Response to reviewers

Title: "Enhancing Accuracy of Indoor Air Quality Sensors via Automated Machine Learning Calibration"

General Response: We would like to thank the editor and reviewers for the positive feedbacks and constructive comments. Below, we've provided a point-by-point response to all comments to clarify revisions and improvements in the manuscript.

Reviewer #1:

Comment 1. Ln 171: As far as I understand, there are eight regression algorithms. Better to mention the exact number of algorithms you used instead of 'multiple'.

Response:

We thank the reviewer. We have revised the Methods to report the exact algorithms used and to clarify wording. Specifically, we now state that AutoML evaluated 30 machine learning algorithms (Table S1).

Revised sentences:

We employed an AutoML framework to develop and select calibration models for the indoor air quality sensors. The AutoML approach generate a variety of (i.e., 30 in this study) candidate models and optimised their hyperparameters. Then the AutoML algorithm would identify a model that best maps the sensor outputs to the reference concentrations by ranking the cross-validation root mean squared error (RMSE). In our implementation, the input features to each model included the sensor's raw readings, indoor temperature, and RH, while the target output was the PM concentration measured by Fidas 200.

Comment 2. Ln 175: How did you allocate the 80% and 20% dataset? Randomly, chronologically, or other methods? Did you also use cross validation in the step?

Response:

Data splitting was random following previous studies. Within each concentration regime (<50 and ≥50 μg m^{-3}), we used H2O's splitFrame with a fixed seed (1014) to allocate 80% of the rows to training and 20% to a held-out test set. During AutoML, we used k-fold cross-validation (5-fold) on the training portion for model selection (sorted by RMSE). The held-out 20% test set was never used for training or tuning; we report both cross-validated training metrics and external test metrics (see Table S1). This choice ensured both train/test and cross-validation folds contained comparable concentration

distributions while avoiding temporal leakage, as the experiment container was well-mixed and emission episodes were interleaved.

We revised the manuscript by adding the above paragraph to reflect these points.

Comment 3. Ln 223: From Figure S1, I can't really see a clear threshold of 50 µg m^{-3}. To me, it looks more like 25 µg m^{-3} is the threshold. It's ok to use 50 µg m^{-3}, but it's better to show more clearly why it is chosen.

Response:

Thank you for the helpful observation. We chose 50 µg m^{-3} because the scatter shows two regimes relative to the 1:1 line. At or below 50 µg m^{-3} the point cloud is tight and lies mostly above the 1:1 line. This indicates a positive sensor bias at low concentrations. Above 50 µg m^{-3} the points shift below the 1:1 line and the fitted trend is flatter than the 1:1 reference. This is consistent with signal compression at higher particle loads. The observations are not evenly distributed. Most data lie below about 25 µg m^{-3} and only a small number of points fall between 25 and 100 µg m^{-3}. A split at 50 µg m^{-3} therefore separates the two behaviours cleanly and avoids further fragmenting ranges that are already sparse. We have clarified this in the manuscript so that the rationale for the chosen split is explicit.

Revised sentences:

Our exploratory analysis (Fig. S1) revealed a clear threshold at 50 µg m^{-3} where the sensor bias flips. We chose this value because the scatter plot of sensor versus reference measurements shows two distinct regimes relative to the 1:1 line. At or below 50 µg m^{-3}, the data cloud is tight and lies mostly above the 1:1 line, which indicates a positive sensor bias (overestimation) at low concentrations. Conversely, above 50 µg m^{-3} the cloud shifts below the 1:1 line, and the fitted trend becomes flatter than the 1:1 reference, a pattern consistent with signal compression and underestimation at higher particle loads. This split is further justified by the data distribution; most data lie below about 25 µg m^{-3}, with only a small number of points between 25 and 100 µg m^{-3}. A split at 50 µg m^{-3} produces two interpretable regimes that align with the observed change in bias, keeps the rare high-concentration events together, and avoids slicing the dense background data into very small groups, which would reduce model stability. Therefore, we applied a stratified calibration strategy, training separate AutoML models for the low (<50 µg m^{-3}) and high (50–600 µg m^{-3}) regimes in both the field-to-drift (f2d) and drift-to-reference (d2r) stages. This allows us to tailor the calibration to the specific bias profile of each regime and thereby minimises systematic error across the sensor's full operating range.

Comment 4. Figure 2: It is a good flow chart, but the arrows with different colors are a bit confusing. What do brown and blue arrows represent respectively?

Thank you for raising this point. We agree that the color encoding should have been explicitly defined in the manuscript. We have added clear explanations in Figure 2 caption and in the caption so that the meaning of each arrow is unambiguous.

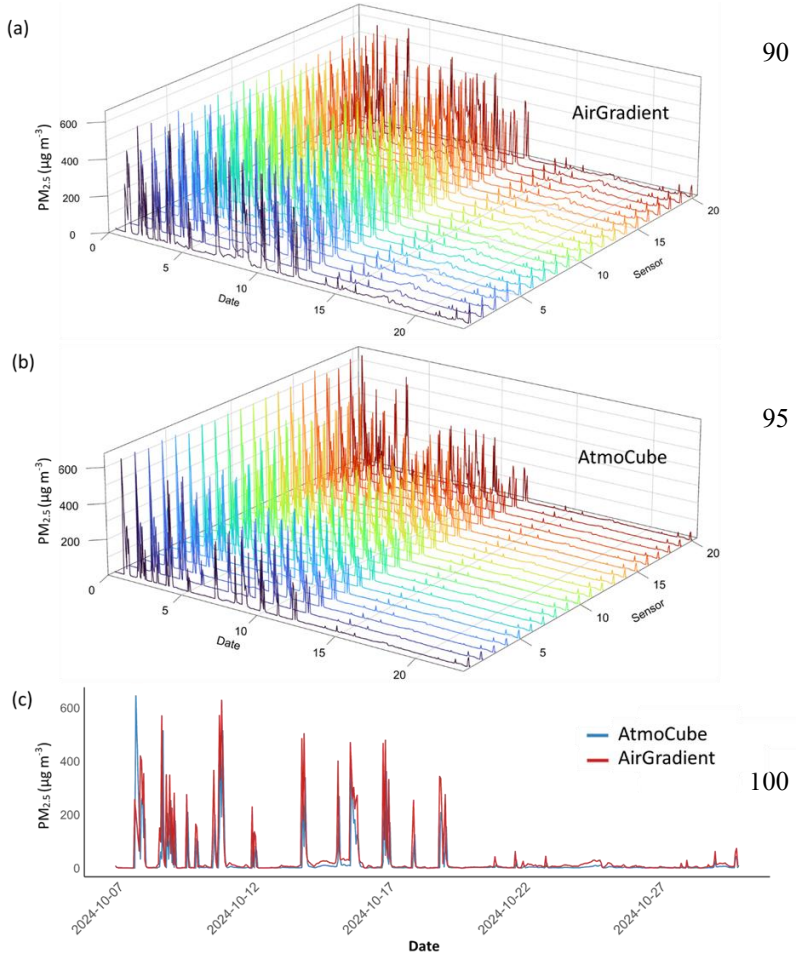
The blue arrows represent the primary data flow used to train and apply the AutoML calibration models across Processes 1 to 3. This path links the field sensors, the drift-reference sensors, and the reference instrument during both training and prediction, and it yields calibrated field-sensor readings relative to the reference instrument. The brown dashed arrow represents the post-deployment recalibration path. After the field sensors are retrieved from the observation, we rebuild the field-to-drift and drift-to-reference models using the post-deployment dataset to correct for sensor drift and then route the updated predictions through Process 3 to obtain post-deployment calibrated readings referenced to the same instrument. The color scheme therefore distinguishes workflow timing rather than sensor type or data source, with blue indicating the main training and prediction workflow and brown dashed indicating the optional recalibration that is activated after deployment.

To make this explicit without modifying the flowchart, we inserted the following text in Figure 2 caption: “We use a fixed color scheme to distinguish the two stages of the workflow. Blue arrows and lines represent the main training and prediction path that spans Processes 1–3. The brown arrows and lines represent the post-deployment recalibration path, which is executed after sensor retrieval to correct drift using the post-deployment dataset. The resulting predictions are passed through Process 3 to obtain calibrated readings mapped to the reference instrument.”

Comment 5. Figure 3 a/b: Are these two subplots 3-dimensional? I see that all the lines for timeseries are not really aligned vertically. It can be misleading as they can mean that they are not in the same timestamp. I understand that in this way you can illustrate the difference between lines more clearly, but at least you need to draw a z-axis to show that there is a third dimension.

Response:

Thank you for flagging this. In the revision we converted panels 3a and 3b into explicit 3D time series plots by adding and labelling a z axis. The x axis shows time, the y axis shows $PM_{2.5}$ concentration in µg m^{-3}, and the z axis indexes the sensors with tick labels corresponding to sensor numbers. All series share the same time coordinate, so timestamps are aligned vertically, and the apparent separation is now clearly along the sensor dimension rather than time. These clarifications improve readability and remove the possibility of misinterpretation, while leaving the analysis and conclusions unchanged. Please see the revised figure below.



Comment 6. Table 1: The problem with using 50 µg m^{-3} as threshold might be that the subset of 'above 50 µg m^{-3}' has too few data points compared to the 'below 50 µg m^{-3}'. Have you considered using another value as a threshold (similar to question 2)?

Response:

We recognize that there are fewer data points above 50 µg m^{-3}. We chose 50 µg m^{-3} as the threshold because there is a clear, empirically observed transition between two distinct response regimes of the sensors in comparison with the reference instruments. As shown in our exploratory plots, data at or below this value forms a tight cluster predominantly above the 1:1 line, indicating a consistent low-end positive bias. In contrast, above 50 µg m^{-3}, the data points rotate below the 1:1 line and the slope flattens. Choosing the threshold at this transition point allows us to model a real change in the sensor's behavior, a rationale we now explicitly state in the Methods.

For data above-50 µg m^{-3}, although relatively small, it is sufficiently informative to train a stable model. The performance of the calibrated models validates this approach, as they retain strong agreement with the reference instrument. In this high-concentration regime, AirGradient achieved an R^2=0.92 and IOA=0.87, while AtmoCube achieved an R^2=0.76 with minor residual bias. The fit is neither noisy nor over-tuned. Further evidence comes from independent train-test statistics, which show that the models generalize well, achieving a test R^2 of 0.78–0.79 and confirming that the sample is large enough to learn a robust mapping specific to that regime.

We indeed considered adjusting the threshold, but the highly skewed nature of the data distribution makes alternatives ineffective. Most observations fall below 25 µg m^{-3}, with very few events between 25 and 100 µg m^{-3}. Lowering the threshold would contaminate the high-range model with low-bias data points, blurring the very change in slope we aim to capture. Conversely, raising the threshold would further diminish the already scarce high-range data. Therefore, keeping the threshold at the empirically observed change point preserves interpretability and reduces model misspecification.

Comment 7. Ln 176-195, Figure S7, and several instances in the paper: You separated two model subsets 'below 50 µg m^{-3}' and 'above 50 µg m^{-3}'. When you compared the results using MAE and RMSE, it's not easy to tell which one has a worse MAE as you have the two subsets with different reference values. It's the most obvious on Figure S7 that you tried to compare the errors in different concentration ranges using MAE. This comparison may not be the most appropriate as a smaller concentration apparently has a smaller MAE. What matters in this case is the error percentage. I suggest authors using mean absolute percentage error (MAPE) instead of MAE when comparing the two subsets. For RMSE, there is also an alternative NRMSE.

Response:

Thank you for this helpful suggestion, which we have implemented in the revised manuscript. In addition to RMSE and MAE, we now report normalized RMSE (NRMSE) and symmetric mean absolute percentage error (sMAPE) for each concentration regime. NRMSE is defined as RMSE divided by the mean Fidas 200 concentration in the corresponding subset and expressed as a percentage, while sMAPE provides a symmetric percentage error that is more stable at very low concentrations than conventional MAPE. Figure S7 and the accompanying text (Lines 176–195) have been updated so that comparisons between the below 50 µg m^{-3} and above 50 µg m^{-3} regimes are based on NRMSE and sMAPE rather than MAE alone. This makes the relative error behavior across the two concentration ranges clearer and directly addresses the reviewer's concern about concentration-dependent absolute errors.

We retain RMSE and MAE to facilitate comparison with guideline values, manufacturer specifications, and previous calibration studies, which almost always report absolute error metrics, and because they summarize complementary aspects of model performance (MAE reflecting the typical deviation and RMSE being more sensitive to occasional large errors). In the revision, NRMSE and sMAPE are therefore used to compare relative performance across concentration regimes, while MAE and RMSE are reported alongside to document the absolute magnitude and structure of the residuals.

Comment 8. Table S1: What are the criteria of ranking the eight regression algorithms? For the best algorithm? For model subset of low conc., the one listed as rank 1 has a higher RMSE/MAE and lower R2 compared to the one listed as rank 2. For high conc., the one ranked as 5th has a better MAE/RMSE/R2 results than the one ranked top. Are you using some other criteria for the ranking? Please list them in the table as well. Also, please clarify what criteria you used in the main text as well.

Response:

We recognize the need to be clear with the criteria. In the original Table S1, "Rank 1" referred to the H2O AutoML leaderboard order, which we configured to sort by cross-validated RMSE computed on the training set (k-fold

cross-validation). In k-fold cross-validation, the training set is split into k parts; each model is trained on k-1 parts and evaluated on the remaining part, and the errors are averaged. RMSE (μg m^{-3}) penalizes large errors more than small ones, which is important for $PM_{2.5}$ where high concentrations matter.

In the same table, we also reported performance on an independent 20% test set that AutoML did not use for training or ranking. Because the leaderboard metric (cross-validated RMSE) and the external test metric (single-split RMSE on the test set) are computed on different data with different protocols, it is expected that the model at Leader Rank 1 may not have the lowest error on that test set. This is especially plausible in the high-concentration subset, where there are fewer data points and variance is higher, and when comparing metrics with different sensitivities (e.g., RMSE vs. MAE vs. R^2).

To remove ambiguity, we revised Table S1 to report two ranks for each subset. Leader Rank is the AutoML order by cross-validated RMSE on the training set and is our model-selection criterion. We use Leader Rank because (i) it preserves the independence of the 20% test set by not using it to choose the winner, avoiding optimistic bias; (ii) averaging errors across folds provides a more stable, lower-variance estimate than a single split; and (iii) RMSE penalizes large deviations, aligning with the scientific and regulatory importance of limiting large $PM_{2.5}$ errors. Alongside this, External Test Rank orders models by RMSE on the independent 20% test set and is included for transparency. In the Methods we now state explicitly that selection follows the leaderboard metric, whereas all performance reported in the Results refers to the test set (with splits created using a fixed random seed for reproducibility).

For transparency we also describe each column exactly as it appears in the revised Table S1. “Model_Subset” identifies whether the model belongs to the low concentration subset below 50 μg m^{-3} or the high concentration subset at or above 50 μg m^{-3}. “Model_ID” is the H2O identifier that allows exact reproduction. “Algorithm” is the model family returned on retrieval. “Test_RMSE”, “Test_MAE”, and “Test_R2” report performance on the independent 20% hold-out set, with RMSE and MAE in μg m^{-3} and R^2 dimensionless. “Train_RMSE”, “Train_MAE”, and “Train_R2” are training-frame summaries shown for context. All columns prefixed with “LB_” are copied directly from the AutoML leaderboard with extra columns enabled and therefore reflect the cross-validation view used to form the leaderboard. In this file they include “LB_rmse”, “LB_mae”, “LB_mean_residual_deviance”, and “LB_ Rank”, where “LB_Rank” is the raw leaderboard order that determines the top model for each subset.

Comment 9. Could this framework also be used outdoors for ambient air concentration? Can this framework also work for aerosols of larger size and gas pollutants?

Response:

Thank you for raising the question of generalizability. The framework is model agnostic and can be transferred to outdoor settings. In practice, the same three stage structure field to drift reference, drift reference to reference instrument, and the composite field to reference transfer can be implemented outdoors by collocating a subset of sensors with a regulatory or research-grade outdoor reference sensor and by using outdoor meteorology as covariates such as ambient temperature, relative humidity, and wind driven dispersion proxies. The concentration regime split is not fixed and should be learned from the outdoor dataset using the same data driven rationale that we applied indoors. However, if there is no clear difference in the trend in different concentration range, one ML model may be sufficient. The manuscript now states that regime boundaries are empirical and should be re-estimated for a new application, e.g., outdoors.

Yes, the workflow should also apply to other particulate matters such as PM_1, PM_4, and PM_{10} as well as if a time aligned reference for the target parameter is available.

Technical comments:

Ln 251: No need to use hyphens for 400-to-500 µg m-3

Ln 261: If you have used 50 µg m-3 as a threshold, then there is no need to use the symbol '~'.

Ln 311: What is ~10? It's very confusing what you referred to without a unit.

Response:

Thank you for these careful technical comments. We have revised the text accordingly. At line 251, "between 400-to-500 µg m^{-3}" has been corrected to "between 400 and 500 µg m^{-3}". At line 261, we removed the tilde so that 50 µg m^{-3} is now written without the approximation symbol, consistent with its use as a threshold. At line 311, we clarified that "~10" refers to an approximate 10% relative error at 600 µg m^{-3} and now state the unit explicitly in the sentence.

