Response to Referee #3, amt-2017-43

General Comments:

1. The work presents the process involved in trying to calibrate a low-cost NO2 sensor for citizen science work. The sensor was collocated near a regulatory monitor for a period of 6 days, deployed in a community for 2 months, and then collocated again for a period of about 9 days. The work explored a number of calibration equations and determined that the best calibration equation would consider the temperature and relative humidity influences and the co-sensitivity to ozone. However, the sensors were not built to also measure ozone and thus, a calibration scheme omitting this factor was selected.

We conclude that the calibration without the ozone signal gives good results e.g. from the agreement of sensor 54200 with the readings of an independent reference station located at 3 km distance from the calibration site (RMSE of 5.2 µg/m^3 and negligible bias, see Figure 12). The collinearity between temperature, RH and ozone solves part of the sensor's cross-sensitivity to ozone. We now include a discussion how this calibration generates a bias at locations where the NO2/O3 ratio deviates from the calibration site. We estimate underestimations of NO_2 concentrations at street sides to be smaller than 2.3 µg/m^3 75% of the time (see response to Referee #1).

2. Unfortunately, the calibration procedure discussed is not novel or state of the art. Based on the title, I expected that it would be one or other or dynamic and easy to apply on the fly in the field. This definitely doesn't fit the bill. I think the manuscript would be better received if it were refocused to include a look at the data from the 2-month citizen science deployment.

In the revision, we 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, but put more attention to our lessons learnt and recommendations on hardware, experimental setup, and data analysis approach, as we believe that many future campaigns will benefit from this information. This is now reflected in the new title "Field calibration of electrochemical NO2 sensors in a citizen science context". An in-depth analysis of the campaign data will be the subject of a following paper.

3. I agree with the comments already posted by other reviews/researchers and have tried simply to add additional information in this review.

Specific Comments:

1. P4, Line 7 – Why was this criteria chosen? 33% of an hour seems rather low and at best arbitrary.

This criterion was found to be a good trade-off between noise reduction by averaging and not losing too many hourly measurements. This is now included in the text.

2. P4, Line 14 – Why was the collocation effort conducted at Vondelpark (urban back- ground) and not Oude Schans (urban)? This might have minimized the differences between the calibration and study periods.

Both Vondelpark as Oude Schans are classified as urban background stations. Vondelpark measures a broad range of species such as NO, NO2, PM2.5 and PM10, whereas Oude Schans only measures NO and NO2. Furthermore, Vondelpark station has better facilities such as accessibility, physical space, power supply, and internet connection.

3. P4, Line 25 – Include the average deployment period/time to the citizen campaign discussion.

Added to text: "In this 1537-hour period the devices produced 1204 valid hourly measurements on average."

4. P4, Line 33 – Throughout this manuscript, be more specific about your descriptors like higher and better. Discuss the metric used to make those determinations. For examples, regarding temperature on this line, the absolute highest temperature nor the mean temperature appears to be higher during both calibration periods so what metric are you looking at?

Our discussion of the distributions is based on the values of the 75th percentile. This is now included in the text. Also added to P6, Line 16: "As the electrochemical NO_2 sensor loses sensitivity at higher temperatures *(see the negative slope in Figure 8(b) for temperatures below 30°C)*"

5. P6, Line 18 – These two paragraphs should be re-visited to try to simplify. The model letters are not in order of best fit and that might help.

We swapped the B and C labels of the calibration models, so model A to E are now in order of increasing performance. We rewrote the mentioned paragraph to:

"From the fit results we see that Model B (including RH) performs better than Model A, but Model C (including T) outperforms Model B. When both RH and T are included (Model D) the results of Model C are improved marginally. This can be understood in terms of a strong sensor dependence on temperature, a weak dependence on RH, and the collinearity between temperature and RH. Note that measuring RH is essential for guarding the data quality of electrochemical sensors, as these sensors are very sensitive to *sudden changes* in RH, see e.g. AAN (2013) and Pang et al. (2016)."

6. P6, Line 26 – Why is ozone considered as a metric if it wasn't routinely measured during the campaign? It reinforces your argument that it should be measured but it's really no good to you in your current work. To really lend weigh to your argument that ozone should also be measured if using this NO2 sensor, you should explore whether the ozone concentrations from the nearest monitor would be a helpful addition and if a sensor based ozone measurement is good enough to help.

Ozone is measured at three locations in Amsterdam: two urban background locations, and one street side location (see www.luchtmeetnet.nl). Due to the chemical lifetime of ozone (which is long compared to NO_2), the ozone gradients over the city are rather smooth, except in the vicinity of NOx sources (such as motorized traffic) where ozone levels are generally lower due to titration by NO. From ozone measurement during the considered three-month period we derive that this reduction in ozone is around 13% (see our response to Referee #1). The relevance of including calibration model E in our study is that it quantifies the cross-sensitivity to ozone and enables us to make an estimation of the introduced bias when the sensor devices are located at a street side. This analysis is now included in the revised text.

7. P6, Line 30 – Discuss the technical differences between these sensor models.

The newer sensor model is designed to have higher sensitivity to NO_2 and less interference of O_3. The old sensor model has indeed smaller coefficients for SWE and larger correction terms for ozone (see the c_1 and c_5 coefficients of model E in the Supplement). This is now included in the text.

8. P7, Line 2 – Use a statistical measure rather than a figure of demonstrate improved performance.

We copy the corresponding results from the Supplement to specify: "R^2 increases from 0.30 to 0.83"

9. P7, Line 5 – What does calibrated but uncorrected mean?

Changed to "Calibrated data without temperature filter".

