Atmos. Meas. Tech. Discuss.,
doi:10.5194/amt-2017-183-AC1, 2017

Atmospheric Measurement Techniques Discussions

Interactive comment on "Intra-urban spatial variability of surface ozone and carbon dioxide in Riverside, CA: viability and validation of low-cost sensors" *by* Kira Sadighi et al.

Kira Sadighi et al.

kira.sadighi@colorado.edu

Received and published: 10 December 2017

Anonymous Referee #1 Received and published: 26 September 2017 The manuscript describes a system of two low-cost sensors, one for CO2 and one for O3, that were calibrated and deployed in Riverside, CA. This area of research is very active right now and this paper is timely as there are not a lot of studies that are published on low-cost CO2 sensors specifically that rigorously evaluate their performance, although many groups are working on this type of sensor. The paper is mostly well-written and organized, although there were a few confusing points to me over the language and relationship between the calibration, collocation, validation, and deployment periods

as described. Some of this confusion prevents the reader from really understanding the uncertainty evaluation made for these sensors. Specifically for the CO2 measurements, there is no clear explanation of large differences between a low- cost sensor that is co-located with a reference sensor over all the time periods. I am not convinced that the stated uncertainty of 15 ppm is valid. The authors also note some issues with large shifts in calibrations of the CO2 sensors due to manufacturer's software, leading to eliminating some of the sensors. Given these issues, I am not sure the CO2 sensor portion of this paper is useful to a reader without further clarification and perhaps additional analysis of the existing data set. The ozone sensors seem to have been more thoroughly evaluated. More on these and other comments are below. I recommend publication only after some major changes to the manuscript addressing the issues.

The availability of the CO2 data, due to technical issues with our certain set of sensors, has made evaluation of that dataset difficult. After further consideration, we have decided to remove the CO2 sensor results from the paper. We feel that it makes the analysis of the ozone sensors less strong. Throughout the paper, we have added sentences that hopefully clarify for the reader which time periods are being discussed, and where the U-Pods are during those time periods. For addressing the specific comments, most or all of the reviewer's suggestions have been incorporated into the text, and these comments are marked with an indent below.

Specific Comments: For all the questions posed below, I would recommend the authors address an answer in the text of the paper itself, and not just answer in the reviewer responses, unless they have reason not to include the information in the text.

Introduction: Page 2 L11: The authors state that AQMS is expensive, but that is of course a relative term - can they estimate a rough cost? Is the instrument itself the source of expense (or for example, is it the cost of maintenance, data retrieval, site access, calibration)? And in general (this may come up elsewhere), I found that sometimes it is not clear in the paper whether the authors are referring to the ozone or co2 sensors. I would think that the AQMS only monitor ozone, so this should be specified

here. In general there is a feeling while reading this paper that the initial focus was on ozone and co2 was thrown in later, so a clean read-through to look for this might be good. We realized that the analysis and further synthesis of CO2 data was not as strong as the ozone analysis and discussion. This is in part a product of the poor ELT CO2 sensor characterization due to manufacturer algorithms. We have decided to remove the CO2 analysis and instead focus this paper on ozone variability. In regards to the monitoring station costs, entire monitoring stations can cost more than 100 thousand USD in addition to the expense to run and maintain them as well as their physical footprint, usually requiring property and indoor conditioning. These details were added for better reader context. More details surrounding the time and cost requirements of the sensors were added as well.

Page 3 L4: The minimum number of sites required by whom? Sites for ozone measurements presumably, not co2? I found this paragraph confusing - are 20 sites not enough to capture the variability in concentrations that is spatially heterogeneous below 10s of km? Are the 20 sites the same as the "current EPA monitoring networks"? The siting requirements are stipulated by the US EPA for ozone monitoring. Many studies before ours, as cited in the manuscript have found variability below the spacing of air quality monitors 10 km apart. Also, although there are 20 sites, they are not spaced evenly throughout the county.

version In the references for other low-cost sensor experiments, there have been studies using low-cost CO2 sensors: Shusterman et al. (ACP: https://www.atmoschem- phys.net/16/13449/2016/acp-16-13449-2016.pdf) and Martin et al. (AMT: https://www.atmos-meas-tech.net/10/2383/2017/amt-10-2383-2017.pdf). These should be mentioned. These are useful resources, thank you for them. Analysis of CO2 has been omitted to further investigate at a later date and these sources will certainly come in handy.

