https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

Atmospheric Measurement Techniques Discussions

Research of Low-cost Air Quality Monitoring Models with Different Machine Learning Algorithms

Gang Wang[1, 2, 3], Chunlai Yu[1, 3], Kai Guo[2], Haisong Guo[1, 3], Yibo Wang[2]
[1]Huanghe Science and Technology College, Zhengzhou 450063, China
[2]Hanwei Electronics Group Corporation, Zhengzhou 450001, China
[3]Zhengzhou Key Laboratory of Intelligent Measurement Techniques and Applications, Zhengzhou, 450063, China
Correspondence to: Gang Wang (wywanggang163@163.com)

Abstract. To improve the prediction for the future air quality trends, the demand for low-cost sensor-based air quality gird monitoring is growing gradually. In this study, a low-cost multi-parameter air quality monitoring system (LCS) based on different machine learning algorithm is proposed. The LCS can measure particulate matter ($PM_{2.5}$ and PM_{10}) and gas pollutants (SO_2, NO_2, CO and O_3) simultaneously. The multi-dimensional multi-response prediction model is developed based on the original signals of the sensors, ambient temperature (*T*) and relative humidity (*RH*), and the measurements of the reference instrumentations. The performance of the different algorithms (RF, MLR, KNN, BP, GA-BP) with the parameters such as determination coefficient R^2 and Root Mean Square Error (RMSE) are compared and discussed. Using these methods, the R^2 of the algorithms (RF, MLR, KNN, BP, GA-BP) for the PM is in the range 0.68 - 0.99; the mean RMSE values of $PM_{2.5}$ and PM_{10} are within 3.96 - 16.16 μgm^{-3} and 7.37 - 28.90 μgm^{-3}, respectively. The R^2 of the algorithms (RF, MLR, KNN, BP, GA-BP) for the gas pollutants (O_3, CO and NO_2) is within 0.70 – 0.99; the mean RMSE values for these pollutants are 4.06 - 16.07 μgm^{-3}, 0.04 - 0.15 mgm^{-3}, 3.25 - 13.90 μgm^{-3}, respectively. The R^2 of the algorithms (RF, KNN, BP, GA-BP, except for MLR) for SO_2 is within 0.27 - 0.97, and the mean RMSE value is in the range 1.05 - 3.22 μgm^{-3}. These measurements are consistent with the national environmental protection standard requirement of China, and the LCS based on the machine learning algorithms can be used to predict the concentrations of PM and gas pollution.

1 Introduction

The development along with increased population and urbanization brings disadvantages, such as decreasing air quality and impact on public and individual health (Ioannis et al., 2020; Khreis et al., 2022; Singh et al., 2021). Among the atmospheric pollutants, the primary pollutant is fine particulate matter, which affects the respiratory system and cardiac activity of humans. The secondary pollutants are SO_2, CO, NO_x, and O_3, which also induce disease or chronic poisoning. To improve the understanding of air pollution exposure and to predict future air quality trends(Zimmerman et al., 2018), air quality assessment and forecasting are the essentials. The conventional air quality monitoring instrumentations are high cost, which has limited the spatial coverage of the monitoring stations(Zimmerman et al., 2018). The development and applications of the low-cost commercially available sensor-based air quality monitoring system (LCS) would considerably reduce both installation and maintenance costs (Spinelle et al., 2017). The larger spatial density of the air quality grid monitoring network becomes possible, which would plays an important role in monitoring pollution trend, locating of pollution source, supporting environmental management(Zhao et al., 2019) and support better epidemiological models(Khreis et al., 2022; Zimmerman et al., 2018). These demands promote the LCS growing gradually(Cui et al., 2021; Wang et al., 2016).

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

The LCS typically utilizes the electrochemical, metal oxide or light scattering sensors for gas-phase or particulate pollutants measurement, such as sulfur dioxide (SO_2), nitrogen oxide (NO_2), carbon monoxide (CO), ozone (O_3), total volatile organic compounds (VOCs), and particulate matters (PM). These electrochemical and metal oxide sensors have intrinsic problems such as uneven quality, signal drift, temperature and humidity impacts, and gaseous cross-sensitivities (Spinelle et al., 2015, 2017; Zimmerman et al., 2018) (Brilli et al., 2020; Guo et al., 2020; Jiao et al., 2016; Magi et al., 2020). For example, limited by the poor selection performance, the NO_2 electrochemical sensor also undergo redox reactions in the presence of O_3 gaseous pollutants. The diffusion coefficient of the electrochemical sensor can be affected by temperature and relative humidity(Hitchman et al., 1997; Masson et al., 2015). The reagent of the electrochemical sensor is consumed over time, which affects the stability of the sensor. These features of the sensors have historically been poorly addressed by laboratory calibrations, limiting the utility for air quality monitoring (Zimmerman et al., 2018).

The de-convolving of cross-sensitivity effect and stability on sensor performance is complex, and this has been increasing interest in multifarious algorithms for low-cost sensor calibration. The linear or multivariate linear calibration models (Alexopoulos, 2010; Khreis et al., 2022; Zoest et al., 2019), have been developed and studied to overcome these sensors weaknesses. The R^2 performance when measuring NO_2 and $PM_{2.5}$ is less than 0.58(Khreis et al., 2022) with mixed results after the calibration and generally deteriorating performance. The accurate and precise calibration of low cost gas sensors still represents a challenge in ambient air quality monitoring. High-dimensional multi-response calibration models(Alexopoulos, 2010) are built for CO, NO, NO_2, and O_3, with the determination coefficient R^2 within 0.39–0.88 between the models result and the reference monitors results. The artificial neural network (ANN) calibration model has the intelligence to process nonlinear data with self-learning and self-memory(Spinelle et al., 2015), which does not need an accurate mathematical model and has been widely utilized in data analysis to meet computational intelligence requirements(Amuthadevi et al., 2021; Janabi et al., 2021). The ANN has been used in calibration models for measuring ozone or nitrogen oxide(Esposito et al., 2016; Spinelle et al., 2015). For example, the ANN calibration model was used to calibrate O_3 and the uncertainty could meet the European data quality objectives; however, meeting these objectives for NO_2 remains a challenge. Dynamic neural network calibrations of NO_2 sensors was demonstrated with the mean absolute error less than 2 ppb; however, the same performance for O_3 was not observed. Furthermore, the model for calibration performance for measuring NO, NO_2, and O_3 was tested on 4 weeks of data. Random-forest-based machine learning algorithm was also used to improve the calibration strategies of low-cost sensors(Zimmerman et al., 2018). The average mean absolute error on the testing data set from the random forest models was 38 ppb for CO (14% relative error), 10 ppm for CO_2 (2 % relative error), 3.5 ppb for NO_2 (29 % relative error), and 3.4 ppb for O_3 (15 % relative error), and Pearson r versus the reference monitors exceeded 0.8 for most units. Multiple linear regression(Ionascu et al., 2021) based temperature and humidity correction and ANN based calibration shown the potential for significant further improvement for leave one out cross validation(Ali et al., 2021). An integrated genetic programming dynamic neural network model was used to accurately estimate the carbon monoxide and nitrogen dioxide pollutant concentrations from the multi-sensor measurement data(Davut et al., 2022). However, these calibrations have only been tested on a short measurement period and small number of sensor matrix, each containing one sensor per pollutant (Cross et al., 2017; Esposito et al., 2016; Spinelle et al., 2015), not have been utilized to evaluate and predict the concentration values of multi pollutants simultaneously, such as $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3.

In this work, the LCS is developed to measure $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3 simultaneously. Taking the original electronic signals of the sensors as input and measurements obtained by the reference instrumentations as output, five calibration strategies are applied and contrasted. The calibration model used are multivariate linear regression (MLR)(Alexopoulos, 2010), genetic-algorithm-back-propagation neural (GA-BP) network (Ning et al., 2019; Wang et al., 2019), BP neural network(Xu et al., 2021), K Nearest Neighbor (KNN)(Zhao et al., 2021) and random-forest (RF)(Breiman, 2001; Liu et al., 2012) based machine learning

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

algorithm. The measurement is implemented in the real-world conditions almost a 12-month period (1 March 2021 and 28 February 2022) spanning multiple seasons and a wide range of meteorological conditions to ensure calibration model robustness. The performance of the different algorithms with the parameters such as R^2 and Root Mean Square Error (RMSE) are compared and discussed.

The rest of this paper is organized as follows. The measurement setup is described in section 2. The principles of the calibration strategies are presented in section 3. The results and discussion are shown in section 4. The conclusion is drawn in section 5.

2 Measurement setup

This section describes the measurement site and data collection, schematic block of the LCS, and the reference instrumentation. The low-cost here is defined as below 150 dollars per pollutant, commercial availability and low maintenance. The sensors typically utilize electro-chemical signal and scattering light intensity for gas-phase pollutants (SO_2, NO_2, CO and O_3) and particle pollutants ($PM_{2.5}$, PM_{10}) measurement.

