Lab & Field Benchmarking of Portable Air Quality Sensors: Parcel A

Study accomplished under the authority of VMM
Reference: 2023/HEALTH/R/2937
January 2023

vito
Vision on technology for a better world
vito.be

Lab and Field Benchmarking of Portable Air Quality Sensors: Parcel A

VITO
Boeretang 200
2400 MOL
Belgium
VAT No: BE0244.195.916
vito@vito.be – www.vito.be
IBAN BE34 3751 1173 5490 BBRUBEBB

Martine Van Poppel
Project Manager
+32 14 33 53 65 /
martine.vanpoppel@vito.be

Jelle Hofman
Researcher
+32 14 33 57 58 / jelle.hofman@vito.be

AUTHORS

Jelle Hofman, VITO
Martine Van Poppel, VITO
Borislav Lazarov, VITO

Distribution: limited I Ref.: 2023/HEALTH/R/2937

DISTRIBUTION LIST

Evelyne Elst, VMM
Christophe Stroobants, VMM
Inge Smets, VMM
Jordy Vercauteren, VMM

Distribution: limited II Ref.: 2023/HEALTH/R/2937

SUMMARY

This report describes the results of an extensive lab- and field benchmarking study of 9 commercially available portable sensor systems, measuring particulate matter (PM) or nitrogen dioxide (NO_2). After an initial inventory study (literature study + market search) and sensor selection from long list (39) to shortlist (12), 9 sensor systems were ultimately selected and benchmarked in a lab test, mobile field test (internal GPS) and a 3-month field co-location campaign.

Results of the considered sensor systems indicate that out-of-the-box is relatively good for PM and BC, but maturity of the tested NO_2 sensors is still low and additional effort is needed in terms of signal noise and calibration. The PM sensors showed similar performances and sensitivities (e.g. to RH) and best sensor performance was reached after conducting a (2-week) field co-location calibration. Moreover, the variability between the sensors (BSU) was very low which is important when comparing data from multiple sensors. Future testing should focus on gaining real-life user feedback (mobile deployments with citizens/employees), evaluate sensors during mobile deployment and assessing urban PM gradients (possibly in combination with other pollutants) to determine the required sensor accuracy (raw and/or field calibrated).

For NO_2, more variation between the considered sensors (3) was observed with issues in terms of stability/repeatability (high noisiness) and sensor response calibration. One sensor was performing best (out-of-the-box), while another showed highest potential due to the high signal stability and association. Future work should focus on testing different calibration approaches or noise cancellation techniques (hardware or post-processing). Moreover, urban NO_2 gradients can be studied by means of mobile monitoring in order to determine the required sensor accuracy.

Besides the quantitative performance metrics, qualitative evaluations were gained throughout this benchmarking study. Main concerns include smartphone application considerations (availability (region/country), iOS/Android, clock sync issues, continuous connectivity) and redundancy of data storage by means of a SD card (no data was lost on SD cards) and clarity about time resolution, potential data compression in cloud portals. We valued sensor systems with internal GPS sensors, SD storage and autonomous operation (no app connectivity needed). Requirements might be different for citizens, employees or other sensor users.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Distribution: limited XI Ref.: 2023/HEALTH/R/2937

1 INTRODUCTION

1.1 Background

During the past decades, air quality has improved significantly in Flanders. Yet air pollution remains a major cause of more than 5,000 premature deaths per year, highlighting the importance of proper air pollution monitoring. Currently, the exposure of the population to air pollution is still determined based on the citizen home address (=static exposure). However, research has shown that people are exposed to the highest air pollution peaks at times when they are mobile (i.e. during transport). Therefore, these movements must also be taken into account when calculating personal exposure to air pollution (=dynamic exposure). Tentative research in the context of CurieuzeNeuzen showed that air pollution hotspots in the city centers of Flemish cities and municipalities also strongly increase the exposure of the population living outside those hotspots (Figure 1).

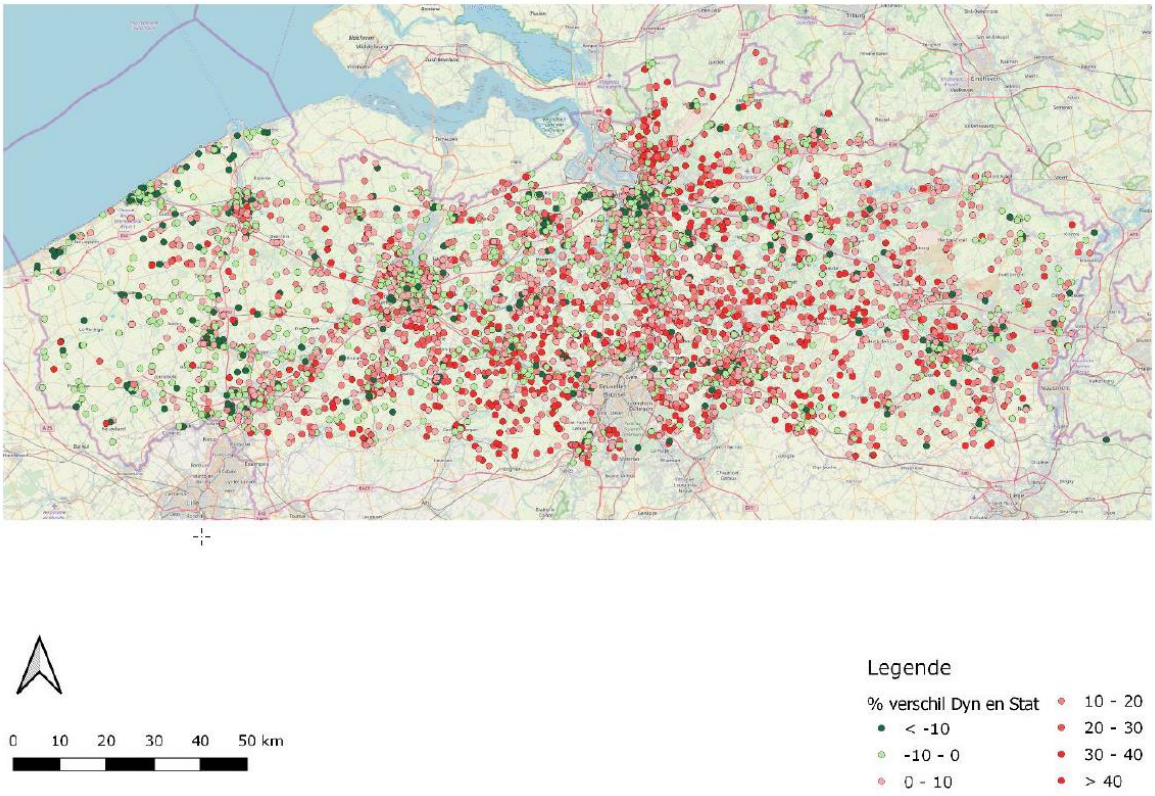


Figure 1 Relative difference between dynamic and static exposure to NO_2 in %.

The Flanders Environmental Agency (VMM) and the Flanders Environmental Planning Agency (VPO), therefore, want to develop a toolset to monitor the personal exposure of every Flemish citizen to air pollution, taking into account the movements of citizens throughout the day. This includes vulnerable groups such as elderly and citizens of all socio-economic status (SES) groups. This toolset will result in a validated, model-based approach for determining dynamic exposure or – as a fallback – a monitoring-based approach. This toolset has been included in the portfolio of the Innovative Public Procurement Program (PIO) and can be found on this web page.

To obtain this toolset, an experiment will be conducted with citizens equipped with portable instruments to dynamically measure air quality for different target (SES) group-pollutant combinations. In the first place, the intention is to evaluate existing air quality models (i.e. ATMO-street) with regard to their usability for determining dynamic exposure (phase 3 to 5 in Figure 2). The monitoring results obtained from the dynamic monitoring experiments will be compared with the modeled results of the same trajectories. At best, we can demonstrate that ATMO-street is already capable in estimating the dynamic exposure with sufficient accuracy. If the model results still prove to be insufficient, we will have to fall back on (larger-scale) measurements with sensors for the time being to estimate the dynamic exposure of citizens. The collected monitoring data can in turn be used again to further improve the air quality model until the model results ultimately prove to be sufficient.

From the derived exposure tools (phase 6 in Figure 2), we can raise awareness among citizens about their personal exposure, facilitate behaviour change (e.g. the healthiest route app), enable ground-breaking health research and inform policymakers on local air pollution hotspots in their city/municipality, allowing for evidence-based policy measures.

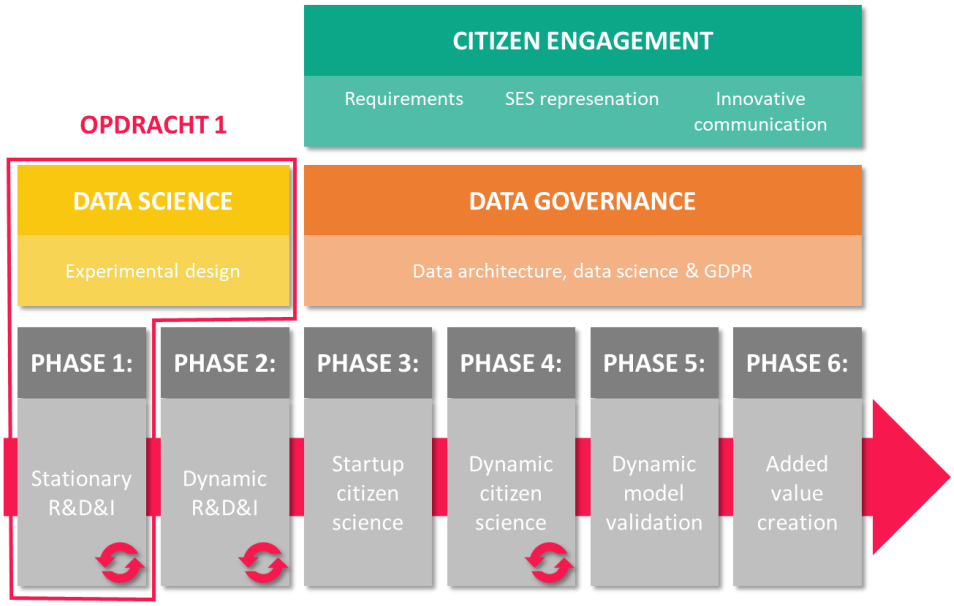


Figure 2 Global project overview

1.1.1 Preliminary study

Under support of the Innovative Public Procurement Program (PIO), a preliminary study has already been completed. This study defined a number of innovative use cases, personae (including lower SES groups) and a possible action plan was proposed after conducting a qualitative market consultation. In short, the conclusion of this preliminary study is that a series of “iterative experiments” must be carried out using “a hybrid set of innovative devices” of types 1 (commercial mid-grade instruments) and 3 (commercial low-cost sensors) or 4 (DIY sensors).

This preliminary phase identified 2 research gaps that are targeted in this study:

- **Parcel A**: To what extent can commercially available technologies be used or should in-house development be used to carry out the innovative dynamic exposure measurements?

- **Parcel B**: What is the ideal experimental design for the next experiments?

Parcel A aims at **benchmarking commercially available, innovative air quality instruments for mobile air quality monitoring, with attention to the innovation requirements** as described in the preliminary study. In addition, the focus for this assignment is narrowed to the pollutants **nitrogen dioxide (NO_2) and particulate matter ($PM_{2.5}$ and PM_{10})**.
Specifically, the following tasks are fulfilled:

- A limited literature study with a focus on portable monitoring devices mentioned or shown in the preliminary study
- The selection and purchase of monitoring devices to be benchmarked worth a maximum of € 20,000 in consultation with VMM
- Benchmark study of the purchased devices under controlled (laboratory) conditions and in real-life (field) conditions, when mobile and co-located at a regulatory air quality monitoring station (AQMS)

The literature study, inventory (longlist), prioritization (shortlist) and purchase of the selected portable air quality instruments are described in a separate report. This report describes the benchmark results obtained during the lab and field campaigns.

2 METHODOLOGY

2.1 Selection portable sensors systems

Based on the preliminary study, an earlier literature study and market survey on air quality sensors conducted by VITO (1), a new literature study with focus on portable air quality sensors (~90 publications), sensor benchmark results from independent research institutes (AIRlab, AQ-SPEC, SamenMeten, EPA Air Sensor Toolbox, SeeTheAir,…) and sensor projects (BelAir, Snuffelfiets,…), a **longlist of 39 sensor candidates** was created. This longlist was further prioritized (scored) based on a set of predefined requirements:

- Price
- Measured pollutants
- Additional variables (e.g. temperature, relative humidity, pressure…)
- Temporal monitoring resolution
- Housing
- GPS availability
- Autonomy (h)
- Data storage
- Size
- Weight
- Display
- Required actions (buttons, smartphone app,…)

This resulted in a **shortlist of 12 promising portable sensor systems** for which quotation requests were send out. Ultimately, 9 sensor systems were purchased (Table 1), of which 8/9 contained a $PM_{2.5}$ and PM_{10} sensor, and 3/9 sensor systems contained an additional NO_2

sensor (SODAQ NO_2, DST Observair and 2BTech PAM). All 9 sensor systems can be regarded as portable air quality sensor systems, with some form of portability, autonomy (battery), data storage/transmission and localization (GPS) (Table 1). The 9 different sensor systems often included similar PM/NO_2 sensors; ultimately resulting in 3 different PM sensors and only one NO_2 sensor for evaluation (Table 1).

As the DST Observair included a black carbon (BC) sensor as well, and we got hold of a low-cost BC sensor system (BCmeter; https://bcmeter.org/) during the field benchmarking, BC (in addition to PM and NO_2), emitted by road traffic and health-related PM constituent, was evaluated as well during the field co-location campaign. It should be noted that the BCmeter can be considered as a research prototype for stationary measurements (wifi, power cable). In order to obtain a portable BCmeter, additional hardware/software development will be needed.

A picture of the purchased sensor systems (10) is provided in Figure 3.

Figure 3 Pictures of the purchased sensor systems (10) with on the upper picture (from left to right): PAM (2BTech), GeoAir, Observair (DST), SODAQ Air (SODAQ), PMscan (TERA), Open Seneca (Open Seneca) and ATMOTube Pro (ATMO). Lower pictures: SODAQ NO2 (left), Habitatmap Airbeam (middle) and BCmeter (right)

Table 1 Specifications of the purchased portable sensor systems for the lab and field benchmarking study

Device	Manufacturer	URL	PRICE (€)	Data Logging	GPS	Battery	Smartphone required?	Portability	Metrics	PM sensor	NO_2 sensor	Temporal resolution	Autonomy
ATMOTube Pro	Atmotube	https://atmotube.de/	249	X	Smartphone	X	X	carabiner	TVOC, **PM_1, $PM_{2.5}$, PM_{10}**, P, Temp, RH	Sensirion SPS30	-	2 sec	24h
Airbeam 3	HabitatMap	https://www.habitatmap.org/airbeam/buy-it-now	232.5	SD	X	X	No	belt clip + carabiner	**PM_1, $PM_{2.5}$, PM_{10}**, Temp RH	Plantower PMS7003	-	1 sec	17h
SODAQ AIR	SODAQ	https://shop.sodaq.com/sodaq-air.html	216	X	X	X	No	bicycle mount	**PM_1, $PM_{2.5}$, PM_{10}**	Sensirion SPS30	-	10 sec – 5 min	?
SODAQ NO2	SODAQ	Prototype	?			X	No	No	**PM_1, $PM_{2.5}$, PM_{10}, NO2,** Temp RH	Sensirion SPS30	Alphasense NO2-A43F	10 sec – 5 min	
PMScan	Tera	https://airparif.shinyapps.io/ChallengeResultsEN/	225	X	Smartphone	X	X	Strap	**PM_1, $PM_{2.5}$, PM_{10}**, Temp RH	TERA Next-PM	-	1 sec	15h
PAM	2B Technologies	https://twobtech.com/docs/manuals/model_PAM_revC-2.pdf	3500	X	X	X	No	No	CO, CO2, **PM_1, $PM_{2.5}$, PM_{10}, NO_2** Temp, Press, RH	Plantower PMS7003	Alphasense NO2-A43F	2 sec	7h
ObservAir®	Distributed Sensing Technologies (DST)	manual	4900	X (USB, SD, web)	X	X	No	No	**BC, NO_2,** CO	-	Alphasense NO2-A43F	2 sec	8h
GeoAir	Jaycon systems	https://www.mdpi.com/1424-8220/21/11/3761	Loan	SD (not included)	X	X	No	Belt clip + fixation hole	**PM_1, $PM_{2.5}$, PM_4, PM_{10}**, tVOC, Temp, RH	Sensirion SPS30	-	1 sec (1 min)	12-15h (1 minute)
Open-Seneca	Open-Seneca	https://www.open-seneca.org/air-quality-monitor	175	SD card	X	X	No	bicycle mount	**PM_1, $PM_{2.5}$, PM_4, PM_{10},**Temp, RH	Sensirion SPS30	-	1 sec	5h
BCmeter	BCmeter	https://bcmeter.org/	?	X (Linux/wifi)	no	no	no	no	**BC**	-	-		no

2.2 Benchmarking protocol

The purchased sensor systems were evaluated under controlled (laboratory) and real-life (field) conditions (Figure 4). Field benchmarking included a mobile test on a cargo bike and a 3-month co-location campaign at a regulatory urban background (R801) air quality monitoring station in Antwerp, Belgium.

Figure 4 Pictures of the laboratory PM exposure chamber (left), mobile field test (middle) and field co-location (right) campaigns.

2.2.1 Laboratory protocol

Laboratory tests were conducted for both PM and NO_2.

For $PM_{2.5}$ and PM_{10}, we evaluated:

- Lack of fit (linearity) of PM sensor when exposed to different concentrations of dolomite dust between 0 and 350 µg/m^3 (PM_{10}). Following setpoints were applied; 0,30, 40, 60,130, 200 and 350 µg/m^3. A Palas Particle dispenser (RBG 100) system connected to a fan-based dilution system in a PM exposure chamber was used.
- Sensitivity of PM sensor to the coarse (2.5-10µm) particle size range. We dosed sequentially 7.750 µm and 1.180 µm-sized monodisperse dust (silica nanospheres with density of 2 g/cm^3) using an aerosolizer (from the Grimm 7.851 aerosol generator system connected to a fan-based dilution system in a PM exposure chamber. This testing protocol is currently under discussion at the CEN TC264 working group (WG42) on performance targets for air quality sensors.

From the lack of fit tests, we evaluated the **comparability of the sensor vs reference** by calculating the associated sensor performance metrics for linearity (R^2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE) and Expanded Uncertainty (U_{exp}). In addition, the sensor stability (µg/m^3) was calculated as the standard deviation of the 1-minute averages at each setpoint (steady-state conditions) and the sensor accuracy (%) at each setpoint was calculated as:

$$accuracy\ (\%) = 100 - \left(\frac{|sensor - REF|}{REF}\right) * 100$$

With $sensor$ and REF the respective average sensor and reference concentrations (µg/m^3) at each setpoint interval. As reference instrument, we used a Grimm 11-D with heated sampling inlet line (EDM 264, Grimm).

The **comparability between the sensors** can be regarded as the observed variability between sensors of the same type and is calculated by the between-sensor-uncertainty (BSU):

$$BSU_{sensor} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(sensor_{ij} - average_i\right)^2}{n(k-1)}}$$

with n the number of sensors (3) and k the number of measurements.

For the sensor systems that included a NO_2 sensor, we evaluated:

- Lack of fit (linearity) test for the NO_2 sensors at concentration setpoints of 0, 40, 100, 140 and 200 µg/m^3.
- Sensor sensitivity to relative humidity at 15, 50, 70 and 90% (±5%) during stable temperature conditions of 20 ± 1°C.
- Sensor sensitivity to temperatures at -5, 10, 20 and 30 °C (±3°C) during stable relative humidity conditions of 50 ± 5%
- Sensor cross-sensitivity to ozone (120 µg/m^3) at zero and 100 µg/m^3 NO_2
- Sensor response time under rapidly changing NO_2 concentrations (from 0 to 200 µg/m^3).

From the lack of fit tests, we evaluated the **comparability of the sensor vs reference** by calculating the associated sensor performance metrics for linearity (R^2), accuracy (%), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Bias Error (MBE) and Expanded Uncertainty (U_{exp}). In addition, we evaluated **sensor stability** (standard deviation at each setpoint) and **comparability between the sensors** by calculating the between-sensor-uncertainty (BSU). As reference instrument, we applied a Thermo Scientific 42iQ-TL chemiluminescence monitor.

2.2.2 Field protocol

2.2.2.1 Mobile test

The small-scale mobile field test aimed at testing the GPS accuracy of the sensor systems at a ~10km trajectory in the heterogeneous urban environment of Antwerp (BE) with a variation of open areas, street canyons, tunnels,…(Figure 5). This GPS accuracy was evaluated by calculating the average horizontal distance (m) of the high-resolution mobile GPS measurements to a reference GPS trajectory.

The reference GPS track was determined by evaluating 3 different GPS systems (TomTom Runner2, Garmin Edge 810 and Komoot smartphone application), and selecting the best performing one (horizontal accuracy with regard to street network) as the reference GPS trajectory.

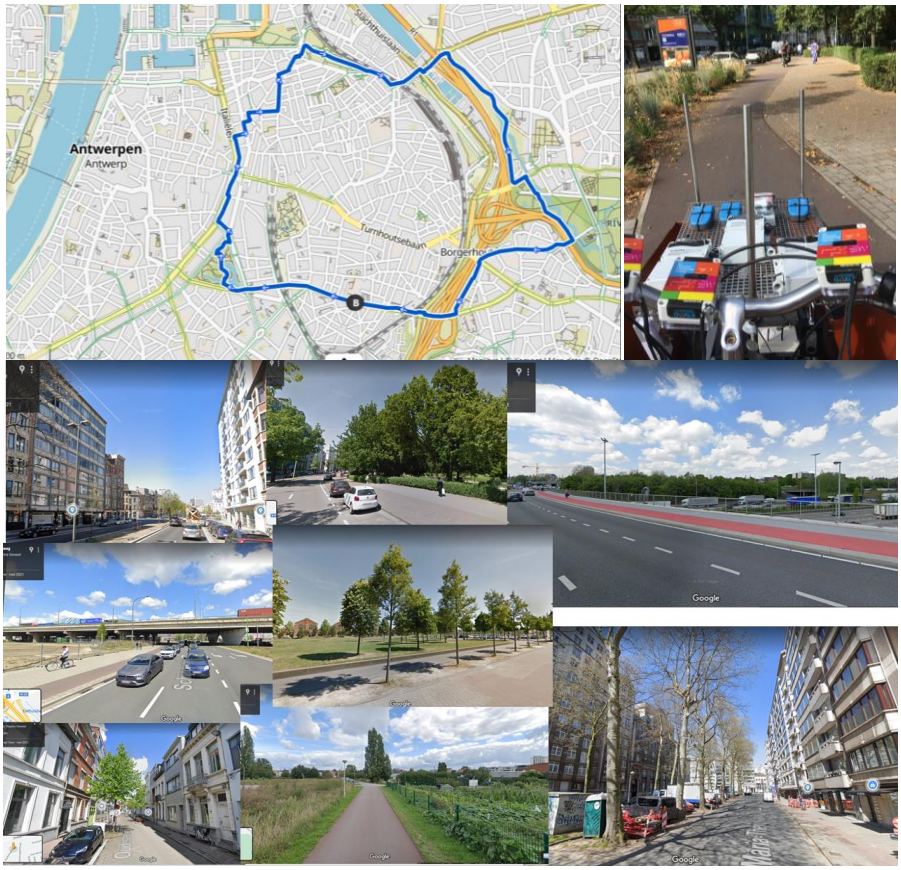


Figure 5 Mobile field trajectory of 10.4km in the city center of Antwerp, Belgium, cycled with the cargo bike (upper), with associated streetview pictures (lower) showing the variety in urban canopy.

2.2.2.2 Field co-location campaign

During the field co-location campaign, the portable sensors were exposed to ambient pollutant concentrations in a dedicated (actively ventilated) outdoor shelter, deployed on top (near the air inlets) of a regulatory urban background monitoring station (R801) in Antwerp, Belgium, for a period of 3 months (7/9/2022-5/12/2022). The collected sensor data was subsequently evaluated for:

- Hourly data coverage (%)
- Timeseries plot: RAW & LAB CAL
- Scatter plot: RAW & LAB CAL
- Comparability between sensors: Between sensor uncertainty (BSU)
- Comparability with reference (hourly): R^2, RMSE, MAE, MBE
- Expanded uncertainty (non-parametric): Uexp (%) at concentrations (±10%) of 50 µg/m³ (PM_{10}), 30 µg/m³ ($PM_{2.5}$), 40 µg/m³ (NO_2) and 1 µg/m³ (BC)

All these metrics were calculated for $PM_{2.5}$, PM_{10}, NO_2 and BC (where applicable). In addition we evaluated the sensitivity of the sensors (R^2, RMSE, MAE, MBE) in the **coarse particulate fraction** (PM_{10}-$PM_{2.5}$) and the impact of a **2-week field co-location calibration** (linear for PM and multilinear for NO_2) on the resulting sensor performance (+comparison with lab calibration).

3 RESULTS & DISCUSSION

3.1 Laboratory benchmark

3.1.1 PM tests

3.1.1.1 Lack-of-fit

For all sensor systems containing a PM sensor (8/9), lack-of-fit tests were conducted on 3 different days (18/7, 11/8 and 30/8) at PM_{10} concentrations ramping between 0 and 350 µg/m³. All sensor data was temporally aggregated (averaged) to a 1-minute resolution and merged with the reference (Grimm 11D) data. Setpoint averages were calculated based on the steady-state conditions (final 15 minutes of each 1-hour setpoint). From these setpoint averages, linearity plots with associated regression coefficients (slope + intercept (y=a*x+b) and slope only (y=a*x)) were derived and sensor accuracy (%) calculated. All results are shown per sensor type and subsequently presented in an overview table.

3.1.1.1.1 ATMOTUBE Pro

The raw Atmotube Pro sensor measurements respond nicely to the increasing concentrations steps (Figure 6), resulting in a good linearity between sensor and reference (R^2>0.99 in Table 2). Nevertheless, sensor readings seem to underestimate the actual (Grimm) $PM_{2.5}$ concentrations. This sensor underestimation is more pronounced for PM_{10}, while PM_1 is slightly overestimating actual concentrations (Figure 7). Mean setpoint accuracy (mean of different accuracies at each setpoint) varied between 82-85% for PM_1, 63-69% for $PM_{2.5}$ and 28-31 for PM_{10}.

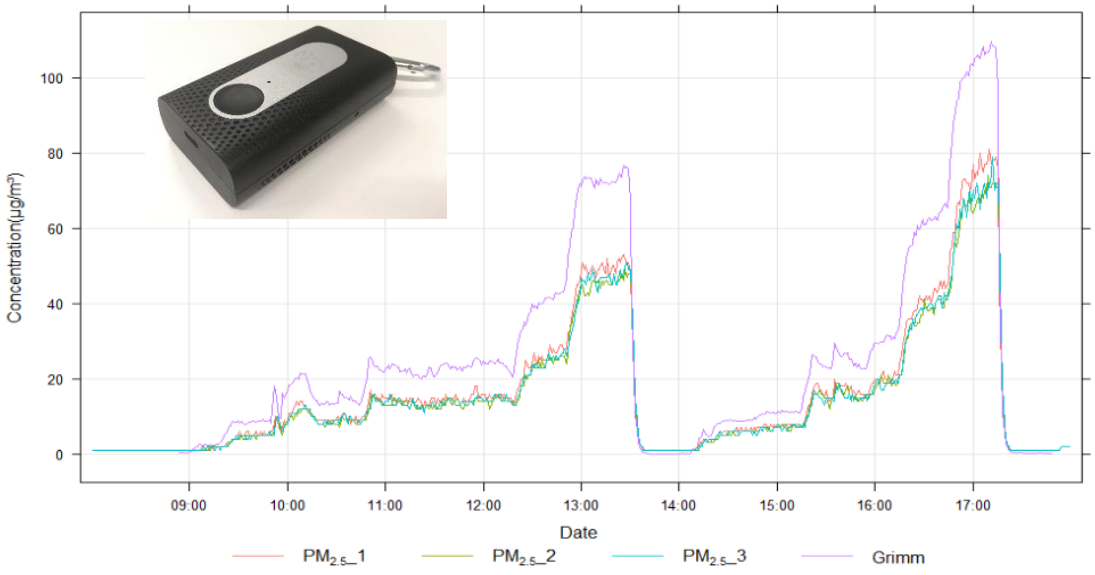


Figure 6 $PM_{2.5}$ concentrations generated during the lack-of-fit test as measured by the 3 Atmotube Pro sensors (1-3) and the reference monitor (Grimm)

PM_1				$PM_{2.5}$				PM_{10}				
PM_1_1	PM_1_2	PM_1_3	PM_1_REF	$PM_{2.5}$_1	$PM_{2.5}$_2	$PM_{2.5}$_3	$PM_{2.5}$_REF	PM_{10}_1	PM_{10}_2	PM_{10}_3	PM_{10}_REF	setpoints
1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.01	2.00	2.00	2.00	0.07	0
2.13	2.00	2.00	2.51	6.27	5.60	5.87	8.90	10.47	9.33	9.60	30.56	30
3.00	2.67	3.00	3.03	7.87	7.07	7.20	11.24	13.13	11.80	11.93	40.46	60
7.00	6.47	6.47	6.03	17.33	16.00	15.80	25.82	29.47	26.40	26.13	102.33	110
7.80	7.60	7.53	6.70	20.27	19.13	18.87	30.64	34.40	32.40	31.80	126.39	160
17.47	16.20	17.00	13.98	43.07	39.73	40.27	63.38	73.13	66.47	66.53	248.61	250
31.73	28.93	30.07	23.90	77.07	69.27	70.87	106.78	130.87	115.53	117.67	395.38	400

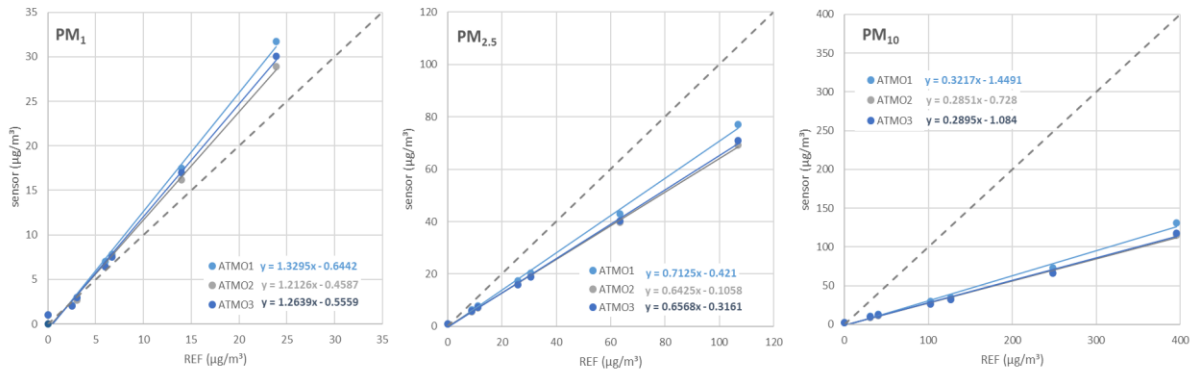


Figure 7 Setpoint averages (upper; µg/m³) and resulting linearity plots (lower) for PM_1, $PM_{2.5}$ and PM_{10}.

Derived regression coefficients (slope + intercept (y=a*x+b) and slope only (y=a*x)) and linearity (R^2) for each size fraction (PM_1, $PM_{2.5}$ and PM_{10}) and sensor (1-3) are provided in Table 2. Intercepts of all sensors and size fractions are relatively small and derived slope + intercept and slope only correspond very well.

Table 2 Regression coefficients (slope + intercept and slope only) for each particle size fraction (PM_1, $PM_{2.5}$ and PM_{10}) sensor (1-3) with associated linearity (R^2)

	Intercept	slope	R^2	slope only
PM_1_1	-0.64	1.33	0.994	1.29
PM_1_2	-0.46	1.21	0.995	1.18
PM_1_3	-0.56	1.26	0.995	1.23
$PM_{2.5}$_1	-0.42	0.71	0.998	0.71
$PM_{2.5}$_2	-0.11	0.64	0.999	0.64
$PM_{2.5}$_3	-0.32	0.66	0.998	0.65
PM_{10}_1	-1.45	0.32	0.992	0.32
PM_{10}_2	-0.73	0.29	0.995	0.28
PM_{10}_3	-1.08	0.29	0.993	0.29

When plotting all 1-minute averaged data of both sensor and reference during the lack-of-fit test, we observe an overall good linearity for $PM_{2.5}$ (R^2=0.98-0.99) and PM_{10} (R^2=0.94-0.96), but low accuracy with mean absolute errors (MAE) ranging between 9 and 11 µg/m³ for $PM_{2.5}$ and 77-80 µg/m³ for PM_{10}. The low accuracy is also reflected by the expanded uncertainty which varies between 79-82% for PM_{10}, and 42-49% for $PM_{2.5}$.[1]

[1] Note that the DQO (Data Quality Objectives) for 'indicative measurements' in the current Directive 2008/50/EC is 50%.

The comparability between the 3 sensors is very good (low variability between sensors), with a between-sensor-uncertainty of 1.52 µg/m³ for $PM_{2.5}$.

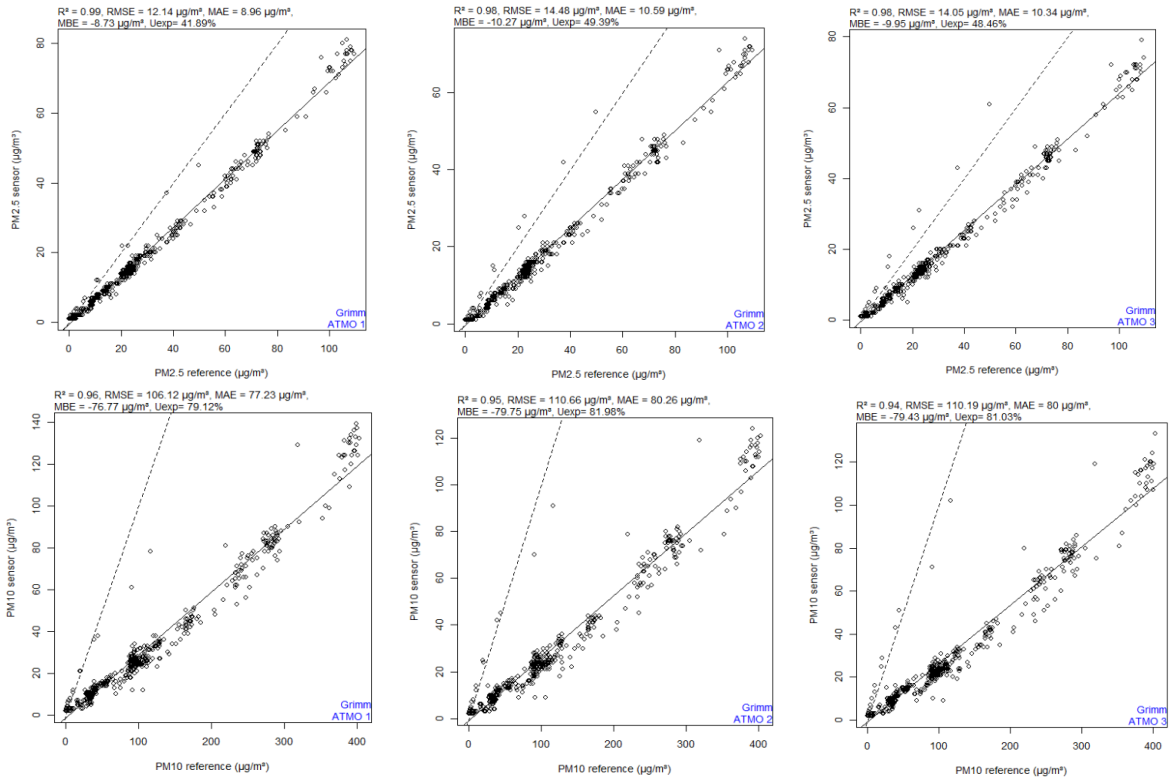
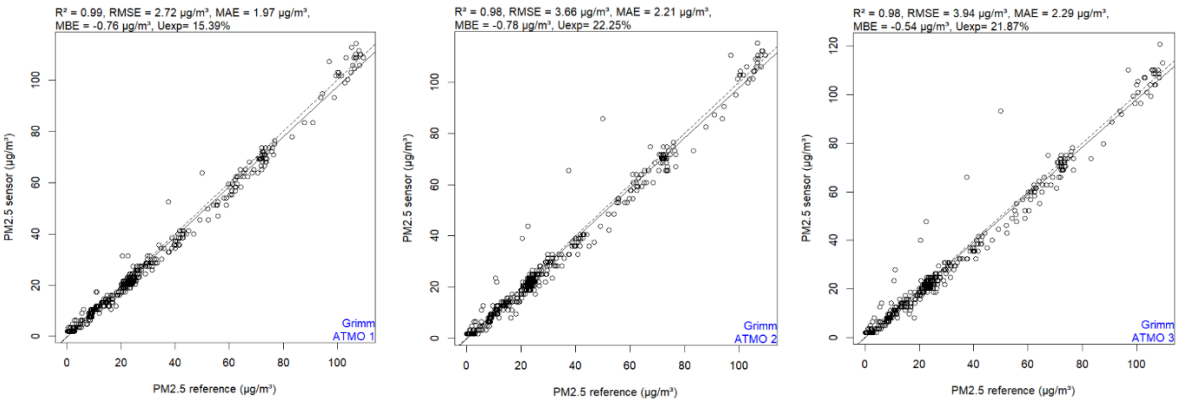


Figure 8 Scatterplots of 1-minute averaged reference (Grimm) and sensor (ATMO1-3) data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp) for $PM_{2.5}$ (upper) and PM_{10} (lower).

After applying a lab calibration based on the derived slopes and intercepts provided in Table 2 ($sensor_{cal}=(sensor_{raw}-b)/a$), the sensor accuracy (MAE) improved for $PM_{2.5}$ (MAE: ~2 µg/m³) and PM_{10} (MAE: 15-16 µg/m³) and the expanded uncertainty falls well below 50% for both $PM_{2.5}$ and PM_{10}. Note that both training (to derive slope and intercept) and test data are identical. Given the fact that training and test data are identical and the lab test was performed with a specific aerosol, we do not expect that the lab calibration will improve the field data.



