February 29, 2020

Dear AMT Editor:

Re: Response to Reviewers' Comments from the Public Discussion - MS No.: amt-2019-393

We thank the comments by the Referee No.3 and No.4 from the public discussion. Our responses are provided below.

<table>
<tr><td>Comment 1 from Referee No. 3</td><td>Lines 147-150, the authors stated that random search is more efficient, which is also a unique part of this study. Please explain why it is more efficient than manual or grid search in principle and if possible, give some quantitative information</td></tr>
<tr><td>Author's response</td><td>
The grid search can be considered as an automated manual search.
Grid search method and manual search method consider every combination of all the hyperparameters to build the learning models and each model needs to be evaluated to find out the one of the highest accuracies for training and prediction. For the XGBoost algorithm used in the manuscript Section 2.3.2, we tuned 7 hyperparameters. Each hyperparameter has 20 difference parameters. The grid looks like the following table:
<table>
<tr><td>Hyperparameter</td><td>P1</td><td>…</td><td>P20</td></tr>
<tr><td>H1</td><td></td><td></td><td></td></tr>
<tr><td>H2</td><td></td><td></td><td></td></tr>
<tr><td>…</td><td></td><td></td><td></td></tr>
<tr><td>H7</td><td></td><td></td><td></td></tr>
</table>
The complete grid search requires 20^7 = 1,280 millions of trials. In this study, we used 10-fold cross validation, which means each trial will run 10 times. So, the total runs will be 1,280 millions * 10 = over 12 billion, which is computationally expensive.
For random search, instead of computing the cases of all possible combinations, random combinations of hyperparameters are selected at each trial. Due to the random nature of sampling, the entire space of the grid could be reached (Zheng 2015).
The higher efficiency of random search can be explained by probability theory: Considering a sample space with a finite
</td></tr>
</table>

	maximum, if we need to find a sample that is within the top 5% of all the samples, 60 random observations would give us 95% probability to find the sample. The value of 60 is calculated as follows: As there are 5% eligible samples in the space, each random observation has 5% of chance to find the eligible sample. On another hand, each random observation has (1-5%) chance not to find the eligible sample. If we take n random observations, the chance of not to get the eligible sample would be $(1-0.05)^n$, or the chance of getting the eligible sample would be $1-(1-0.05)^n$. Let $1-(1-0.05)^n > 95\%$ And we can solve for n = 60 Therefore, random search method would significantly save computation resource but still have a good chance to guess the close-to-optimal combination of hyperparameters.
Author's changes in manuscript	We added a reference below in line 150 to explain the rationale of random search method. Zheng (2015) explained that random search with 60 samples will find a close-to-optimal combination with 95% of probability.

Comment 2 from Referee No. 3	Figure 1, I would appreciate a geographical map showing where is the monitoring location.
Author's response	a geographical map is added as Figure 2
Author's changes in manuscript	Added a new figure - Figure 2.

Comment 3 from Referee No. 3	CRAZ also monitors NOx, NMHC, ozone, and wind data, which may also influence the PM concentration. Why these data were not included in the machine learning?
Author's response	The low-cost sensor evaluated in this study only measured temperature (T) and relative humidity (RH). The ultimate goal of low-cost sensor application is to provide same quality data as the reference method using available information provided by the low-cost sensor. Therefore, we only used the parameters that the sensor measured. Next phase of the study would be testing other types of low-cost sensors, which may provide other parameters than T and RH. In that case, we would include those parameters in machine learning.

Author's changes in manuscript	Not applicable.

Comment 4 from Referee No. 3	Line 195: there is a typo: "SHAPR" should be "SHARP".
Author's response	Corrected
Author's changes in manuscript	Corrected to SHARP

Comment 5 from Referee No. 3	SHARP was used as the reference method for PM monitoring. How often was SHARP calibrated to ensure its data quality
Author's response	The SHARP instrument is regulated by the provincial air monitoring directive. It was calibrated monthly.
Author's changes in manuscript	We added a clarification in Line 171 The instrument was calibrated monthly

Comment 6 from Referee No. 3	Line 228: how the hyperparameters were determined?
Author's response	The hyperparameters were determined by the XGBoost algorithm itself. Detailed explanation of each hyperparameter is provided in the XGBoost documentation https://xgboost.readthedocs.io/en/latest/parameter.html
Author's changes in manuscript	We added the following reference in line 238 Detailed explanation of each hyperparameter is provided in the XGBoost documentation (XGBoost developers, 2019)

Comment 7 from Referee No. 3	Figure 3: will you explain what the shape of erlenmeyer flask means in the plot?
Author's response	The plots outside of the boxplots in Figure 3 is called violin plot. The violin plot is to describe the density of data. More details can be found in the following link:

	https://mode.com/blog/violin-plot-examples/
Author's changes in manuscript	The following sentences were added in Line 264 The violin plot in Figure 3 describes the distribution of the PM2.5 values measured by the low-cost sensor and SHARP using density curve. The width of each curve represents the frequency of PM2.5 values at each concentration level.

Comment 8 from Referee No. 3	One aspect of the uniqueness of this study is that its study covers different seasons. I would like to see a brief discussion how season influence the results of low-cost sensors.
Author's response	We added a section to discuss the seasonal impact in Section 3.5.2 We assessed the seasonal impact on the low-cost sensor by comparing the mean of absolute differences between the daily average of sensor values and the daily average of SHARP values in winter (December 2018 to February 2019) and spring (March 2019 to April 2019). A descriptive statistic is presented in Table 7. We used a two-sample t test to assess if the means of absolute differences for winter and spring were equal. The p value of the t test was 0.754. Because $P = 0.754 > \alpha = 0.05$, we retained the null hypothesis. There was not sufficient evidence at the $\alpha = 0.05$ level to conclude that the means of absolute differences between the low-cost sensor and SHARP PM values were significantly different for winter season and spring season.
Author's changes in manuscript	Added a section 3.5.2

Comment 1 from Referee No. 4	Work with sensor vendor to find out the reason of high equipment disability rate and build a larger sensor network in an area to evaluate the calibration method crossing different sensors
Author's response	We thank the reviewers' recommendation for our future work. We think it might be caused by water damage to the controller board. We planned to carry out a Phase 2 of the study to work with the sensor vendors or may try different sensors to understand what might be the cause that sensors do not last long. In phase 2, we also plan to deploy multiple sensors at multiple Alberta air monitoring stations for a longer time, such as 1 year, so

	we can test sensor precision and bias, as well as transferability of machine learning models.
Author's changes in manuscript	Not applicable

Comment 2 from Referee No. 4	Expand the temperature range to a warm condition, such as 30 C- 38 C to evaluation RH with higher temperature's effect on low cost sensor and calibration method
Author's response	We thank the reviewers' recommendation for our future work. Because of sensor failure, the sensor we used did not last to summer. We plan to set up experiments in phase 2 to cover a boarder ranges of weather condition We added a short discussion about season impacts. We assessed the seasonal impact on the low-cost sensor by comparing the mean of absolute differences between the daily average of sensor values and the daily average of SHARP values in winter (December 2018 to February 2019) and spring (March 2019 to April 2019). A descriptive statistic is presented in Table 7. We used a two-sample t test to assess if the means of absolute differences for winter and spring were equal. The p value of the t test was 0.754. Because $P = 0.754 > \alpha = 0.05$, we retained the null hypothesis. There was not sufficient evidence at the $\alpha = 0.05$ level to conclude that the means of absolute differences between the low-cost sensor and SHARP PM values were significantly different for winter season and spring season.
Author's changes in manuscript	Added a section 3.5.2

Respectful submitted

Calgary, Alberta Canada

Si et al.

Evaluation and Calibration of a Low-cost Particle Sensor in Ambient Conditions Using Machine Learning ~~Technologies~~Methods

Minxing Si[1, 2, a], Xiong Ying[1, a], Shan Du[3], Ke Du[1*]

[1]Department of Mechanical and Manufacturing Engineering, University of Calgary, 2500 University Drive. NW, Calgary, AB, Canada, T2N 1N4
[2]Tetra Tech Canada Inc., 140 Quarry Park Blvd, Calgary AB Canada, T2C 3G3
[3]Department of Computer Science, Lakehead University, 955 Oliver Road, Thunder Bay, ON, Canada, P7B 5E1

[*]*Correspondence to* Ke Du (kddu@ucalgary.ca)

[a.] The authors contributed equally to the work.

Abstract. Particle sensing technology has shown great potential for monitoring particulate matter (PM) with very few temporal and spatial restrictions because of low-cost, compact size, and easy operation. However, the performance of low-cost sensors for PM monitoring in ambient conditions has not been thoroughly evaluated. Monitoring results by low-cost sensors are often questionable. In this study, a low-cost fine particle monitor (Plantower PMS 5003) was co-located with a reference instrument, named Synchronized Hybrid Ambient Real-time Particulate (SHARP) monitor, in Calgary Varsity air monitoring station from December 2018 to April 2019. The study evaluated the performance of this low-cost PM sensor in ambient conditions and calibrated its readings using simple linear regression (SLR), multiple linear regression (MLR), and two more powerful machine learning algorithms using random search techniques for the best model architectures. The two machine learning algorithms are XGBoost and feedforward neural network (NN). Field evaluation showed that the Pearson r between the low-cost sensor and the SHAPR instrument was 0.78. Fligner and Killeen (F-K) test indicated a statistically significant difference between the variances of the $PM_{2.5}$ values by the low-cost sensor and by the SHARP instrument. Large overestimations by the low-cost sensor before calibration were observed in the field and were believed to be caused by the variation of ambient relative humidity. The root mean square error (RMSE) was 9.93 when comparing the low-cost sensor with the SHARP instrument. The calibration by the feedforward NN had the smallest RMSE of 3.91 in the test dataset, compared to the calibrations by SLR (4.91), MLR (4.65), and XGBoost (4.19). After calibrations, the F-K test using the test dataset showed that the variances of the $PM_{2.5}$ values by the NN and the XGBoost and by the reference method were not statistically significantly different. From this study, we conclude that feedforward NN is a promising method to address the poor performance of the low-cost sensors for $PM_{2.5}$ monitoring. In addition, the random search method for hyperparameters was demonstrated to be an efficient approach for selecting the best model structure.

