

<https://doi.org/10.1038/s41612-025-01231-5>

Comprehensive comparison of correction techniques for low-cost air quality sensors: the impact of device type and deployment environment



Idris Hayward^{1,2}, Nicholas A. Martin², Valerio Ferracci^{2,3}, Mohsen Kazemimanesh², Simon Jude³, Christopher Walton³, Zaheer Ahmad Nasir³ & Prashant Kumar^{1,4} ✉

Low-cost air quality sensors have shown great promise as a complement to high-cost reference and equivalent methods. Though not currently as accurate, their low barrier of entry and smaller form factor allow them to be deployed in greater numbers, thus enabling air quality measurements to be made at a far higher spatial and temporal resolution than previously possible. However, their measurements require corrections as they suffer from both short-term biases (e.g., changes in environmental conditions such as temperature and humidity), and long-term measurement drift due to degradation. Many studies have focused on calibration and re-calibration of sensors, but fewer focus on correcting pre-calibrated sensor measurements. Correcting measurements is a likely scenario for people buying off-the-shelf devices, as they will not have access to the raw data that underpins the measurements, such as sensor voltages. Previous studies focused on a small range of correction techniques, without accounting for the variances that can occur between devices or locations. This work aimed to perform a comprehensive assessment of different correction techniques applied to air quality sensor systems. More than 470,000 unique measurement corrections were tested across two sites to determine best practices for correction campaigns going forward, resulting in a far more robust study than previous works. It highlights the large variances in results that occurred between sites, particularly for NO₂, with results often more impacted by device type and location than the regression technique used. Simpler linear models were also found to perform just as well as, and sometimes better than, more complex non-parametric techniques. This study highlights that, though a strong focus is often put on comparing different regression methods, the choice of technique has less impact than the configuration of the device or the conditions of the co-location site. Therefore, future studies should focus less on small-scale comparisons of regression techniques and more on how to improve the transferability and applicability of results from a co-location campaign to another.

Low-cost sensor systems show great promise in complementing current high-cost reference and equivalent instrumentation by increasing the temporal and spatial resolution of air quality measurements^{1–3}. Their cost, size, and power requirements are reduced, making them viable for citizen science projects and spatially dense air quality

networks, where their high-cost counterparts would not be financially or logistically feasible.

However, measurements made by low-cost sensors are usually of lower quality than those made by reference instruments, and can be subject to significant interferences from environmental factors (e.g., humidity), other

¹Global Centre for Clean Air Research (GCARE), School of Engineering, Civil and Environmental Engineering, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, UK. ²Air Quality and Aerosol Metrology Group, National Physical Laboratory (NPL), Teddington, Middlesex, UK. ³Cranfield Environment Centre, Cranfield University, Cranfield, UK. ⁴Institute for Sustainability, University of Surrey, Guildford, UK. ✉e-mail: p.kumar@surrey.ac.uk

atmospheric species, and measurement drift⁴. They therefore require regular corrections to improve their lifetime and the quality of their measurements, to enhance confidence in their reported data^{5,6}.

Low-cost sensors often use electrochemical (EC) sensors to measure trace gas concentrations^{7,8}. For particulate matter (PM), components are classified as nephelometers or optical particle counters (OPCs) by the manufacturer^{9,10}, both of which measure particle concentrations using scattered light. Previous studies have argued that low-cost nephelometers could be better classified as poorly performing OPCs^{11–13}. We will not be discussing the validity of the nomenclature and will focus only on OPCs in this study.

EC sensors are capable of measuring gases in the parts per billion (ppb, or nmol mol⁻¹) range^{14–17}. Previous studies have shown that the measurements are more susceptible to biases from temperature^{15–22}, humidity^{15–22}, and electronic noise^{8,16,17} at these low levels, and are disproportionately affected by baseline drift^{15–23} and loss of sensitivity¹⁵ over time. OPCs are often used to measure the concentration of one or more size fractions of PM. They do not make a direct mass measurement of PM, instead using the intensity and refraction patterns of light passed through an air sample to estimate the concentration of different sizes of PM in the atmosphere^{24,25}. Previous studies have shown that OPC measurements are also susceptible to biases from humidity^{2,3,9,10,23,24,26–28}. Studies on the effect of temperature on OPCs show mixed results, with some reporting an influence^{1–3,6} and others reporting none^{25,27,29}. OPCs are also affected, though not to the same extent as EC sensors, by baseline drift, thought to arise from a combination of degradation of the laser and reduced air flow from fan degradation and dust build-up^{6,29,30}.

Manufacturer (i.e., out-of-the-box) calibration factors of low-cost sensors are provided to end-users, but their use is only described as “a suggestion”²⁹ and they tend to have mixed results. Borrego and co-workers⁴ evaluated a number of sensors and found that manufacturer calibrations have normalised root mean squared errors (nRMSE) ranging from 0.26 to 9.15 when measuring NO₂ (Note: this variance could partially be attributed to the use of RMSE, which is an inappropriate metric when comparing different studies, see Supplementary Note 4, Section 3). Many low-cost air quality sensors are calibrated in a laboratory environment, where they are exposed to a fixed range of controlled conditions⁴. However, these laboratory-based calibrations often cannot cover the wide range of variations in both the characteristics of pollutants (e.g., the chemical composition of ambient PM) and environmental conditions that the devices will encounter in the field^{4,10,16,17,20,23}. Low-cost sensors' performance can also be negatively affected if sensors are deployed in a different environment than the one in which they were calibrated^{2,20,23}. This is particularly true for OPCs, where changes in the chemical composition of particulates can bias measurements due to changes in their refractive index^{2,6,10,25,29,31–33}. These biases are not limited to location changes, as short-term changes in composition from sources such as fireworks and wildfires can also introduce biases in the measurements^{7,29}.

Due to this variation in manufacturer calibration performance and degradation of measurement quality over time, low-cost sensor measurements necessitate correction before, during, and after deployment. However, many studies focus on only calibrating the initial raw sensor measurements, with far fewer covering corrections^{2,6,34,35} and even fewer covering measurements of gaseous pollutants^{23,36}. Many OPCs come with built-in manufacturer calibrations, the quality of which varies depending on the similarities between the environments they were calibrated and deployed in ref. 9. This results in an increased focus on correcting OPC manufacturer calibrations for both pre-built units and bespoke sensor kits.

Many studies have focused on the calibration of low-cost sensor measurements to account for these biases, but when aggregating these studies, few conclusions could be made. Results from individual studies seem to contradict the aggregated results due to variations in many experimental factors such as duration of study, proportion of measurements used for the training set, and which independent variables (e.g., temperature and humidity) were used³⁷. This complicates

determining the best technique to use when calibrating raw low-cost sensor measurements or correcting pre-calibrated ones. The main objective of this study is to compare a wide range of pre-processing and regression techniques, some used previously in the literature, and others novel to this study, to determine the best methodology for correcting low-cost sensor measurements. We focus on corrections instead of calibrations, as it is a requirement for deploying off-the-shelf devices, which do not provide the underlying raw data for their calibrations, such as the sensor voltages. By comparing a very large number of techniques applied to the same datasets, more reliable trends can be uncovered, and recommendations can be made with more confidence.

This study uses sensor data collected in two long-term field co-location campaigns. This enables us to provide comprehensive recommendations on which techniques to use when correcting low-cost sensor measurements, maintaining measurement quality against drift and changes in environmental conditions. The first dataset was collected during the Breathe London Pilot Study, a two-year project that monitored air quality across London using over 100 low-cost AQMesh sensor systems and two mobile measurement platforms^{38,39}. Four of these AQMesh sensors were co-located with reference equipment at the Teddington Bushy Park Automatic Urban and Rural Network (AURN) monitoring site, along with several other low-cost sensors. The second co-location campaign took place over 20 months at the top of the Facility for Airborne Atmospheric Measurements (FAAM) building on the Cranfield University campus, utilising five low-cost Zephyr air quality monitors and one set of reference instruments. The measurements from this type of monitors have been corrected in previous studies^{40–42}.

Multiple combinations of variables were tested when correcting low-cost sensor measurements, including temperature and humidity. A variable called “Time Since Origin” was calculated, representing the cumulative time since deployment. This variable was used to test whether any long-term drifts in the measurement could be accounted for and corrected without recalibration.

A range of regression techniques were applied in this study, including both parametric and non-parametric methods (Table S2). Regression estimators were sourced from scikit-learn⁴³, XGBoost⁴⁴, and PyGAM⁴⁵. They were all chosen due to their compatibility with the scikit-learn API⁴⁶. All available estimators were tested to get maximum coverage in our comparisons. Summaries of the techniques used can be found in Supplementary Note 4.

A range of error metrics was tested in this study to develop a comprehensive view of the performance of each correction method (Table S3). Though a range of techniques were tested, mean absolute error (MAE, and normalised variants) is discussed here as it is an appropriate metric when comparing different models. r^2 and root mean squared error were calculated for each pipeline (Supplementary Note 6; Sections S6.1–6.3) but were not used as they are not appropriate when comparing different models, particularly when measurements of different lengths and magnitudes are used for each one (Supplementary Note 4, Section 3).

Results

Due to the large volume of results obtained in this study, it would be challenging to examine the performance of individual combinations of pre-scaling and correction techniques for each pollutant in detail. To examine them effectively, the results of the two sites were grouped by a pre-defined category (e.g., Pollutant and Scaling Technique). They were then filtered to examine a subset of the results (e.g., NO₂ corrections only). Then, they were filtered to remove any results outside the median $\pm 3 \times$ interquartile range (IQR). This removes any significant outliers, shifting the focus to general performance.

Summary statistics were then calculated from the results to visualise the performance of the different groups, split by site. The first quartile (Q1), median, and third quartile (Q3) were calculated. The minimum and maximum values were calculated by finding the smallest and largest result within the median $\pm 1.5 \times$ IQR.

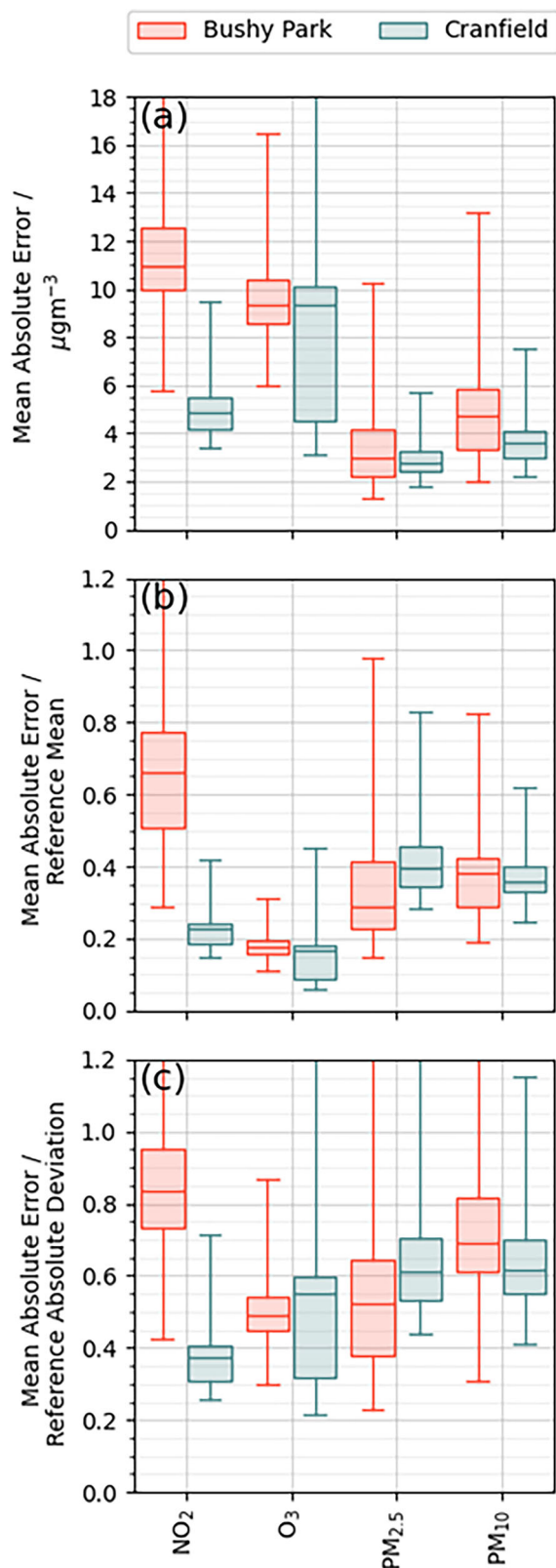


Fig. 1 | Comparison of the performance metrics for each pollutant across the two studies. Performance of corrections split by pollutant, evaluated by **a** mean absolute error, **b** mean absolute error normalised by the mean of reference measurements, and **c** mean absolute error normalised by the absolute deviation of reference measurements (Sections S6.1–6.3). The mean absolute errors of each category are summarised by box plots, with the solid box representing the range between the 1st and 3rd quartiles, with the centre line representing the median MAE. The tails represent the highest and lowest MAE values within the median $\pm 3 \times$ interquartile range (IQR).

