

Evaluating machine learning model performance in a two-step colocation process for TVOC and BTEX sensor calibration

Caroline Frischmon¹, Jack Porter², Ethan Balagopalan², William Senga², Jill Johnston³, and Michael Hannigan¹

¹University of Colorado Boulder Department of Mechanical Engineering, 1111 Engineering Drive Boulder CO 80309

²South Coast Air Quality Monitoring District, 21865 Copley Drive Diamond Bar, CA 91765

³University of California Irvine Department of Environmental and Occupational Health, 856 Health Sciences Quad, Irvine, CA 92697

Correspondence: Caroline Frischmon (caroline.frischmon@colorado.edu)

Abstract. Calibration of low-cost air quality sensors (LCSs) for total volatile organic compound (TVOC) and benzene, toluene, ethylbenzene, and xylenes (BTEX) quantification remains challenging due to the sensors' cross-sensitivity to temperature and humidity and their tendency to drift over time. In this study, we aimed to improve TVOC and BTEX metal oxide (Figaro TGS 2600, 2602, 2611) sensor calibration using a two-step colocation strategy. A two-step colocation places one LCS (the secondary standard) with a reference monitor while others operate in the field, then briefly colocates the field sensors with the secondary standard to address inter-sensor variability and drift. This strategy made it possible to develop the calibration model under environmental conditions closely matching those of the field, which is essential for model transferability from colocation to field conditions. In addition to TVOC and BTEX, we applied the two-step colocation process to NO₂ electrochemical (Alphasense B-4) sensors to demonstrate the broader applicability of our approach beyond TVOC and BTEX quantification.

Next, we compared the performance of multiple machine learning models, including ridge, lasso, random forest, gradient boosting, extreme gradient boosting, support vector regression, and linear regression, to investigate the optimal model choice for calibration. We found that no single model performed best across all pollutants. For example, gradient boosting excelled at capturing peak TVOC concentrations, while linear regression performed best for BTEX. Conversely, linear regression was the worst-performing model for NO₂. Overall, the models showed satisfactory RMSE around 40- 50 ppb for TVOC, 1.25- 1.75 ppb for BTEX, and 4-6 ppb for NO₂. However, all models also overestimated baseline concentrations and underestimated peaks. The severity of this bias depended on the reference concentration distribution, with the most severe peak underestimation occurring in the more heavily skewed TVOC and BTEX data. The systematic bias at baseline and peak concentrations was not evident in the overall mean bias error, which was near zero for all pollutants. This result underscores the need to evaluate model performance across the entire concentration distribution. Finally, we found that calibration performance was sensitive to the choice of training and testing data split. Future research could seek to optimize the training and testing split to ensure robust model transferability to field data.

1 Introduction

Low-cost sensors (LCSs) are an increasingly popular tool used by community groups, scientific researchers, and local governments to document air quality impacts at spatial and temporal scales finer than are possible with regulatory or reference-grade monitoring (Fanti et al., 2021; Clements et al., 2017; Commodore et al., 2017; Okorn and Iraci, 2024). Despite their increasing usage, LCSs face limitations due to their reduced accuracy compared to reference-grade monitoring methods (Karagulian et al., 2019; Castell et al., 2017; Clements et al., 2017). To improve accuracy, LCSs are often calibrated via collocation with a reference instrument, which corrects sensor drift and cross-sensitivity to temperature, humidity, and non-target pollutants (Liang, 2021; Karagulian et al., 2019; Liu et al., 2020).

In a collocation, sensors are placed next to a reference-grade monitor for a set period. The sensor signal is then calibrated to the reference-grade concentration time series via a linear regression or machine learning calibration model (Liang, 2021; Okorn and Iraci, 2024). For certain pollutants, such as ozone and PM_{2.5}, this process can significantly improve the accuracy of LCSs (Barkjohn et al., 2021; Masiol et al., 2018). However, calibrating LCSs for other pollutants, especially volatile organic compounds (VOCs), has had more limited success (Clements et al., 2017; Collier-Oxandale et al., 2019), leading some users to interpret low-cost VOC data only qualitatively (Frischmon et al., 2025a; Raheja et al., 2022).

VOCs are a broad class of chemicals that originate from both natural sources, such as wildfires and vegetation, and anthropogenic sources, such as industrial solvents, consumer products, and fossil fuel extraction, refining, and combustion. VOCs impact respiratory and cardiovascular health and can cause secondary health impacts by contributing to the formation of secondary organic aerosols and tropospheric ozone (Kampa and Castanas, 2008; Haagen-Smit et al., 1953; Laaksonen et al., 2008). Benzene, toluene, ethylbenzene, and xylenes, or BTEX, are a subset of VOCs that are especially relevant to human health due to their carcinogenicity (Kampa and Castanas, 2008). Total VOCs (TVOCs) and subsets like BTEX are challenging to quantify using LCSs in part because of their relatively low ambient concentrations (Spinelle et al., 2017b). Additionally, the chemical makeup of TVOCs can include many different species to which LCSs exhibit varying sensitivities. Low-cost TVOC and BTEX sensors can also show greater sensitivity to temperature and humidity than to VOC species, and they exhibit drift over long-term use (Lewis et al., 2016; Collier-Oxandale et al., 2019). Given these challenges, researchers have had the greatest success in calibrating these sensors for indoor use only, where concentrations tend to be higher and temperature and humidity can be controlled (Robin et al., 2021; Leidinger et al., 2014).

In this paper, we demonstrate a comprehensive calibration approach to better address the challenges of ambient, outdoor TVOC and BTEX quantification using LCSs. This approach involves a two-step collocation and a robust evaluation of machine learning calibration models. We include NO₂ quantification as a comparison to demonstrate how our approach can be used for other pollutants as well.

1.1 Two-step collocation

Calibration models that perform well with collocation data may still produce large error in field predictions if the model is not transferable to the environmental and sensor conditions at field sites. For instance, Malings et al. (2019) showed how

55 sensor drift reduces the accuracy of calibration models when applied to field data collected long after the colocation period. Others have demonstrated how calibration models can overfit to the environmental conditions, source mixtures, and pollutant concentrations present at the colocation site, making the model less reliable when transferred to field sites (Malyan et al., 2024; Casey and Hannigan, 2018; Vikram et al., 2019; Nowack et al., 2021). These impacts are especially important to consider for VOCs, which exhibit higher spatial variability than other pollutants and are composed of a complex mixture of chemicals, 60 further complicating model transferability (Collier-Oxandale et al., 2018).

To ensure a calibration model is transferable to field data, colocations must capture the range of environmental conditions expected during field deployment, though many colocations occur only before and/or after field deployment, potentially during different seasons (Zamora et al., 2023). In contrast, a two-step colocation occurs simultaneously with field deployment, making it possible to capture data under the same seasonal conditions. In a two-step colocation, a single LCS system, referred to here 65 as the secondary standard, is colocated with a reference-grade monitor while the remaining LCS systems collect data at field sites (Sá et al., 2023; Okorn and Hannigan, 2021b). The field sensors are then colocated with the secondary standard for a shorter period, referred to as the harmonization step, to address inter-sensor variability. A two-step colocation can improve calibration transferability, especially when the colocation with the reference monitor occurs close to the field deployment sites, as this increases the likelihood of capturing similar environmental conditions within the same seasonal timeframe. Two-step 70 colocations also allow for more extensive colocation periods without decreasing the amount of field data that can be collected, which is advantageous for complex machine learning models that require extensive data to prevent overfitting. In this study, we improve upon the two-step colocation process developed by Sá et al. (2023) and Okorn and Hannigan (2021b) to further address sensor drift and model transferability.

1.2 Evaluation of machine learning calibration models

75 Machine learning algorithms can improve calibration models by accounting for the non-linear signal response of LCSs (Robin et al., 2021; Liang, 2021). Due to their complexity, machine learning models require more extensive colocation data to prevent model overfitting to the colocation data, as this can make the model less reliable under changing field conditions (Concas et al., 2021). The optimal machine learning model varies by dataset and is typically selected by comparing calibration model performance metrics, such as root mean squared error (RMSE) and mean bias error (MBE), for a portion of colocation data 80 excluded from model training (Wang, 2024). Evaluating the excluded data, called testing data, helps prevent overfitting to the training data.

Studies comparing the performance of various machine learning calibration models, such as random forest, artificial neural networks, support vector machine regressions, and gradient boosting, have relied on PM_{2.5}, CO, NO₂, CO₂, and O₃ datasets (Zimmerman et al., 2018; Casey and Hannigan, 2018; Johnson et al., 2018; Srishti et al., 2023; Considine et al., 2021). Com- 85 pared to these pollutants, TVOC and BTEX data are often much more imbalanced, with the majority of measurements clustered near low parts-per-billion baseline levels and occasional episodic spikes multiple orders of magnitude higher (Edgerton et al., 1989; Ou-Yang et al., 2018). Imbalanced datasets make it more difficult to assess calibration model performance, as models that poorly fit elevated concentrations may still achieve strong overall performance metrics if they accurately predict baseline

concentrations (Okorn and Hannigan, 2021a; Silberstein et al., 2024). In these cases, partitioning the performance metrics by data percentile can reveal significantly underestimated peak concentrations (Zimmerman et al., 2018; Frischmon et al., 2025b). Additionally, certain machine learning models, especially gradient boosting, are generally better suited to address data imbalance (Galar et al., 2011), but calibrations for low-cost TVOC and BTEX sensors have been limited to linear regression, random forest, and neural network models (De Vito et al., 2008; Frischmon et al., 2025b; Spinelle et al., 2017a; Yurko et al., 2019; Hong et al., 2023; Robin et al., 2021; Li et al., 2023; Okorn et al., 2021; Okorn and Hannigan, 2021a). Thus, our study provides the first comparison to our knowledge of a comprehensive suite of machine learning models for TVOC and BTEX sensor calibration, including random forest, gradient boosting, support vector regression, artificial neural network, ridge, and lasso algorithms. We also evaluate model performance for NO₂ calibration to compare how these models perform on a more balanced dataset.

2 Methods

We deployed a network of low-cost sensor platforms, called HAQ-Pods, in South Los Angeles to study the impacts of local oil and gas development. Field data collection is ongoing as of July 2025. However, since this study is focused only on sensor calibration, rather than field data analysis, we limit the present analysis to data collected between 10/4/2023 and 3/28/2024 to use just one pre- and post-harmonization. The pre-harmonization occurred from 10/4/23 to 10/9/23, and the post-harmonization occurred from 3/18/24 to 3/28/24. From 10/10/23 to 3/18/24, the secondary standard HAQ-Pod was colocated with a reference instrument for calibration, while all other pods were deployed at field sites in South Los Angeles (Figure 1).