Keywords: Low-cost sensor, machine learning, TensorFlow, XGBoost, $PM_{2.5}$

1 Introduction

Particular matter (PM), whether it is natural or anthropogenic, has pronounced effects on human health, visibility, and global climate (Charlson et al., 1992; Seinfeld and Pandis, 1998). To minimize the harmful effects of PM pollution, the Government of Canada launched the National Air Pollution Surveillance (NAPS) program in 1969 to monitor and regulate PM and other criteria air pollutants in populated regions, including ozone (O_3), sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2). Currently, PM monitoring is routinely carried out at 286 designated air sampling stations in 203 communities in all provinces and territories of Canada (Government of Canada, 2019). Many of the monitoring stations use Beta Attenuation Monitor (BAM), which is based on the adsorption of beta radiation, or Tapered Element Oscillating Microbalance (TEOM) instrument, which is a mass-based technology to measure PM concentrations. An instrument that combines two or more technologies, such as Synchronized Hybrid Ambient Real-time (SHARP), is also used in some monitoring stations. The SHARP instrument combines light scattering with beta attenuation technologies to determine PM concentrations.

Although these instruments are believed to be accurate for measuring PM concentration and have been widely used by many air monitoring stations worldwide (Chow and Watson, 1998; Patashnick and Rupprecht, 1991), they have common drawbacks: they can be challenging to operate, bulky, and expensive. The instrument costs from 8,000 Canadian dollars (CAD) to tens of thousands of dollars (Chong and Kumar, 2003). The SHARP instrument used in this study as a reference method costs approximately $40,000 (CDNova Instrument Ltd., 2017). Significant resources, such as specialized personnel or technicians, are also required for regular system calibration and maintenance. In addition, the sparsely spread stations may only represent PM levels in limited areas near the stations because PM concentrations vary spatially and temporally depending on local emission sources as well as meteorological conditions (Xiong et al., 2017). Such a low-resolution PM monitoring network cannot support public exposure and health effects studies that are related to PM, because these studies require high spatial- and temporal-resolution of monitoring network in the community (Snyder et al., 2013). In addition, the well-characterized scientific PM monitors are not portable due to their large size and volumetric flow rate, which means they are not practical for measuring personal PM exposure (White et al., 2012).

As a possible solution to the above problems, a large number of low-cost PM sensors could be deployed, and a high-resolution PM monitoring network could be constructed. Low-cost PM sensors are portable and commercially available. They are cost-effective and easy to deploy, operate, and maintain, which offers significant advantages compared to conventional analytical instruments. If many low-cost sensors are deployed, PM concentrations can be monitored continuously and simultaneously at multiple locations for a reasonable cost (Holstius et al., 2014). A dense monitoring network using low-cost sensors can also assist in mapping hotspots of air pollution, creating emission inventories of air pollutants, and estimating adverse health effects due to personal exposure to the PM (Kumar et al., 2015).

However, low-cost sensors present challenges for broad application and installation. Most sensor systems have not been thoroughly evaluated (Williams et al., 2014), and the data generated by these sensors are of questionable quality (Wang et al., 2015). Currently, most low-cost sensors are based on laser light scattering technology (LLS), and the accuracy of LLS is mostly affected by particle composition, size distribution, shape, temperature, and relative humidity (Jayaratne et al., 2018; Wang et al., 2015).

Several studies evaluated LLS sensors by comparing the performance of low-cost sensors with medium- to high-cost instruments under laboratory and ambient conditions. For example, Zikova et al. (2017) used low-cost Speck monitors to measure $PM_{2.5}$ concentrations in indoor and outdoor environments, and the low-cost sensors overestimated the concentration by 200% for indoor and 500% for outdoor, compared to a reference instrument – Grimm 1.109 dust monitor. Jayaratne et al. (2018) reported that PM_{10} concentrations generated by a Plantower low-cost particle sensor (PMS 1003) were 46% greater than a TSI 8350 DustTrak DRX aerosol monitor under a foggy environment. Wang et al. (2015) compared PM measurements from three low-cost LLS sensors – Shinyei PPD42NS, Samyoung DSM501A, and Sharp GP2Y1010AU0F – with a SidePack (TSI Inc.) using smoke from burning incense. High linearity was found with R^2 greater than 0.89, but the responses depended on particle composition, size, and humidity. Air Quality Sensor Performance Evaluation Center (AQ-SPEC) of South Coast Air Quality Management District (SCAQMD) also evaluated the performances of three Purple Air PA-II sensors (model: Plantower PMS 5003) by comparing their readings with two United States Environmental Protection Agency (US EPA) Federal Equivalent Method (FEM) instruments – BAM (MetOne) and Grimm dust monitors in laboratory and field environments in south California (Papapostolou et al., 2017). Overall, the three sensors showed moderate to good accuracy, compared to the reference instrument for $PM_{2.5}$ for a concentration range between 0 to 250 µg m^{-3}. Lewis et al. (2016) evaluated low-cost sensors in the field for O_3, nitrogen oxide (NO), NO_2, volatile organic compounds (VOCs), $PM_{2.5}$, and PM_{10}; only O_3 sensors showed good performance compared to the reference measurements.

Several studies developed calibration models using multiple techniques to improve low-cost sensors' performance. For example, De Vito et al. (2008) tested feedforward neural network (NN) calibration for benzene monitoring and reported a further calibration was needed for low concentrations. Bayesian optimization was also used to search feedforward NN structures for the calibrations of CO, NO_2, and NO_x low-cost sensors (De Vito et al., 2009). Zheng et al. (2018) calibrated Plantower low-cost particle sensor PMS 3003 by fitting a linear least-squares regression model. A nonlinear response was observed when ambient $PM_{2.5}$ exceeded 125 ug m^{-3}. The study concluded that a quadratic fit was more appropriate than a linear model to capture this nonlinearity.

Zimmerman et al. (2018) explored three different calibration models, including laboratory univariate linear regression, empirical MLR, and a more modern machine learning algorithm, random forests (RF), to improve Real-time Affordable Multiple-Pollutant (RAMP) sensor's performance. They found that the sensors calibrated by RF models improved their accuracy and precision over time, with average relative errors of 14% for CO, 2% for CO_2, 29% for NO_2, and 15% for O_3.

The study concluded that combing RF models with low-cost sensors is a promising approach to address the poor performance of low-cost air quality sensors.

Spinelle et al. (2015) reported several calibration methods for low-cost O_3 and NO_2 sensors. The best calibration method for NO_2 was an NN algorithm with feedforward architecture. O_3 could be calibrated by simple linear regression (SLR). Spinelle et al. (2017) also evaluated and calibrated NO, CO, and CO_2 sensors, and the calibrations by feedforward NN architectures showed the best results. Similarly, Cordero et al. (2018) performed a two-step calibration for an AQmesh NO_2 sensor using supervised machine learning regression algorithms, including NNs, RFs, and Support Vector Machines (SVMs). The first step produced an explanatory variable using multivariate linear regression. In the second step, the explanatory variable was fed into machine learning algorithms, including RF, SVM, and NN. After the calibration, the AQmesh NO_2 sensor met the standards of accuracy for high concentrations of NO_2 in the European Union's Directive 2008/50/EC on Air Quality. They highlighted the need to develop an advanced calibration model, especially for each sensor, as the responses of individual sensors are unique.

Williams et al. (2014) evaluated eight low-cost PM sensors; the study showed frequent disagreement between the low-cost PM sensors and FEMs. In addition, the study concluded that the performances of the low-cost sensors were significantly impacted by temperature and relative humidity (RH). Recurrent NN architectures were also tested for the calibrations of some gas sensors (De Vito et al., 2018; Esposito et al., 2016). The results showed that the dynamic approaches performed better than traditional static calibration approaches. Calibrations of $PM_{2.5}$ sensors were also reported in recent studies. Lin et al. (2018) performed two-step calibrations for $PM_{2.5}$ sensors using 236 hourly data collected on buses and road cleaning vehicles. The first step was to construct a linear model, and the second step used RF machine learning for further calibration. The RMSE after the calibrations was 14.76 µg m^{-3}, compared to a reference method. The reference method used in this study was a Dylos DCI1700 device, which is not a US EPA federal reference method (FRM) or FEM. Loh and Choi (2019) trained and tested SVC, k-nearest neighbor, RF, and XGBoost machine learning algorithms to calibrate $PM_{2.5}$ sensors using 319 hourly data. XGBoost archived the best performance with a RMSE of 5.0 µg m^{-3}. However, the low-cost sensors in this study were not co-located with the reference method, and the machine learning models were not tested using unseen data (test data) for predictive power and overfitting.

Although there are studies in calibrating low-cost sensors, most of them focused on gas sensors or used short-term data to calibrate PM sensors. To our best knowledge, no one has reported studies on PM sensor calibration using random search techniques for the best machine learning model's configuration under ambient conditions during different seasons. In this study, a low-cost fine particle monitor (Plantower PMS 5003) was co-located with a SHARP monitor Model 5030 at Calgary Varsity Air Monitoring Station in an outdoor environment from December 7, 2018, to April 26, 2019. The SHARP instrument is the reference method in this study and is a US EPA FEM (US EPA, 2016). The objectives of this stuudy are: (1) to evaluate the performance of the low-cost PM sensor in a range of outdoor environmental conditions by comparing its

$PM_{2.5}$ readings with those obtained from the SHARP instrument; and (2) to assess four calibration methods: a) a SLR or univariate linear regression based on the low-cost sensor values; b) a multiple linear regression (MLR) using the $PM_{2.5}$, RH, and temperature measured by the low-cost sensor as predictors; c) a decision-tree-based ensemble algorithm, called XGBoost or Extreme Gradient Boosting; and d) a feedforward NN architecture with a backpropagation algorithm.

XGBoost and NN are the most popular algorithms used on Kaggle – a platform for data science and machine learning competition. In 2015, 17 winners out of 29 competitions on Kaggle used XGBoost, 11 winners used deep NN algorithm (Chen and Guestrin, 2016).

This study is unique in the following ways:

1) To the best of our knowledge, this is the first comprehensive study using long-term data to calibrate low-cost particle sensors in the field. Most previous studies focused on calibrating gas sensors (Maag et al., 2018). There are two studies on PM sensor calibrations using machine learning, but they used a short-term dataset that did not include seasonal changes in ambient conditions (Lin et al., 2018; Loh and Choi, 2019). The shortcomings of the two studies were discussed above.
2) Although several studies researched the calibration of gas sensors using NN, this study explores multiple hyperparameters to search for the best NN architecture. Previous research configured one to three hyperparameters, compared to six in this study (De Vito et al., 2008, 2009, 2018; Esposito et al., 2016; Spinelle et al., 2015, 2017). In addition, this study tested the Rectified Linear Unit (ReLU) as the activation function in the feedforward NN. Compared to sigmoid and tanh activation functions used in the previous studies for NN calibration models, the ReLU function can accelerate the convergence of stochastic gradient descent to a factor of 6 (Krizhevsky et al., 2017).
3) Previous NN and tree-based calibration models used manual search or grid search for hyperparameters tuning. This study introduced random search method for the best calibration models. Random search is more efficient than traditional manual and grid search (Bergstra and Bengio, 2012) and evaluates more of the search space, especially when search space is more than three dimensions (Timbers, 2017). Zheng (2015) explained that random search with 60 samples will find a close-to-optimal combination with 95% of probability.