We focused on four key criteria in this study. The average performance of correction methodologies was compared with the median result. The best performers were compared using the 1st quartile and the minimum result. The worst performers were compared using the 3rd quartile and the maximum result. The consistency of a technique was also examined using the IQR. A low IQR showed consistent performance, whilst a large IQR showed significant variation.

We focused on the relative performance of different techniques instead of aiming for a set quantitative target (e.g., stating that $\text{MAE} < 5 \mu\text{g m}^{-3}$ indicates a good correction method). Different low-cost air quality sensors provide varying levels of measurement uncertainty. The CEN/TS 17660-1:2021 and CEN/TS 17660-2:2024 (PM_{2.5} only)⁴⁷ technical specifications separate them into three classes by measurement uncertainty. Class 1 devices are at a minimum consistent of indicative air quality measurements in line with data quality objectives; Class 3 devices have a more relaxed uncertainty target not related to a data quality objective. Focusing on relative improvements allows this study to be applied to all classes of low-cost sensors. Given the high variability in performance amongst sensors, it is recommended that end users ensure their sensor complies with the expected performance separately.

We initially focus on the overall results, followed by examining the results split by pollutants, devices, folds, pre-scaling techniques, and independent variables used. This gives a broad overview of the performance of different techniques. We then examine two pollutants (NO₂ and PM_{2.5}) to obtain a more granular view of how the inclusion of multiple features and different scaling techniques affects correction performance. More information can be found in the “Methods” section.

All pollutants

Figure 1 shows the summarised performance of all correction pipelines (defined in section “Building and fitting regression pipelines”) split by pollutant and site using the MAE. Tabular representation of the box plots can be seen in Table S8 in the Supplementary Material. The non-normalised MAEs in Fig. 1a are spread out, with the NO₂ results in particular having no overlap between the boxes representing the Bushy Park and Cranfield sites. The best performing NO₂ correction at the Bushy Park site has a higher MAE than the Q3 MAE at the Cranfield site, meaning over 75% of the Cranfield corrections perform better than the best at Bushy Park. O₃ achieves the same median MAE across both sites but the IQR is significantly higher at the Cranfield site, showing a more varied performance in the correction pipelines. PM_{2.5} had similar averages at both sites but a lower range than the NO₂ results. When normalising the MAE with the mean of the reference measurements, the average error is higher at Cranfield relative to Bushy Park for PM_{2.5}, and the IQR is more consistent between sites. The O₃ measurements are significantly lower than those of others for both sites, achieving a minimum ratio of 0.06 at the Cranfield site. Average MAEs for PM₁₀ results become consistent between sites, but the IQR at the Cranfield site is lower.

Normalising the MAE by the absolute deviation of the reference measurements, referred to as nMAE in this study, not only improves the relative difference in the medians between sites but also between pollutants, so this normalisation method was used for the summaries in the rest of this site. MAE and CvMAE (MAE/Reference Mean) results can be found in the Supplementary Material. MAE results are from Figs. S4–S20 and Tables S20–S36, CvMAE results are from Figs. S21–S37 and Tables S37–S53.

To improve the robustness of the results and reduce the chances of overfitting, five variations on each pipeline were trained, each using 80% of the training data. These subsets of the data are referred to as “folds”. The data was split to ensure an equal distribution of the target pollutant in each fold. A sixth “mean” fold was also generated using the mean of the measurements of the first five folds to test whether this could smooth out overfit estimators, particularly non-parametric ones. The performance of each fold was tested to ensure it remained consistent. There was no significant difference in average nMAE and spread between training set folds seen in Fig. 2, suggesting that the measurements were evenly distributed within training sets.

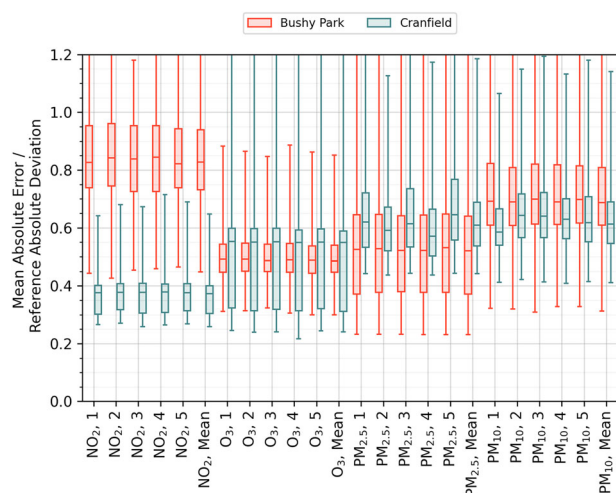


Fig. 2 | Performance of corrections split by fold and pollutant, evaluated by mean absolute error normalised by the absolute deviation of reference measurements.

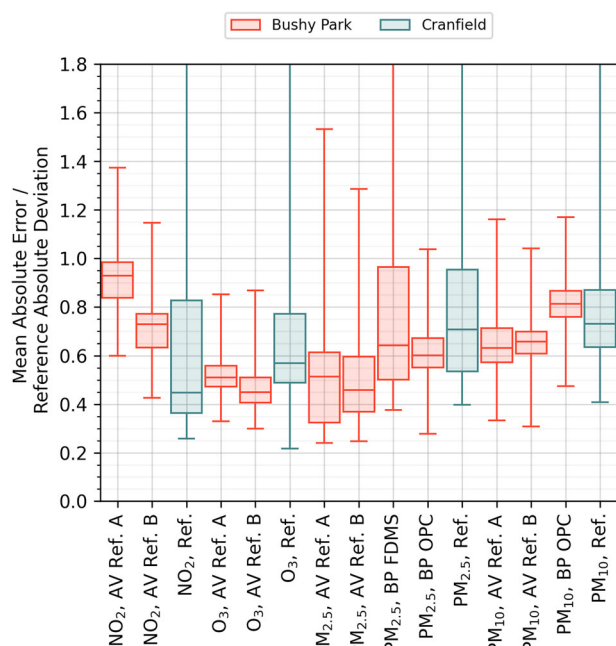


Fig. 3 | Performance of corrections split by reference (or reference equivalent) device and pollutant, evaluated by mean absolute error normalised by the absolute deviation of reference measurements.

There were some variations in performance for $PM_{2.5}$ and PM_{10} at the Cranfield site, suggesting a short, significant episode in one or more measurements (e.g., relative humidity or temperature) that was not shared between all folds, causing inconsistency. Tabular representation of the box plots can be seen in Table S9 in the Supplementary Material.

No significant advantage is clear when using the average predicted measurements of all five folds, further highlighting the consistency between pipelines trained on different folds. Due to the consistency between folds, all folds and the mean result were used when comparing measurements.

At the Bushy Park site, at least two reference or reference equivalent (henceforth referred to collectively as “reference”) devices were used for each pollutant. To test how closely the pipelines were able to correct the low-cost sensor measurements to the reference measurements, the performance of pipelines against each reference device was compared in Fig. 3 (key in Table S1, tabular representation of the box plots can be seen in Table S10 in the Supplementary Material). Corrections against AirView B perform

significantly better than AirView A for NO_2 , achieving a median nMAE of 0.77 vs. 0.98. These are both significantly higher than Cranfield’s reference instrument at 0.45. Corrections against AirView B also perform better than AirView A for O_3 , though the relative difference is smaller at 0.45 vs. 0.51. The Cranfield reference device performed slightly worse at 0.57. AirView A and B both achieved better median nMAE than the Cranfield reference for PM_{10} (0.62 and 0.66 vs. 0.73), though the Bushy Park OPC performed worse (0.81). The four $PM_{2.5}$ reference instruments at Bushy Park (AirView A, AirView B, Bushy Park FDMS, and Bushy Park OPC) achieve a lower median than the one at Cranfield (0.40, 0.42, 0.55, 0.6 vs. 0.70). However, nMAE spread for the Cranfield reference instruments is much higher than that for the Bushy Park devices, with the Bushy Park FDMS being the only reference device with a higher IQR (0.44).

Despite higher averages, the Cranfield reference devices achieve the lowest nMAE for all pollutants except $PM_{2.5}$, showing higher device variability but better performance in some situations. When comparing the Bushy Park data further, it was decided to use the best-performing reference device for each pollutant to maintain parity with the Cranfield site, where only one reference device was present for each pollutant. Therefore, AirView Reference B was used for analysing NO_2 and O_3 , and AirView Reference A was used for $PM_{2.5}$ and PM_{10} analysis.

Due to potential variations not only between sensor system brands but also between devices from the same manufacturer, the performance of individual low-cost air quality systems was compared in Fig. 4 (key in Table S1, tabular representation of the box plots can be seen in Table S11 in the Supplementary Material), split by pollutant. The individual sensor system performance differs substantially among pollutants. The Breathe London AQMesh and AQMesh C pipelines had higher median nMAE for NO_2 corrections (1.02 and 1.23) than AQMesh A and B (0.86 and 0.79). The Cranfield devices also had mixed performance when correcting NO_2 measurements. Zephyr A, B, and D all scored consistently lower than the Bushy Park low-cost air quality sensor systems. Overall, the NO_2 corrections performed far better at the Cranfield site than the Bushy Park site, with the two worst Cranfield low-cost air quality sensor systems achieving similar performance to the best Bushy Park low-cost air quality sensor systems (Zephyr C and E, median nMAEs 0.82 and 0.91 vs. NPL AQMesh A and B, median nMAEs 0.87 and 0.82).

The O_3 correction performance is more consistent than that for NO_2 at Bushy Park, with the Breathe London AQMesh, AQMesh A and B all achieving similar median nMAE scores. AQMesh C’s median nMAE is nearly twice as low as the other low-cost air quality sensor systems for O_3 corrections (0.29 vs. 0.5, 0.5, and 0.47).

Zephyr A, B, and E achieve similar median nMAE scores when correcting O_3 measurements (0.58, 0.53, and 0.58). However, Zephyr E’s IQR is more than double that of A and B (0.20 vs. 0.07 and 0.06). Zephyr D achieves significantly lower nMAEs for O_3 with a median of 0.31. The Q3 nMAE for Zephyr D’s corrected O_3 measurements is 0.38, lower than the Q1 nMAE for all other Zephyr devices. Zephyr C performs even worse when correcting O_3 than NO_2 , with a median nMAE of 1.42, nearly 2.5 times greater than the next highest median. Its best-performing correction pipeline achieves an nMAE of 1.08, worse than the worst performer of the other 4 Zephyr systems (Zephyr E, 0.87). O_3 corrections achieved similar nMAEs across both sites. NPL AQMesh C and Zephyr D are the two best performers, with the other AQMesh devices (Breathe London, A and B) achieving similar median performances to Zephyr A, B, and E. Zephyr C’s corrected O_3 performs the worst across all sites.

$PM_{2.5}$ correction performance is mixed between devices for both sites. The low-cost $PM_{2.5}$ monitors in the AirView cars, Low-Cost A and Low-Cost B, were the two best performers, with median nMAEs of 0.40 and 0.36. As the reference and low-cost $PM_{2.5}$ monitors shared the same inlet on the outside of the AirView car, they would have had minimal variation between $PM_{2.5}$ concentrations, which may explain the lower median nMAE. NPL AQMesh B was the next best performer, with a median nMAE of 0.59. NPL AQMesh A and Nova Sensor A and B have similar performances, with median nMAEs of 0.61, 0.57, and 0.57. NPL AQMesh C and the Breathe

Fig. 4 | Performance of corrections split by device and pollutant, evaluated by mean absolute error normalised by the absolute deviation of reference measurements.