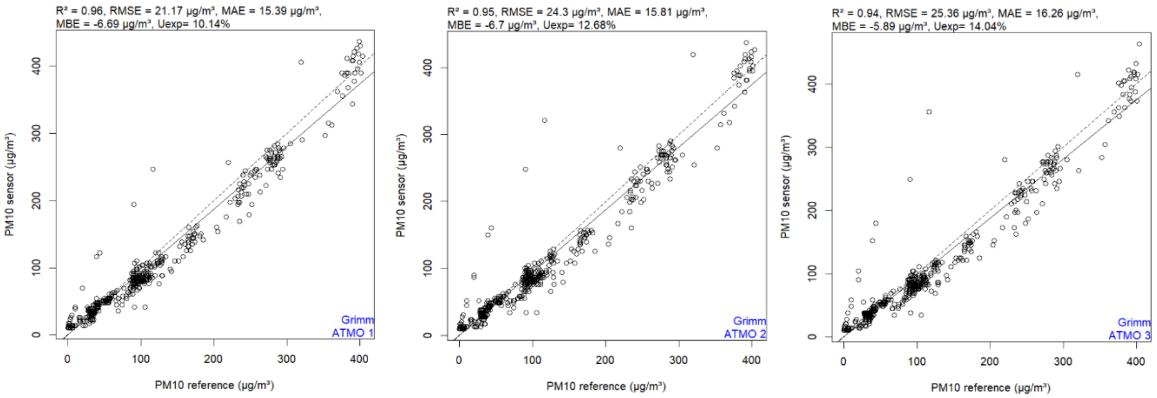


Figure 9 Scatterplots of 1-minute averaged reference (Grimm) and calibrated sensor (ATMO1-3) data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp) for $PM_{2.5}$ (upper) and PM_{10} (lower).

3.1.1.1.2 TERA PMscan

The raw TERA PMscan sensor measurements respond nicely to the increasing concentrations steps (Figure 10), resulting in a good linearity between sensor and reference (R^2>0.99 in Table 3). Sensor readings slightly underestimate the actual (Grimm) $PM_{2.5}$ concentrations. This sensor underestimation is more pronounced for PM_{10}, while PM_1 is significantly overestimating actual concentrations (Figure 11). Mean setpoint accuracy (mean of different accuracies at each setpoint) varied between 12-28% for PM_1, 76-84% for $PM_{2.5}$ and 45-51 for PM_{10}.

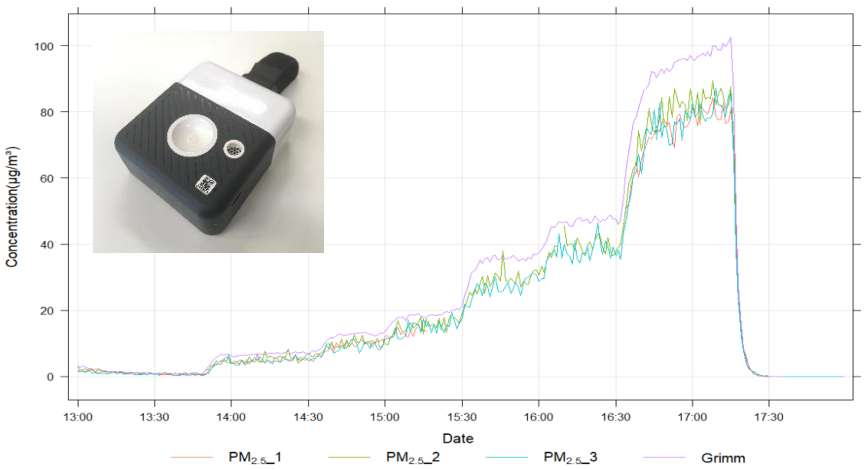


Figure 10 $PM_{2.5}$ concentrations generated during the lack-of-fit test as measured by the 3 TERA PMscan sensors (1-3) and the reference monitor (Grimm)

PM1				PM2.5				PM10				
PM1_1	PM1_2	PM1_3	PM1_REF	PM2.5_1	PM2.5_2	PM2.5_3	PM2.5_REF	PM10_1	PM10_2	PM10_3	PM10_REF	setpoints
0.86	0.73	0.71	0.46	1.36	1.38	1.23	1.38	2.45	2.32	1.68	2.21	0
3.26	3.24	2.84	2.09	5.01	5.83	5.11	7.17	9.89	11.07	10.79	23.08	30
6.34	6.01	5.31	3.46	9.91	10.69	9.44	12.89	18.55	21.50	18.59	40.63	40
	8.93	8.09	4.84		16.13	15.60	19.07		30.69	31.05	61.58	60
	16.58	15.53	8.30		29.96	27.95	35.85		62.17	52.26	122.65	130
	21.86	20.24	10.49		39.61	38.47	47.50		82.59	75.54	166.28	200
49.69	46.42	43.46	22.42	79.64	83.67	80.72	98.20	154.77	171.19	148.77	324.95	350

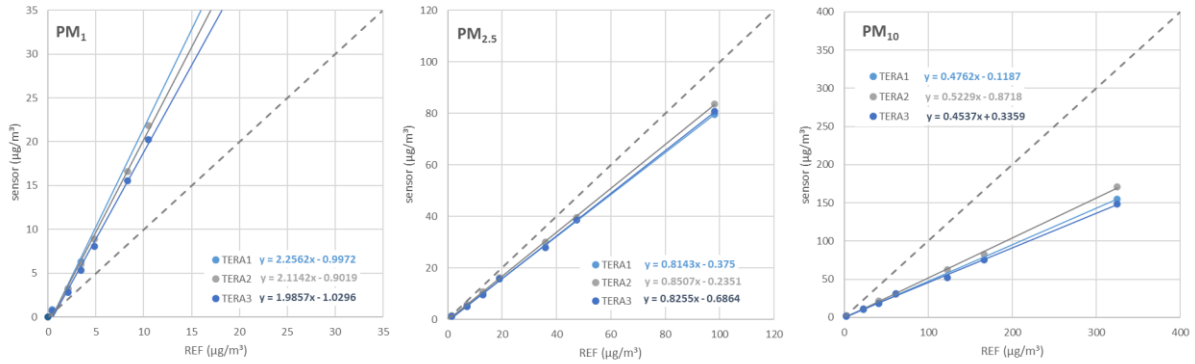


Figure 11 Setpoint averages (upper; µg/m³) and resulting linearity plots (lower) for PM_1, $PM_{2.5}$ and PM_{10}.

Derived regression coefficients (slope + intercept (y=a*x+b) and slope only (y=a*x)) and linearity (R^2) for each size fraction (PM_1, $PM_{2.5}$ and PM_{10}) and sensor (1-3) are provided in Table 3.

Table 3 Regression coefficients (slope + intercept and slope only) for each particle size fraction (PM_1, $PM_{2.5}$ and PM_{10}) sensor (1-3) with associated linearity (R^2)

	intercept	slope	R^2	slope only
PM1_1	-1.00	2.26	0.999	2.20
PM1_2	-0.90	2.11	0.999	2.05
PM1_3	-1.03	1.99	0.999	1.91
PM2.5_1	-0.38	0.81	1.000	0.81
PM2.5_2	-0.24	0.85	1.000	0.85
PM2.5_3	-0.69	0.83	1.000	0.81
PM10_1	-0.12	0.48	1.000	0.48
PM10_2	-0.87	0.52	0.999	0.52
PM10_3	0.34	0.45	0.999	0.46

When plotting all 1-minute averaged data of both sensor and reference during the lack-of-fit test (Figure 12), we observe an overall good linearity for $PM_{2.5}$ (R^2=0.99-1) and PM_{10} (R^2=0.98-0.99), but varying accuracy with mean absolute errors (MAE) ranging between 4-6 µg/m³ for $PM_{2.5}$ and 45-53 µg/m³ for PM_{10}. Expanded uncertainty (Uexp) falls below 34% for $PM_{2.5}$, qualifying the sensors for indicative (class 1 sensors) measurements. For PM_{10}, expanded uncertainty is much higher (60-72%).

The comparability between the 3 sensors is good (low variability between sensors), with a between-sensor-uncertainty of 1.64 µg/m³ for $PM_{2.5}$.

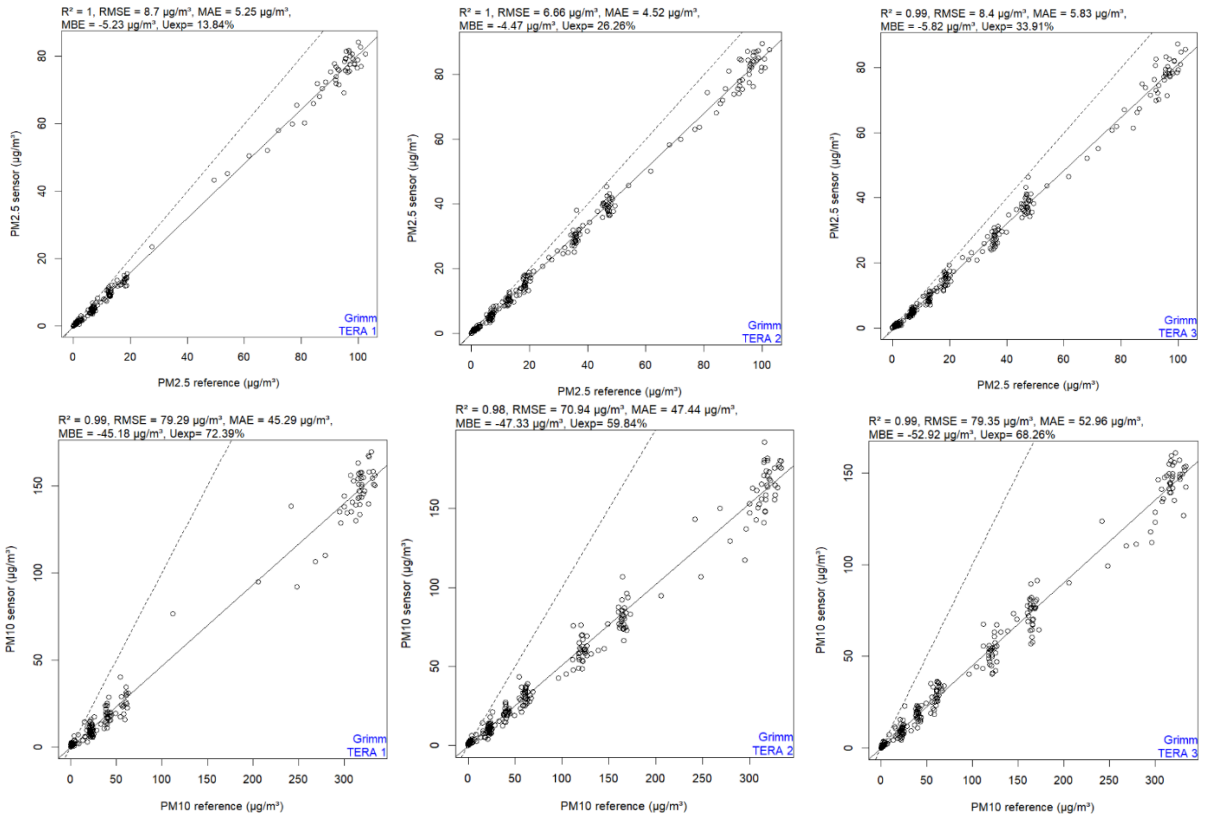
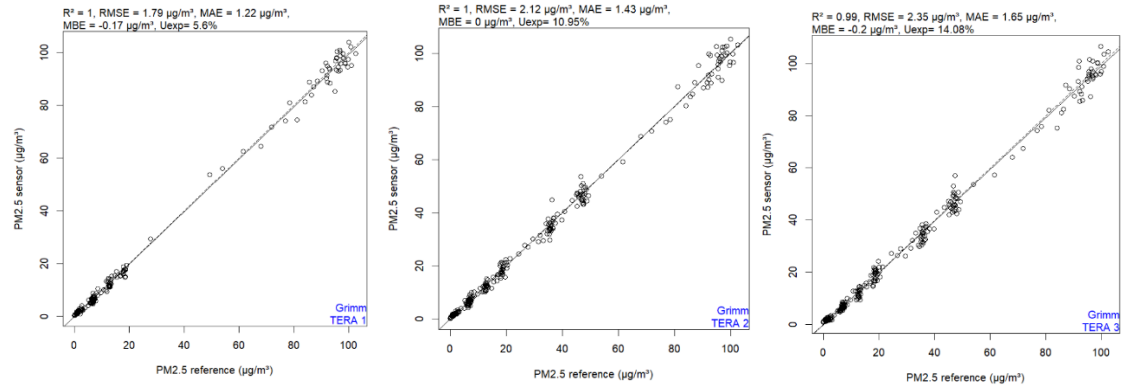


Figure 12 Scatterplots of 1-minute averaged reference (Grimm) and sensor (TERA1-3) data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp) for $PM_{2.5}$ (upper) and PM_{10} (lower).

After applying a lab calibration based on the derived slopes and intercepts provided in Table 3 ($sensor_{cal}=(sensor_{raw}-b)/a$)), the sensor accuracy (MAE) improved for $PM_{2.5}$ (MAE: <2 µg/m³) and PM_{10} (MAE: 8-9 µg/m³) and the expanded uncertainty falls well below 50% for both $PM_{2.5}$ and PM_{10} (Figure 13). Note that both training (to derive slope and intercept) and test data are identical.



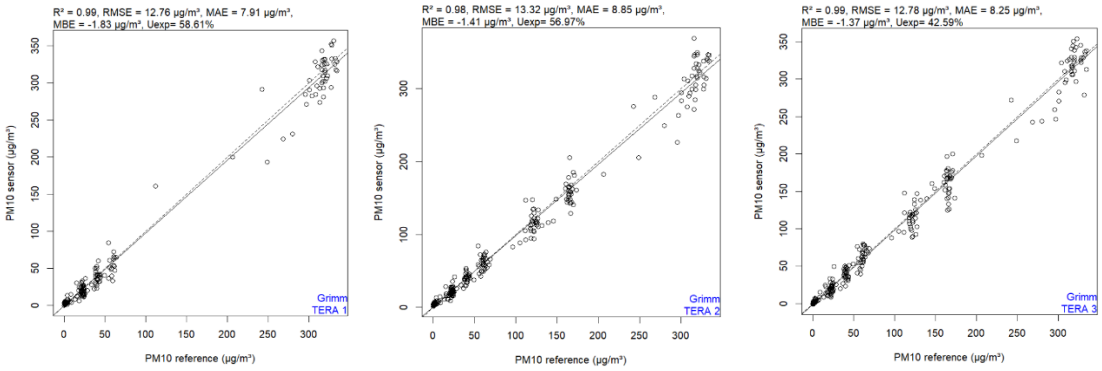


Figure 13 Scatterplots of 1-minute averaged reference (Grimm) and calibrated sensor (ATMO1-3) data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp) for $PM_{2.5}$ (upper) and PM_{10} (lower).

3.1.1.1.3 Open Seneca

The raw Open Seneca sensor measurements respond nicely to the increasing concentrations steps (Figure 14), resulting in a good linearity between sensor and reference (R^2>0.99 in Table 4). Sensor readings slightly underestimate the actual (Grimm) $PM_{2.5}$ concentrations. This sensor underestimation is more pronounced for PM_{10}, while PM_1 is slightly overestimating actual concentrations (Figure 11). Mean setpoint accuracy (mean of different accuracies at each setpoint) varied between 80-86% for PM_1, 53-56% for $PM_{2.5}$ and 22-23 for PM_{10}.

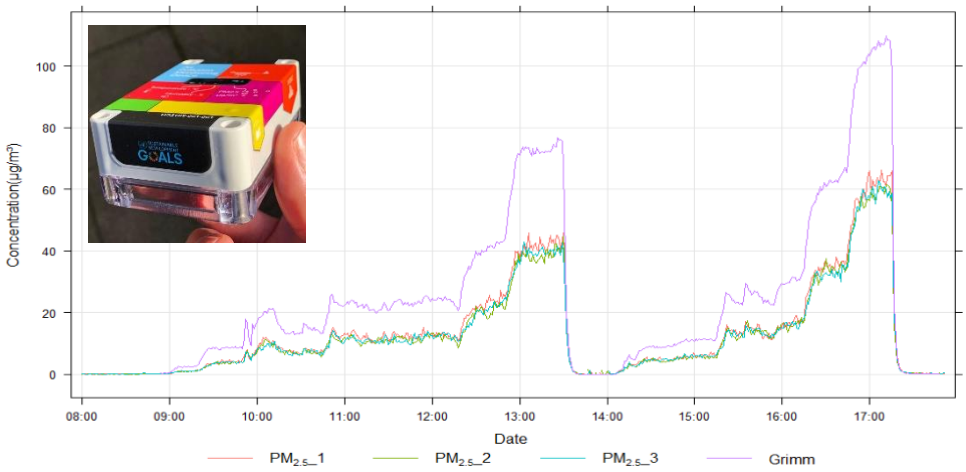


Figure 14 $PM_{2.5}$ concentrations generated during the lack-of-fit test as measured by the 3 Open Seneca sensors (1-3) and the reference monitor (Grimm)

PM_1				$PM_{2.5}$				PM_{10}				
PM_1_1	PM_1_2	PM_1_3	PM_1_REF	$PM_{2.5}$_1	$PM_{2.5}$_2	$PM_{2.5}$_3	$PM_{2.5}$_REF	PM_{10}_1	PM_{10}_2	PM_{10}_3	PM_{10}_REF	setpoints
0.25	0.20	0.18	0.21	0.38	0.30	0.28	0.26	0.54	0.42	0.39	0.42	0
2.94	2.77	2.75	2.51	4.70	4.42	4.74	8.90	6.70	6.28	7.01	30.56	30
3.49	3.25	3.36	3.03	6.21	5.57	5.98	11.24	9.34	8.23	9.01	40.46	60
7.37	6.92	7.09	6.03	14.71	13.90	13.99	25.82	23.28	22.05	22.03	102.33	110
8.16	7.80	8.04	6.70	16.77	16.20	16.08	30.64	26.86	26.05	25.48	126.39	160
16.89	16.34	16.22	13.98	35.74	34.48	33.04	63.38	57.88	55.80	52.73	248.61	250
29.75	28.06	28.66	23.90	62.96	59.43	59.15	106.78	101.97	96.29	94.90	395.38	400

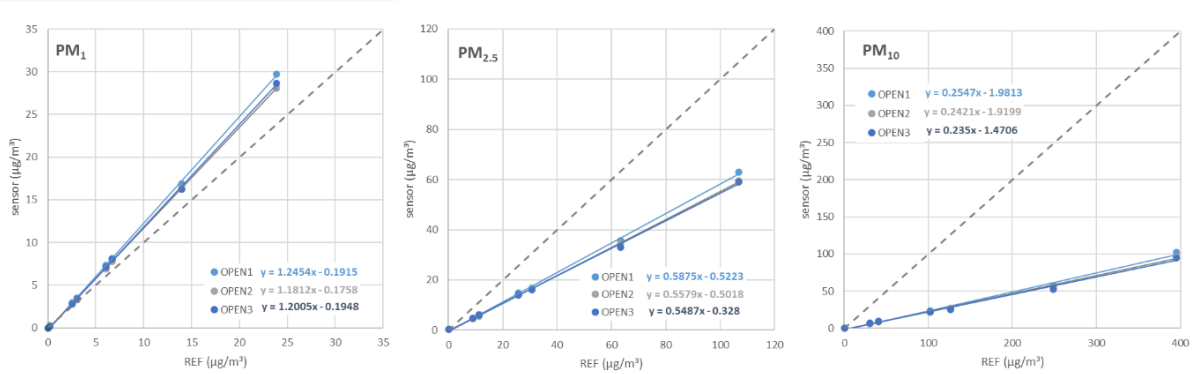


Figure 15 Setpoint averages (upper; µg/m³) and resulting linearity plots (lower) for PM_1, $PM_{2.5}$ and PM_{10}.

Derived regression coefficients (slope + intercept (y=a*x+b) and slope only (y=a*x)) and linearity (R^2) for each size fraction (PM_1, $PM_{2.5}$ and PM_{10}) and sensor (1-3) are provided in Table 4.

Table 4 Regression coefficients (slope + intercept and slope only) for each particle size fraction (PM_1, $PM_{2.5}$ and PM_{10}) sensor (1-3) with associated linearity (R^2)

	intercept	slope	R^2	slope only
PM1_1	-0.19	1.25	1.000	1.2329
PM1_2	-0.18	1.18	1.000	1.1697
PM1_3	-0.19	1.20	1.000	1.1878
PM2.5_1	-0.52	0.59	0.999	0.58
PM2.5_2	-0.50	0.56	1.000	0.55
PM2.5_3	-0.33	0.55	0.999	0.54
PM10_1	-1.98	0.25	0.995	0.25
PM10_2	-1.92	0.24	0.996	0.23
PM10_3	-1.47	0.23	0.993	0.23

When plotting all 1-minute averaged data of both sensor and reference during the lack-of-fit test (Figure 16), we observe an overall good linearity for $PM_{2.5}$ (R^2=0.99) and PM_{10} (R^2=0.96-0.97), but lower accuracy with mean absolute errors (MAE) ranging between 12-13 µg/m³ for $PM_{2.5}$ and 84-87 µg/m³ for PM_{10}. Expanded uncertainty (Uexp) ranges between 50-57% for $PM_{2.5}$ and 88-90% for PM_{10}.

The comparability between the 3 sensors is good (low variability between sensors), with a between-sensor-uncertainty of 1.21 µg/m³ for $PM_{2.5}$.

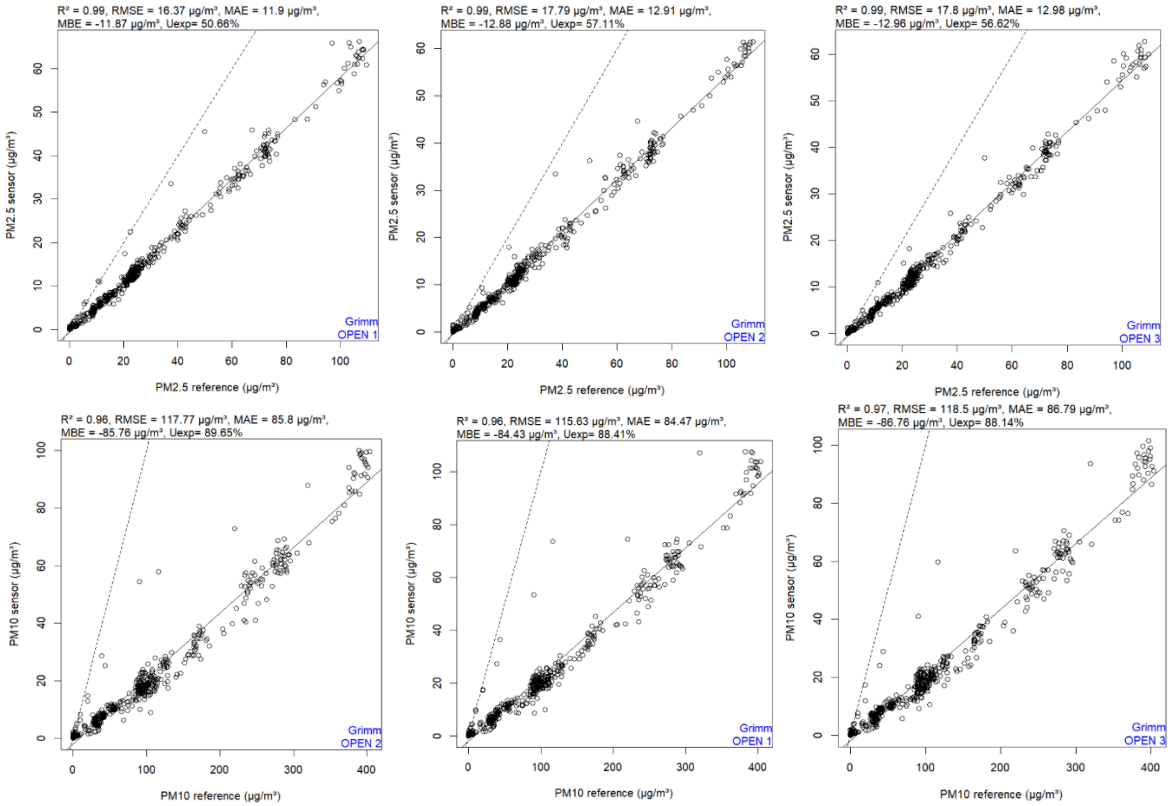


Figure 16 Scatterplots of 1-minute averaged reference (Grimm) and sensor (OPEN1-3) data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp) for $PM_{2.5}$ (upper) and PM_{10} (lower).

After applying a lab calibration based on the derived slopes and intercepts provided in Table 3 ($sensor_{cal}=(sensor_{raw}-b)/a$), the sensor accuracy (MAE) improved for $PM_{2.5}$ (MAE: <2 µg/m³) and PM_{10} (MAE: 13-15 µg/m³) and the expanded uncertainty falls well below 50% for both $PM_{2.5}$ and PM_{10} (Figure 17). Note that both training (to derive slope and intercept) and test data are identical.

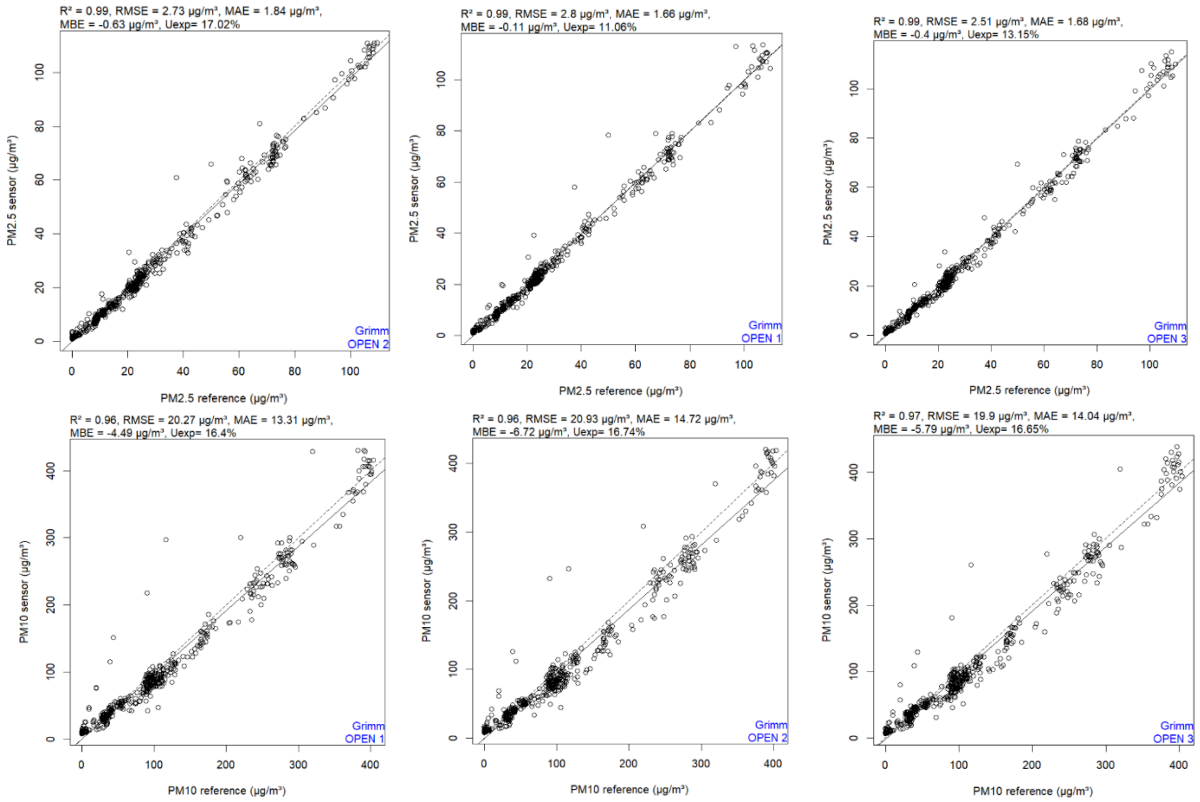


Figure 17 Scatterplots of 1-minute averaged reference (Grimm) and calibrated sensor (ATMO1-3) data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp) for $PM_{2.5}$ (upper) and PM_{10} (lower).

3.1.1.1.4 SODAQ Air

The raw SODAQ Air exhibits a 5 minute resolution when stationary and changes automatically to ~10 seconds when mobile. This results in a lower monitoring resolution, when compared to the other sensor systems (Figure 18). Although the sensors respond nicely to the increasing concentrations steps, the observed variability between the sensors is much larger when compared to the other sensor systems (Figure 18). Sensor readings seem to underestimate the actual (Grimm) $PM_{2.5}$ concentrations. This sensor underestimation is more pronounced for PM_{10}, while PM_1 is overestimating actual concentrations (Figure 19). Mean setpoint accuracy (mean of different accuracies at each setpoint) varied between 31-94% for PM_1, 48-95% for $PM_{2.5}$ and 20-43 for PM_{10}.

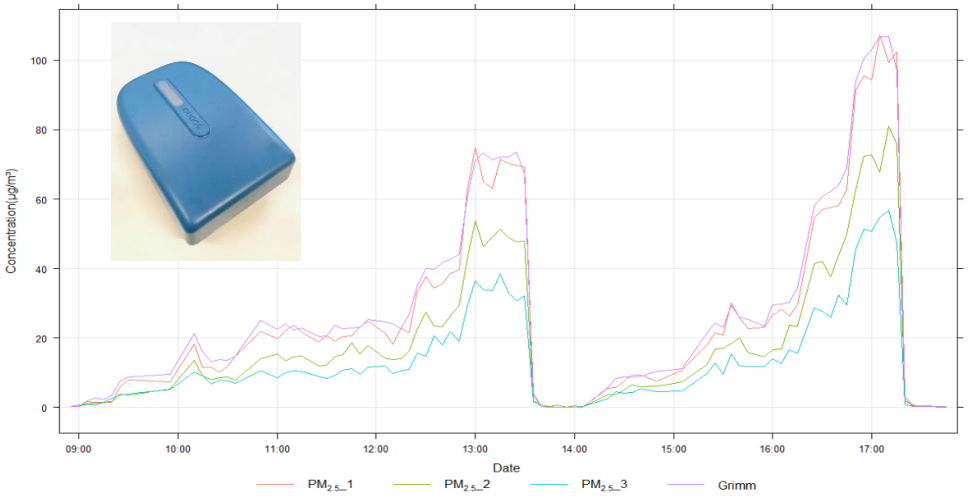


Figure 18 $PM_{2.5}$ concentrations generated during the lack-of-fit test as measured by the 3 SODAQ Air sensors (1-3) and the reference monitor (Grimm)

PM1				PM2.5				PM10				
PM1_1	PM1_2	PM1_3	PM1_REF	PM2.5_1	PM2.5_2	PM2.5_3	PM2.5_REF	PM10_1	PM10_2	PM10_3	PM10_REF	setpoints
0.25	0.21	0.19	0.23	0.30	0.28	0.24	0.29	0.34	0.34	0.27	0.57	0
3.84	3.18	2.30	2.51	8.71	5.84	4.55	8.78	14.45	8.94	7.19	29.62	30
4.85	4.02	3.40	2.97	10.69	7.36	4.69	11.26	17.55	11.24	6.06	37.20	60
10.58	8.08	6.48	6.11	25.64	18.47	12.33	26.22	43.49	30.82	19.15	104.66	110
11.64	8.82	6.52	6.56	27.05	18.98	14.37	29.86	45.24	30.99	23.68	124.38	160
23.22	17.47	13.32	13.75	57.60	41.16	28.70	62.38	98.41	69.40	46.91	250.92	250
41.52	32.23	24.40	23.64	100.30	73.83	54.15	105.63	169.93	123.33	89.45	393.36	400

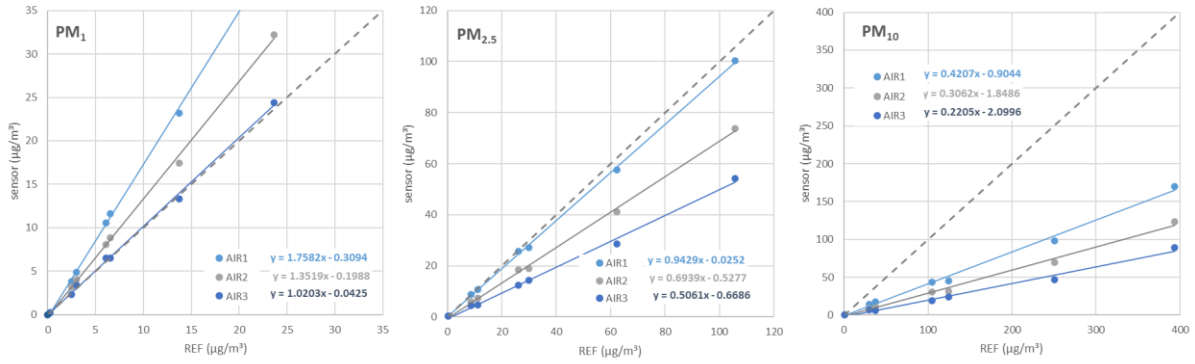


Figure 19 Setpoint averages (upper; µg/m³) and resulting linearity plots (lower) for PM_1, $PM_{2.5}$ and PM_{10}.

Derived regression coefficients (slope + intercept (y=a*x+b) and slope only (y=a*x)) and linearity (R^2) for each size fraction (PM_1, $PM_{2.5}$ and PM_{10}) and sensor (1-3) are provided in Table 5.

Table 5 Regression coefficients (slope + intercept and slope only) for each particle size fraction (PM_1, $PM_{2.5}$ and PM_{10}) sensor (1-3) with associated linearity (R^2)

	intercept	slope	R^2	slope only
PM1_1	-0.31	1.76	0.999	2.2016
PM1_2	-0.20	1.35	0.998	2.0492
PM1_3	-0.04	1.02	0.998	1.9114
PM2.5_1	-0.03	0.94	0.999	0.81

PM2.5_2	-0.53	0.69	0.998	0.85
PM2.5_3	-0.67	0.51	0.996	0.81
PM10_1	-0.90	0.42	0.994	0.48
PM10_2	-1.85	0.31	0.992	0.52
PM10_3	-2.10	0.22	0.986	0.46

When plotting all 1-minute averaged data of both sensor and reference during the lack-of-fit test (Figure 20), we observe an overall good linearity for $PM_{2.5}$ (R^2=0.99) and PM_{10} (R^2=0.96-0.97), and a wide variety in accuracy with mean absolute errors (MAE) ranging between 2-15 µg/m³ for $PM_{2.5}$ and 68-91 µg/m³ for PM_{10}. Expanded uncertainty (Uexp) ranges between 17-61% for $PM_{2.5}$ and 68-83% for PM_{10}.

The comparability between the 3 sensors is low (high variability between sensors) when compared to the other sensor systems, with a between-sensor-uncertainty of 3.96 µg/m³ for $PM_{2.5}$.

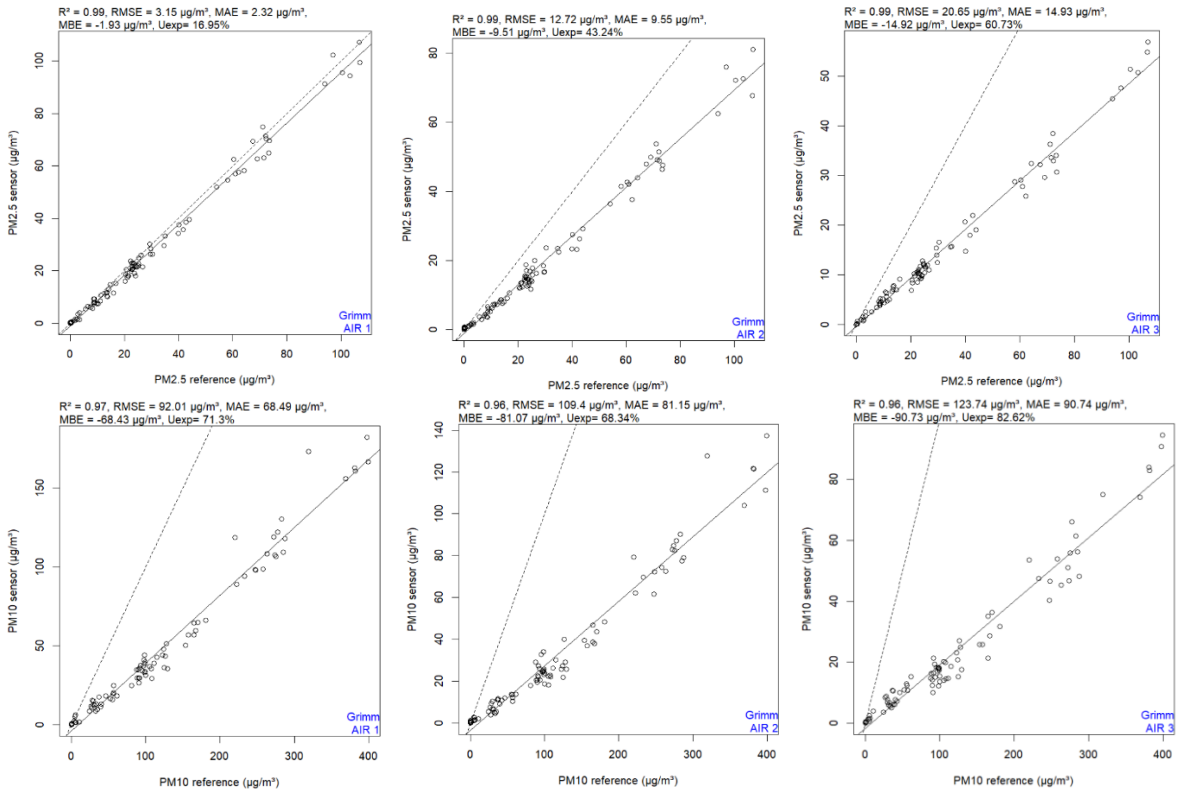


Figure 20 Scatterplots of 1-minute averaged reference (Grimm) and sensor (ATMO1-3) data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp) for $PM_{2.5}$ (upper) and PM_{10} (lower).

After applying a lab calibration based on the derived slopes and intercepts provided in Table 5 ($sensor_{cal}$=($sensor_{raw}$-b)/a)), the sensor accuracy (MAE) improved for $PM_{2.5}$ (MAE: ~2 µg/m³) and PM_{10} (MAE: 12-17 µg/m³) and the expanded uncertainty falls well below 50% for both $PM_{2.5}$ and PM_{10} (Figure 21). Note that both training (to derive slope and intercept) and test data are identical.

