Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2018-216-RC1, 2018

Atmospheric Measurement Techniques Discussions

Interactive comment on "Development of a General Calibration Model and Long-Term Performance Evaluation of Low-Cost Sensors for Air Pollutant Gas Monitoring" *by* Carl Malings et al.

Anonymous Referee #1

Received and published: 24 October 2018

This paper examines various approaches for calibrating low-cost air quality (AQ) sensors, with the goal of making recommendations for effective calibration strategies, especially over relatively long timescales (months to years). This is an important and timely topic in atmospheric chemistry, and is certainly will be of interest to the readership of AMT. The key result, that a "generalized calibration model" - in which a number of sensors are calibrated via collocation at EPA-grade AQ monitoring sites, and the average calibration be used for all sensors – provides adequate accuracy for many applications, is certainly a useful and important one. However, a weakness of this paper is that the analysis approaches taken are not always clearly described (or well-justified) in the manuscript, so it is not always obvious how general the conclusions are. In par-

ticular, the calibration approach over longer terms is not well-described, and appears to involve an test/training approach that is different than what would be used under most conditions. Thus the general validity of the recommendations (that new models should be developed every year) is unclear. These issues, described below, should be addressed before this paper is published in AMT.

(format is "pageNumber_lineNumber")

The way one would calibrate sensors under standard deployment conditions is to collocate at an EPA station for some period of time (or to calibrate in the lab), then deploy the sensor to some other location of interest to make new measurements (possibly returning the sensor to the collocation spot later on for re-calibration). But this doesn't appear to be the approach taken here, where the long-term data (Figure 6 and 7, and accompanying text, p. 14-15) seems to have training/test data taken throughout a given year. (Though this isn't well-described in the manuscript – what were the training and test times? How were these chosen?) If the test data is indeed taken throughout the year, this isn't really a realistic calibration approach, so it is unclear to me how the authors can make recommendations about how or how often calibrations should be done (1_20, 14_29).

Similarly, from Table 4, it appears the training and test sets cover nearly identical ranges in pollutant concentrations – to within 1 ppb for the four gases (CO, NO, NO, O3). How is this possible? On implication of these identical ranges is that the performance of the hybrid approach (discussed in sections 2.3.5) cannot really be distinguished from the non-hybrid approaches. This is mentioned near the very end of the paper (17_17), but really should be discussed sooner, and the hybrid and RF-only models probably should not be discussed as separate approaches. If they are, the number of "crossings" (switches from RF to LR, fraction time evaluated by RF vs time evaluated by LR) should be discussed.

4_24-29: The gRAMP approach involves selecting a subset of the sensors for calibra-

tion and seeing how the others do with this calibration. However very little information was given on which sensors were used/withheld. Presumably these were sampled randomly (via a k-fold cross-validation, etc), to make sure the selection of sensors in the training set did not bias results?

7_22 (also 2-26): The authors describe the work by Hagan et al. as a clustering approach, but this is incorrect – the authors may be confusing k-nearest-neighbors (kNN, used by Hagan) with k-means-clustering (used in this work). kNN is not a clustering method; clustering in k-means-clustering is computationally much less intensive than storing and comparing to every input-output pair (as is done in kNN), but it can also lead to a dramatic degradation of the quality of the training dataset. Thus, the present k-means-clustering results cannot be compared to the approach of Hagan et al.

Overall: the authors may want to reconsider their terminology, given they are trying to make general recommendations for sensor use, including use of non-RAMP AQ sensors (this is the focus of section 4.1). I would recommend using terms to describe the models that are more general and non-sensor-specific than iRAMP, gRAMP, etc.

Minor comments

- 2_20-22: The wording here should probably be softened somewhat; it is challenging (but not impossible) to access all relevant atmospheric concentrations in the lab.

- 3_21: Small typo: the company name is Alphasense, not AlphaSense.

- 4_24: the gRAMP approach has some similarities to the averaging approach taken by Smith et al. (Faraday Discuss. 2017, 200, 621-637); while there are differences in these two techniques, this previous work should certainly be acknowledged here.

- 5_6-10: how were these cutoffs (15 minutes, 21 days) chosen? It's stated the 15 minute averaging was chosen to reduce noise, but results from other time intervals (1 min, 5 min, 1 hour, etc) are not presented. Are the data so noisy that such averaging is necessary? (Or is this just minute-by-minute variability?)

- Figures 2-3: What do the error bars refer to here - the spread among individual sensors? If possible, it might be more useful to show the data from each individual sensor here.

- 7_30-8_1: since this issue is important to all nonparametric models, as the authors state, this point should be made earlier, not just in the section on k-means-clustering.

- 8_30-32: this sentence implies that the hybrid approach was developed by Zimmerman et al. and used by Hagan et al. My understanding from those two papers (and from the timing of the original AMTD submissions) is that Hagan implemented it, and it was mentioned as a potential approach by Zimmerman.

- 9_13 (section 2.4): this is a very useful section, but it should be highlighted that these metrics are used on the test/validation data only.

- p10-14: Here there is a lot of text describing the individual figures. All this detailed information was rather hard to follow, and hard to glean what the major results were; a "bigger-picture" discussion of what the figures tell us might be helpful.

- Figure 5: how many sensors are we talking about here? Were all of them moved?

- 14_10-12: I don't follow this sentence. If the models are trained and tested on data from both years, how can a change in model performance indicate a change in the models? Do the authors mean a change in the sensors themselves (as discussed in the next paragraph)?

- 14_24: If the sensor is degrading, its output signal will probably be lower for a given amount of pollutant. Is this observed? If not, what evidence is there for degradation, other than a change to the calibration?

- Additionally, in 14_28: I think the problem is not that the electrode material (typically some metal) is "used up" but rather that the electrolyte concentration changes over time, by either evaporation or leaking.

- Table 4: this table is very useful, but a bit hard to follow in its current form. Some suggestions/questions: - since it's not relevant to the text, maybe remove CO2, avoiding the ppb/ppm problem - the concentrations (as measured by FEM/FRM monitors) are in the "LR" row, which might suggest they are relate to the linear regression. Maybe move them to the header of each pollutant, to separate the calibration technique used from the data - the column title "Models" isn't clear - a CO concentration of 7ppb is unusually (maybe impossibly) low – remote regions generally have levels of $\sim$100 ppb. It might be worth checking that dataset. - T and RH ranges should be included.

16_11-14: This statement is based only on comparisons of sensors run under different conditions at different times and places. Comparisons like this can only really be made when different sensors are studying the same airmass.

- Citations: twice (Hagan et al., Sadighi et al.) the AMTD citation is used rather than the AMT one.

- SI: making the data publicly available is a really excellent feature of this paper, and a good template for other sensor papers. However the file is almost 14GB! It might make more sense to provide just the raw data, and the scripts used; most users will want the data only. Those that want to examine the model output can run the scripts themselves.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2018-216, 2018.