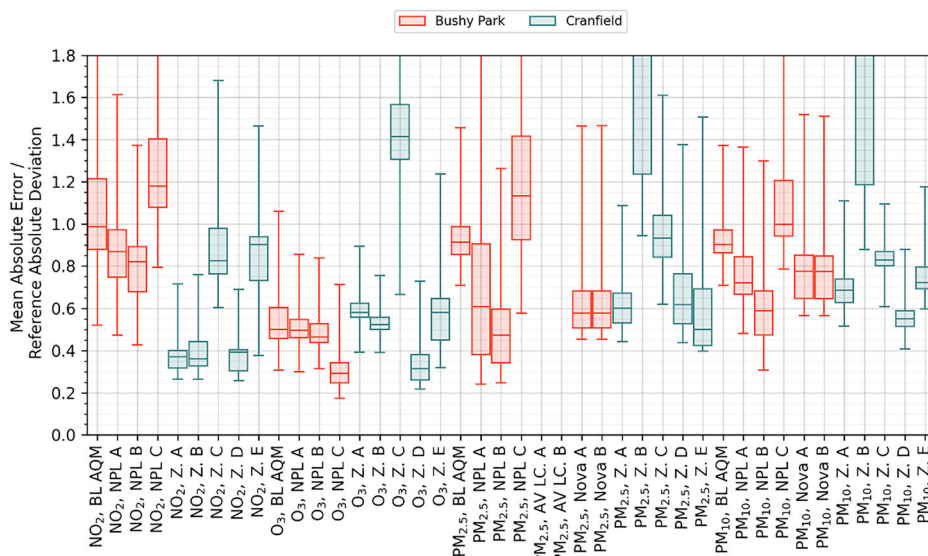
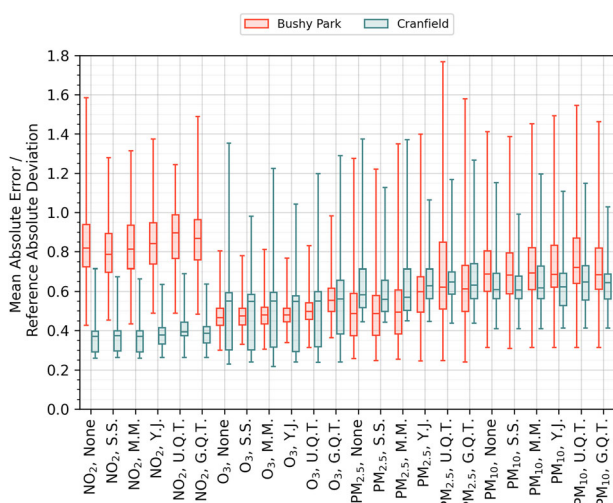


Fig. 5 | Performance of corrections split by scaling technique and pollutant, evaluated by mean absolute error normalised by the absolute deviation of reference measurements.

London AQMesh were again the worst performers at the Bushy Park site, with median nMAEs of 1.13 and 0.91. $PM_{2.5}$ corrections for Zephyr A, D, and E had similar performance, scoring median nMAEs of 0.60, 0.62, and 0.50, matching the performance of NPL AQMesh A and B and Nova Sensor A and B. Zephyr B and C perform noticeably worse, with median nMAEs of 2.17 and 0.93.

PM_{10} correction performance is similarly mixed. NPL AQMesh B records the best median nMAE score with 0.59. NPL AQMesh A, Nova A, and Nova B all achieve similar median nMAE scores at 0.72, 0.77, and 0.77, respectively. The IQR for these 4 devices is also similar, between 0.18 and 0.2. The Breathe London AQMesh device performs worse with a median nMAE of 0.90, and NPL AQMesh C performs worse with a median nMAE of 0.99. At the Cranfield site, Zephyr B is the worst performer this time by a significant margin, with a median nMAE of 1.94. Zephyr D averages the best with a median nMAE of 0.55. The worst average sensor system performance at Cranfield (Zephyr B, 1.94) is significantly worse than Bushy Park (AQMesh C, 0.99). However, the performance of the other sensor systems is similar between sites.

After examining their performance, a sub-set of low-cost sensors was selected for further analysis for the rest of the study. NPL AQMesh A and B,

and Zephyr A, B, and D were used for NO_2 analysis. The Breathe London AQMesh, NPL AQMesh A, B, and C, and Zephyr A, B, D, and E, were used for O_3 analysis. NPL AQMesh A and B, AirView Low-Cost A and B, Nova Sensor A and B, and Zephyr A, C, D, and E were used for $PM_{2.5}$ analysis. NPL AQMesh A and B, Nova Sensor A and B, and Zephyr A, D, and E were used for PM_{10} .

Comparing the pre-scaling methods used prior to regression in Fig. 5 (key in Table S1, tabular representation of the box plots can be seen in Table S12 in the Supplementary Material), no trends are immediately apparent. The non-parametric techniques (Quantile and Yeo-Johnson Transform) result in worse performance on average, with slightly higher median and minimum nMAE scores for NO_2 , O_3 , and $PM_{2.5}$. The difference is most pronounced for $PM_{2.5}$ at the Bushy Park site, with median nMAEs of 0.46 and 0.66 for Standard Scale and Yeo-Johnson Transform, respectively, and least pronounced for O_3 corrections at the Cranfield site, where the median nMAE remains the same between both sets of pre-scaling techniques. As the transformers alter the distribution of the measurements, if there is a linear relationship between the sensor and reference measurement, then this linearity may be altered. If the relationship between the uncorrected sensor measurement and reference measurement is non-linear, linear models such as Linear Regression and Theil-Sen Regression will not be able to model it, accounting for the worse performance of transformers with some of the pollutants. If there is no significant difference between parametric scaling algorithms and non-parametric transformers, this implies the relationship between the uncorrected sensor measurements and reference measurement is already non-linear, or the transformers do not affect the distribution enough to remove the linear relationship.

Looking at this summary, it could be concluded that the pre-scaling technique used has little to no influence on the performance of the correction. However, later analysis of different scaling algorithms for each pollutant on individual regression technique performance shows that it does have an impact on performance in certain situations, both positive and negative (see section “Simpson’s paradox”).

When examining how the number of independent variables used in a regression algorithm can affect correction performance in Fig. 6 (key in Table S1, Tabular representation of the box plots can be seen in Table S13 in the Supplementary Material), some interesting trends emerge. For each pollutant, there are four variations of correcting the measurements. The first run only uses the target variable as measured by the device (e.g., NO_2), the second uses the target variable and all secondary variables (e.g., $NO_2 + RH + NO + O_3 + T + RH^2 + RH \times T + T^2$), the third uses the target variable and time since origin (TSO, a feature representing the time elapsed since the first measurement in the training set) and the fourth uses

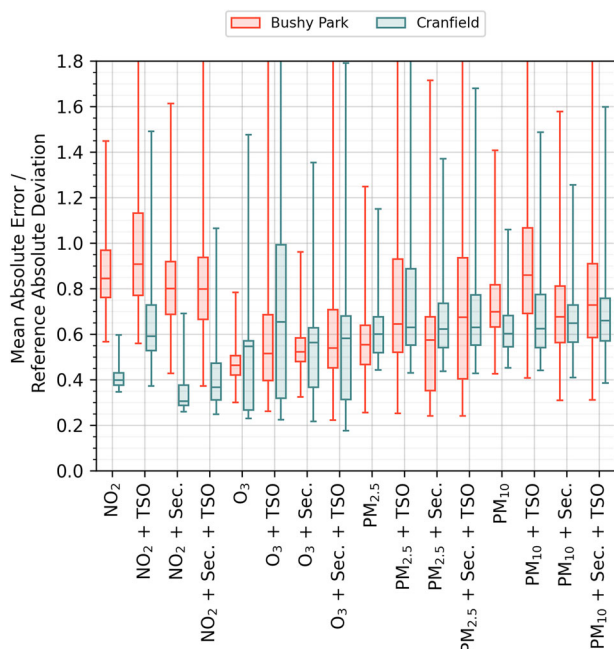


Fig. 6 | Performance of corrections split by independent variables used and pollutant, evaluated by mean absolute error normalised by the absolute deviation of reference measurements.

the target variable, all secondary variables and TSO. TSO was included in the analysis to determine whether baseline drift within the low-cost measurements could be modelled and counteracted, with mixed but mostly negative results.

The inclusion of the secondary variables when correcting NO_2 measurements significantly improved the minimum nMAEs at both Bushy Park and Cranfield (0.57 and 0.35 vs. 0.45 and 0.26). However, the median nMAEs at Bushy Park do not differ much (0.84 vs. 0.80), suggesting only some techniques are able to effectively utilise these secondary variables. The inclusion of TSO has a minimal effect on NO_2 corrections, only slightly improving the minimum measurement when added to the target variable and all secondary variables.

The inclusion of secondary variables in O_3 corrections worsens median nMAE scores at Bushy Park (0.46 vs. 0.52) and Cranfield (0.55 vs. 0.56). It has a much greater effect on the first quartile at the Cranfield site, showing the overall distribution of measurements is pushed higher, even if the median only increases slightly (0.27 vs. 0.37). Utilising only uncorrected O_3 and TSO increased the median nMAE at both sites. It also increases the median when added to O_3 and the secondary variables, but not by as much. The minimum nMAE is lower at both Bushy Park and Cranfield when including all secondary variables and TSO, suggesting that at least some techniques are able to use the TSO feature to model drift.

The inclusion of secondary variables with the uncorrected $\text{PM}_{2.5}$ measurements increases the median nMAE at the Bushy Park site (0.51 vs. 0.57) and at the Cranfield site (0.60 vs. 0.62). The median nMAE increases further when adding TSO as a feature, suggesting drift in measurements was not a significant problem during the co-locations, and the effects of temperature and humidity were already counteracted by the sensor systems in the uncorrected measurements.

The inclusion of secondary variables with the uncorrected PM_{10} measurements slightly decreases the median nMAE at the Bushy Park site (0.69 vs. 0.67) while increasing it at the Cranfield site (0.69 vs. 0.66). However, the minimum nMAE scores are decreased at both by adding secondary variables. The inclusion of the TSO feature worsens the median correction performance at the Bushy Park site while having a limited effect on the median performance at the Cranfield site, suggesting either drift was not an issue with these measurements, or the effects were too complex to model.

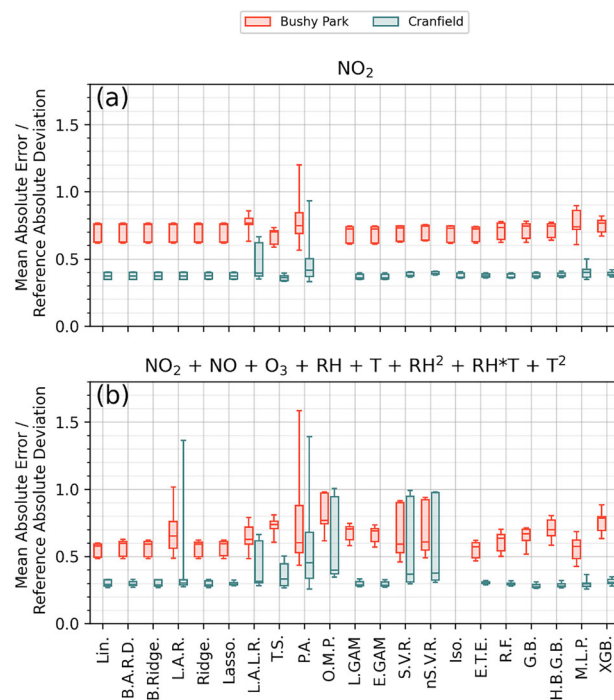


Fig. 7 | Performance of regression techniques correcting NO_2 measurements. Results shown are split by **a** NO_2 only and **b** $\text{NO}_2 + \text{RH} + \text{NO} + \text{O}_3 + \text{T} + \text{RH}^2 + \text{RH} \times \text{T} + \text{T}^2$, evaluated by mean absolute error normalised by the absolute deviation of reference measurements.