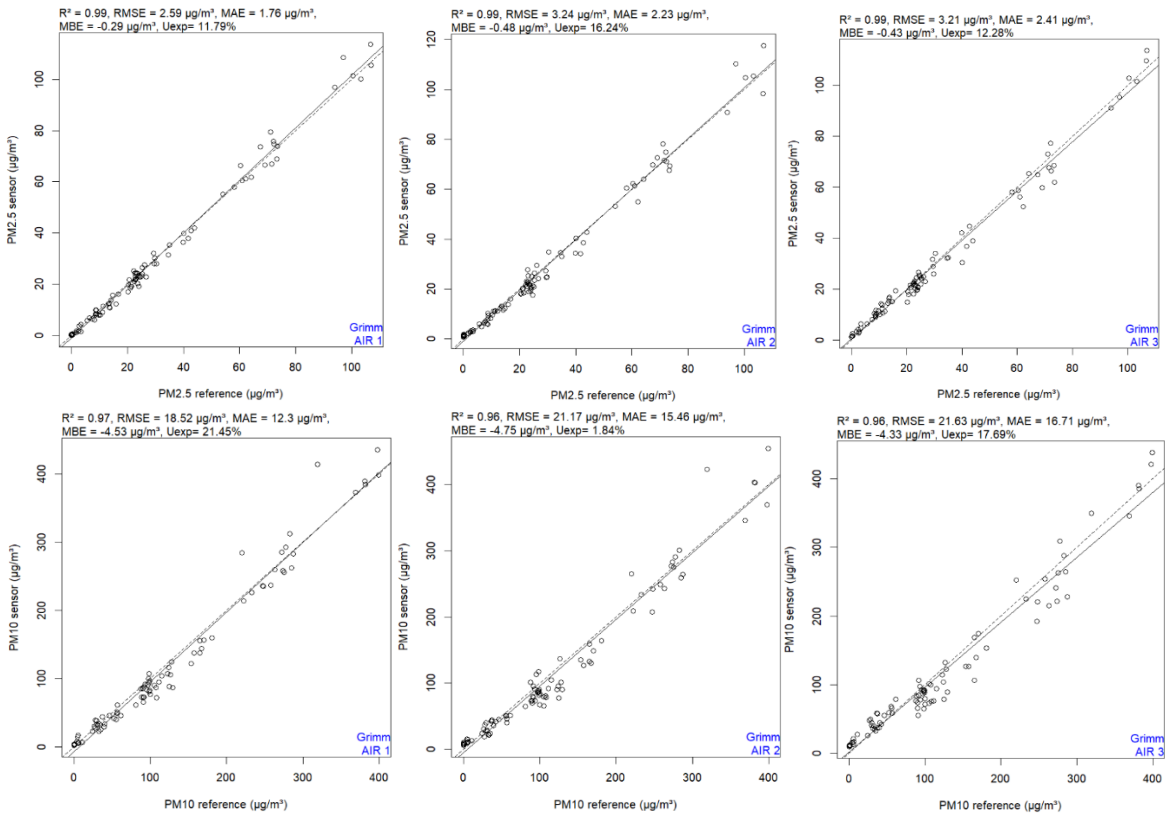


Figure 21 Scatterplots of 1-minute averaged reference (Grimm) and calibrated sensor (AIR1-3) data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp) for $PM_{2.5}$ (upper) and PM_{10} (lower).

3.1.1.1.5 SODAQ NO_2

Just like the SODAQ Air, the SODAQ NO2 (prototype) showed a 5 minute resolution as well when connected. We, however, noticed significant connectivity issues resulting in a very low data coverage (sensor 1>2>3) during the lack-of-fit test (Figure 22). Potential explanations might be connectivity issues within the exposure chamber (Although not observed for other sensors systems relying on GPRS/4G) or electromagnetic interferences of the hardware with other sensors or lab equipment.

To cope with the connectivity issues, we calculated setpoint averages, regression coefficients and accuracies separately for each sensor (Figure 23). Based on the available data, we observed similar sensor behaviour as the SODAQ Air (good linearity, large variation in accuracy due to large between-sensor-uncertainty). This shouldn't surprise as the same PM sensor (Sensirion SPS30) is included in the sensor box. Mean setpoint accuracy varied between 60-77% for PM_1, 35-70% for $PM_{2.5}$ and 13-29 for PM_{10}.

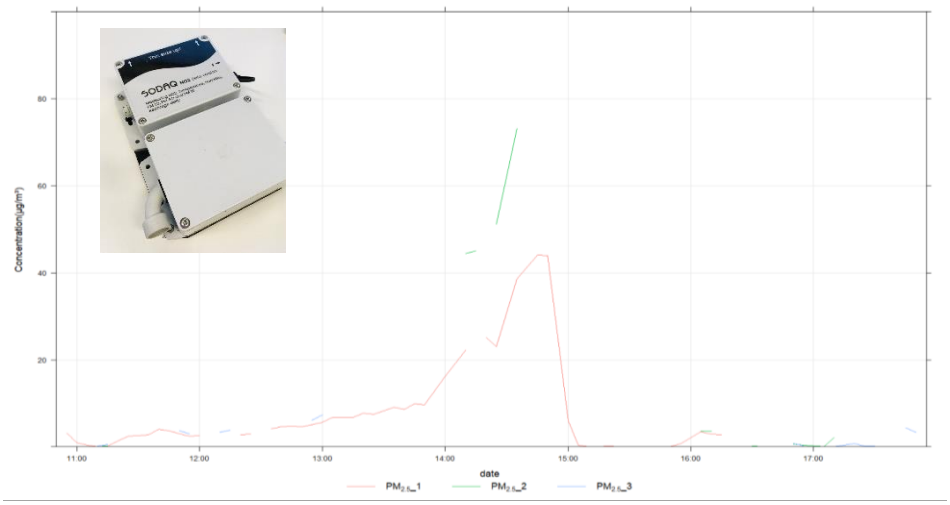


Figure 22 $PM_{2.5}$ concentrations generated during the lack-of-fit test as measured by the 3 SODAQ NO_2 sensors

NO2_1						
PM_1_1	**PM_1_REF**	**$PM_{2.5}$_1**	**$PM_{2.5}$_REF**	**PM_{10}_1**	**PM_{10}_REF**	**setpoints**
0.04	0.00	0.07	0.01	0.11	0.13	0
2.52	3.44	3.07	8.11	3.40	25.52	30
2.34	3.58	2.82	8.95	3.09	27.79	40
3.42	4.54	4.84	14.66	6.09	52.59	60
4.68	5.96	7.31	22.26	9.84	86.73	110
5.91	7.05	9.79	27.97	13.66	113.40	130
12.42	14.48	23.50	61.67	35.18	229.12	200
21.18	25.41	44.03	108.30	68.71	379.41	300
NO2_2						
PM_1_1	**PM_1_REF**	**$PM_{2.5}$_1**	**$PM_{2.5}$_REF**	**PM_{10}_1**	**PM_{10}_REF**	**setpoints**
0.00	0.00	0.00	0.00	0.00	0.00	0
3.32	3.37	4.71	8.54	6.18	26.09	30
						40
						60
						110
11.06	7.15	20.85	28.52	32.04	115.19	130
23.71	14.36	49.13	61.76	78.52	228.20	200
						300

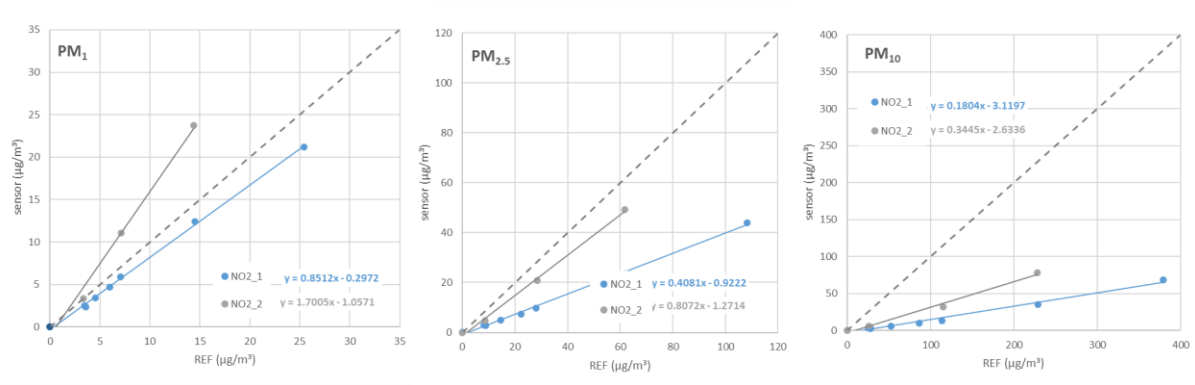


Figure 23 Setpoint averages for sensor 1 and 2 (upper; µg/m³) and resulting linearity plots (lower) for PM_1, $PM_{2.5}$ and PM_{10}. Sensor 3 didn't collect any data during the lack-of-fit experiment

Derived regression coefficients (slope + intercept (y=a*x+b) and slope only (y=a*x)) and linearity (R^2) for each size fraction (PM_1, $PM_{2.5}$ and PM_{10}) and sensor (1-3) are provided in Table 6.

Table 6 Regression coefficients (slope + intercept and slope only) for each particle size fraction (PM_1, $PM_{2.5}$ and PM_{10}) sensor (1-2) with associated linearity (R^2)

	intercept	slope	R^2	slope only
PM1_1	-0.30	0.85	0.998	0.8318
PM1_2	-1.03	1.70	0.991	1.6027
PM2.5_1	-0.92	0.41	0.998	0.3946
PM2.5_2	-1.24	0.81	0.997	0.78
PM10_1	-3.12	0.18	0.985	0.17
PM10_2	-2.60	0.34	0.990	0.33

When plotting all 1-minute averaged data of both sensor and reference during the lack-of-fit test (Figure 24), we observe an overall good linearity for $PM_{2.5}$ (R^2=0.99) and PM_{10} (R^2=0.96-0.99), and mean absolute errors (MAE) ranging between 8-11 µg/m³ for $PM_{2.5}$ and 68-91 µg/m³ for PM_{10}. Expanded uncertainty (Uexp) ranges between 17-61% for $PM_{2.5}$ and 55-89% for PM_{10}.

The comparability between the 2 sensors is low (high variability between sensors) when compared to the other sensor systems. Between-sensor-uncertainty could not be calculated as no simultaneous sensor data was collected.

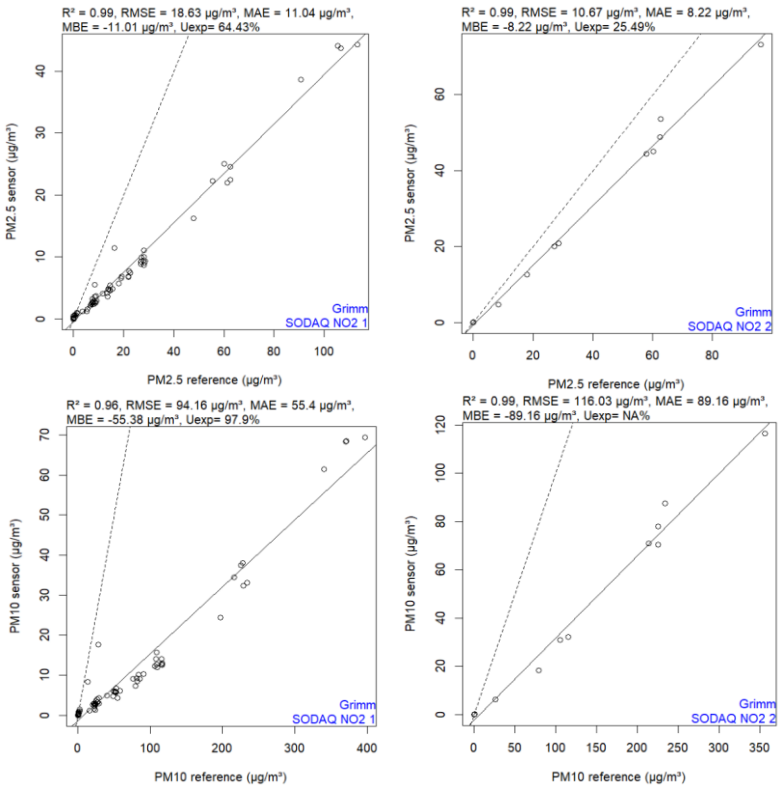


Figure 24 Scatterplots of 1-minute averaged reference (Grimm) and sensor (SODAQ NO2 1-2) data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp) for $PM_{2.5}$ (upper) and PM_{10} (lower).

After applying a lab calibration based on the derived slopes and intercepts provided in Table 6 ($sensor_{cal}$=($sensor_{raw}$-b)/a)), the sensor accuracy (MAE) improved for $PM_{2.5}$ (MAE: ~2 µg/m³) and PM_{10} (MAE: 10-16 µg/m³) and the expanded uncertainty falls well below 50% for both $PM_{2.5}$ and PM_{10} (Figure 25). Note that both training (to derive slope and intercept) and test data are identical.

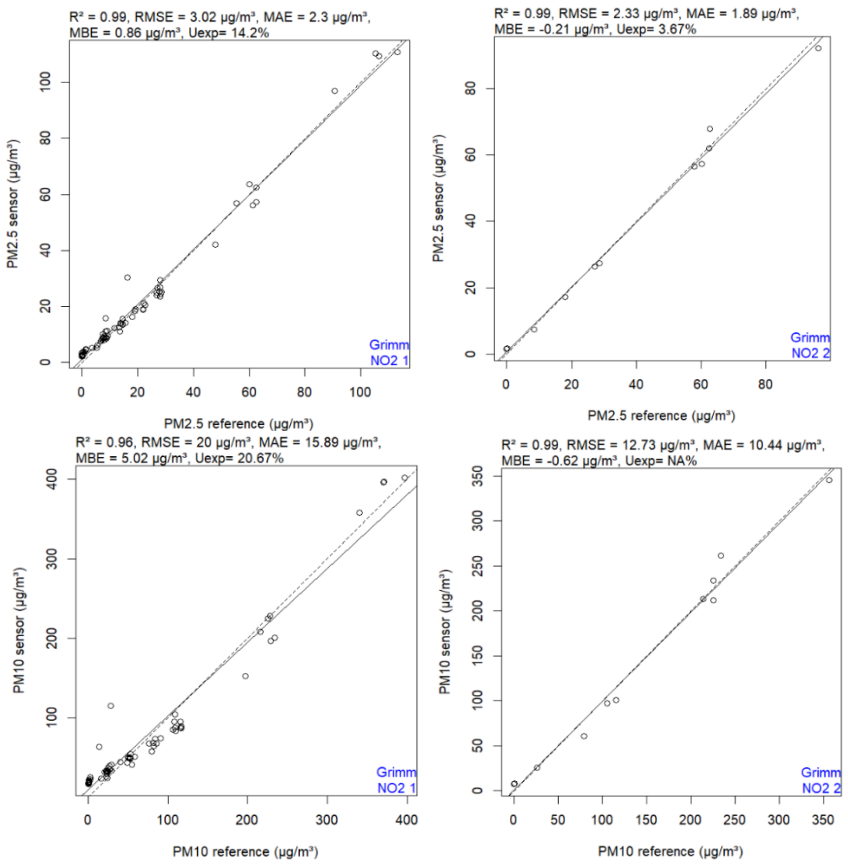


Figure 25 Scatterplots of 1-minute averaged reference (Grimm) and calibrated sensor (NO2 1-2) data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp) for $PM_{2.5}$ (upper) and PM_{10} (lower).

3.1.1.1.6 2BTech PAM

The raw 2BTech PAM data (for which only one sensor system was purchased) responded to the increasing concentration steps, but exhibited more signal noise when compared the other sensor systems (Figure 26). Sensor readings underestimate the actual (Grimm) concentrations for all size fractions in the order; PM_1<$PM_{2.5}$<PM_{10} (Figure 27). Mean setpoint accuracy was 63% for PM_1, 29% for $PM_{2.5}$ and 13% for PM_{10}.

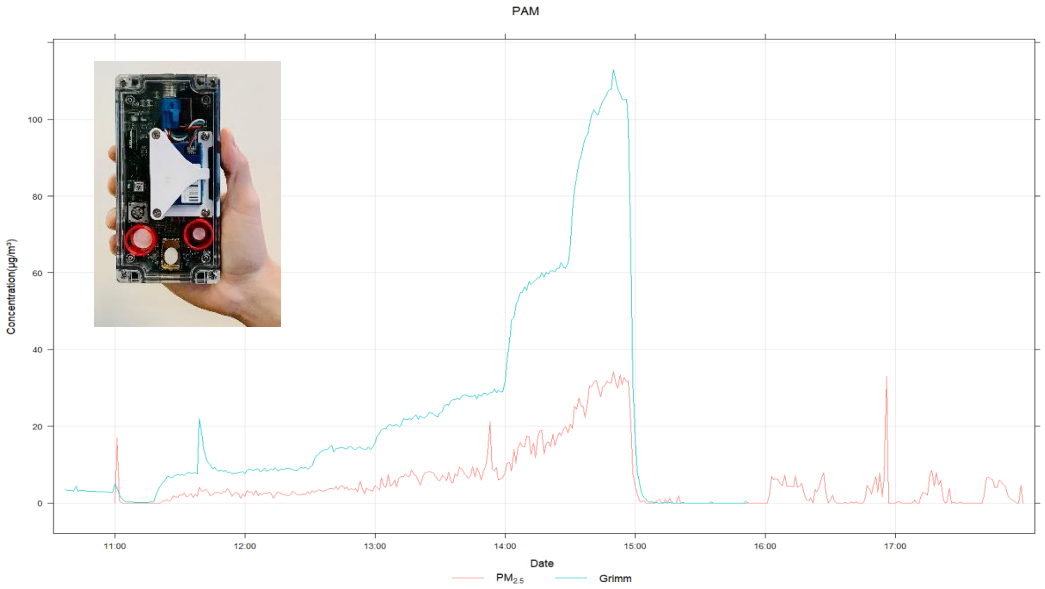


Figure 26 $PM_{2.5}$ concentrations generated during the lack-of-fit test as measured by the 2BTech PAM sensor and the reference monitor (Grimm)

PM_1				$PM_{2.5}$				PM_{10}				
PM_1_1	PM_1_2	PM_1_3	PM_1_REF	$PM_{2.5}$_1	$PM_{2.5}$_2	$PM_{2.5}$_3	$PM_{2.5}$_REF	PM_{10}_1	PM_{10}_2	PM_{10}_3	PM_{10}_REF	setpoints
0.20			0.00	0.20			0.01	0.20			0.08	0
1.79			3.49	2.44			8.16	3.86			25.96	30
1.93			3.65	2.37			8.84	4.18			28.18	40
2.76			4.52	3.70			14.35	6.52			51.32	60
4.75			6.01	7.04			22.59	10.90			89.03	110
6.17			7.15	9.04			28.69	14.42			115.07	130
8.55			14.34	17.34			60.87	29.06			228.86	200
13.66			25.12	31.58			107.13	53.07			376.62	300

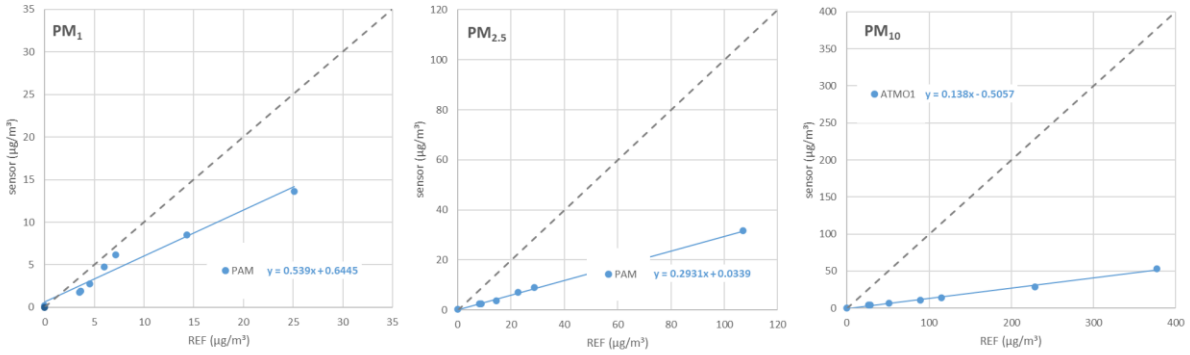


Figure 27 Setpoint averages (upper; µg/m³) and resulting linearity plots (lower) for PM_1, $PM_{2.5}$ and PM_{10}.

Derived regression coefficients (slope + intercept (y=a*x+b) and slope only (y=a*x)) and linearity (R^2) for each size fraction (PM_1, $PM_{2.5}$ and PM_{10}) and sensor are provided in Table 7.

Table 7 Regression coefficients (slope + intercept and slope only) for each particle size fraction (PM_1, $PM_{2.5}$ and PM_{10}) sensor (1-3) with associated linearity (R^2)

	intercept	slope	R^2	slope only
PM_1	0.644	0.539	0.962	0.5817
$PM_{2.5}$	0.034	0.293	0.998	0.2936

PM_{10}	-0.506	0.138	0.995	0.1359

When plotting all 1-minute averaged data of both sensor and reference during the lack-of-fit test (Figure 28), we observe a slightly lower linearity for $PM_{2.5}$ (R^2=0.96) and PM_{10} (R^2=0.95) due to exhibited noise, and fairly low accuracy with mean absolute errors (MAE) of 17 µg/m³ for $PM_{2.5}$ and 78 µg/m³ for PM_{10}. Expanded uncertainty (Uexp) is 79% for $PM_{2.5}$ and 96% for PM_{10}.

The between-sensor-uncertainty could not be evaluated because we only had one sensor system available for evaluation.

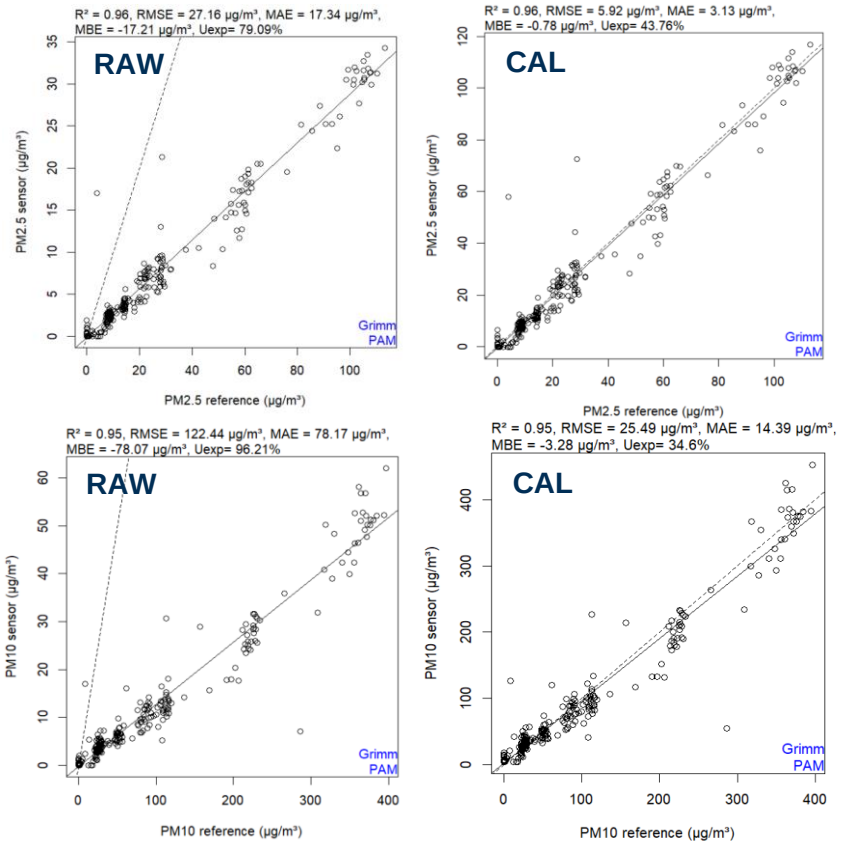


Figure 28 Scatterplots of 1-minute averaged reference (Grimm) and sensor (PAM) data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp) for both raw (left) and calibrated (right) $PM_{2.5}$ (upper) and PM_{10} (lower) data.

After applying a lab calibration based on the derived slopes and intercepts provided in Table 7($sensor_{cal}$=($sensor_{raw}$-b)/a)), the sensor accuracy (MAE) improved for $PM_{2.5}$ (MAE: 3 µg/m³) and PM_{10} (MAE: 14 µg/m³) and the expanded uncertainty falls just below 50% for both $PM_{2.5}$ and PM_{10} (Figure 28). Note that both training (to derive slope and intercept) and test data are identical.

3.1.1.1.7 GeoAir

The GeoAir experienced power supply issues during the lack-of-fit measurements (insufficient amperage from applied USB hubs), resulting in data loss for all sensors. During the coarse

testing, we noticed that only one GeoAir (powered via separate power supply) captured data. We, therefore, rely on the coarse test (1 sensor), mobile test (3 sensors) and the performance results from the field campaign (3 sensors).

3.1.1.2 Coarse test

In previous studies and benchmarking projects (e.g. VAQUUMS), low-cost sensors have showed low sensitivity in the coarse (2.5-10 µm) PM fraction. At VITO, we developed a test procedure to evaluate sensor sensitivity in the coarse fraction. This procedure is currently under debate as well within the CEN WG42 on data quality objectives for sensors.

We expose the sensors to monodisperse dust (silica microspheres) of consecutively 7.750 µm and 1.180 µm (fine) diameters. We experimented with an aerosolizer to reach representative (~100-150 µg/m³) PM_{10} concentrations by generating dust pulses every 30 seconds during a 5 minute period. The idea is to simulate conditions with mainly fine ('Fine test cond.') and mainly coarse aerosol ('Coarse test cond.') respectively. Two representative 5-minute periods (1 coarse test, 1 fine test) were subsequently selected and evaluated by calculating the dust composition (% coarse), PM_{10}, $PM_{2.5}$ and PM_{coarse} sensor/REF ratios and 2 relative change metrics:

- the relative change in sensor/REF ratio between the two tests conditions (fine and coarse) (%)
- the relative change in PM_{10} sensor/REF ratio between fine and coarse test conditions (%)

As can be seen from Figure 29, the generated dust composition is clearly different for the coarse test (mainly composed of coarse-size particles and small amount of $PM_{2.5}$) and fine test (mainly composed of $PM_{2.5}$ and a fraction of coagulated coarse) particle peaks.

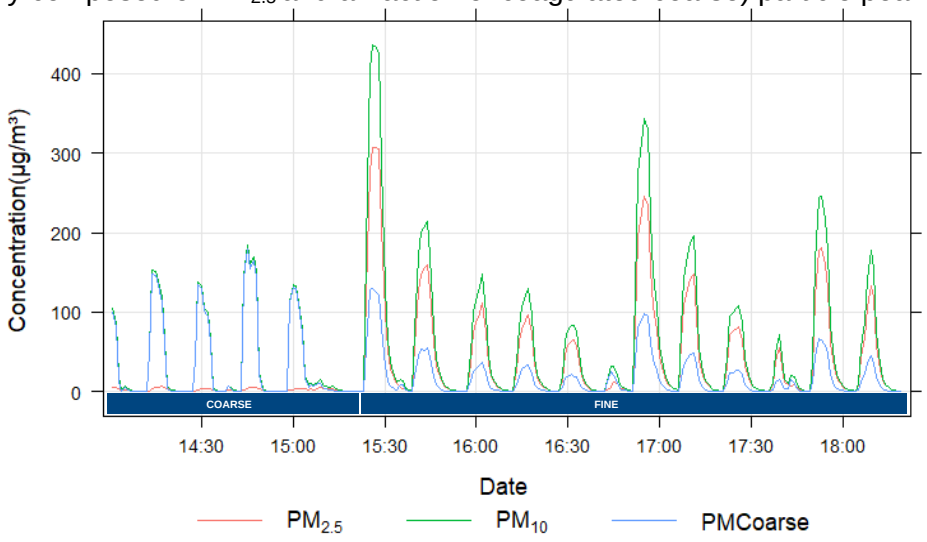


Figure 29 Coarse PM testing procedure with consecutive 5-minute generation periods of coarse (7.750 µm) and fine (1.180 µm) PM peaks.

Coarse tests have been performed for all 8 sensor systems on 2 days (14/7 and 2/9). It should be noted that a higher coarse composition was obtained during the second test day with 97% coarse particles, when compared to 75% during the first test day, which seems to result in higher change ratios as well (Table 8).

During both tests, sensors tend to visually pick up fine (1.180 µm) particle spikes, but appeared far less responsive to the coarse fraction spikes (Figure 30). Note that in both fine and coarse generation spikes, $PM_{2.5}$ is present.

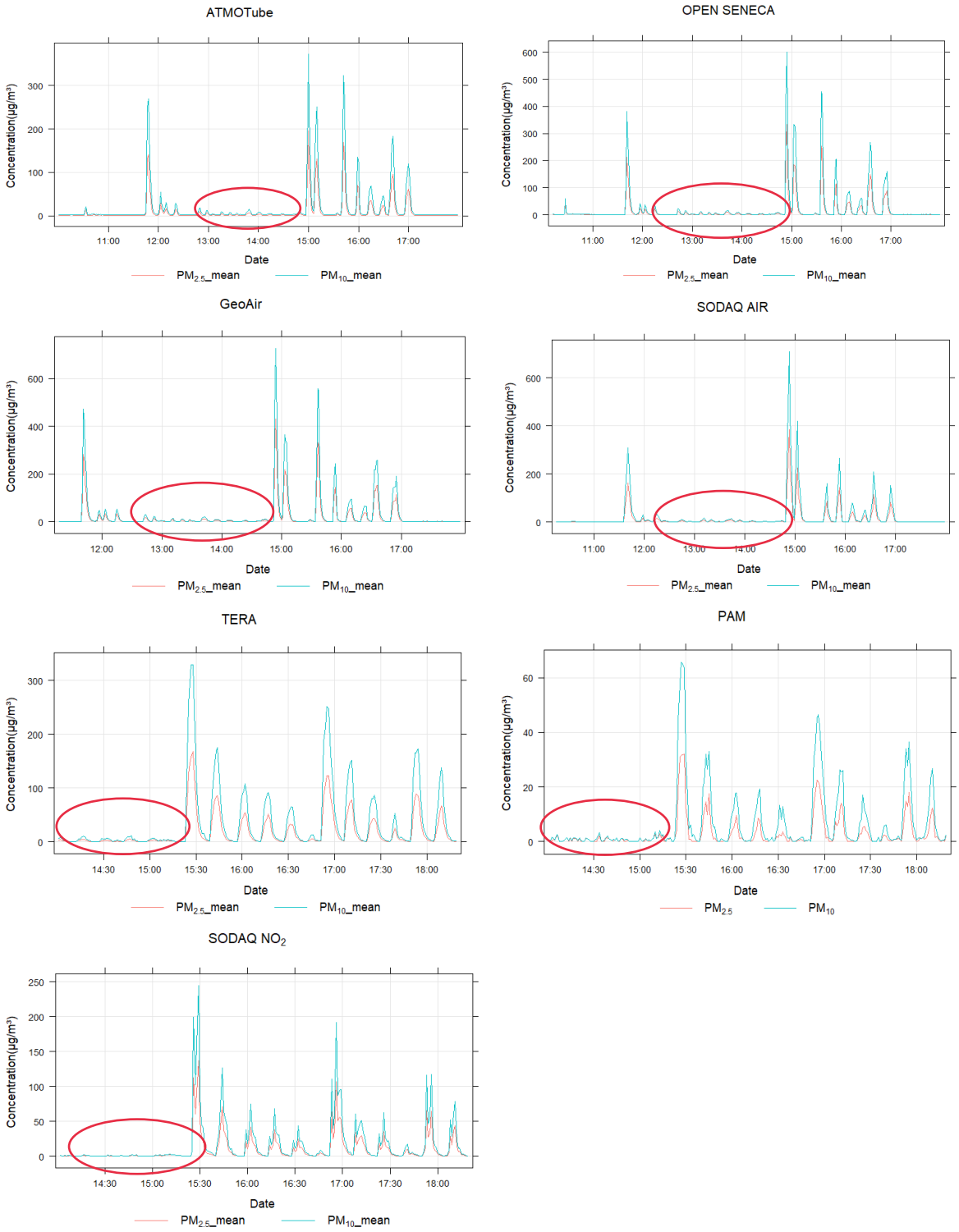


Figure 30 Difference between generated coarse (red circles) and fine PM spike periods, as captured by the different sensor systems on 14/7 (AtmoTube, Open Seneca, GeoAir and SODAQ AIR) and 2/9 (TERA, PAM, SODAQ NO_2)

Moreover, when plotting correlation plots between $PM_{2.5}$ and the PM_{coarse} (PM_{10}-$PM_{2.5}$) size fraction (of fine and coarse test conditions together), no relation is observed for the reference (Grimm) data where two associations (point clouds) are observed reflecting the different particle composition during the respective coarse and fine test conditions. For all sensor systems, significant associations are obtained between the $PM_{2.5}$ and PM_{coarse} fraction, indicating that the changing composition is not picked up by the sensor systems and the coarse PM fraction of the sensors might be derived algorithmically from the measured $PM_{2.5}$ concentrations.

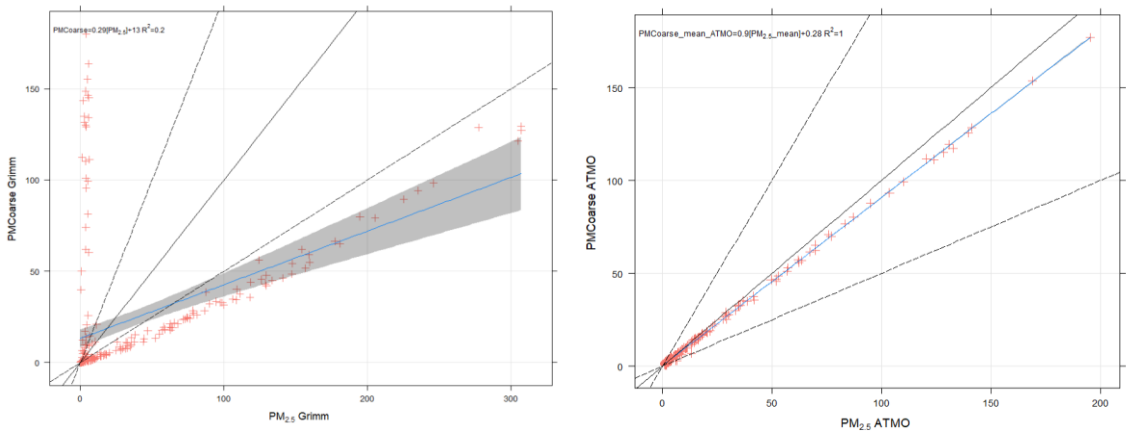


Figure 31 Observed association between measured $PM_{2.5}$ and PM_{coarse} for the reference (Grimm; left) and ATMOTube Pro sensor (right), when exposed to both fine and coarse particle peaks.

From Table 8, we observe very similar change ratios (%) for the considered sensor systems but varying change ratios with changing dust composition (67-76% for test 1, 93-100% for test 2). Similar observations between the sensors is not surprising as all sensors are ultimately based on only 3 particle sensors (Sensirion SPS30, Plantower PMS and TERA next-PM). A slightly better performance (94 vs 99-100%) of the TERA sensor seems to be suggested in terms of coarse PM detection. Deviating results are obtained for the ATMOTube (* in Table 8), which showed variable (4-7 minute) peak mismatch between the different sensors. We suspect that this might be due to clock synchronization issues when connected to the smartphone app.

*Table 8 Coarse test results obtained on 14/7 (AtmoTube, Open Seneca, GeoAir and SODAQ AIR) and 2/9 (TERA, PAM, SODAQ NO_2) with observed coarse composition (% coarse), PM_{10}, $PM_{2.5}$ and PM_{coarse} sensor/REF ratios, fine/coarse change ratio (%; between highlighted columns) and PM_{10} change ratio (%). *faulty results due to peak mismatch.*

	TEST	% COARSE	PM_{10} sensor/PM_{10} REF	$PM_{2.5}$ sensor/$PM_{2.5}$ REF	PM_{coarse} sensor/PM_{coarse} REF	%change in fine/coarse ratio (target=0)	%change in PM_{10} SENSOR/REF ratio (target=0)
ATMO	COARSE (7750nm)	75	0.02	0.05	0.02	**37***	**-19***
	FINE (1180nm)	14	0.02	0.01	0.07		
OPEN	COARSE (7750nm)	75	0.16	0.38	0.09	**-76**	**72**
	FINE (1180nm)	14	0.58	0.38	1.78		
SODAQ AIR	COARSE (7750nm)	75	0.17	0.37	0.10	**-73**	**72**

	FINE (1180nm)	14	0.62	0.39	2.00		
GEOAir	COARSE (7750nm)	75	0.21	0.50	0.11	**-76**	**67**
	FINE (1180nm)	14	0.63	0.44	1.76		
TERA	COARSE (7750nm)	97	0.05	0.73	0.03	**-94**	**93**
	FINE (1180nm)	27	0.72	0.49	1.33		
SODAQ NO2	COARSE (7750nm)	97	0.01	0.18	0.00	**-99**	**97**
	FINE (1180nm)	27	0.23	0.18	0.35		
PAM	COARSE (7750nm)	97	0.00	0.13	0.00	**-100**	**97**
	FINE (1180nm)	27	0.13	0.08	0.26		

3.1.2 NO_2 lab tests

3.1.2.1 Lack-of-fit

For all sensors containing a NO_2 sensor (3/9), lack-of-fit tests were conducted on 3 different days (12/8 14/8 and 15/8) at NO_2 concentrations ramping between 0 and 200 µg/m³. Due to the varying monitoring resolutions of the sensor systems (2 sec - 5min), all data was temporally aggregated (averaged) to a 1 minute resolution and merged with the reference (Thermo NOx analyzer) data. Setpoint averages were calculated based on steady-state conditions (final 1.5-hour considering a 15-minute buffer period before each setpoint change). From these setpoint averages, linearity plots were generated and regression coefficients (slope + intercept (y=a*x+b) and slope only (y=a*x)) calculated. In addition, we calculated the sensor stability as the standard deviation (µg/m³) at each setpoint (steady-state condition) and the sensor accuracy (%) at each setpoint. All results are shown per sensor type and subsequently presented in an overview table.

3.1.2.1.1 SODAQ NO_2

Within the SODAQ NO_2 sensor data, we noticed significant noise (high amplitude periods) and data gap (connectivity) issues. A negative linear association between the measured sensor readings and the increasing NO_2 concentration was observed, significantly blurred by the exhibited sensor noise, indicating that proper sensor calibration was not performed by SODAQ (Figure 34). Due to the low data availability, sensor setpoints and stabilities were derived individually for each sensor and provided in Figure 34. Sensor readings are inversely correlated to the actual NO_2 concentrations and a large deviation was observed between sensor 1 and sensor 2 and 3.

The stability was clearly impacted by the signal noise, resulting in sensor stabilities of 5-80 µg/m³, compared to a stability of 0.2-0.23 µg/m³ for the reference analyzer (Figure 34). Mean

setpoint accuracy (mean of different setpoint accuracies) varied between -113 and -254% for the different sensors.

