Comments on AMT-2017-30 manuscript

The work by Hundt et al. "Spectroscopic real-time monitoring of NO2 for city scale modelling", describe a mobile NO2 QCL based spectrometer which has been installed on a tram roof and used for long-term monitoring in the city of Zurich. The data have been compared with data from AQM stations in the city, and a statistical analysis have been performed for concentration maps.

The science in this manuscript has great potential, however the general presentation of the data, the description of the statistical models and the description results has to be improved. I will also suggest authors to look for a better and clearer structure of the paper.

I believe that this work required major revision before being accepted to AMT.

General comments:

Further effort should be made on the quality of the writing, the English and the way of presenting data. The paper structure should be reviewed: The main structure: Introduction, methods, results and discussion and conclusions is ok, but the sub-structure should be reviewed. The section 2.4 should be partially included in section 2.3 and 2.5. Section 3.1.should not be split into two sections, same for section 3.3. Finally in figure 10 authors could add the name of the stations so that readers don't need to go to the supplementary material to see station's positions. Finally, authors are making data analysis with 3 seconds averaged data while the instrument response time is larger (9.6 seconds reported in page 3 or 11 seconds in page 9). The whole data analysis should be redone accordingly.

Specific comments:

Page 2, Line 8 : « With CLDs, NO2 is determined indirectly as the difference of consecutively measured NOx and NO, making this technique not fast enough for mobile measurements ». To my knowledge CLDs can switch between NO and NOx channel within only few seconds, which I believe is still considered high resolution (<1 hour as mentioned at page 1 line 28). What it can happens is that in case of rapid NOx transitions the value of NO2 can be biased, leading to overestimation or underestimation of NO2 content. Furthermore, since is an indirect approach, it can suffer from artefacts (for instance during the conversion NO2 -> NO in the catalytic oven, other species may also be converted, leading to an overestimation of NO2). I think authors should reformulate this part more in those lines, and please add references.

Page 3, Line 4: « Light passes a 12m … the infrared detector ». Move this below together with the cell description. Could authors comment on the possibility to reduce the cell volume? If my calculation are correct, the height of the cell is 1.8 cm, while usually the beam waist for such a type of propagation is 4-5 mm. It could be possible to reduce the height in order to gain a factor of 2 on the cell volume or there are other factors that will make this volume ideal?

Page 3, Line 17: « We determined the flushing time for a complete gas exchange in the cell to be 9.6 s by exponentially fitting the NO2 signal after switching from outside air to filtered NO2 free air». Is that what expected according to the total volume of the instrument and the gas flow at 100hPa of pressure? Do you see any memory effects that slow down the gas exchange process? Do you observe the same response time while going from outside air to filtered air and on the other-way around?

Page 3, Line 29: « the instrument repeatedly determines the zero-point offset measuring filtered NO 2 -free air ». Mention already here how often this zero-point is determined.

Figure 1a: It could be more interesting to have a picture of the instrument without the cover to be able to see the inside; Figure 1b: Further efforts should be made to provide a clearer and more informative schematic of the setup.

Page 4, Line 13: « The gap between 27 Dec 2015 and 05 Feb 2016 is due to discarded data of reduced quality related to misalignment in the optical system of the instrument. ». It could be interesting to know how many time the instrument required the intervention of the operator over this period of time to show the robustness of the instrument. Any explication on why it took so long (more than 1 month) to realign the optical system?

Page 4, Line 15: « Moreover, data was omitted when the NO2 spectrum was not clearly identified by the processing algorithm ». Authors should specify the percentage of "bad" data that were omitted. Is that related to S/N ratio? Did authors used a threshold on the standard deviation of the fit? More explanations are needed for this point. Replace "data was omitted" with "data were omitted".

Figure 2: « high quality NO2 data ». Authors should be more clear in the manuscript about what good quality data means. Which rules have been used to reject data?

Section 2.3: From making the reading more clear, please specify the different stations (ZUE, SCH, HUE etc) and mention which one are in a background zone and which one in a polluted area.

Page 5, Line 18: « Largest emission sources contributing to ambient NO2 concentrations in Zurich are motorized traffic (47% of NOx) and heating systems (28% of NOx) ». What is the remaining 25%?

Page 6, Line 8: "We used four types:" of what? The object is missing.

Page 8, Line 11 : « Such a correction to the normally calibration free QCLAS method ». Authors should also discuss and if relevant quantifying the effect of missing UV radiation in the inlet of the instrument that would perturb the photo stationary state ratio between NO and NO2. They mention a flushing time of 11sec that is long enough for accounting for an overestimation of the NO2 (due to the reaction of NO with O3 in absence of UV light). A quantification of this artefact should be reported.

Page 8, Line 17: "Discrepancies between the two instruments are mostly due to the higher temporal resolution of the QCLAS and imperfect time synchronization." This should be avoided if QCLAS data are averaged according to the acquisition rate of the CLD (which seems to be the case as mentioned in the caption of figure 4 (1 minute average for both dataset). Were time stamps of the two instruments synchronized before starting the comparison? This statement should not be necessaire if data comparison is done properly.