The IQR increases with additional variables across all sites and pollutants: it is at its minimum when only using the target variable, and at its maximum when using the target variable and TSO. This suggests that only some techniques can effectively model the relationships between the secondary variables and the reference measurement. The following sections (along with S5.2.1 and S5.4.1 in the Supplementary Information) explore more in depth the effect of the inclusion of independent variables with different regression techniques for different pollutants.

NO_2

Comparing regression technique performance for NO_2 corrections against using NO_2 only and $\text{NO}_2 + \text{RH} + \text{NO} + \text{O}_3 + \text{T} + \text{RH}^2 + \text{RH} \times \text{T} + \text{T}^2$, significant variations in performance are found (Fig. 7). When only NO_2 is used (Fig. 7a, key in Table S1, tabular results in Supplementary Material, Table S14a,b), linear regression, the Bayesian methods, least angle regression, ridge regression and Lasso regression all have similar performance. These techniques all optimise their coefficients using least squares loss functions, only differing in how the loss functions are used to select or assign weights to independent variables. As only one variable is present, this step is irrelevant, and all techniques perform the same way.

Least angle Lasso regression performs worse, with similar median nMAE to the other techniques at both sites, but much higher maximum nMAE scores. This technique performs poorly with pre-normalised data (see Fig. S2), so this worse performance comes from the inclusion of pre-scaled correction pipelines utilising Standard Scaling and MinMax Scaling. Once multiple independent variables are added (Fig. 7b, key in Table S1, tabular results in Supplementary Material, Table S14c,d), they all have similar improvements in performance except for least angle and least angle Lasso regression. These techniques are more susceptible to noise, so if one of the secondary variables has a large random error component, the performance of the correction will suffer (see Supplementary Note 3 Section 2).

When only NO_2 is used, Theil–Sen regression performs better than the previously mentioned methods as it is robust to outliers. However, when

multiple independent variables are used, it performs worse at Bushy Park and is more varied at Cranfield, as the proportion of outliers it is tolerant of decreases with the number of independent variables (Supplementary Note 3 Section 7). Though the median nMAE for Theil-Sen decreases from 0.36 to 0.33 and the minimum nMAE decreases from 0.33 to 0.27, the maximum nMAE increases from 0.39 to 0.50, suggesting this technique is not appropriate if multiple independent variables are used and it is difficult to minimise the number of outliers in the dataset.

Passive-aggressive regression is the best and worst performer when correcting NO_2 measurements at the Bushy Park site. It was the best performer at the Cranfield site, both with NO_2 only and with all independent variables, the second worst performer when using NO_2 only (with the least angle Lasso regression performing worse), and the fifth worst performer when using all secondary variables, with Least Angle Lasso Regression and the three support vector regression techniques performing worse. When examining the performance of passive-aggressive regression split by scaling techniques in Fig. S2, it is clear that the large variance in performance depends on whether the variables are pre-scaled or not.

Orthogonal matching pursuit is not utilised for NO_2 alone as it needs at least 2 independent variables. Its performance varies substantially when all features are used. It is suggested to perform better when the data is normalised (see Supplementary Note 3 Section 4), and this is confirmed in Fig. S2. However, its best nMAE score in the Bushy Park is 0.62, equivalent to the worst nMAE score for Bayesian ridge regression and higher than the worst score for linear regression (0.60). Its best nMAE score at the Cranfield site (0.35) is still worse than the worst score for multiple techniques, including linear regression and Bayesian ridge regression (0.33 for both).

Linear and expectile GAMs have similar performances to linear regression and similar methods discussed above when only NO_2 is used as a feature. They both see little improvement in the median nMAE scores when using multiple features (0.72 for both vs. 0.70 and 0.69) at the Bushy Park site. The improvement is greater at the Cranfield site (0.36 vs. 0.29 for both), furthering the already mismatched results between both campaigns.

The two support vector machine methods also have similar performance to Linear Regression and other similar methods mentioned above when only NO_2 is used as a feature. However, their performance is significantly more varied when multiple features are used. This is especially true for the Cranfield site, where some correction pipelines perform worse than the worst Bushy Park pipeline. This is due to the large difference in magnitudes between the independent variables (e.g., NO_2 may be 20 while RH^2 could be 9216 (i.e., 96^2), two orders of magnitude higher). When all independent variables are normalised, the performance of support vector machines is much better, comparable to other techniques (see Fig. S2).

All non-parametric techniques (isotonic regression and all techniques to the right in Fig. 7) perform similarly to linear regression and other techniques discussed previously, aside from multi-layer perceptron regression, which has similar median nMAE performance (0.74 and 0.40 at Bushy Park and Cranfield, respectively) but with higher maximum nMAE values than Linear Regression (0.90 and 0.48 vs. 0.77 and 0.40 for Bushy Park and Cranfield, respectively). This variance is due to pre-scaled measurements performing worse than unscaled ones.

Isotonic Regression can only utilise a single independent variable, so it has no results when all secondary variables are used (Fig. 7b). The other non-parametric techniques improve similarly at the Cranfield site when using all secondary variables, with median nMAE decreasing between 0.05 and 0.1. The improvements are far more varied at the Bushy Park site, with an ensemble of extra trees and a multi-layer perceptron regression improving the most. Their median nMAE scores are 0.72 and 0.74, respectively, when using NO_2 only, decreasing to 0.57 and 0.58 when using all secondary variables. The median nMAE score for XGBoost regression increases with the inclusion of secondary variables at the Bushy Park site (0.77 vs. 0.78). This suggests that there is a significant noise component in some of the variables that affects the non-parametric regression algorithms in different ways, with some able to compensate for it and some not.

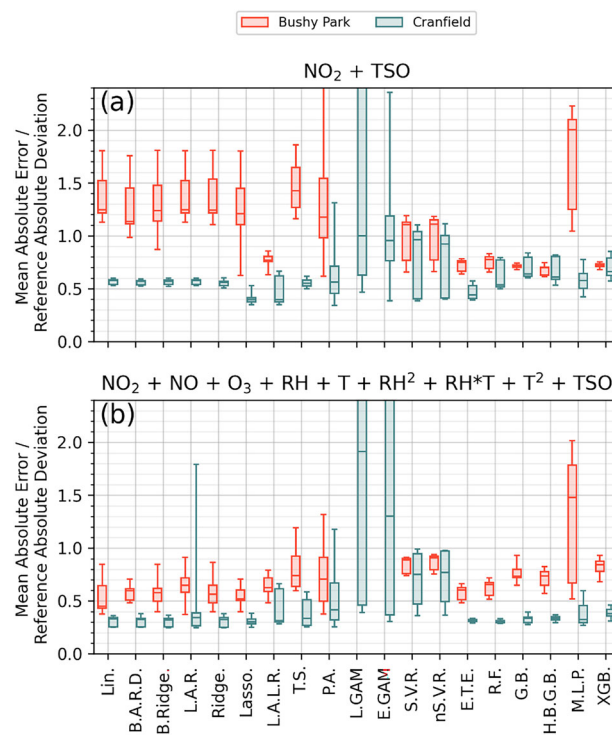


Fig. 8 | Performance of regression techniques correcting NO_2 measurements.

These are split by independent variables used: **a** NO_2 and time since origin feature, and **b** $\text{NO}_2 + \text{RH} + \text{NO} + \text{O}_3 + \text{T} + \text{RH}^2 + \text{RH} \times \text{T} + \text{T}^2$, and time since origin. All techniques are evaluated by mean absolute error normalised by the absolute deviation of reference measurements.

The inclusion of TSO with NO_2 as independent variables has a largely negative impact on NO_2 corrections at both the Bushy Park and Cranfield sites (Fig. 7a vs. Fig. 8, key in Table S1, tabular results in Supplementary Material, Table S15a, c). Parametric techniques significantly worsened across the board when adding TSO as an independent variable, with the exception of cross-validated Lasso regression at the Cranfield site and least-angle Lasso regression at both sites. The lack of change in performance between the two combinations is due to the coefficient for TSO being set to 0, giving it no influence. If there was a drift component within the first 72 days (the size of the training set), there was no linear relationship that could be modelled while environmental features were not present.

However, once the other secondary variables are included with the time feature, correction performance improves significantly. It also brings the two sites closer in performance, with some techniques having overlapping nMAE distributions. These improvements come from more than just the inclusion of environmental features. Linear regression achieves a median nMAE of 0.58 and 0.29 in the Cranfield and Bushy Park sites, respectively, when using all independent variables except TSO. TSO's inclusion improves the nMAE further at Bushy Park while slightly worsening Cranfield performance, achieving median nMAEs of 0.45 and 0.33, respectively. Linear regression is the best performer across all techniques at the Bushy Park site when all variables are used, with a median nMAE of 0.45 and a minimum nMAE of 0.38. This minimum nMAE is only matched by passive-aggressive regression. The larger median associated with passive-aggressive regression is due to the differing performance based on the scaling technique. This shows that, though significant focus is sometimes made on non-parametric techniques like Random Forest, XGBoost, and multi-layer perceptrons, sometimes a simple linear model can outperform them.

Cross-validated Lasso regression achieves the lowest median nMAE score at the Cranfield site when using all features. All parametric techniques, aside from least angle Lasso regression, achieve the same minimum nMAE at 0.25. This highlights that even though a significant number of other

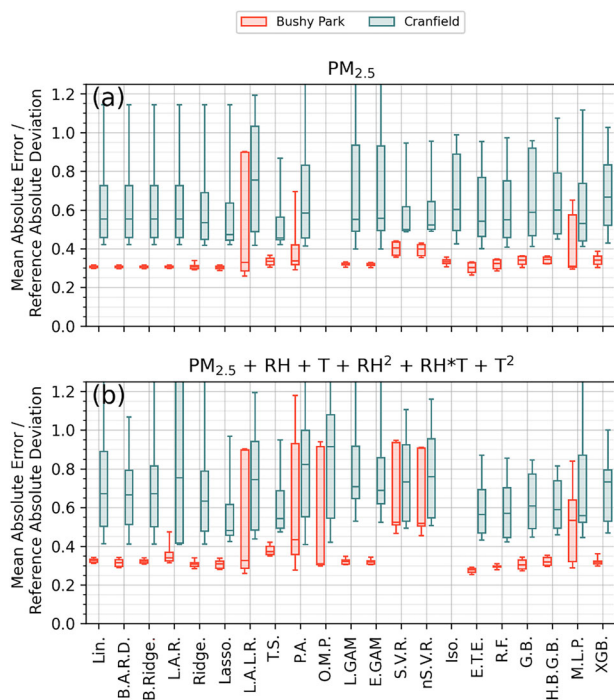


Fig. 9 | Performance of regression techniques correcting $PM_{2.5}$ measurements. These are split by independent variables used: **a** $PM_{2.5}$, and **b** $PM_{2.5} + RH + T + RH^2 + RH \times T + T^2$. All techniques are evaluated by mean absolute error normalised by the absolute deviation of reference measurements.

studies have focused on non-parametric machine learning techniques such as Random Forest and XGBoost, simpler parametric techniques can perform better under the right conditions while being more explainable.

The two GAMs perform particularly poorly when including TSO as an independent variable, with the median nMAE values increasing by over 30 times at the Bushy Park site when only NO_2 and TSO are used.

The nMAE scores worsen at both sites when using NO_2 and TSO in support vector machine-based NO_2 corrections, with the median nMAE scores for support vector regression increasing from 0.73 to 1.11 at the Bushy Park site and from 0.38 to 0.96 at the Cranfield site. This trend is also present when all other secondary variables are included. The performance of non-parametric techniques is mixed when using TSO as a variable in NO_2 corrections. When only using NO_2 and TSO at the Bushy Park site, some techniques, such as XGBoost regression, have lower nMAE scores when using the TSO, while the multi-layer perceptron has a much higher median nMAE, as the weights used in the neural network do not converge when scaling techniques are used, as discussed in Section 2.2.1. However, the only technique to achieve a lower minimum nMAE score is histogram-based gradient boosting regression, decreasing from 0.64 to 0.61. These improvements in the median nMAE scores are not seen at the Cranfield site, with all median nMAE scores increasing for non-parametric techniques when only using NO_2 and TSO.

