Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2019-30-AC2, 2019

Atmospheric Measurement Techniques Discussions

Interactive comment on "Evaluating and Improving the Reliability of Gas-Phase Sensor System Calibrations Across New Locations for Ambient Measurements and Personal Exposure Monitoring" *by* Sharad Vikram et al.

Sharad Vikram et al.

ashley.collier@colorado.edu

Received and published: 18 May 2019

Comment: "General overview: This paper is a well thought through experiment and has some exciting ideas about the building of sensor networks and data processing to improve or understand error and bias in the systems. It uses a coherent approach, and develops a new statistical method which is mostly well described and accessible to the atmospheric scientist reader. There are a few major areas for improvements which are suggested below."

Response: We thank reviewer 2 for their thoughtful and detailed review and suggestions. We have made substantial and detailed revisions to address the comments, as discussed below. We are grateful to reviewer 2 for their help in substantially improving the paper.

Major Comments

Comment: "Presentation of measured data: Measurement data: despite doing a great job in statically analysis of sensor data, this paper lacks a figure with the epa and low cost sensor measurement data, ideally 1 panel with initial error envelope and 1 with final error envelope. For example no2 sensor measurement (+MAE+95th percentile MAE) (-MAE-95th percentile MAE) (I.e. from Figure 9 level2) - the authors may have better suggestions but if the paper and sensor systems data are to be used by scientists, citizens and community groups this is the information which needs to be presented."

Response: Thanks for this idea. We have developed the suggested graphic and integrated it into Section 3.2 before the old Figure 9 (now Figure 10). To implement the before/after comparison, we used the collected voltages for the "before" data. If you have other or different ideas for the presentation and discussion, we are open to it. The new figure has also been attached below (Fig. 1).

Comment: "Use of reference station data Generally the authors seem to treat the data from the reference station as not existing outside of the experiment and repeatedly make comments about the results in this paper showing new information about pollutant variation at the test sites. This shows a lack of forethought and analysis of existing data. The chemical climate for no2 and ozone at the sites is measured by the reference station, presumably for some years. The information to calculate the variability therefore is already collected and reported somewhere (e.g. EPA or San Diego authority reports) . The authors should use existing knowledge to inform their data analysis and interpretation rather than present the data in a knowledge vacuum. For example

the speculation about the chemical climatology (my phrase not used in manuscript) of Shafer shows a clear disregard for existing evidence outside their experiment which is really avoidable."

Response: The authors thank the reviewer for this observation and have revised the manuscript to address this oversight. In Section 2.1, we have added information on the classification of sites as well as expected influences as defined by the respective air pollution control districts. Additionally, the authors have reviewed several documents from the respective air pollution control districts (including recent monitoring plans) in order to better understand the historic and typical pollutant levels and trends at these sites. A discussion of some of this information has also been added to Section 2.1, providing better context for our understanding of the data throughout the paper.

Comment: "Terminology: The error terminology is not applied consistently through the paper. "Error" is used a lot without it being clear whether it is overall error, expanded uncertainty, a replacement for Mean Absolute Error. Terminology is only introduced on p14 in the results section and then it is only sparsely used after that with terms such as centred error and bias error being added in (p15 line 6) This lack of clarity in the terminology leads to more questions than are answered in the manuscript e.g. âËŸA ′c How do the authors assess the total error with the set of benchmarks âËŸA ′c When the difference in error is discussed (e.g. p14 line 11), which difference is being discussed? âËŸA′c In the discussion the authors do not specify which error they are referring to e.g. p16 line 6 onwards. Difference in error is discussed in qualitative terms despite there being quantitative data in the work. Significance of differences are not discussed. Overall the authors should revise with a consistent terminology and perhaps a table glossary. Also only MAE is shown in the main paper (Figure 6), please could the authors add the equivalent plots in the same figure for each of the benchmark errors. It is hard to assimilate the many tables in the Appendix with the results. The tables potentially should be moved into the main manuscript."

Response: We apologize for causing confusion. We have made several edits through-

out the paper to clarify which error we are referring to, when relevant. Although we present several errors in the tables to support direct comparison with multiple previous works, in general, we expect the various errors to track each other in relative, if not absolute, magnitude. Thus, any conclusion that we make using one error could be reached using one of the others.

Comment: "Discussion of cost Interestingly I think there are 2 points from this paper: the new NN splitting method offers improvements for incrementally built networks and the Simple linear algorithmic gives the best understanding of changes. The former involves large amounts of expensive model development and potentially lack of clarity as to why the model works to improve sensor data (to a non-scientist sensor user). The cost implications for what is now not low cost at all (enclosures, powered fans, telemetry, expert algorithm development and maintenance) seem to be not openly explored, rather a nebulous "small cost increment for new nodes" offered as a positive: Could the authors perhaps discuss whether there is unconcious bias in their cost increment assessment?"