Page 9, Line 1: "The standard deviation of the QCLAS measurements from the CLD was found to be 0.96 ppb therefore we conclude that the instrument accuracy is about 1 ppb and that the assumption of a constant relative loss of NO2 was correct." Authors should mention the precision of the CLD measurements. This sentence is not clear: is the 0.96 ppb the standard deviation of QCLAS data with respect to CLD data as plotted in figure 4. If so, the sentence "The standard deviation of the QCLAS measurements from the CLD" should be written more as: "The standard deviation of the QCLAS measurements with respect to CLD data". Further discussions about the sensibility of the instrument should arise here. According to the Allan-Werle analysis (Fig.3) the precision of the instrument should be0.02 ppb. Since that is not the case, this means that the AW-deviation should

start to go up in a longer term. I believe that authors should provide an AWW-deviation plot on a longer term to report this long term drifts that will justify this final precision of 1 ppb.

Page 9, Line 10: "We discarded all the positions with corrections exceeding 10 m". Authors should mention the percentage of rejected date due to inconsistent position. Was 10 m decided because of the optimal distance for high resolution data? Was this threshold fixed also according to the GPS position? Discussion on this choice will be appreciated.

Page 9, Line 13: "we added 11.1 seconds to the timestamps of the matched positions in order to account for the flushing time of the measuring cell and for half of the length of the measurement interval." What it should be taken into account is renewing of the sample in the cell, which corresponds to the residence time of the gas in the inlet plus the residence time in the cell. It is not clear here if the 11.1 sec correspond to this. Why half of the length of the measurement interval is also taken into account? At page 3 the response time was estimated to 9.6 sec. Please change the one that is not right.

Page 9, Line 15: "The NO2 measurements refer to 3 seconds averaging time. Therefore, they are strictly speaking not point measurements but refer to a route segment." Not well written. Please rewrite it with a better English. And as mentioned above 3 seconds average is not ideal because the instrument response time is larger.

Page 9, Line 17: "The length of this segment is the product of the integration time of 3 seconds and the speed of the tram. Tram speed associated with the mobile NO2 measurements was smaller than 5 m/s for 64 % and smaller than 10 m/s for 92 % of the time." What is important here is also the average speed of the tram. This will also question the 10 m threshold that authors selected for discharging data because of bad positioning. If the measurement is done in 3 sec and the maximum tram speed is above 10 m/s then the a threshold of 10 m on the position data is not justified and it should be higher in my opinion (It should corresponds to the maximum speed of the tram times the measurement time. eg. 10m/s x 3 s = 30 m threshold).

Page 10, Line 3: "We analyzed the quality of the data obtained in the mobile application twofold: First, we quantified the measurement uncertainty related to noise and variations of the zero point offset. Second, we compared the tram measurements to measurements from a fixed air quality monitoring site (see next section)." What is the mobile application twofold? The whole sentence is not clear and should be rewritten.

Page 10, Line 14: "We corrected the NO2 measurements by linearly interpolating the zero point offsets derived in two consecutive zeroing periods… … … Zero point offsets are temporally correlated as the standard deviation of the differences of two consecutive zero point offsets amounts to 2.1 ppb and thus is smaller than 4.3 ppb expected for white noise." I think that this part should be better written. I understand the idea of "predicting" the zero point offset using the two neighbour ones. What is not clear is: is this analysis used to remove "bad" zeroing point? If so, it should be expressed more clearly. Where the 2.1 ppb comes from? From my understanding zero point offsets will be temporally correlated if the difference between the predicted and the measured zero point value (which I will not call standard deviation) will fit within the 4.3 ppb expected in case of white noise. The quantiles are highly reported in this work but not enough used in the discussion.

Page 11, Line 8: Why authors compare measurement from the tram and fix site with different time averages? For fig 6 is clear that measurements with 3 sec integration can show much high variability (due probably to local production). To be more consistent, the comparison should be for the same

time window and same averaging time. Authors should discuss about the higher variability observed at shorter time averaging (3 sec).

Page 11, Line 14: The statement "The calibration of sensors using precise measurements from AQM sites is an option for mobile operated sensors" requires better explanation.

Table 1 Caption: replace number of values with number of measurements. Here too, the quantile values are not well used in the discussion. Authors should either used the more critically or remove them from the table.

Page 11, Line 21: Suggestion to replace "aggregated by 50 meters of track length" with "with 50 meters resolution". Same in the Caption of Figure 8. As mentioned in the caption, the fact that in the figure are the results of the full campaign (and not only 1-way from Bellevue and Bahnofquai should be also mention in the text. Authors should give an idea if how many 1-way trips are used for this plot and how much this corresponds in term of time.

Figure 8. "The tram is directly impacted by traffic between the first and the third stop as well as between the fifth and the seventh stop. There is only little traffic between the third and the fifth stop." Not important in the caption, but this discussion should come out in the text, mentioning that there is then a correlation between the zones which are highly impacted by traffic and the NO2 level measured. What are the "extra" dots in the graph? Out-layer data? This should be specify.

Page 11, Line 21: what authors mean for similar locations? This is something that it should be explained more clearly. What are the criteria for making two locations similar?