When using all independent variables, the inclusion of TSO worsens median and minimum nMAE scores for all non-parametric techniques across all sites. Given the improvements gained by parametric techniques, this suggests the non-parametric techniques may have overfit to the relationships within the training set. Using a larger training set so the drift can be modelled over a longer period of time may improve the performance of TSO as an independent variable.

$PM_{2.5}$

The inclusion of secondary variables when correcting $PM_{2.5}$ (Fig. 9, key in Table S1, tabular results in Supplementary Material, Table S16) improves the nMAE scores in select circumstances but often worsens them at both

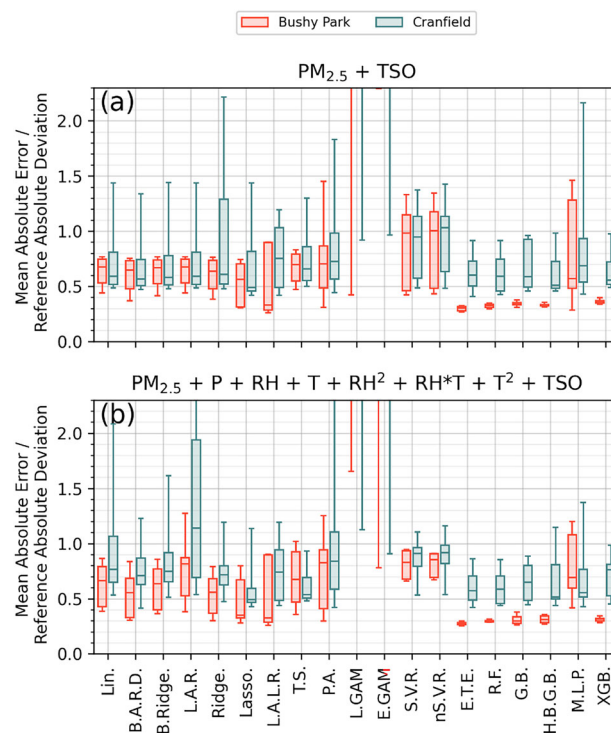


Fig. 10 | Performance of regression techniques using TSO correcting $PM_{2.5}$ measurements. These are split into: **a** techniques using $PM_{2.5}$ and time since origin, and **b** $PM_{2.5}$ with all independent and interaction variables as well as time since origin. All techniques are evaluated by mean absolute error normalised by the absolute deviation of reference measurements.

sites. The minimum nMAE scores very slightly decrease for parametric techniques that apply loss functions to their coefficients. For example, the minimum nMAE score for ridge regression decreases from 0.31 to 0.28 and from 0.42 to 0.41 at Bushy Park and Cranfield, respectively. However, the maximum nMAE scores also increase from 0.32 to 0.33 and from 0.62 to 1.05. This shows that minimising the effects of some coefficients is not enough to minimise the bias introduced by these secondary variables. The performance of GAMs and support vector machines both worsens with the inclusion of secondary variables.

The performance of non-parametric techniques is better when secondary variables are included, with the median and maximum nMAE values decreasing for all techniques at both sites, aside from multi-layer perceptrons, which have some nMAE scores biased high when using MinMax scaling. However, the minimum nMAE scores also increase at the Cranfield site when including secondary variables. As non-parametric techniques improve in performance while parametric techniques worsen when including secondary variables in $PM_{2.5}$ corrections, the interactions between the sensor measurements, relative humidity, and temperature have a high degree of non-linearity. This makes them difficult to model with linear techniques, even with the RH^2 , T^2 , and $RH \times T$ interaction features. The device manufacturers likely already include corrections for relative humidity with their own calibration, so biases would likely only occur at extreme conditions, in this case, very low or high humidity.

Figure 10 (key in Table S1, tabular results in Supplementary Material, Table S17) shows that using both uncorrected $PM_{2.5}$ and TSO as features causes worse performance for most linear models than using $PM_{2.5}$ on its own. Though their average performance is worse, cross-validated Lasso regression, least angle Lasso regression, and passive-aggressive regression do achieve similar performance with or without TSO in some scenarios, with the minimum nMAEs for TSO being 0.31 and 0.44 for Bushy Park and Cranfield, respectively, both landing between the minimum nMAE and the first quartile of nMAEs in their respective sites. These techniques have

proven robust to poor feature selection in NO₂ and O₃ corrections, suggesting they limited the influence of TSO on corrections.

Adding other secondary features does not improve the correction performance significantly. Techniques that apply loss functions to their coefficients, such as ridge regression, do see improvements in their minimum and median nMAE scores, but are both still higher than PM_{2.5} and secondary variables without TSO.

Corrections using GAMS perform significantly worse when adding TSO as a variable, matching trends in NO₂ and O₃. Support vector machine performance also worsens on average. The maximum nMAE for support vector regression increases from 0.44 and 0.51 to 1.33 and 1.37 at Bushy Park and Cranfield, respectively. When adding all secondary features, maximum nMAEs are equivalent whether TSO is used or not. However, the minimum nMAEs are higher when adding TSO, showing that, though TSO does not impact the performance of the worst support vector machine corrections, it does reduce the quality of the best ones.

Outside of multi-layer perceptrons, which have issues when scaling the data, the non-parametric technique's performance does not significantly differ between using TSO with PM_{2.5} and PM_{2.5} on its own. When using all secondary features and TSO, non-parametric techniques generally perform slightly better than all secondary features without TSO, but the performance between the two is similar, suggesting that if baseline drift was effectively modelled, the effects were minimal. Given that the performance of OPCs is particularly susceptible to seasonal effects from changes in particulate composition, it is unlikely that the intricacies of these effects could be captured in 72 days' worth of training data, and a longer training period may yield better results. Overall, there appears to be minimal benefit from correcting baseline drift in PM_{2.5} measurements using TSO as a feature when using only 72 days of training data.

Simpson's paradox

Previous studies examining the performance of sensor system corrections often have a narrow scope, with limited or no variation in pre-processing techniques, independent variables used, and regression techniques. The goal of this study was to examine a broad spectrum of pre-processing and regression techniques to determine the optimal combination to use when deploying pre-calibrated sensors in the field, providing a definitive methodology to the preparation stage of a small-scale study or low-cost sensor network. However, after examining more than 390,000 results in the Bushy Park study and over 80,000 results from the Cranfield study, it became clear that there were significant issues in drawing conclusions from the summarised data.

For example, Fig. 5 shows summaries of the performance of different pre-scaling algorithms used to modify the data prior to the regression step. Though the parametric scaling algorithms tend to have lower minimum and median nMAEs than the non-parametric transformation algorithms, there is often not a significant difference, and it could be concluded that the scaling algorithm does not impact performance significantly.

On the other hand, when examining Fig. S2, showing the effects of no pre-scaling, MinMax scaling, and Yeo-Johnson transform on NO₂ corrections, significant differences appear. For example, passive-aggressive regression in the Bushy Park data performs poorly compared to other linear models when the data is not pre-scaled. However, when MinMax scaling is applied, it has the lowest median nMAE for linear models at Bushy Park and the lowest minimum nMAE across all regression models at both sites. Orthogonal matching pursuit also performs much worse without scaling. Once MinMax Scaling is applied to the measurements, the nMAEs decrease significantly, particularly in the Cranfield data, where the median nMAE decreases from 1.00 to 0.37.

The opposite is true for multi-layer perceptron regression. When no pre-scaling is applied, it achieves the lowest minimum nMAE at both sites. However, once MinMax scaling is applied, its performance becomes much more varied, with the minimum nMAE increasing from 0.43 to 0.57 in the Bushy Park study. Least angle Lasso regression also performs worse when MinMax scaling is applied, though least angle regression improves overall,

and Lasso regression sees no change. Non-parametric techniques, aside from multi-layer perceptrons, see no significant changes when applying pre-scaling techniques, as they are not based on the underlying distribution of the data. They make no assumptions about the structure of the data, so the distribution and scaling of the measurements are irrelevant.

When aggregating the results of all techniques, these trends disappear, and no clear relationship between the effect of the pre-scaling technique and measurement performance is seen. This hides its significance in specific scenarios and may cause poor recommendations depending on the technique used. This is an example of Simpson's paradox, a phenomenon where aggregating multiple sets of results can result in a different trend than one seen within each individual set. In this case, pre-scaling the data has a significant impact on the performance of several regression techniques, but the differences tend to cancel each other out when aggregating the data.

This was an issue when comparing results of prior studies³⁷ and has been shown to be an issue in this study as well. There is one or more confounding variables that appear to affect the measurement. Some of them are compared in this study, but others were not due to the complexity of visualising the results. For example, a comparison of the performance of different regression techniques when modifying both the independent variables used and scaling techniques may yield further insights into how varying magnitudes and distribution shapes between independent variables affect different regression techniques. However, this would be a very complex comparison and would create an overwhelming number of graphs, far beyond the scope of this study.

Though the analytical steps used in this study were controlled and applied consistently to all datasets, such as the pre-scaling technique used, the independent variables used, the regression technique, and the length of the training set, some factors were not. Neither the distribution of the data in the training set nor the quality of the measurements used was controlled, varying between sites, pollutants, and devices. This can introduce multiple sources of variance within results, which makes direct comparison difficult in this study.

Different types of devices were also used, with none shared between the two sites. AQMesh sensors, Nova sensors, and Thermo-Fischer PDR1500s were used in the Bushy Park study, while Earthsense Zephyr sensors were used in the Cranfield study. Uncertainties can be introduced not just by any different sensing components used in the devices but also by the proprietary algorithms used, potentially introducing unpredictable biases.

Variation in results

The correction process was standardised between pollutants and studies, with all pipelines following the same basic processes, with the only variations being the introduction of secondary and interaction variables, the introduction of the TSO feature, the technique used to pre-scale measurements, and regression techniques (see Fig. 11). However, variances still existed between sites. Normalising the MAE by the absolute deviation of the reference measurements in the validation set (nMAE, Fig. 1c) helped minimise differences in average performance between different pollutants, but did not significantly improve variations in performance between the two sites, particularly for NO₂. Normalising the MAE by the mean of the reference measurements in the validation set (CvMAE, Fig. 1b) worsened the difference between the Bushy Park and Cranfield studies for NO₂ relative to non-normalised MAE (Fig. 1a).

This could be due to a number of factors. Variance can be introduced not just between different types of low-cost sensing components, but variations within components of the same type. Components used for secondary measurements, such as temperature and humidity, can also vary, meaning two individual low-cost sensor systems can potentially drastically vary in performance. The measurements were also filtered differently. The AQMesh data and AirView Reference data came with status codes, indicating whether the measurements were valid or whether they were in a rebasing period and/or outside the measurement limits of the device. AQMesh measurements were filtered to only accept "Valid" measurements.

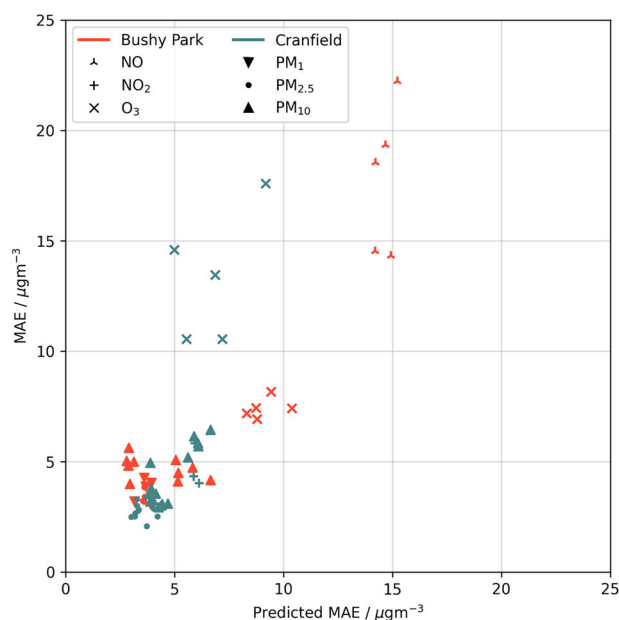


Fig. 11 | Predicted MAE scores vs. actual MAE scores using measurement data.

