KING'S College LONDON

University of London

Breathe London wearable sensor evaluation - Flow

Authors: Dr Benjamin Barratt and Shanon Lim

Background – the importance of portable sensor testing

Personal air quality exposure assessment is a growing area of research whereby small portable sensors are provided to individuals to better estimate the air pollution people breathe. An important aspect of this type of research is the reliability, accuracy and precision of the sensors. Evaluation of portable sensors are typically completed with two tests, the first in a fixed location to test accuracy against an approved reference monitor and the second a mobile test to evaluate their ability to measure air pollution fluctuations in different environments.

Testing is carried out in urban, real world environments so key performance metrics can be determined to evaluate the sensor's fitness of purpose. Sensors are co-located with reference monitors at a dedicated air quality monitoring station for a sustained period of time to provide an assessment of the sensor's performance relative to a reference monitor and in response to changing field conditions. Besides evaluating sensor reliability, this extended evaluation period provides critical temporal data enabling the full characterisation of a sensor's performance in a specific type of dynamic outdoor environment where meteorology and concentrations of target and interfering species are subject to change.

The initial phase of a sensor's testing regime aims to evaluate the sensor's capacity for continuous, un-interrupted data capture, its inter-unit precision and the comparability of its data with reference monitor outputs. Raw data capture rates are used as an indicator of reliable sensor function and the robustness to withstand dynamic urban environments. If the above phase is satisfactorily passed, appropriate correction factors can be derived to calibrate the sensor against the specific reference monitor used during the test. After comparison to a fixed reference monitor, the sensors are also evaluated in a short mobile monitoring campaign to test how they respond in different pollution environments and commuting modes.

This document is one of a series detailing results of portable air quality sensor testing carried out as part of the selection process for the Breathe London Wearables study funded by the Greater London Authority.

Introduction – sensor testing protocol

The Breathe London Wearables study is a public engagement campaign that aims to characterise London school children's exposure to air pollution and present this information in a way that the school community can understand, relate and act upon. In order to achieve the study's objectives, a suitable wearable air pollution sensor had to be identified, tested and selected. The sensor requirements were as follows:

1. Monitor $PM_{2.5}$ pollutant concentrations and GPS position at a time resolution of at least 1 minute. Monitored nitrogen dioxide (NO_2) concentrations were also desirable, but not essential.
2. Small and light enough to be carried by school children aged 5 – 11 years.
3. Battery life of at least 10 hours to cover a full school day.
4. Sufficiently low cost to allow at least 20 units to be deployed within a budget of £20,000.
5. Sufficiently robust and reliable to deliver valid results despite potentially rough treatment by children.
6. Demonstrable accuracy and precision sufficient to allow robust comparison between sensors and illustrate spatial variation in pollutant concentrations.

Six sensors appeared to meet these criteria and were selected for testing; (i) Plume Flow, (ii) Airbeam2, (iii) University of Cambridge PAM and (iv) Dyson wearable sensor. The suppliers of the two remaining sensor units were not able to supply test units in time for the trial, so these were dropped. A predefined testing protocol was followed for each sensor to ensure fair treatment and transferability of outcomes. The purpose of the protocol was to independently verify that the wearable sensor was able to demonstrate performance characteristics to deliver the aims of the project. It also allowed us to identify sensor features and limitations, which would influence the design of the subsequent sensor deployments.

The testing protocol included two phases – a static test and a mobile test. The static test ran from 8 October to 22 October 2018. The initial plan for static testing was for a period of three weeks, however only two weeks could be completed due to time constraints. The sensors were required to be in a low pollution environment (office) for a few days prior to the mobile test so the sensors could calibrate to baselines levels, this requirement also reduced the number of days available for static testing. Three sensor units of each type were placed within a Stevenson's screen within one metre of the inlet of a $PM_{2.5}$ FDMS (Filter Dynamics Measurement System) reference monitor at the Marylebone Road kerbside research monitoring site (www.londonair.org.uk/london/asp/publicdetails.asp?site=MY7). Sensor measurements were extracted from the units and a series of statistical tests performed on the data. The first 24 hours data were excluded to allow a settling in period.

