General comments

This manuscript addresses a highly relevant and timely topic: the evaluation of low-cost sensor performance in real-world exposure settings, with a focus on their potential application in epidemiological studies. The proposed approach is valuable, and the work presents several strengths, including the use of a modular sensor platform, the incorporation of multiple performance metrics, and the attempt to link results to regulatory benchmarks. In this regard, the study has the potential to become a significant contribution to the literature on sensors and personal exposure assessment.

However, the manuscript still requires certain adjustments to reach its full impact. Some elements of the study design are not fully integrated into the discussion of the results, and certain sections would benefit from greater methodological clarity. Overall, the manuscript feels somewhat disorganized, with sections that could be restructured to enhance internal coherence, and the use of technical language could be refined to improve precision. Finally, some conceptual discussions are potentially valuable, but appear somewhat disconnected and would benefit from being more clearly linked to the study's main objectives and findings.

In summary, this is a manuscript with an original approach and strong potential, which could be significantly strengthened through a comprehensive revision aimed at improving its structure, clarity, and argumentative coherence.

Comments by section

Title

The current title partially reflects the rich content of the study. In particular, it would be important to highlight that the study does not exclusively address performance, but also includes the calibration process of the sensors, and that its main value lies within the framework of environmental epidemiology A suggestion could be "Calibration and performance evaluation of low-cost PM2.5 and NO2 air quality sensors for environmental epidemiology".

Abstract

The abstract effectively summarizes the main elements of the manuscript, but some rewording could improve its clarity and alignment with the content of the paper. For instance, the second sentence ("This study aimed to transcend the limitations...") may come across as somewhat broad and ambitious, and could benefit from a more specific description of the limitations being addressed. The final sentence of the abstract reads more as an aspirational statement than a conclusion directly supported by the findings. It may be more appropriate to include this type of reflection in the discussion section.

Introduction

The introduction correctly outlines the distinction between direct and indirect approaches to assessing exposure to air pollutants. However, when discussing the direct approach, the text focuses primarily on mobile personal measurements, which does not accurately represent the methodology used in this study.

Given that a substantial portion of the study focuses on evaluating sensor performance in indoor environments, it would be useful to include in the introduction a more explicit review of previous work in this area. For instance, studies that have assessed the performance of sensors

in domestic settings or proposed calibration/validation strategies specifically tailored to indoor conditions. Not including this indoor background limits the reader's ability to grasp from the outset how underdeveloped the literature is in this domain. It also reduces the opportunity to position this study as an important contribution to a still-emerging field, and to contrast the findings with relevant previous work.

One of the most valuable aspects of the manuscript is the use of regulatory metrics, such as the calculation of relative expanded uncertainty (REU) as defined in European directives, to assess sensor performance. However, these directives, and the related technical specifications, were not designed for personal monitoring. This represents an opportunity for the manuscript to elaborate more thoroughly on an important gap in the literature: the relevance, limitations, and potential of using regulatory standards as quality criteria in epidemiological studies. Addressing this point would strengthen the conceptual justification of the chosen approach and connect it to recent debates on the adaptation of regulatory frameworks to emerging technologies.

Methodology

Section 2.1 includes valuable elements of the study design, such as pulmonary function tests, health surveys. However, these elements are not developed or connected with the results presented. It would be useful to clarify whether these data fall outside the scope of this manuscript or if they will be analyzed in future work. Otherwise, the reader may perceive a disconnect between the protocol and the actual focus of the study.

The description of sensor deployment would benefit from more detailed information on installation conditions, such as the height above ground, orientation and distance from potential sources (e.g., windows, heaters, traffic), and the ventilation characteristics (passive or active?) of the protective enclosure.

The title "Study design and quality assurance" could be revised, as subsections 2.3.1 and 2.3.2 focus on quality control rather than experimental design.