2.1 Measurement site and data collection

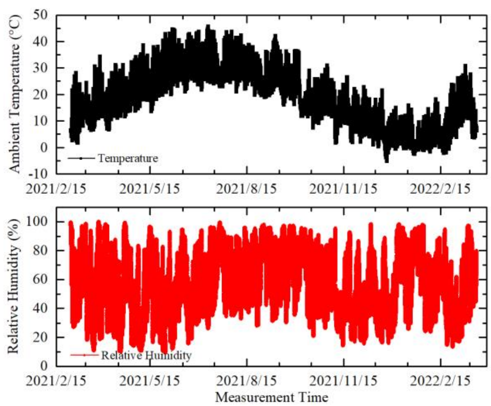

Figure 1. Temperature and relative humidity ranges during the measurement period (1 March 2021 and 28 February 2022)

Measurements for $PM_{2.5}$, PM_{10}, CO, SO_2, NO_2 and O_3 were made continuously between 1 March 2021 and 28 February 2022, which were used as the start and end dates for the analyses. The measurement was made on the 30 Yaochang Streat, Zhongyuan District, Zhengzhou City, Henan Province, where there was an independent reference monitoring system for $PM_{2.5}$, PM_{10}, CO, SO_2, NO_2 and O_3 measurement. The LCS was mounted at a consistent height with the reference monitoring system. The data collection interval of the LCS and reference instruments was 5 minutes. During the measurement period, the ranges of the ambient temperature and relative humidity separately were -5℃ to 50℃ and 10% to 98%, shown in Figure 1. The ambient temperature increased, decreased and fluctuated separately within 1 March 2021 and 30 June 2021, 1 July 2021 and 31 October 2021, 1 November 2021 and 28 February 2022, dividing the whole measurement period into three segments.

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

Atmospheric Measurement Techniques Discussions

2.2 Schematic block of LCS

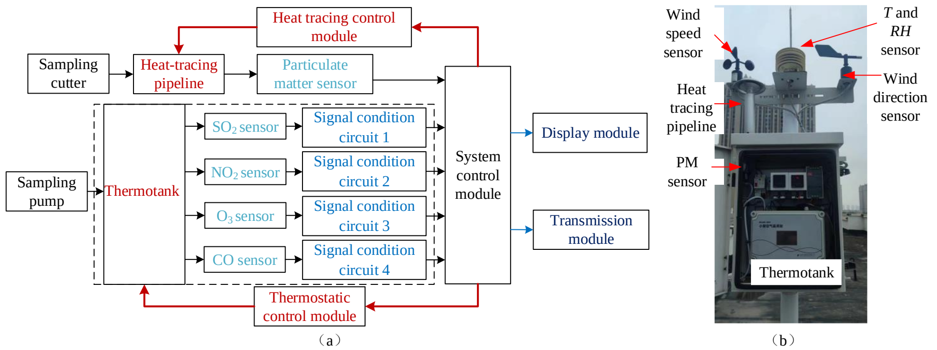

Figure 2. Schematic block and site photo of the LCS. The left panel (a) is the schematic block of the LCS. The system control module can ensure the temperature stability of the heat tracing pipeline and thermo-tank through the heat tracing control module and thermostatic control module, respectively. The sampling cutter is used to filter particles larger than 10 µm. The sampling pump is utilized to deliver ambient air to the surface of the sensors. The right panel (b) is the site photo of the LCS.

In this study, the LCS is developed by Hanwei Electronics Group Corporation, and its schematic block diagram is shown in Figure 2. The LCS uses the commercially available particulate matter sensor and electrochemical sensors from Cubic Ltd and Alphasense Ltd, respectively. The particulate matter sensor device is the laser diode (LD) based particle sensor, using a spectrophotometer to measure the particle scattering light intensity. The PM sensor device (PM3006) can measure size dependent $PM_{2.5}$ and PM_{10} concentration of the particles in the size range of 0.3 to 10 µm. The gas pollution (SO_2\$NO_2$\$O_3$\CO) sensor used are with 4 electrodes (i.e. reference, worker, counter and auxiliary electrodes), where the auxiliary electrode is not exposed to the target analyte to account for changes in the sensor baseline signal under different meteorological conditions(Mead et al., 2013).
The electrochemical sensor outputs are measured using electronic circuitry designed by Hanwei and optimized for signal stability. The circuitry is developed with custom electronics to drive the device, multiple stages of filtering circuitry for specific noise signatures, and an analog-to-digital converter for measurement of the conditioned signal.
Due to the redox reaction on the anode and the cathode of the electrochemical sensor, the movement of charge between the electrodes produces a current proportional to the analyte reaction rate, which can be used to determine the analyte concentration(Mead et al., 2013) and the sensor whether working effectively. The linearity of the gas sensors of SO_2, NO_2, CO and O_3 is examined in laboratory to evaluate the performance of the sensors before used in real-world conditions.
In laboratory, the sensors were tested under steadily increased different single standard gas concentration, which was from 0 - 5 mgm^{-3} for CO sensor, 0 - 0.2 mgm^{-3} for NO_2, 0 - 1.1 mgm^{-3} for O_3 and 0 - 1.4 mgm^{-3} for SO_2. The electrical output signals of the gas pollution sensors, proportional to the concentrations of the single standard substances, shown in Figure 3, verified the sensor working properly and effectively and could be applied to the LCS.

Atmospheric Measurement Techniques Discussions

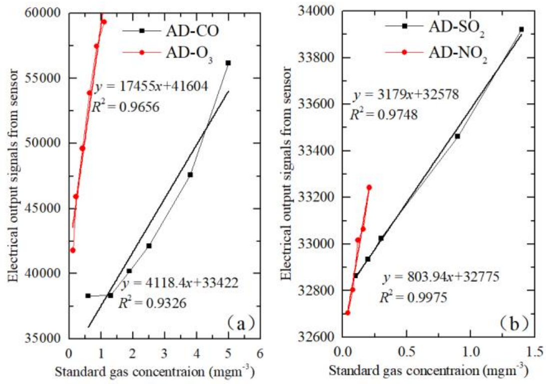

Figure 3. Electrical output signals versus single standard gas concentration in laboratory condition. The left panel (a) and right panel (b) represent the proportional relations between CO and O_3 sensors, SO_2 and NO_2 sensors, respectively. The duration of each measurement is about 30 minutes.

However, even with an auxiliary electrode, electrochemical sensors may insufficiently account for the impacts of temperature and relative humidity. With the purpose to eliminate the influence of the external environment on the sensor as much as possible, the particles flow through a sampling cutter and heat-tracing pipeline to the particulate matter sensor, and the gaseous pollutants are pumped to the electrochemical sensors, which are secured in a thermo-tank. The temperature values of the heat-tracing pipeline and thermo-tank can be maintained at 60 ℃±2 ℃ to reduce the influence of relative humidity and 25 ℃±2 ℃(Wei et al., 2018) to keep the sensor operating at a stable temperature, respectively.

The measurement results of particulate matter sensor and gas pollution sensors, transmitted to the system control module through the data buses, are directly displayed on the local display module and wirelessly transmitted to the corresponding online server through the transmission module. As the uni-variate linear models does not in-corporate any cross-sensitivities to other pollutants or any nonlinearities in the response, we attempt to using the sensor electronic results as the input and the reference measurements as the output, to build multi-dimensional multi-response prediction models to de-convolve the effects of cross-sensitivity and stability on sensor performance utilizing MLR, RF, KNN, BP and GA-BP calibration models.

2.3 Reference instrumentation

In order to reduce the adsorption effect on particle matter and gaseous pollutants, the reference measurements are made on ambient air continuously drawn through Teflon fluorinated ethylene propylene (FEP)(Wei et al., 2018) tubing with a six-port stainless steel manifold for flow distribution to the gas analyzers and particulate monitors(Mead et al., 2013). It should be pointed out that the LCS was mounted at a consistent height with the reference monitoring system during the measurement period.

The reference ambient particulate monitor 5014*i*, which uses beta attenuation of the ambient particulate deposited onto a filter tape, is applied to measure the mass concentration of suspended and refined particulates. The reference NO-NO_2-NO_X monitor 42*i*, using the linear proportional of the chemi-luminescence reaction of NO and O_3 after NO_2 transformed into NO, is utilized to measure the NO_2 concentration. The SO_2 reference analyzer is 43*i* using the ultraviolet light (which is emitted as the excited SO_2 molecules decay to lower energy states) intensity proportional to the SO_2 concentration. The CO reference monitor is 48*i* utilizing the principle that CO absorbs infrared radiation at a wavelength of 4.6 μm and the infrared absorption can be transformed to be proportional to the CO concentration. The 49*i* O_3 analyzer operates on the principle that O_3 molecules absorb UV light at a wavelength of 254 nm, and the absorption intensity of the UV light is directly related to the ozone concentration. All these reference monitors are produced

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

by Thermo Fisher Scientific Inc. The time interval for all reference measurements is 5 minutes. The reference gas and particulate analyzers are checked and calibrated weekly using calibration gas mixtures to correct the baseline drift.

