Cover letter major revision amt-2017-43

First we would like to thank the three referees for their time to evaluate our manuscript and provide useful comments. It helped us to restructure the paper and add additional results, and so improve the quality of the presented work. The detailed changes which lead to this mayor revision can be found in the Track Changes version below.

Outline of the most important changes:

- New title which better captures the presented study: “Field calibration of electrochemical NO2 sensors in a citizen science context”.
- More focus on citizen science context of the described experiment, including more recommendations for improved follow-up experiments.
- Inclusion of predictive regression, showing that the calibration model is able to predict values on a short time scale.
- An improved analysis of the introduced bias when using the sensor devices at sites where the NO2/O3 ratio is different than at the calibration site.
- An improved analysis of the temperature readings in relation to the internal sensor temperature.
- Reduction and update of figures.
- New (better readable) labelling of sensor IDs.
- Revision of the English grammar throughout the manuscript. We are willing to do a stricter language check by native speakers if the paper is selected for publication.

Table of contents

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 first submission, however, the paper focussed more on the technicalities of the calibration we applied, which might have confused 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 have shifted 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 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 $\mu g/m^3$ (4.56 ppb) for their test data, which is comparable to our estimated 7 $\mu g/m^3$ 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 included 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 µg/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 × 13% × 60 = 2.3 µg/m^3.

We included this elaboration on location dependency of the calibration model now in the Discussion section. 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. This is now included in Section 4.6.

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, our study 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 it 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.

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 first submission, however, the paper focussed more on the technicalities of the calibration we applied, which might have confused 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 have shifted 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 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 (Section 6): "To improve the NO_2 measurements further we recommend to include an additional low-cost ozone sensor, e.g. Ox-B431 by Alphasense. 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 included a predictive analysis in Section 4.6 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 are now 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. Added to the text: "This criterion was used for its simplicity and effectiveness. Note that, due to the large offset in the raw S_{WE} and S_{AE} signal, realistic NO_2 peak values are still detectable as the corresponding sensor response is still within a 10% bandwidth."

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 is now 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 (updating the hardware with active ventilation is now 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 is now 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 now mentioned 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.

Loss of sensitivity during lifetime and improved sensor design are now mentioned in Section 4.3: "The two worst performing sensor devices (SD02 and SD01) contain the older NO2-B42F sensor. The newer NO2-B43F model is designed to have higher sensitivity to NO2 and less interference of ozone. The old sensor model has indeed smaller coefficients for S_{WE} and larger correction terms for ozone (see the c_1 and c_5 coefficients of model E in the Supplement). This, however, can also be related to their longer operating time, as both sensors have been used in previous experiments for more than a year."

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 Section 4.3 and in Figure 7(a). However, the adjusted R^2 does not change dramatically from R^2, as the number of observations (n≈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 changed 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. This elaboration is now included in Section 3.1.

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 took 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 included 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 took 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 adjusted 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 took this figure 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 adjusted the figure accordingly, and extended Table 5 with statistic summaries for the 1st and 2nd calibration periods.

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 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 7(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.

Added to Section 4.3: “The two worst performing sensor devices (SD02 and SD01) contain the older NO2-B42F sensor. The newer NO2-B43F model is designed to have higher sensitivity to NO2 and less interference of ozone. The old sensor model has indeed smaller coefficients for S_{WE} and larger correction terms for ozone (see the c_1 and c_5 coefficients of model E in the Supplement). This, however, can also be related to their longer operating time, as both sensors have been used in previous experiments for more than a year. “

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 removed 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 extended 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 included a new panel in Figure 2 showing the position of 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 took 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 included 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 took 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 replaced Figure 12b by a plot of the distribution of residuals and we extended 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.

Field calibration of electrochemical NO_2 sensors in a citizen science context~~Practical field calibration of electrochemical NO_2 sensors for urban air quality applications~~

Bas Mijling[1], Qijun Jiang[2], Dave de Jonge[3], Stefano Bocconi[4]

[1]Royal Netherlands Meteorological Institute (KNMI), Postbus 201, 3730 AE, De Bilt, The Netherlands
[2]Laboratory of Geo-Information Science and Remote Sensing, Wageningen University & Research, Droevendaalsesteeg 3, 6708 PB Wageningen, The Netherlands
[3]Public Health Service of Amsterdam (GGD), Nieuwe Achtergracht 100, 1018 WT, Amsterdam, The Netherlands
[4]Waag Society, Nieuwmarkt 4, 1012 CR, Amsterdam, The Netherlands

Correspondence to: Bas Mijling (mijling@knmi.nl)

Abstract.

In many urban areas the population is exposed to elevated levels of air pollution. However, real-time air quality is usually only measured at few locations. These measurements provide a general picture of the state of the air, but they are unable to monitor local differences. New low-cost sensor technology is available for several years now, and has the potential to extend the official monitoring network significantly even though the current generation of sensors suffer from various technical issues.

Citizen science experiments based on these sensors must be designed carefully to avoid generation of data which is of poor or even useless quality. This study explores the added value of the 2016 Urban AirQ campaign, which focused on measuring nitrogen dioxide (NO_2) in Amsterdam, the Netherlands. 16 low-cost air quality sensor devices were built and distributed among volunteers living close to roads with high traffic volume for a two-month measurement period.

Each electrochemical sensor was calibrated in-field next to an air monitoring station during an 8-day period, resulting in R^2 ranging from 0.3 to 0.7. When temperature and relative humidity are included in a multilinear regression approach, the NO_2 accuracy is improved significantly, with R^2 ranging from 0.6 to 0.9. Recalibration after the campaign is crucial, as all sensors show a significant signal drift in the two-month measurement period. The measurement series between the calibration periods can be corrected in hindsight by taking a weighted average of the calibration coefficients.

Validation against an independent air monitoring station shows good agreement. 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. Stronger ozone titration at street sides causes an underestimation of NO_2 concentrations, which 75% of the time is less than 2.3 µg m^{-3}.

Our findings show that citizen science campaigns using low-cost sensors based on the current generations of electrochemical NO_2 sensors may provide useful complementary data on local air quality in an urban setting, provided that experiments are properly set up and the data are carefully analysed.~~In many urban areas the population is exposed to elevated levels of air pollution. However, air quality is usually only measured at a few locations. These measurements provide a general picture of the state of the air, but they are unable to monitor local differences. Since a few years new low-cost sensor technology is available, which has the potential to extend the official monitoring network significantly. These sensors, however, are still in an experimental stage and suffer from various technical issues which limit their applicability.~~

~~This study explores the added value of alternative air quality measurements, focusing on nitrogen dioxide (NO_2) in Amsterdam, the Netherlands. 16 low-cost air quality sensor devices were built and distributed among volunteers living close to roads with high traffic volume for a two-month measurement campaign.~~

~~Careful calibration of individual sensors is essential to measure ambient concentrations of NO_2 significantly. Field calibration was done next to an air monitoring station during an 8-day period, resulting in R^2 ranging from 0.3 to 0.7. The NO_2 accuracy can be improved by including temperature and humidity measurements from an additional low-cost sensor, R^2 ranging from 0.6 to 0.9. Recalibration is crucial, as all sensors show significant signal drift after the two-month measurement campaign. The measurement series between the calibration periods can be corrected in hindsight by taking a weighted average of the calibration coefficients.~~

~~Validation against an independent air monitoring station shows good agreement. Using our approach, the standard deviation of a typical sensor device for NO_2 measurements was found to be 7 μg m^{-3}. This shows that, if properly treated, low-cost sensors based on the current generations of electrochemical NO_2 sensors may provide useful complementary data on local air quality in an urban setting.~~

1 Introduction

Because air pollution is difficult to measure, instrumental and operational costs of official measurement stations are usually high. Air quality networks in cities, if present at all, are therefore usually sparse. Diffusive sampling is a common addition to these real-time measurements and are successfully used to monitor local differences (see e.g. Cape, 2009). However, these differences are poorly attributed to an emission source due to the long averaging time of these measurements (usually 4-weekly). Emerging low-cost sensor technology has the potential to extend the official monitoring network significantly, and improve our understanding of local urban air pollution. Miniaturized and affordable sensors potentially enable citizens to measure their environment in more detail in space and time (Kumar et al., 2015). ~~However, m~~Most commercially available sensors ~~are still in an experimental stage and~~ , however, suffer from various technical issues which limit their applicability. Despite their limitations many experiments are done with air quality devices containing these sensors, often by motivated but not necessarily scientifically trained people. Comprehensive calibration and validation of these devices is crucial (see e.g.

Lewis and Edwards, 2016; Lewis et al., 2016), but often overlooked. The resulting poor data quality is of concern to health authorities, scientists and citizens themselves.~~The poor data quality is of concern to health authorities, scientists and citizens themselves. Before conclusions can be drawn from the measurements, comprehensive calibration and validation is essential (e.g. Lewis and Edwards, 2016; Lewis et al., 2016).~~

Several studies have been done to explore the performance of low-cost air quality sensors, e.g. Jiao et al., 2016, Duvall et al., 2016; Mead et al., 2013; Moltchanov et al., 2015. For NO_2 monitoring, mostly metal oxide and electrochemical sensors are used (Borrego et al., 2016; Spinelle et al., 2015b; Thompson, 2016). Typical ambient concentrations of NO_2 are at part-per-billion (ppb) level. The main problems encountered in NO_2 sensor evaluations in these real-world environments are low sensitivity, poor selectivity, low precision and accuracy, and drift. Especially metal oxide sensors are not very stable (Spinelle et al., 2015b; Thompson, 2016) and suffer from lower selectivity. Therefore, in this study, we opted for electrochemical sensors to measure NO_2.

Mead et al. (2013) already noted the strong interference of ozone and other ambient factors in electrochemical NO_2 sensors. The performance can be increased significantly when adding additional measurements of e.g. temperature and humidity in a regression model or neural network, as shown by e.g. Piedrahita et al. (2014), Spinelle et al. (2015b), Masson et al. (2015). Coping with sensor degradation remains a serious issue. Some studies, such as Jiao et al. (2016), include an additional temporal term in their linear regression which improves the predicted NO_2 slightly.

In the following sections we assess the data quality of the 2016 Urban AirQ campaign. As many similar initiatives depending on participating citizens, this campaign was not set up as a strictly controllable scientific experiment such as in the previously mentioned studies. However, we will demonstrate that citizen air quality monitoring using the current generation of electrochemical NO_2 sensors may provide useful data of urban air quality, by using a practical method for field calibration and correcting for sensor degradation in hindsight.~~The following sections will further explore the applicability of electrochemical NO_2 sensors for measurements of urban air quality, using a practical method for in-field calibration and regression modelling for assessment of accuracy and sensor degradation.~~

2 The Urban AirQ project

The Urban AirQ project explores the added value of alternative air quality measurements in the city, by addressing citizens' questions about their local air quality. ~~.~~ It focusses on a 2×1 km^2 area around Valkenburgerstraat, a primary road in the East-central part of Amsterdam, see Figure 1. Its dense traffic causes regular exceedances of the European annual limit value for nitrogen dioxide (40 $\mu g\ m^{-3}$).

Two town hall meetings were organized in which residents of this area were invited to raise their concerns about air pollution in their neighborhood and to formulate related research questions. Topics included the relation between traffic density and air pollution, the difference between a main road and a side street, the front side of an apartment compared to its backside, the influence of apartment height, and the influence of cut-through traffic at nighttime. The residents were invited to participate in finding answers to their questions by measuring their outdoor air quality with 16 experimental low-cost sensor devices (labeled SD01 to SD16), built for this purpose by Waag Society.

Measurements were done from June to August 2016. Beforehand, the sensor devices were calibrated using side-by-side measurements next to an official air quality measurement station. With a second calibration period after the campaign, individual sensor drift was assessed and compensated in hindsight.

The Urban AirQ experiment is unique in the sense of the used number of devices, the duration of the experiment, the direct involvement of citizens, and the use of open hardware and generation of open data.

3 Urban AirQ ~~NO_2~~ sensor devices

The concept of the Urban AirQ sensor is building a device with low-cost electronic components which is easy to operate, so citizens can do their own air quality measurements. It builds on the basic design described by Jiang et al. (2016), having an improved power supply, weather resistant housing, WiFi connectivity, and additional sensors for temperature, relative humidity, and particulate matter. The sensor development is part of an open hardware project; detailed technical information can be found at https://github.com/waagsociety/making-sensor.

Central is the microcontroller board (Arduino UNO) which handles the reading of the sensors and sends the data to the WiFi module (ESP8266), see Figure 2.

For NO_2 measurements, an ~~amperometric~~ electrochemical cell is used from Alphasense Ltd (Essex, United Kingdom). The cell contains four electrodes. The target gas, NO_2, diffuses through a membrane where it is chemically reduced at the Working Electrode, generating a current signal. This electric current is balanced by a opposite current from the Counter Electrode. The Reference Electrode sets the operating potential of the Working electrode. The sensor also includes an Auxiliary Electrode, which is used to compensate for baseline changes in the sensor. To get full sensor performance, low noise interface electronic is necessary. An individual sensor board with amperometric circuitry, also provided by Alphasense, is used to guarantee a low noise environment and to optimize the sensor resolution at low ppb levels. The sensor signal is read by a 16-bit analog to digital (A/D) converter (ADS1115). ~~14 sensor devices contain model NO2-B43F for NO_2 measurements, the other two use model NO2-B42F.~~ Two sensor devices (SD01 and SD02) contain model NO2-B42F for NO_2 measurements, the other 14 contain the newer NO2-B43F sensor.

12 of the 16 sensor boxes are also equipped with a Shinyei PPD42NS sensor in order to measure particulate matter optically. The present paper, however, will focus only on the assessment of the NO_2 measurements.

All devices ~~are equipped~~measure internal temperature and relative humidity (RH) with a DHT22 sensor from Aosong Electronics ~~measuring temperature and relative humidity (RH)~~.

~~12 of the 16 sensor boxes are also equipped with a Shinyei PPD42NS sensor in order to measure particulate matter optically. The present paper, however, will focus only on the assessment of the NO_2 measurements.~~The system is supplied with a 7.5V output adapter and a regulator board which generates 5V for the Arduino and the sensors. The microcontroller consumes around 10 mA. The PM sensor needs a 80 mA current, the NO_2 sensor about 10 mA, and the DHT22 less than 1 mA. The WiFi module peaks periodically to 350 mA when establishing an internet connection.

