Long-term Evaluation of Commercial Air Quality Sensors: An Overview from the QUANT Study

Sebastian Diez[1,2], Stuart Lacy[2], Hugh Coe[3], Josefina Urquiza[4,5], Max Priestman[6], Michael Flynn[3], Nicholas Marsden[3], Nicholas A. Martin[7], Stefan Gillott[6], Thomas Bannan[3], Pete Edwards[2]

[1]Centro de Investigación en Tecnologías para la Sociedad, Universidad del Desarrollo, Santiago, Chile, CP 7550000

[2]Wolfson Atmospheric Chemistry Laboratories, University of York, York, YO10 5DD, UK

[3]Department of Earth and Environmental Science, Centre for Atmospheric Science, School of Natural Sciences, The University of Manchester, Manchester, M13 9PL, UK

[4]Grupo de Estudios de la Atmósfera y el Ambiente (GEAA), Universidad Tecnológica Nacional, Facultad Regional Mendoza (UTN-FRM), Cnel. Rodriguez 273, Mendoza, 5501, Argentina

[5]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) Argentina

[6]MRC Centre for Environment and Health, Environmental Research Group, Imperial College, London, W12 0BZ, UK

[7]National Physical Laboratory, Teddington TW11 0LW, UK

Correspondence: Sebastian Diez (sebastian.diez@udd.cl); Pete Edwards (pete.edwards@york.ac.uk)

Abstract. In times of growing concern about the impacts of air pollution across the globe, lower-cost sensor technology is giving the first steps in helping to enhance our understanding and ability to manage air quality issues, particularly in regions without established monitoring networks. While the benefits of greater spatial coverage and real-time measurements that these systems offer are evident, challenges still need to be addressed regarding sensor reliability and data quality. Given the limitations imposed by intellectual property, commercial implementations are often "black boxes", which represents an extra challenge as it limits end-users' understanding of the data production process. In this paper we present an overview of the QUANT (Quantification of Utility of Atmospheric Network Technologies) study, a comprehensive 3-year assessment across a range of urban environments in the United Kingdom, evaluating 43 sensor devices, including 119 gas sensors and 118 particulate matter sensors, from multiple companies. QUANT stands out as one of the most comprehensive studies of commercial air quality sensor systems carried out to date, encompassing a wide variety of companies in a single evaluation and including two generations of sensor technologies. Integrated into an extensive data set open to the public, it was designed to provide a long-term evaluation of the precision, accuracy, and stability of commercially available sensor systems. To attain a nuanced understanding of sensor performance, we have complemented commonly used single-value metrics (e.g., Coefficient of Determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE)) with visual tools. These include Regression plots, Relative Expanded Uncertainty (REU) plots, and Target plots, enhancing our analysis beyond traditional metrics. This overview discusses the assessment methodology, and key findings showcasing the significance of the study. While more comprehensive analyses are reserved for future detailed publications, the results

shown here highlight the significant variation between systems, the incidence of corrections made by manufacturers, the effects of relocation to different environments, and the long-term behaviour of the systems. Additionally, the importance of accounting for uncertainties associated with reference instruments in sensor evaluations is emphasised. Practical considerations in the application of these sensors in real-world scenarios are also discussed, and potential solutions to end-user data challenges are presented. Offering key information about the sensor systems' capabilities, the QUANT study will serve as a valuable resource for those seeking to implement commercial solutions as complementary tools to tackle air pollution.

Keywords: air pollution, commercial sensor systems, QUANT, long-term evaluation.

1. Introduction

Emerging lower-cost sensor systems[1] offer a promising alternative to the more expensive and complex monitoring equipment traditionally used for measuring air pollutants such as $PM_{2.5}$, NO_2, and O_3 (Okure et al., 2022). These innovative devices hold the potential to expand spatial coverage (Malings et al., 2020) and deliver real-time air pollution measurements (Tanzer-Gruener et al., 2020). However, concerns regarding the variable quality of the data they provide still hinder their acceptance as reliable measurement technologies (Karagulian et al., 2019; Zamora et al., 2020).

Sensors[2] face key challenges such as cross-sensitivities (Bittner et al., 2022; Cross et al., 2017; Levy Zamora et al., 2022; Pang et al., 2018), internal consistency (Feenstra et al., 2019; Ripoll et al., 2019), signal drift (A. Miech et al., 2023; Li et al., 2021; Sayahi et al., 2019), long term performance (Bulot et al., 2019; Liu et al., 2020) and data coverage (Brown & Martin, 2023; Duvall et al., 2021; Feinberg et al., 2018). Additionally, environmental factors such as temperature and humidity (Bittner et al., 2022; Farquhar et al., 2021;~~, and humidity~~ Crilley et al., 2018; Williams, 2020) can significantly influence sensor signals.

In recent years, manufacturers of both sensing elements (Han et al., 2021; Nazemi et al., 2019) and sensor systems have made significant technological advances (Chojer et al., 2020). For example, there are now commercial and non-commercial systems equipped with multiple detectors to measure distinct pollutants (Buehler et al., 2021; Hagan et al., 2019; Pang et al., 2021) helping to mitigate the effects of cross-interferences. Additionally, enhancements in electrochemical OEMs have been demonstrated in terms of their specificity (Baron & Saffell, 2017; Ouyang, 2020).

However, the complex nature of their responses, coupled with their dependence on local conditions means sensor performance can be inconsistent (Bi et al., 2020). This complicates the comparison of results or anticipating sensor future performance across different studies. Moreover, assessments of sensor performance found in the academic

---

[1] The term "sensor systems" refers to sensors housed within a protective case, which includes a sampling and power system, electronic hardware and software for data acquisition, analog-to-digital conversion, data processing and their transfer (Karagulian et al., 2019). Unless specified otherwise, the term "sensor" will be used as a synonym of "sensor systems". Other alternative names for "sensor systems" used here are "sensor devices" (or "devices"), "sensor units" (or "units").

[2] In a narrower sense, "sensor" typically denotes the specific component within a sensor system that detects and responds to environmental inputs, producing a corresponding output signal. To distinguish this from the broader use of "sensor" as equivalent to "sensor system" in our text, we will utilise alternative terms such as "detector", "sensing element", or "OEM" (original equipment manufacturer) when referring specifically to this component, thereby preventing confusion.

literature often rely on a range of protocols (e.g., CEN (2021) and Duvall et al. (2021)) and data quality metrics (e.g., Spinelle et al. (2017) and Zimmerman et al. (2018)), with many studies limited to a single-site co-location and/or short-term evaluations that do not fully account for broader environmental variations (Karagulian et al., 2019).

The calibration of any instrument used to measure atmospheric composition is fundamental to guarantee their accuracy (Alam et al., 2020; Long et al., 2021; Wu et al., 2022). Using out-of-the-box sensor data without fit-for-purpose calibration can produce misleading results (Liang & Daniels, 2022). An effective calibration not only involves identifying but also compensating for estimated ~~and correcting~~ systematic effects ~~errors~~ in the sensor readings, a process defined as a correction (for a detailed definition and differentiation of calibration and correction see JCGM, 2012). For standard air pollution measurement techniques, calibration is often performed in a controlled laboratory environment (Liang, 2021)~~, or by sampling gas from a certified standard cylinder in the field. For PM, particles of known density and size are used, controlling the airflow conditions~~. For example, for gases, a known concentration is sampled from a certified standard. Similarly, for PM, particles of known density and size are generated. Both gases and PM calibration are conducted under controlled airflow conditions

Yet, the aforementioned challenges with lower-cost sensor-based devices suggest that such calibrations may not always accurately reflect real-world conditions (Giordano et al., 2021). A frequent approach involves co-locating sensors alongside regulatory instruments in their intended deployment areas and/or conditions and using data-driven methods to match the reference data (Liang & Daniels, 2022). Numerous studies have investigated the effectiveness of calibration methods for sensors e.g. (Bigi et al., 2018; Bittner et al., 2022; Malings et al., 2020; Spinelle et al., 2017; Zimmerman et al., 2018), including selecting appropriate reference instruments (Kelly et al., 2017), the need for regular calibration to maintain accuracy (Gamboa et al., 2023), the necessity of rigorous calibration protocols to ensure consistency (Kang et al., 2022), and transferability (Nowack et al., 2021) of results. Ultimately, the reliability and associated uncertainty of any applied calibration will influence the final sensor data quality.

For end-users to make informed decisions on the applicability of air pollution sensors, a realistic understanding of the expected performance in their chosen application is necessary (Rai et al., 2017). Despite this, there has been relatively little progress in clarifying the performance of sensors for air pollution measurements outside of the academic arena. This is largely due to the significant variability in both the number of sensors and the variety of applications tested, compounded by the proliferation of commercially available sensors/sensor systems with different configurations. ~~as well as the availability of highly accurate measurement instrumentation and/or regulatory networks to those outside of the atmospheric measurement academic field.~~ Furthermore, the access to highly accurate measurement instrumentation and/or regulatory networks remains limited for those outside of the atmospheric measurement academic field (e.g. Lewis and Edwards (2016) and Popoola et al. (2018)). From a UK clean air perspective, this ambiguity represents a major problem. The lack of a consistent message undermines the exploitation of these devices' unique strengths, notably their capability to form spatially dense networks with rapid time resolution. Consequently, there is potential for a mismatch in users' expectations of what sensor systems can deliver and their actual operating characteristics, eroding trust and reliability.

In this work, as part of the UK Clean Air program funded QUANT project, we deployed a variety of sensor technologies (43 commercial devices, 119 gas and 118 PM measurements) at 3 representative UK urban sites — Manchester, London and York— alongside extensive reference measurements, to generate the data for an comprehensive ~~extensive~~ in-depth performance assessment. This project aims to not only evaluate the performance of sensor devices in a UK urban climatological context but also provide critical information for the successful application of these technologies in various environmental settings. To our knowledge, QUANT is the most extensive and longest-running evaluation of commercial sensor systems globally to date. Furthermore, we tested multiple manufacturers' data products, such as out-of-the-box data versus locally calibrated data, for a significant number of these sensors to understand the implications of local calibration. This comprehensive approach offers unprecedented insights into the operational capabilities and limitations of these sensors in real-world conditions. Significantly, some of the insights gathered during QUANT have contributed to the development of the Publicly Available Specification (PAS 4023, 2023), which provides guidelines for the selection, deployment, maintenance, and quality assurance of air quality sensor systems. While this manuscript serves as an initial overview, detailed analyses of the measured pollutants and study phases, offering a more comprehensive perspective on sensor performance, are planned for future publications.

In the following sections, we delve into the methodology and provide an overview of the QUANT dataset, as well as a discussion of some of the key findings and potential considerations for end-users.

2. QUANT study design

To capture the variability of UK urban environments, identical units were installed at three carefully selected field sites. Two of these sites are highly instrumented urban background measurement supersites: the London Air Quality Supersite (LAQS; for more details, refer here: https://uk-air.defra.gov.uk/networks/site-info?site_id=HP1) and the Manchester Air Quality Supersite (MAQS; for more details, see: http://www.cas.manchester.ac.uk/restools/firs/), located in densely populated urban areas with unique air quality challenges. The third site is a roadside monitoring site in York, which is part of the Automatic Urban and Rural Network (AURN; click here for more details: https://uk-air.defra.gov.uk/networks/site-info?uka_id=UKA00524&search=View+Site+Information&action=site&provider=archive), representing a urban environment more influenced by traffic. This selection strategy ensures that the QUANT study's findings reflect the dynamics of urban air quality across different UK settings, while providing comprehensive reference measurements. Further details about each site can be found in Section S1 in the Supp.

2.1 Main study

The Main QUANT assessment study aimed to perform a transparent long-term (19 Dec 2019 - 31 Oct 2022) evaluation of commercially available sensor technologies for outdoor air pollution monitoring in UK urban environments. Four units ~~duplicates~~ of five different commercial sensor devices (Table 1) were purchased in Sept 2019 for inclusion in the study, with the selection criteria being: market penetration and/or previous performance reported in the literature, ability to measure pollutants of interest (e.g. NO_2, NO, O_3, and $PM_{2.5}$), and capacity to run continuously reporting high time resolution data (1-15 min data) ideally in near real-time (i.e., available within minutes of measurement) with data accessible via an API.

Table 1. Main QUANT devices description. The 20 units, all commercially available and ready for use as-is, offered 56 gas and 56 PM measurements in total. For a detailed description of the devices see Section S3~~1~~ in the Supp.

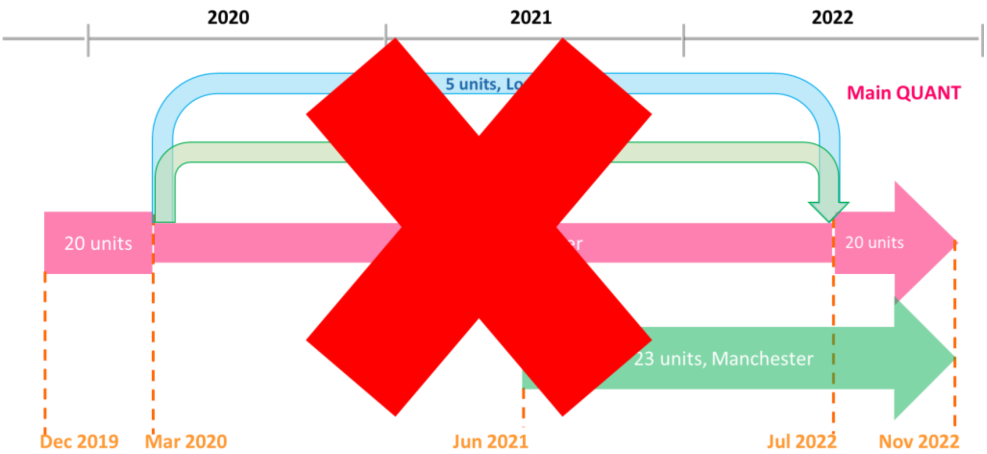
Product*	Measurements	Cost (£)**

(# units)	Company[3]	NO	NO_2	O_3	CO	CO_2	PM_1	$PM_{2.5}$	PM_{10}	
AQY (4)	Aeroqual	-	✓	✓	-	-	-	✓	✓	~4.7K
AQM (4)	AQMesh	✓	✓	✓	-	✓	✓	✓	✓	~8.6K
Ari (4)	QuantAQ	✓	✓	✓	✓	✓	✓	✓	✓	~8.6K
PA (4)	PurpleAir	-	-	-	-	-	✓	✓	✓	~0.3K
Zep (4)	Earthsense	✓	✓	✓	-	-	✓	✓	✓	~7K

*AQY: Aeroqual; AQM: AQMesh; Ari: Arisense; PA: PurpleAir; Zep: Zephyr. **Cost (Sep 2019) per unit including UK taxes and associated contractual costs (i.e., communication, data access, sensor replacement, etc.).

~~To capture the variability of UK urban environments, identical units were installed at three carefully selected field sites. Two of these sites are highly instrumented urban background measurement supersites: the London Air Quality Supersite (LAQS) and the Manchester Air Quality Supersite (MAQS), located in densely populated urban areas with unique air quality challenges. The third site is a roadside monitoring site in York, which is part of the Automatic Urban and Rural Network (AURN, https://uk-air.defra.gov.uk/data/), representing a urban environment more influenced by traffic. This selection strategy ensures that the QUANT study's findings reflect the dynamics of urban air quality across different UK settings, while providing comprehensive reference measurements. Further details about each site can be found in Section S3 in the Supp., and the available reference instrumentation in Section S4.~~

Initially, all the sensors were deployed in Manchester for approximately 3 months (mid-Dec 2019 to mid-Mar 2020) before being split up amongst the three sites (Fig. 1). At least one unit per brand was re-deployed to the other two sites (mid-March 2020 to early-July 2022) leaving two devices per company in Manchester to assess inter-device consistency. In the final 4 months of the study, all the sensor systems were relocated back to Manchester (early July 2022 to the end of October 2022).



[3] Throughout this article, the terms "manufacturers" and "company" are used interchangeably to refer to entities that produce, and/or sell sensor systems or devices. This usage reflects the industry practice of referring to businesses involved in the production and distribution of technology products without distinguishing between their roles in manufacturing or sales.

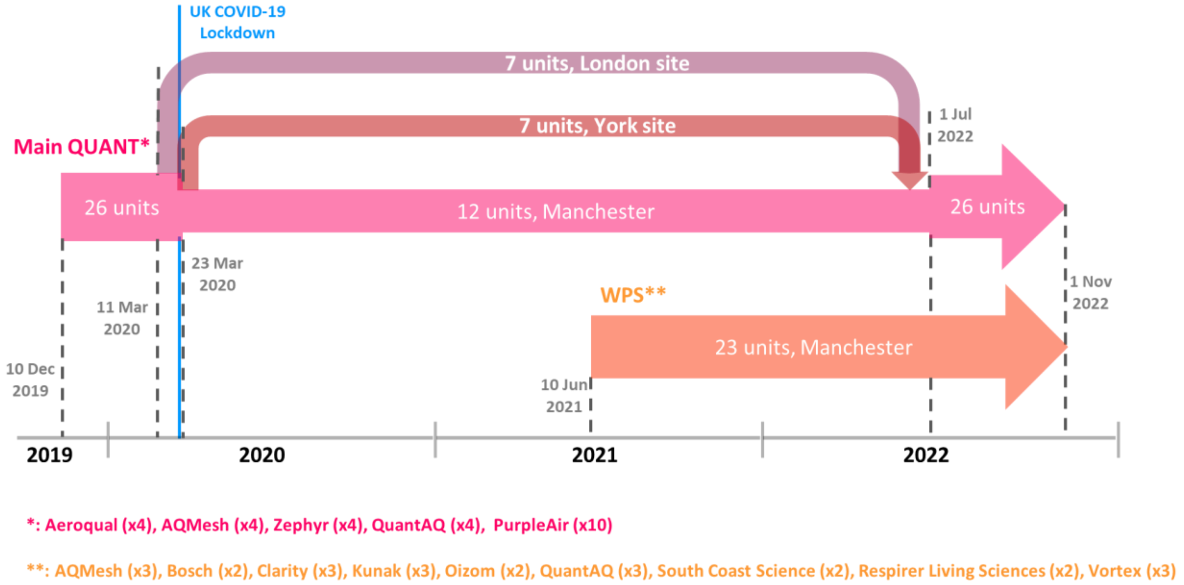


Figure 1. Main QUANT ~~Quant~~ and Wider Participation Study (WPS) timeline.

2.2 Wider Participation Study

The Wider Participation Study (WPS) was a no-cost complementary extension of the QUANT assessment, specifically designed to foster innovation within the air pollution sensors domain. This segment of the study took place entirely at the MAQS from 10th June 2021 to 31st October 2022 (Fig. 1). It included a wider array of commercial platforms (9 different sensor systems brands), and offered manufacturers the opportunity to engage in a free-of-charge impartial evaluation process. Although participation criteria matched those of the Main QUANT study, a key distinction lay in the voluntary nature of participation: manufacturers ~~vendors~~ were invited to contribute multiple sensor devices throughout the WPS study (see Table 2). Participants were able to demonstrate their systems' performance against collocated high-resolution (1-minute) reference data at a state-of-the-art measurement site such as the Manchester supersite.

Table 2. The 23 WPS devices deployed at the Manchester supersite, all commercially available and ready for use as-is, provided 63 gases and 62 PM measurements in total. For a detailed description of the devices see the Section S4~~3~~ in the Supp.