Reviewer #2:

Comment 1. Line 146–152: Here, the authors describe which two types of low-cost sensors were used and against which reference instrument they were compared. I am wondering whether the low-cost sensors have a similar detectable particle size range as the Palas Fidas 200. It could add an additional layer of clarity to mention the particle size range for the reference instrument and - if available - also for the low-cost sensors.

Response:

Thank you for this suggestion.

We do agree. Yes, the detectable particle size range is similar. We have revised the manuscript to include the manufacturer-specified particle size ranges for all instruments. The reference instrument, the Palas Fidas 200, has a certified detection range of 0.18–18 µm. The low-cost sensors, the Plantower PMS5003 and the Sensirion SPS30, both have a specified particle size detection range of 0.3–10 µm.

While addressing this comment, we also took the opportunity to include the manufacturer-specified measurement uncertainties for all instruments to further enhance the technical comparison. The revised text now reflects both points.

Revised sentences:

An aerosol spectrometer (i.e., Palas Fidas 200 (detectable particle size of 0.18-18 µg, ranges from 0 to 10,000 µg m^{-3} with 9.7% uncertainty for $PM_{2.5}$ measurements)) was used as the reference-grade instrument for sensor performance evaluation and calibration. A total of 40 low-cost air quality sensors was deployed within the chamber, settled on a table at near the same height with Fidas 200 to minimize positional variability. Our air quality sensors consisted of two different types, including 20 units of AirGradient ONE (Model I-9PSL) and 20 units of AtmoCube. AirGradient ONE sensors measure $PM_{2.5}$ using a Plantower PMS5003 laser-scattering sensor (manufacturing specification: detectable particle size of 0.3-10 µm, with ±10 µg m^{-3} at 0-100 µg m^{-3} 10% at 100-500 µg m^{-3}), and temperature and RH through a Sensirion SHT40 sensor. AtmoCube sensors detect particulate matter using a Sensirion SPS30 laser-scattering sensor (manufacturing specification: detectable particle size of 0.3-10 µm, with ±5 µg m^{-3} at 0-100 µg m^{-3}, ±10% at 100-1000 µg m^{-3}), temperature using a Sensirion STS35-DIS, and RH using a Sensirion SHTC3.

Comment 2. Line 154 – 155: Hopefully, the team got to enjoy the food afterward. Scientific dedication always deserves a good meal.

Response:

Thank you for this kind remark. The cooking part did indeed come with some well-deserved meals for the team, and it was all conducted in accordance with our laboratory safety and hygiene procedures.

Comment 3. Line 157: The phrase "natural indoor conditions" sounds somewhat contradictory, since indoor environments are by definition artificial. It might be clearer to use "realistic" or "typical".

Response:

Yes, we fully agree that "natural indoor conditions" is by definition artificial and thus we have revised the wording accordingly.

In the revised manuscript, we now write "typical indoor conditions" at the corresponding location (former line 157), so the sentence reads:

"Temperature and RH levels were allowed to exchange passively with the outdoor air with no mechanic ventilations or windows/door opening, mimicking indoor conditions where these parameters may fluctuate."

This change clarifies our intended meaning and avoids the contradiction you highlighted.

Comment 4. Line 223–228: I wonder how the threshold of 50 µg m^{-3} was determined. The text mentions that the exploratory analysis revealed a bias flip at this concentration, but it could strengthen the explanation to briefly clarify why 50 µg m^{-3} was selected as the cutoff.

Response: Please also see response to Comment 3 by reviewer 1.

We appreciate this observation. We selected 50 µg m^{-3} because the scatter plot clearly shows two distinct regimes relative to the 1:1 line. Below 50 µg m^{-3}, the points form a compact cluster mostly above the 1:1 line for AirGradient ONE sensors, indicating a positive sensor bias at low concentrations. Above 50 µg m^{-3} the points shift below the 1:1 line and the fitted regression becomes shallower than the 1:1 reference, which is consistent with signal compression at higher particle loads. The data are also unevenly distributed. Most observations fall below about 25 µg m^{-3} and only a small fraction lie between 25 and 100 µg m^{-3}. Using 50 µg m^{-3} as the threshold therefore separates these two behaviors clearly while avoiding further subdivision of already sparse ranges. We have revised the manuscript to explain this rationale explicitly.

Comment 5. Line 251: In the phrase "between 400-to-500 µg m^{-3}" the hyphen is unnecessary. It would read more clearly as "between 400 and 500 µg m^{-3}."

Response:

Thank you for pointing this out. We have revised the wording accordingly and now write "between 400 and 500 µg m^{-3}" to improve clarity at Line 251.

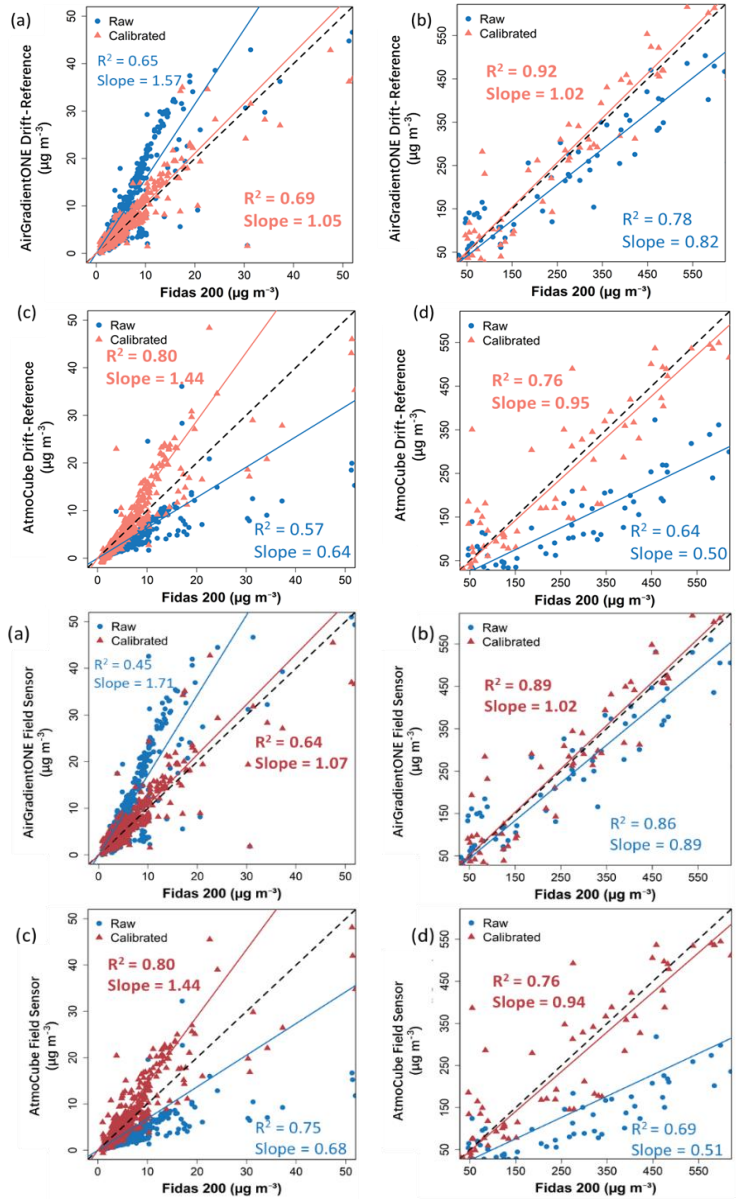
Comment 6. Figure 3: The two time series plots showing the raw data from multiple low-cost sensors are displayed as 3D graphics, which causes a slight misalignment between the sensor readings and the x–y axes, making the visualization somewhat confusing. I am not sure this is the most effective way to present the data, although it is not a major issue. I would generally recommend adding x-axis tick marks to indicate the days (with major ticks for the labeled days and minor ones for each individual day) and including grid lines especially in panels (a) and (b), which would help improve readability and reduce the visual confusion caused by the three-dimensional layout.

Response:

We appreciate you drawing attention to the potential ambiguity in Figure 3. In the revised version we have converted panels 3a and 3b into explicit three-dimensional time series plots by adding and labelling a z axis. The x axis now represents time, the y axis shows $PM_{2.5}$ concentration in µg m^{-3}, and the z axis indexes the sensors, with tick labels corresponding to sensor numbers. All series share a common time coordinate, so timestamps are aligned vertically and the apparent separation is clearly along the sensor dimension rather than time. These changes improve readability and remove the scope for misinterpretation, while leaving the analysis and conclusions unchanged.

Comment 7. Figures 4 and 5 appear to have different sizes and resolutions. Since both figures are quite similar, it would be visually more appealing and consistent to display them at the same size and resolution for better comparison and overall presentation quality.

Response: Thank you for this helpful suggestion. We agree that having Figures 4 and 5 in the same size and resolution improves consistency and makes comparison easier. In the revised manuscript, we have adjusted both figures so that they are now displayed at the same size and resolution to enhance visual clarity and overall presentation quality. Please see the revised figure below.



Comment 8. I really appreciate the thorough and transparent discussion of the limitations of the proposed method. The authors clearly acknowledge the constraints related to environmental conditions, emission sources, and long-term applicability. An analysis of long-term sensor drift would have provided valuable additional insights into the robustness of the calibration over extended periods. However, since this aspect is explicitly discussed as a limitation and identified as an important direction for future work, I find the current scope appropriate and well-justified for this study.

Response:

We thank the reviewer for this positive and encouraging comment on our discussion of limitations. We fully agree that an explicit analysis of long-term sensor drift would provide valuable evidence on the robustness of the calibration over extended deployments. In future studies we plan multi-year field deployments to quantify drift and to test automated recalibration within the proposed framework.

Enhancing Accuracy of Indoor Air Quality Sensors via Automated Machine Learning Calibration

Juncheng Qian[1], Thomas Wynn[1], Bowen Liu[2], Yuli Shan[1], Suzanne E. Bartington[3], Francis D. Pope[1], Yuqing Dai[1,*], Zongbo Shi[1,*]

[1] School of Geography, Earth and Environment Sciences, University of Birmingham, Birmingham, B15 2TT, UK
[2] Department of Management, Birmingham Business School, University of Birmingham, Birmingham, B15 2TT, UK
[3] Institute of Applied Health Research, University of Birmingham, Birmingham, UK

Correspondence to: Zongbo Shi (z.shi@bham.ac.uk), Yuqing Dai (y.dai.2@bham.ac.uk)

Abstract. Indoor fine particles ($PM_{2.5}$) exposure poses significant public health risks, prompting growing use of low-cost sensors for indoor air quality monitoring. However, maintaining data accuracy from these sensors is challenging, due to interference of environmental conditions, such as humidity, and instrument drift. Calibration is essential to ensure the accuracy of these sensors. This study introduces a novel automated machine learning (AutoML)-based calibration framework to enhance the reliability of low-cost indoor $PM_{2.5}$ measurements. The multi-stage calibration framework connects low-cost field sensors to be deployed with intermediate drift-correction reference sensors and a reference-grade instrument, applying separate calibration models for low (clean air environment) and high (pollution events) concentration ranges. We evaluated the framework in a controlled indoor chamber using two different sensor models exposed to diverse indoor pollution sources under uncontrolled natural ambient conditions. The AutoML-driven calibration significantly improved sensor performance, achieving a strong correlation with reference measurements (R^2>0.90) and substantially reducing error metrics (with normalized root-mean-square error (NRMSE) and symmetric mean absolute percentage error (sMAPE) roughly halved relative to uncalibrated data). Bias was effectively minimised, yielding calibrated readings closely aligned with the reference instrument. These findings demonstrate that our calibration strategy can convert low-cost sensors into a more reliable tool for indoor air pollution monitoring. The improved data quality supports atmospheric science research by enabling more accurate indoor $PM_{2.5}$ monitoring, and informs public health interventions and evaluation by facilitating better indoor exposure assessment.

1 Introduction

Air quality monitoring is essential in understanding exposure to pollutants in both outdoor and indoor environments, which informs public health improvement strategies. In particular, indoor air quality (IAQ) has gained attention because people spend the majority of their time indoors, yet historically it has been difficult to measure indoor pollutants continuously (Aix et al., 2023). Traditional approaches for IAQ assessment relied on expensive reference instruments (e.g. filter-based gravimetric samplers with pumps and impactors) that require expert operation and maintenance. These practical challenges made long-term indoor monitoring infeasible in most settings (Levy Zamora et al., 2018). Recently, however, dramatic advances in low-cost sensor technology have transformed this landscape. Compact and affordable low-cost sensors for particulate matter (PM) and gases have made it possible to deploy dense monitoring networks and to track air quality in homes, offices, and other indoor spaces in real-time. For example, a consumer-grade PM sensor "PurpleAir" is now widely used, and over 5,600 devices reporting to an online map, and about 18% of these were deployed indoors as of 2020 (Koehler et al., 2023). This surge in low-cost sensor use highlights their promise for broad IAQ surveillance and community engagement in air quality improvement efforts.

As low-cost sensors proliferate, ensuring their data quality through proper calibration has become a critical concern. These sensors often suffer from biases and interferences that can compromise accuracy. For example, low-cost PM sensors that use optical scattering can be highly sensitive to environmental factors like relative humidity (RH) and aerosol properties. At high

RH (> 80%), condensation on the sensor or particles can lead to overestimation of fine particles ($PM_{2.5}$) concentrations (Crilley et al., 2020; Hagan & Kroll, 2020). Cross-sensitivities are also common, electrochemical gas sensors may respond to non-target gases (e.g. ozone sensors responding to nitrogen dioxide NO_2). Moreover, the performance of air quality sensors can degrade over time due to aging and fouling of components (so-called "drift effect"). Studies have showed that low-cost sensors tend to lose sensitivity or shift baseline after months of use, and electrochemical sensor singles degrades within two years, necessitating periodic recalibration (Zaidan et al., 2022; Zimmerman et al., 2018) .

To address these issues, a variety of calibration techniques have been explored previously, ranging from simple corrections to machine learning (ML) models. Traditional calibration methods typically include collocating low-cost sensors with a reference-grade instrument (such as federal reference methods, FRMs) and deriving a statistical correction (Liang, 2021). The simplest approach is a linear regression or affine transformation that aligns the sensor readings to the reference values. Additional environmental parameters are generally incorporated into multi-variate calibration models, for example, temperature and RH are included as independent variables to account for their influence on sensor response (Kang & Choi, 2024). These methods, including one-point or two-point calibrations and polynomial fits, have been shown to improve sensor accuracy under stable conditions (Cowell et al., 2023). In practice, laboratories or field researchers may perform a pre-deployment calibration by exposing sensors to known pollutant concentrations and fitting a curve. However, a calibration derived in one setting does not necessarily transfer well to another. Studies have noted that calibrations done in controlled lab environments often do not span the full range of real-world conditions, limiting their generality (Kim et al., 2019; Li et al., 2018; Mousavi & Wu, 2021). Different particle compositions also affect the magnitude of the sensor response (Crilley et al., 2020; Zou et al., 2021). Therefore, in situ calibration is often recommended to capture local environmental effects to yield more robust calibration models, allowing necessary adjustments for factors like aerosol composition and meteorological conditions (Raysoni et al., 2023). Although the performance of these traditional methods may be suboptimal when sensor response relationships are highly non-linear or environment-specific, they are still widely used due to their transparency and ease of implementation.

Recently, ML algorithms have been employed to improve calibration accuracy and capture complex sensor behaviours. ML calibration methods can simulate non-linear relationships and interactions that traditional linear methods might neglect (Villanueva et al., 2023). A range of ML approaches has been applied, including artificial neural networks (ANN), support vector regression (SVR), random forests (RF), gaussian process regression (GPR), and even semi-parametric models like generalized additive models (GAM) (Mahajan & Kumar, 2020). These data-driven models leverage not only raw readings from the sensor but often additional features (e.g., RH, temperature, timestamps) to learn the mapping to actual pollutant concentrations. Several studies have presented the effectiveness of ML-based calibration. Nowack et al. (2021) compared a regularized linear model (ridge regression) against non-linear models (random forest and GPR) for calibrating nitrogen dioxide (NO_2) and particulate matter with a diameter less than 10 micrometres (PM_{10}) sensors, finding that the machine learning approaches achieved high out-of-sample accuracy (frequently coefficient of determination R^2>0.8) and outperformed

traditional multiple linear regression models (Nowack et al., 2021). Mahajan et al. (2019) observed that an SVR model provided better calibration performance for PM_{10} sensors than both linear regression and standard neural networks (Munir et al., 2019). Nonetheless, ML-based calibrations also present challenges. They typically require a substantial dataset of sensor as well as reference readings for training, and their predictions can be unreliable outside the range of training data. For instance, an ANN or RF may struggle to extrapolate to pollutant levels higher than it has been seen during calibration, whereas a Gaussian process regression model may handle extrapolation with less bias (Nowack et al., 2021). Additionally, the calibration model learned at one location may not generalize to a new location (i.e., site transferability issue) unless a wide variety of conditions are considered. Despite these limitations, ML-based calibration can significantly improve the performance of low-cost sensors when carefully applied (Liu et al., 2019; Nowack et al., 2021; Villanueva et al., 2023; Zimmerman et al., 2018). While most field calibration studies to date have focused on outdoor deployments, where sensors are co-located with regulatory-grade monitors or used in ambient networks, a critical gap in the current literature is the calibration of low-cost sensors specifically for indoor environment.