2.1 Instrumentation

HAQ-Pods were developed in the Hannigan Air Quality (HAQ) Lab and have previously been used to capture local pollution events (Frischmon et al., 2025a; Silberstein et al., 2024). HAQ-Pods use interchangeable, commercial sensors to quantify various pollutants. The metal oxide and electrochemical sensors used in this study are listed in Table 1. We also used a Bosch BME 680 sensor for temperature and humidity measurements. The sensors are housed in a weather-proof case (10.7" x 9.8" x 4.8") outfitted with fans to draw air across the sensor surfaces. HAQ-Pods feature updated circuitry compared to Y-Pod air monitors, which have been described elsewhere (Collier-Oxandale et al., 2020, 2018).

Table 1. Sensors included in the HAQ-Pods

Sensor	Target pollutant	Measurement principle
Figaro TGS 2600	TVOC and BTEX	Metal oxide
Figaro TGS 2602	TVOC and BTEX	Metal oxide
Figaro TGS 2611	TVOC and BTEX	Metal oxide
Alphasense B-4 series	NO ₂	Electrochemical

For the TVOC calibration, the secondary standard HAQ-Pod was colocated with an Extractive Fourier Transform Infrared Spectrometer operated and maintained by the South Coast Air Quality Management District (South Coast AQMD) using methods similar to those described in Mellqvist et al. (2017). Ambient air is continuously drawn through heated Teflon inlet (1-inch ID) mounted on the roof of the station into a 25-liter cell at ~25 liters per minute. Infrared light is generated with a glow-bar and directed into the cell with a series of mirrors. The light beam is passed through the cell multiple times via a series of curved mirrors to achieve a total path length of ~100m before exiting the cell and being directed into the infrared spectrometer (Bruker, Type: Matrix-M). Light is detected with an Indium/Antimonide detector in the 1800-4000 cm^{-1} region and the spectrometer has a spectral resolution of 0.5 cm^{-1} .

Gas concentrations are calculated using synthetic spectra generated by fitting calibration spectra taken from the HITRAN (Sharpe et al., 2004) and PNNL (Rothman et al., 2005) databases to the raw spectra collected by the spectrometer in a linear least squares type fitting algorithm within the spectral evaluation window of 2725 cm^{-1} to 3007 cm^{-1} (Griffith, 1996; Johansson et al., 2014). The spectral fit includes calibration spectra for H₂O, CH₄, HDO, Propane, Butane, and Octane with TVOC reported as the sum of Propane, Butane, and Octane.

For the BTEX calibration, the secondary standard HAQ-Pod was colocated with an automated gas chromatograph (auto-GC) instrument (TricornTech; MiTAP P320 Series) operated and maintained by South Coast AQMD. The auto-GC operates continuously with a distinct sampling and analysis phase that span 40 min and 20 min, respectively. During the sampling phase, the auto-GC continuously draws ambient air at 5 cm^3/min through a roof mounted inlet into a pre-concentrator for 40 minutes, totaling 200 cm^3 of gas. During the analysis phase, the sample passes through a series of heated GC columns by way of pressurized inert carrier gas (N₂), where compounds are separated by polarity and subsequently detected with a series of photoionization detectors. Digital chromatograms are recorded and used to calculate ambient gas concentration via reference to a set of previously determined calibration chromatograms. Calibration chromatograms are determined by routinely (twice weekly) directing gas from a concentration-certified compressed cylinder into the auto-GC. The method is capable of an accuracy of 0.1 parts per billion by volume (ppbv).

For the NO₂ calibration, the secondary standard HAQ-Pod was colocated with a Teledyne T200 instrument maintained by the South Coast AQMD. The reference instrument was calibrated by the South Coast AQMD every six months and underwent weekly zero and span checks, in addition to nightly PC checks to ensure correct values.

2.2 Two-step colocation

The secondary standard HAQ-pod was colocated with the South Coast AQMD instruments while all other HAQ-pods were deployed at field sites in South Los Angeles from 10/10/23 to 3/18/24. The field HAQ-pods were harmonized with the secondary standard HAQ-pod from 10/4/23 to 10/9/23 and from 3/18/24 to 3/28/24. The physical set-up of the harmonization is shown in Figure 2. Unlike previous two-step colocations, the HAQ-Pods were harmonized with the secondary standard both before and after field deployment to better account for sensor drift across the entire study period, as suggested by Sá et al. (2023).

Sá et al. (2023) and Okorn and Hannigan (2021b) demonstrated different approaches for propagating the calibration model to the field LCSs in a two-step colocation (Figure 3). Okorn and Hannigan (2021b) developed a primary calibration model between

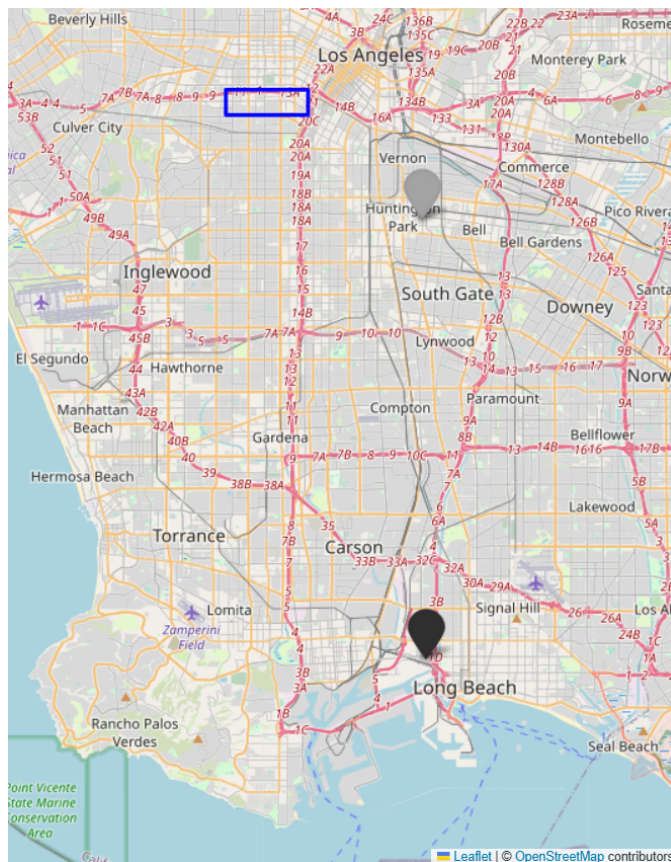


Figure 1. The TVOC and BTEX colocation site is indicated by the black icon, while the NO₂ colocation site is indicated by the gray icon. The harmonization site and field sites are encompassed in the blue box. ©OpenStreetMap contributors 2025. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

the secondary standard and reference instrument. In the harmonization step, the field LCS systems were then calibrated via a secondary calibration model to the concentration timeseries predicted by the calibrated secondary standard. This method is better suited for projects where the colocation and harmonization steps are similar in length, since each step requires equally complex calibration models.

Conversely, Sá et al. (2023) first developed a multi-linear regression model for each sensor signal corrected to the secondary standard's corresponding sensor signal. The multi-linear regression models for the pollutant sensors included temperature and humidity in addition to the pollutant sensor signal. The primary calibration model between the secondary standard and reference monitor could then be applied directly to the corrected field LCS systems. Since this method uses only a linear regression in the harmonization step, it is well-suited for projects where the harmonization is much shorter than the colocation. However, including temperature and humidity in the pollutant sensor regression model may introduce dependencies on environmental conditions that then require the harmonization step to span the full range of conditions expected in the field—thereby reducing



Figure 2. Harmonization set up in Los Angeles

the method’s usefulness for short harmonizations. To avoid this, we included only the pollutant sensor signal and a time elapsed variable in our linear regressions for harmonization, so that the regression corrects only for inter-sensor linear baseline shifts and drift (Equation 1).
 160

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 T_i + \epsilon_i \quad (1)$$

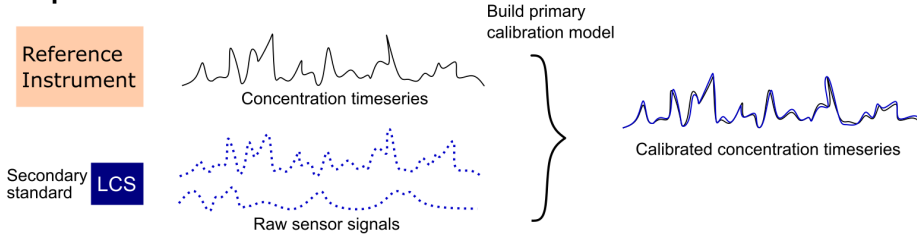
Where β s represent the model coefficients, x_i is the sensor signal for the i^{th} observation, T_i is the time elapsed from the start of data collection for the i^{th} observation, and ϵ_i represents the error term for the i^{th} observation.

2.3 Calibration model development

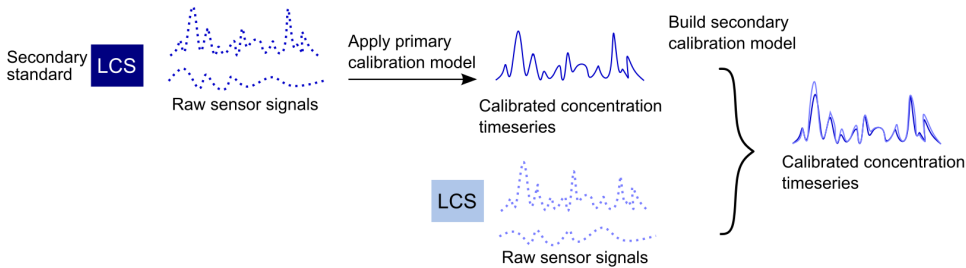
165 Inputs into the TVOC and BTEX colocation calibration models included the signals from the metal oxide sensors, as well as temperature, humidity, and a time elapsed feature to account for sensor drift. The NO_2 calibration model used inputs from the electrochemical NO_2 sensor along with temperature, humidity, and time-elapsed data.

To evaluate model performance, we split the colocation data into training and testing datasets, allocating 80% for training and 20% for testing. While this train/test split is typically performed once, we repeated the calibration process 20 times to assess
 170 model robustness. For each run, we used a different split of the data. Specifically, for each iteration, we randomly selected a starting point in the time series and designated the next 10% of data as testing data. A second 10% segment was selected from a point halfway through the dataset relative to the first chunk, wrapping around if necessary. Figure 4 shows an example test/train split for each pollutant with reference and predicted concentrations (using a linear regression model) to demonstrate how this test/train approach allows the testing data to span different times of the year. This approach ensures the model is not overfit

Step 1: Colocation



Step 2: Harmonization (Okorn et. al).



