Sensor performance and calibration evaluation using reference monitor collocations

The Breathe London project carried out hundreds of collocations at regulatory monitoring sites and in the field at Breathe London sensor locations to calibrate sensors and assess sensor performance. In this report, we use collocations conducted in 2018 and 2019 at three reference-grade monitoring stations in Greater London to characterize the out-of-box (uncalibrated) uncertainty of our AQMesh instruments as compared to ratified reference measurements. We then compare the uncalibrated performance of the sensors to measurements calibrated using a conventional calibration approach (collocation linear regression calibration) and a novel approach (cloud-based network calibration method, described in Appendix 2C). We also use the reference collocation dataset to characterize longer-term uncertainties in NO_2 sensor performance. Lastly, we discuss the effect of reference data ratification procedures on calibration results.

The results presented in this report are unique to the sensors for the specific pollutants that we analyze. We focus mainly on NO_2 and $PM_{2.5}$ which are the two publicly available datasets on the Breathe London website[1]. In the final section of this report we provide supplemental figures for NO and PM_{10} uncertainty results, the two additional pollutants with extensive reference collocation data available.

Key Findings

- Multiple calibration approaches are effective in reducing the error of sensor measurements, but the robustness of applied calibrations over time is compromised by long-term variations in sensor bias
- Maintaining repeat and/or continuous collocations with reference sites for the duration of the campaign is critical to characterizing and mitigating long-term sensor performance issues.

Report Sections

1. Collocations and calibration methods applied
2. Uncalibrated sensor performance
3. Physical collocation calibration evaluation
4. Cloud-based network calibration evaluation
5. Comparison of calibration approaches
6. Transfer standard (“gold pod”) calibration uncertainty
7. Effect of reference data ratification on calibration results
8. Supplemental figures

[1] Note: Analysis results throughout this report for $PM_{2.5}$ and PM_{10} measurements are obtained using a filtered dataset that excludes hours with low visibility (< 10 km) and/or high relative humidity (> 90%). These high humidity/low visibility conditions appeared to correspond with a divergence in AQMesh accuracy compared to reference instruments. Including these periods may lead to increased uncertainty estimates of performance during collocations.

1. Collocations and calibration methods applied

The collocation operational procedures for the Breathe London project are described in the QAQC manual (Appendix 2A). In this report, we evaluate each physical collocation of a sensor pod with a reference instrument by analyzing each pair of collocation timeseries with three aims: (1) to characterize the uncertainty of uncalibrated sensor measurements compared to the reference instrument; (2) to derive a manual calibration factor for the collocated sensor so that its measurements most closely agree with the reference instrument during the collocation period; and (3) to characterize the uncertainty of sensor measurements after application of the manual calibration and an independently derived cloud-based calibration algorithm developed by the University of Cambridge (see Appendix 2C). Below, we describe the collocation analysis process and provide a visual example.

For each collocation, we compare hourly average pollutant concentration data from our AQMesh pods to corresponding ratified measurements from the collocated reference instrument. The duration of collocations is shown in the left panel of **Figure 2**, with the typical collocation lasting between 7-14 days[2]. We calculate a suite of statistics for the collocation timeseries including correlation between the two sets of measurements (R^2), mean bias of the AQMesh sensor measurements, and two estimates of timeseries error including root mean square error (RMSE) and mean absolute error (MAE). We calculate the bias and error associated with the raw sensor measurements, as well as measurements calibrated using two distinct methods: collocation linear regression calibration and cloud-based network calibration (described in Appendix 2C).

Figure 1 (right) shows an example analysis result from a typical reference site collocation. The scatter plot (top panel) shows the linear regression calibration result (blue line) and the cloud-based network calibration result (green line). The three following graphs show the timeseries of reference measurements (black line) compared to uncalibrated sensor measurements, linear regression calibrated sensor measurements, and cloud-based network calibrated sensor measurements

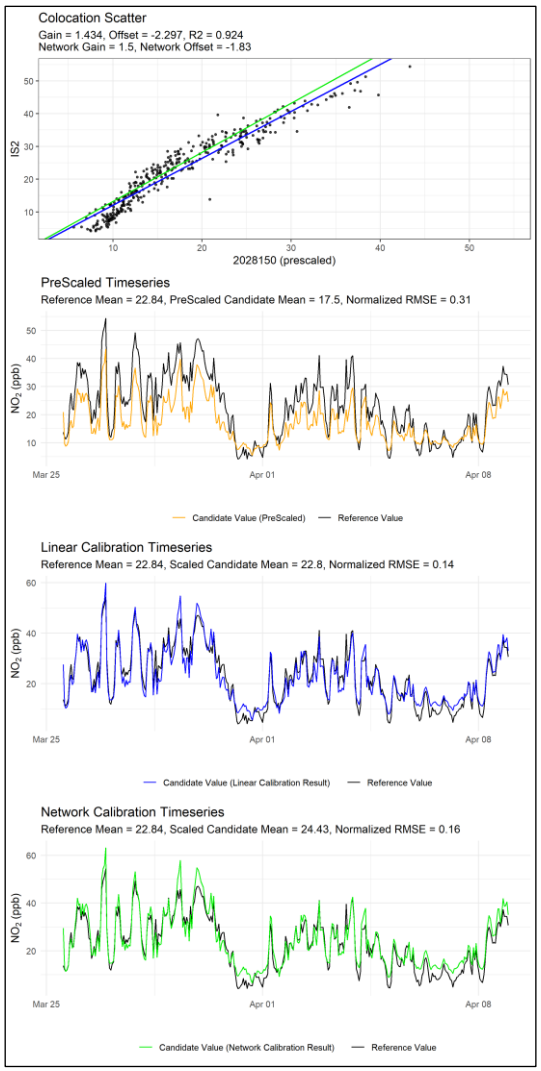


Figure 1 – Example NO_2 collocation analysis result for Pod 2028 at IS2 (Holloway Road) reference site in early 2019

[2] During the beginning of the Breathe London Project, a series of pre-deployment collocations were conducted that lasted as short as 1-3 days. Subsequent rounds of collocations aimed for a 1-2 week duration.

respectively. We analyze the bias and error of these different timeseries compared to ratified reference data to estimate the measurement error associated with the raw and calibrated data sets. In the illustrative **Figure 1**, we apply the linear regression calibration to the time-series used to derive the calibration gain and offset. Because this regression-based approach is designed to minimize the residuals between the reference and test data, the results are not necessarily indicative of the performance of the linear calibration method to other time periods. Consequently, we also evaluate the performance of the linear calibration method using other time periods than the one used to derive the manual calibration factors.

2. Uncalibrated sensor performance

The Breathe London project determined that calibration was an essential part of the QAQC process, due to the substantial uncertainty of uncalibrated pod measurements. In this section, we characterize the performance of hourly-average uncalibrated "out-of-box" pod measurements. **Figure 2** shows the distribution of analysis results of all short-term (<21 days) reference site collocations: n=66 total (n=44 unique sensors and 22 repeats) for NO_2 and n=23 total (n=19 unique sensors and 4 repeats) for $PM_{2.5}$. **Figure 1** shows an example of an individual collocation analysis result. Note that our QAQC procedure would exclude poor calibration results based on statistical criteria for each calibration method (Appendix 2A), so this is not a reflection of the quality of our QAQC'd data, but rather what sensor performance would have been if we had not calibrated the network or redacted any data.

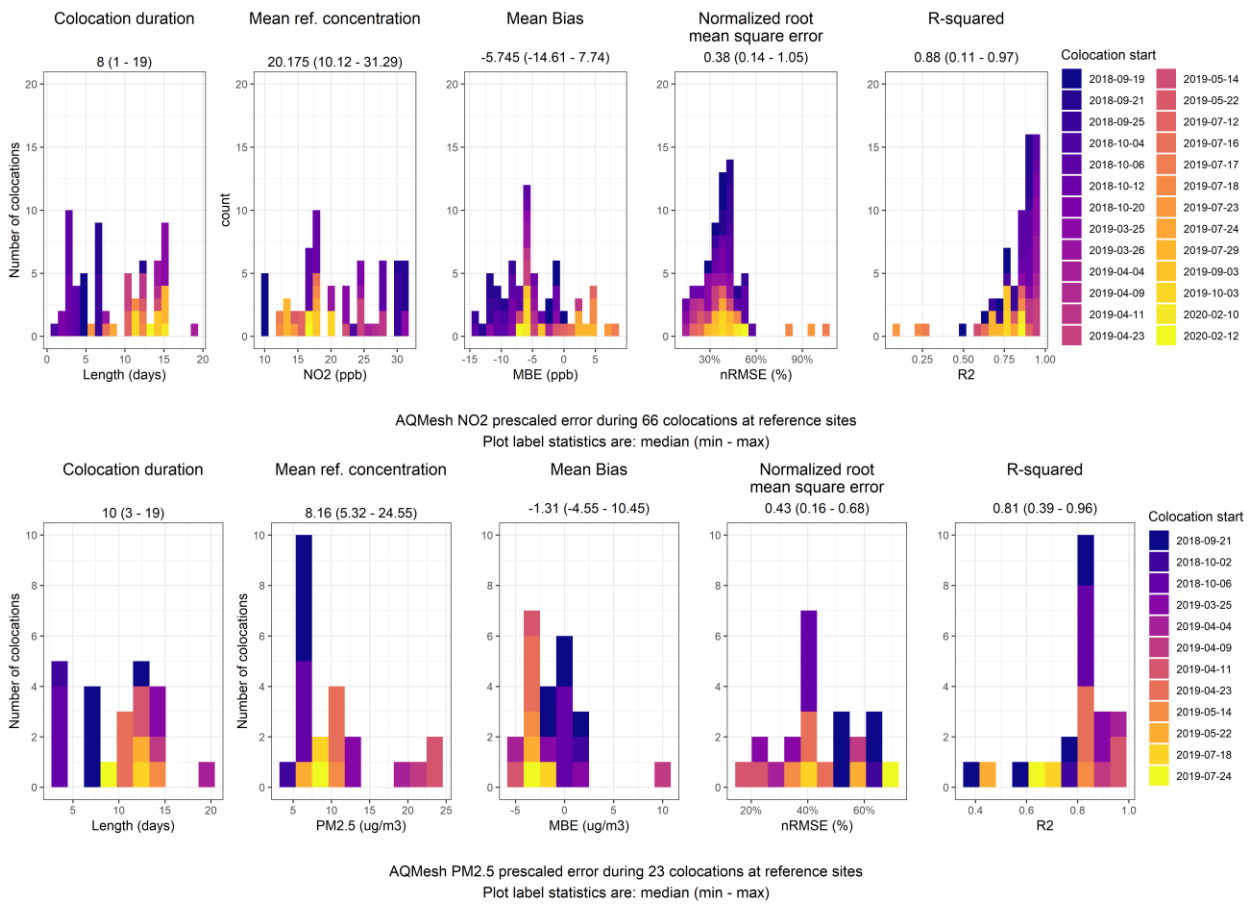


Figure 2 – Distribution of uncalibrated collocation statistics, NO_2 (top) and $PM_{2.5}$ (bottom), for all short-term (<21 days) ratified reference site collocations during Breathe London project. The color scale indicates the start date of each collocation period.

Key Points

- The median R^2 values for NO_2 and $PM_{2.5}$ are 0.88 and 0.81 respectively. These relatively high R^2 values suggest that the sensor response exhibits good linearity over a range of concentrations and supports the use of a linear calibration approach.
- For both pollutants, the median normalized RMSE of pre-scaled measurements is on the order of 40%. This means that during an average collocation, the uncalibrated AQMesh sensor measurement error was about 40% of the mean air pollutant concentration during that collocation.
- The distributions of collocation bias are centered on -5.7 and -1.3 for NO_2 and $PM_{2.5}$ respectively, suggesting that there is an overall low bias in the uncalibrated AQMesh sensors for both pollutants that would be reflected in estimates of uncalibrated network mean concentrations.
- Mean bias results for individual collocations range from -14.6 to +7.7 for NO_2 and -4.6 to +10.5 for $PM_{2.5}$. The observed range in mean bias across the uncalibrated network would effectively

prevent the identification of spatial differences in pollution levels smaller than 10-15 ppb NO_2 and 5-10 µg/m3 $PM_{2.5}$.

