Evaluation of Low-cost Air Quality Sensor Calibration Models

KASIMIR AULA, EEMIL LAGERSPETZ, PETTERI NURMI, and SASU TARKOMA,
University of Helsinki, Finland

We contribute a novel model evaluation technique that divides available measurements into training and testing sets in a way that adheres to the requirements imposed on professional monitoring stations. We perform extensive and systematic experiments with a wide range of state-of-the-art calibration models to demonstrate that our approach provides accurate insights about the performance of calibration models in real-world deployments, while at the same time highlighting issues with evaluation techniques used in previous works. Among others, our results show that although trained and tested in the same location, calibration errors can exhibit deviation up to 116% depending on the evaluation protocol that is being adopted. We also demonstrate that models trained with continuous data can suffer up to 76% greater error when tested with data coming from diverse environmental conditions. In contrast, when models are trained and tested with our method, the variability of errors is significantly reduced and the robustness of calibration models is significantly improved. The overall performance improvements depend on pollutant concentration, ranging from 10% for low concentrations to 90% for high concentrations that represent conditions that are most dangerous for human health.

CCS Concepts: • **Applied computing** → **Environmental sciences**; • **Hardware** → *Sensor applications and deployments*; • **Human-centered computing** → *Ubiquitous and mobile computing systems and tools;*

Additional Key Words and Phrases: Machine learning, sensor calibration, model evaluation, data segmentation

ACM Reference format:
Kasimir Aula, Eemil Lagerspetz, Petteri Nurmi, and Sasu Tarkoma. 2022. Evaluation of Low-cost Air Quality Sensor Calibration Models. *ACM Trans. Sensor Netw.* 18, 4, Article 72 (December 2022), 32 pages.
https://doi.org/10.1145/3512889

1 INTRODUCTION

Low-cost pollution sensors, costing less than $2,500, are rapidly emerging as a powerful solution for acquiring pollutant information at high spatiotemporal resolution [8, 23, 47, 48]. Increasing the

This work was supported in part by the European Union through the Urban Innovative Action Healthy Outdoor Premises for Everyone (UIA03-240), Business Finland via the MegaSense program, and the Academy of Finland through the grants 324576 (MegaSense) and 339614 (Foundations of Pervasive Sensing Systems). The authors thank the Finnish Computing Competence Infrastructure (FCCI) for supporting this project with computational and data storage resources through the Ukko2 infrastructure (persistent identifier urn:nbn:fi:research-infras-2016072533).
Authors' address: K. Aula, E. Lagerspetz, P. Nurmi (corresponding author), and S. Tarkoma, University of Helsinki, P. O. 68 (Pietari Kalmin Katu 5), Helsinki, Finland, FI-00014; emails: {kasimir.aula, eemil.lagerspetz, petteri.nurmi, sasu.tarkoma}@helsinki.fi.

1550-4859/2022/12-ART72 $15.00
https://doi.org/10.1145/3512889

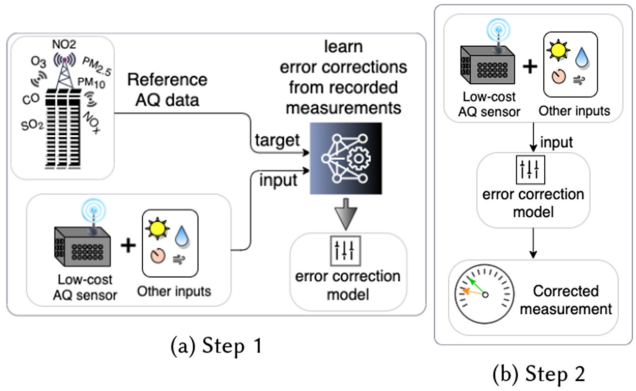


Fig. 1. Machine learning calibration model adjusts low-cost sensor's measurements based on differences with reference measurements. The model is trained in Step 1 by comparing a collection of low-cost sensor measurements to measurements from a reference station. In Step 2, the model can be used for correcting incoming measurements.

scale of monitoring is essential particularly in densely packed urban environments where pollutant concentrations can differ significantly even at a short distance apart [37, 43]. Unfortunately, low-cost sensors come with a significant caveat, as ensuring the validity and quality of the measurements is challenging. Indeed, evaluations of low-cost sensors have shown the measurements to have low correspondence with high-precision pollutant monitoring stations, such as those used by national infrastructure [2, 4]. The accuracy of low-cost sensor measurements can be improved by taking advantage of machine learning–based calibration [8]. The idea in these approaches, illustrated in Figure 1, is to learn a calibration (or correction) function that compensates for errors in the pollutant measurements using information about environmental variables and other pollutants. Machine learning–based calibration of low-cost sensors is currently a highly active research area [8] and a wide range of techniques, ranging from variations of commonly used machine learning techniques, such as neural networks [7, 14, 35, 47], ensemble methods [31, 57], or regression models [9, 33, 46], to sophisticated deep learning–based approaches [53], have been proposed. Such techniques can provide significant improvements to measurement accuracy, e.g., Lee et al. [29] report improvements of over 80%, and Lin et al. [31] improvements between 6.1% and 133% for $PM_{2.5}$ (compared to simple methods) and similar results have been obtained for other pollutants [5, 14, 53, 54].

While the algorithmic aspects of calibration have received significant attention in the literature, unfortunately the *evaluation* of such techniques has received much less attention. Whereas professional measurement instruments and national reference stations are subject to rigorous and standardized evaluation protocols that require testing against low and high concentrations, considering the effect of temperature, humidity, and other potential interferents, and considering multiple devices and testing locations [15], evaluations of low-cost calibration techniques tend to rely on data from proprietary deployments with varying lengths, pollutant concentrations, and other characteristics and use statistical model evaluation protocols, such as holdout evaluation or cross-validation (see Section 8). These protocols do not take into account the characteristics of air quality measurements and fail to account for requirements placed on professional stations. Indeed, common statistical model evaluation protocols rely on time-based splits that can give optimistic impressions about the performance of a model as the measurements can be highly sensitive to short-term autocorrelation, seasonality effects, sensor drift, or other temporal effects. Indeed, neither

Table 1. Regulatory Standards Set by the EU

	EU Regulatory Standard	Temporal limit
CO	10	8-h mean
NO_2	40	annual mean
	200	1-h mean
O_3	120	8-h mean
SO_2	350	1-h mean
	125	24-h mean
	20	annual mean
PM_{10}	50	24-h mean
	20	annual mean
$PM_{2.5}$	25	annual mean

Units are in $\mu g/m^3$, except for CO (mg/m^3). As harmful levels are set in hours, it is crucial that calibrated values from low-cost monitors can be trusted.

short-term continuous training-testing split nor cross-validation are guaranteed to contain data covering diverse environmental conditions [22], and therefore they only indicate performance on the specific evaluation data. This lack of correspondence with standardized evaluation procedures makes it difficult to properly assess the validity of the calibrated measurements, making it difficult to take advantage of the outputs of low-cost sensors in scientific studies, decision making, and pollution monitoring as a whole. Besides impacting the usefulness of the corrected measurements, the limitations of current evaluation models also make it difficult to reliably compare different calibration models or to assess their suitability for different deployment locations and timescales. Finally, ensuring robust performance is essential for ensuring the information is actionable and can be used to derive accurate insights to help mitigate adverse health effects resulting from pollution. Indeed, according to regulatory standards, shown in Table 1, adverse health effects can result even from few hours of exposure to pollutants. Therefore model evaluation should yield true performance estimates and not merely test the model's capability to estimate air quality in very restricted environmental conditions.

We contribute by developing a novel statistical evaluation technique that has been designed to overcome the key limitations of model evaluation techniques used in existing works on the evaluation of low-cost air quality sensors. Our technique has been designed to adhere to evaluation requirements of professional-grade measurement stations and capture key characteristics of measurements. Our approach uses data-driven criteria that are inspired by statistical sampling techniques to split measurements into training and testing sets so that generality of the measurements against temporal, spatial, and distributional characteristics can be assessed. We validate and demonstrate the benefits of our evaluation protocol through extensive benchmarks that have been conducted using a public dataset. In our experiments, we consider 10 calibration models of different complexity level and compare performance implications given by our protocol against representative examples of the different evaluation protocols considered in previous works. Among others, our results show that although trained and tested in the same location, calibration errors can exhibit deviation up to 116% depending on the evaluation protocol that is being adopted. We also demonstrate that models trained with continuous data can suffer up to 76% greater error when tested with data coming from diverse environmental conditions. In contrast, when models are trained and tested with our method, the variability of errors is significantly reduced and the robustness of calibration models is significantly improved. The magnitude of improvements depends on the concentration of pollutants, ranging from 10% for low pollutant concentrations up

to 90% to high pollutant concentrations, which present the most dangerous conditions for human health.

Summary of Contributions:

- **Novel evaluation framework** for machine learning calibration models that adheres to the requirements imposed on professional monitoring stations.
- **Extensive benchmarks** that consider 10 different calibration models and four different evaluation protocols to demonstrate that our approach yields better impressions about the capabilities and restrictions of calibration models used in low-cost air quality sensing.
- **Novel insights** on calibration model performance. Among others, we show that calibration errors can exhibit significant deviation of up to 116% depending on the evaluation protocol that is being adopted. We also demonstrate that models trained with continuous data can suffer up to 76% greater error when tested with data coming from diverse environmental conditions.
- **Improved robustness and training performance** by limiting training data to periods containing differing environmental conditions. Our experiments show this boosts model performance while enhancing the robustness and the generality of the calibration models.

2 MOTIVATION

The performance of professional-grade reference instruments is governed by stringent metrological requirements that require testing against high and low concentrations, assessing performance in at least two different locations, and ensuring sufficient stability over time [16]. This contrasts with research on low-cost sensor calibration where the evaluations are typically carried out using data from a single location and using common statistical model assessment techniques, such as holdout validation or cross-validation, without examining the generality of the techniques [7, 19, 28, 36]. We next conduct small-scale experiments to highlight some of the key issues with these evaluation techniques, thus motivating the need for improved evaluation protocols. The experiments are carried out using a public dataset[1] [51] containing hourly measurements of several gaseous pollutants' measurements from a reference sensor and a low-cost metal-oxide sensor together with measurements from temperature and humidity sensors; see Section 4 for a more detailed description of the data.

Autocorrelation: Both air pollutants and environmental variables are strongly autocorrelated [25], which can cause information to leak from the training data into the testing data and result in overly optimistic performance estimates [22]. This issue is particularly problematic when the measurements cover a short time period or when the environmental variables contain little variation. Information leakage can give the models a strong performance gain in the short-term but degrades performance in long-term predictions due to the model overfitting on short-term correlation patterns between the pollutants and environmental variables. Table 2 shows the autocorrelation for different environmental variables (temperature and relative and absolute humidity) and gaseous pollutants (CO, NO_2, NO_x) in the dataset. From the table, we can observe that the autocorrelation values remain high for several weeks. These correlations need to be accounted for in the evaluation to ensure the results are robust. We later show that simply ensuring temporal separation alone is not sufficient, as it does not guarantee the characteristics of the training and testing sets to differ.

To show how autocorrelation affects the performance of low-cost calibration techniques, we split the data into segments containing 1 or more weeks of measurements and sort the segments

[1] https://archive.ics.uci.edu/ml/datasets/Air+Quality.

Table 2. Features Often Used in Machine Learning Calibration Exhibit Strong Autocorrelation That Persists for Weeks

n=	T	RH	AH	CO	NO_2	NO_x
1	0.86	0.48	0.62	0.44	0.57	0.56
2	0.82	0.41	0.54	0.43	0.55	0.53
4	0.77	0.41	0.47	0.42	0.5	0.53
8	0.54	0.34	0.3	0.34	0.48	0.44

The table shows autocorrelation computed from hourly measurements with lags equal to n weeks for temperature, relative and absolute humidity, CO, NO_2, and NO_x from the reference monitor. This is a data property that must be addressed and separated from model behavior and calibration efficiency.