Step 2: Harmonization (Sa et. al. and this paper)

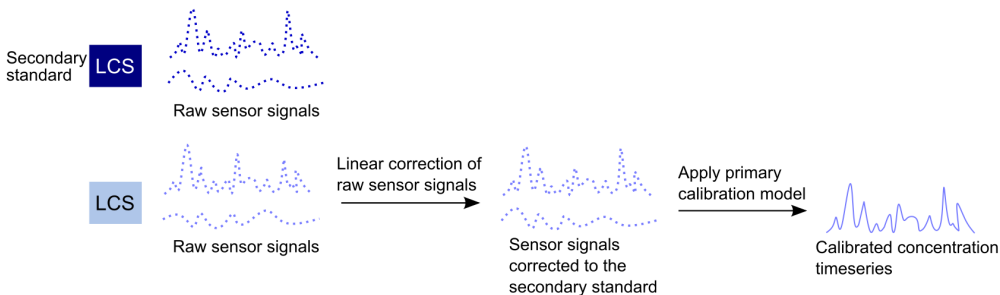


Figure 3. Model propagation methods for two-step colocation. This example has two sensor signal inputs: the target pollutant signal and temperature. Solid lines indicate a concentration timeseries and dotted lines indicate raw sensor signal. Note that Sa et. al. include temperature and humidity in all linear correction models, while we include time elapsed in all linear correction models in this paper.

175 to a specific environmental state or season, while minimizing information leakage from autocorrelation between training and testing datasets (Nowack et al., 2021).

The sensor signals are time averaged to 60-minute intervals prior to model training to reduce noise (Cabello-Solorzano et al., 2023). Signals were then z-scored to a mean zero and standard deviation equal to one based on the training data. During model training, we use k-fold cross-validation (k=5) to tune hyper-parameters to values that minimize the loss function, or mean squared error (MSE), shown in Equation 2. We used the Scikit-learn Python module for all models except the artificial neural network, which used Tensorflow, and extreme gradient boosting, which used the Xgboost package.

180

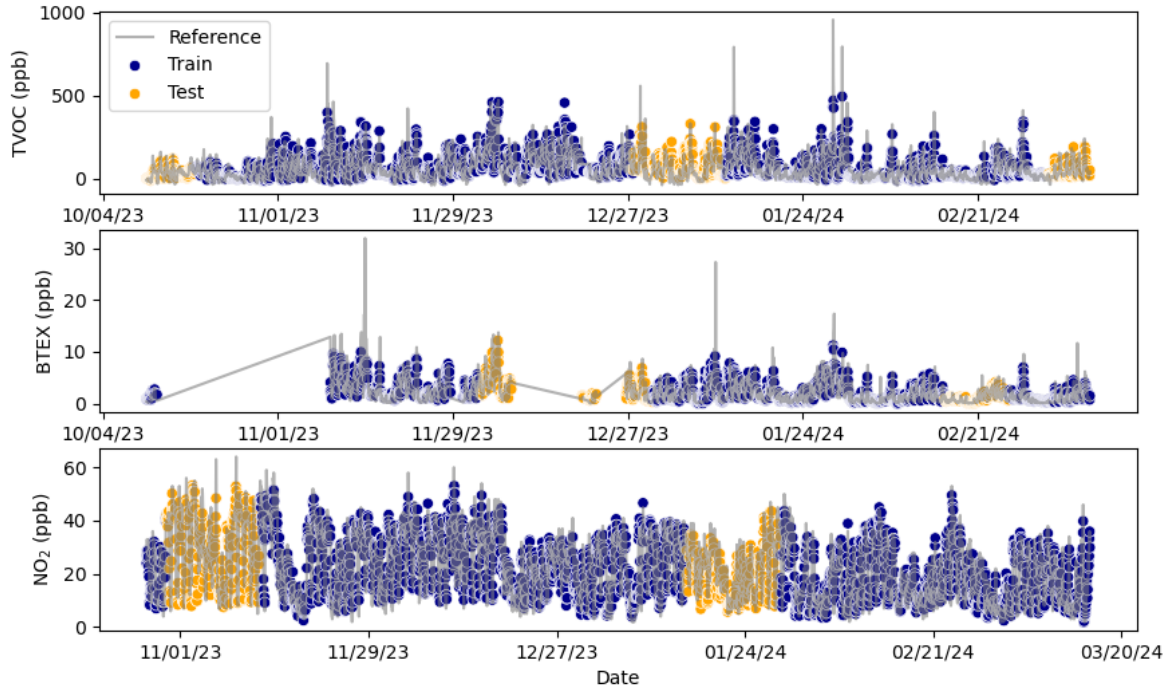


Figure 4. Example of the test/train split for each colocation timeseries.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

Where y_i is the reference concentration for the i^{th} observation and \hat{y}_i is the predicted concentration for the i^{th} observation, and n is the number of observations.

185 2.4 Machine learning model descriptions

2.4.1 Linear regression (LR)

LR employs the following expression to predict the reference concentration using the input sensor signals:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \epsilon_i \quad (3)$$

Where β s represent the model coefficients, y_i is the reference concentration for the i^{th} observation, x_{ij} is the j^{th} feature or
 190 input for the i^{th} observation, p is the number of features or model inputs, and ϵ_i represents the error term for the i^{th} observation.

The model coefficients are tuned by minimizing the model error. There are no hyper-parameters that require tuning via k-fold cross-validation in a linear regression model.

2.4.2 Regularized linear regression: Lasso (L) and ridge (R)

Lasso (L) and ridge (R) are regularized linear regression models that add a penalty term to the loss function to reduce overfitting. Lasso uses L1 regularization, which adds the sum of the absolute values of the regression coefficients to the loss function (Equation 4) (Tibshirani, 1996). This reduces the coefficients of unimportant inputs to zero, removing them from the model. Lasso regression is useful when it is suspected that not all model inputs are important for prediction. In ridge regression (L2 regularization), the penalty term is the sum of the squared values of the model coefficients (Equation 5), which reduces the coefficients towards zero without setting them at zero (Hoerl and Kennard, 1970). Thus, ridge regression simplifies the model while retaining all inputs in the model.

$$\text{Lasso loss Function} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (4)$$

$$\text{Ridge loss Function} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (5)$$

Where y_i is the reference concentration for the i^{th} observation, \hat{y}_i is the predicted concentration for the i^{th} observation, β_j is the coefficient for the j^{th} model input, and λ is the regularization hyper-parameter, which determines the strength of the regularization.

2.4.3 Ensemble decision tree-based models: Random forest (RF), gradient boosting (GB), and Extreme Gradient Boosting (XGB)

Random forest (RF), gradient boosting (GB), and extreme gradient boosting (XGB) are ensemble decision tree-based learning methods. Similarly to a flow chart, a decision tree separates the data into nodes based on the characteristics of the data, eventually reaching a final node that determines the prediction of the data (Song and Ying, 2015). As ensemble methods, these models combine outputs from multiple trees to produce a more robust solution (Breiman, 2001; Natekin and Knoll, 2013). RF uses a bootstrap aggregating approach (bagging) to combine the trees, where each tree is built independently and given equal weight in the final solution (Kumar and Sahu, 2021; Breiman, 2001). In contrast, GB and XGB use a "boosting" technique, where trees are built sequentially and seek to improve upon the error of the prior tree (Kumar and Sahu, 2021; Natekin and Knoll, 2013). XGB is a faster, optimized version of GB that also adds regularization (lasso and ridge) to reduce overfitting (Chen and Guestrin, 2016).

2.4.4 Support vector machine regression (SVR)

Support vector machine regression (SVR) maps the data to a high-dimensional space using a kernel function and determines the best-fitting hyperplane for the transformed data (Smola and Schölkopf, 2004). SVR minimizes the model complexity by
220 only fitting the hyperplane to points falling outside a margin of tolerance. This is different than other machine learning models, which seek to minimize overall error across all data points.

2.4.5 Artificial neural network (ANN)

An artificial neural network (ANN) is a machine learning model designed to mimic the way input signals are processed and transformed into output signals in the brain (Wesolowski and Suchacz, 2012). In an ANN, the input signals pass through
225 interconnected layers of nodes. The connections between the nodes are assigned weights, which are adjusted as the model is trained to optimize the output signal accuracy (Stansbury, 2014; Krogh, 2008).

2.5 Calibration model evaluation

We evaluated the overall fit of the calibration models using three performance metrics: root mean squared error (RMSE) (Equation 6), mean bias error (MBE) (Equation 7), and coefficient of determination (R^2) (Equation 8). We also evaluated
230 RMSE and MBE by percentile groups to assess model fits across the entire range of data. For the percentile-based assessment, the data were partitioned into 0-5th percentile, 5th-25th percentile, 25th-75th percentile, 75-95th percentile, and 95th-100th percentile groups.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

$$\text{MBE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) \quad (7)$$

$$235 \quad R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

Where y_i is the reference concentration for the i^{th} observation and \hat{y}_i is the predicted concentration for the i^{th} observation, and n is the number of observations.

Lastly, we applied the calibration models to sensor data collected from three field sites within the Los Angeles study area. We compare the predictions of each model to understand how they vary when applied to field conditions.

The full HAQ-Pod network included 11 field sites. We show data for just three sites to maintain the focus on sensor calibration in the present study. These three HAQ-Pods, along with the secondary standard HAQ-Pod had 100% data completeness in the harmonization step, resulting in 333 harmonization data measurements for each HAQ-Pod after hourly averaging. Data completeness for HAQ-Pods 1, 2, and 3 in the field was 89%, 86% and 99%, respectively, resulting in 3413, 3289, 3789 field data points for each HAQ-Pod. The secondary standard HAQ-Pod had 90% data completeness, resulting in 3452 data points for the colocation calibration models.

3.1 Two-step colocation

As observed during harmonization, colocated HAQ-Pod raw sensor signals are generally well correlated, with correlation coefficients between the secondary standard HAQ-pod raw signals and corresponding raw signals in the other HAQ-pods ranging from 0.87 to 1 (supplemental). However, some shifted baselines and amplitudes are still visible in the left column of Figure 5, especially for the metal oxide sensors, demonstrating the importance of correcting individual sensor signals rather than relying on a universal or general calibration. After applying the linear correction developed during harmonization (coefficients listed in Tables S2-S7), the signals are visually more aligned (right column of Figure 5 and have correlation coefficients with the colocation HAQ-pod signal ranging from 0.96-1 (supplemental).

