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)

Review -

Casey and Hannigan explore the spatial and temporal transferability of field calibration models (specifically linear models (LMs) and artificial neural networks (ANNs)) for two sensors, O3 (e2vO3) and CO2 (eltCO2), reported by the integrated U-POD sensor package. By 'spatial/temporal transferability' they mean a determination as to whether a calibration model trained from sensor co-location (with reference instrumentation measuring target species) at one location works effectively when that same sensor system is then deployed at a different location. As the authors point out, changing the micro-environment (and local air pollution source contributions to that unique environment) may pose additional complications/challenges when trying to reconcile quantitative measurements with low-cost sensors. The authors make some attempt to separately describe temporal and spatial extension to better understand whether time-alone undermines the accuracy of the calibration models or change of location.

While the topic of sensor calibration and extension of calibration models across a diverse set of deployment scenarios is of fundamental importance to the field of low-cost AQ sensing, the paper, as written, largely fails to pull together a coherent narrative from which active participants in the low-cost AQ measurement space could easily glean useful, actionable information. To be clear, the topic of sensor quantification is inherently complex, and the authors undertake an ambitious analysis spanning 3 years of data from 10 U-POD systems deployed across 4 micro-environments. There are important lessons to be learned from their efforts, but at present these lessons are not brought to the fore of the paper and as a result are easily lost to the reader.

Throughout the manuscript the authors refer back to their published work (Casey et al., 2017). In the vast majority of instances in which this reference is provided, there is little to no contextual detail explicitly drawing the lines of connectivity between the current work and the previous work. Seeking out the exact evidence that exists in the earlier work and relating its relevance to the current work is left entirely up to the reader. Overall, this referencing needs to be done in a manner that is not vague and does not require that the reader be intimately familiar with the previous work. The paper would also be strengthened if the unique and novel insights that result from the current work were more clearly differentiated from the Casey et al., 2017 effort.

There are seemingly contradictory statements throughout the text. These tend to originate from the authors' desire to provide a clear-cut answer as to whether or not a given model 'worked' in a given case study under a given environmental sampling condition. The fact of the matter is, low-cost AQ sensor quantification is extremely convoluted and often times the validity of data can be somewhat ambiguous. Faced with this level of complexity, the current manuscript fails to provide a succinct and systematic

evaluation/reporting approach, and as such main (and important) take-home lessons from their work are lost.

Specific comments:

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.

- L9. Avoid ending sentence with 'to'
- L13: this is one of the core conclusions: the resilience of a given calibration model depends on the circumstances of the deployment for that same sensor system. As such, the paper would be strengthened if the authors focused the narrative on succinctly describing such dependences and circumstances relating these factors back to the sensitivity, selectivity, and stability of each sensor system and sensor type. This language is far too vague, especially for an abstract. What circumstances?
- L15: 'a number' - again, this is too vague. Define exactly how many of the case studies were characterized has having superior AAN models and how many were just as well served with an LM model
- L16: This line suggests that people should model CO2 with ANNs not LMs. The more detailed discussion in the body of the paper contradicts this assertion.
- L19: subscript O3

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.

- L11: What is the difference between supervised learning methods and ANNs? This warrants a more detailed description / definition.
- L15: This sentence (bracketed in red) - is very important, but also very wordy and hard to follow. Related to this assertion, it is not clear how the authors disentangle the temporal and spatial domain from one another, particularly the temporal domain. Time-decay patterns in the data are going to be present whether or not the sensor system has been moved to a different location. How would one ascribe difference in that case to a spatial domain and not temporal domain?
- 'hold up' this language is too casual and used throughout the text. Consider re-wording.

Section 1.2

- L28: 'A number of enclosures..' define the number.
- If Casey et al., 2017 demonstrated the ANN results for CO2 and O3 in the Spring of 2017 in Greeley, CO; is that same data being presented as a portion of this paper (as Figure 1 suggests).
- The concluding sentences of this section nicely frame the motivation/need for the current work, consider bringing this to the fore of the paper / abstract, etc.

Section 1.3

- Final sentence: It's unclear why, if all of the U-POD sensor systems were equipped to measure CO and CH4 alongside CO2 and O3, analogous training/test matrix pairs are unavailable for these other species.