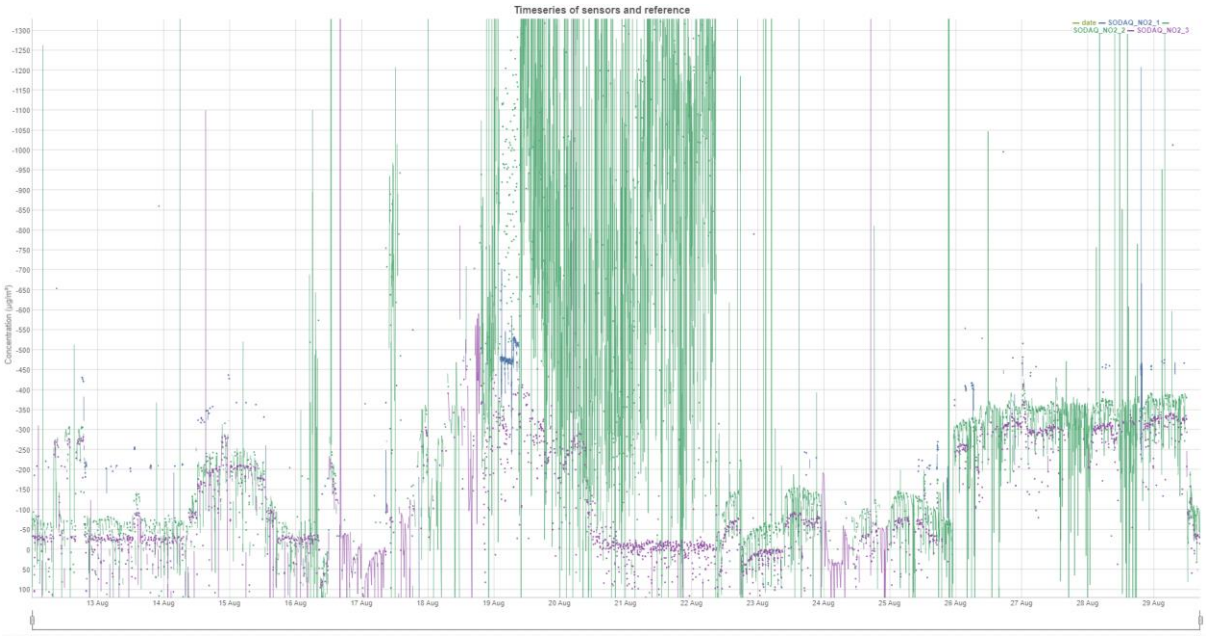


Figure 32 Varying sensor connectivity and noise observed for the SODAQ NO2 (1-3) sensors during the lack-of-fit test

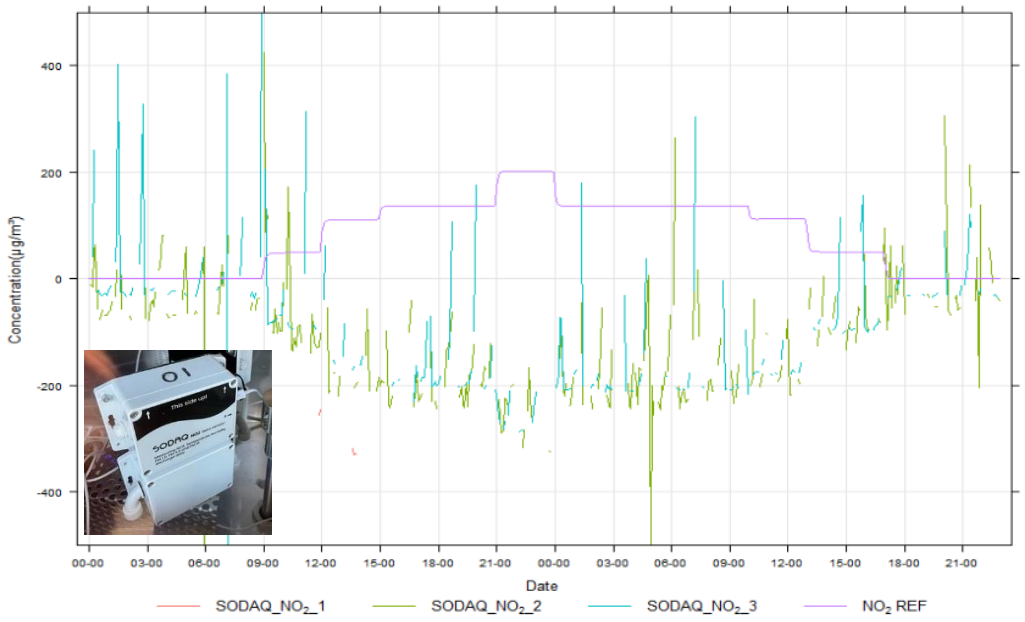


Figure 33 NO_2 concentrations generated during the lack-of-fit test as measured by the 3 SODAQ sensors (1-3) and the reference monitor (NO_2 REF)

SODAQ 1					SODAQ 2					SODAQ 3				
NO2_sensor	NO2_REF	SD_sensor	SD_REF	setpoints	NO2_sensor	NO2_REF	SD_sensor	SD_REF	setpoints	NO2_sensor	NO2_REF	SD_sensor	SD_REF	setpoints

				0	-18.40	-0.02	60.27	0.02	0	-2.00	-0.04	46.22	0.03	0
				40	-66.13	49.55	107.82	0.29	40	-45.57	49.62	110.03	0.27	40
-325.00	111.81	5.29	0.20	100	-156.36	111.75	63.88	0.22	100	-145.78	111.70	51.49	0.16	100
				140	-180.13	136.29	41.10	0.23	140	-150.62	136.29	104.92	0.20	140
-437.00	201.98			200	-238.90	202.05	64.00	0.36	200	-242.36	202.17	85.04	0.47	200
		5.29	**0.20**	**Stability (SD)**			**67.42**	**0.23**	**Stability (SD)**			**79.54**	**0.22**	**Stability (SD)**

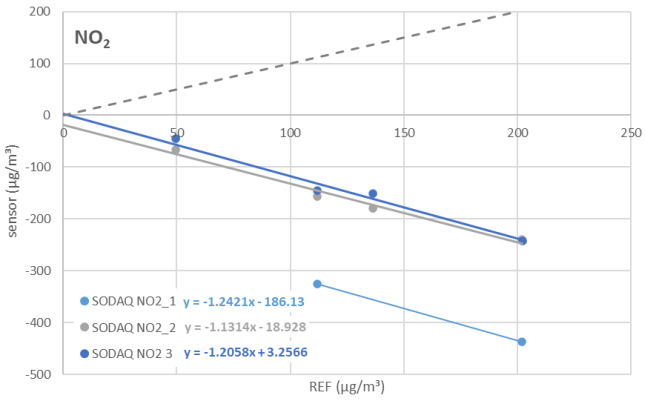


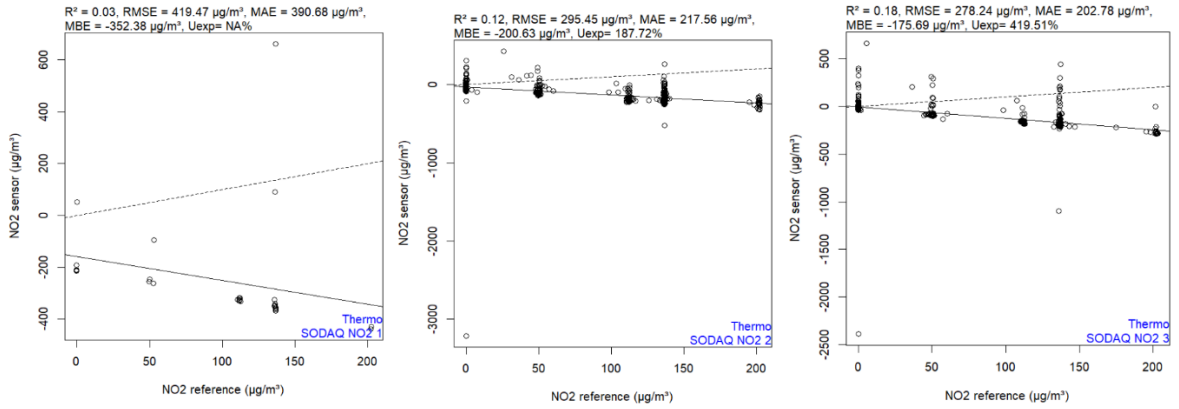
Figure 34 Average sensor (NO2_sensor) and reference (NO2_REF) concentrations (µg/m³), with associated stabilities (SD: standard deviation), derived for each concentration setpoint during the lack-of-fit test (upper) and associated linearity plots (lower).

Derived regression coefficients (slope + intercept (y=a*x+b) and slope only (y=a*x)) and linearity (R^2) for each sensor (SODAQ 1-3) are provided in Table 9.

Table 9 Regression coefficients (slope + intercept and slope only) for each sensor (1-3) with associated linearity (R^2)

	Intercept	Slope	R^2	slope only
SODAQ 1	-186.13	-1.24	1.00	-2.3379
SODAQ 2	-18.93	-1.13	0.99	-1.2586
SODAQ 3	3.26	-1.21	0.99	-1.1839

When plotting all 1-minute averaged data of both sensor and reference during the lack-of-fit test (Figure 35), we observe a low linearity (R^2=0.3-0.18) due to exhibited signal noise, and very high mean absolute errors (MAE: 203-391 µg/m³). Expanded uncertainty (Uexp) varies between 188% and 420%, not qualifying for the indicative (<25%) data quality objective.



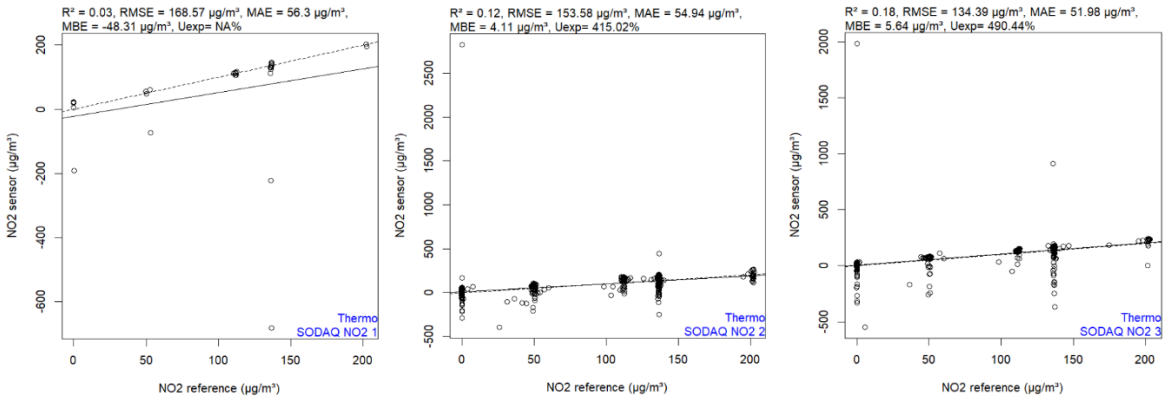


Figure 35 Scatterplots of 1-minute averaged reference (Thermo) and the raw (upper) and calibrated (lower) sensor data for each of the considered sensors (SODAQ 1-3) with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp).

After applying a lab calibration based on the derived slopes and intercepts provided in Table 9 ($sensor_{cal}$=($sensor_{raw}$-b)/a)), the sensor accuracy (MAE) improved to 52-56 µg/m³, while the expanded uncertainty remains at 450-491% (Figure 35). Note that both training (to derive slope and intercept) and test data are identical.

The uncertainty between the sensors (BSU) was 125 µg/m³ which can be considered as very poor.

The SODAQ NO_2 out-of-the-box performance can be considered as inadequate. Potential calibration is hindered by the high signal noise, while sensor boxes showed connectivity issues and high between-sensor-uncertainty (BSU).

3.1.2.1.2 2BTech PAM

The raw 2BTech PAM data (for which only one sensor system was purchased) showed a positive response to the increasing NO_2 concentration steps (Figure 36), but exhibited some signal noise and extremes (peak values with unknown reason) resulting in a low sensor stability of 27 µg/m³, when compared to 0.19 µg/m³ for the reference analyzer (Figure 37). Sensor readings slightly underestimate the actual NO_2 concentrations with a mean setpoint accuracy of 71.5%.

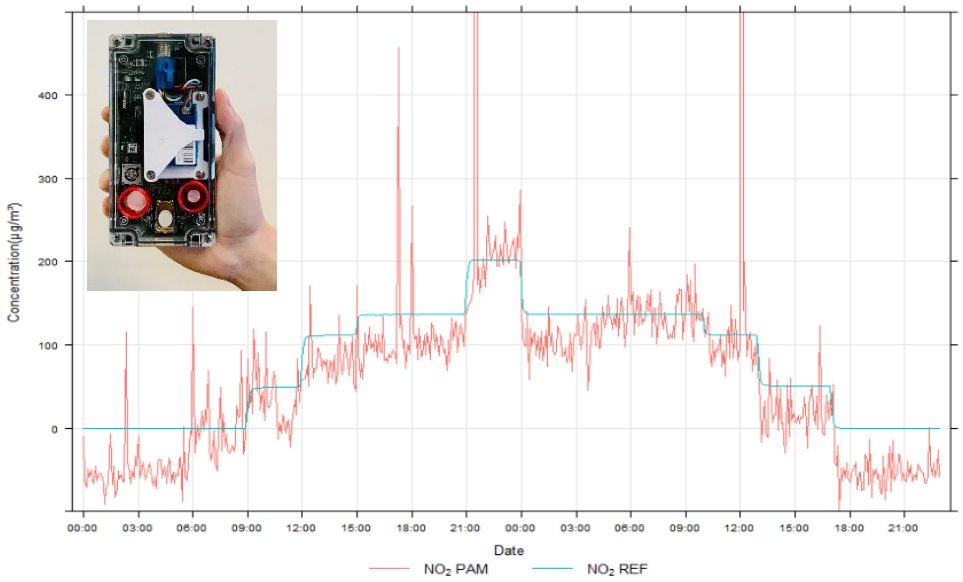


Figure 36 NO_2 concentrations generated during the lack-of-fit test as measured by the PAM sensor and the reference monitor (NO_2 REF)

2BTech PAM					
NO2_sensor	**NO2_REF**	**SD_sensor**	**SD _REF**	**setpoints**	**n**
-9.72	-0.02	35.23	0.02	0	31
20.98	49.61	31.84	0.19	40	31
81.83	111.69	26.31	0.24	100	31
103.57	136.38	19.44	0.19	140	31
212.89	202.07	21.61	0.28	200	31
		26.89	**0.19**	**Stability (SD)**	

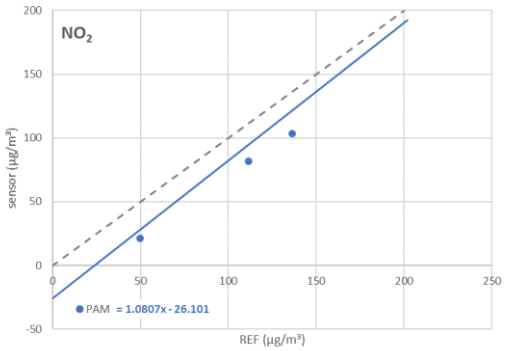


Figure 37 Average sensor (NO2_sensor) and reference (NO2_REF) concentrations (µg/m³), with associated stabilities (SD: standard deviation), derived for each concentration setpoint during the lack-of-fit test (upper) and associated linearity plot (lower).

Derived regression coefficients (slope + intercept (y=a*x+b) and slope only (y=a*x)) and linearity (R^2) are provided in Table 9.

Table 10 Regression coefficients (slope + intercept and slope only) with associated linearity (R^2)

	Intercept	**Slope**	**R^2**	**slope only**
PAM	-26.10	1.08	0.96	0.9053

When plotting all 1-minute averaged data of both sensor and reference during the lack-of-fit test (Figure 38), we observe a low linearity (R^2=0.13) due to some extremes, and mean absolute errors of 49.5 µg/m³ (raw) and 36.6 µg/m³ (calibrated). Expanded uncertainty (Uexp) is 110% and further improves up to 80%, still not qualifying for the indicative (<25%) data quality objective.

The BSU could not be calculated as we only had one PAM available.

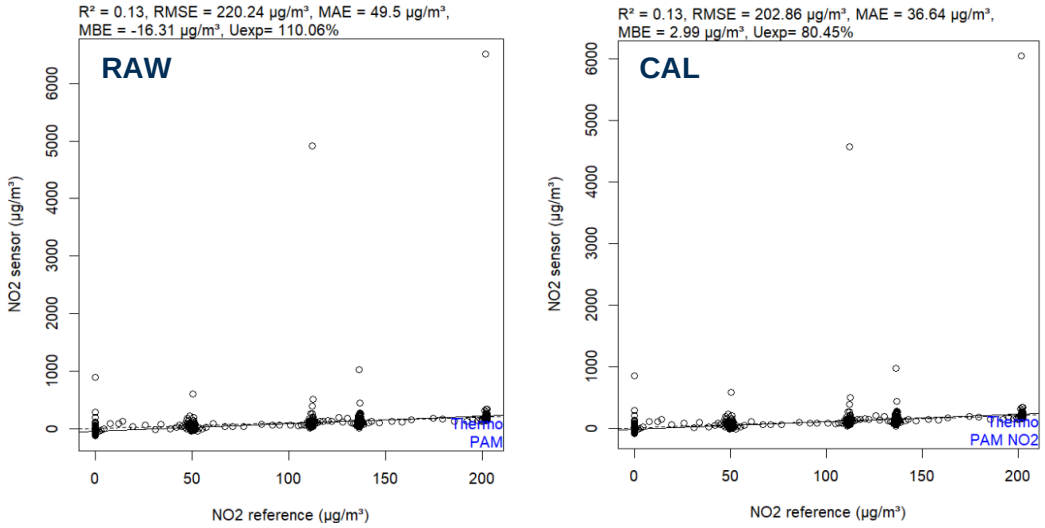


Figure 38 Scatterplots of 1-minute averaged reference (Thermo) and the raw (left) and calibrated (right) sensor data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp).

The PAM out-of-the-box performance is much better than the SODAQ performance, but still suffers from signal noise resulting in low stability and accuracy, insufficient for reaching the data quality objectives for indicative monitoring.

3.1.2.1.3 DST Observair

The raw NO_2 data from the DST Observair (for which only one sensor system was purchased) varied between -0.03 and 0.03 µg/m³ (invisible when plotting against 0-200 µg/m³ reference concentrations) and showed a negative linear response to the increasing NO_2 concentration steps (Figure 39), indicating that proper sensor calibration was not yet performed by DST. DST warned for the out-of-the-box data quality in advance and typically relies on co-located reference measurements to train a sensor calibration model using machine learning techniques. Compared to the SODAQ NO_2 and PAM, the Observair exhibits rather low signal noise, resulting in more pronounced distinction between the exposed concentration steps (Figure 39).

The low NO_2 values (-0.03 and 0.03 µg/m³) resulted in a high sensor stability of <0.01 µg/m³, when compared to 0.20 µg/m³ for the reference analyzer and need for proper calibration (Figure 40).

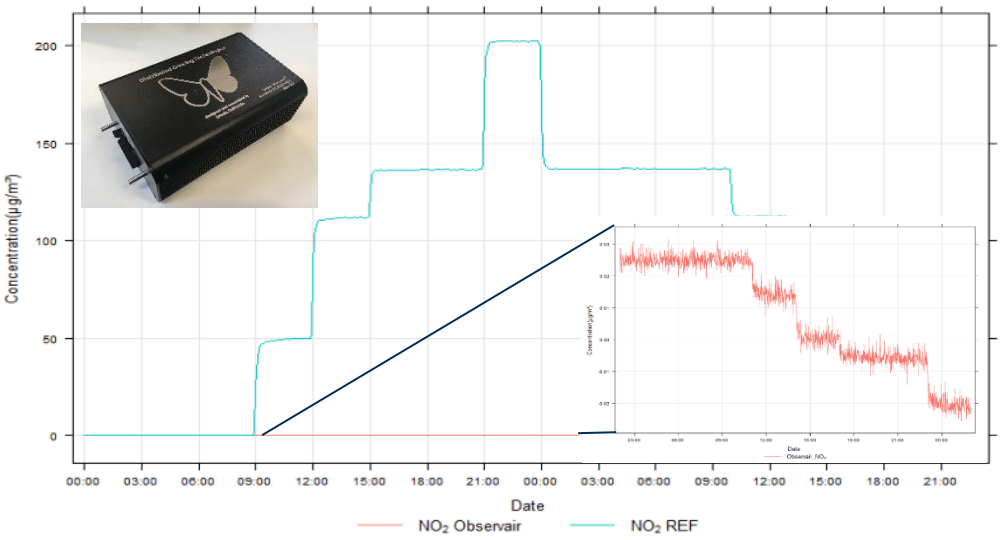


Figure 39 NO_2 concentrations generated during the lack-of-fit test as measured by the PAM sensor and the reference monitor (NO_2 REF)

DST Observair					
NO2_sensor	**NO2_REF**	**SD_sensor**	**SD_REF**	**setpoints**	**n**
0.02	-0.02	<0.01	0.03	0	91
0.01	49.63	<0.01	0.23	40	91
0.00	111.71	<0.01	0.24	100	91
-0.01	136.36	<0.01	0.21	140	91
-0.02	202.07	<0.01	0.31	200	91
		<0.01	**0.20**	**Stability (SD)**	

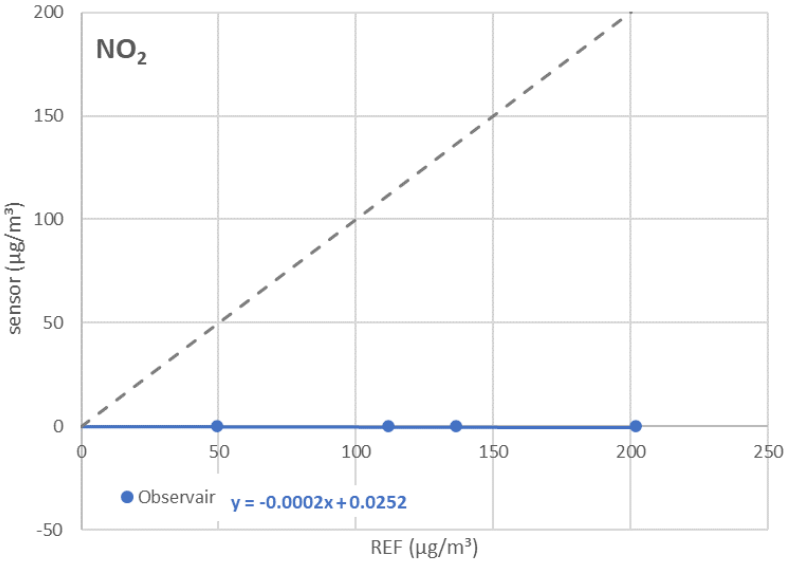


Figure 40 Average sensor (NO2_sensor) and reference (NO2_REF) concentrations (µg/m³), with associated stabilities (SD: standard deviation), derived for each concentration setpoint during the lack-of-fit test (upper) and associated linearity plot (lower).

Derived regression coefficients (slope + intercept (y=a*x+b) and slope only (y=a*x)) and linearity (R^2) are provided in Table 11.

Table 11 Regression coefficients (slope + intercept and slope only) with associated linearity (R^2)

	Intercept	Slope	R^2	slope only
PAM	0.03	<0.01	1.00	-0.00002

When plotting all 1-minute averaged data of both sensor and reference during the lack-of-fit test (Figure 41), we observe a very good linearity (R^2=0.98) and mean absolute error of 79 µg/m³ (raw). After linear calibration based on the slope and intercept provided in Table 11, the accuracy further improves to 13.45 µg/m³. Expanded uncertainty (Uexp) of the raw data is 112% and improves to 65% after linear calibration, still not qualifying for the indicative (<25%) data quality objective.

The BSU could not be calculated as we only had one Observair available.

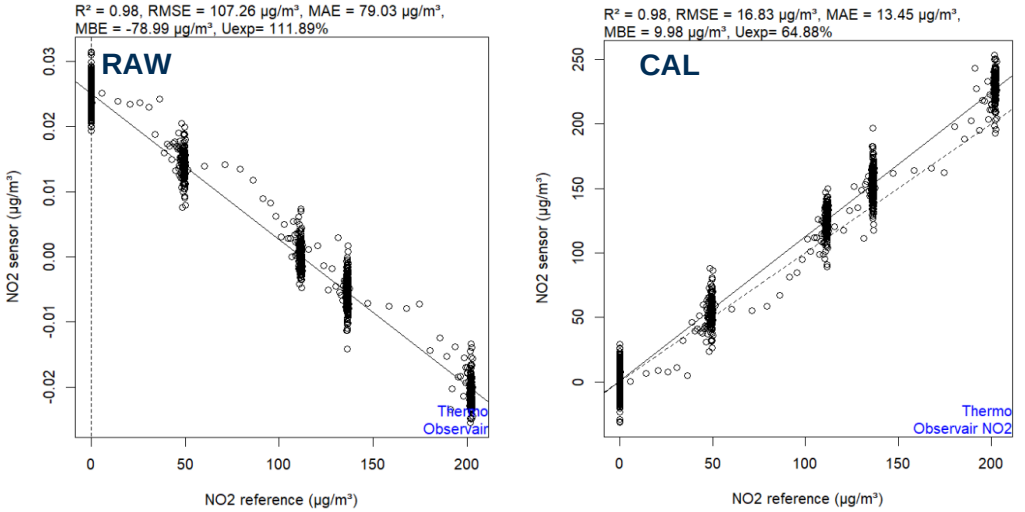


Figure 41 Scatterplots of 1-minute averaged reference (Thermo) and the raw (left) and calibrated (right) sensor data with associated performance metrics (R^2, RMSE, MAE, MBE and Uexp).

The DST Observair out-of-the-box linearity is very good but suffers from a poor accuracy as no factory calibration seemed to have been performed. After calibration, the accuracy of the Observair outperforms the observed accuracies of the SODAQ NO2 and PAM. Nevertheless, it still doesn't reach the 25% data quality objective set for indicative monitoring instruments.

3.1.2.2 Sensitivity to relative humidity

The sensor response to changing relative humidity (0-50-75-90%) was tested at 0 and 200 µg/m³ NO_2 and a stable gas chamber temperature of 20°C. For all sensors, changing RH steps seemed to result in an initial peak response, with subsequent 1- to 2-hour stabilization period. The observed responses are very small for the Observair (-0.05-0.05 µg/m³), and opposite raw responses were observed when comparing the PAM to SODAQ (Figure 42).

When considering the lab-calibrated results, similar responses are observed for the 3 sensor systems (Figure 42), only differing in terms of noisiness/stability (Observair<PAM<SODAQ NO_2). Similar responses can be explained by the fact that all sensor systems rely on the same sensor (Alphasense NO2-B43F). Similar sensor responses are observed at 0 and 200 µg/m³.

Figure 42 Raw (upper) and calibrated (middle + lower) sensor responses to varying relative humidity steps (RH; %) under 0 (upper and middle) and 200 (lower) µg/m³ NO_2 conditions.

Setpoint averages (µg/m³) and stability (µg/m³) were calculated based on the lab-calibrated sensor data (for comparison) under steady-state conditions (final 1.5-hour considering a 15-minute buffer period before each setpoint change) and are provided in Table 12.

Table 12 Setpoint averages (µg/m³) and stabilities (µg/m³) based on the 1-minute averaged calibrated sensor data

CAL		Setpoint average (µg/m³)					Setpoint stability (µg/m³)					
	n	**SODAQ_1**	**SODAQ_2**	**SODAQ_3**	**Observair**	**PAM**	**SODAQ_1**	**SODAQ_2**	**SODAQ_3**	**Observair**	**PAM**	**Setpoint**
NO_2 = 0	61	-14.19	-78.92	-29.16	-11.54	-21.39	10.82	58.72	11.26	10.66	43.01	0
	61	NA	NA	-7.62	12.19	-27.17	NA	NA	25.24	7.62	28.85	50
	61	NA	-124.56	0.07	16.74	-10.77	NA	36.28	48.35	6.06	47.74	75
	61	NA	53.39	105.21	10.07	-10.77	NA	36.28	3.14	8.74	44.11	90
	n	**SODAQ_1**	**SODAQ_2**	**SODAQ_3**	**Observair**	**PAM**	**SODAQ_1**	**SODAQ_2**	**SODAQ_3**	**Observair**	**PAM**	**Setpoint**
NO_2 = 200	61	NA	338.54	210.24	209.90	261.85	NA	628.80	21.76	12.33	77.41	0
	61	NA	870.91	166.35	238.37	294.42	NA	441.08	89.60	11.79	61.63	50
	61	NA	795.02	246.19	262.31	298.69	NA	705.15	27.21	10.75	44.28	75
	61	NA	1257.80	236.36	272.98	298.69	NA	705.15	99.90	12.68	55.48	90

From the setpoint-derived regression plots and changing sensor/REF ratio (Figure 43), similar sensor responses are observed for all 3 sensor systems, once again reflecting the underlying hardware (NO_2 sensor) similarities. Nevertheless, most variability is observed in the SODAQ response, probably due to the experienced sensor noise.

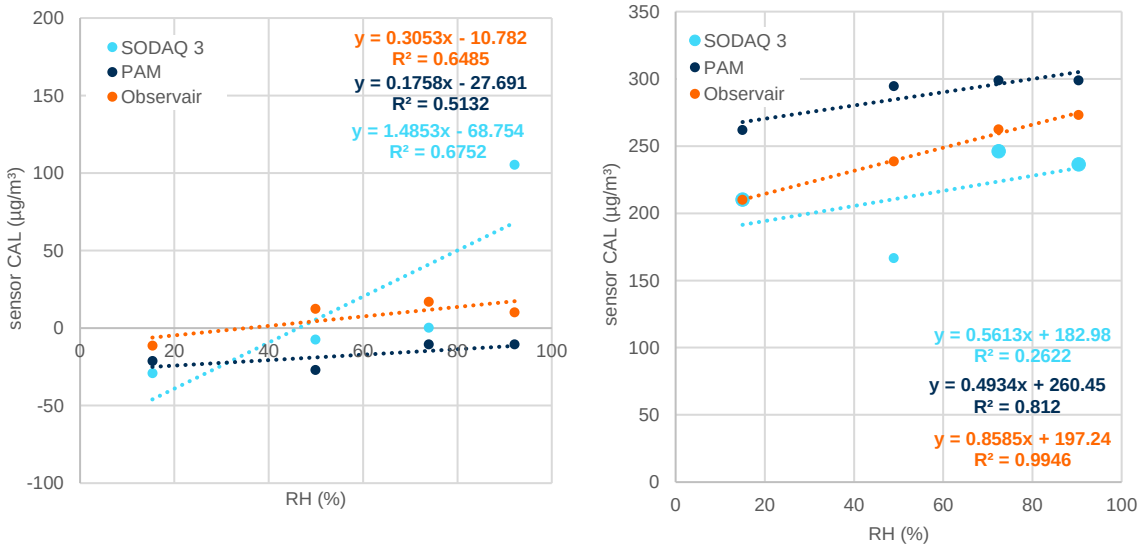


Figure 43 Regression plots showing the setpoint-averaged sensor responses (µg/m³) to changing relative humidity (RH; %) at 0 (left) and 200 (right) µg/m³ NO_2 with associated regression functions and determination coefficients (R^2).

3.1.2.3 Temperature sensitivity

The sensor response to changing temperature (-5, 10, 20 and 30°C) was tested at 0 and 200 µg/m³ NO_2 and a stable gas chamber relative humidity of 50%. For all sensors, changing temperature steps, just like relative humidity, seemed to result in an initial peak response (transient effect), with subsequent 1- to 2-hour stabilization period. The observed responses are very small for the Observair (-0.05-0.05 µg/m³), and opposite raw responses were observed when comparing the PAM to SODAQ (Figure 44).

When considering the lab-calibrated results, sensor noisiness/stability makes it hard to interpret the sensor responses (Figure 44). While the SODAQ and PAM seem to vary around a constant NO_2 concentration when exposed to varying temperatures (<30°C), the Observair

seems to show consistent higher NO_2 concentrations under decreasing temperatures. Similar sensor responses are observed at 0 and 200 µg/m³.

Figure 44 Raw (upper) and calibrated (middle + lower) sensor responses to varying temperatures (Temp; °C) under 0 (upper and middle) and 200 (lower) µg/m³ NO_2 conditions.

Setpoint averages (µg/m³) and stability (µg/m³) were calculated based on the lab-calibrated sensor data (for comparison) under steady-state conditions (final 1.5-hour considering a 15-minute buffer period before each setpoint change) and are provided in Table 13.

Table 13 Setpoint averages (µg/m³) and stabilities (µg/m³) based on the 1-minute averaged calibrated sensor data

CAL		Setpoint average (µg/m³)					Setpoint stability (µg/m³)					
	n	SODAQ_1	SODAQ_2	SODAQ_3	Observair	PAM	SODAQ_1	SODAQ_2	SODAQ_3	Observair	PAM	Setpoint
NO_2 = 0	61	NA	-34.94	21.78	99.11	52.23	NA	44.42	NA	9.27	56.57	-5
	61	NA	NA	-43.24	61.33	-3.57	NA	NA	69.15	3.80	43.51	10
	61	NA	58.84	32.42	26.44	-12.07	NA	18.40	1.77	4.29	31.3	20
	61	NA	44.11	4.36	-56.63	463.18	NA	17.63	3.25	4.77	2893.53	30
	n	SODAQ_1	SODAQ_2	SODAQ_3	Observair	PAM	SODAQ_1	SODAQ_2	SODAQ_3	Observair	PAM	Setpoint
NO_2 = 200	61	NA	285.32	272.92	314.10	207.85	NA	35.51	6.39	5.62	68.48	-5
	61	NA	223.63	NA	274.46	207.20	NA	232.83	NA	5.27	45.54	10
	61	220.49	277.50	257.98	250.04	206.28	NA	67.23	7.80	7.18	35.78	20
	61	NA	255.89	NA	185.72	293.94	NA	83.05	NA	6.24	83.45	30

From the setpoint-derived linearity plots and regression coefficients (Figure 45), varying sensor responses are observed for considered sensor systems, sometimes blurred by the noisiness/stability of the sensor system (e.g. SODAQ and PAM). The Observair seems to show consistent NO_2 reductions with increasing temperatures.

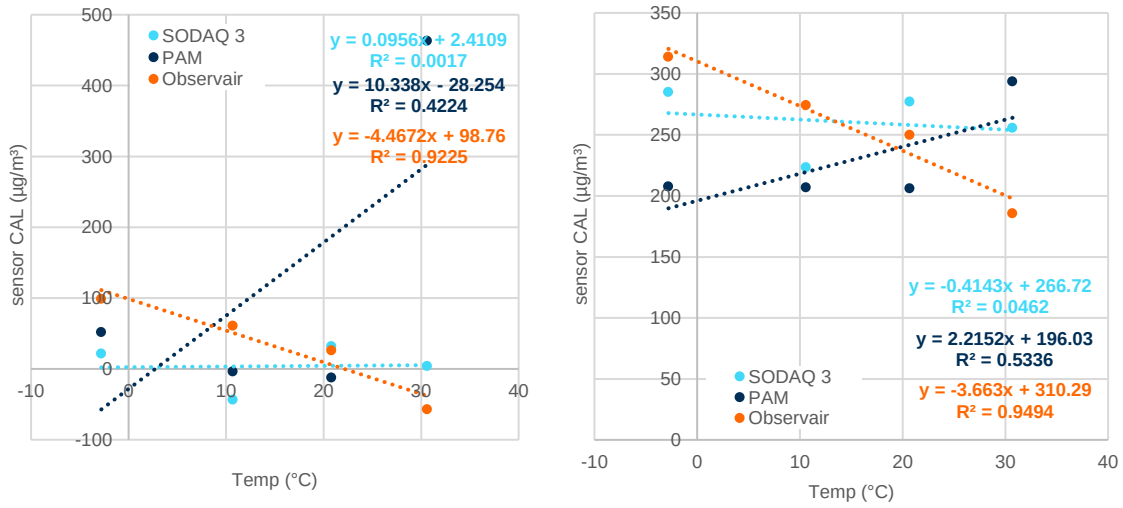


Figure 45 Linearity plots showing the setpoint-averaged sensor responses (µg/m³) to changing temperatures (Temp; °C) at 0 (left) and 200 (right) µg/m³ NO_2 with associated regression functions and determination coefficients (R^2).

3.1.2.4 Response time

To simulate rapidly changing NO_2 concentrations, sensors were placed in glass tubes that allowed for rapid concentration changes between 0-200 µg/m³. The smaller volume of the glass tubes (compared to the NO_2 exposure chamber), only allowed evaluation of the Observair and PAM sensor as the SODAQ NO_2 boxes didn't fit in the glass tubes.

30-minute intervals (0 and 200 µg/m³) were considered and lab-calibrated sensor data was compared to the 1-minute data from the Thermo NOx analyzer. Averages and 90-percentiles (90% of max concentration) concentrations were determined for each 200 µg/m³ plateau, and the associated response time, i.e. time needed to reach 90% concentration was calculated for

each sensor system (and reference analyzer). Again, the noisy signal makes the evaluation of this test not strait forward.

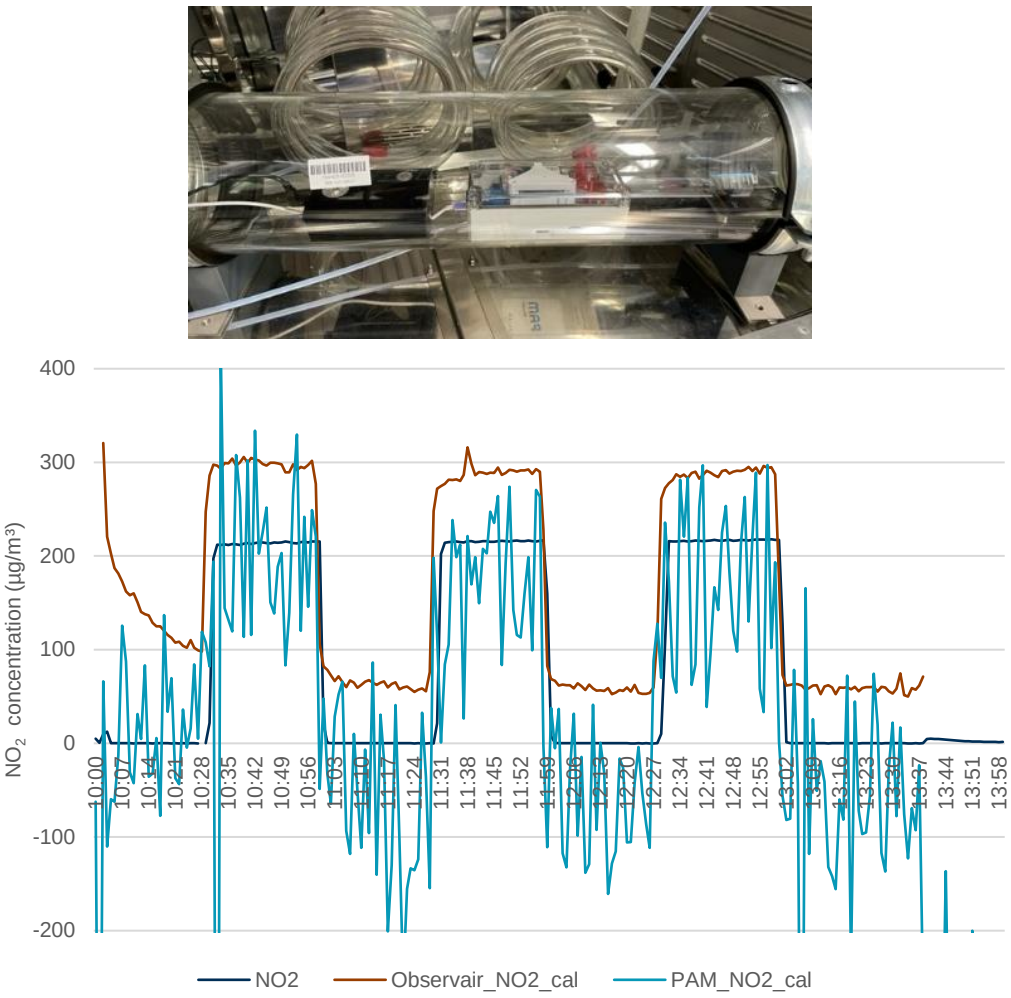


Figure 46 Sensor setup in the glass tubes during response test (upper) and measured NO_2 concentrations (lower) of the lab-calibrated Observair (Observair_NO2_cal), lab-calibrated PAM sensor (PAM_NO2_cal) and Thermo NOx analyzer (NO_2).