Product* (# units)	Company	Measurements							
		NO	NO_2	O_3	CO	CO_2	PM_1	$PM_{2.5}$	PM_{10}
Mod (3)	QuantAQ	-	-	-	-	-	✓	✓	✓
AQM (3)	AQMesh	✓	✓	✓	✓	✓	✓	✓	✓
Atm (2)	RLS**	-	-	-	-	-	✓	✓	✓
IMB (2)	Bosch	-	✓	✓	-	-	-	✓	✓
Poll (2)	Oizom	✓	✓	✓	✓	✓	-	✓	✓
AP (3)	Kunak	✓	✓	✓	✓	✓	✓	✓	✓
SA (3)	Vortex IoT	-	✓	✓	-	-	-	✓	✓
NS (3)	Clarity	-	✓	-	-	-	✓	✓	✓

Prax (2)	SCS***	✓	✓	✓	✓	✓	✓	✓	✓

*Mod: Modulair; AQM: AQMesh; Atm: Atmos, Poll: Polludrone ~~Polludrone: Poll~~; AP: Kunak Air Pro; SA: Silax Air, NS: Node-S, Prax: Praxis. **RLS: Respirer Living Sciences. ***SCS: South Coast Science.

2.3 Sensor deployment and data collection~~, co-located reference data and data products~~

All sensor devices were installed at the measurement sites as per manufacturer recommendations, adhering strictly to manufacturers' guidelines for electrical setup, mounting, cleaning, and maintenance guaranteed proper installation. Since all deployed systems were designed for outdoor use, no additional protective measures were necessary. Each of the systems were mounted on poles acquired specifically for the project or on rails at the co-location sites, without the need for special protections. Following the manufacturer's suggestions, sensors were positioned within 3 metres of the reference instruments' inlets. Custom electrical setups were developed for each sensor type, incorporating local energy sources and weather-resistant safety features, alongside security measures to deter vandalism and ensure uninterrupted operation. Routine maintenance was conducted monthly, although the COVID-19 pandemic necessitated longer intervals between visits. Despite these obstacles, efforts to maintain sensor security and functionality continued unabated, employing both physical safeguards and remote monitoring to preserve data integrity.

In addition to the device supplier's own cloud storage (accessed on-demand via each supplier’s web portals), an automated daily scraping of each company’s API was performed to save data onto a secure server at the University of York to ensure data integrity. ~~PurpleAir units were exempt from this due to a lack of mobile data connection and poor internet signal at the sites; instead, readings were locally collected and manually uploaded.~~ Unlike other brands that utilise mobile data connections, PurpleAir sensors rely on WiFi for data transmission. Due to poor internet signal at the sites, we locally collected and manually uploaded readings for these units. Minor pre-processing was applied at this stage, including temporal harmonisation to ensure that all measurements had a minimum sampling period of 1-minute, ensuring consistency in measurement units and labels, and coercing into the same format to allow for full compatibility across sensor units. No additional modifications to the original measurements were applied; missing values were kept as missing and no additional flags were created based on the measurements beyond those provided by the manufacturers. ~~No outlier checks or data modifications were applied at this stage.~~ For an overview of the sensor measurands and their corresponding data time resolutions as provided by the companies participating in the Main QUANT study and the WPS, please see Seccion S3 and S4 (Table S4 and S5) respectively.

2.4 Data products and co-located reference data

In addition to providing an independent assessment of sensor performance, QUANT also aimed to contribute to device manufacturers to help advance the field of air pollution sensors. During QUANT, device calibrations were performed solely at the discretion of the manufacturers without any intervention from our team, thus limiting the involvement of manufacturers in the provision of standard sensor outputs and unit maintenance as would be required by any standard customer. This approach enabled manufacturers to independently assess and benchmark their sensors' performance, using provided reference data to potentially develop calibrated data products. It's noteworthy that not all manufacturers chose to utilise these data for corrections or enhancements. However, those who did were expected to create and submit calibrated data products, subsequently named as "out-of-box" (initial data product), “cal1” (first calibrated product), and “cal2” (second calibrated product). This differentiation highlighted the varying degrees of engagement and application of the reference data by different manufacturers. Figures S2 and S3 (section S3 and S4 respectively) show a time-line of the different data products.

To this end, three separate 1-month periods of reference data, spaced every 6 months, were shared with each supplier, provisional data soon after each period, and ratified data when available. ~~For an overview of reference instrumentation at each site refer Table S1, and for details on the quality assurance procedures applied to the reference instruments see Table S2.~~ All reference data were embargoed until it was released to all manufacturers simultaneously to ensure consistency across manufacturers. For an overview of reference and equivalent-to-reference instrumentation, as defined in the European Union Air Quality Directive 2008/50/EC (hereafter referred to as EU AQ Directive), at each site, please refer to Section S2 (Table S1). For details on the quality assurance procedures applied to the reference instruments, see Table S2. To see the dates and periods of the shared reference data refer to Table S3. ~~Access to colocated reference data allowed the companies to assess sensors' performance and, if they chose, to generate and provide additional calibrated data products. These products are distinct data versions provided by manufacturers throughout QUANT, before and/or after sharing reference data — for instance, "out-of box", "cal1", "cal2", etc. Figures S1 and S2 show a time-line of the different data products. To see the dates and periods of the shared reference data refer to Table S3. All reference data was embargoed until it was released to all manufacturers simultaneously to ensure consistency across manufacturers. Not every manufacturer opted to use this data to apply corrections or improve calibrations, but if they chose to do so, the updated measurements were treated as a separate data product. Device calibrations were performed solely at the discretion of the manufacturers without any intervention from our team, thus limiting the involvement of vendors/manufacturers in the provision of standard sensor outputs and unit maintenance as would be required by any standard customer.~~

3. Results and discussion

A key challenge in sensor performance evaluation is the high spatial and temporal variability errors that impact the accuracy of their readings, making the application of laboratory corrections more challenging. ~~Furthermore, the overreliance on global performance metrics, such as R^2 (i.e., the Coefficient of Determination), RMSE (i.e., the Root Mean Squared Error), and MAE (i.e., the Mean Absolute Error) is an important issue when assessing sensors. While these metrics provide a general understanding of sensor performance, they can be limiting or even misleading, restricting a comprehensive understanding of the error structure and the measurement information content (Diez et al., 2022).~~ Furthermore, the overreliance on global performance metrics is a significant concern in sensor assessment. The Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are among the most popular single-value metrics for evaluating sensor performance, alongside others (e.g., the bias, the slope and intercept of the regression fit). However, while single-value metrics offer an overview of performance, they can be limiting or misleading. They condense vast amounts of data into a single value, simplifying complexity at the expense of a nuanced understanding of error structures and information content (Diez et al., 2022), potentially overlooking critical aspects of sensor performance (Chai & Draxler, 2014). Visualisation tools (such as Regression plots, Target plots, and Relative Expanded Uncertainty plots) complement these metrics, allowing end users to identify relevant features, which could be beyond the scope of global metrics. For additional details on the metrics utilised in this study, including some of their limitations and advantages refer to section "S5. Performance Metrics". This section also provides a summary of current guidelines and standardisation initiatives, which may offer a foundation for end-users to select appropriate metrics for their own analyses (refer to table S6). For further discussion on metrics and visualisation tools for performance evaluation, readers are directed to Diez et al. (2022).

In response to these challenges, the QUANT assessment represents the most extensive independent appraisal of air pollution sensors in UK urban atmospheres. As the results presented here illustrate, QUANT is dedicated to examining sensor performance through multiple complementary ~~perspectives and~~ metrics and visualisation tools, aiming to integrate these to accurately reflect the complexity of this dataset. This methodology promotes a nuanced understanding of sensor performance, extending beyond the limitations of conventional global single-value metrics.

Furthermore, by providing open access to the dataset, we encourage stakeholders to explore and utilise the data according to their unique needs and contexts, as detailed in the "Data Availability" section. In addition, we have developed a publicly accessible analysis platform (https://shiny.york.ac.uk/quant/), designed for straightforward offline analysis of the QUANT dataset. This platform enables users to interactively visualise the data through various representations, such as time series, regression plots, and Bland-Altman plots. It also offers statistical parameters (including regression equation, R^2, and RMSE) for analysing different pollutants, selecting specific sensors or manufacturers, and comparing across various co-location timeframes.

The following sections aim to provide an overview of the data and provide initial findings, with a focus on those that are most relevant to end-users of these technologies. The majority of examples presented here focus on $PM_{2.5}$ and NO_2 measurements, due to both a larger dataset available for these pollutants and their critical role in addressing the exceedances that predominantly impact UK air quality. All metrics and plots presented here are based on 1-hour averaged data. Unless otherwise specified, a data inclusion criterion of 75% was uniformly applied across our analyses to ensure the reliability and representativeness of the results. This threshold aligns with the EU AQ Directive, which mandates this proportion when aggregating air quality data and calculating statistical parameters. To highlight broad implications and insights into sensor technology, rather than focusing on the performance of specific manufacturers, figures illustrating brand-specific features have been anonymized. This is intended to prevent potential bias and encourage a holistic view of the data, ensuring interpretations remain focused on general trends rather than isolated examples.

3.1 Inter-device precision

Inter-device precision refers to the consistency of measurements across multiple identical devices (i.e., same brand and model) ~~of the same type~~, an important characteristic to ensure the reliability of sensor outputs over time (Moreno-Rangel et al., 2018). During QUANT, all the devices were collocated for the first 3 months and the final 3 months of the deployment to assess inter-device precision and its changes over time. Fig. 2 shows the inter-device precision (as defined by the CEN/TS 17660-1:2021, i.e., the "between sensor system uncertainty" metric: $u_s(b_s, s)$) of $PM_{2.5}$ measurements during these periods. For an overview of NO_2 and O_3 inter-device precision, see the "S6. Complementary plots" section in the supplementary (figures S4 and S5). While most of the companies display a certain level of inter-device precision stability in each period (except for one, with a seemingly upward trend in the final period), there are evident long-term changes. Notably, out of the four manufacturers assessed in the final period (each having 3 devices running simultaneously), three experienced a decline in their inter-device precision compared to two years earlier. This is likely due to both hardware degradation but also drift in the calibration, which at this point had been applied between 16 and 34 months prior (depending on the manufacturer). For extended periods, inconsistencies among devices from the same manufacturer might emerge, leading to varying readings under similar conditions. Consequently, data collected from different devices may not be directly comparable, which could result in inaccuracies or misinterpretations when analysing air quality trends or making decisions.

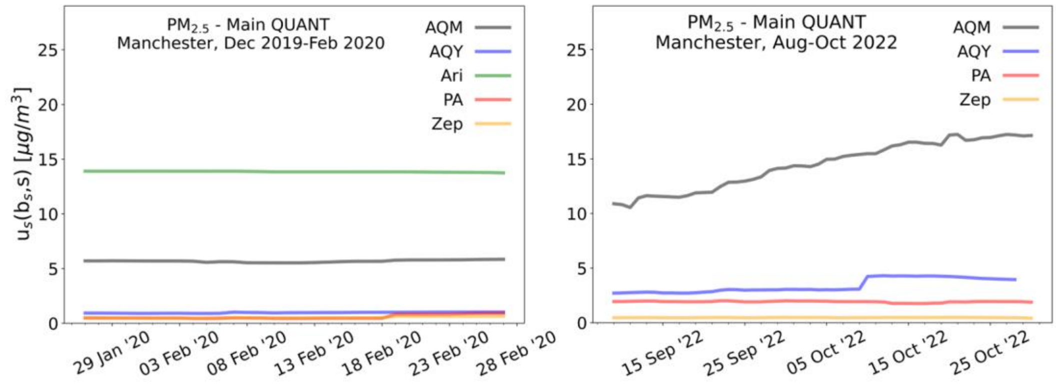


Figure 2. The inter-device precision of $PM_{2.5}$ measurements from "identical" devices across the 5 companies participating in QUANT is assessed using the "between sensor system uncertainty" metric (defined by the CEN/TS 17660-1:2021 as *u(bs, s)*). Each line represents this metric as a composite of all sensors per brand (excluding units with less than 75% data) within a 40-day sliding window.

It is worth noting that the inter-device precision provides no information on the accuracy of the sensor measurements; a batch of devices may provide a highly consistent, but also highly inaccurate measurement of the target pollutant.

The "target plot" (as shown in Fig. 3) is a tool commonly used to depict the bias/variance decomposition of an instrument's error relative to a reference (for more details see Jolliff et al. (2009)). The mean bias error (MBE) is used to characterise accuracy and precision is quantified by the centered Root Mean Squared Error (cRMSE, e.g. Kim et al. (2022) also called unbiased Root Mean Squared Error (uRMSE, e.g. Guimarães et al. (2018)). Fig. 3 visualises the performance of a set of $PM_{2.5}$ sensors of the WPS deployment for the first 2 months (out-of-box data) and the last 3 months of colocation (manufacturer-supplied calibrations). ~~In addition to highlighting which devices are most accurate, Fig. 3 also provides an additional perspective of inter-device precision.~~ In addition to showcasing inter-device precision, Fig. 3 also serves as a transition to accuracy evaluation (the focus of the subsequent section).

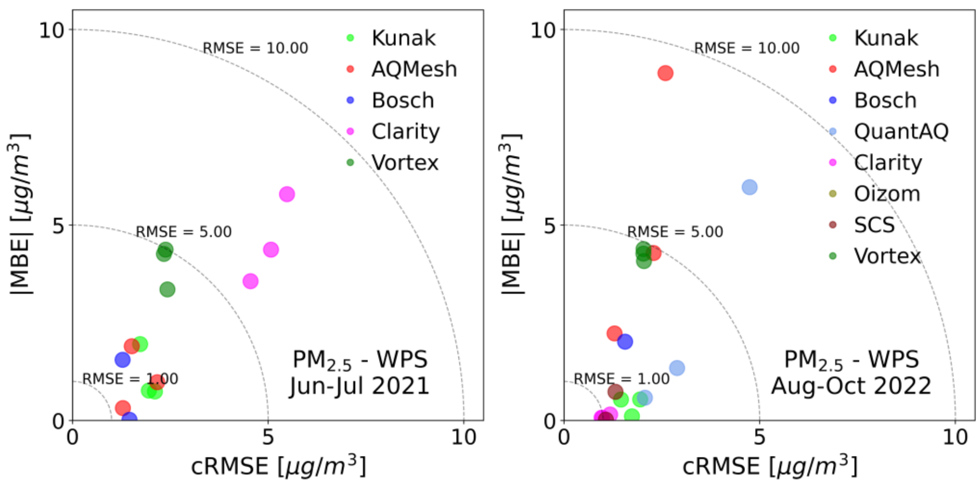


Figure 3. Target diagrams for the WPS $PM_{2.5}$ measurements during the initial co-location period (Jun-Jul 2021, left) and final co-location period (Aug-Oct 2022, right). The error (RMSE) for each instrument is decomposed into the MBE (y-axis) and cRMSE (x-axis). Each point represents an individual sensor device, with duplicate devices having the same colour.

Since only units with more than 75% of the data were considered, the plot on the right shows fewer units than the plot on the left.

3.2 Device accuracy and co-location calibrations

Sensor measurement accuracy denotes how close a sensor's readings are to reference values (Wang et al., 2015). Characterising this feature is imperative for establishing sensor reliability and making informed decisions based on its data. Fig. 4 shows that co-location calibration can greatly impact observed NO_2 sensor performance in a number of ways. Firstly, measurement bias is often, but not always, reduced following calibration, as evidenced by a general trend for devices to migrate towards the origin (RMSE = 0 ppb). Secondly, it can help to improve within-manufacturer precision ~~by grouping sensor systems from the same company closer together.~~, as evidenced by sensor systems from the same company grouping more closely as the right plot in Fig. 4 shows. The figure also highlights a fundamental challenge with evaluating sensor systems: the measured performance can vary dramatically over time —and space— as the surrounding environmental conditions change. To quantify this, 95% Confidence Intervals (CIs) were estimated for each device using bootstrap simulation and are visualised as a shaded region. For the out-of-the-box data, these regions are noticeably larger than in the calibrated results for most manufacturers, suggesting that colocation calibration has helped to tailor the response of each device to the specific site conditions. ~~This is reinforced by the cRMSE component reducing by a greater extent than the MBE; in the terminology of machine learning, the calibration has helped reduce the variance portion of the bias-variance trade-off.~~ This observation suggests that colocation calibration effectively improves each device's response to particular site conditions. This improvement is underscored by the more substantial reduction in the cRMSE component compared to the MBE. The cRMSE, representing the portion of error that persists after bias removal, essentially measures errors attributable to variance within the data space. In the context of out-of-the-box data, this "data space" spans all potential deployment locations used by manufacturers for initial calibration model training (i.e., before shipping the sensors for the QUANT study), thus exhibiting high variability. However, applying site-specific calibration significantly narrows this variability, leveraging local training data to minimise variance.

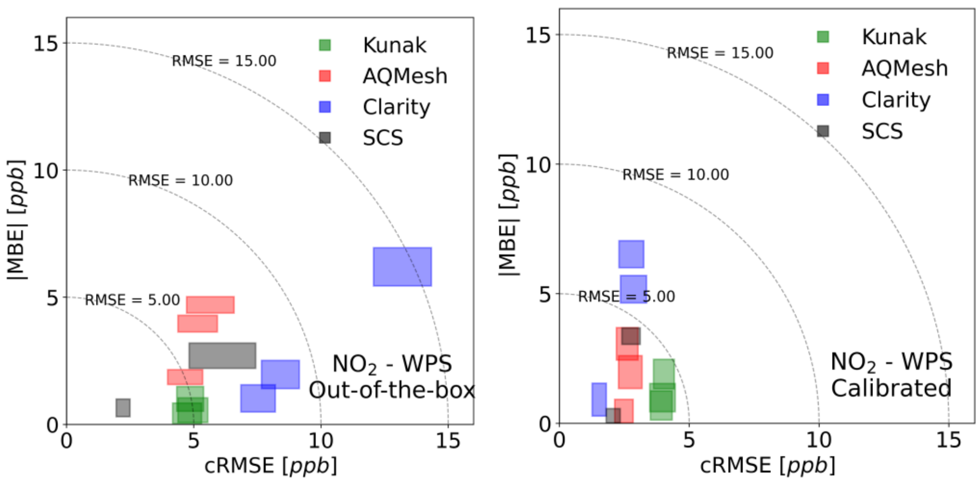


Figure 4. Effect of colocation calibration on NO_2 sensor accuracy. The accuracy is quantified using RMSE, which is decomposed into MBE (y-axis) and cRMSE (x-axis). 95% confidence regions were estimated using bootstrap sampling. The left panel displays results from the period Jun - Jul 2021 ('out-of-the-box' data), while the right-hand panel summarises Aug 2021 when calibrations were applied for all the WPS manufacturers.

However, it is important to note a limitation of Target Plots: they primarily focus on sensor behaviour around the mean. Therefore, the collective improvement evidenced by Fig. 4 might be only partial. For applications where it is important to understand how calibrations impact lower or higher percentiles, considering other metrics or visual tools would be advisable. An example of this is the absolute and Relative Expanded Uncertainty (REU, defined by the Technical Specification CEN/TS 17660-1:202). Unlike the more commonly used metrics such as R^2, RMSE, and MAE, which measure performance of the entire dataset, the REU offers a unique "point by point" evaluation, enabling its representation in various graphical forms, such as time series or concentration space (for the REU mathematical derivation, refer to section "S5. Performance Metrics"). The REU approach also incorporates the uncertainty of the reference method into its assessment, highlighting the intrinsic uncertainty present in all measurements, including those from reference instruments. This consideration of reference uncertainty is crucial for a holistic understanding of sensor performance and calibration effectiveness. For a comprehensive discussion on this, refer to Diez et al. (2022). Fig. 5 illustrates how NO_2 calibrations might not only improve collective performance around the mean (as indicated by the dotted red line in Fig. 5 and previously displayed in the target plot) but across the entire concentration range.