2 Method

2.1 Data preparation

One low-cost sensor unit was provided by Calgary-based company SensorUp and deployed at the Varsity station in the Calgary Reginal Airshed Zone (CRAZ) in Calgary, Alberta, Canada. The unit contains one sensor, one electrical board, and one housing as a shelter. The sensor in the unit is Plantower PMS 5003, and it measured outdoor fine particle ($PM_{2.5}$)

concentrations (μg m^{-3}), air temperature (°C), and RH (%) every six seconds. The minimum detectable particle diameter by the sensor is 0.3 μm. The instrument costs approximately $20 CAD and is referred to as the low-cost sensor in this paper.

The low-cost sensor is based on LLS technology; $PM_{2.5}$ mass concentration is estimated from the detected amount of scattered light. The LLS sensor is installed on the electrical board and then placed in the shelter for outdoor monitoring. The unit has a wireless link to a router in the Varsity station. A picture of the low-cost sensor and the monitoring environment where the low-cost sensor unit and the SHARP instrument were co-located is provided in Fig. 1. The location of the Varsity station is provided in Fig. 2. The router uses cellular service to transfer the data from the low-cost sensor to SensorUp's cloud data storage system. The measured outdoor $PM_{2.5}$, temperature, and RH data at a six-second interval from 00:00 on December 7, 2018, to 23:00 on April 26, 2019, were downloaded from the cloud data storage system for evaluation and calibration.

Figure 1: The low-cost sensor used in the study and the ambient inlet of the reference method – SHARP Model 5030

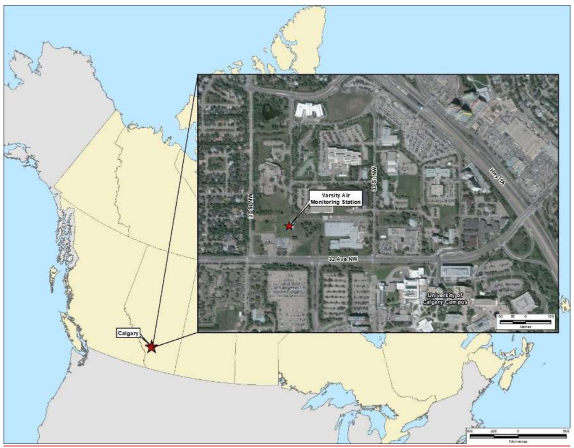

Figure 2: Location of Varsity Air Monitoring Station

Formatted: Caption

The reference instrument used to evaluate the low-cost sensor is a Thermal Fisher Scientific's SHARP Model 5030. The SHARP instrument was installed at the Calgary Varsity station by CRAZ. The SHARP instrument continuously uses two compatible technologies, light scattering and beta attenuation, to measure $PM_{2.5}$ every six minutes with an accuracy of ±5%. The SHARP instrument is operated and maintained by CRAZ in accordance with the provincial government's guideline outlined in Alberta's air monitoring directive. The instrument was calibrated monthly. Hourly $PM_{2.5}$ data are published on the Alberta Air Data Warehouse website (http://www.airdata.alberta.ca/). The Calgary Varsity station also continuously monitors CO, methane, oxides of nitrogen, non-methane hydrocarbons, outdoor air temperature, O_3, RH, total hydrocarbon, wind direction, and wind speed. Detailed information on the analytical systems for the CRAZ Varsity station can be found on their website (https://craz.ca/monitoring/info-calgary-nw/).

The ambient conditions in this study measured by the SHARP instrument are presented in Table 1.

Table 1: Ambient Condition Measured by SHARP

Climate Data	SHARP Value
Temperature	-31.4 ℃ ~ 19 ℃
RH	10% ~ 99%
Wind Speed	4.3 ~ 37.1 km/h 10 m

The following steps were taken to process the raw data from 00:00 on December 7, 2018, to 23:00 on April 26, 2019:

1) The six-second interval data recorded by the low-cost sensor, including $PM_{2.5}$, temperature, and RH, were averaged into hourly data to pair with SHARP data because only hourly SHARP data are publicly available.
2) The hourly sensor data and hourly SHARP data were combined into one structured data table. $PM_{2.5}$, temperature, and RH by the low-cost sensor as well as $PM_{2.5}$ by SHARP columns in the data table were selected. The data table then contains 3,384 rows and four columns. Each row represents one hourly data point. The columns include the data measured by the low-cost sensor and the SHARP instrument.
3) Rows in the data table with missing values were removed – 299 missing values for $PM_{2.5}$ from the low-cost sensor and 36 missing values for $PM_{2.5}$ from the SHARP instrument. The reason for missing data from the SHARP instrument is because of the calibration. However, the reason for missing data from the low-cost sensor is unknown.
4) The data used for NN were transformed by z standardization with a mean of zero and a standard deviation of one.

After the above steps, the processed data table with 3,050 rows and four columns was used for evaluation and calibration. The data file is provided in the supplementary information of this paper. Each row represents one example or sample for the training or testing by the calibration methods.

2.2 Low-cost sensor evaluation

Pearson correlation coefficient was used to compare the correlation for $PM_{2.5}$ values between the low-cost sensor and the SHARP. SHAPR was the reference method. The $PM_{2.5}$ data by the low-cost sensor and SHARP were also compared using root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAE).

Fligner and Killeen test (F-K test) was used to evaluate the equality (homogeneity) of variances for $PM_{2.5}$ values between the low-cost sensor and the SHARP instrument (Fligner and Killeen, 1976). F-K test is a superior option in terms of robustness and power when data are non-normally distributed, the population means are unknown, or outliers cannot be removed (Conover et al., 1981; de Smith, 2018). The null hypothesis of the F-K test is that all populations' variances are equal; the alternative hypothesis is that the variances are statistically significantly different.

Field Code Changed

Field Code Changed

2.3 Calibration

Four calibration methods were evaluated: SLR, MLR, XGBoost, and NN. Some predictions from the SLR, MLR, and XGBoost have negative values because they extrapolate observed values and regression is unbounded. When the predicted $PM_{2.5}$ values generated by these calibration methods were negative, the negative values were replaced with the sensor data.

MLR, XGBoost, and feedforward NN use the $PM_{2.5}$, temperature, and RH data measured by the low-cost sensor as inputs. The $PM_{2.5}$ measured by the SHARP instrument is used as the target to supervise the machine learning process. The processed dataset with 3,050 rows and four columns was randomly shuffled and then divided into a training set, which was the data used to build models and minimize the loss function, and a test set, which was the data that the model has never run with before testing (Si et al., 2019). The test dataset was only used once and gave an unbiased evaluation of the final model's performance. The evaluation was to test the ability of the machine learning model to provide sensible predictions with new inputs (LeCun et al., 2015). The training dataset had 2,440 examples (samples). The test dataset had 610 examples (samples).

Field Code Changed

Field Code Changed

2.3.1 Simple linear regression and multiple linear regression

The calibration by a SLR used Equation 1.

$$\hat{y} = \beta\_0 + \beta\_1 \times 〖PM〗\_2.5 \quad (1)$$

β_0 and β_1 are the model coefficient and were calculated using the training dataset. $\hat{y}$ is model predicted (calibrated) values. $PM_{2.5}$ is the value measured by the low-cost sensor.

The MLR used $PM_{2.5}$, RH, and temperature measured by the low-cost sensor as predictors because the low-cost sensor only measured these parameters. The model is expressed as Equation 2.

$$\hat{y} = \beta_0 + \beta_1 \times PM_{2.5} + \beta_2 \times T + \beta_3 \times RH \quad (2)$$

The model coefficients, β_0 to β_3, were calculated using the training dataset with SHARP provided readings as $\hat{y}$. The outputs of the models generated by the SLR and MLR were evaluated by comparing to the SHARP's readings in the test dataset.

2.3.2 XGBoost

XGBoost is a scalable decision tree-based ensemble algorithm, and it uses a gradient boosting framework (Chen and Guestrin, 2016). The XGBoost was implemented using the XGBoost (Version 0.90) and sklearn (Version 0.21.2) packages in Python (Version 3.7.3). Random search method (Bergstra and Bengio, 2012) was used to tune the hyperparameters in the XGBoost algorithm, and the hyperparameters tuned include

Field Code Changed

Field Code Changed

- Number of trees to fit (n_estimator)
- Maximum depth of a tree (max_depth)
- Step size shrinkage used in update (learning_rate)
- Subsample ratio of columns when constructing each tree (colsample_bytree)
- Minimum loss reduction required to make a further partition on a leaf node of the tree (gamma)
- L2 regularization on weights (reg_lambda)
- Minimum sum of instance weight needed in a child (min_child_weight)

Detailed explanation of each hyperparameter is provided in the XGBoost documentation (XGBoost developers, 2019). Ten-fold cross-validation was used to select the best model with minimum MSE from the random search. The best model was then evaluated against the SHARP $PM_{2.5}$ data using the test dataset.

2.3.3 Neural network

A fully connected feedforward NN architecture was used in the study. In a fully connected NN, each unit (node) in a layer is connected to each unit in the following layer. Data from the input layer are passed through the network until the unit(s) in the output layer is (are) reached. An example of a fully connected feedforward NN is presented in Fig. 3~~2~~. A backpropagation algorithm is used to minimize the difference between the SHARP measured values and the predicted values (Rumelhart et al., 1986).

