Closing the gap on lower cost air quality monitoring: machine learning calibration models to improve low-cost sensor performance

Review of Zimmerman et al.,

Eben S. Cross, Leah R. Williams, Gregory R. Magoon

Aerodyne Research, Inc. Billerica, MA 01821 USA

As written the Zimmerman et al., manuscript is an important contribution to the growing body of low-cost AQ sensor characterization efforts. The tone of the manuscript is a bit over-stated (incl. the title), as pointed out by the other reviewers, and the overall impact of the work could be improved if the authors more carefully addressed the following points of concern:

- Scope of work completed
    - The manuscript strongly emphasizes the unprecedented scale/scope of the completed work, stating that 19 RAMP systems were deployed for 6 months. At face value this would constitute a ~ 24wk interval across which to train & test the model. The actual reported tests appear more selective (both in terms of the number of RAMPS and duration of testing interval). As written, this is somewhat misleading. The authors should make an effort to more clearly state the scope of work as it pertains the results presented in the paper.
        - Pulling data reported in table 3:
        - CO: Test data spanned as few as 10 days, up to 108 days with an average of less than 6 weeks. Figure S7 shows only 16 of 19 RAMPS for evaluation (despite fact that 19 systems were RF-trained)
        - NO2: Test data spanned as few as 2 days up 56 days with an average of 3.4 weeks. Figure S8 shows only 10 of 19 RAMPS were evaluated (despite fact that 19 systems were RF-trained)
        - O3: Test data spanned 11-103 days (average less than 6 weeks) with 16 out of 19 system evaluated
        - CO2 15 out of 19 systems evaluated and the number of days of test data were not tabulated.
        - What is the fraction of training-to-test data for each RAMP system for which statistical metrics were reported?
        - Data displayed for RAMP #4 in Figure 8 shows 15 weeks of test data. From the average number of test sample days reported in Table 3, is RAMP #4 a significant outlier? Did the majority of other RAMP systems run for shorter periods of time?
    - While the authors point out that the limited NO2 training/test data was due to a malfunction in their reference monitor at the co-location site, that does not explain why only 10 out of the 19 RAMP systems which were trained with the ambient RF model were included in the presented results.
        - The authors should comment on the impact of the significantly shorter evaluation period on the NO2 results. Specifically, did the loss of the NO2 reference monitor exclude data sampled over the colder or warmer

seasons in Pittsburgh and if so, how would this impact the range of conditions across which the RF model was found to be robust?

- Laboratory calibrations
  - As the authors' correctly point out, laboratory calibrations have formed the basis for much of the low-cost AQ sensor characterization work completed to-date. The manner in which the laboratory calibration experiments were executed in the current work raises a number of concerns:
    - The authors should justify their laboratory calibration approach, specifically, sampling the sensors under 9 LPM of active flow, under air compositions dominated by (presumably) clean air, doped with single species of interest (excluding O3) under RH conditions that are outside of the specified operating range of the electrochemical sensors being trained. Given that these sensors operate under diffusion limited conditions, active vs passive flow can have a significant effect on the rate with which analyte molecules reach the working electrode surface of each electrochemical sensor. From the picture of the RAMP node, it appears that when fully integrated, the sensors are positioned to sample the air passively. This disconnect between the LAB cal. conditions and the ambient sampling configuration should be addressed if the authors are honestly trying to assess the validity of the LAB model on reconciling ambient concentrations from deployed RAMP monitors.
    - The lack of any systematic logging or control of temperature and RH under these laboratory conditions limits the overall usefulness (and relevance) of the laboratory calibration to reconciling ambient concentrations. While the LAB model is limited in its sophistication, the execution of the lab experiments themselves also presents environmental conditions that do not overlap with their ambient co-location conditions. This apparent disconnect between the LAB and field needs to be explained further.
    - The absence of any O3 lab calibrations needs to be explained further. Why was this species excluded and given the RF model assessment of the Ox-B431 sensor sensitivities to different parameters, do the authors think this sensor type would provide more reasonable LAB-based calibration models, if such experiments had been conducted?

