Dear Reviewer 1,

We would like to offer our sincere thanks for spending your time in the review of our work and helping us to significantly improve the quality and clarity of the manuscript with your very detailed comments and suggestions. Each of your comments is listed below in black text, followed by our response and edits in blue text.

Review 1 Comments:

Overview: In Casey et al., the authors investigated the performance of calibration models developed for ambient O3 and CO2 across time and space using field deployments spanning 2014 – 2017 as case studies. Specifically, they looked at the impact of post-deployment calibration vs pre- and post- calibration, and the impact of applying a calibration model developed in one location on U-Pods deployed in other locations. Calibration models investigated included linear models and artificial neural networks. The size and scope of the study is impressive, and I believe there is a significant quantity of insightful information within this paper.

However, in general, I found the narrative of the paper to be confusing (it is hard to effectively distill such a breadth of research) and the take home points could be made considerably clearer.

Additionally, I think this paper would benefit with a few more analyses of general model performance implications and a closer look at the impact of relative humidity. Following these corrections to comments identified below, I believe the publication is suitable to be published in Atmospheric Measurement Techniques.

Response: We have carefully addressed each of the comments below and carried out the analyses suggested by the reviewer to investigate the implications and impact of relative humidity and sensor drift in time. We have also worked to significantly improve and clarify the narrative of the paper as well as clarified and outlined take home points.

GENERAL COMMENTS

Comment: In general, I found this manuscript a little hard to read, because I felt like a cohesive narrative was missing. A lot of information is presented in a rapid-fire manner such that the results and discussion section reads more like a results section, with limited discussion. Although there is no straight forward solution to this problem, I would suggest that the authors think about the three to five key messages they wish to convey in the manuscript and that they tune and streamline the text to support this narrative.
Response: Thanks very much to the Reviewer for helping point out a way for us to improve a cohesive and clear narrative in this work.
Edits: We have added the following five key points to the conclusion as well as supporting edits in the abstract and throughout the results and discussion section:

"The following findings from this work, and associated recommendations, are made to help inform the logistics of future studies that employ field calibration methods of low-cost gas sensors.

1. **Finding:** For O_3 models, LMs perform better than ANNs when the chemical composition of local emissions sources is significantly different in the model-training location relative to the model-application location. We found that when models were trained in an urban area with significant mobile sources, then tested in a peri-urban area, more strongly influenced by oil and gas emissions, the differences in local sources of pollution were significantly different enough that LMs outperformed ANNs. Alternatively, when models were trained in one oil and gas production region and tested in another the different composition of local emissions (lighter vs. heavier hydrocarbons) was not significant enough for LM performance to surpass the performance of ANNs, though some positive bias was evident in predicted O_3 mole fractions.

Explanation: ANNs are very effective at compensating for the influence of interfering gas species through pattern recognition of a training dataset. However, if different patterns, in terms of the relative abundance of various oxidizing and reducing compounds in the air, are present in the testing location relative to the training location, ANNs may not able to compensate for the influence of interfering gas species as effectively. The relative abundance of interfering oxidizing and reducing compounds are not included as model parameters, but ANN performance is challenged by these circumstances.
Recommendation: When measuring O_3 or other gas species with a metal oxide type sensor, if the nature of dominant emissions sources at the model training location is significantly different than the nature of dominant emissions sources in the model application location, us an LM instead of an ANN. For the best performance, try to train models in locations with similar emissions sources to a desired sampling location. If the nature of dominant emissions sources at the model training and application locations are similar, signals from an array of multiple unique metal oxide sensors will likely augment model performance.

2. **Finding:** For CO_2 models, LMs perform better than ANNs when model training occurs significantly (more than several months) prior to or subsequent to the model application period.
Explanation: CO_2 sensors drift over time in terms of sensitivity and baseline response. When models are extrapolated in time (when training takes place more than several months prior or subsequent to the model application period), ANN performance can be compromised to a greater extent than LM performance because ANNs are able to represent relationships during training very effectively, and with significant more complexity and nonlinear relationships among time and other model inputs than LMs. The more complex the model, the less likely it can be extrapolate effectively. LMs, with no interaction terms like we employ in this work, are not able to fit data and potentially complex patterns inherent in sensor drift over time during training as closely as an ANN, but the simple linear relationships they represent between the time input and the target gas mole fraction over the course of training are more likely to hold prior or subsequent to the training period.
Recommendation: When measuring CO_2 with a NDIR sensor, if model-training data is only available more than several months prior or subsequent to the model application period, use a LM instead of an ANN. For the best model performance, use training data that is collected directly pre or post of the model application period, and preferably data from both pre and post of the model application period. Training models using data from both pre and post of a given model application period helps models to encompass sensor drift over time as well as increases the likelihood of covering the full range of environmental parameter space that occurs during the model application period so that extrapolation of these parameters is avoided.

3. **Finding:** Extrapolation of an O_3 or CO_2 model in time, and especially significant extrapolation in time, can change both the type of model that is most effective, as well as the specific model input signals that are most effective.
Explanation: Low-cost sensors change over time, both in terms of their baseline response and in terms of their sensitivity to target and interfering gas species. Different sensor types drift due to different physical phenomenon so further a generalization across sensor types is difficult.
Recommendation: Use training data collected directly pre and post of the model application period in order to implement a 'best performing model' for each gas species that can be applied using data from different model training and application pairs.

4. **Finding:** ANNs yield less bias and more accurate gas mole fraction quantification than LMs, even when transferred to a new location under the following circumstances: when extrapolation of training parameters is avoided during the model application period, when training takes place for several weeks to a month prior and subsequent to the model

application period, and when the dominant local emissions sources are similar in the model training and application locations.
Explanation: Our previous study and multiple other ambient and laboratory based experiments have shown, arrays of low-cost sensors in combination with ANN regression models can support useful quantification of gases in mixtures and in the ambient environment because ANNs can more effectively represent complex nonlinear relationships among environmental variables and signals in a sensor system like a U-Pod than LMs**.** With this work, we have explored limitations associated with these methods when challenged in different ways, as we present with a number of case studies.
Recommendation: If minimizing error and bias in measurements of gas mole fractions using low-cost sensors systems is a primary goal, design sensor system training and field deployment experiments so that extrapolation of model training parameters is avoided during the model application period, so that training takes place for several weeks to a month directly prior and directly subsequent to the model application period, and so that the dominant local emissions sources are similar in the model training and application locations. When these conditions are satisfied, ANNs can be robustly implemented, with better performance than LMs.

It is also imperative that sensor users keep in mind the primary importance of minimizing extrapolation of temperature, humidity and sensor signal from model training to application."

Comment: Many of the figures are needlessly complicated by an overload of case studies, unintelligible sensor signal labels, and colours. If there is any way of summarizing this data more cohesively, it would significantly improve the paper.
Response: Thank you very much for the feedback and helping us to simplify and clarify figures.
Edits: According to the specific comments below, we have split what was previously Figure 8 into two figures (now Figure 10 and 11) in order to simplify the graphics and highlight the content of each and simplified and clarified Figure 9 (now Figure 12). We have added definitions for the sensor inputs in the Figure captions for what were Figures 8 and 9 (now Figures 10, 11, and 12). We have also updated Figure 1 (now Figure 2) in order to clarify model training and test periods for each case study, as well as how many U-Pods were included in each case study.

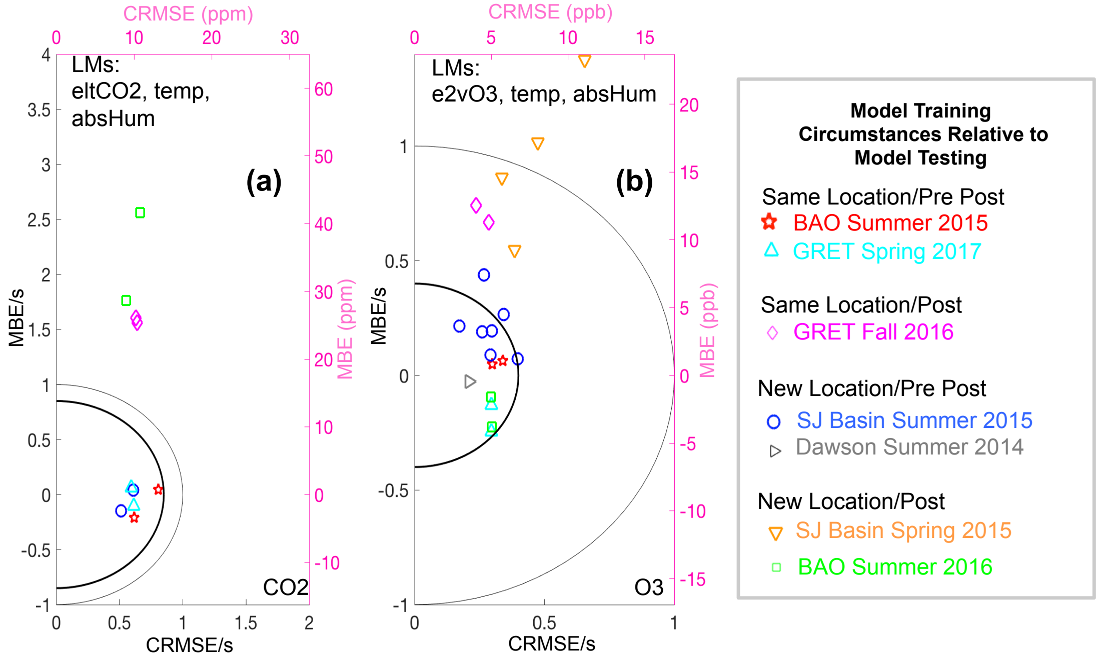


Figure 10: **Target diagrams demonstrating performance of a previously determined best-performing model across all new test datasets. (a) CO2 and (b) O3 LM performance when only the primary gas sensor, temperature and humidity are inputs. (c) CO2 and (d) O3 ANN performance with inputs that were found to perform best at the GRET site in the spring of 2017** (Casey et al., 2017). **Model input definitions: eltCO2 (ELT S300 CO2 sensor), e2vO3 (e2v MiCs-2611 sensor), temp (temperature) , and absHum (absolute humidity).**

