Mobile air quality monitoring and comparison to fixed monitoring sites for instrument performance assessment

Andrew R. Whitehill, Melissa Lunden, Brian LaFranchi, Surender Kaushik, Paul A. Solomon

Response to Reviewers:

Report #1

Reviewer comments for:

Mobile air quality monitoring and comparison to fixed monitoring sites for instrument performance assessment

This review is for the above revised manuscript submitted for publication in Atmospheric Measurement Techniques. The manuscript partially develops and proposes implementation of a new method to evaluate changes in instrument performance in mobile monitoring of air quality. To do that, the authors use high-temporal resolution (O ~ 1s) mobile-monitoring data collected using regulatory-graded instruments from two campaigns conducted in different regions for very different lengths of time for three pollutants, O3, NO2, and NO. The authors then compare stationary referencing of this data during collocations with the regulatory monitor to the referencing of “vehicle-in-motion” concentrations with regulatory monitoring data (based on distance and road type from the regulatory monitor) for one site, and find similar performance evaluations across pollutants for the residential roadtype in their new approach. For the second site, the authors do not conduct stationary referencing and only perform the latter “vehicle-in-motion” referencing to the regulatory monitor to estimate optimal temporal “running windows” to identify instrument issues. They calculate that for a 3 km spatial window, a temporal running window of 40 hours for data would allow detection of a systematic measurement drift or sudden instrument or sensor malfunction over the time scale of 7-9 days. In their revised manuscript, the authors identify and address systematic measurement drift or malfunction by briefly discussing the implementation of this method on another dataset. I recommend publication of this manuscript following the addressal of the following major and minor comments.

Major comments

1. Using r2 as a measure of random variability:

The authors use r2 as a reflection of “the random variability between the mobile and stationary measurements that results from a combination of measurement precision as well as true spatial variability”. However, I think what the authors wanted to instead say is that r2 is a measure of “the random variability between the mobile and stationary measurements that results from a combination of measurement precision as well as r2 also captures some true spatial variability”. The reason is that systematic (and not random) spatial variability occurs not just because of road type but several other factors such as wind direction and turbulence regimes which are affected by things such as emission sources and times of day. While I do see the value of r2 in the main manuscript, especially in the context of Figures 3 and 7, it’s hard to argue r2 even captures random variability when you are not even sampling the same parcel of air. I suggest that authors explicitly acknowledge the true spatial variability captured by r2 as a systematic and not random variability.

In lines 435-438, the authors say, “The smaller bias, in particular for O3 and OX, could be due to better inter-lab comparability in the California dataset, but aggregating data across multiple sites may also explain a reduction in the systematic bias. For example, if one site has a slightly positive bias and another site has a slightly negative bias (due to monitor siting or random calibration variability), those biases will partially cancel each other out.” This sentence made me rethink how clear is r2 a measure of random bias. I am not convinced that r2 in the way it is used (comparison of different air samples at the same time), is a reasonable measure of random spatial variability. I suggest the authors instead use a cleaner approach to separate systematic and random variability such as the comparison of actual bias and absolute bias. You could add those comparisons either by adjusting the current panel plots or as supplementary figures, and briefly discussing them (1-2 sentences) wherever using r2 as a measure of random variability is a deficient way of going about it.

We added text to Section 4.2 (Lines 285 – 295) to clarify what we mean by random spatial variability (which we changed to “spatio-temporal variability”) versus persistent spatial biases over time, which we refer to as systematic spatial bias.