Response: One of the goals of the low-cost sensing community is that eventually both the hardware and software will be available more or less "off the shelf", mitigating their costs for end-users. Today, however, many of the sensors and most of the accompanying software are research prototypes that are designed more for open-ended experimentation than end-use sensing. To support the ongoing transition to practice, we have now published an archival repository containing all of our hardware plans, software, and raw data, and cited it at the end of the Introduction. We have also made several changes to Section 2 to make it clearer what infrastructure is used for research versus calibration versus application. This includes an added picture on the right side of Figure 2 and an extended introduction to Section 2.3 Data Collection. If it is believed that the paper should be more explicit on the intended meaning of low-cost sensors, we could add a footnote to the Introduction.

Minor Comments and Corrections

Comment: “Abstract Could do with quantitative analysis of results in abstract e.g. how much does N-N improve model?”

Response: The results in the abstract have been made more precise both qualitatively and quantitatively, particularly for the split-NN model. In general, however, because so many models and training configurations were evaluated, a concise quantitative summary is difficult, and ultimately we found reference to the box plots served best. If there are ideas for a more concise quantitative summary, we are open to it.

Comment: “Introduction - Well written and readable. P4 lines 20-35: the list of the results does not fit in the introduction”

Response: We have moved the list of results to the conclusion (replacing and integrating the straight prose) and added additional detail supported by the body of the paper.

Comment: “Methods - Section 2.1 sampling sites: the authors describe the sampling sites “expected profile” in descriptive terms. Given that the sites are regulatory local environmental and emission metadata for the site is probably available in the EPA station records. No references for the regulatory station information or data is provided or links to EPA reports using site data.”

Response: As described in response to the major comment regarding the use of reference station data, we have added information from the appropriate air pollution control districts, including references to official documents.

Comment: “P6 line 13: “over the air” what does that mean?”

Response: We apologize, it means wirelessly. We have replaced this phrase with something more descriptive.

Comment: “P6 line 34: the noise on the signals and SD are discussed cf the sensor data. It would be useful to have the statistics of the raw (level 0) data and the “cleaned” or level 1 data for each sensor deployment as per epa site format used in Appendix

B- or in a table in the main paper as it is critical for understanding the data processing effects”

Response: Thanks for noting this omission. We have added details about how much data was filtered to Section 2.4 Preprocessing, in particular that 2.4% of the 5-second data was filtered. It was not meaningful to report this in Appendix B, since that reports minute-level data.

Comment: “P7 line 6 is the sensiron accurate to 0.05% rh or is it just the data resolution on readout. Please could the accuracy/precision be stated. Resolution is not really useful.”

Response: The accuracy of the humidity sensor was added. Thank you for the feedback.

Comment: “Section 2.2.1 and 2.3 : the passive electrochemical samplers are placed in an actively ventilated housing “for this study”. If truly for only this study and the sensors would be deployed differently under a normal operation, how are the results relevant to different setups?”

Response: The expected use case for deployment for the sensors is that they will be exposed to ambient conditions and not placed in a larger enclosure with limited airflow. In real-world use cases, when the sensors are attached to backpacks, bikes, etc., the air flow would be sufficient for the sensors to sample ambient conditions. In our extended deployment, the sensors were placed inside of larger enclosures with small ports, so active ventilation was used to push air into the box. We have clarified these details in section 2.

Comment: “2.3 Data is stored in the cloud: are they available for the public? What is the archive for data (and data identifier)?”

Response: We have created a separate repository that is now linked in the paper in Section 6, after the acknowledgements. It also includes are hardware plans and

software.

Comment: “P9 line 9: was the data from the reference station provisional or final ratified, I.e. regulatory automatic network data is ratified on a cycle. What was the date of data provision and data capture statistics for those periods? It is not enough to just say they came from the EPA. Data should be referenced properly.”

Response: The reference site data utilized in this analysis was not final ratified data as the timing of our study did not allow us to wait for this version of the data. Regarding the reference data, we did remove data collected during calibration periods as well as any data flagged during initial QA/QC by the regulatory agency who supplied the data. We have added these details to the end of Section 2.3 Data Collection. We have also added a note clarifying that the reference data used had not undergone complete QA/QC procedures and therefore is not final data from these stations in Section 6 Code and Data Availability.

Comment: “Results: p14 line 15 is the difference between level 2 error vs level 1 statistically significant? It looks quite small on the Figure.

Response: In this and the following sentence, we are emphasizing the *slight* improvement for Level 2, conveying that we feel that the effect size is small. Although the difference is likely significant given the size of the datasets, it would convey the wrong message since the effect is not especially impressive. However, we’d be glad to add this detail if it’s desired.

Comment: “P9: averaging of minute data: arithmetic mean, time weighted average? How are data gaps in a minute treated?”

Response: To make the sentence in 2.4 Preprocessing clear on these questions, we rewrote it as follows: “For the remaining data, a simple average was computed over each one-minute window so as to match the time resolution of the data from the reference monitors. If an entire minute of data is missing due to a crashed sensor or

preprocessing, no minute-averaged value is generated.”