3 Principles

This section describes the principles of the calibration methods, such as MLR, BP, GA-BP, KNN and RF, and the metrics for performance evaluation. The calibration models are constructed with the sensors (i.e., $PM_{2.5}$, PM_{10}, CO, SO_2, NO_2 and O_3 sensors.) electronic results as the input and the reference measurements as the output.

3.1 Calibration methods

3.1.1 Multiple linear regression model

After the data collected by the LCS, the raw data should be preprocessed. The PM3006 particulate matter sensor can measure particle counters in the size range of 0.3 to 10 μm. By using the particle counters $x_{0.5}$, $x_{1.0}$, $x_{2.5}$, $x_{5.0}$ and $x_{10.0}$ of the sensors, listed in Table 1, the measured particle number concentration is converted to PM mass concentrations in the $PM_{2.5}$ and PM_{10} size fractions.

Table 1. Size range of the particulate matter sensor. The sensor can measure particles with the size range of 0.3~0.5 μm, 0.5~1.0 μm, 1.0~2.5 μm, 2.5~5.0 μm and 5.0~10 μm, simultaneously. The corresponding particle counters are expressed as $x_{0.5}$, $x_{1.0}$, $x_{2.5}$, $x_{5.0}$ and $x_{10.0}$, respectively.

Range (μm)	0.3~0.5	0.5~1.0	1.0~2.5	2.5~5.0	5.0~10.0
Particle counter	$x_{0.5}$	$x_{1.0}$	$x_{2.5}$	$x_{5.0}$	$x_{10.0}$

Taking the particle counters, listed in Table 1, as input and the concentrations $Y_{pm2.5}$ and Y_{pm10} of $PM_{2.5}$ and PM_{10} measured by 5014*i* as output, the multivariate linear regression (MLR) models(Alexopoulos, 2010; Zoest et al., 2019) is built. Due to the previously established influence of ambient temperature (T) and relative humidity (RH) on sensor response(Jiao et al., 2016; Masson et al., 2015), the multi-dimensional multi-response preprocessing and prediction models can be written as Eq. (1).

$$\begin{cases} Y_{pm2.5}=w_{1\_pm2.5}\cdot x_{0.5}+w_{2\_pm2.5}\cdot x_{1.0}+w_{3\_pm2.5}\cdot x_{2.5}+w_{4\_pm2.5}\cdot T+w_{5\_pm2.5}\cdot RH+b_{pm2.5} \\ Y_{pm10}=w_{1\_pm10}\cdot x_{0.5}+w_{2\_pm10}\cdot x_{1.0}+w_{3\_pm10}\cdot x_{2.5}+w_{4\_pm10}\cdot x_{5.0}+w_{5\_pm10}\cdot x_{10.0}+w_{6\_pm10}\cdot T+w_{7\_pm10}\cdot RH+b_{pm10} \end{cases}, \quad (1)$$

The equation (1) can be simplified as,

$$\begin{cases} Y_{pm2.5}=W_{pm2.5}\cdot X_{pm2.5}+b_{pm2.5} \\ Y_{pm10}=W_{pm10}\cdot X_{pm10}+b_{pm10} \end{cases}, \quad (2)$$

Where $W_{pm2.5}$= [$w_{1\_pm2.5}$, $w_{2\_pm2.5}$, $w_{3\_pm2.5}$, $w_{4\_pm2.5}$, $w_{5\_pm2.5}$] and W_{pm10}= [w_{1_pm10}, w_{2_pm10}, w_{3_pm10}, w_{4_pm10}, w_{5_pm10}, w_{6_pm10}, w_{7_pm10}] are the corresponding weight coefficients; the $X'_{pm2.5}$ = [$x_{0.5}$, $x_{1.0}$, $x_{2.5}$, T, RH] and X'_{pm10} = [$x_{0.5}$, $x_{1.0}$, $x_{2.5}$, $x_{5.0}$, $x_{10.0}$, T, RH] represent the input particle counters and values obtained from the PM sensor, the temperature sensor and humidity sensor; the $b_{pm2.5}$ and b_{pm10} are the intercept values of the model.

Due to the poor selection performance and cross interference of the electro-chemical sensors response, the output values from each sensor using net sensor response to the target analyte, such as O_3, NO_2, SO_2, concentration measured by the inference monitor are used to build the MLR model. The CO gaseous pollution is also one of the criteria pollutants, which is must to be measured in China. Thus, the multi-dimensional multi-response preprocessing and prediction model for the 4 gas pollutions, T and RH can be written as Eq. (3).

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

Atmospheric Measurement Techniques
Discussions

$$\begin{cases} Y_{SO_2}= w_{11}\cdot x_{SO_2}+w_{12}\cdot x_{NO_2}+w_{13}\cdot x_{CO}+w_{14}\cdot x_{O_3}+w_{15}\cdot T+w_{16}\cdot RH+b_{SO_2} \\ Y_{NO_2}= w_{21}\cdot x_{SO_2}+w_{22}\cdot x_{NO_2}+w_{23}\cdot x_{CO}+w_{24}\cdot x_{O_3}+w_{25}\cdot T+w_{26}\cdot RH+b_{NO_2} \\ Y_{CO}= w_{31}\cdot x_{SO_2}+w_{32}\cdot x_{NO_2}+w_{33}\cdot x_{CO}+w_{34}\cdot x_{O_3}+w_{35}\cdot T+w_{36}\cdot RH+b_{CO} \\ Y_{O_3}= w_{41}\cdot x_{SO_2}+w_{42}\cdot x_{NO_2}+w_{43}\cdot x_{CO}+w_{44}\cdot x_{O_3}+w_{45}\cdot T+w_{46}\cdot RH+b_{O_3} \end{cases}, \tag{3}$$

The equation (3) can be simplified as,

$$Y_{[SO_2,\, NO_2,\, CO,\, O_3]}=W_{gas}\cdot X_{gas} + B_{gas}, \tag{4}$$

Where $W_{gas}=\begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} & w_{15} & w_{16} \\ w_{21} & w_{22} & w_{23} & w_{24} & w_{25} & w_{26} \\ w_{31} & w_{32} & w_{33} & w_{34} & w_{35} & w_{36} \\ w_{41} & w_{42} & w_{43} & w_{44} & w_{45} & w_{46} \end{bmatrix}$ is the corresponding weight coefficient; the X_{gas} =[x_{SO_2}, x_{NO_2}, x_{CO}, x_{O_3}, T, RH] is the convertor output values of the sensors through the electronic circuitries; the B_{gas}' = [b_{SO_2}, b_{NO_2}, b_{CO}, b_{O_3}] is the intercept value of the model.

Hereto, the multi MLR models for the gas sensor and PM sensor are separately developed. The training data is used to calculate the model regression coefficient and intercept values, and the withheld testing data is utilized to evaluate the performance of the model performance.

3.1.2 BP neural network model

The BP neural network algorithm is one of the most widely used ANN models. It is a multi-layer feed-forward network trained through an error back propagation algorithm by constantly adjusting the weight and intercept of the network. The feed-forward topological structure of the BP neural network model, shown in Figure 4, includes the input layer, hidden layer and output layer. With the purpose to avoid the numerical problems caused by the extreme values of polarization, eliminate the misleading effects for feature extraction and obtain the accurate estimation of pollutant concentrations(Janabi et al., 2021), the collected input sensor date X_I and output date Y_O should be respectively normalized with min-max normalization to limit values in each dimension between 0 and 1(Hande et al., 2021).

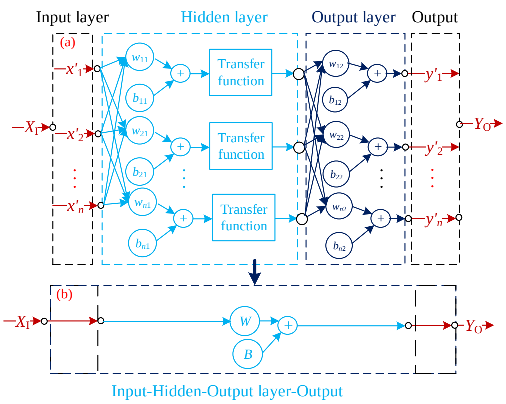

Figure 4. Topological structure of BP neural network model. The up panel (a) is the feed-forward topological structure. The X_I and Y_O are the input data and output data, respectively. The X'_i and Y'_i separately indicate the normalized items of X and Y. The w_{i1} and b_{i1}, w_{j2} and b_{j2} separately represent the weight value and intercept value of the hidden layer and output layer. The down panel (b) is equivalent to panel (a) to simplify the formulas.