P3 L31: should be "high VOC concentrations". (and presumably NOx?). Yes, thanks, this is what was meant. Fixed to include “concentrations” and “NOx”.

P4 L13 "should be there is a large number of vehicles"... Suggested modification made in text.

P5 L8: Some indication of why medians were used rather than means? (to reduce influence of extreme outliers?). Median values were used to reduce the influence of outliers within the minute. This clarification has been added to the text.

P6 L7: semicolon I think should be a colon. This change was made.

P6 L10: It seems that uncertainties and precision should be given for the two reference sensors. For the CO2 standards, who certified them and what is their associated uncertainty? Is the Licor calibrated or drift-corrected in the field at all? What about the ozone sensor - how is it calibrated or drift-corrected? It would be important to assess whether either of these instruments is sensitive to ambient temperature, pressure, humidity, etc, same as the low-cost sensors. If these are not corrected or controlled for these variables, the authors should address whether this fact changes their interpretation of the various correlations and fits, and how. i.e. if the ozone reference concentration is dependent on temperature, would that have resulted in the interaction term that was observed? (I’m not sure, perhaps not.). The uncertainty of the reference ozone monitor is discussed later in the paper during the discussion and is framed within the uncertainty of the sensors. We have added more information regarding temperature and humidity compensation for the reference ozone monitors. Analysis of CO2 has been omitted to further investigate at a later date.

P6 L18 sentence structure awkward Thanks for catching this. Modified sentence for consistency

P6 Eq 2 p1 should have the 1 subscripted. Yes, fixed in text.

P6 bottom, P7 L1: Martin et al. did this for CO2 sensors, but only for a much shorter period of time (2 weeks?). Thank you for turning us to this. Analysis of CO2 has been omitted to further investigate at a later date.

P6 L30 perhaps to add clarification here, note that the comparisons were made between the concentrations from the low-cost sensors after being corrected by the equations 1 and 2 using the coefficients from the initial calibration test (constant coefficients p in time?) and the reference concentrations. To add clarification to this sentence, the first two sentences of this paragraph were combined. Now it is clear that raw sensor data was converted to meaningful gas concentrations using the calibration model coefficients and those values were compared to reference-grade concentration measurements over the ∼3-month validation period.

P7 L5: microenvironmental space? Yes, this refers to the small environmental characteristics of a location (e.g., temperature, humidity, pressure, gas species concentrations etc.) but we have removed this word for clarity.

P7 L12: this is the first reference of the 4T equation - this is equation 1? Yes, this is Eq. 1. Clarified in text.

Table 1: I wonder how different the coefficients Px are between different individual sensors? Can the authors give an idea of this? This is an interesting question and has prompted substantial modification to this manuscript. Although this wasn't initially within the focus of the paper, we thought it would be appropriate to add now that the CO2 analysis will not be included at this point. A figure showing distributions of modeled coefficient values and coefficients of variation for the sensors has been added along with a short discussion on similarity among sensors and potential reasons for these differences.

P7 last line: "during the deployment that were outside the range of those experienced during the calibration time period" might be clearer here. Yes, this would help. This has been clarified in the text

P8 L2: "As such" is not really clear as to which path you chose (assess or avoid). So you went with avoiding any extrapolation and filtering any data points with parameters outside the calibration range? We did not want to extrapolate in order to reduce error.

The second sentence here is correct. This was further explained in the text.

P8 L15-20. Not clear - wouldn't it be best to just eliminate O3 values that were higher than those experienced during the calibration period? How could measurements be over 7% of the highest maximum value, if it was the highest maximum value? Is it because you are looking at the highest value of the reference instrument? This seems odd to me all around. Why filter O3 and not Co2? We did not want to eliminate values that were higher than seen in the calibration because high ozone is of special importance to human health, and those values occur later in the summer. This is one of the reasons we filtered for environmental conditions, to reduce some of the error while still being able to record high values of ozone. In addition, the validation period acts as a way of knowing how well we estimate those higher ozone levels during the deployment period. We have changed the way that we filtered for maximum values, which has been explained in the text, and now makes the maximum 171 ppb. We feel this method is appropriate and did not significantly change results. A total of 110 data points were influenced by this change across all the U-Pods.