Indoor air, however, can differ markedly from outdoor air in composition and dynamics. Factors like indoor-generated particles (from cooking, smoking, etc.), confined space, and higher humidity or temperature fluctuations can all influence sensor readings. For example, cooking can release ultrafine particles and organic aerosols in short bursts, causing sharp concentration spikes. A study reported that indoor $PM_{2.5}$ levels peaking near 488 $\mu g\ m^{-3}$ during cooking in a home, far exceeding typical outdoor concentrations (Cowell et al., 2023). Tobacco smoke similarly produces dense particulate matter and complex chemicals in confined spaces. Also, indoor spaces often have limited ventilation, allowing pollutants to accumulate and humidity to fluctuate in ways not seen outdoors. These conditions test the limits of calibration models. A calibration model trained mostly on moderate outdoor pollution levels may not extrapolate well to the abrupt spikes or ultra-low concentrations encountered indoors (Koehler et al., 2023). Compounding the issue, gathering extensive indoor calibration datasets is difficult, reference-grade indoor measurements are rare because deploying instruments indoors at scale is resource-intensive. As a result, there is a paucity of calibration methods tailored to indoor use, and questions remain about how well the algorithms proven in ambient air translate to indoor settings. This gap is increasingly problematic as the adoption of indoor air quality sensors grows; without reliable calibration, the data from these sensors could mislead users or undermine trust in sensor-based monitoring.

In this study, we aim to bridge the gap by introducing a replicable calibration approach for indoor air quality sensors using Automated Machine Learning (AutoML). AutoML is an emerging technology that automates the selection of machine learning algorithms and hyperparameters to build optimal models (LeDell & Poirier, 2020). Our objective is to develop a calibration framework that can be easily applied to low-cost sensor data in indoor environment to improve its accuracy and reliability. Unlike traditional calibration methods that might rely on fixed formulas or manually crafted ML models, an AutoML-based approach automates the selection and optimization of the calibration model. In our framework, sensor readings (e.g., raw $PM_{2.5}$ concentrations) are combined with environmental variables (mainly indoor temperature and RH), and an AutoML is employed to identify the best-performing calibration model through automated testing of many algorithms and hyperparameter settings.

By allowing the AutoML system to explore a wide range of potential models (from linear regressions to complex ensemble methods), we ensure that the final chosen model is well-suited to the characteristics of the indoor dataset, without requiring the user to have advanced machine learning expertise. The proposed approach is replicable in that it provides a general template that can be applied to other indoor sensor deployments, that is, researchers or practitioners can feed their co-location data into the same AutoML pipeline to obtain a custom calibration model for their specific environment.

The remainder of this paper is structured as follows. Section 2 describes the experimental setup and calibration methodology, including indoor air quality sensors, reference instruments, data collection procedures, and the AutoML workflow employed to generate calibration models. Section 3 presents the calibration results and discusses the implications of the findings. Section 4 summarizes the key findings. We also discuss limitations of our approach and provide recommendations for future research.

2 Method

2.1 Experimental Configurations

A controlled laboratory experiment was conducted within a custom-built container designed to simulate realistic indoor air pollution conditions (Fig. 1(a)). The chamber was equipped with fans to ensure uniform pollutant distribution (Fig. 1(b)), which minimized spatial concentration variations, essential for maintaining stable and reproducible conditions during sensor evaluation. An aerosol spectrometer (i.e., Palas Fidas 200 (detectable particle size of 0.18-18 µg, ranges from 0 to 10,000 µg m^{-3} with 9.7% uncertainty for $PM_{2.5}$ measurements)) was employed as the reference-grade instrument to provide high-precision baseline measurements for sensor performance evaluation and calibration. A total of 40 low-cost air quality sensors was deployed within the chamber, settled on a table at near the same height with Fidas 200 to minimize positional variability. Our air quality sensors consisted of two different types, including 20 units of AirGradient ONE (Model I-9PSL) and 20 units of AtmoCube. AirGradient ONE sensors measure $PM_{2.5}$ using a Plantower PMS5003 laser-scattering sensor (detectable particle size of 0.3-10 µg, with ± 10 µg m^{-3} at 0-100 µg m^{-3}, ±10% at 100-500 µg m^{-3}), and temperature and RH through a Sensirion SHT40 sensor. AtmoCube sensors detect particulate matter using a Sensirion SPS30 laser-scattering sensor (detectable particle size of 0.3-10 µg, with ±5 µg m^{-3} at 0-100 µg m^{-3}, ±10% at 100-1000 µg m^{-3}), temperature using a Sensirion STS35-DIS, and RH using a Sensirion SHTC3.

To generate diverse and realistic indoor air pollution profiles, three indoor emission sources were introduced into the experimental container, including incense sticks, cigarette smoke from 7th to 21st Oct 2024, and cooking emissions (i.e., frying vegetables, bacon, and fries) from 22nd to 30th Oct 2024 (Fig. 1(b) and Fig. 1(c)). All AirGradient ONE and AtmoCube sensors and the Fidas 200 were exposed to the same emission sources simultaneously. Temperature and RH levels were allowed to exchange passively with the outdoor air with no mechanical ventilation or windows/door opening, mimicking indoor conditions where these parameters may fluctuate. ~~Temperature and RH levels were allowed to exchange passively with the ambient environment with no ventilations or windows opening, mimicking natural indoor conditions where these parameters fluctuate freely.~~ Between each emission event, the container was ventilated until pollutant concentrations returned to

background levels (mainly during the night), ensuring that there was no cross-contamination between different test conditions, thus generating a reliable dataset for subsequent sensor performance evaluation and calibration.

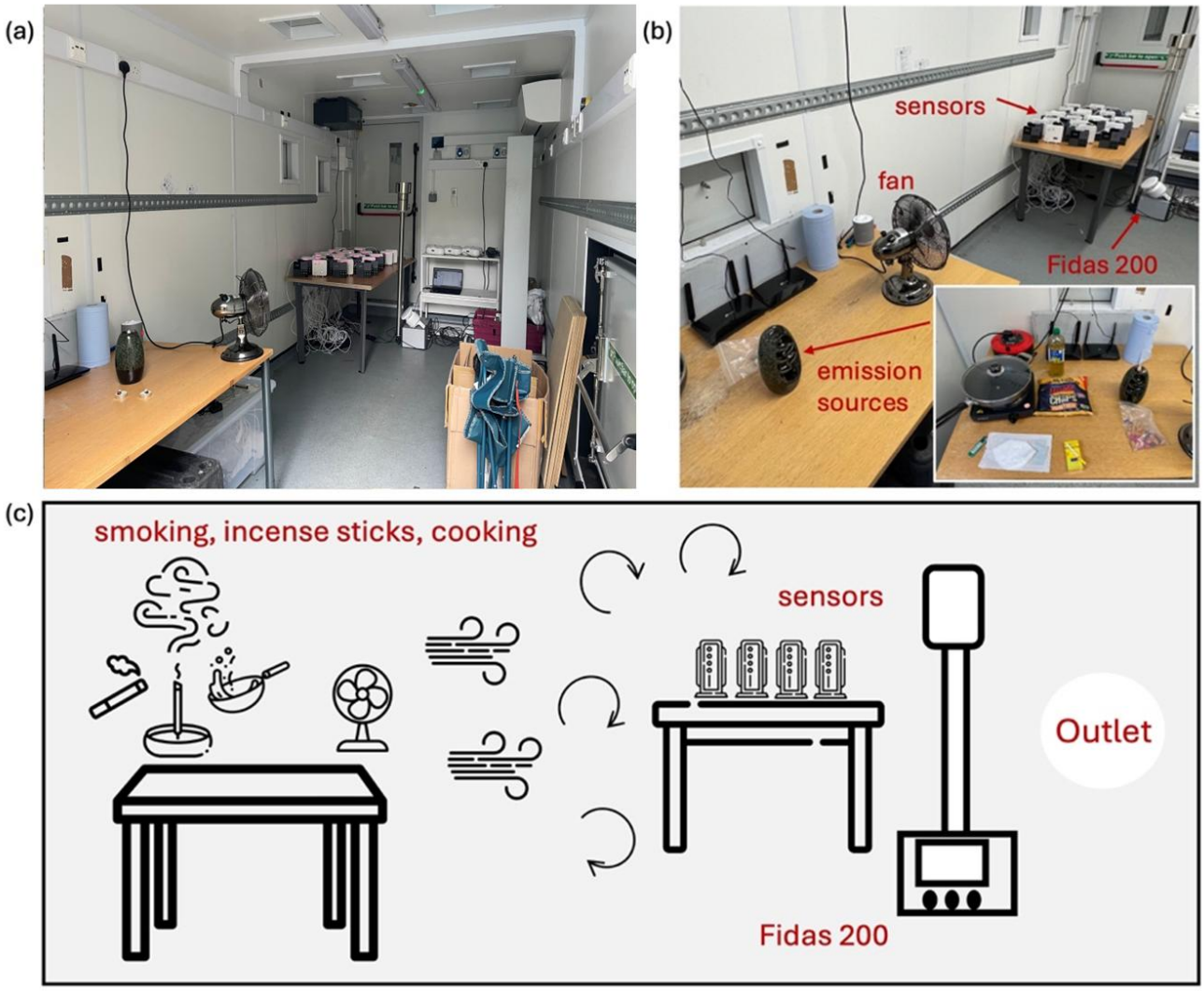


Figure 1: Overview of indoor air quality sensor calibration setup: (a) fully renovated half-size container, (b) emission sources and analytical instrumentation, and (c) schematic of pollutant generation and instrument placement.

2.2 Automated Machine Learning

We employed an AutoML framework to develop and select calibration models for the indoor air quality sensors. The AutoML approach ~~systematically~~ generate a variety of (i.e., 30 in this study) candidate models and optimise~~d~~s their hyperparameters.

Then the AutoML algorithm -would identify a model that best maps the sensor outputs to the reference concentrations. In our implementation, the input features to each model included the sensor's raw readings, indoor temperature, and RH, while the target output was the PM concentration measured by Fidas 200. ~~The AutoML process explored multiple regression algorithms, including gradient boosting machines (GBM), distributed random forest (DRF), and extreme gradient boosting (xgboost), to identify a model that best maps the sensor outputs to the reference concentrations.~~

This study~~A typical training strategy was applied, with 80% of dataset allocated for model training and the remaining 20% reserved for performance testing.~~ used H2O's splitFrame with a fixed seed (1014) to allocate 80% of the rows to training and 20% to a held-out test set. During AutoML, we used k-fold cross-validation (5-fold) on the training portion for model selection (sorted by root mean square error (RMSE)). The held-out 20% test set was never used for training or tuning; we report both cross-validated training metrics and external test metrics (see Table S1). This choice ensured both train/test and cross-validation folds contained comparable concentration distributions while avoiding temporal leakage, as the experiment container was well-mixed and emission episodes were interleaved.

Evaluation metrics were calculated for each candidate to guide the selection of the best model. We primarily used the RMSE, normalized root mean square error (NRMSE), mean absolute error (MAE), symmetric mean absolute percentage error (sMAPE), mean bias error (MBE), index of agreement (IOA), and R^2 as the performance criteria. RMSE quantified the average magnitude of prediction errors in units matching the observed data, with lower values reflecting smaller deviations. We also use NRMSE to provide a dimensionless measure of error that allows model performance to be compared fairly across different concentrations. MAE measured the average absolute difference between observed and predicted values, providing an interpretable measure of accuracy independent of error direction. We also calculated sMAPE because it expresses errors as a bounded percentage relative to both observed and predicted values, making performance more comparable across different concentration ranges and less sensitive to extreme values. MBE provides the average bias in the predictions, where positive or negative values indicated overestimation or underestimation, respectively. IOA indicates the overall level of agreement (from -1 to 1) between reference measurements and predicted values, with 1 denoting perfect agreement (ideal model performance), 0 with no agreement (predictions no better than simply predicting the observed average), and -1 with complete disagreement or systematic inverse relationship (Willmott et al., 2011). R^2 (values in [0, 1]) indicates the proportion of the variance in the reference measurements explained by the model, with values closer to 1 indicating a stronger linear association. The formulas are represented below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2} \quad (1)$$

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2}}{\bar{o}} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|o_i - p_i| \quad (~~2~~3)$$

$$\text{sMAPE} = \frac{100}{n}\sum_{i=1}^{n}\frac{2|o_i - p_i|}{|o_i| + |p_i|} \quad (4)$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(o_i - p_i) \quad (5)$$

$$IOA = \begin{cases} 1 - \frac{\sum_{i=1}^{n}|p_i - o_i|}{c\sum_{i=1}^{n}|o_i - \bar{o}|}, \ when \ \sum_{i=1}^{n}|p_i - o_i| \le c\sum_{i=1}^{n}|o_i - \bar{o}| \\ \frac{c\sum_{i=1}^{n}|o_i - \bar{o}|}{\sum_{i=1}^{n}|p_i - o_i|} - 1, when \ \sum_{i=1}^{n}|p_i - o_i| > c\sum_{i=1}^{n}|o_i - \bar{o}| \end{cases} \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sqrt{\sum_{i=1}^{n}(o_i - \bar{o})^2}} \quad (7)$$

here o_i denotes the i-th value from the reference dataset, p_i is the i-th predicted value from the calibration models, n represents the total number of data points in the dataset, and $\bar{o}$ is the arithmetic average of all reference measurements.

After training the model, AutoML ranks candidates on its leaderboard by the RMSE obtained from k-fold cross-validation on the training set (Table S1). The highest-ranked model (Leader Rank 1) is therefore the model with the smallest cross-validated RMSE among all candidates. We adopt this criterion to (i) keep the 20% test set independent of model selection (avoiding optimistic bias), (ii) obtain a more stable, lower-variance estimate by averaging errors across folds rather than relying on a single split, and (iii) prioritize a loss that penalizes large deviations, which is appropriate for $PM_{2.5}$ calibration (RMSE in $\mu g\ m^{-3}$). After selection, all performance reported in the Results refers to the independent test set. ~~Among all candidate models, stacked ensemble models show superior stability and predictive accuracy and was therefore selected as the final calibration model in this study (Table S1).~~

2.3 Calibration Procedure

To ensure reproducible calibration of the low-cost sensors against the Fidas 200, we first established a three-step protocol that accounts for variability among sensor units while maintaining consistency with reference measurements. The approach is designed to be scalable for large sensor networks in real-world indoor monitoring applications. The key steps include:

(1) **Field sensor-to-"Drift–reference sensor" calibration (f2d).** A subset of five sensors from each sensor type (AtmoCube and AirGradient ONE) was randomly selected to serve as "drift–reference sensors". These drift–reference sensors were used exclusively for calibration purposes and were not deployed for field indoor monitoring. The remaining sensors, referred to as "field sensors", were intended for operational deployment. We employed AutoML to develop calibration models that map the field sensors' raw readings to the corresponding averaged measurements of the drift–reference sensors at each time step:

$$\widehat{d_j}(t) = \mathcal{F}_j^{f2d}\left(x_j(t)\right) \quad (8)$$

$$\mathcal{F}_j^{f2d} = \arg\min_{f\in\mathcal{F}}\sum_{t=1}^{N}\left[f\left(x_j(t)\right) - \bar{d}(t)\right]^2 \quad (9)$$

$$x_j(t) = \left[s_j(t), T_j(t), RH_j(t)\right]^T \quad (\text{~~8~~10})$$

$$\bar{d}(t) = \sum_{k=1}^{K} d_k(t) \quad (\text{~~9~~11})$$

where $\hat{d}_j(t)$ is calibrated PM concentration for field sensor j (1, …, M) at a time index of calibration record t (1, …, N); $x_j(t)$ represents raw sensor reading, temperature, and RH; $\bar{d}(t)$ denotes mean of K(=5) drift–reference sensors; and $\mathcal{F}_j^{f2d}$ represent best-performing model chosen for sensor j (GBM in this study) from pool of AutoML candidate models $\mathcal{F}$ during this f2d process. Note that here $T_j(t)$ and $RH_j(t)$ should be calibrated against averaged values of the drift–reference sensors using a simple univariate transfer function before being used as input features.

(2) **"Drift–reference sensor" to "Reference instrument" calibration (d2r).** The averaged readings from drift–reference sensors were calibrated against Fidas 200 following similar procedure above:

$$\hat{r}(t) = \mathcal{F}^{d2r}\left(z(t)\right) \quad (\text{~~10~~12})$$

$$\mathcal{F}^{d2r} = \arg\min_{f \in \mathcal{F}} \sum_{t=1}^{N} \left[f\left(z(t)\right) - r(t)\right]^2 \quad (\text{~~11~~13})$$

$$z(t) = \left[\bar{d}(t), \bar{T}(t), \overline{RH}(t)\right]^T \quad (\text{~~12~~14})$$

here $\hat{r}(t)$ represents calibrated PM concentration for drift–reference sensors; $r(t)$ is PM concentration measured by the reference instruments (Fidas 200); $z(t)$ represents a vector of $\bar{d}(t)$, and calibrated $\bar{T}(t)$ and $\overline{RH}(t)$ (against Fidas 200); and $\mathcal{F}^{d2r}$ denotes best-performing model for the d2r calibration.

