THE CYPRUS INSTITUTE

Practicalities in Machine Learning Calibration of Air Quality Measurements from Low-Cost Gas Sensors

by

LANGAT KIPKEMOI VINCENT

MASTER OF SCIENCE / MASTER OF PHILOSOPHY

A DISSERTATION SUBMITTED TO THE CYPRUS INSTITUTE TOWARDS THE FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE / MASTER OF PHILOSOPHY IN ENVIRONMENTAL SCIENCE

NICOSIA, MARCH 2022

VALIDATION PAGE

Master of Science/Philosophy Candidate : **Langat Kipkemoi Vincent**

Title of Thesis : **Practicalities in Machine Learning Calibration of Air Quality Measurements from Low-Cost Gas Sensors.**

The dissertation was submitted towards the fulfillment of the requirements for the degree of Master of Science/Master of Philosophy at the Graduate School of The Cyprus Institute and was approved on the 30^{th} of April 2022 by the members of the Academic Committee of The Cyprus Institute.

Examination Committee :

- **Professor George Biskos(Chair)**
- **Professor Jean Sciare**
- **Assistant Professor Mihalis Nicolaou**
- **Associate Research Scientist Spyros Bezantakos**

DECLARATION

This dissertation was submitted towards the fulfillment of the requirements for the award of a Master of Science/ Master of Philosophy Degree from The Cyprus Institute. It is the product of my own original work, unless otherwise mentioned through references, notes or other statements.

Langat Kipkemoi Vincent

Abstract

Low cost sensors (LCSs) that measure the concentrations of gaseous pollutants hold great promises for Air Quality Monitoring (AQM) as they can improve the spatio-temporal resolution of observational networks. The performance of LCSs, however, is affected by a number of factors including temperature and relative humidity of ambient air, as well as cross-sensitivities with gaseous species other than the target gas, thereby deteriorating the quality of their measurements. To address these issues, data from LCSs can be calibrated against reference instruments using machine learning (ML) algorithms. In this work, I have evaluated the performance of a number of ML algorithms for calibrating measurements from CO, NO_2, O_3 and SO_2 LCSs against respective reference measurements. The best model is then used to determine (1) the influence of temporal resolution of the measurements to the calibration performance, (2) the minimum fraction of data needed for model training while maintaining the quality of calibrated measurements within acceptable levels, and (3) the ideal calibration frequency with collocated reference measurements. My results show that the quality of LCS measurements improve significantly for all sensors after ML calibration, with Random Forest (RF) being the best performing algorithm, corroborating previous works. By varying the temporal resolution of the training data from 1 h to 2 min, the performance of the RF model in terms of the normalized root mean squared error (NRMSE) and the relative expanded uncertainty calculated at maximum observed concentration (REU_{max}) improves by 11-21 %. The results also suggest that the minimum fraction of data required for training the ML models depends on the frequency of carrying out collocated measurements with reference instruments and using the resulting datasets for training the calibration model. If the calibrations are carried out on a monthly basis, ca. 50 % of the period is needed for collecting data to train the RF algorithm and qualify the LCSs for indicative measurements as defined by the EU directive (2008/50/EC). If the training is carried out every 3 or 6 months by sampling the training data continuously, then ca. 60 % of the measuring period is required for collecting training data. In those cases, if the sampling of the training data is made over specific periods every month, but the entire training dataset is used to calibrate the measurements over 3 or 6 months, the amount of data required for qualifying the LCSs for indicative measurements can significantly reduce down to 22 %. This, however, would require that the measurements from the LCSs are calibrated retrospectively, which for specific applications is not such of a problem.

Contents

1 Introduction

The adverse effects of air pollution on human health and the environment warrant for continuous monitoring of air quality. Traditionally, air quality monitoring for regulatory purposes is carried out at a number of observational stations equipped with reference-grade air quality monitors. The number of stations that can be operated within a given geographical area, however, is limited by the high installation and operation cost of the instruments (Kumar and Sahu, 2021; Arroyo et al., 2021). This has spurred the development of low-cost air quality gas sensors (Kumar et al., 2015) that are significantly less expensive compared to their reference-grade instrument counterparts, motivating research towards improving their very sensing nanomaterials materials to meet requirements in air quality monitoring (Baranwal et al., 2022; Isaac et al., 2022). As a matter of fact, low-cost sensors (LCSs) are increasingly being used to complement the reference instruments in measuring the concentration of air pollutants, as they can significantly increase spatio-temporal resolution of air quality monitoring (AQM) networks (Lewis et al., 2018; Chen et al., 2018; Zuidema et al., 2021; Nowack et al., 2021; Zimmerman, 2022). Their low cost and ease of operation also allow personal use for monitoring air quality in the indoor and outdoor environments, providing great means for assessing the impacts of air pollutants on human health (Schäfer et al., 2021; Patra et al., 2021).