The resulting response times derived from the 3 consecutive 0-200 plateaus are provided in Table 14 and varied between 1-2 minutes for the sensor systems and 3 minutes for the reference analyzer.

		AVG	90%	t_90
Plateau 1	Thermo	215	193	3 min
	Observair_cal	295	265	2 min
	PAM_cal	197	177	3 min
Plateau 2	Thermo	216	194	3 min
	Observair_cal	286	258	1-2 min
	PAM_cal	177	159	1 min
Plateau 3	Thermo	217	195	3 min
	Observair_cal	291	262	1 min
	PAM_cal	175	158	2 min

Table 14 NO_2 average (AVG) and 90-percentile (90%) concentration and associated response time (t_90), calculated for the Observair and PAM sensor systems and Thermo NOx analyzer.

3.2 Field benchmarking campaign

The field benchmarking campaign consisted of a **mobile field test** with all sensors deployed on top of a cargo bike to evaluate the GPS signal of the sensors in a heterogeneous urban landscape.

Next to the mobile test, the **sensor performance was evaluated in representative urban environmental conditions and pollutant concentrations** by co-locating all sensors on top of an urban background air quality monitoring station (R801) in Antwerp, Belgium, for a 3-month period.

3.2.1 Mobile field test

On September 7th, 2022, a trajectory of 10.4 km through the city center of Antwerp, Belgium, was cycled with a cargo bike (Figure 47), covering various urban topologies, e.g. narrow street canyons, open landscapes, road tunnels, bridges and natural areas (Figure 48).

All sensors were mounted on top (in the free airflow) of a cargo bike (Figure 47). Package sleeves were used during mounting to damp the sensor attachment platform from potential vibrations of the cargo-bike whilst cycling. Besides the sensors, two mid-range instruments namely a Grimm 11D (PM; without heated inlet) and MA200 (BC) were placed inside the cargo bike with air inlets at the height of the sensors. Finally, the cargo bike was equipped with 3 different GPS instruments (Garmin 810 Edge, TomTom Runner 2, Komoot smartphone application) for consideration as reference GPS track.

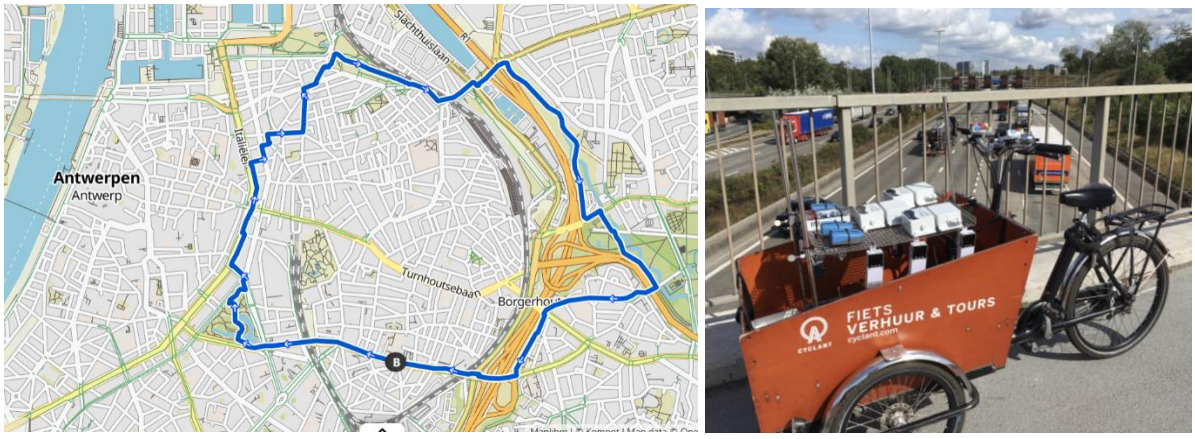
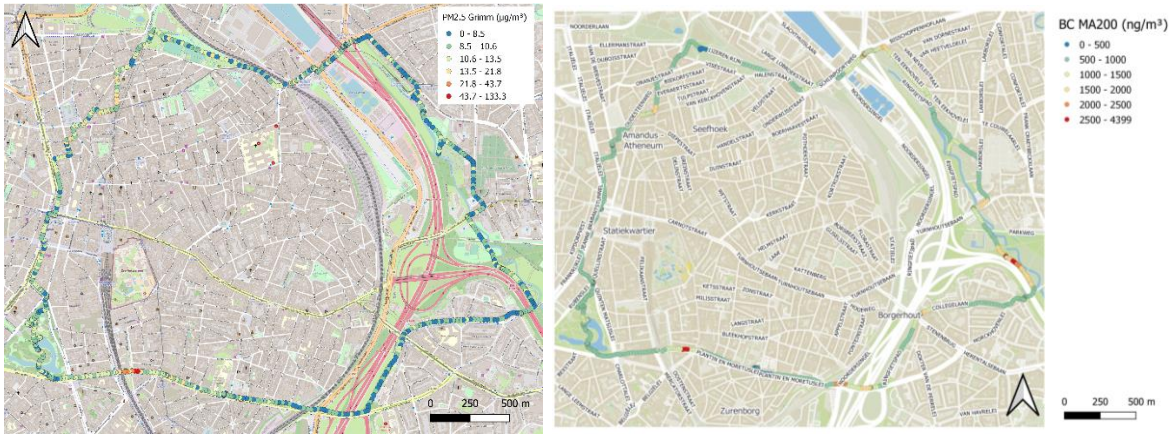


Figure 47 Considered cycling route through Antwerp (left) and instrument setup on the mobile platform (cargo bike; right).

Figure 48 Google Streetview images taken along the considered cycling route

3.2.1.1 Concentration variability

The exhibited mobile $PM_{2.5}$ concentration variability (measured by the Grimm) ranged between 4.8 and 133.3 µg/m³, while the mobile BC (measured by the MA200) varied between 0.4 and 4.4 µg/m³. When both pollutants were plotted on a map, spatial variability could be observed with both common and differing hotspot locations along the cycling trajectory. While highest $PM_{2.5}$ concentrations were observed at a housing façade construction site, highest BC concentrations were obtained when cycling downwind of a busy highway (E313/E34).



3.2.1.2 GPS accuracy

3.2.1.2.1 Reference GPS track

In order to evaluate the GPS accuracy, we started by selecting 1 of the 3 GPS instruments as reference track. We did so by visually comparing the horizontal accuracies of the 3 GPS tracks. The TomTom track showed a higher temporal monitoring resolution (1 sec) and better alignment with our traveled cycling route (horizontal accuracy), when compared to the tracks

of the Garmin Edge or Komoot app. We, therefore, selected the TomTom Runner 2 as reference GPS track for the sensors.

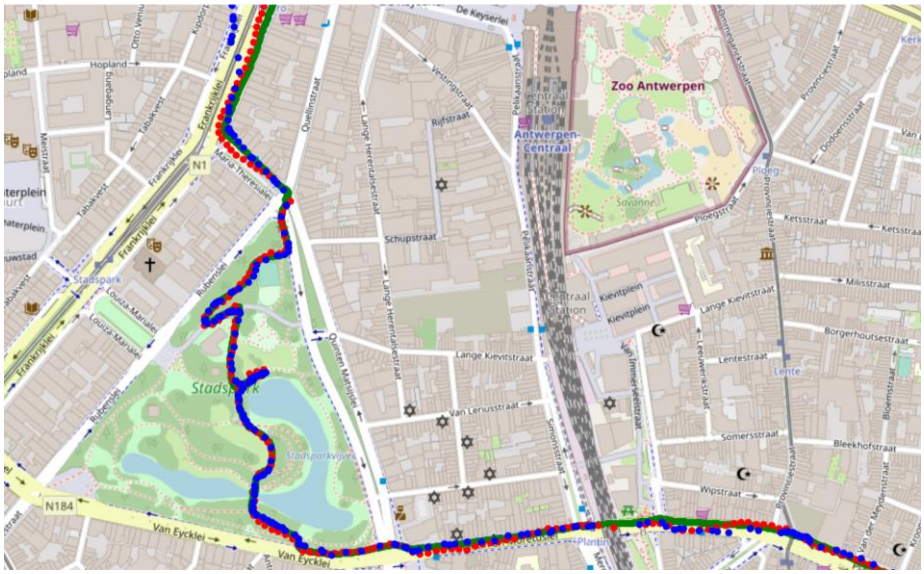


Figure 49 Detail of different GPS tracks with TomTom Runner 2 (green), Garmin Edge 810 (blue) and Komoot app (red) along the traveled cycling route.

3.2.1.2.2 Sensor GPS evaluation

When plotting all sensor tracks (Latitude/Longitude) on a map, it is clear that GPS accuracy performs better in open areas, when compared to narrow and/or high street canyons. More GPS noise is also observed in and around tunnels where the GPS signal might briefly be lost.

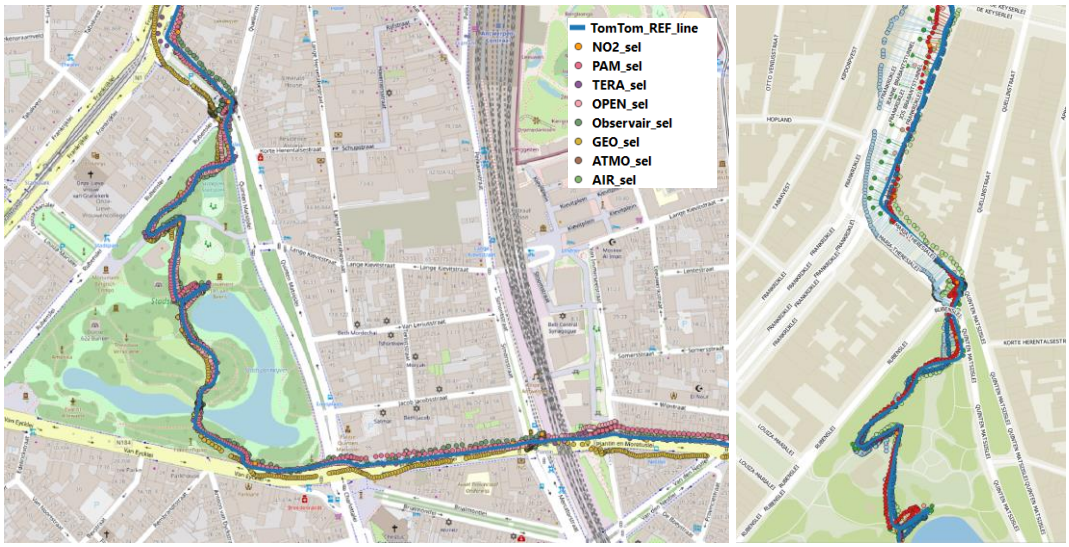


Figure 50 Left: GPS tracks of the considered sensor systems (dots) and reference GPS track (blue line). Right: Application of horizontal distance calculation to reference GPS track.

Next, the “distance to nearest hub (line to hub)” tool was applied in QGIS to calculate the horizontal distance (m) from each sensor datapoint to the reference GPS track. We cleaned 9% of the SODAQ AIR data exhibiting 0’s in the Latitude and Longitude coordinates. All

sensor-specific horizontal distances (m) were averaged over the entire cycling route and provided with associated datapoint counts (n) in Table 15.

Table 15 Average horizontal accuracy of the considered sensor systems.

	Horizontal distance to REF line (m)	Count (n)
TomTom (REF)	*0.00*	*6077*
TERA PMscan	2.28	1008
OPEN SENECA	3.18	489
SODAQ NO2	3.73	78
2BTech PAM	4.20	2476
SODAQ AIR	4.28	449
ATMOTube Pro	5.43	4
DST Observair	7.35	2446
GeoAir	8.15	4616

From Table 15, it becomes clear that the obtained horizontal GPS accuracy along our cycling route was generally good achieving a <10 m spatial resolution for all sensor systems which is good enough for street segment/map matching, or buffer averaging applications, i.e. tools that are commonly applied to map mobile measurements to the street network. Highest horizontal accuracy (2.28 m) was obtained for the TERA PMscan, while the lowest horizontal accuracy (8.15 m) was observed for the GeoAir.

3.2.2 Field performance

All sensor systems were deployed in an actively ventilated exposure shelter on top an urban background monitoring station (R801) in the city center of Antwerp. The co-location campaign lasted for 3 months, from 7/9/2022 until 5/12/2022 (Figure 51). The sensors were evenly distributed across the different shelter levels (3) and powered via USB hubs and additional power plugs (for sensors that required higher amperages; GeoAir, BCmeter). If 3 devices of a sensor system were available, we distributed 1 sensor on each shelter platform (Figure 51).

Figure 51 Location of the exposure shelter on top of R801 (left), detail of the exposure shelter (middle) and operating sensor systems inside the shelter (right).

Different data transmission protocols required different data collection procedures and a dedicated sensor manual was (addendum) created describing the operation and data-offload procedures of the different sensor systems. Some sensor systems automatically uploaded data via GPRS/4G (SODAQ), some sensor systems stored data on an internal SD-card

(GeoAir), while some sensor systems relied on an app to operate (TERA PMscan) or upload data (ATMOTube) or a combination of these data transmission protocols (PAM, OpenSeneca, Airbeam, Observair). Other field experiences included:

- BCmeter (not mobile) relied on a wifi connection for proper (clock synced) operation and data upload.
- Airbeam arrived later than the other systems and was only deployed from 5/10 onwards. It needed dedicated firmware in order to access the Belgian (Proximus) network, only properly operating by 9/11.
- The 2BTech PAM showed a lower temporal data resolution in the dashboard (data storage purposes), when compared to the SD card. We made use of the SD card data.
- The TERA PMscan sensor needed app connectivity in order to collect data. We used smartphones (Samsung Galaxy and OnePlus) and connected the app for continuous data collection. In addition we disabled the automatic app shutdown configuration in the smartphones. However, in practice we noticed the apps automatically shutting down after ~1-3 days.
- The Observair showed a warning requiring a reset (by using magnet which was not supplied). We drained the device instead and conducted a device reset, clock synchronisation before re-deployment in the field (16/11).
- As the considered BC sensors (BCmeter and Observair) required manual filter changes (every ~1-2 days), we performed a dedicated co-location campaign for 1.5 week, from 16/11/2022 until 25/11/2022.

All sensor data was offloaded (remotely via web dashboards and on-site via SD card readout) weekly to avoid data loss and a logbook was created to keep track of that status and encountered issues.

Table 16 Logbook of the co-location field campaign

DATE	TIME	Remarks	ATMOTube	TERA PMScan	Open Seneca	SODAQ AIR	SODAQ NO2	GeoAir	PAM	DST OBSERVAIR	AIRBEAM	BCmeter
07/09/2022	14:30	Sensors installed and running										
09/09/2022		Online check	?	TERA 1 en 2 ok TERA 2 no data?	?	All ok	11, 12 ok 10 no connectivity?	?	OK	?		
12/09/2022		Online check		TERA 1 en 2 ok TERA 2 no data?		All ok	All ok		OK			
15/09/2022		Onsite check	All ok (8B; 9:55: deo)		All ok			All ok		All ok		
26/09/2022		Online check	?	TERA3 no data (>16/9) TERA2 no data (>15/9) TERA1 no data (>20/9)	?	AIR1 & 2 ok AIR3 no data (>15/9)	10, 11,12 ok	?		?		
28/09/2022		Onsite check	All ok	TERA 2 & 3 logged out TERA 1 solved	All ok	AIR3 replaced data cable:ok	ok	All ok	OK	Not sampling; LED blinking yellow/blue : ATN warning -> replace filter		
05/10/2022		Onsite check	All ok	TERA 2 no data >15/9 & 3 no data >16/9 TERA 1 no data >2/10	All ok	All ok	All ok	All ok	OK	replaced filter but doesn't startup	AIRBEAM activated Only 1 connected	
12/10/2022		Onsite check	All Ok	All TERA logged out after =3 days	All ok	All ok	All ok	All ok	OK	replaced filter but doesn't startup	reconnected sensors	
				reconnected app		no data between 8/10-10/10	no data between 8/10-10/10					
26/10/2022		Onsite check	All ok	All TERA logged out after =3 days	All ok	AIR2 and 3 ok! AIR 1 no data, battery? --> Reconnected	All ok	All ok	OK	jammed, reset with magnet failed.	No data	
				reconnected app			NO2_2 shows a lot of noise around 19/10 (see plot)			collected sensor to drain/check at office	collected sensors to conduct firmware update	
28/10/2022		Onsite check (maaien)									Firmware update!	
09/11/2022		Onsite check	All ok	All TERA logged out after =3-6 days	All ok	Data until 28/10?	1 and 2 OK	All ok	OK	NA	AB installed (2/3 transmitting via cellular)	
				reconnected app		AIR 1 data tot 8/10	NO2_3 data tot 30/10			clock sync ok (UTC+2 (1h advance)) charge+app+test ok		
						kapotte USB kabel. Vervangen	reconnected NO2_3					
16/11/2022		Onsite check	All ok	TERA 3 (Oneplus) still connected! But no data? TERA 2 and 3 offline + reconnected	All ok	AIR1 and 3 show data gaps no data AIR2?	NO2_1 and 2 show data gaps no data NO2_3	All ok	OK	Deployed at 12:14	All ok (2/3 online)	Deployed at 12:14
17/11/2022		Onsite filter change Observair/Bcmeter		TERA 3 reconnected						Filter change		Filter change
24/11/2022		Onsite check + filter change	All ok	All disconnected	All ok	AIR3 show data gaps no data AIR1+2?	NO2_1 and 2 show data gaps no data NO2_3	All ok	OK	Filter change	All ok (2/3 online)	Filter change
30/11/2022		Onsite check	All ok		All ok	no data	no data	All ok	OK	Filter change	All ok (2/3 online)	Unplugged sensors
		Stop BC measurements (Bcmeter)				contacted SODAQ	contacted SODAQ					
05/12/2022	9u	Removal shelter	All ok	TERA 2 (no new data)/3 disconnected TERA 1 still connected	All ok	AIR 1 until 11/11 AIR 2 until 27/10 AIR3 until 23/11	only SODAQ NO2_2 NO2_1 untill 22/11 NO2_3 until 30/10	All ok	All ok	All ok	All ok (2/3 online)	

BC CAMPAIGNE

3.2.2.1 Reference data

From R801, we collected NO_2 (Thermo 42C; µg/m³), O_3 (Teledyne API400E; µg/m³), PM_1, $PM_{2.5}$, PM_{10} (Palas FIDAS 200; µg/m³), BC (Thermo MAAP; µg/m³), relative humidity (%) and temperature (°C). The hourly data showed a good data availability with hourly (n=2132) data coverage of 96.7, 96.6 and 92.9% for, respectively, PM, BC and NO_2.

Descriptive statistics show $PM_{2.5}$ concentrations in the range of 1-51 µg/m³ (mean =10.85 µg/m³), while 2-111 µg/m³ is obtained for NO_2 (mean = 26 µg/m³). Atmospheric temperature varied between 1 and 27°C (mean = 13°C), while relative humidity was within 42 and 100% (mean=83.5%).

	NO_2 *µg/m³*	**O_3** *µg/m³*	**PM_1** *µg/m³*	**$PM_{2.5}$** *µg/m³*	**PM_{10}** *µg/m³*	**RH** *%*	**Temp** *°C*
Min	2.00	0.00	1.000	0.94	3.90	42.0	1.40
25%	15.00	9.00	3.500	5.64	12.90	76.5	9.80
Median	23.50	28.75	5.500	7.99	17.40	86.5	13.10
Mean	**26.06**	**28.57**	**8.879**	**10.85**	**20.14**	**83.5**	**12.90**
75%	34.00	45.00	11.000	13.63	24.90	93.0	15.95
Max	110.50	92.00	46.500	50.76	82.90	100.0	26.95
NA's	151	111	70	70	70	70	118

The temporal pollutant variability reflects typical urban pollution dynamics (Figure 52), with morning and evening rush hour peaks for NO_2 and BC, slightly delayed PM peaks with a regional background character and O_3 that is produced photochemically at low NO_2 concentrations and high solar radiation (inversely related to NO_2).

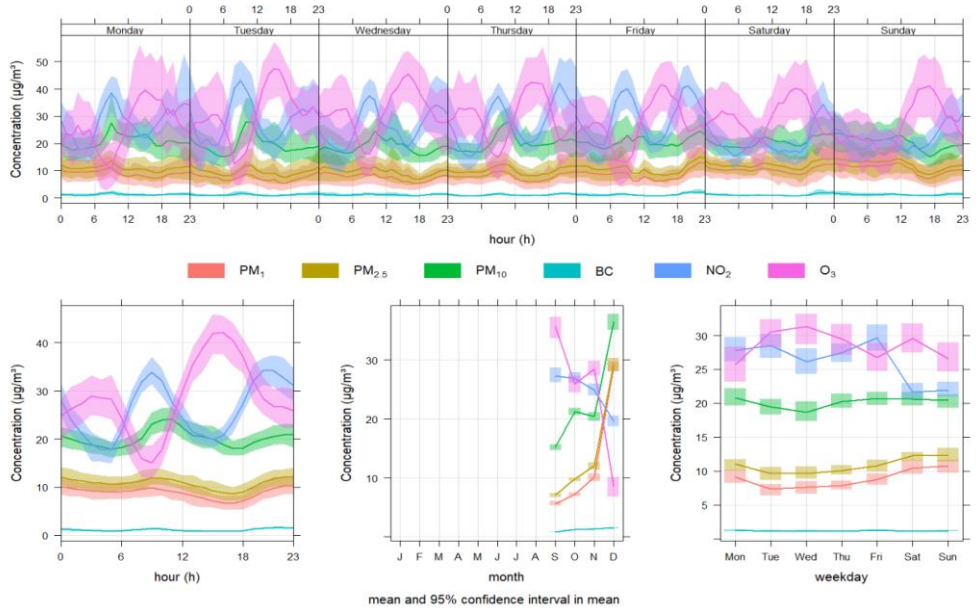


Figure 52 Temporal pollutant variability of PM, BC, NO_2 and O_3 at R801. Highlighted areas denote 95% CI.

3.2.2.2 Sensor data

3.2.2.2.1 ATMOTUBE PRO

The ATMOTube Pro was easy to use in the field and data transmission through the app worked fine, resulting in an overall good data coverage (Figure 53). However, due to a mistake of the user, the data period of 12/10-29/10 was not properly transmitted and lost. In addition, ATMO 3 stopped capturing data from 30/11/2022 around 14:22h. This resulted in an hourly data coverage of 70-79%.

START	STOP	NA	%	SENSOR
07/09/2022 15:00	05/12/2022 09:00	0	79,41	ATMO1
07/09/2022 15:00	05/12/2022 09:00	0	79,41	ATMO2
07/09/2022 15:00	30/11/2022 14:22	0	69,56	ATMO3

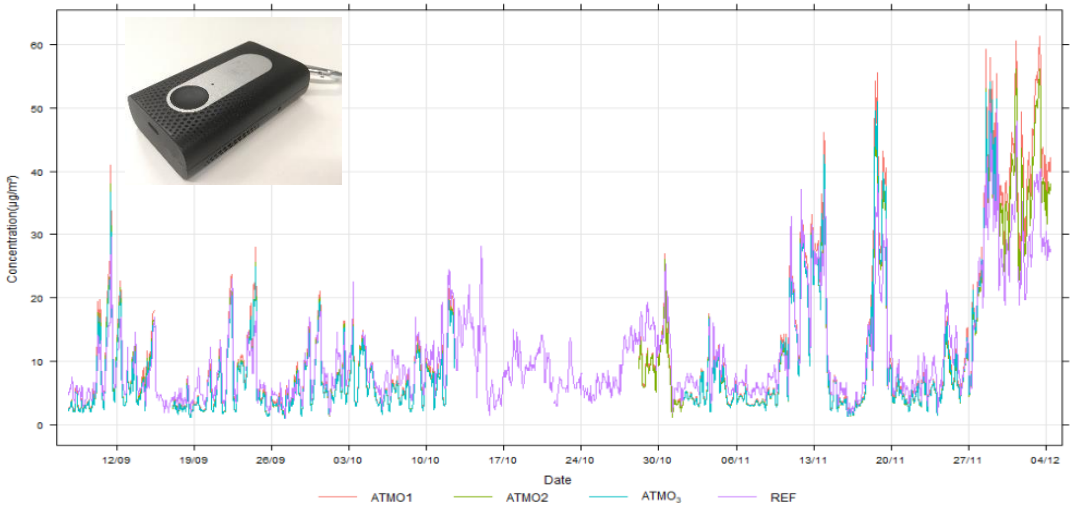


Figure 53 Hourly data coverage (upper) and resulting timeseries of $PM_{2.5}$ concentrations measured by the ATMO 1-3 sensors and the reference $PM_{2.5}$ monitor (µg/m³; lower).

When evaluating the hourly-averaged $PM_{2.5}$ data against the reference data (Figure 54), overall good correlations (R^2=0.87-0.89) are obtained with mean absolute errors of 2.6-3.3 µg/m³. The MBE (-1.27-0.12 µg/m³) indicates that the error varies around the mean and that there is no significant under- or overestimation by the sensor. The expanded uncertainty of the raw data already qualifies for the indicative data quality objective (<50%) for 2 out of 3 sensors (40 and 47%).

As observed during the lab tests, the variation between the sensors is small, with a between sensor uncertainty (BSU) of 0.58 µg/m³.

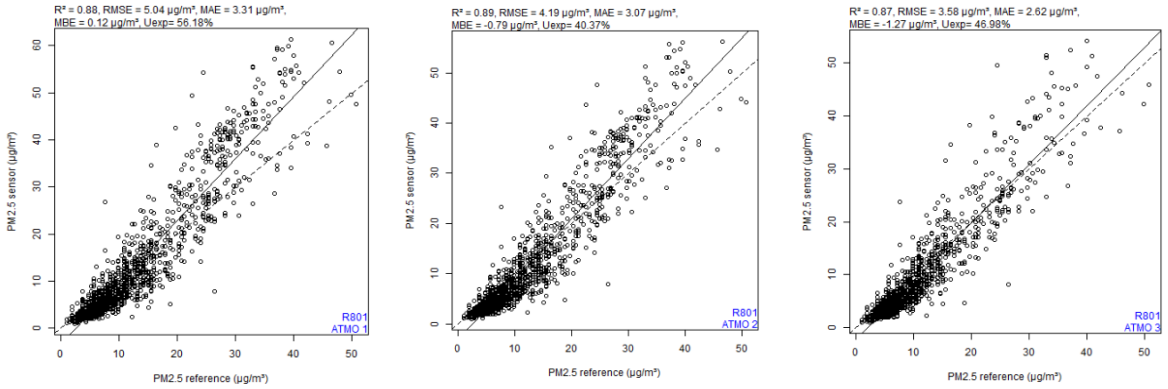


Figure 54 Regression plots of the reference $PM_{2.5}$ measurements against the concentrations measured by the ATMO sensors, with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

The sensor performance decreases for PM_{10} (R^2=0.64-0.66, MAE=8.82-9.1 µg/m³) and the association is entirely lost (R^2=0-0.03) when focusing on the coarse fraction (PM_{coarse}=PM_{10}-$PM_{2.5}$). This confirms our earlier observations during the lab coarse tests. We can, therefore, state that the PM sensor is not able to reliably quantify coarse particles.

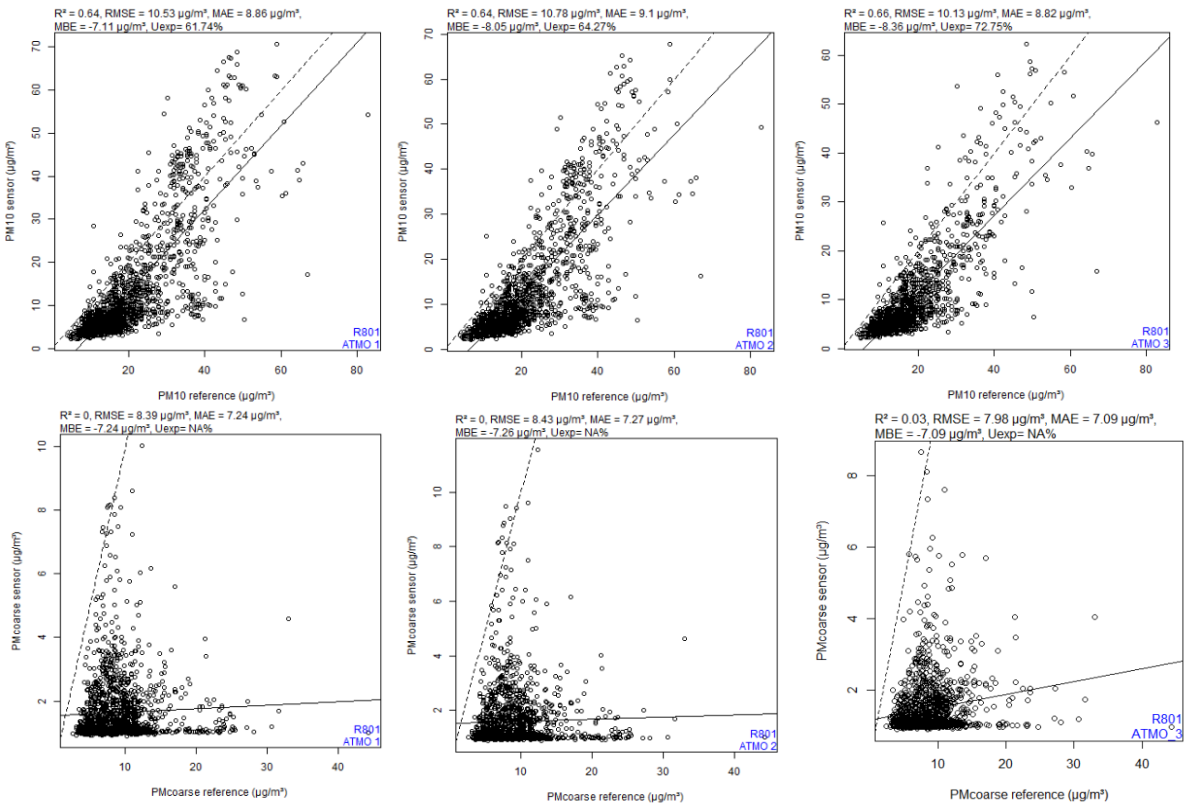


Figure 55 Regression plots of the reference PM_{10} (upper) and PM_{coarse} (lower) measurements against the concentrations measured by the ATMO 1-3 sensors (left-middle-right), with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

- **Sensor calibration**

When applying the lab calibration (lab-derived slope and intercept), the sensor accuracy worsens for both $PM_{2.5}$ (MAE=5.98 µg/m³, U_{exp}=120%) and PM_{10} (MAE=26 µg/m³, U_{exp}=334%). The lab-derived calibration does not seem to hold in field conditions, which is not surprising as field conditions are different in terms of PM composition and environmental conditions (temperature, relative humidity).

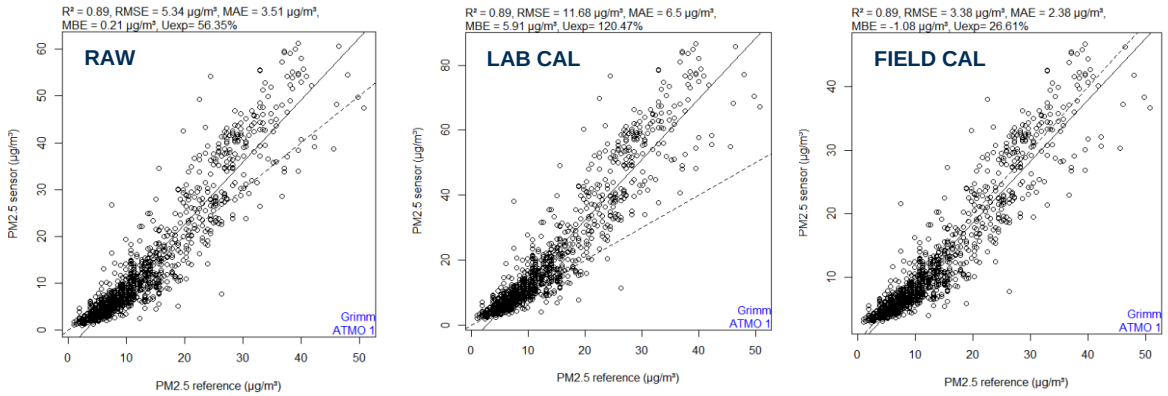


Figure 56 Comparability of ATMO1 $PM_{2.5}$ concentrations against the reference using the raw (left), lab-derived calibration (middle) and field-derived calibration (right).

A field calibration was conducted by deriving sensor- (ATMO 1-3) and PM fraction- (PM_1, $PM_{2.5}$ and PM_{10}) specific linear regression coefficients from a 2-week training period (7/9/2022-21/9/2022). The resulting calibration performance was evaluated based on the remaining 2.5 months of test data (22/9/2022-5/12/2022) and outperfomed the raw and lab-calibrated data with an mean accuracy (MAE) reaching 2.38 µg/m³ and an U_{exp} of 27% for $PM_{2.5}$.

	intercept	slope	R^2	sensor
coeff_PM1_1	-1.740084	1.2769819	0.8234760	PM1_1
coeff_PM1_2	-1.641693	1.1861026	0.8360657	PM1_2
coeff_PM1_3	-1.549820	1.1374006	0.8336587	PM1_3
coeff_PM2.5_1	-2.914829	1.3751440	0.7666170	PM2.5_1
coeff_PM2.5_2	-2.728792	1.2805454	0.7792343	PM2.5_2
coeff_PM2.5_3	-2.568031	1.2332829	0.7787030	PM2.5_3
coeff_PM10_1	-2.391964	0.6726374	0.3589985	PM10_1
coeff_PM10_2	-2.228885	0.6285376	0.3712579	PM10_2
coeff_PM10_3	-2.083355	0.6091420	0.3720475	PM10_3

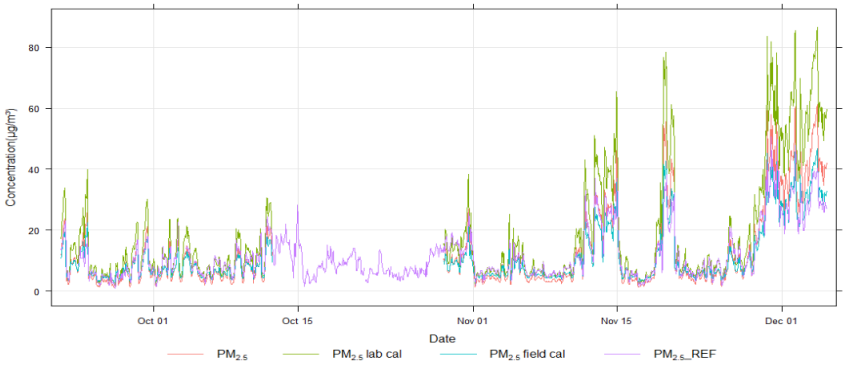


Figure 57 Field-derived intercept, slope and resulting R^2 of the 2-week training data (left), and resulting time series (right) of raw, lab- and field-calibrated $PM_{2.5}$ data (µg/m³) of ATMO1 and $PM_{2.5}$ reference (µg/m³).

- **Sensor drift**

Potential sensor drift was investigated by evaluating the sensor/REF ratio for both the raw and field-calibrated data over time (Figure 58). No distinct or gradual deviation in sensor/REF ratio is observed until December. From December onwards, the sensor/REF ratio seems to jump to 1.5 for the raw data and >1 for the field-calibrated data. This is likely an environmental effect (dust composition), as this period coincides with distinct PM peaks (Figure 57). We can

conclude that no gradual aging (sensor drift) can be observed based on the 3-month co-location, which was also not expected for this relative short period.

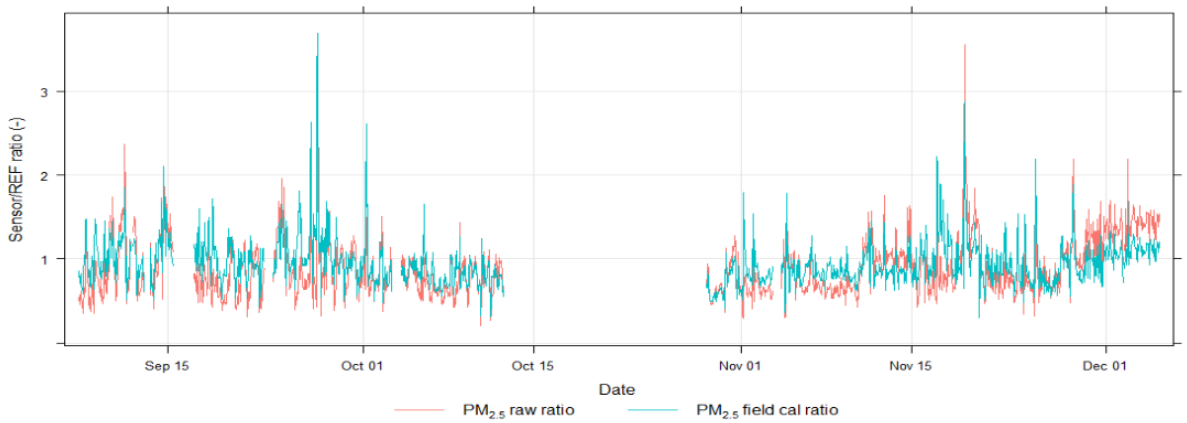


Figure 58 Timeseries of sensor/REF ratio for the raw and field calibrated $PM_{2.5}$ data of ATMO1.

