured concentration as estimated by the RAMP monitor and $c_{\text{true},i}$ is the corresponding true value measured by a regulatory-grade instrument. The variance of the estimation is assessed via the coefficient of variation of the mean absolute error (CvMAE), the average of the absolute differences between the estimated and actual values divided by the mean of the actual values. The estimates used in evaluating the CvMAE are corrected for any bias as determined above:

$$\text{CvMAE} = \frac{\sum_{i=1}^{n}|c_{\text{estimated},i} - n_{\text{bias}} - c_{\text{true},i}|}{\sum_{i=1}^{n}(c_{\text{true},i})}, \tag{8}$$

where:

$$n_{\text{bias}} = \frac{1}{n}\sum_{i=1}^{n}(c_{\text{estimated},i} - c_{\text{true},i}), \tag{9}$$

Correlation between estimated and actual concentrations is assessed using the Pearson linear correlation coefficient (r):

$$r = \frac{\sum_{i=1}^{n}\left(c_{\text{estimated},i} - \frac{1}{n}\sum_{j=1}^{n}c_{\text{estimated},j}\right)\left(c_{\text{true},i} - \frac{1}{n}\sum_{j=1}^{n}c_{\text{true},j}\right)}{\sqrt{\sum_{i=1}^{n}\left(c_{\text{estimated},i} - \frac{1}{n}\sum_{j=1}^{n}c_{\text{estimated},j}\right)^2}\sqrt{\sum_{i=1}^{n}\left(c_{\text{true},i} - \frac{1}{n}\sum_{j=1}^{n}c_{\text{true},j}\right)^2}}, \tag{10}$$

Intuitively, these basic metrics are used to quantify the difference in averages between estimated and true concentrations (MNB), the average of differences between these (CvMAE), and the similarity in their behavior (r).

In addition to the above metrics, EPA methods for evaluating precision and bias errors are used as outlined in Camalier et al. (2007). To summarize, the precision error is evaluated as:

$$\text{Precision} = \sqrt{\frac{n\sum_{i=1}^{n}\delta_i^2 - \left(\sum_{i=1}^{n}\delta_i\right)^2}{n\chi^2_{0.1,n-1}}}, \tag{11}$$

where $\chi^2_{0.1,n-1}$ denotes the 10th percentile of the chi-squared distribution with $n-1$ degrees of freedom and the percent difference in the ith measurement is evaluated as:

$$\delta_i = \frac{c_{\text{estimated},i} - c_{\text{true},i}}{c_{\text{true},i}} \cdot 100, \tag{12}$$

The bias error is computed as:

$$\text{Bias} = \frac{1}{n}\sum_{i=1}^{n}|\delta_i| + \frac{t_{0.95,n-1}}{n}\sqrt{\frac{n\sum_{i=1}^{n}\delta_i^2 - \left(\sum_{i=1}^{n}|\delta_i|\right)^2}{n-1}}, \tag{13}$$

where $t_{0.95,n-1}$ is the 95th percentile of the t distribution with $n-1$ degrees of freedom. Prior to the computation of these

5 precision and bias metrics, measurements where the corresponding true value is below an assigned lower limit are removed from the measurement set to be evaluated, so as not to allow near-zero denominator values in Eq. (12). Lower limits used in this work are based on the guidelines presented by Williams et al. (2014) and are listed in Table 1. Note that this removal of low values is applied only when computing the precision and bias error metrics, and not when evaluating the other metrics described above.

10 Using the EPA precision and bias calculations allows for these values to be compared against performance guidelines for various sensing applications, as presented Williams et al. (2014) and listed in Table 2. For the RAMP monitors, a primary goal is to achieve data quality sufficient for hotspot identification and characterization (Tier II) or personal exposure monitoring (Tier IV), which requires that both precision and error bias metrics be below 30%. A supplemental goal is to achieve performance sufficient for supplemental monitoring (Tier III), requiring precision and bias metrics below 20%.

15 **3 Results**

In this section, we examine the performance of the RAMP gas sensors and the various calibration models applied to their data. We will focus our attention on the CO, NO, NO_2, and O_3 sensors. Results for calibration of measurements by the CO_2 sensors are presented in supplemental figures.

3.1 Performance across Individualized Models on CMU Site Collocation Data

20 Figure 2 presents a comparison of the performance of various calibration models applied to testing data collected at the CMU site during 2017. As described in Sect. 2.3, collocation data are divided into training and testing sets, with the former (always being between three and four weeks in total duration) used for model development and the latter used to test the developed model using the assessment metrics described in Sect. 2.4, as presented in Fig. 2. All models in the figure are of the "iRAMP" category, being developed using only data collected by a single RAMP and the collocated regulatory-grade instruments. In the

25 figure, squares indicate the median performance across all RAMPs for each performance metric, and the error bars span from the 25th to 75th percentiles of each metric across the RAMPs. **For CO, 48 iRAMP models are compared; for NO, 19 models; for NO₂, 62 models; for O₃, 44 models**. Note that only 20 RAMP monitors included an NO sensor. An iRAMP model was not developed for RAMP monitors that had fewer than 21 days of collocation data with the relevant regulatory-grade instrument. The figures are arranged such that the lower-left corner denotes "better" performance (CvMAE close to 0 and r

30 close to 1).

11

Typically, several of the model types provide similar performance for a given gas. For CO and O_3, the simple parametric quadratic regression models perform as well as or better than the non-parametric modelling approaches, and even linear regression models give reasonable results. For NO_2 and NO, while the non-parametric hybrid or neural network models perform best, complete quadratic regression models give comparable performance. Quadratic regression and hybrid models give the most consistently good performance, being among the top four methods across all gases. Bias tends to be low to moderate (depending on the gas) regardless of correction method (MNB less than 1% for CO, less than 2% for O_3, less than 10% for NO, and less than 20% for NO_2 across all methods). Table 2 lists the EPA performance guidelines for various applications, and Table 3 lists the modelling methods which meet these based on performance at the CMU site in 2017. All methods meet at least Tier I (educational monitoring, <50% error) criteria for all gases considered. Most methods fall within the Tier II (hotspot detection) or Tier IV (personal exposure) performance levels (<30% error) for all gases. For CO, quadratic regression methods meet Tier III (supplemental monitoring) criteria (<20% error). In Table 4, the durations of the training and testing periods and the measured concentration ranges during these periods are provided. Finally, in Table 5, additional metrics are presented about these performance results, including un-normalized MAE and bias in the measured concentration units, to allow for direct comparison with the concentrations ranges. More detailed information is also provided in the supplemental information.

3.2 Comparison of Individualized, Best, and General Models on CMU Site Collocation Data

Next, we examine how the performance of the best individual models (bRAMP) and of the general models (gRAMP) applied to all RAMPs compare to the performance of the individualized RAMP (iRAMP) models presented in Sect. 3.1. Evaluation is carried out on the testing data collected at the CMU site in 2017. For simplicity, we restrict ourselves to three models for each gas, chosen from among the better-performing iRAMP models and including at least one parametric and one non-parametric approach. Figure 3 presents these comparisons.

Across all gases and models, iRAMP models tend to perform best, as might be expected since these models are both trained and applied to data collected by a single RAMP monitor, and therefore will account for any peculiarities of individual sensors. Between the bRAMP models, in which a model is trained using data from a single RAMP and applied across multiple RAMPs, and gRAMP models, which are trained on data from a virtual "typical" RAMP (composed of the median signal from several RAMPs) and then applied across other RAMPs, it is difficult to say which approach would be better based on these results, as they vary by gas as well as by modelling approach. For parametric models (i.e., linear and quadratic regression) the bRAMP and gRAMP versions typically have similar performance, although there is less variability in performance for the gRAMP versions. For non-parametric models (i.e., neural network and hybrid models), performance of bRAMP versions is typically better than the gRAMP versions, although in the case of NO_2 and O_3 the performance is comparable. Overall, we find that a bRAMP or gRAMP version of several of the models can give similar performance to its iRAMP version, even though these models are not calibrated to each individual RAMP.

3.3 Performance of Selected Models at Regulatory Monitoring Sites

Figure 4 depicts the performance of calibration models for two RAMP monitors deployed at two EPA monitoring stations operated by the ACHD (one monitor is deployed to each station). Filled markers indicate the performance of the models at these sites, while hollow markers indicate the 2017 testing period performance of the corresponding RAMP when it was at the

5   CMU site for comparison. For each gas type, different calibration models are used, chosen from among the models depicted in Fig. 3. Models trained at the CMU site (as presented in previous sections) are used to correct data collected by the RAMP monitor at the station. Note that all data collected at either deployment site are treated as testing data, and that no data from these other sites are used to calibrate the models. Also note that not all gases monitored by RAMPs are monitored by the stations, hence why only one station may appear in each plot.

10   Overall, there tends to be a change in model performance at either of the deployment sites as compared to the CMU site. This is to be expected to some degree, as the concentration range and mixture of gases (especially at the Parkway East site, which is located next to a major highway) can be different at a new site (where the model was not trained), and thus cross-sensitivities of the sensors may be affected. These differences appear to be greatest for CO, with performance being *better* at the Parkway East site, where overall CO concentrations are higher (both the average and standard deviation of the CO concentration at the

15   Parkway East site are more than double those of the CMU or Lawrenceville sites). Additionally, gRAMP models tend to perform as well as or better than iRAMP models when monitors are deployed to new sites (only the CO results at Parkway East are much better for the iRAMP than the gRAMP models). Furthermore, the performance of the gRAMP models at the training site is typically more representative of the expected performance at other sites than that of the iRAMP models. This is likely because, while the iRAMP models are trained for individual RAMPs at the training site, the gRAMP models are

20   trained across multiple RAMPs at that site, and therefore are more robust to a range of different responses for the same atmospheric conditions. Thus, when a RAMP is moved from one site to another, and its responses change slightly due to a change in the surrounding conditions, the gRAMP model will be more robust against these changes. Based on these results, since the change in performance as a monitor is deployed to a field site is often greater than the gap between iRAMP and gRAMP performance at the calibration site (as assessed in Sect. 3.2), there is no reason to prefer an iRAMP model to a gRAMP

25   model for correction of field data.

To evaluate the performance of these sensors in a different way, EPA-style precision and bias metrics are provided in Fig. 5. Only CO, O_3, and NO_2 are considered, as these are the gases for which performance guidelines have been suggested by the EPA (Table 2). These guidelines are indicated by the dotted boxes in the figure; points falling within the box meet the criteria for the corresponding tier. Also, the range in observed performance at the CMU site, as depicted in Fig. 3, is reproduced here

30   for comparison using black markers with error bars. For CO, by these criteria, the gRAMP model outperforms the iRAMP model, with the gRAMP model meeting Tier II or IV criteria (<30% error) for all locations. Thus, under these metrics, for CO the gRAMP model is more representative of performance at other sites, while for the iRAMP model performance is more varied between sites. For O_3, performance of both models at Lawrenceville is better than assessed at the CMU site, and both

models fall near the boundary between Tiers II/IV and Tier III performance criteria (about 20% error). For NO$_2$, performance at the deployment sites in terms of the bias is always worse than predicted by the CMU performance, although in terms of precision, the gRAMP model at the CMU site better represents the site performance than the iRAMP model (the same trend as was seen using the other metrics).

3.4 Performance of Calibration Models over Time

We now examine the change in performance of calibration models over time. Figure 6 shows the performance of models developed based on data collected at the CMU site in both 2016 and 2017 and tested on data collected in either of these years. Training and testing data for 2017 represent the same training and testing periods as used for previous results. For 2016, training and testing data are divided using the same procedure as was applied for 2017 data, as discussed in Sect. 2.3. For example, the results for "2016 Data, 2017 Models" represent the performance of models calibrated using the training data subset of the 2017 CMU site data when applied to the testing data subset of the 2016 CMU site data. A change in performance between these two models on data from the same year will indicate the degree to which the models have changed from one year to the next; likewise, a change in performance for the same model applied to data from different years will indicate the degree to which sensor responses have changed over time. Note that NO is omitted here because data to build calibration models for this gas were not collected in 2016. Also note that results presented in the rest of this paper only use data collected in 2017 for model training and evaluation.

A drop in performance when models from one year are applied to data collected in the next year is consistently observed for all models and gases, with O$_3$ having the smallest variability from one year to the next. This suggests that degradation is occurring in the sensors, reducing the intensity of their responses to the same ambient conditions and/or changing the relationships between their responses. Thus, a model calibrated on the response characteristics of the sensors in one year will not necessarily perform as well using data collected by the same sensors in a different year. This degradation has also been directly observed, as the raw responses of "old" sensors deployed with the RAMPs since 2016 were compared to those of "new" sensors recently purchased in 2018; in some cases, responses of "old" sensors were about half the amplitude of those of "new" sensors exposed to the same conditions. This is consistent with the operation of the electrochemical sensors, where the electrolyte concentration changes over time as part of the normal functioning of the sensor. To compensate for this, new models should be calibrated for sensors on at least an annual basis, to keep track with changes in signal response. Furthermore, calibration models should preferably be applied to data from sensors with a similar age to avoid effects due to different signal responses of sensors which have degraded to varying degrees. Finally, in comparing model performance from one year to the next, there is no significant increase in error associated with using gRAMP models (trained for sensors of a similar age) rather than iRAMP or bRAMP models.