Table 1 | Results for model predicting the MAE of a passive-aggressive regression correction on MinMax-scaled data using features from the dataset, such as mean and standard deviation, in an extra trees ensemble regressor

Metric	Result	Result without O ₃ , Zephyr C
r^2	0.69	0.83
Root mean squared error	2.54 $\mu\text{g m}^{-3}$	1.76 $\mu\text{g m}^{-3}$
Mean absolute error	1.57 $\mu\text{g m}^{-3}$	1.21 $\mu\text{g m}^{-3}$
Mean bias error	-0.39 $\mu\text{g m}^{-3}$	0.13 $\mu\text{g m}^{-3}$
95% Quantile absolute error	6.43 $\mu\text{g m}^{-3}$	3.51 $\mu\text{g m}^{-3}$

The AirView Reference status codes corresponded to a range of conditions, including flow error, temperature, and humidity outside of limits and automated calibration periods. These were all reported by the individual reference and reference equivalent devices. Only measurements where the devices reported no adverse conditions were used.

As these corrections were not based on the raw sensor outputs, but rather the pre-calibrated measurements, some of the variation may have been removed by a “black-box” calibration algorithm. This may limit the ability to quantify how sensor variance affects the correction performance, as it is at least partially occluded by the manufacturer’s proprietary calibration and correction routines. These contributors to uncertainty are likely abstracted away or reduced. To examine whether these variations in performance can be measured, an experiment was run to test whether statistics such as the mean and absolute deviation of the measurements can be used to predict the MAE score when using MinMax scaling and passive-aggressive regression.

The results were filtered for passive-aggressive regression corrections pre-scaled with MinMax scaling using all secondary features except for TSO. All MAE scores were obtained for NO₂, O₃, PM_{2.5}, and PM₁₀ for both studies. NO, PM₁, and PM₄ MAE scores were also obtained for the Bushy Park site. The mean, median, standard deviation, absolute deviation, range, interquartile range, skew, maximum measurement, 2.5th quantile, 25th quantile, 75th quantile, and skew were obtained for both the uncorrected and reference Measurements. Interaction variables were also calculated by multiplying and dividing all these statistics by each other, similar to $RH \times T$ in the correction pipelines above. Interactions between a variable and itself

were not calculated. This resulted in 715 separate statistics associated with the measurement.

The measurements were split into a train and test set, with ~75% of the measurements used in the training set. All folds from each site were grouped in the same set to minimise the chance of data leakage between the train and test sets.

The features were then MinMax scaled and used in an extra trees ensemble regression algorithm to predict the MAE when using MinMax scaling and passive-aggressive regression. Figure 11 shows the actual MAE against the predicted MAE scores of all studies in the test set. Significant outliers can be seen in Zephyr C corrections for O₃. Removing these measurements from the test set significantly improves the performance of predictions as seen in Table 1. This shows that some relationship can be modelled between the uncorrected and reference measurements, though there are outliers, particularly in the case of Zephyr C’s O₃ measurements. Further work should focus on what causes the variance between devices, as measurement statistics on their own cannot normalise results between studies, making comparisons difficult.

Comparisons to previous work

Comparison between this study and others is difficult due to the breadth of results and the difficulty in comparison between co-location campaigns. However, we can reinforce trends found in other studies and highlight where others could see potential improvements in results.

Cordero and co-workers⁴⁸ focused on correcting NO₂ measurements of four AQMesh pods, utilising the uncorrected NO₂, NO, and O₃ measurements along with temperature as independent variables. They also obtained mixed results when applying correction techniques, with three sensor systems performing best with Random Forest and one with support vector machines based on the mean absolute error. The performance improvements of using Random Forest over linear regression varied, with drops in MAE between 7% and 20% for three systems, and an increase of 17% for the remaining system. Artificial neural networks, an approach similar to multi-layer perceptrons, were also tested. The MAE of the two sensor systems more than doubled compared to linear regression, matching the very varied performance seen in this study. Standard scaling was used to transform the pre-corrected measurements for artificial neural networks, but no comparison was made against unscaled data. Albeit on a smaller scale, the varied performance of different techniques matches the one observed in the present study, and the examination of more techniques and pre-scaling methods may have yielded better results in the study.

Okafor and colleagues⁴⁹ examined the use of univariate and multivariate linear regression and artificial neural networks when correcting NO₂ and O₃ measurements of six sensor systems, though all six were used in a single correction, along with temperature and humidity. The univariate linear regression utilised the first sensor’s measurements only. The addition of multiple variables to linear regression saw an increase in the mean absolute error for O₃ corrections, similar to the increase seen in the median nMAE for linear regression O₃ corrections in this study (Fig. S3). However, when using exhaustive feature selection, a technique used to select only the most appropriate independent variables to use, the performance of linear regression improved significantly. This is similar to trends in this study where linear models that performed feature selection, such as Bayesian automatic relevance detection, showed significantly lower minimum nMAEs than linear regression when using multiple independent variables for O₃ corrections (Fig. S3).

Bush and co-workers⁵⁰ examined the use of Random Forest models to improve low-cost sensor performance. This study used the initial measurements, as well as internal sensor voltages, temperature, humidity, and several engineered features, including the rate of change of temperature and humidity. There are also some controversial features, including the hour of day and day of the week, as well as a flag that highlights measurements made during rush hour that would have no direct impact on biases within the measurement but could improve measurements at a fixed site at the expense

of portability. The Random Forest models gave significant improvements in the mean absolute error, with decreases of 93% in the NO₂ model, 95% in the PM₁₀ model, and 90% in the PM_{2.5} model. However, this is for a single sensor system using a single model with engineered features that potentially reduce its portability. As other models performed better for NO₂ in our study, particularly passive aggressive regression for NO₂, and cross-validated lasso regression and extra trees ensemble for PM_{2.5}, it is possible that other techniques could have improved the measurement quality further.

Discussion

More than 390,000 separate measurement corrections from the Bushy Park site and 80,000 from the Cranfield site were examined as part of this study to compare the best techniques to use when pre-calibrating low-cost air quality sensor measurements. This is several orders of magnitude more results than previous studies, giving a far more robust view of how different correction methodologies perform. By examining the trends across several pollutants, variable combinations, and pre-scaling techniques, a more well-rounded idea of the relative performance of techniques can be formed. For example, in many cases, linear models such as linear regression and passive-aggressive regression perform just as well as, if not better than, non-parametric techniques like Random Forest. There is not a large difference between techniques, particularly within their groupings (linear models, GAMs, SVMs and non-parametric). As these corrections are applied to pre-calibrated measurements, which already account for some biases from external factors (e.g., relative humidity), the influences from these are minimised compared to raw measurements. Applying the same methodology for calibrations of raw sensor readings may yield very different trends.

Looking at trends within the results, though there is significant benefit to some techniques over others, far greater improvements can be made by modifying conditions at the co-location site that were not measured directly here. Examining NO₂ corrections in Fig. 7, though minimum nMAEs were lower when using passive-aggressive regression or multi-layer perceptrons at both sites, far larger differences occurred between Bushy Park and Cranfield. The Cranfield site achieved a median nMAE of 0.29 for linear regression when using all environmental and interaction variables, half that of Bushy Park (0.58). This supports previous findings³⁷ that comparisons cannot easily be made between studies due to a wide range of variations from external factors. Normalising by absolute deviation made direct comparisons between pollutants easier, bringing the median O₃ scores more in line with other pollutants. However, it did not improve the inter-site comparison in NO₂, suggesting other external factors need to be accounted for to be able to compare results properly. A model was set up that predicts the MAE score of passive-aggressive regression using metadata from the low-cost sensor and reference measurements (Fig. 11, Table 1), but it had limited prediction capabilities, particularly with Zephyr C's O₃ measurements.

This variance could be caused by a range of sources previously discussed, including the different devices used, intra-sensor variability, the secondary environmental conditions of the studies, the sizes of the validation sets, and the black-box calibrations applied by the sensor manufacturers. Further work should be undertaken into the causes of the variances between the results of co-location studies to aid in better comparisons between studies that greater inform researchers, citizen scientists, and operators of air quality networks.

Given the difficulty with comparing results from the same co-location campaigns, and more so between the two co-location campaigns in this study, it would be of little consequence to compare the results of this study to others, as the biases that are induced from the conditions of the individual campaigns can sometimes cause more variance in results than the methodologies used to correct the measurements.

Despite a large focus of recent work on non-parametric techniques such as random forest, XGBoost, and neural networks, simpler linear models often performed better at both sites. However, no technique proved itself to be the best-performing overall, with significant variances from a

range of factors. This is similar to findings from previous studies that compare results across the literature^{37,51}. However, this study has the advantage of using more magnitudes of results with less variation in the experimental setup.

As we cannot recommend a single technique to use, we instead recommend that those wishing to correct low-cost sensor measurements test several techniques, including linear models and non-parametric techniques, along with applying different pre-scaling techniques to the measurements. The authors highlight the performance of passive-aggressive regression and extra-trees ensemble across both sites in this study, but other techniques may perform better in different scenarios. The authors also recommend that more studies of similar scope be undertaken with different co-location datasets to enable a proper database of calibration and correction results to be built, informing researchers, citizen scientists, and operators of air quality networks on best practices when correcting measurements made by these devices.

Methods

Co-location studies

The Bushy Park co-location campaign began on 01/01/2018 and ended on 31/12/2021, taking place at the London Teddington Bushy Park AURN monitoring site, an urban background site (51.425286, -0.345606). It featured 10 sensor systems (Table 2) measuring NO, NO₂, O₃, and PM.

Reference data from the Bushy Park AURN site was used, initially measured by a filter dynamic measurement system (FDMS, Table S4) and an optical particle counter (OPC, Table S4) later in the campaign. Two AirView cars, Vauxhall Astras modified with Google StreetView cameras and reference air quality monitoring instrumentation (Table S4), were parked at the Bushy Park site as part of the Breathe London Pilot scheme. The cars were driven around London as mobile air quality monitoring units, but were co-located with the Bushy Park AURN site when not being driven. The measurements were filtered to only use data collected when the cars were stationary in the car park.

The AQMesh sensors were mounted on top of the Bushy Park AURN site (~3 m from the ground) and measured both gaseous and particulate pollution levels with electrochemical sensors and an optical particle counter, respectively. They also recorded relative humidity and temperature data, which are used internally for measurement correction (Table S4). NPL AQMesh A and B were powered with a solar panel and a battery, while NPL AQMesh C and the Breathe London AQMesh were powered from mains electricity.

The duration of this study meant that the AQMesh A and B were deployed for longer than the recommended lifetime of an Alphasense electrochemical sensor. The sensors were serviced during the study by the manufacturer to replace the electrochemical sensors. This was not factored into the study, as we make the assumption that the measurement quality should be maintained by the manufacturer, though in practice, this may not be the case.

The Nova sensors were mounted on top of the Bushy Park AURN site and powered from mains electricity. They measured particulate levels *via* an optical particle counter, which used a heated inlet to reduce humidity, as well as correcting measurements for humidity internally (Table S4).

The Bushy Park monitoring site was used for the PM reference measurements to assess the performance of different correction techniques. All devices, aside from the two AirView cars, were located within 2 m of the Bushy Park AURN site inlet; the AirView cars were parked a short distance away (20 m) as seen in Fig. S1.

Corrections were made to all pollutants measured, but only NO₂ and PM_{2.5} are discussed in detail in this study. Time series plots of the reference measurements for each pollutant, as well as the duration of the training period for each device, can be found in the Supplementary Material, Figs. S37–S40.

