Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

Atmospheric Measurement Techniques Discussions

Testing the performance of field calibration techniques for low-cost gas sensors in new deployment locations: across a county line and across Colorado

Joanna Gordon Casey[1], Michael P. Hannigan[1]

[1]Department of Mechanical Engineering, University of Colorado at Boulder, Boulder, 80309, United States of America

Correspondence to: Joanna Gordon Casey (joanna.casey@colorado.edu)

Abstract. We assessed the performance of ambient ozone (O_3) and carbon dioxide (CO_2) sensor field calibration techniques when they were generated using data from one location and then applied to data collected at a new location. We also explored the sensitivity of these methods to the timing of field calibrations relative to deployments they are applied to. Employing data from a number of field deployments in Colorado and New Mexico that spanned several years, we tested and compared the performance of field-calibrated sensors using both linear models (LMs) and artificial neural networks (ANNs) for regression. Sampling sites covered urban, rural/peri-urban, and oil and gas production influenced environments. Generally, we found that the best performing model inputs and model type depended on circumstances associated with individual case studies. In agreement with findings from our previous study that was focused on data from a single location (Casey et al., 2017), ANNs remained more effective than LMs for a number of these case studies but there were some exceptions. In almost all cases the best CO_2 models were ANNs that only included the NDIR CO_2 sensor along with temperature and humidity. The performance of O_3 models tended to be more sensitive to deployment location than to extrapolation in time while the performance of CO_2 models tended to be more sensitive to extrapolation in time that to deployment location. The performance of O3 ANN models benefited from the inclusion of several secondary metal oxide type sensors as inputs in many cases.

1 Introduction

In places like the Denver Julesburg (DJ) and San Juan (SJ) Basins, along Colorado's Front Range and in the Four Corners Region, oil and gas production activities have been increasing with the advent of horizontal drilling that can be effectively used in conjunction with hydraulic fracturing to produce hydrocarbons from unconventional geologic formations. Public health concerns have arisen about the increasing number of people living alongside these industrial activities and emissions. We previously developed methods to quantify ozone (O_3), carbon dioxide (CO_2), methane (CH_4), and carbon monoxide (CO) using low-cost gas sensors in an area where the ambient mole fractions of these species are influenced by oil and gas production activities (Casey et al., 2017). Such low-cost sensor measurements could enable greater understanding of air quality in oil and gas production basins, informing the spatial and temporal scales that people live and work in a way that

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

Atmospheric Measurement Techniques Discussions

current technologies used by regulatory agencies cannot feasibly accomplish. We tested and compared the performance of direct and inverted linear models (LMs) as well as artificial neural networks (ANNs) as regression tools in the field calibration of low-cost sensor arrays to quantify these target gas species using month-long test datasets, training each model with approximately one month of data prior to and one month of data subsequent to this test period.

1.1 Low-Cost Sensors For Air Quality Measurements

The use of low-cost metal oxide, electrochemical and non-dispersive infrared sensors to characterize air quality is becoming increasingly common across the globe (Clements et al., 2017; Kumar et al., 2015). Field normalization methods to quantify atmospheric trace gases using low-cost sensors have been found to be more effective than lab calibration when sensors are deployed in the ambient environment, and subject to changing temperature and humidity (Cross et al., 2017; Piedrahita et al., 2014; Sun et al., 2016). Our previous study and several others have compared the efficacy of LMs (simple and multiple linear regression) relative to supervised learning methods, 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. These effective supervised learning techniques often incorporate multiple gas sensor signals as inputs in order to quantify each target gas, in addition to environmental variable sensor signals, with the goal of compensating for the effects of interfering gas species and environmental factors. In practice though, the reason multiple gas sensors are able to improve the performance of supervised learning type regression, and linear models for that matter, may be in part the result of correlation, or correlation for some time periods, between mole fractions of target gases themselves that hold for one model training location, but might not hold up at alternative sampling sites or during other time periods.

1.2 Summary of Previous Study

Specifically, our previous study was carried out using sensor measurements collected over the course of several months in the spring of 2017, in Greeley, Colorado, which lies within the Denver Julesburg oil and gas production basin. Others had recently demonstrated the efficacy of machine learning methods in the quantification of atmospheric trace gases using arrays of low-cost sensors in urban (De Vito et al., 2008, 2009; Zimmerman et al., 2017) and rural (Spinelle et al., 2015, 2017) areas. Our previous study tested the relative efficacy of machine learning methods and LMs in the quantification of CH_4, O_3, CO_2, and CO in an area strongly influenced by oil and gas production activities, where enhanced levels of hydrocarbons and other industry related pollutants could potentially confound measurements. Our previous study was also the first time machine learning regression techniques were compared with LMs toward the quantification of CH_4 using arrays of low-cost sensors in any setting. A number of enclosures containing arrays of low-cost gas sensors were co-located with optical gas analysers at a Colorado Department of Public Health and Environment monitoring site. ANNs and LMs were trained using a variety of sensor signal input sets from a month of co-located data collected prior to and following a month long test period. The performance of each model was then evaluated relative to reference instrument measurements during the test period. For quantification of all four trace gases that we tested in this oil and gas-influenced setting, we found that ANNs performed

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

Atmospheric Measurement Techniques Discussions

better than LMs. The better performance of ANNs over LMs was likely largely attributable to the ability of ANNs to more effectively represent complex and nonlinear relationships among sensor responses, environmental variables, and trace gas mole fractions than LMs. However, the performance of these powerful regression methods could be aided by relationships among atmospheric trace gases specific to the training location, which would not necessarily hold at different sampling sites.

1.3 Spatially Distributed Networks of Sensors

Distributed spatial networks of low-cost sensor measurements that motivate our work and a large number of other recent low-cost sensor studies are contingent on the spatial transferability of quantification techniques. In this work we test how well the field calibration of low-cost sensors can inform target gas mole fractions when sensors are deployed in a new location. Measurements we collected with low-cost sensors in the DJ and SJ Basins in recent years support this analysis for O_3 and CO_2 for both LMs and ANNs, including a comparison of models with a number of different input sets. In previous work (Casey et al., 2017) we have additionally addressed the quantification of CO and CH_4 using arrays of low-cost sensors together with field normalization methods, but these species are not included in the present analysis because model training and testing deployment pairs, diverging in location and timing and analogous to those we present for O_3 and CO_2, were not available.

1.4 Oil and Gas Production and Air Quality

Oil and gas production related emissions, namely nitrogen oxides (NO_X) and volatile organic compounds (VOCs), have been shown to influence tropospheric ozone (O_3), which is particularly relevant in regions that are in non-attainment of the United States Environmental Protection Agency (USEPA) National Ambient Air Quality Standards (NAAQS) for ozone, like the Colorado Front Range where the DJ Basin is situated. A number of studies have demonstrated that oil and gas related emissions contribute to increased O_3 in the DJ Basin (Cheadle et al., 2017; Gilman et al., 2013; McDuffie et al., 2016). Very high levels of ozone during winter months have also been observed and attributed directly to oil and gas production emissions in the Upper Green River Basin of Wyoming and Utah's Uinta Basin (Ahmadov et al., 2015; Edwards et al., 2013, 2014; Field et al., 2015; Oltmans et al., 2016; Schnell et al., 2009). Additionally, a modeling study concluded that oil and gas production activities could significantly impact ozone near emissions sources (Olaguer, 2012).

Abeleira and Farmer show that ozone production throughout much of the Front Range, outside of downtown Denver, is likely to be NO_X limited implying that local NO_X sources are likely influencing ozone on small spatial scales (Abeleira and Farmer, 2017). Oil and gas industry related NO_X sources, like well pad combustion and diesel truck traffic, could lead to pockets of elevated O_3 throughout the DJ Basin. Low-cost O_3 sensors could augment the few and far apart regulatory sites that currently monitor O_3 levels in places like the DJ Basin, which has better coverage than many other production basins in the United States. While elevated ambient CO_2 levels are not directly harmful to human health, continuous CO_2

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

Atmospheric Measurement Techniques Discussions

measurement can provide information about nearby combustion-related pollution and atmospheric dynamics that lead to the pooling of potentially harmful compounds associated with the oil and gas production industry.

2 Methods

2.1 Sensors and U-pods

Ten U-Pods (mobilesensingtechnology.com) were populated with a number of low-cost gas sensors, as in our previous study (Casey et al., 2017). The gas sensors are listed in Table 1 along with model input codes we assigned to each. A RHT03 sensor was used in each U-Pod to measure temperature (temp) and relative humidity (rh). A Bosch BMP085 sensor was used to measure pressure in each U-Pod.

2.2 Deployment Locations and Timelines

These ten U-Pods were deployed at a number of sampling sites in and around the DJ and SJ Basins over the course of several years, from 2014 - 2017. Deployments generally consisted of co-location with reference measurements prior to and following a period of spatially distributed measurements. During some the distributed measurement periods, a subset of U-Pods remained co-located with reference instruments where the field calibrations took place. During other distributed measurement periods, U-Pods were deployed in new locations that were equipped with reference measurements. We opportunistically employ data from a number of these sensor deployments, treating them as case studies in order to characterize the performance of field calibration models when they are extended to new locations. For each case study, data was divided into training and test periods.

Table 2 lists the O_3 and CO_2 reference instruments that were co-located with U-Pods at each sampling site, along with instrument operators, calibration procedures, and reference data time resolution. The first distributed measurement campaign took place during the summer of 2014 when five U-Pods were sited at locations around Boulder County, with four distributed along the eastern boundary of the county, adjacent to Weld County where dense oil and gas production activities were underway. A background site, further from oil and gas production activities was also included to the west, near a busy traffic intersection on the north end of the City of Boulder. Co-locations with reference measurements toward field calibration of sensors took place at the Continuous Ambient Monitoring Program (CAMP) Colorado Department of Health and Environment (CDPHE) air quality monitoring site in downtown Denver. One of the distributed sampling sites, Dawson School, was also equipped with an optical O_3 instrument, operated by Detlev Helmig's research group from the Institute for Artic and Alpine Research (INSTAAR). This study was funded by Boulder County with the combined aims of gaining a better understanding of how oil and gas emissions affect air quality in Boulder County and learning more about how low-cost air quality measurement methods can help inform air quality in this context.