2. Using non-highway as a class versus residential for “mobile to stationary” referencing

If you look carefully at Figures 3 and 7, it is clear that residential roads are showing stable behavior regardless of distance in terms of mean and median bias. This suggests that they are able to capture a systematic instrument bias that perhaps other road types cannot. The authors identify this in lines 311-312 as “For the Residential road class, the bias between mobile collocation and parked collocation changes very little as the distance buffer increases for all species.” Additionally, in Lines 462-465, I appreciate the authors’ effort to highlight the value of data on residential roads. It is then surprising that the authors want to use road type data other than residential to determine detection thresholds of systematic instrument bias/drift/malfunction. I suggest that authors not club residential and other road types, or at the least show in the supplement that just using residential data does not dramatically lower the detection thresholds of instrument malfunction. Otherwise, that residential roads are a close proxy of stationary collocation is a major finding, is easy to understand, and all figures and discussion should be orientated around that aspect (e.g. Figure 4). This also makes sense in other ways, since health exposure studies naturally sample large sections of residential areas. This will also address another issue I had with the manuscript which was the lack of results associated with Scenario 1 identified in Section 6. I suggest showing Scenario 1 in the Supplement similar to the analysis showing in Section 6.1, and at least briefly discussing it at the relevant places.

Additional analysis on Scenario 1 is outside of the scope of the paper and would require more than just minor revisions to the manuscript. The comparison of two scenarios in Section 6 were intended primarily to introduce the linkage between number of collocation data points and uncertainty in ΔX, contrasting the 500m Residential only scenario with the 3000m Non-Highway scenario. However, the subsequent analysis is focused exclusively on the 3000m Non-Highway data set because it requires a sufficiently large data set to do the sub-sampling. The Non-Highway data set also is expected to result in a more conservative estimate of the uncertainty in ΔX than the Residential only scenario, which is likely more typical across most complex urban environments. We have updated the text in the introduction to Section 6 to reduce the emphasis on introducing two distinct scenarios (1 and 2), which inadvertently raised an expectation of a contrasting analysis of the two for the reader. Instead, we focus the discussion on the differences in the size of the data set between the two examples, and later provide justification for choosing the 3000m Non-Highway scenario for the subsequent analysis.

Minor comments

1. In addition, the authors say in the responses that “We revised and improved all the maps in the manuscript, including adding shading to indicate parking areas and adding wind roses and scale bars.” However, just looking at Figure 1, while I do see parking areas, I neither see wind roses nor scale bars. The authors need to address this issue before publication. Also, please check that you have actually incorporated all aspects that you claim in responses before submitting the final revisions.

We remade the maps so that the wind roses and scale bars are clearer (Figures 1, 5, and 6).

2. Figure 3: add number of points in the parked colocations aspect as well. There are no black dots in the bottom subfigures.

Updated figure as suggested (Figure 3).

3. Figures 5 and 6 do not seem to be particularly useful for the main manuscript. Move them to the supplement.

Figure 5 and 6 serve to demonstrate the range of roads driven and the type of driving patterns, which is important to the interpretation of the resulting data. Therefore, we have chosen to keep these images in the main text.

4. Figure 8 axis labels should clearly state the use of “running medians” in the y-axis.

Updated figure as suggested (Figure 8).

5. Lines 412-415 “This is in contrast to Sections 3 and 4 where the mean was used to aggregate the one second data up to one minute or one hour. In general, using the median versus the mean produce similar results for O3, NO2, and OX; however, using the hourly medians versus means significantly reduces the impact of high NO outliers (peaks) on the NO aggregation.” Please add supplementary figures showing the difference of mean versus median based figures for Sections 3 and 4. Otherwise, I recommend using consistent underlying central tendency metric across sections as it unnecessary complicates this manuscript for an average reader.

Either measure of central tendency (mean versus median) work for the hourly data aggregation. We wanted to demonstrate examples of using both the mean and the median and believe that either can be used in the resulting analysis with only minimal differences on the results.

Report #2

In their manuscript “Mobile air quality monitoring and comparison to fixed monitoring sites for instrument performance assessment”, Whitehill and coauthors present analyses from stationary and mobile comparison measurements of ozone, NO, NO2, and Ox (O3 + NO2), measured on-board mobile platforms and in air quality network stations to be used to identify instrument bias during ongoing field measurements with the mobile measurement platforms. This is a strongly revised version of a previous manuscript, in which the data analysis was taken a lot further, compared to the initial submission.