~~O~~Our exploratory analysis (Fig. S1) revealed a clear threshold at 50 µg m^{-3} where the sensor bias flips. We chose this value because the scatter plot of sensor versus reference measurements shows two distinct regimes relative to the 1:1 line. At or below 50 µg m^{-3}, the data cloud is tight and lies mostly above the 1:1 line, which indicates a positive sensor bias (overestimation) at low concentrations. Conversely, above 50 µg m^{-3} the cloud shifts below the 1:1 line, and the fitted trend becomes flatter than the 1:1 reference, a pattern consistent with signal compression and underestimation at higher particle loads. This split is further justified by the data distribution; most data lie below about 25 µg m^{-3}, with only a small number of points between 25 and 100 µg m^{-3}. A split at 50 µg m^{-3} produces two interpretable regimes that align with the observed change in bias, keeps the rare high-concentration events together, and avoids slicing the dense background data into very small groups, which would reduce model stability. Therefore, we applied a stratified calibration strategy, training separate AutoML models for the low (<50 µg m^{-3}) and high (50–600 µg m^{-3}) regimes in both the field-to-drift (f2d) and drift-to-reference (d2r) stages. This allows us to tailor the calibration to the specific bias profile of each regime and thereby minimises systematic error across the sensor's full operating range.

~~ur exploratory analysis (Fig. S1) revealed a clear threshold at 50 µg m^{-3} where the sensor bias flips. The sensors tend to overestimate Fidas 200 measurements below but underestimate them above the threshold. Therefore, we applied stratified calibration strategy, training separate AutoML models for the low (<50 µg m^{-3}) and high (50–600 µg m^{-3}) regimes in both the field-to-drift (f2d) and drift-to-reference (d2r) stages. It allows us to tailor the calibration to the specific bias profile of each regime and thereby minimises systematic error across the sensor's full operating range.~~

(3) Field sensor-to-"Reference instrument" calibration (f2r). For every time stamp t, the field sensor's raw reading is first converted to a drift–reference proxy as in Step (1) f2d. That proxy, combined with calibrated temperature and RH (against Fidas 200), is then fed into the calibration models in Step (2) d2r to calculate concentrations directly comparable to the reference dataset:

$$\widetilde{r_j}(t) = \mathcal{H}_j\left(x_j(t)\right) \equiv \left(\mathcal{F}^{d2r} \circ \mathcal{F}_j^{f2d}\right)\left(x_j(t)\right) \quad (\text{~~13~~}15)$$

where $\widetilde{r_j}(t)$ denotes final PM concentration of sensor j aligned to Fidas 200; and $\mathcal{H}_j$ represents shorthand for the overall transfer function $\mathcal{F}^{d2r} \circ \mathcal{F}_j^{f2d}$.

The sensor performance drift over long deployments, the calibration derived pre-deployment gradually becomes less reliable. After retrieval we therefore rebuild the f2d and d2r models with the post-deployment dataset, obtaining a second set of predictions $\widetilde{r_j'}(t)$. For any timestamp t within the deployment period $0 \le t \le D$ (with D the total duration), we fuse the two predictions with a simple linear weight that shifts emphasis from the pre- to the post-deployment model:

$$r_j*(t) = \left(1 - \frac{t}{D}\right) \times \widetilde{r_j}(t) + \frac{t}{D} \times \widetilde{r_j'}(t) \quad (\text{~~14~~}16)$$

thus, $r_j*(t)$ equals the pre-deployment estimate at the campaign start ($t = 0$), the post-deployment estimate at the end ($t = D$), and a smoothly blended value in between, providing a first order correction for drift.

The overall calibration framework is shown schematically in Fig. 2.

3 Results and Discussions

3.1 Low-cost sensor raw readings

Figure 3 compares the timeseries responses of the two sensor types, from AirGradient ONE and AtmoCube to indoor emission events. During the combustion episodes (cigarette smoking and incense-burning) that occurred between 12th and 22nd October 2024, the AirGradient ONE sensors repeatedly recorded uncalibrated $PM_{2.5}$ concentrations exceeding 500 µg m^{-3}, and all units tracked those peaks almost identically, showing high intra-sensor coherence and a high sensitivity to combustion-derived particles. The AtmoCube sensors followed the same temporal pattern but with systematically lower maximum concentrations compared to the AirGradient ONE sensors, with peak readings **between 400 and 500 µg m^{-3}** ~~between 400 to 500 µg m^{-3}~~. Cooking activities generated far lower PM concentrations. Routine meal preparation produced brief excursions of ~30 µg m^{-3} on both sensor types, while a single spike of 80 µg m^{-3} on 30th October consistent with braise and fry high-fat foods that known to generate abundant aerosols (Xu et al., 2024). Therefore, although both AirGradient ONE and AtmoCube sensors correctly identified the timing of each emission episode, AirGradient ONE consistently reported higher absolute concentrations, particularly for the most intense combustion plumes than those of AtmoCube sensors.

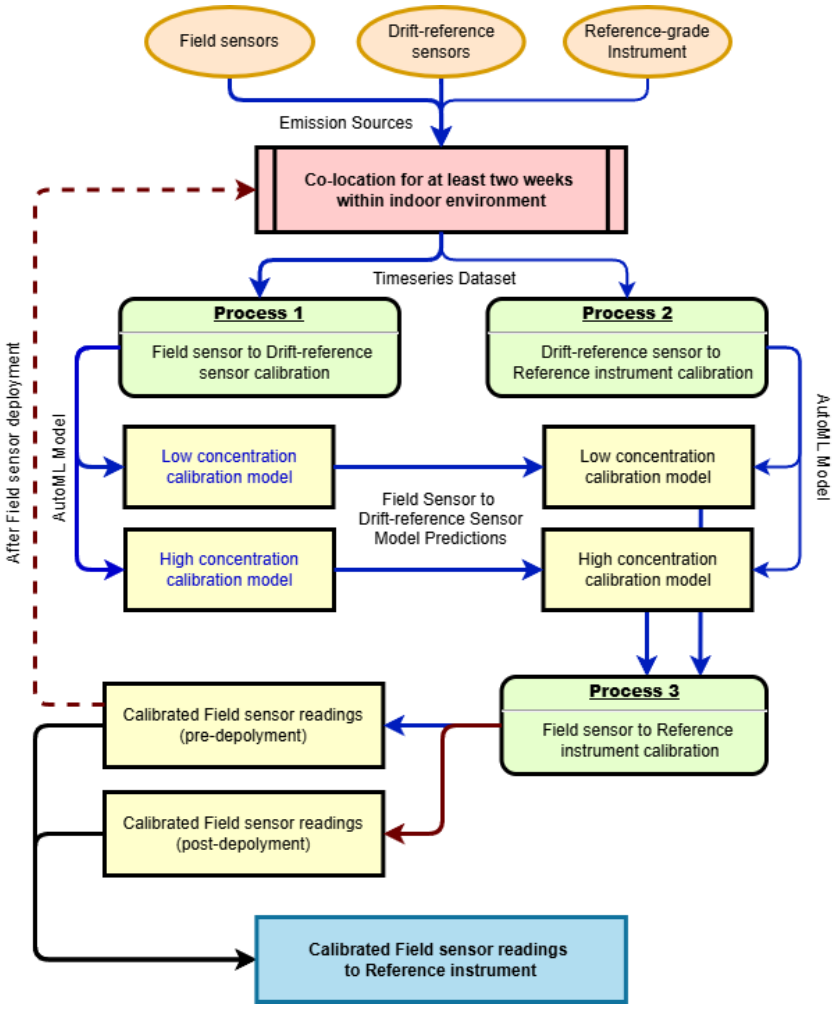
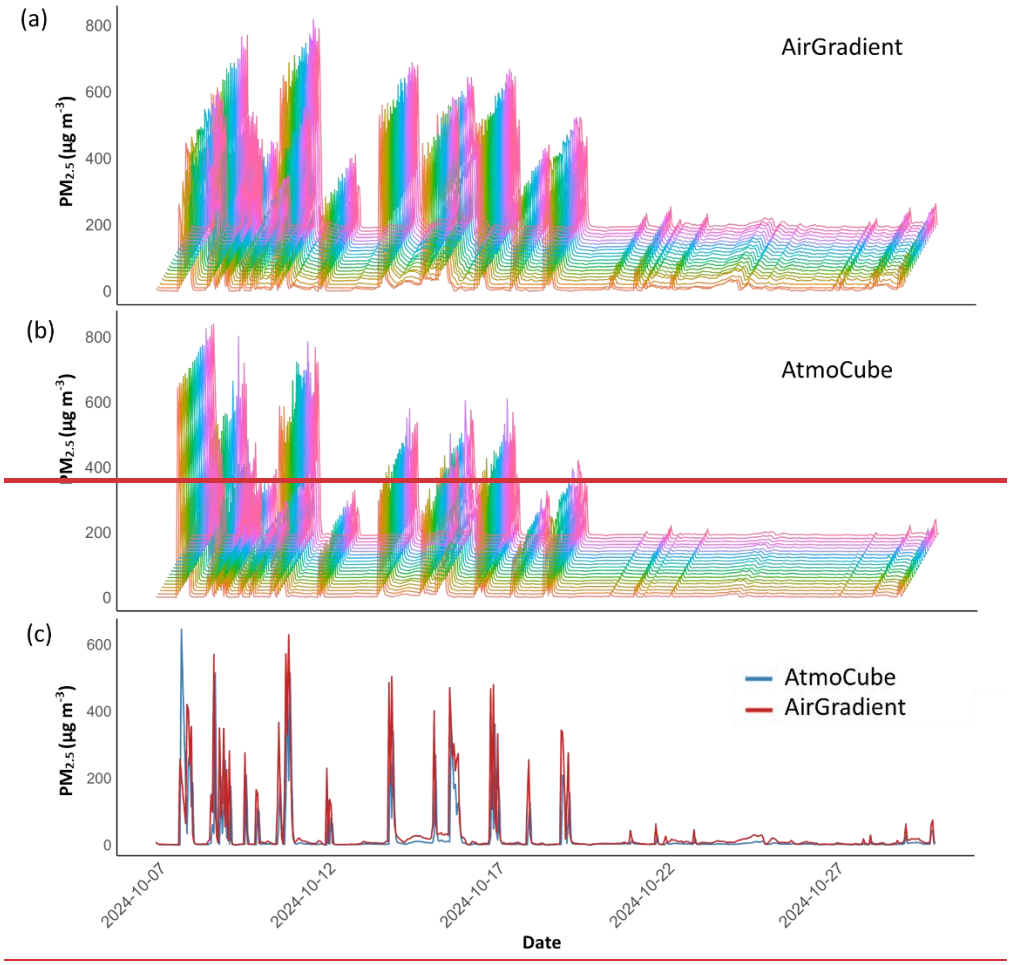


Figure 2: Flowchart of the indoor air quality sensor calibration strategy. The flowchart used a fixed colour scheme to distinguish the two stages of the workflow. Blue arrows and lines represent the main training and prediction path that spans Processes 1–3. The brown arrows and lines represent the post-deployment recalibration path, which is executed after sensor retrieval to correct drift using the post-deployment dataset. The resulting predictions are passed through Process 3 to obtain calibrated readings mapped to the reference instrument.

The inter-type relationship is summarised in Fig. S2, showing the averaged drift-reference $PM_{2.5}$ measurements from AirGradient ONE and AtmoCube. At concentrations below ~50 µg m^{-3} (hereafter denotes as "below-50") (Fig. S2(a)), AirGradient ONE readings lay predominately above the 1:1 reference line, showing a positive bias relative to AtmoCube sensors. Once concentrations exceeded ~50 µg m^{-3} (denotes as "above-50") (Fig. S2(b)), this coherence vanished and the paired data became more scattered, indicating that the two sensor types diverge progressively with increasing particle load.

610 Calibration that reconciles these type- (brand) specific sensitivities is therefore essential for any application that requires accurate absolute $PM_{2.5}$ values.

Sensor-measured environmental parameters exhibited similar systematic offsets (Fig. S3 for temperature and Fig. S4 for RH). Throughout the calibration, AirGradient ONE temperatures were 1.2–1.8°C higher than those from AtmoCube (Fig. S3(a) and
615 S3(b)), where paired data cluster above the identity line (slope=1.01, R^2=0.94). AirGradient ONE measured 4–7 % lower than AtmoCube sensors for RH maxima, whereas at minima AirGradient ONE read 3–5 % higher, as in Fig. S4(a) and S4(b). Intra-type variability reached ~2°C for AirGradient ONE sensors but was ≤1.5°C for AtmoCube sensors, and both types recorded the same diurnal trend (Fig. S3(c) and S3(d)). RH measurements ranged from 47% to 89% (Fig. S4(c) for AirGradient ONE and 4(d) for AtmoCube). AirGradient ONE sensors exhibited tighter clustering (intra-type variability ≤5%) than AtmoCube
620 (≤10%), but they showed a systematic pattern.


(a)
AirGradient
(b)
AtmoCube
(c)
AtmoCube
AirGradient
PM2.5 (µg m-3)
0
200
400
600
800
2024-10-07
2024-10-12
2024-10-17
2024-10-22
2024-10-27
Date

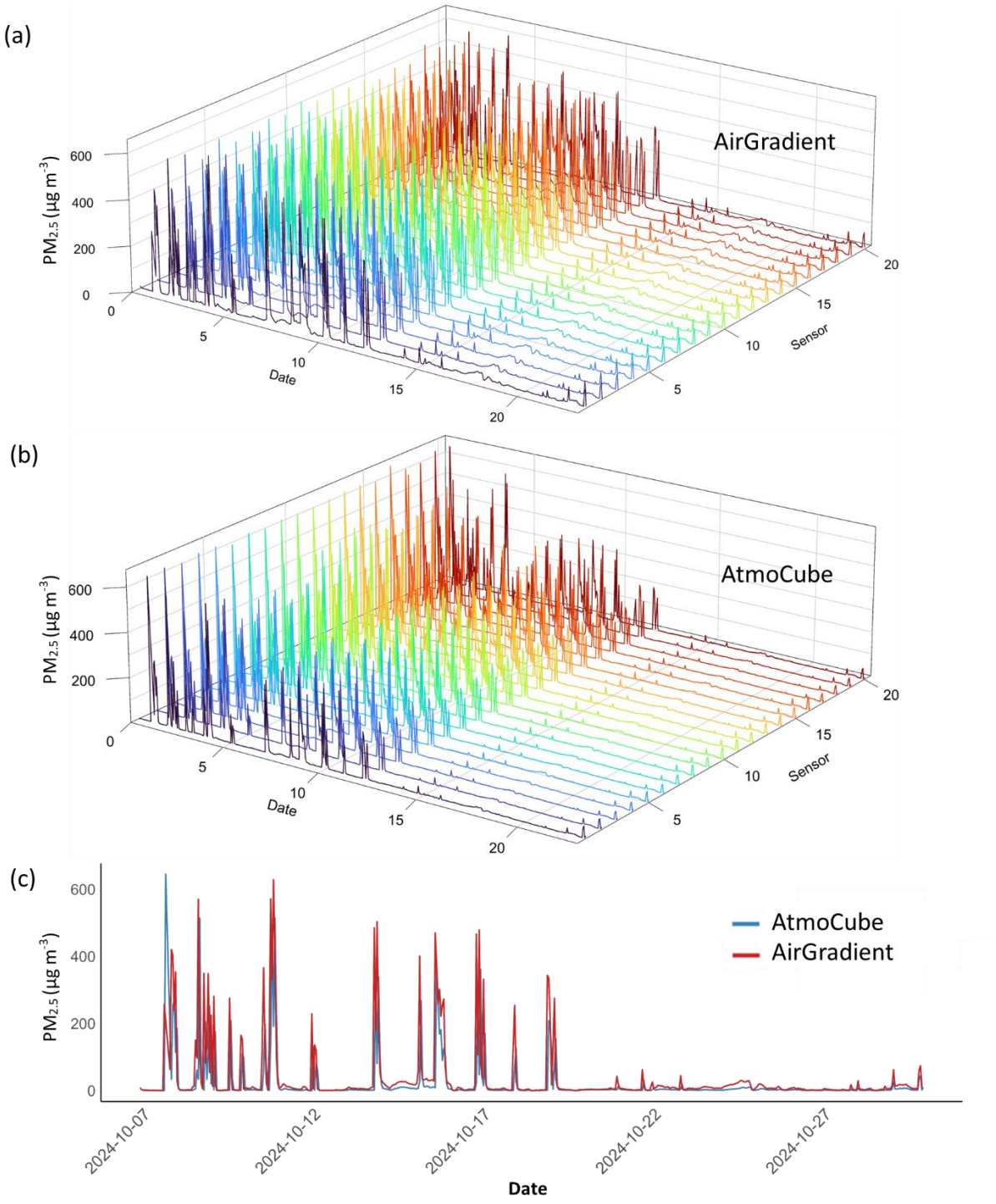


Figure 3: Timeseries of (a) $PM_{2.5}$ from AirGradient ONE sensors, (b) $PM_{2.5}$ from AtmoCube sensors, and (c) Averaged concentration from drift-reference sensors.