Additionally, calibration model performance was assessed as a function of averaging time. Note that the calibration models discussed in this paper are developed using RAMP data averaged over 15-minute intervals, as discussed in Section 2.3. However, these models may be applied to raw RAMP signals averaged over longer or shorter time periods. Furthermore, the

14

calibrated data can also be averaged over different periods. To investigate the effects of averaging time on calibration model performance, we assess the performance of RAMPs calibrated with gRAMP models for CO, O_3, and CO_2 at the CMU site in 2017, with averaging performed either before or after the calibration. Results are provided in the supplemental materials (Figures S4 and S5). Overall, we find little variation in calibration model performance with respect to averaging periods

5   between 1 minute and 1 hour.

3.5 Changes in Field Performance over Time

Finally, we track the performance of RAMPs over time at specific deployment locations, as depicted in Fig. 7, to evaluate changes in calibration model field performance over time. This is done using three RAMP monitors; one was deployed at the ACHD Lawrenceville station from January through September of 2017, as well as during November and December 2017. A

10   second RAMP was kept at the CMU site year-round, where it was collocated with regulatory-grade instruments intermittently between May and October. The third RAMP was deployed at the ACHD Parkway East site beginning in November of 2017. The same gRAMP calibration models as depicted in Fig. 4 (using the training data collected at the CMU site in 2017) are used; note that the RAMP present at the CMU site was a part of the training set of RAMPs for the gRAMP model, while the other two RAMPs were not. Performance of CO, NO_2, and O_3 sensors are depicted, as both CO and NO_2 were continuously

15   monitored at both ACHD sites and intermittently monitored at the CMU site, and O_3 was consistently monitored at ACHD Lawrenceville as well as being intermittently monitored at the CMU site. Performance is assessed on a weekly basis. For CO, the limited quadratic regression gRAMP model is used, and for O_3, the hybrid gRAMP model is used, as these showed the least variability in performance of the gRAMP models in Fig. 3. For NO_2, a hybrid gRAMP model is used, which provided the same performance as the neural network gRAMP model in Fig. 3.

20   Performance of the calibrated O_3 measurements shows almost uniformly high correlation and low CvMAE throughout the year. For CO and (to a lesser degree) NO_2, while CvMAE is relatively consistent, periods of lower and more variable correlation occurred from July to September (for CO) or October (for NO_2). These periods of lower correlation do not appear to coincide with periods of atypical concentrations, nor with periods of excessive pollutant variability at the site, nor with any unusual pattern in the other factors measured by the RAMP. Periods of lower performance appear to roughly coincide for the

25   CMU and Lawrenceville sites for the time during which both sites were active, and observed pollutant concentration ranges were comparable for both the CMU and Lawrenceville sites during these periods. There does not appear to be a clear seasonal or temporal trend to this performance, as low correlations occur in the late summer but not early summer, when conditions were similar. Thus, while sensor performance is observed to fluctuate from week to week, there does not appear to be a seasonal degradation in performance.

4 Discussion and Conclusions

Based on the results presented in Sect. 3.1, complete quadratic regression and hybrid models give the best and most consistent performance across all gases. Of these, the hybrid models, combining the complicated non-polynomial behaviors of random forest models (capable of capturing unknown sensor cross-sensitivities) with the generalization performance of parametric linear models, tend to generalize best for NO, NO_2, and O_3 when applied to data collected at new sites. For CO, quadratic regression models generalize better. Neural networks perform well for NO and NO_2 but not for CO; limited quadratic regression models perform well for CO and O_3 but not for NO and NO_2. Linear regression, Gaussian processes and clustering are the worst overall models for these gases, never being in the top two best performing models, and only rarely being one of the top three. 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. The fact that for most gases a variety of calibration approaches show similar (and for typical uses cases, acceptable) performance may reflect better underlying performance from the RAMP monitor, as similar studies for other low-cost sensor packages showed a wider variability in performance between calibration approaches (see e.g. the summary provided by Zimmerman et al., 2018). This suggests that the primary difference between these monitors, i.e. the internal circuitry which is unique to the RAMP, is the cause for this consistency; however, determination of this is beyond the scope of this paper.

Overall, in most cases the generic bRAMP and generalized gRAMP calibration models perform worse than the individualized iRAMP models at the calibration site, but the decline in performance may be manageable and acceptable depending on the use case. For example, for NO_2 (Fig. 3), median performance of bRAMP and gRAMP neural network and hybrid models are only 15% worse in terms of CvMAE and 5% worse is terms of r. For O_3, median performance of all models is above 0.8 for r and below 0.25 for CvMAE, indicating a high level of correlation and relatively low estimation error. For CO, limited quadratic models meet the same criteria. Furthermore, in examining the generalization performance of the models when applied to new sites, as depicted in Figs. 4 and 5, for NO_2 (in terms of the r and CvMAE metrics) and for CO (in terms of the EPA precision and bias metrics), the gRAMP models show more consistent performance between the calibration and deployment locations than the iRAMP models. For O_3, performance of iRAMP and gRAMP models at the Lawrenceville site is comparable, while for NO, performance of the gRAMP model at the Parkway East site is actually better than the iRAMP model. This may indicate that the NO sensors are more affected by changes in ambient conditions than the other electrochemical sensors, and the gRAMP model is better able to average out these sensitivities than the other model categories considered.

Based on comparisons between the performance of models from one year to the next, as well as the analysis of changes in performance of the RAMP monitors collocated with regulatory-grade instruments for long periods, some sensors, such as the O_3 sensor, are quite stable over time. For the CO sensor, performance seemed variable over time, and performance noticeably degraded from one year to the next, although no seasonal trends were apparent, and overall performance may be acceptable (CvMAE < 0.5). The NO_2 sensor also exhibited some degradation from one year to the next, although performance was stable over time in 2017, with minimal changes in overall performance during this long deployment period.

16

It can generally be expected that RAMP monitors will at least meet Tier II or Tier IV EPA performance criteria (<30% error) for O_3 (with hybrid random forest/linear regression bRAMP and possibly gRAMP models) and CO (with limited quadratic regression gRAMP models). Individualized calibration models are more likely to meet Tier II/IV criteria for NO_2 (with iRAMP or bRAMP hybrid models), and localized calibration may also be required. For NO, while no specific target criteria are established, neural network iRAMP or gRAMP models appear to perform best.

4.1 Recommendations for Future Low-Cost Sensor Deployments

In comparing different methods for the calibration of electrochemical sensor data, it was found that in some cases, e.g. for CO, simple parametric models, such as quadratic functions of a limited subset of the available inputs, were sufficient to transform the signals to concentration estimates with a reasonable degree of accuracy. For other gases, e.g. NO_2, more sophisticated nonparametric models performed better, although parametric quadratic regression models making use of all sensor inputs were still among the best performing models for all gases. Depending on the application, therefore, different methods might be appropriate, e.g., using simpler parametric models such as quadratic regression to calibrate measurements and provide real-time estimates, while using more sophisticated non-parametric methods such as random forest models when performing long-term analysis for exposure studies. Of the non-parametric methods considered, hybrid random forest/linear regression models gave the best general performance across all the gas types. These models, along with the quadratic models, should therefore be considered for situations where it is desirable to use the same type of calibration across all gases, e.g. to reduce the "overhead" of programming multiple calibration approaches. The hybrid model, which combines the flexibility of the random forest with the generalizability of the linear model, is most theoretically promising for general application. Future work will also investigate other forms of hybrid models. For example, combinations of neural network and linear regression models may work well for NO, where neural networks provided better performance than hybrid models using random forests. Also, for CO and Ozone, hybrid models combining random forests with quadratic regression might perform better than those with linear models, since quadratic models perform better than linear models for these gases overall.

Although there is a reduction in performance as a result of not using individualized monitor calibration models when these are calibrated and tested at the same location, the use of a single calibration model across multiple monitors, representing either the best of available individualized models or a general model developed for a "typical" monitor, tends to give more consistent generalization performance when tested at a new site. This suggests that variability in the responses of individual sensors for the same gas when exposed to the same conditions (such as would be accounted for when developing separate calibration models for each monitor) tends to be lower than the variability in the response of a single sensor when exposed to different ambient environmental conditions and a different mixture of gases (such as is experienced when the monitor is moved to a new site). Models that are developed and/or applied across multiple monitors will avoid "overfitting" to the specific response characteristics of a single sensor in a single environment. Thus, considering that it is impractical to perform a collocation for each monitor at the location where it is to be deployed, there is little benefit to developing individualized calibration models

17

for each monitor when their performance will be similar to (if not worse than) that of a generalized model when the monitor is moved to another location.

There are several additional qualitative advantages to using generalized models. First, the effort required to calibrate models is reduced, since not every monitor needs to be present for collocation and separate models do not have to be created for every monitor. For example, while for CO only 48 RAMPs had sufficient data to calibrate individualized models based on data collected at the CMU site in 2017, general models can be calibrated and applied for all 68 RAMPs which were at the CMU site during this period, as well as for additional RAMPs which were never collocated at the CMU site but have the same gas sensors installed. Second, collocation data collected from multiple monitors at different sites can be combined in the creation of a generalized model, whereas individualized models would require each monitor to be present at each collocation site. This means that a wider range of ambient gas concentrations can be reflected in the training data, allowing for better generalization. Finally, the use of generalized models allows for robustness against noise of individual sensors, which can lead to mis-calibration of individualized models but is less likely to do so if data from multiple sensors are averaged. Therefore, for future deployments, generalized models applicable across all monitors should be used.

For long-term deployments, it is recommended that new models be developed each year, due to the noticeable change in performance when models for one year were used for processing data collected in the subsequent year. If generalized models are used, model development can be performed using only a representative subset of monitors collecting data across a range of temperature and humidity conditions, allowing most monitors to remain deployed in the field (although periodic "sanity checks" should be made for field-deployment monitors to ensure all on-board sensors are operating properly). Another option is to maintain a few "gold standard" monitors collocated with regulatory-grade instruments year-round and to use these monitors for the development of generalized models to be used with all field-deployed monitors over the same period. Determination of how many monitors are necessary to develop a sufficiently robust generalized model is a topic of ongoing work.

Data Availability

All data (reference monitor data, RAMP raw signal data, calibrated RAMP data for both training and testing), and codes (in MATLAB language) to recreate the results discussed here are provided online at https://doi.org/10.5281/zenodo.1302030 (Malings, 2018a). Additionally, an abridged version of the dataset (without the calibrated data or models, but still including the codes to generate these) is available at https://doi.org/10.5281/zenodo.1482011 (Malings, 2018b).

References

5 Afshar-Mohajer, N., Zuidema, C., Sousan, S., Hallett, L., Tatum, M., Rule, A. M., Thomas, G., Peters, T. M. and Koehler, K.: Evaluation of low-cost electro-chemical sensors for environmental monitoring of ozone, nitrogen dioxide, and carbon monoxide, Journal of Occupational and Environmental Hygiene, 15(2), 87–98, doi:10.1080/15459624.2017.1388918, 2018.

Aleksander, I. and Morton, H.: An introduction to neural computing, 2. ed., International Thomson Computer Press, London., 1995.

10 Camalier, L., Eberly, S., Miller, J. and Papp, M.: Guideline on the Meaning and the Use of Precision and Bias Data Required by 40 CFR Part 58 Appendix A, U.S. Environmental Protection Agency. [online] Available from: https://www3.epa.gov/ttn/amtic/files/ambient/monitorstrat/precursor/07workshopmeaning.pdf, 2007.

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?, Environment International, 99, 293–302, doi:10.1016/j.envint.2016.12.007, 2017.

15

Cross, E. S., Williams, L. R., Lewis, D. K., Magoon, G. R., Onasch, T. B., 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, Atmospheric Measurement Techniques, 10(9), 3575–3588, doi:10.5194/amt-10-3575-2017, 2017.

English, P. B., Olmedo, L., Bejarano, E., Lugo, H., Murillo, E., Seto, E., Wong, M., King, G., Wilkie, A., Meltzer, D., Carvlin, G., Jerrett, M. and Northcross, A.: The Imperial County Community Air Monitoring Network: A Model for Community-based Environmental Monitoring for Public Health Action, Environmental Health Perspectives, 125(7), doi:10.1289/EHP1772, 2017.

20

Hacker, K.: Air Monitoring Network Plan for 2018, Allegheny County Health Department Air Quality Program, Pittsburgh, PA. [online] Available from: http://www.achd.net/air/publiccomment2017/ANP2018_final.pdf, 2017.

Hagan, D. H., Isaacman-VanWertz, G., Franklin, J. P., Wallace, L. M. M., Kocar, B. D., Heald, C. L. and Kroll, J. H.: Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments, Atmospheric Measurement Techniques, 11(1), 315–328, doi:10.5194/amt-11-315-2018, 2018.

25

Hornik, K.: Approximation capabilities of multilayer feedforward networks, Neural Networks, 4(2), 251–257, doi:10.1016/0893-6080(91)90009-T, 1991.

Jerrett, M., Burnett, R. T., Ma, R., Pope, C. A., Krewski, D., Newbold, K. B., Thurston, G., Shi, Y., Finkelstein, N., Calle, E. E. and Thun, M. J.: Spatial Analysis of Air Pollution and Mortality in Los Angeles, Epidemiology, 16(6), 727–736, doi:10.1097/01.ede.0000181630.15826.7d, 2005.

30

Karner, A. A., Eisinger, D. S. and Niemeier, D. A.: Near-Roadway Air Quality: Synthesizing the Findings from Real-World Data, Environmental Science & Technology, 44(14), 5334–5344, doi:10.1021/es100008x, 2010.