The title of subsection 2.3.1, "Sensor calibration before deployment," also seems inappropriate (same with the title of Table 1), since the content primarily addresses the co-location experiments necessary for calibration but does not include the full calibration process, which is described in other sections (e.g., 2.3 Sensor correction models). Reorganizing these subsections could improve clarity and support reader comprehension.

Some passages in subsection 2.3.1 could be edited for clarity, as they currently lack precision. For example, statements like "The experiments in the particle chamber took about an hour and were repeated at least twice" or "the sensors were co-located in the laboratory for some days" are somewhat vague.

Since the selected co-location site (Hauptstätter Straße) is a known urban hotspot with elevated pollution levels, a brief methodological justification for this choice would be useful.

Throughout the manuscript, the term "reference instruments" (RIs) is used generically, but under European legislation there are two levels of "fixed measurements": reference instruments and equivalent to reference. It would be helpful to specify which category applies to each instrument used in this study.

Further clarification is also recommended in section 2.3.2: Which sensors were validated against which official stations? How far apart were they? How was agreement with the passive diffusion tubes quantified? What interpretive criteria were applied for indoor PM2.5 records? Table 2 indicates that the validation with diffusion tubes was quantitative, how was the quality of those measurements ensured?

The title of section “2.3 Sensor correction models” does not accurately reflect the content, as the methods described (MLR, SVR, RFR, ANN) are in fact calibration procedures. The manuscript uses the terms “correction” and “calibration” somewhat interchangeably. Since the study has a strong methodological focus on performance evaluation and the application of statistical calibration models, more precise terminology is recommended. I suggest referring to the International Vocabulary of Metrology (https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf). It would also be helpful to include a summary table that, for each pollutant and environment combination, specifies: the model used, the explanatory variables, the reference instrument type, and details on training, testing, and validation. What temporal resolution was used to train the models?

Section 2.4 indicates that the models were trained using 15-minute aggregated data, while section 2.5 reports performance evaluated at 1, 5, and 10 minutes. It is important to clarify whether models trained at 15-minute resolution were directly applied to data at other resolutions, and what the limitations of this approach are.

In section 2.5, where the use of target diagrams is described, a more stringent performance criterion is defined using a threshold of 0.5. However, this threshold appears arbitrary. It would be helpful to justify its selection, or to clarify that it is an exploratory criterion adopted for this study.

This section combines references to technical standards CEN/TS 17660-1 (2021) and 17660-2 (2025), which are specific to sensors, with DQOs from EU directives 2008/50/EC and 2024/2881, which were not designed for low-cost devices or indoor measurements. This could mislead readers (especially those unfamiliar with EU regulation) into thinking that sensors are expected to comply with the stricter DQOs from the 2024/2881 directive, which is not the case in current practice or within the scope of existing CEN/TS guidance. It is also important to note that both the CEN/TS standards and the cited directives are aimed at outdoor ambient air quality and set minimum temporal resolutions of one hour or longer (depending on the pollutant). Therefore, their direct application to shorter averaging intervals, like those used in this study, is not formally covered. It is suggested to explicitly state whether the inclusion of new DQOs is intended as an exploratory or forward-looking exercise, and to clarify that the current regulatory framework remains anchored in the 2008/50/EC directive and its associated DQOs. Simplifying the language in this section could also help improve accessibility for non-specialist readers.

The statement in line 341 regarding the need for “enough data points” to train the models is relevant but not further developed in the text. It would be valuable to include a more explicit discussion of the actual number of data points used in this study for calibrating each model (especially the NO2 models based on machine learning), and whether that amount is adequate given the complexity of the algorithms. Additionally, the expression “low uncertainties” may be ambiguous to some readers, as it refers here to the uncertainty associated with the calibration model (i.e., how well the functional relationship between raw and calibrated data reflects the actual relationship), and not to total measurement uncertainty (of which calibration

uncertainty is only one component). Clarifying this distinction would help avoid potential misinterpretation.