While in the initial version of the manuscript, the authors have mainly based their analysis on linear regression analysis and have presented multiple correlation plots, they picked up my suggestion to show the quality of agreement between air quality network results and their mobile platform measurement results as a function of distance between both measurements for different road types separately. They have taken this further and focus now on average or median bias instead of slope and intercept. In addition, the authors have added a discussion on how well their approach of mobile comparisons could be used for other types of pollutants, depending on their nature and homogeneity of their concentrations in the environment.

The revised version of the manuscript goes well beyond the first version in analysis depth and coherence. In addition, the authors have addressed all relevant points raised in my first review to an acceptable degree. There are several minor points to be worked on, as detailed below. After these points are addressed, I suggest publishing the manuscript in AMT.

Detailed comments:

L24: “… highways showing the most variance.” – Shouldn’t it be “… highways showing the strongest biases.”?

Changed from “most variance” to “largest differences” (Line 24)

L53: “In addition, natural variability …” □ “In addition, natural spatial variability …”

Changed to “natural spatial variability” (Line 55)

L82: “… commonly measured and air quality …” □ “… commonly measured in air quality …”

Changed to “commonly measured at air quality” (Line 84)

L121: “… next-generation air quality instruments”: This sounds like very sophisticated air quality instruments. Don’t you just mean “low-cost sensors”?

Changed to “low-cost sensors” (Line 123)

L141: The drivers were instructed to park facing into the wind when possible. Was assessed, whether the measurements were affected by the own exhaust under still conditions or when the wind was from the back (which both could also occur during the mobile measurements, e.g., when stopping the vehicle). Were such self-sampling intervals removed from the data sets?

Added: “A visual screening did not reveal any suspected self-sampling periods, so no additional attempts were made to identify or remove such periods.” (Line 144 – 145)

L151, Figure 1: From the satellite view and the scale at the bottom of the image, the maximum distance for the CAMP site seems to be rather 35 m and not 85 m.

Added text to caption:

“(Not shown are two periods where the cars parked at the CAMP site just north of the map due to a lack of street parking closer to the site).” (Line 155 – 156)

L162: The QA evaluations showed instrument bias of 3%-6% for O3 and 3%-8% for NO. How do these percentage biases translate into ppbv biases? If they are calculated from the span gas concentrations, 80 and 360 ppbv, this would mean an observed bias of up to 5 ppbv for O3 and up to 29 ppbv for NO. This is larger than the minimum necessary bias for observation as stated in the abstract and in the results section (4 and 8 ppbv). How can this be?

Clarified by adding: “Field bias measurements are based on span checks assuming the linearity of the instrumentation response across the measurement range, which was confirmed in laboratory multi-point calibrations. At the mean observed concentrations during the study (33 ppbv O3 and 53 ppbv NO), the error in span measurements translates to an average bias and precision of 2 ppbv for O3 and an average precision and bias less than 4 ppbv for NO.” (Lines 166 – 171)

L206f: The mean or median DeltaX values are claimed to be a more direct assessment of systematic bias than slope and intercept of a linear regression. This is true for an offset in the instrument data. However, a bias due to a change in instrument sensitivity would rather be detected by linear regression than using mean or median DeltaX values.

Clarified that the bias discussed is “offset bias”. Also clarified that “This is due to the high sensitivity of OLS regression statistics on outlier points.” And “a few extreme outlier points might occur that will skew OLS statistics but have less influence on ΔX statistics”. (Lines 215 – 221)

L230: “… differences in concentrations due to the distance between the locations …” - The differences in concentrations are less due to the distance between the locations but rather due to the differences in distance to the sources of emission plumes. Otherwise, it would only result in more scatter, but not in a systematic bias.

Clarified that the difference are due to the “relative proximity of the two instruments to passing emission plumes” (Line 239)

L239: Why does the influence of local traffic emissions reduce the applicability of a linear regression approach but not of a measurement bias approach?