After the normalization process, the BP network can be established. To optimize the best parameters of the network, the number of hidden layer, the transfer functions of the layers, and the end conditions should be determined. If the parameters are inappropriate, the BP-model will be over trained or insufficient. In this study, a shallow structure with a single hidden layer is chosen, as extensive

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

testing did not show any noticeable improvement in calibration performance with deeper structure consisting of multiple hidden layers(Ali et al., 2021). This also reduced the complexity and the training time.

3.1.3 Genetic algorithm-BP model

In the traditional BP neural network, the initial weights and thresholds are randomly generated. The results often fall into a local minimum rather than a global minimum, and would lead to the distortion of the prediction result. In addition, the convergence speed of the BP neural network is usually slow. To solve these problems, the genetic algorithm (GA) (Liang et al., 2018) with BP algorithm is also used to avoid the inherent defects of BP algorithm. The GA method is essentially a direct search method that does not rely on specific problems and gradient information. It follows the survival and elimination rule of biological evolution, generates the following hypotheses by mutating and reconstructing the best existed hypothesis and makes it possible to solve the problem(Ning et al., 2019). Generally, the GA is used to find an optimal initial weight and a threshold value for the model, so that the model could converge in the direction of minimum value(Wang et al., 2019). The GA-BP hybrid algorithm is used to reduce the time for the BP neural network to adjust the weight and threshold itself and achieve the goal of improving work efficiency.

3.1.4 K nearest neighbor model

The *k* nearest neighbor (KNN) is also one of the simplest method for classification as well as regression problem(Kumar, 2015; Zhao and Lai, 2021). The KNN is a supervised method that uses estimation based on values of neighbors, which can automatically adapt to the supervised learning problems with arbitrary Bayes decision boundaries(Zhao and Lai, 2021). From the supervisor dataset, the KNN solution utilizes the values of given dependent variable y_i to approximate the dependent variable y^*, which is closest with respect to distance between their corresponding model parameters. For regression problem, the mean of the observed labels of k nearest neighbors of independent variable $\boldsymbol{X}$ is assigned to be the predicted label. In this study, the k is set to 10 with the performance having no obvious difference from other numbers.

3.1.5 Random forest model

The random forest (RF) model is used for solving regression or classification problems(Breiman, 2001; Liu et al., 2012). It works by constructing an ensemble of decision trees using a training data set; the mean value from that ensemble of decision trees is then used to predict the value for new input data(Zimmerman et al., 2018). With the purpose to establish a RF model, the maximum number of decision trees of the forest should be specified. Each tree is constructed using a bootstrapped random sample from the training data set. By considering a random subset of the possible explanatory variables with the strongest predictor of the response, the origin node of the decision tree can be split into sub-nodes. The node splitting process is repeated until a terminal node is reached. The terminal node can be specified using the maximum number of sub-nodes or the minimum number of data points in the node. To illustrate the method, consider building a random forest model for one LCS using a single decision tree and a subset of 20490 data points to build a calibration model, shown in Figure 5. The RF model can predict data with variable parameters within the training range. Therefore, a larger and more variable training data set should create a better final model. To avoid missing any spikes during the training window, a 5-fold cross-validation approach(Zimmerman et al., 2018) is also used to maximize utilization of the training data set. This approach helps to minimize bias in training data selection when predicting new data and ensures that every point in the training window is used to build the model.

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

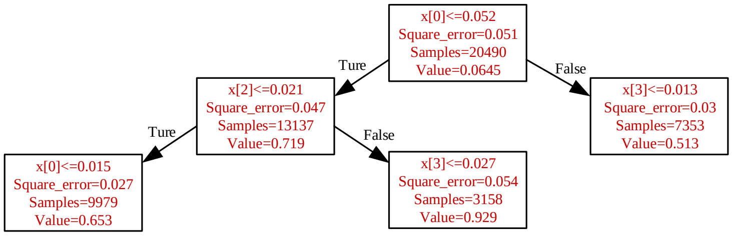

Figure 5. Simplified illustration of the RF with a single decision tree and a subset. The *x*[0], *x*[2], *x*[3] represent the CO, SO_2 and O_3 pollutants. At the first split, points with normalized CO sensor signal ≤0.052 are sent to a terminal node; the remaining points go to the other splitting node. The Samples is the number of data points in each terminal node. The Value is the average in each terminal node.

3.2 Metrics for performance evaluation

To quantitatively compare the MLR, BP, GA-BP, KNN and RF model output to the reference monitor concentrations, the determination coefficient R^2 and root mean square error (RMSE) (Janabi et al., 2021) are utilized. Where, the R^2 is obtained as Eq. (5).

$$R^2=1-\left[\sum_{i=1}^{n}(y_i-f_i)^2\right]/\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right], \quad (5)$$

Where y_i and $\bar{y}$ are the real-time data and mean data obtained by reference instrumentations, respectively. The f_i is the model output data according to the model algorithm. The R^2 reflects the fit degree between the model output data and the reference monitor measurement. The measurement results should meet the requirements of environmental standards of China(Jiao et al., 2016). The RMSE measures how much error there is between the predicted values and the reference measurements, which is calculated by the following Eq. (6).

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{i,\ p}-y_i)^2}, \quad (6)$$

Where $y_{i,p}$ and y_i represent the *i*th model output data form the algorithm-based LCS system and the reference data from the reference instrumentations, the *n* is the number of the measurement data in the dataset.

4 Result and discussion

Following the model building, the goodness of regression and root mean square error between the model output concentrations and the reference monitor concentrations are evaluated for all calibration model approaches. The plots for the $PM_{2.5}$, PM_{10}, O_3, CO, NO_2 and SO_2 illustrating the time series and goodness of fit of the models are provided in the Figure 9 - Figure 14. The R^2 and RMSE values are listed in Table 2 – Table 5.

4.1 Parameters of the model

For the BP and GA-BP models, the parameters are the functions for the hidden layer and output layer, the type of the hidden layer, the number of iteration times, and the number of the nerve units(Wang et al., 2013). The functions for the hidden layer and output layer in this study respectively are the default tansig and the purelin functions. With the more complex type of the hidden layer and less obvious improvement, the hidden layer is single type to achieve the goal of work efficiency.

To determine the best number of iteration times and nerve units, the measurement from the LCS and reference monitor between 1 March 2021 and 30 June 2021 is used. The number of iteration time is optimized using the mean squared error (MSE) between the model value from the model and the reference monitor output value. The tendency of the MSE is shown in Figure 6. The training is performed for 500 iterations. It is observed that the MSE decreases with the number of iteration time increasing, the rate of

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

decrease and the variation of the MSE is negligible beyond 100 iterations. More iterations incur higher computational cost for the training and small performance improvement. There is also the risk of overtraining resulting in poor generalization capability.

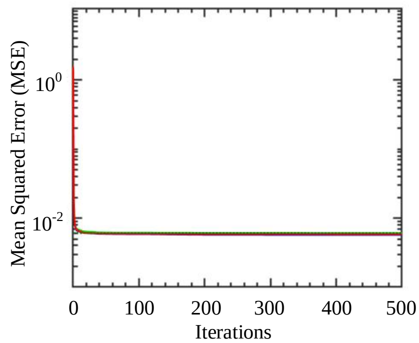

Figure 6. The MSE with the number of iterations.

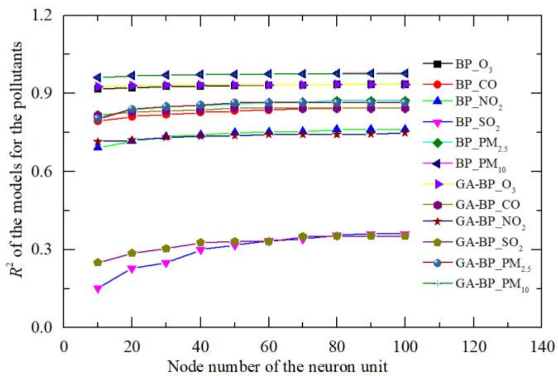

Figure 7. The R^2 with different node number of the neuron for the pollutants.

The node number of the nerve units is determined by the contrast results of determination coefficient R^2 for different gas and PM pollutants within 1 March 2021 and 30 June 2021. The results are shown in Figure 7. The R^2 is improved as the number of nerve units increasing. The rate of increase and the variation of R^2 is negligible beyond 70 units. More units incur higher computational cost and time for the training and small performance improvement.

For the RF model, the number of tree is determined by the contrast results of determination coefficient R^2 for different gas pollutants within 1 March 2021 and 30 June 2021. The results are shown in Figure 8. The R^2 is improved as the number of tree increasing. The rate of increase and the variation of R^2 is negligible beyond 20. The terminal node is specified using a maximum number of sub-node points per node. The R^2 is also improved as the number of sub-node increasing under the same tree number. The rate of increase and the variation of R^2 is negligible beyond 100. More number of the tree or the sub-node incur higher computational cost and time for the training and small performance improvement.