The Cranfield co-location campaign began on 01/01/2022 and ended on 31/08/2023, taking place at the top of the facility for airborne atmospheric measurements (FAAM) building on the Cranfield University campus

Table 2 | Devices used in Bushy Park and Cranfield co-location studies and their deployment dates

Device	Ref.	Study	Start date	End date	PM ₁	PM _{2.5}	PM ₄	PM ₁₀	NO	NO ₂	O ₃	T	RH
Bushy Park FDMS	x	Bushy Park	01/01/2018	01/01/2020	–	x	–	–	–	–	–	–	–
Bushy Park OPC	x	Bushy Park	01/01/2020	08/06/2021	–	x	–	x	–	–	–	–	–
AirView A	x	Bushy Park	29/08/2018	25/08/2019	x	x	x	x	x	x	x	–	–
AirView B	x	Bushy Park	28/08/2018	25/08/2019	x	x	x	x	x	x	x	–	–
Nova A	–	Bushy Park	27/02/2019	24/06/2020	–	x	–	x	–	–	–	–	–
Nova B	–	Bushy Park	27/02/2019	24/06/2020	–	x	–	x	–	–	–	–	–
NPL AQMesh A	–	Bushy Park	01/06/2018	17/12/2021	x	x	x	x	x	x	x	x	x
NPL AQMesh B	–	Bushy Park	21/08/2018	11/01/2021	x	x	x	x	x	x	x	x	x
NPL AQMesh C	–	Bushy Park	30/04/2019	01/01/2022	x	x	x	x	x	x	x	x	x
Breathe London AQMesh	–	Bushy Park	13/07/2018	03/03/2021	x	x	x	x	x	x	x	x	x
AirView A (Low-Cost)	–	Bushy Park	29/08/2018	25/08/2019	–	x	–	–	–	–	–	–	–
AirView B (Low-Cost)	–	Bushy Park	28/08/2018	25/08/2019	–	x	–	–	–	–	–	–	–
Cranfield Reference	x	Cranfield	01/06/2022	01/05/2023	x	x	x	x	x	x	x	x	x
Zephyr A	–	Cranfield	01/06/2022	04/04/2023	x	x	x	x	x	x	x	x	x
Zephyr B	–	Cranfield	03/10/2022	04/04/2023	x	x	x	x	x	x	x	x	x
Zephyr C	–	Cranfield	03/10/2022	07/12/2022	x	x	x	x	x	x	x	x	x
Zephyr D	–	Cranfield	04/10/2022	17/01/2023	x	x	x	x	x	x	x	x	x
Zephyr E	–	Cranfield	01/06/2022	04/04/2023	x	x	x	x	x	x	x	x	x

Devices with an x sign in the Ref. column were used as reference or reference equivalent measurements. An x in a pollutant column, indicating that the device was measuring this pollutant during its deployment.

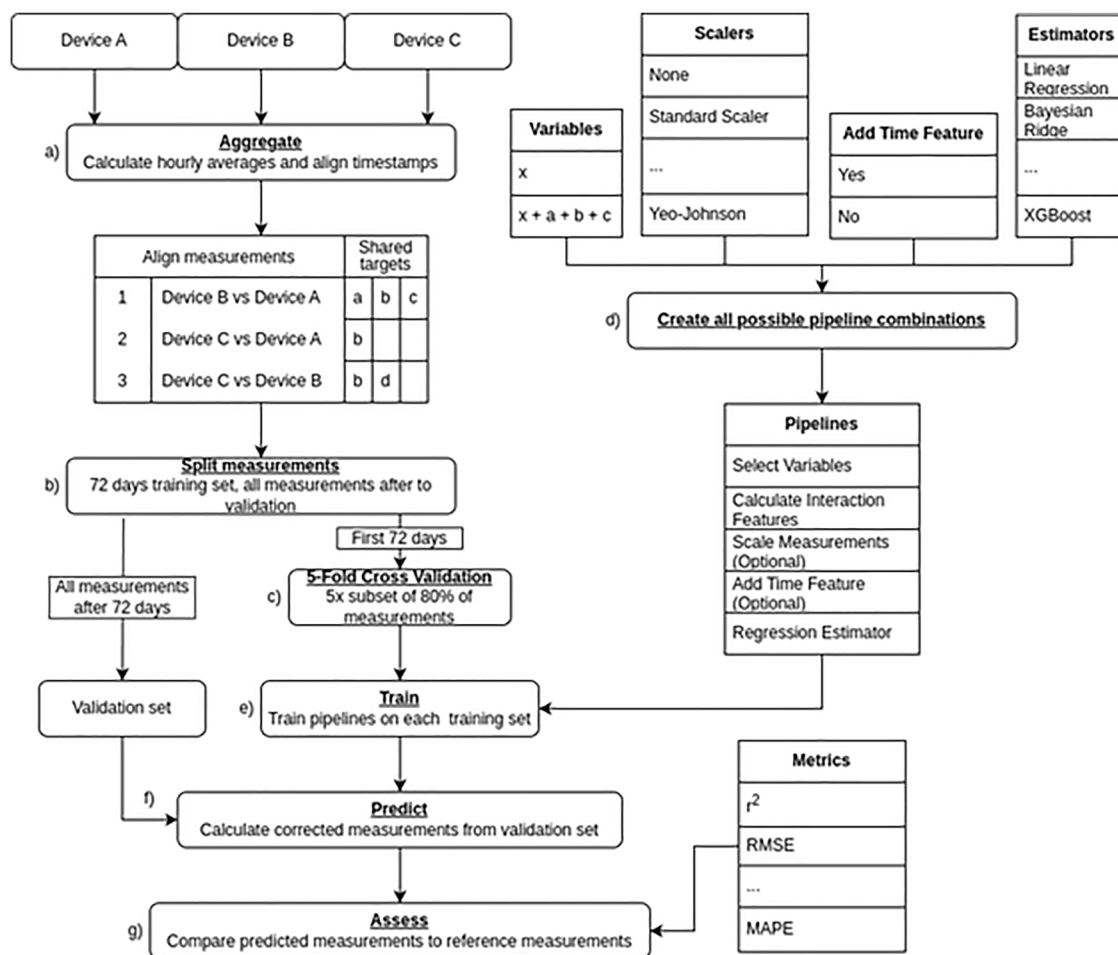


Fig. 12 | Correction methodology used in this study. All key steps are labelled (a–g). Collection, aggregation, and splitting of measurements are shown in the procedure starting in the top left. The building of all possible combinations of pipelines from the specified components is shown in the procedure starting in the top right.

(52.074357, −0.625291) as part of the Cranfield Living Laboratory/Urban Observatory initiative (<https://www.livinglab.ac.uk>). This site is comparable to an urban background site. It featured five low-cost Zephyr sensors and reference instruments, measuring NO, NO₂, O₃, and PM (Table 2). All devices were co-located within 2 m of the reference inlets⁵².

Corrections were made to NO₂, O₃, PM_{2.5}, and PM₁₀, but only NO₂ and PM_{2.5} are discussed in detail in this study. Time series plots of the reference measurements for each pollutant, as well as the duration of the training period for each device, can be found in the Supplementary Material, Figs. S41–S44.

Building and fitting regression pipelines

The target pollutants and secondary variables (see Table 2) were downloaded from a database and averaged to hourly intervals (hour ending). Timestamps were adjusted to align all measurements to the hour ending where necessary. A comparison for each set of measurements involving a co-located low-cost sensor and reference device was then created (Fig. 12a).

For each site, the measurements in each comparison were aligned by timestamp, excluding any measurements that a low-cost sensor made while the reference device was not measuring and vice versa. Any measurements made at a specific time with missing data were removed, as many regression algorithms cannot work with missing data.

The first 72 days of measurements from each comparison were used as the training set, with all remaining measurements being used as the validation set. Any comparisons with less than 72 days of measurements were discarded. To attempt to minimise the impacts of overfitting, the training set was divided into five shuffled sets via 5-fold cross-validation, each

containing 80% of the measurements. Each split was made so that the distribution of reference measurements was equivalent.

Correction “pipelines” were then generated. Each pipeline was composed of a series of transformation steps. Each step processed the provided measurements in turn. The final regression step then transforms the final set of measurements by minimising the error between the training measurements and the reference measurements.

The pipelines contained three to five steps: (1) select variables to use; (2) calculate interaction features for temperature and relative humidity (i.e., T^2 , RH^2 and $T \times RH$); (3) (optional) scale and/or transform the input measurements (e.g., Yeo-Johnson transform); (4) (optional) add time from origin column; and (5) perform regression to predict corrected measurements.

A big advantage of using pipelines is the standardised approach they provide. By using a standardised workflow, we can ensure that one device at Bushy Park and another at Cranfield were corrected in the same way. This ensures that any differences in results between two corrections performed with the same configuration are due to the conditions of the site, not variations in correction methodology.

The first step of the pipeline is used to filter out all independent variables that the user may deem not relevant. In this study, this was used to either select only the target pollutant (i.e., NO₂) or the target pollutant and all available secondary features (e.g., NO₂, RH, T, NO, and O₃, see Table 3).

The study was restricted to these two variations to limit computational intensity. These two separate iterations were run to test the effectiveness of secondary features when correcting low-cost sensor measurements. Some regression techniques can only operate with a

Table 3 | Fields and secondary variables are used as features when correcting measurements

Target pollutants	Secondary variables
PM _{2.5}	RH, <i>T</i>
PM ₁₀	RH, <i>T</i>
NO ₂	RH, <i>T</i> , NO, O ₃
O ₃	RH, <i>T</i> , NO, NO ₂

specific number of features, so these were not run for both iterations of the pipeline. For example, isotonic regression can only work with a single feature, and would not work if NO₂ is paired with relative humidity or temperature. On the other hand, orthogonal matching pursuit can only work with more than one feature, so it would fail if used with only NO₂.

The second step calculated “interaction features” from the features that were selected in the previous step. These features were calculated from relative humidity and temperature measurements if they were present, creating three additional features: RH², *T*², and RH × *T*. These interaction features were calculated to model the non-linear impact of relative humidity and temperature on low-cost air quality sensor measurements as identified in previous studies²⁰.

The third step scales and/or transforms the measurements to normalise the distribution of measurements. These scaling or transformation operations utilise a variety of techniques; a description of each can be found in Table S5. These techniques were tested as some regression techniques can benefit from rescaling all features to be equivalent⁵³.

The non-parametric techniques were included to test whether modifying the distribution of air quality measurements, which are often non-normal, to a more normal one, can improve the performance of regression techniques.

The fourth step adds a column representing the time elapsed since the co-location campaign began. It assigns a value of 0 to the first measurement in the training set and a value of 1 to the last in the training set and linearly interpolates between them. The values in the validation set are all greater than one, calculated based on the time elapsed since the first measurement. This additional feature was included to test whether a variable representing time could be used to counteract baseline drift in low-cost air quality sensors. Some other studies have utilised a timestamp feature when calibrating low-cost air quality sensors, such as Astudillo et al.⁵⁴.

As these devices use black-box algorithms to convert from raw sensor voltages to the final measurement, it is unknown whether corrections for baseline drift are already made by the device. TSO not improving the measurement does not conclusively demonstrate there is no drift within the raw sensor measurements, but could also be the result of drift corrections already being implemented by the manufacturer, or the training set being too limited to isolate long-term drift from the short-term biases caused by changes in environmental conditions.

A range of regression techniques was then tested using the measurements processed by the pipelines (Fig. 12e). A separate pipeline was trained on data from each fold, then all pipelines were used to predict measurements in the same validation set. The use of the same validation measurements for each set of pipelines for a correction candidate ensured all pipelines could be compared fairly.

The validation set of each correction technique was transformed by each pipeline (Fig. 12f) to generate “corrected” measurements. The resulting measurements in each validation set were then compared against the corresponding reference measurements using a wide range of error metrics (Table S3, Fig. 12g). Each fold was tested individually. A mean set of measurements was also calculated by taking the average of the predicted measurements across all five folds. The mean, standard deviation, absolute deviation, range, and interquartile range of the reference measurements

were also calculated to provide several different methods of normalising the results at different sites.

Data availability

The datasets generated and/or analysed during the current study are not publicly available due to mixed licensing of some parts of the underlying data limiting redistribution, but some of the low-cost sensor measurements and all correction results are available from the corresponding author on reasonable request. The code used to generate these comparisons can be found here: [https://github.com/CaderIdris/Graddnodi].