3.1 Averaging and filtering

Raw sensor measurements are stored in a central database on a one minute base. However, the calibration analysis is based on hourly averages to enable direct comparison between the ground truth (also provided as hourly values), and to improve the signal to noise ratio.

The NO_2 sensor measurements are done at the Working Electrode (S_{WE}) and the Auxiliary Electrode (S_{AE}). They are provided as counts from the A/D converter. Sensor readings of temperature and RH are converted according to the indication of the manufacturer to degrees Celsius and percentages respectively.

Raw, hourly averaged, sensor data is shown in Figure 3. 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 see that the internal sensor temperatures are 2-5°C higher than ambient temperature. The devices are not actively ventilated, which means that the energy dissipation of the electronics influences their internal temperature. The variable position of the temperature sensors with respect to these heat sources further explain the variance in temperature and relative humidity.

Careful filtering is needed before the data can be further processed. We have applied the following rules:

- Raw, minute-based, S_{WE} and S_{AE} measurements outside a ±10% range of their mean value during the entire measuring period are considered outliers. This ~~affects~~ filters out 0.33% of all measurements. This criterion was used for its simplicity and effectiveness. Note that, due to the large offset in the raw S_{WE} and S_{AE} signal, realistic NO_2 peak values are still detectable as the corresponding sensor response is still within a 10% bandwidth.
- All readings at sensor temperatures above 30°C are discarded to avoid non-linear temperature dependence of the electrochemical NO_2 sensor (see Sect. 4.4). This ~~affects~~ filters out 4.53% of the measurements during the entire period.
- At least 20 valid minute-based measurements are required to calculate a representative hourly mean. This criterion was found to be a good trade-off between noise reduction by averaging and not losing too many hourly measurements.

During the first calibration period, the sensors were measuring 79% of the time on average. After applying the criteria above, this resulted in 70% valid hourly measurements. During the measurement campaign, the sensors produced 79% valid hourly measurements on average, with the uptime dropping to 50% in places were sensors experienced connectivity problems due to limited range of the participant’s WiFi network.

3.2 Calibration periods

Calibration of the sensors devices have been done by placing the 16 sensors side by side on the rooftop of the air quality station at Vondelpark, operated by the Public Health Service of Amsterdam (GGD). This station is classified as a city background station. It measures nitrogen dioxide, nitrogen monoxide (NO), ozone (O_3), particulate matter (PM_{10}, $PM_{2.5}$, particle number and size distribution), black carbon, and carbon monoxide (CO). For NO and NO_2 measurements, GGD alternates Teledyne API 200E and Thermo Electron 42I NO/NO_x analysers, both based on chemiluminescence. The validated measurements used in this study are considered to be the ground truth. The calibration period spanned several days to be able to test the sensors under a wide range of ambient conditions. To assess the stability of the calibration, the sensors were brought back after the two-month measurement campaign to the calibration facility for a second calibration period. The Urban AirQ campaign consisted therefore of three phases.

The first field calibration period at GGD Vondelpark station started at 2 June 2016, 00h LT (local time), and ended at 10 June 2016, 10h (8.5 days; 204 hours). Due to connectivity problems sensor data was missing between 4 June 19h and 6 June 9h.

During the following citizen campaign, 15 sensors were distributed among the participants. One sensor (~~55303~~SD03) was kept at the Vondelpark station as a reference. The first sensor was installed and connected at 13 June 2016, 18h, the last sensor connected at 17 June 2016, 17h. At 15 August 2016, 9h, the first sensor was disconnected, at 16 August 2016, 18h, the last sensor was disconnected. In this 1537-hour period the devices produced 1204 valid hourly measurements on average.

The second field calibration period at GGD Vondelpark station started at 18 August 2016, 15h, and ended at 29 August 2016, 00h (10.4 days; 249 hours) . Due to connectivity problems sensor data was missing between 26 August 12h and 27 August 11h.

Figure 4 shows the distribution of temperature, relative humidity, NO_2, and O_3 during the different periods. Looking at the 75th percentile of the distributions, ~~T~~the calibration periods are characterized by higher temperatures and ozone levels than the campaign period. The range of ~~hourly~~ NO_2 concentrations at the Vondelpark station in the calibration periods is larger than in the campaign, reaching more frequently higher NO_2 values. During the campaign the sensors are ~~more closely located~~closer to the GGD station at Oude Schans, where measured~~.~~ NO_2 values ~~measured here~~ are generally a few µg m^{-3} higher than at Vondelpark. The Oude Schans site does not measure ozone.

4 NO_2 calibration

Electrochemical sensors such as the Alphasense NO2-B series~~,~~ are known to be sensitive to interfering species and ambient factors. Especially ozone, temperature, and relative humidity influence the sensor reading (see e.g. Spinelle et al., 2015a).

4.1 Explaining the NO_2 sensor signal

To understand better the behavior of the NO_2 sensor, we study its sensitivity to different ambient factors. We use the first calibration period to test the correlation of the measured S_{WE} and S_{AE} signal with NO_2, ozone, temperature and humidity by making a best fit though the hourly time series, e.g.

$$S_{\mathrm{WE}}(t) = c_0 + c_1 \mathrm{NO}_2(t) \tag{1}$$

~~-~~Temperature and RH were not available from the obtained GGD Vondelpark station data. ~~Instead of taking from ambient air measurements, w~~We take temperature and RH from the average readings from the DHT22 sensors instead, which~~, as these reflect~~ better reflect the internal sensor conditions than ambient air measurements.

Figure 5 shows scatter plots for an average performing sensor and the R^2, the coefficient of determination. The measured S_{WE} signal can be explained by ambient NO_2 (R^2=0.20), but better by its anti-correlation with ozone (R^2=0.49). Temperature alone is an even better predictor for the sensor signal (R^2=0.73), probably because of direct temperature dependence of the sensor, and indirect dependence (temperature being a reasonable proxy for both NO_2 and O_3 concentrations). Also the correlation with humidity is very strong (R^2=0.73). The measured S_{WE} signal can best be explained as a linear combination of NO_2, O_3, T, and RH together, resulting in a correlation of 0.98 (R^2=0.96).

The S_{AE} signal is practically insensitive to NO_2. This suggests that a combination of S_{WE} and S_{AE} is more sensitive to NO_2 and less to the other interfering factors, as intended by the manufacturer.

4.2 NO_2 calibration models

For NO_2 measurements, the~~The~~ sensor manufacturer suggests to correct both Working Electrode and Auxiliary Electrode for a zero-offset with $S_{\mathrm{WE,0}}$ and $S_{\mathrm{AE,0}}$ respectively. Then a sensitivity constant s is applied to convert from mV to ppb NO_2:

$$\mathrm{NO}_2[\mathrm{ppb}] = \frac{(S_{\mathrm{WE}} - S_{\mathrm{WE,0}}) - (S_{\mathrm{AE}} - S_{\mathrm{AE,0}})}{s} \tag{2}$$

In practice, the factory-supplied constants $S_{\mathrm{WE,0}}$, $S_{\mathrm{AE,0}}$, and s do not result in realistic values of NO_2, see e.g. Cross et al. (2017). As an alternative, we propose a linear combination of signal S_{WE} and S_{AE} (calibration model A):

$$\mathrm{NO}_2[\mu\mathrm{g\ m}^{-3}] = c_0 + c_1 S_{\mathrm{WE}} + c_2 S_{\mathrm{AE}} \tag{3}$$

~~with t~~The coefficients c_1 and c_2 ~~to be~~are determined with data from the calibration period using ordinary least squares (OLS). ~~Table 1 shows the fit results and the corresponding correlation with true NO_2 signal.~~ As can be seen from the fit results in Table 1, within the batch of sensors there is a large variability of direct sensitivity to ambient NO_2.

During the calibration period, hourly ozone values (also taken from the Vondelpark station) happened to be a good proxy for the ambient NO_2 concentration: $NO_2(t) = 44.6 – 0.40·O_3(t)$ in [µg m^{-3}], with R^2 of 0.49~~ (see Figure 6)~~.

When compared with Table 1, one can see that direct sensor readings from a fair part of the sensors cannot outperform this result. To improve the results we use additional measurements and their statistical relation to NO_2. We fit different calibration models with multiple linear regression (using OLS). The calibration models which were tested are listed in Table 2.

Temperature and RH are taken from the DHT22 sensor. Note that ~~T~~there is no need to calibrate the individual T and RH sensor ~~values~~ signals beforehand; the calibration coefficients for NO_2 are determined for the specific set of all sensors in the box. However, this means that if an individual sensor is replaced, new calibration parameters for the sensor box have to be derived.

4.3 Calibration results

A complete overview of ~~fit results~~the regression coefficients and their error estimates for all models can be found in the supplement. The sign of the calibration parameters can be easily understood. As the electrochemical NO_2 sensor loses sensitivity at higher temperatures (see the negative slope in Figure 7(b) for temperatures below 30°C), coefficients c_3 are positive to compensate for this effect. The additional sensor response due to cross-sensitivity with ozone is compensated by negative values for c_5.

~~From the fit results we see that Model C (including RH) performs better than Model A, but model B (including T) outperforms model C. Model D (including both RH and T) only marginally improves the results of Model B. This can be understood from the strong sensor dependence on temperature directly, and indirectly on temperature as a proxy for ozone. The better performance of model C with respect to model A can be explained by RH being a reasonably proxy for temperature. 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).~~ 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 marginally improved. 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-110 (2013) and Pang et al. (2016).

The best calibration results (i.e. R^2 values closer to 1) are obtained by including ozone (Model E). The ozone values were obtained from the GGD Vondelpark station, as the sensor devices do not measure ozone themselves.

As local ozone measurements were only available during the calibration periods, we used Model D for the Urban AirQ campaign, i.e. generating an NO_2 value based on a linear combination of S_{WE}, S_{AE}, T, and RH. The regression analysis of Model D and correlation with the NO_2 ground truth can be found in Table 3.

The two worst performing sensor ~~boxes~~ devices (~~14560051~~SD02 and ~~1184206~~SD01) contain the older NO2-B42F sensor. The newer NO2-B43F model is designed to have higher sensitivity to NO2 and less interference of ozone. The old sensor model has indeed smaller coefficients for S_{WE} and larger correction terms for ozone (see the c_1 and c_5 coefficients of model E in the Supplement). This, however, can also be related to their longer operating time, as both sensors have been used in previous experiments for more than a year. ~~It is not clear if their poor performance can be attributed to the different sensor~~

~~model, or to their longer operating time (both sensors have been used in previous experiments for more than a year).~~ Again, one can see that even within the same batch of sensors, there is a significant spread in performance, around a median value for R^2 of 0.83. Figure ~~7~~6 shows the results for the different calibration models for ~~an~~ the average performing sensor SD15. The time series in Figure 6~~7~~(b) shows clearly how the performance of a typical sensor device improves when temperature and humidity are included in the calibration analysis. The adjusted R^2, which corrects R^2 for the number of explanatory variables, increases from 0.29 to 0.82. Note that R^2_{adj} is only slightly smaller than R^2, as the number of observations (n≈150) is relatively high compared to the number of regression variables (k=2…5).

4.4 Dependency on temperature

Calibrated data without temperature filter ~~Calibrated, but uncorrected, data~~ show occasionally strong negative values, see Figure ~~8~~ 7 below. These negative peaks coincide with internal sensor temperatures exceeding 30 °C. This behavior can be explained from the dependency of the electrochemical sensor on temperature becoming non-linear, see Figure ~~8~~7(b): the sensitivity of the NO_2 sensor decreases linearly with temperature up to around 30 degrees, while above 40 degrees the sensor gains sensitivity with rising temperatures. In these~~is~~ regimes, the response of the sensor cannot be described well with our multilinear regression approach. As temperatures during the measurement period only rose occasionally above 30 °C, we decided to filter these measurements out.

4.5 Startup time

When a sensor device is switched on for service, the electrochemical cell must be stabilized by the potentiostatic circuit which can take a few hours due to the high capacitance of the working electrode (AAN-105, 2009). 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., 2017).~~When the sensors are switched on after an unused period they need time to stabilize. Figure 9 give some examples of 4 sensors which are switched on at their campaign sites after being offline for a couple of days.~~ The startup-effect is translated by the calibration model as a strong positive NO_2 peak, which should be filtered out. From our sensor data we estimate a stabilization time of 4 hours.~~After 4 hours most sensors are sufficiently stabilized.~~ Note that this startup effect should not be confused with the response time, which is determined to be less than 2 minutes in Mead et al. (2013) and Spinelle et al. (2015a).

4.6 Predictivity, ~~Sensor~~ sensor drift~~t, aging~~, and uncertainty estimation

~~Almost all electrochemical sensors have some degree of drift because of aging and poisoning (Di Carlo et al., 2011; Hierlemann and Gutierrez-Osuna, 2008). This becomes a serious complication when the drift is in the order of the strength of the signal of interest. The idea of keeping sensor 55303 next to the reference station during the whole campaign was to study sensor degradation in more detail. Unfortunately, the sensor was removed temporarily from 10 to 14 July for service,~~

~~which introduced a sudden and unexplained offset in measurements. By introducing a second calibration period after the measurement campaign, we have another possibility to assess the stability of the sensors, and calibrate the measurements in hindsight. All sensors were brought back to the GGD station at the Vondelpark. In Figure 10, the sensor signals (calibrated with coefficients from the first calibration period) are compared to the official station measurements. As can be seen in Figure 10(b), most sensors have been drifting in the intermediate two-month period. Note that part of the drift could also be partly related to the aging of the DHT22 temperature and RH sensor.~~

Almost all electrochemical sensors have some degree of drift because of aging and poisoning (Di Carlo et al., 2011; Hierlemann and Gutierrez-Osuna, 2008). This becomes a serious complication when the drift is in the order of the strength of the signal of interest. The idea of keeping sensor SD03 next to the reference station during the whole campaign was to study sensor degradation in more detail. Unfortunately, the sensor was removed temporarily from 10 to 14 July for service, when it was decided to add a PM module to the device. The increased energy dissipation after the modification (the Shinyei PPD42NS module uses a heater resistor to force a convective flow of sampling air) caused an increase of the internal device temperature by 2.5°C on average. This sudden jump in temperature disrupted the reference time series.