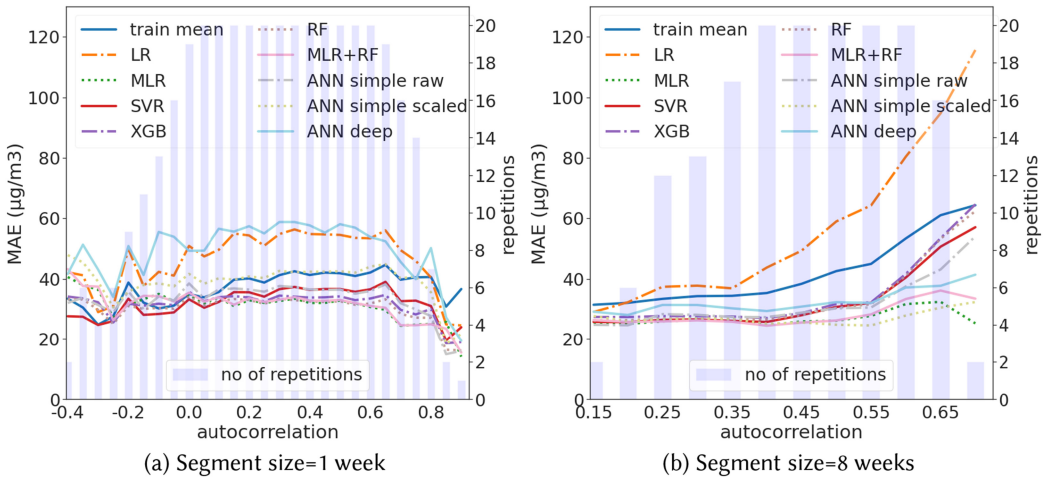
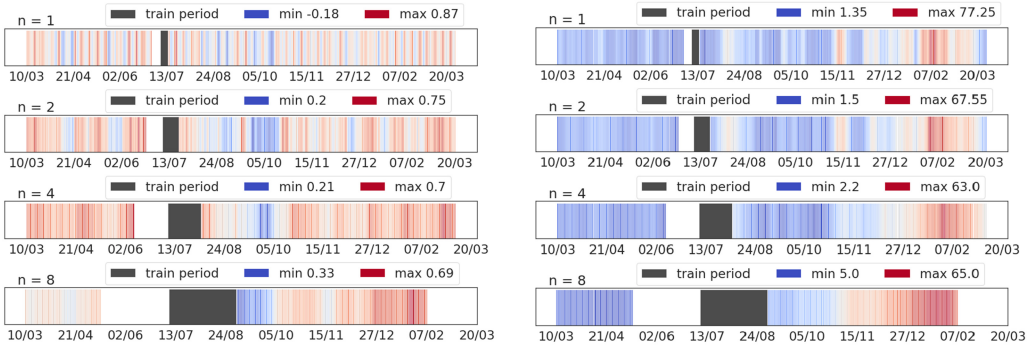


Fig. 2. The accuracy of calibration varies as target variable's (NO_2) autocorrelation between training and testing data changes. When the segment size of training and testing sets increases, i.e., the longer the continuous period of measurements that is considered, model predictions tend to reflect more general pollution levels than the actual required corrections, which results in higher calibration errors. The error in the y-axis is mean absolute error.

using autocorrelation.[2] We then construct training and testing datasets from these segments and evaluate the performance of different machine learning techniques used in the literature (see Section 4 for details of the techniques). Figure 2 shows the performance of the calibration models as a function of autocorrelation between training and testing data. In Figure 3(a) we see that, when the segments contain 1 week of data, the model errors are rather stable until the autocorrelation exceeds 0.8 and the errors decrease significantly when the autocorrelation is higher. Figure 3(b) shows that when the segments cover 8 weeks of data the autocorrelation's effect to model error is more significant. Note that autocorrelation measures whether the changes in the segments follow

[2]Autocorrelation has been calculated with a lag of one but using segments instead of points. Thus, for segment size of 1 week, the autocorrelation measures the similarity between successive weeks, and for segment size 8 weeks the autocorrelation measures similarity of successive 2-month periods.



(a) Autocorrelation (b) Distribution difference

Fig. 3. Autocorrelation (a) and distribution difference (difference between first and third quartiles) (b) between training and testing sets for NO_2 with different segment (number of continuous weeks) lengths. The color scale ranges from dark blue to dark red. For autocorrelation dark blue indicates correlation of −1 and dark red correlation of 1. For distribution differences, dark blue indicates the smallest difference and dark red the greatest difference (in $\mu g/m^3$) between training and testing data.

the same trend, not whether they are similar. As a result, the data can manifest similar patterns yet come from different distributions as they cover different seasons. This effect can be seen from higher autocorrelation in measurements resulting in higher error when the segment size is several months long. On shorter segment sizes these effects have more immediate effect and decrease the autocorrelation, thus also reducing model performance as a function of autocorrelation. We also see that the range of autocorrelation values is much smaller, even if the data span an entire year (split into segments of 8 weeks). These patterns are only partially explained by seasonal changes and highlight how both the total length of data and the length of the segments considered in the evaluation affect performance. Ensuring the evaluation results of low-cost calibration techniques are robust thus necessitates considering both the autocorrelation and the time window of the correlation patterns.

Temporal difference: A naïve solution for reducing the effect of autocorrelation is to use measurements from two periods that are sufficiently far apart. We next demonstrate that this is not sufficient, as it does not guarantee meaningful distributional differences to guarantee robustness of the models. We consider NO_2 as the target variable and use four segment lengths (1, 2, 4, and 8 weeks). We separately consider (i) the autocorrelation between training and testing data using a fixed lag of 1 week and (ii) the difference in distribution between the training and the testing data. For the latter, we choose one segment as the training period, sort the remaining segments by their similarity with the training segment, and progressively use the remaining measurements as testing measurements. We measure similarity using the average distance of the first and third quartile of the respective distributions. This effectively corresponds to comparing the difference in data spread (or variance) between two distributions. As most machine learning models use the interquartile range (or variance) for normalizing inputs, difference in data spread thus captures the effect data variance has on the inputs for the machine learning models. An alternative would be to use a full distributional metric, such as Kullback–Leibler divergence or Hellinger distance, but these measures compare the entire distribution and thus are more sensitive to outliers.

The results with four segment sizes (of 1, 2, 4, and 8 weeks) are shown in Figure 3. We observe that, in most cases, extending the time gap between training and testing data increases the difference in corresponding distributions but that the magnitude of autocorrelation varies much more

erratically. This is particularly the case for short segments that contain 1 week of data. While increasing temporal distance between testing and training sets typically increases difference in the training and testing distributions, it does not guarantee that sets are not correlated. Indeed, as environmental parameters have strong autocorrelations, periods that are temporally apart but share similar environmental characteristics result in short-term correlation patterns leaking from the training data into the the testing data. Note that correlation does not necessarily lead to a better performance for the models, but it easily leads the model to learn the temporal dependency instead of the actual relationship between pollutants, environmental variables, and other variables used in calibration. The result also implies that the naïve solution of enforcing a time gap is insufficient for ensuring a robust evaluation, highlighting how non-trivial it is to select training and testing datasets that are good at assessing robustness of the model performance.

Similarity in Distributions: The evaluation process is further complicated by irregular pollution events that change the target pollutant's value distribution. Simpler machine learning models with one-dimensional input may result only in modest accuracy due to cross-sensitivities, i.e., interference from non-target gases or aerosols affecting sensor readings, and feature cross-correlations [8, 30], whereas complex models with multidimensional input require a large amount of data to avoid overfitting on the most common feature and target values. Even in the simplest case of training and testing a calibration model for a deployment in a single site, the calibration model may achieve low test error if the testing data are similar to training data. However, this model may fail drastically when changes occur, e.g., due to infrequent or extreme weather events. This is due to the training data only partially capturing the true distribution of values, and thus the model has limited generality on the parts of the distribution that are not part of the training data. These issues are further exacerbated when distribution changes are greater between training and testing environment, e.g., when going from lab calibration to real-world deployment, or attempting to generalize the model from one city to another with very different pollution levels or distributions. Previous work on evaluating air quality calibration models has mainly used (short-term) continuous data periods for training, testing, or both phases [5, 29, 45, 47, 57]. Table 3 shows how the magnitude of standard deviation increases as we choose a larger dataset, which highlights how short-term data tend to have high similarity and thus limited distributional differences. This means that evaluations with short time periods are unlikely to reflect performance with a feature's true variation or to satisfy variability criteria placed on professional-grade measurement instruments.

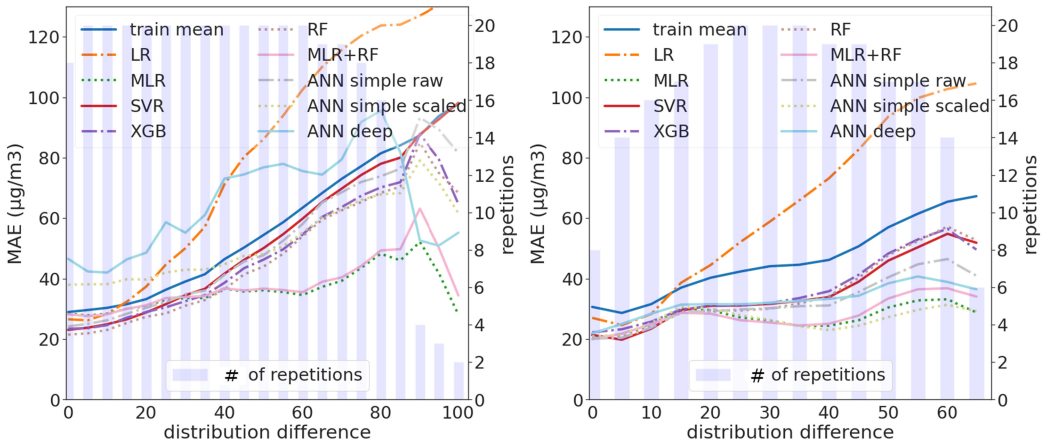
Also changes in target distribution affect model performance. Figure 4 shows how calibration errors increase as the training data's target distribution differs from the target distribution of the testing data. Larger distributional differences can be observed when the segment size is smaller in Figure 4(a), but using a larger segment size of 8 weeks (Figure 4(b)) also shows a similar relationship. This means that simply increasing the size of the testing set alone is not sufficient as the distributional differences also need to be taken into account. Note that standard machine learning techniques, such as sample reweighting and bootstrapping [39, 41], are also insufficient, as they focus on balancing the training set to give equal weights of different parts of the distribution rather than assessing model performance in diverse conditions. Unlike our approach, these techniques also do not account for temporal dependencies.

Summary: Taken together, the evaluation results highlight—in line with previous research—how autocorrelation, temporal variations, and distributional differences have a significant impact on the accuracy and robustness of low-cost calibration models. While real-world data offer no control over the distribution of target values or the environmental or other events that may impact the values of pollutants or environmental parameters, at a minimum the calibration evaluation should

Table 3. Averaged SD for an n Week Long Dataset for Temperature (°C), Relative Humidity (%), CO (mg/m^3), NO_x (ppb), and NO_2 ($\mu g/m^3$)

n =	T	RH	CO	NO_x	NO_2
1	4.15	13.38	1.32	145.04	38.69
2	4.4	14.48	1.33	147.11	39.74
3	4.57	14.46	1.35	151.1	40.78
4	4.63	15.02	1.36	151.23	41.3
5	4.8	14.76	1.32	148.9	40.67
6	4.88	15.38	1.36	154.59	41.72
7	5.04	15.33	1.39	158.99	42.18
8	5.3	15.35	1.38	154.62	43.04
9	5.04	15.37	1.35	157.33	41.61
10	5.52	15.36	1.39	164.6	43.59
11	5.44	15.57	1.33	155.06	42.15
12	5.51	15.9	1.4	159.63	43.32
All data	8.84	17.31	1.46	211.51	48.66

For comparison purposes, the lowest line in the table shows SD for all available data. The values were produced by computing standard deviations from consecutive non-overlapping periods and averaging the results.



(a) Segment size=1 week (b) Segment size=8 weeks

Fig. 4. The performance of calibration decreases as target variable's (NO_2) distribution in testing data differs from the distribution in training data. The error in the y-axis is mean absolute error.

be subjected to testing where all three aspects are taken into consideration. While previous research has identified these effects to have an influence on evaluation [11, 44, 49], thus far no unified evaluation procedure that can address all of these aspects in model evaluation, and also improve model training, has been proposed. We next introduce an evaluation framework that has been designed to help minimize the impact of these factors and improve the robustness of evaluations.

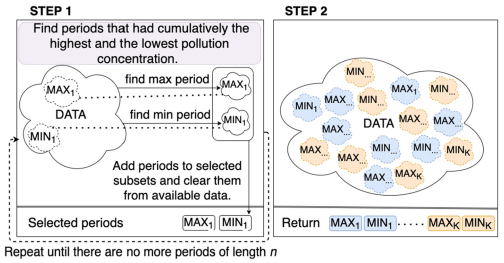


Fig. 5. Illustration of Diverse Data Selector (Algorithm 1). Blue indicates training data and orange testing data.

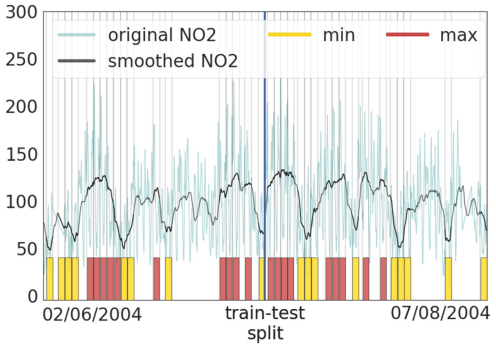


Fig. 6. Demonstration of Diverse Data Selector in practice. The number of windows in a set can be limited to ensure the sets remain descriptive to their concentration level. The image shows 8 windows for each set.

3 DIVERSE EVALUATION FRAMEWORK

The previous section highlighted key issues that standard model evaluation protocols face when applied to air quality calibration. We next present an improved evaluation framework, coined the *Diverse Evaluation Framework*, that addresses these issues. The framework has been designed to consider the requirements for professional-grade stations and especially the effect of high and low target pollutant concentrations. The mechanism can be applied to both target variables and features considered by the calibration model, such as temperature, humidity, and other pollutants. Thus, our framework offers a generic procedure for creating training and testing datasets that can be used to evaluate model generality. As part of our evaluation, we demonstrate that it captures short-term seasonal variations but breaks long-term autocorrelations, while allowing us to consider the effect of distributional differences in data. The mechanism for creating the splits can be understood as an adaptive and hierarchical cluster sampling technique where the data are progressively split into clusters according to the current characteristics being evaluated. The training and testing datasets are then created by sampling measurements from these clusters. The overall idea is illustrated in Figure 5, and a pseudocode describing how to implement the method is shown in Algorithm 1.

Diverse Data Selector: Our algorithm creates a pool of segments that are scored according to a target criterion. From this pool, it is then possible to create different types of sets, e.g., separate between low and high concentrations or create balanced sets that integrate both high and low concentrations. By operating on segments, we can preserve short-term autocorrelations and support models that incorporate temporal dependencies, while at the same time being able to omit long-term correlations. In the following, we use distributional differences as a running example of the target criteria for evaluation. The algorithm can be applied to other target criteria, such as autocorrelation or the values of a covariate, and this example is only used to illustrate the key ideas of the algorithm.