Loh, M., Sarigiannis, D., Gotti, A., Karakitsios, S., Pronk, A., Kuijpers, E., Annesi-Maesano, I., Baiz, N., Madureira, J., Oliveira Fernandes, E., Jerrett, M. and Cherrie, J.: How Sensors Might Help Define the External Exposome, International Journal of Environmental Research and Public Health, 14(4), 434, doi:10.3390/ijerph14040434, 2017.

Malings, C.: Supplementary Data for "Development of a General Calibration Model and Long-Term Performance Evaluation of Low-Cost Sensors for Air Pollutant Gas Monitoring." [online] Available from: https://doi.org/10.5281/zenodo.1302030, 2018a.

Malings, C.: Supplementary Data for "Development of a General Calibration Model and Long-Term Performance Evaluation of Low-Cost Sensors for Air Pollutant Gas Monitoring" (abridged version). [online] Available from: https://doi.org/10.5281/zenodo.1482011, 2018b.

Marshall, J. D., Nethery, E. and Brauer, M.: Within-urban variability in ambient air pollution: Comparison of estimation methods, Atmospheric Environment, 42(6), 1359–1369, doi:10.1016/j.atmosenv.2007.08.012, 2008.

Masson, N., Piedrahita, R. and Hannigan, M.: Quantification Method for Electrolytic Sensors in Long-Term Monitoring of Ambient Air Quality, Sensors, 15(10), 27283–27302, doi:10.3390/s151027283, 2015.

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, Atmospheric Environment, 70, 186–203, doi:10.1016/j.atmosenv.2012.11.060, 2013.

Nabney, I.: NETLAB: algorithms for pattern recognitions, Springer, London ; New York., 2002.

Popoola, O. A. M., Stewart, G. B., Mead, M. I. and Jones, R. L.: Development of a baseline-temperature correction methodology for electrochemical sensors and its implications for long-term stability, Atmospheric Environment, 147, 330–343, doi:10.1016/j.atmosenv.2016.10.024, 2016.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian processes for machine learning, MIT Press, Cambridge, Mass., 2006.

Sadighi, K., Coffey, E., Polidori, A., Feenstra, B., Lv, Q., Henze, D. K. and Hannigan, M.: Intra-urban spatial variability of surface ozone in Riverside, CA: viability and validation of low-cost sensors, Atmospheric Measurement Techniques, 11(3), 1777–1792, doi:10.5194/amt-11-1777-2018, 2018.

Smith, K. R., Edwards, P. M., Evans, M. J., Lee, J. D., Shaw, M. D., Squires, F., Wilde, S. and Lewis, A. C.: Clustering approaches to improve the performance of low cost air pollution sensors, Faraday Discussions, 200, 621–637, doi:10.1039/C7FD00020K, 2017.

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, Environmental Science & Technology, 47(20), 11369–11377, doi:10.1021/es4022602, 2013.

Spinelle, L., Aleixandre, M., Gerboles, M., European Commission, Joint Research Centre and Institute for Environment and Sustainability: Protocol of evaluation and calibration of low-cost gas sensors for the monitoring of air pollution., Publications Office, Luxembourg. [online] Available from: http://dx.publications.europa.eu/10.2788/9916 (Accessed 13 February 2018), 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 and Actuators B: Chemical, 215, 249–257, doi:10.1016/j.snb.2015.03.031, 2015.

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, Environmental Science & Technology, 48(24), 14186–14194, doi:10.1021/es5034074, 2014.

Turner, M. C., Nieuwenhuijsen, M., Anderson, K., Balshaw, D., Cui, Y., Dunton, G., Hoppin, J. A., Koutrakis, P. and Jerrett, M.: Assessing the Exposome with External Measures: Commentary on the State of the Science and Research Recommendations, Annual Review of Public Health, 38(1), 215–239, doi:10.1146/annurev-publhealth-082516-012802, 2017.

Williams, R., Vasu Kilaru, Snyder, E., Kaufman, A., Dye, T., Rutter, A., Russel, A. and Hafner, H.: Air Sensor Guidebook, U.S. Environmental Protection Agency, Washington, DC. [online] Available from: https://cfpub.epa.gov/si/si_public_file_download.cfm?p_download_id=519616, 2014.

Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L. and R. Subramanian: A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring, Atmospheric Measurement Techniques, 11(1), 291–313, doi:10.5194/amt-11-291-2018, 2018.
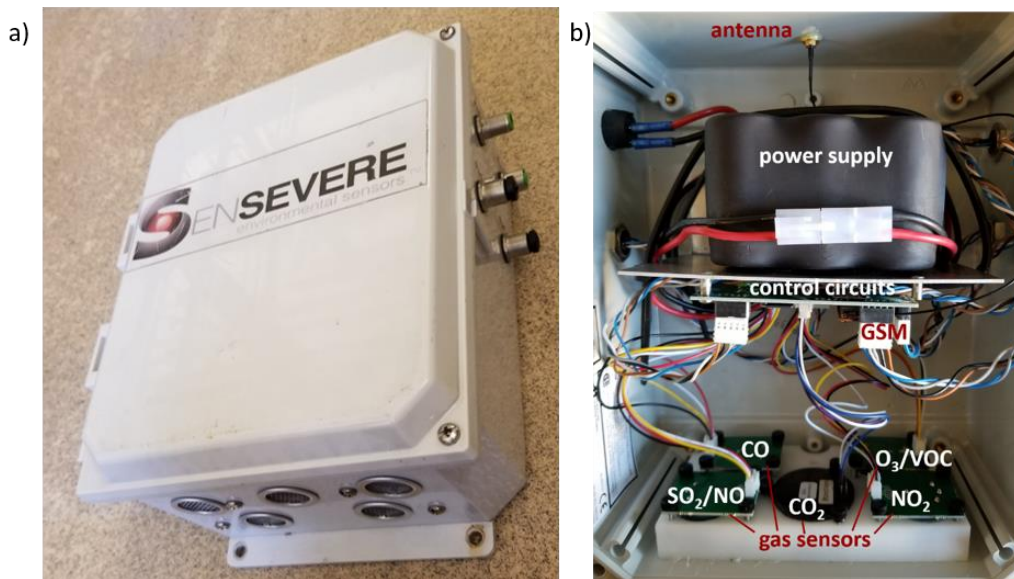
Figure 1: External (a) and internal (b) configuration of the RAMP monitor.
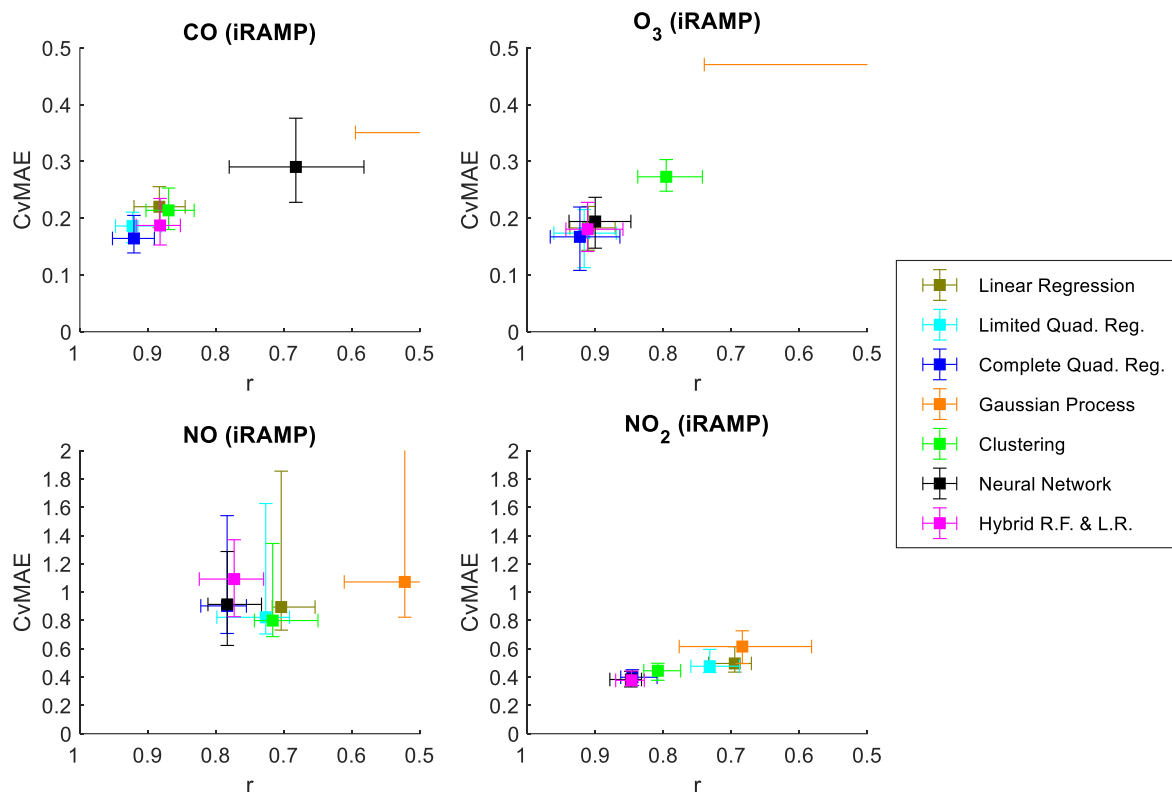
Figure 2. Comparative performance of various individualized RAMP calibration models across gases measured by the RAMPs. Models are trained and tested on distinct subsets of collocation data collected at the CMU site during 2017; performance shown is based on the testing data set only. Proximity to the lower-left corner of each figure indicates better performance. Note the differing vertical axis scales.
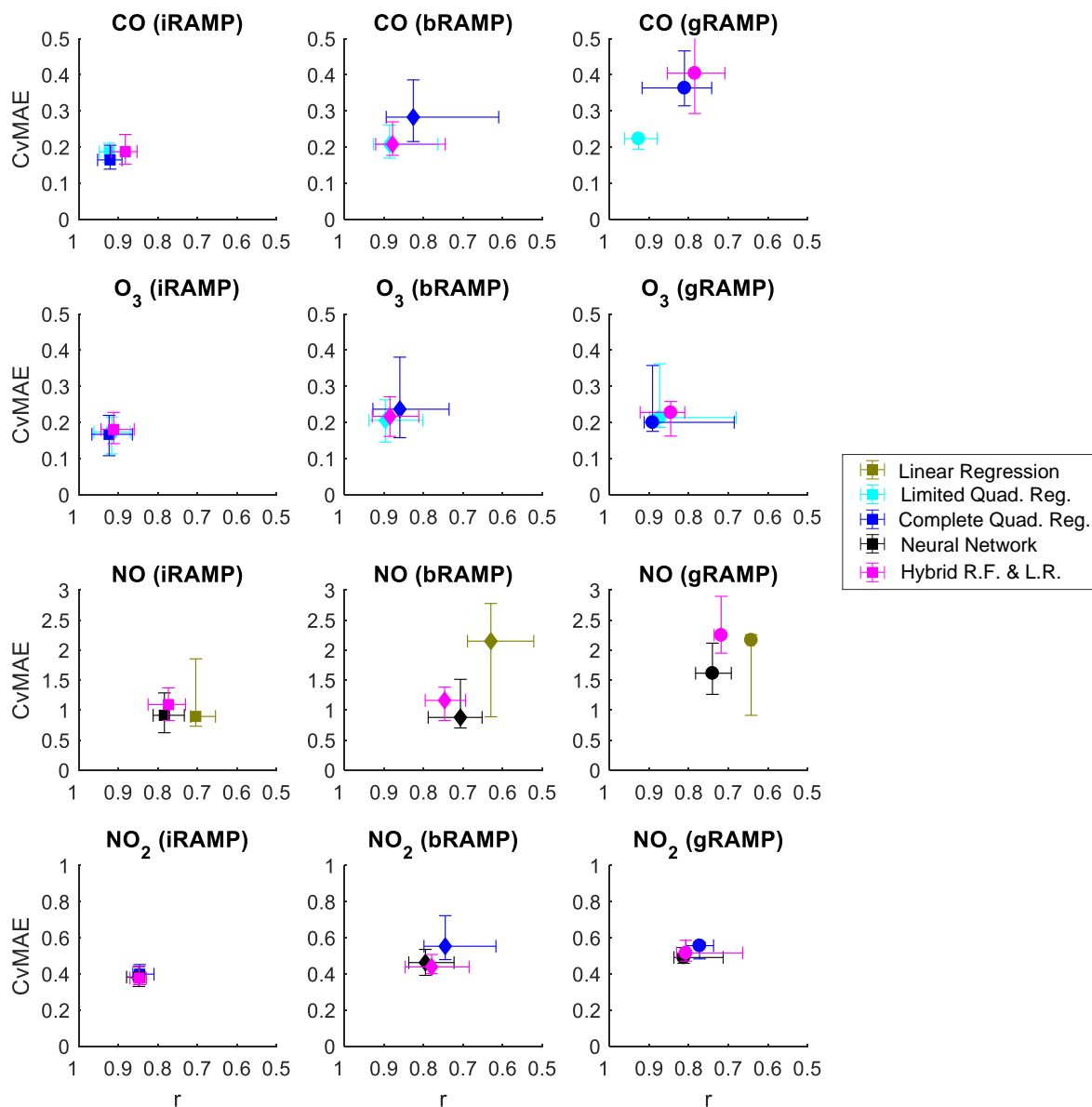
Figure 3. Comparative performance of individualized (iRAMP - square), best individual (bRAMP - diamond), and general (gRAMP - circle) model categories across gases measured by the RAMPs. The modelling algorithms used for each gas corresponds to three of the better-performing algorithms identified among the individualized models. Models are trained and tested on distinct subsets of collocation data collected at the CMU site during 2017; performance shown is based on the testing data set only. Proximity to the lower-left corner of each figure indicates better performance.
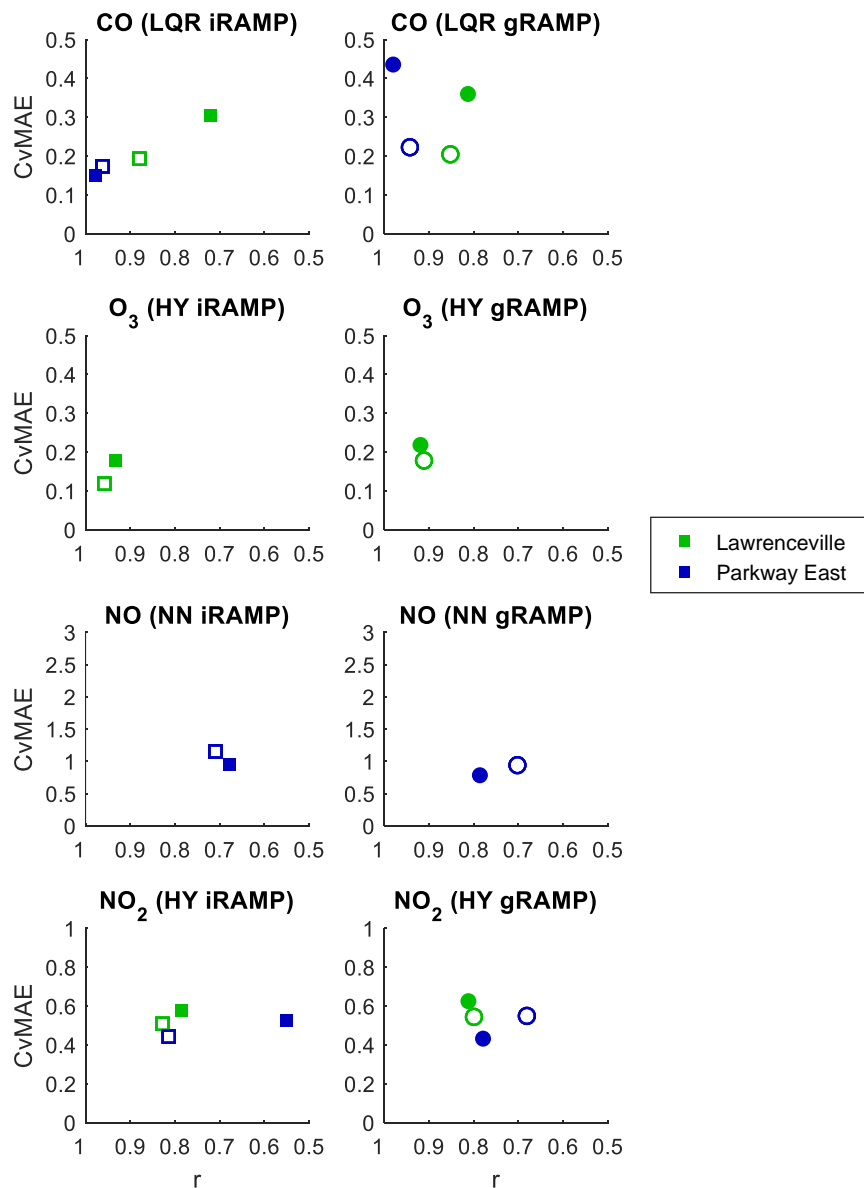
Figure 4. Comparative performance of individual and general models for RAMPs deployed to ACHD monitoring stations (filled makers), compared to the performance of the same RAMPs at the CMU site (hollow markers). For example, a filled green marker indicates the performance of a RAMP at the Lawrenceville site, while a hollow green marker indicates the performance of that same RAMP when it was at the CMU site. The modelling algorithm used for each gas corresponds to the most consistent algorithm identified among the models depicted in Fig. 3: limited quadratic regression (LQR) for CO, neural network (NN) for NO, and hybrid random forest/linear regression models (HY) for NO_2 and O_3. Models are trained on data collected at the CMU site during 2017;