The left column of Figure 5 also reveals why performing both pre- and post-harmonization is essential for some sensors, especially Figaro TGS 2600 and 2611 in this case. These sensors show different behavior compared to the secondary standard HAQ-pod signal in the pre- and post- harmonization, indicating sensor drift. We addressed this drift by including a time-elapsd feature in the harmonization correction, which approximates the drift as a linear function. It is likely the sensor drift is not perfectly linear, but the approximation appears to be sufficient here based on performance at both the start and end of harmonization. Conducting both pre- and post- harmonizations was essential for checking the drift approximation, as otherwise the model may extrapolate drift beyond what is reasonably possible.

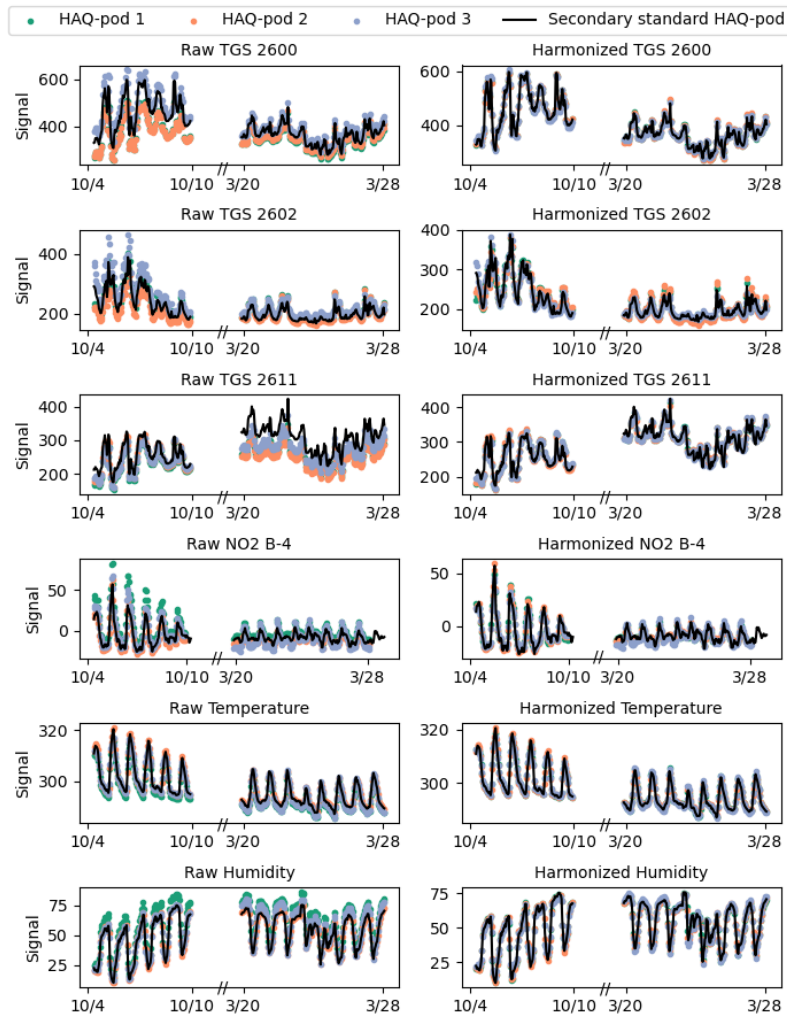


Figure 5. Raw sensor signal collected during harmonization (left column) and corrected signals after harmonization with the colocation HAQ-pod (right column).

265 The temperature and humidity ranges recorded during the field deployment show strong overlap with the temperature and humidity ranges recorded at the HAQ-pod colocation site (Figure 6, note that measurements are uncalibrated sensor temperature and humidity). This overlap, which is important for colocation model transferability to field data, is possible because the two-step colocation allows for simultaneous colocation and field data collection. If colocation had occurred only before and after field deployment, as is typical for many low-cost sensor deployments, the overlap would look more like that of the field and harmonization phases in Figure 6. Conditions during the harmonization phase trended towards higher temperature and humidity because the harmonizations occurred in fall and spring, while the field deployment was mainly in winter. This seasonal difference could lead to poor transferability if temperature and humidity impacts on the sensors were corrected in

270 the harmonization phases because the model would have to extrapolate the impacts at low temperature and humidity ranges. Instead, our two-step colocation strategy corrects for sensor cross-sensitivity to temperature and humidity in the colocation phase only, where there is sufficient overlap of environmental conditions.

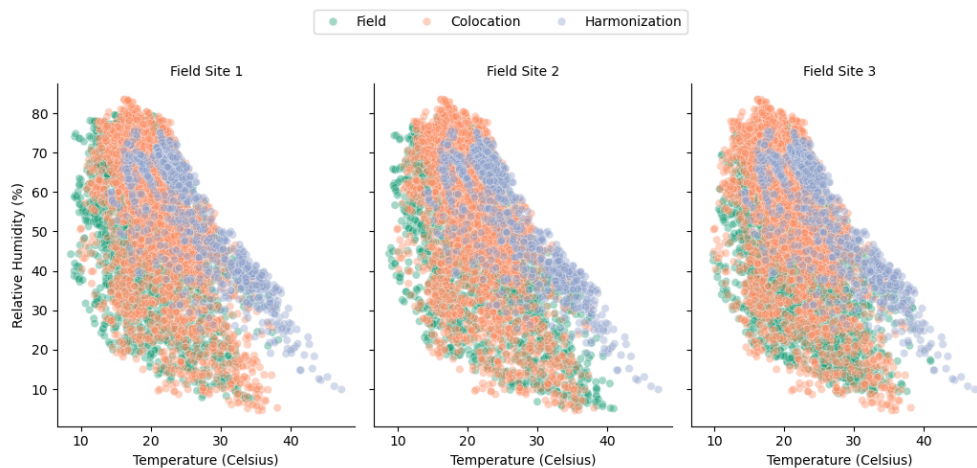


Figure 6. Relative humidity (RH) and temperature (T) observed during field (green), colocation (orange), and harmonization (purple) phases. These plotted values are harmonized sensor signals not calibrated to a true T or RH measurement.

3.2 Evaluation of machine learning calibration models

275 Figure 7 highlights how the distribution of TVOC and BTEX concentrations measured by the reference instruments during colocation are more skewed than NO_2 , as 75% and 78% of the total concentration range falls within the 95-100th percentile group for TVOC and BTEX, respectively, compared with just 31% for NO_2 . Thus, the inclusion of NO_2 in our analysis provides a useful comparison of machine learning calibration performance on a more balanced dataset.

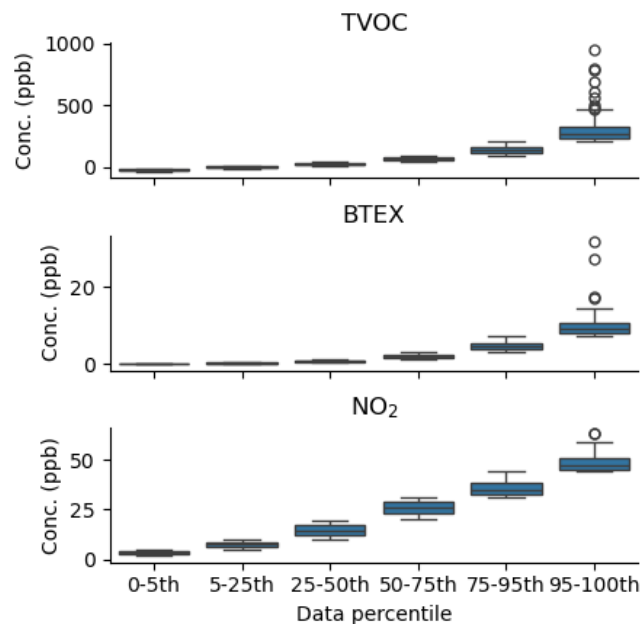


Figure 7. Reference concentrations, divided into data percentile ranges, of each pollutant collected during colocation.

In the following sections, we show the calibration model performance for each pollutant across all 20 calibration repetitions (Figures 8, 11, 14). R^2 values for these models are available in Figure S1. To better visualize how these models actually predict testing and field data concentrations, we show predicted data from one calibration run in Figures 9, 10, 12, 13, 15, and 16. The calibration run we selected to visualize in these plots is one that was close to the median overall performance, based on RMSE and MBE, for each pollutant.

3.2.1 TVOC calibration

Depending on the test/train split of the data, the overall RMSE and MBE for each model varied by 25-35 ppb (Panels A, C in Figure 8). Within each run, the models showed similar overall testing RMSE and MBE, indicating the calibration performance depended more on test/train split than model choice. R^2 for all the models were also generally similar, with a median value of around 0.65 across all runs (Figure S1). All the models tended to overestimate baseline concentrations in the 0–5th percentile, with the linear models showing the least overestimation bias (Figure 8, panel d). The models also tended to underestimate peak concentrations in the 95–100th percentile, though gradient boosting (GB) performed best for these peak values based on its median MBE.

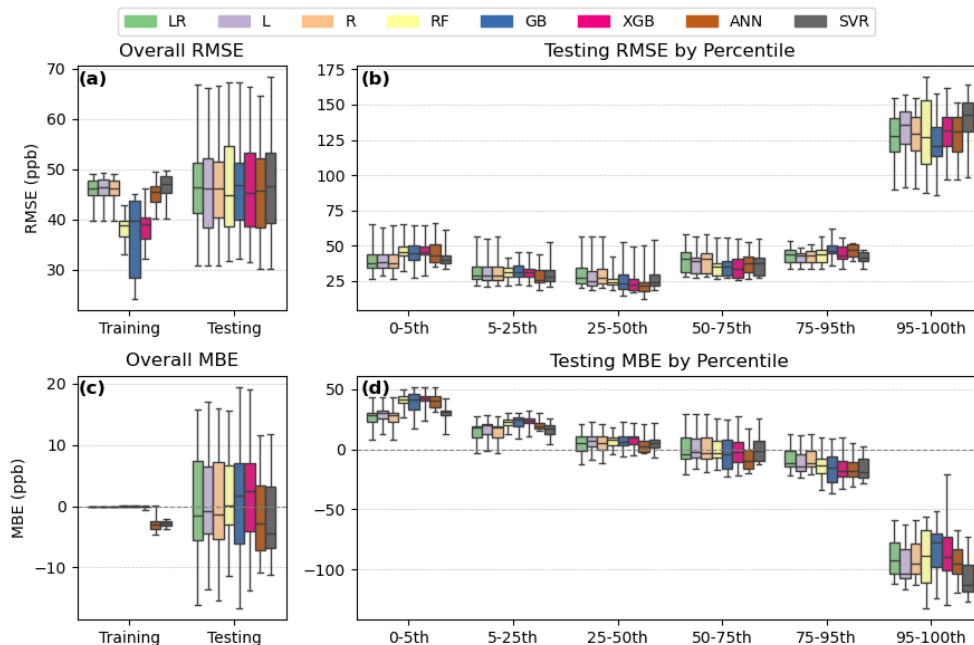


Figure 8. Performance statistics for TVOC calibration models, including overall RMSE (a), RMSE by percentile group (b), overall MBE (c), and MBE by percentile group (d). The calibration models are linear regression (LR), lasso (L), ridge (R), random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), artificial neural network (ANN), and support vector regression (SVR).

