Response to Referee #2, amt-2017-43

Decoupling the interference between NO2 and O3 with Alphasense sensors is a difficult task, as highlighted throughout the literature. However, after reading through the manuscript several times, it does not appear the author's goal was accomplished based on their thesis: to describe a "practical method for in-field calibration and regression modeling" of electrochemical NO2 sensors. Several major concerns including the use of a reference instrument (ozone) as an independent variable within the model and lack of rigorous validation data must be addressed.

The best-performing model includes data from a reference ozone monitor which does not constitute a "practical method" for using low-cost NO2 sensors, and the regression modeling nearly completely describes how well these sensors performed in the past, without properly withholding validation data to describe how they will hold up in the future (predictive versus descriptive modeling). The modeling approach (multivariate linear regression using WE and AE) is not novel in the literature concerning Alphasense electrochemical sensors, especially when considering species other than NO2 (see Lewis 2015[1]) as an example that uses both linear regression and other statistical models).

New low-cost sensor technology for air quality 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, 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 greatly from this information. This is now reflected in the new title "Field calibration of electrochemical NO2 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 addition to a few major corrections, many minor corrections should be addressed as well (outlined below). Therefore, publication of this manuscript in AMT should only be considered after the comments below have been addressed.

Major Comments

P. 6, line 24: Including a reference ozone measurement as an independent variable in the linear model is inappropriate for low-cost sensing. If the goal is to describe a method by which you can use low-cost NO2 sensors to obtain a decent NO2 concentration, then including data from a $5000+ instrument in the analysis simply cannot be included. I understand that there is a strong cross-

sensitivity to ozone, but claiming even a poor ozone measurement would improve results without any evidence to support the claim is invalid. This should be removed completely from the analysis.

Cross-sensitivity to ozone is an important sensor issue, and should be corrected for to get more accurate low-cost NO2 measurements. We think it is appropriate to include it in the analysis to get a better understanding of cross-sensitivity to ozone. We show that the accuracy of the low-cost measurements increase when ozone is included in the correction. This does not mean that the sensor devices should be equipped with a $5000+ instrument. We soften our claim that the performance of the device will improve significantly when low-cost ozone sensors are included: "To further improve accuracy of electrochemical NO2 measurements in low-cost sensor devices we recommend to include an additional low-cost ozone sensor. It is likely that the linear regression approach is able to resolve a significant part of the cross-sensitivity to ozone and NO_2." .

To show the model is predictive (rather than descriptive), previously withheld validation data should be used to evaluate the model. Currently, this work only shows that these sensors can reasonably describe what has been measured in the past, but provides no insight into well they will hold up in the future.

Good point. 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.

All fit parameters in the tables (and throughout the paper) should have error estimates/confidence intervals.

The standard deviations of the regression coefficients will be included in the tables of the Supplement.

A focus on the absolute RMSE, rather than just the bias-corrected RMSE should be highlighted in the abstract

Our claim that "the standard deviation of a typical sensor device for NO_2 measurements was found to be 7 µg m^{-3}" in the abstract is based on the assumption that the weighted calibration approach (described in Section 4.7) removes the sensor bias largely, which is supported by our findings in Section 4.8.

Minor Comments

There are many English language errors (mostly grammar) that need to be worked 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.

P. 2, line 6: These sensors are commercial, not experimental, despite their quality. Stating otherwise supports the idea that they are not currently on the market, which they are.

Adjusted to "many low-cost air quality sensors suffer from various technical issues which limit their applicability."

P. 4, lines 3-4: Rather than just throwing away data based on arbitrary filters, a digital filter could be used. Throwing away data that is not within 10% of the mean is probably not the best methodology; one gives up the ability to measure higher concentrations if a local source were to emerge!

This filter criteria was selected after carefully studying the raw (1-minute) data. As can be seen in Figure 3, the +- 10% bandwidth is wide enough to contain all valid measurements in the linear regime. The filter criterion is simple, yet effective. We did tests with advanced noise filtering using a Fourier transform, but this did not result in significant improvement of the hourly data quality.