- **Impact RH**

In order to evaluate a potential impact from relative humidity, we plotted the sensor/REF $PM_{2.5}$ ratio of ATMO 1 against the exhibited relative humidity (%). An increase in relative humidity seems to result in higher sensor/REF ratios, exponentially increasing for relative humidities above 80%.

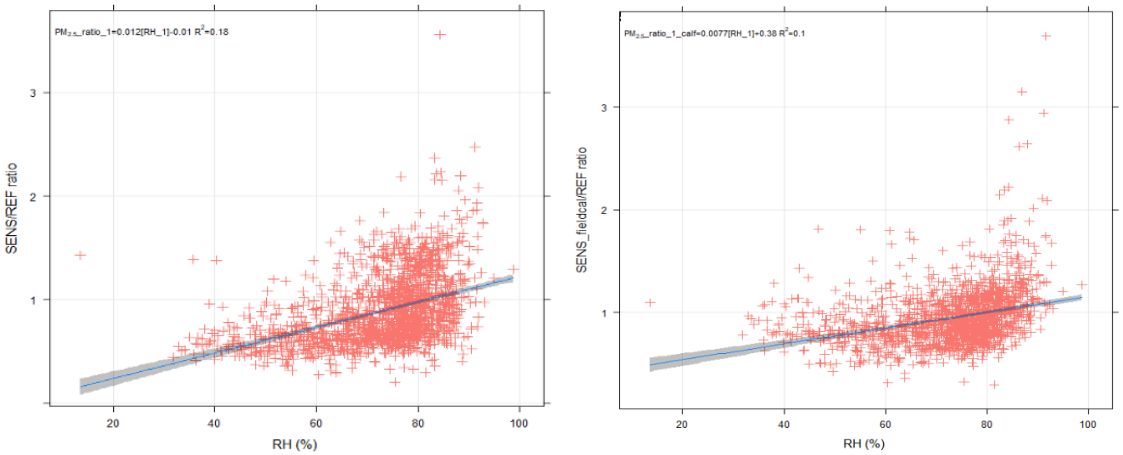


Figure 59 Impact of relative humidity (%) on sensor/REF ratio of raw (left) and field-calibrated (right) $PM_{2.5}$ data of ATMO1

3.2.2.2.2 Open Seneca

The Open Seneca was easy to use in the field and data offload via the SD card worked seamlessly (Figure 60). This resulted in an hourly data coverage of 100% for all 3 sensors.

START	STOP	NA	%	SENSOR
07/09/2022 15:00	05/12/2022 09:00	0	99,95	OPEN1
07/09/2022 15:00	05/12/2022 09:00	0	99,95	OPEN2
07/09/2022 15:00	05/12/2022 09:00	0	99,95	OPEN3

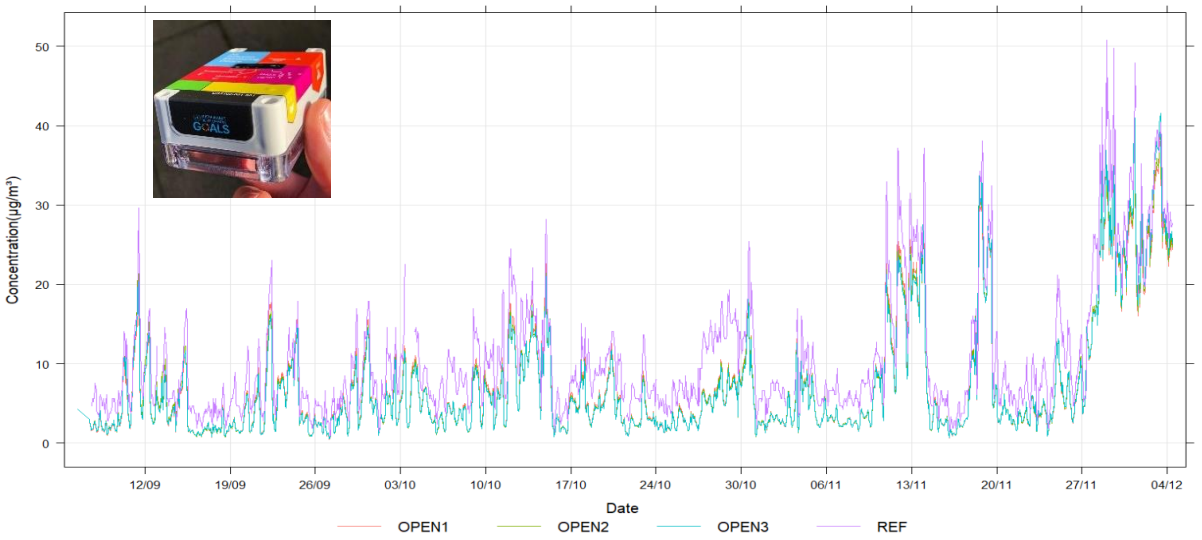


Figure 60 Hourly data coverage (upper) and resulting timeseries of $PM_{2.5}$ concentrations measured by the Open Seneca (OPEN1-3) sensors and the reference $PM_{2.5}$ monitor (µg/m³; lower).

When evaluating the hourly-averaged $PM_{2.5}$ data against the reference data (Figure 61), overall good correlations (R^2=0.89-0.90) are obtained with mean absolute errors of 3.64-3.73 µg/m³. The MBE (-3.52- -3.58 µg/m³) indicates that the sensor slightly underestimates reference concentrations. The expanded uncertainty of the raw data already qualifies for the indicative data quality objective (<50%) for 3 out of 3 sensors (34-35%).

As observed during the lab tests, the variation between the sensors is small, with a between sensor uncertainty (BSU) of 0.33 µg/m³.

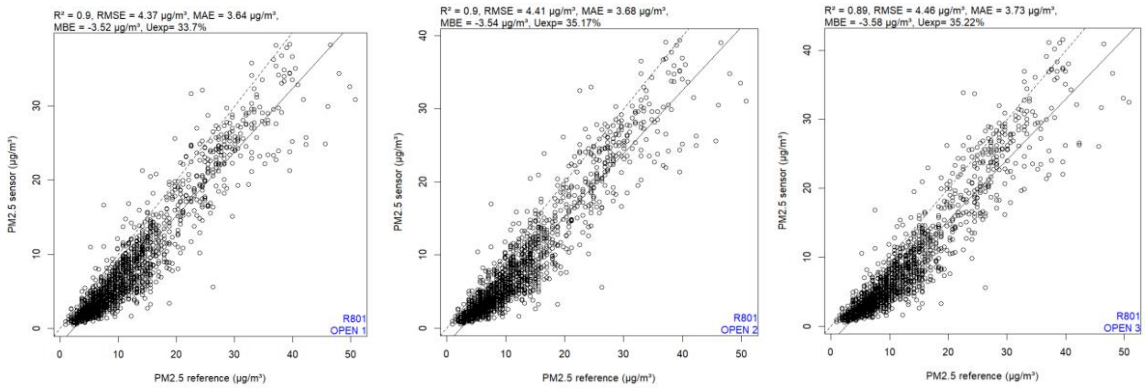


Figure 61 Regression plots of the reference $PM_{2.5}$ measurements against the concentrations measured by the Open Seneca sensors, with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

The sensor performance decreases for PM_{10} (R^2=0.6-0.62, MAE=12.6 µg/m³) and the association is entirely lost (R^2=0-0.01) when focusing on the coarse fraction (PM_{coarse}=PM_{10}-$PM_{2.5}$) as can be observed from Figure 62. This confirms our earlier observations during the lab coarse tests. We can, therefore, state that the PM sensor is not able to reliably quantify coarse particles.

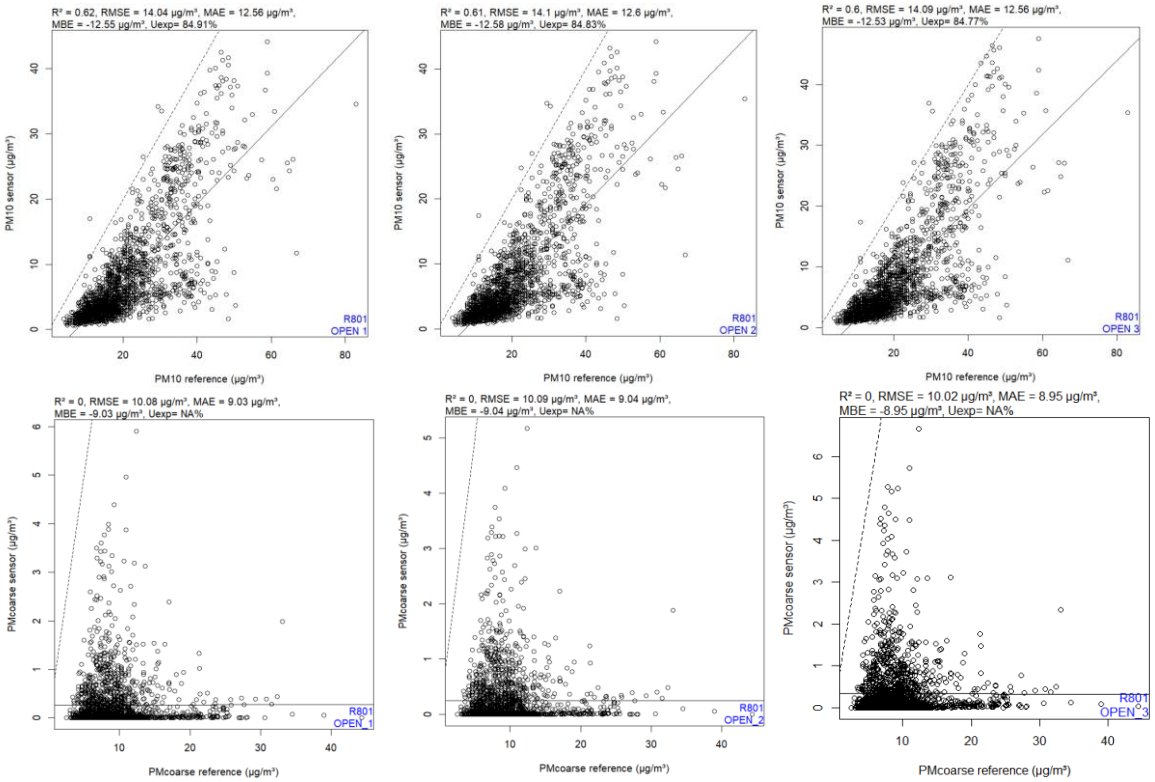


Figure 62 Regression plots of the reference PM_{10} (upper) and PM_{coarse} (lower) measurements against the concentrations measured by the Open Seneca sensors (OPEN1-3; left-middle-right), with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

- **Sensor calibration**

When applying the lab calibration (lab-derived slope and intercept; Figure 63), the sensor accuracy worsens for both $PM_{2.5}$ (MAE=3.58 µg/m³, U_{exp}=67%) and PM_{10} (MAE=18 µg/m³, U_{exp}=245%). The lab-derived calibration does not seem to hold in field conditions, which is not surprising as field conditions are different in terms of PM composition and environmental conditions (temperature, relative humidity).

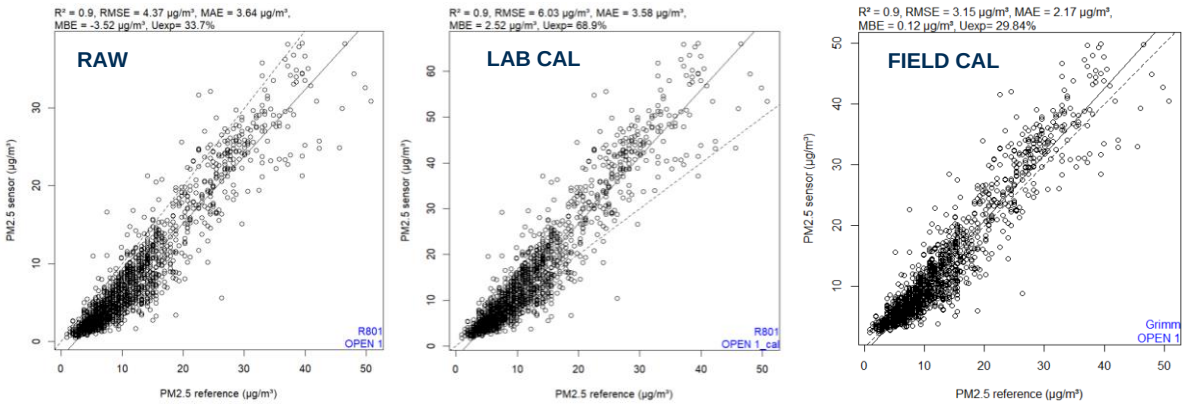


Figure 63 Comparability of Open Seneca $PM_{2.5}$ concentrations against the reference using the raw (left), lab-derived calibration (middle) and field-derived calibration (right) on the test data (22/9/2022-5/12/2022).

A field calibration was conducted by deriving sensor- (OPEN1-3) and PM fraction- (PM_1, $PM_{2.5}$ and PM_{10}) specific slopes and intercepts (Figure 64) based on a 2-week training period (7/9/2022-21/9/2022). The resulting calibration performance was evaluated based on the remaining 2.5 months of test data (22/9/2022-5/12/2022) and outperfomed the raw and lab-calibrated data with a mean accuracy (MAE) reaching 2.17 µg/m³ and an U_{exp} of 30% for $PM_{2.5}$ (Figure 64).

	intercept	slope	R^2	sensor
coeff_PM1_1	-0.2089451	0.7711502	0.8596761	PM1_1
coeff_PM1_2	-0.4206327	0.8640786	0.8523179	PM1_2
coeff_PM1_3	-0.3270691	0.8043029	0.8541878	PM1_3
coeff_PM2.5_1	-1.4073876	0.7976602	0.7999471	PM2.5_1
coeff_PM2.5_2	-1.7581784	0.8897583	0.7823791	PM2.5_2
coeff_PM2.5_3	-1.5403564	0.8296911	0.7875436	PM2.5_3
coeff_PM10_1	-1.5139479	0.3674068	0.3925708	PM10_1
coeff_PM10_2	-2.2288853	0.6285376	0.2368245	PM10_2
coeff_PM10_3	-1.4921193	0.3737799	0.3653570	PM10_3

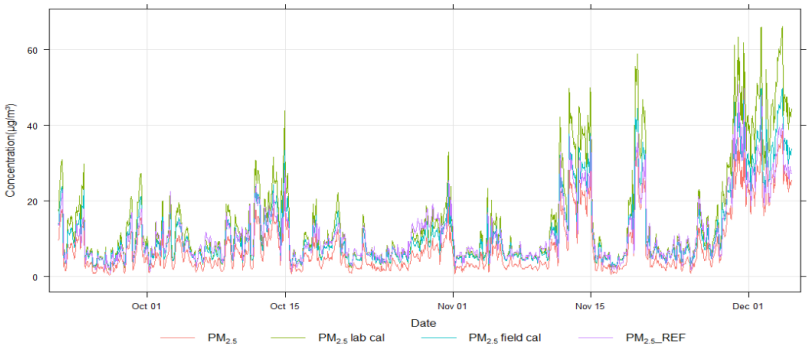


Figure 64 Field-derived intercept, slope and resulting R^2 of the 2-week training data (left), and resulting time series (right) of raw, lab- and field-calibrated $PM_{2.5}$ data (µg/m³) of OPEN1 and $PM_{2.5}$ reference (µg/m³).

- **Sensor drift**

Potential sensor drift was investigated by evaluating the sensor/REF ratio for both the raw and field-calibrated data over time (Figure 65). No distinct or gradual deviation in sensor/REF ratio is observed based on the 3-month co-location period.

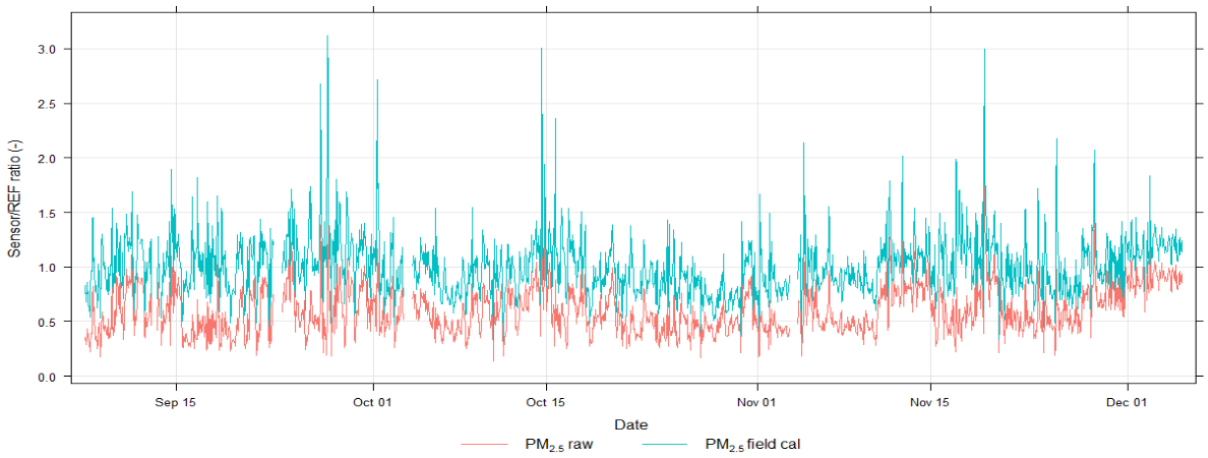


Figure 65 Timeseries of sensor/REF ratio for the raw and field calibrated $PM_{2.5}$ data of Open Seneca 1.

- **Impact RH**

In order to evaluate a potential impact from relative humidity, we plotted the sensor/REF $PM_{2.5}$ ratio of OPEN 1 raw and field-calibrated data against the exhibited relative humidity (%). An increase in relative humidity seems to result in higher sensor/REF ratios, exponentially increasing for relative humidities over 80% (Figure 66).

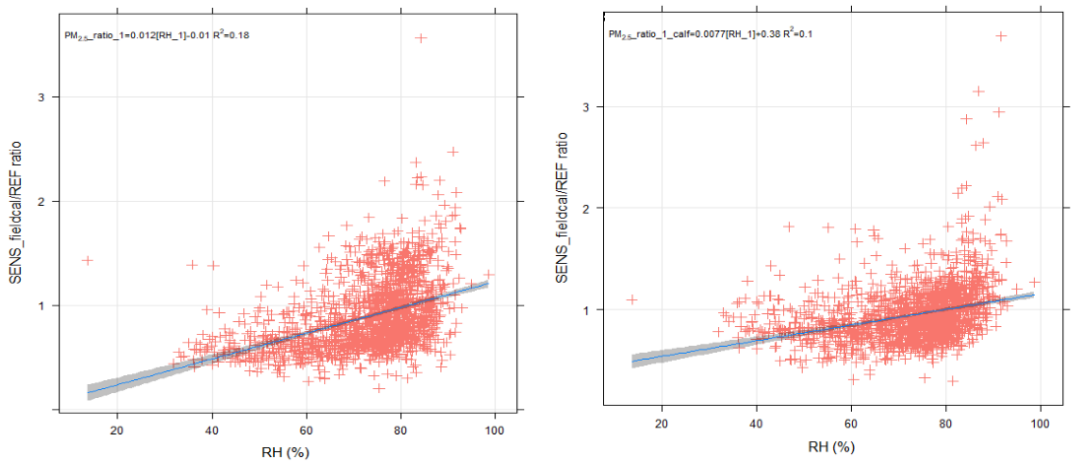


Figure 66 Impact of relative humidity (%) on sensor/REF ratio of raw (left) and field-calibrated (right) $PM_{2.5}$ data of OPEN1

3.2.2.2.3 TERA PMscan

The TERA PMscan needed app connectivity in order to log measurements and smartphone application automatically shut down after 1-3 days in continuous operation (Figure 67). This resulted in a low hourly data coverage of 7-33%. Although performant in short-term monitoring campaigns (developed for personal exposure mapping when on the move), instrument design is, therefore, not suitable for continuous long-term monitoring.

START	STOP	NA	%	SENSOR
07/09/2022 15:00	05/12/2022 08:59	0	33,4	TERA1
07/09/2022 15:00	24/11/2022 12:27	0	11,1	TERA2
07/09/2022 15:00	01/12/2022 22:42	0	6,5	TERA3

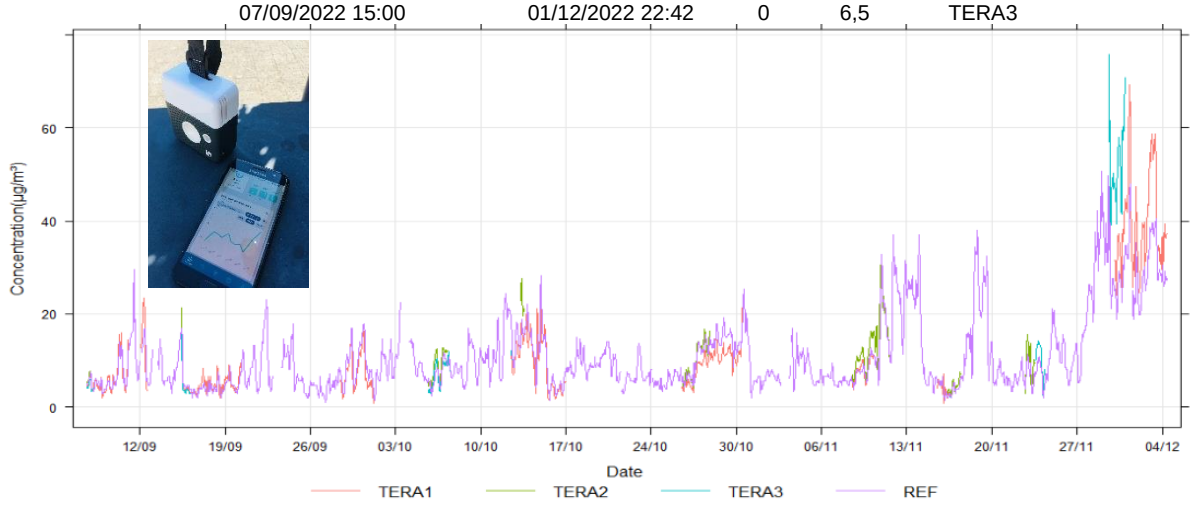


Figure 67 Hourly data coverage (upper) and resulting timeseries of $PM_{2.5}$ concentrations measured by the Open Seneca (TERA1-3) sensors and the reference $PM_{2.5}$ monitor (µg/m³; lower).

When evaluating the available hourly-averaged $PM_{2.5}$ data against the reference data (Figure 68), overall good correlations (R^2=0.81-0.90) are obtained with mean absolute errors of 2.16-7.24 µg/m³. The MBE (0.7-5.82 µg/m³) indicates that the sensor slightly overestimates

reference concentrations. The expanded uncertainty of the raw data already qualifies for the indicative data quality objective (<50%) for 2 out of 3 sensors (18 and 49.9%).

As observed during the lab tests, the variation between the sensors is small, with a between sensor uncertainty (BSU) of 0.11 µg/m³.

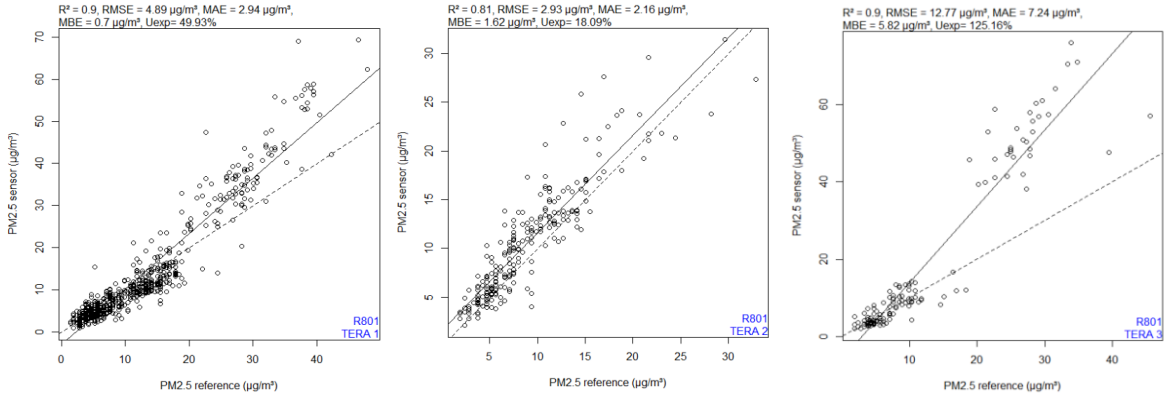


Figure 68 Regression plots of the reference $PM_{2.5}$ measurements against the concentrations measured by the TERA1-3 sensors, with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

The sensor performance decreases for PM_{10} (R^2=0.51-0.68, MAE=5.22-10.86 µg/m³) and the association decreases further (R^2=0.18-0.4) when focusing on the coarse fraction (PM_{coarse}=PM_{10}-$PM_{2.5}$) as can be observed from Figure 69. This confirms our earlier observations during the lab coarse tests. We can, therefore, state that the PM sensor is not able to reliably quantify coarse particles, although best PM_{coarse} performance is obtained when compared to the other PM sensors.

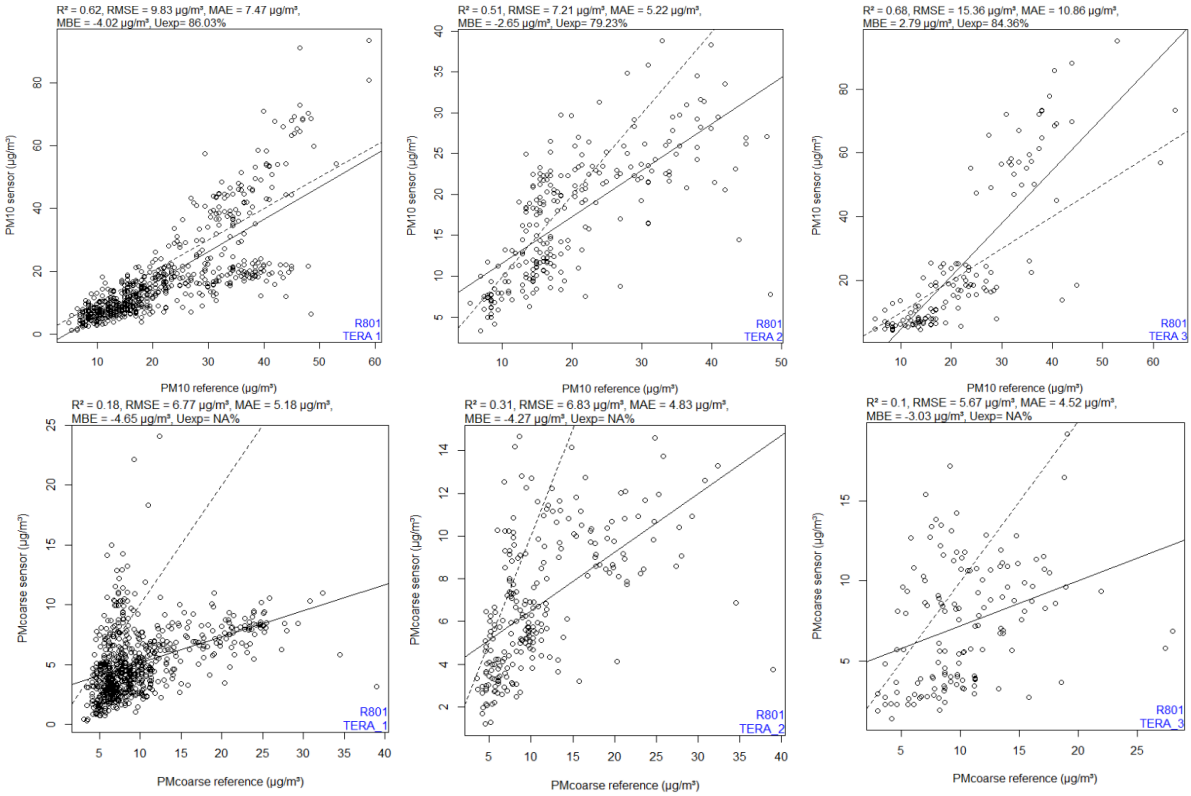


Figure 69 Regression plots of the reference PM_{10} (upper) and PM_{coarse} (lower) measurements against the concentrations measured by the TERA PMscan sensors (TERA1-3; left-middle-right), with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

- **Sensor calibration**

When applying the lab calibration (lab-derived slope and intercept; Figure 63), the sensor accuracy worsens for both $PM_{2.5}$ (MAE=5.63 µg/m³, U_{exp}=88%) and PM_{10} (MAE=16 µg/m³, U_{exp}=241%). The lab-derived calibration does not seem to hold in field conditions, which is not surprising as field conditions are different in terms of PM composition and environmental conditions (temperature, relative humidity).

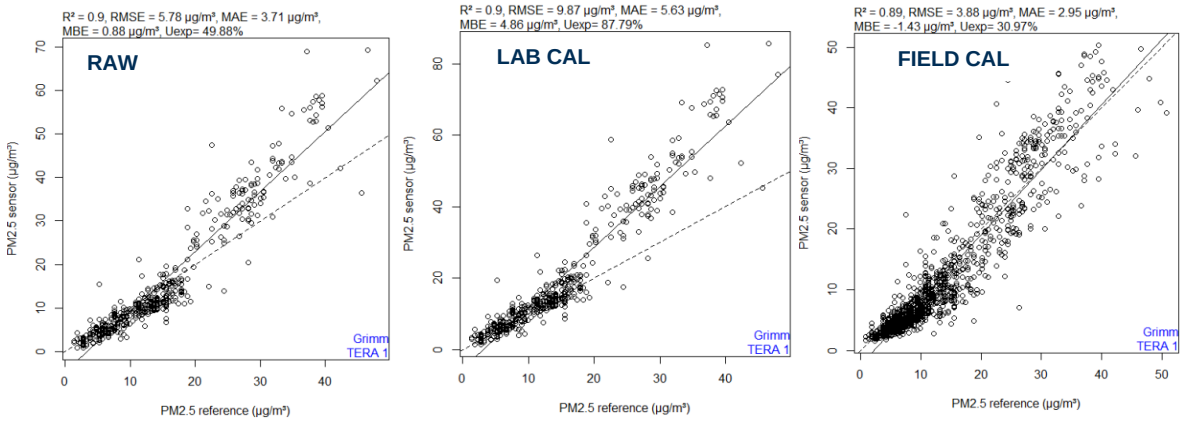


Figure 70 Comparability of TERA $PM_{2.5}$ concentrations against the reference using the raw (left), lab-derived calibration (middle) and field-derived calibration (right) on the test data (22/9/2022-5/12/2022).

A field calibration was conducted by deriving sensor- (TERA1-3) and PM fraction- (PM_1, $PM_{2.5}$ and PM_{10}) specific slopes and intercepts based on a 2-week training period (7/9/2022-21/9/2022). The resulting calibration performance was evaluated based on the remaining 2.5 months of test data (22/9/2022-5/12/2022) and outperfomed the raw and lab-calibrated data with a mean accuracy (MAE) reaching 2.95 µg/m³ and an U_{exp} of 31% for $PM_{2.5}$ (Figure 70).

3.2.2.2.4 SODAQ NO2

The SODAQ NO2 automatically transmits data via GPRS/4G. No data offload actions were therefore necessary. Although convenient, the sensors seemed to experience connectivity issues, independently of each other, ultimately resulting in low hourly data coverages between 34 and 51%. Moreover, NO_2 signal noise was exhibited for sensor 2 and 3, although much less frequent than observed during the lab tests. These noise events occur independently at times that other sensors are operating normally, suggesting that the cause of these events is rather sensor-specific and not caused by external confounders (e.g. electromagnetic interferents).

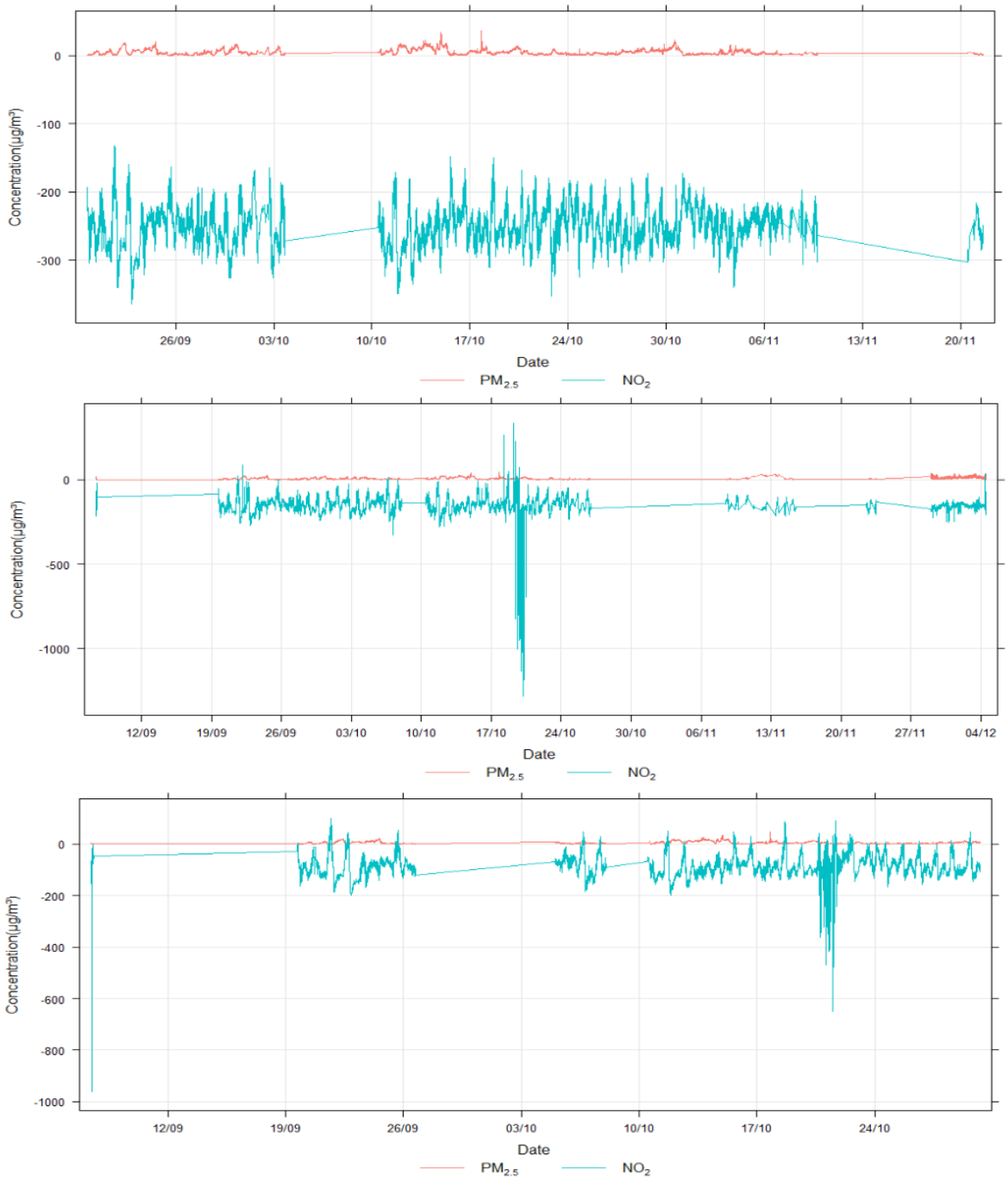


Figure 71 Raw $PM_{2.5}$ and NO_2 sensor data collected by SODAQ NO_2 1 (upper), 2 (middle) and 3 (lower) during the 3-month co-location campaign

START	STOP	NA	%	SENSOR
19/09/2022 16:41	22/11/2022 13:20	0	50,5	SODAQ NO2_1
19/09/2022 16:56	05/12/2022 09:00	0	49,3	SODAQ NO2_2
19/09/2022 16:57	30/10/2022 05:31	0	33,5	SODAQ NO2_3

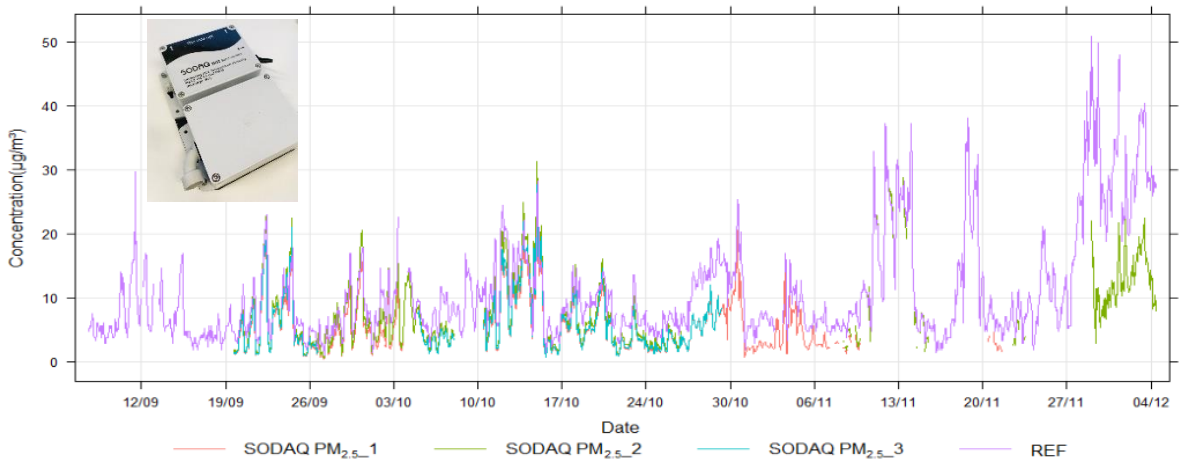


Figure 72 Hourly data coverage (upper) and resulting timeseries of $PM_{2.5}$ concentrations measured by the SODAQ (1-3) sensors and the reference $PM_{2.5}$ monitor (µg/m³; lower).

- **PM performance**

When evaluating the available hourly-averaged $PM_{2.5}$ data against the reference data (Figure 73), lower correlations are observed when compared to the other PM sensors (R^2=0.53-0.75) are obtained with mean absolute errors of 3.44-4.37 µg/m³. The MBE (-3.1- -3.8µg/m³) indicates that the sensor slightly underestimates reference concentrations. The expanded uncertainty of the raw data already qualifies for the indicative data quality objective (<50%) for 2 out of 3 sensors (14 and 24%).

As observed during the lab tests, the variation between the sensors is higher as well when compared to the other PM sensors, with a between sensor uncertainty (BSU) of 0.44 µg/m³.