25

performance shown for the CMU site (hollow marker) is based on the testing data for the corresponding RAMPs collected at that site. Proximity to the lower-left corner of each figure indicates better performance.
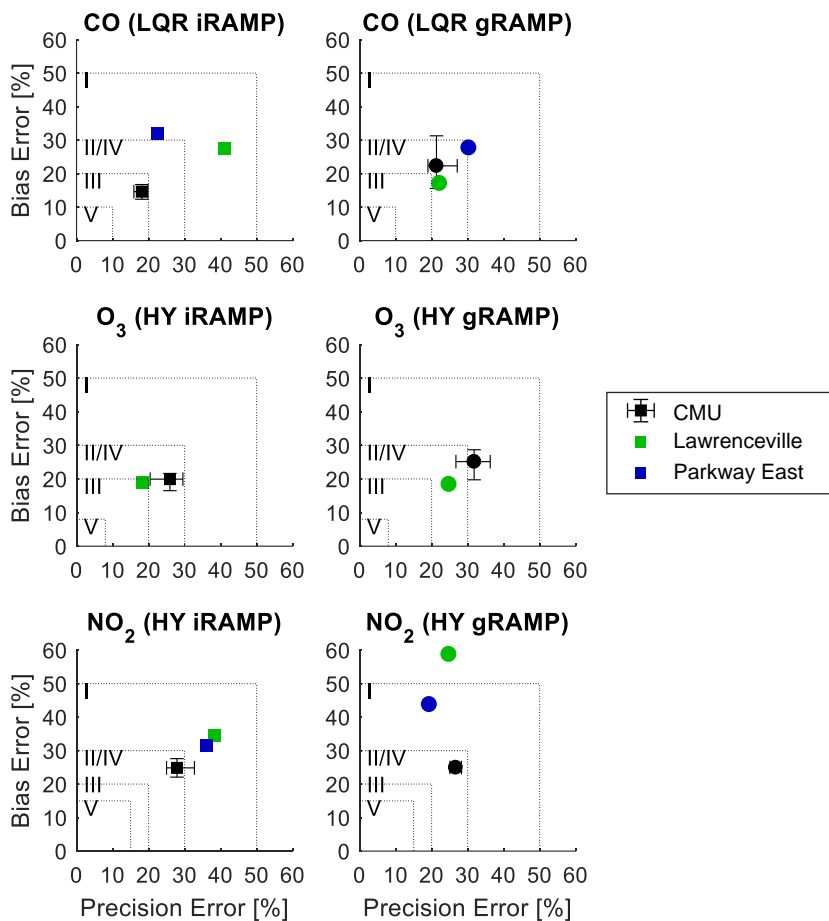

Figure 5. Comparative performance of individual and general models for RAMPs deployed to ACHD monitoring stations using EPA performance criteria. Dotted lines indicate the outer limits of each performance tier. Performance shown for the CMU site is based on performance across all RAMPs at that site based on testing data only. Proximity to the lower-left corner of each figure indicates better performance.
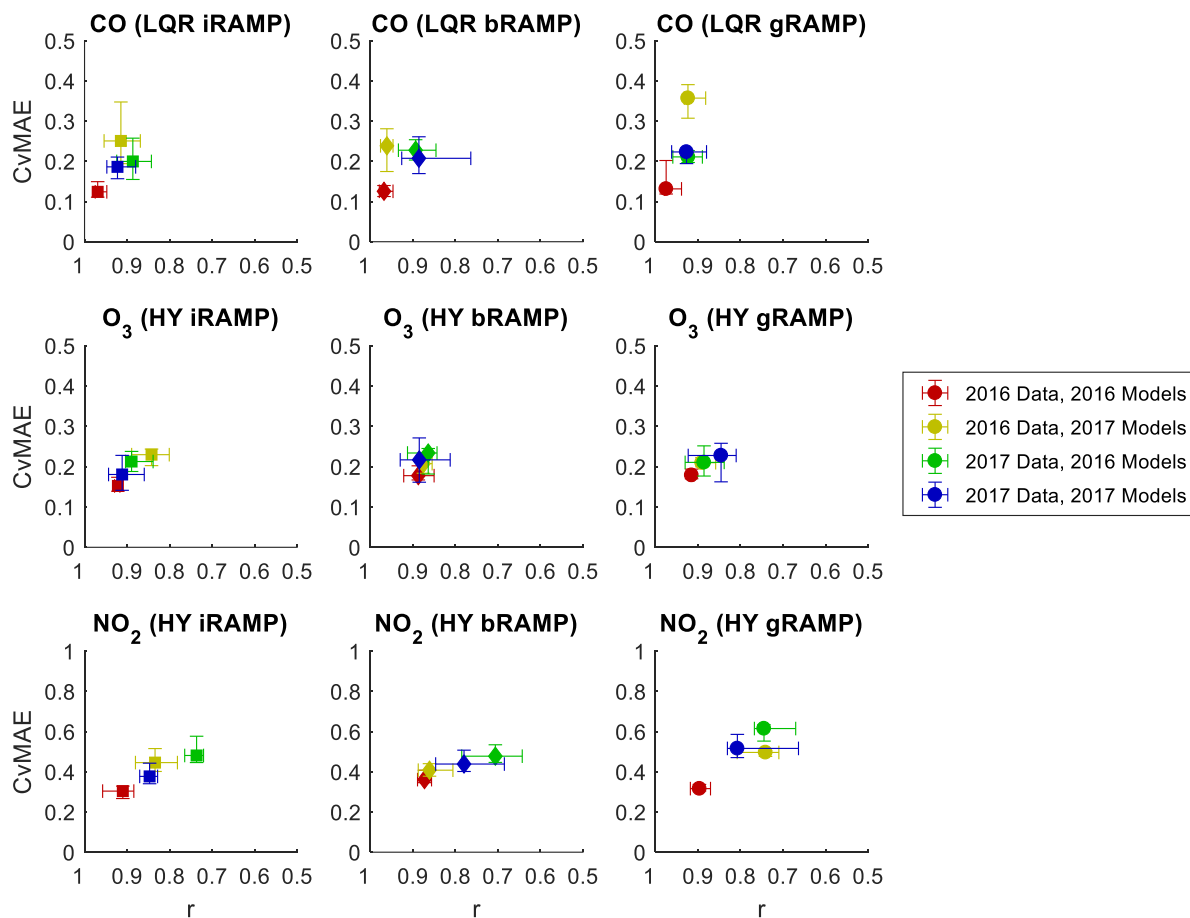
Figure 6. Comparative performance of models in 2016 and 2017. The "Models" year indicates the year from which training data collected at the CMU site are used to calibrate the model; the "Data" year indicates the year from which testing data collected a the CMU site are used to evaluate the model.
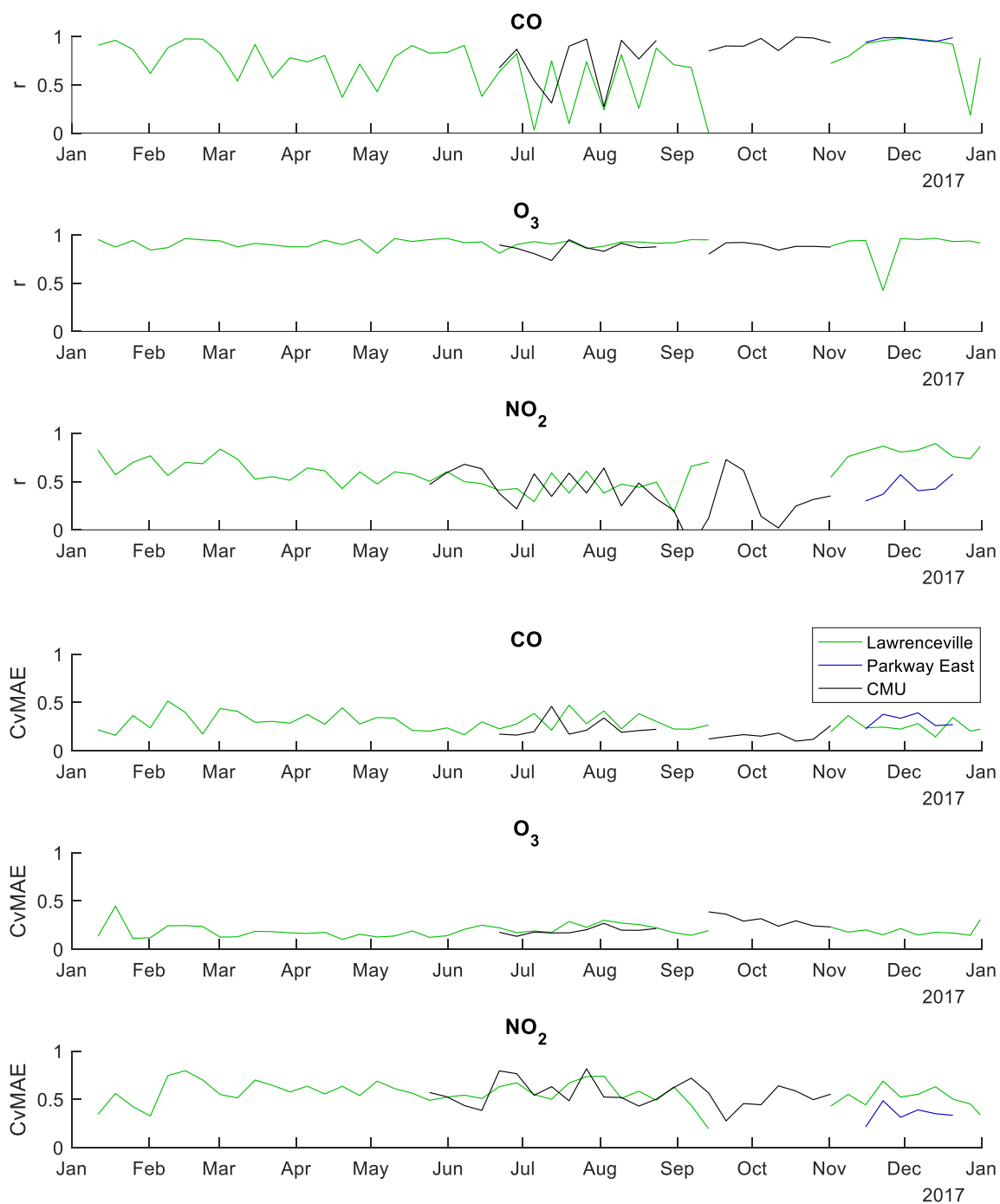
Figure 7. Tracking the performance of RAMP monitors deployed to ACHD Lawrenceville, ACHD Parkway East, and CMU over time. Statistics are computed for each week. Results shown correspond to those of models trained using data collected at the CMU

site during 2017. For CO, the generalized limited quadratic regression model is used; for NO_2 and O_3, generalized hybrid random forest/linear regression models are used.