If the analysis is going to be based on the "more linear" regime of these sensors (dropping all data > 30C), it should be more pronounced in the abstract and introduction (page 4, lines 5-6). This is a huge limitation and one of the most important research topics for electrochemical sensors (as used for ambient measurements).

This will now be included in the abstract: "Using our approach, the standard deviation of a typical sensor device for NO_2 measurements was found to be 7 µg m^{-3}, *provided that temperatures are below 30°C*.". This limitation is further addressed in the Conclusion/Outlook.

P. 6 line 10: If the DHT22 sensor does not need to be individually calibrated, the authors should explain why they observed such large variance between DHT22 sensors and how this affects their model results

The spread in temperature and RH displayed in the raw data is partly explained by the sensor-to-sensor variability. By looking at nighttime temperatures (to eliminate the effect of local heating by exposure to direct sunlight) we discovered that all derived sensor temperatures are 2-5 degrees higher than the ambient temperature. The devices are not actively ventilated (this will be included in the recommendations!), which means that the energy dissipation of the device influences their internal temperature. The variable position of the temperature sensors with respect to these heat sources further explain the variance in temperature. This analysis will be included in the analysis of the raw measurement in Section 3.1.

P. 6, line 15: Comments suggest the sensor loses sensitivity at higher temperatures. This seems counterintuitive given that diffusion across the membrane should be faster at higher T. What is the explanation for this effect?

From the technical data sheet shown in Figure 8(b), one can see that sensitivity of the NO_2 sensor decreases linearly with temperature up to around 30 degrees. Above 40 degrees the sensor gains sensitivity with rising temperatures. This is added to the text in Section 4.4. The application of a detailed temperature dependency model to describe this non-linear behavior was considered outside the scope of our research.

P. 11, lines 5-6: Diverging results for two different models of Alphasense NO2 sensors are discussed; Alphasense explains why the newer version of the sensor obtains better selectivity towards NO2 and has a reference (Hossain 2016 [2]) that should be examined/discussed.

A more detailed analysis will be added with reference to Hossain 2016.

Equations 6, 7: A time-based interpolation for back-calculation of NO2 is used without sufficient evidence the decay in sensitivity/accuracy is linear in time.

We feel that the assumption that the degradation is linear in time is the best to be made, given the limited data of our experiment and the absence of relevant scientific literature assessing electrochemical sensor degradation.

P. 9, line 22: Is there a reason the authors decide to use r2 rather than adjusted-r2 for comparing to adjusted-r2?

Thank you for pointing this out. We now include an analysis of the adjusted R^2 in the analysis of the NO_2 calibration models in Section4.3 and in Figure 7(a). However, the adjusted R^2 does not change dramatically from R^2, as the number of observations ($n \approx 150$) is relatively high compared to the number of regression variables (k=2…5).

*The median value throughout the campaign is 15 ugm-3 and the stated 95% CI is 14 ugm-3 (2*RMSE); what is signal and what is noise?*

From our error estimation of the sensor devices one can conclude that for low NO_2 values the noise dominates the signal. However, from Figure 4 can be seen that about 25% of the measurements at Oude Schans station were above 25 ug/m3 during the campaign (one is usually more interested in detecting occurrences of high pollution levels). At these levels the signal to noise is significantly better.

What makes a measurement "good enough" (page 10, line 15)?

Gradients in NO2 over the city are often too local that all features can be captured by the limited amount of official air quality stations. When looking at the difference between Vondelpark station and Oude Schans station (both classified as city background stations) between June and August 2016, 22% of the hourly measurements differ more than 7 ug/m3, and 6% of the hourly measurements differ more than 14 ug/m3. These ratios increase further when considering road side stations. From this perspective, sensor devices with an accuracy around 7 ug/m3 can contribute to an improved understanding of spatial patterns.

Claiming the calibration period should be "as long as possible" isn't very helpful. Eventually, the sensitivity of the sensor would begin to decay and one would lose valuable time to move the device and measure other places! Is there a quantitative way to phrase "as long as possible"?