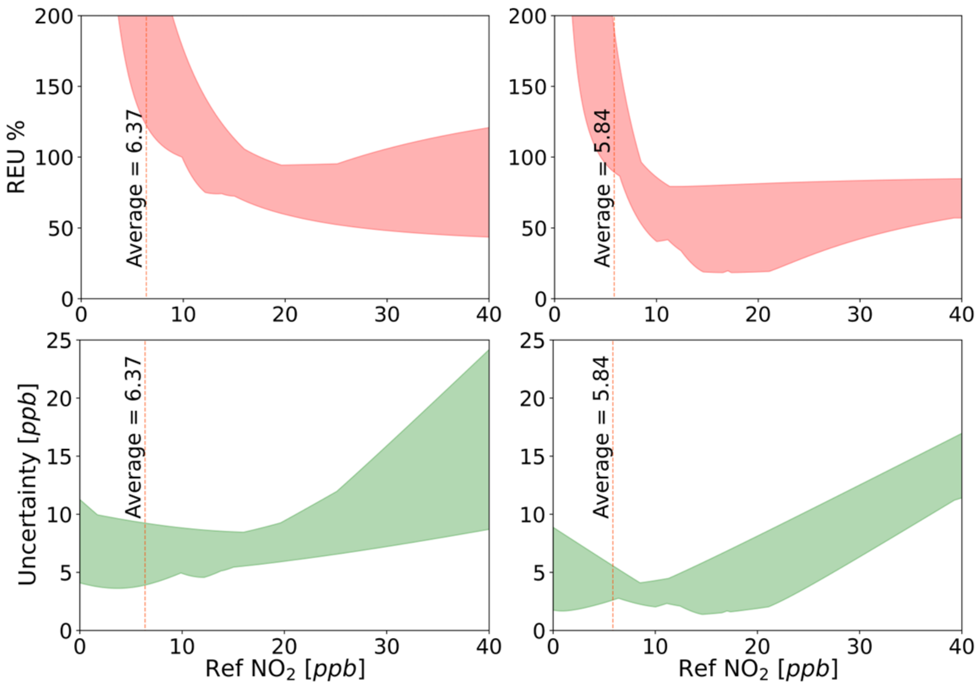


Figure 5. The top plots display the REU (%) across the concentration range, while the bottom plots depict the Absolute Uncertainty (ppb) —both before (left plots) and after (right plots) calibrating NO_2 WPS systems. The shaded areas represent the collective variability evolution (all sensors from all companies) of both metrics. These plots were constructed using the minimum and maximum value of the REU and the Absolute Uncertainty for the entire concentration range.

However, a note of caution when interpreting results from observational studies such as these is that it is impossible to ascertain a direct causal relationship between calibration and sensor performance as there are numerous other confounding factors at play (Diez et al., 2022). Notably these two data products are being assessed over different periods when many other factors will have changed, for example, the local meteorological conditions as well as human-made factors such as reduced traffic levels following the COVID-19 lockdown that commenced in March 2020.

3.3 Reference instrumentation is key

A common assumption when evaluating the performance of sensors is that the metrological characteristics of the sensor predominantly influence discrepancies detected in co-locations. While this presumption can often be justified due to both devices' (sensor and the reference method) relative scales of measurement errors, it is not always the case. Since every measurement is subject to uncertainties, it is crucial to consider those associated with the reference when deriving the calibration factors of placement.

Fig. 6 (left plots) displays the performance of a NO_2 reference instrument (Teledyne T200U) specifically installed for QUANT, located next to the usual instrument at the Manchester supersite (Teledyne T500). Although they use different analytical techniques (chemiluminescence for the T200U and Cavity Attenuated Phase Shift Spectroscopy for the T500), their measurements are highly correlated (R^2~0.95). However, it's possible to identify a proportional bias (slope=0.69), attributed to retaining the initial calibration (conducted in York) without subsequent adjustments, a situation exacerbated by an unnoticed mechanical failure of one of the instrument's components. The REU demonstrates that, under these circumstances, an instrument designated as a reference does not meet the minimum requirements (REU ≤ 15% for NO_2 reference measurements) set out by the Data Quality Objectives (DQOs) of the EU AQ ~~Air Quality~~ Directive~~ 2008/50/EC~~. Figure S6~~3~~ shows a unique sensor evaluated against both the T500 and the T200U. The comparison against the T200U yields better results, suggesting that, in a hypothetical scenario where it was the only instrument at the site, this could lead to misleading conclusions. This situation reinforces the idea that instruments should not only be adequately characterised but also undergo rigorous quality assurance and data quality control programs, as well as receive appropriate maintenance (Pinder et al., 2019). All of this must be performed before and during the use of any instrument.

For PM monitoring, the current EU reference method is the gravimetric technique (CEN EN 12341, 2023), which is a non-continuous monitoring method that requires weighing the sampled filters and off-line processing of the results. Techniques that have proven to be equivalent to the reference method (called "equivalent to reference" in the EU AQ ~~Air Quality~~ Directive) are very often used in practice. In the UK context, the Beta Attenuated Monitor (BAM) and FIDAS (optical aerosol spectrometer) are equivalent-to-reference methods commonly used as part of the Urban AURN Network (Allan et al., 2022). To illustrate these differences in practice, Fig. 6 compares these two equivalent-to-reference $PM_{2.5}$ measurements obtained with a BAM (AURN York site, located on a busy avenue), and a FIDAS unit specifically installed for QUANT. During this specific period, they show a strong linear association ~~do not fully agree~~ (R^2 = 0.87). Although the bias is not extremely pronounced (slope=0.80), the FIDAS measurements are, on average, systematically lower compared to BAM. ~~Despite a not very pronounced bias (slope=0.80), the dispersion of points around the best-fit line is noticeable, limiting the linearity of the FIDAS compared to the BAM.~~

In the hypothetical case that the BAM were to be considered the reference method (arbitrarily chosen for this example as it is the current instrument at the AURN York site) when assessing the FIDAS under these test conditions, it would only meet the criterion stipulated by the EU DQOs for indicative measurements (REU ≤ 25% for $PM_{2.5}$), but not for fixed (i.e., reference) measurements (REU ≤ 50% for $PM_{2.5}$).~~ Of course,~~ This example is primarily intended to illustrate the magnitude of differences between both methods for this particular application, and by no means does this observation imply that the FIDAS measurements are inherently problematic.

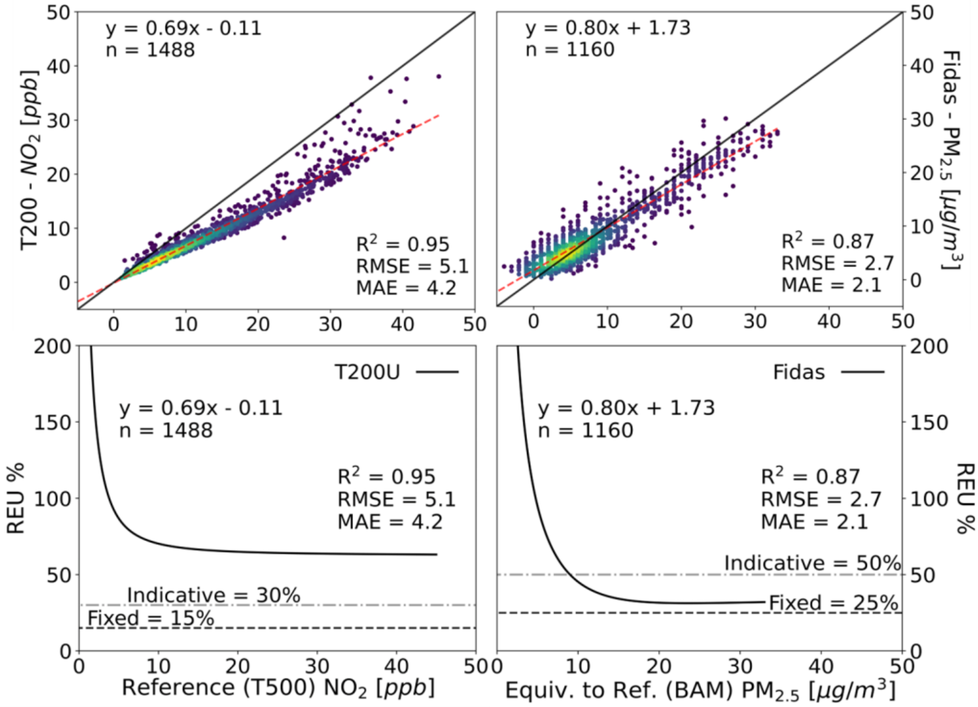
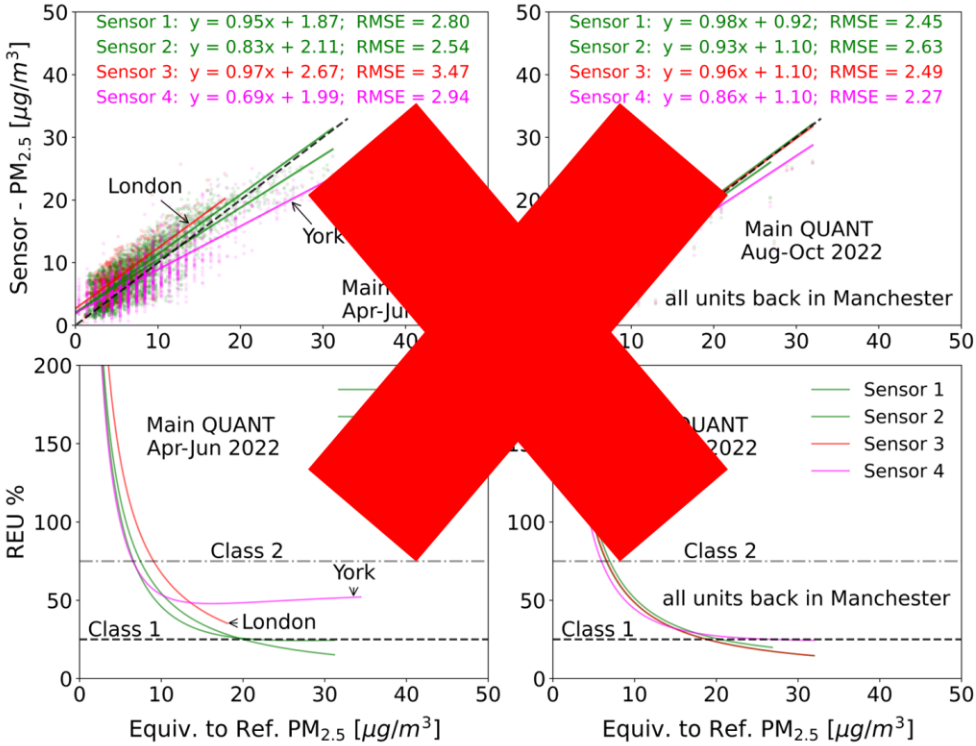


Figure 6. The left plots depict the comparison between the Teledyne T200U (chemiluminescence analyzer) and the reference method (Teledyne T500 CAPS analyzer) at the Manchester supersite. The plots to the right illustrate $PM_{2.5}$ measurements in York, taken with a FIDAS instrument (optical aerosol spectrometer) and a BAM 1020 (beta attenuation monitor), both equivalent-to-reference methods. While the top plots show the regression (including some typical single-value metrics), those on the bottom present the REU alongside the DQOs defined by the EU AQ Directive ~~European Directive 2008/50/EC~~.

Although these two instruments (BAM and Fidas) show a greater concordance between themselves than with sensors (for the comparison of two sensor systems against the BAM and the Fidas, refer to Fig. S74), the choice of the measurement method can have a considerable impact on evaluations of this type. This underscores the importance of adequately characterising the uncertainties of the reference monitor when evaluating sensors.

3.4 Inter-location performance ~~Systems performance after location transfer~~

An extreme example of sensor performance varying due to environmental conditions is when sensors are moved between locations, as their apparent performance may vary drastically. Fig. 7 displays the REU and regression plots for four of the same $PM_{2.5}$ sensor system in two periods: April-June 2022 when the devices were working across the 3 sites (York, Manchester and London), and August-October 2022 when they were all reunited in Manchester. The RMSE remains reasonably consistent (range 2.27 to 3.47 ppb) between the devices across the periods and locations. However, for the device that moved from York to Manchester, a change in slope from 0.69 to 0.86 was observed. Because this device's slope is consistent with the other units while running in Manchester, this is likely due to the different sensor responses in the specific environments. The precise cause of this change is not immediately evident and will be the focus of a follow-up study, but could be due to changes in local conditions (e.g., weather, emissions, etc.) impacting sensor calibration and/or differences in actual $PM_{2.5}$ sources and particle characteristics at the sites (Raheja et al., 2022).



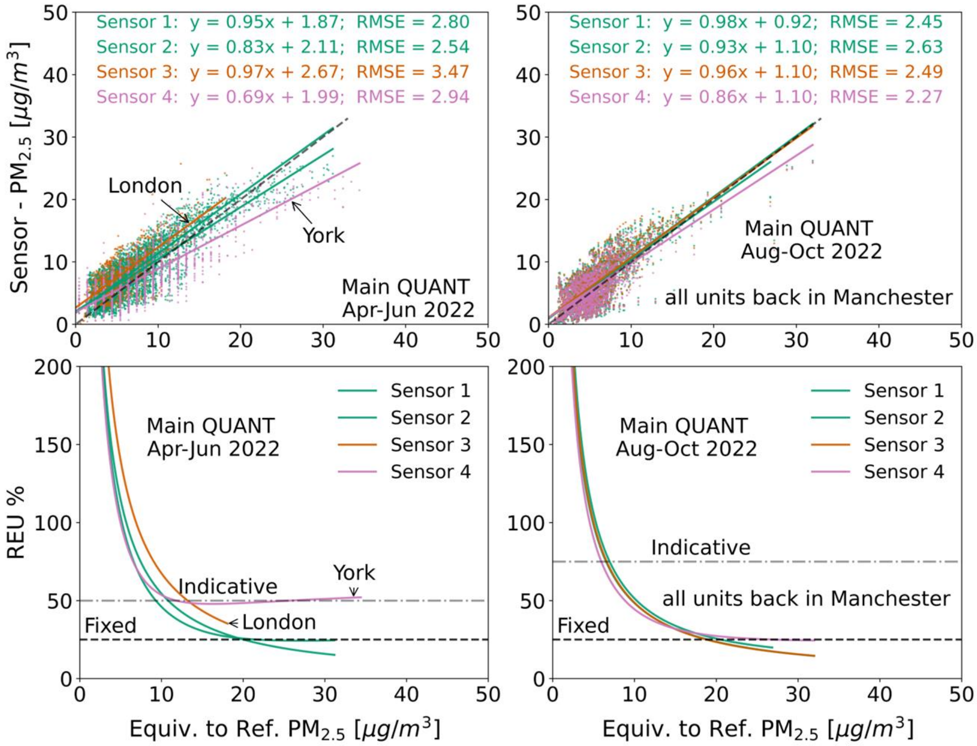
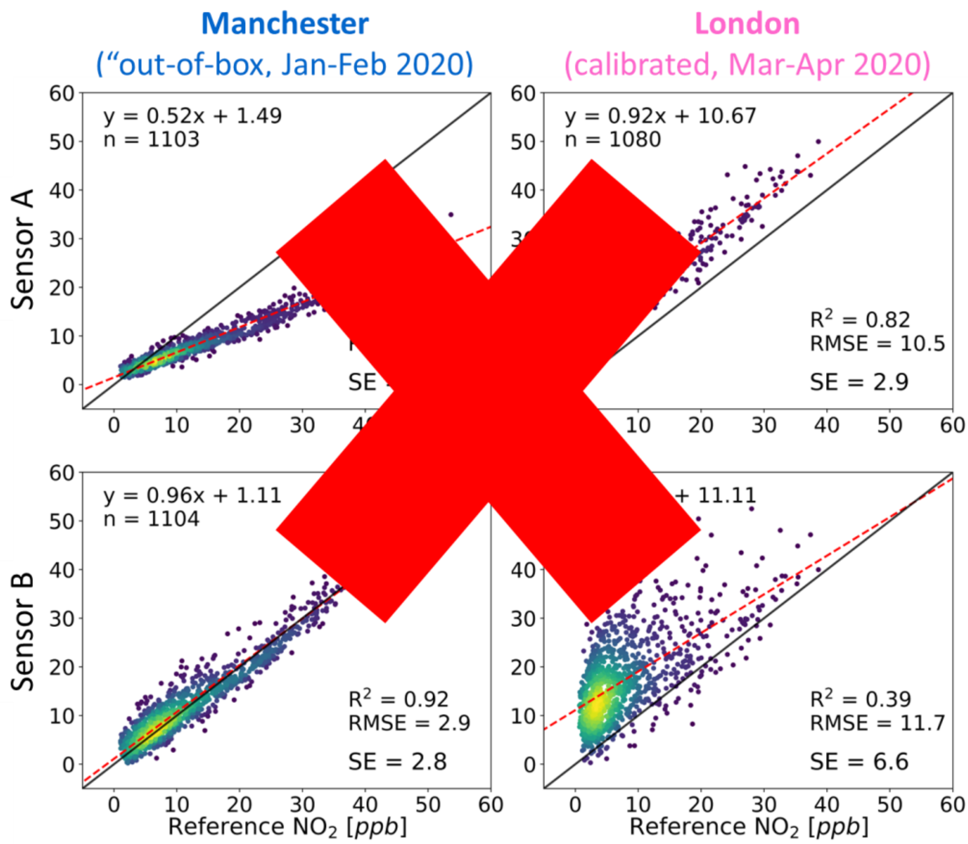
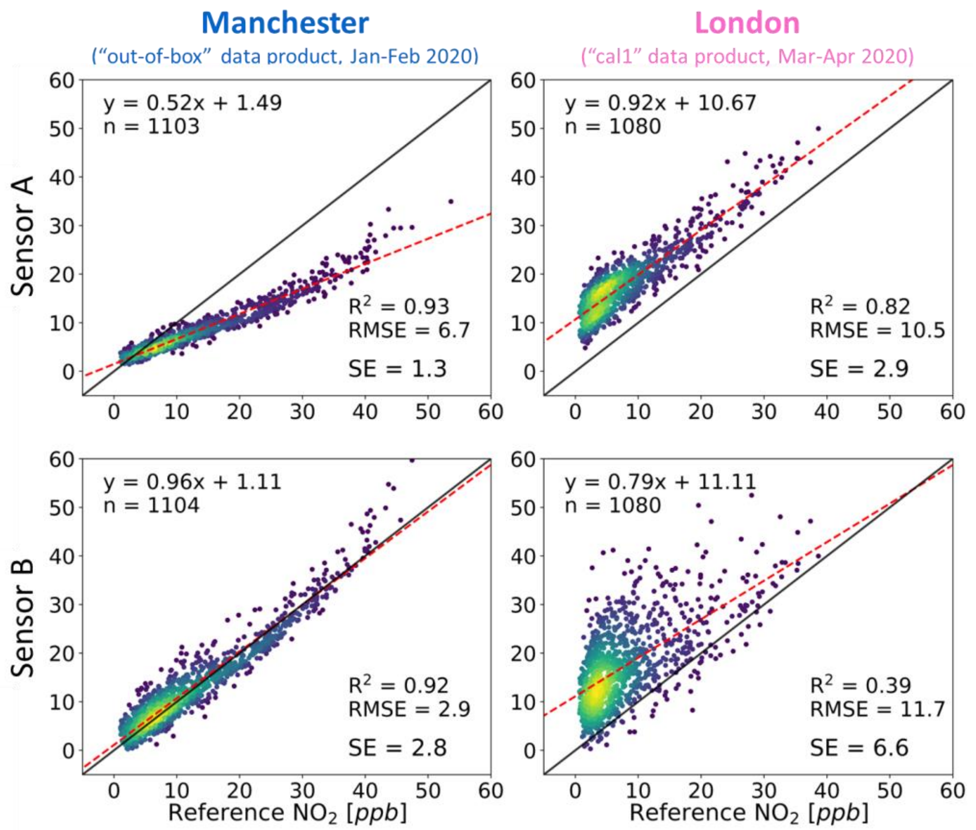


Figure 7. Regression (top) and REU (bottom) plots showing data from four $PM_{2.5}$ sensors (same manufacturer) over 2 time periods: Apr-Jun 2022 and Aug-Oct 2022. The four devices were in separate locations in the first period, but all deployed in Manchester in the second. The horizontal dashed lines represent a reference for the $PM_{2.5}$ DQOs as defined by the EU AQ Directive (for "fixed" $PM_{2.5}$ measurements, REU < 25%; for "indicative" $PM_{2.5}$ measurements, REU < 50%). Readers are encouraged to consult the specified standard for further details.

