After major revisions I am still not convinced that this paper and the supporting research meet the standards of AMT. I think there are a number of major flaws the authors must address in order to make their paper suitable for publication.

In this revision of the paper I feel the authors have revealed a potential fatal flaw of their study:

Page 23 line 12-15: This is very surprising that volunteers just looked at their sensors 3 times in an hour and reported the value to the app. So the concentrations are not hourly averages from each monitor. This makes me question the usefulness of this study. How often does the screen update with the concentration? Is it a 1-minute average? Rolling average of the past hour? Something else? If you pulled 3 random minutes of data from an hour you would expect they might be significantly different than the average for the full hour.

Were all the performance metrics calculated in the same way just reading 3 random minutes off the screen? If not the performance metrics (Figure 1 and +/- 20% reported by Tsinghua) are not really good indicators for the expected accuracy of this work. I think the implications of this sampling methodology on the results need to be discussed.

See Zheng et al. 2018 for discussion of averaging time and low-cost sensor performance: https://www.atmos-meas-tech.net/11/4823/2018/

My other major concern is the overestimate by the sensors during period 1, and the over estimate of the sensors during period 2. During one of the evaluations periods the RH was 94% at maximum while during one of their sampling campaigns the RH was reported from 95% to 98%. Much previous work has shown that nephelometer type devices greatly over estimate during high RH events and in a nonlinear way (including: https://www.atmos-meas-tech.net/11/4823/2018/). There is no evaluation data at the RH that occurred during this project and they present no figures showing the relationship between RH and sensor error and no past work showing how this sensor performs at high RH. Maybe having accurate sensor readings is not important for the main goal of this paper (generating different models) but they have tried to draw a number of large conclusions based on the higher average from the sensor data (people should stay indoors even when government monitors say it is clean out, pollution at road level is higher). I have a similar concern during the second period where the monitors underestimate since they have presented no evaluation data and included no references that show that the sensor provides a linear response above 160 ug/m3. Low-cost sensors may show non-linear responses (http://www.aaqr.org/article/detail/AAQR-17-10-OA-0418) and it is very important to evaluate them under the conditions they will be evaluated under. Any supporting literature the authors can find about the monitor they used or the internal sensing component may support their findings but currently the authors have not cited any previous work with their monitor.

Additional comments:

Figure 1. It's not clear to me what is on the right scatter plot versus the left? Does each plot show the data from all three stations? How was the data recorded for these tests? Since for the experiment it was read off the screen is this also what was done for these? What is the averaging

time for the points shown on this plot. It would be helpful to label them as c and d so you can reference them in the text.

Page 22 line 4: please define small difference.

Abstract line 20: Not sure what all the numbers are in the parenthesis

Abstract Line 16: suggest mentioning how many hours each period was.

Abstract line 22: I don't think you can make any generalizations based on this dataset suggest: "During this project", OK interpolation performs… or something similar

Page 20 line 24: What do you mean by "fine" exposure control?

Page 21: Line 9 and 10: suggest including a citation about underestimation/mis calculation of risk

Page 21: Lines 22-27: These results from Tsinghua are super important since you are reporting data up to 260 ug/m3 but you only have a comparison range up to 160 ug/m3. Please include a citation or a figure in the supplement. Is this an ambient comparison or in the lab?

Page 21 Lines 29-30: Is this conversion factor calculated by you or by the company?

Page 21 line 25: At what averaging time is this +/- 20%

Page 27 line 32: Suggest citing previous literature that has seen significant gradients from 2 to 15 m.

Page 25 Line 15: I don't think it's helpful/scientific to state "Generally, the statistics differed". I think this is not needed

Page 27 Line 20-22: To explore the spatial variation in the… This sentence is unclear I don't understand any point you are trying to make

Page 28 line 14: You can't say this "guaranteed" their reliability

Page 28: Lines 4-7: I think this is too large of a finding based on the results you have presented here