Instead, to assess the short-term stability of the calibration model, we use the first 60% of the measurements from the calibration period (2-7 June) to derive the regression coefficients, and predict the NO_2 values for the remaining 40% (8-10 June), see Table 4. The average RMSE increases from 6.5 to 7.0 µg m^3 when the regression is used for prediction.

We assess the long-term stability of the sensors with a second calibration period after measurement campaign, again at the Vondelpark calibration site. As can be seen from the distribution of the residuals in Figure 8, most sensors drift significantly in the intermediate two-month period. We describe this degradation effect as a bias b between the mean of the hourly estimated NO_2 values $\hat{x}_i$ and the mean of the hourly true NO_2 x_i during the calibration period:

$$b = \frac{1}{N}\sum_{i=1}^{N}\hat{x}_i - \frac{1}{N}\sum_{i=1}^{N}x_i \qquad (4)$$

and the root-mean-square error (RMSE) of the difference between the bias corrected calibrated measurement and the ground truth. The latter is the same as the standard deviation of the residuals (SDR) $\hat{x}_i - x_i$:

$$\mathrm{SDR} = \sqrt{\frac{1}{N}\sum_{i}\left((\hat{x}_i - b) - x_i\right)^2} \qquad (5)$$

As can be seen in Table ~~4~~ 5 below, the bias is mostly positive. Note that sensor ~~54911~~SD16 and ~~1184206~~SD01 had a limited uptime in the second period, which makes their bias and RMS calculation not very representative.

The strongest bias after two months is found for ~~14560051~~SD02 and ~~1184206~~SD01, both of model NO2-B42F and having been used in others experiments for more than one year. These sensors have also the largest RMSE in the first calibration period (see also Table 3), another indication of their poor performance. The range in RMSE of the remaining sensors is 4.5 – 7.2 µg m^{-3} for the first period. The bias corrected RMSE increases to 5.3 – 9.3 µg m^{-3} for the second period. The latter is a more conservative yet more realistic estimation of the precision of the NO_2 estimates, as they are based on measurements which were not used for calibration. Based on our results listed in the last columns of Table 4 and 5, we take 7 µg m^{-3} as a typical uncertainty for the estimated NO_2 values.

The increase of SDR is also due to a loss of sensitivity over time. The aging of the sensors can be further investigated by recalibrating the devices, i.e. determining the coefficients of regression model D, using the data of the second calibration period (see the Supplemental Material). All calibration coefficients of S_{WE} (the only component which has direct sensitivity to NO_2) decrease in value, showing that all sensors suffer from sensitivity loss to NO_2. This results in lower R^2 values, although the performance loss is partly compensated by the other components in the regression. The older Alphasense models NO2-B42F suffer the largest sensitivity loss, which (although the regression tries to compensate with an increased temperature dependence) result in the worst performance loss in terms of R^2.

~~The panels in Figure 11 show how the calibration coefficients change after two months of deployment. Having in mind that the S_{WE} signal is the only component which has direct sensitivity to NO_2, one can see in Figure 11(b) (all dots below the $y=x$ line) that all sensors suffer from sensitivity loss to NO_2. This results in lower R^2 values in Figure 11(f), although the performance loss is partly compensated by the other components in the multivariate linear regression.~~

~~The older Alphasense models NO2-B42F (red dots in Figure 11(b)) are the most insensitive to NO_2, and have the largest sensitivity loss, which the regression tries to compensate with an increased temperature dependence (Figure 11(d)), although this can not avoid that they have the worst performance and the worst performance loss in terms of R^2.~~

4.7 Weighted calibration

Taking 18 µg m^{-3} as a typical NO_2 concentration in an urban environment (Figure 4), the sensor drift as listed in Table ~~4~~ 5 is a significant error component, even after a two month period. It is impossible to predict the progressing bias for an individual sensor. However, using the second calibration period we can compensate for signal drift in hindsight. If $\hat{x}_1(t)$ represents the estimated NO_2 value at time t based on the first calibration period (starting at t_1), and $\hat{x}_2(t)$ the estimated NO_2 value based on the second calibration period (ending at t_2), the we take for intermediate times $t_1 \le t \le t_2$ a weighted average of both calibrations:

$$\hat{x}(t) = \left(1 - f(t)\right)\hat{x}_1(t) + f(t)\hat{x}_2(t) \qquad (6)$$

Assuming that the sensor degradation is linear in time we select

$$f(t) = (t - t_1)/(t_2 - t_1) \qquad (7)$$

such that $f(t_1)=0$ and $f(t_2)=1$.

4.8 Validation against an independent reference station~~Oude Schans station~~

Citizen science can be unpredictable, and we were fortunate that ~~From 14 June to 16 August,~~ sensor ~~54200~~SD04 was ~~placed at~~handed over to an Urban AirQ participant living at Korte Koningsstraat (ground floor~~/street side~~), which happens to be 120m from another GGD station at Oude Schans~~, also classified as an urban background station~~(see Figure 1). The Korte Koningsstraat ~~characterizes as~~ is a side street~~,~~ away from traffic arteries, whereas Oude Schans also classifies as an urban background location. The proximity to a reference station enables us to perform an independent validation of the sensor measurements, as the calibration of the sensor is based on side-by-side measurements with Vondelpark station, at 3 km distance. As can be seen from ~~Table 5~~Figure 9, the sensor readings agree very well with the official measurements. ~~Figure 12(a) and 12(b) show the time series and the scatter plot.~~

Using the weighted calibration explained in the previous section, the measurement bias largely disappears (Table 6). The RMSE (5.3 µg m^{-3}) is comparable to the RMSE found during the calibration period~~ (see Table 4)~~. The results give confidence that our calibration method ~~is independent of~~remains valid for similar urban locations, and that our assumption of sensor degradation being linear in time is acceptable.

5 Discussion

The Alphasense NO2-B4 sensor is used in many low-cost air quality applications for measuring ambient NO_2. As all electrochemical NO_2 sensors, the Alphasense NO2-B4 sensor is not very selective to the target gas. The sensor response can best be explained as a linear combination of NO_2, O_3, temperature and relative humidity signals ($R^2 \approx 0.9$).

As a consequence, a linear combination of the Working Electrode and the Auxiliary Electrode alone give poor indication of ambient NO_2 concentrations. The accuracy varies greatly between different sensors (R^2 between 0.3 and 0.7). For the Urban AirQ campaign, temperature and relative humidity were included in a multilinear regression approach. The results improve significantly with R^2 values typically around 0.8. This corresponds well with the findings of Jiao et al. (2016), who find an adjusted R^2=0.82 for the best performing electrochemical NO_2 sensor in their evaluation, when including T and RH.

Best results are obtained by also including ozone measurements in the calibration model: R^2 increases to 0.9. Spinelle et al. (2015b) used a similar regression and found R^2 ranging from 0.35 to 0.77 for 4 electrochemical NO_2 sensors during a two-week calibration period, but dropping to 0.03—0.08 when applied to a successive 5-month validation period. Low NO_2 values at their semi-rural site partly explains this poor performance, but most likely also unaccounted effects such as changing sensor sensitivity and signal drift.

The sensor devices were tested in an Amsterdam urban background in summertime, with NO_2 values ranging from 3 µg m^{-3} to 78 µg m^{-3}, and median values around 15 µg m^{-3}. During the 3-month period most sensors show loss of sensitivity and significant drift, ranging from -9 to 21 µg m^{-3}. After bias correction we found a typical value for the accuracy of the NO_2 measurements of 7 µg m^{-3}.

This error consists of several components. The reference measurements by the NO/NO_x analysers have an estimated hourly error of 3.65% (certified validation at a 200 μg m^{-3} NO_2 concentration), which would contribute to 0.5 μg m^{-3} under typical conditions. The low-cost DHT22 sensor has a reported error of 0.5 °C for temperature and 2–5% for RH. For a single measurement, this would contribute to a propagated regression~~n~~ error of approximately 1 μg m^{-3} and 0.5 μg m^{-3}, respectively ~~(Figure 11(d) and 11€)~~. It should be noted, however, that binning minute-based measurements to hourly averages removes large part of the variability, while determining the best fitting regression model for each sensor device removes large part of the remaining systematical biases. The largest part of the error term is therefore introduced by the linear regression model itself, which does not include all interfering species or meteorological quantities, and is not able to describe non-linear dependencies of its variables. One should therefore be careful to ~~extrapolating~~ extrapolate the calibration model for conditions different than the calibration period.

The validation results from Section 4.8 show that the calibration holds well for urban locations with similar NO_2/O_3 ratios. Neglecting O_3 as regression parameter, however, will introduce a bias at locations with different NO_2/O_3 ratios found e.g. 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 relation can best be described by $[O_3]_{VDS} = 0.87\ [O_3]_{NDD} + 0.85$ (with 0.93 correlation), which means that ozone levels at the street station are typically 13% lower, due to titration of O_3 with NO. Due to the sensor's cross-sensitivity for ozone, larger values must be subtracted from its signal when the ozone concentration increases. This explains the negative sign of the ozone coefficient c_5 of model E (see Supplement). 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. Using typical values c_5=-0.3 and $[O_3]$=60 μg/m^3 (75th percentile of the distribution during the measurement camping, according to Figure 4) we estimate the underestimation of NO_2 at street side as 0.3 × 13% × 60 = 2.3 μg/m^3.

The found sensor accuracy after ~~two calibrations and corrections~~ weighted calibration is good enough to provide some complementary spatial information ~~to complement official measurements by providing additional information~~ on local air quality between reference stations~~, and detect unexpected hot spots (or low values) of urban NO_2~~.~~.~~ When looking at the difference between Vondelpark station and Oude Schans station (both classified as city background stations) in the period June-August 2016, 22% of the hourly measurements differ more than 7 μg m^{-3}, and 6% of the hourly measurements differ more than 14 μg m^{-3}. These differences increase further when considering road side stations. From this perspective, even sensor devices with an accuracy around 7 μg m^{-3} can contribute to an improved understanding of spatial patterns. However, it must be further investigated if the ~~regression~~ calibration method used here would provide realistic estimates for peak values (such as the EU hourly limit value, 200 μg m^{-3}).

6 Conclusions and outlook

In this study, we examined low-cost electrochemical air quality sensors for citizen urban air quality monitoring. In other words, we evaluated an imperfect air quality sensor in an imperfect scientific experiment. In general, we found that low-cost electrochemical sensors have the potential to complement official environmental monitoring data to help answer questions from the public, which usually cannot be fully answered from official data alone. To reach the potential, however, proper measurement set-up, calibration and recalibration, and data analysis should be guaranteed.

The current generation of low-cost NO_2 sensors has some serious issues which trouble straightforward application. To make electrochemical NO_2 sensor measurements accurate, careful filtering of the raw data is necessary. There is a strong spread in sensor performance, even if the sensors come from the same batch, which make individual calibration essential. A practical calibration method is measuring side-by-side to an air monitoring station. The accuracy of the measurements can be improved by including temperature and humidity measurements from other low-cost sensors in a multilinear regression approach. ~~A practical calibration method is measuring side-by-side to an air monitoring station.~~ It is worth noting that more advanced calibration algorithms such as by Cross et al. (2017) and Mueller et al. (2017) could give better results, but this is not the focus of this paper. It is hard to quantify an optimal length of a calibration period without having a proper understanding of the sensor degradation rate beforehand. ~~This~~ The measurement period should be ~~as long as possible (but~~ 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.

Startup time of sensors is estimated 4 hours. To avoid nonlinear response of the electrochemical sensor at elevated temperatures, we filter out measurements above 30 °C. This is not a serious restriction for applicability in moderate climates such as in the Netherlands, provided that the sensor is protected from direct sunlight. However, for warmer regions or during heat waves this may reduce the data stream considerably, unless the temperature dependencies are better captured by more advanced regression models.

The calibration ~~coefficients~~ seems to be location independent, as long as the NO_2/O_3 ratio is comparable. Application at a street side is likely to introduce a small positive bias. Calibration~~as independent validation in the proximity of a second monitoring station suggests. However, calibration~~ coefficients are not constant in time. During the 3-month period most sensors suffer from significant sensitivity loss and drift. ~~The standard deviation of the random error is estimated 7 μg m^{-3} for a typical sensor.~~ The strongest drift and largest uncertainty are found for the older NO2-B42F sensors. It remains unclear if the ~~poorer~~ worse performance is related to the sensor model or the longer usage in field experiments.

~~Individual sensor drift can be compensated in hindsight by taking a weighted average of the calibration coefficients determined before and after the campaign, assuming that the sensor degradation is linear in time.~~ The sensor degradation troubles practical applications in operational urban networks. ~~makes it necessary to think about s~~Smart re-calibration programs are essential: bringing back sensors to a calibration facility on a regular basis, or recalibrating on the spot by a travelling reference instrument. New data driven techniques, such as Bayesian networks (e.g. Xiang et al., 2016), might offer a solution for this problem. ~~when one wants to use electrochemical sensors operationally in a low-cost urban networks. More research is needed to gain better insight of how sensors age in field applications. This will provide better calibration strategies which improve data quality.~~

On the hardware side we recommend to include active ventilation to guarantee a constant air flow over the gas sensor and suppresses unwanted internal temperature changes due to heating of electronical components. To improve the NO_2 measurements further~~To further improve accuracy of electrochemical NO_2 measurements in low-cost sensor devices~~ we recommend to include an additional low-cost ozone sensor, e.g. Ox-B431 by Alphasense. 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 better resolve cross-sensitivity issues. Even imperfect ozone measurements will improve the NO_2 estimation, as large part of the sensor's cross-dependency issues are solved by the linear regression approach.~~ The RH sensor signal should be used more cleverly to detect and filter for sudden changes in relative humidity. Adding a local data logger is also recommended, to be able to recover data for periods when the WiFi connection to the central database is lost.