Response: “Data filtering: filtering for “the realm of reasonable values” probably needs explaining more completely. Please list the QA filter steps in the appendix. Just to note, short lived plumes do not give reasonable values when you are normally used to looking at average values, but they may be real and relevant. What does +5V represent for each parameter? Probably the filtering is fine, but from the paper I cannot tell that.”

Response: We believe we have sufficiently described these steps in section 2.4 Preprocessing, as they are simple threshold filters. The filtered data are not just spikes, but values that are simply not possible, either physically not possible or literally out of range for the sensor and represent a hardware or software failure.

Comment: “P10 lines 1-5: you can tell that that would be overlap from the reference site data. Not necessary to confirm it with sensor measurements.

Response: We apologize for the confusion generated by this section. Here we intended to state that the hypothesis (i.e., that the pollutant trends would vary between the different reference sites) had been verified by examining the distributions of the reference data. Furthermore, that the expected trends seemed to be reflected during the period of our deployment. The wording in Section 2.4 has been adjusted to clarify this point.

Comment: “P15: It is interesting so see the change in bias between level 1 and 2 and I feel it should warrant further discussion in the manuscript. How many extra levels would be need to achieve an acceptable bias?”

Response: This is a great observation by the reviewer, one that we feel emphasizes an important take-away in the paper. Based on the reviewer’s comment we have enhanced the discussion in Section 3.1, following Figure 7 to better highlight the importance of error due to bias vs overall error. Regarding the question of how many levels would be necessary to achieve an acceptable bias, we feel that determining the precise number

of levels is somewhat beyond the scope of this paper - given that we do not have enough reference sites to continue exploring the question beyond 2 levels. That being said, this would be a valuable question to explore in future work.

Comment: “Figure 7: this figure needs a more complete caption to describe the graphs”

Response: A more detailed description of the target plots has been added to Figure 7.

Comment: “P16 line 10: which error metric?”

Response: We added “in MAE” to clarify that error is reported in MAE.

Comment: “Figure 8: no x-axis label”

Response: Fixed.

Comment: “Figure 9: please match scales on the two NO2 graphs and the two O3 graphs for ease of comparison. Also putting zero or an integer at the origin would be good practice.”

Response: Thanks for catching this mistake. We have normalized the axes and 0-based them.

Comment: “P20 line 16. The authors mention using the 5s data to get more information and improve data quality despite the response time of the system likely to be not 5s (not quantified in this paper) and no reference given to show that this would a likely significant improvement rather than addition of more noise. It would be useful if the authors expanded on why they are optimistic about this.”

Response: We didn’t mean to express optimism, only potential. We have added a comment about the possible impacts of noise and the system response time, citing back to section 2.1.1.

Comment: “Fig A1-A3: Figure captions could be more explanatory. What is the line vs the bars?

Response: These have been clarified in the text. Each bar represents the total proportion of measurements at the given temperature or humidity (a histogram plot). The lines are a visualization of the kernel density estimation of the raw measurements.

Comment: “Appendix B too many decimal places in the tables!”

Response: Good point. We have trimmed them down to three decimal places to match the later tables and removed the decimals for the integer values.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2019-30, 2019.

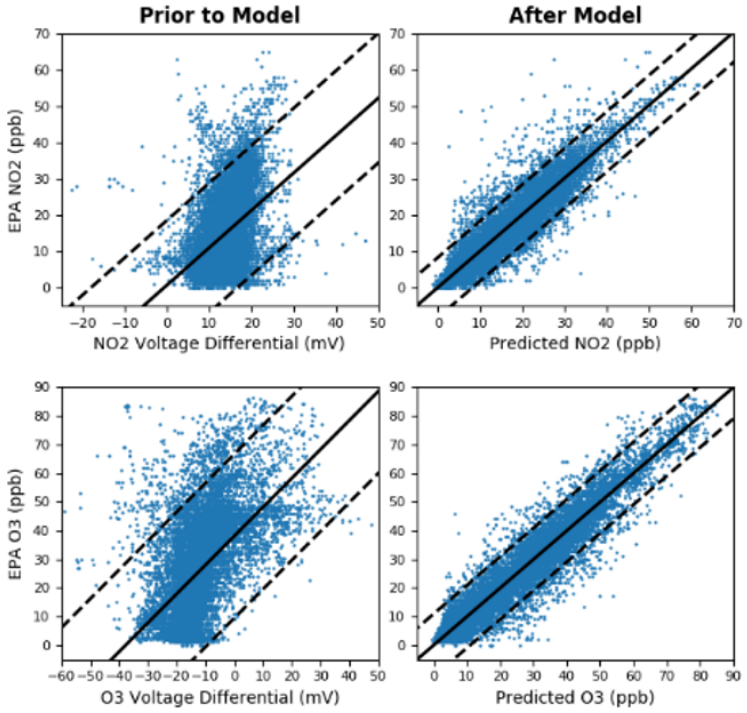

Fig. 1. A single board comparison of the relationship between the raw sensor values and target pollutant concentration.