A second example of inter-location performance ~~changing between locations~~ is presented in Fig. 8, showing NO_2 data from two sensor systems (from two different manufacturers, identified as Systems A and B) ~~(different brands, one~~

~~shown on top of the other)~~ before (left plots) and after (right plots) they were moved from Manchester to London in March 2020. Both sensors saw a reduction in agreement with the reference instrument at the London site compared to Manchester, despite both these sites being classified as urban-background with reference instrument performance regularly audited by the UK National Physical Laboratory.





~~**Figure 8. Comparison of NO_2 measurements for two systems (A and B) that were moved between Manchester (left plots) and London (right plots). The Manchester deployment was from January - February 2020, and the London data were recorded from April - May 2020.**~~

Figure 8. Comparative analysis of NO_2 measurements from two systems (A and B), across two urban settings. The left plots display Manchester "out-of-box" data product (January to February 2020), while the right plots show London "cal1" data product (April to May 2020). This "cal1" label does not indicate corrections specific to London's conditions but denotes a data product from a specific period (as detailed in Figures S2 and S3). The colour gradient represents the density of data points, with darker shades indicating lower densities and brighter shades signifying higher densities.

~~The primary distinction between both systems' behaviour lies in the fact that the sensor located in the top row, even after being relocated to London, maintains a linear response (albeit slightly more degraded than that observed in Manchester, as the R^2 and RMSE show). In contrast, in the second system (bottom row), the response is notably noisier as the Standard Error (SE) — which is the dispersion of the data around the best-line fit line, i.e., the remaining error after bias correction. In scenarios akin to this latter, where there is a high variance in the residuals, a linear correction will not provide a significant improvement. While more sophisticated corrections could be applied, these will be limited by domain knowledge of the end-user, and potentially by other complex data sources that might be available. However, it is important to remember that additional post-processing could increase the risk of overfitting (Aula et al., 2022). On the other hand, for cases like the top plots, users might benefit from trying to correct them using simple linear correction (e.g. using reference instruments if available) or other approaches that could provide means for zero and span correction. A straightforward and cost-effective example could be the use of diffusion tubes for the case of NO2, as discussed in Section 3.6.~~ The primary distinction between both systems' behaviour lies in the fact that the sensor located in the top row (Sensor A), even after being relocated to London, maintains a linear response (albeit slightly more degraded than that observed in Manchester, as indicated by the R^2 and RMSE). In contrast, Sensor

B's response becomes significantly noisier upon relocation to London, as highlighted by the Standard Error (SE) — which represents the remaining error after applying a perfect bias correction. Despite both systems utilising identical sensing elements, the variance in residuals between them may stem from the distinct calibration approaches applied by the respective companies.

For cases resembling Sensor A, users might find it beneficial to implement simple linear correction methods (e.g., using reference instruments if available) or explore other strategies for zero and span correction. A practical and cost-effective approach, for example, is using diffusion tubes for NO_2 measurements, as discussed in Section 3.6. Conversely, in scenarios characterised by high variance in residuals, such as those observed with Sensor B, a-posteriori attempts to apply a simple linear correction are unlikely to result in significant improvement. While more sophisticated corrections are theoretically feasible, their effectiveness is limited by the end-user's domain knowledge and the availability of additional complex data sources. Furthermore, it is important to consider that excessive post-processing may lead to overfitting —a situation where a model excessively conforms to specific patterns in the training data, resulting in poor performance on new, unseen data (Aula et al., 2022).

3.5 Long-term stability

The long-term stability of sensor response is also an important facet of its performance, especially for certain use cases such as multi-year network deployments. There can be multiple causes of long-term changes to sensor response, for example, particles settling inside the sampling chamber in optical-based sensors(e.g. Hofman et al. (2022)), or the gradually changing composition of electrochemical cells (e.g. Williams (2020)). How these changes manifest themselves in the data must be identified if ways to account for them are to be implemented.

Fig. 9 shows the temporal nature of the O_3 and NO_2 errors (MBE, cRMSE and RMSE) from a sensor system between February 2020 and October 2022. The O_3 shows (Fig. 9a) a gradual increase in the overall measurement error, largely due to an increase in the MBE. It also shows a distinct seasonality MBE, increasing by a factor of 3-4 between March and July compared to the August-February period. The cRMSE component shows fluctuations during the study but only has a small increasing trend. The NO_2 system (Fig. 9b) demonstrates a consistently increasing overall error, with a less pronounced seasonal influence. The bias contributes greatly to the total error (see Section 3.6 for NO_2 sensor correction, Fig. 9c).

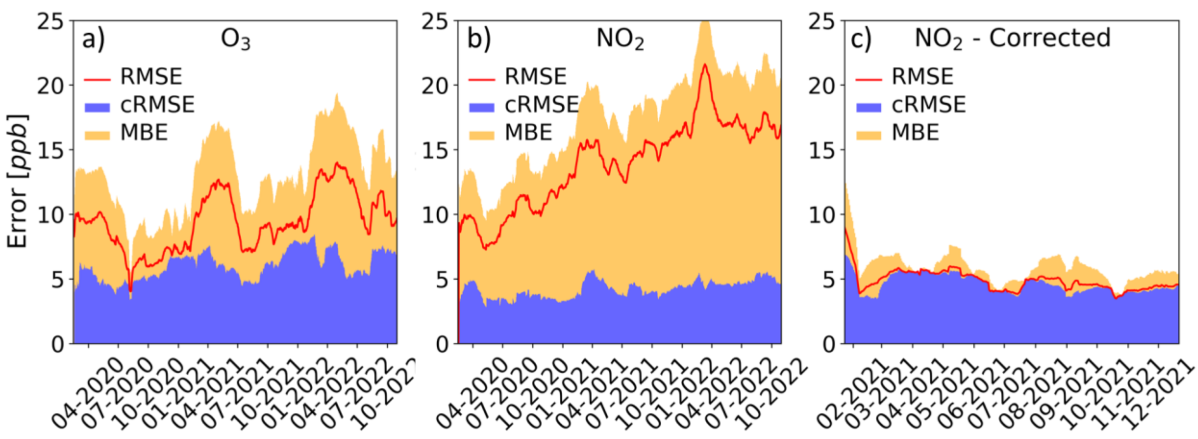
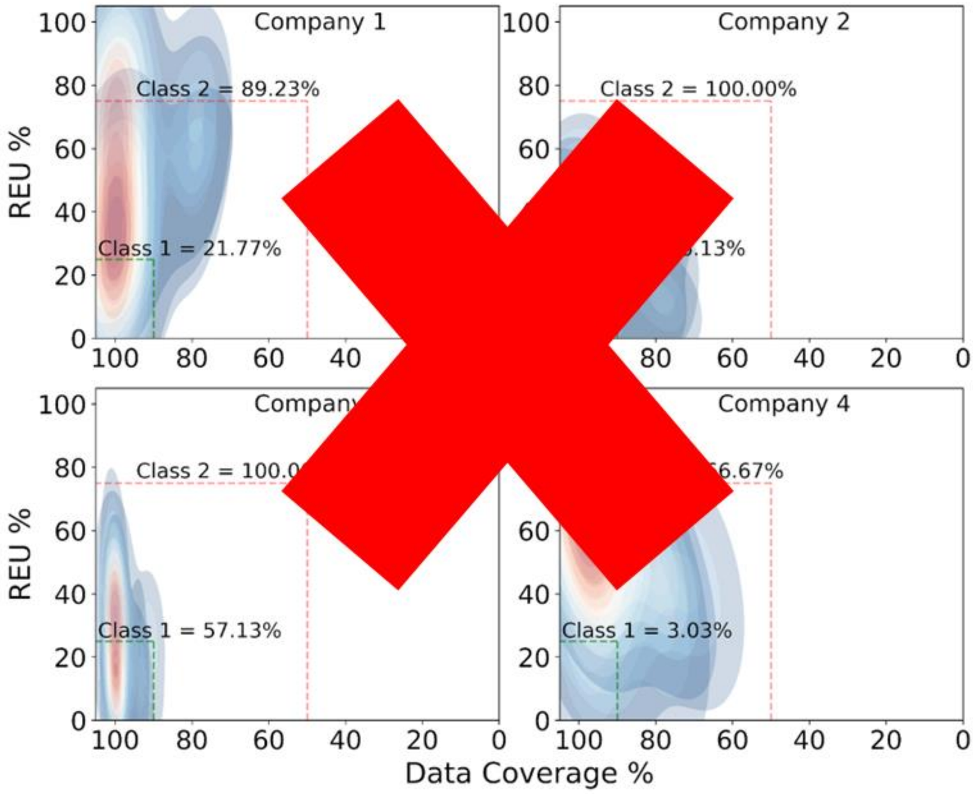


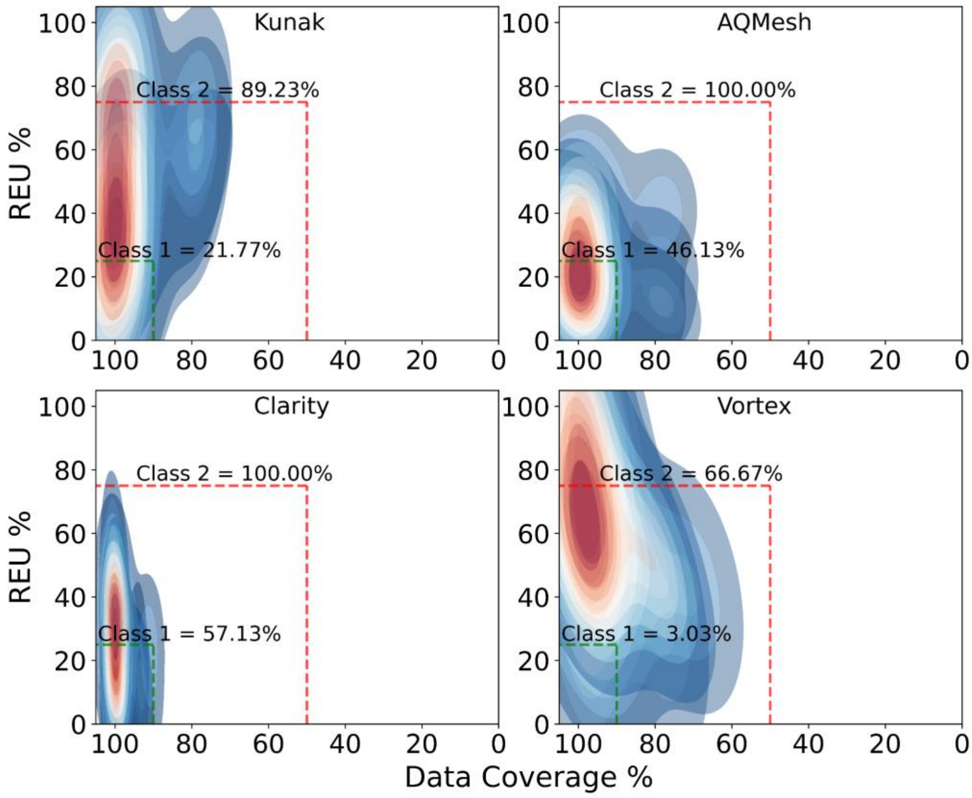
Figure 9. **Seasonal variation of error (as RMSE, red line) of one of the systems belonging to the Main QUANT, decomposed into cRMSE (in blue) and MBE (in yellow) estimated based on a 40-day (aligning with the sample size recommendation by the CEN/TS 17660-1:2021 standard for on-field tests) moving window approach with a 1-day slide (i.e., advancing the calculation 1 day at a time) ~~(1-day slide) moving window~~. Panel a) is for O_3 measurements, and panel b) is for NO_2 (April 2020-Oct 2022). Panel c) is also for NO_2, this time showing the effect of a linear correction using diffusion tubes (see next section for more details).**

3.6 Informing end-use applications

Ultimately, for any air pollution monitoring application, the requirements of the task should dictate the measurement technology options available. For example, if the requirement for a particular measurement is to assess legal compliance, then lower measurement uncertainty must be a key consideration as the reported values need to be compared to a limit value. In contrast, if an application aimed to look at long-term trends in pollutants, then absolute accuracy may not be as important as the long-term stability of sensor response. ~~In order to realise the potential of air pollution sensor technologies, end users need to be provided with the information required to critically assess the strengths and weaknesses of potential candidate sensor devices, ideally in an easy-to-access and interpret manner.~~ To realise the potential of air pollution sensor technologies, end users need to align their specific measurement needs with the capabilities of available devices. Achieving this necessitates access to unbiased performance data, such as long-term stability and accuracy across varying conditions, ideally in an easy-to-access and interpret manner.

Understanding the uncertainty associated with a ~~measurement~~ instrument is essential for recognizing its capabilities and limitations. Accurate instruments are crucial, especially in areas like public health decision-making, where inaccurate data can have profound implications (Molina Rueda et al., 2023). Furthermore, instruments that operate autonomously ensure consistent, uninterrupted data collection, making them more efficient and cost-effective in terms of maintenance and calibration. ~~Figure 10 shows the REU (y-axis) and Data Coverage (DC, x-axis) of companies measuring NO_2 with more than 2 systems running to avoid ambiguity in the results. Using multiple systems, not only avoids ambiguity in results but also enhances the robustness of the data collected.~~ Figure 10 illustrates the collective behaviour of NO_2 sensors from each of the four companies with more than two working systems, showcasing their REU (y-axis) versus Data Coverage (DC, x-axis). Both parameters were calculated for each sensor system using a 40-day moving window approach and then aggregated by brand, ensuring a comprehensive analysis. This methodology leverages overlapping data from multiple sensors to provide a robust representation of company-wide sensor performance and aims to prevent biassed interpretations. Both REU and DC are key criteria within the EU scheme (EU 2008/50/EC) for evaluating the performance of measurement methods, and are complemented by the CEN/TS 17660-1:2021 specifically for sensors. The latter ~~This~~ document defines three different sensor system tiers. Class 1 NO_2 sensors, bounded by the green rectangle (REU < 25% and DC > 90%), offer higher accuracy than Class 2 sensors (REU < 75% and DC > 50%), delimited ~~highlighted~~ by the red rectangle (Class 3 sensors have no set requirements). Presenting the REU and DC ~~data~~ like in Fig. 10 ~~this~~ helps users anticipate the performance of sensor systems —under the assumption that all sensors from the same brand will behave similarly in equivalent environmental conditions—providing more insight into selecting the appropriate instrument for a given project or study.





~~**Figure 10. The REU vs. Data Coverage (DC) for 4 systems companies was evaluated during the WPS for the period Nov 2021-Oct 2022 (after all companies had at least one calibrated product). Both the REU and the DC were estimated based on a 40-day size (which is the number of days used by CEN/TS 17660-1:2021 for on-field tests) moving window (1-day slide). While the green rectangle represents the DQOs for Class 1 sensors, the red one limits the DQOs for Class 2 sensors (Class 3 sensors have no requirements).**~~

Figure 10. REU vs. DC for 4 sensor system companies measuring NO_2, with more than two units working simultaneously during the WPS (period Nov 2021-Oct 2022, after all companies provided at least one calibrated product). Each heat map plot (cooler colours for lower densities and warmer colours for higher densities) aggregates the REU and DC from sensors of the same brand working concurrently. The calculation of these two parameters employ a 40-day (aligning with the sample size recommendation by the CEN/TS 17660-1:2021 standard for on-field tests) moving window approach with a 1-day slide (i.e., advancing the calculation 1 day at a time). The green dashed rectangle limits the Data Quality Objectives (DQOs) for Class 1 sensors, and the red dashed rectangle outlines the DQOs for Class 2 sensors.

Depending on the nature of the sensor data uncertainty, methods can be implemented to improve certain aspects of the data quality for a particular application. One such example is the use of distributed networks to estimate sensor measurement errors, such as that described by (J. Kim et al., 2018). ~~Depending on the application, simpler methods could also be available to reduce the magnitude of the changing bias, and thus significantly improve the accuracy of an individual sensor system, but also that of broader sensor networks. For the case shown in Fig.9b, one possible way to do this would be using supporting observations of NO_2 made via diffusion tubes.~~ Depending on the application and available options, users can access alternative methods to reduce bias, thus enhancing the accuracy of sensor systems and networks. For example, "Indicative methods", as defined by the EU AQ Directive, such as diffusion tubes (e.g., NOx, SO_2, VOCs, etc.), can be an option. Specifically, our study leverages diffusion tube data for NO_2, illustrating one effective approach to bias correction using supporting observations, as exemplified in Fig. 9b. These measurements are widely used to monitor NO_2 concentrations in UK urban environments, due to their lower cost (~£5 per tube) and ease of deployment, but only provide average concentrations over periods of weeks to months (Butterfield et al., 2021). During QUANT, NO_2 diffusion tubes were deployed at the 3 colocation sites (see Section S7 at the Supp. for more details). Combining these measurements offers the possibility of quantifying the average sensor bias, thus reducing the error on the sensor measurement whilst maintaining the benefits of its high time-resolution observations. It is important to note that while bias correction has been applied to the sensor data, the NO_2 diffusion tube concentrations used for comparison purposes must also be adjusted (e.g. following Defra ~~DEFRA~~ (2022)). Fig. 9c shows the accuracy of the same NO_2 sensor data shown in Fig. 9b but applies a monthly offset calculated as the difference between its monthly average measurement and that from the diffusion tube (see Figure S8~~5~~). This shows a dramatic reduction in overall error largely driven by its bias correction. What remains largely resulting from the cRMSE, i.e. the error variance that might arise from limitations from the sensing technology itself and/or the conversion algorithms used to transform the raw signals into the concentration output. To validate the efficacy and reliability of this bias correction method, further long-term studies are warranted.

The development and communication of methods that improve sensor data quality, ideally in accessible ~~digestible~~ case studies, would likely increase the successful application of sensor devices for local air quality management. There is also a need for similar case studies showcasing the successful application of sensor devices for particular monitoring tasks. An example of this from the QUANT dataset is the use of sensor devices to successfully identify change points in a pollutant's concentration profile. These are points in time where the parameters governing the data generation process are identified to change, commonly the mean or variance, and can arise from human-made or natural phenomena (Aminikhanghahi and Cook, 2017). Determining when a specific pollutant has changed its temporal nature is a challenging task as there are a large number of confounding factors that influence atmospheric concentrations ~~a pollutant's concentration at a specific point in time~~, including but not limited to seasonal factors, environmental conditions (both natural and arising from human behaviour), and meteorological factors. This challenge has lead to several "deweathering" techniques being proposed in the literature (Carslaw et al., 2007; Grange and

Carslaw, 2019; Ropkins et al., 2022). While change point detection is highlighted here as a promising application of sensor data, it represents just one of many potential methodologies that could be explored with the QUANT dataset.