In the spring of 2015 we augmented our original fleet of five U-Pods with five more and deployed these sensor systems in the SJ Basin while a targeted field campaign was underway to understand more about a CH_4 'hot spot' that was discovered from satellite based remote sensing measurements (Frankenberg et al., 2016; Kort et al., 2014). The primary goal of this sensor deployment was to inform spatial and temporal patterns in atmospheric trace gases like CH_4, O_3, CO, and CO_2 across the SJ Basin. The majority of U-Pods were located at existing air quality monitoring sites operated by the New Mexico Air Quality Bureau (NM AQB), the Southern Ute Indian Tribe Air Quality Program (SUIT AQP), and the Navajo Environmental Protection Agency (NEPA), which supported validation of sensor measurements for O_3 and CO_2.

We deployed a similar distributed network of sensor systems throughout the DJ Basin in the summer of 2015 and the winter/spring of 2016 to explore spatial and temporal patterns in air quality in another region influenced by oil and gas production activities. For these 2015 and 2016 deployments, we co-located U-Pods with reference instruments operated by National Oceanic and Atmospheric Administration (NOAA) researchers at the Boulder Atmospheric Observatory (BAO) Tower toward field calibration, though none of the distributed sampling sites were equipped with reference instruments to support validation. In August of 2016 the ten U-Pods were moved to the Greeley Tower (GRET) CDPHE air quality monitoring site in Greeley, Colorado, a location which is also strongly influenced by DJ Basin oil and gas production activities; the U-Pods remained there for a year. For the GRET co-location period, CDPHE shared reference measurements for O_3. Additionally, Jeffrey Collett and Katherine Benedict of Colorado State University (CSU) shared CO_2 reference measurements from an instrument they operated at the site before October 1st in 2016 and after March 7th in 2017, when the instrument was located at the GRET site.

The work presented here is aimed at supporting methods to quantify atmospheric trace gases during the distributed deployments described above as well as future distributed sensor network measurements. Fig. 1 shows the timeline of each of these deployments, highlighting the training and testing periods defined for both O_3 and CO_2. Fig. 2 shows sampling site locations in context with urban areas and oil and gas production wells.

2.3 Reference and Sensor Data Preparation

Each of the U-Pod sensor signals was logged to an onboard micro SD card. For metal oxide type sensors, voltage signals were converted into resistance, and then normalized by the resistance of the sensor in clean air. Relative humidity, temperature, and pressure measured in each U-Pod were used to calculate absolute humidity. Over the course of several field deployments, relative humidity sensors in a few of the U-Pods drifted down, causing the lower humidity levels to be cut off or 'bottomed out'. For measurements collected in the spring and summer of 2015 and the spring of 2017, we replaced the relative humidity (RH) signal of U-Pods with malfunctioning humidity sensors with signals from nearby U-Pods with good humidity sensors and complete data coverage as noted in Table S1.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

When the U-Pods were initially deployed at the GRET site, on August 23rd of 2016, the RH sensors in all ten U-Pods malfunctioned, logging an error code of -99 instead of the relative humidity. This malfunction seemed to coincide with the implementation of radio communication from each U-Pod to a central node in an effort to reduce trips to the field site to download data and to identify issues with data acquisition promptly. RH signals in the U-Pods began logging correctly again in November when we stopped remote communication. We replaced RH values for the U-Pods during this time period by utilizing data from the Picarro Cavity Ring-Down Spectrometer that was co-located at GRET with the U-Pods. 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 then replaced the absolute 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.

To perform regressions toward field calibration of sensors, the reference and U-Pod data needed to be aligned. When reference measurements with minute time resolution were available for both training and corresponding testing periods, minute median data from the U-Pods were used. Medians were used as opposed to averages in order to reduce the potential influence of sensor noise as well as to remove short duration spikes in the reference and sensor data that resulted from air masses that may not have been well mixed across the reference instrument inlets and the U-Pod enclosures. When reference data were instead available with only 5-minute or 60-minute time resolution, U-Pod medians were calculated for the same time step. Medians were also calculated for reference measurements with finer time resolution to match the time resolution of corresponding training/testing data. The first 15 minutes of data after any period that the U-Pods had not recorded data for the previous 5 minutes was removed in order to filter transient behavior associated with sensor warm-up.

2.4 Calibration Techniques

Field calibration of low-cost gas sensors to quantify ambient trace gas mole fractions has been shown to work more effectively than lab calibration methods in several previous studies (Piedrahita et al., 2014; Sun et al., 2016; Zimmerman et al., 2017). Low-cost sensors are often sensitive to many aspects of the complex ambient environment in terms of temperature, humidity, and atmospheric chemistry variations. While trace gases that sensors are specifically designed to respond to elicit a sensor response, so can many environmental and chemical confounding influences. Field calibrations have been shown to more effectively mitigate the effects of these confounding influences in the quantification of atmospheric trace gases in the ambient environment relative to lab calibrations. In this work, we explore how well field calibration methods hold up in new locations, a topic which has not yet been sufficiently addressed by the scientific community.

Using methods described previously (Casey et al., 2017), 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. Subsequently, the

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

resulting quantification model performance was evaluated on test datasets. Generally, model inputs were normalized so that they ranged in magnitude from -1 to 1 since this practice is recommended for the ANN optimization algorithm used (Hagan et al., 1997). The LMs employed here consisted of multiple variable linear regressions, where the determined coefficients for a linear combination of sensor inputs minimized residuals between model predictions and reference instrument target gas mole fractions. We used the same set of inputs in ANNs, employing the Levenberg Marquardt optimization algorithm with Bayesian Regularization. Bayesian Regularization is designed to guard against over fitting. In preliminary ANN tests we found that over fitting occurred even when Bayesian Regularization was used, so we additionally implemented early stopping, which showed successfully reduced over fitting. Each ANN had one hidden neuron in each hidden layer. 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 (Casey et al., 2017). For ANNs, training datasets were divided in half on an alternating 24-hr basis, with half used for training and half used for early stopping.

3 Results and Discussion

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. As in our previous work (Casey et al., 2017), we compared performance of LMs and ANNs with a number of different sets of inputs for each train/test data pair. 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. Presented below is an analysis and comparison of the best-performing model for each species as determined in our previous work, as well as performance metrics for the best performing model associated with each new training/testing dataset pairs described in section 2.2.

3.1 Summer 2015 BAO and SJ Basin

The set of deployments we conducted in the summer of 2015 is particularly useful to the objective of characterizing how well field calibration models can be extended to a new location relative to their performance where they were trained. During the testing period, two U-Pods were located at BAO, where training took place, while eight U-Pods were moved to sampling sites in the SJ Basin, across Colorado and over the state line in New Mexico. Seven of these SJ Basin sites were equipped with O_3 reference measurements and two of the sites were equipped with CO_2 reference measurements. This testing period

supports a direct comparison of model performance at the training location relative to a number of distributed sampling sites. With this dataset, we investigate how sensor field calibrations can hold up, both across a large geographic distance, as well as in an environment with different atmospheric chemistry of reducing and oxidizing gases that can influence sensor signals. Sampling sites at BAO, in the DJ Basin, and throughout the SJ Basin were all influenced by oil and gas production activities and their associated emissions to some extent, but the predominant chemistry in the production stream is different in each basin. In the SJ Basin, particularly the northern portion of the basin where all our sampling sites were located production is dominated by coalbed methane. In contrast, many wells in the DJ Basin produce both oil and gas leading to greater relative abundance of heavier hydrocarbons in emissions. The DJ Basin air shed is also more strongly impacted by urban emissions than the SJ Basin air shed, and is more strongly influenced by mobile sources with Denver, Boulder, Fort Collins, Greeley, and many other smaller communities in its midst and along its borders. The Four Corners region, where the SJ Basin is situated, has a much smaller population density. Additionally, while there are some agricultural activities and associated emissions in and around the SJ Basin, there is a significantly larger agricultural industry in and around the DJ Basin. SJ Basin sampling sites spanned a range of elevations, including some that were higher and some that were lower than the BAO Tower, coinciding with a wide range of atmospheric pressure at the distributed sampling sites.

The U-Pod CO_2 data presented in Fig. 3 and Fig. 4 were quantified using ANNs that were trained using data from the BAO Tower with the following inputs from each U-Pod: eltCO2, temp, and absHum. This set of model inputs were found to be the best ANN inputs that we highlighted in our previous study, using data from the GRET site in the spring of 2017 (Casey et al., 2017). Fig. 3 shows scatter plots of U-Pod CO_2 vs. reference CO_2 during the test data period for sensors located at BAO as well as sensors that were located at distributed sampling sites throughout the SJ Basin. The scatter plots show that while there was generally a smaller dynamic range of CO_2 at the SJ Basin sites relative to BAO, model performance did not appear to be impacted or degraded by spatial extension to these locations in the SJ Basin. The line of best fit for Fort Lewis site (periwinkle) is even closer to the 1:1 than the lines of best fit for two U-Pods located at BAO (black and grey). Overlaid histograms of residuals in the bottom right corner of Fig. 3 show that CO_2 residuals from each of the SJ Basin U-Pods are generally centered and evenly distributed about zero with similar spread.

U-Pod CO_2 average residuals from the same data presented in Fig. 3, quantified using ANNs with eltCO2, temp, and absHum signals as inputs, are plotted according to time of day and date in Fig. 4. While the use of ANNs in place of LMs was shown to reduce U-Pod CO_2 residuals significantly with respect to temperature, some daily periodicity in the residuals for all four U-Pods is apparent in the upper plot in Fig. 4 that shows residuals by date. The lower plot in Fig. 4, showing residuals by time of day, demonstrates that CO_2 from all four U-Pods was generally under predicted during early hours of the morning and generally over predicted during afternoon and evening hours. Interestingly, this trend in residuals by time of day is more pronounced for the two U-Pods that remained at BAO. The majority of U-Pods stopped logging data, unfortunately, at one point or another during these deployments. The periods of missing data are reflected in the plots of

residuals by date in Fig. 4 for CO_2 and in Fig. 6 for O_3. Fortunately, no drift over the course of the deployment period was observed in these plots.