The mobile test was carried out on 29 October 2018. This comprised a one-hour test journey on a prescribed route across London from Marylebone Road to Waterloo, incorporating contrasting environments (parkland and busy congested traffic routes). The first half of the journey was carried out by foot, the second half in a diesel taxi. The sensors were assessed based on the inter-unit comparability and how the sensors responded in different pollution environments compared to expected spatial patterns.

To provide an overall assessment, each sensor was given a rating for aesthetics, bulk, setup, reliability, usability, precision, accuracy, GPS and cost. Double weighting was applied to precision and accuracy categories reflecting their importance. A separate report was produced for each unit type detailing performance against each test and their overall assessment rating.

This report details the results for the evaluation of Flow $PM_{2.5}$ sensor units by Plume Labs. As the device also measured NO_2, a short note on the NO_2 sensor is included at the end of the report.

Results

Capture rates (reliability)

This table describes the percentage of valid one-minute readings logged by the sensors. Data loss may be caused by breakdown of sensor, logging or communication system. The target is 100%.

Table 1: Valid data capture rates (% based on 1-minute readings). Capture rates less than 90% are highlighted in red.

Week Commencing	FLOW001 / %	FLOW002/ %	FLOW003 / %
08-Oct-18	43	34	63
15-Oct-18	99	53	99
Full period	71	43	81

The capture rates for the static test were relatively poor, with all sensors having a capture rate of less than 81% over the full monitoring period. However, FLOW001 and FLOW003 had very high capture rates (99%) in the second week of monitoring. The reason behind the poor capture rates are that the Flow sensors are not designed to be left recording without an individual's interaction. They require a smartphone to upload data to the app and this action was only possible once a week during the static testing period. As such the sensors did not have a high capture rate, it is likely if they were used as intended, i.e. an individual using their smartphone continuously and interacting with the device, the capture rate would be substantially higher.

Inter-unit correlations (precision)

Table 2 indicates the degree of correlation between the three sensors tested, describing the level of inter-unit precision. Precision is important to assess the likelihood that additional untested units perform in the same way as tested units and transferability of derived correction/scaling factors. Inter-comparability between devices is particularly important when comparing exposures between different individuals in studies. Results are presented as Reduced Major Axis correlation (RMA) coefficient (R^2). The target is 1.00.

Table 2: Correlation coefficient (R^2) between units. Coefficients of less than 0.75 are highlighted in red.

R^2 (RMA)	FLOW001 / %	FLOW002/ %	FLOW003 / %
FLOW001		0.68	0.76
FLOW002	0.68		0.60
FLOW003	0.76	0.60	

All PM sensors demonstrated a precision between 60 to 76%. This suggests that two sensors monitoring the same environmental conditions would report substantially different concentrations. Direct comparison of results between sensors is important to be able to compare PM concentrations between individuals, this relatively poor precision would limit the usefulness of results.

Correlation coefficient against reference monitor (accuracy)

This table describes the degree of agreement between each sensor unit and the reference $PM_{2.5}$ monitor. The target is 1.00, which would indicate that 100% of the variation in $PM_{2.5}$ was described by the sensor unit.

Table 3: Correlation coefficient (R^2) in comparison with reference monitors. R^2 values less than 0.75 are highlighted in red.

Week Commencing	FLOW001 / %	FLOW002/ %	FLOW003 / %
08-Oct-18	0.01	0.72	0.45
15-Oct-18	0.54	0.76	0.29
Full period	0.54	0.71	0.37

Accuracy (in terms of correlation against the reference monitor) varied from week to week, but was poor overall. FLOW002 was the best performing sensor but due to the poor capture rate of the device, caution should be taken when assessing the accuracy of this sensor. The accuracy of the other two devices was under 55%.

Scaling factor relative to reference monitor (scale correction)

This table shows the multiplication factor required to scale the sensor to the reference monitor. This is calculated using linear regression (y = mx + c, where m is the scaling factor, x is the reference monitor, c is the offset and y is the portable sensor). The target for m is 1.0.