~~The necessity for recalibration troubles practical applications in operational urban networks. Sensors must be brought back to a calibration facility on a regular basis, or must be recalibrated on the spot by a travelling reference instrument. New data driven techniques, such as Bayesian networks (e.g. Xiang et al., 2016), might offer a solution for this problem.~~

Data availability

A complete overview of fit results for all models can be found in the supplement. The hourly Urban AirQ sensor data, calibrated in hindsight by interpolating the calibration in time between two calibration periods, can be downloaded at https://github.com/waagsociety/making-sensor.

Competing interests.The authors declare that they have no conflict of interest.

Acknowledgements.The Urban AirQ project was partly funded by a 2016 Stimulus Grant from AMS (Advanced Metropolitan Solutions). The project is also part of Making Sense, funded by European Union's Horizon 2020 research and innovation programme. Qijun Jiang is supported by the China Scholarship Council for his Ph.D. research. The authors would like to thank Emma Pareschi from Waag Society who was responsible for the hardware development.

References

AAN 110, Alphasense Application Note on Environmental Changes: Temperature, Pressure, Humidity, 2013, http://www.alphasense.com/WEB1213/wp-content/uploads/2013/07/AAN_110.pdf

AAN 105-03, Alphasense Application Note: Designing a potentiostatic circuit, March 2009, http://www.alphasense.com/WEB1213/wp-content/uploads/2013/07/AAN_105-03.pdf

ADS, Alphasense Data Sheet for NO2-B43F, April 2016, http://www.alphasense.com/WEB1213/wp-content/uploads/2016/07/NO2-B43F.pdf

Borrego, C., Costa, A. M., Ginja, J., Amorim, M., Coutinho, M., Karatzas, K., . . . Penza, M. (2016). Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise. Atmospheric Environment, 147, 246-263. doi: http://dx.doi.org/10.1016/j.atmosenv.2016.09.050

Cape, J.N. (2009): The Use of Passive Diffusion Tubes for Measuring Concentrations of Nitrogen Dioxide in Air, Critical Reviews in Analytical Chemistry, Vol. 39, pp 289-310, Iss. 4, doi: 10.1080/10408340903001375

Cross, E. S., Williams, L. R., Lewis, D. K., Magoon, G. R., Onasch, T. B., Kaminsky, M. L., Worsnop, D. R., and Jayne, J. T. (2017): Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements, Atmos. Meas. Tech., 10, 3575-3588, https://doi.org/10.5194/amt-10-3575-2017.

Di Carlo, S., Falasconi, M., Sanchez, E., Scionti, A., Squillero, G., & Tonda, A. (2011). Increasing pattern recognition accuracy for chemical sensing by evolutionary based drift compensation. Pattern Recognition Letters, 32(13), 1594-1603. doi: http://dx.doi.org/10.1016/j.patrec.2011.05.019

Duvall, R., Long, R., Beaver, M., Kronmiller, K., Wheeler, M., & Szykman, J. (2016). Performance Evaluation and Community Application of Low-Cost Sensors for Ozone and Nitrogen Dioxide. Sensors, 16(10), 1698.

Hierlemann, A., & Gutierrez-Osuna, R. (2008). Higher-Order Chemical Sensing. Chemical Reviews, 108(2), 563-613. doi: 10.1021/cr068116m

Jiang Q., Frank Kresin, Arnold K. Bregt, et al. (2016), “Citizen Sensing for Improved Urban Environmental Monitoring”, Journal of Sensors, vol. 2016, Article ID 5656245, 9 pages, 2016. doi:10.1155/2016/5656245

Jiao, W., G. Hagler, R. Williams, R. Sharpe, R. Brown, D. Garver, R. Judge, M. Caudill, J. Rickard, M. Davis, L.Weinstock, S. Zimmer-Dauphinee, K. Buckley (2016), Community Air Sensor Network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern United States, Atmospheric Measurement Techniques, 9, 11, pp 5281-5292, doi 10.5194/amt-9-5281-2016

Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., . . . Britter, R. (2015). The rise of low-cost sensing for managing air pollution in cities. Environment International, 75, 199-205. doi: http://dx.doi.org/10.1016/j.envint.2014.11.019

Lewis, A. and P. Edwards (2016), “Validate personal air-pollution sensors”, Nature 535, 29–31, doi:10.1038/535029a

Lewis, A. C., Lee, J. D., Edwards, P. M., Shaw, M. D., Evans, M. J., Moller, S. J., . . . White, A. (2016). Evaluating the performance of low cost chemical sensors for air pollution research. Faraday Discussions, 189(0), 85-103. doi: 10.1039/c5fd00201j

Masson, N.; Piedrahita, R.; Hannigan, M. Quantification Method for Electrolytic Sensors in Long-Term Monitoring of Ambient Air Quality (2015). Sensors, 15, 27283-27302.

Mead, M. I., Popoola, O. A. M., Stewart, G. B., Landshoff, P., Calleja, M., Hayes, M., . . . Jones, R. L. (2013). The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. Atmospheric Environment, 70, 186-203. doi: 10.1016/j.atmosenv.2012.11.060

Moltchanov, S., Levy, I., Etzion, Y., Lerner, U., Broday, D. M., & Fishbain, B. (2015). On the feasibility of measuring urban air pollution by wireless distributed sensor networks. Science of The Total Environment, 502, 537-547. doi: http://dx.doi.org/10.1016/j.scitotenv.2014.09.059

Mueller, M., J. Meyer, and C. Hueglin (2017). Design of an ozone and nitrogen dioxide sensor unit and its long-term operation within a sensor network in the city of Zurich, Atmos. Meas. Tech., 10, 3783-3799, https://doi.org/10.5194/amt-10-3783-2017.

Pang, X., Shaw, M. D., Lewis, A. C., Carpenter, L. J., & Batchellier, T. (2016). Electrochemical ozone sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air-quality monitoring. Sensors and Actuators B: Chemical, 240, 829-837. doi: http://dx.doi.org/10.1016/j.snb.2016.09.020

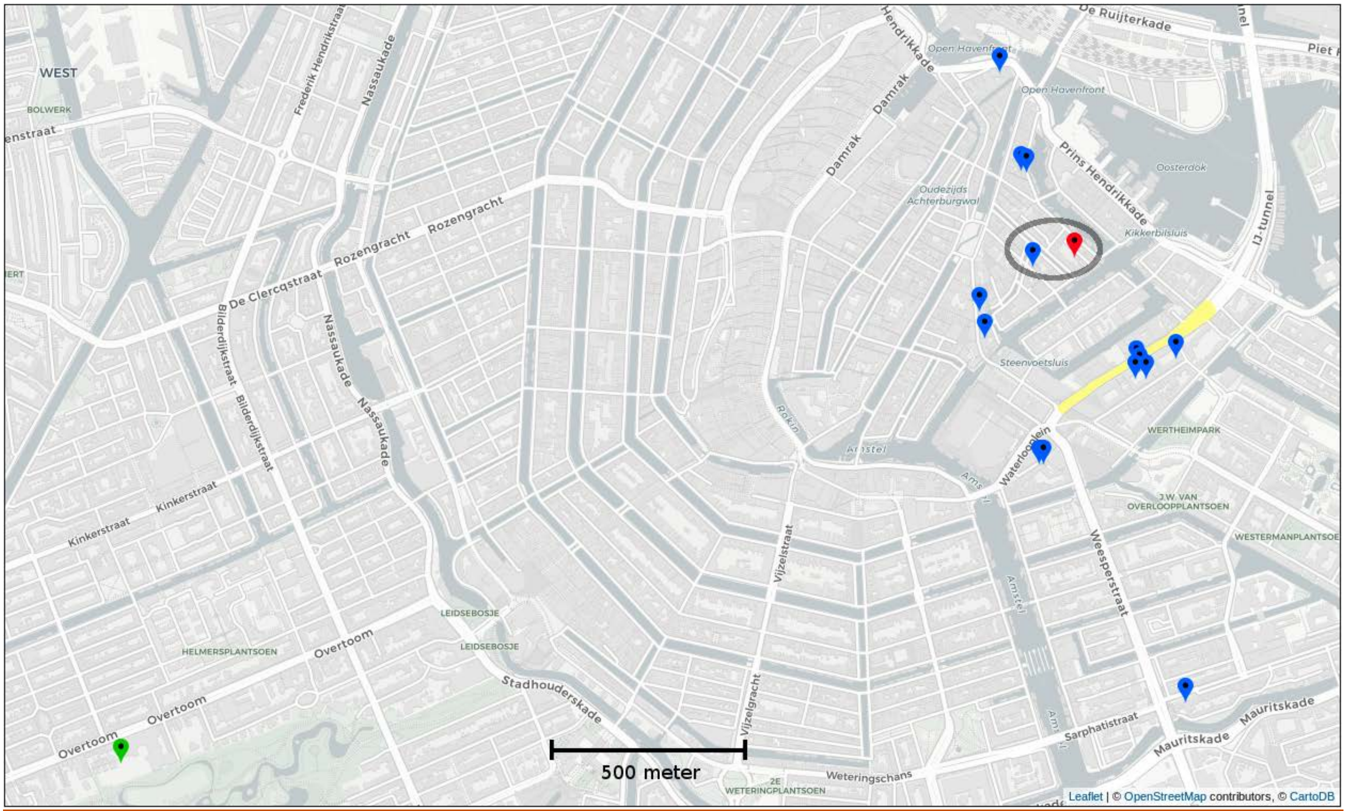
Piedrahita, R., Xiang, Y., Masson, N., Ortega, J., Collier, A., Jiang, Y., Li, K., Dick, R. P., Lv, Q., Hannigan, M., and Shang, L. (2014): The next generation of low-cost personal air quality sensors for quantitative exposure monitoring, Atmos. Meas. Tech., 7, 3325-3336, doi:10.5194/amt-7-3325-2014.

Spinelle, L., M. Gerboles, M. Aleixandre (2015a), EUROSENSORS 2015: Performance evaluation of amperometric sensors for the monitoring of O3 and NO2 in ambient air at ppb level, Procedia Engineering 00 (2015) 000-000

Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M., & Bonavitacola, F. (2015b). Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide. Sensors and Actuators B: Chemical, 215, 249-257. doi: http://dx.doi.org/10.1016/j.snb.2015.03.031

Thompson, J. E. (2016). Crowd-sourced air quality studies: A review of the literature & portable sensors. Trends in Environmental Analytical Chemistry, 11, 23-34. doi: http://dx.doi.org/10.1016/j.teac.2016.06.001

Xiang, Y., Tang, Y., and Zhu, W. (2016). Mobile sensor network noise reduction and recalibration using a Bayesian network, Atmos. Meas. Tech., 9, 347-357, doi:10.5194/amt-9-347-2016.



Leaflet | © OpenStreetMap contributors, © CartoDB

Figure 1 Locations of the sensor devices during the citizen measurement campaign. The green marker indicates the calibration location at GGD Vondelpark. In the circle the location of SD04 and the GGD station at Oude Schans (in red). ~~The red marker indicates the GGD station at Oude Schans. Not shown is the GGD Vondelpark station, 2.5 km in south-west direction.~~The location of Valkenburgerstraat is highlighted in yellow.

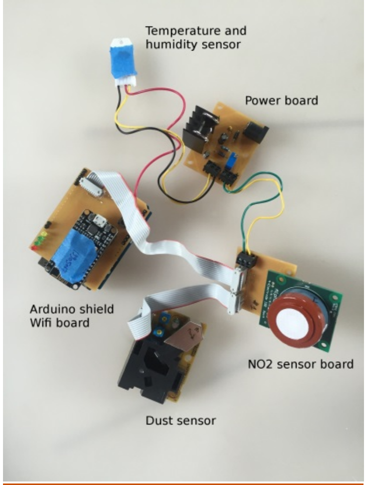
 Temperature and humidity sensor Power board Arduino shield Wifi board NO2 sensor board Dust sensor		

Figure 2 Hardware ~~components~~ modules of a sensor device (left), and the integration in the casing: open (middle) and closed (right). ~~and sensors in their housing (right)~~

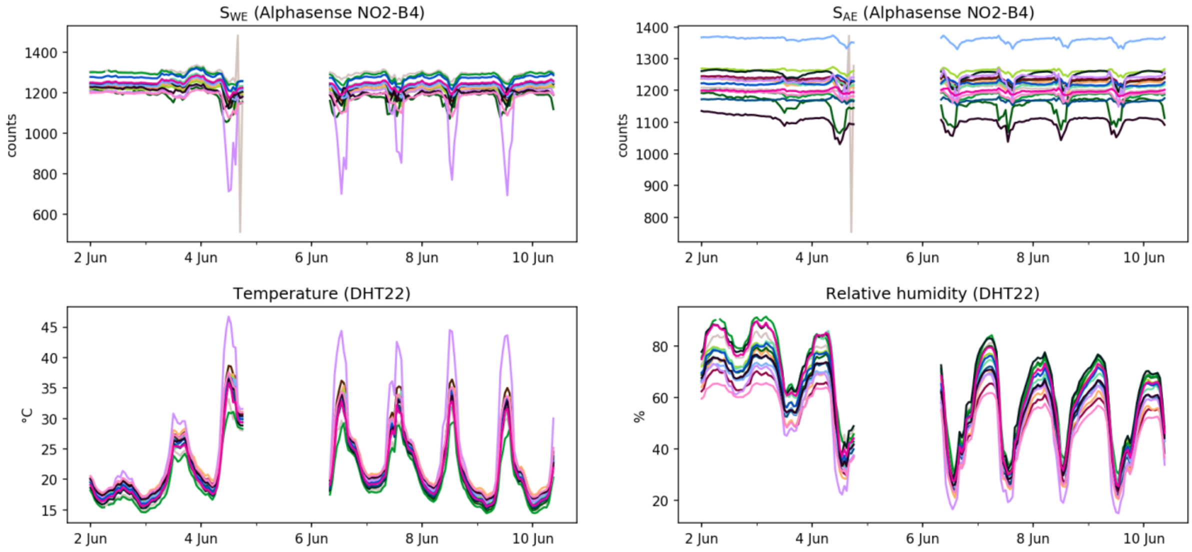


Figure 3 Raw sensor data, unfiltered but hourly averaged, from the 16 sensors during the first calibration period, 2-10 June 2016. The data gap around 5 June is due to a connectivity problem to the central database.