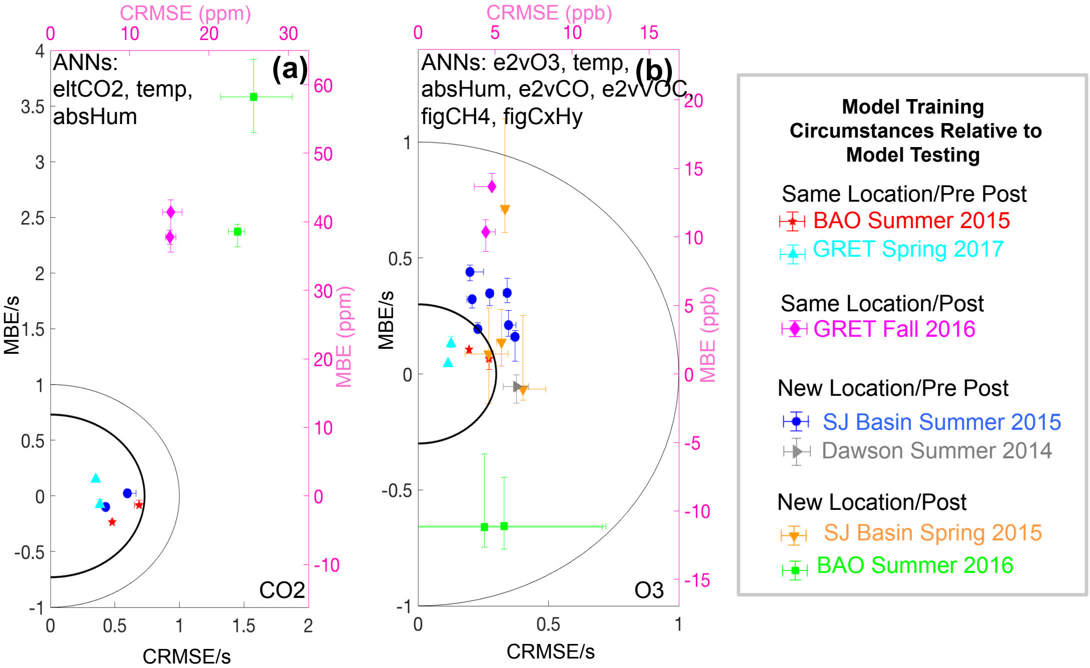
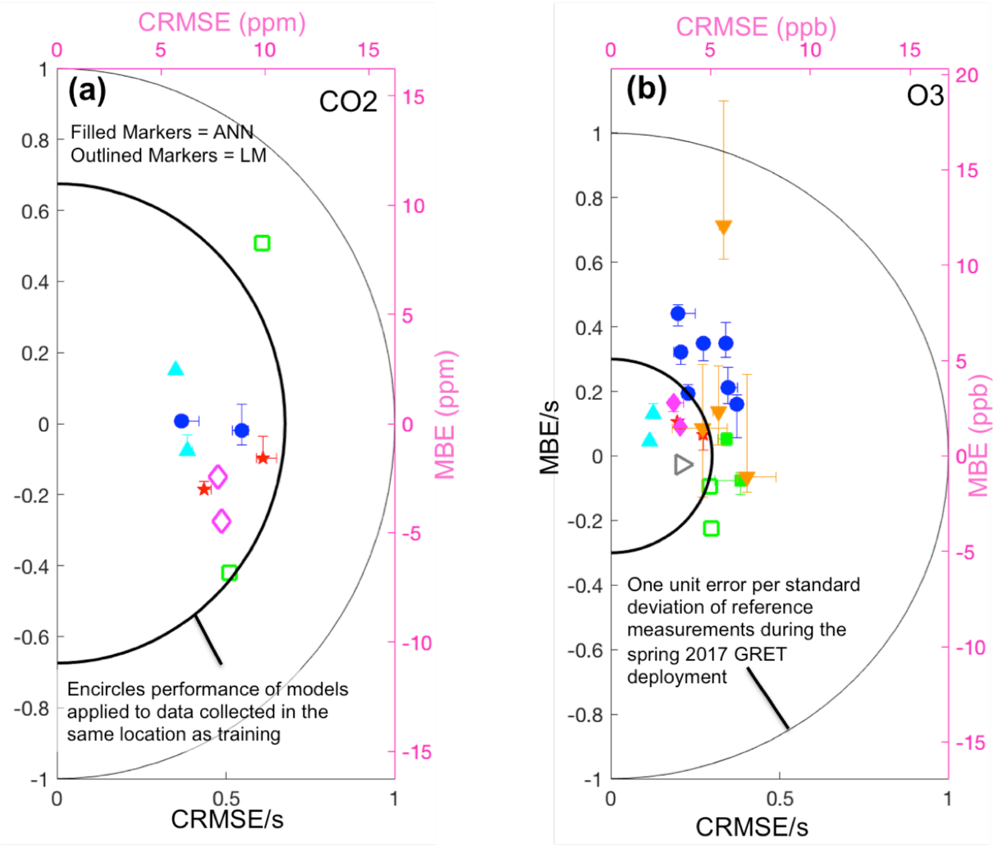


Figure 11: **Target diagrams demonstrating performance of a previously determined best-performing model across all new test datasets (a) CO2 and (b) O3 ANN performance with inputs that were found to perform best at the GRET site in the spring of 2017** (Casey et al., 2017). **Model input definitions: eltCO2 (ELT S300 CO2 sensor), e2vCO (e2v MiCs-5525 sensor), e2vVOC (e2v MiCs-5521 sensor), e2vO3 (e2v MiCs-2611 sensor), figCH4 (Figaro TGS 2600 sensor), figCxHy (Figaro TGS 2602 sensor), temp (temperature) , and absHum (absolute humidity).**



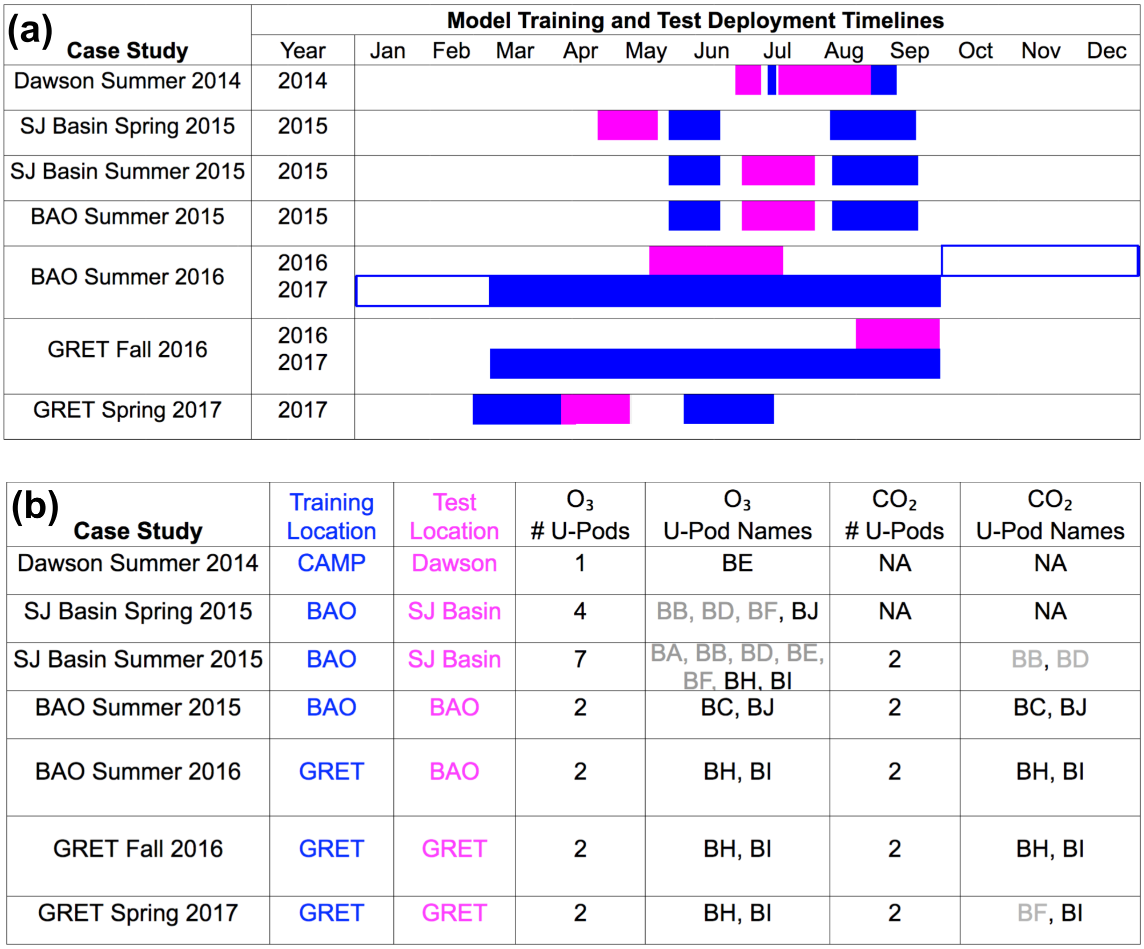
Case Study		CO2 Model Inputs		O3 Model Inputs	Description
BAO Summer 2015	★	eltCO2 temp rh e2vVOC alphaCO	★	e3vO3 temp absHum e2vVOC e2vCO figCH4 figCxHy	Same Location/Pre Post
SJ Basin Summer 2015	●	eltCO2 temp rh e2vVOC alphaCO	●	e3vO3 temp absHum e2vVOC e2vCO figCH4 figCxHy	New Location/Pre Post
SJ Basin Spring 2015			▼	e3vO3 temp absHum e2vVOC e2vCO figCH4 figCxHy	New Location/Post
GRET Spring 2017	▲	eltCO2 temp absHum	▲	e3vO3 temp absHum e2vVOC e2vCO figCH4 figCxHy	Same Location/Pre Post
BAO Summer 2016	□	eltCO2 temp absHum time	□ ■	e3vO3 temp absHum	New Location/Post
GRET Fall 2016	◇	eltCO2 temp absHum rh e2vVOC alphaCO figCH4 figCxHy e2vCO time	◆	e3vO3 temp e2vVOC e2vCO figCH4 figCxHy	Same Location/Post
Dawson Summer 2014			▷	e3vO3 temp absHum	New Location/Pre Post

Figure 12: Target diagrams for (a) CO_2 and (b) O_3 calibration model performance for the best performing model for each particular case when tested on data from a number of field deployments. Model input definitions: eltCO2 (ELT S300 CO2 sensor), e2vCO (e2v MiCs-5525 sensor), e2vVOC (e2v MiCs-5521 sensor), e2vO3 (e2v MiCs-2611 sensor), figCH4 (Figaro TGS 2600 sensor), figCxHy (Figaro TGS 2602 sensor), alphaCO (Alphasense CO-B4 sensor) temp (temperature), absHum (absolute humidity), rh (relative humidity), and time (absolute time).