Table 4: Scaling factor relative to reference monitors (based on hourly readings).
Sensor reporting capture rates less than 50% or with a correlation coefficient less than 0.5 are marked 'n.a.'

Week Commencing	FLOW001	FLOW002	FLOW003
08-Oct-18	n.a.	n.a.	n.a.
15-Oct-18	0.70	0.32	n.a.
Full period	0.69	n.a.	n.a.

Correction factors (offset and scaling), are a normal part of an instrument scaling procedure, but to be effective, they must be stable over time and across a range of ambient conditions. The greater the accuracy (correlation against reference monitor) the more reliable the scaling correction factor will be.

All sensors under-read in comparison with the reference monitor. Most weeks failed to meet the criteria of greater than 50% capture rate and correlation coefficient greater than 0.5, reflecting poor accuracy and reliability of the sensors.

Offset from reference monitor (offset correction)

This table shows the mean offset difference between the sensor and the reference monitor calculated using linear regression (y = mx + c, where c is the offset). The target for c is 0.

Table 5: Offset from reference monitors (based on hourly mean readings).

Week Commencing	FLOW001	FLOW002	FLOW003
08-Oct-18	n.a.	n.a.	n.a.
15-Oct-18	-3.1	-2.3	n.a.
Full period	-3.8	n.a.	n.a.

Offset correction is the second component of the instrument correction procedure. These values indicate that the sensor units have a consistent positive or negative offset from zero. The greater the accuracy (correlation against reference monitor) the more reliable the offset correction factor will be. The sensor units had a baseline under-estimation for $PM_{2.5}$ of around 3 µg m^{-3}, however, due to the poor accuracy of the sensors, appropriate interpretation of the offset values is difficult.

Hourly mean time series

A time series chart comparing each sensor against the reference monitor over the two-week testing period is presented prior to and following application of the full period correction factors.

