Response to Referee #1, amt-2017-43

First I want to say that I appreciate the hard work that goes into this. You've selected a good sensor with a good reputation, and you're methodology for a neighborhood study is at a high-level the right approach–colocation calibration, a few weeks in the field, and then colocation calibration. I think this kind of work in the citizen sensing community is important, and I'm glad that your methodology incorporates good sensor technology and recent best practice. That said, I'm not sure what the precise contribution of this paper is.

New low-cost sensor technology for air quality application is available for several years now, and is used in many experiments often done by motivated but not necessarily scientifically trained people. This can result in gathering of data which, due to their poor quality, is unusable for quantifying air pollution. Our study shows that, if proper attention is payed to calibration, such experiments with low-cost sensors can result in useful measurements.

In its present form, however, the paper focusses on the technicalities of the calibration we applied, which might confuse the reader (or reviewer) that we are dealing with a strict scientific experiment in which all variables can be controlled. On the contrary, as our study deals with data which is generated in a citizen science campaign, one has to be creative to make sense of the gathered data.

Therefore, we propose to shift the focus to how to deal with the analysis of air quality data which is collected with imperfect sensors under imperfect conditions (e.g. in a citizen science campaign). We will still explain our calibration approach, but put more attention to our lessons learnt and recommendations on hardware, experimental set-up, and data analysis approach, as we believe that many future campaigns will benefit greatly from this information. This is now reflected in the new title "Field calibration of electrochemical NO_2 sensors in a citizen science context". We left the "Practical" out, as the sensor degradation issue prevent a really practical calibration scheme which can be used for similar initiatives.

In the realm of calibration technique and design, this is not state-of-the-art, nor is the methodology the right one if the point is the verification of a calibration algorithm. See this paper [http://www.atmos-meas-tech-discuss.net/amt-2017-138/amt-2017-138.pdf] for an example of the latest techniques and best practice– here HDMR takes into account more complex relationships than linear dependence and more complex variable interactions. In the linked submission, superior techniques with a longer co-location periods are applied to the Alphasense NO2 sensor.

The mentioned paper, Cross et al. (2017), was submitted to AMT on April 28, while our paper was submitted more than two months earlier. HDMR might be a more sophisticated method than the widely understood linear regression method used in our study. For NO_2, however, the authors find a RMSE of 8.6 µg/m^3 (4.56 ppb) for their test data, which is comparable to our estimated 7 µg/m^3 when applying our weighted calibration method based on multilinear regression. Unfortunately, Cross et al. do not give insight in their optimal HDMR model for NO_2 nor the sensitivity indices for input pairs (maybe because of the propriety nature of the ARISense device?); it remains unrevealed which signal relation best describes the NO2 concentration in their study.

They use training data which is distributed over a 4.5 month interval to derive a calibration model.

Given the fact that all devices must be calibrated individually, this is an impractical long period before they can be deployed at locations where no reference data is available. Furthermore, by using training data which is distributed over the entire period, sensor degradation within that period cannot be detected. Our study shows that, using training data from a consecutive period, degradation during a successive multiple-month period is significant.

Their methodology is also strong– instead of fitting their calibrations to their entire colocation dataset, they train a calibration on part of it and validate it on a holdout set. This is the proper methodology if your contribution is about multilinear calibration for electrochemical sensors.

In the revised version, we will include a predictive analysis in which the calibration is based on the first half of the calibration period, and the second half of this period is used for validation. The results show that the regression model describes well the measurements on short term, but loses predictability on the long term (e.g. two months) due to sensor degradation.

I presume the intended contribution has more to do with the installation/campaign and data collection ***between*** *co-located calibration, but I have some reservations here as well. While I do believe your data is likely reasonable given the calibration process/sensor selection/hour averaging, you haven't provided strong evidence to substantiate this belief, other than anecdotal evidence about one sensor located near another reference device. You also allude to the fact that (1) your colocation measurement has a lower normal ambient NO2 level than your campaign area, and (2) you don't measure O3 in your campaign area though it more strongly affects your measurement signal than NO2. This combination of facts leaves me quite concerned– the ratio of NO2/O3 might be consistent in your calibration area, and slightly different in your campaign area, and leave you with a systematic bias that you haven't properly accounted for. I don't think assuming the relative contribution of these two components is constant when you know that NO2 levels are different in the campaign area is a safe/fair assumption.*

We believe that the good agreement of sensor 54200 with the readings of an independent reference station OS (located at 3 km distance from the calibration site at Vondelpark) is more than anecdotal evidence. As can be shown in Table 5, RMSE of this sensor is 5.2 µg/m^3 during the two-month campaign period. From Figure 4 can be seen that ozone levels were generally lower than during the calibration period, but still the bias is acceptably small (-0.09 µg/m^3) meaning that the collinearity between temperature and ozone holds for both locations.