Comment: It would be good if the authors could elaborate on which U-Pods were where over all these campaigns. Given that temporal degradation / time was investigated in detail in this paper, some assessment of UPod changes over the three years of campaigns would be helpful if possible.

Response: Thanks very much to the reviewer for this helpful comment.

Edits: Figure 2 (previously Figure 1) has been updated to clearly state the number of U-Pods included in each case study, for both O_3 and CO_2, as well as names of the specific U-Pods that were used during each case study. We have also performed an assessment of U-Pod sensor drift from the summer of 2015 through the summer of 2017, shown below in Figure s26, that we have added to the Supplemental Materials. Figure s26 had been sited in section 3.2.2 of the manuscript, and described in the following text: "While we did not measure and record metal oxide sensor heater resistance for sensors included in U-Pods, we have investigated eltCO2 and e2vO3 sensor signal drift from the summer of 2015 through the summer of 2017. These data are presented in Fig. S26. Systematic downward drift in all eltCO2 sensor signals is apparent over this time frame. A clear and consistent pattern of systematic drift over this time period is less apparent for e2vO3 sensors. Since the training data was collected immediately after, the test data period, and since the test data period was relatively short (approximately one month) sensor drift could be negligible across the combined training/testing time frame."



Case Study	Training Location	Test Location	O_3 # U-Pods	O_3 U-Pod Names	CO_2 # U-Pods	CO_2 U-Pod Names
Dawson Summer 2014	CAMP	Dawson	1	BE	NA	NA
SJ Basin Spring 2015	BAO	SJ Basin	4	BB, BD, BF, BJ	NA	NA
SJ Basin Summer 2015	BAO	SJ Basin	7	BA, BB, BD, BE, BF, BH, BI	2	BB, BD
BAO Summer 2015	BAO	BAO	2	BC, BJ	2	BC, BJ
BAO Summer 2016	GRET	BAO	2	BH, BI	2	BH, BI
GRET Fall 2016	GRET	GRET	2	BH, BI	2	BH, BI
GRET Spring 2017	GRET	GRET	2	BH, BI	2	BF, BI

Figure 2: (a) ANN and LM training and test deployment timelines. The Dawson, BAO, and GRET sampling sites are all located in the DJ Basin. Model training periods for each test deployment are shown in blue, and model test periods are shown in magenta. For the BAO Summer 2016 case study, the period outlined in blue shows data that was used to train O_3 model, but not CO_2 models since CO_2 reference data was not available during winter months. (b) Information about each of the case studies presented in the above timelines, including model training and testing locations, as well as the number and names of U-Pods included in each case study for both O_3 and CO_2 models. The U-Pods with names shown in grey were constructed and deployed starting in May of 2014. The U-Pods with names shown in black were constructed and deployed starting in April of 2015.

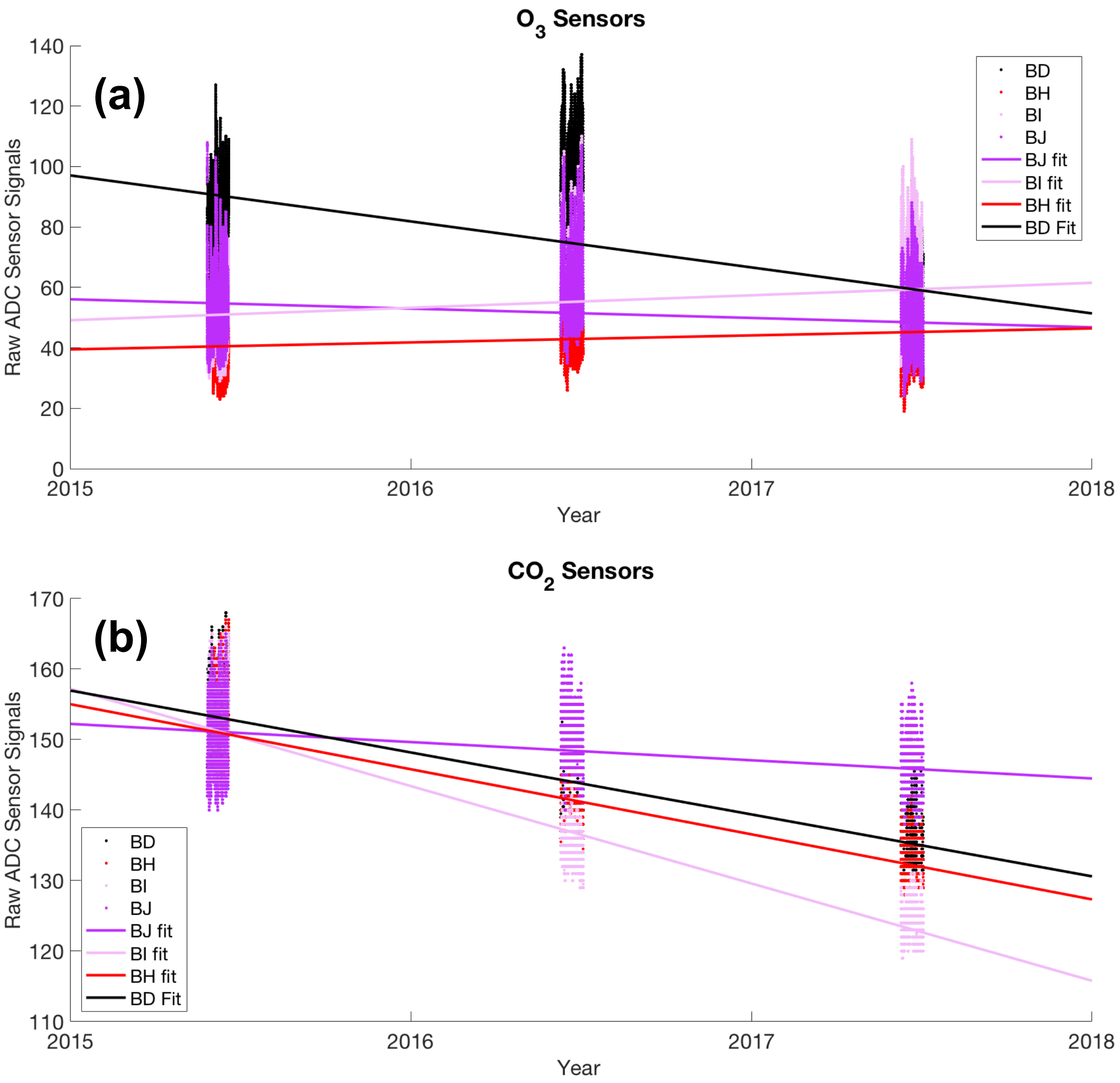


Figure S26 U-Pod sensor drift from 2015 – 2017 for (a) e2v MiCs-2611 O_3 sensors and (b) ELT S300 CO_2 sensors. Data presented are from 23-day periods when U-Pods were co-located together from the summers of 2015, 2016, and 2017. Raw ADC sensor signals were smoothed with rolling hourly medians during these periods, in order to track representative sensor responses across this time period, without the influence of exceptional events. Measurements from summer each year were used to capture sensor response under similar weather conditions.

Comment: Also, when comparing sensor performance spatially, are the U-Pods that are compared the same age?

Response: Thanks to the Reviewer for bringing up this important point.

Edits: We have added the following text accordingly: "Some U-Pods used included in these case studies (indicated in grey font in Fig. 2) were constructed, populated with sensors, and deployed at field sites in the spring of 2014, approximately a year before the rest of the U-Pods were constructed, populated with sensors, and deployed at field sites in the spring of 2015. The relative age of sensor systems included in some case study comparisons could have contributed to some discrepancy in model performance, though systematic differences based on U-Pod age is not apparent."

Comment: Directly addressing the size of the training and testing windows should be included. It is hard to make generalizable conclusions from the study when there is so much variability in training and testing window size. Is there a reason why some training windows are shorter than others? This should be directly addressed in the manuscript.
Response: Thanks to the Reviewer for the helpful feedback.
Edits: We have added the following explanatory text accordingly: "As available data from each case study allowed, we used approximately one month of training data before and after (pre and post of) a given approximately month-long test period. When training data was not available within several months of a test period, significantly longer training datasets were used in order to attempt capture and effectively represent trends in sensor drift over time, as well as to avoid extrapolation of model parameters (particularly temperature) during the test data period. As a result, model-training durations varied across case studies and sometimes significantly exceeded model-testing durations. Each case study is similar in representing approximately one month-long deployment of sensor systems. This study design serves a primary goal of this work, which is to help support the quantification atmospheric trace gases from low-cost gas sensor data in new locations, relative to model training locations, for periods of approximately one month at a time."