Clarified that “Temporal aggregation can smooth some of the outlier points and make the results more reflective of real measurement differences; however, parametric regression statistics will still be biased by the influence of outlier points.” (Lines 247 – 249)

L244-246: The authors state that mobile collocations allow to sample a larger amount of air in the same sampling duration. Mobile measurements sample from the amount of air that passes by the moving vehicle (i.e., depends on difference of velocities of ambient air and vehicle). Stationary measurements sample from the amount of air that passes by the station due to the ambient air movement. On average (i.e., when moving with and against the ambient wind the same amount of time), both are similar in size. In the case that the vehicle moves with the air (in wind direction), the mobile measurement would even probe less volume of air, compared to the stationary measurement.

Added explanation:

“For example, a car traveling 25 meters per second will “sweep” an additional area of 1500 linear meters in one minute compared to the stationary sampling. Thus, regardless of the wind speed, a moving platform will integrate each measurement over a larger area than a stationary platform, making each emission point source have less direct influence on the entire integrated measurement.” (Lines 256 – 259)

L257-258: “While travelling on high traffic roads, the cars are more likely to be impacted by direct emission plumes.” - This is likely true. Nevertheless, I think that the critical point is the traffic density in the vicinity of the mobile measurement vehicle, not only the road classification. There can also be several cars in front of the mobile measurement on residential roads, e.g., at intersections or traffic lights. This would also strongly affect the measurements of NO and O3. While the road type classification is an easy to perform start, a more sophisticated approach would be desirable for the future.

We agree, but this is beyond the scope of the present study. Added the following text;

“Although more sophisticated methods are possible to identify and remove high traffic roads, OSM road classifications are a general proxy that can be applied algorithmically over a large portion of the Earth. In contrast, local traffic count data are more sporadic and not always available or easily accessible for the region of interest.” (Lines 274 – 276)

L271: “… bias values can reflect …” □ “… bias values from in-motion collocations can reflect …”

Clarified that “bias values from both stationary and in-motion collocations can reflect…” (Line 286)

L272: I would remove the “random” in “… generally reflect the random variability …” because non-random variability, e.g., due to persistent spatial differences, would also be reflected in r2.

We clarified Section 4.2 to better define what we mean by “random” variability (Section 4.2)

L288: “the relationship” seems to be not the right expression since there is no interaction between the stationary and the mobile collocation measurements. Better something like “in comparison to”.

Replace “the relationship” with “the similarity”

L292: The larger magnitude of bias for the Highway road class, compared to the other road classes, does not indicate larger spatial variability (this would be shown by larger r2 values) but larger (or smaller in case of O3) average or median concentrations.

Changed to “… indicating a larger influence of direct emission plumes increasing the variability in concentrations measured on Highways…” (Line 315)

L305, Figure 3: Please use consistently in the text and in the Figures either r2 or R2.

Corrected to use r^2 consistently throughout the manuscript.

L306: What does "maximum distance" mean in this context: are the median and mean bias values taken from all measurements up to the respective distance buffer shown, i.e., for larger distance buffers also the measurements from closer to the fixed site are included; or are only the increments in distance buffer used as basis for the respective data? The former provides only very indirect information on how the measured bias depends on distance to the fixed site. It should be more clearly stated, which data are included in the individual data points.

Added clarification (L308-311):

“Note here that for each distance D we include all datapoints within a distance of D from the stationary site, so we are not explicitly showing how bias varies with distance from the stationary monitors.”

This approach is not very helpful in the analysis of how strongly spatial variability affects the comparison measurement, because it is strongly affected by the number of data points in the individual distance buffer increments. E.g., for residential roads there seem to be barely any data points available beyond a distance of 2 km. This results in barely any change in bias values beyond this distance, which is not a result of spatial homogeneity (as it seems) but a result of a lack in data points in this distance range. Normalizing the contribution from individual distance buffer increments by the number of data points within the respective increments would provide a more direct information on the influence of distance on the comparability of the measurements and would therefore allow to apply the results also to other environments, where the distribution of various road types might be different.

Because our method specifically looks at all datapoints within a set distance (maximum distance) from the site while averaging we felt it was more appropriate to show the data as we have in the manuscript.

L326-331: In order to judge whether the observed bias values are significant and whether they would allow a reasonable identification of measurement bias, it would be interesting to also see the average absolute concentration values and their variability during the measurements. This is shown in Figure S4 and S5, but this information would be essential in the main text as well.