Field Code Changed

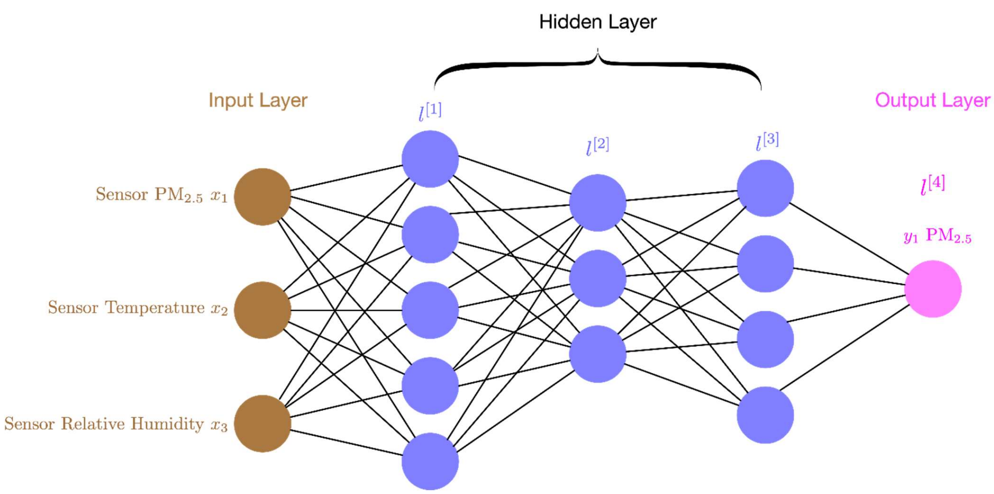

Figure 3~~2~~: Example of a Neural Network Structure

The NN was implemented using the Keras (Version 2.2.4) and TensorFlow (Version 1.14.0) libraries in Python (Version 3.7.3). Keras and TensorFlow were the most referenced deep learning framework in scientific research in 2017 (RStudio, 2018). Keras is the front end of TensorFlow.

Learning rate, L2 regularization rate, numbers of hidden layers, number of units in the hidden layers, and optimization methods were tuned using random search method provided in the scikit-learn machine learning library. Ten-fold cross-validation was used to evaluate the models. The model with the minimum MSE was considered to be the best-fit model and then used for model testing.

3 Results and Discussion

3.1 Sensor evaluation

3.1.1 Hourly data

The RMSE, MSE, and MAE between the low-cost sensor and SHARP for the hourly $PM_{2.5}$ data were 10.58, 111.83, and 5.74. The Pearson correlation coefficient r value was 0.78. The $PM_{2.5}$ concentrations by the sensor ranged from 0 µg m^{-3} to 178 µg m^{-3} with a standard deviation of 14.90 µg m^{-3} and a mean of 9.855 µg m^{-3}. The $PM_{2.5}$ concentrations by SHARP ranged from 0 µg m^{-3} to 80 µg m^{-3} with a standard deviation of 7.80 and a mean of 6.55 µg m^{-3}. Both SHARP and the low-cost sensor dataset had a median of 4.00 µg m^{-3} based on hourly data (Fig. 4~~3~~). The violin plot in Figure 4 descirbes the distribution of the $PM_{2.5}$ values measured by the low-cost sensor and SHARP using density curve. The width of each curve represents the frequency of $PM_{2.5}$ values at each concentration level. The p-value from the F-K test was less than 2.2×10^{-16}, indicating that the variance of the $PM_{2.5}$ values measured by the low-cost sensor was statistically significantly different from the variance of the $PM_{2.5}$ values measured by the SHARP instrument.

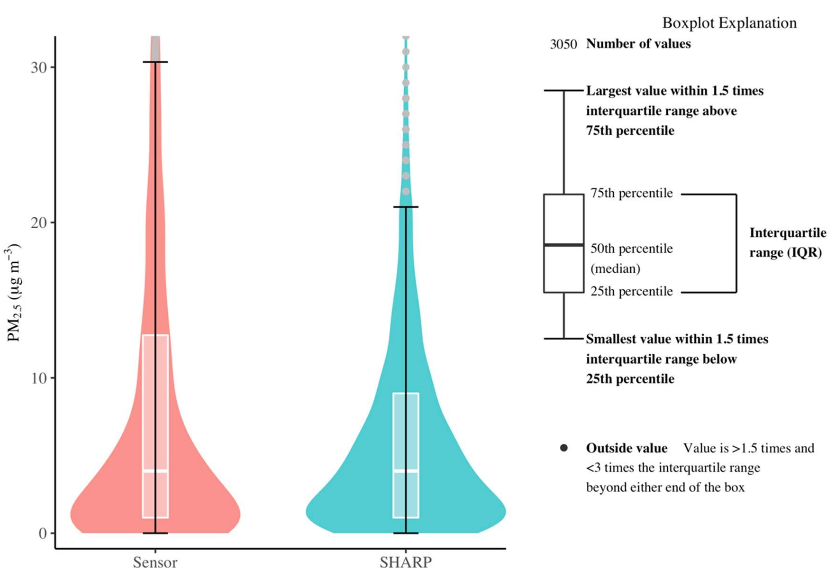

Figure 4~~3~~: Comparison of the Hourly $PM_{2.5}$ Values between the Low-Cost PM Sensor and SHARP**.** Based on 3,050 hourly paired data. The low-cost sensor has 250 hourly data greater than 30 µg m^{-3}. SHARP has 174 hourly data greater than 20 µg m^{-3}. Bars indicate the 25th and 75th percentile values, whiskers extend to values within 1.5 times IQR, and dots represent values outside of the IQR. The boxplot explanation on the right is adjusted from DeCicco (2016)

Field Code Changed

3.1.2 24 Hour rolling average data

Over 24 hours, the median value for SHARP was 5.38 µg m^{-3} and for the low-cost sensor was 5.01 µg m^{-3}. Over five months (December 2018 to April 2019), the low-cost sensor tended to generate higher $PM_{2.5}$ values compared to the SHARP monitoring data (Fig. 5~~4~~)

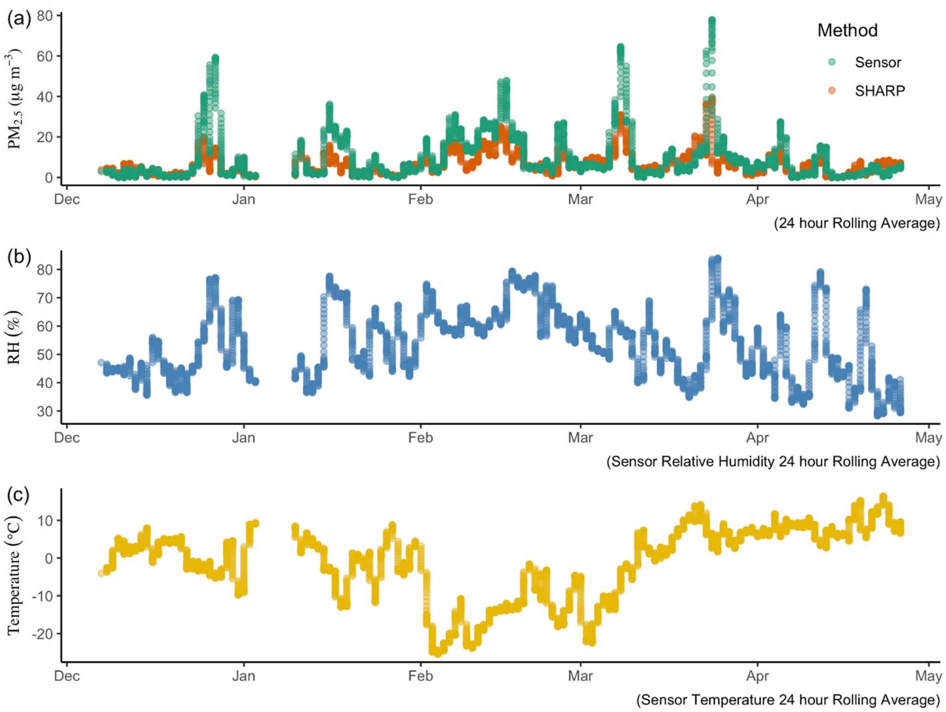

Figure 5~~4~~: $PM_{2.5}$, Relative Humidity, and Temperature data on the basis of 24 hour rolling average

When $PM_{2.5}$ concentrations were greater than 10 µg m^{-3}, the low-cost sensor consistently produced values that were higher than the reference method (Fig.6~~5~~). When the concentrations were less than 10 µg m^{-3}, the performance of the low-cost sensor was closed to the reference method producing slightly smaller values (Fig. 6~~5~~)

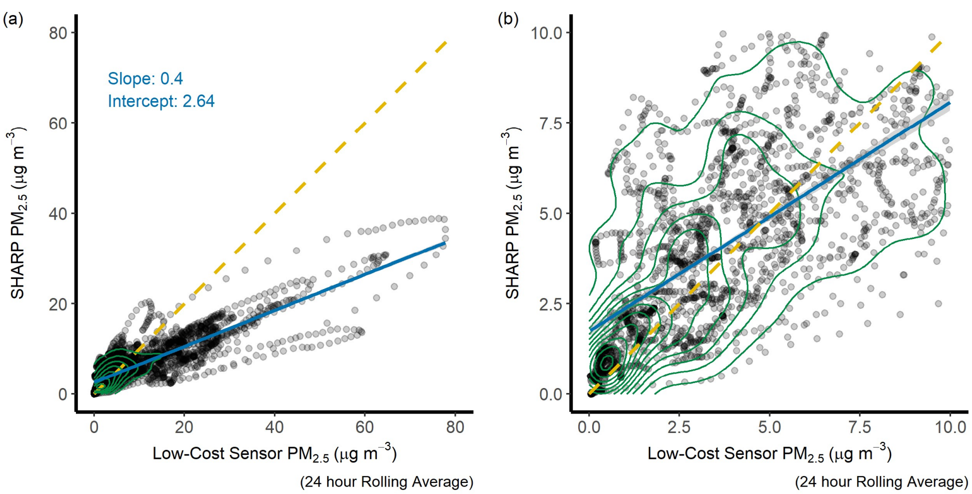

Figure 6~~5~~: SHARP verse Low-Cost Sensor $PM_{2.5}$ Concentration (µg m^{-3}). The yellow dashed line is a 1:1 line. The solid blue line is a regression line. (a) plot is in full scale, (b) plot is a zoom-in plot of plot a. The green circle represents data density.

3.2 Calibration by simple linear regression and multiple linear regression

The RMSE was 4.91 calibrated by SLR and 4.65 by MLR (Table 2). The r value was 0.74 by the SLR and 0.77 by MLR . The p-values in the F-K test by the SLR and MRL were less than 0.05, which suggested that the variances of the $PM_{2.5}$ values were statistically significantly different.

Table 2: Calibration Results by SLR and MLR using Test Dataset

Criteria	Low-Cost Sensor	SLR	MLR
RMSE	9.93	4.91	4.65
MSE	98.62	24.09	21.61
MAE	5.63	3.21	3.09
Pearson r	0.74	0.74	0.77
p-value in the F-K test	7.062×10^{-09}	5.81×10^{-13}	9.90×10^{-10}
β_0	-	2.49	8.47

β_1	0.41	0.46
β_2		-0.12
β_3		-0.0055

Note: The test dataset contains 660 examples.