The algorithm for creating a data split, shown in Algorithm 1, takes as input the entire dataset that is used for evaluation and a segment length w determining the length of continuous periods that are considered. In the algorithm, the operator $S \backslash \{-\infty\}$ denotes set difference, i.e., S excluding $-\infty$. The output of the algorithm consists of a ranked list of segments, i.e., continuous periods of measurements.

The algorithm proceeds through each measurement index $d_1, \ldots, d_n$, considering a segment of length w starting from that index. For each index, we calculate a segment score that summarizes the segment with respect to the desired target criteria. In the case of our running example, one-dimensional-convolution can be used to sum the pollutant values within the segment, and this sum can be used as the score. Once all segments are scored, we extract the segment with the highest value and assign it to the pool of measurements. The scores of the measurements are then cleared, and all the segment sums are recomputed. Recomputing the segment scores ensures that overlapping periods are not included in the pool of segments, avoiding individual high- or low-pollution events dominating the selection. This procedure is continued until there are no more segments of the given length w, i.e., the algorithm identifies the maximum number of segments that can be chosen without having any of the segments overlapping. After the algorithm finishes, it has selected segments by the score in a descending order, and the data points that were not in any of the segments are omitted from evaluation. Low segments are determined analogously and simply require changing arg max to arg min (line 5) and reversing the sign of the reinitialized values (line 10). We can also alternate between arg max and arg min, which produces a combination of high and low periods, which we refer to as diverse data. Figure 5 illustrates the algorithm's functionality with diverse data selection, and Figure 6 shows an example of applying it to real data.

ALGORITHM 1: construct-dataset

input : $\mathbf{d}, w$ where $\mathbf{d} \ni d_1, \ldots, d_n$ and $w \in \mathbb{N}$
output: indexes

1 **begin**
2 indexes = $\emptyset$
3 **while** *True* **do**
4 $S = \bigcup_{i=0}^{N} \sum_{n=1}^{w} d_{i+n} K(n)$ where $K = \begin{cases} 1, & \text{when } n \leq w \\ 0, & \text{otherwise} \end{cases}$
5 $i = \arg\max_i (S \backslash \{-\infty\})$
6 **if** $i = \emptyset$ **then**
7 break
8 **end**
9 indexes.push$(d_i, \ldots, d_{i+w})$
10 $d_i, \ldots, d_{i+w} = -\infty$
11 **end**
12 **end**

Stream-based version of Diverse Data Selector: Air quality sensor deployments typically need to operate on data arriving as a stream of measurements instead of having access to all of the data at one time. Our approach can also be executed in stream mode to support re-calibration of the sensors or to transfer existing calibration models to new environments [37]. Stream-based implementation simply requires a buffer for storing segments and their scores. Incoming measurements are buffered until a full segment is available, and then we calculate a score for the segment as before. If the score is higher than the lowest segment score in the buffer, then we add the new segment into the buffer and remove the segment with the lowest score from the buffer. The process for identifying low concentration events is analogous, sorting the segments by ascending order instead of descending. The size of the buffer determines the amount of data that is available for model evaluation. The more data are available for training, the more costly the training, and thus the size of the buffer also controls how often re-calibration should be applied. By adapting the

buffer size, complexity of the calibration model, and frequency of re-calibration, our approach can be used to support a wide range of different devices. For example, low-end devices with limited memory and storage can simply use a smaller buffer size and re-train the model whenever sufficiently many segments have changed. Higher-end devices, however, can use larger buffer to train a more complex model that typically requires less frequent re-calibration than a simpler model.

Practical Example: Assume that we have a collection of data that could potentially all be used for evaluation purpose. This kind of collection could possibly be achieved by splitting a long dataset from the middle (like we have done in Figure 6). We start by selecting an input feature, a desired feature level (low, high, or diverse), and a time window length for using Diverse Data Selector. The algorithm then produces a sorted set that contains the majority of indexes in it. We can then select a subset of segments according to our needs, e.g., to match a desired testing set size or to ensure a proper level of variation. For example, with a segment size equal to one day, selecting the 30 first segments corresponds to selecting 30 days with lowest or highest feature levels. We can now use this set for evaluation to get an improved understanding about model performance in specific feature conditions. We can the rerun Diverse Data Selector with different parameters, e.g., change the feature level from low to high or to use autocorrelation instead of total sum to score the segments, and get another view of model performance in differing feature conditions. By systematically going through different target criteria, we establish a more complete understanding about model behavior and generality in real-world conditions.

4 EXPERIMENTAL SETUP

We validate and demonstrate the benefits of our evaluation framework through extensive experiments carried out using a publicly available air quality dataset. The experiments focus on showing how understanding about model performance and generality increases with our diverse dataset selection approach and on demonstrating pitfalls of using conventional model evaluation techniques. We next describe the data, evaluation methods, and calibration models used in these experiments. All experiments were implemented in Python using common libraries (numpy, pandas, tensorflow, keras, xgb, scikit, statsmodels). We repeat all experiments 20 times and average the results to reduce the effect of randomness in training.

Experiment Data: We conduct our experiments using a public dataset[3] released by DeVito et al. [52]. The dataset contains measurements for several gaseous pollutants from a low-cost metal oxide sensor and a reference sensor. The dataset also contains measurements from temperature and humidity sensors. The dataset contains hourly measurements from March 10, 2004 to April 4, 2005. We use this dataset despite its age as it presents the only openly accessible dataset that contains measurements for an entire year. Naturally, more recent low-cost air quality sensors have better correspondence with reference stations [38]. This would mean that the calibration errors are likely to be smaller with modern sensors; however, the general issue with data diversity and temporal dependencies would affect the measurements [8, 37].

In total there are 9,357 samples. We consider two pollutants as the target variables for calibration, CO and NO_2. We follow the data processing and guidelines in the original work of DeVito et al. [52] as closely as possible. For both pollutants, we select the suggested feature sets to predict reference station's values. Specifically, for calibrating CO we use NO_2, O_3, and CO low-cost sensors together with relative humidity and temperature, and for NO_2 we use all the measurements available from the low-cost sensor array. The low-cost sensor is a single integrated multisensor device produced by Pirelli labs (which has since been discontinued). Reference measurements are also included

[3] https://archive.ics.uci.edu/ml/datasets/Air+Quality.

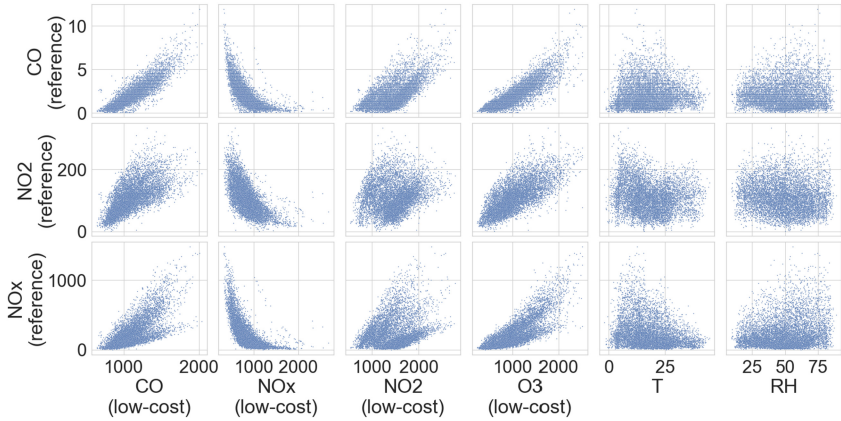


Fig. 7. Correlogram comparing the measurements provided by the reference instrument (y-axis) to those provided by the low-cost sensors (x-axis).

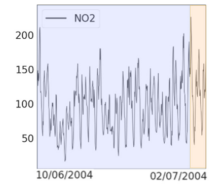
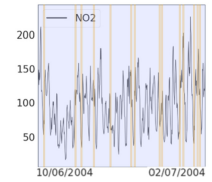
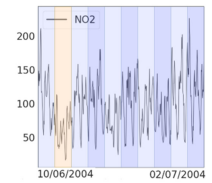
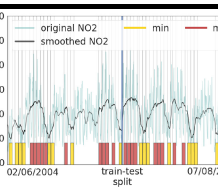
in the same dataset and were provided by a co-located reference certified analyzer [51]. Figure 7 shows a correlogram comparing the low-cost measurements with the reference station. As the figure shows, the measurements rarely align with the diagonal. Calibration algorithms operate by correcting the low-cost sensor values to better align with the diagonal in the correlogram.

Evaluation methods: We compare our method with the three most commonly applied model evaluation techniques in machine learning–based air quality sensor calibration research [5, 21, 33, 45, 47, 52, 53, 57]: ordered holdout, shuffled holdout, and k-fold cross-validation. These techniques together with their advantages and disadvantages are summarized in Table 4.

Calibration models: We consider nine machine learning models covering approaches used by previous works of air quality sensor calibration and representing different levels of model complexity [6, 26, 31, 33, 45, 52, 53, 57]. In addition, we consider a naïve baseline that uses the mean of the training data as prediction. This baseline allows us to estimate the degree to which the models improve on properties of the data itself. The machine learning models we consider are as follows:

- *Training Mean:* Uses the mean value of the target variable in the training data as the predictor. This model is used as a naïve baseline to assess the overall benefits of the calibration.
- *Linear Regression:* A linear regression that uses a single input feature and minimizes the Ordinary Least Squares between predictions and target values [33].
- *Multiple Linear Regression:* A multi-dimensional version uses multiple features as input but otherwise the same as the previous model [9, 33, 45, 53, 57].
- *Support Vector Regression:* Uses a kernel function (Radial Basis Function) to map the input data to a high-dimensional space and then finds a hyperplane in the transformed space that minimizes the distance between the hyperplane and the data points [9, 20, 53].
- ***Extreme Gradient Boosting (XGB)*** [26, 53].
- *RF:* A random forest regressor [57].
- ***Multidimensional linear regression (MLR)****+RF:* A two-phase model introduced in Reference [31] that combines multiple linear regression and a random forest model to separately estimate linear and non-linear corrections.
- ***Artificial neural networks (ANN)*** *simple w/scaling:* A simple feed-forward neural network model used by DeVito et al. [52] with 10 neurons on each of the two hidden layers; uses scaled input values.

Table 4. Most Commonly Applied Model Evaluation Techniques in Machine Learning Air Quality Sensor Calibration Compared to Our Method

Method	Description	Pros	Cons	Method illustration
ordered holdout [7, 31, 33, 47, 52, 53]	Splits the available data into two sets where one set is used for training and the other for testing. The split is done by taking the tail from an (temporally) ordered dataset.	Mimics post-deployment scenario, gives an impression about the functionality in real-world conditions	Sensitive to testing data distribution variation, the trained models tend to overfit to features such as general pollution level and temporal dependencies, especially with short-term data. Technique can only be used reliably if there is a mechanism that can ensure the training and testing sets contain no dependencies [13, 40]	
shuffled holdout [20, 21, 21, 26, 36, 45, 56, 57]	Same as above, but the split is done by randomly sampling the testing samples from all available data.	Gives model performance estimations when training and testing data are very similar; describes model flexibility.	Overfits the data very heavily as theoretically all testing data points might have an adjacent data point in training set, i.e. testing data is unrealistically similar to training data, and it does not reflect real-world performance. [40]	
k-fold CV [5, 57]	Splits all available data into k folds and makes several holdout splits to test the model on all available data. Each fold uses k-th fraction for testing and the rest for training. Performance estimations of each fold are averaged to conclude an average performance.	Uses all data (separately) for training and testing and so might give an indication if a model causes drastically high errors with some data.	Uses adjacent folds for training and testing which leaks information about general pollution level and temporal dependencies. Also for k-1 folds, the technique uses future data to infer errors at current testing fold. Final output is an average of averages which is not very descriptive when finding reasons for model errors. [22]	
Diverse Evaluation Framework	From a collection of data and by using Diverse Data Selector forms subsets that are descriptive to certain feature conditions, e.g. environmental conditions (further explanation in Section 3).	Improves understanding about model performance in wide range of conditions by using various testing sets each corresponding to certain feature characteristics.	Requires a fair amount of data for creating descriptive sets for all features, the benefits of the framework decrease when short-term data is used for evaluation.	

We list description, pros, cons, and illustration from each evaluation method. In conventional method illustrations blue section presents training data and orange section testing data. In the Diverse Evaluation Framework the high and low sets have been selected according to the target variable.

- *ANN simple w/o scaling:* As the previous model but without scaling the input values.
- *ANN deep:* Feed-forward neural network with 5 hidden layers, 32 neurons per hidden layer [45], and scaled input values.

Handling Missing Data: Air quality measurements are characterized by missing measurements, and the public dataset we use is no exception. Of the 9,357 samples, 4% have missing values for the low-cost metal oxide, temperature, and humidity sensors, and the reference variables (CO and NO_2) are missing for 18% of the samples. The way the missing values are handled can have a significant impact on the calibration model [27], with particularly complex models being vulnerable to the way missing data are handled. We tested three methods for handling missing data and selected linear interpolation as this resulted in best alignment with the results of previous works, see below.