10. P7, Line 13 – What factors do you think affect the stabilization time. You mention 'most' sensors stabilized within this time. How many is most? Why not provide a range? What was different about the outliers?

When the device is switched on, the electrochemical cell must be stabilized by the potentiostatic circuit which takes a few hours (Alphasense Application Note AAN-105) due to the high capacitance of the working electrode. Furthermore, when the sensor is transported to another environment the sudden change in RH causes an equilibrium distortion with a relaxation time of about 2h (Mueller et al., Atmos. Meas. Tech., amt-10-3783-2017).

11. P7, Line 26 – Aging of temp and RH sensor is not widely reported as a problem. I realize the sensor was measuring in-box temperature and RH rather than ambient but is there really no available data (nearby temp and RH station) by which to but some bounds on this potential affect. Are you considering testing that hypothesis?

We assessed the possible degradation of DHT22 temperatures by comparing nighttime temperatures with temperature measurements of the GGD Vondelpark station (thus avoiding the effect of local heating by exposure to direct sunlight). Apart from device 55303 (which was modified halfway the campaign), all DHT22 sensor maintain a stable offset with regard to ambient temperature before and after the campaign. In the revision, we therefore remove our suspicion that "part of the drift could also be partly related to the aging of the DHT22 temperature and RH sensor".

12. P10, Line 17 – I think it might also be worth noting what this method would not be able to detect like transient spikes from nearby sources (because you are eliminating any spike outside of 10% of the mean). Because of this exclusion criteria, why do you think you could use this model to provide realistic estimates of peak values?

Due to the large offset in the raw S_{WE} and S_{AE} signal (around 1200, see Figure 3), realistic NO_2 peak values are still detectable as the corresponding sensor response is within the 10% bandwidth around the average raw sensor signal. We added this remark in the description of the filter criteria in Section 3.1

13. Figure 1 – I would like to see the Vondelpark station on this map to better appreciate the distance and variation in the urban environment. It would also help to see how large of an area this study area is in comparison with the city of Amsterdam.

We agree and extend the map accordingly.

14. Figure 2 – Rather than the photo of the sensor boxes charging, I think it would be helpful to see how they sit within this housing to better understand the appropriateness of the temperature and relative humidity measurement, etc.

We include in Figure 2 a new photo showing the components in their housing.

15. Figure 3 – it appears that one sensor, in particular, appears to be an outlier in most of this figures. Was its removal from the study ever considered? Why/why not?

Temperature and RH are converted from mV according to the specs of the DHT-22 sensor manufacturer. The spread in temperature and RH displayed in the raw data is partly explained by the sensor-to-sensor variability. However, the devices are not actively ventilated (this will be included in the recommendations!), which means that they are susceptible for direct sunlight and heat generation from the electronic modules. For the apparent outlier this occasionally happens in the strong non-linear regime of the NO2 sensor, which explains the corresponding strong dips in the S_{WE} signal. After temperature filtering (explained in Section 4.4) and calibration, its performance gave no reason to exclude it from our study.

16. Figure 4 – Please check the text to make reference to Vondelpark and Oude Schans (OS) more consistent and clear. I believe at one post one of the stations is just referred to as GGD.

Ambiguous references in the text to '*GGD station*' have been changed to '*GGD Vondelpark station*'.

17. Figure 6 – Graph is not needed, equation and R2 in the text is sufficient for making this point.

We agree. We will take this plot out and explain textually.

18. Figure 7b – A Figure is not the best way to support the assertion that improved performance is clearly shown. It appears to me to be true only about 50% of the time from this figure.

Figure 7b should be interpreted as an illustration how the improved scatter of Figure 7(a) (panel D versus panel A) represents as time series. The series show that, apart from 7 June, model D (blue lines) is closer to the ground truth (grey line). We added in the text to further specify: "R^2 increases from 0.30 to 0.83".

19. Figure 8a – I would remove this Figure. If you leave it, include temperature.

We will include a second y-axis in this plot with the internal sensor temperatures to better illustrate the non-linear temperature effect.

20. Figure 8b – Just reference the data sheet.

We prefer to keep this Figure, as we think it illustrates the direct cause of the non-linear temperature dependence, and we are not sure if the manufacturer will still provide this NO2-B43F data sheet on their website once they release a new sensor model.

21. Figure 9 – Figure, in this format not needed. If you want a figure, it would more useful to show error between measurements vs. time and for each sensor as it starts.

We agree. We will take this plot out and explain textually.

22. Figure 10 – Using similar scales would help illustrate the drift.

We decided to replace this figure with a plot showing the distribution of the residuals during the two co-location periods.

23. Figure 11 – Error bars/estimates for the coefficients before and after would be a helpful comparison in this Figure.

We decided to leave this figure out (see Referee #2).

24. Figure 12b – Present R2.

We replace Figure 12b by a plot of the distribution of residuals and we extend Table 5 with statistic summaries for the first and second calibration periods (see Referee #2).

25. Tables – Find a way to visually note the older sensors by ID number.

To increase readability, we decided to rename all device IDs to SD*nn*, with *nn* from 01 to 16. A table is added in the Supplement with the relation between old and new IDs. The older NO2-B42F sensors are now labelled SD01 and SD02. To make a better distinction between the different models we highlight SD01 and SD02 in grey in Table 1, Table 3 and Table 4.