- For NO_2, there is a clear temporal trend in the bias of individual collocations. Sensor measurements early in the campaign (Fall 2018 - dark purple) tend to have a low bias against reference measurements, and in Summer 2019 (orange) have a high bias. The seasonal variation in bias likely reflects the effects of ozone cross-interference. This time-dependent bias effect broadens the overall range of NO_2 bias in pre-scaled data and also has implications for calibrated measurements which we discuss further below.

3. Physical collocation calibration evaluation

In this section, we evaluate the measurement uncertainty of sensors when they are calibrated with a linear regression from a reference site collocation (Figure 1, blue lines).

3.1 Short-term sensor performance during reference site calibration

We first analyze sensor performance during short-term reference site collocations when sensors are calibrated using the linear regression result from the same collocation. Our results in Section 3.2 (below) will demonstrate that the measurement error of pods during "calibration" periods (i.e. when we calculate error from the same reference data that the pod was calibrated against) is not representative of the realistic applied or long-term measurement uncertainty of the sensors. However, these calibration period error results are still useful as an indication of the short-term accuracy and precision of the sensor under ideal calibrated conditions with bias eliminated.

Figure 3 shows that calibrated NO_2 and $PM_{2.5}$ sensors have a median normalized RMSE of 15% and 33% respectively during the collocation periods that calibrations were derived from. Sensor bias during these periods is essentially eliminated by the least squares regression used to calibrate the sensors directly against the reference measurements. The results demonstrate the short-term accuracy of the sensors during collocation calibration periods.

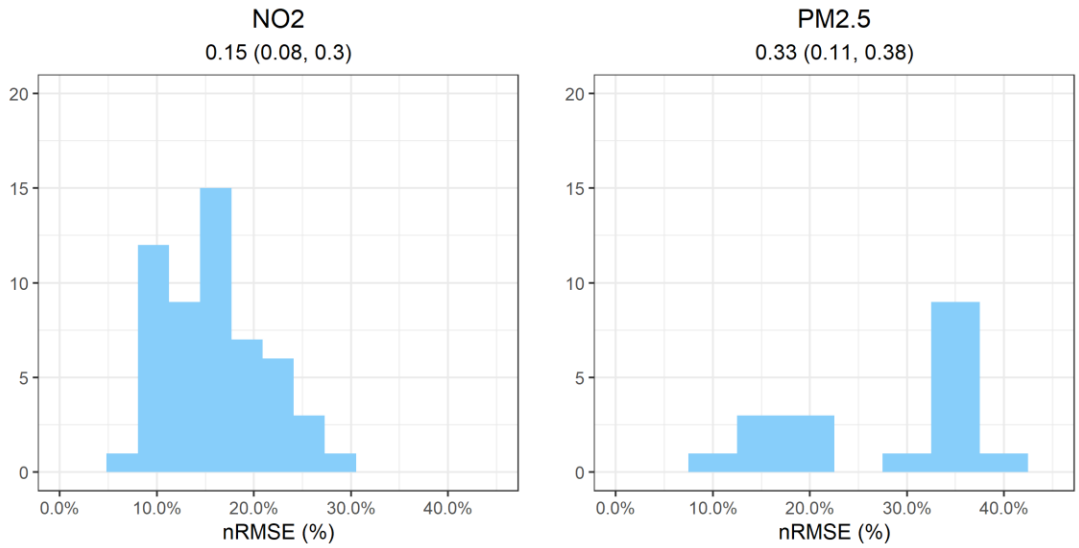


*Figure 3 – Linear regression calibrated error results during reference site collocation calibration periods; NO_2 (left) and $PM_{2.5}$ (right). Statistics shown are median (min, max). n = 54 collocations, 39 unique sensors for NO_2 and 18 collocations, 16 unique sensors for $PM_{2.5}$. Collocations shown here pass QAQC criteria (**Appendix 2A**) of $R^2 > 0.7$ and nRMSE < 0.5 for collocation calibrations.*

3.2 Long-term performance of calibrated sensors

During the project, physical collocation calibrations at reference sites were typically 2 weeks or less (**Figure 2**, left panel). The resulting calibration result was used to scale pod measurements for an extended period while the pod was subsequently deployed in the field. Therefore, it is necessary to understand how robust the calibration result is from a short-term collocation when it is applied to a long-term timeseries. We explore the long-term uncertainties of collocation-calibrated sensors below.

We use two different datasets to study how robust a physical calibration is when applied to a longer timeseries:

1. **Long-term collocations** – Where a sensor has been collocated for an extended time period (>6 weeks) at a reference monitor
2. **Serial collocations** – Where the same sensor has been collocated during multiple time periods

Due to limited data availability for long-term and serial $PM_{2.5}$ collocations[3], the following results are for NO_2 sensors only. In **Section 8** we present supplemental results for NO and PM_{10}.

[3] Of the three regulatory monitoring stations (CD1, IS2, and SK6) where BL collocations were performed, only one (CD1) measured $PM_{2.5}$.

3.2.1 Sensor performance evaluation during long-term reference site collocations

Methods

We test the long-term performance of NO_2 sensor measurements using 4 long-term collocations at reference sites. Rather than calibrating the pods using the entire collocation timeseries, we simulate the calibration we would have obtained from a short-term collocation (Section 3.1) by treating the first two weeks of the long-term collocation as the "calibration" period. Like in the project dataset, this calibration is then applied to the rest of that pod's timeseries of measurements. We then use all non-calibration "testing" periods (that is, all bi-weekly periods except for the initial period that was used to calibrate the timeseries) to calculate the normalized RMSE and mean bias of AQMesh measurements compared to the collocated reference instrument (i.e. for each bi-weekly period, nRMSE is calculated from hourly AQMesh measurements vs. hourly reference measurements).

Results

Table 1 summarizes the measurement bias and error of AQMesh NO_2 measurements during all the long-term collocations. The timeseries of bi-weekly error and bias results are shown in Figures 4 and 5. The summary table and timeseries plots show that when an initial reference site calibration is applied to long-term sensor measurements, the error during bi-weekly "testing" periods is typically significantly higher than during the initial calibration period, with the median long-term nRMSE during non-calibration periods ranging from 15.7% to 42.4% (Table 1). Additionally, each of the sensors had a high bias during the long-term collocation when compared to the reference, with median "testing period" biases ranging from +2.0 ppb to + 3.9 ppb. Notably, biases during individual bi-weekly testing periods reached even higher (+5 to 7 ppb).

Figures 4 and 5 show that long-term trends in bias and error appear to mirror one another (i.e. error fluctuations may be driven largely by changes in bias), and long-term bias trends appear to be consistent across sensors (Figure 5), highlighting a time-dependent oscillation in sensor bias which is believed to be associated with ozone cross-interference for our batch of NO_2 sensors. The seasonal bias in Figure 5 shows an apparent increase into Spring and Summer and a reduction during the winter.

Table 1 and Figure 4 emphasize that even with a robust initial reference site collocation, long-term uncertainties in sensor performance can lead to significantly expanded error (>3× the initial calibration period nRMSE during the worst-case testing periods).

Table 1: Long-term bias and normalized RMSE of AQMesh NO_2 measurements during calibration and long-term testing periods. Note that Pod 83, which exhibits the highest nRMSE during testing and calibration periods, had consistently lower R^2 values than other pods (see Figures 4 and 5) indicating suspect sensor performance.

	Long-term "testing" period			Initial "calibration" period	
	Number of unique bi-weekly "testing" periods	nRMSE median (min, max)	Mean bias (ppb) median (min, max)	nRMSE	Mean bias (ppb)
Pod 17 at IS2	25	21.4% (12.7%, 54.5%)	2.0 (0.9, 5.5)	12.4%	0.0
Pod 79 at SK6	8	19.6% (14.1%, 55.3%)	2.9 (-0.8, 4.6)	17.5%	0.0
Pod 83 at SK6	25	42.4% (21.5%, 103.1%)	3.0 (-2.2, 7.3)	29.6%	0.0
Pod 99 at CD1	8	15.7% (10.2%, 40.4%)	3.9 (1.7, 7.1)	10.2%	0.0

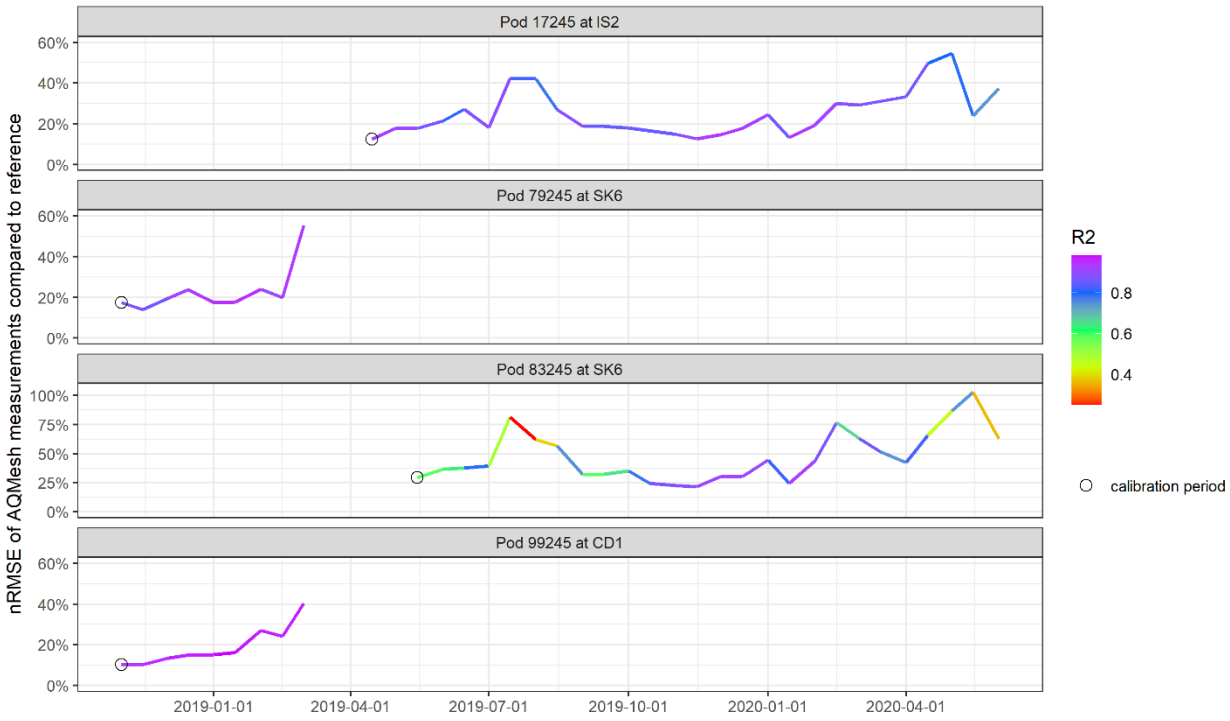


Figure 4 – Bi-weekly nRMSE (normalized RMSE) of AQMesh NO_2 measurements compared to reference measurements, for four long-term collocations at three different reference monitors. Reference data is ratified through Feb 2020. Bi-weekly R^2 of measurements symbolized by color. Each AQMesh sensor is calibrated using data from the first bi-weekly collocation period, as indicated by the hollow circle. Note that the y-axis scale for Pod 83245 at SK6 (third panel from top) differs from other panels.