Replicating previous work: To ensure our implementations of the calibration models are correct, we replicated the results of the previous works where possible. Of the 10 models, 5 have been applied to the public dataset we consider (LR, MLR, XGB, SVR, and ANN) [52, 53], and we focus on comparing the results for these 5 models. The other models (RF, MLR+RF, ANN w/o scaling, ANN deep) have largely been evaluated with proprietary datasets, and exact replication

Table 5. The Mean Absolute Errors from Replicating a CO Calibration Experiment

Model	Previous work (MAE)	Our implementation (MAE)	Interpolation method
LR	0.63 [53]	0.61	drop missing
		0.63	**linear**
		0.61	iterative imputer
MLR	0.39 [53]	0.36	drop missing
		0.39	**linear**
		0.36	iterative imputer
XGB	0.42 [53]	0.41	drop missing
		0.43	linear
		0.41	iterative imputer
SVR	0.34 [53]	0.32	drop missing
		0.33	**linear**
		0.31	iterative imputer
ANN	0.35 [52]	0.42	drop missing
		0.53	linear
		0.36	iterative imputer

The errors are close to the original publications [52, 53]. The best alignment is obtained with linear interpolation, and hence we use it to handle missing values in our experiments.

was not possible. Nevertheless, for these models, we verified that the performance differences between different methods were in line with those published. The original articles for the dataset we consider provide only partial details of how missing data were handled. While replicating the results, we need to identify also the method that was used to handle missing values. We first create the same training-testing partitions as in the original works. We then evaluate the models proposed in the works by considering three methods for handling missing data: dropping missing measurements, linearly interpolating measurements, and using an iterative imputer to fill in missing values. The mean absolute errors of these models when used with the different missing data processing techniques are shown in Table 5. The results closely align with the original publications and linear interpolation results in the closest match with the original results. Hence, we use it for handling missing data in our experiments.

5 EVALUATION RESULTS

We use the data and experimental setup described in the previous section to systematically assess the benefits of our diverse data selection technique while at the same time comparing it against conventional model evaluation techniques. An example of data splits created using our algorithm is shown in Table 4, and examples of distribution statistics are given in the individual subsections. The main results of our evaluation are as follows:

- Our diverse data framework results in training and testing sets that capture the distributional and statistical characteristics of data better than conventional model evaluation techniques (Section 5.1).
- The segment length parameter w offers control over data continuity, allowing our approach to evaluate tradeoff between variations in data and continuity of measurements (Section 5.2).
- Conventional model evaluation techniques easily result in training and testing datasets having similar characteristics, which results in optimistic performance bounds for calibration models (Section 5.3).

Table 6. Dataset Specifications for CO and NO_2 Listed Separately for Each Evaluation Method

	CO						NO_2					
	mean		deviation		ac (lag = 24)		mean		deviation		ac (lag = 24)	
	training	testing	training	testing	training	testing	training	testing	training	testing	training	testing
ord. holdout	1.72	2.9	1.07	1.47	0.59	0.63	92.1	107.4	37.74	38.68	0.62	0.64
shuf. holdout	2.02	2.03	1.23	1.25	0.0	-0.01	94.76	94.64	34.61	34.73	0.0	-0.03
CV	2.02		1.28		0.67		95.92		38.55		0.64	
low	-	0.97	-	0.64	-	0.07	-	78.35	-	28.53	-	-0.04
high	-	4.16	-	1.86	-	-0.02	-	179.69	-	49.83	-	0.16
diverse	-	2.71	-	2.37	-	0.0	-	130.27	-	75.62	-	-0.06
4 months (2880 data points w/ 75:25 training-testing split)												
All data	1.89	2.37	1.18	1.61	0.59	0.58	92.66	126.6	35.29	49.97	0.61	0.66
13 months (9357 data points w/ 50:50 training-testing split)												

The proposed method successfully removes high autocorrelation in the data and provides descriptive data for high and low concentration levels.

- Current calibration models are highly sensitive to data heterogeneity, yet current model evaluation techniques have failed to demonstrate this (Section 5.4).
- Using our diverse data framework for training calibration models improves their generality and reduces the results dependency on general pollution levels (Section 5.5).
- Diverse data framework can additionally improve model robustness while at the same time decreasing the risk of overfitting and reducing computational requirements by focusing the evaluation on periods that have the most variation (Section 5.6).
- Increasing temporal distance between training and testing set to at least 4 months can offer improved robustness for continuous (conventional) model evaluation techniques, but using diverse data is overall the best approach for model evaluation (Section 5.7).

5.1 Improved Coverage of Distributional and Statistical Characteristics without Autocorrelation

We begin by demonstrating that our framework improves the understanding of model performance as each testing set is linked with metrological criteria governing the performance requirements of reference instruments [15]. We accomplish this by using the four different evaluation methods to create training and testing splits from the overall dataset and comparing the distributional and statistical characteristics of these splits.

We split the measurements into two and consider this as a preliminary training–testing split. We then choose a 3-month continuous period from the first half's end to represent continuous training data and select the first month from the testing half to represent a continuous testing evaluation. We also consider these 4 months together as the source for randomly shuffling training data points and for performing a fourfold cross-validation (i.e., leave-one-month out). Finally, we apply our Diverse Data Selector algorithm with a segment length w of one day to the latter half of the data and form low, high, and diverse datasets. The characteristics of the created sets are summarized in Table 6.

From the results, we can see that our method creates a combination of continuous time windows that covers a wide range of conditions, e.g., ranging from low CO concentrations below 1.0 $\mu g/m^3$ with low deviation to high concentrations above 4.0 $\mu g/m^3$ with deviation 1.86, while breaking autocorrelation in the data. In contrast, the datasets formed with conventional evaluation methods have highly autocorrelated data, the differences in training and testing sets are minor, or both. As all the data are collected from the same physical location, this means that there are highly varying periods within the overall data but that conventional evaluation methods fail to identify and exploit them in evaluating model performance.

The impact of this result is also highly significant for human health. Pollutants have different safety limits that are governed by standards, e.g., EU regulatory standards set a 10 mg/m^3 limit for CO for 8-hour exposure and 200 $\mu g/m^3$ for NO_2 for 1-hour exposure (see Table 1). It is therefore important that the performance indicators are always within acceptable limits, not just during days of average pollution levels. For example, looking at the mean and deviation values of NO_2 in all of the testing sets created by the conventional evaluation methods, we see a mean value around 100 $\mu g/m^3$ and deviation around 35 $\mu g/m^3$. In contrast, the high dataset created by our method shows a mean close to 180 $\mu g/m^3$ and deviation close to 50 $\mu g/m^3$. In other words, there are periods where the pollutant concentrations exceed the safety standards, but these are not reflected accurately in the training or testing sets created by conventional methods. For example, taking a real dataset with 90% low pollution values, the selected methods will predict low values to minimize error. When faced with high pollution values from the low-cost sensor, they will still predict conservatively, underestimating the pollution level. As days with high pollution levels occur, possible false implications given by calibrated sensors become actually hazardous to health. For example, a mean error close to conventional testing set deviation (35 $\mu g/m^3$) would fail to inform citizens when conditions exceed hazardous levels. As negative health effects can result even from short time exposures (1h or 8h) and as reference data are not expected to be available for continuous error estimation of each low-cost sensor, ensuring the calibrated models can indeed detect and react to hazardous situations is essential.

For air quality monitoring, it is vital that model performance is investigated under different environmental (i.e., weather or seasonal effects) or other conditions (e.g., human activity) instead of on a temporal basis as calibration models are expected to function in a highly varying real-world environment [8, 51]. The key finding from our evaluation is that conventional methods fail to characterize generality of model performance and provide no indication of how well the model meets performance requirements placed on air quality monitoring solutions. Our diverse selection framework improves the situation by creating splits that can be used to assess performance in differing environmental conditions by considering criteria that link directly with metrological requirements.

RESULT

The proposed diverse framework offers a mechanism to assess calibration performance in varying environmental conditions, unlike conventional model evaluation methods that underestimate the effects of environmental conditions and temporal effects.

5.2 Control over Data Continuity

The previous section showed how our diverse data selection framework successfully forms sets that are descriptive of the distribution of pollutants without being strongly correlated across time. We next demonstrate how the segment length parameter w can be used to control the continuity and characteristics of the training and testing sets. Increasing the segment length, while keeping the size of the training and testing data fixed, results in the distributional characteristics of the testing set progressively resembling those of continuous data. Continuity is essential for models that incorporate temporal dependencies, such as recurrent neural networks or Gaussian process regression, and being able to fine-tune the tradeoff between distributional characteristics and data continuity offers a way to evaluate the generality and time dependency of such techniques.

We use our approach to create datasets with different segment lengths w, ranging from 1 day to 15 days. We compare distributional characteristics of these sets against those extracted from all data (i.e., the full 13 months included in the original dataset). Table 7 shows the statistical

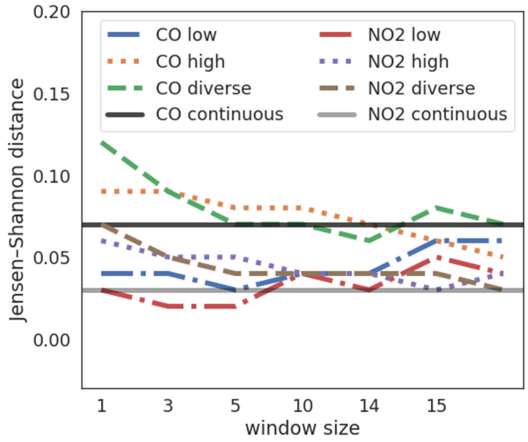


Fig. 8. JS-distance determines how similar two distributions are. Decreasing differences show that a larger window reduces the special characteristics of subsets.

Table 7. Increasing Segment Length while Keeping Set Size Fixed Selects Data Points That Are More Similar to Common Pollution Levels

	low		high		diverse		
window	mean	sd	mean	sd	mean	sd	set size
1	0.97	0.64	4.16	1.86	2.71	2.37	720
3	1.15	0.72	3.96	1.76	2.61	2.15	720
5	1.27	0.93	3.78	1.83	2.58	1.92	720
10	1.42	1.1	3.48	1.74	2.7	1.97	720
14	1.54	1.31	3.23	1.73	2.32	1.71	672
15	1.54	1.33	3.17	1.83	2.3	1.7	720
All data	2.13	1.43	2.13	1.43	2.13	1.43	9357

(a) CO (in mg/m^3)

	low		high		diverse		
window	mean	sd	mean	sd	mean	sd	set size
1	78.35	28.53	179.69	49.83	130.27	75.62	720
3	82.87	32.21	172.85	51.87	130.04	70.93	720
5	85.44	34.0	170.86	55.62	131.73	69.22	720
10	90.26	32.81	163.74	58.77	139.92	64.81	720
14	91.91	28.96	161.23	59.12	129.89	62.94	672
15	93.04	31.47	158.96	54.24	128.19	63.25	720
All data	109.63	46.46	109.63	46.46	109.63	46.46	9357

(b) NO_2 (in $\mu g/m^3$)

characteristics of these sets. For both target pollutants, increasing the segment length results in the mean and standard deviation progressively resembling those of the continuous data. Note that this result is not due to coverage of the data increasing as the size of the testing sets is fixed. The sole exception is segment length $w = 14$ where the size of the testing data decreases due to segment overlap preventing additional segments being chosen.

We also computed the Jensen-Shannon distance of distributions between the testing sets created with our technique and the entire data. The similarities are shown in Figure 8 as a function of segment length. The figure provides further support to the finding that testing sets of similar size but with longer segment lengths generally resemble the overall data more closely. The low concentration datasets for both target pollutants (CO and NO_2) behave differently than the high and diverse datasets, as can be seen from Figure 8. This is due to the low pollution values dominating the data, resulting in the mean pollution levels being biased toward lower concentrations.

A dataset of a fixed size built from longer segments results in closer resemblance to continuous data and emphasizes general pollution levels, whereas shorter segments allow us to assess the impact of variations in environmental conditions. From a calibration model training perspective, a longer time frame enables the use of modelling techniques that incorporate temporal dependencies, but from an evaluation perspective the testing conditions become less restrictive and closer to general pollution levels. This can limit insights about model performance and result in the model performance following the mean pollution level too closely. Our framework offers a potential way

Table 8. Mean Absolute Errors (CO in mg/m^3, NO_2 in µg/m^3) of Calibration Models When Evaluated with Conventional Evaluation Techniques

	(=continuous) ordered holdout	shuffled holdout	CV
Training mean	1.36	1.27	1.32
LR	0.71	0.70	0.73
MLR	0.66	0.61	0.68
SVR	**0.64**	0.62	**0.67**
XGB	0.82	0.58	0.82
RF	0.76	0.54	0.77
MLR+RF	0.75	**0.52**	0.76
ANN simple raw	0.72	0.64	0.72
ANN simple scaled	0.67	0.59	0.71
ANN deep	0.74	0.55	0.80
4 months (2,880 data points w/ 75:25 training-testing split)			

(a) CO (in mg/m^3)

	(=continuous) ordered holdout	shuffled holdout	CV
Training mean	40.35	37.51	38.61
LR	38.76	34.05	37.03
MLR	**26.11**	22.65	**24.91**
SVR	29.29	24.21	26.75
XGB	30.34	20.74	28.04
RF	27.96	19.51	26.46
MLR+RF	27.47	**19.48**	26.17
ANN simple raw	26.73	25.50	25.87
ANN simple scaled	27.81	22.26	25.98
ANN deep	28.61	19.70	26.55
4 months (2,880 data points w/ 75:25 training-testing split)			

(b) NO_2 (in µg/m^3)

The errors are similar to each other, besides the shuffled holdout. Lowest errors for each evaluation method have been emphasized.

to control this effect and assess the impact that temporal dependencies have on the measurements. Tuning the segment length parameter w controls whether the datasets should emphasize continuity or break the temporal structures. A potential way, then, to evaluate techniques that incorporate temporal dependencies is to use our framework to create testing sets with different segment lengths, as this allows us to assess the model's sensitivity to temporal dependencies and the overall level of pollutants.