Linear regression (LR) and ridge regression (R) have nearly identical performance statistics, indicating that ridge, or L2, regularization of the linear model is not necessary for this data. Lasso, or L1, regularization removed humidity and time elapsed from the model inputs in 16 of 20 runs, removed temperature in one run, temperature and time elapsed in one run, and just time elapsed in two runs. None of these combinations significantly changed the model's performance compared to LR and R.

Visual inspection of the scatter plots in Figure 9 shows that all models except the two boosting models, GB and XGB, severely underpredict the five highest concentrations. Boosting models are known to better predict extreme values in a regression model. These results suggest that boosting models are a good choice for applications where accurately capturing extreme concentrations is critical, such as pollution episode detection.

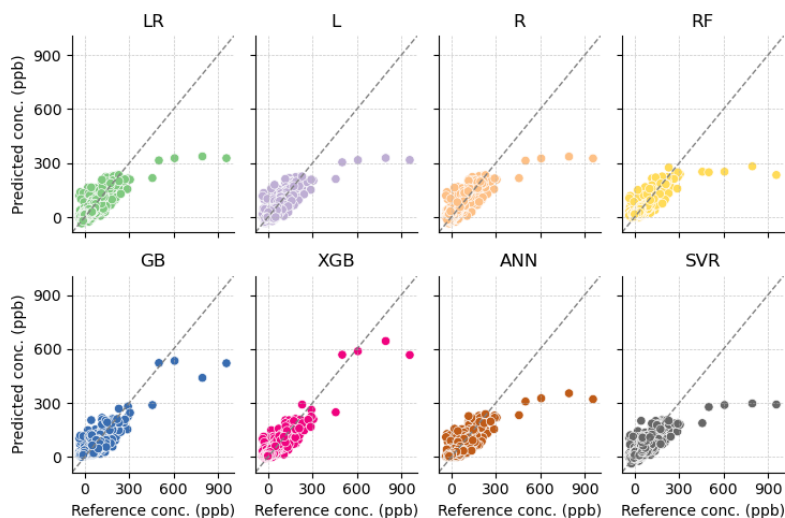


Figure 9. Predicted concentration versus reference concentration for TVOC calibration models: linear regression (LR), lasso (L), ridge (R), random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), artificial neural network (ANN), and support vector regression (SVR).

300 When applied to field data, the models predict similar mean concentrations, within a few parts per billion, but vary more in their predictions of extreme concentrations (baseline and peak) (Figure 10). For the most part, this variation reflects the models performance on testing data. For example, LR, L, R, and SVR predict a negative baseline, which is also evident in the testing data predictions for these models, though less pronounced (Figure 9). The boosting models (GB, XGB) consistently predict the highest maximum concentration, which is reflective of these models' performances in the 95-100th percentile of testing data.

305 Overall, the models are correlated in their predictions, with correlation coefficients ranging from 0.88 to 1 between each model (supplemental).

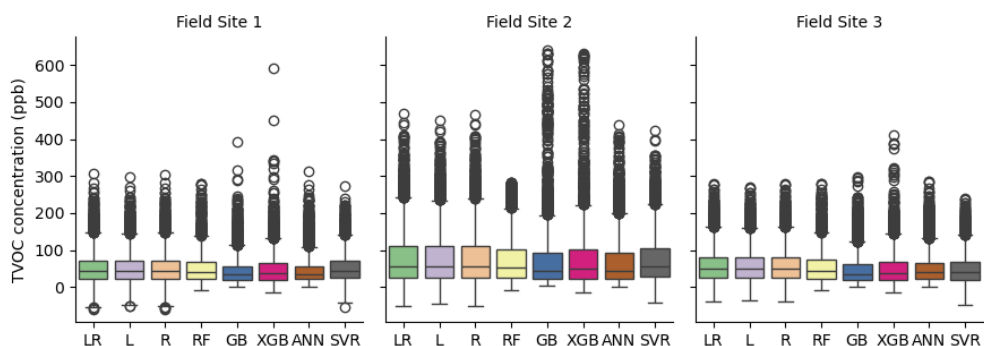


Figure 10. VOC concentration predictions at three field sites for each calibration model: linear regression (LR), lasso (L), ridge (R), random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), artificial neural network (ANN), and support vector regression (SVR)

3.2.2 BTEX calibration

The BTEX calibration models show little variability in performance across different test/train splits, except for data above the 95th percentile, where RMSE varied more than 5 ppb and MBE varied by more than 3 ppb for all models (Figure 11). R^2 was the lowest of the three pollutants, with a median value of around 0.55 across all runs and models (Figure S2). Lasso (L) had the worst performance overall, while the other two linear models (LR, R) show slightly better performance at the highest percentile and have the best overall RMSE and MBE.

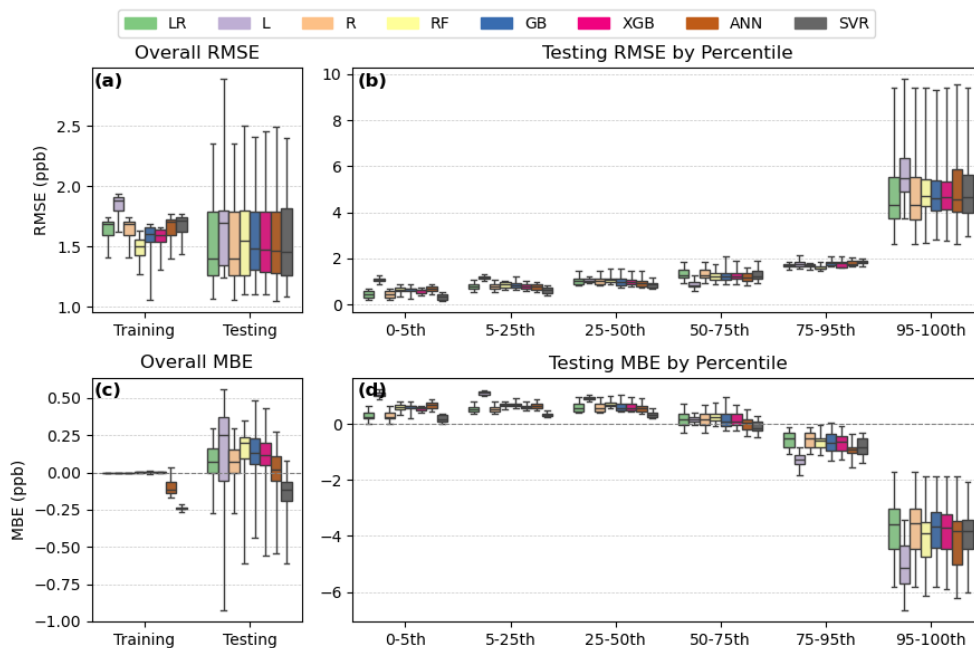


Figure 11. Performance statistics for BTEX calibration models, including overall RMSE (a), RMSE by percentile group (b), overall MBE (c), and MBE by percentile group (d). The calibration models are linear regression (LR), lasso (L), ridge (R), random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), artificial neural network (ANN), and support vector regression (SVR)

SVR shows less overestimation of baseline concentrations (0-25th percentile) than the other models. Comparing this model to ANN in Figure 11 reveals an important issue regarding overall MBE as a performance statistic. Both models show similar bias above the 50th percentile. However, ANN overpredicts lower concentrations (0-50th percentile) much more than SVR. As a result, ANN has a better overall MBE compared to SVR. This is important to consider because in a practical sense, a model that shows less overall bias by equally overpredicting the baseline and underpredicting peak values may not be more useful than a model that accurately predicts baseline concentrations but underpredicts peaks.

Figure 12 shows how the L model predicts far less variability than other models, with few data points predicted above 5 ppb. The L1 regularization used in L removed all inputs except TGS 2602, resulting in a poor model. The L2 regularization in R did not significantly change the model compared to LR.

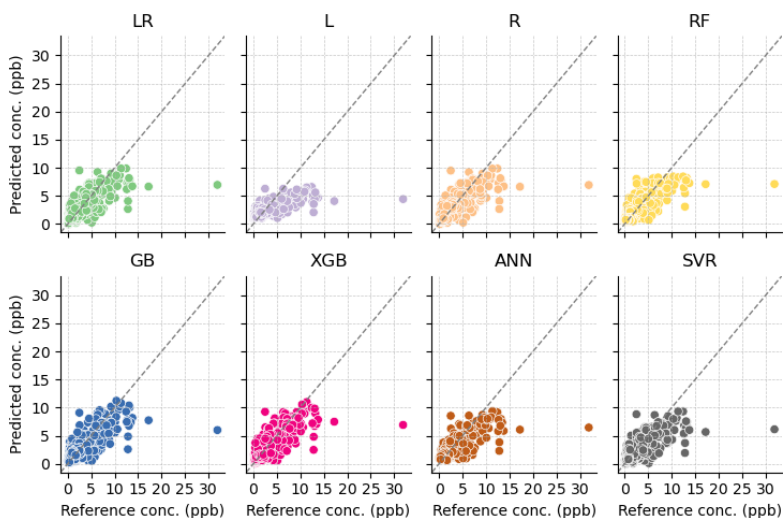


Figure 12. Predicted concentration versus reference concentration for BTEX calibration models: linear regression (LR), lasso (L), ridge (R), random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), artificial neural network (ANN), and support vector regression (SVR).

Like TVOC, the mean concentrations predicted at each field site show little variation between models, but variation is higher when comparing the entire range of data (Figure 12). L consistently has the smallest interquartile range, reflecting the lower variability predicted in the testing data. At sites 1 and 3, all models except L predict similar maximum concentrations. However, at site 2, RF, GB, and XGB predict much lower maximums than LR, R, ANN, and SVR. It is not clear why this inconsistency is occurring; however, it may indicate issues with transferability of the models for BTEX given their changing behavior across field sites. Overall, the models are still highly correlated with each other, with correlation coefficients between models at each field site ranging from 0.89 to 1 (supplementary).