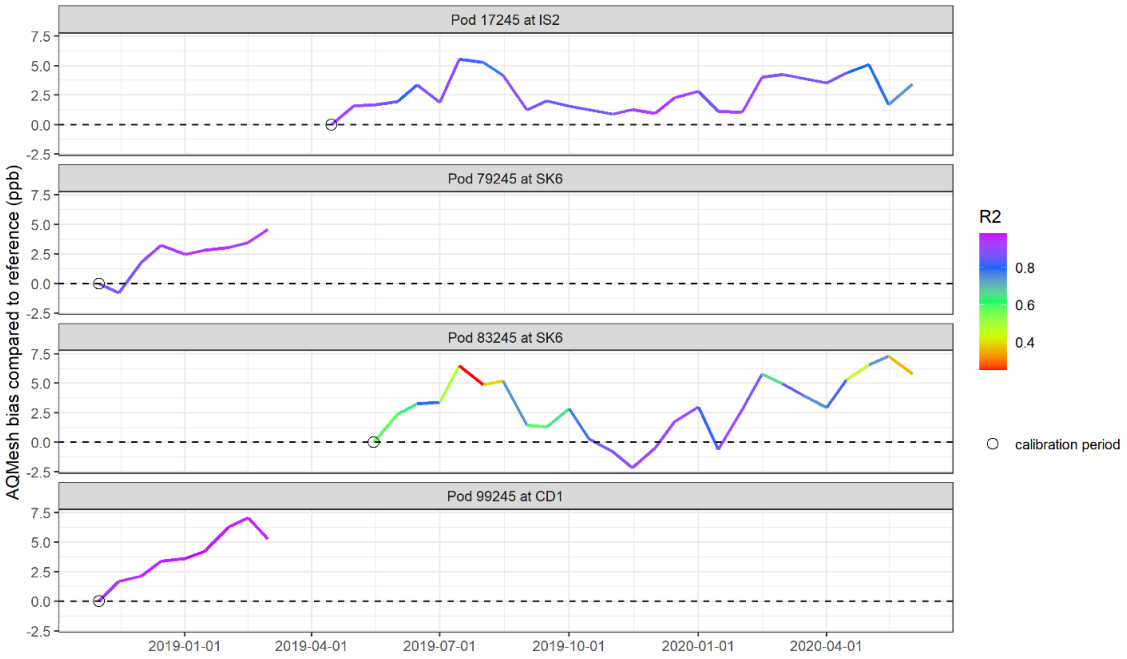


Figure 5 – Bi-weekly bias of AQMesh NO_2 measurements (ppb) compared to reference measurements, for four long-term collocations at three different reference monitors. Reference data is ratified through Feb 2020. Bi-weekly R^2 of measurements symbolized by color. Each AQMesh sensor is calibrated using data from the first bi-weekly collocation period, as indicated by the hollow circle. Note that the y-axis scale for Pod 83245 at SK6 (third panel from top) differs from other panels.

3.2.2 Long-term evaluation of sensor calibration based on serial collocations

Methods

We use sensors (n=10) that have been collocated two or more times at a reference site to analyze the robustness of an initial NO_2 collocation calibration when applied to subsequent collocation periods. The median time gap between the initial and repeat collocations is 18 weeks, with a minimum gap of 1 week and a maximum gap of 47 weeks.

To study the uncertainty of calibrated data during the "repeat" collocations, we "anchor" a sensor to its first calibration and scale the data of subsequent collocation periods using this original calibration. We examine the bias and error statistics to see how the uncertainty and error during the subsequent periods vary compared to the "anchor" period where the calibration was derived from. **Figure 6** shows an example, where the top timeseries is the calibrated timeseries during the collocation period during which the calibration was derived and the second and third panels are subsequent collocations with pod data scaled using the calibration from the first. We exclude n=3 subsequent collocations with R^2 values <0.5 between the candidate and reference values, which would indicate sensor malfunction.

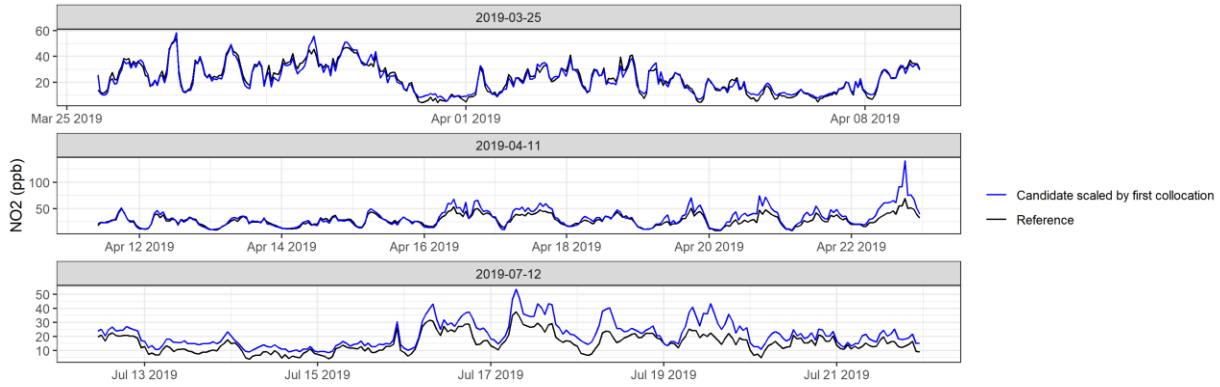


Figure 6 – Pod 2027150 serial collocation timeseries for NO_2 (ppb). The blue line is the AQMesh timeseries and the black line is the reference site timeseries. In each of three periods, the blue line is scaled using the linear regression calibration result from the first (top) collocation. Note that x- and y-axis scales are different for each panel.

Results

Table 2 and Figure 7 show that when the calibration from the first collocation is used to scale AQMesh measurements for subsequent collocations, the error statistics increase substantially compared to the initial collocation calibration period. This finding agrees with the results in **Table 1** where "testing periods" during long-term collocations had higher median nRMSE than calibration periods. We find that the median normalized RMSE among the 26 subsequent "testing" collocations is 31.3%, which is more than double the median RMSE during the initial "calibration" period for the 10 unique sensors of 13.5%. Also similar to the bias trends in long-term collocations, **Table 2** shows that there is a +3.7 ppb median bias in the 23 subsequent collocations, showing a systematic positive bias in this group of sensors. This case study suggests that changes in pod performance may substantially increase sensor bias and error compared to the time period when the calibration was derived.

Table 2: Bias and normalized RMSE of AQMesh NO_2 measurements during serial collocations. All "testing" collocations are subsequent collocations that are calibrated by the first available collocation for the specific sensor.

		Subsequent "testing" collocations		Initial "calibration" period	
Number of unique sensors and initial sensor calibration periods	Number of subsequent collocations for n=10 unique sensors	nRMSE median (min, max)	Mean bias (ppb) median (min, max)	nRMSE median (min, max)	Mean bias[4] (ppb) median (min, max)
10	23	31.3% (11.2%, 78.9%)	3.7 (-2.9, 10.2)	13.5% (10.0%, 19.0%)	0.0 (-0.3, 2.0)

[4] We note that while a linear regression calibration by design eliminates all bias during the calibration period, our results include small biases during the initial collocation periods because we calibrate using a subset of the initial period with outliers removed.

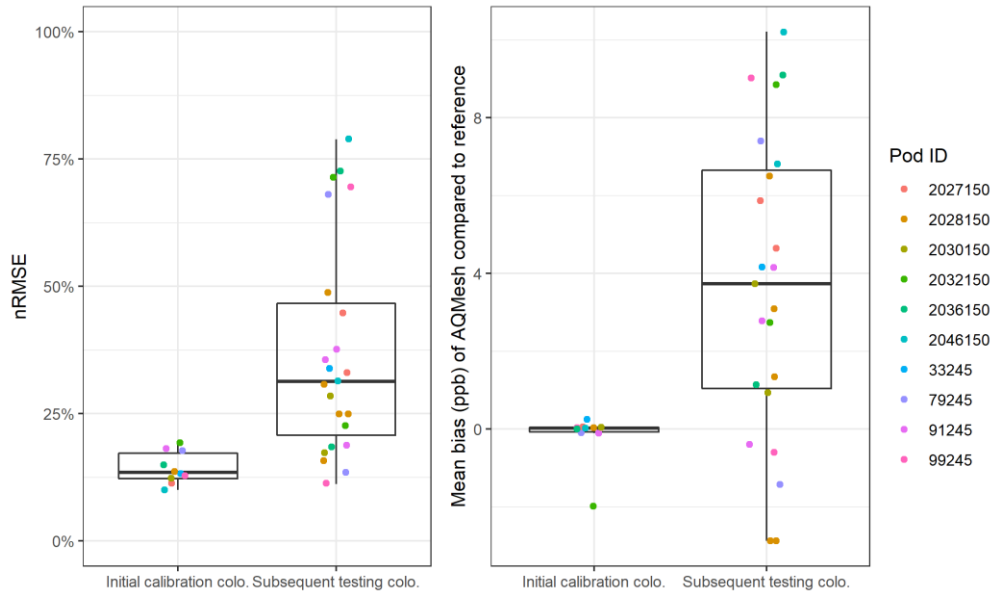


Figure 7 – Normalized root mean square error (nRMSE) and bias of n=23 repeat NO_2 collocations, when calibrated using result from first collocation for each of n=10 unique sensors.

4. Cloud-based network calibration evaluation

The cloud-based network calibration method is a novel approach developed and applied by the University of Cambridge project team that remotely derives calibrations for the entire sensor network without the need for extensive collocation campaigns. The methodology is detailed in Appendix 2C. Cloud-based calibrations for Breathe London sensors were derived from several months of data during 2019: May – Dec 2019 for NO_2 and Apr – June 2019 for $PM_{2.5}$.

The results in **Section 3** showed that our NO_2 instruments can develop significant biases on a timescale of a few months or less after being calibrated (see **Figure 5**). Therefore, it is important to emphasize that our estimates below for the uncertainty of network-calibrated sensors reflect calibration method uncertainty *and* long-term uncertainty in sensor performance.

Methods

We evaluate the performance of the network-based calibration method based on the distribution and summary statistics of error results from the entire batch of short-term (< 3 weeks) reference site collocations. We include only collocations that meet the network method QAQC criteria (Appendix 2A) of covariance > 0.5. We exclude n=3 collocations with R^2 values <0.5 between the candidate and reference values, which would indicate sensor malfunction.

Results

Figure 8 and 9 show the distributions of bias and error collocation statistics for NO_2 and $PM_{2.5}$ measurements respectively when calibrated with the network method.