3.3 Calibration by XGBoost

The hyperparameters selected by the random search for the best model using XGBoost is presented in Table 3.

Table 3: Hyperparameters for the Best XGBoost Model

XGBoost Hyperparameters	Values
Number of trees to fit (n_estimator)	37
Maximum depth of a tree (max_depth)	9
Step size shrinkage used in update (learning_rate)	0.33
Subsample ratio of columns when constructing each tree (colsample_bytree)	0.83
Minimum loss reduction required to make a further partition on a leaf node of the tree (gamma)	6.36
L2 regularization (Ridge Regression) on weights (reg_lambda)	33.08
Minimum sum of instance weight needed in a child (min_child_weight)	25.53

In the training dataset, the RMSE was 3.03, and the MAE was 1.93 by the best XGBoost model. The RMSE in the test dataset reduced by 57.8% using the XGBoost from 9.93 by the sensor to 4.19 (Table 4). The p-value in the F-K test using the test dataset was 0.7256, which showed no evidence that the $PM_{2.5}$ values varied with statistical significance between the XGBoost predicted values and SHARP measured values.

Table 4: Calibration Results by XGBoost using Test Dataset

Criteria	Low-Cost Sensor	XGBoost
RMSE	9.93	4.19
MSE	98.62	17.61
MAE	5.63	2.63
Pearson r	0.74	0.82
p-value in the F-K test	7.062×10^{-09}	0.7256

Note: The test dataset contains 610 examples.

3.4 Calibration by neural network

The hyperparameters for the best NN model are presented in Table 5.

Table 5: Hyperparameters for the Best Neural Network Model

NN Hyperparameters	Values

Learning_rate	0.001
L2 regularization	0.01
Numbers of hidden layer(s)	5
Numbers of units in the hidden layer(s)	32-32-32-32-32
Optimization method	Nadam
Epochs	100

In the training dataset, the RMSE was 3.22, and the MAE was 2.17 by the best NN-based model. The RMSE reduced by 60% using the NN from 9.93 to 3.91 in the test dataset (Table 6). The p-value in the F-K test was 0.43, which suggested that the variances in the $PM_{2.5}$ values were not statistically significantly different between the NN predicted values and SHARP measured values.

Table 6: Calibration Results by Neural Network using Test Dataset

Criteria	Low-Cost Sensor	Neural Network
RMSE	9.93	3.91
MSE	98.62	15.26
MAE	5.63	2.38
Pearson r	0.74	0.85
p-value in the F-K test	7.062×10^{-09}	0.43

Note: the test dataset includes 610 examples.

3.5 Discussion

3.5.1 Relative humidity impact

RH has significant effects on the low-cost sensor's responses. The RH trend matched the low-cost sensor's $PM_{2.5}$ trend closely. The spikes in the low-cost sensor's $PM_{2.5}$ trend corresponded with the increases of RH values, and the low-cost sensor intended to produce inaccurate high $PM_{2.5}$ values when RH suddenly increased (Fig. 5~~4~~). However, the relationship between $PM_{2.5}$ and RH was not linear (Fig. 7~~6~~)

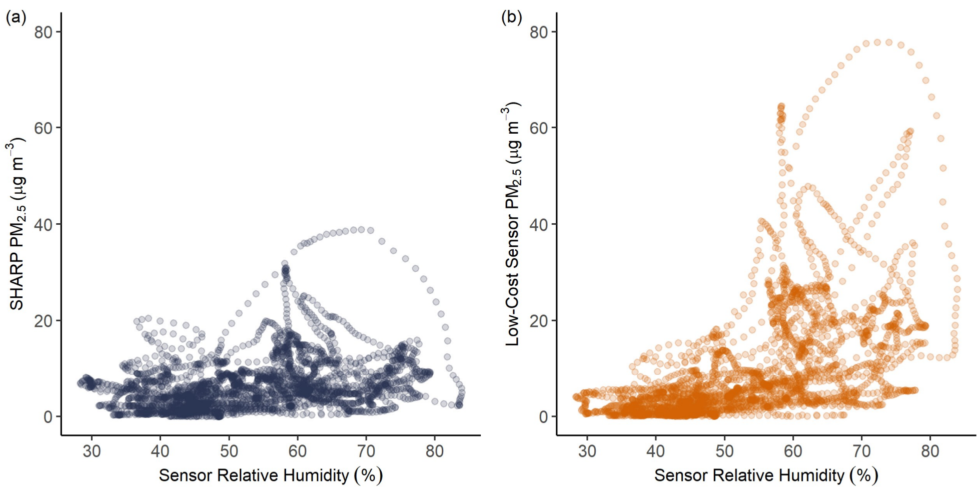

Figure 76: PM$_{2.5}$ verse Relative Humidity

3.5.2 Seasonal impact

We assessed the seasonal impact on the low-cost sensor by comparing the mean of absolute differences of daily average between the sensor values and the SHARP values in winter (December 2018 to February 2019) and spring (March 2019 to April 2019). A descriptive statistics is presented in Table 7.

Table 7: Descriptive Statistics by Seasons

Season	Sample Size (n)	Mean[1]	Standard Deviation
Winter	78	5.13	6.95
Spring	57	4.76	6.45

Note: 1) Mean is calculated by $\sum_{i=1}^{n}(|(sensor_{daily} - SHARP_{daily})|)/n$

We used a two-sample t test to assess if the average differences for winter and spring were statistically significant. The p value of the t test was 0.754. Because P = 0.754 > α = 0.05, we retained the null hypothesis. There was not sufficient evidence at the α = 0.05 level to conclude that the means of absolute differences between the low-cost sensor and SHARP PM values were siginicantly different for winter season and spring season.

Formatted: Caption
Formatted: Caption
Formatted: Superscript
Formatted Table
Formatted: Font: 9 pt
Formatted: Font: 9 pt
Formatted: Font: 9 pt
Formatted: Font: 9 pt
Formatted: Font: 9 pt
Formatted: Font: 9 pt
Formatted: Font: 9 pt
Formatted: Font: 9 pt
Formatted: Font: 9 pt
Formatted: Font: 9 pt
Formatted: Font: 9 pt
Formatted: Font: 9 pt
Formatted: Font: 9 pt

3.5.3~~2~~ Calibration assessment

Descriptive statistics of the $PM_{2.5}$ concentrations in the test dataset for SHARP, low-cost sensor, XGBoost, NN, SLR, and MLR are presented in Table 8~~7~~. The arithmetic mean of the $PM_{2.5}$ concentrations measured by the low-cost sensor was 9.44 µg m^{-3}. In contrast, the means of the $PM_{2.5}$ concentrations were 6.44 µg m^{-3} by SHARP, 6.40 µg m^{-3} by XGBoost, and 6.09 µg m^{-3} by NN.

Table 88~~7~~: Descriptive statistics of $PM_{2.5}$ Concentrations using the Test Dataset

PM2.5 Concentration (µg m^{-3})	**SHARP**	**Low-Cost Sensor**	**XGBoost**	**NN**	**SLR**	**MLR**
Minimum	0.00	0.00	0.00	0.19	2.49	0
1st quartile	2.00	0.083	2.09	1.78	2.83	3.27
Median	4.00	4.00	4.98	4.16	4.13	4.79
Mean	6.44	9.44	6.40	6.09	6.37	6.42
3rd quartile	8.00	11.94	8.61	8.20	7.39	7.18
Maximum	49.00	103.33	39.94	47.19	44.97	48.56
SD	7.32	13.53	6.03	6.23	5.57	5.67

NN and XGBoost produced data distributions that were similar to SHARP (Fig. 88~~7~~). SLR had the worst performance. Fig. 99~~7~~ shows that SLR could not predict low concentrations. The predictions made by NN and XGBoost ranged from 0.19 µg m^{-3} to 47.19 µg m^{-3} and from 0.00 µg m^{-3} to 39.94 µg m^{-3}.

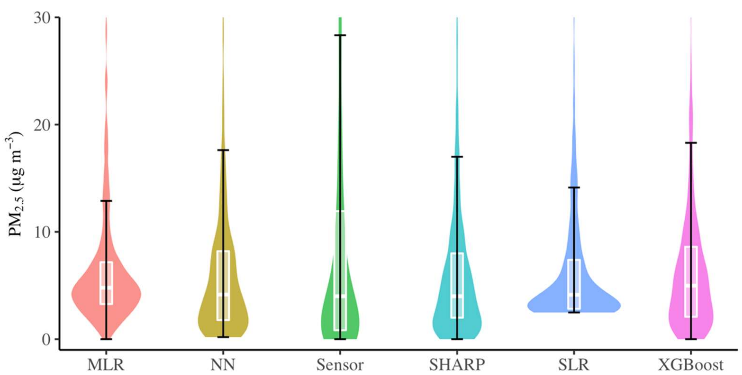

Figure 8: Data Density Comparison in the Test Dataset. Based on 610 Test Examples. NN: neural network, MRL: Multiple Linear Regression, SLR: Simple Linear Regression. $PM_{2.5}$ data greater than 30 µg m^{-3} are not shown in the figure. See the boxplot explanation in Figure 3.

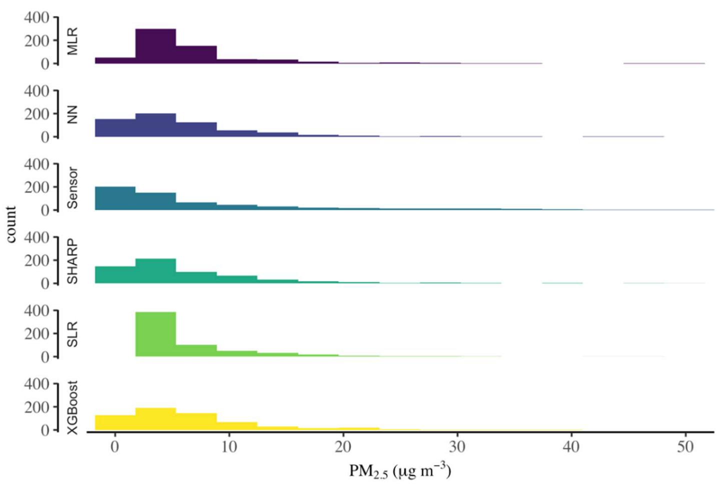

Figure 9: Data Distribution Comparison. Based on 610 Test Examples. NN: neural network, MRL: Multiple Linear Regression, SLR: Simple Linear Regression.