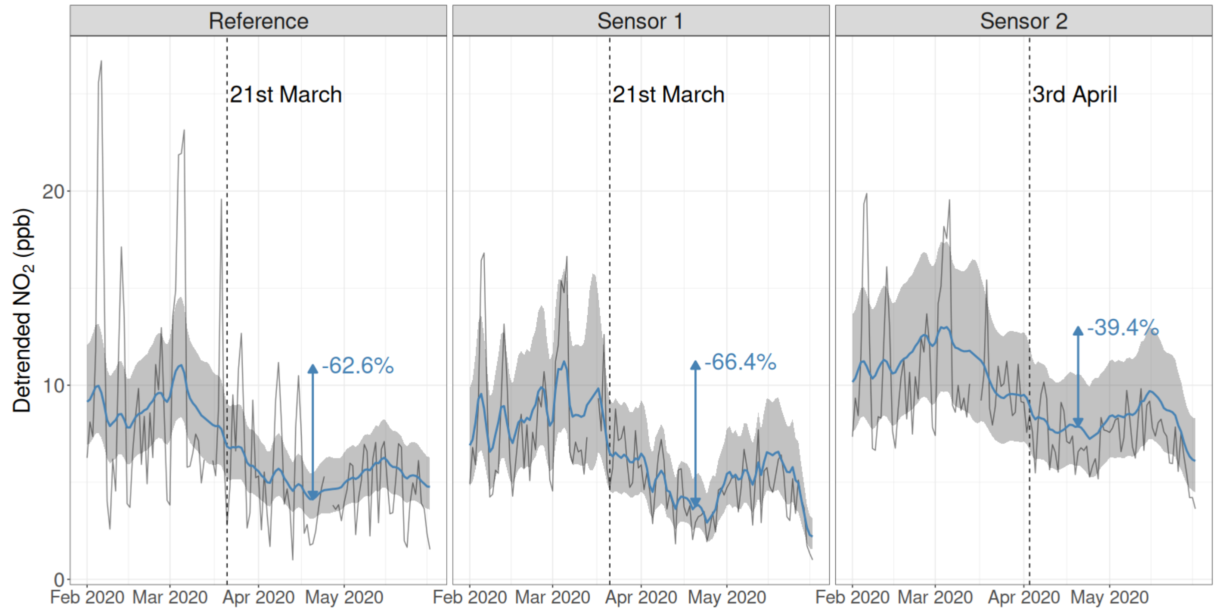


Figure 11. NO_2 measurements (black solid line) and detrended estimates (blue solid line with 95% confidence interval in the shaded grey region) from the reference instrument (left panel) and 2 sensor systems (middle and right panels) from Manchester in 2020. Vertical dashed lines and their corresponding dates indicate identified change points, which correspond to the introduction of the first national lockdown due to COVID-19 on the 23rd of March 2020. The percentage in blue represents the relative peak-trough decrease from 5th March to 20th April.

~~A novel statistical approach to smoothing air quality measurements was applied, accounting for these external factors (Lacy & Moller). This method was applied to NO2 concentrations determined from the sensor systems that had remained in Manchester throughout 2020, aiming to identify whether the well-documented reduction in ambient NO2 concentrations could be observed due to changes in travel patterns associated with COVID-19 restrictions. To provide an objective quantification of whether a change point had occurred, the Bayesian online change point detection (Adams & MacKay, 2007) was applied. Of the 8 devices that measured NO2, clear changepoints corresponding to the introduction of a lockdown were identified in 2 (Fig.11). While this is an unsupervised analysis, it demonstrates the potential of these devices to identify long-term trends with appropriate processing, even with only having had 3 months of training data to fit the model to. This is especially aided by the given algorithm's ability to use reference data as a prior allowing sensor systems to fine-tune the model.~~

A state-space based deweathering model was applied to NO_2 concentrations measured from the sensor systems that had remained in Manchester throughout 2020 to remove these confounding factors, with the overarching objective to identify whether the well-documented reduction in ambient NO_2 concentrations due to changes in travel patterns associated with COVID-19 restrictions could be observed in the low-cost sensor systems. To provide a quantifiable measure of whether a meaningful reduction had occurred, the Bayesian online change-point detection (Adams & MacKay, 2007) was applied. Of the 8 devices that measured NO_2, clear change points corresponding to the introduction of a lockdown were identified in 2 (Fig.11), demonstrating the potential of these devices to identify long-term trends with appropriate processing, even with only 3 months of training data.

4. Conclusions

Lower-cost air pollution sensor technologies have significant potential to improve our understanding and ability to manage air pollution issues. Large-scale uptake in the use of these devices for air quality management has, however, been primarily limited by concerns over data quality and a general lack of a realistic characterisation of the measurement uncertainties making it difficult to design end uses that make the most of the data information content. ~~Developments in the field of air pollution sensor technology are also developing rapidly, with advances in both the measurement technology and particularly in the data post-processing and calibration.~~ Advances are occurring rapidly, in both the measurement technology and particularly in the data post-processing and calibration. A challenge with the use of sensor-based devices is that many of the end-use communities do not have access to extensive reference-grade air pollution measurement capability (Lewis & Edwards, 2016), or in many cases, expertise in making atmospheric measurements or the technical ability for data post-processing. For this reason, reliable information on expected sensor performance needs to be available to aid effective end-use applications. Large-scale independent assessments of air sensor technologies are non-trivial and costly, however, making it difficult for end users to find relevant performance information on current sensor technologies. The QUANT assessment is a multi-year study across multiple locations, that aims to provide relevant information on the strengths and weaknesses of commercial air pollution sensors in UK urban environments.

The QUANT sensor systems were installed at two highly instrumented urban background measurement sites, in Manchester and London, and one roadside monitoring station in York. The study design ensured that multiple devices were collocated to assess inter-device precision, and devices were also moved between locations and able to test additional calibration data products to assess and enable developments in sensor performance under realistic end-use scenarios. A wider participation component of the Main QUANT assessment was also run at the Manchester site to expand the market representation of devices included in the study, and also to assess recent developments in the field.

A high-level analysis of the dataset has highlighted multiple facets of air pollution sensor performance that will help inform their future usage. Inter-device precision has been shown to vary, both between different devices of the same brand and model ~~types~~ and over different periods of time, with the most accurate devices generally showing the highest levels of inter-device precision. The accuracy of the reported data for a particular device can be impacted by a variety of factors, from the calibrations applied to its location or seasonality. This has important implications for the way sensor-based technologies are deployed and supports the case made by others (Bittner et al., 2022; Farquhar et al., 2021; Crilley et al., 2018; Williams, 2020; Bi et al., 2020) that practical methods to monitor sensor bias will be crucial in uses where data accuracy is paramount. Ultimately, this work shows that sensor performance can be highly variable between different devices and end-users need to be provided with impartial performance data on characteristics such as accuracy, inter-device precision, long-term drift and calibration transferability in order to decide on the right measurement tool for their specific application.

In addition to these findings, this overview lays the groundwork for more detailed research to be presented in future publications. Subsequent analyses will focus on providing a more nuanced understanding of the uncertainty in air pollution sensor measurements, thus equipping end-users with better insights into the capability of sensor data. Future studies will delve into specific aspects of air pollution sensor performance: 1) a comprehensive performance evaluation of $PM_{2.5}$ data, assessing their accuracy and reliability under different environmental conditions; 2) an in-depth analysis of NO_2 measurements, examining their sensitivity and response in various urban environments; and 3) a detailed investigation into the detection limits of these sensor technologies, targeting their optimised application in low concentration scenarios. These focused studies are basic steps needed to further advance our understanding of sensors' capabilities and limitations, ensuring informed and effective application in air quality monitoring.

Supplementary

The supplement related to this article is available online at:

Data availability

~~The data for this study can be found at the Centre for Environmental Data Analysis (CEDA): Lacy et al. (2023): Quantification of Utility of Atmospheric Network Technologies: (QUANT): Low-cost air quality measurements from 52 commercial devices at three UK urban monitoring sites. NERC EDS Centre for Environmental Data Analysis, date of citation (https://catalogue.ceda.ac.uk/uuid/ae1df3ef736f4248927984b7aa079d2e).~~

The QUANT dataset, accessible at the Centre for Environmental Data Analysis (CEDA) (Lacy et al., 2023; https://catalogue.ceda.ac.uk/uuid/ae1df3ef736f4248927984b7aa079d2e), is the most extensive collection to date assessing air pollution sensors' performance in UK urban settings. It encompasses gas and PM sensor data recorded in the native reporting frequency of each device. The reference data from the three monitoring sites can be found at:

- MAQS: https://data.ceda.ac.uk/badc/osca/data/manchester;
- LAQS: https://www.londonair.org.uk/london/asp/datadownload.asp);
- YoFi: https://uk-air.defra.gov.uk/data/data_selector.

A comprehensive data descriptor manuscript, detailing the QUANT dataset's collection methods, processing protocols, accessibility features, and overall structure—including variables, data reporting frequencies, and QA/QC practices—has been submitted for publication. At the time of this writing, the manuscript is still under review.

A GitHub repository at https://github.com/wacl-york/quant-air-pollution-measurement-errors provides access to Python and R scripts designed for generating diagnostic visuals and metrics related to the QUANT study, along with sample analyses using the QUANT dataset.

Author contributions

The initial draft of the manuscript was created by SD, PME, and SL. The research was conceptualised, designed, and conducted by PME and SD. Methodological framework and conceptualization were developed by SD, PME, and SL. Data analysis was primarily conducted by SD and SL. The software tools for data visualisation and analysis were developed by SD and enhanced by SL. MF, MP and NM supplied the reference data critical for the study. TB, HC, DH, SG, NAM and JU made substantive revisions to the manuscript, enriching the final submission.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

This work was funded as part of the UKRI Strategic Priorities Fund Clean Air program (NERC NE/T00195X/1), with support from Defra. We would also like to thank the OSCA team (Integrated Research Observation System for Clean Air, NERC NE/T001984/1, NE/T001917/1) at the MAQS, for their assistance in data collection for the regulatory-grade instruments. The authors wish to acknowledge Dr. Katie Read and the Atmospheric Measurement and Observation Facility (AMOF), a Natural Environment Research Council (UKRI-NERC) funded facility, for providing the Teledyne 200U used in this study and for their expertise on its deployment. Special thanks are due to Dr David Green (Imperial College London) for granting access and sharing the data from LAQS (NERC NE/T001909/1). Special thanks to Chris Anthony, Killian Murphy, Steve Andrews and Jenny Hudson-Bell from WACL for the help and support to the project. Our acknowledgment would be incomplete without mentioning Stuart Murray and Chris Rhodes from the Department of Chemistry Workshop for their technical assistance and advice. Further, we acknowledge Andrew Gillah, Jordan Walters, Liz Bates and Michael Golightly from the City of York Council, who were instrumental in facilitating site access and regularly checking on instrument status. We acknowledge the use of ChatGPT to improve the writing style of this article.

References

Adams, R. P. and MacKay, D. J. C.: Bayesian Online Changepoint Detection, https://doi.org/10.48550/arXiv.0710.3742, 19 October 2007.

Alam, M. S., Crilley, L. R., Lee, J. D., Kramer, L. J., Pfrang, C., Vázquez-Moreno, M., Ródenas, M., Muñoz, A., and Bloss, W. J.: Interference from alkenes in chemiluminescent NOx measurements, Atmospheric Meas. Tech., 13, 5977–5991, https://doi.org/10.5194/amt-13-5977-2020, 2020.

Allan, J., Harrison, R., and Maggs, R.: Measurement Uncertainty for PM2.5 in the Context of the UK National Network, 2022.

Aminikhanghahi, S. and Cook, D. J.: A survey of methods for time series change point detection, Knowl. Inf. Syst., 51, 339–367, https://doi.org/10.1007/s10115-016-0987-z, 2017.

A. Miech, J., Stanton, L., Gao, M., Micalizzi, P., Uebelherr, J., Herckes, P., and P. Fraser, M.: In situ drift correction for a low-cost NO 2 sensor network, Environ. Sci. Atmospheres, 3, 894–904, https://doi.org/10.1039/D2EA00145D, 2023.

Aula, K., Lagerspetz, E., Nurmi, P., and Tarkoma, S.: Evaluation of Low-cost Air Quality Sensor Calibration Models, ACM Trans. Sens. Netw., 18, 72:1-72:32, https://doi.org/10.1145/3512889, 2022.

Baron, R. and Saffell, J.: Amperometric Gas Sensors as a Low Cost Emerging Technology Platform for Air Quality Monitoring Applications: A Review, ACS Sens., 2, 1553–1566, https://doi.org/10.1021/acssensors.7b00620, 2017.

Bi, J., Wildani, A., Chang, H. H., and Liu, Y.: Incorporating Low-Cost Sensor Measurements into High-Resolution PM2.5 Modeling at a Large Spatial Scale, Environ. Sci. Technol., 54, 2152–2162, https://doi.org/10.1021/acs.est.9b06046, 2020.

Bigi, A., Mueller, M., Grange, S. K., Ghermandi, G., and Hueglin, C.: Performance of NO, NO2 low cost sensors and three calibration approaches within a real world application, Atmospheric Meas. Tech., 11, 3717–3735, https://doi.org/10.5194/amt-11-3717-2018, 2018.

Bittner, A. S., Cross, E. S., Hagan, D. H., Malings, C., Lipsky, E., and Grieshop, A. P.: Performance characterization of low-cost air quality sensors for off-grid deployment in rural Malawi, Atmospheric Meas. Tech., 15, 3353–3376, https://doi.org/10.5194/amt-15-3353-2022, 2022.

Brown, R. J. C. and Martin, N. A.: How standardizing ‘low-cost’ air quality monitors will help measure pollution, Nat. Rev. Phys., 5, 139–140, https://doi.org/10.1038/s42254-023-00561-8, 2023.

Buehler, C., Xiong, F., Zamora, M. L., Skog, K. M., Kohrman-Glaser, J., Colton, S., McNamara, M., Ryan, K., Redlich, C., Bartos, M., Wong, B., Kerkez, B., Koehler, K., and Gentner, D. R.: Stationary and portable multipollutant monitors for high-spatiotemporal-resolution air quality studies including online calibration, Atmospheric Meas. Tech., 14, 995–1013, https://doi.org/10.5194/amt-14-995-2021, 2021.

Bulot, F. M. J., Johnston, S. J., Basford, P. J., Easton, N. H. C., Apetroaie-Cristea, M., Foster, G. L., Morris, A. K. R., Cox, S. J., and Loxham, M.: Long-term field comparison of multiple low-cost particulate matter sensors in an outdoor urban environment, Sci. Rep., 9, 7497, https://doi.org/10.1038/s41598-019-43716-3, 2019.

Butterfield, D., Martin, N. A., Coppin, G., and Fryer, D. E.: Equivalence of UK nitrogen dioxide diffusion tube data to the EU reference method, Atmos. Environ., 262, 118614, https://doi.org/10.1016/j.atmosenv.2021.118614, 2021.

Carslaw, D. C., Beevers, S. D., and Tate, J. E.: Modelling and assessing trends in traffic-related emissions using a generalised additive modelling approach, Atmos. Environ., 41, 5289–5299, https://doi.org/10.1016/j.atmosenv.2007.02.032, 2007.

CEN: CEN/TS 17660-1 Air quality - Performance evaluation of air quality sensor systems - Part 1: Gaseous pollutants in ambient air, 2021.

CEN EN 12341: Ambient air - Standard gravimetric measurement method for the determination of the PM10 or PM2,5 mass concentration of suspended particulate matter, 2023.

Chojer, H., Branco, P. T. B. S., Martins, F. G., Alvim-Ferraz, M. C. M., and Sousa, S. I. V.: Development of low-cost indoor air quality monitoring devices: Recent advancements, Sci. Total Environ., 727, 138385, https://doi.org/10.1016/j.scitotenv.2020.138385, 2020.

Crilley, L. R., Shaw, M., Pound, R., Kramer, L. J., Price, R., Young, S., Lewis, A. C., and Pope, F. D.: Evaluation of a low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring, Atmospheric Meas. Tech., 11, 709–720, https://doi.org/10.5194/amt-11-709-2018, 2018.

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

DEFRA: Technical Guidance (TG22). Local Air Quality Management, 2022.

Diez, S., Lacy, S. E., Bannan, T. J., Flynn, M., Gardiner, T., Harrison, D., Marsden, N., Martin, N. A., Read, K., and Edwards, P. M.: Air pollution measurement errors: is your data fit for purpose?, Atmospheric Meas. Tech., 15, 4091–4105, https://doi.org/10.5194/amt-15-4091-2022, 2022.

Diez, S., Lacy, S., Read, K., Pete, E., and Josefina, U.: QUANT: A Three-Year, Multi-City Air Quality Dataset of Commercial Air Sensors and Reference Data for Performance Evaluation, https://doi.org/10.5281/zenodo.10775692, 2024.

Duvall, R. M., Clements, A. L., Hagler, G., Kamal, A., Kilaru, V., Goodman, L., Frederick, S., Barkjohn, K. K., Greene, D., and Dye, T.: Performance Testing Protocols, Metrics, and Target Values for Fine Particulate Matter Air Sensors, 2021.

Farquhar, A. K., Henshaw, G. S., and Williams, D. E.: Understanding and Correcting Unwanted Influences on the Signal from Electrochemical Gas Sensors, ACS Sens., 6, 1295–1304, https://doi.org/10.1021/acssensors.0c02589, 2021.

Feenstra, B., Papapostolou, V., Hasheminassab, S., Zhang, H., Boghossian, B. D., Cocker, D., and Polidori, A.: Performance evaluation of twelve low-cost PM2.5 sensors at an ambient air monitoring site, Atmos. Environ., 216, 116946, https://doi.org/10.1016/j.atmosenv.2019.116946, 2019.

Feinberg, S., Williams, R., Hagler, G. S. W., Rickard, J., Brown, R., Garver, D., Harshfield, G., Stauffer, P., Mattson, E., Judge, R., and Garvey, S.: Long-term evaluation of air sensor technology under ambient conditions in Denver, Colorado, Atmospheric Meas. Tech., 11, 4605–4615, https://doi.org/10.5194/amt-11-4605-2018, 2018.

Gamboa, V. S., Kinast, É. J., and Pires, M.: System for performance evaluation and calibration of low-cost gas sensors applied to air quality monitoring, Atmospheric Pollut. Res., 14, 101645, https://doi.org/10.1016/j.apr.2022.101645, 2023.

Giordano, M. R., Malings, C., Pandis, S. N., Presto, A. A., McNeill, V. F., Westervelt, D. M., Beekmann, M., and Subramanian, R.: From low-cost sensors to high-quality data: A summary of challenges and best practices for effectively calibrating low-cost particulate matter mass sensors, J. Aerosol Sci., 158, 105833, https://doi.org/10.1016/j.jaerosci.2021.105833, 2021.

Grange, S. K. and Carslaw, D. C.: Using meteorological normalisation to detect interventions in air quality time series, Sci. Total Environ., 653, 578–588, https://doi.org/10.1016/j.scitotenv.2018.10.344, 2019.

Guimarães, U. S., Narvaes, I. da S., Galo, M. de L. B. T., da Silva, A. de Q., and Camargo, P. de O.:

Radargrammetric approaches to the flat relief of the amazon coast using COSMO-SkyMed and TerraSAR-X datasets, ISPRS J. Photogramm. Remote Sens., 145, 284–296, https://doi.org/10.1016/j.isprsjprs.2018.09.001, 2018.