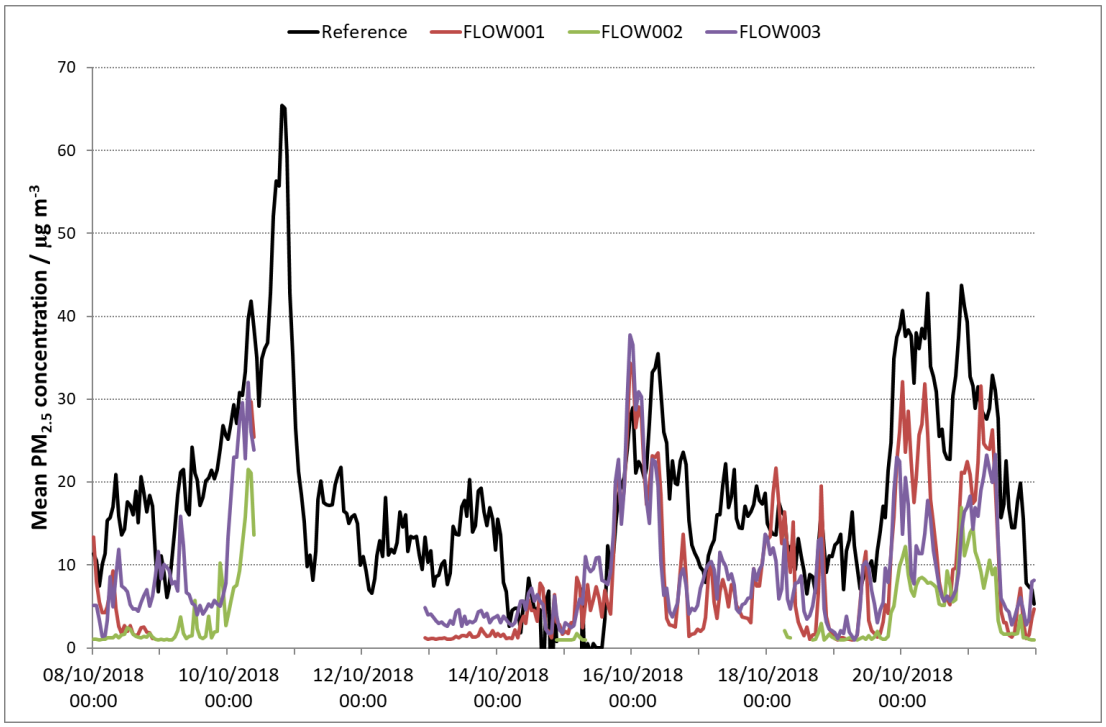

Figure 1: Time series chart of hourly mean sensor and reference $PM_{2.5}$ concentrations over the two-week test period prior to application of scaling factors.

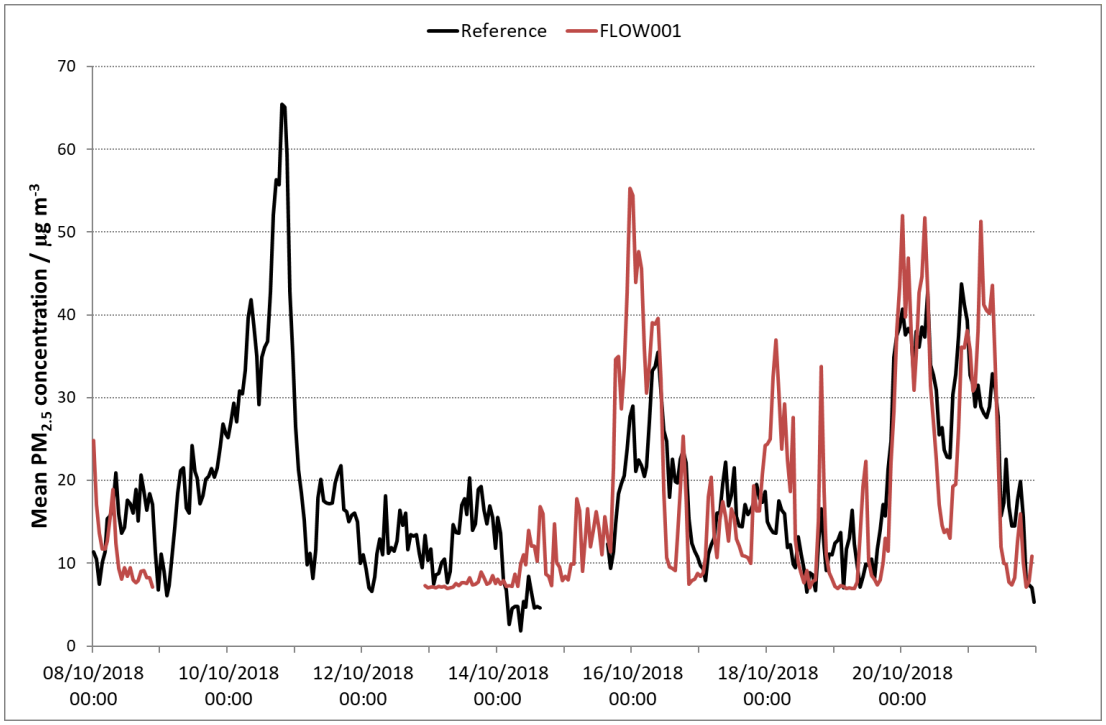

Figure 2: Time series chart of hourly mean sensor and reference $PM_{2.5}$ concentrations over the two-week test period following application of scaling factors. Only FLOW001 was scaled, due to the other sensors not meeting accuracy and reliability requirements.

Only FLOW001 could be corrected for as this was the only sensor which passed the requirements for accuracy and reliability highlighted in Table 4. Values were scaled using the correction factors for the full period (Scaled = (Raw - c) / m). It can be seen from Figure 2 that these factors do not produce a good representation of $PM_{2.5}$, particularly as they tend to over-read at higher concentrations.