Differing from our previous findings, for this group of training and testing data pairs from the summer of 2015 at the BAO and SJ Basin sites, the inclusion of the e2vVOC and alphaCO signals noticeably improved the RMSE in the quantification of CO_2. While the inclusion of these two secondary sensor signals didn't result in the best performance in our previous study, using data from the GRET site (Casey et al., 2017), we found that their inclusion did not degrade performance relative to the models that included just eltCO2, temp, and absHum signals as inputs. Generally, using rh vs. absHum signals as ANN inputs did not seem to make a big difference in model performance, though linear models were sometimes found perform better when the absHum signal is used instead of the rh signal. From Fig. S2, it is apparent that inputs including e2vCO2, temp, rh, e2vVOC, and alphaCO sensor signals as model inputs resulted in the lowest RMSE for U-Pods at BAO as well as at the two SJ Basin sites. Plots analogous to those presented in Fig. 3 and Fig. 4, but with this best performing set of inputs for the present data set pairs are presented in the SM, in Fig. S24 and Fig. S25 respectively.

O_3 was quantified for all the U-Pods deployed at BAO and SJ Basin sampling sites using an ANN with the following inputs: e2vO3, temp, absHum, e2vCO, e2vVOC, figCH4, and figCxHy. ANNs with this configuration were found to perform best in the quantification of O_3 in our previous study (Casey et al., 2017). These same inputs and model configuration were also found to be the best performing among others tested for SJ Basin 2015 dataset pairs as noted in Fig. S2. Interestingly though, LMs with this same set of inputs performed competitively well for a number of the U-Pods in the SJ Basin. For three of seven U-Pods in the SJ Basin, LMs even outperformed ANNs in terms of RMSE. When the BAO trained U-Pods field calibrations for O_3 were extended to sites in the SJ Basin, we found that U-Pods at some of the sites performed better than others across all models that were tested, as seen in Fig. S2.

Scatter plots and trends in residuals are presented in Fig. 5 and Fig. 6 for O_3. These plots show the performance of U-Pods at BAO relative to those at SJ Basin sites in the quantification of O_3 during the test data period. U-Pod O_3 measurements at Fort Lewis, Navajo Lake, and the Sub Station did not agree with reference measurements as well as U-Pod O_3 measurements from the other four SJ Basin sites. U-Pods at Navajo Lake and Sub Station had bad humidity sensor data, as noted in section 6.2.3 and Table S1, so humidity from the U-Pod located at the Ignacio site was used in place of their humidity signals. Since the Ignacio site was located relatively far away from the Navajo Dam and Sub Station sites, 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). The Fort Lewis site had a different reference instrument than those used at other sites, which may have contributed to observed discrepancies. Fig. S1 shows that differences among U-Pod O_3 performance during the test deployment period were larger than those observed during the training phase among the same U-Pods; therefore, the incongruous field calibration performance phenomena we observed

seems to be connected to the sampling sites or the circumstances discussed previously as opposed to the quality of individual sensors in each of those U-Pods.

All SJ Basin U-Pod O_3 measurements systematically over estimate lower levels of O_3 each night, a trend apparent in the scatter plots in Fig. 5 and in the residuals by time of day plot in Fig. 6. Upon examination of the scatter plots in Fig. 5, U-Pods at some sampling sites had positive bias for higher O_3 measurements as well (Shiprock, Ignacio, Sub Station, and Bloomfield), while for others, bias at the higher end of O_3 distributions did not appear to be present (Navajo Dam, Fort Lewis, and Bondad). The residuals by time of day plot in Fig. 6 shows that the two U-Pods at BAO did not have significant trends in their residuals according to the time of day, but that U-Pods deployed at SJ Basin sites consistently over estimated nighttime O_3. The residuals are also plotted with respect to temperature in Fig. 6, where all U-Pods, even those at BAO to a lesser extent, appear to over predict O_3 at lower temperatures, which generally occurred at night. The times of day that generally correspond to the highest O_3 levels generally had the lowest residuals, with some exceptions at the Fort Lewis and Navajo Dam sites.

Fig. 6 includes a plot of the residuals across the duration of the deployment period, showing no significant sensor drift in measurements for any of the U-Pods. This plot also shows that the highest residuals observed generally occurred over short periods in time, particularly for the Fort Lewis (blue) and Sub Station (magenta) sites. In order to further explore the performance of field calibration models for O_3 at SJ Basin sites relative to BAO, the combined parameter space of temperature with O_3 reference mole fractions and temperature with absolute humidity are presented in Fig. 7. The combined temperature and reference O_3 parameter space appears to be similar for all of the U-Pods, both at BAO and the SJ Basin sites. However, there appears to be some outlying combined temperature and humidity parameter space at the Sub Station site and at the Navajo Dam site. Brief excursions of high humidity may be connected to some of the large short-term residuals observed at these two sampling sites.

3.2 Insight from Additional Case Studies of Field Calibration Extension to New Locations

3.2.1 Urban calibration moved to rural/peri-urban setting

The Boulder County deployment in the summer of 2014 was used to test how well a field calibration generated in a busy urban area (at CAMP in downtown Denver) could be extended to a peri-urban setting (at Dawson School in eastern Boulder County). Training took place at CAMP for several days each month, before and after each approximately month-long deployment period at Dawson School over the course of four months. Fig. S7 shows the performance of a number of ANN and LM-based CAMP field calibrations with different sets of inputs at this Dawson School test site. In this case study, LMs performed better than ANNs across all sets of sensor inputs, when models were trained using data from CAMP and tested using data from Dawson. Unlike findings from our previous study (Casey et al., 2017), including secondary metal oxide

type sensors as inputs didn't help to improve model performance. The best performing set of inputs included just e2vO3, temp, and absHum signals. The very different relative abundance of various oxidizing and reducing compounds in downtown Denver relative to the Dawson School site, surrounded by open grassy fields, and in closer proximity to oil and gas production activities, may be the reason why including additional gas sensors as model inputs and the use of ANNs failed to improve the quantification of U-Pod O_3 in this case. Relatively short training durations could also contribute to this finding, based on findings from our previous work (Casey et al., 2017). Notably though, neither ANNs nor the LMs captured the highest levels of O_3 at Dawson School well. We attribute the poor performance at high levels of O_3 at this site, those in exceedance of about 70 ppb, to extrapolation of the O_3 mole fractions encompassed during the training period. The models generally performed well within the O_3 levels covered during the training period. Keeping in mind this finding about O_3 extrapolation, for ambient measurements in the DJ Basin, for subsequent deployments, we selected field calibration sites that were more representative of distributed sampling site locations, outside of the dense urban environment in downtown Denver, where O_3 did not get as high, likely due to increased titration of O_3 at night in connection with abundant NO_X compounds.

3.2.2 Post only calibration moved across the state

We also examined model performance that was subject to extrapolation in time (when training data consisted of only post test deployment co-locations) and temperature. In the spring of 2015 we co-located U-Pods with reference instruments in the SJ Basin, at the Navajo Dam, Sub Station, and Bloomfield sites. Fig. S4 shows the performance of a number of ANN and LM-based BAO field calibrations with different sets of inputs at this SJ Basin test sites in the spring of 2015, just prior to the summer 2015 BAO training period. U-Pod O_3 was quantified for these deployments using training data from the same co-location period at BAO that was used toward quantification of the summer 2015 SJ Basin deployment, described in section 3.1. Interestingly, the addition of time as a model input didn't seem to improve the performance of either ANNs or LMs and ANNs generally outperformed LMs. U-Pods experienced colder temperatures during this spring deployment than were encompassed subsequently in the BAO training period. Linear models generally resulted in more bias than ANNs. Unlike our previous findings, (Casey et al., 2017), with this training/testing dataset pair, we found that including some secondary metal oxide type sensors augmented model performance, but not the figCxHy sensor signal. The Sub Station site is close to two large coal-fired power plants, indicated in Fig. 2 by orange markers in the SJ Basin pane. It is possible that emissions from the San Juan Generating Station (north) and/or the Four Corners Power Plant (south) uniquely influenced the response of this particular Figaro sensor in ways that are not well represented at BAO in the DJ Basin, or at other SJ Basin sampling sites. Temperature or time extrapolation could also be responsible for negating the utility of the figCxHy sensor signal in model performance.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

3.2.3 Post only calibration moved 40 miles across the DJ Basin

In testing how field calibrations that were generated using data collected at the GRET site in 2017 could inform the quantification of O_3 at BAO in the 2016, across the DJ Basin, we were interested to find that again, the inclusion of time as a model input did not yield any improvements in calibration equation performance, even though model training took place so long after the test period. Fig. S5 shows the performance of a number of ANN and LM-based GRET field calibrations with different sets of inputs at this BAO test site the previous summer. Another interesting finding from this training/testing dataset pair was that the addition of secondary metal oxide type gas sensors, didn't seem to help improve the performance of field calibration equations either. Fig. S5 shows that ANNs performed better than LMs and that the most useful set of inputs included just e2vO3, temp, and absHum. Similarly, the performance of field calibration equations for CO_2 generated at GRET in 2017 and applied to data from BAO in the summer of 2016, did not seem to be augmented by the inclusion of additional gas sensor signals, though the inclusion of time as a predictor was useful. In the case of CO_2, LMs outperformed ANNs with notable instability associated with the performance of ANNs when time was included as an input.

3.2.4 Post only calibration applied to the same location

To investigate if reduced performance from these GRET to BAO field calibration tests were more connected to the new deployment location or to the significant extrapolation with respect to time of the calibration models, we generated calibration equations based on similarly long training periods at GRET and applied them to data collected prior to the training period at GRET in the fall of 2016. Fig. S6 shows the performance of a number of ANN and LM-based GRET field calibrations with different sets of inputs at the GRET test site during fall of the previous year. The best performing ANN inputs for this dataset pair were the same ones that we found in our previous study (Casey et al., 2017), with the exception of the humidity signal. The fall 2016 GRET test period coincided with the time period U-Pod absolute humidity was replaced with mixing ratios from a co-located Picarro due to missing humidity sensor data. Interestingly, when humidity was removed as an input, the model performance markedly increased and became competitive with other 'same location' test deployment case studies. 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.