Although LCSs hold great promises for expanding existing AQM networks, they have a number of technical limitations including high limits of detection, low precision and accuracy, and signal drift, which at the moment prohibit their widespread use (Papaconstantinou et al., 2023). Some of these technical limitations are related to environmental conditions they are operated under (i.e., temperature and Relative Humidity; RH), and cross-sensitivities to other gaseous pollutants. As a result, when LCSs are deployed for field measurements at ambient conditions where these factors vary substantially, the deviation between the measurements they provide and those reported by reference instruments can become large, failing to meet the requirements defined by environmental regulatory agencies (Samad et al., 2020; Schäfer et al., 2021). According to the European Union (EU) directive on ambient and cleaner air (2008/50/EC), for air quality measurements to qualify as reference measurements, they should exhibit relative expanded uncertainties (REUs) lower than 15 %; i.e., something that is achieved by reference-grade instruments typically used in AQ monitoring. If the REUs are higher than that and lower than 25 % for CO, NO_2 and SO_2, and 30 % for O_3, the measurements can be qualified as indicative (EU-directive, 2008; Equivalence, 2010), which can be useful for a number of applications; e.g., for identifying pollution sources and hotpots in cities, and determining the spatio-temporal variability in the concentration of air pollutants (Schäfer et al., 2021).

The poor agreement between LCS and reference-grade instrument measurements is not a surprise considering that the former typically come calibrated by the manufacturers at laboratory conditions, which do not capture the complexity encountered in the field (Zuidema et al., 2021). To address this gap, previous efforts have focused on the post calibration of LCS measurements using machine learning (ML) models that take into account the variability of temperature, relative humidity and interfering pollutants, as well as other factors that may influence measurements carried out in the field (Zimmerman et al., 2018; Ferrer-Cid et al., 2020; Okafor and Delaney, 2020; Song et al., 2020; Nowack et al., 2021; Patra et al., 2021; Kumar and Sahu, 2021; Vajs et al., 2021). To

do so, one needs to collocate the LCSs with respective reference instruments for a certain amount of time in order to gather data for training the calibration models.

In this work, the concentration measurements of atmospheric pollutants recorded by LCSs and reference instruments over a period of 6 months in a traffic station in the city of Nicosia, Cyprus is used to train and evaluate the performance of five ML algorithms; namely Linear Regression (LR: (Kumar and Sahu, 2021)); Support Vector Regression (SVR: (Kumar and Sahu, 2021)); Random Forest Regressor (RF: (Zimmerman et al., 2018)); Artificial Neural Network (ANN: (Spinelle et al., 2015)); and Extreme Gradient Boosting (XGBoost: (Chen and Guestrin, 2016)). The best model is then used to determine the effects of a number of practical parameters including the temporal resolution of the training data, how the training data is sampled and the frequency of training/calibration on the performance of the algorithm, as well as the minimum fraction of data needed for training without compromising significantly their quality (i.e., the accuracy of the resulted post calibrated measurements).

2 Experimental

In this work, I have employed four electrochemical sensors manufactured by Alphasense to measure the concentrations of CO, NO_2, O_3 and SO_2. More information about the technical specifications of each sensor is provided by Papaconstantinou et al. (2023) and in Table 2 in the Appendix. Each sensor provides two raw analog voltage signals: one from the working electrode and the other from an auxiliary electrode that serves as the zero background signal against which the signal of the working electrode is compared (Masic et al., 2018; ANN803-05, 2019; Arroyo et al., 2021). These signals are then converted to corresponding concentrations, expressed in ppb, using calibration equations derived by the manufacturer under laboratory conditions (ANN803-05, 2019). For the purpose of my analysis, I will refer to the concentrations calculated based on such equations as laboratory (LAB) calibrated concentrations.

Figure 1: Satellite image showing the location of the traffic station where the measurements were carried out. The station is located next to one of the busiest avenues, i.e., Strovolou Avenue, in Nicosia, Cyprus, and its coordinates are $35°09'07.2''$ N and $33°20'52.0''$ E.