Mobile monitoring evaluation

Mobile monitoring was conducted for a period of just over an hour. A map of the route and concentrations is shown below (Figure 3). To provide different modes of travel and environments the first half of the journey was completed by walking through a park and congested street canyon and the second half completed by taxi. In the absence of a mobile reference monitor only a sensibility check and inter-unit comparisons (precision) could be made (Table 6).

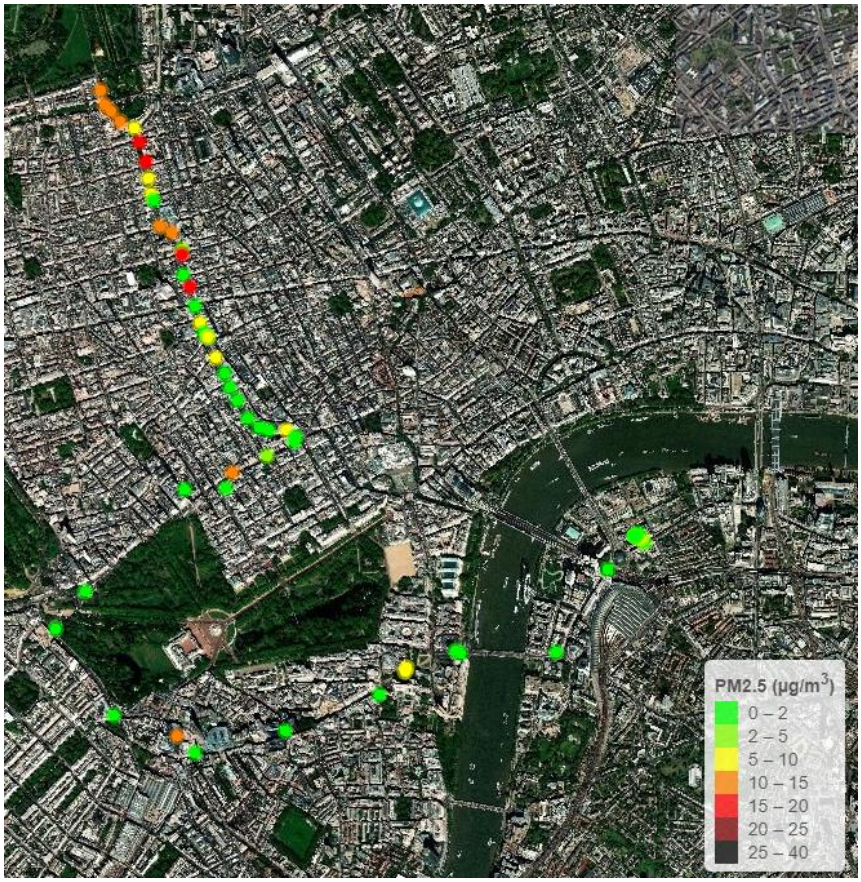

Figure 3: Map of mobile monitoring test for FLOW003, from Marylebone Road (North) to Stamford Street (South), London. Each dot represents one-minute of concentration.

The spatial pattern recorded by the units was not consistent with expectations, with high and low concentrations appearing sporadically during the test. It appeared that FLOW002 was defective and reported results substantially higher compared to the other two sensors.

Table 6: $PM_{2.5}$ correlation coefficient (R^2) between units for mobile monitoring campaign. Coefficients of less than 0.75 are highlighted in red.

R^2 (RMA)	FLOW001 / %	FLOW002/ %	FLOW003 / %
FLOW001		0.03	0.10
FLOW002	0.03		0.01
FLOW003	0.10	0.01	

All PM sensors demonstrated a very poor degree of precision in the mobile monitoring test, with a maximum of 10% of the variation in one unit was explained by any other unit. This poor precision is further observed in the timeseries of the mobile monitoring test presented in Figure 4. No trends in concentrations could be observed when walking in a busy street canyon around 15:10 to 15:30 and while traveling by taxi at 15:30 to 15:50.