In the test dataset, the NN produced the lowest MAE of 2.38 (Fig. 10~~9~~). The MAEs were 2.63 by XGBoost, 3.09 by MLR, and 3.21 by SLR, when compared with the $PM_{2.5}$ data measured by the SHARP instrument. The NN also had the lowest RMSE score in the test dataset. The RMSEs were 3.91 for the NN, 4.19 for XGBoost, and 9.93 for the low-cost sensor (Fig. 10~~9~~). The Pearson r value by the NN was 0.85, compared to 0.74 by the low-cost sensor.

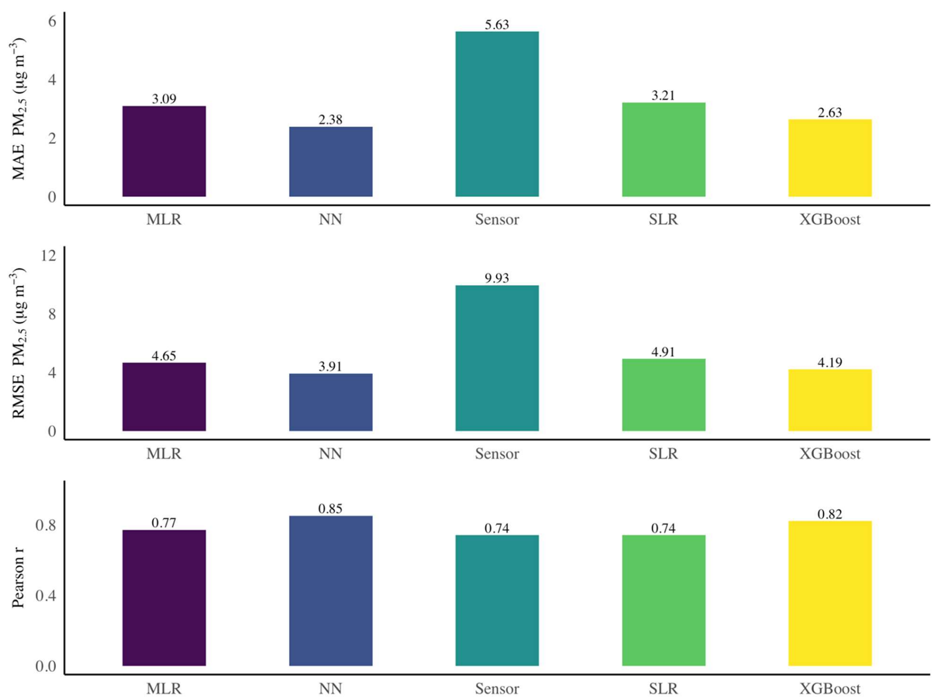

Figure 10: Performances of Different Calibration Methods. Based on 610 Test Examples. NN: neural network, MRL: Multiple Linear Regression, SLR: Simple Linear Regression.

The XGBoost and NN machine learning algorithms have a better performance, compared to traditional SLR and MRL calibration methods. NN calibration reduced RMSE by 60%. Both NN and XGBoost demonstrated the ability to correct the bias for high concentrations made by the low-cost sensor (Fig. 11~~10~~ and Fig. 12~~11~~). Most of the values that were greater than

10 µg m^{-3} in the NN model fall closer to the yellow 1:1 line (Fig. 11~~10~~). NN had slightly better performance for low concentrations compared to XGBoost.

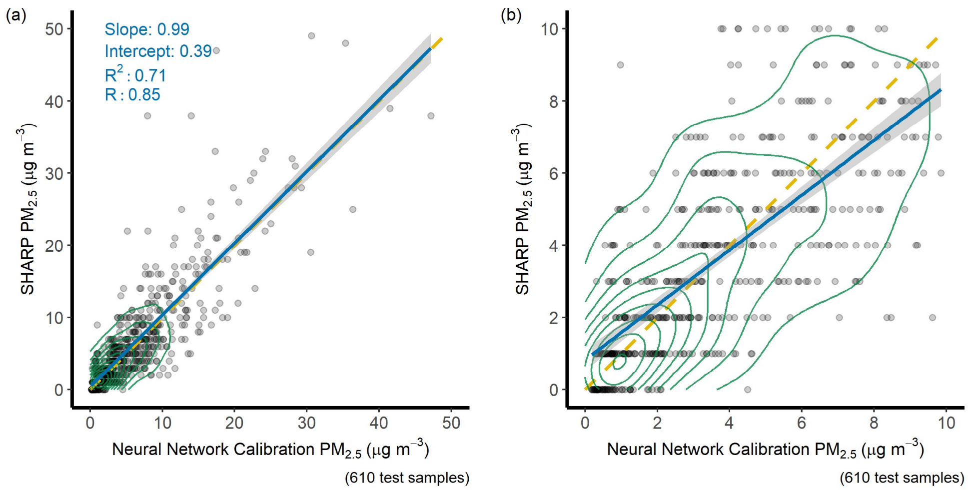

Figure 11: Comparison between the NN predictions and SHARP. Based on 610 test examples. Plot (a) is in full scale. Plot (b) is a zoom-in plot of plot (a). The solid blue line is a regression line. The yellow dashed line is a 1:1 line. The green circle represents data density. The grey area along the regression line represents 1 standard deviation.

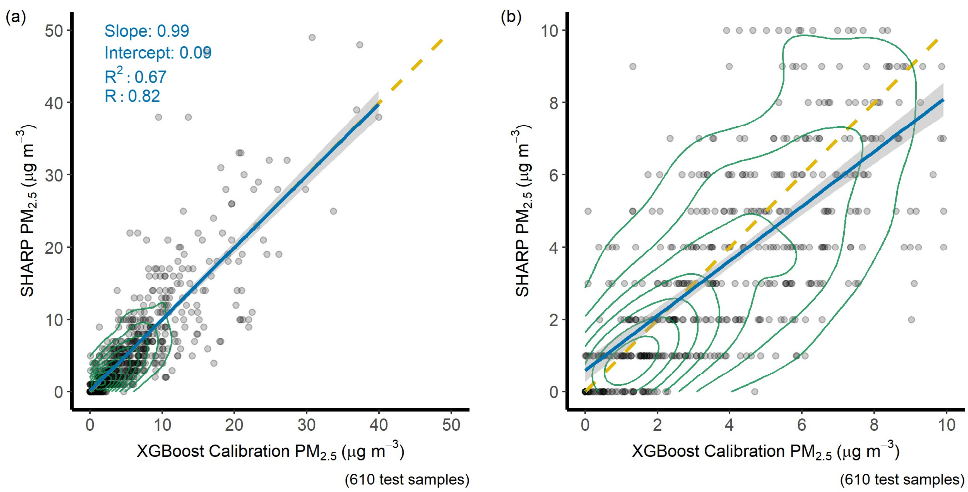

Figure 12: Comparison between the XGBoost predictions and SHARP. Based on 610 test examples. NN: Neural Network. Plot (a) is in full scale. Plot (b) is a zoom-in plot of plot (a). The solid blue line is a regression line. The yellow dashed line is a 1:1 line. The green circle represents data density. The grey area along the regression line represents 1 standard deviation.

4 Conclusions

In this study, we evaluated one low-cost sensor against a reference instrument – SHARP – using 3,050 hourly data from 00:00 on December 7, 2018, to 23:00 on April 26, 2019. The p-value from the F-K test suggested that the variances in the $PM_{2.5}$ values were statistically significantly different between the low-cost sensor and the SHARP instrument. Based on the 24-hour rolling average, the low-cost sensor in this study tended to report higher $PM_{2.5}$ values compared to the SHARP instrument. The low-cost sensor had strong bias when $PM_{2.5}$ concentrations were greater than 10 μg m^{-3}. The study also showed that the sensor's bias responses are likely caused by the sudden changes of RH.

Four calibration methods were tested and compared, including SLR, MLR, NN, and XGBoost. The p-values from the F-K tests for the XGBoost and NN were greater than 0.05, which indicated that, after calibration by the XGBoost and the NN, the variances of the $PM_{2.5}$ values were not statistically significantly different from the variance of the $PM_{2.5}$ values measured by the SHARP instrument. In contrast, the p-values from the F-K tests for the SLR and MLR were still less than 0.05. The NN generated the lowest RMSE score in the test dataset with 610 samples. The RMSE by NN was 3.91, the lowest of the four methods. RMSEs were 4.91 by SLP, 4.65 by MLR, and 4.19 by XGBoost.

However, a wide installation of low-cost sensors may still face challenges, including

- Durability of low-cost sensor. The low-cost sensor used in the study was deployed in ambient environment. We installed four sensors between December 7, 2018, and June 20, 2019. Only one sensor lasted approximately five months; the data from this sensor was used in this study. The other three sensors only lasted two weeks to one month and collected limited data. These three sensors did not collect enough data for machine learning and, therefore, were not used in this study.
- Missing data. In this study, the low-cost sensor dataset has 299 missing values for $PM_{2.5}$ concentrations. The reason for the missing data is unknown.
- Transferability of machine learning models. The models, developed by the two more powerful machine learning algorithms and used to calibrate the low-cost sensor data, tend to be sensor-specific because of the nature of machine learning. Further research is needed to test the transferability of the models for broader use.

Data availability. The hourly sensor data and hourly SHARP data are provided online at 10.5281/zenodo.3473833

Author Contribution: MS conducted evaluation and calibrations. YX installed the sensor and monitored and collected the sensor data. MS and YX wrote the manuscript together and have equal contribution. SD edited the machine learning methods. DK secured the funding and supervised the project. All authors discussed the results and commented on the manuscript.

Competing interests. The authors declare no competing interest.

Disclaimer. Reference to any companies or specific commercial products does not constitute endorsement or recommendation by the authors.

Acknowledgments. The authors wish to thank SensorUp for providing the low-cost sensors, and Calgary Region Airshed Zone's air quality program manager Mandeep Dhaliwal for helping with the installation of the PM sensors and a 4G LTE router, as well as the collection of the SHARP data. The authors would also like to thank Jessica Coles for editing this manuscript.

The project was funded by Natural Sciences and Engineering Research Council of Canada (NSERC) Engage Program (No. EGP 521823–17) and NSERC Collaborative Research and Development Program (No. CRDPJ 535813-18).

References

Bergstra, J. and Bengio, Y.: Random Search for Hyper-Parameter Optimization, J. Mach. Learn. Res., 13, 281–305, 2012.

CDNova Instrument Ltd.: SHARP Cost Estimate, 2017.