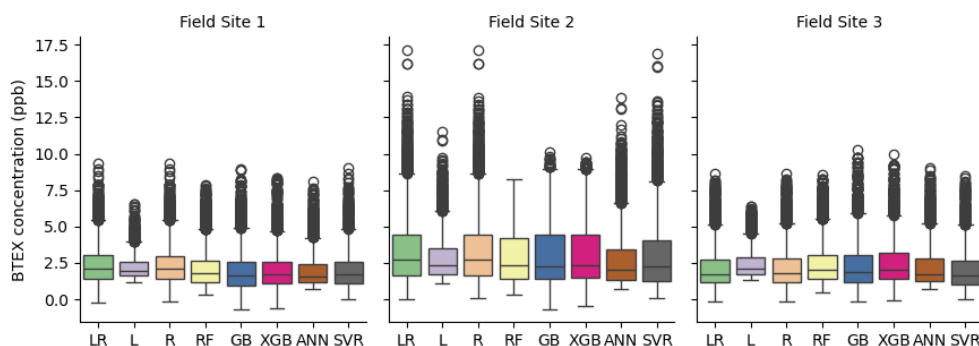


Figure 13. BTEX concentration predictions at three field sites for each calibration model: linear regression (LR), lasso (L), ridge (R), random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), artificial neural network (ANN), and support vector regression (SVR)

3.2.3 NO₂ calibration

330 Overall, all the NO₂ calibration models tend to overpredict concentrations below the 25th percentile and underpredict concentrations above the 50th percentile (Figure 14, panel d). However, underprediction of data above the 95th percentile is much less severe than TVOC and BTEX predictions, likely because the distribution of reference NO₂ concentrations is more balanced than the BTEX and TVOC concentration distributions, resulting in a much smaller range of peak concentrations to predict. RMSE remains much more consistent across percentiles for NO₂ compared to TVOC and BTEX (Figure 14, panel b).

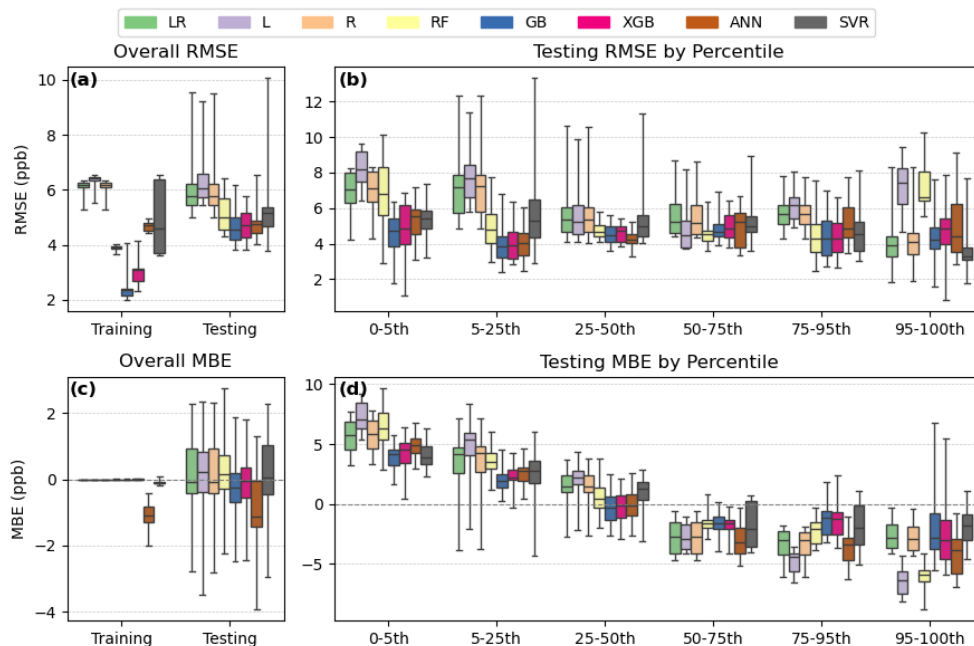


Figure 14. Performance statistics for NO₂ calibration models, including overall RMSE (a), RMSE by percentile group (b), overall MBE (c), and MBE by percentile group (d). The calibration models are linear regression (LR), lasso (L), ridge (R), random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), artificial neural network (ANN), and support vector regression (SVR)

335 The three linear-based models (LR, L, R) perform poorly for NO₂, particularly in the low percentile range of concentrations, where the models predict negative concentrations (Figure 15). Mean R² values were below 0.8 for these models, but mean R² for the non-linear models were above 0.8 (Figure S3). L1 regularization in the L model removes temperature. The poor performance of L across the high and low percentile ranges compared to LR indicates that all inputs (NO₂ target sensor, temperature, humidity, and time elapsed) are necessary for this calibration.

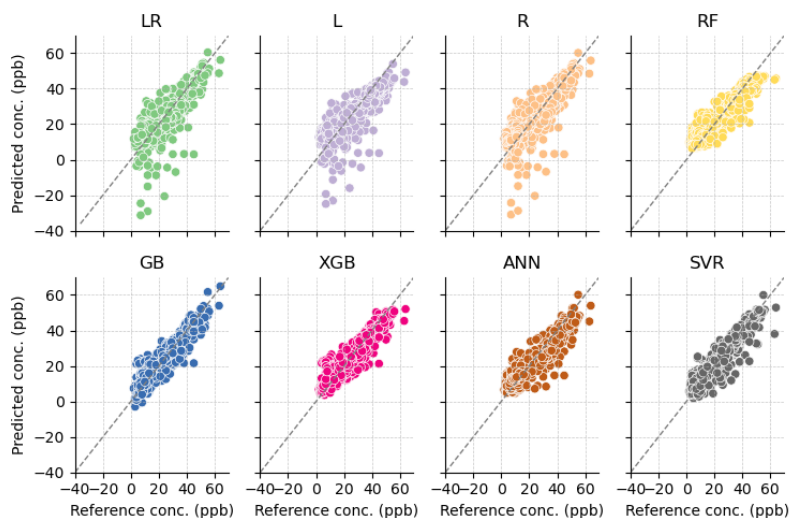


Figure 15. Predicted concentration versus reference concentration for NO₂ calibration models: linear regression (LR), lasso (L), ridge (R), random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), artificial neural network (ANN), and support vector regression (SVR).

340 Once again, differences in the field predictions between each model exist mainly in the extreme values (Figure 16). As
 341 expected from testing data performance, the linear models predict some highly negative concentrations. RF is also consistent
 342 with its testing performance, specifically its baseline overestimation, as it shows the highest minimum concentration for all
 343 three sites. The correlation coefficients between a linear model (LR, R, L) and a non-linear model (RF, GB, XGB, ANN, SVR)
 344 are lower (0.73-0.93) than the correlation coefficients between two linear or two non-linear models (0.89-1) (supplementary).
 345 This is likely because of the unique negative concentrations predicted by the linear models.

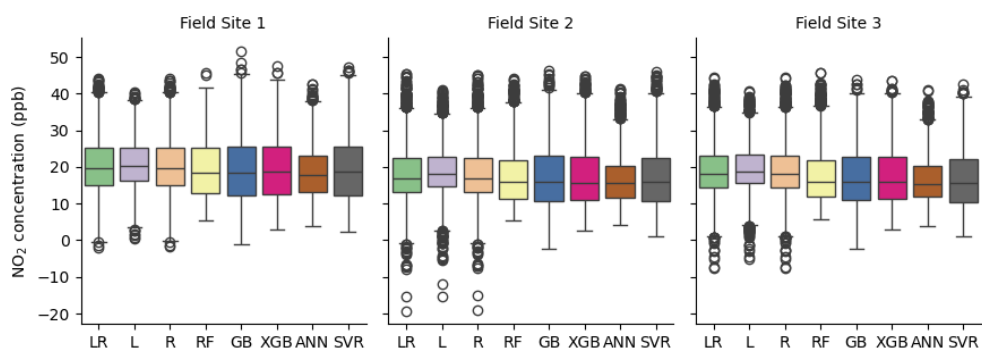


Figure 16. NO₂ concentration predictions at three field sites for each calibration model: linear regression (LR), lasso (L), ridge (R), random forest (RF), gradient boosting (GB), extreme gradient boosting (XGB), artificial neural network (ANN), and support vector regression (SVR)

3.3 Discussion

Across the three pollutants, the calibration models generally underestimated peak concentrations and overestimated baseline levels, though the extent of this bias varied with the distribution of the reference data. For TVOC and BTEX, which had more imbalanced distributions, peak underestimation was more pronounced, likely due to the wider range of high concentrations. 350 These results highlight the importance of evaluating model performance across the full concentration distribution, rather than relying solely on overall metrics. In many cases, biases at the extremes were masked in the overall statistics because the baseline overestimation and peak underestimation often offset each other around the median, resulting in overall MBE values near zero.

While there was broad variability in model performance across pollutants, some patterns still emerged. For example, our results also showed nearly identical performance for linear regression (LR) and ridge regression (R) across the three pollutants, 355 suggesting that L2 regularization was not beneficial for these datasets. Lasso regression also did not lead to improved predictive performance for linear regression, especially for BTEX and NO₂, where removing features through L1 regularization actually worsened model performance. To better understand the importance of each feature in response to the lasso performance, we plot the feature importances from the ensemble decision tree-based models (RF, GB, and XGB) in Figure 17. The feature importances were calculated using the Mean Decrease in Impurity using the Scikit-learn Python module. Based on Figure 360 17, it is clear that the target pollutant sensor features were more important than the temperature, humidity, and time elapsed features across all pollutants. However, given the reduced performance of lasso regression for BTEX and NO₂, it appears that even "less important" features can still play an essential role in model training. The increased importance of TGS 2602 for BTEX compared to TVOC aligns with previous work, which found TGS 2602 to be more sensitive to BTEX than TGS 2600 (Collier-Oxandale et al., 2019).