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

Atmospheric Measurement Techniques
Discussions

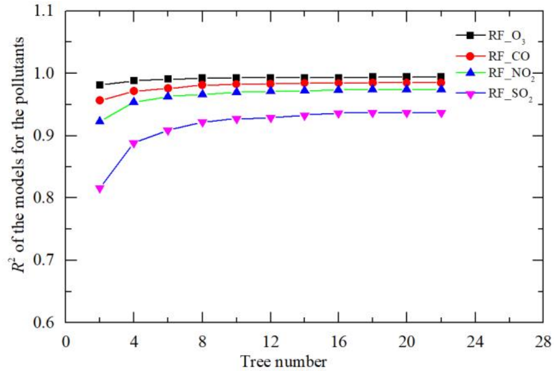

Figure 8. The R^2 with different tree number of the RF model for the pollutants.

4.2 Measurement results of PM

With the results from 1 March 2021 to 28 February 2022 and according the trend of the ambient temperature, shown in Figure 1, the whole data is divided into three segments. The three segments (I, II, and III) separately are within 1 March 2021 and 30 June 2021, 1 July 2021 and 31 October 2021, 1 November 2021 and 28 February 2022. The time series data and regressions of five modes for PM from reference monitor and LCS calibration output are shown in Figure 9 and Figure 10. With the purpose of avoiding over-fit in the five models, the randomly divide parameters of train ratio and test ratio are 80% and 20%, respectively.

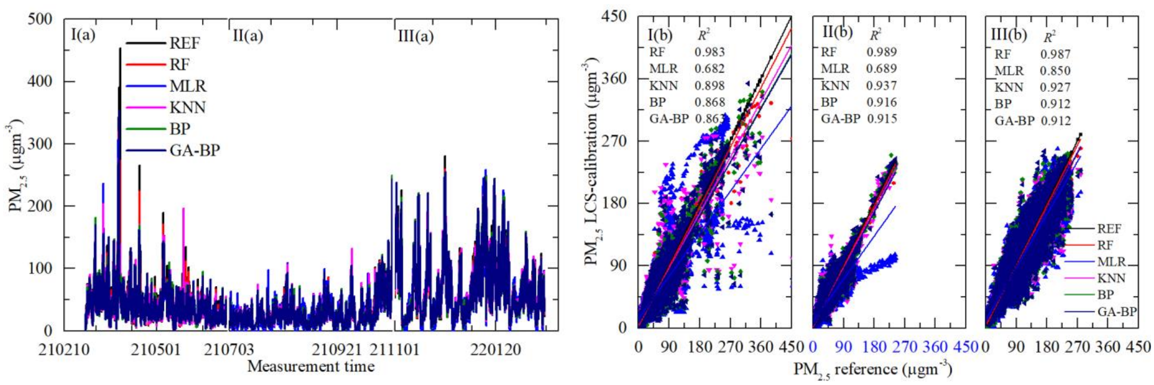

Figure 9. Time series and regressions comparing the reference monitor PM$_{2.5}$ data (black) to five calibration model PM$_{2.5}$ results. Where red, blue, magenta, olive and navy represent RF, MLR, KNN, BP, GA-BP, respectively. The left panel (a) shows the whole time series data of the measurement period. The right panel (b) shows the regressions of the five calibration models.

As shown in Figure 9 (a) and Figure 10 (a), the general tendency of the data fluctuation between the reference monitor and the RF, MLR, KNN, BP, GA-BP based algorithm of the LCS are consistent with each other. The best performance is RF model, the next are KNN, BP and GA-BP, the worst is MLR. The regressions consistence between the reference data and the five model data are also shown in Figure 9 (b) - Figure 10 (b), and listed in Table 2.

The R^2 of RF for the PM is better than 0.98. The R^2 of MLR for the PM is less than 0.91, and even less than 0.7. The R^2 of the other three model are within 0.86 and 0.98. The performance of different calibration models for the PM against reference monitor is also evaluated using RMSE, and the results are listed in Table 3.

Using the data listed in Table 3, the RMSE values from the first (I) and third (III) stages are large than the one from the second (II) stage, the main reason maybe the large fluctuation range of the PM for the climatic factors in winter and spring resulting in the poor model fit. The mean RMSE values of PM$_{2.5}$ between the reference data and the RF, MLR, KNN, BP, GA-BP-based algorithms

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

Atmospheric Measurement Techniques
Discussions

dada are calculated as 3.96 μgm^{-3}, 16.16 μgm^{-3}, 9.48 μgm^{-3}, 10.67 μgm^{-3} and 10.74 μgm^{-3}, respectively. The mean RMSE values of PM_{10} between the reference data and the RF, MLR, KNN, BP, GA-BP-based algorithms dada are calculated as 7.37 μgm^{-3}, 28.90 μgm^{-3}, 18.50 μgm^{-3}, 18.04 μgm^{-3} and 18.17 μgm^{-3}, respectively.

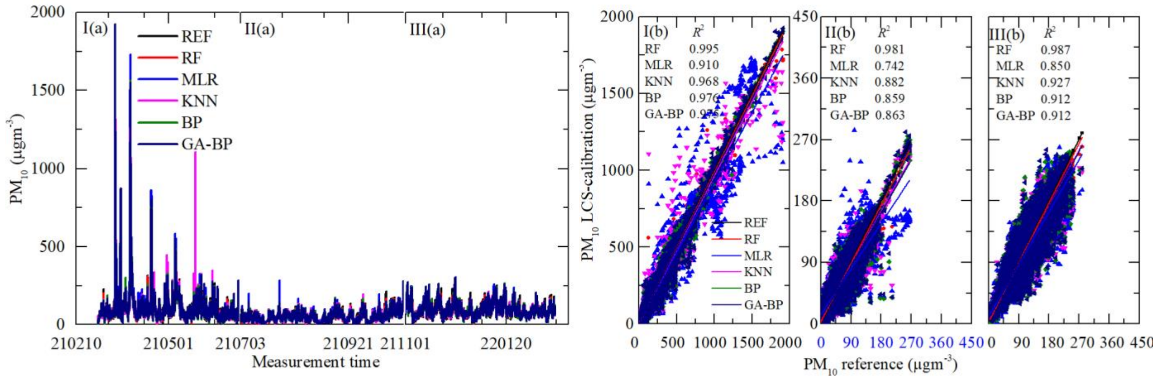

Figure 10. Time series and regressions comparing the reference monitor PM_{10} data (black) to five calibration model PM_{10} results. Where red, blue, magenta, olive and navy represent RF, MLR, KNN, BP, GA-BP, respectively. The left panel (a) shows the whole time series data of the measurement period. The right panel (b) shows the regressions of the five calibration models.

Table 2. Performance of different calibration models for the $PM_{2.5}$ and PM_{10} against reference monitor. The determination coefficient R^2 (higher is better, maximum of 1) of different calibration models (RF, MLR, KNN, BP, GA-BP) versus reference monitor.

R^2 / Model	$PM_{2.5}$			PM_{10}		
	I	II	III	I	II	III
RF	**0.983**	**0.989**	**0.987**	**0.995**	**0.981**	**0.987**
MLR	0.682	0.689	0.850	0.910	0.742	0.850
KNN	0.898	0.937	0.927	0.968	0.882	0.927
BP	0.868	0.916	0.912	0.976	0.859	0.912
GA-BP	0.863	0.915	0.912	0.975	0.863	0.912

Table 3. Performance of different calibration models for the $PM_{2.5}$ and PM_{10} against reference monitor. The RMSE errors (lower is better) of different calibration models (RF, MLR, KNN, BP, GA-BP) versus reference monitor.

RMSE / Model	$PM_{2.5}$ (μgm^{-3})				PM_{10} (μgm^{-3})			
	I	II	III	**MEAN**	I	II	III	**MEAN**
RF	4.03	2.36	5.49	**3.96**	10.37	4.55	7.19	**7.37**
MLR	17.18	12.63	18.68	**16.16**	45.05	16.43	25.22	**28.90**
KNN	9.73	5.67	13.05	**9.48**	27.08	11.14	17.29	**18.50**
BP	11.09	6.56	14.35	**10.67**	23.10	12.15	18.88	**18.04**
GA-BP	11.27	6.61	14.35	**10.74**	23.65	11.99	18.87	**18.17**

4.3 Measurement results of gas pollution

With the results from 1 March 2021 to 28 February 2022 and according the trend of the ambient temperature, shown in Figure 1, the whole data is also divided into three same segments as section 4.2. The time series data and regressions of five modes for gas pollution from reference monitor and LCS calibration output are shown in Figure 11 - Figure 14. With the same purpose of avoiding over-fit in the five models, the randomly divide parameters of train ratio and test ratio are also 80% and 20%, respectively.