Table 1: Assigned lower limits for censoring small measurement values.

Quantity	Assigned Lower Limit
CO	200 ppb
NO_2	10 ppb
O_3	10 ppb

5 **Table 2: EPA air quality sensor performance guidelines for various applications. Reproduced from (Williams et al., 2014).**

Tier	Application	Error Metrics
I	Education	< 50%
II	Hotspot Identification and Characterization	< 30%
III	Supplemental Monitoring	< 20%
IV	Personal Exposure Monitoring	< 30%
V	Regulatory Monitoring	< 7% for O_3
		< 10% for CO
		< 15% for NO_2

Table 3: Performance of iRAMP calibration models with respect to EPA air quality sensor performance guidelines as assessed at the CMU site. Entries in the table denote which models meet the corresponding guidelines for each gas (LR = linear regression; LQR = limited quadratic regression; CQR = complete quadratic regression; GP = Gaussian process; CL = clustering; NN = neural
10 **network; HY = hybrid random forest/linear regression).**

Gas	Tier I	Tier II/IV	Tier III
CO	NN	LR, GP, CL, HY	LQR, CQR
O_3	CL	LR, LQR, CQR, GP, NN, HY	
NO_2	GP	LR, LQR, CQR, CL, NN, HY	

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). Ranges of lower values, average values, and upper values across RAMPs for each period are presented.

Gas	Training Period				Testing Period			
	Duration	Concentration			Duration	Concentration		
	[days]	[ppb]			[days]	[ppb]		
	Range	*lower range*	*average range*	*upper range*	*Range*	*lower range*	*average range*	*upper range*
CO	21 - 28	57-118	193-356	923-3750	3 - 75	57-120	145-451	235-3750
NO	26 - 28	0-1	1-4	21-66	4 - 93	0-2	1-3	11-66
NO_2	22 - 28	0-1	5-9	19-31	4 - 110	0-1	4-9	15-32
O_3	21 - 28	1-3	21-36	62-128	2 - 76	1-23	22-48	54-128
T	[°C]	2-16	18-26	32-42	[°C]	0-18	14-27	27-42
RH	[%]	26-52	56-71	66-94	[%]	25-52	50-73	64-94

5

30

Table 5: Performance data for iRAMP models at CMU in 2017 (Avg. is the average, SD is the standard deviation). The "#" sub-column under "Model" indicates the total number of iRAMP models developed for each gas. Slope and r^2 are presented for the best-fit-line between the calibrated RAMP measures and those of the regulatory monitor.

| Gas	Model Type	#	Slope		r^2		MAE [ppb]		Bias [ppb]	
			Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD
CO	LR	48	1.08	0.36	0.75	0.16	60	19	-6	18
	LQR	48	1.01	0.18	0.83	0.10	48	11	-5	14
	CQR	48	0.96	0.16	0.85	0.09	46	16	-3	20
	CL	48	1.06	0.24	0.74	0.11	58	17	1	22
	NN	48	1.33	1.11	0.46	0.23	84	34	-2	34
	HY	48	0.94	0.20	0.77	0.12	52	15	11	22
NO	LR	19	1.31	0.56	0.15	0.07	2.3	1.1	0.26	0.80
	LQR	19	1.15	0.52	0.25	0.15	2.3	1.1	0.26	0.80
	CQR	19	0.97	0.36	0.36	0.14	2.1	1.0	0.35	0.82
	CL	19	0.67	0.33	0.18	0.10	2.2	1.2	0.08	0.80
	NN	19	0.90	0.35	0.30	0.13	2.0	1.0	0.09	0.55
	HY	19	0.65	0.37	0.32	0.14	2.3	0.8	0.76	0.66
NO_2	LR	62	0.89	0.31	0.17	0.08	3.4	0.7	0.16	0.88
	LQR	62	0.79	0.23	0.21	0.09	3.3	0.6	0.18	0.93
	CQR	62	0.85	0.15	0.47	0.11	2.6	0.5	0.07	0.73
	CL	62	0.77	0.17	0.37	0.12	2.9	0.5	0.27	0.70
	NN	62	0.93	0.16	0.49	0.12	2.6	0.5	0.07	0.59
	HY	62	0.83	0.13	0.48	0.10	2.6	0.4	0.51	0.63
O_3	LR	44	0.98	0.06	0.80	0.12	5.1	1.7	-0.05	1.6
	LQR	44	0.96	0.05	0.83	0.11	4.6	1.7	0.12	1.4
	CQR	44	0.93	0.07	0.82	0.12	4.6	1.7	-0.08	1.1
	CL	44	0.89	0.11	0.62	0.11	7.3	1.3	-0.47	2.4
	NN	44	0.98	0.21	0.73	0.26	5.8	2.8	0.09	1.3
	HY	44	0.93	0.06	0.81	0.09	4.9	1.3	0.40	1.5

5
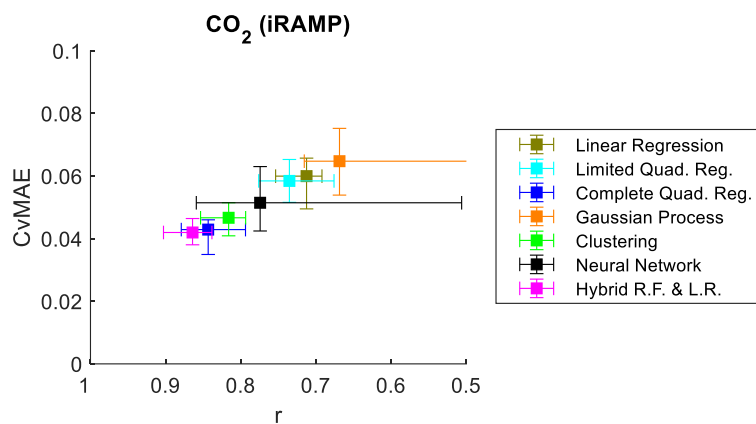
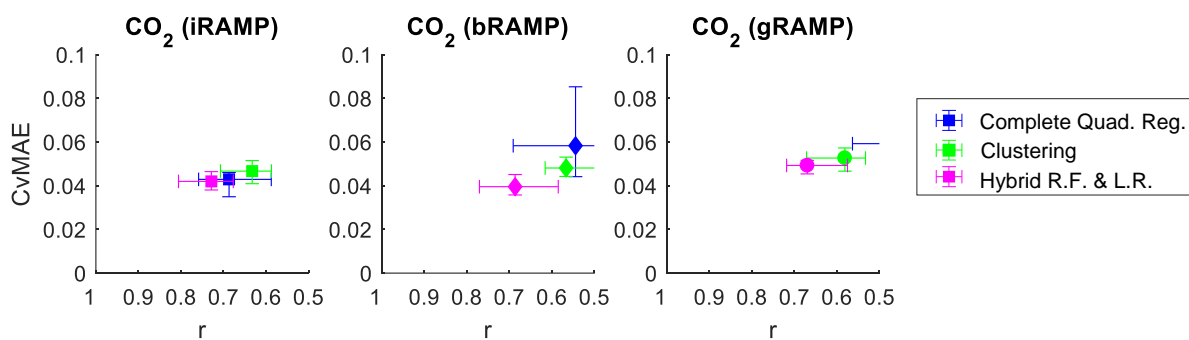**Figure S1. Results corresponding to Fig. 2 for CO₂.**

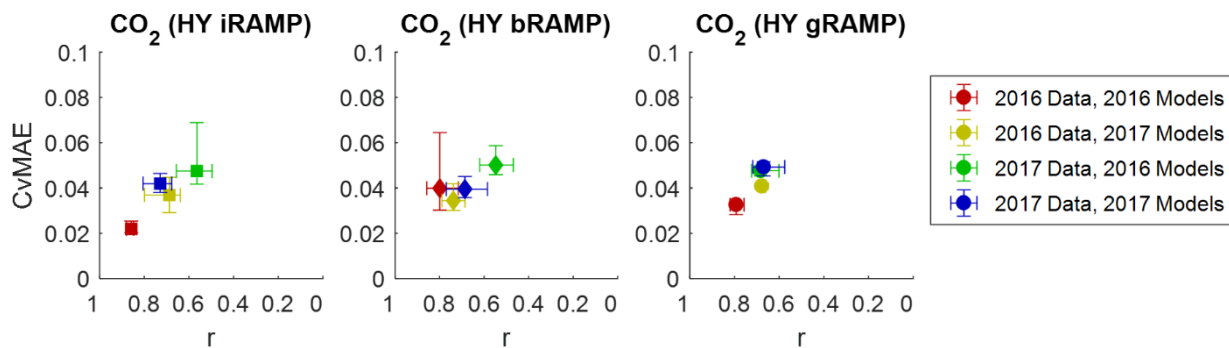**Figure S2. Results corresponding to Fig. 3 for CO₂.**