Hagan, D. H., Gani, S., Bhandari, S., Patel, K., Habib, G., Apte, J. S., Hildebrandt Ruiz, L., and Kroll, J. H.: Inferring Aerosol Sources from Low-Cost Air Quality Sensor Measurements: A Case Study in Delhi, India, Environ. Sci. Technol. Lett., 6, 467–472, https://doi.org/10.1021/acs.estlett.9b00393, 2019.

Han, J., Liu, X., Jiang, M., Wang, Z., and Xu, M.: A novel light scattering method with size analysis and correction for on-line measurement of particulate matter concentration, J. Hazard. Mater., 401, 123721, https://doi.org/10.1016/j.jhazmat.2020.123721, 2021.

Hofman, J., Nikolaou, M., Shantharam, S. P., Stroobants, C., Weijs, S., and La Manna, V. P.: Distant calibration of low-cost PM and NO2 sensors; evidence from multiple sensor testbeds, Atmospheric Pollut. Res., 13, 101246, https://doi.org/10.1016/j.apr.2021.101246, 2022.

JCGM: The international vocabulary of metrology—basic and general concepts and associated terms (VIM), 3rd edn. JCGM 200:2012, 2012.

Jolliff, J. K., Kindle, J. C., Shulman, I., Penta, B., Friedrichs, M. A. M., Helber, R., and Arnone, R. A.: Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, J. Mar. Syst., 76, 64–82, https://doi.org/10.1016/j.jmarsys.2008.05.014, 2009.

Kang, Y., Aye, L., Ngo, T. D., and Zhou, J.: Performance evaluation of low-cost air quality sensors: A review, Sci. Total Environ., 818, 151769, https://doi.org/10.1016/j.scitotenv.2021.151769, 2022.

Karagulian, F., Barbiere, M., Kotsev, A., Spinelle, L., Gerboles, M., Lagler, F., Redon, N., Crunaire, S., and Borowiak, A.: Review of the Performance of Low-Cost Sensors for Air Quality Monitoring, Atmosphere, 10, 506, https://doi.org/10.3390/atmos10090506, 2019.

Kelly, K. E., Whitaker, J., Petty, A., Widmer, C., Dybwad, A., Sleeth, D., Martin, R., and Butterfield, A.: Ambient and laboratory evaluation of a low-cost particulate matter sensor, Environ. Pollut., 221, 491–500, https://doi.org/10.1016/j.envpol.2016.12.039, 2017.

Kim, H., Müller, M., Henne, S., and Hüglin, C.: Long-term behavior and stability of calibration models for NO and NO2 low-cost sensors, Atmospheric Meas. Tech., 15, 2979–2992, https://doi.org/10.5194/amt-15-2979-2022, 2022.

Kim, J., Shusterman, A. A., Lieschke, K. J., Newman, C., and Cohen, R. C.: The BErkeley Atmospheric CO2 Observation Network: field calibration and evaluation of low-cost air quality sensors, Atmospheric Meas. Tech., 11, 1937–1946, https://doi.org/10.5194/amt-11-1937-2018, 2018.

Lacy, S., Diez, S., and Edwards, P.: Quantification of Utility of Atmospheric Network Technologies:

(QUANT): Low-cost air quality measurements from 52 commercial devices at three UK urban monitoring sites., 2023.

Levy Zamora, M., Buehler, C., Lei, H., Datta, A., Xiong, F., Gentner, D. R., and Koehler, K.: Evaluating the Performance of Using Low-Cost Sensors to Calibrate for Cross-Sensitivities in a Multipollutant Network, ACS EST Eng., 2, 780–793, https://doi.org/10.1021/acsestengg.1c00367, 2022.

Lewis, A. and Edwards, P.: Validate personal air-pollution sensors, Nat. News, 535, 29, https://doi.org/10.1038/535029a, 2016.

Li, J., Hauryliuk, A., Malings, C., Eilenberg, S. R., Subramanian, R., and Presto, A. A.: Characterizing the Aging of Alphasense NO2 Sensors in Long-Term Field Deployments, ACS Sens., 6, 2952–2959, https://doi.org/10.1021/acssensors.1c00729, 2021.

Liang, L.: Calibrating low-cost sensors for ambient air monitoring: Techniques, trends, and challenges, Environ. Res., 197, 111163, https://doi.org/10.1016/j.envres.2021.111163, 2021.

Liang, L. and Daniels, J.: What Influences Low-cost Sensor Data Calibration? - A Systematic Assessment of Algorithms, Duration, and Predictor Selection, Aerosol Air Qual. Res., 22, 220076, https://doi.org/10.4209/aaqr.220076, 2022.

Liu, X., Jayaratne, R., Thai, P., Kuhn, T., Zing, I., Christensen, B., Lamont, R., Dunbabin, M., Zhu, S., Gao, J., Wainwright, D., Neale, D., Kan, R., Kirkwood, J., and Morawska, L.: Low-cost sensors as an alternative for long-term air quality monitoring, Environ. Res., 185, 109438, https://doi.org/10.1016/j.envres.2020.109438, 2020.

Long, R. W., Whitehill, A., Habel, A., Urbanski, S., Halliday, H., Colón, M., Kaushik, S., and Landis, M. S.: Comparison of ozone measurement methods in biomass burning smoke: an evaluation under field and laboratory conditions, Atmospheric Meas. Tech., 14, 1783–1800, https://doi.org/10.5194/amt-14-1783-2021, 2021.

Malings, C., Tanzer, R., Hauryliuk, A., Saha, P. K., Robinson, A. L., Presto, A. A., and Subramanian, R.: Fine particle mass monitoring with low-cost sensors: Corrections and long-term performance evaluation, Aerosol Sci. Technol., 54, 160–174, https://doi.org/10.1080/02786826.2019.1623863, 2020.

Molina Rueda, E., Carter, E., L’Orange, C., Quinn, C., and Volckens, J.: Size-Resolved Field Performance of Low-Cost Sensors for Particulate Matter Air Pollution, Environ. Sci. Technol. Lett., 10, 247–253, https://doi.org/10.1021/acs.estlett.3c00030, 2023.

Moreno-Rangel, A., Sharpe, T., Musau, F., and McGill, G.: Field evaluation of a low-cost indoor air quality monitor to quantify exposure to pollutants in residential environments, J. Sens. Sens. Syst., 7, 373–388, https://doi.org/10.5194/jsss-7-373-2018, 2018.

Nazemi, H., Joseph, A., Park, J., and Emadi, A.: Advanced Micro- and Nano-Gas Sensor Technology: A Review, Sensors, 19, 1285, https://doi.org/10.3390/s19061285, 2019.

Nowack, P., Konstantinovskiy, L., Gardiner, H., and Cant, J.: Machine learning calibration of low-cost NO2 and PM10 sensors: non-linear algorithms and their impact on site transferability, Atmospheric Meas. Tech., 14, 5637–5655, https://doi.org/10.5194/amt-14-5637-2021, 2021.

Okure, D., Ssematimba, J., Sserunjogi, R., Gracia, N. L., Soppelsa, M. E., and Bainomugisha, E.: Characterization of Ambient Air Quality in Selected Urban Areas in Uganda Using Low-Cost Sensing and Measurement Technologies, Environ. Sci. Technol., 56, 3324–3339, https://doi.org/10.1021/acs.est.1c01443, 2022.

Ouyang, B.: First-Principles Algorithm for Air Quality Electrochemical Gas Sensors, ACS Sens., 5, 2742–2746, https://doi.org/10.1021/acssensors.0c01129, 2020.

Pang, X., Shaw, M. D., Gillot, S., and Lewis, A. C.: The impacts of water vapour and co-pollutants on the performance of electrochemical gas sensors used for air quality monitoring, Sens. Actuators B Chem., 266, 674–684, https://doi.org/10.1016/j.snb.2018.03.144, 2018.

Pang, X., Chen, L., Shi, K., Wu, F., Chen, J., Fang, S., Wang, J., and Xu, M.: A lightweight low-cost and multipollutant sensor package for aerial observations of air pollutants in atmospheric boundary layer, Sci. Total Environ., 764, 142828, https://doi.org/10.1016/j.scitotenv.2020.142828, 2021.

PAS 4023: Selection, deployment, and quality control of low-cost air quality sensor systems in outdoor ambient air – Code of practice, 2023.

Pinder, R. W., Klopp, J. M., Kleiman, G., Hagler, G. S. W., Awe, Y., and Terry, S.: Opportunities and challenges for filling the air quality data gap in low- and middle-income countries, Atmos. Environ., 215, 116794, https://doi.org/10.1016/j.atmosenv.2019.06.032, 2019.

Raheja, G., Sabi, K., Sonla, H., Gbedjangni, E. K., McFarlane, C. M., Hodoli, C. G., and Westervelt, D. M.: A Network of Field-Calibrated Low-Cost Sensor Measurements of PM2.5 in Lomé, Togo, Over One to Two Years, ACS Earth Space Chem., 6, 1011–1021, https://doi.org/10.1021/acsearthspacechem.1c00391, 2022.

Rai, A. C., Kumar, P., Pilla, F., Skouloudis, A. N., Di Sabatino, S., Ratti, C., Yasar, A., and Rickerby, D.: End-user perspective of low-cost sensors for outdoor air pollution monitoring, Sci. Total Environ., 607–608, 691–705, https://doi.org/10.1016/j.scitotenv.2017.06.266, 2017.

Ripoll, A., Viana, M., Padrosa, M., Querol, X., Minutolo, A., Hou, K. M., Barcelo-Ordinas, J. M., and Garcia-Vidal, J.: Testing the performance of sensors for ozone pollution monitoring in a citizen science approach, Sci. Total Environ., 651, 1166–1179, https://doi.org/10.1016/j.scitotenv.2018.09.257, 2019.

Ropkins, K., Walker, A., Philips, I., Rushton, C., Clark, T., and Tate, J.: Change Detection of Air Quality Time-Series Using the R Package Aqeval, https://doi.org/10.2139/ssrn.4267722, 4 November 2022.

Sayahi, T., Butterfield, A., and Kelly, K. E.: Long-term field evaluation of the Plantower PMS low-cost particulate matter sensors, Environ. Pollut., 245, 932–940, https://doi.org/10.1016/j.envpol.2018.11.065, 2019.

Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M., and Bonavitacola, F.: Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2, Sens. Actuators B Chem., 238, 706–715, https://doi.org/10.1016/j.snb.2016.07.036, 2017.

Tanzer-Gruener, R., Li, J., Eilenberg, S. R., Robinson, A. L., and Presto, A. A.: Impacts of Modifiable Factors on Ambient Air Pollution: A Case Study of COVID-19 Shutdowns, Environ. Sci. Technol. Lett., 7, 554–559, https://doi.org/10.1021/acs.estlett.0c00365, 2020.

Wang, Y., Li, J., Jing, H., Zhang, Q., Jiang, J., and Biswas, P.: Laboratory Evaluation and Calibration of Three Low-Cost Particle Sensors for Particulate Matter Measurement, Aerosol Sci. Technol., 49, 1063–1077, https://doi.org/10.1080/02786826.2015.1100710, 2015.

Williams, D. E.: Electrochemical sensors for environmental gas analysis, Curr. Opin. Electrochem., 22, 145–153, https://doi.org/10.1016/j.coelec.2020.06.006, 2020.

Wu, T. Y., Horender, S., Tancev, G., and Vasilatou, K.: Evaluation of aerosol-spectrometer based PM2.5 and PM10 mass concentration measurement using ambient-like model aerosols in the laboratory, Measurement, 201, 111761, https://doi.org/10.1016/j.measurement.2022.111761, 2022.

S1. Co-location sites

For the main QUANT deployment, 3 field sites were chosen: Manchester, London, and York, all providing extensive reference measurements across a range of chemical environments representative of UK urban atmospheres. On the other hand, only the Manchester site was used for the WPS colocation.

~~The Manchester Air Quality Supersite (MAQS, 53° 26' 39.2"N, 2° 12' 51.9"W) is one of the largest air quality research facilities in the UK, and also because it is located in the south of the city of Manchester (the second largest metropolitan area in the UK, with approx. 3.3 million inh.) in an urban background environment (avg. temp. in winter of about 4-5 °C and RH ~87 %, avg. temp. in summer around 16-17 °C and RH ~88 %. MAQS reference instrumentation details can be found in the Section S4. All the data provided by MAQS was 1-min time resolution.~~

The Manchester Air Quality Supersite (MAQS, 53° 26' 39.2"N, 2° 12' 51.9"W) stands as one of the largest air quality research facilities in the UK. Situated in an urban background setting approximately four kilometres south of Manchester city center — the UK's second-largest metropolitan area with around 3.3 million residents — MAQS benefits from a strategic location on the University of Manchester's Fallowfield Campus. This location is notably distanced from direct traffic emissions, surrounded by student accommodations, university administrative buildings, and sports facilities. The campus's vicinity to shops, bars, and restaurants introduces a range of human activities, including varying levels of foot traffic and associated vehicular movement. Additionally, the presence of these commercial and recreational spaces, alongside residential buildings, contributes to the area's ambient air quality through emissions from heating and cooking, among other sources. For a visual representation of MAQS's surroundings, please refer to Figure S1 (panel a). The site experiences an average winter temperature of approximately 4-5°C with relative humidity around 87%, and an average summer temperature of about 16-17°C with relative humidity near 88%. Detailed information on MAQS's reference instrumentation and the methodologies employed for air quality measurements can be found in section S2. Data from MAQS are provided with a 1-minute time resolution, facilitating a granular temporal analysis of air quality metrics.

~~The London Air Quality Supersite (LAQS) is an urban background monitoring site located at Honor Oak Park (51° 26' 58.9"N 0° 02' 14.6"W) in Greater London, the third biggest European urban conglomeration with approx. 14.8 million inh. (avg. temp. in winter ~5 °C and RH ~84 %, avg. temp. in summer ~17 °C and RH ~72 %). All gas data provided by LAQS is 1-min time resolution and 15-min for PM.~~

The London Air Quality Supersite (LAQS, 51° 26' 58.9"N 0° 02' 14.6"W) serves as an urban background monitoring site, nestled within Honor Oak Park in Greater London. Situated 9 km southeast of the city center of the third-largest European urban conglomeration, LAQS offers a unique window into the air quality challenges of an area inhabited by approximately 14.8 million people. Nestled within the serene King's College sports grounds, is surrounded by middle-class neighbourhoods, abundant parks, and green spaces. This tranquil setting, is distanced from major roads and pollution sources, provides a representative snapshot of the ambient air quality typical of residential London. LAQS's surroundings are marked by a low level of commercial activity, with local shops and restaurants contributing minimally to the area's overall noise and bustle. Figure S1 (panel b) offers an aerial view of LAQS, illustrating the overall urban layout. The area is characterised by a temperate climate, experiencing average winter temperatures of around 5°C with RH of approx. 84%, and milder summers with temperatures averaging 17°C and RH of around 72%. Gas measurements at LAQS are conducted with a 1-minute time resolution, while PM data are collected at a 15-minute resolution (see section S2 for more details).

~~The York Fishergate roadside site (YoFi), located in the city of York (~210,000 inh., avg. temp. in winter of ~4°C and RH ~87 %, avg. temp. in summer around 15 °C and RH ~80 %). This site is a self-contained air quality monitoring station located very close to the city centre on a traffic island (53° 57' 06.9"N 1° 04' 33.1"W) surrounded by a residential/commercial area. This site was chosen to evaluate the LCS responses to a greater pollutant variability typical of traffic-related sites (in contrast with urban background monitoring stations as in the case of MAQS and LAQS). While PM and NOx data from YoFi are 1-hr time resolution, the O3 data is 1-min (deployed on the 15th of May 2020, specifically as part of the QUANT study).~~

The York Fishergate roadside site (YoFi, 53° 57' 06.9"N, 1° 04' 33.1"W), in the historic city of York, which is home to approximately 210,000 inhabitants (avg. temp. in winter of ~4°C and RH ~87 %, avg. temp. in summer around 15 °C and RH ~80 %). Situated just about 1 km from the city center on a traffic island, YoFi stands amidst a predominantly residential area that also encompasses commercial and light industrial elements. Unique to its location, the site is sandwiched between two lanes of Fishergate Road, a major avenue that bifurcates to facilitate traffic flow into and out of the city's southern part. Directly across from YoFi, a primary school adds to the daily human activity around the site, while the nearby River Ouse, located merely 300 metres to the west, contributes to the area's environmental characteristics. A vibrant commercial zone, featuring pubs and restaurants, is found just 100 metres to the north. Moreover, the site is flanked by Walmgate Stray, an expanse of recreational fields, located about 300 metres to the southeast, offering a green respite amidst the urban setting. Additional details can be visualised in Figure S1 (panel c), providing an aerial perspective of the site's key features and its urban context. This self-contained air quality monitoring station was specifically selected for the QUANT study to assess sensors' responses to the greater pollutant variability typical of traffic-related sites, contrasting with the urban background settings of MAQS and LAQS. YoFi provides data on PM and NOx with a 1-hour time resolution. Additionally, in a targeted effort to enhance our understanding of air quality dynamics, O_3 measurements (deployed on the 15th of May 2020, specifically as part of the QUANT study), utilising a 1-minute time resolution to offer detailed insights into temporal variations (refer to section S2 for more details).

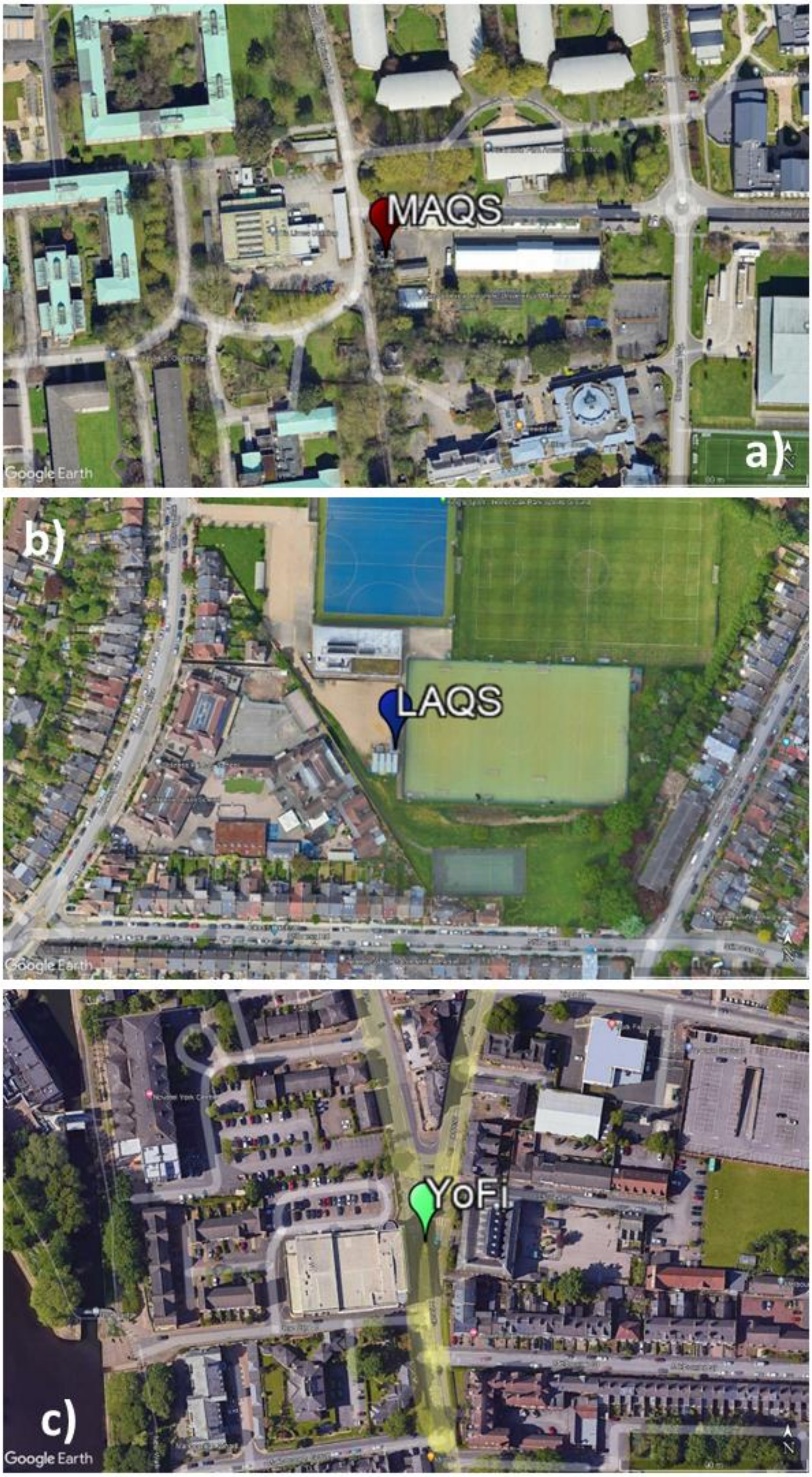

MAQS
a)
Google Earth
b)
LAQS
Google Earth
YoFi
c)
Google Earth

Figure S1: Aerial views of the air quality monitoring sites: a) MAQS, b) LAQS, and c) YoFi, captured from Google Earth. These images illustrate the diverse urban settings of each site, emphasising aspects such as their proximity to traffic sources, presence of green spaces, and the general urban layout. Image credits: Google Earth.