P9 L 13: So "calibration validation" refers to the deployment period from the previous section? Would be good to clarify that the validation period is the same as the deployment period, since both words are used here. This paragraph makes it sound like the best model was chosen based on how well each model did during the validation period, not using the same coefficients and model necessarily that were chosen during the calibration? The authors should clarify - I would have thought that the model was built using the calibration data set (including specific coefficients) and then applied during the deployment/validation, and then that corrected data would be compared with the reference sensor. Can the authors confirm that the coefficients from the calibration period were used in the validation period, and that the validation period is the same period as the deployment period? Thanks for pointing out this confusion. Figure 1 has been updated to make this clearer. Validation refers to U-Pods that were still located with a reference station during their deployment. The coefficients were arrived through

the collocation with the reference station during the calibration time period. We tried to fit several different models to the calibration period data (Table S1). Then we took those coefficients (from multiple models) and applied them to data during the validation time period, which is the beginning of the deployment. The model that performed the best on the validation data was applied to all the raw data for both calibration and deployment.

P9 L24: What does a precision check entail exactly? This should be stated in the text. The details of the precision check were added to the text.

P9 L25: wording: should say "5% from expected values (corresponding to a concentration of about 5 ppb), subsequent data would be flagged ...". Agreed. Change made.

P9 L26 awkward again: "Values within 5 ppb of the expected value would not be flagged". Changed this sentence to be more clear.

P9 L31: How large was the bias on this D45 UPod, and was this included in the statistics given above for ranges for the mean and median residuals? L34: This is confusing, as statistics were already given above. Is this 1-2 ppb bias based on mean or median residual? The mean bias residual for D5 was 5 - 6.4 ppb. This may have been confusing due to the placement of Table 2. We have moved it ahead of this paragraph and clarified the types of statistics. Because D5 was later omitted from the analysis due to the electrical modification, these statistics were not included in the overall uncertainty of the sensors. However, we made sure to show that bias to be transparent about potential issues of using these sensors.

P10 L1-2: Only one CO2 sensor was co-located with a reference for CO2 during the validation period? Maybe this can be re-stated here for those of us who got confused as to why only one sensor was used to assess this uncertainty. Analysis of CO2 has been omitted to further investigate at a later date.

P10 L1-2: were these higher concentrations and higher humidity values within the

range observed in the calibration period, or where they extrapolations of those fits? These were not extrapolations, as explained in comments above. Filtering for the environmental variable space had taken place.

Figure S6: These are plots made for the validation period, not the calibration period, so the fits shown as lines are not used to correct the data, just for informational visuals, is that right? is the red line in (a) the 1:1 line or the linear fit? (same comment for S5). Yes, the red lines are to show where the 1:1 is for viewing purposes. Changes were made to the captions of those figures to make this clear.

Table 2: for CO2, the RMSE is of the 1-minute data, while the mean residual over the whole period (how long was this period again and during what season?), is much lower at 3 ppm. What would the RMSE be for 1-hour means? Later in the paper hourly means or medians are used to look at differences/trends/etc., so this is the more relevant metric. If averaging the sensor data even further to 1-hour averages comparing the 1-hour medians reduces the RMSE that would be useful to know. Analysis of CO2 has been omitted to further investigate at a later date.

P11: Deployment: Is this the same as the validation period? Also in this first paragraph the collocation period is referred to - please confirm and state clearly that this is what was used earlier as the "calibration" period, i.e. the period when all the sensors were collocated and the coefficients and models derived. This has been addressed in previous comments. Validation just refers to the U-Pods that were deployed to places where there were reference monitors. We added a sentence at the beginning to clarify, as well as at the beginning of 3.3.