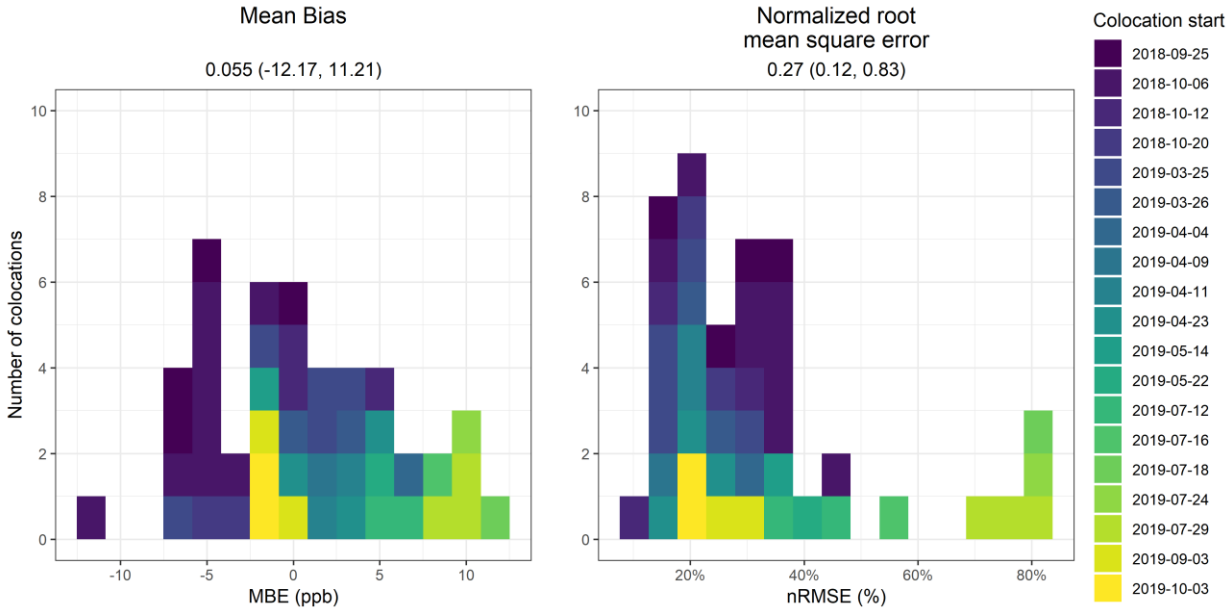


Figure 8 – NO_2 network-method calibrated error and bias results from during reference site collocation periods; mean bias *(left) and* normalized root mean square error *(right). Summary statistics shown are median (min, max). n = 46 collocations, 28 unique sensors. Collocations shown here pass QAQC criteria (**Appendix 2A**) of network calibration covariance > 0.5. Any collocations with R^2 < 0.5 were eliminated as suspect sensor performance.*

For NO_2, the median normalized RMSE of n=46 collocations was 27%, and the median bias across all collocations was +0.1 ppb. During individual collocations, network calibrated sensors still exhibited a large range of biases from -12 to +11 ppb. However, part of this spread in individual collocation bias is likely driven by the time-varying bias effects of ozone cross-interference on our NO_2 sensors. The histogram of NO_2 collocation bias results (**Figure 8**, left) shows that in Fall of 2018 (dark blue), the network-calibrated AQMesh measurements are biased systematically low, in Spring 2019 (blue-green) they are centered near 0 but still display a substantial range of individual collocation biases, and in Summer 2019 (green-yellow), the few collocations are biased quite high. This oscillation in the bias and error of network-calibrated sensors mirrors the seasonal patterns of bias we have shown in prescaled and collocation-calibrated data in prior sections.

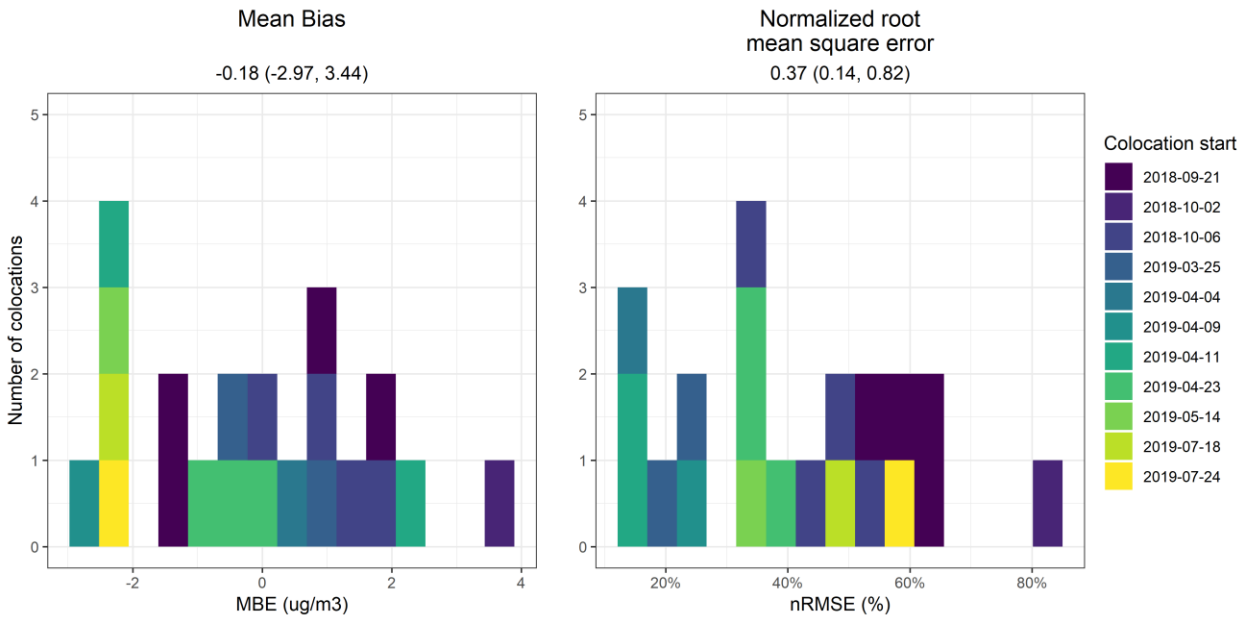


Figure 9 – $PM_{2.5}$ network-method calibrated error and bias results during reference site collocation periods; mean bias *(left) and* normalized root mean square error *(right). Summary statistics shown are median (min, max). n = 21 collocations, 18 unique sensors. Collocations shown here pass QAQC criteria (**Appendix 2A**) of network calibration covariance > 0.5. Any collocations with R^2 < 0.5 were eliminated as suspect sensor performance.*

For $PM_{2.5,}$ the median bias from n=21 collocations was -0.2, with individual collocation biases ranging from -3.0 to 3.4 µg/m3. The median normalized RMSE was 37%, with nRMSE in individual collocations ranging from 12% to 83%.

5. Comparison of calibration approaches

A direct comparison of the effectiveness of the calibration methods applied during the Breathe London project is challenging due to limitations of our data. Our analyses above have shown that long-term fluctuation in NO_2 sensor bias introduces significant uncertainty into measurements, regardless of the calibration method. This temporal component of sensor bias and error complicates comparisons between the cloud-based network calibration method and physical collocation calibration method, especially because these methods derive calibrations from different time windows. Here, we present our best effort to analyze the limited available data to develop a reasonably indicative comparison between these two calibration approaches.

Methods

To compare the methods, we analyze the results of "serial" NO_2 collocations (**Section 3**) compared with network-calibrated NO_2 data and pre-scaled NO_2 data. We identify a subset of n=18 NO_2 collocations where a valid calibration is available from a previous physical collocation calibration for that sensor. This class of data, the "repeat physical" collocation result, reflects the uncertainty of data scaled by a linear regression result from a past physical collocation. We compare the error results of this group to the same collocations calibrated with the network-based approach as well as to the uncalibrated, pre-scaled data.

Lastly, we include a group that is network-calibrated and ozone-corrected. This is similar to the cloud-based network calibration group, except a first-order correction has been applied to mitigate effects of ozone cross-interference. We include this result to show the potential benefits of such a correction on error results but note that this correction has not been applied to any of the other groups.

Results

The network calibration and repeat physical calibration both improve median normalized RMSE to 27% across 18 collocations, compared to 34% for pre-scaled data (**Figure 10 and Table 3**). This subset of collocations has a negative bias in pre-scaled data, consistent with the larger set of pre-scaled data (**Figure 2**). The non-ozone corrected network-calibrated sensors have a positive median bias of 3.1 ppb. Similarly, sensors calibrated using a previous physical collocation result also have a median bias of 2.9 ppb. With implementation of the ozone correction, which we only have available for network-scaled data, the median bias of the group of collocations shifts to -0.7, nearly centering the distribution. Sensors calibrated using both methods still display a wide range of individual pod biases. This is expected based on long-term variations in sensor bias shown in **Figure 5** for the physical collocation method and **Figure 8** for the network calibration method.

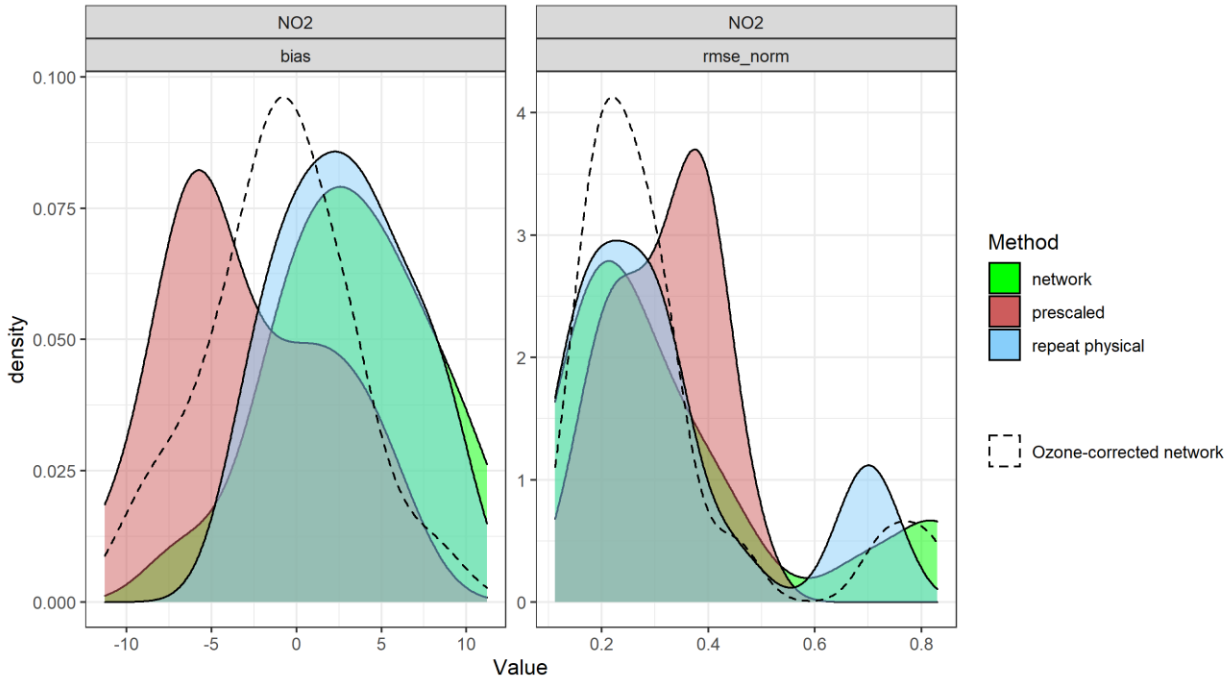


Figure 10 – Collocation error distribution density for two calibration methods: network (cloud-based network calibration method), physical (calibrated using physical collocation linear regression result from previous sensor collocation period at reference), and pre-scaled data (uncalibrated sensor measurements). N=18 collocations. The number of collocations is reduced compared to other figures because collocations must meet network method and physical collocation criteria (R^2>0.7, nRMSE < 0.5, and covariance >0.5), as well as have a previous valid physical collocation result to apply for "repeat physical" scaling. Units: ppb. Statistics evaluated from left to right are mean bias (ppb) and normalized root mean square error.

Table 3 - Collocation error summary statistics by calibration method for n=18 collocations. The number of collocations is reduced compared to other figures because collocations must meet network method and physical collocation criteria (R^2>0.7, nRMSE < 0.5, and covariance >0.5), as well as have a previous valid physical collocation result to apply for "repeat physical" scaling. Values shown are: median (min, max).

Calibration method	Normalized RMSE	Bias (ppb)
Prescaled	34% (14%, 48%)	-4.7 (-11.3, 5.2)
Cloud-based network calibration	27% (13%, 83%)	3.1 (-6.7, 11.21)
Cloud-based network calibration (with ozone correction)	26% (15%, 81%)	-0.7 (-9.2, 7.9)
Physical calibration from previous collocation	27% (11%, 72%)	2.9 (-2.9, 9.1)

6. Transfer standard ("gold pod") calibration uncertainty

In the context of the Breathe London project, which expanded the spatial insight of the current air quality network by monitoring in about a hundred new locations, it was not logistically possible for most pods to be collocated at a reference site, especially if they needed to be calibrated after initial deployment (reasons could include a sensor replacement or rebasing).

Therefore, an alternative calibration approach was undertaken, called the "gold pod" approach. Described in more detail in Appendix 2A, the gold pod approach uses a transfer standard method, first calibrating a gold sensor by means of a reference site collocation before moving that calibrated sensor to calibrate candidate sensors at various BL pod sites.