3.3 Evaluation of models across training/testing dataset pairs

Fig. 8 is a graphical summary of the performance of quantification models for O_3 and CO_2 for each of the training/testing dataset pairs included in this study to explore how these models hold up when applied to data collected at new locations and in new time frames. The top panels in Fig. 8 show the performance of relatively simple linear models with only three inputs, including just the primary gas sensor for each species, temperature, and humidity. The bottom panels in Fig. 8 show the performance of the ANN regression models and inputs that were found to perform the best on data collected at the GRET site in Greeley, Colorado for O_3 and CO_2 in our previous study (Casey et al., 2017). Fig. 9 is a similar graphical summary of

model performance for each dataset pair examined in this study, but in this figure, the best model for each specific deployment is shown. Fig. 8 and Fig. 9 contain target plots showing the MBE and CRMSE of models from each dataset pair in terms of absolute mole fractions and mole fractions normalized uniformly by the standard deviation of reference data during the spring 2017 GRET deployment. In the SM, Fig. S23 contains target diagrams equivalent to those presented in Fig. 8, but with individually normalized MBE and CRMSE, according to the standard deviation of reference measurements during each individual test period.

The outer circle's radius in each of these target diagrams denotes an error-to-signal ratio of 1. The inner circle's radius in each of these target diagrams encompasses the performance of models when they were tested at the same location that they were trained and when training data bookended the test period, so that there was no extrapolation of the model across time or deployment location. We present our findings in the form of these target diagrams in order to compare our findings with those presented in several particularly relevant previous studies focused on the field calibration of low-cost sensors (Spinelle et al., 2015, 2017; Zimmerman et al., 2017). Table 3 lists the relative circumstances between model training and testing periods for dataset pairs, highlighting instances of incomplete coverage during training that led to model extrapolation during testing.

Fig. 8 shows that for CO_2, ANN models generally performed slightly better than LM models with the same set of inputs, though models that were extrapolated significantly in time were the exception. For O_3, ANNs that included multiple secondary metal oxide sensor signals as inputs were also found to generally perform slightly better than the relatively simple LMs that didn't include any secondary gas sensors as inputs over all (with exceptions for individual case studies). This can be seen in Fig. 8, with all plot markers falling within the outer radius in the bottom panel (ANNs) but some plot markers falling outside the outer radius in the top panel (LMs). Models that were not moved to a new location for the test period gained the most benefit in their performance when ANNs were used instead of LMs, resulting in a smaller inner radius in the target plots on the bottom panel relative to the top panel for both O_3 and CO_2. The target diagrams in Fig. 8 show some degradation in performance when models were applied to data in new locations and when training data took place only before or after the test period. All four of the target plots in Fig. 8 demonstrate that bias was introduced when field calibration models were extrapolated in terms of time, when training periods only encompassed data after the test data period and not prior. Interestingly, there are noticeable similarities between the target plots for CO_2 in the left panels and the target plots for O_3 in the right panels. The relative performance of models among each training/test dataset pair remained fairly consistent across the different models employed in data quantification. These systematic trends highlight the importance of model training and testing circumstances relative to specific field calibration model types and inputs. In Fig. 8, all models that were not extrapolated in time all result in error that falls within the outer circle radius, meeting performance standards framed by previous studies (Spinelle et al., 2015, 2017; Zimmerman et al., 2017).

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

For the GRET 2017 to BAO 2016 dataset pairs, CO_2 and O_3 were both best represented by LMs as opposed to ANNs. CO_2 and O_3 models did not benefit from additional gas sensors added as inputs. The best performing models for each included temperature and absolute humidity. CO_2 models performed much better with the inclusion of time, but O_3 models did not improve with time added as an input. In Fig. 8, both ANN and LM markers are included (each with the same inputs: e2vO3, temp, and absHum). LMs had smaller random error but ANNs had smaller bias, highlighting an important consideration in the application of one or the other to inform specific research or measurement goals. One significant difference between the O_3 and CO_2 GRET to summer 2016 BAO and the 2017 GRET to fall 2016 GRET field calibration tests periods is that the field calibration for O_3 at GRET was based on a longer dataset. Since reference measurements were not available for CO_2 from November 1st of 2016 through March 7th of 2017, there was significantly less data and less environmental variable space coverage in CO_2 field calibration training relative to O_3 calibration equation training.

Fig. 9 again shows target diagrams, this time displaying the best performing model for each training/testing deployment pair. The best field calibration model performances for each dataset all fall within the outer radius, showing good performance. Fig. 9 shows that incomplete coverage of parameter space in terms of atmospheric chemistry, weather patterns, sampling location, and sampling timing, can be addressed to some extent by tailoring field calibration models and their inputs to specific training/testing datasets pairs.

For CO_2 we found that field calibration models generally extended with good performance to new locations. ANNs outperformed LMs when training took place pre and post of a test deployment. When training only took place after a test deployment LMs performed better. LMs seem to be better at extrapolating in time. Over time, ELT NDIR CO_2 sensors seem to lose sensitivity and/or drift. When CO_2 models were extended back in time, significant bias resulted when time was not included as an input. ANNs were not able to extrapolate in time with any success and their performance became very unstable when time was added as an input, an occurrence that is apparent in Fig. S5 and Fig. S6. Models performed better when they were extended spatially, all the way across Colorado from the DJ Basin to the SJ Basin, than they did when they were extended back in time. Extension of a LM back in time and across the DJ Basin (from GRET in 2017 to BAO in 2016) resulted in significant MBE relative to the other case studies. The inclusion of multiple additional gas sensors augmented model performance when extended back in time at the same location as training took place, but not at a new location.

For O_3 we found that ANNs with the same set of inputs worked best across a number of case studies, informed by all the metal oxide sensor signals as well as temperature and humidity. The extension of models to new locations often resulted in increased MBE or systematic error, and in some cases increased CRMSE or random error. Some observed bias in the extension of models to new locations could be attributable to different reference instruments with different operators and/or different calibration and data quality measures employed. O_3 model extension to new locations seemed to be more impactful than extension back in time. Interestingly, additional metal oxide sensor signals remained helpful when models were

extended all the way across Colorado, from BAO to the SJ Basin, but these additional gas sensor signals did not remain helpful when O_3 models were extended across a county line, from Adams County (CAMP) to Boulder County (Dawson) or from Weld County (GRET) to Boulder County (BAO). ANNs generally performed better than LMs for O_3, with the exception of these two Front Range case studies. We found in our previous study that shorter training times led to decreased performance in ANNs and sometimes increased performance in LMs. The training time used in the CAMP to Dawson case study was relatively short, which could have contributed to the superior performance of LMs over ANNs.

4 Conclusions

Several previous studies have shown that multiple gas sensor signals and the implementation of supervised learning techniques can improve the performance of field calibration of low-cost sensors in the quantification of a number of atmospheric trace gas mole fractions. We investigated how well these supervised learning techniques hold up when sensors are moved to a new location, different from where calibration model training took place. We tested the spatial and temporal transferability of field calibration models for O_3 and CO_2 under a number of different circumstances using data from multiple reference instrument co-locations, using the same sensors over the course of several years. We found that the best performing field calibration models for both O_3 and CO_2 were not consistent across all training and testing deployment pairs, though some patterns emerged in terms of model type and inputs in association with the spatial and temporal extension of calibration equations, from training to testing performed in oil and gas production areas. The performance of O_3 models generally benefited from the inclusion of multiple metal oxide sensor signals in addition to the primary e2vO3 sensor signal, while the performance of CO_2 models relied more heavily on temperature and humidity inputs. CO_2 model performance was impacted more by temporal extension than spatial extension. In contrast, O_3 model performance was impacted more by spatial extension than temporal extension.

While ANNs and other supervised learning techniques have been shown to consistently out perform linear models in previous studies when training and testing took place in the same location, we find that this trend does not always hold when field calibration models are moved in terms of spatial and temporal coverage. We found that the implementation of calibration models that were well suited to specific training and test data pairs resulted in generally good test performance in terms of centered root mean squared error and mean biased error, scaled by reference measurement standard deviation, reported in target diagrams in previous studies. For example, when models were significantly extrapolated in time, a well-suited set of sensor inputs would generally not include secondary gas sensor signals.

LMs with just one primary gas sensor signal as well as temperature and humidity were found to outperform ANNs when models were applied to a location with different dominating sources of pollution, like Downtown Denver relative to eastern Boulder County. These three-input LMs also outperformed ANNs when models were significantly extrapolated in time.

While these LMs seemed to be more stable under circumstances of significant extrapolation in terms of local air chemistry and timing, we found that they were not able to fully represent some of the complex nonlinear response behavior exhibited by the arrays of sensors.

Field calibration models tested in new locations often resulted in the introduction of additional bias relative to field calibration models that were tested in the same location they were trained in. This trend is nicely shown in Fig. 9, as the data is almost a band running vertically in a range of CRMSEs. Finding ways to effectively mitigate bias associated with new field deployment locations would further improve the performance of field calibrations toward quantification of atmospheric trace gases using arrays of low-cost sensors. Such improvements in the field of low-cost sensors will help to enable dense distributed networks of low-cost sensors to inform air quality in oil and gas production basins.

We show that field normalization trace gas quantification models can more readily be transferred across a large state from one oil and gas production to another, than from an urban to oil and gas production basin that are in closer proximity to each other. We also show that pre and post model training, directly prior to and after field site deployment, is generally much more effective that pre or post model training alone, especially when the training takes place significantly before or after the deployment period. Along with these findings and general guidelines, we recommend further validation efforts in the extension of quantification of atmospheric trace gases using low-cost gas sensor arrays in oil and gas production basins and toward other ambient measurement applications. The findings presented here may be applicable and generalizable in the use of low-cost metal oxide, electrochemical, and non-dispersive infrared sensor arrays in various configurations and sampling regions to characterize mole fractions of a number of atmospheric trace gases. Future studies exploring the sensitivity of our findings to these factors are recommended.

The authors declare that they have no conflict of interest.