P11 L9-10 - I agree on the usefulness of comparing variability during calibration period vs. during deployment - except for the additional uncertainty caused by calibration drift over time, which cannot be assessed with the current data. This should be noted as a caveat - the true uncertainty during deployment might be larger because of the drift in the coefficients and model that is used. This is one of the key questions about use of

low-cost sensors in the field - how often do they need to be re-assessed or calibrated? As mentioned earlier this is an important topic for sensors. It should be noted that the calibration model we used incorporates time so re-calibrations were not done over the deployment period. Rather, the validation serves as a reference for how well the pods that weren't collocated with reference monitors during the deployment are performing. We added more analysis of model performance over time (via the results summarized in Figure 7) and discussed possible reasons for drift.

P11 lines 16-20 - please mention the time period, time of year of these measurements. Also this section is a bit repetitive with the next paragraph on P12, lines 6-10, which states the same information about how we would expect the diurnal cycle to look. Perhaps merge? We have included more details about the time periods that Gao used to show the diurnal cycle of ozone, and excluded a repetitive sentence from P12, lines 6-10.

P12 L13 and later in the text, when examining pair-wise RËĘ2 values, are the pairs of sensors that are in the same location excluded, so that we are only evaluating sensors that are in different locations? My understanding from earlier in the text was that there were 2 ozone sensors in each location even during the "deployment" period. [I am now re-reading the earlier text and realize that there were two ozone sensors in many of the U-Pods - in this case, which sensor's data is being used?]. But still, during the validation/deployment period, some sensors (D0 and D5) were at the same spot - are they shown in blue in Figure 5, rather than part of the red boxes? Because we excluded U-Pods D4, D5, D6, D8, DD, and DF, each of the remaining sites had only 1 U-Pod each. A sentence of this description was added to section 3.2. Blue boxes contain only data from the calibration period, when all the U-Pods were together. Red boxes have only data from the deployment, when no U-Pod is collocated with another. Also, we included a brief description of how we decided which sensor to use in the analysis in that same section.

P12 L14: "The larger the spread and magnitude of the RËĘ2 values, the more spatial

variability...". This seems backwards - the lower the RËĘ2, the more spatial variability there is. Reword? Correct, larger spread and smaller magnitude is indicative of more variability. This has been made clearer in the text.

Figure 5 caption ends with "U-Pod"? An unfortunate mistake, this has been removed.

P13 L6: "The U-Pods are more correlated" should be "the U-Pod O3 measurements, after the correction using the LT4 model, are more correlated ..." This change adds to the clarity; it has been incorporated. Discussion paper P13 L15: "between pairs of U-Pods". Changed.

Figure 6: X-axis should indicate (here and elsewhere) that this is local time. Added a sentence under figure 4 to clarify this.

P14 - the description jumps around in time a bit here (discussion of morning, then 15-17, then back to morning again...).This is a nice analysis and necessary to accompany the RËĘ2 analysis - two values might correlate within an hour, but that could be simply because O3 is increasing across the whole basin during that hour because of PBL changes, whereas the absolute differences indicate real spatial variability in the signal. This paragraph is organized by looking at the inverse relationship between Rˆ2 and differences, instead of by time of day. A sentence was added to try and guide the reader. More in-depth synthesis of the time-of-day effects are discussed later.

P15, Lines 14-15: Was the U-Pod O3 measurement at D7 calibrated against the reference sensor during the deployment phase? i.e. the models and coefficients were re-calculated for the second phase as well? This sensor could give an idea of my earlier question which would tell us how well the calibration does over time. I.e. you could correct the U-Pod data using the calibration from the calibration period, and apply it to the deployment period, and then look at the errors relative to the reference. Re-reading table 2, it seems this is exactly what Table 2 is showing. Was there a trend in the error between the reference and U-Pod measurements over time during the validation collocated phase? Again, it appears that this topic was not explained well enough,

so text was added to make this clear. The sensor calibration model for ozone (4T) incorporates time within the model itself so the intent of the validation using D7 was to show how well this model performed over a range of temperature, humidity, ozone concentration and time for the entirety of the deployment period. Model coefficients were not “regenerated” during the validation (deployment period for D7) as this would change the meaning of the validation statistics in the context of the pods that were not collocated with a reference monitor. Table 2 shows the distributions of the validation statistics for a random 10% of the validation minute-level data iteratively selected.