- RF model
  - With access to 1s reference monitor data it is not clear why the authors chose to use 15 min averages to train and test their RF model. Were shorter or longer time-averages tested and found to be measurably worse than the 15-min averages? What are the implications of using 15-min average data vs 1 or 5-min average data when resolving heterogeneity in local pollution gradients?

    - The authors should expand on their discussion regarding the lack of any extrapolation in the RF model.
        - (related) Figure 5. For RAMPS #9,12,13,18 the authors should explain the straight vertical and horizontal at the ~ (50,50) x,y position on each scatter plot.
    - It would be informative if the authors could comment on the computational cost of running the model. Does this computational cost place constraints on the time-averaging used to train the model in the first place?

- P13 discussion of explanatory variables
    - What do the authors mean by permuting? Replace with another dataset that's not related to the current dataset? A more thorough explanation of this process is warranted as this process appears critical to evaluating the importance of various interfering factors on each sensor type.
    - Figure 9. Why is CO2 more sensitive to CO than CO2?
    - The authors state that SO2 concentrations were below detection limits for the duration of the ambient co-location study and therefore not discussed further in the manuscript. While it is true that the SO2 concentrations in Pittsburgh are very low, the extent to which the SO2-B4 sensor output informed the RF model is in fact statistically significant according to the data presented in Figure 9 which indicates that the MSE can change by ~ 20-40% when the SO2 sensor parameter (presumably differential voltage?) is permuted? A more robust assessment of the importance of the SO2-B4 sensor data to the resulting RF model may be to exclude it altogether from the available input parameters used to train the model.

- All goodness of fit discussions relative to Cross et al., 2017 need to be revised according to the results published in the final accepted version of that manuscript.

Additional comments –

- P11 L15: The figure caption does not indicate this…
    - Figure 4 shows the calibrated RAMP #1 output regressed against the reference monitor concentration for the entire testing period for all three calibration models (LAB, MLR, and RF).
- P12 L20: The text states that the MAE comparison is against the number of points, but Figure 9 displays this data versus the number of weeks, not number of points.
- First paragraph of section 2.2 is unnecessarily repetitive
- 95 sensor measurements (should be 76)..
- P7 L6 ‘beta4’ should be ‘beta3’ according to the formula above
- P9 L20 missing ‘resolution’ following ‘temporal’
- P11 L13 Figure 2 should read Figure 3
- P18 L7 missing ‘this’ - as written: ‘demonstrate that degree’
- P20 L30 Levy 2014 reference is the same as Moltchanov et al., 2015 reference.
- Figure 2 caption should specify units as ‘a.u.’ following >255.9

- Figure 4 (left) – why do the four different pollutant times series all have unique time-periods? If environmental parameters impact the sensors differently (RH, T) then it would be important to keep these parameters self-similar across the evaluation-framework presented here (even though it's only 48-hours worth, should be the same 48 hours for all sensors).
- Figure 7. It's not clear why there are ~ 10 or fewer data points displayed when data from 19 RAMPS are reportedly presented
- Figure 8 caption. 'long periods' is relative. Data displayed is for 15 weeks. Lifetime of the sensors is significantly longer than this (~100-150 weeks). Language should be revised accordingly.
    - The extent to which the model improves over time should be quantified with 95% confidence intervals on the linear fits. By eye, it looks like this confidence interval would include 0.
- Table 3. Rather than identifying the number of days of sampling/evaluation – it would be more appropriate to identify the total number of data points used in each case study.
    - Add an extra column that identifies the time resolution – as this is an important factor that drives signal-to-noise and accuracy and precision metrics as well as various end-use cases of interest.
- Section 4.4. As written, this section oversimplifies the reality of the situation. When analyzing various lower-cost AQ sensor systems it is important to recognize that the combined hardware and software configuration impacts the performance metrics, not the software alone. The authors shouldn't gloss over this fact.