As shown in Figure 11 (a) - Figure 14 (a), the general tendency of the data fluctuation between the reference monitor and the RF, MLR, KNN, BP, GA-BP based algorithm of the LCS are consistent with each other. The best performance is RF model, the next are KNN, BP and GA-BP, the worst is MLR. The regressions consistence between the reference data and the five model data are also shown in Figure 11 (b) - Figure 14 (b), and listed in Table 4. The RMSE values between the reference data and the five model data are listed in Table 5.

Atmospheric Measurement Techniques
Discussions

Table 4. Performance of different calibration models for the gaseous pollutant (SO_2, CO, NO_2, O_3) against reference monitor. The determination coefficient R^2 (higher is better, maximum of 1) of different calibration models (RF, MLR, KNN, BP, GA-BP) versus reference monitor.

R^2 / Model	O_3			CO			NO_2			SO_2		
	I	I	I	I	II	III	I	II	III	I	II	III
RF	**0.995**	**0.994**	**0.980**	**0.989**	**0.989**	**0.978**	**0.981**	**0.981**	**0.967**	**0.962**	**0.969**	**0.939**
MLR	0.900	0.898	0.745	0.729	0.807	0.710	0.456	0.530	0.570	0.065	0.333	0.029
KNN	0.965	0.960	0.874	0.921	0.934	0.861	0.866	0.878	0.786	0.686	0.797	0.447
BP	0.932	0.927	0.829	0.837	0.858	0.815	0.756	0.775	0.716	0.332	0.609	0.278
GA-BP	0.935	0.934	0.831	0.841	0.871	0.816	0.742	0.782	0.708	0.341	0.622	0.286

For the O_3 model, the R^2 of RF is better than 0.98. The R^2 of MLR is less than 0.90, and even less than 0.8. The R^2 of the other three models are within 0.82 and 0.97. The performance of different calibration models for the O_3 against reference monitor is also evaluated using RMSE, and the results are listed in Table 5. Using the data listed in Table 5, the RMSE values from the first (I) and third (III) stages have little difference with the one from the second (II) stage, indicating the O_3 electrochemical sensor suitable for the ambient ozone measurement. The mean RMSE values of O_3 between the reference data and the RF, MLR, KNN, BP, GA-BP-based algorithms dada are calculated as 4.06 μgm^{-3}, 16.07 μgm^{-3}, 10.23 μgm^{-3}, 13.35 μgm^{-3} and 13.0 μgm^{-3}, respectively.

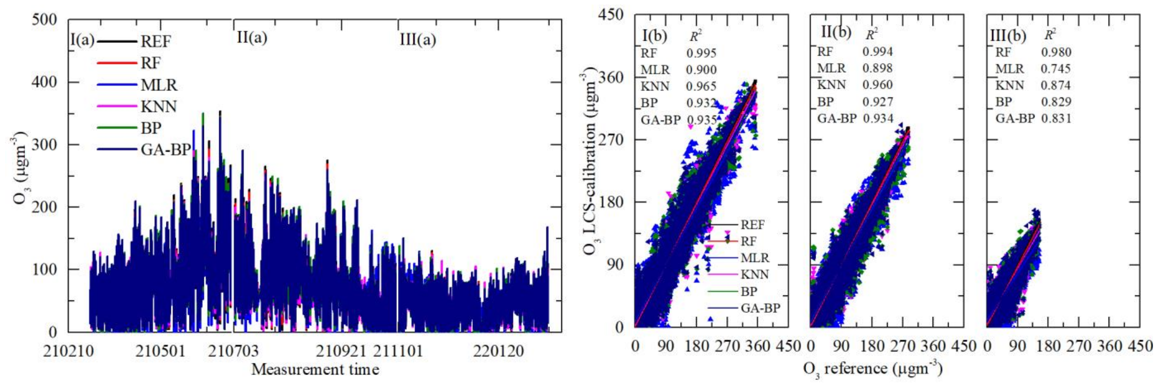

Figure 11. Time series and regressions comparing the reference monitor O_3 data (black) to five calibration model O_3 results. Where red, blue, magenta, olive and navy represent RF, MLR, KNN, BP, GA-BP, respectively. The left panel (a) shows the whole time series data of the measurement period. The right panel (b) shows the regressions of the five calibration models.

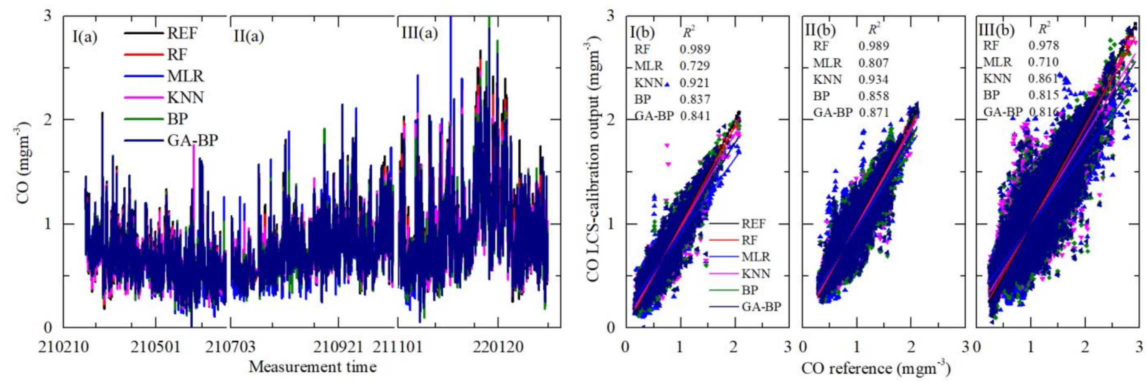

Figure 12. Time series and regressions comparing the reference monitor CO data (black) to five calibration model CO results. Where red, blue, magenta, olive and navy represent RF, MLR, KNN, BP, GA-BP, respectively. The left panel (a) shows the whole time series data of the measurement period. The right panel (b) shows the regressions of the five calibration models.

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

Atmospheric Measurement Techniques Discussions

For the CO model, the R^2 of RF is better than 0.97. The R^2 of MLR is less than 0.81, and even less than 0.7. The R^2 of the other three model are within 0.81 and 0.94. The performance of different calibration models for the CO against reference monitor is also evaluated using RMSE, and the results are listed in Table 5. The RMSE values from the first (I) and third (III) stages have little difference with the one from the second (II) stage, indicating the CO electrochemical sensor also suitable for the ambient CO measurement. The mean RMSE values of CO between the reference data and the RF, MLR, KNN, BP, GA-BP-based algorithms dada are calculated as 0.04 mgm^{-3}, 0.15 mgm^{-3}, 0.10 mgm^{-3}, 0.12 mgm^{-3} and 0.12 mgm^{-3}, respectively.

Table 5. Performance of different calibration models for the gaseous pollutant (SO_2, CO, NO_2 and O_3) against reference monitor. The RMSE errors (lower is better) of different calibration models (RF, MLR, KNN, BP, GA-BP) versus reference monitor.

RMSE / Model	O_3 (μgm^{-3})				CO (mgm^{-3})				NO_2 (μgm^{-3})				SO_2 (μgm^{-3})			
	I	II	III	**MEAN**	I	II	III	**MEAN**	I	II	III	**MEAN**	I	II	III	**MEAN**
RF	4.05	4.06	4.08	**4.06**	0.02	0.03	0.06	**0.04**	2.88	2.88	3.99	**3.25**	0.83	0.64	1.68	**1.05**
MLR	17.79	16.42	14.00	**16.07**	0.12	0.12	0.23	**0.15**	14.54	13.54	13.61	**13.90**	3.53	2.69	5.37	**3.86**
KNN	10.57	10.28	9.84	**10.23**	0.06	0.07	0.16	**0.10**	7.25	6.93	9.61	**7.93**	2.06	1.49	4.05	**2.53**
BP	14.67	13.91	11.46	**13.35**	0.09	0.10	0.18	**0.12**	9.75	9.37	11.07	**10.06**	2.98	2.06	4.63	**3.22**
GA-BP	14.40	13.19	11.41	**13.00**	0.09	0.10	0.18	**0.12**	10.02	9.21	11.21	**10.15**	2.97	2.03	4.60	**3.20**

For the NO_2 model, the R^2 of RF is better than 0.96. The R^2 of MLR is less than 0.60, and even less than 0.5. The R^2 of the other three model are within 0.70 and 0.90. The performance of different calibration models for the NO_2 against reference monitor is also evaluated using RMSE, and the results are listed in Table 5. The RMSE values from the first (I) and third (III) stages have little difference with the one from the second (II) stage, indicating the NO_2 electrochemical sensor still suitable for the ambient NO_2 measurement. The mean RMSE values of NO_2 between the reference data and the RF, MLR, KNN, BP, GA-BP-based algorithms dada are calculated as 3.25 μgm^{-3}, 13.90 μgm^{-3}, 7.93 μgm^{-3}, 10.06 μgm^{-3} and 10.15 μgm^{-3}, respectively.

