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.

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.

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.

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

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

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

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.