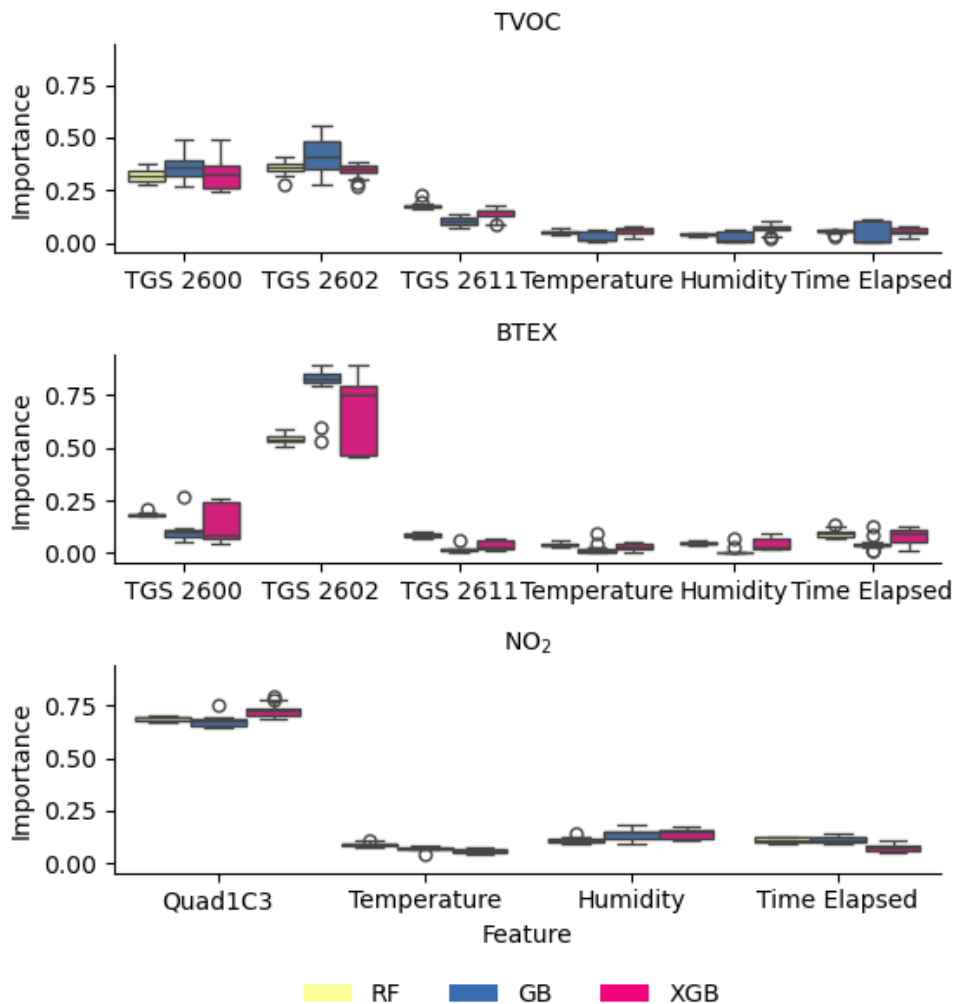


Figure 17. Feature importance for ensemble decision tree-based models (RF, GB, XGB) models for TVOC, BTEX, and NO₂.

365 Boosting methods, which are known for their improved ability to handle imbalanced data, reduced the underestimation of
 peak concentrations for TVOC. However, for BTEX, linear-based models (LR and R) actually outperformed boosting methods
 and other complex machine learning models for both peak concentrations and overall performance. For NO₂, the linear models
 performed the worst, especially at baseline concentrations, highlighting that no single model consistently performs best across
 all collocation datasets. When applied to field data, the models predicted similar mean concentrations for each site but had more
 370 variability in baseline and peak concentration predictions. Generally, the variability followed patterns evident in the testing
 data performance.

We built each calibration model multiple times, with differing test/train splits, to assess the robustness of each model.
 The performance metrics showed some variability, especially in the 95-100th percentile for BTEX and TVOC, based on the

test/train split. This was not surprising, given that the range of concentrations in the 95-100th percentile for the BTEX and
375 TVOC testing data could vary dramatically depending on the test/train split. These results motivate further study into how data
should be split into testing and training sets to ensure transferability and robustness. At a minimum, we suggest ensuring that
the testing data encompass the full range of concentrations measured during colocation.

4 Conclusions

Combining the two-step colocation process with a robust evaluation of multiple machine learning models led to satisfactory
380 calibration of TVOC, BTEX, and NO₂ sensors in this study. Our two-step colocation strategy supported model transferability
by ensuring the colocation encompassed more similar environmental conditions than was possible with only a pre- and post-
calibration. We also showed why a two-step colocation is preferable to a universal calibration model, as the harmonization
step was able to correct for the variability and drift unique to each individual sensor in a network. We found that calibration
model performance is not one-size-fits-all and is affected by the distribution of reference concentrations and the test/train split.
385 Even for models with little overall bias, percentile-based analysis revealed systemic overestimation of baseline concentrations
and underestimation of peaks across all eight models: linear regression, lasso, ridge, random forest, gradient boosting, extreme
gradient boosting, artificial neural network, and support vector regression.

Data availability. Reference data were provided courtesy of the South Coast Air Quality Management District. These data have not passed
through the normal review process and are therefore not quality-assured, and they are thus unofficial data. The HAQ-Pod data are available
390 at [doi].

Author contributions. Conceptualization: MH, JJ, CF; methodology: CF; software: CF; formal analysis: CF; investigation: CF, JP, EB, WS;
resources: JP, EB, WS, MH; writing - original draft: CF; writing - review & editing: JP, EB, WS, JJ, MH; visualization: CF; supervision: JJ,
MH; funding acquisition: JJ, MH.

Competing interests. The authors declare no competing interests. This research was funded by NIEHS 1R01ES033478.

395 *Acknowledgements.* Thank you to the South Coast Air Quality Monitoring District for providing reference instrumentation and to the Los
Angeles pod hosts for allowing us to collect field data in your neighborhood. Thank you also to Venezia Ramirez, Brittney Lu, Nancy Lam,
and Alexander Silverman for your support with pod calibration and maintenance.

References

- Barkjohn, K. K., Gantt, B., and Clements, A. L.: Development and application of a United States-wide correction for PM 2.5 data collected with the PurpleAir sensor, *Atmospheric Measurement Techniques*, 14, 4617–4637, 2021.
- Breiman, L.: Random forests, *Machine learning*, 45, 5–32, 2001.
- Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L., and J. Tallón-Ballesteros, A.: The impact of data normalization on the accuracy of machine learning algorithms: a comparative analysis, in: *International conference on soft computing models in industrial and environmental applications*, pp. 344–353, Springer, 2023.
- Casey, J. G. and Hannigan, M. P.: Testing the performance of field calibration techniques for low-cost gas sensors in new deployment locations: across a county line and across Colorado, *Atmospheric Measurement Techniques*, 11, 6351–6378, 2018.
- Castell, N., Dauge, F. R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., and Bartonova, A.: Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?, *Environment international*, 99, 293–302, 2017.
- 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*, pp. 785–794, 2016.
- Clements, A. L., Griswold, W. G., Rs, A., Johnston, J. E., Herting, M. M., Thorson, J., Collier-Oxandale, A., and Hannigan, M.: Low-cost air quality monitoring tools: from research to practice (a workshop summary), *Sensors*, 17, 2478, 2017.
- Collier-Oxandale, A., Casey, J. G., Piedrahita, R., Ortega, J., Halliday, H., Johnston, J., and Hannigan, M. P.: Assessing a low-cost methane sensor quantification system for use in complex rural and urban environments, *Atmospheric Measurement Techniques*, 11, 3569–3594, 2018.
- Collier-Oxandale, A., Wong, N., Navarro, S., Johnston, J., and Hannigan, M.: Using gas-phase air quality sensors to disentangle potential sources in a Los Angeles neighborhood, *Atmospheric environment*, 233, 117 519, 2020.
- Collier-Oxandale, A. M., Thorson, J., Halliday, H., Milford, J., and Hannigan, M.: Understanding the ability of low-cost MOx sensors to quantify ambient VOCs, *Atmospheric Measurement Techniques*, 12, 1441–1460, 2019.
- Commodore, A., Wilson, S., Muhammad, O., Svendsen, E., and Pearce, J.: Community-based participatory research for the study of air pollution: a review of motivations, approaches, and outcomes, *Environmental monitoring and assessment*, 189, 1–30, 2017.
- Concas, F., Mineraud, J., Lagerspetz, E., Varjonen, S., Liu, X., Puolamäki, K., Nurmi, P., and Tarkoma, S.: Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis, *ACM Transactions on Sensor Networks (TOSN)*, 17, 1–44, 2021.
- Considine, E. M., Reid, C. E., Ogletree, M. R., and Dye, T.: Improving accuracy of air pollution exposure measurements: Statistical correction of a municipal low-cost airborne particulate matter sensor network, *Environmental Pollution*, 268, 115 833, 2021.
- 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, *Sensors and Actuators B: Chemical*, 129, 750–757, 2008.
- Edgerton, S. A., Holdren, M. W., Smith, D. L., and Shah, J. J.: Inter-urban comparison of ambient volatile organic compound concentrations in US cities, *JAPCA*, 39, 729–732, 1989.
- Fanti, G., Borghi, F., Spinazzè, A., Rovelli, S., Campagnolo, D., Keller, M., Cattaneo, A., Cauda, E., and Cavallo, D. M.: Features and practicability of the next-generation sensors and monitors for exposure assessment to airborne pollutants: a systematic review, *Sensors*, 21, 4513, 2021.
- Frischmon, C., Crosslin, J., Burks, L., Weckesser, B., Hannigan, M., and Duderstadt, K.: Detecting air pollution episodes and exploring their impacts using low-cost sensor data and simultaneous community symptom and odor reports, *Environmental Research Letters*, 2025a.