The measurements were carried out between 2 October 2019 and 31 March 2020 at one of the national regulatory air quality monitoring stations in Nicosia. As shown in Figure 1, the station is located approximately 10 m away from one of the busiest avenues, and is equipped with reference-grade instruments for measuring the concentration of CO, NO_2, O_3 and SO_2. These instruments are calibrated at least once every 3 months, and after maintenance. More detailed information on the technical specifications of these instruments is provided in Table 3 in the Appendix.

2.1 ML calibration procedure

Figure 2 shows the steps followed for ML calibration of the LCS measurements. The data from the reference instruments, as well as the temperature and RH recorded by the sensors located close to the reference station, had a time resolution of 2 min, whereas the measurements from the LCSs were recorded every 2 s. To align the dataset, the LCS measurements were averaged every 2 min and concatenated with the respective reference data, temperature and RH. Subsequently, data cleaning was done by dropping all rows with the missing values and those with negative net sensor signals (i.e., obtained by subtracting the signals reported by the auxilliary electrode from those reported by working electrode). From the cleaned dataset, the net sensor signals (NSS), temperature, relative humidity (RH), month, day of week and hour were selected as input variables (also referred to as features) for each algorithm, whereas the respective reference concentrations were used as response (also

referred to as dependent) variables. The temporal variables (month, day of week, and hour) were included as input to unravel the importance of monthly, weekly and diurnal variabilities. For the ML calibration of the NO_2 and O_3 LCSs we also included the O_3 and NO_2 reference concentrations, respectively, as input variables in order to account for cross-sensitivities. According to manufacturer and literature reports, the O_3 LCS is affected by NO_2 (ANN803-05, 2019; Pang et al., 2017), and vice versa (Maag et al., 2016).

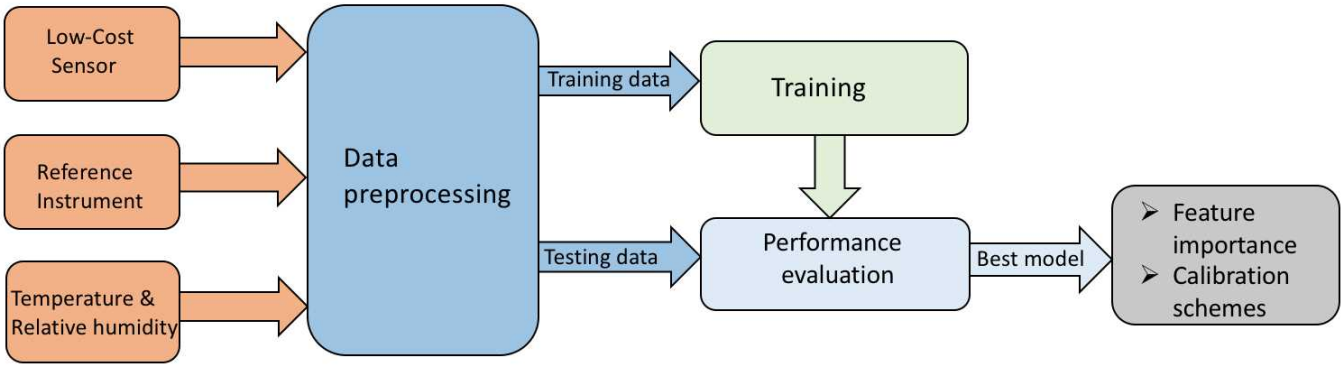

Figure 2: Flow diagram showing the stages of the ML calibration process employed in this work. First we merge the measurements from the LCSs, the reference instruments and the meteorological sensors. The resulting dataset was then pre-processed and segmented into training and testing sets. The training sets were used to train the ML models while the testing sets were used to evaluate their performance. The best-performing model was then used to assess (1) the significance of input variables to model performance, and (2) how different calibration practices affect the fraction of training data required.