Compared to a reference site collocation calibration, there are two primary additional uncertainties related to the transfer standard approach:

1. **Transfer uncertainty –** how much additional uncertainty is introduced simply from the "transfer" of calibration from reference site -> gold pod -> candidate pod (rather than reference site -> candidate pod as analyzed above)
2. **Temporal and spatial robustness** – How much uncertainty is introduced by the passing of time and the change in environmental conditions between the gold pod "gilding" at the reference site and the calibration of the candidate pod in the field?

6.1 Transfer uncertainty

Methods

During the Breathe London project, there are 29 periods where two or more AQMesh sensors are collocated at the same reference monitor. We analyze the "transfer uncertainty" of the gold pod calibration method for NO_2 using these periods with multiple simultaneously collocated pods (**Figure 11**). We randomly designate one pod (n=1) as the "gold pod," and perform a linear regression calibration

against the reference measurements. This pod, in effect, becomes the calibrated gold pod. Normally the gold pod would be redeployed in collocations with pods in the field to calibrate candidates. Here, we treat the remaining collocated pods in the group as "candidate pods" (n=1-4 candidate pods, depending on the group). Using the exact same time period, we calibrate the candidate pods using the calibrated gold pod (this was calibrated using the reference monitor directly). We compare the bias and error resulting from the "transfer" of calibration through the gold pod compared to direct reference monitor calibration (Figures 12 and 13). Because we are using identical time periods and locations, the resulting changes in error and bias results should reflect the direct effect of the "transfer uncertainty."

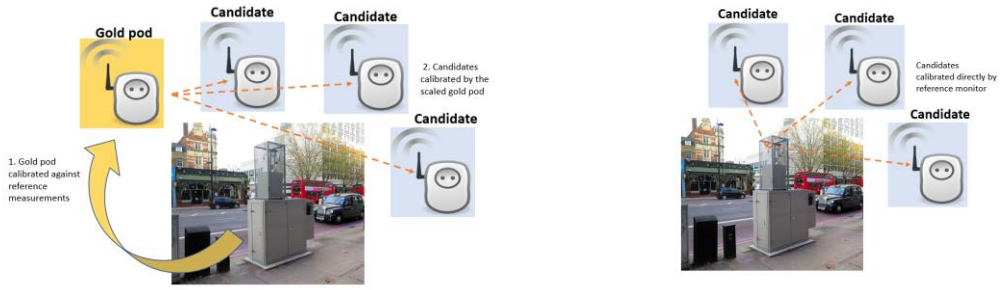


Figure 11 – Transfer uncertainty experiment schematic. Left image shows transfer calibration, right shows reference calibration. Candidate results from each method are analyzed against reference measurements to assess the additional uncertainty related to transfer of calibration

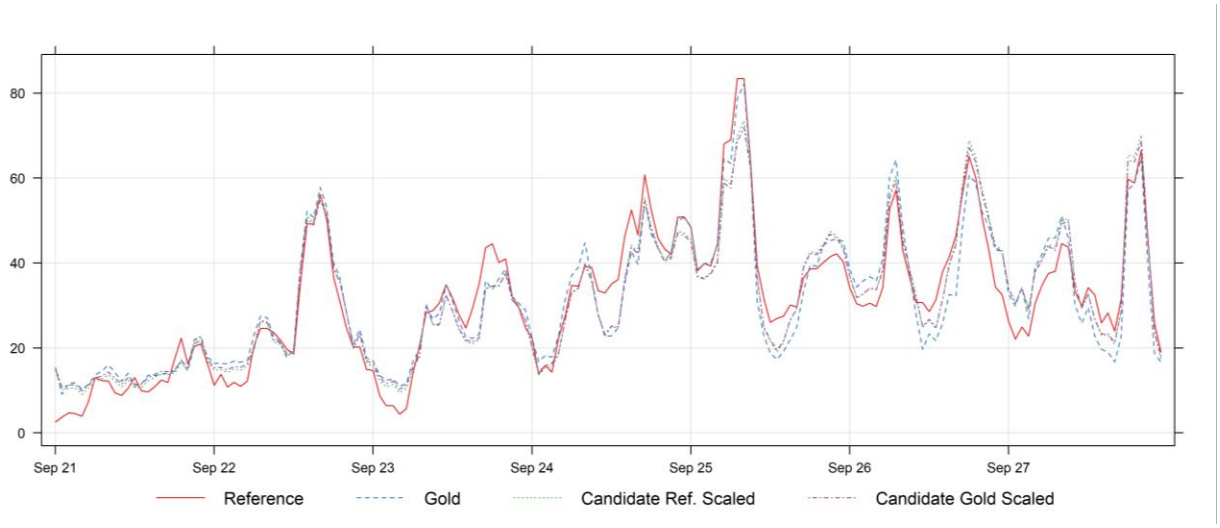


Figure 12 – Example NO_2 (ppb) timeseries of reference monitor ("Reference"), reference-calibrated gold pod ("Gold"), reference-calibrated candidate pod ("Candidate Ref. Scaled"), and gold pod-calibrated candidate pod ("Candidate Gold Scaled"). Comparison of the two latter results allows direct comparison of reference site calibration vs. transfer standard calibration.

Results

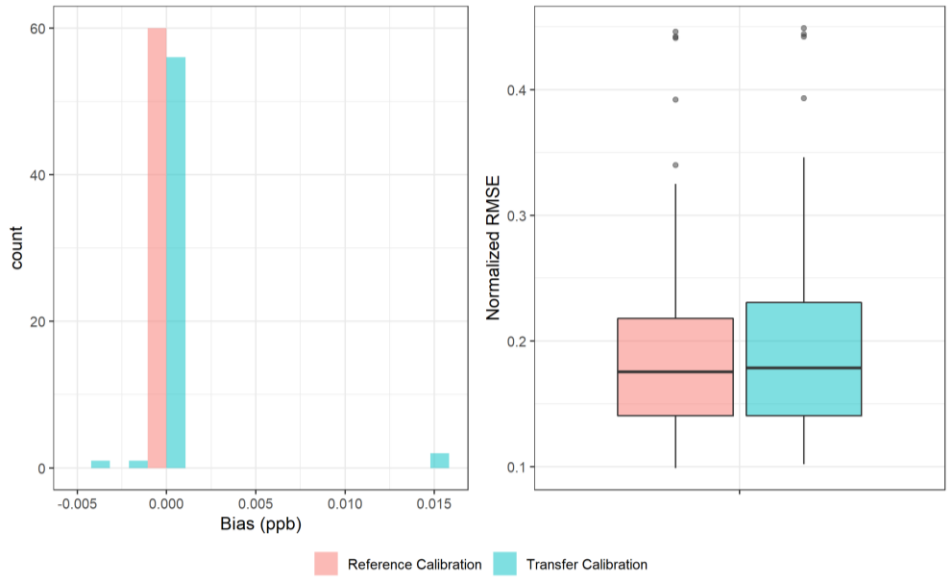


Figure 13 – NO_2 bias (left) and normalized root mean square error (right) results when sensors are calibrated with a gold pod that was calibrated by the reference monitor (blue), compared with direct calibration using the reference monitor, for n=60 trials with n=29 unique sensors during n=29 group reference site collocations.

Figure 13 shows the distribution of error and bias results from the experiment. The results suggest that the uncertainty introduced by the "transfer" process of gold-pod calibration is negligible when the time period, site, and conditions are identical. The median NO_2 RMSE result of candidate pods is 3.3 ppb using the two-step transfer, compared to 3.2 ppb using direct calibration to the reference measurements, only a 3% increase. The median bias remains 0 using the gold pod method, with only a few directional biases introduced in individual sensors that are negligible in magnitude (Figure 13, left panel). The median R^2 value between gold pods and candidate sensors is 0.95, compared to 0.86 between reference measurements and candidate sensors.

Performing the same experiment for $PM_{2.5}$ yields a similar result. The median $PM_{2.5}$ RMSE result of candidate pods is 2.6 µg/m3 using the two-step transfer, compared to 2.5 µg/m3 using direct calibration to the reference measurements, only a 4% increase. The median bias remains 0 using the gold pod method. The median R^2 value between gold pods and candidate sensors is 0.97, compared to 0.84 between reference measurements and candidate sensors, indicating highly reproducible performance between sensors despite poorer agreement to reference measurements.

The results suggest that **within our batch of AQMesh sensors, the responses and error mechanisms relative to the reference instrument are highly reproducible between sensors.** If these results apply broadly, they point to the efficacy of the gold pod method in faithfully reproducing the calibration of the reference site with minimal additional uncertainty.

6.2 Temporal and spatial robustness of gold pod calibrations

The experiment above showed that the process of a two-step calibration from reference -> gold pod -> candidate pod does not introduce significant uncertainty for NO_2 and $PM_{2.5}$ sensors when both calibration steps (gold pod "gilding" and candidate pod calibration against gold pod) are performed on the exact same dataset (same time, conditions, location). However, gold pod collocation calibrations, by design, take place in the field at the candidate pod site at a different time and potentially differing conditions than when the gilding occurred. Unfortunately, there is no reference instrument present to validate against. In this section, we attempt an estimate of the uncertainty introduced in the field by extension of other analyses of reference collocation results presented earlier in this report.

- When AQMesh "gold pod" sensors were gilded using a first reference collocation, and then redeployed one or more times to a reference site, the average nRMSE increased from 13.5% during the initial calibration period to 31.3% during subsequent reference site redeployments, and the AQMesh sensors developed a systematic high bias compared to reference measurements (Table 2).
    - The additional error and bias that developed in these gold pods over time would be propagated to any candidate pod that they were scaled against
- Our long-term collocation analysis in Table 1 and Figure 4 similarly demonstrates that calibrated sensors can develop biases on the order of 8ppb in short (<3-month) timescales and that error of calibrated pods can more than double when the initial calibration is applied to the long-term timeseries.

These observations indicate a higher degree of uncertainty regarding the effectiveness of transfer standard calibration given clear issues with long-term sensor performance. The results suggest that the "gold pod" transfer standard calibration collocations should take place as close to gold pod "gilding" (calibration at reference site) as possible, and gold pods should be frequently re-gilded.

These insights are specific to the technology used in this project and could change with data corrections that improve long-term consistency of sensor measurements and calibrations. We focus here on NO_2, but the supplemental tables and figures in Section 8 present equivalent results for NO and PM_{10} and similarly provide evidence of expanded uncertainty during repeat and long-term application of reference collocation calibration results.

7. Effect of data ratification on calibration results

A logistical constraint of the Breathe London collocation calibration procedures was the ratification timeline of reference monitor data. There is no exact ratification schedule and timing depends on which reference network a monitor belongs to, but in general data ratification for the previous year is performed annually in the first few months of the following year (i.e. 2018 data would become ratified in the first few months of 2019). This meant that for the duration of the project, reference site collocations relied at least partly on un-ratified reference data, creating some additional uncertainty regarding the

robustness of calibration results. Our experience and findings here suggest that projects relying on reference data as a source of calibrations should consider the reference data ratification practices and timeline in project planning and treat calibrations from un-ratified reference data as provisional and subject to change.

In 2020, following ratification of 2019 reference data, we analyzed the ratification-related changes in reference $PM_{2.5}$ and NO_2 data and the corresponding effects on calibration parameters obtained from collocations (Figures 5-7). In summary, we found that:

- Ratification affected some reference NO_2 measurements at CD1 much more significantly than at other sites (**Figure 14**).
- $PM_{2.5}$ ratification changes were negligible, except for redaction of some negative values (**Figure 14**).
- NO_2 ratification generally improved collocation error (nRMSE) and agreement (R^2) slightly.
- Significant NO_2 offset changes during ratification for select 2019 time periods changed collocation results drastically **(Figures 14 and 15)**.
- NO_2 collocation calibration gain changes for most pods were relatively small (<5%) (**Figure 15**).