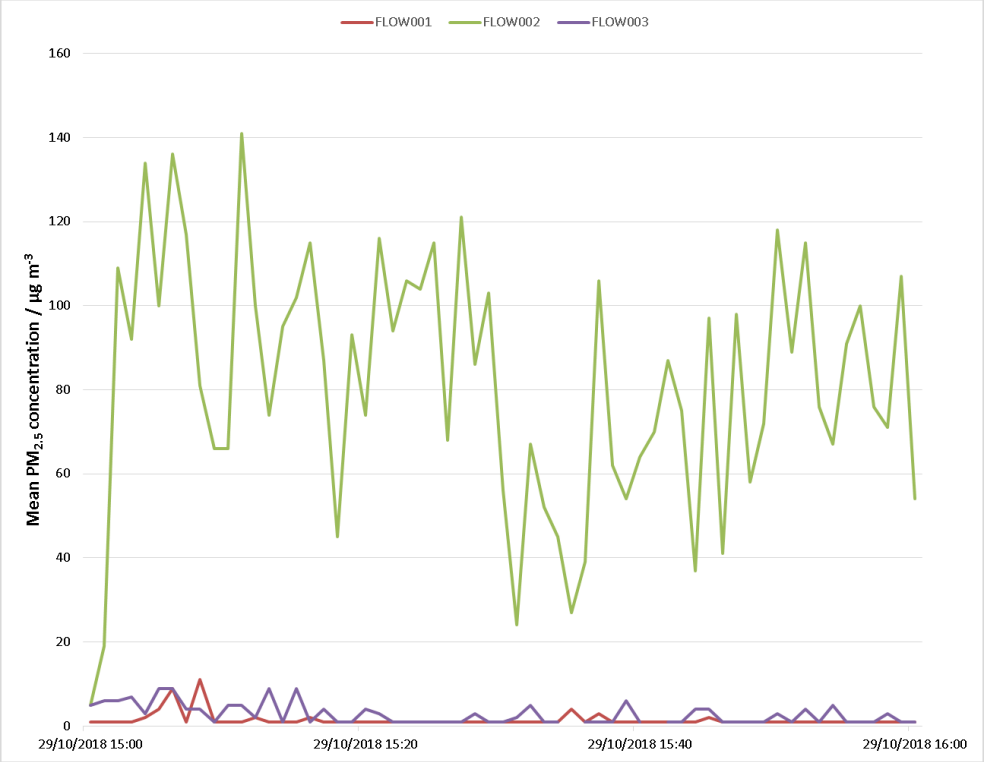

Figure 4: Time series chart of one-minute sensor $PM_{2.5}$ concentrations over the one-hour mobile monitoring campaign.

Overall sensor evaluation for $PM_{2.5}$

Table 7: Scoring results out of 55 (0 – 5, 0 is low, accuracy and precision given double weighting):

Sensor	Aesthetics	Bulk	Setup	Reliability	Usability	Precision	Accuracy	GPS	Cost	Total Score
Flow	**4**	**4**	**1**	**1**	**1**	**2**	**2**	**3**	**4**	**22**

The Flow $PM_{2.5}$ sensor performed poorly in both static and mobile testing. The most important aspects of an air pollution sensor, reliability, precision and accuracy all produced poor results. Usability also had a low rating due to difficulty with accessing the data during the static monitoring campaign. Aesthetics and the small size of the device were its best features.

The device requires a smartphone to use, this does reduce its practicality for use in academic mobile monitoring campaigns. The unit is commercially available and can easily be used by members of the public who are interested in air pollution monitoring. However, the accuracy and precision of the sensor needs to be improved to be able to realise the devices' full potential.

NO_2 sensor evaluation

As the NO_2 sensor evaluation was not the primary aim of this report, only summary results are presented. Fixed testing for the device could not be conducted as the manufacturer's instructions requiring the device to be placed in a low NO_2 pollution environment for at least an hour each day, which could not be done during static testing. Despite this, the mobile monitoring test for the NO_2 sensor was conducted.