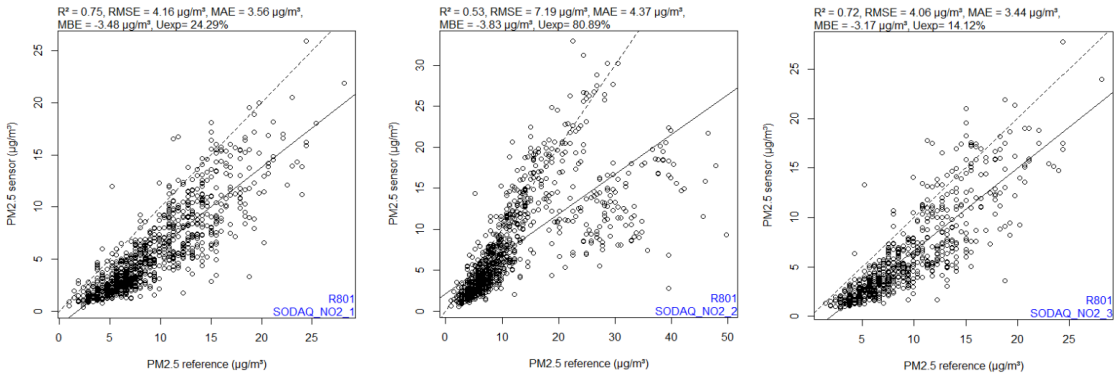


Figure 73 Regression plots of the reference $PM_{2.5}$ measurements against the concentrations measured by the TERA1-3 sensors, with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

The sensor performance decreases for PM_{10} (R^2=0.26-0.46, MAE=12.8-14.1 µg/m³) and the association decreases further (R^2=0-0.06) when focusing on the coarse fraction (PM_{coarse}=PM_{10}-$PM_{2.5}$) as can be observed from Figure 74. This confirms our earlier observations during the lab coarse tests. We can, therefore, state that the PM sensor is not able to reliably quantify coarse particles.

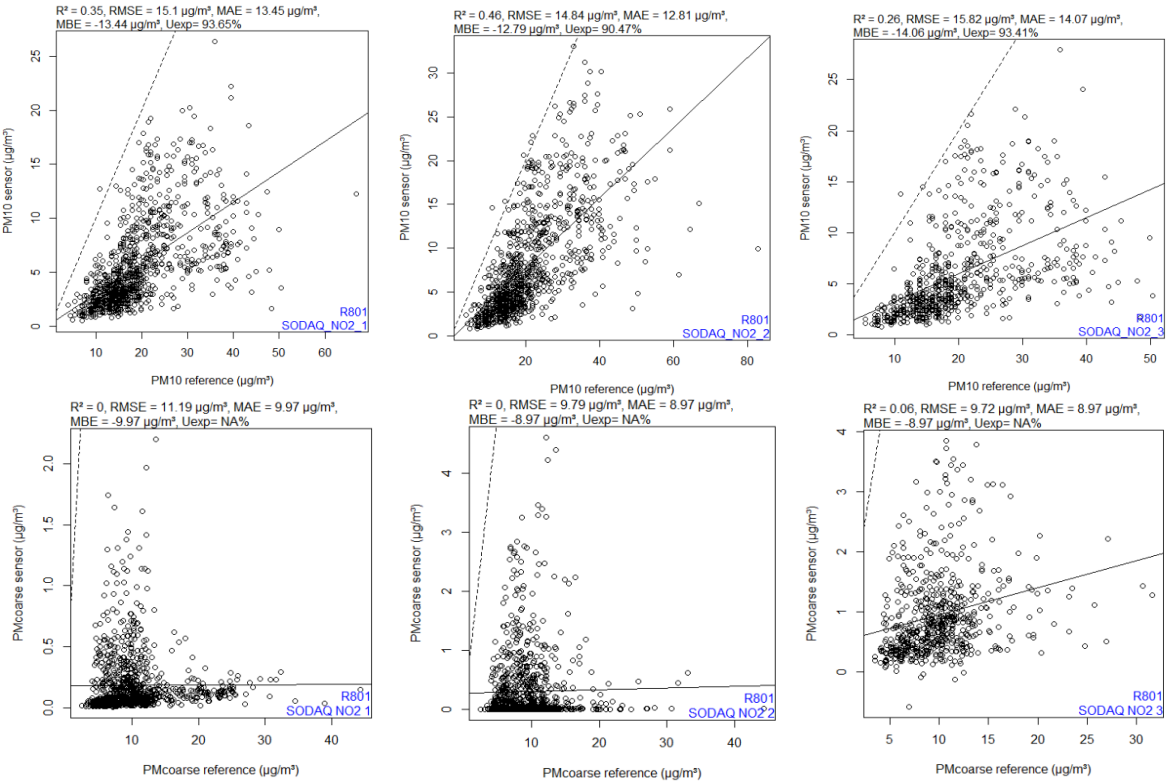


Figure 74 Regression plots of the reference PM_{10} (upper) and PM_{coarse} (lower) measurements against the concentrations measured by the SODAQ NO2 sensors (SODAQ_NO2_1-3; left-middle-right), with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

- **NO_2 performance**

When evaluating the hourly raw performance for NO_2 (Figure 75), negative linear associations (R^2=0.18-0.62) are observed between the raw NO_2 sensor data and the reference concentrations, with very high errors (MAE=112-277 µg/m³ and Uexp=420-826%). This confirms the lab results and indicates that the sensor is not properly calibrated.

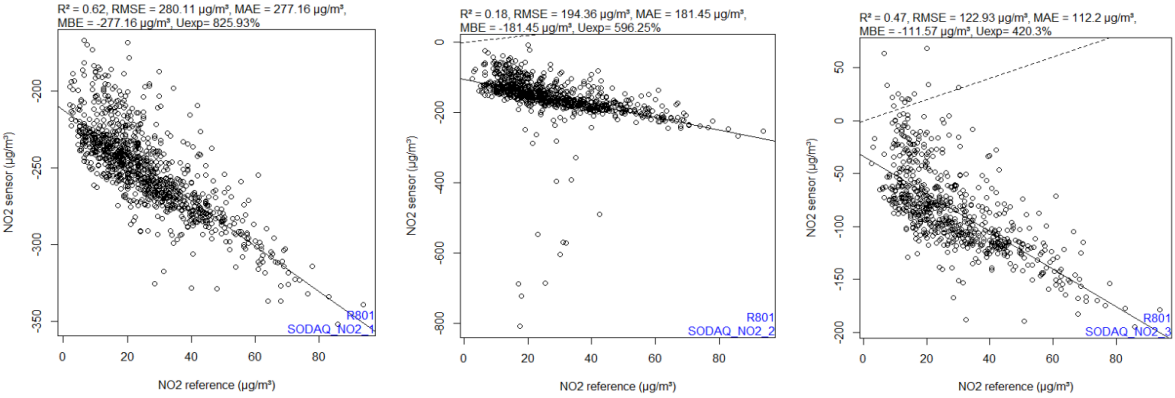


Figure 75 Regression plots of the reference NO_2 measurements against the concentrations measured by the SODAQ NO2 sensors (SODAQ_NO2_1-3; left-middle-right), with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

- **Sensor calibration**

When applying the lab calibration (lab-derived slope and intercept;Figure 76), the sensor accuracy worsens for both $PM_{2.5}$ (MAE=6.63 µg/m³, U_{exp}=91%) and PM_{10} (MAE=29 µg/m³, U_{exp}=73%), while sensor accuracy improves for NO_2 (MAE=27 µg/m³ and Uexp=108%). The lab-derived calibration for PM does not seem to hold in field conditions, which is not surprising as field conditions are different in terms of PM composition and environmental conditions (temperature, relative humidity). For NO_2, the lab-derived calibration yields an initial sensor response to NO_2 and already shows to perform better than the raw sensor readings.

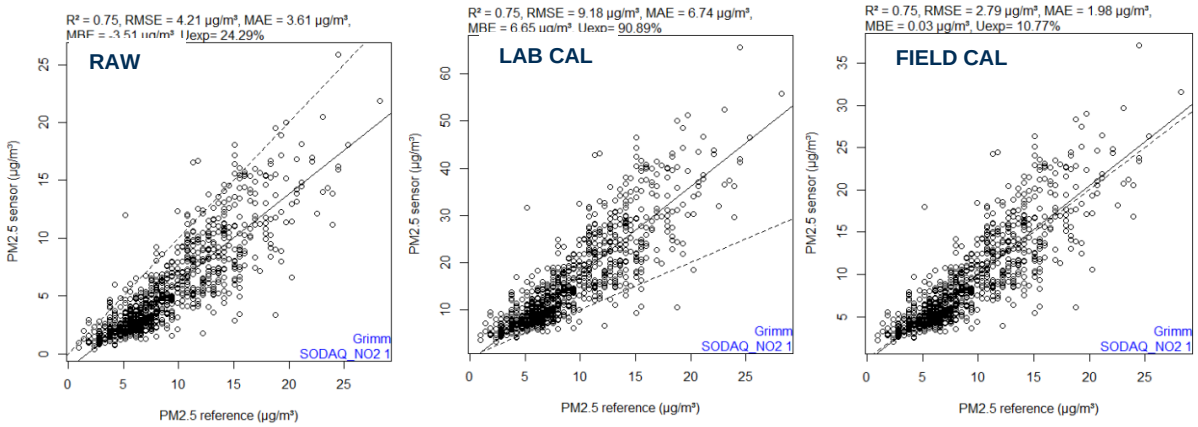


Figure 76 Comparability of TERA $PM_{2.5}$ concentrations against the reference using the raw (left), lab-derived calibration (middle) and field-derived calibration (right) on the test data (22/9/2022-5/12/2022).

A field calibration was conducted by deriving sensor- (SODAQ_NO2 1-3) and PM fraction- (PM_1, $PM_{2.5}$ and PM_{10}) specific slopes and intercepts based on a 2-week training period (7/9/2022-21/9/2022). The resulting calibration performance was evaluated based on the remaining 2.5 months of test data (22/9/2022-5/12/2022) and outperfomed the raw and lab-calibrated data with a mean accuracy (MAE) reaching 1.96 µg/m³ and an U_{exp} of 10.77% for $PM_{2.5}$ (Figure 76).

To test whether a field calibration performed better than the lab calibration for NO_2, a multilinear calibration model was trained with covariates for sensor response, temperature, RH and O_3 following earlier calibration studies (2-4). Model training was based on 2 weeks of co-location data (19/9/2022-4/10/2022) and the calibration performance was tested on the remaining 2 months of test data (Figure 77). The multilinear field calibration outperformed the raw and lab-derived calibration with R^2=0.82 and a MAE of 5.63 µg/m³. The expanded uncertainty improved significantly as well, from 826 to 37%, however, still not qualifying the indicative data quality objective (<25%).

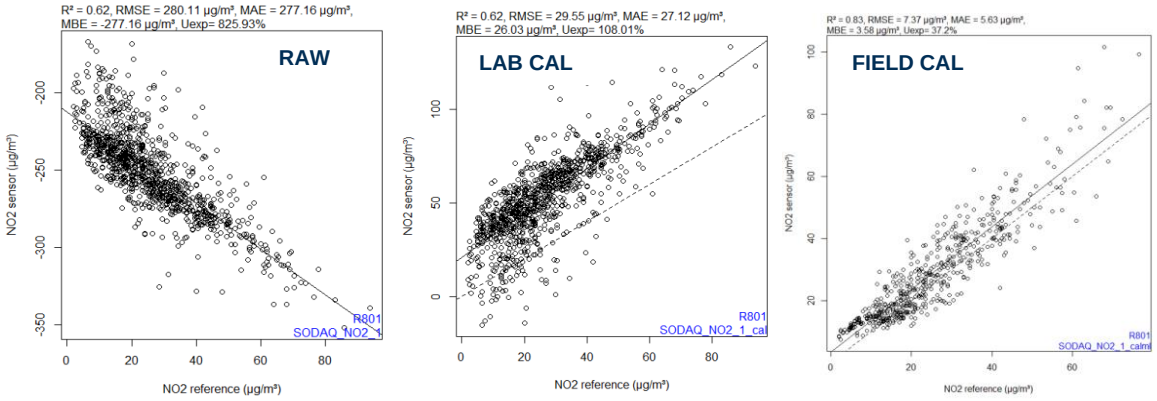


Figure 77 Comparability of SODAQ NO_2 1 concentrations against the reference using the raw (left), lab-derived calibration (middle) and field-derived calibration on 2 months of test data (right).

3.2.2.2.5 SODAQ AIR

The SODAQ AIR automatically transmits data via GPRS/4G. No data offload actions were therefore necessary. Although convenient, the sensors seemed to experience connectivity issues, independently of each other, ultimately resulting in low hourly data coverages between 34 and 53%.

START	STOP	NA	%	SENSOR
08/09/2022 14:39	11/11/2022 14:59	0	34,43	SODAQ AIR_1
08/09/2022 14:36	27/10/2022 22:36	0	52,81	SODAQ AIR_2
08/09/2022 14:36	23/11/2022 15:40	0	43,29	SODAQ AIR_3

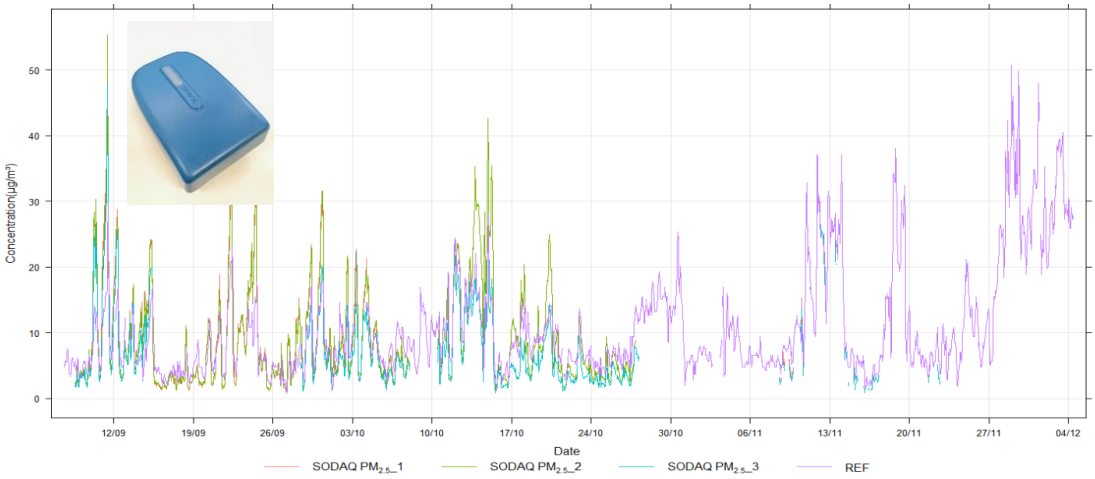


Figure 78 Hourly data coverage (upper) and resulting timeseries of $PM_{2.5}$ concentrations measured by the SODAQ AIR sensors (1-3) and the reference $PM_{2.5}$ monitor (µg/m³; lower).

When evaluating the available hourly-averaged $PM_{2.5}$ data against the reference data (Figure 79), fair correlations (R^2=0.68-0.69) are obtained with mean absolute errors of 3.1-3.2 µg/m³. The MBE (-1.8-0.9 µg/m³) indicates that the error varies around the reference concentration, implying that the sensors do not significantly over- or underestimate reference concentrations. The expanded uncertainty of the raw data already qualifies for the indicative data quality objective (<50%) for all sensors (3.7 - 23%).

The variation between the sensors is rather high, compared to the other PM sensors, with a between sensor uncertainty (BSU) of 0.71 µg/m³.

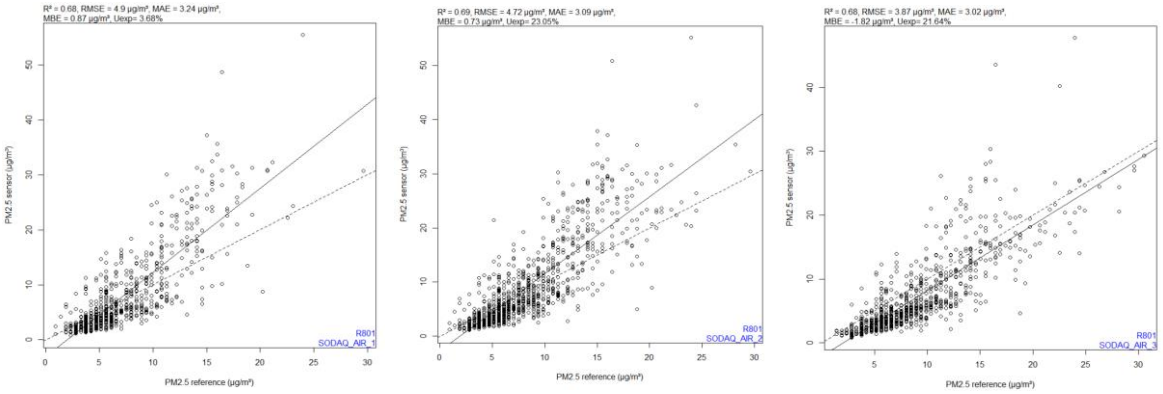


Figure 79 Regression plots of the reference $PM_{2.5}$ measurements against the concentrations measured by the SODAQ AIR1-3 sensors, with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

The sensor performance decreases for PM_{10} (R^2=0.3-0.32, MAE=8.4-11.5 µg/m³) and the association decreases further (R^2=0-0.01) when focusing on the coarse fraction (PM_{coarse}=PM_{10}-$PM_{2.5}$) as can be observed from Figure 80. This confirms our earlier observations during the lab coarse tests. We can, therefore, state that the PM sensor is not able to reliably quantify coarse particles, although best PM_{coarse} performance is obtained when compared to the other PM sensors.

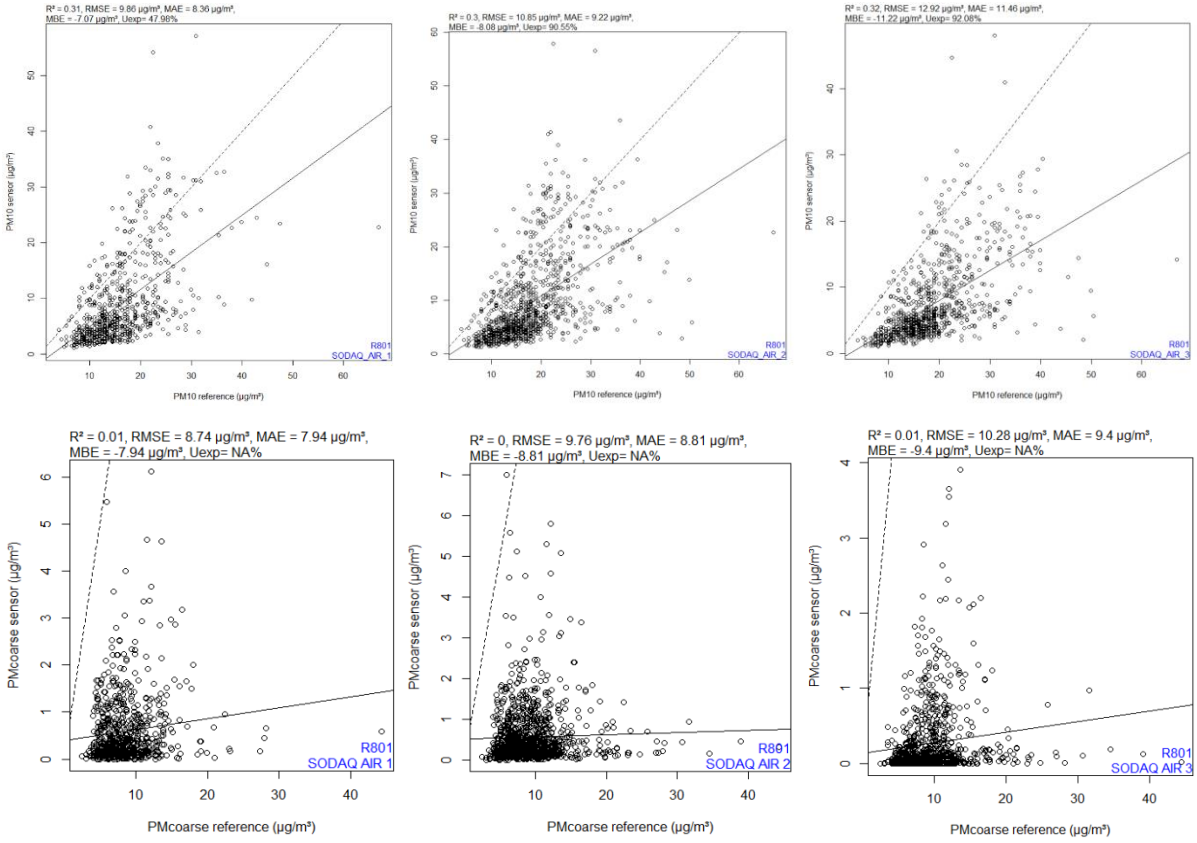


Figure 80 Regression plots of the reference PM_{10} (upper) and PM_{coarse} (lower) measurements against the concentrations measured by the SODAQ AIR sensors (SODAQ_AIR_1-3; left-middle-right), with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

- **Sensor calibration**

When applying the lab calibration (lab-derived slope and intercept; **Error! Reference source not found.**), the sensor accuracy worsens for both $PM_{2.5}$ (MAE=3.25 µg/m³) and PM_{10} (MAE=18 µg/m³). The lab-derived calibration does not seem to hold in field conditions, which is not surprising as field conditions are different in terms of PM composition and environmental conditions (temperature, relative humidity).

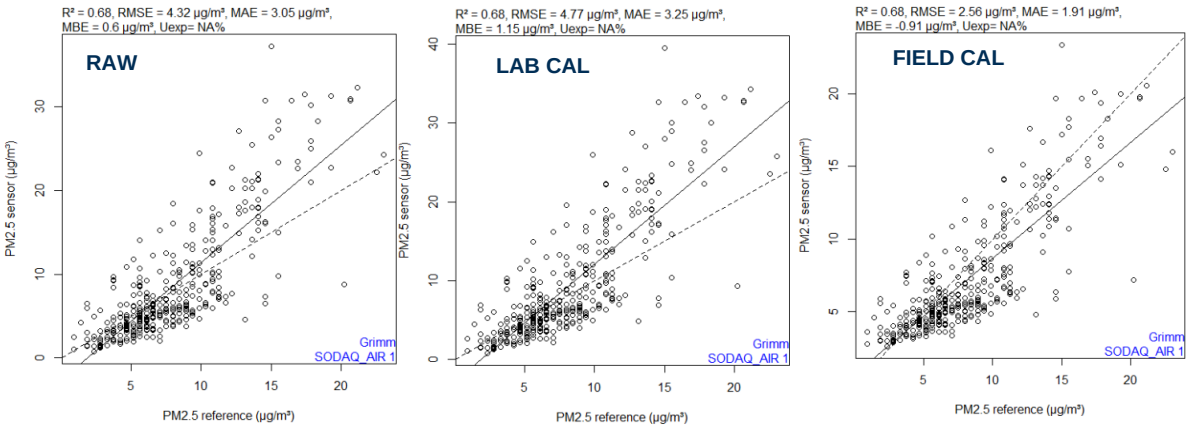


Figure 81 Comparability of SODAQ AIR $PM_{2.5}$ concentrations against the reference using the raw (left), lab-derived calibration (middle) and field-derived calibration (right) on the test data (22/9/2022-5/12/2022).

A field calibration was conducted by deriving sensor- (SODAQ_AIR 1-3) and PM fraction- (PM_1, $PM_{2.5}$ and PM_{10}) specific slopes and intercepts based on a 2-week training period (7/9/2022-21/9/2022). The resulting calibration performance was evaluated based on the remaining 2.5 months of test data (22/9/2022-5/12/2022) and outperfomed the raw and lab-calibrated data with a mean accuracy (MAE) reaching 1.91 µg/m³ for $PM_{2.5}$ (Figure 76).

3.2.2.2.6 AIRBEAM

The AIRBEAM arrived later and was, therefore, co-located later (5/10/2022) than the other PM sensors. AIRBEAM automatically transmits its data via GPRS/4G, while redundancy is foreseen via SD. During the co-location campaign, we experienced connectivity issues to the Proximus network, requiring a firmware update to resolve the issue (and additional data loss between 26/10 and 9/11). Data was ultimately offloaded via the SD cards and resulted in an hourly data coverage of 52-53% (Figure 82).

START	STOP	NA	%	SENSOR
05/10/2022 16:51	05/12/2022 08:53	0	52,3	AIRBEAM 1
05/10/2022 16:53	05/12/2022 08:59	0	53,19	AIRBEAM 2
05/10/2022 17:06	05/12/2022 08:59	0	53,28	AIRBEAM 3

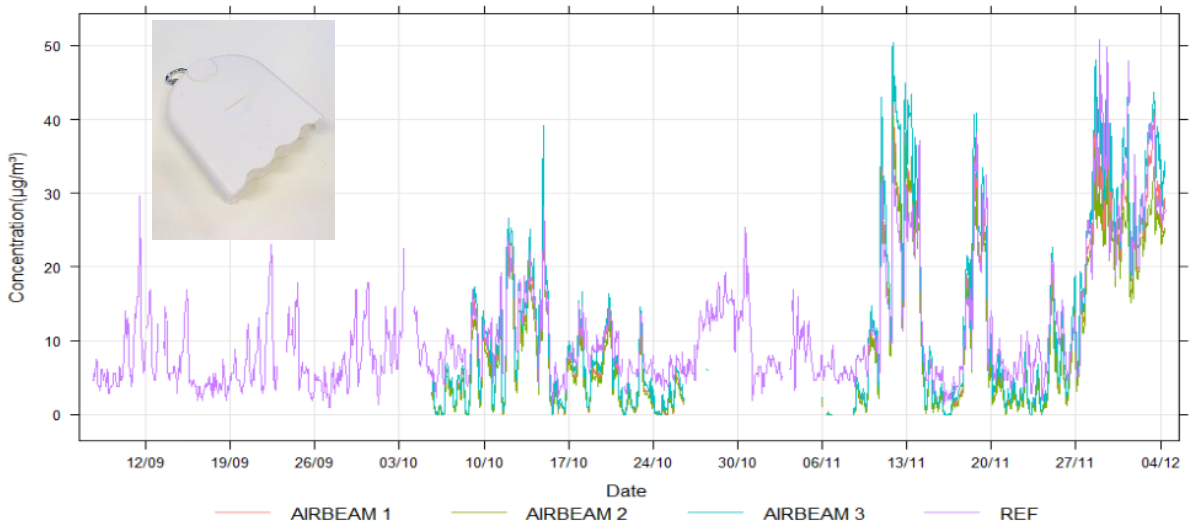


Figure 82 Hourly data coverage (upper) and resulting timeseries of $PM_{2.5}$ concentrations measured by the SODAQ AIR sensors (1-3) and the reference $PM_{2.5}$ monitor (µg/m³; lower).

When evaluating the available hourly-averaged $PM_{2.5}$ data against the reference data (Figure 83), good correlations (R^2=0.86-0.89) are obtained with mean absolute errors of 3.6-4.2 µg/m³. The MBE (-0.6- -3.5 µg/m³) indicate that the sensors slightly underestimate reference concentrations. The expanded uncertainty of the raw data already qualifies for the indicative data quality objective (<50%) for 2 out of 3 sensors (27 and 29%).

The variation between the sensors is rather high, compared to the other PM sensors, with a between sensor uncertainty (BSU) of 0.71 µg/m³.

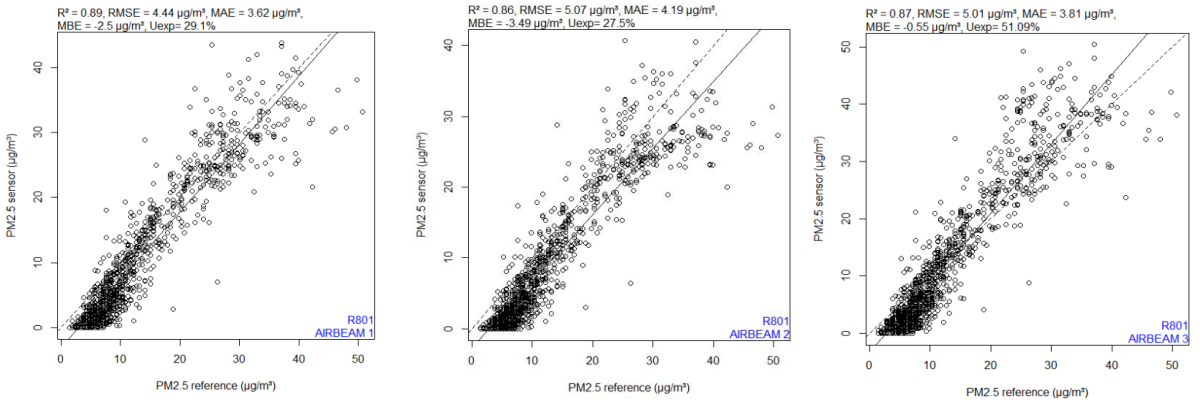


Figure 83 Regression plots of the reference $PM_{2.5}$ measurements against the concentrations measured by the AIRBEAM sensors (1-3), with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

The sensor performance decreases for PM_{10} (R^2=0.69-0.72, MAE=10.8-11.6 µg/m³) and the association decreases further (R^2=0-0.01) when focusing on the coarse fraction (PM_{coarse}=PM_{10}-$PM_{2.5}$) as can be observed from Figure 84. This confirms our earlier observations during the lab coarse tests. We can, therefore, state that the PM sensor is not able to reliably quantify coarse particles, although best PM_{coarse} performance is obtained when compared to the other PM sensors.

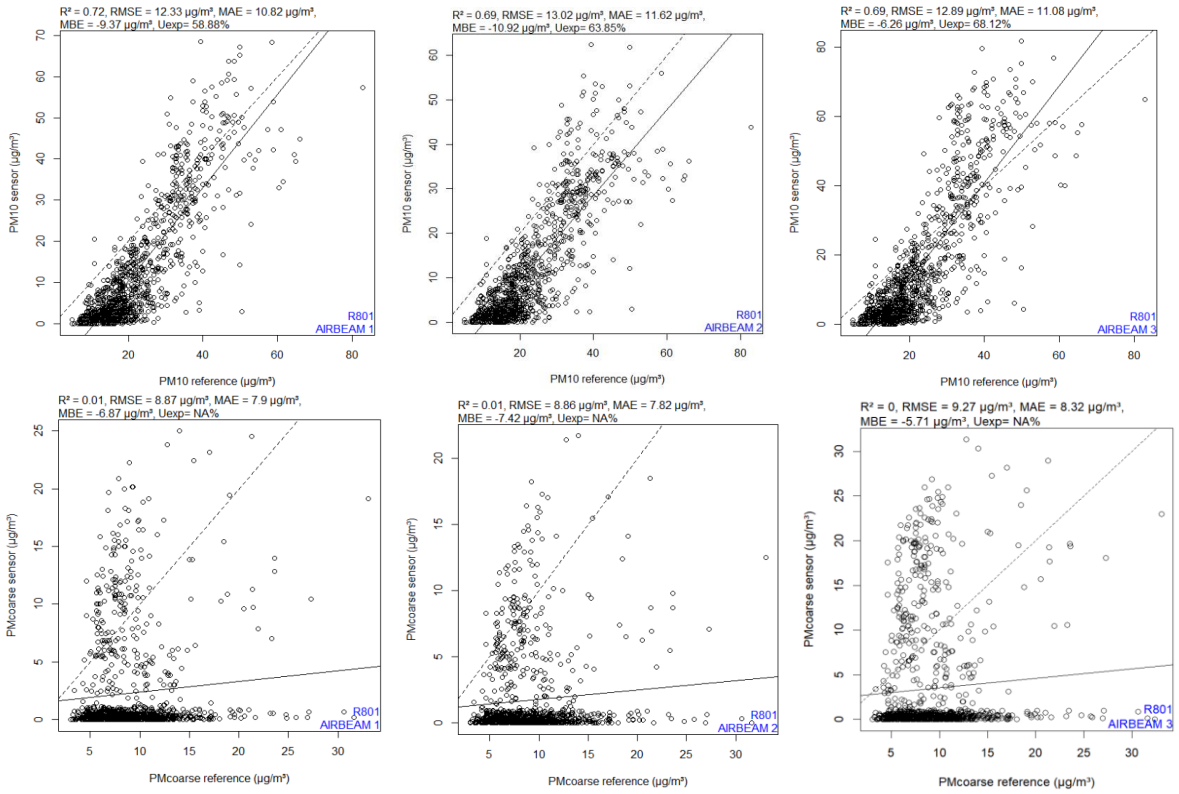


Figure 84 Regression plots of the reference PM_{10} (upper) and PM_{coarse} (lower) measurements against the concentrations measured by the SODAQ AIR sensors (SODAQ_AIR_1-3; left-middle-right), with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

- **Sensor calibration**

Lab calibration could not be applied here, as the sensors were not tested in the lab.

Figure 85 Comparability of AIRBEAM $PM_{2.5}$ concentrations against the reference using the raw (left), lab-derived calibration (middle) and field-derived calibration (right) on the test data (22/9/2022-5/12/2022).

A field calibration was conducted by deriving sensor- (AIRBEAM 1-3) and PM fraction- (PM_1, $PM_{2.5}$ and PM_{10}) specific slopes and intercepts based on the first available 2-week training period (1/10/2022-14/10/2022). The resulting calibration performance was evaluated based on the remaining test data (15/10/2022-5/12/2022) and outperfomed the raw and lab-

calibrated data with a mean accuracy (MAE) reaching 2.36 µg/m³ and U_{exp} of 22% for $PM_{2.5}$ (Figure 86).

3.2.2.2.7 GeoAIR

The GeoAIR stores all 1 second measurements on an integrated SD card and was convenient to use in the field. Data was offloaded weekly via the SD cards and resulted in a good hourly data coverage of 95-96% (Figure 86).

START	STOP	NA	%	SENSOR
07/09/2022 15:00	05/12/2022 09:00	0	96,06	GeoAir 1
07/09/2022 15:00	05/12/2022 09:00	0	96,58	GeoAir 2
07/09/2022 15:02	05/12/2022 09:00	0	95,4	GeoAir 3

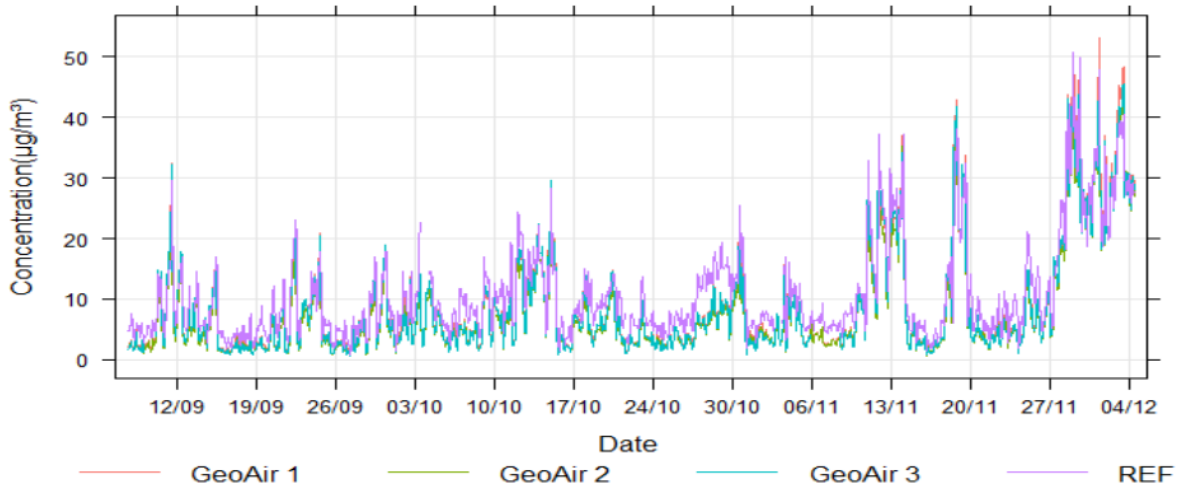


Figure 86 Hourly data coverage (upper) and resulting timeseries of $PM_{2.5}$ concentrations measured by the SODAQ AIR sensors (1-3) and the reference $PM_{2.5}$ monitor (µg/m³; lower).

When evaluating the available hourly-averaged $PM_{2.5}$ data against the reference data (Figure 87), good correlations (R^2=0.89) are obtained with mean absolute errors (MAE) of 2.9-3.2 µg/m³. The MBE (-2.1- -2.8 µg/m³) indicates that the sensors slightly underestimate reference concentrations. The expanded uncertainty of the raw data already qualifies for the indicative data quality objective (<50%) for all sensors (27-30%).

The variation between the sensors is rather high, compared to the other PM sensors, with a between sensor uncertainty (BSU) of 0.61 µg/m³.

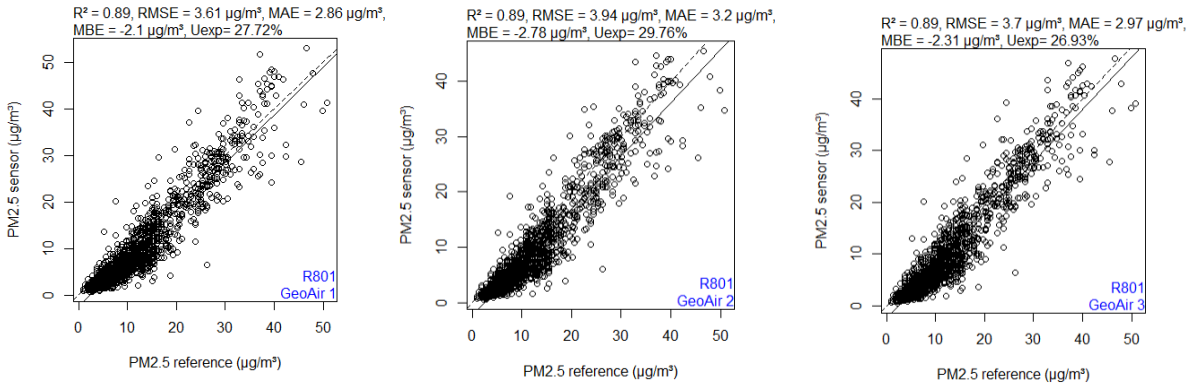


Figure 87 Regression plots of the reference $PM_{2.5}$ measurements against the concentrations measured by the GeoAir sensors (1-3), with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

The sensor performance decreases for PM_{10} (R^2=0.69-0.72, MAE=10.8-11.6 µg/m³) and the association decreases further (R^2=0) when focusing on the coarse fraction (PM_{coarse}=PM_{10}-$PM_{2.5}$) as can be observed from Figure 88. This confirms our earlier observations during the lab coarse tests. We can, therefore, state that the PM sensor is not able to reliably quantify coarse particles, although best PM_{coarse} performance is obtained when compared to the other PM sensors.