- 435 Frischmon, C., Silberstein, J., Guth, A., Mattson, E., Porter, J., and Hannigan, M.: Improving the quantification of peak concentrations for air quality sensors via data weighting, *EGUsphere*, 2025, 1–19, 2025b.
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42, 463–484, 2011.
- 440 Griffith, D. W.: Synthetic calibration and quantitative analysis of gas-phase FT-IR spectra, *Applied spectroscopy*, 50, 59–70, 1996.
- Haagen-Smit, A. J., Bradley, C., and Fox, M.: Ozone formation in photochemical oxidation of organic substances, *Industrial & Engineering Chemistry*, 45, 2086–2089, 1953.
- Hoerl, A. E. and Kennard, R. W.: Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, 12, 55–67, 1970.
- Hong, G.-H., Le, T.-C., Lin, G.-Y., Cheng, H.-W., Yu, J.-Y., Dejchanchaiwong, R., Tekasakul, P., and Tsai, C.-J.: Long-term field calibration
445 of low-cost metal oxide VOC sensor: Meteorological and interference gas effects, *Atmospheric Environment*, 310, 119955, 2023.
- Johansson, J. K., Mellqvist, J., Samuelsson, J., Offerle, B., Lefer, B., Rappenglück, B., Flynn, J., and Yarwood, G.: Emission measurements of alkenes, alkanes, SO₂, and NO₂ from stationary sources in Southeast Texas over a 5 year period using SOF and mobile DOAS, *Journal of Geophysical Research: Atmospheres*, 119, 1973–1991, 2014.
- Johnson, N. E., Bonczak, B., and Kontokosta, C. E.: Using a gradient boosting model to improve the performance of low-cost aerosol
450 monitors in a dense, heterogeneous urban environment, *Atmospheric environment*, 184, 9–16, 2018.
- Kampa, M. and Castanas, E.: Human health effects of air pollution, *Environmental pollution*, 151, 362–367, 2008.
- Karagulian, F., Barbieri, M., Kotsev, A., Spinelle, L., Gerboles, M., Lagler, F., Redon, N., Crunaire, S., and Borowiak, A.: Review of the performance of low-cost sensors for air quality monitoring, *Atmosphere*, 10, 506, 2019.
- Krogh, A.: What are artificial neural networks?, *Nature biotechnology*, 26, 195–197, 2008.
- 455 Kumar, V. and Sahu, M.: Evaluation of nine machine learning regression algorithms for calibration of low-cost PM_{2.5} sensor, *Journal of Aerosol Science*, 157, 105809, 2021.
- Laaksonen, A., Kulmala, M., O'Dowd, C., Joutsensaari, J., Vaattovaara, P., Mikkonen, S., Lehtinen, K., Sogacheva, L., Dal Maso, M., Aalto, P., et al.: The role of VOC oxidation products in continental new particle formation, *Atmospheric Chemistry and Physics*, 8, 2657–2665, 2008.
- 460 Leidinger, M., Sauerwald, T., Reimringer, W., Ventura, G., and Schütze, A.: Selective detection of hazardous VOCs for indoor air quality applications using a virtual gas sensor array, *Journal of Sensors and Sensor Systems*, 3, 253–263, 2014.
- 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., et al.: Evaluating the performance of low cost chemical sensors for air pollution research, *Faraday discussions*, 189, 85–103, 2016.
- Li, Z., Ma, Z., Zhang, Z., Zhang, L., Tian, E., Zhang, H., Yang, R., Zhu, D., Li, H., Wang, Z., et al.: High-density volatile organic compound
465 monitoring network for identifying pollution sources, *Science of The Total Environment*, 855, 158872, 2023.
- Liang, L.: Calibrating low-cost sensors for ambient air monitoring: Techniques, trends, and challenges, *Environmental Research*, 197, 111163, 2021.
- Liu, X., Jayaratne, R., Thai, P., Kuhn, T., Zing, I., Christensen, B., Lamont, R., Dunbabin, M., Zhu, S., Gao, J., et al.: Low-cost sensors as an alternative for long-term air quality monitoring, *Environmental research*, 185, 109438, 2020.
- 470 Malings, C., Tanzer, R., Haurlyliuk, A., Kumar, S. P., Zimmerman, N., Kara, L. B., Presto, A. A., and Subramanian, R.: Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring, *Atmospheric Measurement Techniques*, 12, 903–920, 2019.

- Malyan, V., Kumar, V., Moni, M., Sahu, M., Prakash, J., Choudhary, S., Raliya, R., Chadha, T. S., Fang, J., and Biswas, P.: Assessing the spatial transferability of calibration models across a low-cost sensors network, *Journal of aerosol science*, 181, 106437, 2024.
- 475 Masiol, M., Squizzato, S., Chalupa, D., Rich, D., and Hopke, P.: Evaluation and field calibration of a low-cost ozone monitor at a regulatory urban monitoring station. *Aerosol Air Qual. Res* 18, 2029–2037, 2018.
- Mellqvist, J., Samuelsson, J., Isoz, O., Brohede, S., Andersson, P., Ericsson, M., and Johansson, J.: Emission measurements of VOCs, NO₂ and SO₂ from the refineries in the south coast air basin using solar occultation flux and other optical remote sensing methods, *FluxSense/SCAQMD-2015*, 2017.
- 480 Natekin, A. and Knoll, A.: Gradient boosting machines, a tutorial, *Frontiers in neurorobotics*, 7, 21, 2013.
- Nowack, P., Konstantinovskiy, L., Gardiner, H., and Cant, J.: Machine learning calibration of low-cost NO₂ and PM₁₀ sensors: Non-linear algorithms and their impact on site transferability, *Atmospheric Measurement Techniques*, 14, 5637–5655, 2021.
- Okorn, K. and Hannigan, M.: Applications and limitations of quantifying speciated and source-apportioned vocs with metal oxide sensors, *Atmosphere*, 12, 1383, 2021a.
- 485 Okorn, K. and Hannigan, M.: Improving Air Pollutant Metal Oxide Sensor Quantification Practices through: An Exploration of Sensor Signal Normalization, Multi-Sensor and Universal Calibration Model Generation, and Physical Factors Such as Co-Location Duration and Sensor Age, *Atmosphere*, 12, 645, 2021b.
- Okorn, K. and Iraci, L. T.: An overview of outdoor low-cost gas-phase air quality sensor deployments: current efforts, trends, and limitations, *Atmospheric Measurement Techniques*, 17, 6425–6457, 2024.
- 490 Okorn, K., Jimenez, A., Collier-Oxandale, A., Johnston, J., and Hannigan, M.: Characterizing methane and total non-methane hydrocarbon levels in Los Angeles communities with oil and gas facilities using air quality monitors, *Science of the Total Environment*, 777, 146194, 2021.
- Ou-Yang, C.-F., Liao, W.-C., Chang, C.-C., Hsieh, H.-C., and Wang, J.-L.: Guided episodic sampling for capturing and characterizing industrial plumes, *Atmospheric Environment*, 174, 188–193, 2018.
- 495 Raheja, G., Harper, L., Hoffman, A., Gorby, Y., Freese, L., O’Leary, B., Deron, N., Smith, S., Auch, T., Goodwin, M., et al.: Community-based participatory research for low-cost air pollution monitoring in the wake of unconventional oil and gas development in the Ohio River Valley: Empowering impacted residents through community science, *Environmental Research Letters*, 17, 065006, 2022.
- Robin, Y., Amann, J., Baur, T., Goodarzi, P., Schulteaibert, C., Schneider, T., and Schütze, A.: High-performance VOC quantification for IAQ monitoring using advanced sensor systems and deep learning, *Atmosphere*, 12, 1487, 2021.
- 500 Rothman, L. S., Jacquemart, D., Barbe, A., Benner, D. C., Birk, M., Brown, L., Carleer, M., Chackerian Jr, C., Chance, K., Coudert, L. e. a., et al.: The HITRAN 2004 molecular spectroscopic database, *Journal of quantitative spectroscopy and radiative transfer*, 96, 139–204, 2005.
- Sá, J., Chojer, H., Branco, P., Alvim-Ferraz, M., Martins, F., and Sousa, S.: Two step calibration method for ozone low-cost sensor: Field experiences with the UrbanSense DCUs, *Journal of Environmental Management*, 328, 116910, 2023.
- 505 Sharpe, S. W., Johnson, T. J., Sams, R. L., Chu, P. M., Rhoderick, G. C., and Johnson, P. A.: Gas-phase databases for quantitative infrared spectroscopy, *Applied spectroscopy*, 58, 1452–1461, 2004.
- Silberstein, J., Wellbrook, M., and Hannigan, M.: Utilization of a Low-Cost Sensor Array for Mobile Methane Monitoring, *Sensors*, 24, 519, 2024.
- Smola, A. J. and Schölkopf, B.: A tutorial on support vector regression, *Statistics and computing*, 14, 199–222, 2004.

- 510 Song, Y.-Y. and Ying, L.: Decision tree methods: applications for classification and prediction, *Shanghai archives of psychiatry*, 27, 130, 2015.
- Spinelle, L., Gerboles, M., Kok, G., Persijn, S., and Sauerwald, T.: Performance evaluation of low-cost BTEX sensors and devices within the EURAMET key-VOCs project, in: *Proceedings*, vol. 1, p. 425, MDPI, 2017a.
- Spinelle, L., Gerboles, M., Kok, G., Persijn, S., and Sauerwald, T.: Review of portable and low-cost sensors for the ambient air monitoring
515 of benzene and other volatile organic compounds, *Sensors*, 17, 1520, 2017b.
- Srishti, S., Agrawal, P., Kulkarni, P., Gautam, H. C., Kushwaha, M., and Sreekanth, V.: Multiple PM low-cost sensors, multiple seasons' data, and multiple calibration models, *Aerosol and Air Quality Research*, 23, 220-428, 2023.
- Stansbury, D.: *A gentle introduction to artificial neural networks*, The Clever Machine, 2014.
- Tibshirani, R.: Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society Series B: Statistical Methodology*,
520 58, 267–288, 1996.
- Vikram, S., Collier-Oxandale, A., Ostertag, M. H., Menarini, M., Chermak, C., Dasgupta, S., Rosing, T., Hannigan, M., and Griswold, W. G.: Evaluating and improving the reliability of gas-phase sensor system calibrations across new locations for ambient measurements and personal exposure monitoring, *Atmospheric Measurement Techniques*, 12, 4211–4239, 2019.
- Wang, Z.: Evaluating the efficacy of machine learning in calibrating low-cost sensors, *Applied and Computational Engineering*, 71, 30–38,
525 2024.
- Wesolowski, M. and Suchacz, B.: Artificial neural networks: theoretical background and pharmaceutical applications: a review, *journal of aoac international*, 95, 652–668, 2012.
- Yurko, G., Roostaei, J., Dittrich, T., Xu, L., Ewing, M., Zhang, Y., and Shreve, G.: Real-time sensor response characteristics of 3 commercial metal oxide sensors for detection of BTEX and chlorinated aliphatic hydrocarbon organic vapors, *Chemosensors*, 7, 40, 2019.
- 530 Zamora, M. L., Buehler, C., Datta, A., Gentner, D. R., and Koehler, K.: Identifying optimal co-location calibration periods for low-cost sensors, *Atmospheric measurement techniques*, 16, 169, 2023.
- Zimmerman, N., Presto, A. A., Kumar, S. P., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., et al.: A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring, *Atmospheric Measurement Techniques*, 11, 291–313, 2018.