Section 1.4

- L20: 'Very high levels of ozone' – specify the actual concentration or concentration range
- L23: 'a modeling study' – is there really only one modeling study that shows this?
- Final sentence: 'pooling' avoid using words with common association different from the intended meaning. Consider re-wording. 'accumulating'?

Section 2.1

- L5: “with a number of low-cost gas sensors” – specify the actual number of sensors integrated 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.

- The authors identify that 10 U-POD systems were used in the previous and current work, but the vast majority of case studies (outlined in Figure 1.) utilize just 2 U-PODs at each location. The authors need to more clearly describe in the text how the U-PODs were distributed throughout the work and whether all 10 U-PODs used in the current work had the same characteristic O3 and CO2 response when measuring the same air. The sensor system age (time since manufacture date) and environmental-hysteresis (lifetime environmental exposure of a given UPOD system) is not mentioned anywhere in the text. Do these factors not matter when analyzing the temporal extension of a given calibration model? When considering the fundamental measurement principles of these particular gas sensors, does degradation occur due to gradual (or rapid) deposition of material onto active catalytic sites within the sensors? If so, then the age of a given sensor and what’s it’s been exposed to over its lifetime, ought to factor in.. or at least deserve a mention.
- The explanation of the training vs test sampling periods is confusing as written. Given the nature of the experiment, doesn’t each UPOD system have to be co-located with reference

instrumentation for the full duration of the period of study? It sounds as though the authors aimed to bookend the distributed network measurements ('testing period' with a period of co-location at a reference site in the general vicinity of the deployment ('training period') – but in order evaluate their models, they would have to retain a co-located reference measurement of O3 and CO2 at all times in all locations. Looking at the deployment timelines displayed in Figure 1, it is also evident from the Figure (but not from the text) that the vast majority (~75% or greater) of the total deployment time was used to train the nodes not test the resultant calibration models (~25% of the total time). These train-to-test ratios appear to undermine the general applicability of the models to longer duration, distributed sensor measurements in which no co-located reference measurements are available. The authors should make an effort to bridge the gap between how they were able to execute their experiments and how distributed low-cost AQ sensor systems will ultimately be deployed.

- Highlighted in passage above:
  - L10: define the number of sampling sites. Eliminate vague language in the text.
  - L15: same comment.
  - L21: 5 UPOD systems are purportedly used in the Boulder / CAMP 2014 work. Figure 1 lists 1 UPOD system as being active during that test. Reconcile this.
  - L27: Identify the actual ref. O3 measurement in the text here
  - Last sentence: is this relevant to the current paper/study? Not clear what 'study' the authors are referring to in this sentence.

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.

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.

- L9: The authors claim that the SJ Basin network was similarly executed for the DJ Basin. DJ Basin is absent from Figure 1., replaced presumably by BAO. It is unclear how many UPODs were deployed to the DJ Basin. It's very confusing trying to track in time and location the distribution of the 10 UPODs. If I try and decipher the information in Figure 1, either 2 or 4 UPOD units were deployed to the DJ Basin, which on the face of it, does not constitute a similar network deployment of 10x UPODs deployed to the SJ Basin (although, it seems that only 4 and/or 7 UPOD units were deployed to the SJ Basin..
- L13: The authors identify the BAO site as the relevant co-location site for the DJ Basin-deployed UPODS, but then point out that there were NO co-located reference instrumentation accessible for any of the distributed sampling sites. What does this mean for evaluating / testing their models in the distributed network application?
- L14-16:The authors state the GRET site housed all 10x UPOD systems for a year, but Figure 1 indicates that only 2-6? UPOD systems were used at this location and only for shorter periods of time. Again, the text is extremely hard to follow and the information in Figure 1 does not make it any clearer.
- L26: The only metal oxide sensor that's relevant to the current work is the e2vO3 sensor. The operational fundamentals of this sensor should be described: the raw signal processing, circuitry

considerations, and known theoretical operational conditions that undermine the sensitivity, selectivity, and/or stability of the e2vO3 metal oxide sensor.

- L29: 'in a few' Quantify the number of UPODs with faulty RH sensors
- L31: 'nearby': Define the exact position relative to the faulty UPOD

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.