Mobile monitoring was conducted at the same time as the PM sensor. Prior to testing, sensors were placed in a low NO_2 pollution environment, as per manufacturer's instructions. A map of the route and NO_2 concentrations is shown below (Figure 5). Like the PM sensor the performance of the devices was relatively poor, and the spatial pattern recorded by the units was not consistent with expectations. The inter-comparability between the devices was better than the $PM_{2.5}$ sensor, but each sensor still reported substantially different concentrations, as observed in Figure 6.

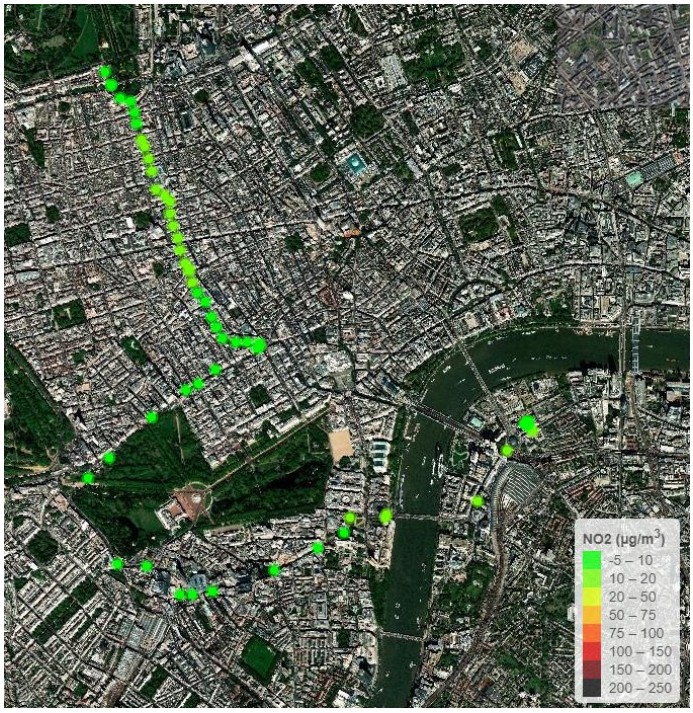

Figure 5: Map of NO_2 mobile monitoring test for FLOW001, from Marylebone Road (North) to Stamford Street (South), London. Each dot represents one-minute of concentration.