5

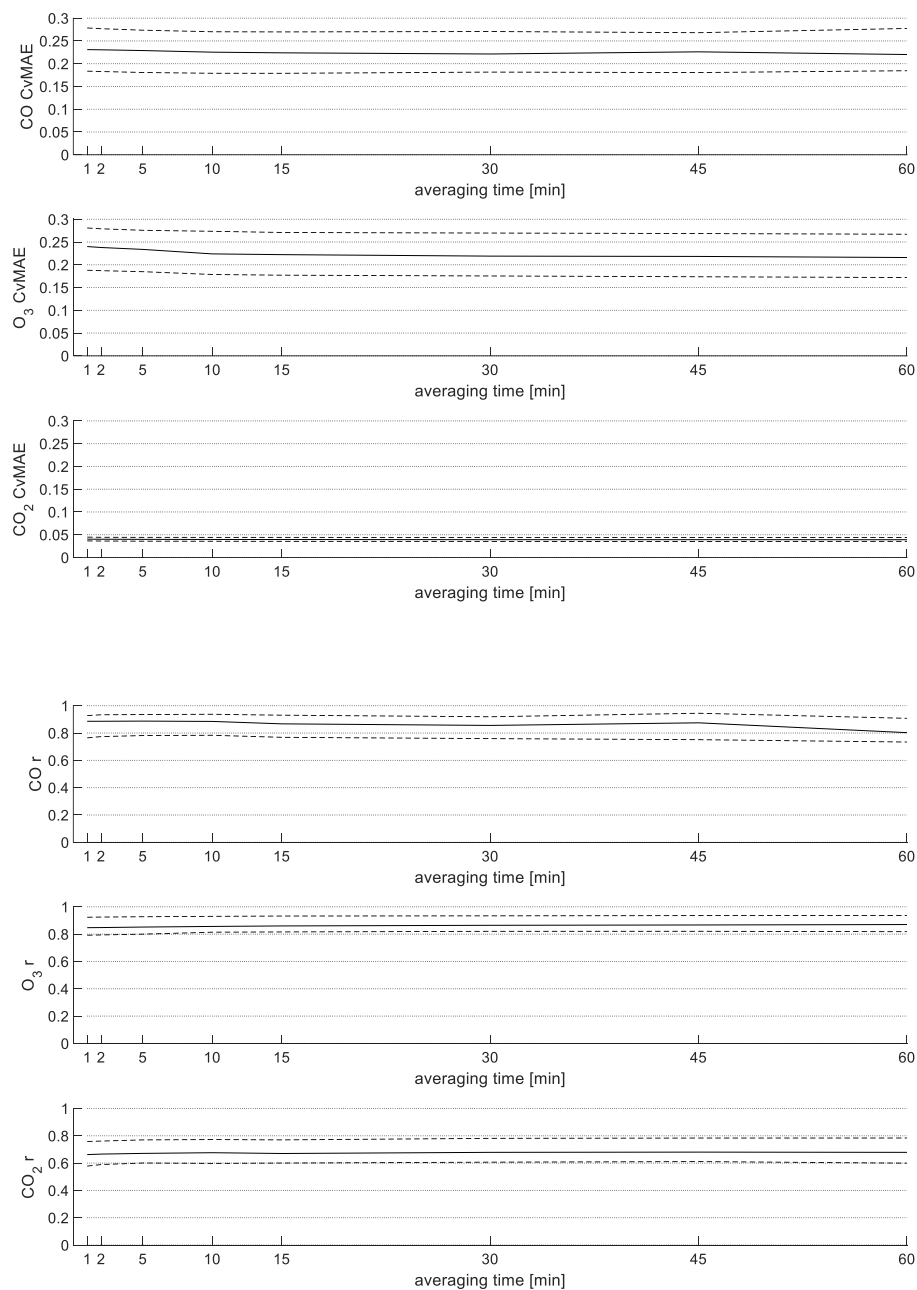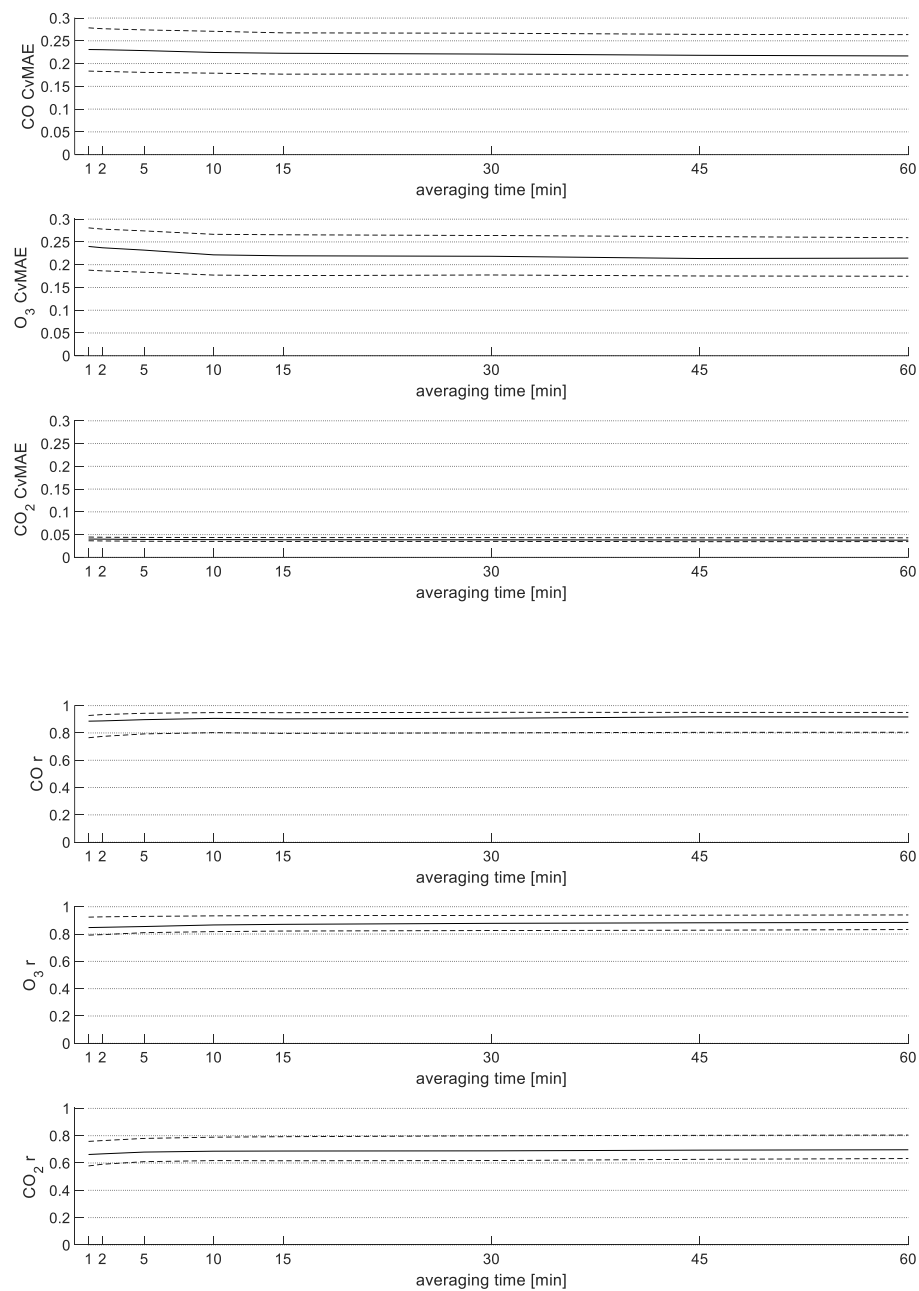**Figure S3. Results corresponding to Fig. 6 for CO₂.**

Figure S4. An evaluation of the performance of the calibration algorithms as a function of the averaging period applied to the raw RAMP data. All models are trained using data collected at the CMU site in 2017. Performance is also evaluated at the CMU site in 2017. Solid lines indicate median performance across RAMPs, dashed lines indicate 25th and 75th percentiles of performance. For CO, the gRAMP LQR model is used; for O₃ the gRAMP HY model is used; for CO₂ the gRAMP HY model is used. Note that all models were originally developed using data averaged at 15 minutes. Results are presented for CvMAE and Pearson r, for averaging times ranging from 1 minute up to 1 hour. These results indicate that the performance of the calibration approaches are fairly stable

for data averaged over periods ranging up to 1 hour. At longer averaging periods, the use of time-averaged environmental variables (such as temperature and relative humidity) in the calibration model appears to reduce performance.

Figure S5. An evaluation of the performance of the calibration algorithms as a function of the averaging period; in contrast to the previous figure, this presents the results when averaging is performed after calibration, rather than before. In terms of CvMAE, performance improves as averaging time increases. In terms of Pearson r, results can be worse with longer averaging, due to the

reduction in the number of points used to evaluate correlation (since there are fewer time periods overall to compare) and to the reduction in the variability (although accuracy is improving as averaging time increases, the variability in the data are also being reduced, and so correlation is decreasing).
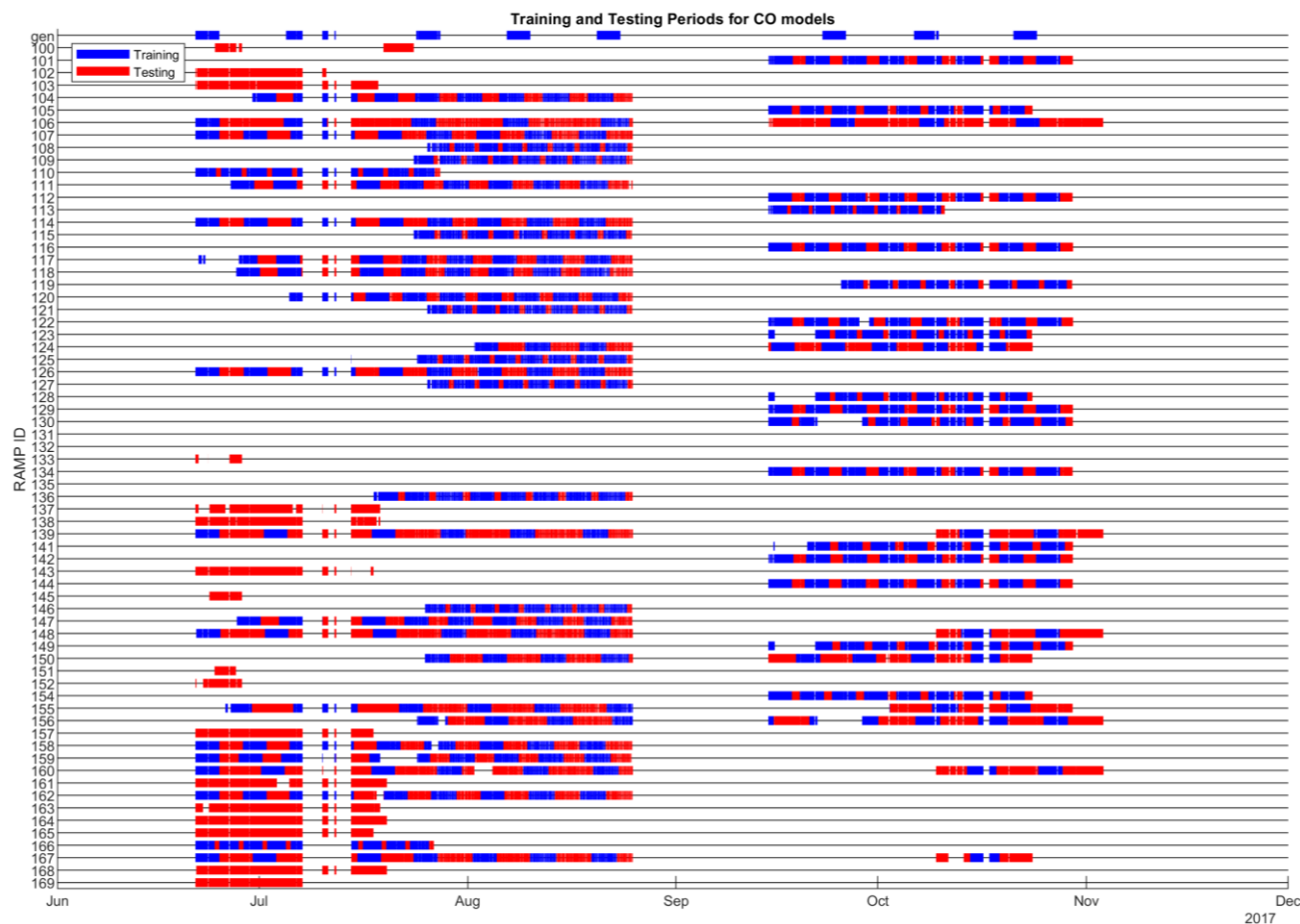
Figure S6. Description of the training and testing periods used for CO models. Blue bars indicate periods used for training data, while red bars denote periods set aside for testing. Time divisions for individual RAMPs (with numeric IDs) are presented corresponding to data used for iRAMP and bRAMP models. Divisions for the "gen" RAMP indicate the training data periods used for gRAMP models, derived from the median of data from the training set of RAMPs collected during these periods; testing data for gRAMP models is drawn from RAMPs which are not part of the training set of RAMPs.