LCS calibration was performed using five ML models as mentioned in the introduction: LR, SVR, RF, ANN and XGBoost. All calculations were carried out in the anaconda environment using python version 3.8.6. To train and evaluate the performance of these models, the 6-month datasets were split into a training and a testing sets. The training sets were derived from the first 80 % of the data of each month and used to train and tune the parameters of the models. The testing sets, which were used to evaluate the performance of the models, comprised of the remaining 20 % of the data. The hyper-parameters for the SVR and ANN ML models were tuned through a 5-fold cross-validation and grid search, whereas those of the RF and XGBoost algorithms were auto-tuned through the AutoML library (FLAML) developed by Microsoft (Wang et al., 2021). The optimal values for the main hyper-parameters obtained for each model are provided in Table 4 in the Appendix. The other parameters for each model were kept at their default values. The calibration performance of the tuned models were then evaluated based on a number of statistical indicators and the best-performing model was used to determine the importance of each feature in the model and assess the effect of (1) the temporal resolution of the training data, (2) how the training data are sampled, and (3) the calibration frequency on the fraction of data required for training the ML algorithms. We should note that the importance of each input variable in the model was determined using the permutation feature importance function within the Scikit-learn ML library.

2.1.1 Data splitting schemes

To investigate the effect of the maner in which the training data is sampled on calibration frequency and the amount of data required for training, I have performed calibration at a frequency of 1, 3 and 6 months using data from CO, NO_2 and O_3 LCSs by following the data splitting schemes shown in Figure 3. For 1 month calibration, the training data is obtained at the start of each month (cf. Figure 3a), while for 3 and and 6 month calibrations, two data splitting cases were considered. The first case is where the training data is obtained at the start of each calibration period (cf. Figure 3b,d) while in the second case, the training data is sampled from all the months (i.e., at the start of each month; cf. Fig.3e,f). We should note here that the cases where the training data is derived at the middle and the end of each calibration period were considered (cf. Fig. 13 and 14 in the Appendix). The results obtain when using these data splitting schemes, were the same as those obtained by using data splitting schemes in Fig.3, and are thus are not considered in this work.