P15 L16: "as well as hourly trends by pod (Fig. 8)." Changed. figure 8 (& Fig 9) caption says "Each scatterplot is four hours of the day". But it’s not - each plot shows all hours of the day in black. We have re-worded this caption to be more clear about what data is represented in the plot.

P18: would a time series plot of these hours help interpret this weekend feature? A time series plot did not reveal any interesting features, but residuals helped to narrow down their time frame. Figure S4 was an investigation into what temperature and humidity corresponded to those data points.

P19: L16: "carbon dioxide distributions" should be "CO2 distribution". Analysis of CO2 has been omitted to further investigate at a later date.

Figure 10: Looking at this figure, knowing that D7 is in the same spot as Rubi, I wonder how they compare. (similar to my previous comment). I realize that the minute data showed a 15 ppm RMSE, which seems consistent with Figure 10 left panel. In Table 2, this 3.0 ppm median residual is for the validation period - but in Figure 10 it seems to be for the Calibration period. But on the right panel, it would be nice somehow to do an evaluation of the D7 sensor during the deployment, using the calibration from the calibration period. The bias here seems much larger, with the median of the D7 sensor significantly higher than the Rubi Licor. Perhaps this is one of the cases of the large baseline shift? (p19 L 18 says that D7 observed higher carbon dioxide -

but the Licor did not observe that same high level, so this is not an accurate way to characterize what appears to be sensor drift). P20, L1. The uncertainty of 15 ppm for CO2 was determined as an RMSE of 1-minute data, where the median difference was only 3 ppm. 15 ppm does not seem to be the correct uncertainty on the median of a distribution of hourly data over a several-monthlong (?) period. Analysis of CO2 has been omitted to further investigate at a later date.

Without more analysis or elimination of sensors that had large shifts in calibration, this claim does not seem to be supported - it is still not obvious that the sensors can determine spatial variability in CO2. Moreover, noting that some sensors are giving hourly mean values (albeit outliers, granted) that are close to 300 ppm (after the calibration correction!) makes me very doubtful as to their performance. Analysis of CO2 has been omitted to further investigate at a later date.

Back to Figure 1: the time line indicates "post-deployment" - is this referred to anywhere else in the text? Initially we had tried to collocate all the U-Pods together again at the end of the deployment. However, by this time the U-Pods had been untouched for some time and the data was degrading. We lost several Temperature and RH sensors during this time, and did not get enough data to include. This time period has been removed from figure 1, and is not used in analysis.

Figure 12. Looking again at the D7 CO2 data, this diurnal plot does not seem to match the plot in Figure 11 on the right side for this sensor, whose overall median reading was _460 ppm. Is there an error here? Analysis of CO2 has been omitted to further investigate at a later date.

P21 L10 - the time period and season should be mentioned here again, as obviously the time of year affects these diurnal cycles. Yes, thank you, we have added that information to the text.

P 21 L15. Again here the 15 ppm RMSE for CO2 is on 1-minute data, when the spatial variability etc is being evaluated using hourly medians for the most part. Analysis of

CO2 has been omitted to further investigate at a later date.

P22, L1-8 - these are very good points to make here, as I think these costs are often overlooked in the context of the low-cost sensors. Thank you.

P22 L15 should be "as high correlations with each other" Yes, the change was made.

P22 L15-17. I do not think this has been shown here. Analysis of CO2 has been omitted to further investigate at a later date.

P22 L19: How often is frequently? This remains to be seen in future work perhaps. Is there a way to determine this frequency from the data collected during the validation period? (i.e. is there a drift relative to the standard with time?). The authors did choose subsets of the validation data in order to do the evaluation in a more robust way, but an investigation here of the time-dependence of the errors would be useful. The question of how often sensors need to be recalibrated is indeed an important and widely discussed topic. Suggested durations between calibrations are highly dependent on the environment and gas species of interest. Due to the interest in this topic, an analysis and paragraph presenting the results of validation from the first week (directly after the calibration) and the last week independently. Residuals through time are discussed to inform the frequency of calibrations pointing to potential sources of this drift.

P23 L3: "undergone" should be "undergo". Fixed this, thanks.

Interactive comment on Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2017-183, 2017.