Figure S7. Description of the training and testing periods used for NO models. Blue bars indicate periods used for training data, while red bars denote periods set aside for testing. Time divisions for individual RAMPs (with numeric IDs) are presented corresponding to data used for iRAMP and bRAMP models. Divisions for the "gen" RAMP indicate the training data periods used for gRAMP models, derived from the median of data from the training set of RAMPs collected during these periods; testing data for gRAMP models is drawn from RAMPs which are not part of the training set of RAMPs.
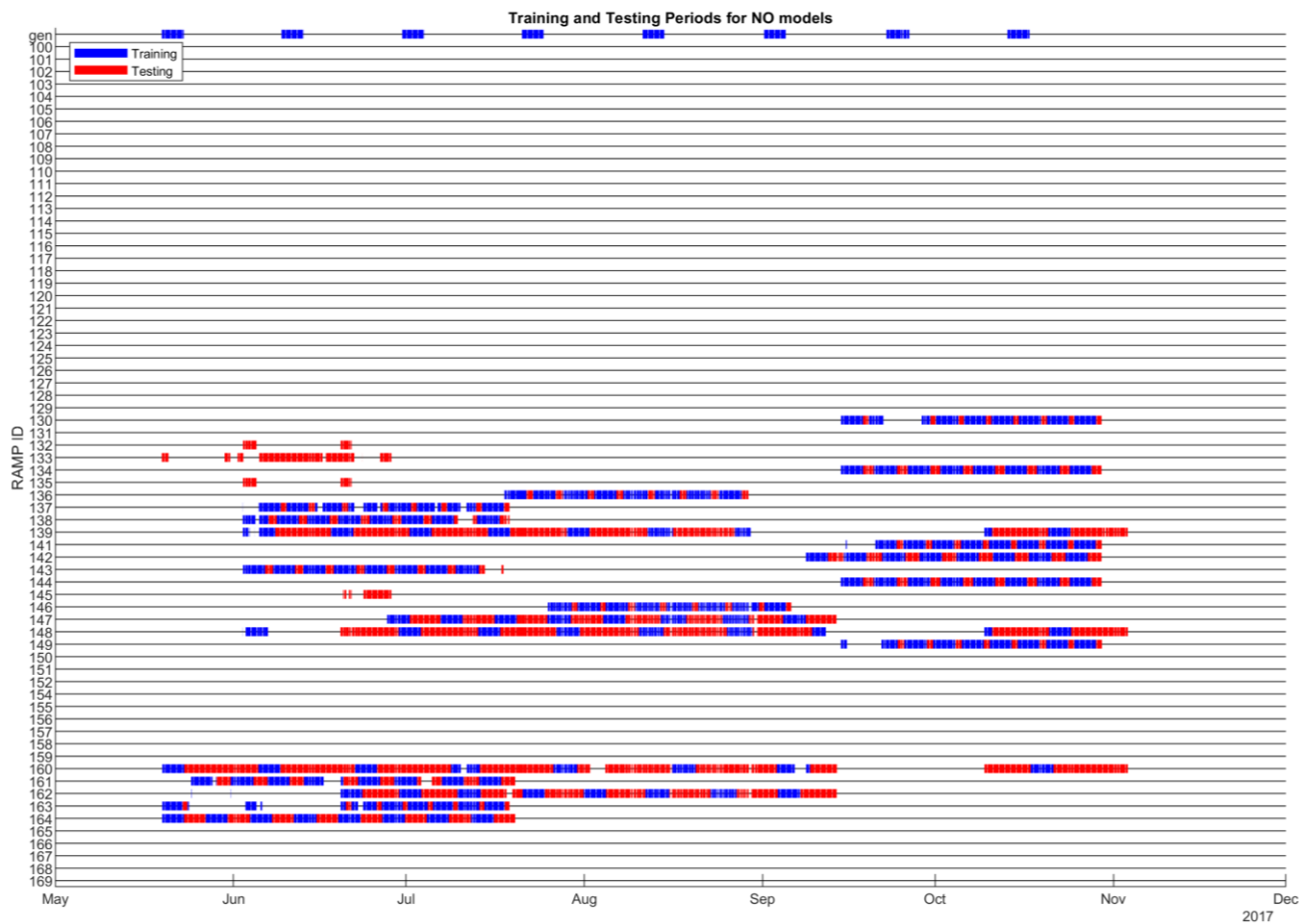
Figure S8. Description of the training and testing periods used for NO₂ models. Blue bars indicate periods used for training data, while red bars denote periods set aside for testing. Time divisions for individual RAMPs (with numeric IDs) are presented corresponding to data used for iRAMP and bRAMP models. Divisions for the "gen" RAMP indicate the training data periods used for gRAMP models, derived from the median of data from the training set of RAMPs collected during these periods; testing data for gRAMP models is drawn from RAMPs which are not part of the training set of RAMPs.
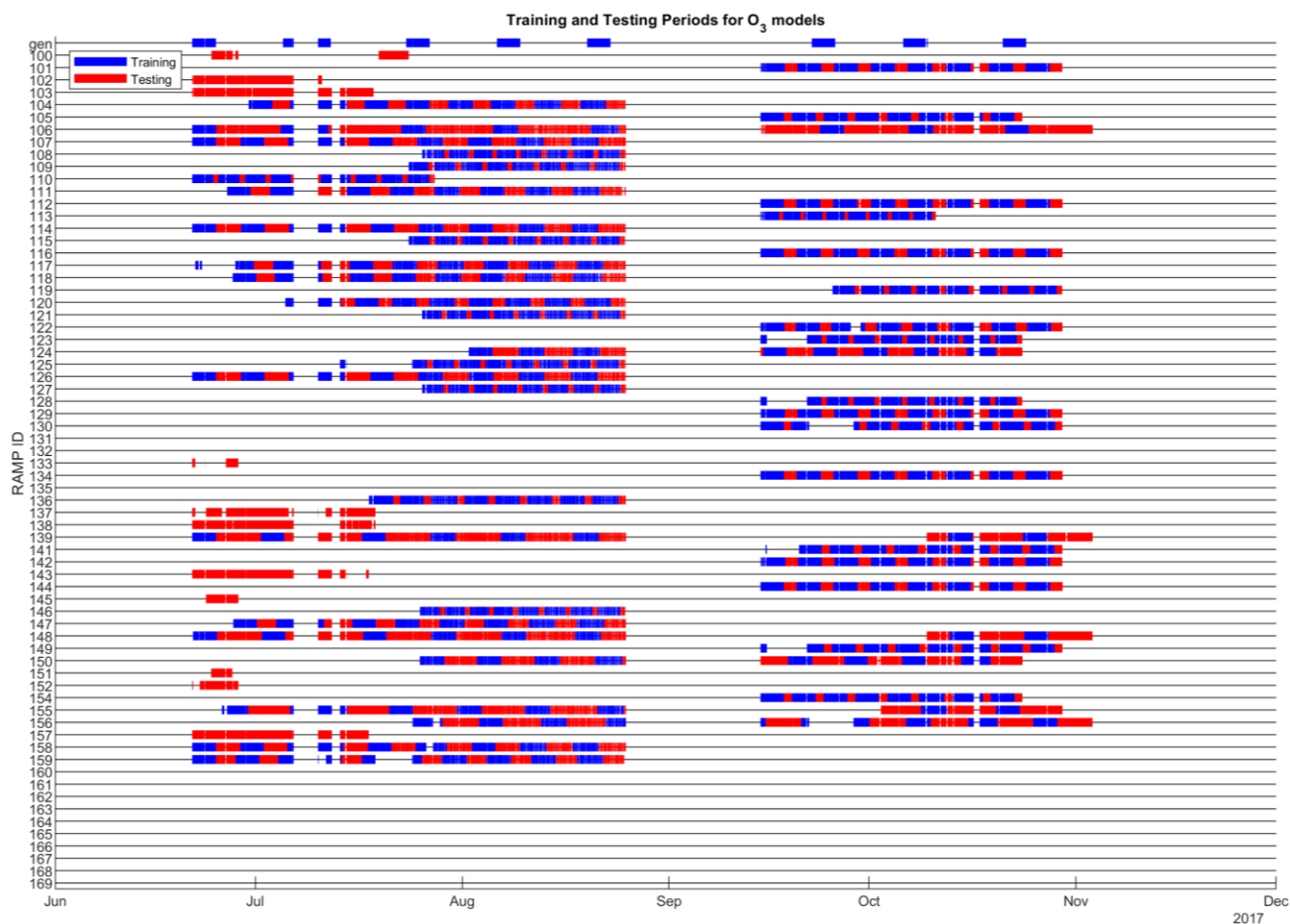
Figure S9. Description of the training and testing periods used for O₃ models. Blue bars indicate periods used for training data, while red bars denote periods set aside for testing. Time divisions for individual RAMPs (with numeric IDs) are presented corresponding to data used for iRAMP and bRAMP models. Divisions for the "gen" RAMP indicate the training data periods used for gRAMP models, derived from the median of data from the training set of RAMPs collected during these periods; testing data for gRAMP models is drawn from RAMPs which are not part of the training set of RAMPs.
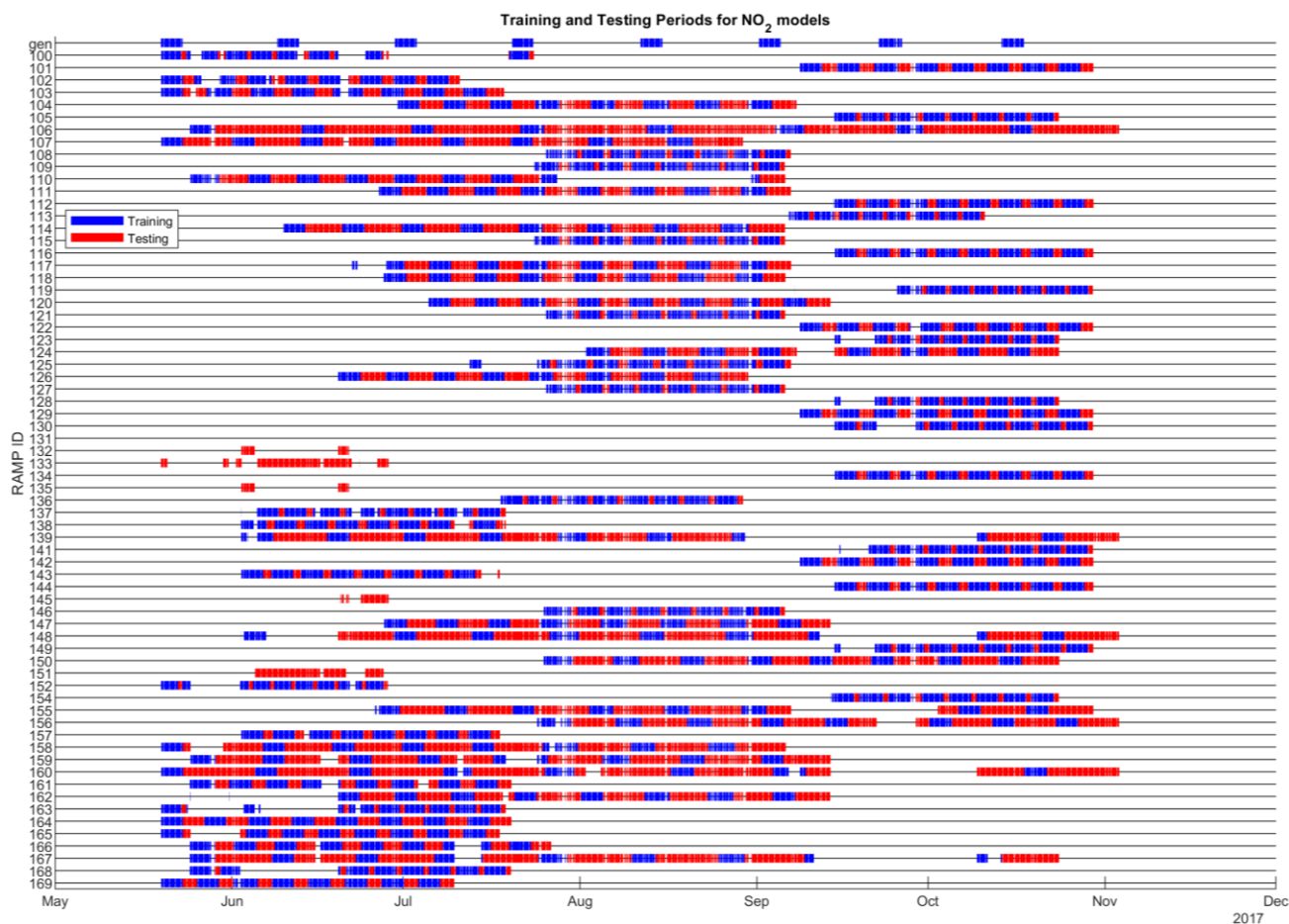
Figure S10. Description of the training and testing periods used for CO₂ models. Blue bars indicate periods used for training data, while red bars denote periods set aside for testing. Time divisions for individual RAMPs (with numeric IDs) are presented corresponding to data used for iRAMP and bRAMP models. Divisions for the "gen" RAMP indicate the training data periods used for gRAMP models, derived from the median of data from the training set of RAMPs collected during these periods; testing data for gRAMP models is drawn from RAMPs which are not part of the training set of RAMPs.

Figure S11. Depiction of the range of CO concentrations experienced during training and testing. Blue boxplots indicate training ranges, while red boxplots denote testing ranges. Dots with circles indicate the midpoint, thicker bars indicate the interquartile range, thinner bars show 1st and 99th percentiles, and colored dots depict outliers. The horizontal axis shows the RAMP ID number (or "gen", which depicts the concentration range used for training gRAMP models).

Figure S12. Depiction of the range of NO concentrations experienced during training and testing. Blue boxplots indicate training ranges, while red boxplots denote testing ranges. Dots with circles indicate the midpoint, thicker bars indicate the interquartile range, thinner bars show 1st and 99th percentiles, and colored dots depict outliers. The horizontal axis shows the RAMP ID number (or "gen", which depicts the concentration range used for training gRAMP models).
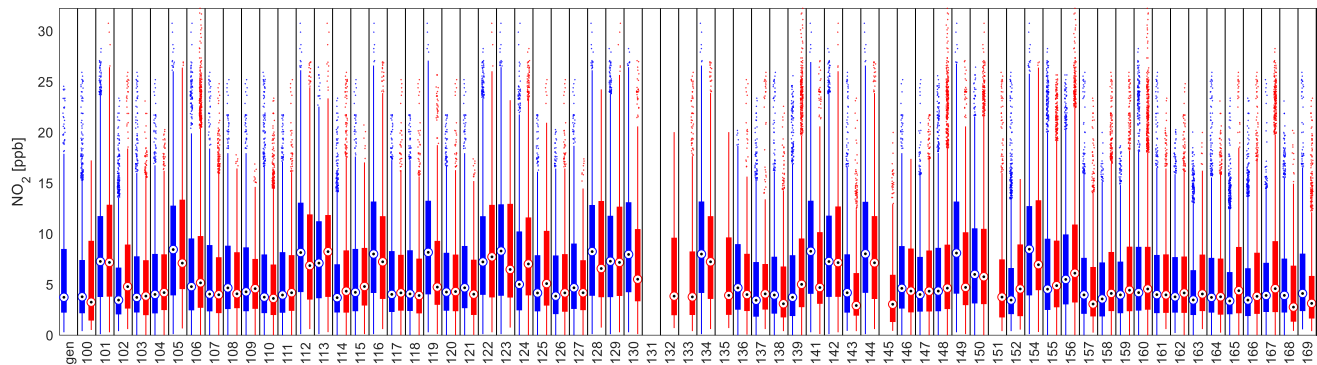
Figure S13. Depiction of the range of NO₂ concentrations experienced during training and testing. Blue boxplots indicate training ranges, while red boxplots denote testing ranges. Dots with circles indicate the midpoint, thicker bars indicate the interquartile range, thinner bars show 1ˢᵗ and 99ᵗʰ percentiles, and colored dots depict outliers. The horizontal axis shows the RAMP ID number (or "gen", which depicts the concentration range used for training gRAMP models).
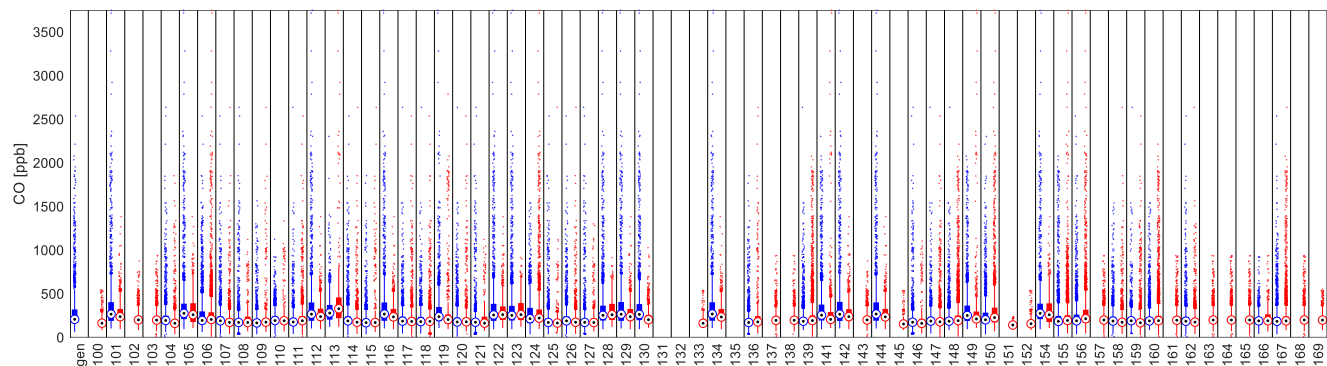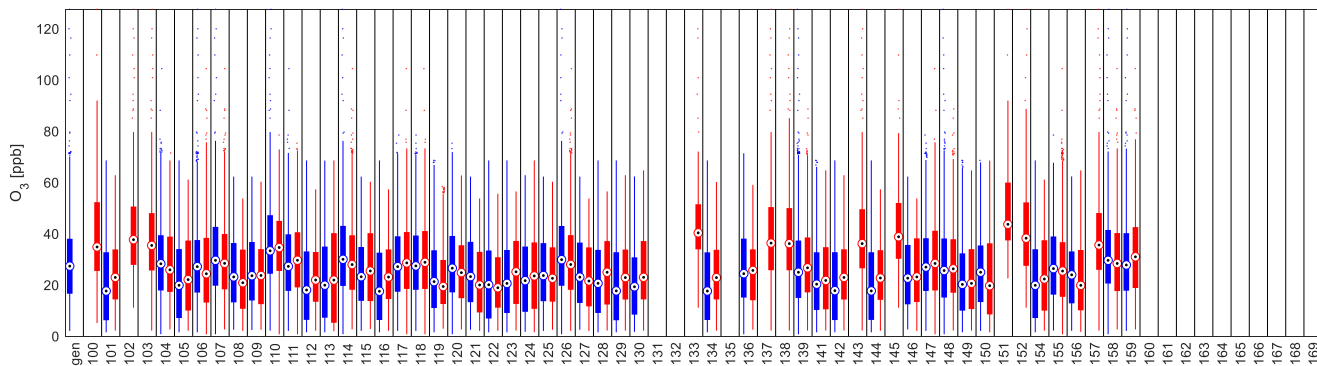