S2. Reference instrumentation, QA/QC, and data-sharing periods

Table S1 summarises the reference instrumentation at each site, Table S2 describes some of the QA/QC processes at the supersites, and Table S3 shows the data periods shared with the suppliers.

Table S1. Research grade instrumentation used for the QUANT study.

Analyte	Manchester	London	York
NO	Thermo 42i-y (Chem)	Teledyne T200U (Chem)	Teledyne T200UP (Chem)
NO_2	*Teledyne T500U (CAPS)	*Teledyne T500U (CAPS)	
O_3	*Thermo 49i (UV)	*Teledyne 400E (UV)	*2B 205 (UV)
PM	*Palas FIDAS200 (OAS)	*Palas FIDAS200 (OAS)	*Met One BAM 1020 (BA)

*Equivalent to reference (as defined in the European Air Quality Directive 2008/50/EC)

Acronyms: Chem: Chemiluminescence; CAPS: Cavity Attenuated Phase Shift Spectroscopy; UV: Ultraviolet; OAS: Optical aerosol spectrometer; BA: Beta attenuation.

Table S2. Summary of Quality Assurance processes in MAQS and LAQS

Instrument	Frequency	*Process
NO_y	At least monthly	Zero and span checks using standard cylinder and scrubber. Corrections to zero and span values.
NO_2	Daily	Automatic zero and span checks using internal NO_2 diffusion tube and scrubber. Zero corrections, span monitored.
O_3	Daily	Automatic zero and span checks using internal O_3 lamp and scrubber. Corrections to zero, span monitored.
CO	Every three hours & monthly	Zero checks every three hours and span checks monthly using onsite cylinder. Adjustments to zero and span values.
CO_2 and CH_4	Regular	Stability checks using onsite cylinder, no corrections made.
*PM	Semiannual	Sizing response verified with Mono dust, flow rate checked with Gilibrator.

*Checked with external standards by NPL every 6 months. These external standards are also used to provide a certification of the on-site standard cylinders. Final corrections to the data are provided by using the audit data to define the concentration of the on-site standards, with zero and span values interpolated between the calibration points.

**Sizing and flow checked every 6-month NPL audit process.1

Table S3. Reference data is shared with the sensor manufacturers.

QUANT main study			Wider Participation Study		
Reference dataset	**Period**	**Released**	**Reference dataset**	**Period**	**Released**
1	10-12-2019 - 17-02-2020	15-04-2020	**1**	17-06-2021 - 16-07-2021	23-07-2021
2	18-02-2020 - 17-08-2020	27-10-2020	**2**	01-12-2021 - 31-12-2021	26-01-2022
3	18-08-2020 - 17-02-2021	15-04-2021	**3**	01-05-2022 - 31-05-2022	15-06-2022

S3. QUANT main study devices

In this section, a brief description of the QUANT main study systems' components is offered.

PurpleAir (PA) (https://www2.purpleair.com) devices (PA-II-SD model, firmware v4.11) reports particulate matter (PM_1, $PM_{2.5}$, and PM_{10}), and it was chosen for its penetration around the world. Two identical Plantower PMS5003 (Plantower) sensors (channels A and B) are found in each PA. It offers two data products (2-min avg. time): the "cf_atm" (for outdoor applications) and the "cf_1" (for indoor or controlled environment applications). The PMS behaves like a nephelometer rather than an optical particle counter to measure the light scattered by the PM (Ouimette et al., 2022) and is composed of a laser, a photodiode, a fan, and a microprocessor control unit. They also measure temperature (Temp), relative humidity (RH), and atmospheric pressure (Pres) (Bosch). The data can be communicated via Wi-Fi or stored locally (microSD card), which was the preferred way during the colocation. No calibrated products are offered by the company.

*Note: For this study, only Channel A and the data product "cf_atm" were included in the analysis and shown in the plots.

AQMesh (https://www.aqmesh.com) reports NO_2, NO, O_3 using electrochemical (EC) sensors (Alphasense), CO_2 with a non-dispersive infrared sensor (NDIR, Alphasense), PM_1, $PM_{2.5}$, and PM_{10} through a light-scattering sensor (Nephelometer, Environmental Instr.) with 1-minute time resolution (algorithm v5.1 for gases and v3.0 for PM). This instrument also registers Temp, RH, and Pres (Solid-State sensors) (Zauli-Sajani et al., 2022) and the sampling mechanism employs a pump. The collected data is sent to the company server via a cellular network and post-processed (Temp, RH, and cross-interference correction) in the cloud by a proprietary algorithm. Finally, the data is released to the final user via secure web login or through its Application Programming Interface (API). Although the first 4 months of the deployment the data had a 15-min resolution, since then the provided resolution is 1-min average.

AQY (v.1.0) is also a multi-species device (https://www.aeroqual.com) and measures O_3, NO_2, $PM_{2.5}$, PM_{10}, Temp, and RH. This is the only device system that does not use Alphasense sensors for gases. While O_3 is quantified using a metal oxide sensor (WO_3-based, Aeroqual Ltd), the NO_2 is measured by an EC sensor (Membrapore type O_3/M5, Aeroqual Ltd) (Weissert et al., 2019). For PM it uses a light scattering method (Nova) to convert size and particle count to a mass fraction and behaves like a nephelometer (Myklebust et al., 2022). These LCS devices send their data (1-min time resolution) to the Aeroqual server via cellular (WiFi could also be used for this purpose) or stored locally (microSD card). The non-local data access is through a web portal or via API.

Zephyr units (https://www.earthsense.co.uk) measure PM (Nephelometer, Plantower), Temp & RH (Sensirion), and Press (Bosch) (the sample uptake uses a fan). As most of the commercial units tested here, it used Alphasense EC sensors (the “A series”, a smaller version than the B series) for gases (NO, NO_2, and O_3). These devices send their raw data to the server via a cellular network, where they pre-process the raw signals. We have secure access to the measurements with a time resolution of 1-min per species through the website or via its API.

ARIsense v200 devices (https://quant-aq.com) measure NO, NO_2, O_3, CO (EC, Alphasense), CO_2 (NDIR, Alphasense), Temp & RH (Sensirion), and Press (Bosch) (Cross et al., 2017). Of all the devices tested, this is the only one that uses an Optical Particle Counter (OPC) for PM (Particles Plus). Communication is carried out through a cellular network and the data products are accessed through a web portal or API (1-minute time resolution). According to the company policy, only the gas data products are subjected to calibrations (if colocation data is available).

Table S4. Summary of sensor measurements and the time resolution data provided by participating companies in the Main QUANT study.

System	Measurands	Time Resol.
PA	PM_1, $PM_{2.5}$, PM_{10}	2min
AQM	PM_1, $PM_{2.5}$, PM_{10}, NO, NO_2, O_3, CO_2	1min/15min
AQY	$PM_{2.5}$, PM_{10}, NO_2, O_3	1min
Zep	PM_1, $PM_{2.5}$, PM_{10}, NO, NO_2, O_3	1min
Ari	PM_1, $PM_{2.5}$, PM_{10}, NO, NO_2, O_3, CO; CO_2	1min

Figure S2~~1~~. Data product for each of the participating companies during Main QUANT. The top panels are for NO_2, the middle panels for O_3 and the bottom panels for $PM_{2.5}$. The y-axis represents the different products: "out-of-box", cal1 and cal2. The x-axis shows the dates for which each company provided the mentioned products.

S4. WPS devices

A short description of the WPS devices' components is shown in this section

<u>Modulair-PM</u> instruments (https://quant-aq.com) employ two different techniques to obtain PM mass concentration (it samples the air using a fan), an OPC (Alphasense, OPC-N3) and a nephelometer (Plantower, PMS5003). This system provides 1-min time resolution data for PM_1, $PM_{2.5}$, and PM_{10}, plus size-resolved particle number concentration (range 350 nm to 40 µm) (Meyer et al., 2022; Westgate and Ng, 2022). Temp, RH, and Press are also measured, but no data was found about the sensing elements it uses. The post-processed data can be accessed locally (microSD card) or through its server (cellular network comm) via its web portal or API.

<u>AQMesh</u> (see earlier description).

The <u>Atmos</u> device (http://urbansciences.in/) reports PM_1, $PM_{2.5}$, PM_{10} (Plantower, PMS7003) plus Temp and RH (Adafruit), employing a fan as a means to sample the air. The system transmits the data (1-min time resolution) to a cloud server (only via Wi-Fi) and also stores it locally (Puttaswamy et al., 2022). The data can be accessed via a web dashboard or API. Unfortunately, and due to the meteorological conditions at the Manchester supersite these co-located devices only survived for about 2 months.

The <u>IMB</u> instrument (https://www.bosch-mobility-solutions.com) measures NO_2, O_3 $PM_{2.5}$ and PM_{10}, (Alphasense sensors), plus Press, RH an Temp (no details were found about the brand and model). The raw data is transmitted to their cloud using cellular connectivity (3G or LTE). The final data is 1-min resolution (accessed only via API).

Polludrone (https://oizom.com) uses Alphasense sensors for gas measurements (B4 series for NO, NO_2, O_3. No data available about CO, CO_2 and SO_2) and a Wuhan Cubic PM3006S for PM ($PM_{2.5}$ and PM_{10}) (Oizom - Polludrone Smart, 2023). It also registers RH and Temp, but no data was found in regards to sensor model/brand. The sampling mechanism uses a fan and data transmission is wireless. The final product (time res is 10-min) can be obtained through the Oizom webpage and/or via API.

Kunak Air Pro (https://www.kunak.es/) uses a fan for sampling and all sensors are from Alphasense (EC, B series for CO, NO, NO_2 and O_3; an NDIR sensor for CO_2; and an OPC-N3 for PM_1, $PM_{2.5}$, and PM_{10}) (Hofman et al., 2022). It also provides Temp, RH, and Press (no data was found in regards to environmental sensor model/brand). The raw data is transmitted via a multi-band network, and the final data (time res is 5-min) can be accessed through their website or via API.

The Silax Air (https://vortexiot.com) system measures NO_2, O_3, PM_{10} and $PM_{2.5}$. Their webpage mentions that for PM an optical scattering sensor is used and EC sensors for the gases. Further details weren't found. The raw data is transmitted via 4G or WiFi and the final user accesses the final product (5-min time res) through API or website.

The Node-S system (https://www.clarity.io) holds a nephelometer (Plantower PMS6003) to measure 3 PM size cuts (PM_1, $PM_{2.5}$, PM_{10}) (Liu et al., 2022) and EC sensors for NO_2 (Alphasense) (Miech et al., 2021). The air is dragged into the system by a fan and a Bosch sensor is used for press, RH, and temp. The data is communicated to Clarity's cloud via cellular signal (4G) and the final product is ~3-min time res (something unusual for sensor systems). Access to the final data is via the web portal or through API.

Praxis/Urban (https://www.southcoastscience.com) system employs EC sensors for NO, NO_2, O_3 (Alphasense, A series), an NDIR for CO_2 (Alphasense), and particle counter (Alphasense, OPC-N3) for PM_1, PM_{10} and $PM_{2.5}$. The Temp/RH is Sensirion and the Press sensor is TDK. The raw data is communicated to the company server using 4G and the user can access it and post-processed data through an API (1-min time res).

Table S5. Summary of sensor measurements and the time resolution data provided by participating companies in the WPS study.

System	Measurands	Time Resol.
Mod	PM_1, $PM_{2.5}$, PM_{10}	1min
AQM	PM_1, $PM_{2.5}$, PM_{10}, NO, NO_2, O_3, CO; CO_2	15min
Atm	PM_1, $PM_{2.5}$, PM_{10}	2min
IMB	PM_1, $PM_{2.5}$, PM_{10}, NO_2, O_3	1min
Poll	PM_1, $PM_{2.5}$, PM_{10}, NO, NO_2, O_3	10min
AP	PM_1, $PM_{2.5}$, PM_{10}, NO, NO_2, O_3, CO; CO_2	5min

SA	PM_1, $PM_{2.5}$, PM_{10}, NO_2, O_3	5min
NS	PM_1, $PM_{2.5}$, PM_{10}, NO_2	~5min
Prax	PM_1, $PM_{2.5}$, PM_{10}, NO, NO_2, O_3, CO; CO_2	1min

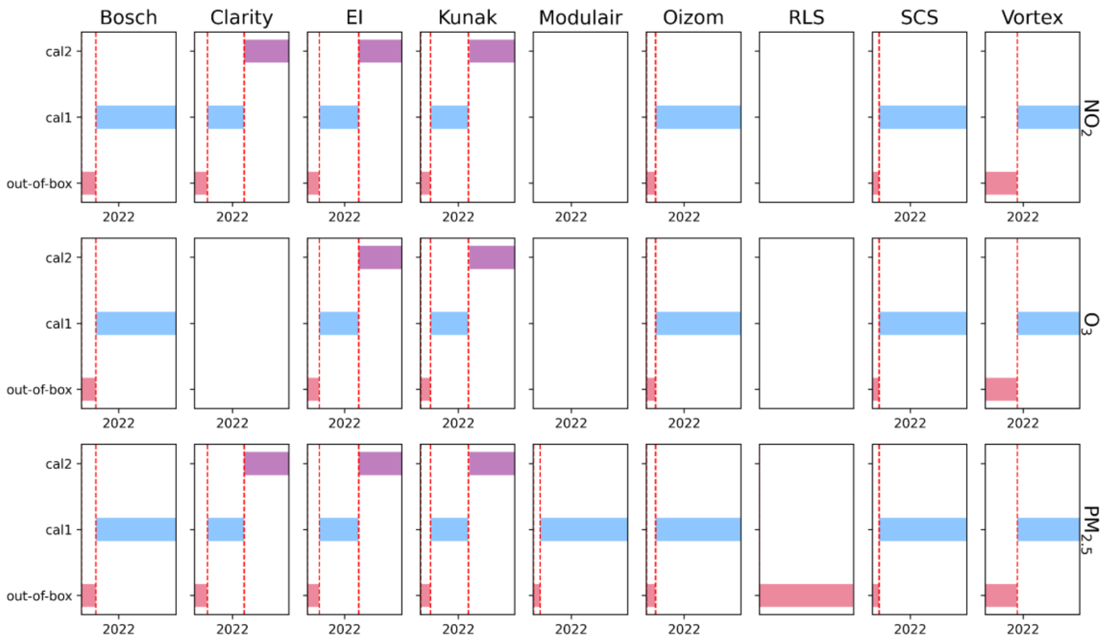


Figure S32. Data product for each of the participating companies in the WPS. The top panels are for NO_2, the middle panels for O_3 and the bottom panels for $PM_{2.5}$. The y-axis represents the different products: "out-of-box", cal1 and cal2. The x-axis shows the dates for which each company provided the mentioned products.

S5. Performance Metrics

In the assessment of sensor measurement error, it is standard practice to employ a linear additive model, described by the following equation:

$$y_i = b_1 x_i + b_0 + \varepsilon_i \quad (1)$$

In this model, the dependent variable "y" represents the sensor measurements, while the independent variable "x" denotes the reference measurements. The coefficient b_1 corresponds to the slope of the regression line (the response sensitivity of the sensor relative to the reference) and b_0 is the ordinate at the origin (the sensor's output when the reference measurement is zero). ε_i, assumed to have a mean of zero and a standard deviation of σ_ε, captures the portion of "y" that cannot be explained by "x". For a sensor to perfectly match the reference measurements (i.e., y = x), b_1 would equal one, with both b_0 and ε_i being zero.

Coefficient of Determination (R^2)

R^2 is an adimensional metric that quantifies the proportion of variance in the sensor measurements ("y") that can be explained by its linear relationship with the reference measurements ("x"):

$$R^2 = \frac{\sum_{i=1}^{n} (x_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y})^2} \tag{2}$$

As a bounded metric, R^2 varies between zero and one ($0 \leq R^2 \leq 1$), where a value closer to one indicates a stronger linear association between the sensor and reference data. Despite being one of the most widely used metrics in sensor evaluation, as highlighted by Karagulian et al. (2019), R^2 comes with limitations that warrant careful consideration. Notably, R^2 does not account for bias in the data; a regression line diverging from the ideal 1:1 relationship between "x" and "y" does not affect its value. Additionally, R^2 is influenced by the dynamic range of the measurements, which can skew its interpretation. Given these nuances, it is prudent to report R^2 alongside complementary metrics that can offer a more rounded view of sensor performance. For a more in-depth analysis of the limitations and proper use of R^2, readers are directed to the discussion in Legates and McCabe Jr. (1999).

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

MAE and RMSE (both dimensional metrics, expressed in the same units as the measured variable), also stand as very popular metrics for performance evaluation, as they offer insights into the accuracy of sensors, presenting a fuller picture than the R^2 alone. These metrics can be estimated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i| \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - x_i)^2} \tag{4}$$

However, both MAE and RMSE quantify average errors. MAE does so by calculating the average magnitude of errors without directionality, utilising absolute differences, while RMSE gauges the standard deviation of these differences, highlighting the squared differences between sensor readings and reference grade measurements. Although MAE and RMSE are both valued for their measure of accuracy, they bear distinct implications in practice. MAE treats all errors equally, allocating proportional weight across the board. Conversely, RMSE disproportionately penalises larger errors due to its squaring of difference values, an aspect noted by (Willmott and Matsuura, 2005). This characteristic makes RMSE particularly sensitive to outliers, shaping its utility in identifying and rectifying significant deviations.

Mean Bias Error (MBE)

The MBE quantifies the average bias in sensor measurements relative to reference values. Expressed in the same units as the variable being measured, MBE reflects the systematic error, offering a straightforward indication of a sensor's tendency to overestimate or underestimate the reference:

$$MBE = \frac{1}{n}\sum_{i=1}^{n} (y_i - x_i) \tag{5}$$

A zero value of MBE indicates no consistent over- or underestimation, while positive or negative values signal systematic bias in measurement. This simplicity in interpretation makes MBE particularly valuable for initial assessments of sensor accuracy and for guiding calibration efforts to correct for systematic bias. However, the MBE does not capture the precision of the measurements. For this reason, MBE is most effective when used in conjunction with other metrics, such as RMSE and MAE, to gain a comprehensive understanding of sensor performance, encompassing both systematic and random errors.