Charlson, R. J., Schwartz, S. E., Hales, J. M., Cess, R. D., Coakley, J. A., Hansen, J. E. and Hofmann, D. J.: Climate Forcing by Anthropogenic Aerosols, Science, 255(5043), 423–430, doi:10.1126/science.255.5043.423, 1992.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, pp. 785–794, ACM Press, San Francisco, California, USA., 2016.

Chong, C.-Y. and Kumar, S. P.: Sensor networks: Evolution, opportunities, and challenges, Proc. IEEE, 91(8), 1247–1256, doi:10.1109/JPROC.2003.814918, 2003.

Chow, J. C. and Watson, J. G.: Guideline on Speciated Particulate Monitoring, [online] Available from: https://www3.epa.gov/ttn/amtic/files/ambient/pm25/spec/drispec.pdf, 1998.

Conover, W. J., Johnson, M. E. and Johnson, M. M.: A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data., 1981.

Cordero, J. M., Borge, R. and Narros, A.: Using statistical methods to carry out in field calibrations of low cost air quality sensors, Sens. Actuators B Chem., 267, 245–254, doi:10.1016/j.snb.2018.04.021, 2018.

De Vito, S., Massera, E., Piga, M., Martinotto, L. and Di Francia, G.: On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sens. Actuators B Chem., 129(2), 750–757, doi:10.1016/j.snb.2007.09.060, 2008.

De Vito, S., Piga, M., Martinotto, L. and Di Francia, G.: CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization, Sens. Actuators B Chem., 143(1), 182–191, doi:10.1016/j.snb.2009.08.041, 2009.

De Vito, S., Esposito, E., Salvato, M., Popoola, O., Formisano, F., Jones, R. and Di Francia, G.: Calibrating chemical multisensory devices for real world applications: An in-depth comparison of quantitative machine learning approaches, Sens. Actuators B Chem., 255, 1191–1210, doi:10.1016/j.snb.2017.07.155, 2018.

DeCicco, L.: Exploring ggplot2 boxplots - Defining limits and adjusting style, Explor. Ggplot2 Boxplots - Defin. Limits Adjust. Style [online] Available from: https://owi.usgs.gov/blog/boxplots/, 2016.

Esposito, E., De Vito, S., Salvato, M., Bright, V., Jones, R. L. and Popoola, O.: Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems, Sens. Actuators B Chem., 231, 701–713, doi:10.1016/j.snb.2016.03.038, 2016.

Formatted: Bibliography, Automatically adjust right indent when grid is defined, Widow/Orphan control, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Fligner, M. A. and Killeen, T. J.: Distribution-Free Two-Sample Tests for Scale, J. Am. Stat. Assoc., 71(353), 210–213, doi:10.1080/01621459.1976.10481517, 1976.

Government of Canada: National Air Pollution Surveillance (NAPS) Network - Open Government Portal, Natl. Air Pollut. Surveill. NAPS Netw. [online] Available from: https://open.canada.ca/data/en/dataset/1b36a356-defd-4813-acea-47bc3abd859b (Accessed 17 September 2019), 2019.

Holstius, D. M., Pillarisetti, A., Smith, K. R. and Seto, E.: Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in California, Atmospheric Meas. Tech., 7(4), 1121–1131, doi:10.5194/amt-7-1121-2014, 2014.

Jayaratne, R., Liu, X., Thai, P., Dunbabin, M. and Morawska, L.: The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog, Atmospheric Meas. Tech., 11(8), 4883–4890, doi:10.5194/amt-11-4883-2018, 2018.

Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet classification with deep convolutional neural networks, Commun ACM, 60(6), 84–90, 2017.

Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L. and Britter, R.: The rise of low-cost sensing for managing air pollution in cities, Environ. Int., 75, 199–205, doi:10.1016/j.envint.2014.11.019, 2015.

Laerd Statistics: Spearman's Rank-Order Correlation - A guide to when to use it, what it does and what the assumptions are., Laerd Stat. [online] Available from: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php (Accessed 24 February 2020), 2020.

LeCun, Y., Bengio, Y. and Hinton, G.: Deep learning, Nature, 521(7553), 436–444, doi:10.1038/nature14539, 2015.

Lewis, A. C., Lee, J. D., Edwards, P. M., Shaw, M. D., Evans, M. J., Moller, S. J., Smith, K. R., Buckley, J. W., Ellis, M., Gillot, S. R. and White, A.: Evaluating the performance of low cost chemical sensors for air pollution research, Faraday Discuss., 189, 85–103, doi:10.1039/C5FD00201J, 2016.

Lin, Y., Dong, W. and Chen, Y.: Calibrating Low-Cost Sensors by a Two-Phase Learning Approach for Urban Air Quality Measurement, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 2(1), 1–18, doi:10.1145/3191750, 2018.

Loh, B. G. and Choi, G.-H.: Calibration of Portable Particulate Matter–Monitoring Device using Web Query and Machine Learning, Saf. Health Work, S2093791119302811, doi:10.1016/j.shaw.2019.08.002, 2019.

Maag, B., Zhou, Z. and Thiele, L.: A Survey on Sensor Calibration in Air Pollution Monitoring Deployments, IEEE Internet Things J., 5(6), 4857–4870, doi:10.1109/JIOT.2018.2853660, 2018.

Myers, L. and Sirois, M. J.: Spearman Correlation Coefficients, Differences between, in Encyclopedia of Statistical Sciences, edited by S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, and N. L. Johnson, p. ess5050.pub2, John Wiley & Sons, Inc., Hoboken, NJ, USA., 2006.

Papapostolou, V., Zhang, H., Feenstra, B. J. and Polidori, A.: Development of an environmental chamber for evaluating the performance of low-cost air quality sensors under controlled conditions, Atmos. Environ., 171, 82–90, doi:10.1016/j.atmosenv.2017.10.003, 2017.

Patashnick, H. and Rupprecht, E. G.: Continuous PM-10 Measurements Using the Tapered Element Oscillating Microbalance, J. Air Waste Manag. Assoc., 41(8), 1079–1083, doi:10.1080/10473289.1991.10466903, 1991.

RStudio: Why Use Keras?, [online] Available from: https://keras.rstudio.com/articles/why_use_keras.html, 2018.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning representations by back-propagating errors, Nature, 323(6088), 533–536, doi:10.1038/323533a0, 1986.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric chemistry and physics: from air pollution to climate change, Wiley, New York., 1998.

Si, M., Tarnoczi, T. J., Wiens, B. M. and Du, K.: Development of Predictive Emissions Monitoring System Using Open Source Machine Learning Library – Keras: A Case Study on a Cogeneration Unit, IEEE Access, 7, 113463–113475, doi:10.1109/ACCESS.2019.2930555, 2019.

de Smith, M.: Statistical Analysis Handbook, 2018 Edition., The Winchelsea Press, Drumlin Security Ltd, Edinburgh. [online] Available from: http://www.statsref.com/HTML/index.html?fligner-killeen_test.html, 2018.

Snyder, E. G., Watkins, T. H., Solomon, P. A., Thoma, E. D., Williams, R. W., Hagler, G. S. W., Shelow, D., Hindin, D. A., Kilaru, V. J. and Preuss, P. W.: The Changing Paradigm of Air Pollution Monitoring, Environ. Sci. Technol., 47(20), 11369–11377, doi:10.1021/es4022602, 2013.

Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M. and Bonavitacola, F.: Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide, Sens. Actuators B Chem., 215, 249–257, doi:10.1016/j.snb.2015.03.031, 2015.

Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M. and Bonavitacola, F.: Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2, Sens. Actuators B Chem., 238, 706–715, doi:10.1016/j.snb.2016.07.036, 2017.

Timbers, F.: Random Search for Hyper-Parameter Optimization | Finbarr Timbers, [online] Available from: https://finbarr.ca/random-search-hyper-parameter-optimization/ (Accessed 4 October 2019), 2017.

US EPA: LIST OF DESIGNATED REFERENCE AND EQUIVALENT METHODS, [online] Available from: https://www3.epa.gov/ttnamti1/files/ambient/criteria/AMTIC%20List%20Dec%202016-2.pdf (Accessed 7 October 2019), 2016.

Wang, Y., Li, J., Jing, H., Zhang, Q., Jiang, J. and Biswas, P.: Laboratory Evaluation and Calibration of Three Low-Cost Particle Sensors for Particulate Matter Measurement, Aerosol Sci. Technol., 49(11), 1063–1077, doi:10.1080/02786826.2015.1100710, 2015.

White, R., Paprotny, I., Doering, F., Cascio, W., Solomon, P. and Gundel, L.: Sensors and "apps" for community-based: Atmospheric monitoring, EM Air Waste Manag. Assoc. Mag. Environ. Manag., 36–40, 2012.

Williams, R., Kaufman, A., Hanley, T., Rice, J. and Garvey, S.: Evaluation of Field-deployed Low Cost PM Sensors, U.S. Environmental Protection Agency,. [online] Available from: https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&DirEntryId=297517 (Accessed 17 September 2019), 2014.

de Winter, J. C. F., Gosling, S. D. and Potter, J.: Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data., Psychol. Methods, 21(3), 273–290, doi:10.1037/met0000079, 2016.

XGBoost developers: XGBoost Parameters — xgboost 1.0.0-SNAPSHOT documentation, [online] Available from: https://xgboost.readthedocs.io/en/latest/parameter.html (Accessed 24 January 2020), 2019.

Xiong, Y., Zhou, J., Schauer, J. J., Yu, W. and Hu, Y.: Seasonal and spatial differences in source contributions to PM2.5 in Wuhan, China, Sci. Total Environ., 577, 155–165, doi:10.1016/j.scitotenv.2016.10.150, 2017.

Zheng, A.: Evaluating Machine Learning Models, First Edition., O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA., 2015.

Zheng, T., Bergin, M. H., Johnson, K. K., Tripathi, S. N., Shirodkar, S., Landis, M. S., Sutaria, R. and Carlson, D. E.: Field evaluation of low-cost particulate matter sensors in high- and low-concentration environments, Atmospheric Meas. Tech., 11(8), 4823–4846, doi:10.5194/amt-11-4823-2018, 2018.

Zikova, N., Hopke, P. K. and Ferro, A. R.: Evaluation of new low-cost particle monitors for PM2.5 concentrations measurements, J. Aerosol Sci., 105, 24–34, doi:10.1016/j.jaerosci.2016.11.010, 2017.

Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L. and R. Subramanian: A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring, Atmospheric Meas. Tech., 11(1), 291–313, doi:10.5194/amt-11-291-2018, 2018.