Comment: I found the discussion of ANN and LM model building to be significantly under-developed, especially considering that this is a measurement techniques journal. This paper relies too heavily on the prior 2017 study, and has too much assumed knowledge that should be summarized in Section 2.4. The resulting LM and coefficients should be provided. As well as some mention of model performance metrics like MAE or r2.
Response: Thanks very much to the Reviewer for this helpful feedback. We have developed the discussion of ANN and LM model building significantly, through the addition of the following text in section 2.4 and have added a new subsection 2.5 describing the calibration model evaluation and testing we implemented in this work. We have also added a table summarizing model performance metrics for our previous work, as a case study, among other case studies. Since LM coefficients are unique to individual case studies, and within those groups, unique to gas sensors in individual U-Pods, and since we carried out an analysis of model sensitivity to inputs in our previous work, we have not included LM coefficients in this work.

Edits: "As in [1], direct LMs and ANNs were trained with a number of different sensor input sets to map those inputs to target gas mole fractions measured by reference instruments. Direct LMs implemented were multiple linear regression models given by

$$r = p_1 + p_2 s_1 + p_3 s_2 + \ldots + p_n s_{n-1} \qquad (1)$$

where r is the target gas mole fraction (measured by a reference instrument) $s_1 - s_{n-1}$ are sensor signals from U-Pods that are included as model predictor variables, and $p_1 - p_n$ are corresponding predictor coefficients. ANNs designed for regression tasks, like those employed in this work, generally consist of artificial neuron nodes that are connected with weights. Weights are initiated with randomly assigned values. An optimization algorithm is then employed to map a given set input values to one or more corresponding target values. An example of a very simple feed forward neural network, and how weights are propagated through it are depicted in Fig. 3. In this work, ANNs were designed by assigning U-Pod sensor signals to artificial neurons in an input layer and assigning target gas mole fractions for an individual gas species, measured by a reference instrument to a single output neuron. Nonlinear, tansig, artificial neurons in one or two hidden layers and a layer of linear neurons were then added between input layer and the network output neuron. Additionally, bias neurons, each assigned a value of 1, were connected to neurons in the hidden layer(s) so that individual connecting weights could be activated or deactivated during the optimization process. The number of neurons in each hidden layer was set equal to the number of inputs included in a given ANN.

For ANN training we employed the Levenberg Marquardt optimization algorithm with Bayesian Regularization [2]. The Levenberg-Marquardt algorithm provides a combination of Gauss-Newton and Gradient Decent methods, towards incremental minimization of a cost function (the summed squared error between the ANN output and target values as a function of all of the weights in the network). Training begins according to the Gauss-Newton method, in which the Hessian matrix (the second order Taylor series representation of the error surface) is approximated as a function of the Jacobian matrix and its transpose, significantly reducing required training time. Network weights are adjusted accordingly each training step to reduce error. If the cost function is not reduced in a given training step, an algorithm parameter is adjusted so that optimization more closely approximates the gradient decent method (a first order Taylor series representation of the cost function), providing a guarantee of convergence on a cost function minimum. Since local minima may exist across the error surface, it is important to train the same network multiple times (with different randomly assigned starting weights), in order to access the stability of ANN performance. In this work each ANN was trained 5 times. Fig. 4 shows a diagram of an ANN architecture employed in this work, when there were five inputs.

In the implementation of Bayesian Regularization, a term is added to the sum of squared error cost function as a penalty for increased network complexity in order to guard against over fitting. A two level Bayesian inference framework is employed, operating on the assumptions the noise in the training data is independent, normally distributed, and also that all of the weights in the ANN are small, normally distributed, and unbiased [2]. In preliminary ANN tests we found that over fitting occurred even when Bayesian Regularization was used, so we additionally implemented early stopping, which proved to be effective in the reduction of over fitting. To implement early stopping, a portion of training data is set aside as validation dataset, and during training, an ANN is applied to this validation data after each training step. Training continues so long as the error associated with the validation dataset is reduced. When the error associated with the validation dataset is no longer being reduced, training stops early. For ANNs, training datasets were divided in half on an alternating 24-hr basis, with half used for training and half used as validation data for early stopping. ANNs with two hidden layers were used for CO_2 and ANNs with one hidden layer were used for O_3, in accordance with our earlier findings for each target gas species [1]. Input signals for both LMs and ANNs were normalized so that they ranged in magnitude from -1 to 1 since this practice is recommended for the ANN optimization algorithm used [2].

Calibration Model Evaluation and Testing

LM and ANN performance was evaluated on test datasets. To evaluate the performance of each of the ANN and LM models that were generated using training data then applied to test datasets, we used residuals, the coefficient of determination (r^2), root mean squared error (RMSE), mean bias error (MBE), and centered root mean squared error (CRMSE). The CRMSE is an indicator of the distribution of errors about the mean, or the random component of the error. The MBE, alternatively, is an indicator of the systematic component of the error. The sum of the squares of the CRMSE and the MBE is equal to the square of the total error, the square root of which is defined by the RMSE.

First, we generated and applied the best performing model, as determined in our previous work (presented in Table 4), to data from each new case study. Each new case study was selected to challenge models in different ways in order to evaluate the resiliency of the findings from our previous study when challenged by different circumstances.

Next, we generated, applied, and evaluated the performance of a number of LMs and ANNs with different sets of inputs for each case study in order to see which specific model performed the best for each individual case study. The r^2, RMSE, and MBE for each of these alternative models when applied to test data are presented in the supplemental materials (SM) in Fig. S2 through Fig. S7, along with representative scatter plots and time series comparing the performance LMs and ANNs for a given set of inputs. In Fig. S2 through Fig. S7, the best performing model inputs for each train/test data pair are shaded in purple. The type of model that performed the best (ANN vs. LM) is indicated in the caption of each figure. We discuss both the performance of the previously determined best fitting model (generated using data from the GRET Spring 2017 case study) when applied and generated to data from new case studies, and the performance of models that were tuned to perform the best for each individual case study. From these comparisons, we draw insight into circumstances that challenge model performance in terms of relative local emissions characteristics, location, and timing between model training and testing pairs."

"For each of the case studies, we present the performance three groups of models. The first of these are linear models with only the primary gas sensor signal, along with temperature, and absolute humidity as inputs. The next group of models includes those that were found to perform best in our previous work. The third group of models tested for each case study includes models that were optimized specifically for each case study. Tables 5 and 6 show the mean and standard deviation of model performance metrics for each of the case studies presented."

Table 5: O_3 model performance metrics.

Case Study	N	R^2	RMSE (ppb)	MBE (ppb)	Standard Deviation R^2	Standard Deviation RMSE	Standard Deviation MBE
O_3 Models							
Best O_3 Model (Casey et al., 2017) **ANN with inputs: e2vO3 temp absHum e2vVOC e2vCO FigCH4 FigCxHy**							
Dawson Summer 2014	1	0.83	6.46	-0.91	0.00	0.00	0.00
SJ Basin Spring 2015	4	0.86	7.74	3.69	0.05	3.82	5.78
SJ Basin Summer 2015	7	0.85	7.03	4.89	0.10	1.10	1.73
BAO Summer 2015	2	0.93	4.26	1.45	0.00	0.31	0.07
BAO Summer 2016	2	0.92	12.21	-11.14	0.00	0.31	0.07
GRET Fall 2016	2	0.96	12.87	12.02	0.01	2.30	2.35
GRET Spring 2017	2	0.98	2.59	1.49	0.00	0.69	1.02
Simple Model (Single Gas Sensor) **LM with inputs: e2vO3 temp absHum**							
Dawson Summer 2014	1	0.95	3.59	-0.46	0.00	0.00	0.00
SJ Basin Spring 2015	4	0.83	17.95	16.09	0.06	6.10	5.83
SJ Basin Summer 2015	7	0.86	6.30	3.53	0.06	1.40	2.06
BAO Summer 2015	2	0.87	5.50	0.94	0.00	0.78	1.56
BAO Summer 2016	2	0.89	5.78	-2.71	0.00	0.78	1.56
GRET Fall 2016	2	0.93	12.73	11.92	0.01	0.62	0.88
GRET Spring 2017	2	0.89	6.00	-3.19	0.00	0.73	1.38
Models Optimized For Case Studies							
Dawson Summer 2014	1	0.95	3.59	-0.46	0.00	0.00	0.00

SJ Basin Spring 2015	4	0.86	7.74	3.69	0.05	3.82	5.78
SJ Basin Summer 2015	7	0.85	7.03	4.89	0.10	1.10	1.73
BAO Summer 2015	2	0.93	4.26	1.45	0.02	0.51	1.54
BAO Summer 2016	2	0.87	6.25	-0.20	0.02	0.51	1.54
GRET Fall 2016	2	0.95	3.99	2.14	0.00	0.28	0.89
GRET Spring 2017	2	0.98	2.59	1.49	0.00	0.69	1.02

Table 6: CO_2 model performance metrics.