Acknowledgements

The many low-cost sensor and reference instrument measurements that facilitated this study were made possible with the gracious help of a number of agencies and individuals. We'd like to thank Bradley Rink and Erick Mattson of CDPHE, Katherine Benedict and Jeff Collett of CSU, Detlev Helmig and Jacques Hueber of INSTAAR, Gaby Pétron, Jon Kofler, Audra McClure, Bruce Batram, and Daniel Wolfe of NOAA, Michael King of NEPA, Christopher Ellis and Andrew Switzer of the SUIT AQP, along with Joe Cotie and Roman Szkoda of the NM AQB for sharing reference instrument data with us and facilitating our U-Pod measurements. We thank Jana Milford for assistance in the preliminary analysis and reporting of results from the Boulder County Study. Thanks to John Ortega for preparing U-Pods and arranging permissions and logistics for the Boulder County project in 2014, including the Dawson School site we use data from in this work. Thanks are also due to include Brianna Yepa, Victoria Danner, Tasha Nez, Rebecca Bullard, Madeline Polmear, and Bryce

Goldstien for helping to maintain U-Pods in the field as well as downloading and organizing data. Bryce Goldstien made some maps in ArcMap of the SJ and DJ Basins with data from the Colorado Oil and Gas Conservation Commission that were adapted and presented here. It was a pleasure working with all these interesting and helpful people. We thank Jana Milford for assistance in the preliminary analysis and reporting of results from the Boulder County Study as well as useful feedback and suggestions in review of this manuscript. Many thanks also to Shelly Miller, Marina Vance, and Christopher Ellis for kindly reviewing this work which was funded by Boulder County and the National Science Foundation Air Water Gas Sustainability Research Network under grant number CBET-1240584.

References

Abeleira, A. J. and Farmer, D. K.: Summer ozone in the northern Front Range metropolitan area: Weekend-weekday effects, temperature dependences, and the impact of drought, Atmos. Chem. Phys., 17(11), 6517–6529, doi:10.5194/acp-17-6517-2017, 2017.

Ahmadov, R., McKeen, S., Trainer, M., Banta, R., Brewer, A., Brown, S., Edwards, P. M., De Gouw, J. A., Frost, G. J., Gilman, J., Helmig, D., Johnson, B., Karion, A., Koss, A., Langford, A., Lerner, B., Olson, J., Oltmans, S., Peischl, J., Pétron, G., Pichugina, Y., Roberts, J. M., Ryerson, T., Schnell, R., Senff, C., Sweeney, C., Thompson, C., Veres, P. R., Warneke, C., Wild, R., Williams, E. J., Yuan, B. and Zamora, R.: Understanding high wintertime ozone pollution events in an oil- and natural gas-producing region of the western US, Atmos. Chem. Phys., 15(1), 411–429, doi:10.5194/acp-15-411-2015, 2015.

Casey, J. G., Collier-Oxandale, A. and Hannigan, M. P.: Performance of Artificial Neural Networks and Linear Models To Quantify 4 Trace Gas Species In an Oil and Gas Production Region with Low-Cost Sensors, Submitted to Sensor. Actuat. B-Chem.

Cheadle, L. C., Oltmans, S. J., Pétron, G., Schnell, R. C., Mattson, E. J., Herndon, S. C., Thompson, A. M., Blake, D. R. and Mcclure-begley, A.: Surface ozone in the Northern Front Range and the influence of oil and gas development on ozone production during FRAPPE / DISCOVER-AQ, , 6, 2017.

Clements, A. L., Griswold, W. G., RS, A., Johnston, J. E., Herting, M. M., Thorson, J., Collier-Oxandale, A. and Hannigan, M.: Low-Cost Air Quality Monitoring Tools: From Research to Practice (A Workshop Summary), Sensors, 17(11), 2478, doi:10.3390/s17112478, 2017.

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

Edwards, P. M., Young, C. J., Aikin, K., deGouw, J. a., Dubé, W. P., Geiger, F., Gilman, J. B., Helmig, D., Holloway, J. S., Kercher, J., Lerner, B., Martin, R., McLaren, R., Parrish, D. D., Peischl, J., Roberts, J. M., Ryerson, T. B., Thornton, J., Warneke, C., Williams, E. J. and Brown, S. S.: Ozone photochemistry in an oil and natural gas

extraction region during winter: simulations of a snow-free season in the Uintah Basin, Utah, Atmos. Chem. Phys. Discuss., 13(3), 7503–7552, doi:10.5194/acpd-13-7503-2013, 2013.

Edwards, P. M., Brown, S. S., Roberts, J. M., Ahmadov, R., Banta, R. M., deGouw, J. a., Dubé, W. P., Field, R. a., Flynn, J. H., Gilman, J. B., Graus, M., Helmig, D., Koss, A., Langford, A. O., Lefer, B. L., Lerner, B. M., Li, R., Li, S.-M., McKeen, S. a., Murphy, S. M., Parrish, D. D., Senff, C. J., Soltis, J., Stutz, J., Sweeney, C., Thompson, C. R., Trainer, M. K., Tsai, C., Veres, P. R., Washenfelder, R. a., Warneke, C., Wild, R. J., Young, C. J., Yuan, B. and Zamora, R.: High winter ozone pollution from carbonyl photolysis in an oil and gas basin, Nature, doi:10.1038/nature13767, 2014.

Field, R. A., Soltis, J., McCarthy, M. C., Murphy, S. and Montague, D. C.: Influence of oil and gas field operations on spatial and temporal distributions of atmospheric non-methane hydrocarbons and their effect on ozone formation in winter, Atmos. Chem. Phys., 15(6), 3527–3542, doi:10.5194/acp-15-3527-2015, 2015.

Frankenberg, C., Thorpe, A. K., Thompson, D. R., Hulley, G., Kort, E. A., Vance, N., Borchardt, J., Krings, T., Gerilowski, K., Sweeney, C., Conley, S., Bue, B. D., Aubrey, A. D., Hook, S. and Green, R. O.: Airborne methane remote measurements reveal heavy-tail flux distribution in Four Corners region, Proc. Natl. Acad. Sci., 113(35), 201605617, doi:10.1073/pnas.1605617113, 2016.

Gilman, J. B., Lerner, B. M., Kuster, W. C. and De Gouw, J. A.: Source signature of volatile organic compounds from oil and natural gas operations in northeastern Colorado, Environ. Sci. Technol., 47(3), 1297–1305, doi:10.1021/es304119a, 2013.

Hagan, M. T., Demuth, H. B., Beale, M. H. and De Jesus, O.: Neural Network Design, PWS Publishing Co., Boston, MA., 1997.

Kort, E. a., Frakenberg, C., Costigan, K. R., Lindenmaier, R., Dubey, M. K. and Wunch, D.: Four corners: The largest US Methane anomaly viewed from space, , 6898–6903, doi:10.1002/2014GL061503.Received, 2014.

Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L. and Britter, R.: The rise of low-cost sensing for managing air pollution in cities, Environ. Int., 75, 199–205, doi:10.1016/j.envint.2014.11.019, 2015.

McClure-Begley, A., Petropavlovskikh, I. and Oltmans, S.: NOAA Global Monitoring Surface Ozone Network: BAO Tower, , doi:http://dx.doi.org/10.7289/V57P8WBF, 2016.

McDuffie, E. E., Edwards, P. M., Gilman, J. B., Lerner, B. M., Dubé, W. P., Trainer, M., Wolfe, D. E., Angevine, W. M., DeGouw, J., Williams, E. J., Tevlin, A. G., Murphy, J. G., Fischer, E. V., McKeen, S., Ryerson, T. B., Peischl, J., Holloway, J. S., Aikin, K., Langford, A. O., Senff, C. J., Alvarez, R. J., Hall, S. R., Ullmann, K., Lantz, K. O. and Brown, S. S.: Influence of oil and gas emissions on summertime ozone in the Colorado Northern Front Range, J. Geophys. Res., 121(14), 8712–8729, doi:10.1002/2016JD025265, 2016.

Olaguer, E. P.: The potential near-source ozone impacts of upstream oil and gas industry emissions, J. Air Waste Manage. Assoc., 62(8), 966–977, doi:10.1080/10962247.2012.688923, 2012.

Oltmans, S. J., Karion, A., Schnell, R. C., Pétron, G., Helmig, D., Montzka, S. A., Wolter, S., Neff, D., Miller, B. R., Hueber, J., Conley, S., Johnson, B. J. and Sweeney, C.: O3, CH4, CO2, CO, NO2 and NMHC aircraft measurements in the Uinta Basin oil and gas region under low and high ozone conditions in winter 2012 and 2013, Elem. Sci. Anthr., (2), 12, doi:10.12952/journal.elementa.000132, 2016.

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

Schnell, R. C., Oltmans, S. J., Neely, R. R., Endres, M. S., Molenar, J. V. and White, A. B.: Rapid photochemical production of ozone at high concentrations in a rural site during winter, Nat. Geosci., 2(2), 120–122, doi:10.1038/ngeo415, 2009.

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

Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M. and Bonavitacola, F.: Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2, Sensors Actuators, B Chem., 238(2), 706–715, doi:10.1016/j.snb.2016.07.036, 2017.

Sun, L., Wong, K. C., Wei, P., Ye, S., Huang, H., Yang, F., Westerdahl, D., Louie, P. K. K., Luk, C. W. Y. and Ning, Z.: Development and application of a next generation air sensor network for the Hong Kong marathon 2015 air quality monitoring, Sensors (Switzerland), 16(2), doi:10.3390/s16020211, 2016.

De Vito, S., Massera, E., Piga, M., Martinotto, L. and Di Francia, G.: On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors Actuators, B Chem., 129(2), 750–757, doi:10.1016/j.snb.2007.09.060, 2008.

De Vito, S., Piga, M., Martinotto, L. and Di Francia, G.: CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization, Sensors Actuators, B Chem., 143(1), 182–191, doi:10.1016/j.snb.2009.08.041, 2009.

Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L. and Subramanian, R.: Closing the gap on lower cost air quality monitoring: machine learning calibration models to improve low-cost sensor performance, Atmos. Meas. Tech. Discuss., (2), 1–36, doi:10.5194/amt-2017-260, 2017.

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

Atmospheric Measurement Techniques Discussions

Table 1: Gas sensors included in U-Pods along with the model input codes we assigned each.