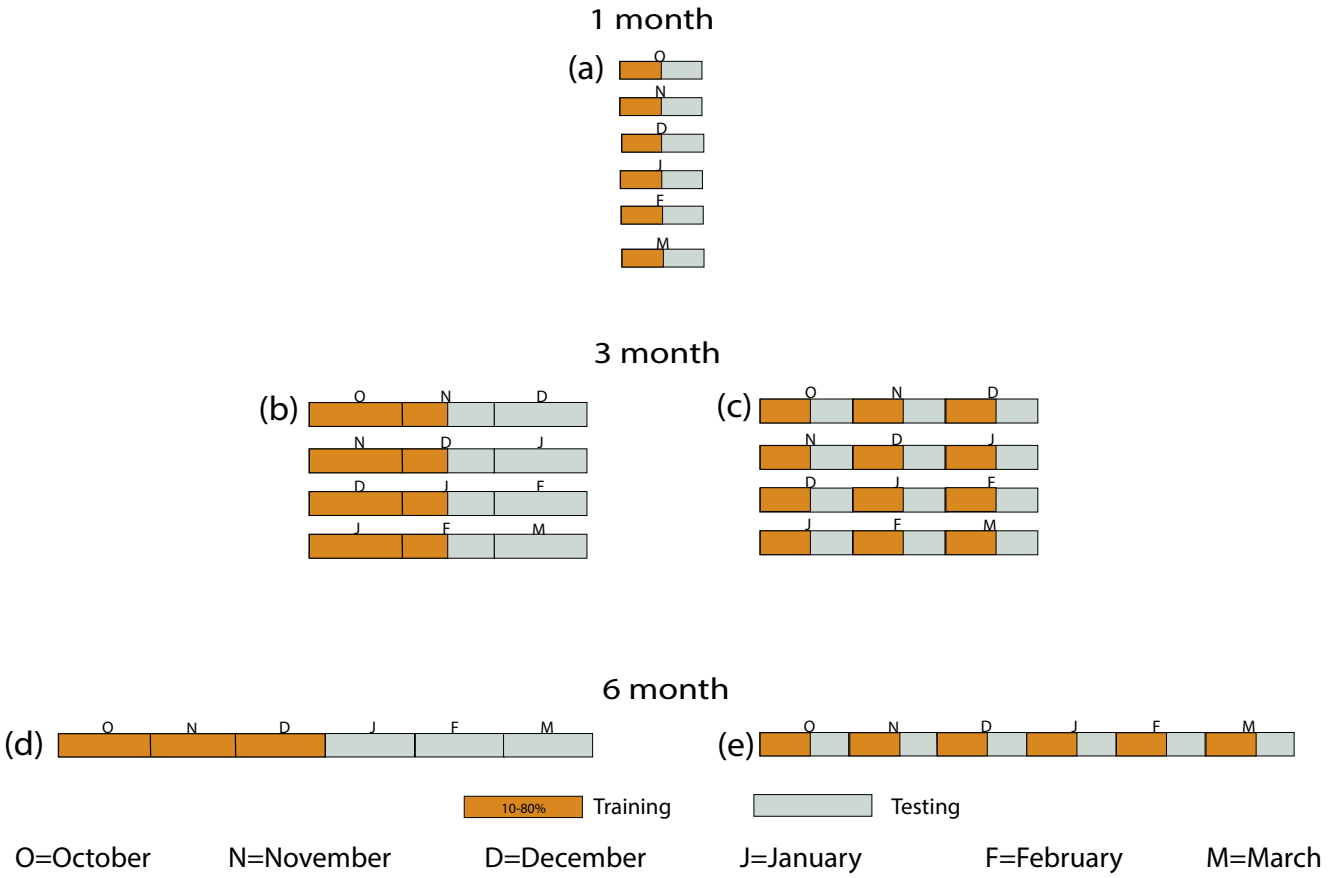

Figure 3: Schemes of splitting the data for training (orange) and testing (gray) the ML models over periods of 1, 3 and 6 months. For 1 month calibration the training data is obtained continuously at the start of each month. For the 3 and 6 month calibrations, two cases are considered: 1. the training data is obtained continuously at the start of each calibration period (b,d), and 2. the training data is sampled at the start of each month (c,e).

2.2 Model evaluation

Performance evaluation of the models was done in two stages. In the first stage we evaluated the accuracy of their calibration based on Pearson correlation coefficient (r), coefficient of determination (R^2) and normalized root mean squared error (NRMSE), given respectively as:

$$r = \frac{\sum_{t=1}^{n}(Cal_t - \overline{Cal})(Ref_t - \overline{Ref})}{\sqrt{\sum_{t=1}^{n}(Cal_t - \overline{Cal})^2 \sum_{t=1}^{n}(Ref_t - \overline{Ref})^2}} \tag{2.1}$$

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(Cal_t - Ref_t)^2}{\sum_{t=1}^{n}(Ref_t - \overline{Ref})^2} \tag{2.2}$$

$$NRMSE = \frac{\sqrt{\frac{\sum_{t=1}^{n}(Cal_t - Ref_t)^2}{n}}}{\overline{Ref}} \tag{2.3}$$

Here Cal_t, $\overline{Cal}$, Ref_t, $\overline{Ref}$ and n denote respectively the calibrated concentration at time t, the mean of calibrated concentrations, the reference concentration at time t, the mean of the reference concentrations, and the total number of data points. The Pearson correlation coefficient, r, provides information on the strength of associations between the calibrated concentrations and their corresponding reference concentrations. R^2 is a measure of the extend to which the input variables used in the models explain the variation in the dependent variable, whereas NRMSE expresses the error between the calibrated and reference concentrations normalized by the mean of the reference concentrations.

In the second stage,target diagrams were created to infer the level of bias and variance in the model, which are key parameters for diagnosing whether a model is over-fitting or under-fitting the data. An under-fitted model has high bias and low variance, whereas the opposite is true for an over-fitted model (Kumar and Sahu, 2021; Yu et al., 2006). For LCS calibration, the root mean square error (RMSE) of the calibration model captures both the bias and the variance in the data provided by LCSs, and is typically used to create target diagrams (Jolliff et al., 2009) and to visualise the performance of different models. The RMSE is determined as:

$$RMSE^2 = MBE^2 + CRMSE^2 \tag{2.4}$$

Here, MBE and CRMSE are the mean bias error and the centred root mean square error (Zimmerman et al., 2018; Thunis et al., 2012), expressed as:

$$MBE = \frac{1}{n}\sum_{t=1}^{n}(Cal_t - Ref_t) \tag{2.5}$$

$$CRMSE = \sqrt{\sigma_{Cal}^2 + \sigma_{Ref}^2 - 2r\sigma_{Cal}\sigma_{Ref}} \tag{2.6}$$

where σ_{Cal} and σ_{Ref} are the standard deviation of calibrated and reference concentrations, respectively.