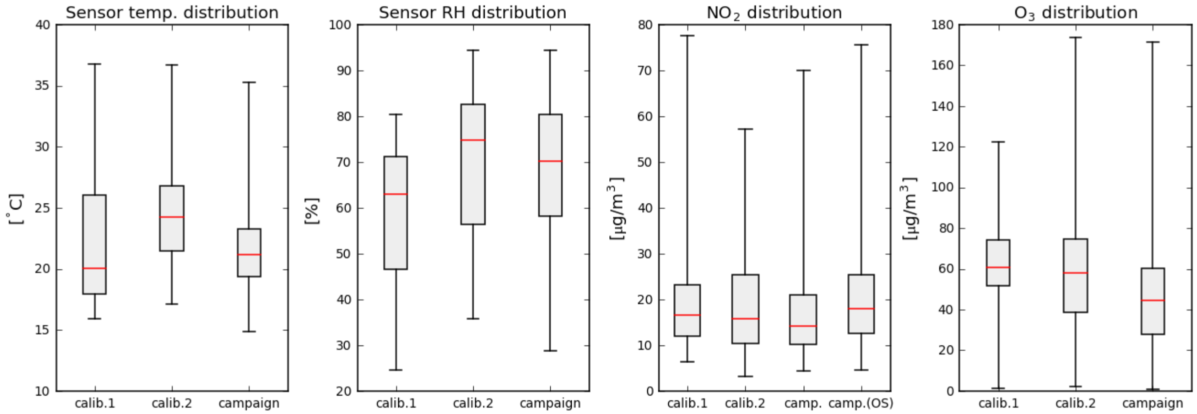


Figure 4 Box whisker diagrams of hourly ambient parameters during the two calibration periods and the measurement campaign. The box edges indicate the 25th – 75th ~~-~~percentile; the whiskers the minimum and maximum values. The median is indicated in red. Temperature and RH are based on the average values of all sensors devices, NO_2 and ozone are taken from the reference station at Vondelpark. For comparison, NO_2 from the reference station at Oude Schans (OS) is also shown.

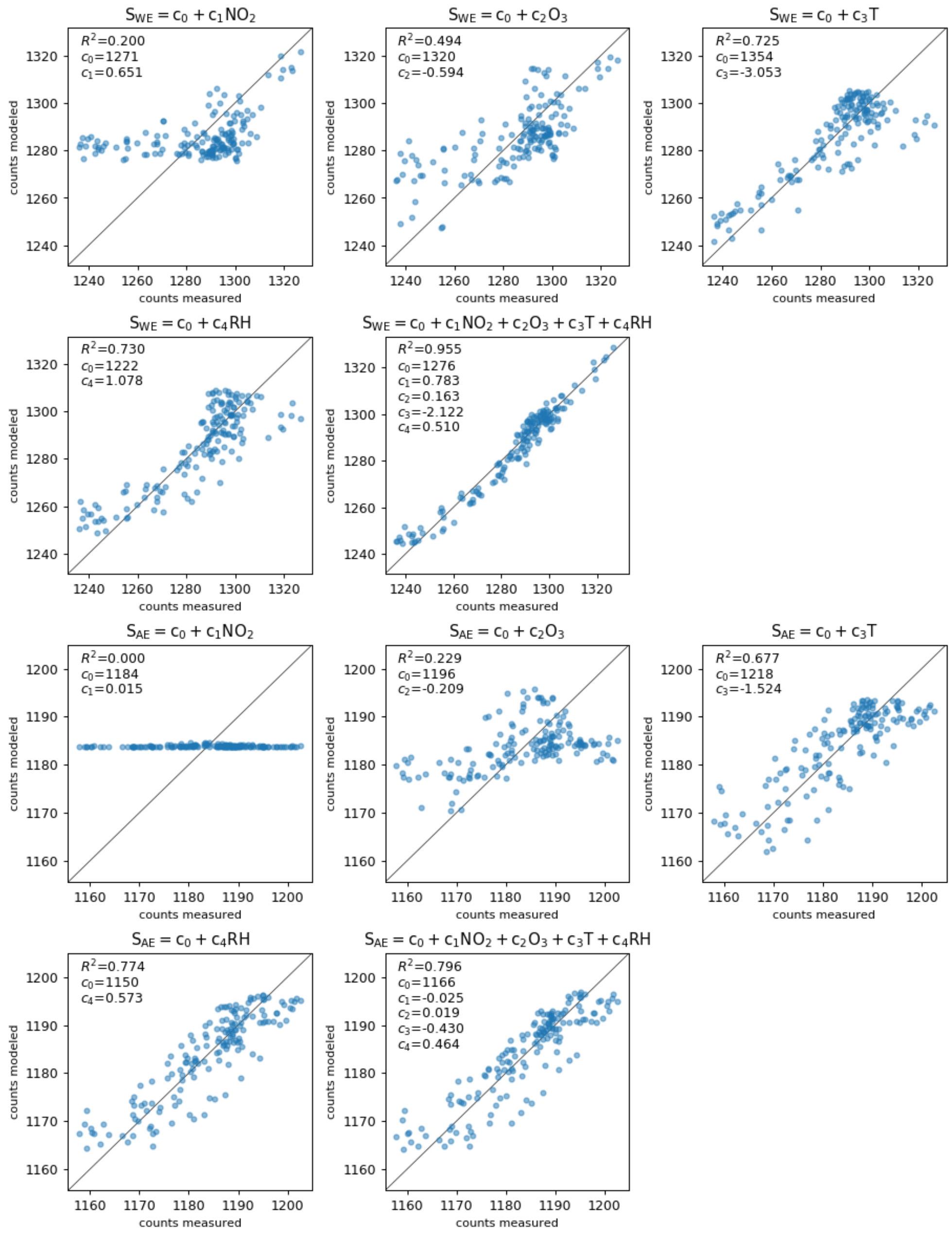


Figure 5 The reading of a typical performing NO2-B43F sensor (~~ID 1185325~~SD10) explained as a linear regression of respectively NO$_2$, O$_3$, T, RH, and all variables. The top two rows show the results for the Working Electrode~~-~~; the bottom two rows for the Auxiliary Electrode. ~~On t~~The ~~axis~~ axes represent the A/D converter counts, which are proportional to the currents generated by the sensor at the corresponding electrode~~can be considered as arbitrary units~~.

~~**Figure 6 Ozone as a proxy of ambient NO$_2$.**~~

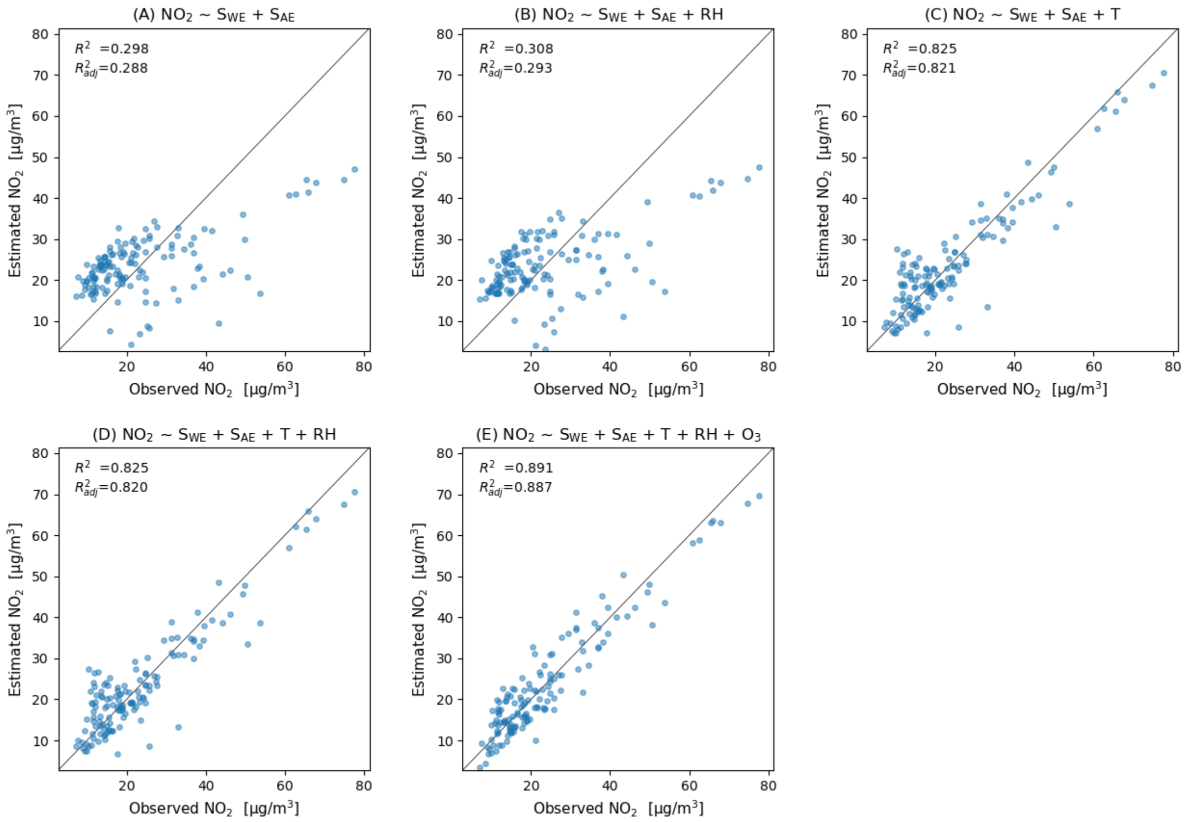


Figure 6~~7~~(a) Calibration model results for an average performing sensor (~~ID 1184838~~SD15). Bottom row shows the recommended calibration by Model D (left), and the results when ozone would be included (right).

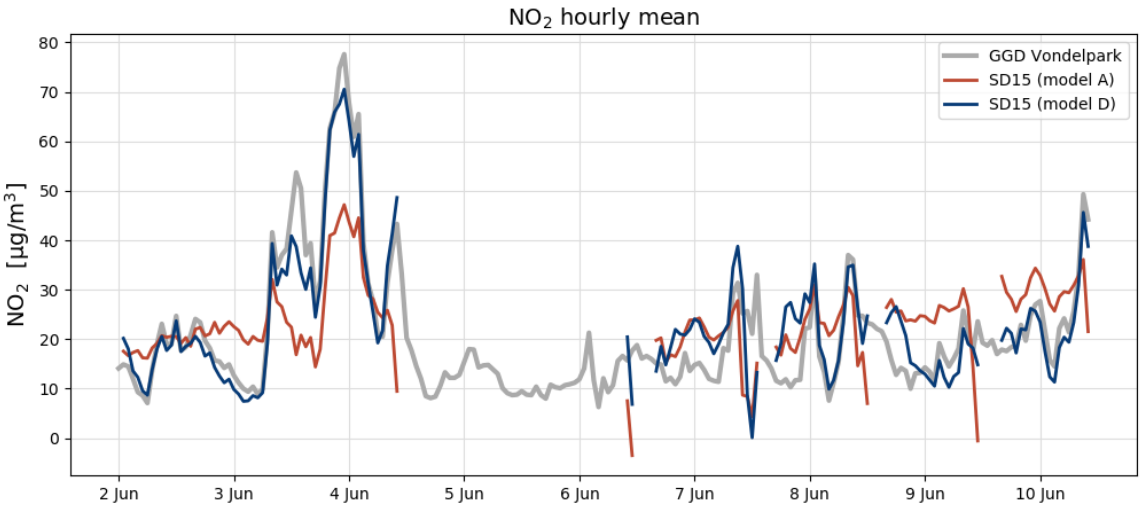


Figure ~~7~~6(b) Time series compared to ground truth with calibration parameters of Model A and D.

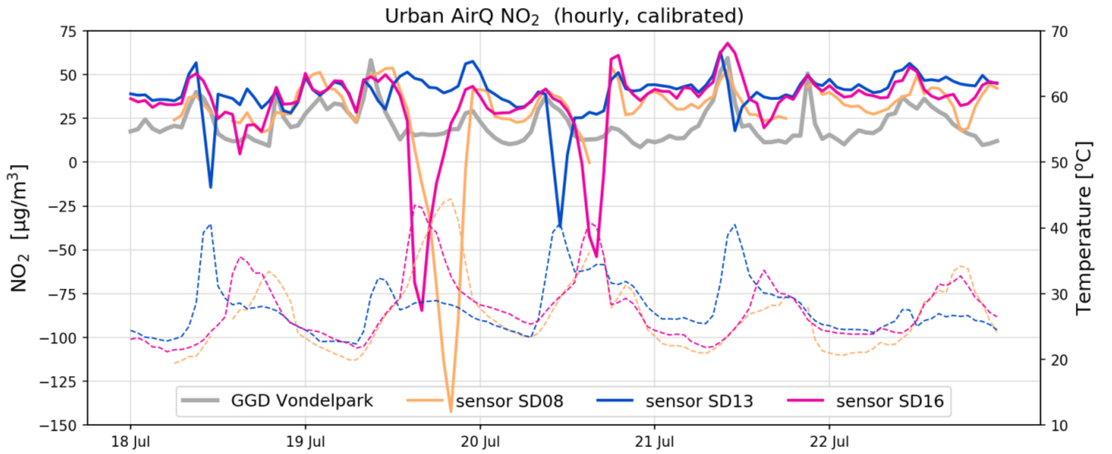


Figure 7~~8~~(a) Examples of negative spikes in the calibrated NO_2 measurements (solid line) due to internal sensor temperatures (dotted line) exceeding 30 °C.

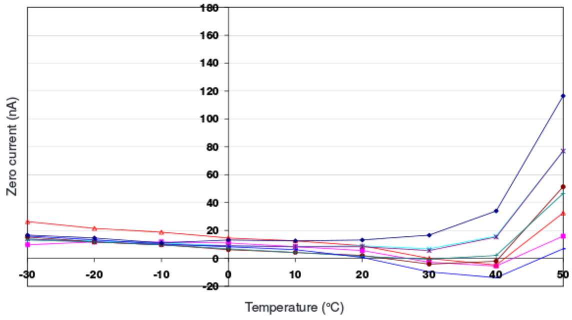


Figure 7~~8~~(b) Variation of zero output of the working electrode caused by changes in temperature for a typical batch of electrochemical sensors. Image taken from Alphasense Data Sheet for NO2-B43F (ADS, 2016).