Sensor Type	NDIR	Metal Oxide					Electro-chemical
Model	S300	TGS 2600	TGS 2602	MiCs-2611	MiCs-5521	MiCs-5525	CO-B4
Make	ELT	Figaro	Figaro	e2v/SGX	e2v/SGX	e2v/SGX	Alphasense
Code	eltCO2	figCH4	figCxHy	e2vO3	e2vVOC	e2vCO	alphaCO

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

Atmospheric Measurement Techniques Discussions

Table 2: **Reference instrument measurements at U-Pod sampling sites**

Deployment	Reference Instrument	Calibration	Operator	Res
Ozone				
CAMP	Teledyne API 400E	quarterly cal/nightly quality checks	CDPHE	1
Dawson	Thermo Electron 49	pre cal/post cal check	INSTAAR	5
BAO*	Thermo Scientific 49c	annual cal/monthly quality checks	NOAA	60
Navajo Dam	Thermo Scientific 49i	quartertly cal/weekly quality checks	NM AQB	1
Bloomfield	Thermo Scientific 49i	quartertly cal/weekly quality checks	NM AQB	1
Sub Station	Thermo Scientific 49i	quartertly cal/weekly quality checks	NM AQB	1
Ignacio	Thermo Scientific 49is	monthly cal/weekly quality checks	SUIT AQP	1
Bondad	Thermo Scientific 49is	monthly cal/weekly quality checks	SUIT AQP	1
Shiprock	Teledyne API T400	quarterly cal/monthly quality checks	NEPA	60
Fort Lewis	2b Technologies 202	factory cal/post cal check	CU Boulder	1
GRET	Teledyne API T400E	quarterly cal/nightly quality checks	CDPHE	1
Carbon Dioxide				
BAO	Picarro G2401		NOAA	1
SJ Basin	LI-COR LI-840A	pre + post cal: zero precision span	CU Boulder	1
GRET	Picarro G2508	periodic zero stability checks	CSU	1

*(McClure-Begley et al., 2016) Res = Time resolution of measurements in minutes

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

Atmospheric Measurement Techniques Discussions

Table 3: **Relative circumstances between model training and testing dataset pairs**

CO_2	O_3	Deployment Time	Deployment Location	Training Location	Training Timing	Incomplete Coverage During Training
◆	◆	Summer 2015	BAO	BAO	Pre Post	
◆	◆	Summer 2015	SJ Basin	BAO	Pre Post	location, pressure
	◆	Spring 2015	SJ Basin	BAO	Post	location, time temp pressure
◆	◆	Spring 2017	GRET	GRET	Pre Post	
◆	◆	Summer 2016	BAO	GRET	Post	location, time
◆	◆	Fall 2016	GRET	GRET	Post	time
	◆	Summer2014	Dawson	CAMP	Pre Post	location, O_3

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

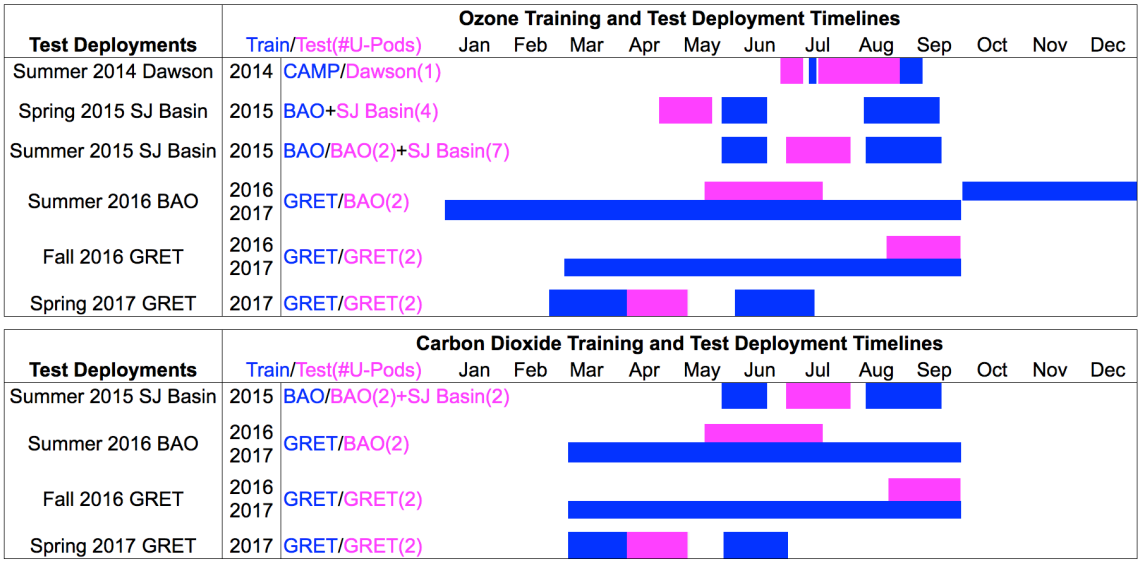

Figure 1: **ANN and LM training and test deployment timelines**

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

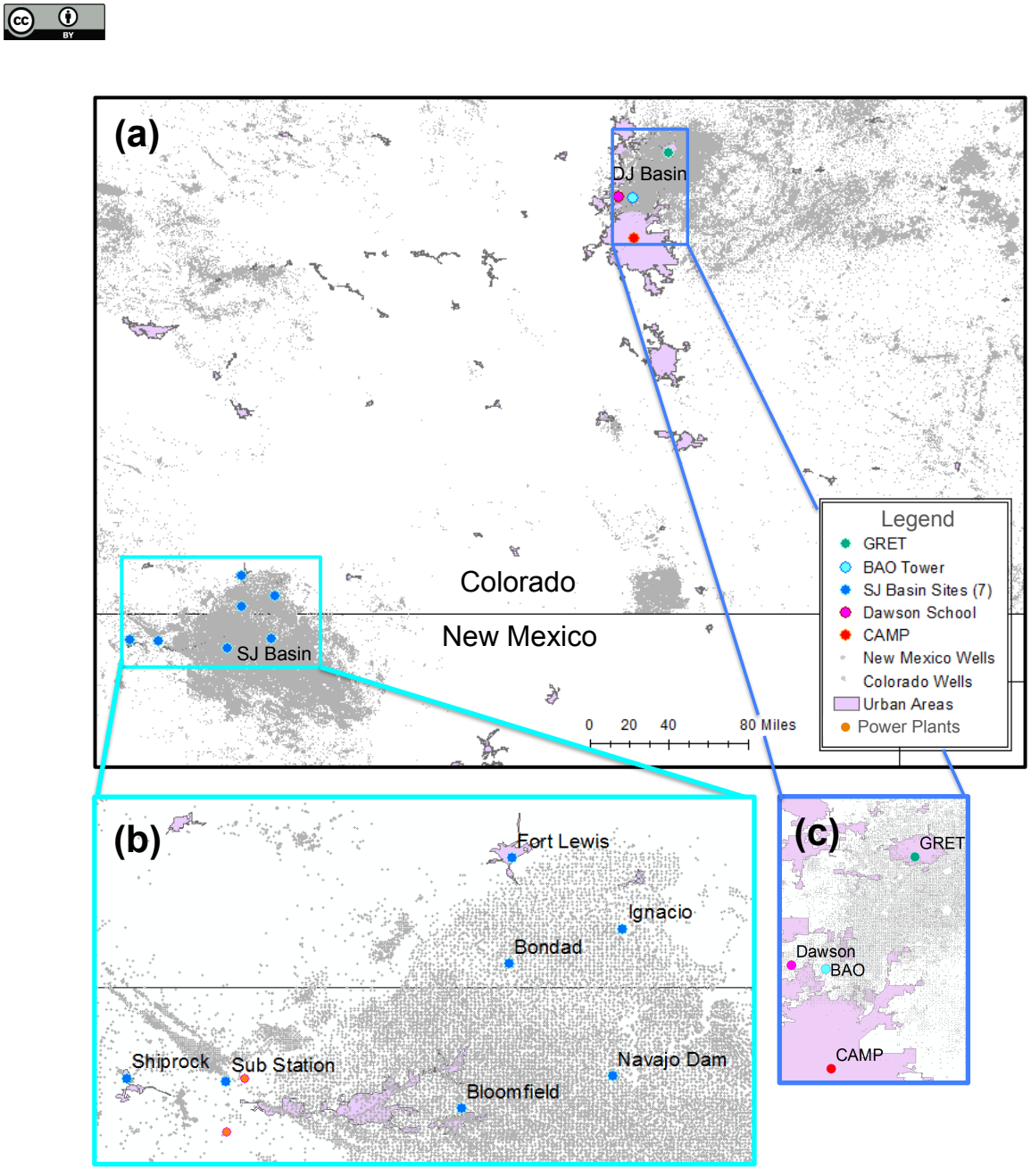

Figure 2: (a) Training and test deployment locations are identified in the SJ and DJ Basins in context with urban centers and oil and gas production wells. (b) Panel zoomed in on the SJ Basin, covering approximately 4,250 square miles (85x50 miles). (c) Panel zoomed in on the DJ Basin covering approximately 1,540 square miles (28x55 miles).