Figure S14. Depiction of the range of O₃ concentrations experienced during training and testing. Blue boxplots indicate training ranges, while red boxplots denote testing ranges. Dots with circles indicate the midpoint, thicker bars indicate the interquartile range, thinner bars show 1ˢᵗ and 99ᵗʰ percentiles, and colored dots depict outliers. The horizontal axis shows the RAMP ID number (or "gen", which depicts the concentration range used for training gRAMP models).
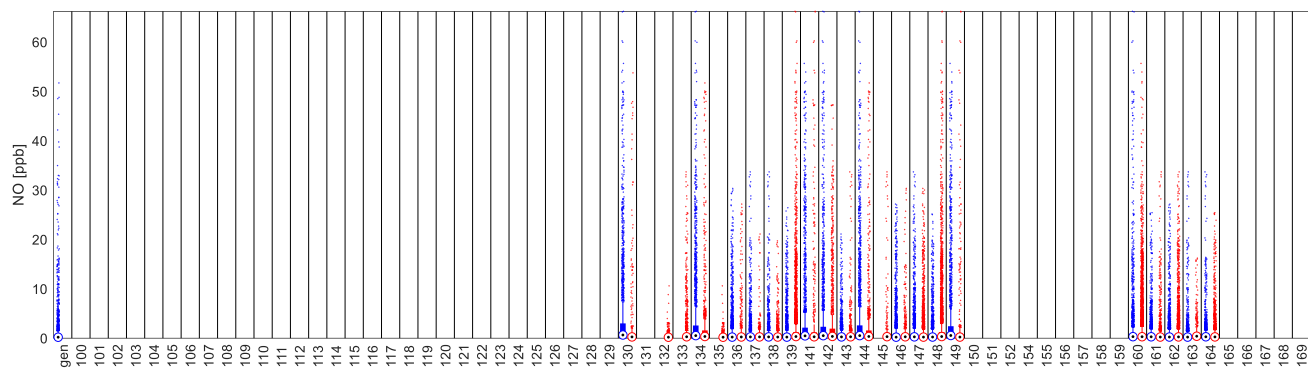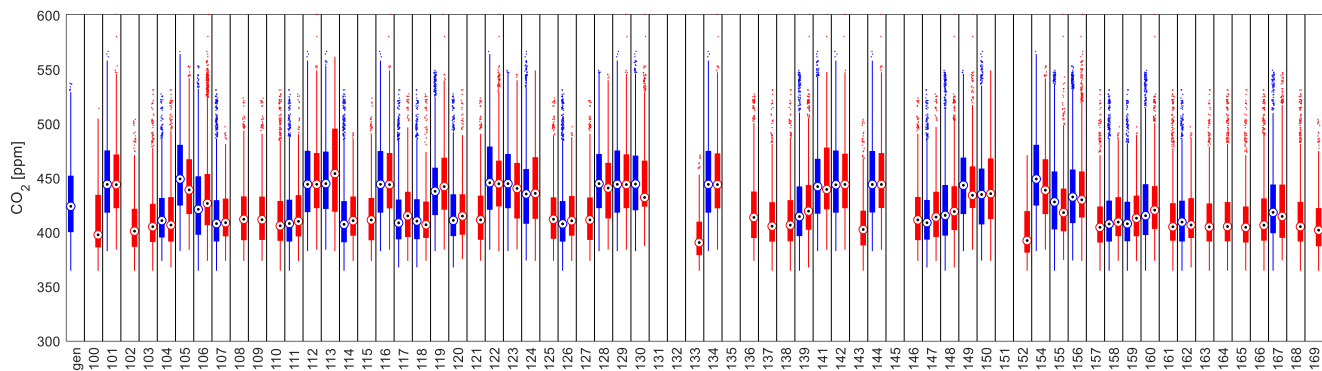
Figure S15. Depiction of the range of CO_2 concentrations experienced during training and testing. Blue boxplots indicate training ranges, while red boxplots denote testing ranges. Dots with circles indicate the midpoint, thicker bars indicate the interquartile range, thinner bars show 1st and 99th percentiles, and colored dots depict outliers. The horizontal axis shows the RAMP ID number (or "gen", which depicts the concentration range used for training gRAMP models).

Table S1: Results corresponding to Table 4 for CO_2.

Gas	Training Period				Testing Period			
	Duration	Concentration			Duration	Concentration		
	[days]	[ppm]			[days]	[ppm]		
		lower range	*average range*	*upper range*		*lower range*	*average range*	*upper range*
	Range				*Range*			
CO_2	21 - 28	365-384	413-454	528-567	2 - 50	365-388	399-458	471-601

Table S2: Results corresponding to Table 5 for CO_2.

Gas	Model Type	#	Testing Performance							
			Slope		r^2		MAE [ppm]		Bias [ppm]	
			Avg.	*SD*	*Avg.*	*SD*	*Avg.*	*SD*	*Avg.*	*SD*
CO_2	LR	38	0.74	0.23	0.21	0.09	24	5	0.6	6.1
	LQR	38	0.62	0.22	0.23	0.12	25	5	1.0	8.1
	CQR	38	0.74	0.20	0.47	0.16	18	3	-1.0	4.8
	CL	38	0.76	0.13	0.43	0.13	20	3	1.4	4.4
	NN	38	0.47	1.53	0.28	0.25	23	7	-2.1	6.5
	HY	38	0.79	0.26	0.53	0.15	19	4	3.2	6.0

5

Table S3: Results corresponding to Table 5 for bRAMP models.

Gas	Model Type	#	Slope		r^2		MAE [ppb]		Bias [ppb]	
			Avg.	*SD*	*Avg.*	*SD*	*Avg.*	*SD*	*Avg.*	*SD*
CO	LR	1	0.68	0.28	0.66	0.23	80	38	54	138
	LQR	1	0.75	0.25	0.71	0.21	59	28	56	119
	CQR	4	0.62	0.29	0.58	0.29	143	337	156	568
	CL	4	0.83	0.31	0.64	0.23	66	29	63	78
	NN	4	1.09	0.58	0.46	0.20	81	31	42	90
	HY	4	0.74	0.26	0.69	0.24	63	38	98	90
NO	LR	1	0.60	0.67	0.09	0.09	8.3	20.6	8.9	19.6
	LQR	1	0.69	0.53	0.14	0.14	3.8	7.2	1.9	6.5
	CQR	2	0.55	0.44	0.19	0.17	8.0	23.2	5.4	19.2
	CL	2	0.38	0.35	0.09	0.13	3.1	1.7	1.1	1.4
	NN	2	1.00	0.57	0.21	0.16	2.2	1.1	0.2	2.0
	HY	2	0.58	0.31	0.26	0.15	2.5	1.8	1.4	2.2
NO_2	LR	1	0.75	0.42	0.14	0.10	4.7	9.6	-0.2	7.0
	LQR	1	0.64	0.27	0.18	0.12	3.5	0.9	-1.4	1.8
	CQR	4	0.44	0.35	0.25	0.19	4.2	2.1	2.1	4.4
	CL	4	0.58	0.29	0.21	0.14	3.5	0.6	2.1	2.9
	NN	4	0.86	0.37	0.33	0.18	3.1	0.7	0.5	2.4
	HY	4	0.78	0.30	0.32	0.19	3.2	1.2	1.4	2.5
O_3	LR	1	0.76	0.24	0.70	0.23	10.5	24.8	4.7	21.5
	LQR	1	0.85	0.22	0.72	0.24	6.1	3.1	-3.1	8.9
	CQR	2	0.75	0.27	0.65	0.27	9.8	17.1	2.9	15.5
	CL	2	0.91	0.30	0.50	0.18	8.7	2.1	-1.7	8.2
	NN	2	0.90	0.53	0.65	0.26	7.0	3.9	-2.6	7.4
	HY	2	1.06	0.21	0.75	0.13	5.8	1.8	0.3	6.5
CO_2	LR	1	0.65	0.40	0.18	0.14	23	5	13	19
	LQR	1	0.41	0.32	0.16	0.16	27	9	15	25
	CQR	4	0.43	0.27	0.31	0.21	58	171	17	148
	CL	4	0.70	0.18	0.32	0.15	21	4	9	17
	NN	4	0.79	0.53	0.29	0.18	31	32	-9	46
	HY	4	0.95	0.27	0.47	0.16	18	3	12	17

Table S4: Results corresponding to Table 5 for gRAMP models.

Gas	Model Type	#	Slope		r²		MAE [ppb]		Bias [ppb]	
			Avg.	*SD*	*Avg.*	*SD*	*Avg.*	*SD*	*Avg.*	*SD*
CO	LR	1	1.03	0.24	0.80	0.11	68	12	26	109
	LQR	1	0.90	0.08	0.85	0.09	56	8	6	93
	CQR	1	0.69	0.18	0.66	0.20	106	41	60	95
	CL	1	1.02	0.19	0.72	0.10	80	12	21	57
	NN	1	0.67	0.22	0.51	0.17	134	59	88	111
	HY	1	0.75	0.11	0.61	0.11	110	41	75	54
NO	LR	1	1.51	0.92	0.07	0.03	3.8	1.8	0.1	1.8
	LQR	1	0.67	0.34	0.08	0.04	7.1	9.6	3.5	8.5
	CQR	1	0.15	0.09	0.06	0.04	5.9	6.4	2.7	6.3
	CL	1	0.43	0.13	0.13	0.03	3.1	1.0	-0.6	0.5
	NN	1	0.49	0.23	0.22	0.12	4.1	3.0	2.3	4.9
	HY	1	0.40	0.22	0.17	0.08	13.2	22.7	9.9	20.1
NO₂	LR	1	1.07	0.40	0.14	0.05	3.9	0.7	-1.2	1.5
	LQR	1	0.86	0.31	0.18	0.07	3.8	0.7	-1.1	1.9
	CQR	1	0.67	0.21	0.30	0.09	3.5	0.4	-1.0	2.7
	CL	1	0.67	0.19	0.26	0.12	3.6	0.5	-0.1	2.4
	NN	1	0.88	0.21	0.34	0.15	3.3	0.4	-0.3	2.9
	HY	1	0.76	0.27	0.30	0.17	3.4	0.5	0.2	2.6
O₃	LR	1	0.89	0.27	0.72	0.22	6.4	2.8	2.2	4.9
	LQR	1	0.77	0.29	0.66	0.24	7.5	4.1	4.4	6.8
	CQR	1	0.76	0.28	0.67	0.25	7.2	4.2	3.5	6.5
	CL	1	0.91	0.13	0.45	0.19	8.9	1.8	-0.8	5.7
	NN	1	0.92	0.18	0.73	0.18	5.9	2.4	1.7	5.1
	HY	1	1.00	0.12	0.73	0.12	5.9	1.5	0.9	2.9
CO₂	LR	1	0.63	0.14	0.21	0.06	26	4	-2	12
	LQR	1	0.55	0.14	0.20	0.06	27	4	-6	12
	CQR	1	0.39	0.11	0.22	0.11	27	5	-9	15
	CL	1	0.71	0.15	0.37	0.13	23	2	3	16
	NN	1	0.30	0.25	0.15	0.12	38	25	-12	19
	HY	1	0.80	0.16	0.43	0.11	21	2	4	15