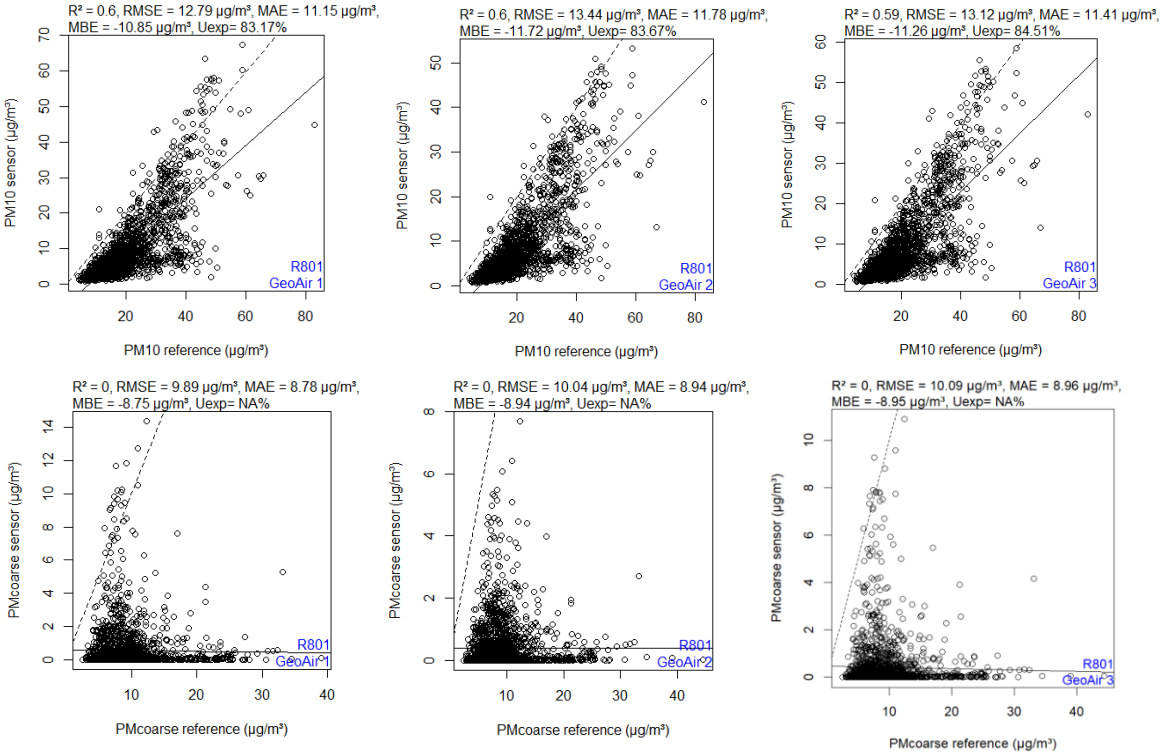


Figure 88 Regression plots of the reference PM_{10} (upper) and PM_{coarse} (lower) measurements against the concentrations measured by the GeoAIR sensors (GeoAir_1-3; left-middle-right), with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

- **Sensor calibration**

Lab calibration could not be applied here, as the sensors did not gather data during the lab tests.

Figure 89 Comparability of AIRBEAM $PM_{2.5}$ concentrations against the reference using the raw (left), lab-derived calibration (middle) and field-derived calibration (right) on the test data (22/9/2022-5/12/2022).

A field calibration was conducted by deriving sensor- (GEOAIR 1-3) and PM fraction- (PM_1, $PM_{2.5}$ and PM_{10}) specific slopes and intercepts based on a 2-week training period (7/9/2022-21/9/2022). The resulting calibration performance was evaluated based on the remaining test data (22/9/2022-5/12/2022) and outperfomed the raw and lab-calibrated data with a mean accuracy (MAE) reaching 2.07 µg/m³ and U_{exp} of 27% for $PM_{2.5}$ (Figure 91).

3.2.2.2.8 2BTech PAM

The PAM included 1 sensor system which automatically transmitted ~2 sec PM and NO_2 data via GPRS/4G (where date is downgraded to ~5 min) and has additional redundancy via an internal SD card where 2 sec data is being stored. Data was offloaded weekly from the SD cards and showed full data coverage (100%).

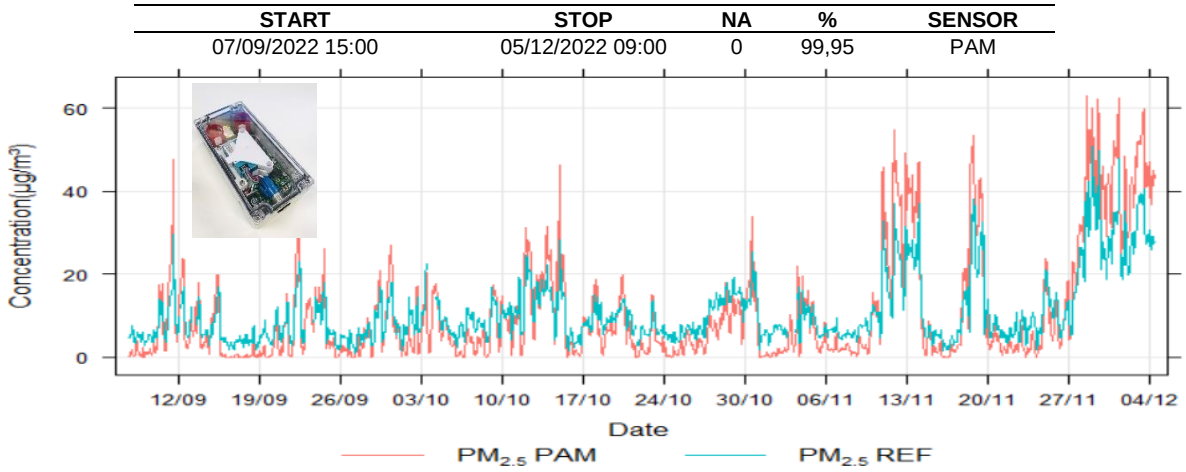


Figure 90 Hourly data coverage (%; upper) and resulting timeseries of $PM_{2.5}$ concentrations measured by the PAM sensor and the reference $PM_{2.5}$ monitor (µg/m³; lower).

- **PM performance**

When evaluating the hourly-averaged raw $PM_{2.5,}$ PM_{10} and PM_{coarse} data against the reference data (Figure 91), similar performance is observed when compared to the other PM sensors with good correlation (R^2=0.89) and MAE of 4.68 µg/m³ for PM2.5, a lower performance for

PM_{10} (R^2=0.62 and MAE=10.28µg/m³) and a lacking association for PM_{coarse} (R^2=0). We can, therefore, state that the PM sensor is not able to reliably quantify coarse particles.

The between sensor variability (BSU) could not be assessed as only one sensor was available.

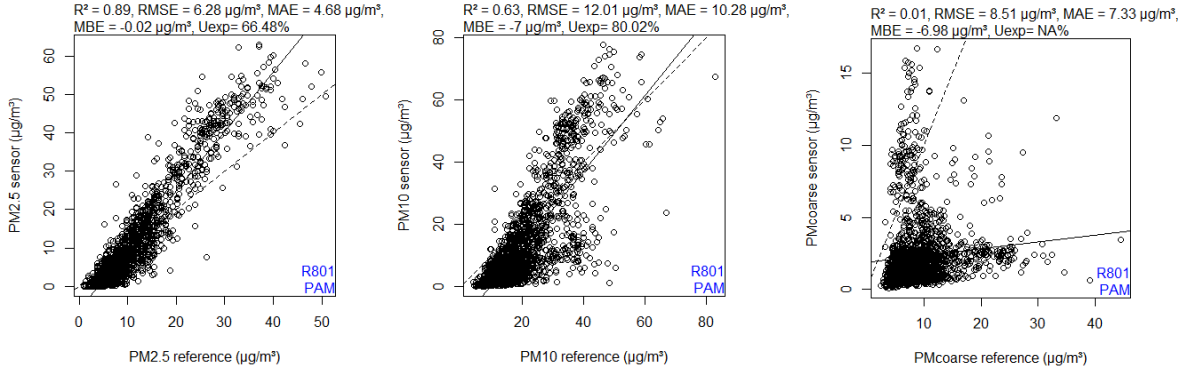


Figure 91 Comparability of the reference $PM_{2.5}$ (left), PM_{10} (middle) and PM_{coarse} (right) measurements against the concentrations measured by the PAM sensor, with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

- **NO_2 performance**

When evaluating the hourly raw performance for NO_2 (Figure 92), moderate correlation is observed (R^2=0.55) between the raw NO_2 sensor data and the reference concentrations, with high error (MAE=84 µg/m³ and Uexp=284%) and significant underestimation (MBE=-84 µg/m³) of actual concentrations.

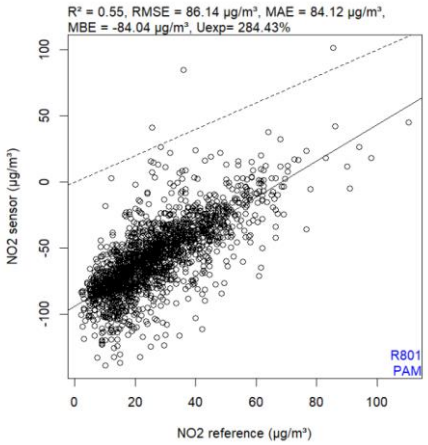


Figure 92 Regression plots of the reference NO_2 measurements against the concentrations measured by the PAM sensor, with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp})

- **Sensor calibration**

When applying the lab calibration (lab-derived slope and intercept; **Error! Reference source not found.**), the sensor accuracy worsens significantly for $PM_{2.5}$ (MAE=30 µg/m³, U_{exp}=478%), PM_{10} (MAE=79 µg/m³, U_{exp}=999%) and NO_2 (MAE=349 µg/m³ and Uexp=1225%). The lab-

derived calibration clearly does not hold in field conditions, which is not surprising as field conditions are different in terms of PM composition and environmental conditions (temperature, relative humidity).

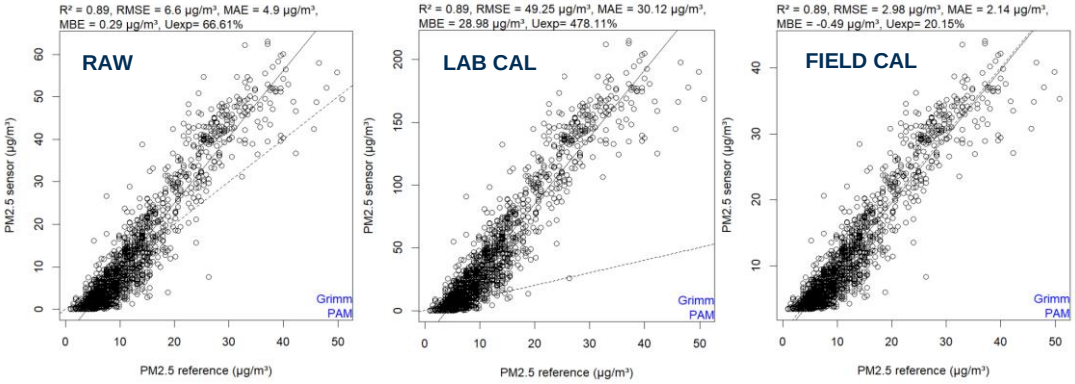
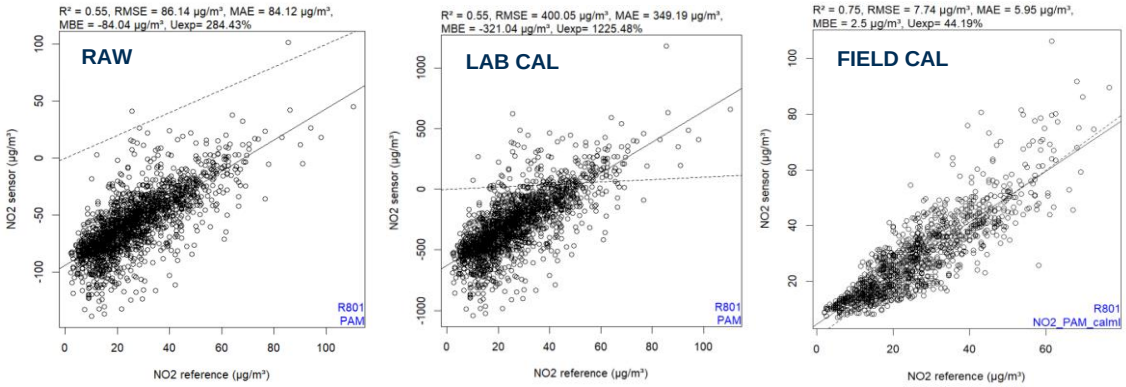


Figure 93 Comparability of PAM $PM_{2.5}$ concentrations against the reference using the raw (left), lab-derived calibration (middle) and field-derived calibration (right) on the test data (22/9/2022-5/12/2022).

A field calibration was conducted by deriving PM fraction- (PM_1, $PM_{2.5}$ and PM_{10}) specific slopes and intercepts based on a 2-week training period (7/9/2022-21/9/2022). The resulting calibration performance was evaluated based on the remaining test data (22/9/2022-5/12/2022) and outperfomed the raw and lab-calibrated data with a mean accuracy (MAE) reaching 2.14 µg/m³ and U_{exp} of 20% for $PM_{2.5}$ (Figure 93).

For NO_2, a multilinear calibration model was trained with covariates for sensor response, temperature, RH and O_3 following earlier calibration studies (2-4). Model training was based on 2 weeks of co-location data and the calibration performance was tested on the remaining 2 months of test data (Figure 94). The multilinear field calibration outperformed the raw and lab-derived calibration with R^2=0.75 and a MAE of 5.95 µg/m³. The expanded uncertainty improved significantly as well, from 284 (raw) to 44%, however, still not qualifying the indicative data quality objective (<25%).



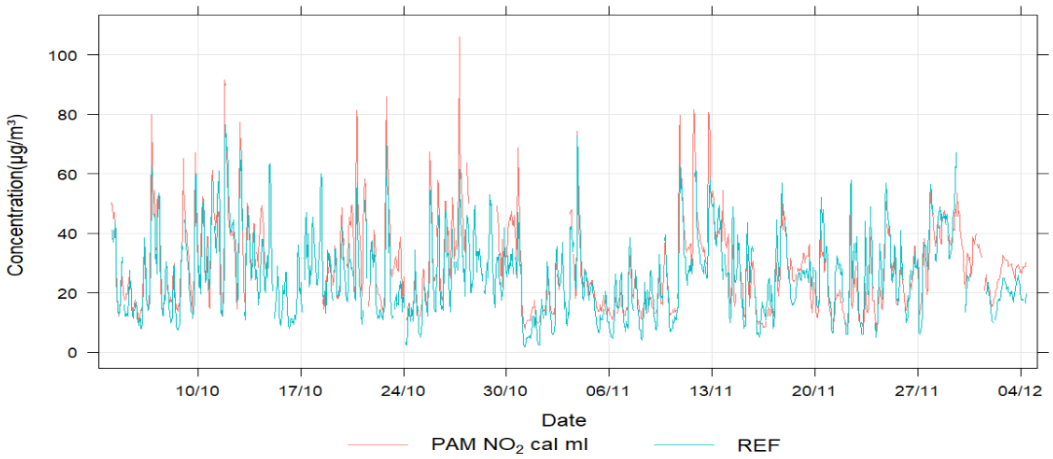


Figure 94 Upper: Comparability of PAM NO_2 concentrations against the reference when applying the raw (left), lab-derived calibration (middle) and field-derived calibration on 2 months of test data (right). Lower: Time series of field-calibrated (multilinear) NO_2 sensor test data and NO_2 reference data.

3.2.2.2.9 DST Observair

The Observair includes a filterstrip to quantify black carbon via an attenuation measurement. This filterstrip, however, saturates and needs manual replacement. The same issue was experienced for the BCmeter. Moreover, after saturation, the Observair is in error mode and does not measure NO_2 either. We, therefore, conducted a dedicated 1.5 week campaign with daily filter changes to evaluate the Observair and BCmeter. The Observair stores its 2 second data on an internal SD card and showed a fair hourly data coverage of 78% (due to a 2-day data loss).

START	STOP	NA	%	SENSOR
16/11/2022 13:00	30/11/2022 12:59	0	77,68	OBSERVAIR

Figure 95 Reported NO_2 (lab-calibrated) and BC concentrations (µg/m³) by the DST Observair sensor during the dedicated monitoring campaign.

We performed additional data cleaning to exclude extreme BC values observed during manual filter changes (Figure 95) and negative values. In the end 94% of the raw hourly-averaged data (n=214) was retained.

- **BC performance**

When evaluating the hourly-averaged and cleaned BC data against the reference data (Figure 96), good correlation (R^2=0.82) and accuracy (MAE=0.25 µg/m³) are observed. The Uexp near 1 µg/m³ amounts 40% at the hourly level. The Observair seems to slightly underestimate actual concentrations (MBE=-0.24 µg/m³).

The between sensor variability (BSU) could not be assessed as only one sensor was available.

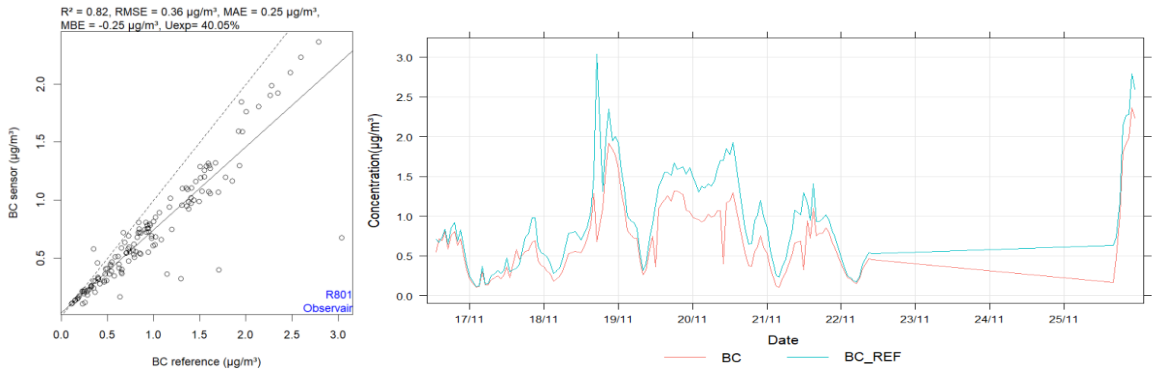


Figure 96 Left: Comparability of the cleaned (n=214) PAM BC readings (µg/m³) against the reference, with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp}). Right: Timeseries of hourly-averaged BC concentrations measured by the DST Observair (red) and air quality monitoring station (green)

- **NO_2 performance**

When evaluating the hourly raw performance for NO_2 (Figure 97), a weak negative linear association is observed (R^2=0.38) between the raw NO_2 sensor data and the reference concentrations, with low accuracy (MAE=28µg/m³) and significant underestimation (MBE=-28 µg/m³). The observed association is significantly lower than the association observed earlier in the lab test (R^2=0.98). We hypothesize that this might be due to environmental confounders (temperature, RH, O_3). DST warned for the out-of-the-box data quality in advance and typically relies on co-located reference measurements to train a sensor calibration model using machine learning techniques (not the scope of this out-of-the-box performance evaluation).

When applying the lab calibration for NO_2 (lab-derived slope and intercept; Figure 97), the sensor association (R^2) remains unchanged, but instead of underestimating, the sensor now overestimates the actual concentrations by ~29µg/m³. The lab-derived calibration, therefore, does not seem to hold in field conditions with many confounders. As we had only 214 hours of data available, we did not train and test a multilinear field calibration on this data.

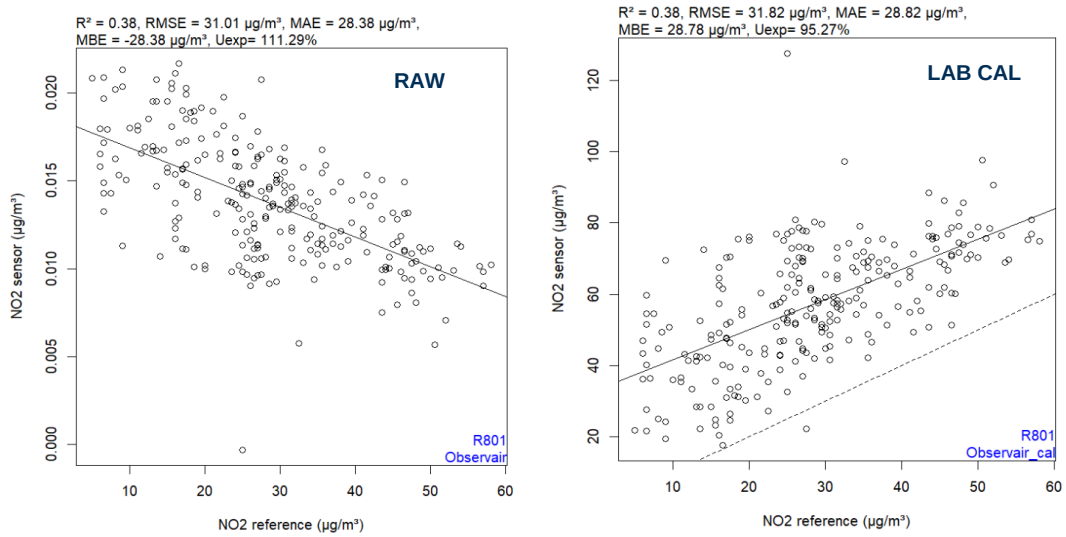


Figure 97 Regression plots of the reference NO_2 measurements against the raw (left) and lab-calibrated (right) concentrations measured by the Observair NO_2 sensor with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp}).

3.2.2.2.10 BCmeter

The BCmeter had a 5minute monitoring resolution and includes a filterstrip to quantify black carbon via an attenuation measurement (as for the Observair). It was included in a dedicated 1.5 week campaign with daily filter changes. The Observair transmits its 5 minute data via wifi (which was set up via a local modem). They can be considered as low-cost DIY devices (currently not aimed at mobile monitoring) and 1 (BC_100) of the 3 sensors broke down during the co-location campaign (data not accessible anymore via wifi). The remaining sensors reached a data coverage of 68 and 89%.

START	STOP	NA	%	SENSOR
16/11/2022 13:03	25/11/2022 23:58	73	67,84	BC_101
16/11/2022 13:02	25/11/2022 23:55	25	88,99	BC_102

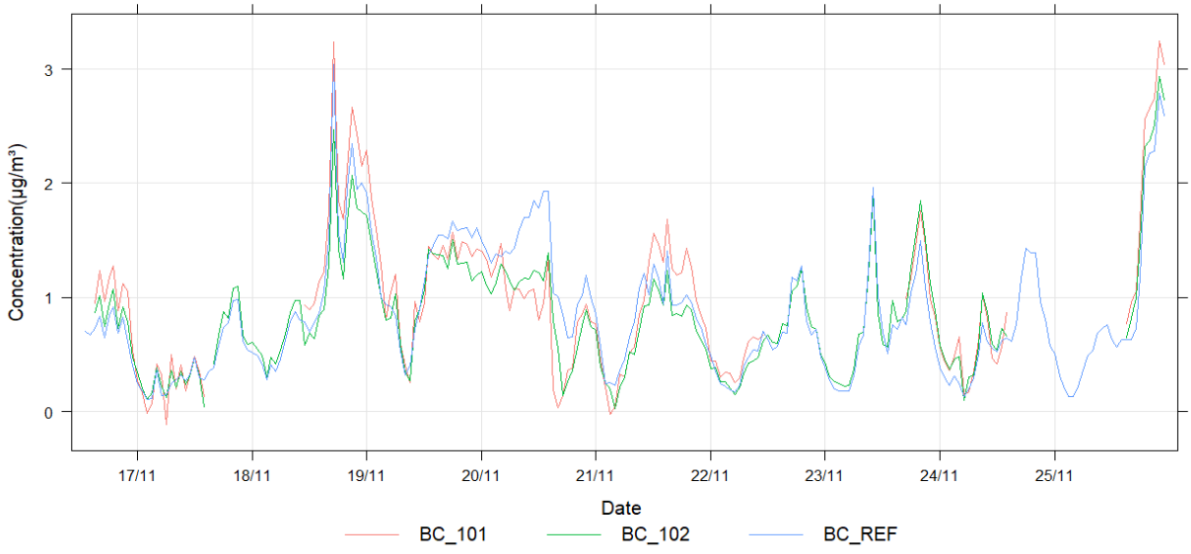


Figure 98 Hourly data coverage (%; upper) and resulting timeseries of $PM_{2.5}$ concentrations measured by the PAM sensor and the reference $PM_{2.5}$ monitor (µg/m³; lower).

Jonas Dahl (sensor supplier) advised us to look at the filter loading (bcmSens) which shouldn't exceed 35%. We evaluated the filter loading after the first day, resulting in loadings of 32 to 35%, which confirmed the need for daily filter changes.

Sensor	Beginning	End (after24h)	%
bcmeter101	11700	7990	-0.32
bcmeter102	15000	9702	-0.35
bcmeter100	9400	6400	-0.32

In addition, we noticed that the sensors experienced a time lag of 2 hours, which was corrected for.

- **BC performance**

When evaluating the hourly-averaged BC data against the reference data (Figure 99), good correlations (R^2=0.79-0.87) and high accuracies (MAE=0.15-0.24 µg/m³) are observed. The MBE (-0.05-0.03) error indicates that the observed error varies around the reference concentration meaning that the sensors not significantly under- or overestimate the actual concentrations. The U_{exp} near 1 µg/m³ amounts 41-84% at the hourly level.

The variability between the sensors was low with a between sensor variability (BSU) of 0.14 µg/m³.

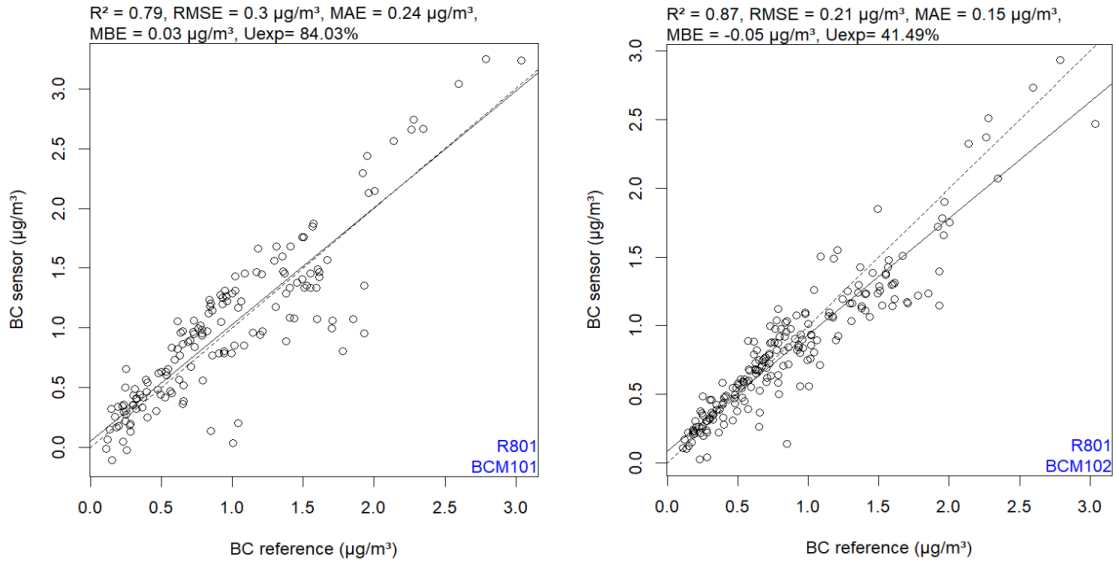


Figure 99 Comparability of the hourly-averaged BC readings (µg/m³) of BCmeter 101 (left) and 102 (right) against the reference, with associated performance metrics (R^2, RMSE, MAE, MBE, U_{exp}).

4 CONCLUSIONS

During this lab and field benchmarking campaign, we collected a lot of quantitative and qualitative performance data and practical user experiences. In Table 17, we provide an overview of the observed data quality performance metrics (hourly coverage, accuracy, R^2, MAE, BSU, stability, Uexp), averaged over the 3 sensors of each brand (when available).

Table 17 Overview of quantitative performance metrics collected during the lab and field testing and averaged per sensor system. For the field results, data quality performance after lab- and field calibration are denoted for each sensor system as "_labcal" and "_fieldcal". The number of sensors used to calculate the performance metrics are provided in brackets.

LAB RESULTS (min)		SETPOINTS				$PM_{2.5}$ LINEARITY TEST			
		Accuracy (%)			R^2	MAE	R^2	Uexp	BSU
		PM_1	$PM_{2.5}$	PM_{10}	-	µg/m³	-	%	µg/m³
PM	ATMOTUBE (3)	84	65	29	1.00	10.0	0.98	47	1.5
	OPEN SENECA (3)	83	54	22	1.00	12.6	0.99	55	1.2
	TERA (3)	18	79	47	1.00	5.2	1.00	25	1.6
	SODAQ AIR (3)	64	70	31	1.00	8.9	0.99	40	4.0
	SODAQ NO2 (3)	68	52	21	0.99	10.9	0.99	45	NA
	GeoAir (3)	NA	NA	NA	NA	NA	NA	NA	NA
	PAM (1)	63	29	13	0.99	17.3	0.96	79	NA
		Accuracy	Stability		R^2	MAE	R^2	Uexp	BSU
		%	µg/m³		-	µg/m³	-	%	µg/m³
NO_2	SODAQ NO2 (3)	-166	51		0.99	270.3	0.11	304	124.7
	PAM (1)	72	27		0.96	49.5	0.13	110	NA
	Observair (1)	<0.01	<0.01		1.00	79.0	0.98	112	NA

FIELD RESULTS (hourly)		Data coverage	MAE	R^2	Uexp	BSU
		%	µg/m³	-	%	µg/m³
$PM_{2.5}$	ATMOTUBE (1)	76	3.5	0.89	56	0.6
	ATMOTUBE_labcal (1)	76	6.5	0.89	120	0.6
	ATMOTUBE_fieldcal (1)	**76**	**2.4**	**0.89**	**27**	**0.6**
	OPEN SENECA (1)	100	3.6	0.90	34	0.3
	OPEN SENECA_labcal (1)	100	3.6	0.90	69	0.3
	OPEN SENECA_fieldcal (1)	**100**	**2.2**	**0.90**	**30**	**0.3**
	TERA (1)	17	3.7	0.90	50	0.1
	TERA_labcal (1)	17	5.6	0.90	88	0.1
	TERA_fieldcal (1)	**17**	**3.0**	**0.90**	**31**	**0.1**
	SODAQ AIR (1)	44	3.1	0.68	-	0.7
	SODAQ AIR_labcal (1)	44	3.3	0.68	-	0.7
	SODAQ AIR_fieldcal (1)	**44**	**1.9**	**0.68**	**-**	**0.7**
	SODAQ NO2 (1)	44	3.6	0.75	24	0.4
	SODAQ NO2_labcal (1)	44	6.7	0.75	91	0.4
	SODAQ NO2_fieldcal (1)	**44**	**2.0**	**0.75**	**11**	**0.4**
	AIRBEAM (1)	53	3.7	0.89	29	0.7
	AIRBEAM_labcal (1)	-	-	-	-	-
	AIRBEAM_fieldcal (1)	**53**	**2.4**	**0.89**	**29**	**0.7**
	GeoAir (1)	96	3.0	0.89	26	0.6
	GeoAir_labcal (1)	-	-	-	-	-
	GeoAir_fieldcal (1)	**96**	**2.1**	**0.89**	**27**	**0.6**
	PAM (1)	100	4.9	0.89	67	-
	PAM_labcal (1)	100	30.1	0.89	478	-
	PAM_fieldcal (1)	**100**	**2.1**	**0.89**	**20**	**-**
		Data coverage	MAE	R^2	Uexp	
		%	µg/m³	-	%	
NO_2	SODAQ NO2_raw (1)	44	277	0.62	826	
	SODAQ NO2_cal (1)	44	27.1	0.62	108	
	SODAQ NO2_mlcal (1)	**44**	**5.6**	**0.83**	**37**	
	PAM (1)	100	84.1	0.55	284	
	PAM_cal (1)	100	349.0	0.55	1225	
	PAM_calml (1)	**100**	**44.2**	**0.75**	**44**	
	Observair_raw (1)	78	28.4	0.38	111	
	Observair_cal (1)	78	28.8	0.38	95	
	Observair_mlcal (1)	-	-	-	-	

		Data coverage %	MAE µg/m³	R^2 -	Uexp
BC	Observair (1)	78	0.3	0.82	40
	Bcmeter (2)	78	0.2	0.83	63

From this quantitative performance data, we can conclude the following for the PM sensors:

- Out-of-the-box performance is already quite good and close to the indicative data quality objective (<50%). **Whether this accuracy is sufficient to quantify urban PM gradients (which are not that steep) should be investigated**.
- Best PM accuracy is observed for PM_1, followed by $PM_{2.5}$ and PM_{10}. All PM sensors do not reliably detect PM_{coarse}. TERA is the only PM sensor that seems to pick up some coarse particles (R^2=0.3), while all other sensors show R^2 of ~0.
- The observed variability between PM sensors (BSU) is typically low (<0.4 in the lab and <0.6µg/m³ in the field). Comparability between sensors is an important requirement when considering data from multiple mobile sensors.
- The accuracy of PM sensors can be further improved using a linear calibration (slope + intercept). Lab calibrations, in this case based on the lack-of-fit (response) test, do not hold in the field due to varying PM composition and environmental conditions. **We therefore stress the need for local field calibrations (in representative pollutant environments and under representative environmental conditions).** Field calibrations result in best data quality performance for all sensor systems and pollutants (Table 17).
- The TERA sensors (3 patents) shows slightly better PM performance for $PM_{2.5}$ and PM_{10}, while lower performance is observed for PM_1
- In general the assessed **PM performance and observed sensitivities (drift/RH) are very similar between the benchmarked sensor systems**, which can be explained by similar underlying hardware (Sensirion SPS30, Plantower, TERA) and lack of applied factory-alghorithms.
- All PM sensors show a sensitivity towards relative humidity (%), with a higher sensor/REF ratio at higher relative humidity.

PM sensors can be regarded as mature and are in reach of the indicative (<50%) data quality objective. Next implementation steps (Parcel B) would be:

- Experiments with citizens/employees to obtain user experience feedback
- Experiments in urban environments to evaluate whether real-life PM gradients can be assessed with the observed raw (or field calibrated) sensor accuracy.

For the considered NO_2 sensors most important take-aways include:

- NO_2 sensors **have shown not to perform out-of-the-box** (negative associations, negative data).
- SODAQ experiences significant noise and connectivity issues.
- The observed association (R^2) and stability of the Observair sensor are significantly better than the PAM and SODAQ NO2 (lower noise), but a calibration step is required.
- After conducting a linear lab calibration based on the lack-of-fit (response) test, Observair results in the best lab performance.

- **Lab-derived NO2 calibrations of the sensor response do not hold in the field.** Note that for the 'lab calibration' we only considered the lack-of-fit test and did not take into account RH, T, or ozone interference, observed in the lab.
- In the field, a multilinear calibration compensating for temperature, relative humidity and ozone confounders significantly improves the association and accuracy of the sensors.

Compared to the PM sensors, contemporary commercial NO_2 sensor systems (for mobile use) can be considered as immature and additional effort is needed in terms of noise reduction & calibration. During this benchmark study, we applied lab and field calibration approaches in order to compare the sensor systems and showed the potential of the sensor systems in terms of association and accuracy. Moreover, considering the general steeper NO_2 gradients in urban environments (compared to PM), we believe in the relevancy of NO_2 for mobile sensor applications.
Next implementations steps could include:

- Testing different calibration approaches: lab vs field co-location vs field continuous (network calibration)
- Test calibration potential of Observair: training based on machine learning
- Real-life experiments to evaluate whether NO_2 gradients in urban environments can be picked up by noisy sensor signals

As one of the sensor systems included BC (Observair) and VMM got hold of a low-cost version (BCmeter), we benchmarked the BC sensor systems as well in the field. Both BC sensor systems showed similar performances in terms of association and accuracy. The form factor of the BCmeter needs iteration and robustness in order to be suitable for mobile applications (battery, GPS, housing).

The mobile test showed **reliable performance in terms of GPS accuracy for all sensors** with mean horizontal accuracies <10m. The considered sensor systems are, therefore, suitable for mapping purposes (map matching with road network, buffer areas). In the city center of Antwerp, the GPS accuracy was clearly impacted by the urban canopy (street canyons, tunnels, landscape openness,…).

In addition to the quantitative metrics, we gained practical user experiences when implementing the sensors systems in the lab and the field. We listed the strengths and weaknesses of specific sensor systems and/or properties:

- New firmware needed to connect to Belgian Proximus network (ESP32 Developer Board in Arduino IDE) → Airbeam
- New drivers to create COM ports for the devices (Arduino IDE) → Airbeam
- Observair (NO2) shuts down when BC filter saturates → Observair → not for continuous monitoring
- Clock sync issue → reset required → Observair
- Regional availability app (FR) → TERA

\-

- App versions (Android/iOS)
- Time resolution: Mobile vs fixed (SODAQ) – SD vs cloud portal (PAM)
- Slow online portal for data offload (SODAQ)
- TERA needs continuous app connection → automatically shut down by iOS/Android → Not for continuous monitoring
- App connectivity issues Airbeam → firmware update to connect to Proximus
- BC meter/Observair: daily filter changes? → not feasible during field test → dedicated 1 week campaign: **16/11 - 25/11**

\+

- Good coverage/reliability: SD card storage/redundancy → Open Seneca, PAM, GeoAir, Airbeam, Observair
- Redundancy: local storage + GPRS/app connectivity → Open Seneca, ATMO, PAM
- Easy to use (interface, on/off, logging, SD): Open Seneca, GeoAir, PAM > TERA >ATMO/SODAQ /Observair/Airbeam

Main concerns include **smartphone application** considerations (availability (region/country), iOS/Android, clock sync issues, continuous connectivity) and **redundancy of data storage** by means of a SD card (no data was lost on SD cards) and **clarity** about time resolution, potential data compression in cloud dashboards. For the purpose of this benchmark study, **we valued sensor systems with internal GPS sensors, redundant SD storage and autonomous operation (no app connectivity needed).** User requirements might be different for citizens, employees or other sensor users.

REFERENCES

1. J. Peters, M. Van Poppel (2020) LITERATUURSTUDIE, MARKTONDERZOEK EN MULTICRITERIA-ANALYSE BETREFFENDE LUCHTKWALITEITSSENSOREN EN SENSORBOXEN. (VITO).
2. J. Hofman *et al.*, Distant calibration of low-cost PM and NO2 sensors; evidence from multiple sensor testbeds. *Atmos Pollut Res* **13**, 101246 (2022).
3. B. Mijling, Q. Jiang, D. de Jonge, S. Bocconi, Field calibration of electrochemical NO2 sensors in a citizen science context. *Atmos Meas Tech* **11**, 1297-1312 (2018).
4. F. Karagulian *et al.* (2020) Calibration of AirSensEUR boxes during a field study in the Netherlands. (European Commission, Ispra).

vision on technology
for a better world