This information is shown in Figures S6 – S9 in the supplement and referenced in the main text.

Furthermore, according to the time series, e.g., for NO it looks like that rather than average or median values the minimum or better something like the 5% percentile would represent the measurements not affected by local plumes (i.e., those of the stationary sites) much better.

It is possible that a 5th percentile might work well for some species, such as NO, but is unlikely to work for others, such as O_3. Exploration of alternative metrics is beyond the scope of this manuscript.

L426: "… random spatial biases." □ "… random spatial variability."

Replace "random spatial biases" with "random spatio-temporal variability"

L433: I agree that the higher r2 for NO2 could be due to the larger dataset used, however, it also may be due to the fact that only 1-hour averages were used there, where short concentration peaks are largely averaged out.

The comparisons here are made to the 1-hour averaged data in Figure 4, which uses the same 1-hour averaging period. Clarified that the comparisons are made to the "equivalent hour-averaged results for the Denver study". (Line 458)

L439f: This statement again shows that without taking the number of data points per road type or distance increment into account, the results reflect to a certain degree the distribution of road types and not necessarily the actual spatial variability of pollutant concentrations.

Added the following sentence:

"Variances in traffic patterns, road type distributions, and other factors could also influence differences in biases in different geographic regions or using different driving patterns, so the range of biases must be measured for each individual study region and study design." (Lines 464 – 467)

L458: Indeed, it seems advantageous to remove Highway road segments from the data. However, it also would probably be advantageous to remove all data points which are from short-time peaks (i.e., from plumes of nearby sources) – also from the data of the other road types.

Added the following clarification:

“Although more complex peak-removal algorithms can achieve similar goals, they add complexity without necessarily improving the data and add additional arbitrary bias (e.g., “cherry-picking”) to the resulting comparisons.” (Lines 489 – 491)

L491-496: It is not really clear to me how the upper and lower traces were calculated. Are these the lowest and highest 1-hour medians within the running median? Or is this the minimum and maximum of the running medians (over different window sizes) of the 1-hour medians? It would be desirable, if this is explained a bit more clearly.

Added the following sentence to clarify:

“For each running median window size N, the upper and lower traces reflect the maximum and minimum of the set of running N-hour medians from this dataset.” (Lines 525 – 526)

L519-525: So, this means that under typical conditions, the response time would be likely several weeks, correct? In this case, wouldn’t it be easier to have a quick calibration check every couple of weeks where a calibration gas mixture is probed for a couple of minutes by the instrument setup, compared to the ongoing analysis of mobile collocations with their higher bias uncertainty

Added clarification to introduction:

“While the use of fleets facilitates scaling of mobile monitoring to large geographic scales, such as multiple counties in an urban area or multiple cities across a large state, coordinating vehicles and drivers to across these geographies makes route laboratory-based calibrations costly, time consuming, and impractical.” (Lines 38 – 40)

L567: Why are here averages of data instead of medians (as in the rest of the manuscript) used?

We recognize that using means instead of medians could create confusion. However, we wanted to demonstrate that running medians of hourly means versus hourly medians both work and produce the desired outcome.

L577; Figure 10: Why does the 40-hour running median start (and end) at the same time as the data points start (and end)? Shouldn't there be a lag in the start with the running median starting after 40 hours only (as in the previous figure)?

Figure was updated based on this suggestion. (Figure 10)

L589: How does the analysis has implications for the spatial heterogeneity? I guess the latter one is unaffected by the analysis.

Replaced “has implications for” with “provides information about”

L608: Removing highway-related data does not reduce the spatial heterogeneity of pollutant concentrations, it reduces the influence of local emission plumes onto the measurement data.

Replaced “spatial heterogeneity” with “influence of local emission plumes” (Line 643)

L633: Not "the influence of highways" needs to be removed, but the influence of emission plumes on the measurements – which are more frequent on highways, compared to other road types.

Clarified that the purpose is to “reduce the influence of emission plumes, which are most abundant on highways” (Lines 668 – 669)