3.2 Raw readings from drift–reference sensors vs. Fidas 200 measurements

Figures 4(a) and 4(c) shows scatter plots of raw and calibrated averaged $PM_{2.5}$ concentrations from AirGradient ONE and AtmoCube drift-reference sensors against the Fidas 200 measurements in the below-50 regime, representative of relatively low air pollution. Before calibration, both AirGradient ONE and AtmoCube sensors exhibited moderate linear correlations with the Fidas 200, with R^2 values of 0.65 for the AirGradient ONE and 0.57 for the AtmoCube, respectively (Table 1). Although both sensor types clustered close to the 1:1 reference line, their slopes reveal systematic biases. AirGradient ONE readings lay predominately above the line with a regression slope of 1.57, producing an average 20% overestimation relative to the Fidas 200, while AtmoCube readings fell below with a slope of 0.64, corresponding to a 55.6% underestimation. Extending the analysis to the above-50 regime (Figs. 4(b) and 4(d)) highlights further divergence. Here, AirGradient ONE sensors had a stronger correlation with the reference (R^2=0.78), but its slope decreased to 0.82, reflecting a slight 3.1% underestimation during high pollution episodes. In contrast, AtmoCube sensors had a lower slope of 0.50 and an R^2 of 0.64, showing a substantial 38.8% underestimation. Therefore, both types of sensor experience signal compression at higher particle loads, yet the magnitude of this non-linearity is sensor specific.

RH can significantly influence the measurement accuracy of particles from indoor air quality sensors (Fig. S5). For AirGradient ONE (Fig. S5(a) and S5(b)), $PM_{2.5}$ readings above the 1:1 reference line at low concentrations consistently associated with periods of high RH, implying that hygroscopic growth of particles at high humidity is a primary driver of AirGradient ONE's low end overestimation (Liang, 2021). Conversely, AtmoCube showed no systematic RH pattern (Fig. S5(c) and S5(d)); its scatter remained broadly uniform across the humidity spectrum, indicating lower RH sensitivity. This disparity may reflect differences in internal RH-compensation algorithms implemented by each manufacturer.

Table 1: Statistical performance of raw and calibrated AirGradient ONE and AtmoCube drift–reference sensors relative to the Fidas 200 measurements for $PM_{2.5}$, stratified by concentration regime (below-50, above-50) and for the combined dataset.

Sensor	Subset	Stage	n (sample size)	R^2	RMSE (NRMSE)	MAE (sMAPE)	MBE	IOA
AirGradient ONE	Below 50 µg m^{-3}	Raw	483	0.65	6.4 (98.5)	3.7 (46.1)	-1.8	0.49
		Calibrated	483	0.69	3.8 (32.6)	1.5 (22.8)	-0.1	0.80
	Above 50 µg m^{-3}	Raw	64	0.78	91.3 (32.5)	69.6 (36.5)	-40.9	0.80
		Calibrated	64	0.92	59 (23.9)	44.6 (31.6)	-45.4	0.87
	All concentration range	Raw	547	0.95	31.8 (82.3)	11.4 (44.9)	-3.2	0.90
		Calibrated	547	0.97	20.5 (59.6)	6.5 (23.8)	-0.6	0.94
AtmoCube	Below 50 µg m^{-3}	Raw	499	0.57	12.4 (140.3)	5.1 (82.2)	-0.044.89	0.63
		Calibrated	499	0.80	7.4 (122)	2.8 (80.8)	0.03-0.27	0.79

Above 50 µg m^{-3}	Raw	48	0.64	182.7 (52.5)	160.5 (62.6)	-150.8	0.48
	Calibrated	48	0.76	91.1 (23.3)	72.3 (25.5)	~~2.4~~-27.7	0.76
All concentration range	Raw	547	0.90	55.4 (143)	18.7 (80.4)	-17.7	0.84
	Calibrated	547	0.94	27.9 (67.7)	8.9 (75.9)	~~0.3~~-2.68	0.92

3.3 Calibrated readings from drift-reference sensor vs. Fidas 200 measurements

In the below-50 regime, calibrated AirGradient ONE drift-reference readings show slight stronger correlation with Fidas 200 measurements (R^2=0.69) compared to their raw values (Fig. 4(a)), and errors are relatively small and have been improved (NRMSE=~~3.8 µg m^{-3}~~32.6%, ~~MAE~~s~~S~~MAPE=~~1.5 µg m^{-3}~~22.8%) as shown in Table 1. The residuals present negligible systematic bias (MBE=-0.1 µg m^{-3}), indicating great improvements from systematic overestimation under low $PM_{2.5}$ concentration before calibration. After calibration, the sensor performance meets the recommended criteria of $R^2 \geq 0.70$ and RMSE ≤ 7 µg m^{-3} (Zamora et al., 2022). At above-50 concentrations (Fig. 4(b)), the improvement in the performance of calibrated AirGradient ONE sensors was even more significant, with R^2 and IOA achieving about 0.92 and 0.87, respectively. ~~Absolute errors increase~~The errors also ~~has~~have been improved (NRMSE = ~~59 µg m^{-3}~~23.9%, ~~MAE~~sMAPE = ~~44.6 µg m^{-3}~~31.6%) as expected ~~but~~ and remain proportionally reasonable (e.g., ~10% uncertainty at 600 µg m^{-3}). A slight negative bias (MBE = -~~5~~4.4 µg m^{-3}) indicating a small underestimation tendency at extreme high concentrations, but high IOA value (0.87) show accurate tracking of both timing and magnitude.

Figure S6 show the impact of RH on calibrated readings of AirGradient ONE sensors for the below-50 (Fig. S6(a)) and the above-50 (Fig. S6(b)) concentration regimes, respectively. Across both concentration ranges the residuals show no systematic humidity bias, indicating that the AutoML model (using RH and temperature as covariates) mitigated hygroscopic growth influences that typically inflate optical counts above 70–80% RH (Ko et al., 2024). The small scatter evident at extreme high RH levels likely reflects limited training data but does not compromise agreement with the reference, corroborating reports that RH-aware calibration can suppress sensor error by around 20% (Liang, 2021).

Calibration likewise improved AtmoCube agreement with the Fidas 200 across the full concentration range (Figs. 4(c) and 4(d)). Overall AtmoCube sensors achieved R^2=0.94 and IOA=0.92 (Table 1). In the below-50 clean air conditions, the calibrated AtmoCube sensors have R^2=0.80, and such slightly lower correlation relative to those of high pollution levels is expected as sensor signals approach the noise floor at very low pollution levels (Johnson et al., 2018). RMSE (7.4 µg m^{-3}) and MAE (2.8 µg m^{-3}) are relatively small, and the mean bias is negligible, indicating that the calibration mitigates the pronounced low-end under-reading observed pre-calibration. At high $PM_{2.5}$ levels, calibrated AtmoCube sensors still show good agreement with Fidas 200 as data points distribute along the 1:1 line but with slightly reduced R^2 (0.76). A possible explanation is that at very high particle loading the sensor's optical detector response starts to become non-linear or approaches a saturation point (Kelly et al., 2017), introducing larger random errors. The residual bias is minor (MBE=~~2.4~~27.7 µg m^{-3}), indicating a small

over-read under very high pollution. Figure S6(c–d) shows that, after calibration, AtmoCube residuals remain almost flat across the full RH ranges in both low and high concentration regimes. Even during episodes exceeding 80 % RH, no coherent over- or under-reading trend was found, indicating that the calibration has effectively reduced humidity interference.

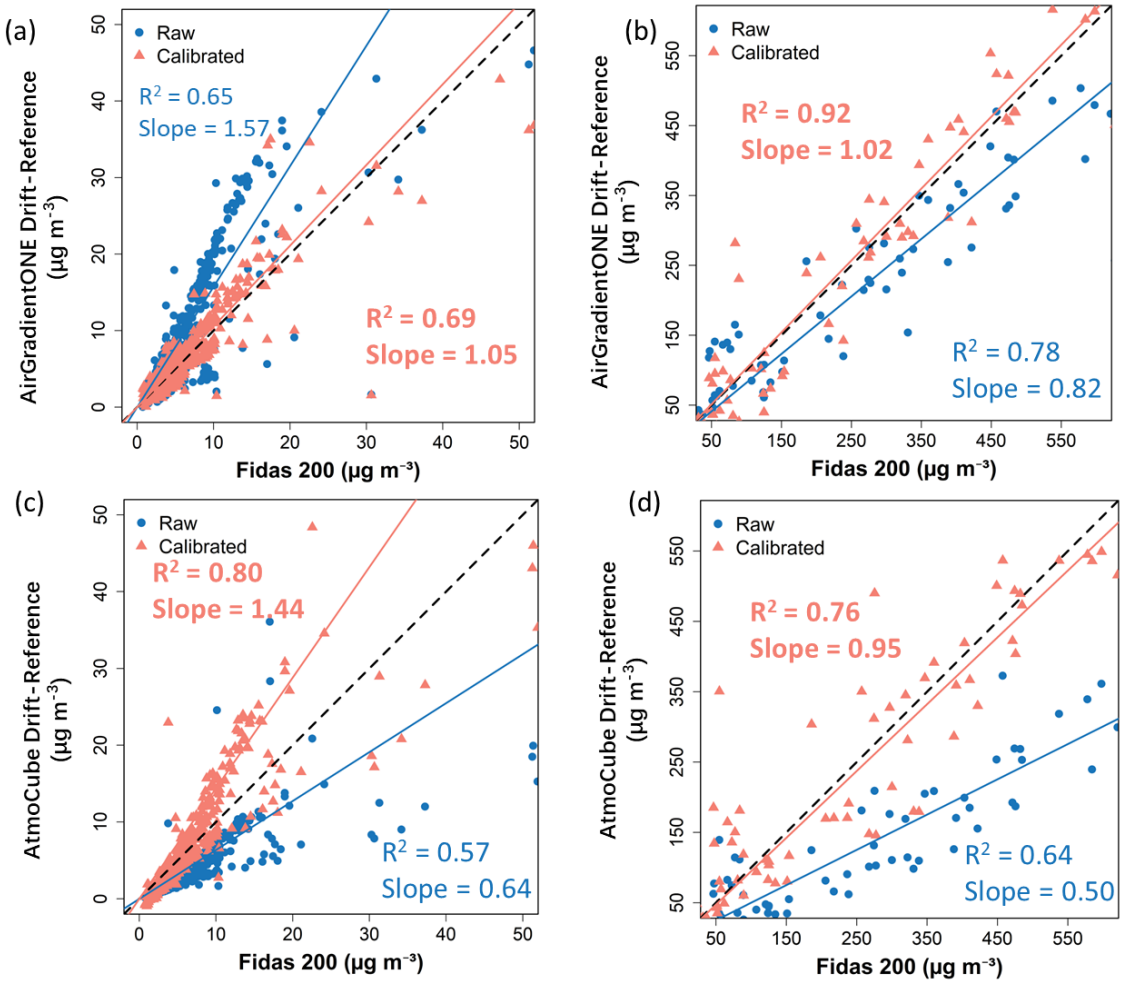


Figure 4: Raw and calibrated $PM_{2.5}$ of drift-reference sensors compared with the Fidas 200 measurements, (a) AirGradient ONE sensors within below-50 regime; (b) AirGradient ONE sensors within above-50 regime; (c) AtmoCube sensors within below-50 regime; (d) AtmoCube sensors within above-50 regime.

3.4 Calibrated readings from field sensors vs. Fidas 200 measurements

The multi-stage calibration strategy effectively improved the performance of field sensors against the reference-grade instrument Fidas 200 (Fig. 5 and Table 2). Within the below-50 regime, AirGradient ONE sensors showed a RMSE of 4 µg m^{-3} and MAE of 1.70 µg m^{-3}, and their correlation R^2 increased from 0.45 to 0.64. By contrast, AtmoCube sensors achieved a

stronger linear match (R^2=0.80) despite relatively higher residual scatter (RMSE=7.5 µg m^{-3}) (Fig. 5(c)), consistent with their finer baseline sensitivity to subtle particulate variations. Performance at above-50 concentration regime indicated that both types of indoor air quality sensor synchronised well with the timing of pollution events while their error signatures differed. AirGradient ONE sensors showed moderate overestimation (MBE=3.9 µg m^{-3}, RMSE=67.1 µg m^{-3}, NRMSE=23.9%), while AtmoCube sensors displayed ~~similar~~ higher systematic bias (MBE=~~3.7~~28.6 µg m^{-3}) ~~but~~ and higher variability (RMSE=91.5 µg m^{-3}, NRMSE=24.5%). These differences may arise from different sensor components, for example, AtmoCube units employed shorter optical path length and proprietary firmware averaging while AirGradient ONE sensors used longer path and raw count reporting of the Plantower PMS5003. Importantly, our calibration strategy reconciled hardware-driven disparities between sensor types. Both types of sensors agreed well with Fidas 200 measurements after calibration, with IOA increasing from 0.90 to 0.94 for AirGradient ONE and from 0.84 to 0.92 for AtmoCube sensors.

To evaluate the multi-step calibration strategy itself rather than the choice of models, we compared AutoML models with multivariate regressions (Fig. S7). Figure S7~~(a)~~ and Figures S8 show~~s~~ that AutoML models produced better performance statistics, showing enhanced predictive accuracy and reliability, particularly when evaluating error distribution across different $PM_{2.5}$ concentration regimes. ~~MAE (Fig. S7(b)) reduced by 15-40% across different concentration ranges, with the largest improvement happened in the 25-50 µg m^{-3}.~~ Such improvements could be due to the ability of AutoML to incorporate interaction terms (RH, temperature) that influence the sensor light-scattering response (Liang, 2021). However, there is only one exception for AtmoCube sensors in the over 100 µg m^{-3}, in which the linear model has a smaller sMAPE. This is might due to the limited number of data in the high concentration range.

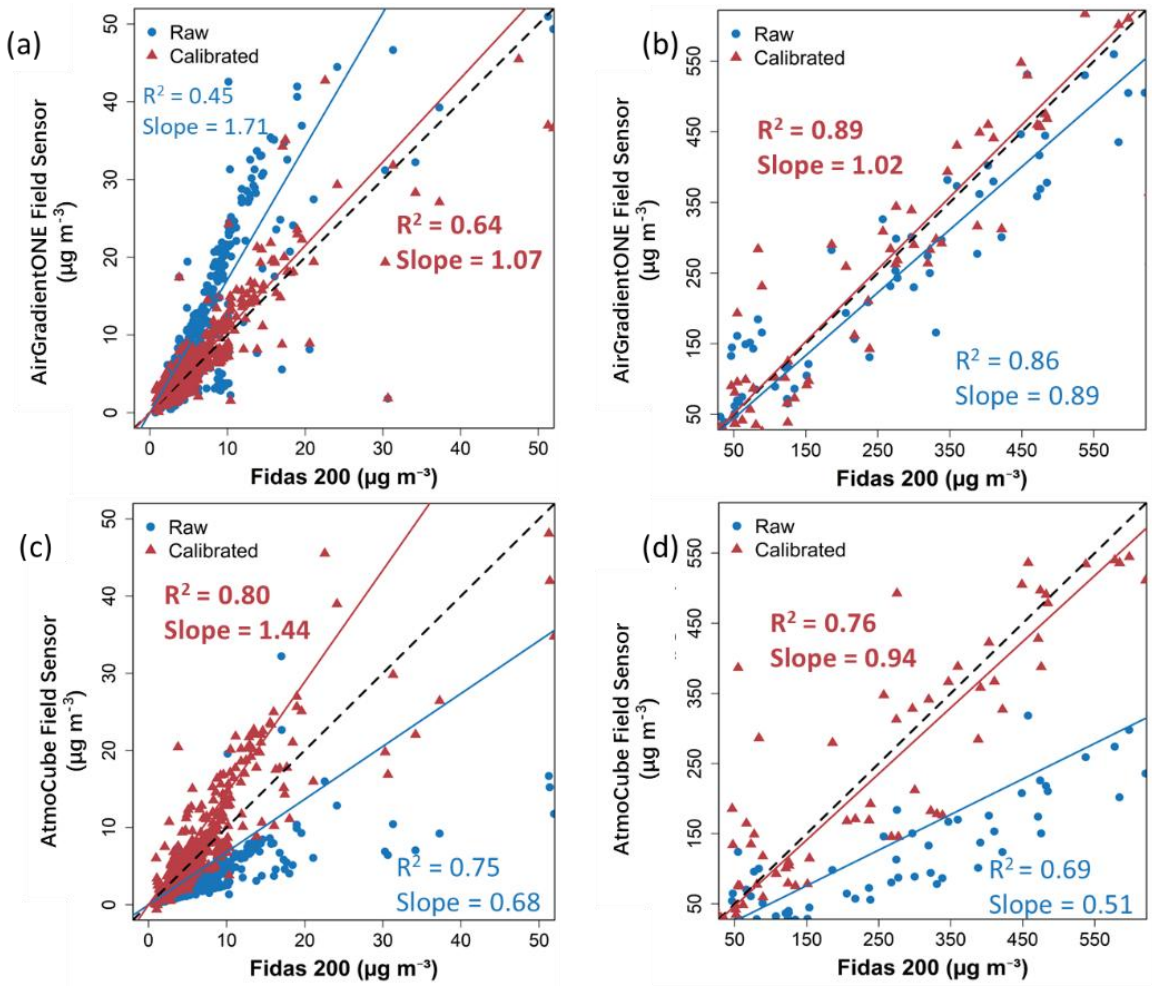


Figure 5: Raw and calibrated $PM_{2.5}$ of field sensors compared with the Fidas 200 measurements, (a) AirGradient ONE sensors within below-50 regime; (b) AirGradient ONE sensors within above-50 regime; (c) AtmoCube sensors within below-50 regime; (d) AtmoCube sensors within above-50 regime.