Received: 24 February 2025; Accepted: 9 September 2025;

Published online: 01 December 2025

References

- Borghi, F. et al. Precision and accuracy of a direct-reading miniaturized monitor in PM(2.5) exposure assessment. *Sensors (Basel, Switzerland)* **18**, 3089 (2018).
- Malings, C. et al. Fine particle mass monitoring with low-cost sensors: corrections and long-term performance evaluation. *Aerosol Sci. Technol.* **54**, 160–174 (2020).
- Considine, E. M., Reid, C. E., Ogletree, M. R. & Dye, T. Improving accuracy of air pollution exposure measurements: statistical correction of a municipal low-cost airborne particulate matter sensor network. *Environ. Pollut.* **268**, 115833 (2021).
- Borrego, C. et al. Assessment of air quality microsensors versus reference methods: the EuNetAir joint exercise—part II. *Atmos. Environ.* **193**, 127–142 (2018).
- Liang, Y. et al. Field comparison of electrochemical gas sensor data correction algorithms for ambient air measurements. *Sens. Actuators B: Chem.* **327**, 128897 (2021).
- Bulot, F. M. et al. Long-term field comparison of multiple low-cost particulate matter sensors in an outdoor urban environment. *Sci. Rep.* **9**, (2019).
- Gerboles, M. & Buzica, D. *Evaluation of Micro-Sensors to Monitor Ozone in Ambient Air* (Publications Office of the European Union, 2009).
- Baron, R. & Saffell, J. Amperometric gas sensors as a low cost emerging technology platform for air quality monitoring applications: a review. *ACS Sens.* **2**, 1553–1566 (2017).
- Alfano, B. et al. A review of low-cost particulate matter sensors from the developers’ perspectives. *Sensors* **20**, 6819 (2020).
- Miskell, G., Salmond, J. & Williams, D. E. Solution to the problem of calibration of low-cost air quality measurement sensors in networks. *ACS Sens.* **3**, 832–843 (2018).
- Ouimette, J. et al. Fundamentals of low-cost aerosol sensor design and operation. *Aerosol Sci. Technol.* **58**, 1–15 (2024).
- Rueda, E. M., Carter, E., L’Orange, C., Quinn, C. & Volckens, J. Size-resolved field performance of low-cost sensors for particulate matter air pollution. *Environ. Sci. Technol. Lett.* **10**, 247–253 (2023).
- Kaur, K. & Kelly, K. E. Laboratory evaluation of the Alphasense OPC-N3, and the Plantower PMS5003 and PMS6003 sensors. *J. Aerosol Sci.* **171**, 106181 (2023).
- Hossain, M., Saffell, J. & Baron, R. Differentiating NO₂ and O₃ at low cost air quality amperometric gas sensors. *ACS Sens.* **1**, 1291–1294 (2016).
- Hagan, D. H. et al. Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments. *Atmos. Meas. Tech.* **11**, 315–328 (2018).
- Smith, K. R. et al. Clustering approaches to improve the performance of low cost air pollution sensors. *Faraday Discuss.* **200**, 621–637 (2017).
- Smith, K. R. et al. An improved low-power measurement of ambient NO₂ and O₃ combining electrochemical sensor clusters and machine learning. *Atmos. Meas. Tech.* **12**, 1325–1336 (2019).

18. Sun, L., Westerdahl, D. & Ning, Z. Development and evaluation of a novel and cost-effective approach for low-cost NO₂ sensor drift correction. *Sensors* **17**, 1916–1930 (2017).
19. Zimmerman, N. et al. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmos. Meas. Tech.* **11**, 291–313 (2018).
20. Malings, C. et al. Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. *Atmos. Meas. Tech.* **12**, 903–920 (2019).
21. Weissert, L. F. et al. Low-cost sensors and microscale land use regression: data fusion to resolve air quality variations with high spatial and temporal resolution. *Atmos. Environ.* **213**, 285–295 (2019).
22. Wei, P. et al. Development and evaluation of a robust temperature sensitive algorithm for long term NO₂ gas sensor network data correction. *Atmos. Environ.* **230**, 117509 (2020).
23. Zoest, V. V., Osei, F. B., Stein, A. & Hoek, G. Application of a sensor network of low cost optical particle counters for assessing the impact of quarry emission on its vicinity. *Atmos. Environ.* **211**, 29–37 (2019).
24. Crilley, L. R. et al. Evaluation of a low-cost optical particle counter (alphasense OPC-N₂) for ambient air monitoring. *Atmos. Meas. Tech.* **11**, 709–720 (2018).
25. Johnson, K. K., Bergin, M. H., Russell, A. G. & Hagler, G. S. W. Field test of several low-cost particulate matter sensors in high and low concentration urban environments. *Aerosol Air Qual. Res.* **18**, 565–578 (2018).
26. Liu, L.-J. S., Slaughter, J. C. & Larson, T. V. Comparison of light scattering devices and impactors for particulate measurements in indoor, outdoor, and personal environments. *Environ. Sci. Technol.* **36**, 2977–2986 (2002).
27. Spinazzè, A. et al. Field comparison of instruments for exposure assessment of airborne ultrafine particles and particulate matter. *Atmos. Environ.* **154**, 274–284 (2017).
28. Hagler, G. S. W., Williams, R., Papapostolou, V. & Polidori, A. Air quality sensors and data adjustment algorithms: when is it no longer a measurement? *Environ. Sci. Technol.* **52**, 5530–5531 (2018).
29. Sayahi, T., Butterfield, A. & Kelly, K. E. Long-term field evaluation of the Plantower PMS low-cost particulate matter sensors. *Environ. Pollut.* **245**, 932–940 (2018).
30. Collingwood, S. et al. Investigating measurement variation of modified low-cost particle sensors. *J. Aerosol Sci.* **135**, 21–32 (2019).
31. Northcross, A. L. et al. A low-cost particle counter as a realtime fine-particle mass monitor. *Environ. Sci. Process. Impacts* **15**, 433–439 (2012).
32. Yi, W. Y. et al. A survey of wireless sensor network based air pollution monitoring systems. *Sensors* **15**, 31392–31427 (2015).
33. Njallson, T. & Novosselov, I. Design and optimization of a compact low-cost optical particle sizer. *J. Aerosol Sci.* **119**, 1–12 (2018).
34. Barkjohn, K. K., Gantt, B. & Clements, A. L. Development and application of a United States-wide correction for PM data collected with the PurpleAir sensor. *Atmos. Meas. Tech.* **14**, 4617–4637 (2021).
35. Stavroulas, I. et al. Field evaluation of low-cost PM sensors (PurpleAir PA-II) under variable urban air quality conditions, in Greece. *Atmosphere* **11**, (2020).
36. Miech, J. A. et al. Calibration of low-cost NO₂ sensors through environmental factor correction. *Toxics* **9**, (2021).
37. Hayward, I., Martin, N. A., Ferracci, V., Kazemimanes, M. & Kumar, P. Low-cost air quality sensors: biases, corrections and challenges in their comparability. *Atmosphere* **15**, (2024).
38. Padilla, L. E. et al. New methods to derive street-scale spatial patterns of air pollution from mobile monitoring. *Atmos. Environ.* **270**, 118851 (2022).
39. Peters, D. R. et al. Evaluating uncertainty in sensor networks for urban air pollution insights. *Atmos. Meas. Tech.* **15**, 321–334 (2022).
40. Schmitz, S., Caseiro, A., Kerschbaumer, A. & Schneidmesser, E. von Do new bike lanes impact air pollution exposure for cyclists?—A case study from Berlin. *Environ. Res. Lett.* **16**, 084031 (2021).
41. Munir, S. et al. The impact of smart traffic interventions on roadside air quality employing machine learning approaches. *Transp. Res. Part D Transp. Environ.* **110**, 103408 (2022).
42. Schmitz, S. et al. Calibrating low-cost sensors to measure vertical and horizontal gradients of NO₂ and O₃ pollution in three street canyons in Berlin. *Atmos. Environ.* **307**, 119830 (2023).
43. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
44. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds. Krishnapuram, B. et al.) 785–794 (Association for Computing Machinery, 2016).
45. Servén, D. & Brummitt, C. pyGAM: Generalized additive models in Python. *Zenodo* <https://doi.org/10.5281/zenodo.1208724> (2018).
46. Buitinck, L. et al. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning* (eds. Bruno, C., Luc, D. R., Paolo, F. & Tias, G.) 108–122 (2013).
47. CEN/TC 264. CEN/TS 17660-2:2024 (E) Vol. 1 (European Committee for Standardisation, 2024).
48. Cordero, J. M., Borge, R. & Narros, A. Using statistical methods to carry out in field calibrations of low cost air quality sensors. *Sens. Actuators B: Chem.* **267**, 245–254 (2018).
49. Okafor, N. U., Alghorani, Y. & Delaney, D. T. Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine learning approach. *ICT Express* **6**, 220–228 (2020).
50. Bush, T. et al. Machine learning techniques to improve the field performance of low-cost air quality sensors. *Atmos. Meas. Tech.* **15**, 3261–3278 (2022).
51. Kang, Y., Aye, L., Ngo, T. D. & Zhou, J. Performance evaluation of low-cost air quality sensors: a review. *Sci. Total Environ.* **818**, 151769 (2022).
52. Li, J., Nasar, Z., Ferracci, V., Harris, N. & Xu, Z. Selection and aggregation of low-cost particle sensors for outdoor particulate matter measurement. *2024 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)* (eds Niewczas, P. & Uttamchandani, D.) 1–6. <https://doi.org/10.1109/I2MTC60896.2024.10561189.48> (2024).
53. Marquardt, D. W. & Snee, R. D. Ridge regression in practice. *Am. Stat.* **29**, 3–20 (1975).
54. Astudillo, G. D., Garza-Castañón, L. E. & Minchala Avila, L. I. Design and evaluation of a reliable low-cost atmospheric pollution station in urban environment. *IEEE Access* **8**, 51129–51144 (2020).

Acknowledgements

We thank the University of Surrey and its Department of Civil and Environmental Engineering for an EPSRC Ph.D. studentship (2018–2021; grant reference #2126120), awarded through the University Research Scholarship scheme, to support I.H.’s Ph.D. research. P.K. acknowledges the support received through the UKRI (NERC, EPSRC, AHRC) funded RECLAIM Network Plus (EP/W034034/1), GP4Streets (UKRI1281), GREENIN Micro Network Plus (APP55977), the NERC-funded GreenCities (NE/X002799/1), and the UGPN-funded (UGPN-NBS and GREENICON) projects. We also thank Georgie Coppin and Stefan Bell for their work maintaining equipment at the Bushy Park co-location site. The authors would also like to thank the Breathe London Pilot Scheme consortium led by Environmental Defence Fund Europe and funded by the Clean Air Fund for generating and providing public access air quality data during the course of the project, part of which formed part of the Bushy Park co-location study. The Cranfield University Urban Observatory is funded by the UKCRIC: Urban Observatories (Strand B)—(EPSRC EP/P016782/1) and the UKCRIC—CORONA: City Observatory Research platform for iNnova-tion and Analytics projects (EPSRC EP/R013411/1). These are supported by the UKCRIC

Coordination Node (EPSRC EP/R017727/1), which funds UKCRIC's ongoing coordination. It is also supported by the Cranfield University Net Zero Research Airport project, funded by Research England's UK Research Partnership Investment Fund (UKRPIF) Net Zero Funding Scheme. The authors are grateful to Prof. Neil R.P. Harris for his support for the Cranfield University Urban Observatory and related projects.

Author contributions

I.H. wrote the code to import and analyse the data and wrote the first draft of the paper, then provided editorial oversight. N.A.M., V.F. and M.K. assisted with data collection in the Bushy Park co-location study, data analysis, result interpretation, and made edits to the paper. V.F., S.J., C.W., and Z.A.N. performed data collection in the Cranfield co-location study, assisted with results interpretation, and made edits to the paper. P.K. assisted with the study conceptualisation, supervision, funding acquisition, interpretation of results, and made edits to the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41612-025-01231-5>.

Correspondence and requests for materials should be addressed to Prashant Kumar.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025



DSPACE

<https://dspace.org/>

Comprehensive comparison of correction techniques for low-cost air quality sensors: the impact of device type and deployment environment

Hayward, Idris; Martin, Nicholas A.; Ferracci, Valerio; Kazemimanesh, Mohsen; Jude, Simon ...

2025-12-01

Springer

<https://doi.org/10.1038/s41612-025-01231-5>

Hayward I, Martin NA, Ferracci V, et al., (2025) Comprehensive comparison of correction techniques for low-cost air quality sensors: the impact of device type and deployment environment. npj Climate and Atmospheric Science, Volume 8, December 2025, Article number 389