- Did the implementation of radio communication for the UPODs have any impact on any of the other measurements in the system, beyond RH?
- At the beginning of the paragraph, the authors state that the radio communication was active until November, but the substitute RH values from the Picarro were only applied up to October 1 (later part of the paragraph). This is confusing.
- Generally speaking, faulty or absent RH measurements on-board the UPOD (or any low-cost AQ sensor system that suffers from environmental interference) is a potentially widespread issue across the emerging field. I think the authors missed an opportunity to discuss their work-around in more detail and comment on the importance of maintaining stable RH measurements within any given low-cost AQ sensor system.
    - The completely unusable radio communication RH values and the drifting RH values mentioned in section 2.3 beg the question – do the authors think this is a failure on the RHT component itself or the circuitry of the UPODs. Again, if the evidence suggests the former, that is useful empirical data for others in the field.
    - Where is RH measured specifically within each UPOD. Is the measurement internal to the box or positioned in a manner to provide a true ambient RH measurement? What are

the implications of using alternative RH data sources that are not on-board the same UPOD?

- If median values were used for the co-located reference instruments, but the data from those instruments was 1-min averages, how did the authors obtain reference measurement medians at 1-min (the vast majority of temporal resolution used in the current work).
- L19: What % of the total data used in training/testing each UPOD was removed due to this 5-min null data condition?

Section 2.4

- L32 'using methods described previously', given the importance of the LMs and ANNs in the current work, each model should be described in more detail in the manuscript.
- P7L6 – need reference for Bayesian Regularization
- The concepts of early stopping, hidden neurons, and hidden layers need to be described

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.

- Highlighted sentence is confusing as written. How can there be multiple 'best' preforming models?
- Does section 2.2 really succinctly describe each training/testing dataset pair?

Section 3.1

- This is the first place in the text of the manuscript where the limited extent of co-location upon distributed field deployment is described and how the 10 UPODs are reconciled against such limitations.
- For the purposes of the current study, if there is no co-location with reference, is it still a relevant data point? Can the authors effectively 'test' their model under these circumstances?

- This section P8L3 is also the first mention of reducing/oxidizing interfering gas species – this potential deserves a more detailed explanation in the context of the specific micro-environment source contributions
- The overall discussion of factors impacting differences between the two Basin depolyments is fairly scattered. It would be more beneficial to the reader if the authors could draw more specific lines of connectivity between environmental or pollution source contributions and the robustness (or lack of robustness) in the model.

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

- eltCO2, temp, absHum should be human readable, this is the first time these parameters appear in the text. I understand that they were listed in the table describing UPOD guts, but they should be spelled out here.
- L18: it is unclear to what extent the current work and the previous work are duplicated here? Does the previous work form the basis for determining the optimal set of input parameters to train the ANN model and those same set of input parameters were found to be optimal again in this second application or are the actual applications overlapping and therefore the result is redundant? This is an example where I find the self-referential context to Casey et al., 2017 confusing (and lacking specific differentiating information).
- The under-prediction / over-prediction behavior of all four UPODs warrants more discussion. What environmental conditions are pushing the model beyond its limits? What is the fundamental (under-the-hood) reason for the interference in the first place (based on sensor fundamentals)?
- Why did the majority of UPODs stop logging data during the deployment? Did the system over-heat? What fraction of the total possible sample time was missed?

Section 3.1 continued..

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

Highlighted above:

- L6-9: Discussion is confusing and language is too casual: "did not make a big difference" – too vague. Quantify based on the statistical analysis of the model test data. When considering the benefit of including extra sensor inputs in the training matrix for their models, again the Authors are drawing comparisons to their earlier work (Casey et al, 2017) but it's not really clear how this improves/informs the current work – besides stating that the inclusion of the parameters didn't make the data product worse.
- L10 e2vCO2 does not exist as a sensor metric in the UPODs.
- L15: 'all the UPODS' how many is this again?