Case Study	N	R^2	RMSE (ppm)	MBE (ppm)	Standard Deviation R^2	Standard Deviation RMSE	Standard Deviation MBE
CO_2 Models							
Best CO_2 Model from (Casey et al., 2017) **ANN with inputs: eltCO2 temp absHum**							
SJ Basin Summer 2015	2	0.65	8.42	-0.62	0.00	1.81	1.41
BAO Summer 2015	2	0.75	9.98	-2.60	0.05	13.00	13.89
BAO Summer 2016	2	0.69	54.38	48.37	0.05	13.00	13.89
GRET Fall 2016	2	0.74	42.37	39.58	0.02	2.44	2.57
GRET Spring 2017	2	0.83	6.31	0.59	0.03	0.13	2.61
Simple Model (Single Gas Sensor) **LM with inputs: eltCO2 temp absHum**							
SJ Basin Summer 2015	2	0.71	7.84	0.27	0.01	1.43	0.42
BAO Summer 2015	2	0.69	10.62	-1.26	0.06	1.52	10.67
BAO Summer 2016	2	0.73	11.82	0.73	0.06	1.52	10.67
GRET Fall 2016	2	0.82	8.62	-3.46	0.00	0.69	1.45
GRET Spring 2017	2	0.55	9.88	-0.33	0.03	0.29	1.91
Models Optimized For Case Studies							
SJ Basin Summer 2015	2	0.72	7.45	-0.11	0.04	2.06	0.31
BAO Summer 2015	2	0.80	8.85	-2.29	0.10	6.47	7.08
BAO Summer 2016	2	0.73	11.82	0.73	0.06	1.52	10.67
GRET Fall 2016	2	0.82	8.62	-3.46	0.00	0.69	1.45
GRET Spring 2017	2	0.83	6.31	0.59	0.03	0.13	2.61

Comment: It would be good to include an explicit discussion of % reduction in error by using established models vs. "best fit" models. Can we generalize? What is the quantitative impact of using your prior models vs making a new model every time. My interpretation from this paper is that we need a new model for every U-Pod for every deployment – is there any way around this? I feel there is a significantly missed opportunity to be quantitative here. Section 3.3 could be substantially enhanced using some sort of summary figure/table (other than a target diagram) that gives percent change in bias, random error, r2, mae etc. by switching from pre/post to just post, or by switching location. Given that there are many pairs of sensors looking at impact of pre/post vs. just post or impact of location switching, you could show average % change in model fitting statistics as well as confidence intervals or standard deviations to show the spread across the case studies. This might be a helpful way of streamlining the paper.

Response: Thanks very much to the Reviewer for this helpful comment, which will help us be more quantitative as well as clarify and focus the narrative and results we present.

Edits: Accordingly, we have added a table showing the percent change in R2, RMSE, and MBE, when one set of models is used instead of another, as well as the following text: "Table 7 shows the percent change in model performance metrics when one model-training paradigm is used in place of another, highlighting relative benefits associated with the implementation of different models for O_3 and CO_2."

Table 7: **Relative benefits associated with the implementation of different models for O_3 and CO_2.**

Case Study	Mean % Increase in R^2	Mean % Decrease in RMSE	Mean % Decrease in MBE	Mean % Increase in R^2	Mean % Decrease in RMSE	Mean % Decrease in MBE
	CO_2 Models			O_3 Models		
Benefit of Models Optimized For Case Studies Over The Best Models from (Casey et al., 2017)						
Dawson Summer 2014				14.51	44.42	50.00
SJ Basin Spring 2015				0.00	0.00	0.00
SJ Basin Summer 2015	10.56	11.52	82.60	0.00	0.00	0.00
BAO Summer 2015	5.84	11.27	11.95	0.00	0.00	0.00
BAO Summer 2016	5.72	78.27	98.49	-5.01	48.82	98.19
GRET Fall 2016	11.17	79.66	108.73	-0.54	68.99	82.22
GRET Spring 2017	0.00	0.00	0.00	0.00	0.00	0.00
Benefit of The Best Models from (Casey et al., 2017) Over Simple Linear Models						
Dawson Summer 2014				-12.67	-79.92	-99.99
SJ Basin Spring 2015				3.20	56.88	77.09
SJ Basin Summer 2015	-8.41	-7.29	331.39	-1.34	-11.53	-38.41
BAO Summer 2015	8.70	6.05	-106.48	6.79	22.48	-53.85
BAO Summer 2016	-5.41	-360.09	-6543.84	2.57	-111.22	-310.71
GRET Fall 2016	-10.05	-391.73	1244.99	2.88	-1.12	-0.86
GRET Spring 2017	51.92	36.13	278.55	10.00	56.90	146.65
Benefit of Models Optimized For Case Studies Over Simple Linear Models						
Dawson Summer 2014				0.00	0.00	0.00
SJ Basin Spring 2015				3.20	56.88	77.09
SJ Basin Summer 2015	1.26	5.06	140.25	-1.34	-11.53	-38.41
BAO Summer 2015	15.04	16.64	-81.80	6.79	22.48	-53.85
BAO Summer 2016	0.00	0.00	0.00	-2.57	-8.10	92.59

GRET Fall 2016	0.00	0.00	0.00	2.33	68.64	82.07
GRET Spring 2017	51.92	36.13	278.55	10.00	56.90	146.65

Comment: This is mentioned in the specific comments, but I would like to see a quantitative assessment of the impact of swapping out RH data if a U-Pod failed. You could accomplish this by taking a U-Pod with valid RH data, replacing it with the Picarro or nearby station RH data, and quantitatively assessing the impact on model performance. That way, you could transition from hypotheticals about the impact of this data swapping to some actual numbers.

Response: Thanks very much to the reviewer for helping us to be less hypothetical about the impact of this data swapping. We have carried out a dummy experiment, testing the effect of this humidity data swapping on data collected during the GRET Spring 2017 case study. A figure, showing the relative performance of models when the humidity data was taken from the U-Pods directly and replaced with measurements from the Picarro CRDS, has been added to the Supplemental Materials. This figure, and associated implication have been cited in the main text.

Edits: "In our previous work, we showed that O_3 models were very sensitive to the humidity signal input (Casey et al., 2017). In this case study, it seems that replacing actual humidity signals with closely approximated humidity signals, negatively influenced model performance. In order to investigate this observation further, we tested the influence of replacing humidity data in the same manner, using mixing ratios from the same co-located Picarro, on test data from the GRET Spring 2017 case study. A comparison of model performance under normal and this 'borrowed RH' circumstance are presented in Fig. S27 in the SM. O_3 model performance was negatively impacted when 'borrowed' RH values based on Picarro data replaced U-Pod RH sensor signals. From these findings, it seems likely that the inclusion of multiple metal oxide type sensors as inputs in the model, which all respond strongly to humidity fluctuations, helped the ANN to effectively represent the influence of humidity in the system, more so than including a 'borrowed RH' signal from another instrument. We tested models with multiple gas sensor signals and no humidity signal as inputs for a number of other case studies as well (as seen in Fig. S2, Fig. S4, and Fig. S5), when good humidity data from U-Pod enclosures was available, but they did not turn out to be the best performing model in any of these other tests."

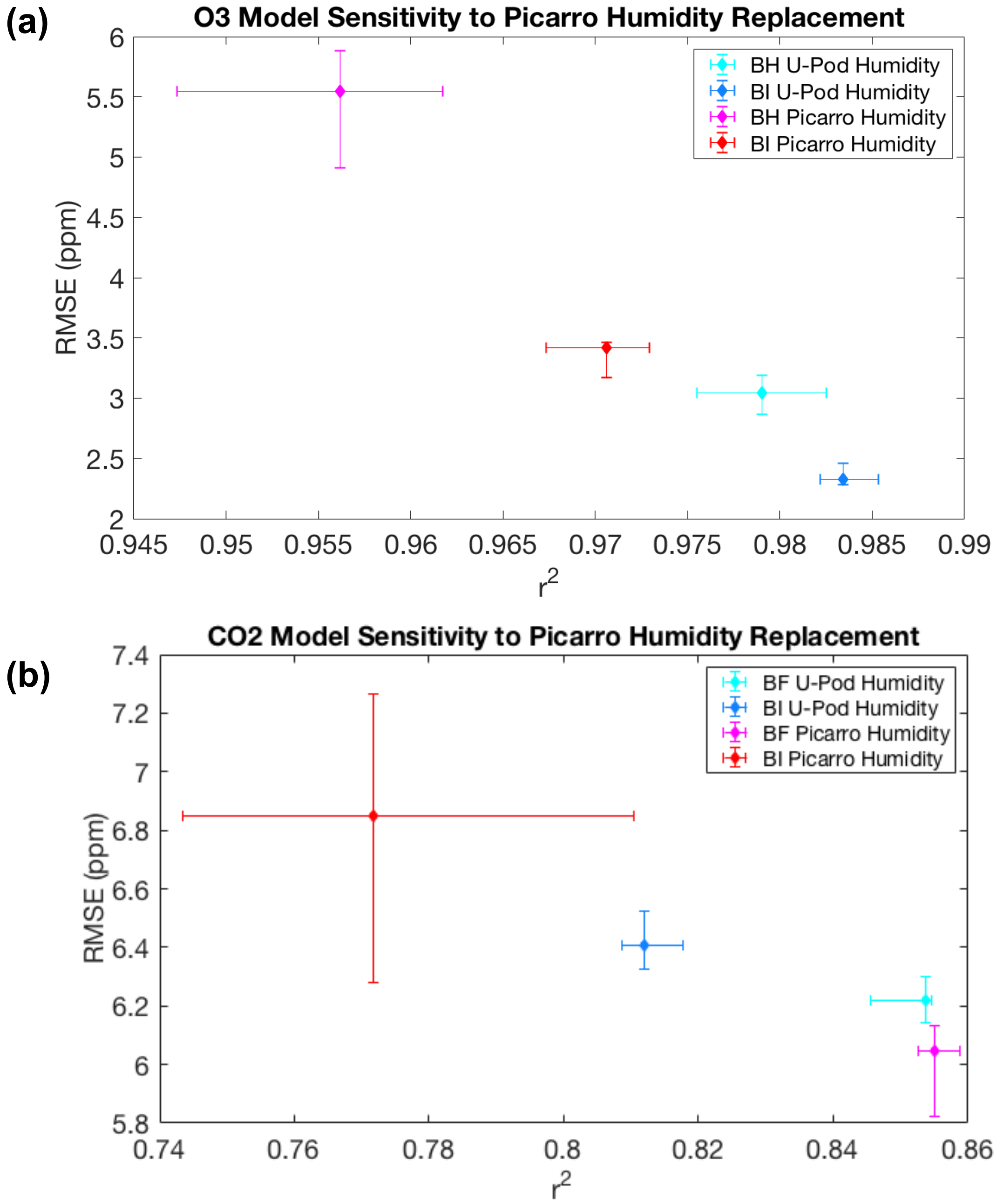


Figure S27 A comparison of model performance when humidity inputs are taken from sensor measurements collected within a given U-Pod sensor system enclosure, vs the performance of models when humidity inputs are replaced using data from a Picarro CRDS for (a) O_3 (b) CO_2

SPECIFIC COMMENTS

Comment: P1 - L13-14: Seems like an oxymoron to say "Generally" if the circumstances for best model performance are case study specific. Recommendation to remove the word "generally".