Relative Expanded Uncertainty (REU)

In contrast to single-value metrics such as R^2, RMSE, and MAE, which assess data sets as a whole, REU offers a "point by point" metric. This allows for graphical representations (like the REU in the concentration space or as a time series), offering detailed insights into measurement performance variability. The REU's mathematical framework is outlined in the "Guidance for the Demonstration of Equivalence of Ambient Air Monitoring Methods" (European Commission, 2010), as follows:

$$U(y_i) = \sqrt{\frac{RSS}{n-2} - u^2(x_i) + (y_i - b_0 - b_1 x_i)^2} \quad (6)$$

$$REU(y_i) = \frac{k.U(y_i)}{\hat{x}} \quad (7)$$

$$RSS = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \quad (8)$$

here, U(yi) represents the measurement uncertainty [concentration units]; REU(yi) denotes the REU [percentage]; u(xi) is the random uncertainty of the reference monitor [concentration units]; "n" stand for the number of collocated data points considered; RSS is the Residual Sum of Squares; k is the coverage factor (set at 2 for a 95% confidence level).

A distinctive feature of REU is its incorporation of the uncertainty associated with the reference method (i.e., $u(x_i)$). This aspect recognizes that all measurements, including those from reference methods, are subject to inherent uncertainties. While calculating REU is more complex than traditional metrics, it's essential to acknowledge that, like any metric, REU is based on specific assumptions and considerations. These factors must be thoughtfully evaluated when interpreting data to ensure that conclusions are firmly rooted in the context of the study.

Current guidance and normalisation efforts

Table S6 summarises the key metrics addressed in some of the most recent guidance documents and technical standards. These metrics have been categorised under various labels: linearity, bias, error, uncertainty, data coverage, and inter-sensor precision. Each of these guidelines and regulations has its own set of procedures, protocols, and thresholds. Therefore, it is advisable for readers to consult the original documents for a detailed understanding of these specificities.

Table S6. Summary of field evaluation metrics for sensors according to different guidelines and technical standards.

Feature	EPA[1&2]	CEN[3]	ASTM[4&5]

Pollutants covered	$PM_{2.5}$ & O_3	NO_2, O_3, CO, SO_2 & Bencene	$PM_{2.5}$, PM_{10} NO_2, O_3, CO & SO_2
Linearity	R^2	----	R^2
Bias	Slope	Slope	Slope
	Intercept	Intercept	Intercept
Error	----	----	MAE
	RMSE	----	RMSE
	NRMSE	----	NRMSE
Uncertainty	----	REU	----
Data coverage	Data completeness	Data Capture	Data Capture Rate
Inter–sensor precision	SD	$u_{(bs,s)}$	$S_{r,f}$
	CV	----	----

References in the table:

[1]EPA/600/R-20/279 Performance Testing Protocols, Metrics, and Target Values for Ozone Air Sensors.

[2]EPA/600/R-20/280 Performance Testing Protocols, Metrics, and Target Values for Fine Particulate Matter Air Sensors.

[3]CEN/TS 17660-1: Air quality - Performance evaluation of air quality sensor systems - Part 1 Gaseous pollutants in ambient air.

[4]ASTM D8406-22: Standard Practice for Performance Evaluation of Ambient Outdoor Air Quality Sensors and Sensor-based Instruments for Portable and Fixed-point Measurement.

[5]ASTM WK74812: Standard Specification for Ambient Outdoor Air Quality Sensors and Sensor-based Instruments for Portable and Fixed-Point Measurement.

Acronyms: EPA: U.S. Environmental Protection Agency; CEN: European Committee for Standardization; ASTM: American Society for Testing and Material. CV: Coefficient of Variation; SD: Standard Deviation (see the definition in the EPA Performance Testing Protocols); $u_{(bs,s)}$: Between sensor system uncertainty (see the definition in the CEN TS 17660-1); $S_{r,f}$: field reproducibility standard deviation (see the definition in the ASTM protocols).

S6. Complementary plots

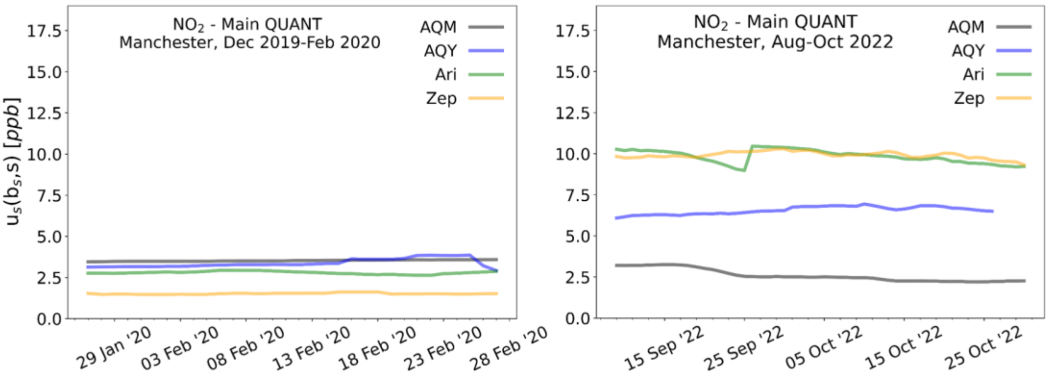


Figure S4. Inter-device precision of NO_2 measurements from "identical" devices across the 4 companies participating in QUANT is assessed using the "between sensor system uncertainty" metric (defined by the CEN/TS 17660-1:2021 as *u(bs, s)*). Each line represents this metric as a composite of all sensors per brand (excluding units with less than 75% data) within a 40-day sliding window.

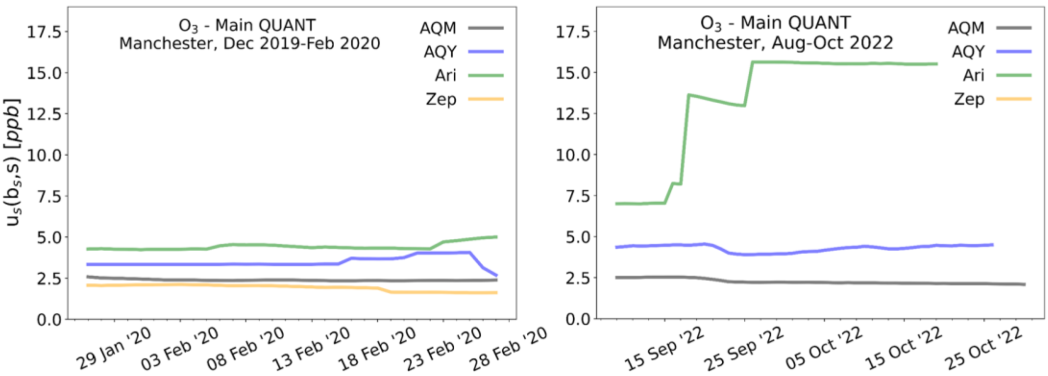


Figure S5. The inter-device precision of O_2 measurements from "identical" devices across the 4 companies participating in QUANT is assessed using the "between sensor system uncertainty" metric (defined by the CEN/TS 17660-1:2021 as *u(bs, s)*). Each line represents this metric as a composite of all sensors per brand (excluding units with less than 75% data) within a 40-day sliding window.

~~S5. Sensor performance estimated using different reference methods~~

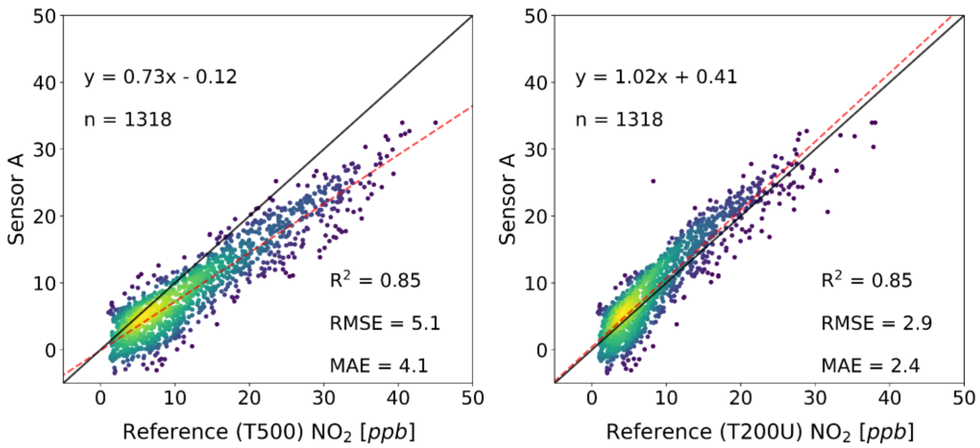


Figure S6~~3~~. Comparative analysis of "Sensor A" performance against two reference instruments for NO_2 measurements. The left plot shows the correlation with the Teledyne T500 (Cavity Attenuated Phase Shift Spectroscopy), while the right plot is against the Teledyne T200U (chemiluminescence) and specifically installed at the Manchester supersite for the QUANT study. The dashed red line represents the line of best fit for the sensor data against each reference, indicating a closer agreement with the T200U (slope=1.02) compared to the T500 (slope=0.73).

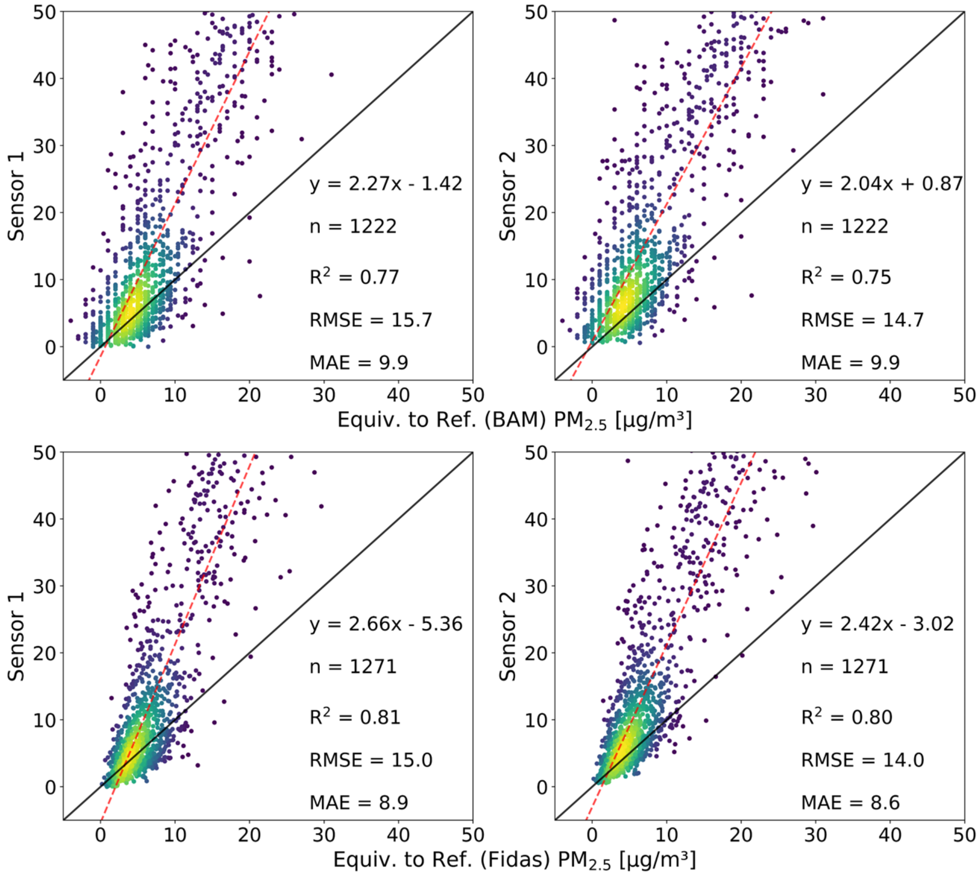


Figure S7~~4~~. Comparative regression analysis and performance metrics of two distinct $PM_{2.5}$ sensor systems benchmarked against a BAM for the top plots and a Fidas for the bottom plots. Each plot demonstrates the correlation and agreement between the sensor readings and the two equivalent-to-reference instruments in a roadside site located in York.

S7~~6~~. NO_2 Diffusion tubes

A diffusion tube co-location study was carried out between November 2020 and November 2021 at the MAQS, LAQS and York sites, using two types of diffusion tubes: the conventional (also known as LAQM, for Local Air Quality Management) and UUNN (for UK Urban NO2 Network). LAQM tubes have an open end and capture NO_2 which is converted to nitrite when reacting with triethanolamine (TEA) for subsequent analysis. On the other hand, UUNN tubes, similar in the sampling process to LAQM, include an amorphous polyethylene filter at the open end to further mitigate the effect of wind on NO_2 measurements. For more details refer to (Butterfield et al., 2021). Both types of tubes (conventional and UUNN) were installed in duplicates, either in shelters (to limit the incidence of wind) or directly exposed without protection in mounting blocks. Figure S5 illustrates the performance comparison of traditional diffusion tubes and a sensor system in Manchester. The data from these diffusion tubes have been used to correct the sensor shown here and explained in detail in Section 3.6 (Figures 9b and 9c).

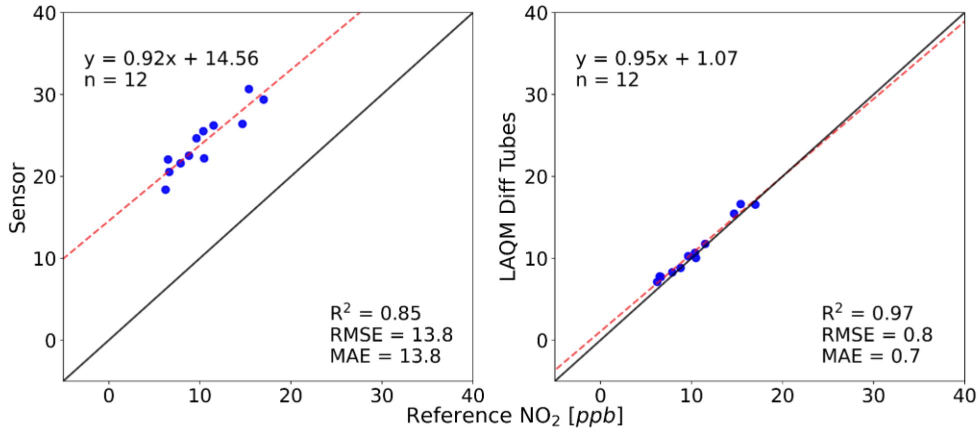


Figure S8~~5~~. The left plot displays the correlation between an air quality sensor's readings and those from a reference monitor for NO_2, while the right plot demonstrates the LAQM diffusion tube performance. The LAQM plot shows a tighter correlation with the 1:1 line, indicating a higher accuracy in measuring NO_2 concentrations for the period Nov 2020 - Nov 2021 at the Manchester supersite (blue dots represent monthly averages).

References

Butterfield, D., Martin, N. A., Coppin, G., and Fryer, D. E.: Equivalence of UK nitrogen dioxide diffusion tube data to the EU reference method, Atmos. Environ., 262, 118614, https://doi.org/10.1016/j.atmosenv.2021.118614, 2021.

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

European Commission: Guide to the demonstration of equivalence of ambient air monitoring methods, Report by an EC Working, Group on Guidance. European Commission, 2010.

Hofman, J., Peters, J., Stroobants, C., Elst, E., Baeyens, B., Van Laer, J., Spruyt, M., Van Essche, W., Delbare, E., Roels, B., Cochez, A., Gillijns, E., and Van Poppel, M.: Air Quality Sensor Networks for Evidence-Based Policy Making: Best Practices for Actionable Insights, Atmosphere, 13, 944, https://doi.org/10.3390/atmos13060944, 2022.

Karagulian, F., Barbiere, M., Kotsev, A., Spinelle, L., Gerboles, M., Lagler, F., Redon, N., Crunaire, S., and Borowiak, A.: Review of the Performance of Low-Cost Sensors for Air Quality Monitoring, Atmosphere, 10, 506, https://doi.org/10.3390/atmos10090506, 2019.

Legates, D. R. and McCabe Jr., G. J.: Evaluating the use of “goodness-of-fit” Measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, https://doi.org/10.1029/1998WR900018, 1999.

Liu, G., Moore, K., Su, W.-C., Delclos, G. L., Gimeno Ruiz de Porras, D., Yu, B., Tian, H., Luo, B., Lin, S., Lewis, G. T., Craft, E., and Zhang, K.: Chemical explosion, COVID-19, and environmental justice: Insights from low-cost air quality sensors, Sci. Total Environ., 849, 157881, https://doi.org/10.1016/j.scitotenv.2022.157881, 2022.

Meyer, M., Afshar-Mohajer, N., Cross, E., and Mudgett, P.: Feasibility of using Low-Cost COTS Sensors for Particulate Monitoring in Space Missions, 2022.

Miech, J. A., Stanton, L., Gao, M., Micalizzi, P., Uebelherr, J., Herckes, P., and Fraser, M. P.: Calibration of Low-Cost NO2 Sensors through Environmental Factor Correction, Toxics, 9, 281, https://doi.org/10.3390/toxics9110281, 2021.

Myklebust, H., Aarhaug, T. A., and Tranell, G.: Use of a Distributed Micro-sensor System for Monitoring the Indoor Particulate Matter Concentration in the Atmosphere of Ferroalloy Production Plants, JOM, 74, 4787–4797, https://doi.org/10.1007/s11837-022-05487-7, 2022.

Oizom - Polludrone Smart: http://www.aqmd.gov/aq-spec/sensordetail/oizom---polludrone-smart, last access: 27 January 2023.

Ouimette, J. R., Malm, W. C., Schichtel, B. A., Sheridan, P. J., Andrews, E., Ogren, J. A., and Arnott, W. P.: Evaluating the PurpleAir monitor as an aerosol light scattering instrument, Atmospheric Meas. Tech., 15, 655–676, https://doi.org/10.5194/amt-15-655-2022, 2022.

Puttaswamy, N., Sreekanth, V., Pillarisetti, A., Upadhya, A. R., Saidam, S., Veerappan, B., Mukhopadhyay, K., Sambandam, S., Sutaria, R., and Balakrishnan, K.: Indoor and Ambient Air Pollution in Chennai, India during COVID-19 Lockdown: An Affordable Sensors Study, Aerosol Air Qual. Res., 22, 210170, https://doi.org/10.4209/aaqr.210170, 2022.

Weissert, L. F., Alberti, K., Miskell, G., Pattinson, W., Salmond, J. A., Henshaw, G., and Williams, D. E.: Low-cost sensors and microscale land use regression: Data fusion to resolve air quality variations with high spatial and temporal resolution, Atmos. Environ., 213, 285–295, https://doi.org/10.1016/j.atmosenv.2019.06.019, 2019.

Westgate, S. and Ng, N. L.: Using in-situ CO2, PM1, PM2.5, and PM10 measurements to assess air change rates and indoor aerosol dynamics, Build. Environ., 224, 109559, https://doi.org/10.1016/j.buildenv.2022.109559, 2022.

Willmott, C. J. and Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Clim. Res., 30, 79–82, https://doi.org/10.3354/cr030079, 2005.

Zauli-Sajani, S., Marchesi, S., Boselli, G., Broglia, E., Angella, A., Maestri, E., Marmiroli, N., and Colacci, A.: Effectiveness of a Protocol to Reduce Children's Exposure to Particulate Matter and NO2 in Schools during Alert Days, Int. J. Environ. Res. Public. Health, 19, 11019, https://doi.org/10.3390/ijerph191711019, 2022.