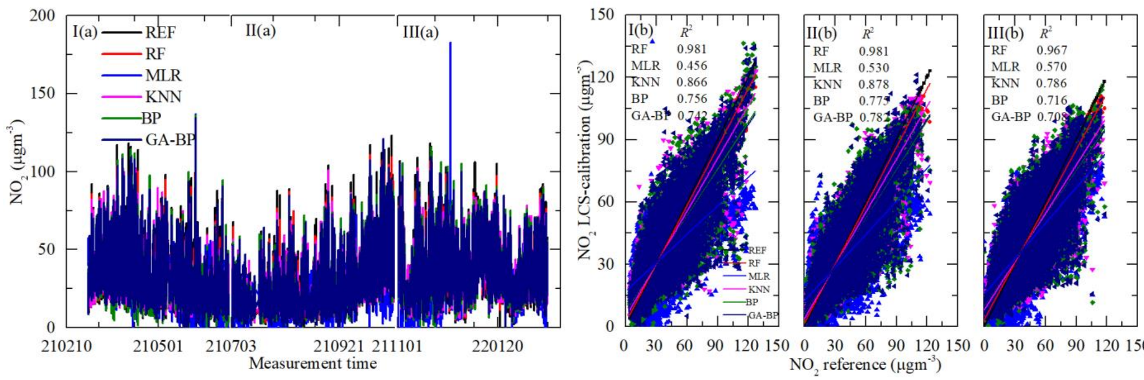

Figure 13. Time series and regressions comparing the reference monitor NO_2 data (black) to five calibration model NO_2 results. Where red, blue, magenta, olive and navy represent RF, MLR, KNN, BP, GA-BP, respectively. The left panel (a) shows the whole time series data of the measurement period. The right panel (b) shows the regressions of the five calibration models.

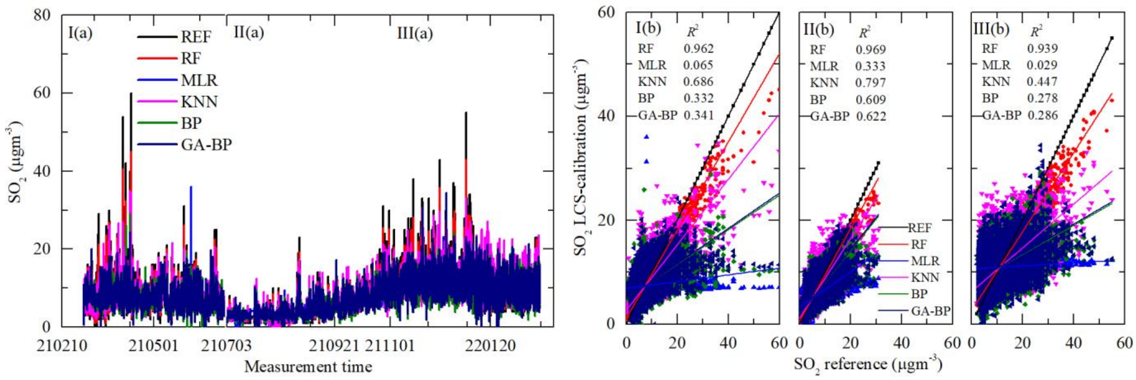

Figure 14. Time series and regressions comparing the reference monitor SO_2 data (black) to five calibration model SO_2 results. Where red, blue, magenta, olive and navy represent RF, MLR, KNN, BP, GA-BP, respectively. The left panel (a) shows the whole time series data of the measurement period. The right panel (b) shows the regressions of the five calibration models.

For the SO_2 model, the R^2 of RF is better than 0.93. The R^2 of MLR is less than 0.40, and even less than 0.1. The R^2 of the other three model are within 0.27 and 0.80. The performance of different calibration models for the SO_2 against reference monitor is also evaluated using RMSE, and the results are listed in Table 5. The RMSE values from the first (I) and third (III) stages have little difference with the one from the second (II) stage, indicating the SO_2 electrochemical sensor with the RF calibration can be used for the ambient SO_2 measurement. The mean RMSE values of SO_2 between the reference data and the RF, MLR, KNN, BP, GA-BP-based algorithms dada are calculated as 1.05 μgm^{-3}, 3.86 μgm^{-3}, 2.53 μgm^{-3}, 3.22 μgm^{-3} and 3.20 μgm^{-3}, respectively.
As shown in Figure 11 - Figure 14 and listed in Table 4 – Table 5, the models (except the RF model) perform poorly for SO_2, especially during the spring and winter. There maybe three reasons for this phenomenon. The first one is the cross interference effect from NO_2 and O_3. The second one is the reaction products (SO_3) dissolved in the electrolyte, which can affect the ion concentration. The last one is the stability of electrode material, which is easily affected by external interference.

5 Conclusion

A low-cost air quality monitoring system based on RF, MLR, KNN, BP, GA-BP algorithms are proposed. The system can measure $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3, simultaneously. The PM prediction model is proposed by taking the particle counters $x_{0.5}$, $x_{1.0}$, $x_{2.5}$, $x_{5.0}$, $x_{10.0}$ of the sensors, ambient temperature (T) and relative humidity (RH) as input and the concentrations $Y_{pm2.5}$ and Y_{pm10} of $PM_{2.5}$ and PM_{10} measured by the reference instrumentation as output. The gas pollutant predictions model is also proposed by taking results of the electro-chemical sensors, T and RH as input and the measurements by the reference monitors as output. The experimental results show that the R^2 of RF for the PM is better than 0.98; the R^2 of MLR for the PM is less than 0.91; the R^2 of the other three model are within 0.86 and 0.98. The mean RMSE values of $PM_{2.5}$ and PM_{10} between the reference data and the RF, MLR, KNN, BP, GA-BP-based algorithms dada are calculated as 3.96 μgm^{-3}, 16.16 μgm^{-3}, 9.48 μgm^{-3}, 10.67 μgm^{-3}, 10.74 μgm^{-3}, and 7.37 μgm^{-3}, 28.90 μgm^{-3}, 18.50 μgm^{-3}, 18.04 μgm^{-3}, 18.17 μgm^{-3}, respectively. For the gas pollutants (SO_2, NO_2, CO and O_3), the R^2 of RF for is better than 0.93; the R^2 of KNN, BP and GA-BP for the gas pollutants (SO_2, NO_2, CO and O_3) is within 0.27 to 0.97; the R^2 of MLR for the NO_2, CO and O_3 is within 0.46 to 0.90, but for SO2 less than 0.40, and even less than 0.1. The mean RMSE values of O_3, CO, NO_2 and SO_2 between the reference data and the RF, MLR, KNN, BP, GA-BP-based algorithms dada are calculated as 4.06 μgm^{-3}, 16.07 μgm^{-3}, 10.23 μgm^{-3}, 13.35 μgm^{-3}, 13.0 μgm^{-3}; 0.04 mgm^{-3}, 0.15 mgm^{-3}, 0.10 mgm^{-3}, 0.12 mgm^{-3}, 0.12 mgm^{-3}; 3.25 μgm^{-3}, 13.90 μgm^{-3}, 7.93 μgm^{-3}, 10.06 μgm^{-3}, 10.15 μgm^{-3}; and 1.05 μgm^{-3}, 3.86μgm^{-3}, 2.53 μgm^{-3}, 3.22 μgm^{-3} and 3.20 μgm^{-3}, respectively. These measurements are consistent with the national environmental protection standard requirement of China. Therefore, the low-cost multi-parameter air quality monitoring system, based on RF, MLR, KNN, BP, GA-BP algorithms, can be used to predict the concentrations of PM and gas pollution. In the next research, we will focus on improving the generalization of the algorithms in more applications, and the performance of the SO_2 sensor.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Acknowledgment

Funding for this study was provided by the National Key Research and Development Program of China (Assistance Agreement No. 2021YFB3200403) and Zhengzhou Education Department (No. 23B413006). The authors also wish to thank Minghui, Li, Hongbiao Liu, and Jinlong Wang for helpful conversations.