The Referee is right in his remark that sensor degradation would interfere with long calibration times. We change the text to: "It is hard to quantify an optimal length of a calibration period without having a proper understanding of the sensor degradation rate beforehand. The measurement period should be at least a few days to capture the sensors behavior under a wide range of pollution levels and meteorological conditions. Very long calibration periods (in the order of months) will cause sensor degradation issues to interfere with the calibration results."

The description of the in-field co-location (when an NO2 sensor is compared to a closeby reference sensor) is quite confusing. It took several read-throughs to really under- stand when and where everything was taking place. This could be greatly simplified by adapting the map figure with notes.

Thank you for this suggestion. We added more information on the map in Figure 1, which now should explain better the set-up of our study.

P. 9, lines 6-14: The authors claim an in-field co-located NO2 sensor stays calibrated at another site, but the error bars on those measurements are the same as the absolute value of those measurements. How can one be sure they are not just looking at noise?

The correlation with measurements of the nearby site (Oude Schans) is 0.88 (Table 5), showing that the sensor device is measuring NO_2 reasonably well (see also Figure 12). If we were looking at noise, correlations would be close to 0.

Figure Comments

Each figure should be able to stand alone and tell a story; many of the figures do not contribute substantially to the paper and could be omitted. Specific comments include:

Figure 1 needs labels for the co-location stations (text) to make it easier to understand what was taking place

We added more information on the map in Figure 1, which now should explain better the set-up of our study.

Figure 3 demonstrates a large absolute error on some of the RH and T measurements (15 C swing on Temp and 20% on RH). Why? Should counts be converted to volts to ease comparisons with existing literature? What is going on with the clear outlier?

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. As explained above, all derived internal sensor temperatures were found to be 2-5 degrees higher than the ambient temperature, indicating that the energy dissipation of the device influences its internal temperature. During daytime, the exposure to direct sunlight (the devices were places the rooftop of the monitoring station) contributes further to the temperature outliers seen in Figure 3. These happen in the strong non-linear regime of the NO2 sensor, which explains the corresponding strong dips in the SWE signal.

Figure 5 should have more descriptive axis labels – using just the title to describe the plot makes it hard for the reader to understand what is going on. Many of these plots are not needed ([row 2, col 2], [row 3, col 1], [row 3, col 2], [row 4, col 3]). The authors claim ozone is correlated with AE response, but clearly, that is just a temperature effect. Otherwise, the authors need to describe how ozone can diffuse across the analyte and undergo a redox reaction at the AE surface.

We clarified the plots by adjusting the title and including the regression coefficients. We decided not to leave out panels, as we feel that all panels illustrate a different aspect of the behavior of the sensor. However, we improved our description of this figure in the text.

Figure 6 is not needed. It does not add anything to the paper and is well known through basic photochemistry.

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

Figure 7a should not include the model with ozone in the regression (row 2, col 2)

We think it is appropriate to leave it in the analysis to get a better understanding of cross-sensitivity to ozone.

Figure 8a does not do a good job at conveying the point (that transient temperature spikes affect the signal) since temperature is not shown anywhere.

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

Figure 8b is not needed. These details are in the technical spec sheet and previous literature – just cite those.

We feel that this figure is illustrative for better understanding the non-linear temperature behavior, losing sensitivity with increasing temperatures, followed by a strong gain in sensitivity for higher temperatures.

Figure 9 is not needed – simply describing the start-up/warm-up period in the methods section along with other filtering methodology was sufficient.

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

Figure 10a and 10b do not seem to convey what you are trying to convey – plotting a distribution of the residuals during the two co-location periods would be much more helpful.

Good suggestion. We will adjust the figure accordingly.

Figure 11 was already described in a Table – no need for a plot as well. They are very confusing and don't add anything in terms of advancing the story. It just makes it seem like the linear model is not very robust or repeatable. It also appears to suggest the is a drift in the y-intercept of nearly 1000 ugm3 in some instances!

We agree and take this plot out.

Figure 12b could also be plotted as a distribution of residuals – one would then be able to see clear overlap (or not) if there is/isn't bias.

Good suggestion. We will adjust the figure accordingly, and we will extend Table 5 with statistic summaries for the 1st and 2nd calibration periods.