Table 2: Statistical performance of raw and calibrated AirGradient ONE and AtmoCube field sensors relative to the Fidas 200 measurements for $PM_{2.5}$, stratified by concentration regime (below-50, above-50) and for the combined dataset.

Sensor	Subset	Stage	n	R^2	RMSE (NRMSE)	MAE (sMAPE)	MBE	IOA
AirGradient ONE	Below 50 µg m^{-3}	Raw	483	0.45	7.3 (102)	4.3 (44.6)	2.32.02	0.41
		Calibrated	483	0.64	4 (60.9)	1.7 (26.9)	0.1	0.77
	Above 50 µg m^{-3}	Raw	64	0.86	83.1 (33.8)	62.2 (37.9)	-22.242.5	0.82
		Calibrated	64	0.89	67.1 (23.9)	48.7 (29.1)	3.9	0.86
	All range	Raw	547	0.94	29.2 (85.6)	11 (43.8)	-3.190.6	0.90
		Calibrated	547	0.96	23.3 (60.2)	7.2 (27.1)	0.5	0.94
AtmoCube	Below 50 µg m^{-3}	Raw	499	0.75	12.4 (141.6)	4.9 (77.4)	-4.617.8	0.64
		Calibrated	499	0.80	7.5 (82.1)	3.6 (22.7)	0.2-0.15	0.77
	Above 50 µg m^{-3}	Raw	48	0.69	180.3 (53.5)	158.2 (64.1)	-15247	0.48
		Calibrated	48	0.76	91.5 (24.5)	72.6 (26.7)	3.7-28.6	0.74
	All concentration range	Raw	547	0.88	54.7 (146)	18.3 (76.2)	-17.81	0.84
		Calibrated	547	0.94	28.1 (67.9)	9.6 (23.1)	0.5-2.65	0.92

3.5 Limitations and implications

Our framework significantly improved the low-cost sensors performance under different concentrations. But there are still some limitations, and further research is needed on the generalizability of the model and calibration strategies. First, the training data were collected in a single experimental container under temperate-climate humidity (with RH between 45–85%) and may not capture sensor behaviour in very moist interiors. Second, the present study did not capture every indoor emission source, particularly those with moderate emission levels. We do not know whether the sensors will be sensitive to particle types (e.g., particles from different sources). Third, evaluating sensor drift demands the months-to-years timescales of real deployments and was not evaluated. Future work should gather data from warmer, high-humidity homes to capture sensor behaviour at elevated RH conditions, consider additional moderate emission sources such as off-gassing materials, and run multi-year field trials to quantify drift and test automated recalibration. These steps will increase the robustness and evaluate long-term accuracy of the calibration strategy. However, the thresholds delineating “low” and “high” categories are derived from empirical observations within the analysed dataset. Accordingly, researchers are encouraged to initially assess their own data and adapt this strategy as necessary to ensure its applicability.

The implications of our findings are significant for atmospheric science and indoor air quality management, especially in the context of the growing use of low-cost sensors for exposure assessment and public health applications. By showing that inexpensive sensors can be calibrated to yield high-quality data indoors, this study helps bridge the important gap between indoor and outdoor air pollution monitoring. Furthermore, the application of AutoML in sensor calibration showcases the value

of advanced data-driven techniques in atmospheric measurements. AutoML could be used to periodically re-calibrate hundreds of sensors automatically as new reference data become available, maintaining network accuracy with minimal human intervention. This is particularly relevant for community science projects or indoor air quality campaigns where resources for manual calibration are limited. By improving the reliability of indoor air measurements, the study contributes to a future where continuous indoor air quality monitoring is feasible on a large scale, driving better-informed strategies to safeguard public health in the spaces where people live and work.

The regime thresholds used in this study were derived empirically from our indoor dataset and should not be assumed for other indoor cases, outdoors, or for other pollutants. Users should re-estimate cut points from their own co-located data and retrain the staged models with environment appropriate features.

4 Summary

In this work, we introduced an automated machine learning (AutoML) calibration framework for enhancing the performance of low-cost indoor air quality sensors. The AutoML-calibrated sensors met or exceeded study objectives by significantly improving measurement accuracy for fine particles ($PM_{2.5}$) across all concentration regimes. The multi-stage calibration workflow achieved tight agreement with reference measurements (from Fidas 200), evidenced by substantial increases in coefficient of determination (R^2) and reductions in error metrics. In the low-concentration regime (below 50 µg m^{-3}), R^2 improved from moderate values (~0.6 pre-calibration) to approximately 0.85 post-calibration, with root-mean-square error (RMSE) dropping by roughly half (e.g., from ~5 to ~3 µg m^{-3}), as well as the NRMSE. At higher concentrations (above 50 µg m^{-3}), gains were even more pronounced, with R^2 approaching or exceeding 0.90 (near reference-grade performance) and RMSE falling from tens of µg m^{-3} to single digits. Similarly, mean absolute error (MAE) and symmetric mean absolute percentage error (sMAPE) declined markedly, and mean bias error (MBE) was effectively eliminated, shifting from significant systematic biases (e.g., 5–10 µg m^{-3} over- or underestimation) to nearly zero. These results show that the calibrated sensors reliably resolve indoor particulate levels at background concentrations and during elevated pollution events, closely tracking the reference instrument across the full range. These findings confirm that our multistage calibration effectively eliminated sensor bias under varied indoor conditions and emission sources. The initial stage corrected baseline drift. Subsequent stages used AutoML to address scatter caused by relative humidity and nonlinear responses at high particle concentrations. These factors are often overlooked in simpler methods. AutoML efficiently selected the best models for each phase, removed the need for manual tuning, and revealed subtle patterns in the data. By integrating AutoML into a structured multistage process, we achieved robust bias correction across scenarios, yielding accurate, precise measurements well-suited for indoor air quality monitoring.

Author Contributions

Juncheng Qian: Writing – original draft, Writing – review & editing, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Thomas Wynn:** Writing – review & editing. **Bowen Liu:** Writing – review & editing, Supervision. **Yuli Shan:** Writing – review & editing, Supervision. **Suzanne E. Bartington:** Writing – review & editing. **Francis D. Pope:** Writing – review & editing. **Yuqing Dai:** Writing – review & editing, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization, Supervision. **Zongbo Shi:** Conceptualization, Interpretation, Visualization, Writing – review & editing, Supervision.

Acknowledgement:

We would like to thank Joseph Day and Nana Wei, ~~and~~ the wider WM-NetZero research team, and INHABIT research team~~m~~ for their support, including comments on the manuscript. This work was supported by Wellcome Trust grant number: 227150_Z_23_Z, and UKRI-MRC grant number: MR/Z506680/1.

Declaration of competing interest

Some authors are members of the editorial board of journal Atmospheric Measurement Techniques.

Data availability

Data is available at https://github.com/DandE9996/sensor_calibration

References

Aix, M.-L., Schmitz, S., & Bicout, D. J. (2023). Calibration methodology of low-cost sensors for high-quality monitoring of fine particulate matter. *Science of The Total Environment*, *889*, 164063. https://doi.org/10.1016/j.measurement.2024.114529

Cowell, N., Chapman, L., Bloss, W., Srivastava, D., Bartington, S., & Singh, A. (2023). Particulate matter in a lockdown home: evaluation, calibration, results and health risk from an IoT enabled low-cost sensor network for residential air quality monitoring. *Environmental Science: Atmospheres*, *3*(1), 65-84. https://doi.org/10.1039/d2ea00124a

Crilley, L. R., Singh, A., Kramer, L. J., Shaw, M. D., Alam, M. S., Apte, J. S., Bloss, W. J., Hildebrandt Ruiz, L., Fu, P., Fu, W., Gani, S., Gatari, M., Ilyinskaya, E., Lewis, A. C., Ng'ang'a, D., Sun, Y., Whitty, R. C. W., Yue, S., Young, S., & Pope, F. D. (2020). Effect of aerosol composition on the performance of low-cost optical particle counter correction factors. *Atmospheric Measurement Techniques*, *13*(3), 1181-1193. https://doi.org/10.5194/amt-13-1181-2020

Hagan, D. H., & Kroll, J. H. (2020). Assessing the accuracy of low-cost optical particle sensors using a physics-based approach. *Atmospheric Measurement Techniques Discussions*, *2020*, 1-36. https://doi.org/10.5194/amt-13-6343-2020

Johnson, K. K., Bergin, M. H., Russell, A. G., & Hagler, G. S. W. (2018). Field Test of Several Low-Cost Particulate Matter Sensors in High and Low Concentration Urban Environments. *Aerosol Air Qual Res*, *18*(3), 565-578. https://doi.org/10.4209/aaqr.2017.10.0418

Kang, J., & Choi, K. (2024). Calibration methods for low-cost particulate matter sensors considering seasonal variability. *Sensors*, *24*(10), 3023. https://doi.org/10.1016/j.scitotenv.2023.164063

Kelly, K. E., Whitaker, J., Petty, A., Widmer, C., Dybwad, A., Sleeth, D., Martin, R., & Butterfield, A. (2017). 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

Kim, S., Park, S., & Lee, J. (2019). Evaluation of performance of inexpensive laser based PM2. 5 sensor monitors for typical indoor and outdoor hotspots of South Korea. *Applied Sciences*, *9*(9), 1947. https://doi.org/10.3390/app9091947

Ko, K., Cho, S., & Rao, R. R. (2024). Evaluation of calibration performance of a low-cost particulate matter sensor using collocated and distant NO2. *Atmospheric Measurement Techniques*, *17*(10), 3303-3322. https://doi.org/10.5194/amt-17-3303-2024

Koehler, K., Wilks, M., Green, T., Rule, A. M., Zamora, M. L., Buehler, C., Datta, A., Gentner, D. R., Putcha, N., & Hansel, N. N. (2023). Evaluation of calibration approaches for indoor deployments of PurpleAir monitors. *Atmospheric Environment*, *310*, 119944. https://doi.org/10.1016/j.atmosenv.2023.119944

LeDell, E., & Poirier, S. (2020). H2o automl: Scalable automatic machine learning. Proceedings of the AutoML Workshop at ICML,

Levy Zamora, M., Xiong, F., Gentner, D., Kerkez, B., Kohrman-Glaser, J., & Koehler, K. (2018). Field and laboratory evaluations of the low-cost plantower particulate matter sensor. *Environmental science & technology*, *53*(2), 838-849. http://dx.doi.org/10.1021/acs.est.8b05174

Li, J., Li, H., Ma, Y., Wang, Y., Abokifa, A. A., Lu, C., & Biswas, P. (2018). Spatiotemporal distribution of indoor particulate matter concentration with a low-cost sensor network. *Building and Environment*, *127*, 138-147. https://doi.org/10.1016/j.buildenv.2017.11.001

Liang, L. (2021). 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

Liu, H.-Y., Schneider, P., Haugen, R., & Vogt, M. (2019). Performance assessment of a low-cost PM2. 5 sensor for a near four-month period in Oslo, Norway. *Atmosphere*, *10*(2), 41. https://doi.org/10.3390/atmos10020041

Mahajan, S., & Kumar, P. (2020). Evaluation of low-cost sensors for quantitative personal exposure monitoring. *Sustainable Cities and Society*, *57*, 102076. https://doi.org/10.1016/j.scs.2020.102076

Mousavi, A., & Wu, J. (2021). Indoor-generated PM2. 5 during COVID-19 shutdowns across California: application of the PurpleAir indoor–outdoor low-cost sensor network. *Environmental science & technology*, *55*(9), 5648-5656. https://doi.org/10.1021/acs.est.0c06937

- Munir, S., Mayfield, M., Coca, D., Jubb, S. A., & Osammor, O. (2019). Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities—A case study in Sheffield. *Environmental monitoring and assessment*, *191*, 1-22. https://doi.org/10.1007/s10661-019-7231-8

Nowack, P., Konstantinovskiy, L., Gardiner, H., & Cant, J. (2021). Machine learning calibration of low-cost NO 2 and PM 10 sensors: Non-linear algorithms and their impact on site transferability. *Atmospheric Measurement Techniques*, *14*(8), 5637-5655. https://doi.org/10.1016/j.atmosenv.2023.119692

Raysoni, A. U., Pinakana, S. D., Mendez, E., Wladyka, D., Sepielak, K., & Temby, O. (2023). A review of literature on the usage of low-cost sensors to measure particulate matter. *Earth*, *4*(1), 168-186. https://doi.org/10.3390/earth4010009

Villanueva, E., Espezua, S., Castelar, G., Diaz, K., & Ingaroca, E. (2023). Smart multi-sensor calibration of low-cost particulate matter monitors. *Sensors*, *23*(7), 3776. https://doi.org/10.3390/s23073776

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2011). A refined index of model performance. *International Journal of Climatology*, *32*(13), 2088-2094. https://doi.org/10.1002/joc.2419

Xu, X., Hu, K., Zhang, Y., Dong, J., Meng, C., Ma, S., & Liu, Z. (2024). Experimental evaluation of the impact of ventilation on cooking-generated fine particulate matter in a Chinese apartment kitchen and adjacent room. *Environ Pollut*, *348*, 123821. https://doi.org/10.1016/j.envpol.2024.123821

Zaidan, M. A., Motlagh, N. H., Fung, P. L., Khalaf, A. S., Matsumi, Y., Ding, A., Tarkoma, S., Petäjä, T., Kulmala, M., & Hussein, T. (2022). Intelligent air pollution sensors calibration for extreme events and drifts monitoring. *IEEE Transactions on Industrial Informatics*, *19*(2), 1366-1379. https://doi.org/10.1109/TII.2022.3151782

Zamora, M. L., Buehler, C., Lei, H., Datta, A., Xiong, F., Gentner, D. R., & Koehler, K. (2022). Evaluating the Performance of Using Low-Cost Sensors to Calibrate for Cross-Sensitivities in a Multipollutant Network. *AtmoCubeS ES T Eng*, *2*(5), 780-793. https://doi.org/10.1021/acsestengg.1c00367

Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., & Robinson, A. L. (2018b). A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, *11*(1), 291-313. https://doi.org/10.5194/amt-11-291-2018

Zou, Y., Clark, J. D., & May, A. A. (2021). Laboratory evaluation of the effects of particle size and composition on the performance of integrated devices containing Plantower particle sensors. *Aerosol Science and Technology*, *55*(7), 848-858. https://doi.org/10.1080/02786826.2021.1905148

Supplementary Materials

In this file, AirGriadient is denoted as AG, and AtmoCube is denoted as AC.