~~**Figure 9 Examples of sensor startup effects when switched on.**~~

~~**Figure 10(a) Time series of a batch of sensors, calibrated with model D, compared with the reference measurements (grey line).**~~

~~Figure 10(b) Comparison of the time series of the same batch of sensors with the reference measurements (grey line), after two months of operation.~~

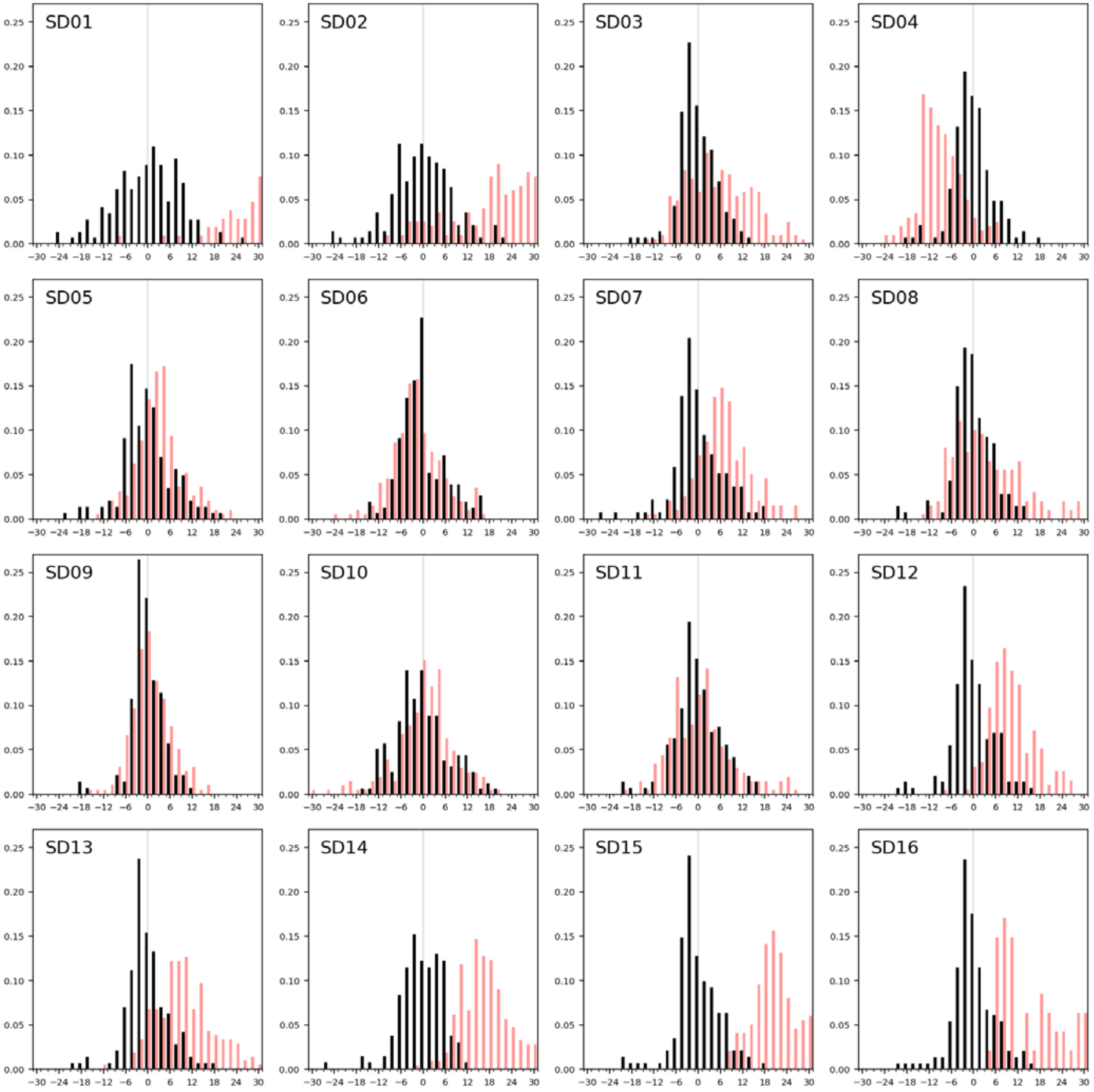


Figure 8 Sensor drift during two months of operation, shown as the distribution of residuals with the reference measurements during the first calibration period (black bars) and during the second period (red bars).

~~Figure 11 Change in calibration coefficients of model D from the first calibration period (horizontal axis) when recalibrating after two months of deployment (vertical axis). The red dots correspond to sensor devices containing the Alphasense NO2-B42F.~~

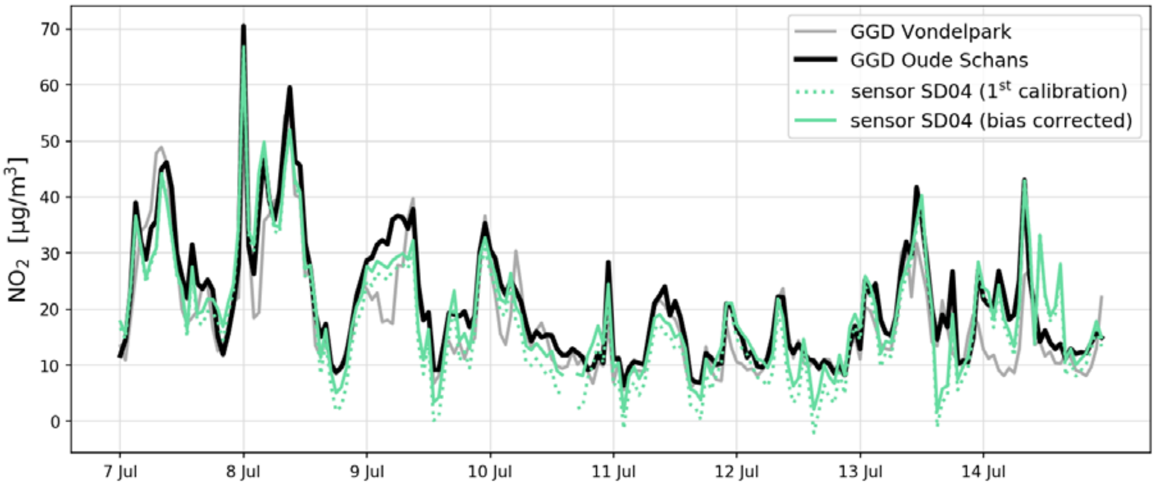


Figure ~~12~~9(a) Comparison of sensor ~~54200~~SD04 NO$_2$ time series with the nearby Oude Schans station (8-day snap shot), and the effect of bias correction. For comparison, measurements of Vondelpark station are also shown.

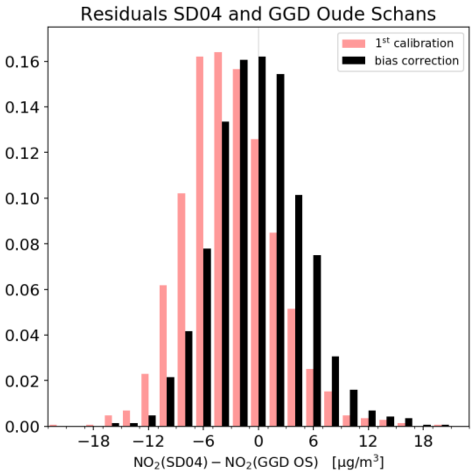


Figure ~~12~~9(b) ~~Scatterplot of sensor 54200 against Oude Schans station NO2 measurements during the campaign period.~~Distribution of residuals of NO$_2$ measurements between sensor SD04 and Oude Schans station during the campaign period, with and without bias correction.

Table 1 Fit results for regression model A. Older NO2-B42F sensors highlighted in grey.~~, sorted from best to worst sensor~~

Sensor ID	c_0	c_1 (S_{WE})	c_2 (S_{AE})	R^2
SD01	455.4	0.6977	-1.0835	0.47
SD02	355.9	0.8862	-1.2633	0.62
SD03	-228.6	1.0877	-0.8029	0.72
SD04	-968.2	0.9138	-0.1237	0.69
SD05	-155.1	0.8368	-0.6841	0.48
SD06	-141.9	0.6136	-0.5241	0.44
SD07	-576.4	0.9615	-0.4811	0.57
SD08	231.4	1.0802	-1.2514	0.68
SD09	100.5	0.8669	-0.8952	0.56
SD10	342.0	0.8221	-1.1629	0.50
SD11	338.4	0.9823	-1.2246	0.61
SD12	-375.2	0.7775	-0.4837	0.54
SD13	-1703.4	0.8218	0.5544	0.60
SD14	162.6	0.8156	-0.9075	0.46
SD15	1211.2	0.9008	-1.8984	0.30
SD16	-594.3	0.8007	-0.3192	0.49

Table 2 Regression models for NO_2

Model A	$NO_2 = c_0 + c_1 \cdot S_{WE} + c_2 \cdot S_{AE}$	Linear combination of Working Electrode and Auxiliary Electrode
~~Model B~~	~~$NO_2 = c_0 + c_1 \cdot S_{WE} + c_2 \cdot S_{AE} + c_3 \cdot T$~~	~~Temperature correction~~
Model B~~C~~	$NO_2 = c_0 + c_1 \cdot S_{WE} + c_2 \cdot S_{AE} + c_4 \cdot RH$	Relative humidity correction
Model C	$NO_2 = c_0 + c_1 \cdot S_{WE} + c_2 \cdot S_{AE} + c_3 \cdot T$	Temperature correction
Model D	$NO_2 = c_0 + c_1 \cdot S_{WE} + c_2 \cdot S_{AE} + c_3 \cdot T + c_4 \cdot RH$	Temperature and RH correction
Model E	$NO_2 = c_0 + c_1 \cdot S_{WE} + c_2 \cdot S_{AE} + c_3 \cdot T + c_4 \cdot RH + c_5 \cdot O_3$	~~Adding also c~~Correction for temperature, RH, and ozone cross-sensitivity

Table 3 Fit results for regression model D~~, ordered from best to worst sensor~~. Older NO2-B42F sensors highlighted in grey.

Sensor ID	c_0	c_1 (S_{WE})	c_2 (S_{AE})	c_3 (T)	c_4 (RH)	R^2
SD01	790.9	0.8707	-1.5645	-0.5051	0.4513	0.62
SD02	589.2	0.8618	-1.4742	0.2142	0.4204	0.67
SD03	-1272.1	1.2045	-0.1492	1.2690	-0.2944	0.87
SD04	-1613.3	1.1499	0.1818	0.3200	-0.4442	0.85
SD05	-1623.1	1.1235	0.2088	1.7161	-0.4430	0.75
SD06	-824.8	1.1850	-0.5839	1.6737	-0.3069	0.81
SD07	-1217.6	1.1305	-0.1642	1.9435	0.0000	0.79
SD08	-1129.7	1.1835	-0.2705	2.2559	-0.2704	0.86
SD09	-586.3	1.1794	-0.6738	2.0415	-0.2192	0.90
SD10	-1152.7	1.1668	-0.3120	2.9112	-0.2147	0.72
SD11	-1109.8	1.1055	-0.2339	3.3191	-0.1693	0.81
SD12	-1074.9	1.0961	-0.2346	1.4954	-0.2799	0.84
SD13	-1074.6	1.1294	-0.3058	1.8671	-0.1561	0.83
SD14	8.1	1.1860	-1.1889	2.5401	0.0268	0.84
SD15	-104.5	1.8111	-1.7939	4.8373	0.0596	0.83
SD16	-1215.5	1.2551	-0.3038	2.1742	-0.1333	0.84

Table 4 Descriptive and short-term predictive error of model D in μg m^{-3}

	2-7 June (descriptive)		8-10 June (predictive)	
Sensor ID	**Uptime**	**RMSE**	**Uptime**	**RMSE**
SD01	92h	9.25	54h	9.31
SD02	89h	7.95	53h	13.74
SD03	88h	5.58	53h	4.37
SD04	90h	6.00	54h	4.94
SD05	90h	7.62	53h	8.75
SD06	97h	6.36	57h	5.57
SD07	85h	7.09	52h	6.26
SD08	88h	5.95	52h	6.59
SD09	88h	4.94	52h	3.69
SD10	99h	7.44	59h	8.09
SD11	91h	6.78	53h	5.42
SD12	93h	6.08	52h	5.07
SD13	89h	6.25	54h	5.31
SD14	83h	3.96	48h	14.61
SD15	89h	6.75	52h	4.52
SD16	93h	6.06	55h	5.61

Table 45 Bias and random error in μg m^{-3} when calibrated in the first period with model D

	1st calibration period			2nd calibration period		
Sensor ID	**Uptime**	**Bias**	**SDR**	**Uptime**	**Bias**	**SDR**
SD01	146h	-0.1	8.8	106h	40.1	18.2
SD02	142h	0.0	8.2	199h	21.4	12.8
SD03	141h	0.0	5.1	205h	5.6	9.3
SD04	144h	0.0	5.5	202h	-9.2	5.8
SD05	143h	0.0	7.0	192h	3.0	6.3
SD06	154h	0.0	6.0	197h	-2.1	6.8
SD07	137h	0.0	6.6	196h	6.6	6.8
SD08	140h	0.0	5.4	199h	3.1	9.1
SD09	140h	0.0	4.5	196h	0.7	5.3

SD10	158h	0.0	7.2	206h	0.2	7.9
SD11	144h	0.0	6.3	205h	0.5	8.5
SD12	145h	0.0	5.7	194h	10.1	6.0
SD13	143h	0.0	5.8	206h	9.8	7.7
SD14	131h	0.0	5.9	211h	16.6	6.9
SD15	141h	0.0	6.0	198h	21.3	6.8
SD16	148h	0.0	5.7	47h	15.6	8.7

Table 6~~5~~ Comparison of sensor ~~54200~~SD04 with Oude Schans station during the campaign period, according to different calibrations