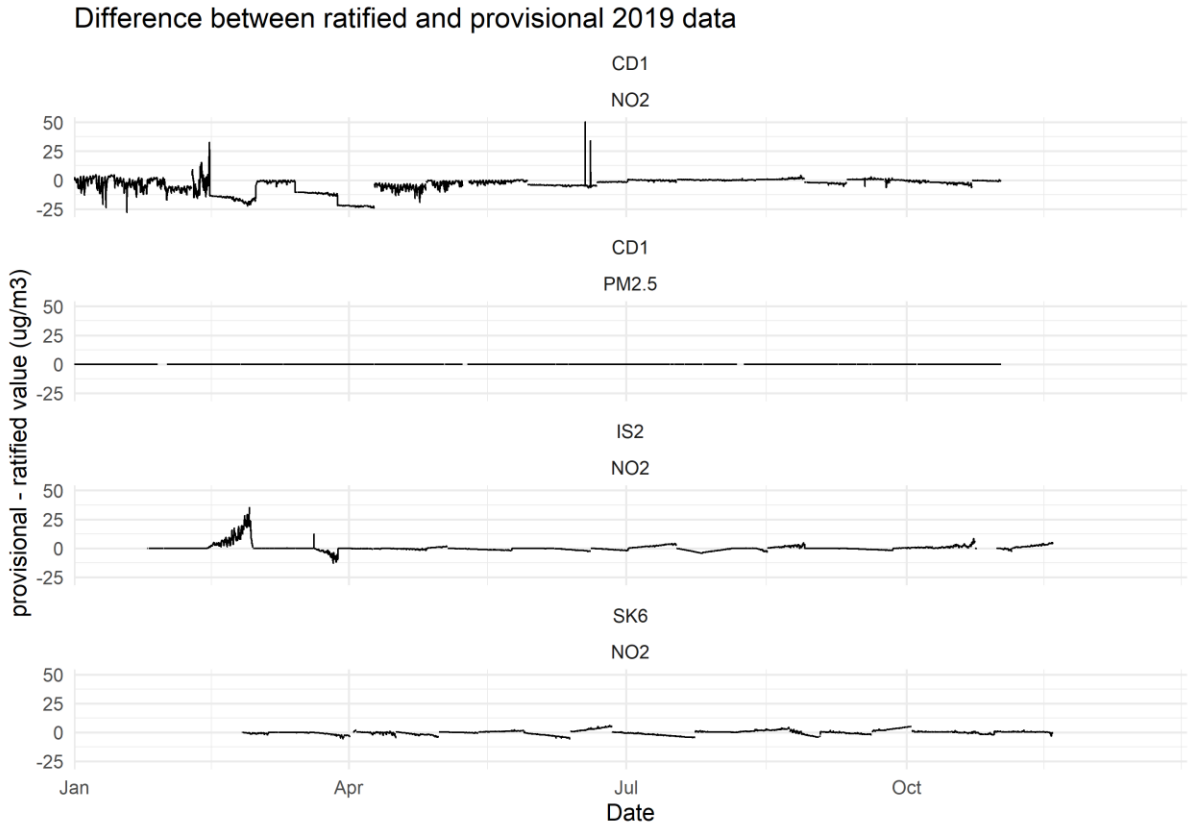


Figure 14 – Difference timeseries of provisional – ratified reference data at three London reference sites where Breathe London collocations were carried out: CD1 – Swiss Cottage, IS2 – Holloway Road, and SK6 – Elephant and Castle.

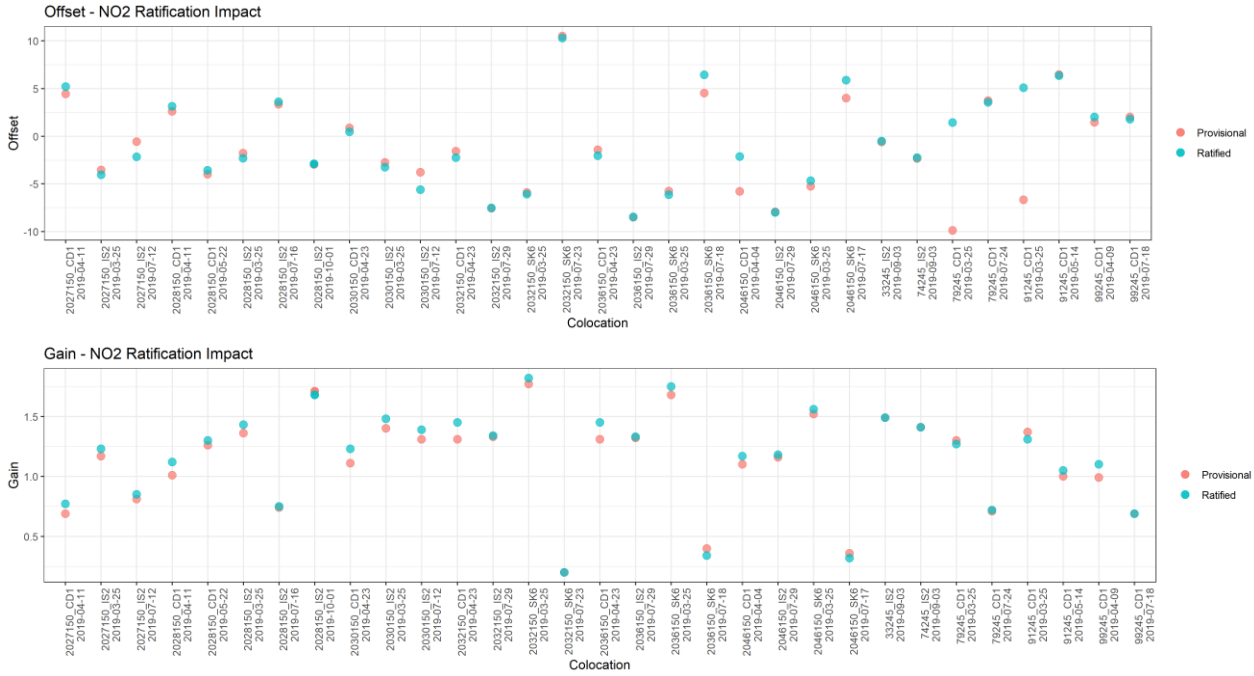
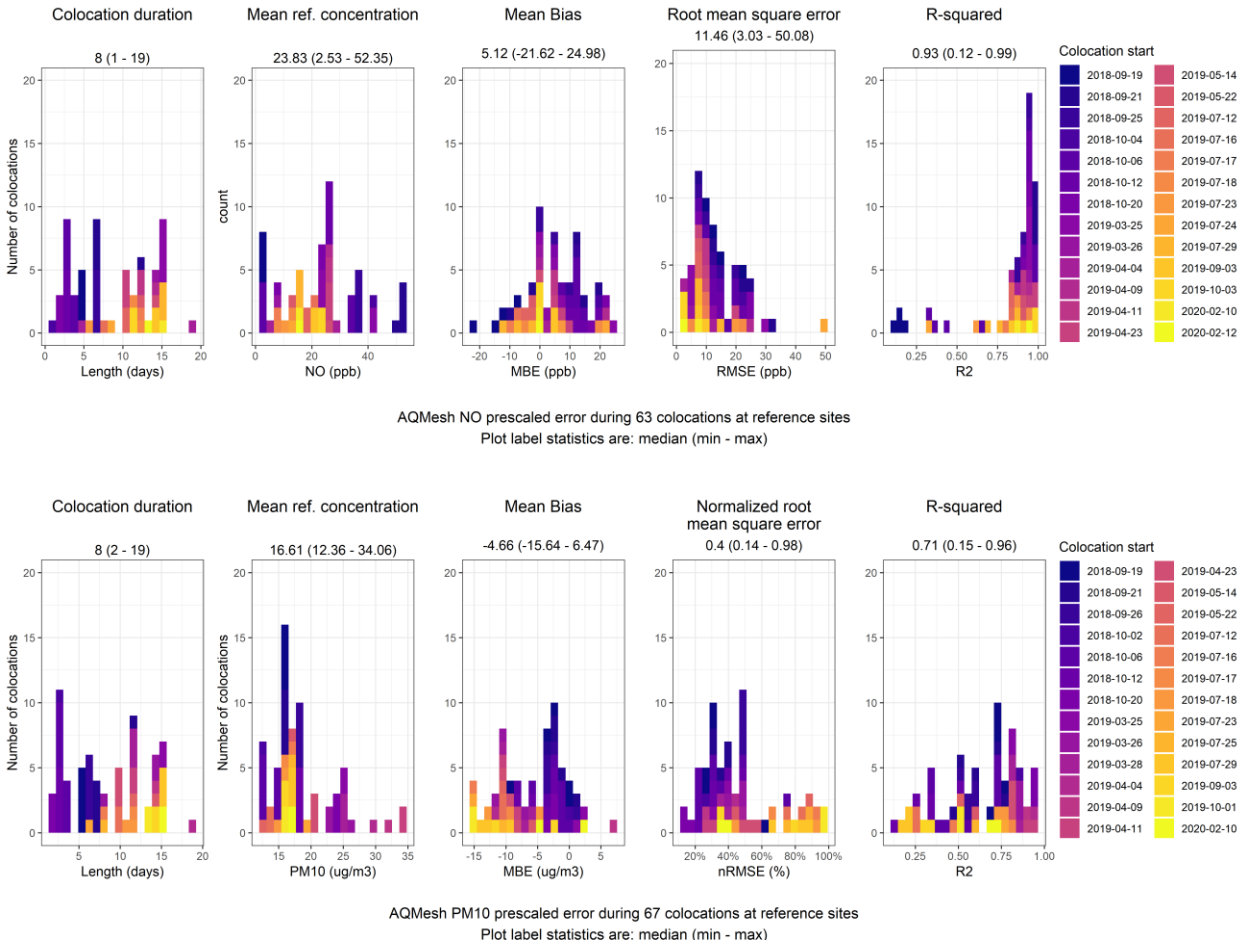
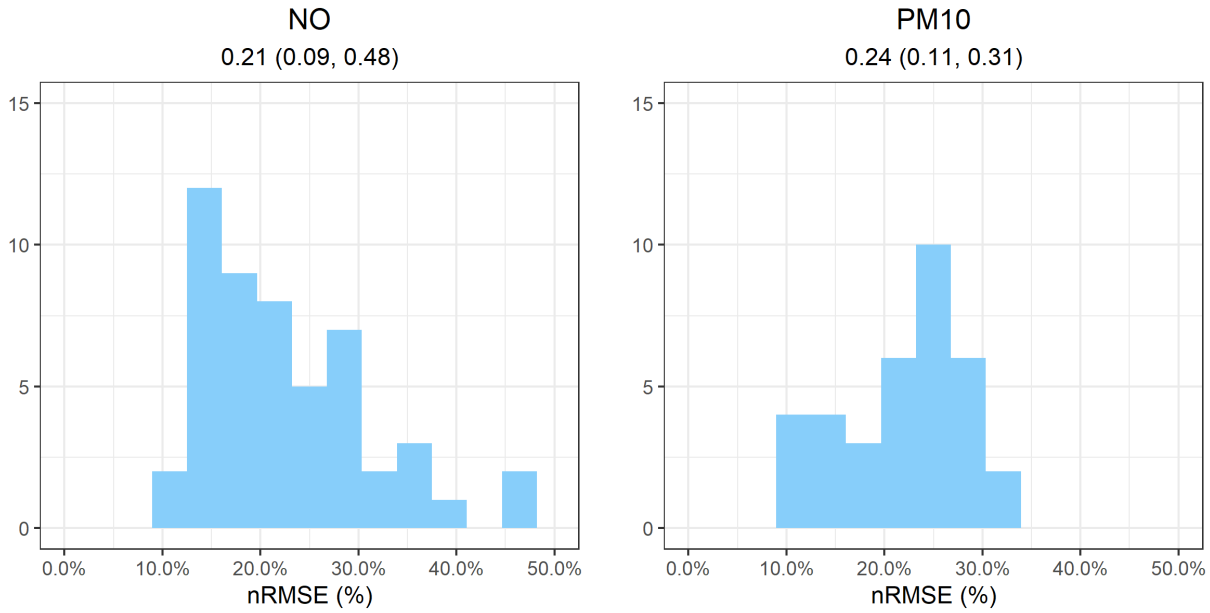


Figure 15 – NO_2 calibration offset and gain results, comparing provisional reference data results (red dots) with ratified (blue).