References

Alexopoulos, E. C.: Inroduction to Multivariate Regression Analysis, Hippokratia, 14(1), 23-28, 2010.
Ali, S., Glass, T., Parr, B., Potgieter, J., and Alam, F.: Low Cost Sensor With IoT LoRaWAN Connectivity and Machine Learning-Based Calibration for Air Pollution Monitoring, IEEE Transactions on Instrumentation and Measurement, 70, 5500511, doi:10.1109/TIM.2020.3034109, 2021.
Amuthadevi, C., Vijayan, D. S., and Ramachandran, V.: Development of air quality monitoring (AQM) models using different machine learning approaches, Journal of Ambient Intelligence and Humanized Computing, 33(2022), doi:10.1007/s12652-020-02724-2, 2021.
Breiman, L.: Random Forests, Machine Learning, 45(1), 5-32, doi:10.1023/A:1010933404324, 2001.
Brilli, L., Berton, A., Carotenuto, F., Gioli, B., Gualtieri, G., Martelli, F., and Profeti, S.: Innovative low-cost air quality stations as a supporting means for road traffic regulations in urban areas, 15th International Conference on Atmospheric Sciences and Applications to Air Quality (ASSAQ), 489(2020), 012023, doi:10.1088/1755-1315/489/1/012023, 2020.
Cross, E. S., Williams, L. R., Lewis, D. K., Magoon, G. R., Onasch, T. B., Kaminsky, M. L., and Worsnop, D. R.: Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements, atmos. Meas. Tech., 10(9), 3575-3588, doi:10.5194/amt-10-3575-2017, 2017.
Cui, H., Zhang, L., Li, W., Yuan, Z., Wu, M., Wang, C., and Ma, J.: A new calibration system for low-cost Sensor Network in air pollution monitoring, Atmospheric Pollution Research, 12(5), 101049, doi:10.1016/j.apr.2021.03.012, 2021.
Davut, A., and Baykant, A. B.: An effective integrated genetic programming and neural network model for electronic nose calibration of air pollution monitoring application, Neural Computing and Applications, 34(15), 12633-12652, doi:10.1007/s00521-022-07129-0, 2022.
Esposito, E., De, V. 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, Sensors Actuators B-Chemical, 231, 701-713, doi:10.1016/j.snb.2016.03.038, 2016.
Guo, Y., Gao, J., Wang, W., Qin, X., Ren, J., and Ni, D.: Research on outdoor environmental performance evaluation of low-cost atmospheric particulate sensors, China Environmental Science, 40(12), 5133-5141, doi:10.19674/j.cnki.issn1000-6923.2020.0565, 2020.
Hande, B., and Selda, G.: Estimation of Concentration Values of Different Gases Based on Long Short-Term Memory by Using Electronic Nose Biomedical Signal Processing and Control, 69, doi:10.1016/j.bspc.2021.102908, 2021.
Hitchman, M. L., Cade, N. J., Gibbs, T. K., and Hedley, N. J. M.: Study of the factors affecting Mass Transport in Electrochemical Gas Sensors, Analyst, 122(11), 1411-1417, doi:10.1039/a703644b, 1997.
Ioannis, M., Elisavet, S., Agathangelos, S., and Eugenia, B.: Environmental and Health Impacts of Air Pollution: A Review, Frontiers in Public Health, 8(14), 1-13, doi:10.3389/fpubh.2020.00014, 2020.
Ionascu, M. E., Castell, N., Boncalo, O., Schneider, P., Darie, M., and Marcu, M.: Calibration of CO, NO_2, and O_3 Using Airify: A Low-Cost Sensor Cluster for Air Quality Monitoring, Sensors, 21(23), 7997, doi:10.3390/s21237977, 2021.
Janabi, S. A., Alkaim, A., Al-Janabi, E., Aljeboree, A., and Mustafa, M.: Intelligent forecaster of concentrations ($PM_{2.5}$, PM_{10}, NO_2, CO, O_3, SO_2) caused air pollution (IFCsAP), Neural Computing and Applications, 33(21), 14199-14229, doi:10.1007/s00521-021-06067-7, 2021.

https://doi.org/10.5194/amt-2023-163
Preprint. Discussion started: 13 October 2023

Jiao, W., Hagler, G., Willianms, R., Sharpe, R., Brown, R., Garver, D., and Judge, R.: Community Air Sensor Network (CAIRSENSE) project: evaluation of low-cost sensor performance in a suburban environment in the southeastern United States, atmos. Meas. Tech., 9(11), 5281-5292, doi:10.5194/amt-9-5281-2016, 2016.

Khreis, H., Johnson, J., Jack, K., Dadashova, B., and Park, E. S.: Evaluating the Performance of Low-Cost Air Quality Monitors in Dallas, Texas, International Journal of Environmental Research and Public Health, 19(3), 1647, doi:10.3390/ijerph19031647, 2022.

Kumar, T.: Solution of Linear and Non Linear Regression Problem by K Nearest Neighbour Approach: By Using Three Sigma Rule, 2015 IEEE International Conference on Computational Intelligence & Communication Technology, 197-201, doi:10.1109/CICT.2015.110, 2015.

Liang, Y., Ren, C., Wang, H., Huang, Y., and Zheng, Z.: Research on soil moisture inversion method based on GA-BP neural network model, International Journal of Remote Sensing, 40(5-6), 2087-2103, doi:10.1080/01431161.2018.1484961, 2018.

Liu, Y., Wang, Y., and Zhang, J.: New machine learning algorithm: Random forest, Information Computing and Applications. Proceedings of the Third International Conference, ICICA 2012, 246-252, doi:10.1007/978-3-642-34062-8_32, 2012.

Magi, B. I., Cupini, C., Francis, J., Green, M., and Hauser, C.: Evaluation of $PM_{2.5}$ measured in an urban setting using a low-cost opticall particle counter and a Federal Equivalent Method Beta Attenuation Monitor, Aerosol Science & Technology, 54(2), 147-159, doi:10.1080/02786826.2019.1619915, 2020.

Masson, N., Piedrahita, R., and Hannigan, M.: Quantification method for electrolytic sensors in long-term monitoring of ambient air quality, Sensors, 15(10), 27283-27302, doi:10.3390/s151027283, 2015.

Mead, M. I., Popoola, O. A. M., Stewart, G. B., and Landshoff, P.: The use of electro-chemical sensors for monitoring urban air quality in Low-cost, high-density networks, Atmos. Environ., 70, 186-203, doi:10.1016/J.ATMOSENV.2012.11.060, 2013.

Ning, M., Guan, J., Liu, P., Zhang, Z., and O'Hare, G. M. P.: GA-BP Air Quality Evaluation Method Based on Fuzzy Theory, CMC-Computers Materials & Continua, 58(1), 215-227, doi:10.32604/cmc.2019.03763, 2019.

Singh, A., Ng'ang'a, D., Gatari, M. J., Kidane, A. W., Alemu, Z. A., Derrick, N., and Webster, M. J.: Air quality assessment in three East African cities using calibrated low-cost sensors with a focus on road-based hotspots, Environmental Research Communications, 3(7), 075007, doi:10.1088/2515-7620/ac0e0a, 2021.

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, Sensors 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 CO_2., Sensors and Actuators B-Chemical, 238, 706-715, doi:10.1016/j.snb.2016.07.036, 2017.

Wang, C., Pan, B., Wu, X., Song, Y., Zhang, L., Ma, J., and Sun, K.: Research on Quality Control of Atmospheric Grid Monitoring Based on Large Data Analysis, Environmental Monitoring in China 32(6), 1-6, doi:10.19316/j.issn.1002-6002.2016.06.01, 2016.

Wang, S., Li, L., Ma, W., and Chen, X.: Trajectory analysis for on-demand services: A survey focusing on spatial-temporal demand and supply patterns, Transportation Research Part C: Emerging Technologies, 108, 74-99, doi:10.1016/j.trc.2019.09.007, 2019.

Wang, X., Shi, F., Yu, L., and Li, Y. (2013). *MATLAB neural network analysis*: Beihang University Press.

Wei, P., Ning, Z., Ye, S., Sun, L., and Yang, F.: Impact Analysis of Temperature and Humidity Conditions on Electrochemical Sensor Response in Ambient Air Quality Monitoring, Sensors, 18(2), doi:10.3390/s18020059, 2018.

Xu, X., Guo, H., and Fan, J.: Water quality evaluation of Xiaoshan water quality station in eastern Zhejiang Water Diversion Project Based on BP network, Journal of Physics: Conference Series, 1732, 012035, doi:10.1088/1742-6596/1732/1/012035, 2021.

Zhao, C., Wang, Y., Shi, X., Zhang, D., Wang, C., Jiang, J., and Zhang, Q.: Estimating the Contribution of Local Primary Emissionsto Particulate Pollution Using High‐Density Station Observations, J. Geophys. Res.: Atmospheres, 124, 1648-1661, doi:10.1029/2018JD028888, 2019.

Zhao, P., and Lai, L.: Minimax Rate Optimal Adaptive Nearest Neighbor Classification and Regression, IEEE Transactions on Information Theory, 67(5), 3155-3182, doi:10.48550/arXiv.1910.10513, 2021.

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

Zoest, V. v., Osei, F. B., Stein, A., and Hoek, G.: Calibration of low-cost NO_2 sensors in an urban air quality network, Atmospheric Environment, 210, 66-75, doi:10.1016/j.atmosenv.2019.04.048, 2019.