Figure 9. There is a miss of general explanation about this analysis. I have the feeling that all for plots at figure 9 don't bring a lot of information from the statistical point of view. I think authors should do a stronger effort in reporting data which are important for the understanding of the work. In the text this figure is not well descripted, ant the readers don't really get the point of what is plotted.
Few more details: The really last red dot at the top right is suspicious… For this measurement period there should be a black dot (MAXTi) on top of the Q095Ti dot.
Modelling period is a bit confusing because to me this is a simple data analysis, and there is not modelling behind. I suggest to change it with measurement period.
The use of "Frequency" on the y-axes to me is also ambiguous, I would suggest to put "number of occurrences". I found weird that in figure c and d the y-axes does not extend to include the whole data set.

Page 16, Line 9: "The frequency a particular variable is effectively used for splitting the 30 minutes data set depends only on the measurements". I guess authors wanted to say that "The frequency **of** a particular variable is effectively used for splitting the 30 minutes data set depend**ing** only on the measurements". Anyway this sentence should be rewritten more clearly.

Page 16, Line 11: The whole paragraph: "The number and spatial distribution of measurements in a modelling period (approximately 500)… …Regression trees based on tram measurements have about 6 location types Ri on average." should entirely be rewritten. I don't see how the number of measurements in the modelling period will limit the possibility of separate different location types?

Section 3.3.1: I am sceptic on meaning of the model version Vt, especially on a such a short time scale. I guess that this approach could be les redundant by increasing the time intervals. For instance using the 30 min intervals to predict other 30 min intervals. Add "respectively" at the end of sentence (line 23).

Page 16, Line 28: "The difference in performance between VAQ and VSTA is small due to the similarity of NO2 concentration encountered at site STA and by the tram-based instrument." Why? VAQ and VSTA are both based on fixed site. How come their similarity should be due to the similarity between NO2 measurement at STA and on the tram?

Table2: How the variables to be used on each model have been chosen? Why there are not variables for VSTA? Same question for section 3.3.2. From what authors describe it seems that variables are chosen to achieve the best overall fit to the measurements at AQM sites. But those variables have a specific meaning and they should not be selected to making the measurement better matching but more according to their physical meaning.

Figure 11 and its discussion: In my opinion, Figure 11 shows that model prediction is less effective at stations far away from the tram track (BLU and HEU). As expected, predictions that use data from the stations (WAQ and WT+A) provide better RMSE but still slope values are more far from unit. I don't think difference between models can be really discussed here. What about error bars in figure 11?

Section 3.4: "Our results suggest that the number of required mobile or static QCLAS instruments for comprehensive NO2 measurement in Zurich is not smaller than the current number of fixed AQM sites." How authors get to this conclusion?

Section 3.4: What about taxis? I guess those could provide a more random and spread-out data set with respect to public transportation which will follow a fix path.

Page 19, Line 19: "The model type the maps are based on is irrelevant in this context." Meaningless.

Page 19, Line 23: "within the same footprint." What does it mean? Same source?

Technical comments:

Page 2, Line 10: « Similarly, electrochemical sensors are currently not suited for such applications as their response time is not short enough ». Add a reference

Page 2, Line 25: « a cylindrical multipass cell ». Put the reference Mangold 2016 also here.

Page 2, Line 27: replace 1.1 with 2.1

Page 2, Line 30: remove Additionally

Page 3, Line 14: change "it is not sensitive" with "it is less sensitive".

Page 3, Line 17: « while the air flux is limited by a 100 µm diameter orifice ». Specify to which flow.

Page 3, Line 22: « nano PC ». Specify which one.

Page 3, Line 34: Say what GSM stays for.

Page 4, Line 8: « altering intervals ». I suggest to replace altering intervals with varying intervals.

Page 6, Line 5: replace "in the environment of this location" with "in a specific location".

Page 6, Line 17: replace « for every three seconds » with "every 3 seconds ".

Page 6, Line 21: « First, we focused on small-scale features of the pollutant field. This… …yields a benchmark for the performance of the mobile sensor network » This part should be rewritten more clearly.

Page 7, Line 21: Replace "und" with "and".

Page 7, Line 23: Replace Allan with Allan-Werle everywhere in the manuscript. Is now officially called in this way by the community.

Page 7, Line 30: replace « Such behavior is expected for white noise » with "Such behaviour is expected in case of white noise limited regime".

Page 9, Line 9: replace third with 3.

Page 10, Line 3: Replace "rain" with "bad weather conditions".

Page 10, Line 12: replace "in 20 minutes" with "within a time interval of less than 20 minutes". Replace "omitted" with "rejected".

Page 11, Line 13: Suggestion to replace "is not as good" with "is less pronounced".

Page 12, Line 7: in the sentence "…in 30 minutes modelling periods" add (T_i) which refers then to figure 9.

Page 17, Line 5 and 11: Replace "utilizing" with "using"

Page 17, Line 14: Why authors go back to to VT, VAQ and VSTA described in previous section? The way of presenting things is extremely confusing.

Page 17, Line 19: "The figure shows that the spatial coverage is not optimal, yet, in any version." Why? How authors get to this conclusion?

Arfire et al reference: add the journal.