8. Supplemental figures



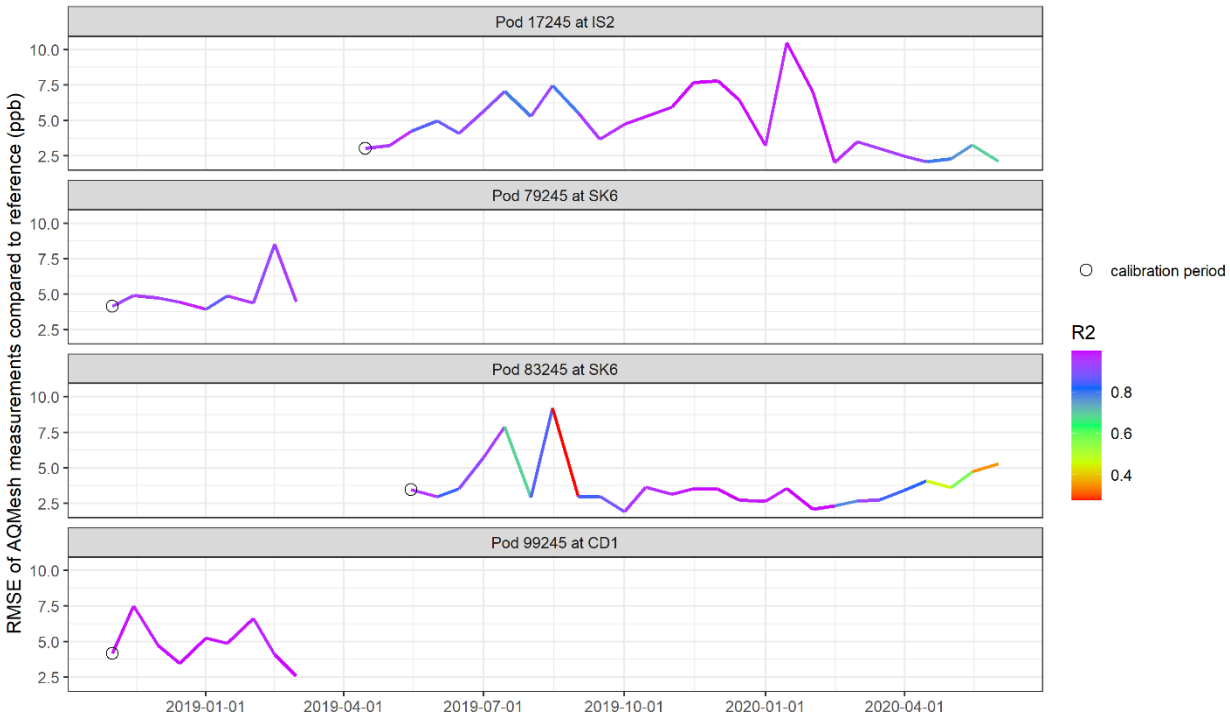
Supplemental figure 1 – Distribution of uncalibrated collocation statistics, NO (top) and PM_{10} (bottom), for short-term (<21 days) ratified reference site collocations during Breathe London project. The color scale indicates the start date of each collocation period. Note that NO RMSE (top, panel 4) is reported in absolute units (ppb), rather than normalized, due to low (<5 ppb) average NO concentrations during certain collocations (see panel 2 for mean reference concentrations). Note that we also excluded n=3 extreme outliers from NO collocations that had $R^2 < 0.1$.



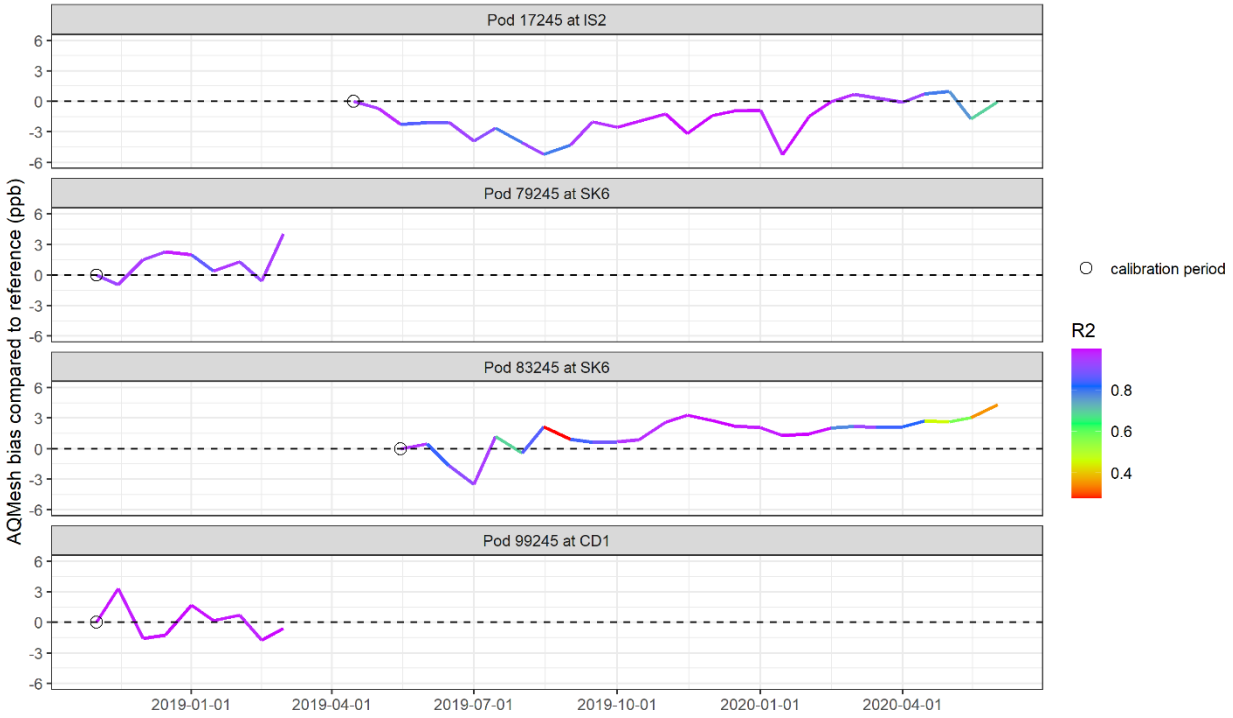
*Supplemental figure 2 – Linear regression calibrated error results during reference site collocation calibration periods; NO (left) and PM_{10} (right). Statistics shown are median (min, max). n = 51 collocations, 34 unique sensors for NO and 35 collocations, 25 unique sensors for PM_{10}. Collocations shown here pass QAQC criteria (**Appendix 2A**) of $R^2 > 0.7$ and nRMSE < 0.5 for collocation calibrations. The number of PM_{10} collocations is significantly lower because many collocations do not satisfy R^2 criteria (see Supplemental figure 1, right-most panel, for full distribution of collocation R^2 results).*

Supplementary table 1: Long-term bias and RMSE of AQMesh NO measurements during calibration and long-term testing periods. Note that Pod 83 tends to have consistently lower R^2 values than other pods (see Supplemental figures 3 and 4) indicating suspect sensor performance. Note that we analyze absolute RMSE (ppb) here, instead of normalized RMSE used for other pollutants, because the normalized statistic for NO is affected by periods with low (<5 ppb) NO concentrations.

	Long-term "testing" period			Initial "calibration" period	
	Number of unique bi-weekly "testing" periods	RMSE (ppb) median (min, max)	Mean bias (ppb) median (min, max)	RMSE (ppb)	Mean bias (ppb)
Pod 17 at IS2	25	4.7 (2.1, 10.5)	-1.7 (-5.2, 1.0)	3.0	0.0
Pod 79 at SK6	8	4.6 (3.9, 8.5)	1.4 (-1.0, 4.0)	4.1	0.0
Pod 83 at SK6	25	3.4 (1.9, 9.2)	2.1 (-3.5, 4.3)	3.5	0.0
Pod 99 at CD1	8	4.8 (2.6, 7.5)	-0.2 (-1.7, 3.3)	4.2	0.0



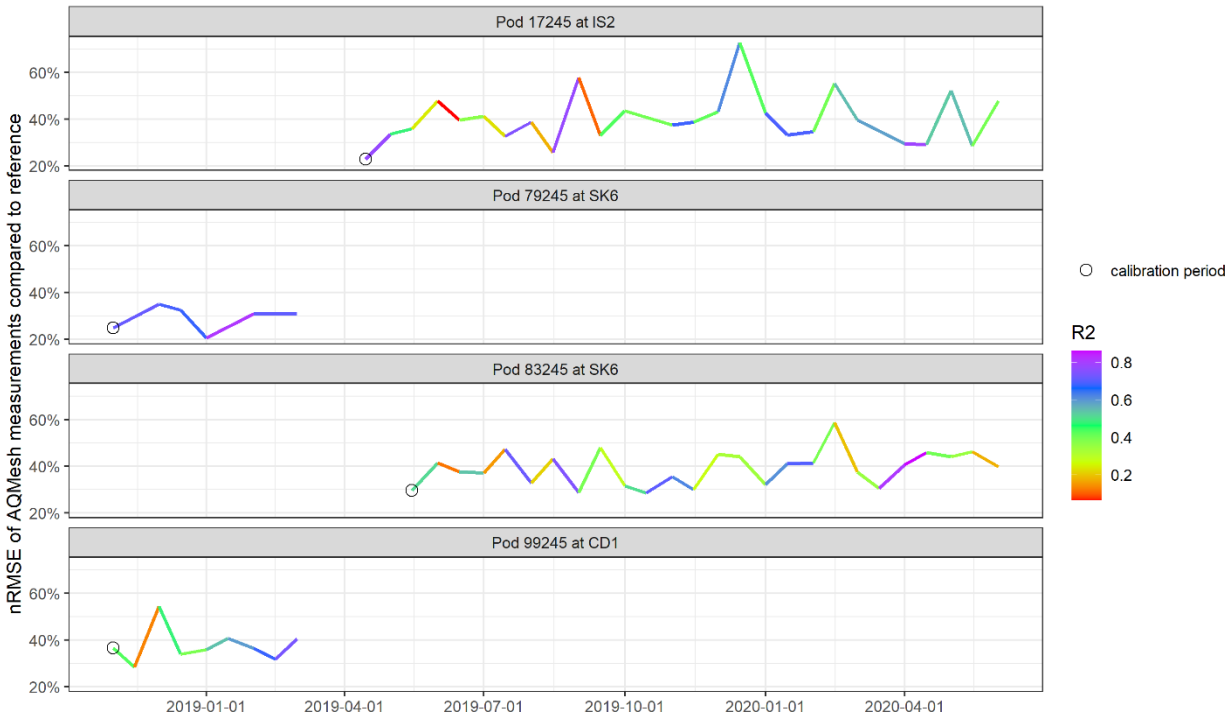
Supplemental figure 3 – Bi-weekly RMSE (ppb) of AQMesh NO measurements compared to reference measurements, for four long-term collocations at three different reference monitors. Reference data is ratified through the beginning of 2020. Bi-weekly R^2 of measurements symbolized by color. Each AQMesh sensor is calibrated using data from the first bi-weekly collocation period, as indicated by the hollow circle.