~~Bergstra, J. and Bengio, Y.: Random Search for Hyper-Parameter Optimization, J. Mach. Learn. Res., 13, 281–305, 2012.~~

~~CDNova Instrument Ltd.: SHARP Cost Estimate, 2017.~~

~~Charlson, R. J., Schwartz, S. E., Hales, J. M., Cess, R. D., Coakley, J. A., Hansen, J. E. and Hofmann, D. J.: Climate Forcing by Anthropogenic Aerosols, Science, 255(5043), 423–430, doi:10.1126/science.255.5043.423, 1992.~~

~~Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, pp. 785–794, ACM Press, San Francisco, California, USA., 2016.~~

~~Chong, C.-Y. and Kumar, S. P.: Sensor networks: Evolution, opportunities, and challenges, Proc. IEEE, 91(8), 1247–1256, doi:10.1109/JPROC.2003.814918, 2003.~~

~~Chow, J. C. and Watson, J. G.: Guideline on Speciated Particulate Monitoring, [online] Available from: https://www3.epa.gov/ttn/amtic/files/ambient/pm25/spec/drispec.pdf, 1998.~~

~~Conover, W. J., Johnson, M. E. and Johnson, M. M.: A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data., 1981.~~

~~Cordero, J. M., Borge, R. and Narros, A.: Using statistical methods to carry out in field calibrations of low cost air quality sensors, Sens. Actuators B Chem., 267, 245–254, doi:10.1016/j.snb.2018.04.021, 2018.~~

~~De Vito, S., Massera, E., Piga, M., Martinotto, L. and Di Francia, G.: On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sens. Actuators B Chem., 129(2), 750–757, doi:10.1016/j.snb.2007.09.060, 2008.~~

~~De Vito, S., Piga, M., Martinotto, L. and Di Francia, G.: CO, NO2, and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization, Sens. Actuators B Chem., 143(1), 182–191, doi:10.1016/j.snb.2009.08.041, 2009.~~

~~De Vito, S., Esposito, E., Salvato, M., Popoola, O., Formisano, F., Jones, R. and Di Francia, G.: Calibrating chemical multisensory devices for real world applications: An in-depth comparison of quantitative machine learning approaches, Sens. Actuators B Chem., 255, 1191–1210, doi:10.1016/j.snb.2017.07.155, 2018.~~

~~DeCicco, L.: Exploring ggplot2 boxplots - Defining limits and adjusting style, Explor. Ggplot2 Boxplots - Defin. Limits Adjust. Style [online] Available from: https://owi.usgs.gov/blog/boxplots/, 2016.~~

~~Esposito, E., De Vito, S., Salvato, M., Bright, V., Jones, R. L. and Popoola, O.: Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems, Sens. Actuators B Chem., 231, 701–713, doi:10.1016/j.snb.2016.03.038, 2016.~~

~~Fligner, M. A. and Killeen, T. J.: Distribution-Free Two-Sample Tests for Scale, J. Am. Stat. Assoc., 71(353), 210–213, doi:10.1080/01621459.1976.10481517, 1976.~~

~~Government of Canada: National Air Pollution Surveillance (NAPS) Network - Open Government Portal, Natl. Air Pollut. Surveill. NAPS Netw. [online] Available from: https://open.canada.ca/data/en/dataset/1b36a356-defd-4813-acea-47bc3abd859b (Accessed 17 September 2019), 2019.~~

~~Holstius, D. M., Pillarisetti, A., Smith, K. R. and Seto, E.: Field calibrations of a low-cost aerosol sensor at a regulatory monitoring site in California, Atmospheric Meas. Tech., 7(4), 1121–1131, doi:10.5194/amt-7-1121-2014, 2014.~~

~~Jayaratne, R., Liu, X., Thai, P., Dunbabin, M. and Morawska, L.: The influence of humidity on the performance of a low-cost air particle mass sensor and the effect of atmospheric fog, Atmospheric Meas. Tech., 11(8), 4883–4890, doi:10.5194/amt-11-4883-2018, 2018.~~

~~Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet classification with deep convolutional neural networks, Commun ACM, 60(6), 84–90, 2017.~~

~~Kumar, P., Morawska, L., Martani, C., Biskos, G., Neophytou, M., Di Sabatino, S., Bell, M., Norford, L. and Britter, R.: The rise of low-cost sensing for managing air pollution in cities, Environ. Int., 75, 199–205, doi:10.1016/j.envint.2014.11.019, 2015.~~

~~LeCun, Y., Bengio, Y. and Hinton, G.: Deep learning, Nature, 521(7553), 436–444, doi:10.1038/nature14539, 2015.~~

~~Lewis, A. C., Lee, J. D., Edwards, P. M., Shaw, M. D., Evans, M. J., Moller, S. J., Smith, K. R., Buckley, J. W., Ellis, M., Gillot, S. R. and White, A.: Evaluating the performance of low cost chemical sensors for air pollution research, Faraday Discuss., 189, 85–103, doi:10.1039/C5FD00201J, 2016.~~

~~Lin, Y., Dong, W. and Chen, Y.: Calibrating Low-Cost Sensors by a Two-Phase Learning Approach for Urban Air Quality Measurement, Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 2(1), 1–18, doi:10.1145/3191750, 2018.~~

~~Loh, B. G. and Choi, G.-H.: Calibration of Portable Particulate Matter–Monitoring Device using Web Query and Machine Learning, Saf. Health Work, S2093791119302811, doi:10.1016/j.shaw.2019.08.002, 2019.~~

~~Maag, B., Zhou, Z. and Thiele, L.: A Survey on Sensor Calibration in Air Pollution Monitoring Deployments, IEEE Internet Things J., 5(6), 4857–4870, doi:10.1109/JIOT.2018.2853660, 2018.~~

~~Papapostolou, V., Zhang, H., Feenstra, B. J. and Polidori, A.: Development of an environmental chamber for evaluating the performance of low-cost air quality sensors under controlled conditions, Atmos. Environ., 171, 82–90, doi:10.1016/j.atmosenv.2017.10.003, 2017.~~

~~Patashnick, H. and Rupprecht, E. G.: Continuous PM-10 Measurements Using the Tapered Element Oscillating Microbalance, J. Air Waste Manag. Assoc., 41(8), 1079–1083, doi:10.1080/10473289.1991.10466903, 1991.~~

~~RStudio: Why Use Keras?, [online] Available from: https://keras.rstudio.com/articles/why_use_keras.html, 2018.~~

~~Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning representations by back-propagating errors, Nature, 323(6088), 533–536, doi:10.1038/323533a0, 1986.~~

~~Seinfeld, J. H. and Pandis, S. N.: Atmospheric chemistry and physics: from air pollution to climate change, Wiley, New York., 1998.~~

~~Si, M., Tarnoczi, T. J., Wiens, B. M. and Du, K.: Development of Predictive Emissions Monitoring System Using Open Source Machine Learning Library – Keras: A Case Study on a Cogeneration Unit, IEEE Access, 7, 113463–113475, doi:10.1109/ACCESS.2019.2930555, 2019.~~

~~de Smith, M.: Statistical Analysis Handbook, 2018 Edition., The Winchelsea Press, Drumlin Security Ltd, Edinburgh. [online] Available from: http://www.statsref.com/HTML/index.html?fligner-killeen_test.html, 2018.~~

~~Snyder, E. G., Watkins, T. H., Solomon, P. A., Thoma, E. D., Williams, R. W., Hagler, G. S. W., Shelow, D., Hindin, D. A., Kilaru, V. J. and Preuss, P. W.: The Changing Paradigm of Air Pollution Monitoring, Environ. Sci. Technol., 47(20), 11369–11377, doi:10.1021/es4022602, 2013.~~

~~Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M. and Bonavitacola, F.: Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide, Sens. Actuators B Chem., 215, 249–257, doi:10.1016/j.snb.2015.03.031, 2015.~~

~~Spinelle, L., Gerboles, M., Villani, M. G., Aleixandre, M. and Bonavitacola, F.: Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO, and CO2, Sens. Actuators B Chem., 238, 706–715, doi:10.1016/j.snb.2016.07.036, 2017.~~

~~Timbers, F.: Random Search for Hyper-Parameter Optimization | Finbarr Timbers, [online] Available from: https://finbarr.ca/random-search-hyper-parameter-optimization/ (Accessed 4 October 2019), 2017.~~

~~US EPA: LIST OF DESIGNATED REFERENCE AND EQUIVALENT METHODS, [online] Available from: https://www3.epa.gov/ttnamti1/files/ambient/criteria/AMTIC%20List%20Dec%202016-2.pdf (Accessed 7 October 2019), 2016.~~

~~Wang, Y., Li, J., Jing, H., Zhang, Q., Jiang, J. and Biswas, P.: Laboratory Evaluation and Calibration of Three Low-Cost Particle Sensors for Particulate Matter Measurement, Aerosol Sci. Technol., 49(11), 1063–1077, doi:10.1080/02786826.2015.1100710, 2015.~~

~~White, R., Paprotny, I., Doering, F., Cascio, W., Solomon, P. and Gundel, L.: Sensors and "apps" for community-based: Atmospheric monitoring, EM Air Waste Manag. Assoc. Mag. Environ. Manag., 36–40, 2012.~~

~~Williams, R., Kaufman, A., Hanley, T., Rice, J. and Garvey, S.: Evaluation of Field-deployed Low Cost PM Sensors, U.S. Environmental Protection Agency. [online] Available from: https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&DirEntryId=297517 (Accessed 17 September 2019), 2014.~~

~~Xiong, Y., Zhou, J., Schauer, J. J., Yu, W. and Hu, Y.: Seasonal and spatial differences in source contributions to PM2.5 in Wuhan, China, Sci. Total Environ., 577, 155–165, doi:10.1016/j.scitotenv.2016.10.150, 2017.~~

~~Zheng, T., Bergin, M. H., Johnson, K. K., Tripathi, S. N., Shirodkar, S., Landis, M. S., Sutaria, R. and Carlson, D. E.: Field evaluation of low-cost particulate matter sensors in high- and low-concentration environments, Atmospheric Meas. Tech., 11(8), 4823–4846, doi:10.5194/amt-11-4823-2018, 2018.~~

~~Zikova, N., Hopke, P. K. and Ferro, A. R.: Evaluation of new low-cost particle monitors for PM2.5 concentrations measurements, J. Aerosol Sci., 105, 24–34, doi:10.1016/j.jaerosci.2016.11.010, 2017.~~

~~Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L. and R. Subramanian: A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring, Atmospheric Meas. Tech., 11(1), 291–313, doi:10.5194/amt-11-291-2018, 2018.~~