Results

In section 3.1.1, while PM2.5 was modeled using a linear approach with modest data volumes (200-600 data points per sensor), the NO2 models rely on machine learning algorithms (MLR, SVR, RFR, ANN), which typically require significantly larger datasets to avoid overfitting and ensure model stability. Since no table equivalent to Table S5 is provided for NO2, and the data volumes used are not explicitly reported, it is difficult to assess the validity and robustness of the applied models. See also the related comments on section 2.5.

In section 3.1.2, all NO2 calibration models fall on the left-hand side of the target diagram. It would be relevant to consider whether this reflects a systematic underestimation of the actual variability by the calibrated sensors, or a limitation in the dynamic response range of the devices.

The inclusion of multiple additional performance metrics in this section (MAE, MEF, slope/intercept, r) does not appear to be clearly justified, especially given that robust tools such as REU and target diagrams (which integrate error and bias) are already applied. Without a clear explanation of their added value (particularly from the perspective of potential users such as researchers, agencies, or communities) these metrics may come across as a statistical exercise without a defined purpose. Moreover, their inclusion could obscure the central message of the manuscript. It is recommended to focus this section on the key metrics relevant for practical decision-making and consider moving the rest to the Supplement.

In section 3.2.1, passive samplers are used to validate sensor performance; however, no information is provided about the quality/uncertainty of the data they produce.

Section 3.2.3 presents the use of activity logs as a strategy to contextualize sensor data and explore individual exposure patterns. It is suggested to frame this as an exploratory or complementary tool, and to consider whether previously cited studies offer methodological frameworks that could support a more systematic integration of metadata in future research on personal exposure.

Discussion

The final paragraph of section 4.1.1 introduces the “EASE” methodology, but it is unclear what exactly the authors are referring to, which hinders its interpretation. Additionally, this paragraph includes some generalizations that do not appear to be directly supported by the study’s results. It is recommended to revise the paragraph to more accurately reflect the exploratory nature of this reflection.

Section 4.1.2 would benefit from being restructured to clearly distinguish between: (i) the review of previous studies, (ii) the authors’ own findings, and (iii) the methodological decisions taken. It would also be useful to discuss whether the use of a linear model to incorporate ozone was the most appropriate choice, considering that sensor cross-sensitivity to O3 is not necessarily linear or constant.

Section 4.1.3 addresses the effect of averaging time on NO2 sensor calibration, but the empirical evidence is limited due to the short duration of the calibration phase (two weeks). It is recommended to report how many data points were used for each averaging interval (1, 5,

10, 15 min) and to discuss how variability in sample size affects the comparison. The statement "Our study demonstrated…" could be softened, as the available data do not support a generalizable conclusion. Moreover, the reduction in MAE is not contextualized in relation to other performance metrics, nor is its practical relevance evaluated, which makes it difficult to interpret the actual benefit of using a 10-minute average.

One of the recommendations in section 4.3 states: "we recommend adjusting the averaging time based on the available data...". This may be confusing or lead to methodologically questionable interpretations. The choice of temporal resolution should primarily be guided by the study's objectives (e.g., capturing acute exposure events), not just by the amount of data available. It is suggested to rephrase this recommendation or to justify its applicability in more detail.

The closing paragraph of this section suggests that, due to the poor performance of sensors at low concentrations, they may be more suitable for environments with high pollution levels. However, this statement is not supported by empirical evidence in the present study. It is recommended to revise this conclusion to more directly acknowledge the limitation observed, rather than inferring a potential use case that was not investigated here.

The reflection in section 4.4 on the meaning of the term "low-cost" is valid, but feels disconnected from the rest of the manuscript. If this content is to be retained, it is suggested to move it to the conclusions and, most importantly, to explicitly connect it to the study's findings. For instance, the authors could discuss whether the calibration, validation, and performance requirements observed in this study still justify labeling these sensors as "low-cost" in real-world applications. Otherwise, the section may come across as tangential or anecdotal.