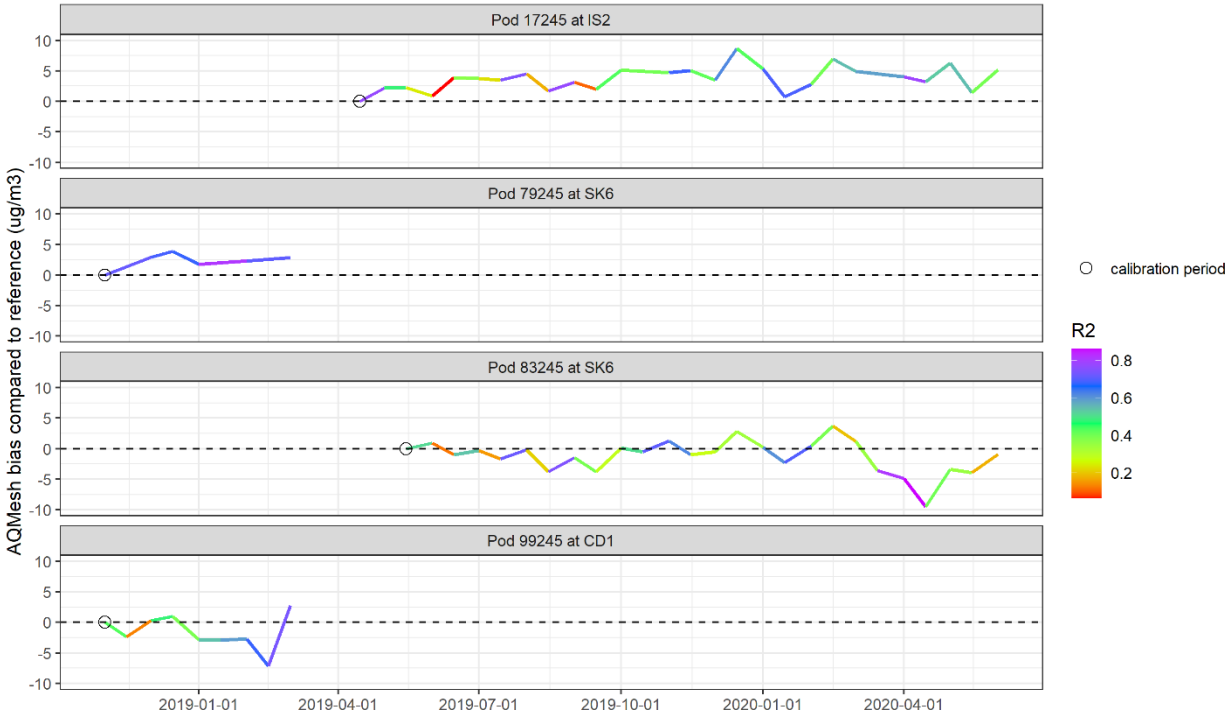
Supplemental figure 4 – Bi-weekly bias of AQMesh NO measurements (ppb) compared to reference measurements, for four long-term collocations at three different reference monitors. Reference data is ratified through the beginning of 2020. Bi-weekly R^2 of measurements symbolized by color. Each AQMesh sensor is calibrated using data from the first bi-weekly collocation period, as indicated by the hollow circle.

Supplementary table 2: Long-term bias and RMSE of AQMesh PM_{10} measurements during calibration and long-term testing periods.

	Long-term "testing" period			Initial "calibration" period	
	Number of unique bi-weekly "testing" periods	nRMSE median (min, max)	Mean bias (µg/m3) median (min, max)	nRMSE	Mean bias (µg/m3)
Pod 17 at IS2	25	38.7% (25.7%, 72.9%)	3.7 (0.8, 8.6)	22.9%	0.0
Pod 79 at SK6	5	31.2% (20.5%, 35.1%)	2.9 (1.8, 3.9)	25.0%	0.0
Pod 83 at SK6	25	40.7% (28.6%, 58.8%)	-0.9 (-9.5, 3.7)	29.6%	0.0
Pod 99 at CD1	8	36.1% (28.3%, 54.4%)	-2.5 (-7.2, 2.7)	36.5%	0.0



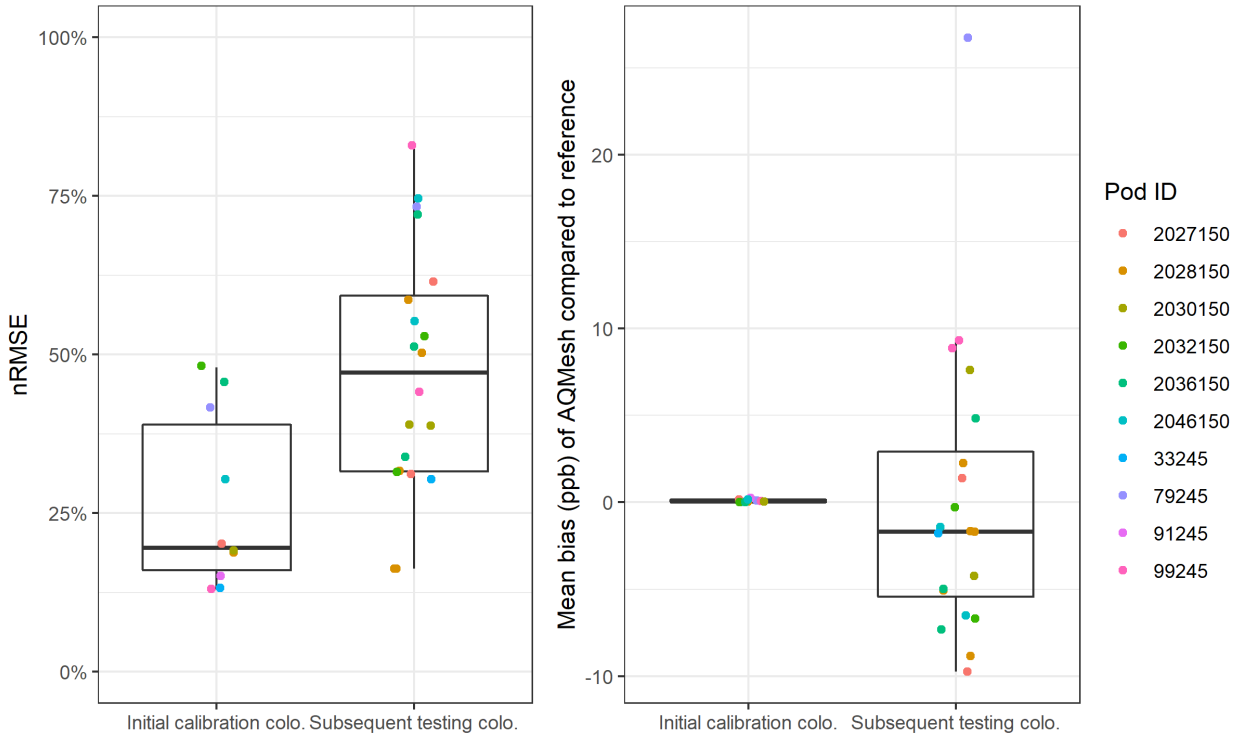
Supplemental figure 5 – Bi-weekly nRMSE (%) of AQMesh PM_{10} measurements compared to reference measurements, for four long-term collocations at three different reference monitors. Reference data is ratified through the beginning of 2020. Bi-weekly R^2 of measurements symbolized by color. Each AQMesh sensor is calibrated using data from the first bi-weekly collocation period, as indicated by the hollow circle.



Supplemental figure 6 – Bi-weekly bias of AQMesh PM_{10} measurements (µg/m3) compared to reference measurements, for four long-term collocations at three different reference monitors. Reference data is ratified through the beginning of 2020. Bi-weekly R^2 of measurements symbolized by color. Each AQMesh sensor is calibrated using data from the first bi-weekly collocation period, as indicated by the hollow circle.

Supplemental table 3: Bias and normalized RMSE of AQMesh NO measurements during serial collocations. All "testing" collocations are subsequent collocations that are calibrated by the first available collocation for the specific sensor.

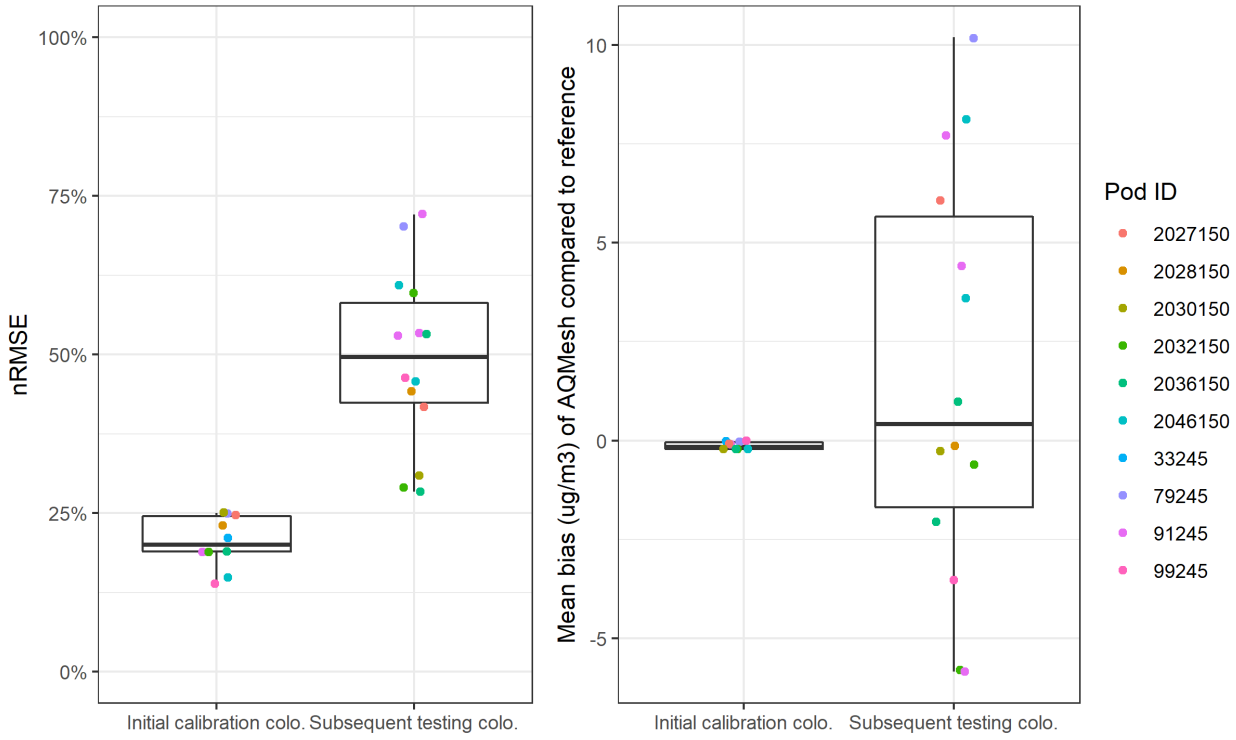
		Subsequent "testing" collocations		Initial "calibration" period	
Number of unique sensors and initial sensor calibration periods	Number of subsequent collocations for n=10 unique sensors	nRMSE median (min, max)	Mean bias (ppb) median (min, max)	nRMSE median (min, max)	Mean bias (ppb) median (min, max)
10	20	47.2% (16.3%, 83.0%)	-1.7 (-9.7, 26.7)	19.5% (13.0%, 48.0%)	-0.1 (-0.3, 0.0)



Supplemental figure 7 – Normalized root mean square error (nRMSE) and bias of n=20 repeat NO collocations, when calibrated using result from first collocation for each of n=10 unique sensors.

Supplemental table 4: Bias and normalized RMSE of AQMesh PM_{10} measurements during serial collocations. All "testing" collocations are subsequent collocations that are calibrated by the first available collocation for the specific sensor.

		Subsequent "testing" collocations		Initial "calibration" period	
Number of unique sensors and initial sensor calibration periods	Number of subsequent collocations for n=10 unique sensors	nRMSE median (min, max)	Mean bias (µg/m3) median (min, max)	nRMSE median (min, max)	Mean bias (µg/m3) median (min, max)
10	14	49.6% (28.3%, 72.1%)	0.4 (-5.8, 10.2)	20.0% (14.0%, 25.0%)	0.2 (0.0, 0.2)



Supplemental figure 8 – Normalized root mean square error (nRMSE) and bias of n=14 repeat PM_{10} collocations, when calibrated using result from first collocation for each of n=10 unique sensors.