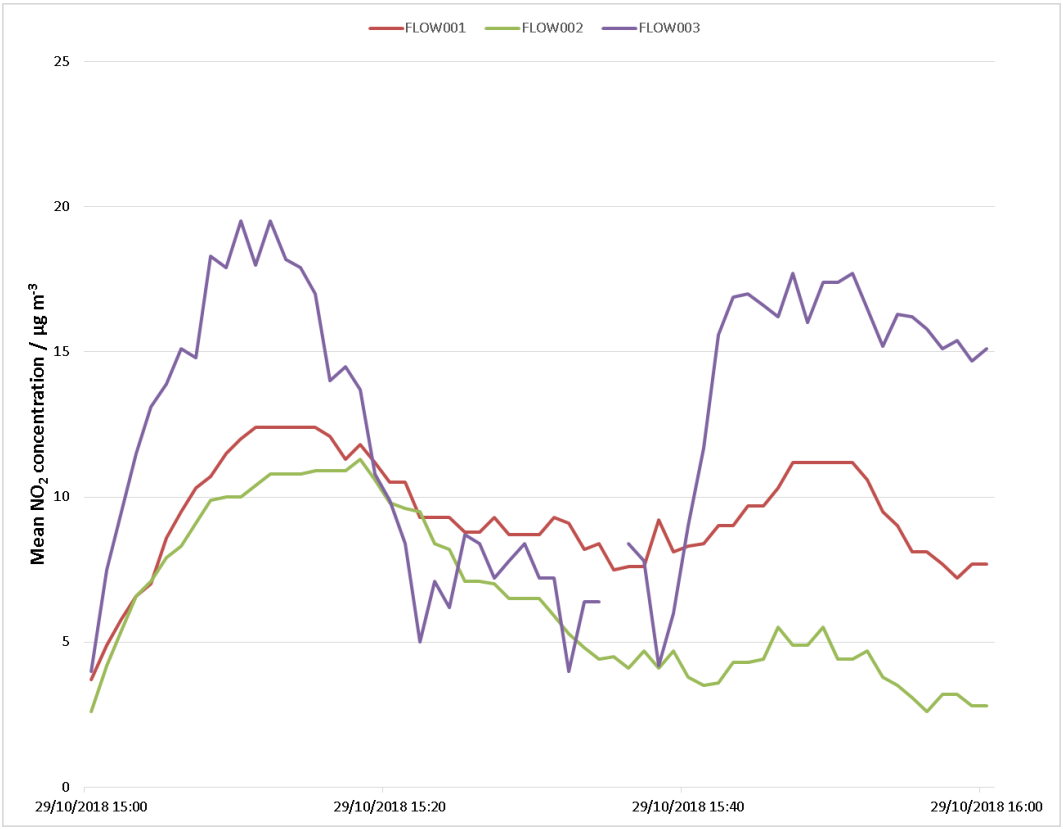

Figure 6: Time series chart of one-minute sensor NO_2 concentrations over the one-hour mobile monitoring campaign

Conclusion

Three portable Flow sensors by Plume Labs were put through rigorous testing to assess their suitability for personal exposure monitoring campaigns. The evaluation consisted of two tests, comparison to a fixed reference monitor for a period of two weeks and a one-hour mobile monitoring test assessing the portable sensors performance in different pollution environments.

The sensor unit revealed relatively poor precision between sensors and very poor accuracy when measuring PM compared to the fixed reference monitor. There were issues with reliability as there were gaps in data recorded over the two-week static testing period, although these gaps may have been a result of the method of use. The mobile monitoring test revealed a poor response to PM levels in different pollution environments. The small size and aesthetics of the device were its best features. From these tests conducted in October 2018, we do not recommend the use of these $PM_{2.5}$ sensors for Breathe London and similar studies, as the accuracy and precision of the device needs to be substantially improved.

The NO_2 sensors could not be compared to a fixed reference monitor for extended periods, therefore we cannot comment on the accuracy of these sensors. However, in the mobile monitoring test the sensors performed slightly better than the PM sensors, but the spatial pattern recorded by the units was not consistent with the expected fluctuation in NO_2 pollution levels.

It is important to note that these tests were only performed in a London pollution environment and it is likely in different regions and cities with different pollution sources the correction factors and accuracy would be different. For use in other environments similar tests would need to be run against reference monitors located in similar environments. To develop these devices further, additional post processing of raw data could be used to improve the accuracy and precision of the sensors.

Note: Results applicable to the version of the sensor tested at that time (October 2018). Any changes in the software algorithm used to convert sensor signals into pollutant concentrations would require retesting.

Note from manufacturer: We are grateful for the opportunity for Flow, our personal air quality sensor, to be tested as part of the Breathe London study. We wish to bring to attention the following important caveats.

1. Use case: Flow is designed to be used as a personal exposure tracker, in continuous Bluetooth connection to a smartphone. The static study in this report couldn't provide such regular interaction with a smartphone, which would have drastically increased the percentage of valid one-minute readings logged by the sensor. Due to the ensuing low data capture rate, caution should be taken when assessing the performance of the sensors on the sole basis of this first experiment.

2. Dysfunctional sensor: one of the three sensors the mobile sensing experiment relies on (Flow #002) was unfortunately non-operational. This is likely due to dust build-up (which a regular end-user could have fixed with help from customer support), and readings from this device should hence not be taken into account and are not representative of its nominal performance.

3. Product improvements: these results apply to the early versions of the sensor hardware and software available at the time of the study (October 2018). Since then, important improvements and corrections on both hardware, manufacturing and software were implemented, which would likely significantly improve results if tested in a comparable context.

Testing a set of fully functional sensors with a data logging setup that more adequately mirrors their regular operations is required to accurately evaluate device performance.