RESULT

The diverse framework offers control over data continuity and temporal dependencies, unlike existing model evaluation methods. The segment length parameter w controls the tradeoff between data continuity and descriptiveness of the dataset. Repeating the evaluation with multiple segment lengths thus allows assessing the impact of both variations and continuity on performance.

5.3 Diverse Data and Calibration Model Performance

Thus far we have considered how our framework affects the characteristics of the datasets that are used for training and testing. We next assess the impact these characteristics have on model performance and demonstrate how our evaluation framework helps to obtain a significantly better understanding about the calibration model's performance compared to existing machine learning validation techniques.

We first carry out performance evaluation with three conventional evaluation schemes (ordered holdout, shuffled holdout, and cross-validation; see Table 4). We train all the calibration models (see Section 4) with 3 months of data and then use a 1-month testing set to evaluate them. The results in Table 8 show that these evaluation methods give similar implications for overall model performance. Shuffled holdout always produces the the lowest error, which is due to the fact that the entire distribution is shown in the training phase and the testing data points might be close to identical to training data points. The performance of ensemble methods supports this, as they are known to easily overfit to properties of the data.

Table 9 shows the same results when the calibration models are run on three different datasets generated with our method: low, high, and diverse, containing low, high, and mixed concentrations of target pollutants, respectively. In the table, we have also included ordered holdout (continuous) as a baseline to highlight how model performance changes. The diverse and high concentration

Table 9. Mean Absolute Errors of Calibration Models When Evaluated with Different Testing Sets

	cont.	low	high	diverse
Training mean	1.5	0.88 (-41%)	2.57 (+71%)	2.02 (+35%)
LR	0.97	**0.55 (-43%)**	1.34 (+38%)	**1.17 (+21%)**
MLR	**0.95**	0.72 (-24%)	1.3 (+37%)	1.25 (+32%)
SVR	**0.95**	0.63 (-34%)	1.44 (+52%)	1.3 (+37%)
XGB	1.07	0.61 (-43%)	1.57 (+47%)	1.29 (+21%)
RF	0.98	0.63 (-36%)	1.39 (+42%)	1.25 (+28%)
MLR+RF	0.97	0.84 (-13%)	**1.24 (+28%)**	1.21 (+25%)
ANN simple raw	0.96	0.88 (-8%)	1.47 (+53%)	1.45 (+51%)
ANN simple scaled	0.98	1.11 (+13%)	1.45 (+48%)	1.52 (+55%)
ANN deep	1.01	0.6 (-41%)	1.54 (+52%)	1.24 (+23%)
AVG	1.03	0.74 (-28%)	1.53 (+49%)	1.37 (+33%)
4 months (2,880 data points w/ 75:25 training-testing split				

(a) CO (in mg/m^3)

	cont.	low	high	diverse
Training mean	32.77	25.29 (-23%)	88.25 (+169%)	65.0 (+98%)
LR	31.4	27.07 (-14%)	113.8 (+262%)	77.21 (+146%)
MLR	30.05	27.86 (-7%)	**42.77 (+42%)**	43.02 (+43%)
SVR	27.61	25.55 (-7%)	64.97 (+135%)	54.5 (+97%)
XGB	28.6	23.63 (-17%)	73.17 (+156%)	55.21 (+93%)
RF	27.32	**21.74 (-20%)**	66.77 (+144%)	51.83 (+90%)
MLR+RF	**27.31**	22.7 (-17%)	52.61 (+93%)	44.32 (+62%)
ANN simple raw	32.56	28.63 (-12%)	52.92 (+63%)	48.22 (+48%)
ANN simple scaled	30.77	29.01 (-6%)	43.73 (+42%)	44.03 (+43%)
ANN deep	29.35	30.65 (+4%)	43.76 (+49%)	**40.57 (+38%)**
AVG	29.77	26.21 (-12%)	64.28 (+116%)	52.39 (+76%)
4 months (2,880 data points w/ 75:25 training-testing split)				

(b) NO_2 (in $\mu g/m^3$)

Lowest errors for each testing set have been emphasized. Comparison with continuous data highlights that calibration models suffer from large error variation when testing data changes, which is why model performance estimations should be better separated from data properties.

testing sets selected with our method result in significantly higher errors than continuous or low concentration data, highlighting how other validation methods give optimistic views of model performance and are biased toward lower concentration levels that tend to dominate the distribution of pollutants. The performance of calibration models changes between −12 and +116 % for NO_2 and between −28 and +49 % for CO. The significance of these changes is further highlighted by the fact that for both target pollutants merely predicting the mean value from training data results in lower error than using a simple ANN calibration. Changes in the evaluation sets also affect the performance differences between the methods with the best-performing model being highly dependent on the way the training and testing sets are constructed. Our approach allows testing against variations in the pollutant concentrations and assessing the generality and sensitivity of the model to these and other changes in the characteristics of the training and testing sets.

The performance variations indicate that the calibration model's performance is sensitive to changes in the pollution distribution, which makes the calibration model's performance highly sensitive to the conditions where the training and testing data have been collected. Training and testing with similar target distributions intuitively results in lowest error rates that can be seen as the best case performance. In the data considered in our evaluation, high pollution concentrations are infrequent, resulting in continuous evaluation methods being biased toward lower concentrations and the errors being strongly dependent on whether higher concentration periods occurred during the training and testing periods or not. As we showed in Figure 3, simply enforcing temporal separation is not sufficient for ensuring the measurements are sufficiently varied for assessing model generality. Conventional evaluation methods are not reliable for evaluating a calibration model's performance, as they fail to account for distributional characteristics and benefit from similarity of training and testing data. As a result, they provide only limited insights about the generality of the calibration models and easily offer an optimistic view of true model performance. Our diverse data selection framework overcomes this issue, ensuring differences in distributional characteristics are explicitly modeled and considered as part of the evaluation.

RESULT

Our evaluation framework avoids testing data to be very similar to training data, overcoming the main issue with continuous evaluation methods and offering a method that can assess the impact of different pollution levels.

Table 10. Averaged Model Errors (MAE) and Their Deviations w.r.t. Each Training–Testing Dataset Combination with Two Lowest Values Emphasized

training level	testing level	CO mean	CO sd	T mean	T sd	RH mean	RH sd
low	low	**0.49**	**0.04**	**0.73**	**0.20**	**0.59**	**0.13**
high	low	1.80	0.48	0.91	0.22	0.90	0.27
low	high	2.93	0.23	1.29	0.30	**0.87**	0.23
high	high	**0.97**	**0.16**	**0.61**	**0.13**	0.92	**0.18**

(a) CO (in mg/m^3)

training level	testing level	NO_2 mean	NO_2 sd	T mean	T sd	RH mean	RH sd
low	low	**18.53**	**2.49**	**28.78**	**7.87**	**22.44**	7.10
high	low	61.11	23.44	55.26	20.08	**27.61**	**4.77**
low	high	86.74	14.82	38.74	20.66	30.04	**4.53**
high	high	**29.97**	**5.38**	**22.17**	**4.51**	28.89	5.47

(b) NO_2 (in µg/m^3)

Performance with different combinations of low and high concentration datasets shows that calibration models can be emphasized to learn a given distribution's properties. Although this brings out the issue that less frequent environmental phenomena are less likely to be handled accurately, it also sheds light on the effect of correct training data selection.

5.4 Performance with Homogeneous and Heterogeneous Data

We showed that conventional model validation techniques easily result in training and testing data being highly similar. This, in turn, results in the evaluation process merely characterising the calibration model's adaptability to general pollution levels without assessing its capability to learn a generic mapping from low-cost measurements to those provided by reference monitors. We next address in more detail how the calibration models perform with homogeneous and heterogeneous measurements. Besides addressing model generality, this experiment also evaluates the model's ability to transfer across locations with differing environmental conditions.

We create a total of 12 datasets for the two target pollutants (CO and NO_2), comprising low and high sets created using the level of pollutant, or the level of humidity or temperature as the criteria for creating the datasets (6 per pollutant). We train the calibration models with one set and then evaluate them with another. When the levels are the same, we apply **cross-validation (CV)** without breaking any window structures. We use a fixed 1-day segment length when forming the high and low sets in this experiment. Training and testing data were sampled from 10 months of data, and the final sets covered 75 to 79 days for training and 26 to 30 days for testing. The results are shown in Table 10.

The results show that data homogeneity has (as expected) a significant effect on the performance of the calibration models. Besides few exceptions, similar training and testing environments result in the lowest errors and deviations. With heterogeneous data, training models with data from a lower level and testing against a higher level tends to cause higher errors than reversing the roles. This indicates that models are better at adapting from higher pollution levels to lower levels than the other way around. A sole exception in this is the simple ANN, which uses scaled input values. The errors vary considerably due to the challenge of scaling the feature values. Indeed, the scaler is trained with training data, and therefore its scaling is limited to the values that appeared in the training data.

Field testing requirements for reference monitors [15] state that performance needs to be tested in different pollutant concentrations and against differing environmental conditions, whereas our results have shown that current validation techniques tend to result in similar data being used for evaluation. Our results have also shown how low-cost calibration models are highly sensitive to variations in the level of pollutant concentrations or those of environmental variables, while highlighting how current model evaluation techniques fail to characterize the models' sensitivity to these variations. Indeed, our results show that the calibration model's performance exhibits substantial variation according to the relationship of training and testing environment. Our proposed framework allows using different training and testing splits, which can identify potential issues with the models and provide a better understanding of model performance. In

Table 11. Training Calibration Models with Diverse Training Data Improves Model Generality Compared to Training with Continuous Data (Table 9) or When Performance Is Contrasted against Different Concentration Levels

	cont.	low	high	diverse
Training mean	1.24	1.05 (-15%)	2.41 (+94%)	2.02 (+63%)
LR	0.7	0.52 (-26%)	1.4 (+100%)	1.2 (+71%)
MLR	0.68	0.55 (-19%)	1.3 (+91%)	1.15 (+69%)
SVR	0.68	0.54 (-21%)	1.36 (+100%)	1.19 (+75%)
XGB	0.86	0.63 (-27%)	1.49 (+73%)	1.29 (+50%)
RF	0.8	0.58 (-28%)	1.46 (+82%)	1.24 (+55%)
MLR RF	0.8	0.64 (-20%)	1.32 (+65%)	1.19 (+49%)
ANN simple raw	0.77	0.76 (-1%)	1.29 (+68%)	1.25 (+62%)
ANN simple scaled	0.75	0.56 (-25%)	1.44 (+92%)	1.25 (+67%)
ANN deep	0.91	0.61 (-33%)	1.43 (+57%)	1.17 (+29%)
AVG	0.82	0.64 (-22%)	1.49 (+82%)	1.29 (+57%)

(a) CO (in mg/m^3)

	cont.	low	high	diverse
Training mean	46.65	25.56 (-45%)	87.57 (+88%)	64.91 (+39%)
LR	67.92	28.67 (-58%)	116.67 (+72%)	79.18 (+17%)
MLR	26.04	26.12 (+0%)	38.17 (+47%)	39.03 (+50%)
SVR	35.05	25.23 (-28%)	63.36 (+81%)	53.12 (+52%)
XGB	37.42	24.42 (-35%)	64.37 (+72%)	51.53 (+38%)
RF	36.74	24.97 (-32%)	62.06 (+69%)	51.14 (+39%)
MLR RF	27.93	23.99 (-14%)	42.81 (+53%)	39.14 (+40%)
ANN simple raw	30.09	24.51 (-19%)	50.38 (+67%)	45.11 (+50%)
ANN simple scaled	26.19	26.81 (+2%)	37.8 (+44%)	39.16 (+50%)
ANN deep	28.96	31.17 (+8%)	33.81 (+17%)	34.89 (+20%)
AVG	36.3	26.14 (-28%)	59.7 (+64%)	49.72 (+37%)

(b) NO_2 (in $\mu g/m^3$)

case a single performance metric is desired, the results can be averaged across different splits, as long as the process for establishing the splits is standardized.

RESULT

The diverse data framework creates splits that are in line with field testing requirements for reference monitors, unlike conventional model evaluation techniques that tend to result in training and testing data being similar, which limits their potential to characterize model performance in line with field testing requirements.

5.5 Diverse Data Improves Generality

We next demonstrate an added benefit of our proposed data selection method by showing how the use of diverse data to train calibration models helps to improve model generality compared to training with continuous data. We select two different diverse datasets that we use separately for training and for testing. We achieve this by dividing all available data into two equal parts (6.5 months in each) temporally from the middle and running the subset selection individually on both of these sets. The continuous training and testing data contained 4,678 data points (6.5 months) and the sampled low, high, and diverse subsets 1,800 data points (2.5 months) for training and 720 data points (1 month) for testing. Table 11 shows model errors when trained with diverse training data, and Table 12 shows performance differences between individual model errors when training either with continuous or diverse data. Both performance levels are estimated with multiple testing set runs.

Table 11 shows that the low and the high concentration testing sets result in the lowest and highest errors in 85% (17 of 20 rows in the table) of the cases. Similarly to training calibration models with continuous data (Table 9), the errors with diverse testing data are higher than those with continuous data, indicating that continuous data are less challenging and do not give a good understanding about performance in a real-world deployment where distributional differences may arise from infrequent pollution events. Comparison between these tables shows that models trained with continuous data suffer a 76% greater error when tested with diverse data instead of continuous, but when trained with diverse data and tested against it, the error is only 37% greater than with continuous data training and testing. Diverse training therefore reduces the original model's error by 39 percentage points, improving robustness against diverse conditions.