	1st calibration	2nd calibration	Weighted calibration
Mean NO_2, GGD Oude Schans	19.96 µg m^{-3}	19.96 µg m^{-3}	19.96 µg m^{-3}
Mean NO_2, sensor ~~54200~~SD04	17.02 µg m^{-3}	22.21 µg m^{-3}	19.87 µg m^{-3}
Bias	-2.94 µg m^{-3}	2.25 µg m^{-3}	-0.09 µg m^{-3}
RMSE ~~residuals~~	6.10 µg m^{-3}	5.25 µg m^{-3}	5.20 µg m^{-3}
Correlation	0.89	0.89	0.88

Supplement: NO_2 regression model coefficients

Units c_0 (Intercept):	µg m^{-3}
Units c_1 (S_{WE}):	µg m^{-3}/count
Units c_2 (S_{AE}):	µg m^{-3}/count
Units c_3 (T):	µg m^{-3}/°C
Units c_4 (RH):	µg m^{-3}/%
Units c_5 (O3):	µg m^{-3}/µg·m^{-3}

Table S1 Relation sensor ID and its network ID, which is used as reference in raw data

Sensor device ID	WiFi chip ID
SD01	1184206
SD02	14560051
SD03	55303
SD04	54200
SD05	1184527
SD06	1184739
SD07	1183931
SD08	53788
SD09	26296
SD10	1185325
SD11	1184453
SD12	717780
SD13	55300
SD14	13905017
SD15	1184838
SD16	54911

Table S2 Regression results for sensor devices

SD01 [a]		**Intercept**	S_{WE}	S_{AE}	**T**	**RH** [b]	O_3
Model A	1st period	455.38 ± 55.18	0.6977 ± 0.0649	-1.0835 ± 0.0970			
	2nd period [c]	-6.04 ± 36.69	0.2475 ± 0.0488	-0.2343 ± 0.0604			
Model B	1st period	715.45 ± 59.71	0.8394 ± 0.0592	-1.4811 ± 0.1001		0.5326 ± 0.0743	
	2nd period [c]	2.24 ± 43.51	0.2469 ± 0.0490	-0.2431 ± 0.0654		0.0280 ± 0.0782	
Model C	1st period	827.92 ± 87.54	0.8688 ± 0.0680	-1.5498 ± 0.1262	-1.6344 ± 0.3130		
	2nd period [c]	-173.77 ± 64.95	0.3000 ± 0.0499	-0.1698 ± 0.0618	1.5927 ± 0.5177		
Model D	1st period	790.88 ± 82.04	0.8707 ± 0.0635	-1.5645 ± 0.1178	-0.5051 ± 0.3778	0.4513 ± 0.0958	
	2nd period [c]	-178.93 ± 64.10	0.3133 ± 0.0497	-0.2007 ± 0.0628	2.1055 ± 0.5715	0.1650 ± 0.0827	
Model E	1st period	274.85 ± 78.12	0.3186 ± 0.0703	-0.4805 ± 0.1346	-0.5447 ± 0.2820	-0.4744 ± 0.1126	-0.5349 ±
	2nd period [c]	56.69 ± 54.19	0.2864 ± 0.0371	-0.3343 ± 0.0490	1.4917 ± 0.4309	-0.1120 ± 0.0686	-0.3883 ±

[a] Alphasense NO2-B42F sensor, used in previous experiments for more than one year

[b] RH sensor overestimates and often saturated at 100%

[c] Only 42% uptime in 2nd calibration period.

SD02 [a]		**Intercept**	S_{WE}	S_{AE}	**T**	**RH**	O_3
Model A	1st period	355.92 ± 65.74	0.8862 ± 0.0621	-1.2633 ± 0.0921			
	2nd period	303.68 ± 86.54	0.2770 ± 0.0667	-0.5599 ± 0.1034			
Model B	1st period	624.53 ± 85.42	0.8686 ± 0.0583	-1.5077 ± 0.1017		0.3916 ± 0.0863	
	2nd period	629.53 ± 97.17	0.3356 ± 0.0624	-0.9477 ± 0.1159		0.3625 ± 0.0615	
Model C	1st period	502.09 ± 109.36	0.9007 ± 0.0624	-1.4001 ± 0.1229	-0.5684 ± 0.3410		
	2nd period	68.85 ± 147.75	0.2973 ± 0.0671	-0.3864 ± 0.1357	0.8454 ± 0.4327		
Model D	1st period	589.20 ± 105.35	0.8618 ± 0.0596	-1.4742 ± 0.1174	0.2142 ± 0.3720	0.4204 ± 0.1000	
	2nd period	34.28 ± 123.80	0.4429 ± 0.0584	-0.6025 ± 0.1161	2.8976 ± 0.4263	0.5956 ± 0.0651	
Model E	1st period	-87.90 ± 101.40	0.3690 ± 0.0645	-0.2424 ± 0.1460	0.1739 ± 0.2770	-0.6170 ± 0.1234	-0.5754 ±
	2nd period	-174.15 ± 107.47	0.4075 ± 0.0496	-0.3524 ± 0.1023	3.8518 ± 0.3769	0.2585 ± 0.0672	-0.3428 ±

[a] Alphasense NO2-B42F sensor, used in previous experiments for more than one year

SD03		**Intercept**	S_{WE}	S_{AE}	**T**	**RH**	O_3
Model A	1st period	-228.65 ± 137.58	1.0877 ± 0.0578	-0.8029 ± 0.1113			
	2nd period	-470.06 ± 98.31	0.8521 ± 0.0388	-0.4193 ± 0.0772			
Model B	1st period	-1335.96 ± 157.68	1.2551 ± 0.0482	-0.1132 ± 0.1127		-0.6560 ± 0.0686	
	2nd period	-991.61 ± 161.21	0.8898 ± 0.0386	-0.0591 ± 0.1168		-0.1618 ± 0.0404	
Model C	1st period	-972.80 ± 115.40	1.1445 ± 0.0410	-0.3343 ± 0.0878	1.7279 ± 0.1455		
	2nd period	-913.18 ± 132.27	0.8192 ± 0.0375	-0.0765 ± 0.1031	0.8840 ± 0.1867		
Model D	1st period	-1272.13 ± 137.05	1.2045 ± 0.0425	-0.1492 ± 0.0979	1.2690 ± 0.1867	-0.2944 ± 0.0798	

	2nd period	-1050.59 ± 159.66	0.8448 ± 0.0410	0.0095 ± 0.1172	0.6707 ± 0.2328	-0.0758 ± 0.0497	
Model E	1st period	-818.09 ± 120.96	0.8961 ± 0.0487	-0.1706 ± 0.0782	0.5898 ± 0.1678	-0.5387 ± 0.0695	-0.2749 ±
	2nd period	-728.05 ± 108.84	0.8202 ± 0.0275	-0.1908 ± 0.0795	1.0731 ± 0.1579	-0.2465 ± 0.0350	-0.3029 ±

SD04		**Intercept**	S_{WE}	S_{AE}	**T**	**RH**	**O_3**
Model A	1st period	-968.20 ± 145.13	0.9138 ± 0.0538	-0.1237 ± 0.1254			
	2nd period	-371.22 ± 144.45	0.9786 ± 0.0500	-0.6833 ± 0.1329			
Model B	1st period	-1729.95 ± 119.61	1.1641 ± 0.0430	0.2736 ± 0.0939		-0.5386 ± 0.0444	
	2nd period	-1190.28 ± 141.99	1.0625 ± 0.0413	-0.0659 ± 0.1236		-0.4225 ± 0.0414	
Model C	1st period	-1044.89 ± 110.06	1.0490 ± 0.0427	-0.2245 ± 0.0954	1.4562 ± 0.1412		
	2nd period	-864.22 ± 116.48	0.9909 ± 0.0378	-0.3182 ± 0.1048	1.5499 ± 0.1269		
Model D	1st period	-1613.28 ± 153.33	1.1499 ± 0.0445	0.1818 ± 0.1204	0.3200 ± 0.2638	-0.4442 ± 0.0896	
	2nd period	-1055.65 ± 131.76	1.0203 ± 0.0384	-0.1723 ± 0.1144	1.1527 ± 0.1844	-0.1639 ± 0.0561	
Model E	1st period	-1129.35 ± 115.34	0.8046 ± 0.0426	0.1830 ± 0.0848	-0.3285 ± 0.1936	-0.7627 ± 0.0685	-0.3671 ±
	2nd period	-848.14 ± 97.58	0.8909 ± 0.0298	-0.1992 ± 0.0836	1.5326 ± 0.1378	-0.3227 ± 0.0427	-0.2241 ±

SD05		**Intercept**	S_{WE}	S_{AE}	**T**	**RH**	**O_3**
Model A	1st period	-155.10 ± 197.19	0.8368 ± 0.0743	-0.6841 ± 0.1768			
	2nd period	475.82 ± 194.53	0.9137 ± 0.0542	-1.2719 ± 0.1730			
Model B	1st period	-1953.53 ± 246.66	1.1485 ± 0.0672	0.5047 ± 0.1881		-0.9840 ± 0.1050	
	2nd period	-805.01 ± 261.61	1.0611 ± 0.0538	-0.3549 ± 0.2090		-0.6526 ± 0.0988	
Model C	1st period	-1056.05 ± 162.02	1.0371 ± 0.0562	-0.1946 ± 0.1340	2.3488 ± 0.2045		
	2nd period	-983.97 ± 191.54	0.9821 ± 0.0414	-0.2015 ± 0.1588	2.3771 ± 0.1997		
Model D	1st period	-1623.07 ± 222.70	1.1235 ± 0.0592	0.2088 ± 0.1715	1.7161 ± 0.2649	-0.4430 ± 0.1245	
	2nd period	-1162.98 ± 221.80	1.0114 ± 0.0452	-0.0756 ± 0.1771	2.1686 ± 0.2386	-0.1564 ± 0.0989	
Model E	1st period	-1079.04 ± 158.48	0.7104 ± 0.0522	0.2328 ± 0.1174	0.5648 ± 0.2032	-0.8305 ± 0.0906	-0.4053 ±
	2nd period	-1067.82 ± 174.06	0.8927 ± 0.0371	-0.0218 ± 0.1389	2.4442 ± 0.1887	-0.4412 ± 0.0818	-0.2397 ±

SD06		**Intercept**	S_{WE}	S_{AE}	**T**	**RH**	**O_3**
Model A	1st period	-141.88 ± 158.37	0.6136 ± 0.0607	-0.5241 ± 0.1168			
	2nd period	437.30 ± 151.50	0.8025 ± 0.0589	-1.2130 ± 0.1582			
Model B	1st period	-931.37 ± 123.99	1.2158 ± 0.0619	-0.4780 ± 0.0800		-0.7288 ± 0.0555	
	2nd period	-300.44 ± 174.06	0.9395 ± 0.0566	-0.7145 ± 0.1600		-0.4714 ± 0.0692	
Model C	1st period	-639.87 ± 102.28	1.0652 ± 0.0470	-0.6367 ± 0.0721	2.3781 ± 0.1504		
	2nd period	-581.47 ± 122.97	0.9636 ± 0.0413	-0.5853 ± 0.1151	2.6484 ± 0.1756		
Model D	1st period	-824.79 ± 106.47	1.1850 ± 0.0529	-0.5839 ± 0.0695	1.6737 ± 0.2198	-0.3069 ± 0.0728	
	2nd period	-666.44 ± 134.13	0.9811 ± 0.0427	-0.5242 ± 0.1212	2.4866 ± 0.2035	-0.0941 ± 0.0604	
Model E	1st period	-463.82 ± 73.02	0.8150 ± 0.0426	-0.4419 ± 0.0459	0.8318 ± 0.1531	-0.5519 ± 0.0499	-0.3402 ±
	2nd period	-592.51 ± 107.94	0.8732 ± 0.0358	-0.4531 ± 0.0976	2.6967 ± 0.1647	-0.2927 ± 0.0522	-0.2249 ±

SD07		Intercept	S_{WE}	S_{AE}	T	RH [a]	O_3
Model A	1st period	-576.41 ± 188.25	0.9615 ± 0.0716	-0.4811 ± 0.1520			
	2nd period	-239.15 ± 155.74	0.8866 ± 0.0486	-0.6834 ± 0.1418			
Model B	1st period	-576.41 ± 188.25	0.9615 ± 0.0716	-0.4811 ± 0.1520			
	2nd period	-239.15 ± 155.74	0.8866 ± 0.0486	-0.6834 ± 0.1418			
Model C	1st period	-1217.57 ± 144.34	1.1305 ± 0.0528	-0.1642 ± 0.1110	1.9435 ± 0.1678		
	2nd period	-977.93 ± 145.57	0.8717 ± 0.0393	-0.0987 ± 0.1284	1.6673 ± 0.1647		
Model D	1st period	-1217.57 ± 144.34	1.1305 ± 0.0528	-0.1642 ± 0.1110	1.9435 ± 0.1678		
	2nd period	-977.93 ± 145.57	0.8717 ± 0.0393	-0.0987 ± 0.1284	1.6673 ± 0.1647		
Model E	1st period	-578.07 ± 142.70	0.7891 ± 0.0606	-0.3243 ± 0.0934	1.7254 ± 0.1405		-0.2656 ±
	2nd period	-495.36 ± 120.13	0.7724 ± 0.0316	-0.3963 ± 0.1025	2.3365 ± 0.1401		-0.2254 ±

[a] RH sensor not working

SD08		Intercept	S_{WE}	S_{AE}	T	RH	O_3
Model A	1st period	231.44 ± 103.68	1.0802 ± 0.0639	-1.2514 ± 0.1086			
	2nd period	428.20 ± 110.91	1.0221 ± 0.0609	-1.3582 ± 0.1103			
Model B	1st period	-521.55 ± 174.37	1.1806 ± 0.0618	-0.7141 ± 0.1443		-0.4831 ± 0.0937	
	2nd period	141.16 ± 175.82	1.0604 ± 0.0631	-1.1578 ± 0.1454		-0.0651 ± 0.0311	
Model C	1st period	-798.25 ± 114.09	1.1319 ± 0.0454	-0.5061 ± 0.0995	2.4721 ± 0.2100		
	2nd period	-941.92 ± 168.22	0.9603 ± 0.0505	-0.2244 ± 0.1480	2.5145 ± 0.2593		
Model D	1st period	-1129.69 ± 139.87	1.1835 ± 0.0454	-0.2705 ± 0.1136	2.2559 ± 0.2085	-0.2704 ± 0.0716	
	2nd period	-983.10 ± 189.26	0.9685 ± 0.0534	-0.1975 ± 0.1586	2.4876 ± 0.2659	-0.0127 ± 0.0265	
Model E	1st period	-725.55 ± 113.06	0.8481 ± 0.0478	-0.2249 ± 0.0860	1.2801 ± 0.1849	-0.4709 ± 0.0577	-0.2966 ±
	2nd period	-685.96 ± 131.35	0.8376 ± 0.0377	-0.2914 ± 0.1089	2.5194 ± 0.1824	-0.1211 ± 0.0196	-0.2898 ±