- L19: ‘For a number of UPODs’: state the number.
- L21: ‘for some of the sites..’: which sites?
- L15-22: this paragraph seems to say that the ANN training matrix determined to be optimal in Casey et al., 2017 was also found to be optimal in the current work, with inclusion of all peripheral sensors to the input training matrix for O3. But they also state that the LMs data products were just as good (or better) when compared to the ANN models. This result seems important, but not really discussed further. The results are left vague. Conclusions as to why this might be the case are absent.
- L27: ‘had bad RH data’ – as noted in a section that doesn’t exist. What is bad RH data?
- L29: ‘relatively far away’ – how far? Again. These details matter.
- L30-31: Apparently one of the major results from Casey et al., 2017 is an extreme sensitivity to RH when using ANN’s to quantify O3. Given the failure of the RH sensor throughout much of the work presented in the current work, it seems critically important that this RH-sensitivity be discussed in much greater detail in the current work, not simply stated in an off—handed matter with a reference to the prior work.
- L32: ‘had a different reference instrument’ what was the instrument and why do the authors think that this particular reference instrument was in error, subsequently disrupting the validity of their calibration model?
- L34 – carried into highlighted passage below: The authors indicate that the sampling sites or the circumstances discussed previously are the reason for the poor model performance, not the sensors comprising the UPODs. First, WHAT circumstances specifically, and what specifically about the sampling sites? This level of non-explanation is unacceptable.

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.

- L22: brief excursions of high humidity – how brief? How high?
- Can the authors comment on the role that humidity transients play in fundamental sensor response? The description of the high and low bias resulting from the models at different locations and different times of day is difficult to follow. What are the common response characteristics and failings of the model that manifest across the case studies featured here? What are the lessons learned and how can these lessons better inform ANN model development moving forward?

Section 3.2.1

- Extrapolation of the ANN and LM models is problematic. Why? If the full-span of O3 (or CO2) concentration encountered in the field deployment is not covered in the training set for the model, is the model incapable of reasonably extrapolating?

Section 3.2.2

- L15: post-test deployment co-locations: It's unclear what is meant by 'post-test', please clarify.
- L16: state the # of UPODs
- The concept of extrapolation in time is confusing. Please clarify what is meant by this? Generating a model at time X and then applying that same model to time X-Y?
- The authors identify coal-fired power plants as an important near-field ('close-by') pollutant source that could contribute a specific (unique) pollutant signature that could render the utility of the Figaro sensor useless. Did the CO2 response of the UPODs or reference instruments or CO

response of the sensor measurements indicate a near-field power plant plume across the deployment area?

Section 3.2.3

- How specifically was 'time' included as a raw input vector in the training matrix? Absolute time? Time since start of deployment? Time since calibration? Time since sensor manufacture?
- L11-12: "…LMs outperformed ANNs with notable instability associated with the performance of ANNs when time was included as an input." In the previous sentence the authors stated that time was useful predictor of CO2.. but the last sentence appears to contradict this assertion. The fact that LMs outperformed ANNs for CO2 also contradicts general assertions made in the abstract.
- The authors should comment on the notion that time-sensitive response patterns in sensors indicates that some level of time-decay. Is this the case with the CO2 sensor and that's why time as a input parameter in the model makes such a big difference? Is there some fundamental reason why the ANNs would be poorly suited to model time-decay patterns in the sensors?

Section 3.2.4

- L23-24 – final sentence in this section is very important. Where the faulty RH (and necessity of substituting RH from alternate sources) degraded the models, if enough RH variability was captured with the suite of peripheral metal oxides sensors, the RH-interference could be effectively modeled without explicit RH inputs. It would seem important to emphasize this point a bit more prominently and discuss further – especially in the context of overcoming some of the RH-measurement shortfalls elsewhere in the manuscript through similar means.

4. Conclusions

- Supervised learning techniques – generally, the manuscript lacks a description of what is meant by this -
- L19-20 the concepts of temporal and spatial extension are still a bit confusing here. Earlier statements to clarify exactly what is meant by each condition would be helpful.
- L24: how does one move something in terms of its temporal coverage?
- L1-3P16: LMs appear to be more robust when applied to a changing deployment condition – but then the authors hedge and say that they "… were not able to fully represent some of the complex nonlinear response behavior exhibited by the arrays of sensors." So a linear model can't model nonlinear behavior? The statement needs to be more specific.
- L7: "..data is almost a band running vertically in a range of CRMSEs." Data running in 'a band' doesn't aid in the interpretation of the data. Re-phrase to address the statistical product that results from the bias that was encountered.
- Final paragraph: how 'generalizable' are the models developed here? It would seem that despite having done an exhaustive amount of work, each individual UPOD system still required its own ANN or LM based on co-located data and raw sensor data from that individual sensor system. While the input matrix of raw sensor signals may be more generalizable, the models themselves appear to be very much node-specific, at least in so far as what has been shown in the paper.
- It is unclear how the extension of the model frameworks discussed in the current paper can be used in the context of low-cost electrochemical sensors