Table 12. A Comparison of MAEs Shows That Feeding Calibration Models Diverse Training Data also Improves Models to Perform Better in a Wide Range of Concentrations, Compared to Using Merely Continuously Selected Data in Training

	cont.	low	high	diverse
Training mean	-0.0	0.03	-0.03	0.0
LR	0.0	-0.0	0.01	-0.0
MLR	-0.04	0.07	-0.23	-0.09
SVR	-0.01	0.07	-0.07	0.0
XGB	0.03	0.14	-0.12	0.01
RF	0.05	0.08	-0.03	0.03
MLR RF	0.01	0.14	-0.21	-0.03
ANN simple raw	0.04	0.14	-0.1	-0.0
ANN simple scaled	-0.08	-0.09	-0.1	-0.08
ANN deep	0.15	0.09	0.02	0.0
AVG	0.02	0.07	-0.09	-0.02
improvement (in %)	30.0	10.0	80.0	30.0

(a) CO (in mg/m^3)

	cont.	low	high	diverse
Training mean	-0.08	0.06	-0.16	-0.02
LR	3.6	2.17	3.73	2.64
MLR	-0.69	0.31	-4.06	-2.97
SVR	0.48	2.29	-2.93	-0.56
XGB	-9.44	2.51	-20.28	-9.41
RF	-1.65	5.86	-7.41	-0.44
MLR RF	-2.35	4.56	-9.85	-3.74
ANN simple raw	-2.81	1.77	-8.45	-3.71
ANN simple scaled	-1.07	-0.0	-4.48	-3.32
ANN deep	0.53	4.36	-7.64	-4.24
AVG	-1.35	2.39	-6.15	-2.58
improvement (in %)	70.0	0.0	90.0	90.0

(b) NO_2 (in µg/m^3)

Green color indicates lower error with diverse training data than with continuous data. We also show how many models (in %) improved their performance when trained with diverse training data.

Table 13. Using the Subset Selection Algorithm to Create a Training Dataset with More Complete Range of Possible Phenomena also Improves Model Generality and Results in Lower Error

testing set / training set	low	high	diverse	continuous	mean	sd
low	**0.54**	2.58	1.81	1.17	1.52	0.76
high	0.73	1.52	1.35	0.85	1.11	**0.33**
diverse	0.64	**1.49**	**1.29**	0.82	**1.06**	0.34
continuous	0.58	1.58	1.31	**0.8**	1.07	0.4
mean	**0.62**	1.79	1.44	0.91		
sd	**0.07**	0.46	0.21	0.15		

(a) CO (in mg/m^3)

testing set / training set	low	high	diverse	continuous	mean	sd
low	**19.87**	88.18	59.54	47.56	53.79	24.52
high	32.11	**54.46**	51.84	**34.21**	43.16	**10.07**
diverse	26.14	59.7	**49.72**	36.3	**42.97**	12.78
continuous	23.76	65.86	52.3	37.65	44.89	15.76
mean	**25.47**	67.05	53.35	38.93		
sd	4.44	12.85	**3.7**	5.13		

(b) NO_2 (in µg/m^3)

From Table 12(b) we see that models trained with diverse data mostly have a smaller error than those using continuous training data. Only when evaluating models with low testing data does the use of long continuous data seem beneficial, which further provides evidence that lower values dominate the data and hence are overrepresented when conventional model evaluation techniques are used to partition the data. Using diverse training data is highly advantageous for improving model performance against infrequently occurring pollution levels, while also improving performance in continuous long-term evaluation.

RESULT

Using data splits created by the proposed diverse framework to train calibration models improves model generality when tested in differing pollution concentrations and reduces the dependency of evaluation results on general pollutant levels.

5.6 Other Criteria

We next further emphasize the effect of data selection by computing an average over all model errors. The averaged model performances for each training and testing data combination are shown in Table 13. The lowest errors are highlighted in the table. As expected, and in line with our earlier findings in Section 5.4, which only considered low and high levels, the lowest error results from having similar training and testing environments. Testing models trained with continuous data

against high and diverse data shows how the continuous data does not cover high pollution levels sufficiently well. Diverse training data are more advantageous with these testing sets, showing better or similar performance rate in 80% of the models.

Model Robustness: The mean and standard deviation columns of testing data in Table 13 show how model performance changes when trained with a fixed datasets and evaluated separately with each testing set. This is indicative of model robustness, i.e., performance under various pollution conditions, since if some testing data causes significantly lower or higher errors, the our training data overrepresent certain data properties. In contrast, small differences in performance with various testing sets would indicate stable and predictable performance rate that would build confidence about the performance level. The lowest value in testing data's mean column and the second lowest value in standard deviation column indicate that in general diverse training data improves model robustness to changes in the testing data. This is due to the fact that diverse training data covers well various pollution phenomena that can occur in the testing data, i.e., data encountered in a real-world deployment.

Overfitting: Training data's mean and standard deviation rows in Table 13 show how model performance changes when trained separately with each data set and evaluated with a fixed testing set. This demonstrates the results sensitivity to the testing set's coverage of various phenomena. For example, in both tables high testing set emphasizes model performance during days with high pollution levels, and using dissimilar data (from days with low pollution levels) for training causes 70% (CO) and 62% (NO_2) higher errors than using similar high pollution level data for training. The mean and deviation thus depict how much influence training data variation has on the model performance. Both parts of Table 13 indicate that although diverse training data are the optimal choice for highest performance with diverse testing data, the sensitivity of diverse testing data to training data selection is small in case of both pollutants.

Computational efficiency: The general consensus is that the more data complex machine learning algorithms are given the better they will perform. This statement carries the assumption that all data are equally valuable learning material, which does not hold in our case. If the data distribution is highly skewed toward the most common environmental feature values, then we might have copious numbers of very similar data points to the extent that their presence diminishes the model's attention on the less common phenomena.

Besides improving robustness, diverse data selection can potentially also reduce computational needs. From Tables 12 and 13, we see that using a balanced training dataset, i.e., diverse training data, passes more descriptive information to the model and results in a very comparable model than using training dataset that has a great number of the most common values, i.e., continuous training data. The continuous training dataset has 4,678 data points (6.5 months of hourly-based measurements), whereas the diverse training dataset has only 1,800 data points (2.5 months). As the data contains many periods with low CO levels (see Section 5.2), using diverse data for training can reduce the use of redundant low pollution periods for training the models. As diverse training data has less than 40% of the continuous training dataset, this offers an opportunity for devices with limited computation capacity to (re-)train calibration models.

RESULT

Besides improving model robustness and reducing overfitting, the diverse framework offers potential for improving computational efficiency by reducing the effect of redundant periods with similar pollution concentrations.

Table 14. Averaging over All Calibration Models We See Deteriorating of Calibration as We Increase the Distance between Training and Testing Data

months apart	errors		mean		deviation	
	model avg	model std	training	testing	training	testing
0	0.61 (44%)	0.19	2.22	1.97	1.52	1.38
1	0.68 (49%)	0.21	2.52	1.97	1.7	1.38
2	0.88 (64%)	0.19	2.72	1.97	1.77	1.38
3	1.00 (74%)	0.19	2.67	1.97	1.59	1.38
4	0.86 (64%)	0.14	2.24	1.97	1.3	1.38
5	0.80 (59%)	0.14	1.7	1.97	1.09	1.38
6	0.88 (65%)	0.17	1.64	1.97	1.04	1.38
7	0.88 (65%)	0.19	1.83	1.97	1.15	1.38
8	0.74 (55%)	0.14	1.96	1.97	1.2	1.38
AVG	0.82 (60%)	0.14	2.17	1.97	1.37	1.38
	3 months (2,160 data points w/ 2:1 training-testing split)					

(a) CO (in mg/m^3)

months apart	errors		mean		deviation	
	model avg	model std	training	testing	training	testing
0	29.15 (56%)	6.16	136.01	144.67	49.11	51.59
1	34.16 (65%)	6.83	122.81	144.67	44.57	51.59
2	42.13 (81%)	7.67	113.73	144.67	46.13	51.59
3	45.30 (86%)	10.87	105.05	144.67	41.9	51.59
4	53.22 (101%)	12.78	95.86	144.67	36.72	51.59
5	50.89 (94%)	19.27	88.8	144.67	36.94	51.59
6	51.38 (95%)	23.18	92.27	144.67	38.45	51.59
7	51.68 (96%)	20.27	95.16	144.67	37.89	51.59
8	53.13 (100%)	17.13	91.74	144.67	33.39	51.59
AVG	45.67 (86%)	13.32	104.60	144.67	40.57	51.59
	3 months (2,160 data points w/ 2:1 training-testing split)					

(b) NO_2 (in $\mu g/m^3$)

However, with a gap of 3 months, the error stops increasing and shows some signs of stability. We also compare averaged model errors to the diverse errors from Table 9. In terms of model generality, this is a better implication from the performance level that calibration models have actually achieved.

5.7 Improved Holdout Validation

As the final step, we use our results to provide a simple extension to ordered holdout validation that improves its suitability for evaluating calibration models. The improvement results from separating model performance from inherited data properties, and thus yielding a better understanding about the performance implications in a real-world deployment. We also show how this method compares to our diverse data selection framework. We investigate this by selecting the testing data with increasing temporal distance from the training data. We start by choosing adjacent training and testing datasets and then gradually increasing the temporal distance between the two sets by moving the training set to appear earlier in time. The dataset specifications and model errors are shown in Table 14. We compute average errors and deviations over all calibration models with respect to the testing sets to show how model behavior stabilizes when the temporal distance between training and testing data is increased. We also compare the averaged model errors from this experiment to the diverse errors in Table 9. The results of these comparisons are shown in Table 14.

The table highlights the optimistic impressions about model performance that are to be expected if we choose the testing data to follow directly the training period. This, however, tells more about the property of the relation between training and testing data than about the function learned by any model. From the results for CO in Table 14(a), we see that all errors are smaller compared to errors indicated by diverse dataset in Table 9(a), and on average the lagged holdout testing sets reach around 60% of the error given by the diverse testing data. However, with adjacent training and testing data (lag = 0) the error magnitudes are only 44% of the diverse errors. With NO_2 in Table 14(b) instead the errors are more similar to those indicated by diverse testing data. On average, the adjacency of training and testing data reduces the error by 44% while in general the error behavior seems more stable from 4 months onwards. This implies that compared to consecutive training and testing periods, a calibration model's level of generality can better be approximated with conventional holdout validation method by leaving a sufficiently long temporal gap between training and testing data. In practice, it is hard to estimate how long of a gap should be left between training and testing data, and this is highly sensitive to seasonal variations at the location. Nevertheless, evaluating the performance as a function of temporal difference can significantly improve ordered holdout and other continuous evaluation techniques, even if diverse data are better at capturing variations in the data.

RESULT

Using the diverse framework for evaluation produces the most insights about model performance and thus is the preferred option for evaluating model performance. Ordered holdout (continuous) and other conventional evaluation methods are not reliable methods to evaluate a calibration model's goodness, since they benefit too much from similarity of training and testing data.

6 GUIDELINES AND BEST PRACTICES FOR EVALUATION

Our results have highlighted several issues and effects that can affect the performance of machine learning–based calibration techniques and that should be considered in their evaluations. Building on our findings, we draw the following six key observations that may be helpful to future evaluations of machine learning–based calibration techniques and for establishing a common pattern of best practices for evaluation.

Know what your calibration model is set to learn: The results highlighted how the models performed best when the training and the testing data were similar (see Tables 10 and 13) and had high errors when this was not the case. Careless partitioning of training and testing data can result in significant bias and give misleading results of model performance. It is vital to separate the performance of the algorithm from the characteristics and limitations of the data that are used for evaluation. Real-world deployments provide no guarantees on the distribution of data, and thus it is essential also to derive bounds on the data values where a specific performance can be achieved.

Use diverse testing data for model evaluation: The distributions of target variables in the training and testing data should always be investigated to reflect on the calibration model's performance and generality. Our results have highlighted that model behavior might be very different when infrequent environmental conditions occur (e.g., Table 9). Performing robustly in such conditions is vital, as they often represent the highest pollution concentrations and therefore situations that are most dangerous to human health. Statistical tests should therefore be used to ensure that the deployment location's estimated distribution is adequately covered. Otherwise, the suitability of calibration techniques, regardless of level of sophistication, is challenging to evaluate and next to impossible to guarantee.

Test with and without correlations in data: Autocorrelation of environmental data can be beneficial for modeling [12, 25] even if it can be dangerous in model evaluation. In Section 2, we showed that strong autocorrelation benefits calibration models with short-term data and results in low error. However, as the training and testing set size increase autocorrelation turns into a disadvantage. We also showed how conventional evaluation methods tend to form datasets with high autocorrelation (Table 6). To obtain the best insights into model performance, it is therefore important to understand to what degree the performance is a function of autocorrelation. This can be accomplished using our diverse evaluation framework with different segment lengths or using our diverse evaluation framework together with a conventional model evaluation technique.

Divide data into discontinuous sets: Selecting data points according to time (ordered holdout in Table 4) or by pollutant concentration thresholds [21] is problematic, since either the data are not descriptive to environmental conditions or then there is a possible issue of adjacent data points and information leakage. We have identified an information leakage issue with conventional evaluation methods and proposed a method that breaks overall continuity in the data without suffering from the drawbacks that existing alternatives, such as shuffled holdout evaluation, suffer (Tables 6 and 8).

By doing so, we can conduct a fair evaluation where calibration model performance reflects actual changes in environmental conditions.