SD09		Intercept	S_{WE}	S_{AE}	T	RH	O_3
Model A	1st period	100.52 ± 221.21	0.8669 ± 0.0671	-0.8952 ± 0.1979			
	2nd period	407.81 ± 127.41	0.9154 ± 0.0458	-1.1897 ± 0.1159			
Model B	1st period	-1138.92 ± 172.00	1.1781 ± 0.0498	-0.1707 ± 0.1407		-0.8205 ± 0.0609	
	2nd period	-132.85 ± 146.09	1.0685 ± 0.0488	-0.8851 ± 0.1171		-0.2933 ± 0.0477	
Model C	1st period	-332.23 ± 109.76	1.1460 ± 0.0353	-0.8613 ± 0.0965	2.4841 ± 0.1183		
	2nd period	-504.18 ± 113.38	1.0011 ± 0.0334	-0.5837 ± 0.0943	2.0206 ± 0.1492		
Model D	1st period	-586.25 ± 132.75	1.1794 ± 0.0358	-0.6738 ± 0.1103	2.0415 ± 0.1799	-0.2192 ± 0.0687	
	2nd period	-688.42 ± 119.56	1.0694 ± 0.0368	-0.4885 ± 0.0944	1.8326 ± 0.1522	-0.1460 ± 0.0380	
Model E	1st period	-383.42 ± 107.85	0.8973 ± 0.0424	-0.5253 ± 0.0892	1.1754 ± 0.1726	-0.4695 ± 0.0613	-0.2518 ±
	2nd period	-498.89 ± 100.31	0.9728 ± 0.0319	-0.5403 ± 0.0778	2.1983 ± 0.1309	-0.2250 ± 0.0323	-0.1837 ±

SD10		**Intercept**	S_{WE}	S_{AE}	**T**	**RH**	$\mathbf{O_3}$
Model A	1st period	342.04 ± 94.07	0.8221 ± 0.0657	-1.1629 ± 0.1206			
	2nd period	417.68 ± 78.62	0.8047 ± 0.0546	-1.2119 ± 0.1009			
Model B	1st period	-89.45 ± 187.91	0.9168 ± 0.0738	-0.8859 ± 0.1583		-0.2824 ± 0.1071	
	2nd period	103.71 ± 118.52	0.8641 ± 0.0558	-0.9951 ± 0.1164		-0.2487 ± 0.0717	
Model C	1st period	-847.45 ± 133.34	1.1001 ± 0.0566	-0.5102 ± 0.1108	2.9678 ± 0.2803		
	2nd period	-784.93 ± 122.97	0.8745 ± 0.0432	-0.3272 ± 0.1113	3.2652 ± 0.2889		
Model D	1st period	-1152.70 ± 175.33	1.1668 ± 0.0611	-0.3120 ± 0.1325	2.9112 ± 0.2760	-0.2147 ± 0.0820	
	2nd period	-862.03 ± 131.60	0.8947 ± 0.0449	-0.2759 ± 0.1154	3.1490 ± 0.2968	-0.0950 ± 0.0593	
Model E	1st period	-825.25 ± 115.40	0.7707 ± 0.0478	-0.1058 ± 0.0867	1.8251 ± 0.1930	-0.4975 ± 0.0564	-0.3808 ±
	2nd period	-622.53 ± 103.17	0.8094 ± 0.0352	-0.3689 ± 0.0890	3.2492 ± 0.2283	-0.2528 ± 0.0475	-0.2555 ±

SD11		**Intercept**	S_{WE}	S_{AE}	**T**	**RH**	$\mathbf{O_3}$
Model A	1st period	338.42 ± 80.88	0.9823 ± 0.0665	-1.2246 ± 0.1025			
	2nd period	748.59 ± 74.96	0.9642 ± 0.0547	-1.5368 ± 0.0924			
Model B	1st period	0.26 ± 133.88	1.0444 ± 0.0675	-0.9995 ± 0.1229		-0.2995 ± 0.0961	
	2nd period	752.43 ± 95.23	0.9629 ± 0.0587	-1.5387 ± 0.0973		0.0038 ± 0.0575	
Model C	1st period	-962.71 ± 126.96	1.0735 ± 0.0485	-0.3309 ± 0.1070	3.4356 ± 0.2980		
	2nd period	30.62 ± 145.29	1.0385 ± 0.0526	-1.0668 ± 0.1198	1.8190 ± 0.3228		
Model D	1st period	-1109.75 ± 139.25	1.1055 ± 0.0495	-0.2339 ± 0.1128	3.3191 ± 0.2972	-0.1693 ± 0.0709	
	2nd period	33.02 ± 143.00	0.9974 ± 0.0539	-1.0453 ± 0.1182	2.2205 ± 0.3501	0.1582 ± 0.0580	
Model E	1st period	-480.10 ± 118.32	0.7539 ± 0.0490	-0.3363 ± 0.0839	1.6813 ± 0.2670	-0.3806 ± 0.0560	-0.3277 ±
	2nd period	99.69 ± 109.82	0.9454 ± 0.0416	-1.0242 ± 0.0907	2.4400 ± 0.2692	-0.0973 ± 0.0494	-0.2625 ±

SD12		**Intercept**	S_{WE}	S_{AE}	**T**	**RH**	$\mathbf{O_3}$
Model A	1st period	-375.21 ± 197.57	0.7775 ± 0.0611	-0.4837 ± 0.1851			
	2nd period	-406.98 ± 191.77	0.8879 ± 0.0500	-0.5767 ± 0.1841			
Model B	1st period	-1332.74 ± 156.87	1.1032 ± 0.0497	0.0257 ± 0.1345		-0.6993 ± 0.0561	
	2nd period	-1248.39 ± 178.05	0.9608 ± 0.0414	0.0870 ± 0.1644		-0.4312 ± 0.0437	
Model C	1st period	-819.17 ± 126.64	1.0416 ± 0.0420	-0.4203 ± 0.1154	2.0988 ± 0.1400		
	2nd period	-800.71 ± 148.10	0.9405 ± 0.0379	-0.3286 ± 0.1402	1.6465 ± 0.1364		
Model D	1st period	-1074.88 ± 140.40	1.0961 ± 0.0430	-0.2346 ± 0.1219	1.4954 ± 0.2136	-0.2799 ± 0.0770	
	2nd period	-1012.78 ± 166.26	0.9545 ± 0.0377	-0.1466 ± 0.1541	1.2583 ± 0.1985	-0.1562 ± 0.0589	
Model E	1st period	-595.45 ± 113.66	0.7813 ± 0.0435	-0.2757 ± 0.0908	0.8578 ± 0.1697	-0.4865 ± 0.0605	-0.2965 ±
	2nd period	-701.86 ± 121.46	0.8586 ± 0.0280	-0.3051 ± 0.1111	1.6906 ± 0.1460	-0.2922 ± 0.0434	-0.2300 ±

SD13		**Intercept**	S_{WE}	S_{AE}	**T**	**RH**	$\mathbf{O_3}$
Model A	1st period	-1703.40 ± 201.83	0.8218 ± 0.0583	0.5544 ± 0.1554			
	2nd period	-1008.31 ± 189.21	0.8631 ± 0.0504	-0.0632 ± 0.1732			

Model B	1st period	-1826.17 ± 148.81	1.1334 ± 0.0515	0.3588 ± 0.1156		-0.5732 ± 0.0523	
	2nd period	-1161.34 ± 190.56	0.8856 ± 0.0497	0.0550 ± 0.1729		-0.1936 ± 0.0589	
Model C	1st period	-872.76 ± 146.63	1.1012 ± 0.0437	-0.4577 ± 0.1269	2.3418 ± 0.1732		
	2nd period	-968.33 ± 167.16	0.8761 ± 0.0445	-0.1315 ± 0.1532	1.1078 ± 0.1454		
Model D	1st period	-1074.57 ± 179.99	1.1294 ± 0.0458	-0.3058 ± 0.1490	1.8671 ± 0.3032	-0.1561 ± 0.0822	
	2nd period	-999.93 ± 174.21	0.8800 ± 0.0450	-0.1057 ± 0.1584	1.0664 ± 0.1587	-0.0381 ± 0.0582	
Model E	1st period	-594.35 ± 134.76	0.7795 ± 0.0444	-0.2874 ± 0.1062	1.0126 ± 0.2282	-0.4704 ± 0.0645	-0.3327 ±
	2nd period	-505.72 ± 107.36	0.8246 ± 0.0271	-0.4485 ± 0.0964	2.1700 ± 0.1113	-0.2329 ± 0.0363	-0.3003 ±

SD14		**Intercept**	***S_{WE}***	***S_{AE}***	**T**	**RH**	**O_3**
Model A	1st period	162.64 ± 165.94	0.8156 ± 0.0903	-0.9075 ± 0.1248			
	2nd period	-3.20 ± 202.78	0.8580 ± 0.0540	-0.8237 ± 0.1811			
Model B	1st period	369.33 ± 139.19	1.0602 ± 0.0807	-1.2825 ± 0.1134		-0.6434 ± 0.0819	
	2nd period	-1011.65 ± 198.00	1.0253 ± 0.0480	-0.1369 ± 0.1663		-0.4382 ± 0.0452	
Model C	1st period	19.56 ± 91.93	1.1888 ± 0.0544	-1.1987 ± 0.0709	2.4905 ± 0.1454		
	2nd period	-1147.64 ± 153.38	0.9569 ± 0.0366	-0.0244 ± 0.1311	2.1478 ± 0.1342		
Model D	1st period	8.09 ± 97.95	1.1860 ± 0.0552	-1.1889 ± 0.0766	2.5401 ± 0.2039	0.0268 ± 0.0770	
	2nd period	-1278.51 ± 159.07	0.9905 ± 0.0383	0.0621 ± 0.1333	1.8680 ± 0.1693	-0.1217 ± 0.0460	
Model E	1st period	114.64 ± 71.51	0.8144 ± 0.0527	-0.8532 ± 0.0635	1.2001 ± 0.1929	-0.4387 ± 0.0705	-0.3356 ±
	2nd period	-844.54 ± 120.58	0.9049 ± 0.0287	-0.1972 ± 0.0992	2.2316 ± 0.1266	-0.2564 ± 0.0350	-0.2176 ±

SD15		**Intercept**	***S_{WE}***	***S_{AE}***	**T**	**RH** [a]	**O_3**
Model A	1st period	1211.20 ± 242.16	0.9008 ± 0.1180	-1.8984 ± 0.2883			
	2nd period	1455.17 ± 155.20	1.2443 ± 0.0810	-2.4648 ± 0.1843			
Model B	1st period	911.69 ± 319.97	0.9893 ± 0.1330	-1.7240 ± 0.3122		-0.2561 ± 0.1797	
	2nd period	1455.17 ± 155.20	1.2443 ± 0.0810	-2.4648 ± 0.1843			
Model C	1st period	-166.53 ± 139.22	1.8265 ± 0.0748	-1.7541 ± 0.1448	4.8106 ± 0.2373		
	2nd period	-438.20 ± 143.92	1.4576 ± 0.0516	-1.1488 ± 0.1363	3.6043 ± 0.2039		
Model D	1st period	-104.50 ± 169.26	1.8111 ± 0.0786	-1.7939 ± 0.1576	4.8373 ± 0.2413,	0.0596 ± 0.0921	
	2nd period	-438.20 ± 143.92	1.4576 ± 0.0516	-1.1488 ± 0.1363	3.6043 ± 0.2039		
Model E	1st period	-56.70 ± 134.13	1.2676 ± 0.0865	-1.2255 ± 0.1397	3.1038 ± 0.2705	-0.3717 ± 0.0871	-0.3226 ±
	2nd period	-217.54 ± 133.72	1.2729 ± 0.0539	-1.1467 ± 0.1228	3.7105 ± 0.1844		-0.1401 ±

[a] RH sensor breaks down after July 25

SD16		**Intercept**	***S_{WE}***	***S_{AE}***	**T**	**RH**	**O_3**
Model A	1st period	-594.31 ± 220.12	0.8007 ± 0.0704	-0.3192 ± 0.1976			
	2nd period [a]	-254.68 ± 307.78	0.3469 ± 0.0885	-0.1361 ± 0.2747			
Model B	1st period	-1537.42 ± 194.12	1.1674 ± 0.0655	0.1164 ± 0.1584		-0.5503 ± 0.0550	

	2nd period [a]	-1053.52 ± 346.39	0.5320 ± 0.0926	0.3510 ± 0.2752		-0.2220 ± 0.0601	
Model C	1st period	-1045.41 ± 129.96	1.2206 ± 0.0476	-0.4227 ± 0.1144	2.4971 ± 0.1466		
	2nd period [a]	-1118.84 ± 294.51	0.5547 ± 0.0805	0.3426 ± 0.2357	1.3564 ± 0.2612		
Model D	1st period	-1215.51 ± 146.15	1.2551 ± 0.0490	-0.3038 ± 0.1229	2.1742 ± 0.1972	-0.1333 ± 0.0555	
	2nd period [a]	-1156.53 ± 316.09	0.5629 ± 0.0846	0.3693 ± 0.2498	1.2518 ± 0.3962	-0.0290 ± 0.0819	
Model E	1st period	-623.06 ± 135.29	0.8844 ± 0.0575	-0.3786 ± 0.0993	1.5146 ± 0.1753	-0.2937 ± 0.0482	-0.2883 ±
	2nd period [a]	-553.67 ± 329.07	0.7349 ± 0.0897	-0.2996 ± 0.2928	1.7739 ± 0.3817	-0.2115 ± 0.0894	-0.2733 ±

[a] Only 18% uptime in 2nd calibration period