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

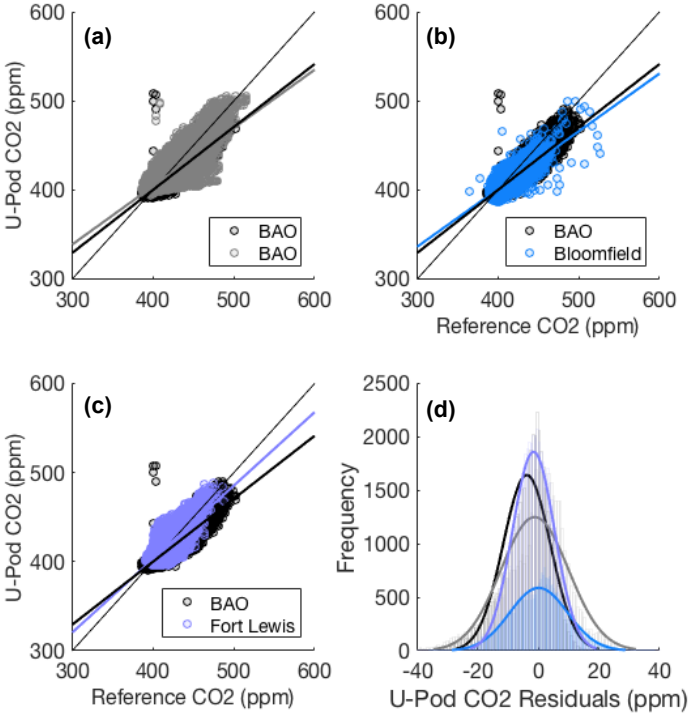

Figure 3: **Scatter plots of U-Pod CO_2 vs. reference CO_2 and overlaid histograms of U-Pod CO_2 residuals for (a) BAO and BAO (b) BAO and Bloomfield (c) BAO and Fort Lewis. A 1:1 single-weight reference line is included in each scatter plot along with double-weight lines of best fit for U-Pods at each sampling location. Data from U-Pod BC at BAO is plotted in black along with U-Pods BJ, BB, and BD at BAO, Fort Lewis, and Bloomfield, respectively. Sensor signal inputs include eltCO2, temp, and absHum. (d) Overlaid histograms of model residuals with respect to reference CO_2.**

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

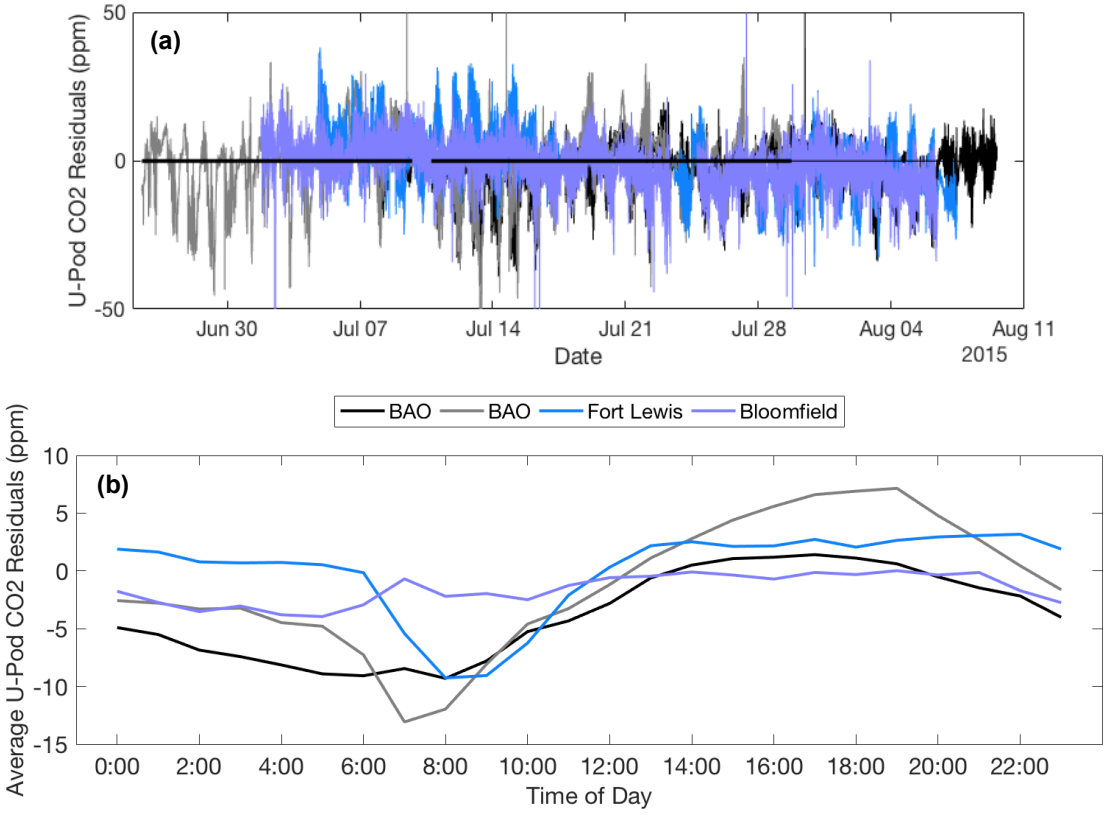

Figure 4: **U-Pod CO_2 residuals by (a) data and (b) time of day and throughout the duration of the deployment period. Sensor signal inputs include eltCO2, temp, and absHum.**

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

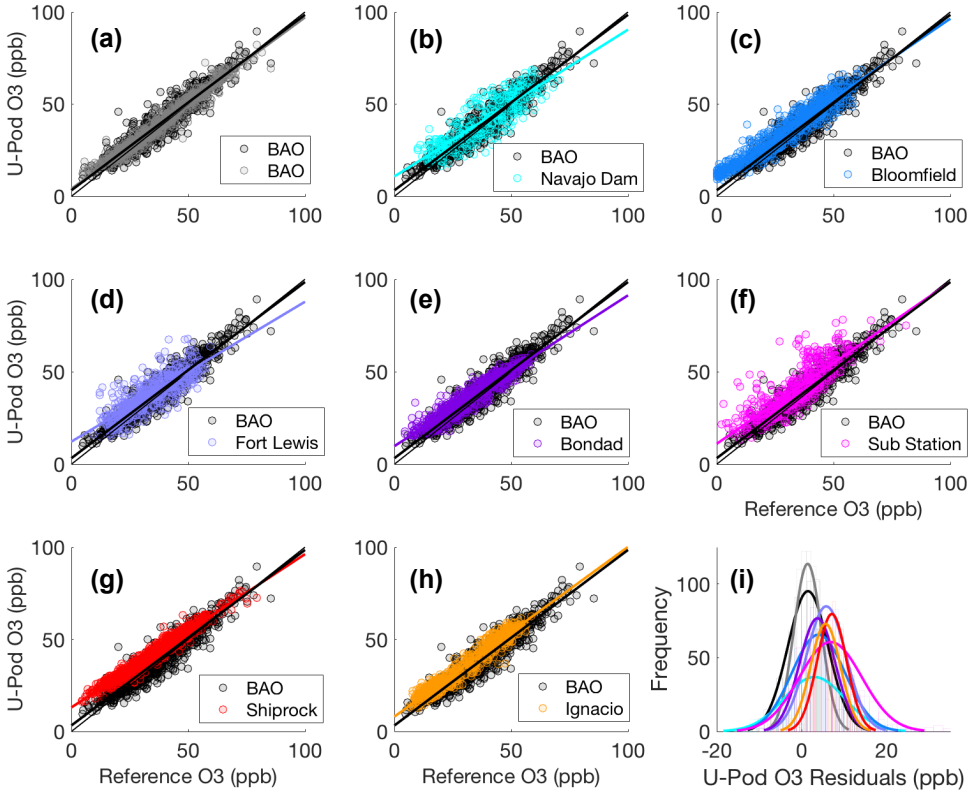

Figure 5: **Scatter plots of U-Pod vs reference O_3, comparing U-Pod BC at BAO, in black, with (a) U-Pod BJ at BAO (b) U-Pod BA at Navajo Dam (c) U-Pod BB at Fort Lewis (d) U-Pod BD at Bloomfield (e) U-Pod BE at Bondad (f) U-Pod BF at the Sub Station (g) U-Pod BH at Shiprock and (h) U-Pod BI at Igniacio in cyan. (i) Overlaid histograms of model residuals with respect to reference O_3.**

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

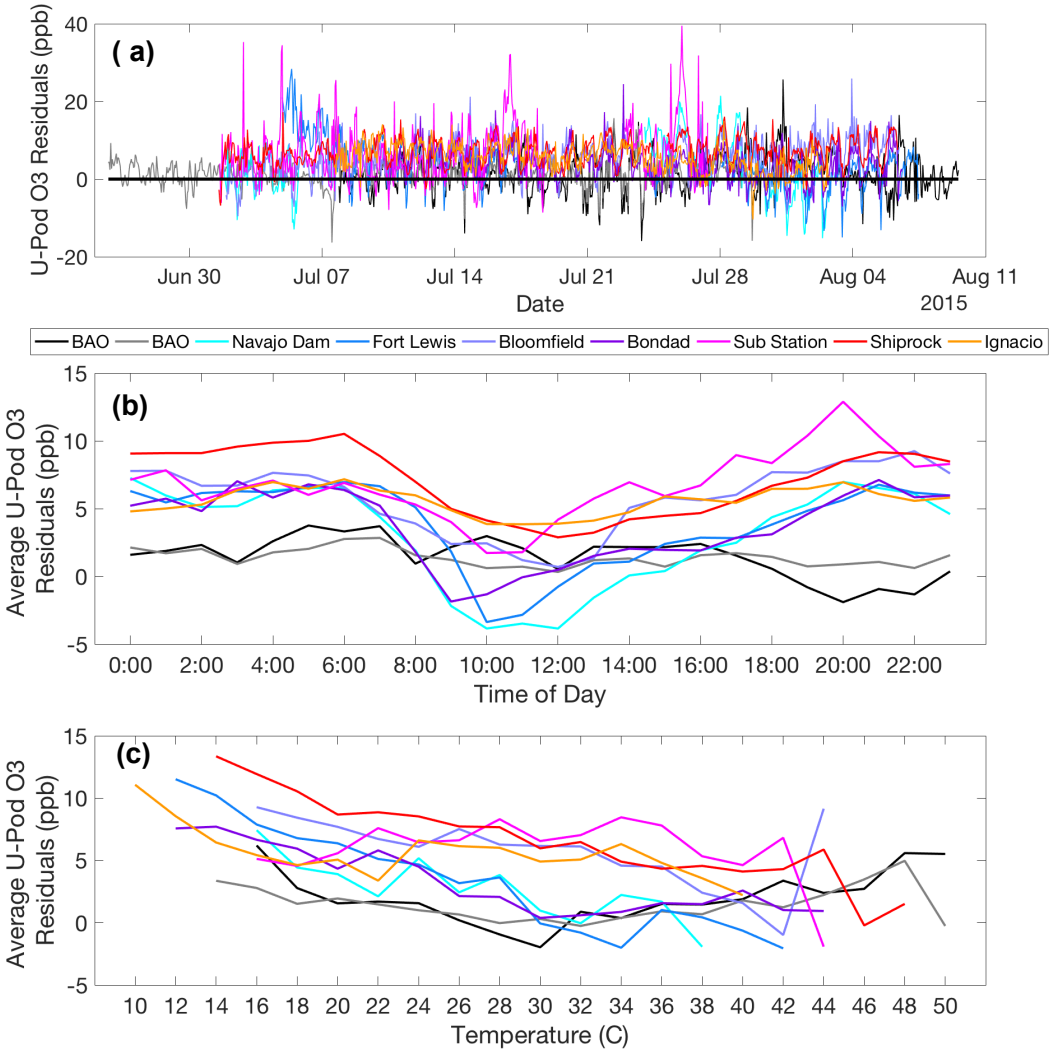

Figure 6: **Residuals of U-Pod O_3 spanning of the deployment period, by (a) date (b) time of day and (c) temperature.**