Response: Thank you for the helpful feedback**.**

Edits: We have removed the word 'Generally'.

Comment: P3 - L19-24: Please discuss why ozone is elevated near O&G production.

Response: Thank you for helping us clarify why ozone can be elevated near oil and gas production activities.
Edits: We have added the following text to augment this discussion: "NO_X and VOC emissions, including those from oil and gas production activities, react in the atmosphere in the presence of sunlight to form tropospheric O_3."
"Emissions of industry related air pollutants, including O_3 precursors, NO_X and VOCs are expected to occur on spatially distributed scales, across multiple individual components on individual well pads, transmission lines, transportation routes, and gathering stations that are each distributed throughout production basins (Litovitz et al. 2013; Mitchell et al. 2015; Allen et al. 2013). Spatially distributed networks of low-cost sensors have the potential to better inform spatial variability of air quality than existing Regulatory air quality monitoring stations which feasibly cover such spatially resolved measurements continuously, and may not be representative of air quality across smaller spatial scales (Bart et al., 2014; Jiao et al., 2016; Moltchanov et al., 2015)."

Comment: P3 – L27: Can you quantify "small spatial scales" in this context? Is well pad combustion and diesel traffic really contributing so much that it is universally increasing ozone? Most of the construction traffic would occur during active drilling and less so during production when well pad sites are very quiet. I think some further thinking or elaboration on this train of thought it warranted.
Response: Thank you for the helpful comment.
Edits: We have added the following detail, regarding spatial scales that ozone may be influcened near oil and gas emissions sources: " a modeling study concluded that oil and gas production activities could significantly impact ozone near emissions sources, beginning 2 and 8 km downwind of compressor engine and flaring activities, respectively [3]."
We have also added the following text to address how emissions may change across the lifetime of a given oil and gas production well: "While emissions from truck traffic (and in some cases drilling rig generators), at a given well pad are highest during the drilling, stimulation, and completion phases, industry truck traffic often persists as produced water and condensate tanks are collected from storage tanks on a well pad throughout the life a the well, as do emissions from flaring and compressor engines."

Comment: P4 – L1-2: What do you mean by "pooling" of compounds - I am not sure I understand this sentence.
Response: Thank you for helping us clarify this statement.
Edits: We have made the following edits accordingly: "While elevated ambient CO_2 levels are not directly harmful to human health, continuous CO_2 measurement can provide information about nearby combustion-related pollution and atmospheric dynamics that lead to the accumulation of potentially harmful compounds associated with the oil and gas production industry during periods of atmospheric stability."

Comment: P4 – L5-8: I think some short discussion of the operating principles of the sensors would be helpful here.
Response: Thank you for the helpful feedback.
Edits: We have added a discussion of the operating principles of the sensors to section 1.1 accordingly:
"While low-cost sensors have been emerging on the market with sufficient sensitivity to resolve variations in ambient mole fractions of target gases of interest, they are also sensitive to temperature and humidity variations that occur in the ambient environment. NDIR sensors, like the ELT s300 CO_2 sensor employed in this study, have good selectivity, but, since pressure and temperature are not controlled in the optical cavity of ELT s300 CO_2 sensors, the influence of temperature on sensor signals plays an important role. The influence of humidity is also important to address because changes in water vapor are known to influence NDIR measurements of CO_2 in terms of spectral cross-sensitivity due to absorption band broadening (Licor, 2010).

Both metal oxide and electrochemical type sensors operate on the principle of oxidizing or reducing reactions at sensor surfaces. For electrochemical sensors, like the Alphasense CO-B4 sensor employed in this study, oxidizing or reducing compounds react at the working electrode, resulting in the transfer of ions across an electrolyte solution from the working electrode to the counter electrode, balanced by the flow of electrons across the circuit connecting the working electrode to the counter electrode. A linear relationship is expected between this current and the target gas mole fraction. Electrochemical sensors can be tuned to respond more or less strongly to specific gases by adjusting the materials properties of the working electrode. A membrane is located between the working electrode and the exterior of the sensor in order to control redox reaction rates. Gases diffusion through the membrane to reach the working electrode and the electron transfer rates have been shown to increase at higher temperatures (Xiong and Compton, 2014), and since chemical reaction rates are also influenced by temperature, electrochemical sensor responses can be influenced by sensor operating temperature. Changes in ambient humidity levels can cause sensors to loose or gain of the electrolyte solution, by mass, also influencing electrochemical sensor response (Xiong and Compton, 2014).

For metal oxide sensors, and to a lesser extent for electrochemical sensors, resolving the response of a sensor attributable to the target gas species can also pose a challenge in the presence of interfering gas species. Metal oxide sensors, like those used in this study, have a resistive heater circuit that warms up the sensor surface, causing O_2 molecules to adsorb to the sensor surface, which leads to increased resistance across the surface of the sensor. In the presence of an oxidizing compound, like O_3, more oxygen molecules are adsorbed to the sensor surface and the resistance across the sensor surface in increased further. In the presence of a reducing compound, like CO, oxygen molecules are removed from the sensor surface, allowing electrons to flow more freely, resulting in decreased resistance across the sensor surface. For metal oxide sensors, the resistance across the sensor surface can then be used to determine the mole fraction of a given oxidizing or reducing compound, often according to a nonlinear relationship. Exposure to humidity has been shown to significantly lower the sensitivity of metal oxide gas sensors making it an important parameter to address in a gas quantification model (Wang et al., 2010). Metal oxide sensor operating temperature has also been shown to strongly influence sensor sensitivity and selectivity to different gas species (Wang et al., 2010). Metal oxide type sensors can be tuned to respond differently from one another to oxidizing and reducing gas species by using different metal oxide materials and doping agents for the sensor surface, but selectivity is difficult to achieve."

We have also added a section to the introduction, section 1.2, entitled "Low-Cost Air Quality Sensor Quantification:

"Because low-cost gas sensor signals are influenced, sometimes significantly, by interfering gas species and changing weather conditions in the ambient environment, field normalization methods to quantify atmospheric trace gases using low-cost sensors have been found to be more effective than lab calibration (Cross et al., 2017; Piedrahita et al., 2014; Sun et al., 2016). Our previous study and several others have compared the efficacy field calibration models generated using LMs (simple and multiple linear regression) relative to supervised learning methods (including ANNs and random forests), all finding that ANNs (Casey et al., 2017; Spinelle et al., 2015, 2017) and random forests (Zimmerman et al., 2017) outperformed LMs in the ambient field calibration of low-cost sensors. Like earlier laboratory based studies (Brudzewski, 1999; Gulbag and Temurtas, 2006; Huyberechts and Szeco, 1997; Martín et al., 2001; Niebling, 1994; Niebling and Schlachter, 1995; Penza and Cassano, 2003; Reza Nadafi et al., 2010; Srivastava, 2003; Sundgren et al., 1991), ANN-based calibration models, incorporating signals from an array of gas sensors with overlapping sensitivity as inputs, have been able to effectively compensate for the influence of interfering gas species and resolve the target gas mole fraction.

ANNs are known to be able to very effectively represent complex, nonlinear, and collinear relationships among input and output variables in a system (Larasati et al., 2011). ANNs are useful in the field calibration of low-cost sensors because, through pattern recognition of a training dataset, they are able to effectively represent the complex processes and relationships among sensors and the ambient environment that would be very challenging to represent analytically or based on empirical representation of individual driving relationships. In practice though, the reason multiple gas sensors are able to improve the performance of calibration models may be in part the result of correlation between mole fractions of target gases themselves that hold for one model training location, but might not remain effective at alternative sampling sites or during other time periods."

Comment: P4 – L25: Not sure what is meant by "toward" here
Response: Thank you for the helpful feedback.
Edits: We have replaced "toward" with "that were used for"

Comment: P5 – L27: Is this "clean air" normalization done dynamically/in real-time in parallel with the actual measurement? Or is the clean air measurement established during some calibration/maintenance? Please clarify.
Response: Thank you for helping us clarify.
Edits: We have added the following text accordingly: "For metal oxide type sensors, voltage signals were converted into resistance, and then normalized by the resistance of the sensor in clean air, R_0. A single value for R_0 was used for each sensor across the study duration. This R_0 value was taken as the resistance of each sensor at the GRET field deployment site when the target pollutant had approached background levels (at night for the metal oxide O_3 sensors and midday for all other metal oxide sensors), and when the ambient temperature was approximately 20° C and relative humidity of approximately 25%."

Comment: P5 – L30-32: Is there expected to be spatial variability of RH?
Response: Thank you for helping us to clarify.
Edits: We have added the following text to section 2.3: "The closest U-Pod with good humidity sensors ranged from several feet, when U-Pods were co-located during deployments in the DJ Basin, to approximately fifty miles during deployments in the San Juan Basin."

In Section 3.1 of we have added this text also: "Since the Ignacio site was located approximately twenty-two and fifty miles away from the Navajo Dam and Sub Station sites respectively, this could have introduced some additional error into the application of a calibration equation, particularly since we showed earlier that O_3 ANNs like the ones we employed here are very sensitive to humidity inputs (Casey et al., 2017). Spatial variability in humidity across tens of miles could be significant as isolated storms (which are on average 15 miles in diameter) propagate throughout the region in the summer."