Conclusions

The conclusions section brings together various elements of the study, but does not clearly articulate a central message regarding the value, applicability, and limitations of the evaluated sensors for epidemiological research. It is recommended to revise this section to: (i) avoid generalized claims that are not supported by the evidence presented, (ii) strengthen the critical analysis of the results (e.g., the limited transferability of laboratory calibrations, or the actual utility of machine learning models with small datasets), and (iii) better guide the reader regarding practical implications and future directions. Reformulating the conclusions based on the study's initial objectives would help close the manuscript with greater clarity and strength.

Specific comments

Line 29: Consider adding "fine fraction" when referring to PM2.5.

Lines 45-46: Suggest adding a reference to support the statement.

Line 62: It would be useful to acknowledge that, although the high temporal resolution offered by sensors is valuable for capturing dynamic exposures, it also increases instrumental noise, which directly impacts measurement uncertainty, particularly important in epidemiological studies.

Line 82: It may be appropriate to introduce that this is a study attempting to apply a direct measurement approach.

Lines 88-89: Consider rephrasing for improved clarity.

Lines 95-96 and 103-104 could be rewritten for a more natural and concise expression.

In section 2.2 (and in other parts of the manuscript), the terms “air quality monitor”, “sensor system” and “AQSS” are used interchangeably without clarifying whether they are synonymous.

Check the numbering of the “Study design and quality assurance” section (2.3?).

In Table 1, clarify what is meant by “low concentrations”.

Lines 161-162: Clarify what the authors are referring to.

Lines 249-250: The wording could be improved for clarity and precision.

Line 253: This statement could be reconsidered or qualified: “In theory, the higher the number of data points, the better the model performs”, as model performance does not necessarily with more data points.

Lines 270-271: “Higher performance” could be replaced with “better performance” or a more neutral formulation. Also, “in general” suggests there may be exceptions, but that’s not applicable here, where the interpretation of target diagrams is systematic.

Line 287: “Resolution” appears twice.

Line 288: The phrase “especially in mobile measurements” seems out of place in the context of this study, since no mobile measurements were performed. In the following sentence, there may be a typo (“long” instead of “longer”?).

Consider including a brief explanation of the nomenclature used for sensor IDs.

Figures are in general a bit difficult to interpret, are different models, averaging intervals, or measurement phases being shown? It is suggested to include a complete legend in each figure or, alternatively, clearly describe in the text what each marker represents.

Figure 8 does not show results for participants P1 and P3, and the text in section 3.2.1 does not explain their exclusion. The sampling period (14 or 15 days, as mentioned in the text) is also not indicated. Recommend expanding or clarifying.

The term “passive samples” is used throughout the text. A less ambiguous expression could be “passive sampler measurements” or “passive sampling data”.

Line 455: It seems to refer to the data collected by the monitoring station, not the station itself. Check for clarity.

Line 456: It is recommended to specify more clearly that the sensor was located outdoors, outside the second-floor window, to avoid ambiguity.

Lines 464-465: It is stated that this is a “clear example” of the effect of temperature and relative humidity on electrochemical sensors. However, the presented data do not allow us to determine whether the observed error is due to sensor limitations or extrapolation of the model beyond the training range.

In Figure 9b (PM2.5), relative humidity is included as a contextual variable, which may be misleading since, according to the methodology, RH was not used as a variable in the PM calibration model. Given that RH was used for NO2 models (Figure 9a), this graphic choice could be misleading and would benefit from clarification.

Line 470: The statement that machine learning models “consistently overestimated” concentrations compared to passive tubes may be overstated, as only one biweekly average value per participant is available, and the comparison is limited to a few cases (Figure 8).

The use of the term “addition of ozone” in lines 531 and 535 could be misleading, as in the context of atmospheric chemistry or sensor testing it is often interpreted as the physical addition of ozone gas to an experimental mixture.