Table S1. Top 10 AutoML model training statistics for AirGradient sensors

Model_Subset	Model_ID	Test_RMSE	Test_MAE	Test_R^2	Train_RMSE	Train_MAE	Train_R^2	LB_RMSE	LB_MAE	LB_mean_residual_deviance	LB_Rank
Low (<50 µg/m³)	StackedEnsemble_BestOfFamily_1_AutoML_1_20251008_144707	3.452	1.45	0.569	3.668	1.14	0.721	4.562	1.558	20.816	1
Low (<50 µg/m³)	DeepLearning_grid_2_AutoML_1_20251008_144707_model_4	3.414	1.505	0.579	4.185	1.312	0.637	4.607	1.604	21.227	2
Low (<50 µg/m³)	DeepLearning_grid_2_AutoML_1_20251008_144707_model_3	3.673	1.955	0.512	4.825	2.03	0.517	4.642	1.548	21.544	3
Low (<50 µg/m³)	DeepLearning_grid_1_AutoML_1_20251008_144707_model_3	3.765	1.492	0.488	4.682	1.462	0.546	4.649	1.558	21.612	4

Low (<50 µg/m³)	DeepLearning_grid_1_AutoML_1_20251008_144707_model_2	3.54	1.724	0.547	4.535	1.584	0.574	4.673	1.946	21.838	5
Low (<50 µg/m³)	DeepLearning_grid_3_AutoML_1_20251008_144707_model_4	3.491	1.681	0.56	3.937	1.373	0.679	4.683	1.735	21.934	6
Low (<50 µg/m³)	DeepLearning_grid_2_AutoML_1_20251008_144707_model_1	3.59	1.53	0.534	1.352	0.93	0.962	4.685	1.681	21.946	7
Low (<50 µg/m³)	DeepLearning_grid_3_AutoML_1_20251008_144707_model_3	3.583	1.815	0.536	4.691	1.862	0.544	4.689	1.78	21.982	8
Low (<50 µg/m³)	DeepLearning_grid_1_AutoML_1_20251008_144707_model_4	3.553	1.52	0.544	4.532	1.398	0.574	4.698	1.504	22.067	9
Low (<50 µg/m³)	StackedEnsemble_AllModels_1_AutoML_1_20251008_144707	3.41	1.406	0.58	3.463	1.092	0.751	4.749	1.54	22.556	10

High (≥50 µg/m³)	DeepLearning_grid_3_AutoML_2_20251008_145912_model_3	85.74	67.492	0.73	77.232	61.739	0.862	85.468	67.948	7304.795	1
High (≥50 µg/m³)	StackedEnsemble_BestOfFamily_1_AutoML_2_20251008_145912	81.534	62.367	0.755	54.901	45.028	0.93	86.229	65.222	7435.452	2
High (≥50 µg/m³)	DeepLearning_grid_1_AutoML_2_20251008_145912_model_2	65.676	55.946	0.841	73.635	54.655	0.875	86.233	66.143	7436.123	3
High (≥50 µg/m³)	DeepLearning_grid_2_AutoML_2_20251008_145912_model_3	84.987	65.181	0.734	85.45	67.694	0.831	87.139	67.996	7593.221	4
High (≥50 µg/m³)	StackedEnsemble_AllModels_1_AutoML_2_20251008_145912	66.179	56.598	0.839	35.681	29.452	0.971	87.799	66.987	7708.709	5
High (≥50 µg/m³)	DeepLearning_grid_1_AutoML_2_20251008_145912_model_1	95.343	77.827	0.666	43.161	34.733	0.957	88.113	69.388	7763.824	6

High (≥50 µg/m³)	DeepLearning_grid_3_AutoML_2_20251008_145912_model_1	140.436	84.304	0.274	18.061	13.396	0.992	90.677	72.281	8222.326	7
High (≥50 µg/m³)	DeepLearning_grid_1_AutoML_2_20251008_145912_model_3	66.326	51.31	0.838	67.328	50.082	0.895	91.621	70.071	8394.498	8
High (≥50 µg/m³)	DeepLearning_grid_1_AutoML_2_20251008_145912_model_5	112.725	94.296	0.533	90.516	66.322	0.811	92.224	72.416	8505.28	9
High (≥50 µg/m³)	DeepLearning_grid_1_AutoML_2_20251008_145912_model_4	66.452	52.316	0.838	57.167	40.525	0.925	93.509	74.121	8743.909	10

Table S2. Top 10 AutoML model training statistics for AtmoCube sensors

Model_Subset	Model_ID	Test_RMSE	Test_MAE	Test_R2	Train_RMSE	Train_MAE	Train_R2	LB_rmse	LB_mae	LB_mean_residual_deviance	LB_ Rank
Low (<50 µg/m³)	DeepLearning_grid_3_AutoML_1_20251008_161954_model_1	6.57	2.895	0.803	2.828	1.368	0.972	6.782	2.529	45.989	1
Low (<50 µg/m³)	StackedEnsemble_BestOfFamily_1_AutoML_1_20251008_161954	6.253	2.94	0.822	2.995	1.663	0.969	7.231	2.478	52.285	2
Low (<50 µg/m³)	DeepLearning_grid_2_AutoML_1_20251008_161954_model_4	13.264	3.79	0.197	6.022	1.886	0.874	7.353	2.559	54.068	3
Low (<50 µg/m³)	StackedEnsemble_AllModels_1_AutoML_1_20251008_161954	6.242	2.704	0.822	3.936	1.599	0.946	7.432	2.364	55.233	4
Low (<50 µg/m³)	DeepLearning_grid_1_AutoML_1_20251008_161954_model_3	3.968	1.801	0.928	7.361	2.067	0.812	7.961	2.299	63.38	5

Low (<50 µg/m³)	DeepLearning_grid_1_AutoML_1_20251008_161954_model_4	4.318	1.876	0.915	7.481	1.986	0.805	8.109	2.385	65.755	6
Low (<50 µg/m³)	DeepLearning_grid_2_AutoML_1_20251008_161954_model_1	6.4	2.973	0.813	3.578	1.6	0.955	8.159	2.991	66.577	7
Low (<50 µg/m³)	GBM_grid_1_AutoML_1_20251008_161954_model_1	6.54	2.881	0.805	6.186	2.093	0.867	8.213	3.02	67.453	8
Low (<50 µg/m³)	DeepLearning_grid_3_AutoML_1_20251008_161954_model_3	15.851	14.642	-0.147	16.739	15.26	0.026	8.27	3.637	68.386	9
Low (<50 µg/m³)	DeepLearning_grid_3_AutoML_1_20251008_161954_model_4	5.592	2.483	0.857	5.623	1.889	0.89	8.296	2.648	68.831	10
High (≥50 µg/m³)	StackedEnsemble_AllModels_1_AutoML_2_20251008_162752	121.022	100.61	0.629	44.53	34.225	0.939	89.695	74.541	8045.133	1

High (≥50 μg/m³)	DeepLearning_grid_1_AutoML_2_20251008_162752_model_2	136.761	118.11	0.526	73.709	60.697	0.834	94.745	80.044	8976.538	2
High (≥50 μg/m³)	DeepLearning_grid_3_AutoML_2_20251008_162752_model_3	129.325	109.921	0.576	56.071	45.928	0.904	96.485	78.438	9309.44	3
High (≥50 μg/m³)	StackedEnsemble_BestOfFamily_1_AutoML_2_20251008_162752	129.279	110.052	0.577	64.196	52.846	0.874	99.53	79.996	9906.26	4
High (≥50 μg/m³)	DeepLearning_grid_2_AutoML_2_20251008_162752_model_3	129.605	106.901	0.575	48.165	36.582	0.929	100.524	82.119	10105.011	5
High (≥50 μg/m³)	DeepLearning_grid_1_AutoML_2_20251008_162752_model_3	114.367	96.043	0.669	68.377	54.058	0.857	102.895	80.471	10587.304	6
High (≥50 μg/m³)	DeepLearning_grid_3_AutoML_2_20251008_162752_model_1	137.281	114.282	0.523	35.036	20.896	0.963	106.762	87.51	11398.144	7

High (≥50 μg/m³)	GBM_lr_annealing_selection_AutoML_2_20251008_162752_select_model	139.195	111.078	0.509	67.122	53.151	0.862	109.135	87.475	11910.463	8
High (≥50 μg/m³)	DeepLearning_grid_1_AutoML_2_20251008_162752_model_4	114.042	97.226	0.671	57.249	40.911	0.9	111.219	83.622	12369.733	9
High (≥50 μg/m³)	GBM_grid_1_AutoML_2_20251008_162752_model_7	139.607	101.849	0.506	17.735	14.743	0.99	115.542	91.521	13350.025	10

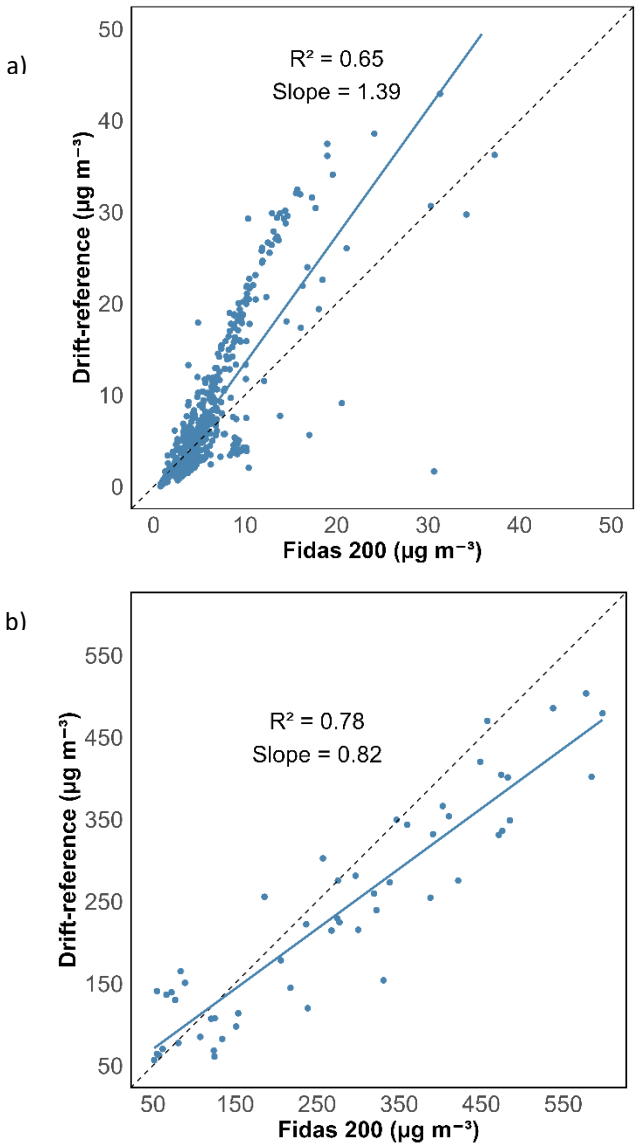


Figure S1. Relationship between all drift-reference sensors average concentration and Fidas 200, a) below 50 µg m^{-3}, b) above 50 µg m^{-3}.

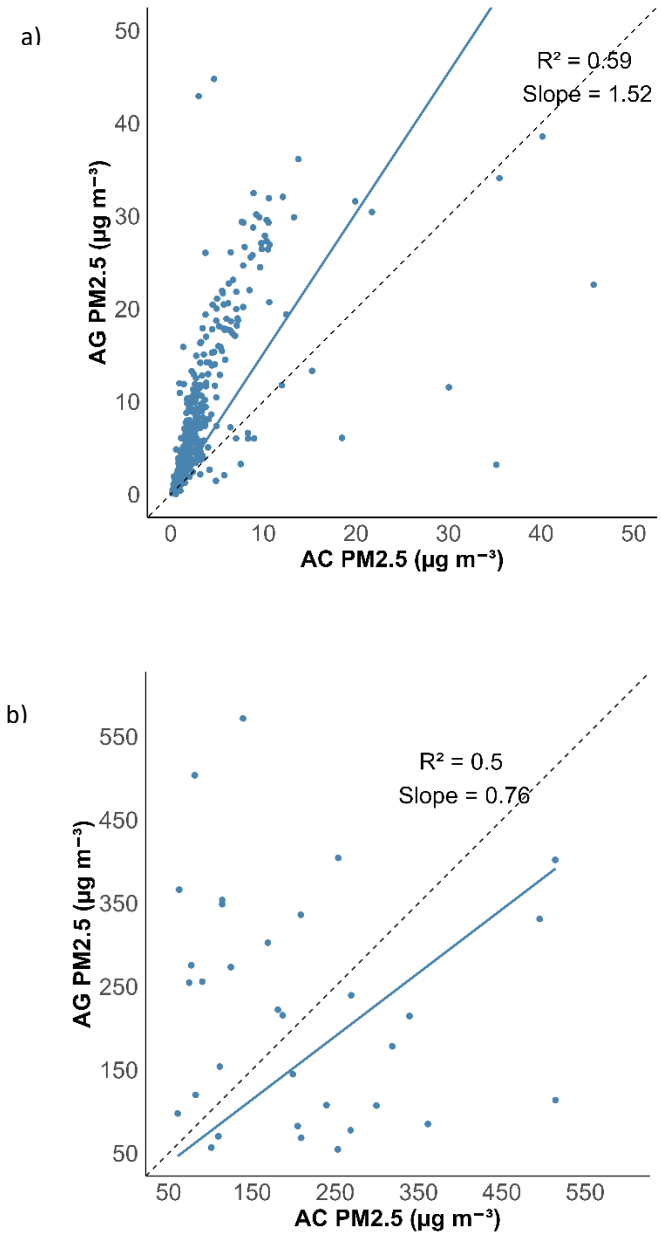


Figure S2. Inter-relationships between AG and AC sensors, a) below 50 µg m^{-3}, b) above 50 µg m^{-3}.

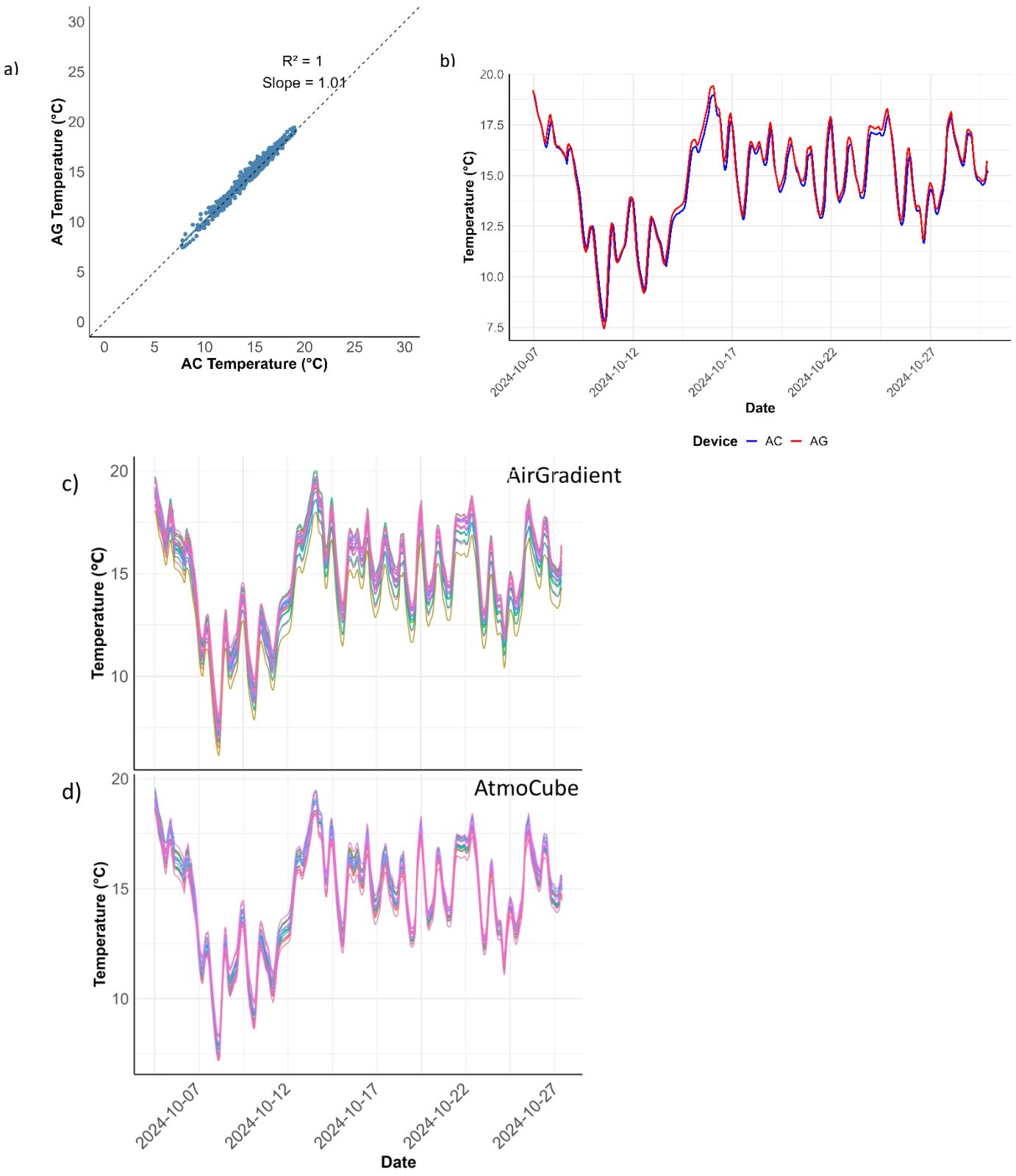


Figure S3. Inter-relationship of measured temperature of AC and AG, a) scatter plot, b) timeseries, c) timeseries of AG temperature measurements, d) timeseries of AC temperature measurements