Characterize the error types of calibrated values: From the end-user's perspective it is vital to recognize the operational range for calibrated measurements [21]. As we showed in Table 10, variation of features such as temperature, humidity, and a target or other pollutant can disadvantage calibration models and decrease their performance. Conventional evaluation methods do not bring out these disadvantages (Table 4). As the magnitude of errors can vary considerably across environmental conditions, it should be possible to identify environmental conditions where the model accuracy is not guaranteed to be within an acceptable level. This is particularly critical when the information is presented to decision makers as any actions should be based on accurate data. Existing work has proposed ways to visualize these errors, e.g., by calculating the expanded relative uncertainty across the entire pollutant range [32, 46]. Our work offers a mechanism to take these effects directly account as part of the evaluation and to improve the model's generalization performance against such effects.

Using a single short-term continuous period for evaluation is insufficient: Small distributional differences (Figure 3), strong autocorrelation (Table 2), and uncertain environmental feature and pollution level variations are among the key reasons to avoid using a single short-term continuous period for evaluating calibration models. If the amount of data is not sufficient for forming diverse sets, then the next best option is to have a sufficiently large gap between the training and the testing period. The gap period should not be used as a validation period for optimal parameter selection, and the similarity of training and testing data should still be investigated. Results in Section 5.3 imply that 4 months might be a suitable choice with the dataset used in this work, but naturally this depends on the location and its pollution characteristics. While having such a gap alleviates some issues, it does not solve the main problems with conventional evaluation techniques, and using the diverse data selection is a better choice for deriving insights about model robustness.

7 DISCUSSIONS

Naturally, there is room for further work and improvements. We discuss a few points here.

Improved benchmarking: Low-cost air quality sensors and their machine learning calibration has been an active research topic for over a decade [8, 34, 51], but highly varying experimental setups, datasets, metrics, and sensor technologies make it difficult to reliably compare the performance of different calibration models from previous studies. Similarly to other fields, such as computer vision [42], air quality sensor calibration research would benefit from common datasets and standardized evaluation protocols that provide a basis for replicating and comparing different studies and methods. Our work paves the road toward such development by offering guidelines for ensuring the evaluation protocol addresses model robustness and performance in real-world conditions.

Amount of data: The diverse data selector performs best when there is a sufficiently long period of data, as this guarantees the best coverage of different environmental conditions. If sufficient storage is not available, then our approach can be used in stream-based mode with multiple different segment lengths. Model training and testing can then be re-run whenever sufficient amounts of data for a given segment length is available. The algorithm for creating data partitions is computationally efficient, and, as we have shown, using diverse data can reduce model training costs by reducing redundancy in the data and offering a more balanced training data for the calibration models. Our results have shown that the use of diverse data significantly improves model

generality and robustness, and hence we would recommend using diverse data even if the length of the dataset is smaller but re-evaluating model performance as further data become available.

Sensor limits and evaluation: In our evaluation, we used values of target pollutants and environmental variables for partitioning data into training and testing sets. In practice, the sensor values are dependent on the sensor hardware and the measurement technology used for monitoring pollutants. For example, some sensors have limited capability in detecting low pollution concentrations, whereas others are vulnerable to certain environmental conditions [8]. This suggests that limitations of sensor technology should also be taken into consideration during evaluation, e.g., by placing constraints on the values that are considered reliable or partitioning data into sets that are within and outside manufacturer specified sensitivity limits.

On deep learning: In our experiments, we included a wide range of machine learning techniques for sensor calibration, but naturally there are many others that could be considered. For example, deep learning solutions based on convolutional or recurrent neural networks are increasingly being adopted for low-cost air quality sensors and have shown promising results [53]. The key challenge with deep learning is that it tends to require large amounts of data due to the inherent model complexity, and understanding their performance limits is difficult to verify without data from multiple deployments with varying conditions. Recurrent networks additionally require sufficiently continuous data to be able to learn temporal dependencies in the data, and most successful deep learning models for air quality data leverage these dependencies. For example, some models even use an input window of 1 week [53], which means that segment length should be at least 1 month to ensure that dependencies between adjacent weeks can be captured. As we have shown in the article, these long continuous segments easily result in the model emphasising background pollution levels instead of being able to work robustly across different environmental conditions. Overcoming these issues requires larger datasets from multiple locations, improved regularisation and other overfitting control techniques within the deep learning model, and further studies to ensure the deep learning models are not overfitting or emphasising performance in very limited set of environmental conditions.

Potential in other application domains: The focus of our work is on the evaluation of calibration models for low-cost air quality sensors, which is a representative example of sensing domains where robust evaluation is critical yet difficult. However, the main principle of our solution is generic and applicable in many other sensing domains where the data have a strong temporal component and there is a key target variable whose distributional characteristics vary across time. Indeed, our diverse data selection framework focuses on identifying information rich periods from continuous measurements and constructing training and testing sets where temporal dependencies between such periods are broken. Examples of other such domains where our framework is likely to be highly beneficial include activity recognition [22], continuous object recognition from video [17], acoustic sensing [1], and wireless sensing [55].

8 RELATED WORK

Calibration models: Machine learning calibration focuses on training a model that can mitigate the effects of environmental conditions and pollutant cross-sensitivities to learn a complex mapping from measurements of low-cost sensors to those given by reference instruments [8, 34]. Many machine learning models have been proposed, with MLR being one of the most popular choices for the task [24, 33, 46]. More complex and non-linear calibration function can overcome MLR's restriction to linear mapping and currently the state of the art in machine learning air quality sensor calibration models cover ANN [7, 45, 47, 52, 53], Support Vector Regression [9, 20, 53], Non-linear

Autoregressive network with exogenous inputs [54, 56], XGB [26, 53], and Random Forest with its variants [31, 57]. Comparing the approaches is currently challenging, as the evaluations of these techniques have not been systematic, with the papers considering different validation techniques, datasets, and environmental conditions [8]. Most papers have used time-based evaluation (holdout or cross-validation). As we have shown, this is not a good way to evaluate machine learning model performance. Our article addresses the evaluation of calibration techniques, providing a systematic approach for validating models in a way that adheres to regulatory standards and metrological requirements.

Model evaluation: Previous work has mainly considered the performance of calibration models through conventional model validation methods. These methods include time-based training-testing split (ordered holdout) [7, 31, 33, 47, 52, 53], non-temporal training-testing splits (shuffled holdout) [20, 21, 26, 45, 56], and CV [5, 26, 57]. Instead of selecting a random period for testing, the evaluation should emphasize the performance in different environmental conditions. Only few studies have addressed this challenge and most of them have relied on data from multiple deployment sites. Hagan et al. [21] investigate calibration accuracy in two deployment sites and consider the effect of changes in the target value's distribution on calibration errors. They also consider how using data only from a certain concentration range affects model performance. However, coarse splitting into sets containing values that are greater or smaller than a selected threshold does not remove the problem of adjacent data points leaking information from training data. DeVito et al. [11] highlight concept drift as one of the main factors for calibration degradation and propose a potential solution thorough adaptive network calibration [10]. The main issue with this approach is that evaluating learning with a short time span can result in the model not learning enough and resulting in the model overfitting on the specific environmental conditions. Vikram et al. [50] conduct thorough evaluation by using data from three different deployment sites to analyze the effect of distribution changes in training and testing site. However, when assessing single-site performances the best results might also be caused by their temporally random selection of testing data points, which is not the correct way to evaluate against time-series data. Their work also presents a two-stage model structure that compensates pollution concentrations at global and local levels to achieve higher model generality.

There are also some works that highlight issues with model performance, even if the solutions that have been proposed have only been evaluated with very short-term data. Zimmerman et al. [57] analyze calibration accuracy as a function of ambient concentration and as a function of humidity, which stands out in comparison to the more dominant performance estimation as a function of time. However, similarly to other works, they use randomly selected testing measurements and k-fold cross-validation with small folds, which is known to give optimistic results [22]. Their follow-up work suffers from the same issues [36]. Maag et al. [35] emphasize system performance in different environmental conditions but only consider short-term data comprising less than 30 days of measurement. Gu et al. [20] highlight that the data are autocorrelated and propose short-, mid-, and long-term performance estimations as different angles to consider in evaluation. However, their dataset covers only 1 week's time span, providing limited insights into long-term performance. Our work builds on these works, providing a new evaluation framework that allows for assessing the model's sensitivity to environmental conditions and distributional differences, in line with requirements imposed by regulatory standards. The procedure we develop is generic and can be used equally with single-site and multi-site deployments.

Evaluation issues with time-series data: Our work is also inspired by previous works highlighting evaluation issues with time-series data. The machine learning model evaluation techniques covered in this work (see Section 4) are among the most common techniques to assess model

performance. However, their applicability to a domain requires careful consideration. Raschka [40] points out that these evaluation methods generally make the assumption that all data are independent and identically distributed, which does not hold for environmental data. Bergmeir et al. [3] found techniques like shuffled holdout or cross-validation to be biased, since they effectively peek into the testing data distribution. Our experiment results in Table 8 are in line with their findings. Hammerla and Plötz showed how overlapping segments have to be carefully divided into training and testing data in activity recognition tasks to avoid correlations between them [22]. Although the domain and problem are different, it highlights the issue of feature information passing from training to testing data through autocorrelation. Forman et al. [18] show how averaging of less common error types (e.g., high pollution phenomena in our case) is a problem in model evaluation. These factors have inspired and driven our work to create a more sophisticated evaluation method that can assess machine learning calibration model performance with environmental time-series data.

9 CONCLUSIONS AND SUMMARY

Low-cost sensors have significant potential to support air quality research by increasing the resolution of data. Before scientific studies can rely on low-cost sensor data, researchers need to be convinced about the quality and validity of measurements. This requires transparent evaluation protocols that are able to highlight not only the overall performance of calibration models but also their limitations and performance bounds. We have contributed a novel evaluation protocol for machine learning calibration of low-cost air quality sensors that offers a rigorous evaluation framework that provides better insights into the performance of calibration models in real-world deployments and can be used to identify conditions where the accuracy of models is sufficient (or insufficient) for studies. We have also shown how existing approaches, based on conventional model evaluation techniques, fail in achieving this as they tend to result in test data that are close to the training data, either temporally or in terms of distributional characteristics, limiting the insights that can be garnered from model evaluation. Our model overcomes these limitations, creating testing conditions that better capture the statistical and distributional characteristics of data without being correlated over time, while at the same time providing better insights into model performance. Our model can also improve training of calibration models, offering improved robustness, reduced risk of overfitting, and potential for faster runtime performance. Based on our results, we derived guidelines for future evaluations to establish a roadmap for the way forward.

ACKNOWLEDGMENTS

The authors are grateful to the anonymous reviews and editor for their constructive comments.

REFERENCES

[1] Yang Bai, Li Lu, Jerry Cheng, Jian Liu, Yingying Chen, and Jiadi Yu. 2020. Acoustic-based sensing and applications: A survey. *Comput. Netw.* 181, 9 (2020), 107447. https://doi.org/10.1016/j.comnet.2020.107447

[2] Petra Bauerová, Adriana Šindelářová, Šitěpán Rychlík, Zbyněik Novák, and Josef Keder. 2020. Low-cost air quality sensors: One-year field comparative measurement of different gas sensors and particle counters with reference monitors at Tušimice observatory. *Atmosphere* 11, 5 (2020), 492. https://doi.org/10.3390/atmos11050492

[3] Christoph Bergmeir and José Manuel Benítez. 2012. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* Data Mining for Software Trustworthiness. 191, 15 (2012), 192–213. https://doi.org/10.1016/j.ins.2011.12.028

[4] C. Borrego, A. M. Costa, J. Ginja, M. Amorim, M. Coutinho, K. Karatzas, Th Sioumis, N. Katsifarakis, K. Konstantinidis, S. De Vito, et al. 2016. Assessment of air quality microsensors versus reference methods: The EuNetAir Joint Exercise. *Atmos. Environ.* 147 (2016), 246–263. DOI: https://doi.org/10.1016/j.atmosenv.2016.09.050

[5] C. Borrego, J. Ginja, M. Coutinho, C. Ribeiro, K. Karatzas, Th Sioumis, N. Katsifarakis, K. Konstantinidis, S. De Vito, E. Esposito, M. Salvato, P. Smith, N. André, P. Gérard, L. A. Francis, N. Castell, P. Schneider, M. Viana, M. C. Minguillón, W. Reimringer, R. P. Otjes, O. von Sicard, R. Pohle, B. Elen, D. Suriano, V. Pfister, M. Prato, S. Dipinto, and M. Penza.