It must be said that both OS and Vondelpark station qualify as a city background station, which implies that they have similar NO_2/O_3 ratios. The Referee is right in his concern about the influence of different NO_2/O_3 ratios found at locations closer to emission sources. To get a better understanding of the possible impact, we compared hourly ozone measurements from the GGD authorities at Van Diemenstraat (VDS, classified as street station) against Nieuwdammerdijk (NDD, classified as urban background station) during June-August 2016. The location of these stations can be found at www.luchtmeetnet.nl. The relation can best be described by $[O_3]_{VDS} = 0.87\ [O_3]_{NDD} + 0.85$, which means that ozone levels at the street station are typically 13% lower that at the background station, due to titration of O_3 with NO. As the electrochemical NO_2 sensor is cross-sensitive for ozone, larger values must be subtracted from the sensor signal when the ozone concentration increases. This explains the negative ozone coefficient c_5 we find with calibration model E. According to the regression results in the Supplement a typical value for c_5 is -0.3.

Calibration with model D will overcorrect (i.e. subtract too much) for locations which have lower ozone concentrations than at the calibration site, resulting in an underestimation of NO_2 concentrations. For $[O_3]=60 \mu g/m^3$ (75 percentile of the distribution during the measurement camping, according to Figure 4) we estimate the underestimation in NO_2 at street side as $0.3 \times 13\% \times 60 = 2.3 \mu g/m^3$.

In our revised paper we will include this elaboration on location dependency of the calibration model. As already indicated in the Conclusions and Outlook, we believe that the inclusion of an additional low-cost ozone sensor (e.g. Ox-B431 by Alphasense) in an updated version of the device will reduce the bias due to different NO_2/O_3 ratios at different locations.

The 'sudden and unexplained' offset in the only sensor you kept colocated with your reference is also slightly concerning, and deserves more explanation/treatment than your paper provides.

We further analyzed the data of the reference sensor (55303) and found the cause of this sudden jump. Initially, this device not equipped with a PM10 module. Half-way the campaign, the technical operators decided to add this module, and removed the sensor between 10 and 14 July for service. Once placed back, temperature measurements by its DHT-22 sensor show that the internal device temperature increased by 2.5 degree on average. This can be attributed to increased power dissipation: after the periodic WiFi connection (350 mA peak), the PM module is the largest consumer of electricity (80 mA). This sudden jump in temperature is the main cause for the disrupted reference series.

There are many papers published that look at citizen science installations like this, and present novel work in other regards–things like spatio-temporal models that are validated against slightly better reference devices ('AirCloud', Sensys 2014), interesting UI for citizen interaction ('HazeWatch', Sensys 2013), etc. They are generally explicit about their contribution as a user interaction or have a slightly more compelling story around validation of their campaign data. They are also typically in human-interaction focused conferences.

More focus has been put on the citizen science aspect of our experiment. Unlike the mentioned projects, it focusses on low-cost NO2 sensing, which due to its specific calibration issues needs special attention to be successfully applied in a citizen science setting.

I'm not convinced that having a citizen campaign by itself warrants a publication, though it forms a strong foundation to experiment/build work on top of.

The revised paper is now stronger on the 'lessons learnt' side, so that the paper also can be read as a guide for more successfully setting up similar citizen science campaigns.

I do commend you on the open-sourcing of your data, and I think perhaps there is a case to be made that this aspect of it is worth publishing, but I'm still a little wary that validation of your data and key assumptions should be a little tighter (that NO2/O3 in your calibration/measurement region are similar, that your calibration technique is the proper one in the location of your measurement, etc). The lack of quantification of error in the locations you are measuring and the weak/qualitative claims about usefulness of the data are also a little disconcerting in this regard.

We feel that the inclusion of new analyses in the revised paper adds to the validity of our results. However, the set-up of the experiment limits the possibilities to answer some of these questions in detail at this stage, but gives us directions how to better organize future experiments.

Finally, there are several grammatical issues floating around the paper. (…) More in depth grammatical review is definitely required.

Thank you for pointing this out. We revised the grammar throughout the paper. We are willing to do a stricter language check by native speakers if the paper is selected for publication, and it is still considered necessary.