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

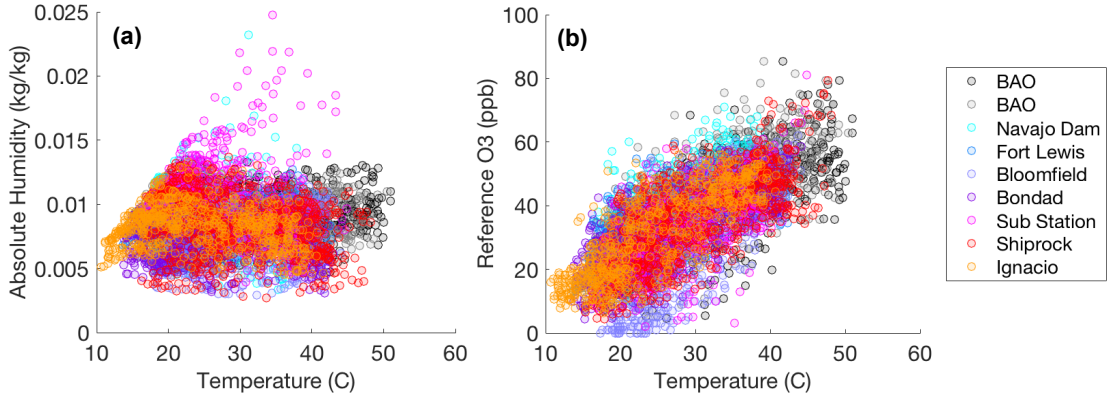

Figure 7: **Scatter plots showing the combined parameter space of (a) absolute humidity with temperature and (b) reference O_3 with temperature for each of the U-Pod sampling sites at BAO and the SJ Basin.**

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

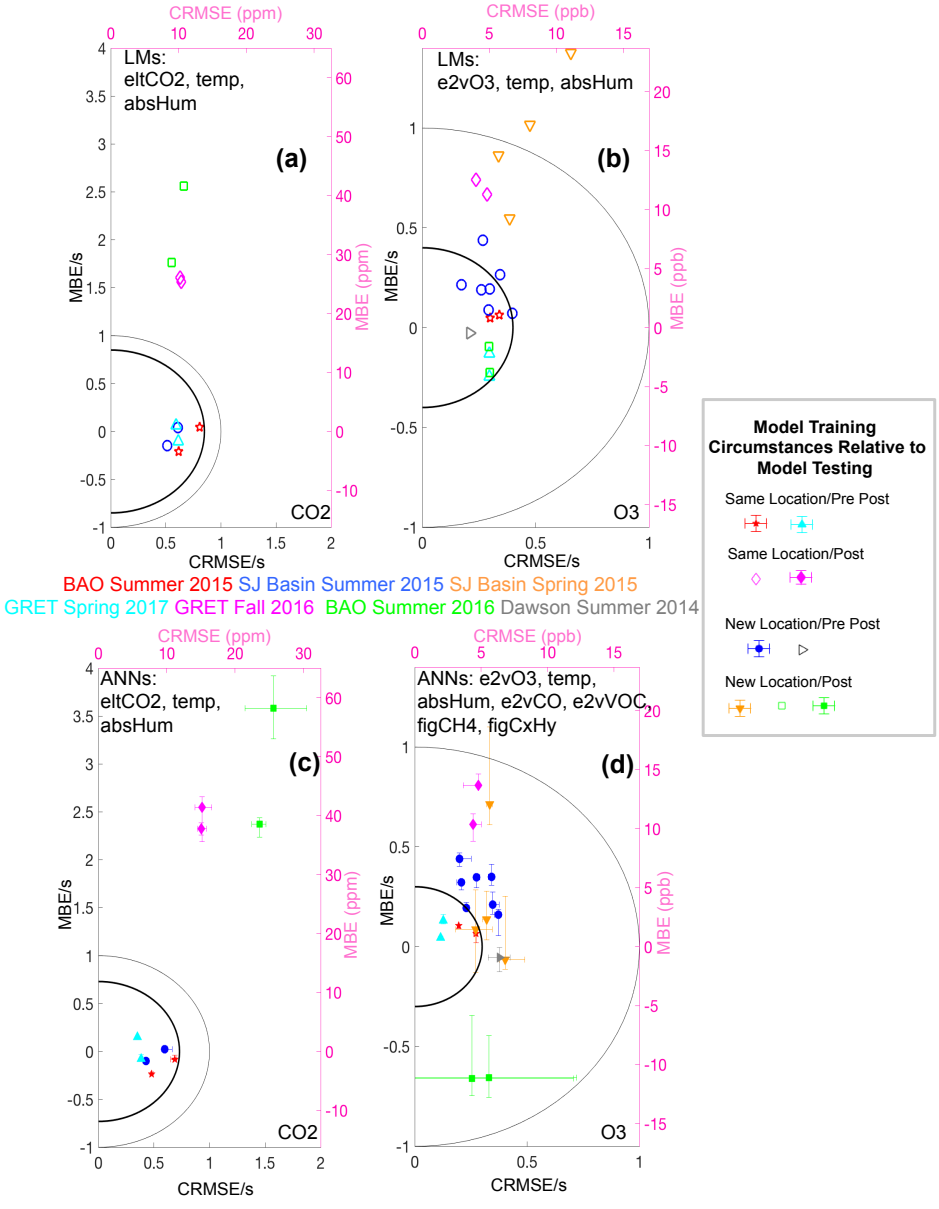

Figure 8: **Target diagrams demonstrating performance of a previously determined best-performing model across all new test datasets. (a) CO2 and (b) O3 LM performance when only the primary gas sensor, temperature and humidity are inputs. (c) CO2 and (d) O3 ANN performance with inputs that were found to perform best at the GRET site in the spring of 2017** (Casey et al., 2017)**.**

Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2018-81
Manuscript under review for journal Atmos. Meas. Tech.
Discussion started: 22 May 2018

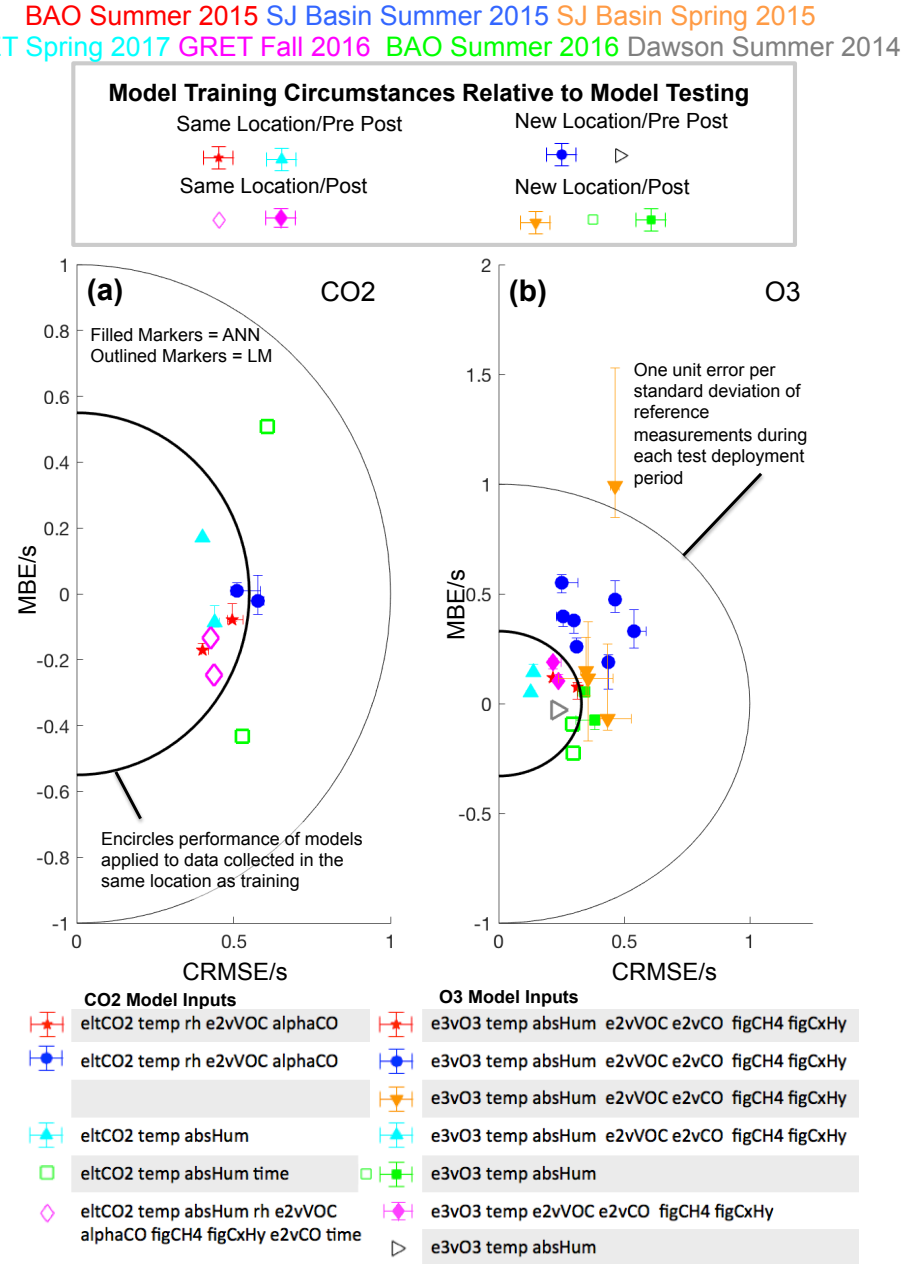

Figure 9: **Target diagrams for (a) CO_2 and (b) O_3 calibration model performance for the best performing model for each particular case when tested on data from a number of field deployments.**