Comment: P5 – L30-32: Why not just replace the RH sensors directly?
Response: Good question.
Edits: In answer, we have added the following text: "RH sensors were not replaced during field deployments in order to preserve consistency across different deployment periods, allowing for the possibility of a single comprehensive model to apply to all data from a single U-Pod. After some experimentation in generating a 'master model' that could be applied to data from a given U-Pod for all collected field measurements, across several years, we determined that individual models for each deployment would be more effective, and replacing RH sensors that had drifted down would have been appropriate in support of the methods presented here. We have since upgraded to Sensirion AG SHT25 sensors, which appear to be more robust and consistent over the course of long-term field deployments."

Comment: P7 – Section 3.0 first paragraph – Are there some general conclusions from the SM that you can discuss here? Some discussion of model performance is warranted vs just describing what figures are in the SM.
Response: Thanks very much for the feedback.
Edits: We have moved the paragraph in question to the methods section and have added the following sentence, letting the reader know that these plots are discussed in the results and

discussion section in context with each case study presented: "The best-performing model for each case study are highlighted below in the Results and Discussion section."

Comment: P8 – L17: What is eltCO2?? Can you better define all the model parameter inputs? This comes up in Figure 9 as well.
Response: Yes, thank you for the feedback.
Edits: Description added here and at the first mention of other model input codes in the manuscript in the text: "eltCO2 (ELT S300 CO2 sensor)"
We have also defined these model input codes in the caption for Figure 9 (now Figure 12) as well as Figure 8 (now Figure 10 and Figure 11): "Model input definitions: eltCO2 (ELT S300 CO2 sensor), e2vCO (e2v MiCs-5525 sensor), e2vVOC (e2v MiCs-5521 sensor), e2vO3 (e2v MiCs-2611 sensor), figCH4 (Figaro TGS 2600 sensor), figCxHy (Figaro TGS 2602 sensor), alphaCO (Alphasense CO-B4 sensor), temp (temperature), absHum (absolute humidity), rh (relative humidity), and time (absolute time)."

Comment: P8 – L30-33: Is this early morning under prediction really true? Bloomfield doesn't look like it is exhibiting any diurnal variation in residual error at all… I feel like given the small number of U-Pods, it is hard to make this conclusion definitively.
Response: We agree, thank you for pointing the trend in Bloomfield out.
Edits: We have edited the text to say "three of four U-Pods" instead of "all four U-Pods"

Comment: P9 – L4-13: I am confused now – why did you use the model with three inputs (eltCO2, abshum, and temp) if the best performing model had more variables? I feel like the model selection discussion is substantially underdeveloped. There could be many good reasons to not choose a more complex model, but any discussion of this seems to be completely omitted, or the reasoning is too difficult to follow.
Response: Thank you for the feedback and for helping us to clarify. You have helped us see that some important details were missing from the methods section regarding model selection and testing procedures.
Edits: We have added a subsection 2.5 to the end of the methods section entitled "Calibration Model Evaluation and Testing". In this section, we first define the r2, RMSE, MBE, and CRMSE metrics that are used to evaluate the performance of a given model when it is applied to a test dataset. Next we added a paragraph describing how we first tested models that were found to perform best for each gas species in our previous work, and then evaluated the performance of the best model for each specific case study. We then describe the methodology behind model selection and testing for each case study, in the following text and in the newly added Table 4:

"First, we generated and applied the best performing model, as determined in our previous work (presented in Table 3), to data from each new case study. Each new case study was selected to challenge models in different ways in order to evaluate the resiliency of the findings from our previous study when challenged by different circumstances.

Table 3: Best performing models, as determined for each gas species, in the previous study (Casey et al., 2017)

Gas Species	Model Type	Sensor Signal Model Inputs	
CO_2	ANN	eltCO2 temp absHum	(ELT S300 CO2 Sensor) (temperature) (absolute humidity)
O_3	ANN	e2vO3 e2vCO e2vVOC figCH4 figCxHy	(e2v MiCs-2611) (e2v MiCs-5525) (e2v MiCs-5521) (Figaro TGS 2600) (Figaro TGS 2602)

		temp (temperature) absHum (absolute humidity)

Next we tested LMs for CO2 and O3 that contained only the primary target gas sensor for each species, as well as temperature and absolute humidity as inputs. Finally, we generated, applied, and evaluated the performance of a number of LMs and ANNs with different sets of inputs for each case study in order to see which specific model performed the best for each individual case study. The r^2, RMSE, and MBE for each of these alternative models when applied to test data are presented in the supplemental materials (SM) in Fig. S2 through Fig. S7, along with representative scatter plots and time series comparing the performance LMs and ANNs for a given set of inputs. In Fig. S2 through Fig. S7, the best performing model inputs for each train/test data pair are shaded in purple. The type of model that performed the best (ANN vs. LM) is indicated in the caption of each figure. We discuss both the performance of the previously determined best fitting model (generated using data from the GRET Spring 2017 case study) when applied and generated to data from new case studies, and the performance of models that were tuned to perform the best for each individual case study. From these comparisons, we draw insight into circumstances that challenge model performance in terms of relative local emissions characteristics, location, and timing between model training and testing pairs. Table 4 lists the relative timing and parameter coverage between model training and testing periods for dataset pairs, highlighting instances of incomplete coverage during training that led to model extrapolation during testing."

Comment: P9 – L28-30: It would be good to do a more comprehensive assessment of the impact of replacing RH sensor signal on model performance. Could you conduct a dummy experiment where you replace RH data you actually logged with that of a nearby or alternate monitor and then quantify the impact on model outcome? Given that it seems that a) RH/abshum is an important variable and b) that you had significant data loss issues, I feel a more quantitative assessment of the impact of these data substitutions is needed.

Response: Thanks very much to the reviewer for helping us to be less hypothetical about the impact of this data swapping. We have carried out a dummy experiment, testing the effect of this humidity data swapping on data collected during the GRET Spring 2017 case study. A figure, showing the relative performance of models when the humidity data was taken from the U-Pods directly and replaced with measurements from the Picarro CRDS, has been added to the Supplemental Materials. This figure, and associated implication have been cited in the main text.

Edits: "The fall 2016 GRET test period coincided with the time period U-Pod absolute humidity was replaced using mixing ratios from a co-located Picarro due to missing humidity sensor data. Interestingly, when this 'borrowed' humidity signal was not included as an input, the model performance markedly increased and became competitive with other 'same location' test deployment case studies. In our previous work, we showed that O_3 models were very sensitive to the humidity signal input (Casey et al., 2017). In this case study, it seems that replacing actual humidity signals with closely approximated humidity signals, negatively influenced model performance. In order to investigate this observation further, we tested the influence of replacing humidity data in the same manner, using mixing ratios from the same co-located Picarro, on test data from the GRET Spring 2017 case study. A comparison of model performance under normal and this 'borrowed RH' circumstance are presented in Fig. S27 in the SM. O_3 model performance was negatively impacted when 'borrowed' RH values based on Picarro data replaced U-Pod RH sensor signals. From these findings, it seems likely that the inclusion of multiple metal oxide type sensors as inputs in the model, which all respond strongly to humidity fluctuations, helped the ANN to effectively represent the influence of humidity in the system, more so than including a 'borrowed RH' signal from another instrument. We tested models with multiple gas sensor signals and no humidity signal as inputs for a number of other case studies as well (as seen in Fig. S2, Fig. S4, and Fig. S5), when good humidity data from U-Pod enclosures was available, but they did not turn out to be the best performing model in any of these other tests."