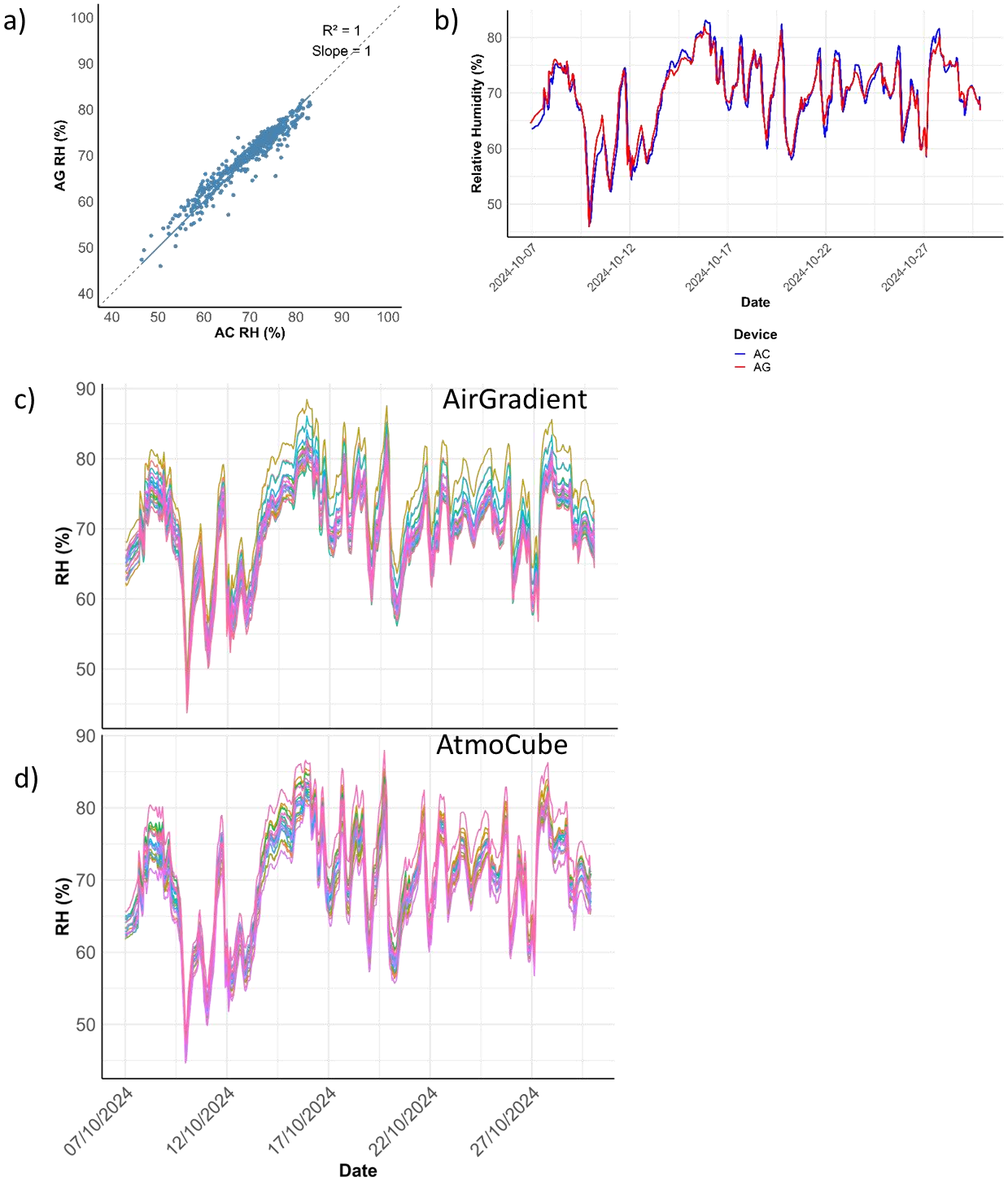


Figure S4. Inter-relationship of measured relative humidity of AC and AG, a) scatter plot, b) timeseries, c) timeseries of AG relative humidity measurements, d) timeseries of AC relative humidity measurements

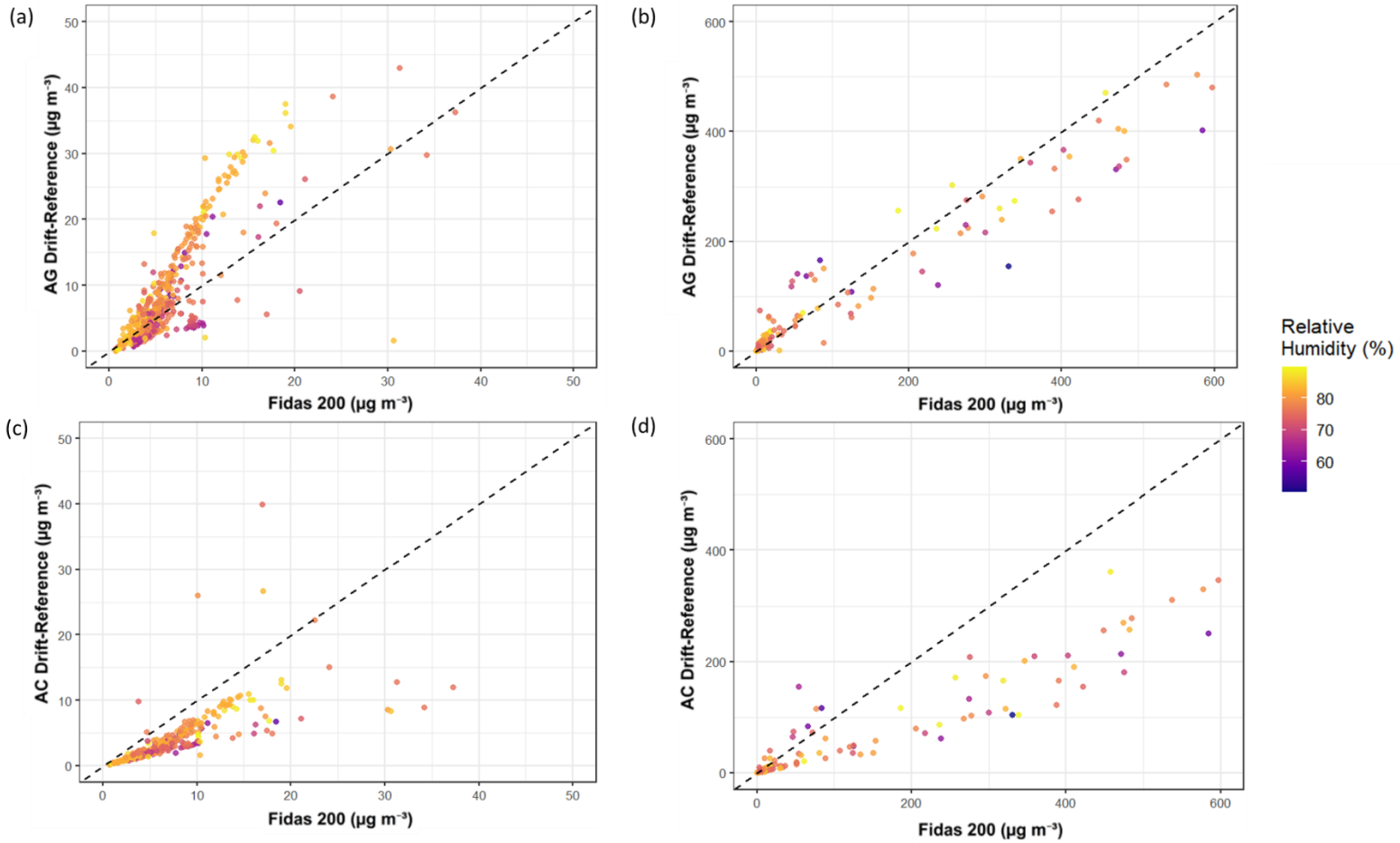


Figure S5. Pre-calibration $PM_{2.5}$ readings with relative humidity levels, a) AG below 50 µg m^{-3}, b) AG above 50 µg m^{-3}, c) AC below 50 µg m^{-3}, and d) AC above 50 µg m^{-3}.

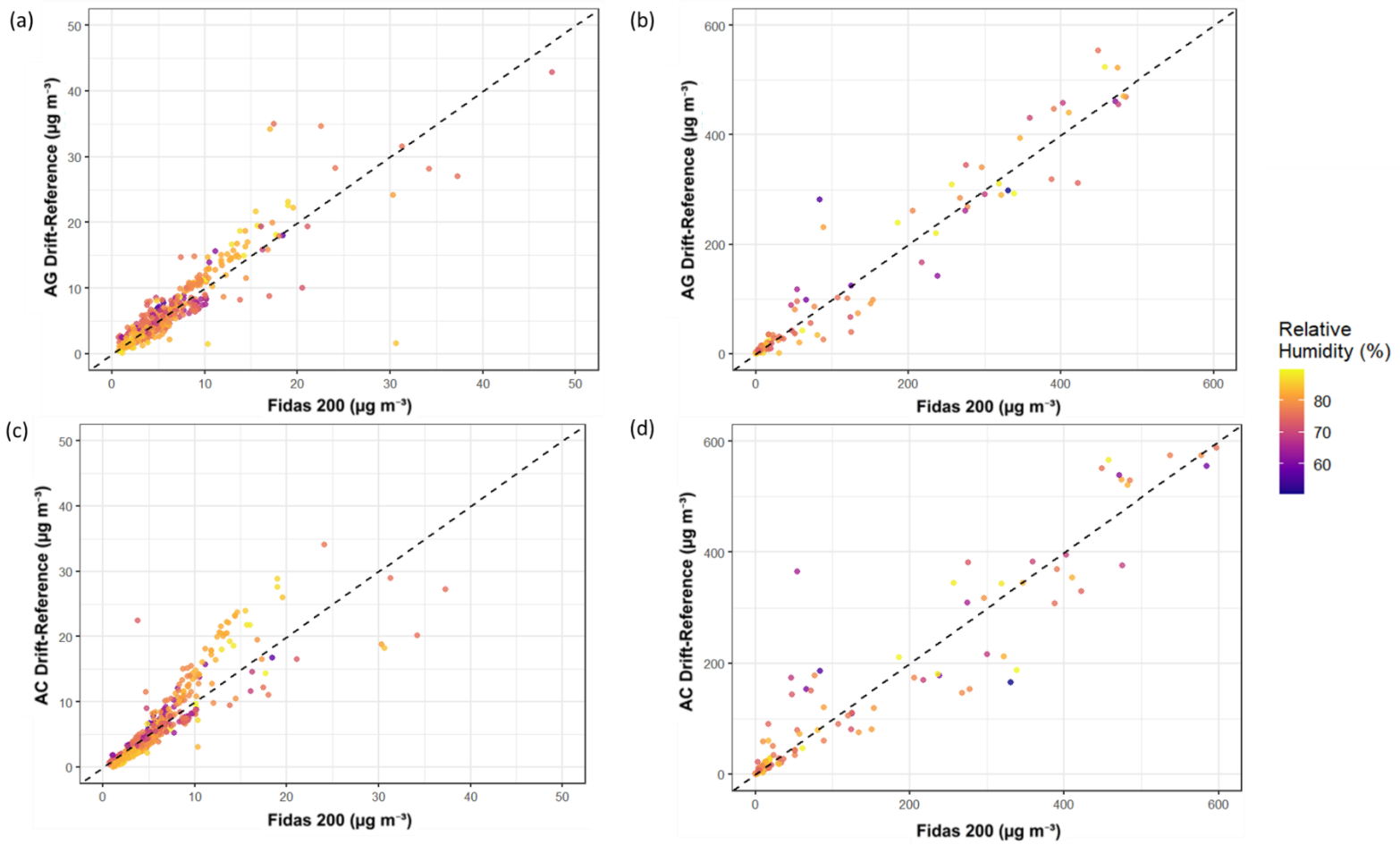
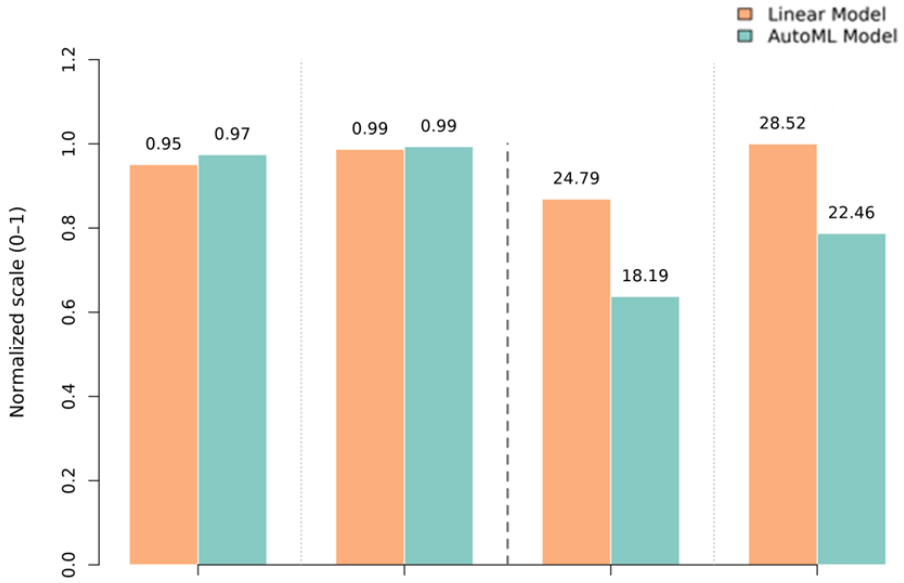


Figure S6. Post-calibration $PM_{2.5}$ readings with relative humidity levels, a) AG below 50 µg m^{-3}, b) AG above 50 µg m^{-3}, c) AC below 50 µg m^{-3}, and d) AC above 50 µg m^{-3}.

(a)



(b)

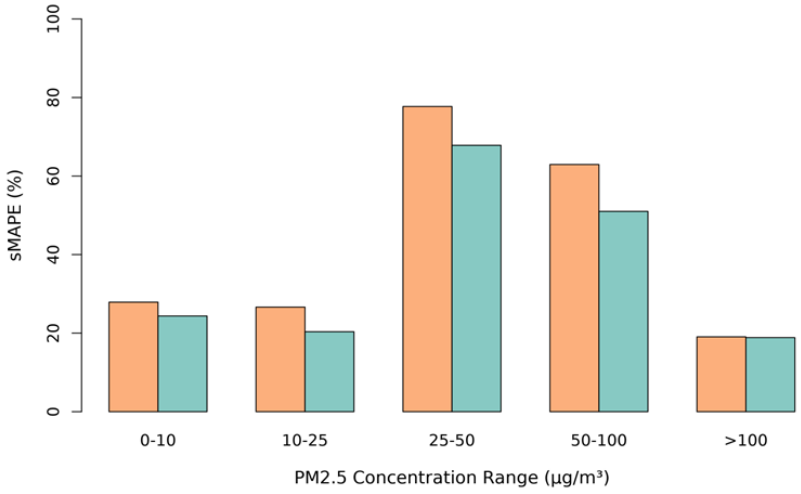
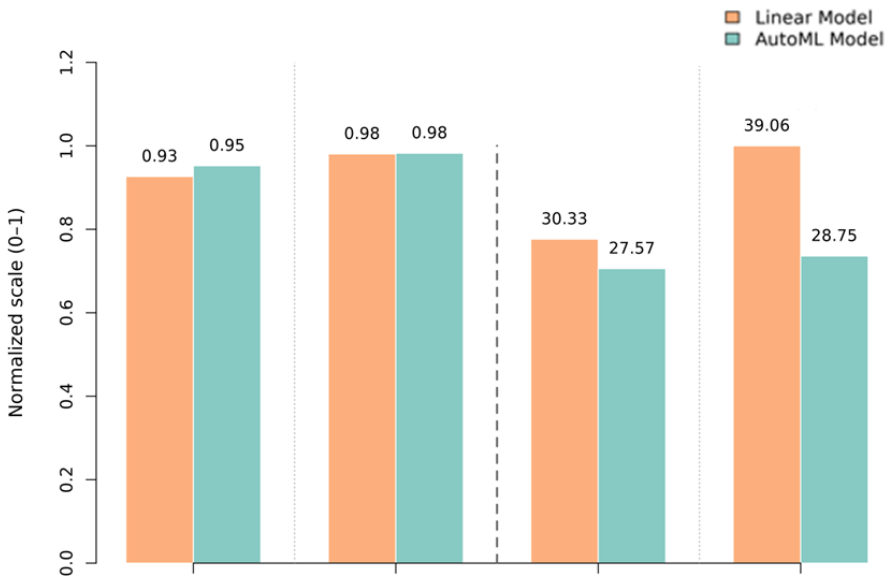


Figure S7. Comparison of AutoML model and the multivariate linear regression model for AirGradient ONE, a) performance metrics, and b) error by concentration range.

5

(a)



(b)

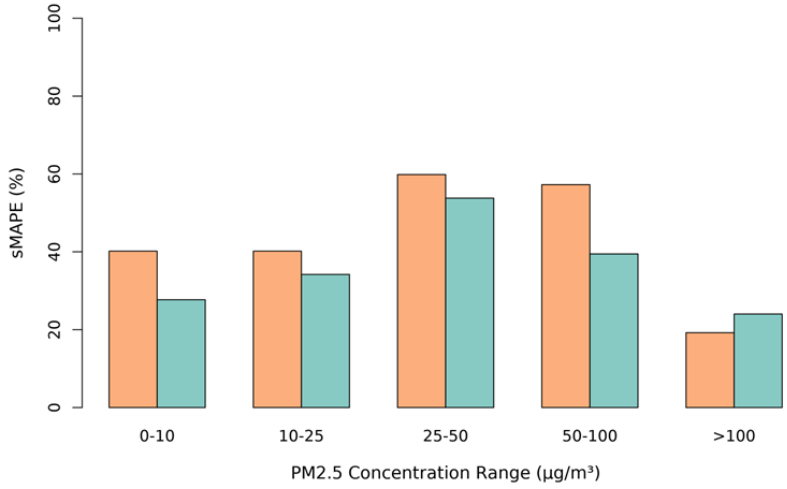


Figure S8. Comparison of AutoML model and the multivariate linear regression model for AtmoCube, a) performance metrics, and b) error by concentration range.