2018. Assessment of air quality microsensors versus reference methods: The EuNetAir Joint Exercise—Part II. *Atmos. Environ.* 193 (2018), 127–142. https://doi.org/10.1016/j.atmosenv.2018.08.028

[6] Chen-Chia Chen, Chih-Ting Kuo, Ssu-Ying Chen, Chih-Hsing Lin, Jin-Ju Chue, Yi-Jie Hsieh, Chun-Wen Cheng, Chieh-Ming Wu, and Chun-Ming Huang. 2018. Calibration of low-cost particle sensors by using machine-learning method. In *Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems (APCCAS'18)*. IEEE, 111–114. https://doi.org/10.1109/APCCAS.2018.8605619

[7] Yun Cheng, Xiucheng Li, Zhijun Li, Shouxu Jiang, Yilong Li, Ji Jia, and Xiaofan Jiang. 2014. AirCloud: A cloud-based air-quality monitoring system for everyone. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems (SenSys'14)*. ACM, New York, NY, 251–265. http://doi.acm.org/10.1145/2668332.2668346

[8] Francesco Concas, Julien Mineraud, Eemil Lagerspetz, Samu Varjonen, Xiaoli Liu, Kai Puolamäki, Petteri Nurmi, and Sasu Tarkoma. 2021. Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis. *ACM Trans. Sens. Netw.* 17, 2 (2021), 20:1–20:44. https://doi.org/10.1145/3446005

[9] José María Cordero, Rafael Borge, and Adolfo Narros. 2018. Using statistical methods to carry out in field calibrations of low cost air quality sensors. *Sens. Actuat. B: Chem.* 267 (2018), 245–254. https://doi.org/10.1016/j.snb.2018.04.021

[10] Saverio De Vito, Girolamo Di Francia, Elena Esposito, Sergio Ferlito, Fabrizio Formisano, and Ettore Massera. 2020. Adaptive machine learning strategies for network calibration of IoT smart air quality monitoring devices. *Pattern Recogn. Lett.* 136 (2020), 264–271. https://doi.org/10.1016/j.patrec.2020.04.032

[11] Saverio De Vito, Elena Esposito, Nuria Castell, Philipp Schneider, and A. Bartonova. 2020. On the robustness of field calibration for smart air quality monitors. *Sens. Actuat. B: Chem.* 310 (2020), 127869. https://doi.org/10.1016/j.snb.2020.127869

[12] Luis A. Díaz-Robles, Juan C. Ortega, Joshua S. Fu, Gregory D. Reed, Judith C. Chow, John G. Watson, and Juan A. Moncada-Herrera. 2008. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmos. Environ.* 42, 35 (2008), 8331–8340. https://doi.org/10.1016/j.atmosenv.2008.07.020

[13] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. The reusable holdout: Preserving validity in adaptive data analysis. *Science* 349, 6248 (2015), 636–638. https://doi.org/10.1126/science.aaa9375

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

[15] European Environment Agency. 2010. Guide To The Demonstration of Equivalence of Ambient Air Monitoring Methods. Retrieved from http://ec.europa.eu/environment/air/quality/legislation/pdf/equivalence.pdf.

[16] European Environment Agency. 2018. Air Quality in Europe—2018 Report. Retrieved from https://www.eea.europa.eu/publications/air-quality-in-europe-2018.

[17] Huber Flores, Naser Hossein Motlagh, Agustin Zuniga, Mohan Liyanage, Monica Passananti, Sasu Tarkoma, Moustafa Youssef, and Petteri Nurmi. 2021. Toward large-scale autonomous marine pollution monitoring. *IEEE IoT Mag.* 4, 1 (2021), 40–45.

[18] George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.* 12, 1 (November 2010), 49–57. https://doi.org/10.1145/1882471.1882479

[19] Yi Gao, Wei Dong, Kai Guo, Xue Liu, Yuan Chen, Xiaojin Liu, Jiajun Bu, and Chun Chen. 2016. Mosaic: A low-cost mobile sensing system for urban air quality monitoring. In *Proceedings of the 35th Annual IEEE International Conference on Computer Communications (INFOCOM'16)*. IEEE, 1–9. https://doi.org/10.1109/INFOCOM.2016.7524478

[20] K. Gu, J. Qiao, and W. Lin. 2018. Recurrent air quality predictor based on meteorology- and pollution-related factors. *IEEE Trans. Industr. Inf.* 14, 9 (2018), 3946–3955. https://doi.org/10.1109/TII.2018.2793950

[21] D. H. Hagan, G. Isaacman-VanWertz, J. P. Franklin, L. M. M. Wallace, B. D. Kocar, C. L. Heald, and J. H. Kroll. 2018. Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments. *Atmos. Meas. Techn.* 11, 1 (2018), 315–328. https://doi.org/10.5194/amt-11-315-2018

[22] Nils Y. Hammerla and Thomas Plötz. 2015. Let's (Not) stick together: Pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'15)*. ACM, New York, NY, 1041–1051. https://doi.org/10.1145/2750858.2807551

[23] David Hasenfratz, Olga Saukh, Silvan Sturzenegger, and Lothar Thiele. 2012. Participatory air pollution monitoring using smartphones. In *Proceedings of the 2nd International Workshop on Mobile Sensing*. Academic Press,1–5.

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

[25] Joel Horowitz and Samir Barakat. 1979. Statistical analysis of the maximum concentration of an air pollutant: Effects of autocorrelation and non-stationarity. *Atmos. Environ.* 13, 6 (1979), 811–818. https://doi.org/10.1016/0004-6981(79)90272-5

[26] Nicholas E. Johnson, Bartosz Bonczak, and Constantine E. Kontokosta. 2018. Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment. *Atmos. Environ.* 184 (2018), 9–16. https://doi.org/10.1016/j.atmosenv.2018.04.019

[27] Heikki Junninen, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen, and Mikko Kolehmainen. 2004. Methods for imputation of missing values in air quality data sets. *Atmos. Environ.* 38, 18 (2004), 2895–2907.

[28] Fadi Kizel, Yael Etzion, Rakefet Shafran-Nathan, Ilan Levy, Barak Fishbain, Alena Bartonova, and David M. Broday. 2018. Node-to-node field calibration of wireless distributed air pollution sensor network. *Environ. Pollut.* 233 (2018), 900–909. https://doi.org/10.1016/j.envpol.2017.09.042

[29] Hoochang Lee, Jiseock Kang, Sungjung Kim, Yunseok Im, Seungsung Yoo, and Dongjun Lee. 2020. Long-term evaluation and calibration of low-cost particulate matter (PM) sensor. *Sensors* 20, 13 (2020), 24 pages. https://doi.org/10.3390/s20133617

[30] Alastair Lewis and Peter Edwards. 2016. Validate personal air-pollution sensors. *Nature* 535, 7410 (2016), 29–31. DOI: https://doi.org/10.1038/535029a

[31] Yuxiang Lin, Wei Dong, and Yuan Chen. 2018. Calibrating low-cost sensors by a two-phase learning approach for urban air quality measurement. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 2, 1, Article 18 (March 2018), 18 pages. https://doi.org/10.1145/3191750

[32] Hai-Ying Liu, Philipp Schneider, Rolf Haugen, and Matthias Vogt. 2019. Performance assessment of a low-cost PM2.5 sensor for a near four-month period in Oslo, Norway. *Atmosphere* 10, 2 (2019), 41. DOI: https://doi.org/10.3390/atmos10020041

[33] Balz Maag, Olga Saukh, David Hasenfratz, and Lothar Thiele. 2016. Pre-deployment testing, augmentation and calibration of cross-sensitive sensors. In *Proceedings of the 2016 International Conference on Embedded Wireless Systems and Networks (EWSN'16)*. 169–180.

[34] B. Maag, Z. Zhou, and L. Thiele. 2018. A survey on sensor calibration in air -pollution monitoring deployments. *IEEE IoT J.* 5, 6 (2018), 4857–4870. https://doi.org/10.1109/JIOT.2018.2853660

[35] Balz Maag, Zimu Zhou, and Lothar Thiele. 2018. W-Air: Enabling personal air pollution monitoring on wearables. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 2, 1, Article 24 (March 2018), 25 pages. https://doi.org/10.1145/3191756

[36] C. Malings, R. Tanzer, A. Hauryliuk, S. P. N. Kumar, N. Zimmerman, L. B. Kara, A. A. Presto, and R. Subramanian. 2019. Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. *Atmos. Meas. Techn.* 12, 2 (2019), 903–920. https://doi.org/10.5194/amt-12-903-2019

[37] Naser Hossein Motlagh, Eemil Lagerspetz, Petteri Nurmi, Xin Li, Samu Varjonen, Julien Mineraud, Matti Siekkinen, Andrew Rebeiro-Hargrave, Tareq Hussein, Tuukka Petäjä, Markku Kulmala, and Sasu Tarkoma. 2020. Toward massive scale air quality monitoring. *IEEE Commun. Mag.* 58, 2 (2020), 54–59.

[38] Naser Hossein Motlagh, Martha A. Zaidan, Pak L. Fung, Eemil Lagerspetz, Kasimir Aula, Samu Varjonen, Matti Siekkinen, Andrew Rebeiro-Hargrave, Tuukka Petäjä, Yutaka Matsumi, Markku Kulmala, Tareq Hussein, Petteri Nurmi, and Sasu Tarkoma. 2021. Transit pollution exposure monitoring using low-cost wearable sensors. *Transport. Res. D: Transport Environ.* 98 (2021), 102981. DOI: https://doi.org/10.1016/j.trd.2021.102981

[39] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. IEEE Computer Society, 2233–2241. https://doi.org/10.1109/CVPR.2017.240

[40] Sebastian Raschka. 2020. Model evaluation, model selection, and algorithm selection in machine learning. arXiv:1811.12808 [cs.LG]. Retrieved from https://arxiv.org/abs/1811.12808.

[41] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 4331–4340. http://proceedings.mlr.press/v80/ren18a.html.

[42] Olga Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, Michael S. Bernstein, A. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115 (2015), 211–252. DOI: https://doi.org/10.1007/s11263-015-0816-y

[43] Joshua S. Apte, Kyle Messier, Shahzad Gani, Michael Brauer, Thomas Kirchstetter, Melissa Lunden, Julian D. Marshall, Christopher J. Portier, Roel Vermeulen, and Steven P. Hamburg. 2017. High-resolution air pollution mapping with Google Street View cars: Exploiting big data. *Environ. Sci. Technol.* 51, 12 (2017), 6999–7008. https://doi.org/10.1021/acs.est.7b00891

[44] Ravi Sahu, Ayush Nagal, Kuldeep Kumar Dixit, Harshavardhan Unnibhavi, Srikanth Mantravadi, Srijith Nair, Yogesh Simmhan, Brijesh Mishra, Rajesh Zele, Ronak Sutaria, et al. 2021. Robust statistical calibration and characterization of portable low-cost air quality monitoring sensors to quantify real-time O_3 and NO_2 concentrations in diverse environments. *Atmos. Meas. Techn.* 14, 1 (2021), 37–52. DOI: https://doi.org/10.5194/amt-14-37-2021

[45] M. Si, Y. Xiong, S. Du, and K. Du. 2020. Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods. *Atmos. Meas. Techn.* 13, 4 (2020), 1693–1707. https://doi.org/10.5194/amt-13-1693-2020

[46] Laurent Spinelle, Michel Gerboles, Maria Gabriella Villani, Manuel Aleixandre, and Fausto Bonavitacola. 2015. Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide. *Sens. Actuat. B: Chem.* 215 (2015), 249–257. https://doi.org/10.1016/j.snb.2015.03.031

[47] Laurent Spinelle, Michel Gerboles, Maria Gabriella Villani, Manuel Aleixandre, and Fausto Bonavitacola. 2017. Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO_2. *Sens. Actuat. B: Chem.* 238 (2017), 706–715. https://doi.org/10.1016/j.snb.2016.07.036

[48] Bartosz Szulczyński and Jacek Gębicki. 2017. Currently commercially available chemical sensors employed for detection of volatile organic compounds in outdoor and indoor air. *Environments* 4, 1 (3 2017), 21. https://doi.org/10.3390/environments4010021

[49] Vera van Zoest, Frank B. Osei, Alfred Stein, and Gerard Hoek. 2019. Calibration of low-cost NO2 sensors in an urban air quality network. *Atmos. Environ.* 210 (2019), 66–75. DOI: https://doi.org/10.1016/j.atmosenv.2019.04.048

[50] S. Vikram, A. Collier-Oxandale, M. H. Ostertag, M. Menarini, C. Chermak, S. Dasgupta, T. Rosing, M. Hannigan, and W. G. Griswold. 2019. Evaluating and improving the reliability of gas-phase sensor system calibrations across new locations for ambient measurements and personal exposure monitoring. *Atmos. Meas. Techn.* 12, 8 (2019), 4211–4239. https://doi.org/10.5194/amt-12-4211-2019

[51] S. De Vito, E. Massera, M. Piga, L. Martinotto, and G. Di Francia. 2008. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sens. Actuat. B: Chem.* 129, 2 (2008), 750–757. http://www.sciencedirect.com/science/article/pii/S0925400507007691.

[52] Saverio De Vito, Marco Piga, Luca Martinotto, and Girolamo Di Francia. 2009. CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic Bayesian regularization. *Sens. Actuat. B: Chem.* 143, 1 (2009), 182–191. https://doi.org/10.1016/j.snb.2009.08.041

[53] H. Yu, Q. Li, R. Wang, Z. Chen, Y. Zhang, Y. Geng, L. Zhang, H. Cui, and K. Zhang. 2020. A deep calibration method for low-cost air monitoring sensors with multi-level sequence modeling. *IEEE Trans. Instrum. Meas.* 69, 9 (2020), 7167–7179. DOI: https://doi.org/10.1109/TIM.2020.2978596

[54] Martha Arbayani Zaidan, Naser Hossein Motlagh, Pak L. Fung, David Lu, Hilkka Timonen, Joel Kuula, Jarkko V. Niemi, Sasu Tarkoma, Tuukka Petäjä, Markku Kulmala, et al. 2020. Intelligent calibration and virtual sensing for integrated low-cost air quality sensors. *IEEE Sens. J.* 20, 22 (2020), 13638–13652. DOI: https://doi.org/10.1109/JSEN.2020.3010316

[55] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. 2018. CrossSense: Towards cross-site and large-scale WiFi sensing. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom'18)*. ACM, New York, NY, 305–320. https://doi.org/10.1145/3241539.3241570

[56] Y. Zhou, S. De, G. Ewa, C. Perera, and K. Moessner. 2018. Data-driven air quality characterization for urban environments: A case study. *IEEE Access* 6 (2018), 77996–78006. https://doi.org/10.1109/ACCESS.2018.2884647

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

Received 29 May 2021; revised 24 November 2021; accepted 19 January 2022