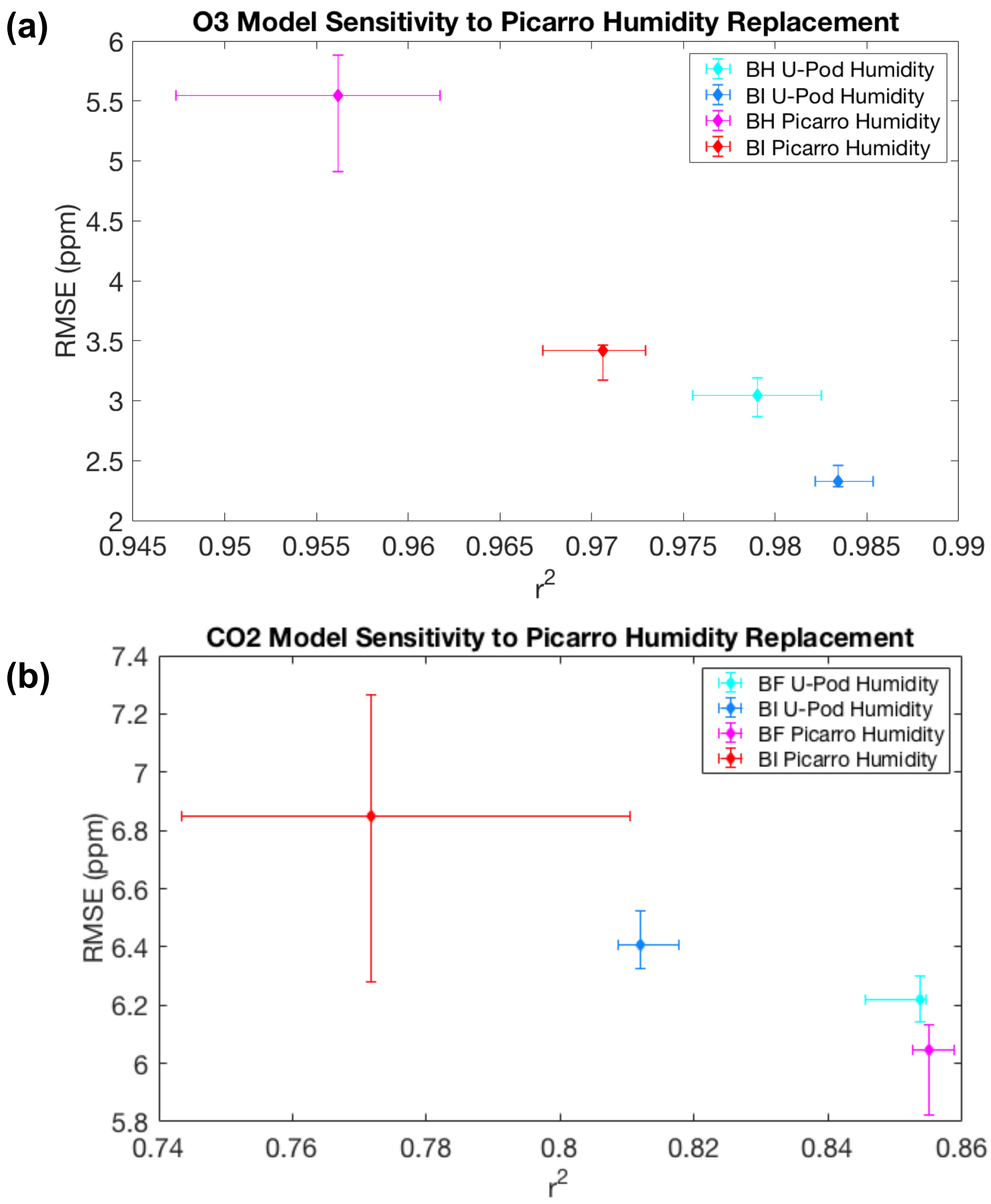


Figure S27 A comparison of model performance when humidity inputs are taken from sensor measurements collected within a given U-Pod sensor system enclosure, vs the performance of models when humidity inputs are replaced using data from a Picarro CRDS for (a) O_3 (b) CO_2

Comment: P11 – L2-13: Can you comment on the quality of the fit at Dawson vs CAMP in addition to the ideal model. The discussion is fairly qualitative. Also, the LM should be much better at extrapolating vs the ANN (which cannot extrapolate I think…not sure) – can you comment on this difference? Does LM perform better because it can extrapolate?

Response: Thank you for this useful comment. We agree it is true that LMs should accommodate extrapolation more effectively than ANNs, and that a quantitative description of model performance is warranted.

Edits: We have added the following text, accordingly: "The fact that LMs performed better than ANNs in this case (with an r2 of .95 and RMSE of 0.35 ppb for LMs, as opposed to an r2 of .9 and an RMSE of 5.1 ppb for ANNs) may have to do with the general expectation that LMs be more resilient to extrapolation than ANNs."

Comment: P11 – Section 3.2.2: If the calibration is immediately after deployment, I am not surprised that there wasn't much of an effect. Do you anticipate there should be a significant time effect on such short time scales? What is the lifespan of the U-Pods?
Response: Thank you for the useful comment.
Edits: We have added the following text accordingly: "Gas sensor manufactures don't clearly define sensor lifetimes, but sensors are generally expected to loose sensitivity over time. For example, Alphasense CO-B4 electrochemical sensors are expected to have 50% of their original sensitivity after two years (Alphasense, 2015). The heater resistance in a give metal oxide type sensor is expected to drift over time, influencing sensor measurements (e2v Technologies Ltd., 2007). Masson and colleagues observed a significant drift in a metal oxide sensor heater resistance over the course of a 250 day sampling period in a laboratory setting (Masson et al., 2015). While we did not measure and record metal oxide sensor heater resistance for sensors included in U-Pods, we have investigated eltCO2 and e2vO3 sensor signal drift from the summer of 2015 through the summer of 2017. These data are presented in Fig. S26. Systematic downward drift in all eltCO2 sensor signals is apparent over this time frame. A clear and consistent pattern of systematic drift over this time period is less apparent for e2vO3 sensors. Since the training data was collected immediately after, the test data period, and since the test data period was relatively short (approximately one month) sensor drift could be negligible across the combined training/testing time frame."

Comment: P11 – L25: I am confused by the introduction of discussion around figCxHy – should we expect this sensor to play an important role?
Response: Thank very much to the reviewer for helping us to clarify.
Edits: We have added the following text: "Again the model for O_3 that was found to perform best in our previous (Casey et al., 2017), an ANN with temp, absHum and all metal oxide sensor signals as inputs, performed the best at sites included in this case study, with one exception. At the Sub Station site the inclusion of the figCxHy sensor signal decreased model performance. Additionally, the performance of all models tested at the Sub Station site during the SJ Basin Spring 2015 deployment was significantly worse in terms of MBE than model performance at other sites, both LMs and ANNs with different sets of inputs. Since this sensor signal input augmented model performance at the same sampling location during the summer deployment period, this finding could be attributable to the extrapolation with respect to temperature that occurred during the test period of this case study. As discussed in the introduction, metal oxide sensor sensitivity to different gas species can vary along with sensor surface temperature. Models were trained to use the figCxHy sensor signal, across the ambient temperatures in encompassed by the training data, to help account for the influence of confounding gas species at the BAO site. We think it is possible that the different temperatures in combination with the unique mix of gas species present at the Sub Station site, which the figCxHy sensors are highly sensitive to, caused the ANN to perform worse."

Comment: P12 – L5: How long is "so long"? This is related to my comment on
Response: Thank you for helping us to be more specific.
Edits: We have changed "so long" to "several months".

P11 – Section 3.2.2.

Comment: P12 – L21-24: I am confused – did you switch to humidity measured by Picarro or omit humidity entirely? It is not clear to me what happened here.
Response: Thank you for pointing out this confusion.
Edits: We have more clearly described the humidity replacement process (if any) for each individual case study dataset pair in the methods section. We have additionally added to the text in section 2.3 as follows: "Water mole fractions measured by the Picarro were converted into mass-based mixing

ratios to match the units of the absolute humidity signal in the U-Pod data. We applied an adjustment to this absolute humidity signal so that it matched observations in U-Pods during the following month when good RH sensor data was available, to account for the fact that temperatures were higher in U-Pod enclosures than the ambient environment. We then replaced the relative humidity signal in each U-Pod from August 23rd through October 1st in 2016 with the mixing ratios derived from Picarro measurements. Using the temperature and pressure logged in each U-Pod along with the absolute humidity from the Picarro, relative humidity was calculated for each U-Pod during this period."

And to section 3.2.4: "Interestingly, when humidity this 'borrowed humidity signal was not included as an input, the model performance markedly increased and became competitive with other 'same location' test deployment case studies."

Comment: P13 – L13: I don't really understand what is meant by "relative circumstances" – could you be more explicit about each of the case studies? Perhaps a table that outlines case study, with a one sentence description, and a column describing limitations would be more appropriate (and should be introduced at the beginning of the paper).
Response: Thank you for the feedback.
Edits: We have changed "relative circumstances" to "relative timing and parameter coverage". We have also adapted Table 4 according to this feedback and described the 'relative circumstances' present in each case study much more thoroughly in the methods section.

Comment: P13 – L18: What is meant by "extrapolated significantly?" Can you be specific?
Response: Sure, thank you for the comment.
Edits: We have changed "extrapolated significantly" to "extrapolated more than several months".

Comment: Table 1: I find Table 1 almost impossible to follow. It is not very clear which sensor measures which pollutant, as the input codes are frequently indecipherable. I am honestly not sure what I am supposed to get out of this table.
Response: Thank you for this useful feedback that will help us improve the quality and clarity of Table 1.
Edits: We have added a row at the top of the table indicating the target gases for each sensor. We have added to the Table caption an explanation of the input codes for the sensors: "Gas sensors included in U-Pods along with the model input codes we assigned each. The input code for each gas sensor is simply an abbreviation for the make of the sensor, followed by the target gas species(s)."

Comment: Table 3: I find it difficult to interpret this table. What do the black diamonds mean? What do you mean by "relative circumstances"??
Response: Thank you for the helpful comment.
Edits: We have updated Table 4 by replacing the first column so it is more clearly an indication of which case studies covered which target gases (O_3 and CO_2, or just O_3). We have also updated the caption of Table 4 to be more descriptive and informative, and less confusing: "Relative timing and parameter coverage between model training and test deployment dataset pairs. Incomplete coverage of time occurred if training only took place before or after the test data period and not before and after (pre and post). Incomplete coverage of location occurred when training took place in one location and testing took place in another. Incomplete coverage of parameters, including the target gas mole fraction, temperature, time, and pressure occurred when the values observed during training did not encompass the values observed during testing."

Comment: Figure 8: There is way too much going on in this figure, it is almost difficult to look at. Is there a more streamlined way of presenting the findings that is less complicated? I feel there is valuable information in the Figure, but it's hard to determine what that is, due to information overload. Ditto for

Figure 9, though it isn't as bad. Also you would do well to remind the reader what each variable represents.
Response: Thank you very much for the feedback and helping us to simplify and clarify figures.
Edits: We have split what was previously Figure 8 into two figures (now Figure 10 and 11) in order to simplify the graphics and highlight the content of each. We have also added definitions for the sensor inputs in the Figure captions for what were Figures 8 and 9 (now Figures 10, 11, and 12).

TECHNICAL CORRECTIONS

Comment: P1 – L18: "in time than to…" vs. in time that to
Response: Thanks very much for catching this typo.
Edits: We have changed 'that' to 'than' accordingly.

Comment: P2 – L18: change informal language "hold up" to something more scientific
Response: Thank you for the helpful feedback**.**
Edits: We have replaced 'hold up' with 'remain effective'.

Comment: P2 – L20: Delete word "Specifically"
Response: Thank you for the helpful edit.
Edits: We have deleted the word "Specifically"

Comment: P7 – L8: Rephrase "…showed successfully reduced over fitting"
Response: Thank you for the helpful comment.
Edits: We have replaced "…showed successfully reduced over fitting" with proved to be effective in the reduction of over fitting"

Comment: Figure 1: Enhance figure caption to explicitly state that blue is training, pink is testing
Response: Thank you, we will do this.
Edits: "Model training periods for each test deployment are shown in blue, and model test periods are shown in magenta."