2.3 Data quality assessments

The goal of calibrating measurements from LCSs using ML models is to improve their accuracy in order to meet the data quality levels required for specific applications. Here we used the EU directive 2008/50/EC as a guideline to assess the quality of the data produced by the LCSs and the different calibration schemes tried. According to the directive, the LCSs can primarily be used for indicative measurements (i.e measurements with less stringent uncertainty requirements in comparison to regulatory measurements), in which case the uncertainty limits are set to 25 % for CO, NO_2 and SO_2, and 30 % for O_3. These uncertainty requirements have to be met at specific limit values for CO, NO_2 and SO_2, and at the target value for O_3. Limit values are specific concentrations below which the harmful effects of gaseous pollutants on human health and the environment are reduced. Similarly, target values are levels fixed, below which the long-term effects of air pollutants on human health and the environment are minimized (Equivalence, 2010). Limit or target values are determined for specific periods over which the measurements are averaged depending on the pollutant. For example, the limit values for NO_2 and SO_2 are determined on an hourly basis and have values of 200 and 350 ppb, respectively, whereas the limit value for CO and the target value for O_3 are determined based on a 8 hour averaging, having respective values of 10 ppm and 125 ppb (EU-directive, 2008). We should note here that the concentrations for CO, NO_2, O_3 and SO_2 observed at the station used in this study were lower than the above-mentioned limit or target values. Considering that, the uncertainties reported in this work are calculated at the maximum concentrations reported by the reference instruments.

The EU directive uses the relative expanded uncertainty (REU) as a data quality indicator (DQI), calculated based on the guidelines provided by Walker and Schneider (2020). According to those, the calibrated LCS concentrations are assumed to have a linear relationship with the reference concentrations given as:

$$\mathrm{Cal} = \theta_0 + \theta_1 \cdot \mathrm{Ref} \tag{2.7}$$

where Cal and Ref are calibrated and reference concentrations, respectively, whereas, θ_0 and θ_1 are regression parameters which are obtained by a two-step adjusted orthogonal regression. The relative expanded uncertainty at a given concentration is then calculated at 95% confidence interval as:

$$\mathrm{REU}(\mathrm{Cal}_i) = \frac{2\sqrt{\frac{\mathrm{RSS}}{n-2} + (\lambda - (\theta_1 - 1)^2) \cdot \mathrm{U}^2_{\mathrm{Ref}_i} + (\theta_0 + (\theta_1 - 1) \cdot \mathrm{Ref}_i)^2}}{\mathrm{Ref}_i} \cdot 100\% \tag{2.8}$$

where,

$$\mathrm{RSS} = \sum_{i=1}^{n} ((\mathrm{Cal}_i - \theta_0 - \theta_1 \cdot \mathrm{Ref}_i)^2 - (\theta_1^2 + \lambda) \cdot U^2_{\mathrm{Ref}_i}) \tag{2.9}$$

$$\lambda = \frac{(\mathrm{U}_{\mathrm{Cal}_i})^2}{(\mathrm{U}_{\mathrm{Ref}_i})^2} \tag{2.10}$$

Here RSS, $\mathrm{REU}_{\mathrm{max}}$, $\mathrm{Ref}_{\mathrm{max}}$, $\mathrm{Cal}_{\mathrm{max}}$, $\mathrm{U}_{\mathrm{Cal}_{\mathrm{max}}}$ and $\mathrm{U}_{\mathrm{Ref}_{\mathrm{max}}}$ are, respectively, the residual sum of squares, the relative expanded uncertainty at the maximum concentration observed, the maximum reference concentration, the maximum calibrated concentration, the random uncertainty for maximum calibrated concentration, and the

random uncertainty for maximum reference concentration. For CO, NO_2 and SO_2, the $U_{Ref_{max}}$ values were obtained from the technical specification of each reference instrument as provided by the manufacturers (cf. Table 3 in the Appendix). For the O_3 reference measurements, for which no information about the accuracy of the reference instrument is provided by the manufacturer, I assumed $U_{Ref_{max}}$ to be 0.1%. For $U_{Cal_{max}}$, which is difficult to determine considering that its value depends on uncertainty of the reference concentrations, I have assumed $U_{Cal_{max}} = U_{Ref_{max}}$ and thus $\lambda = 1$ in my calculations following recommendation provided by Walker and Schneider (2020).

By squaring both sides of Eq. (2.8) and rearranging, two effects contributing to the $REU(Cal_i)$ are derived; the random uncertainty effect and the bias effect. These two effects are used to create target digrams to visualize the level of uncertainty.

$$REU^2(Cal_i) = RUE^2(Cal_i) + BE^2(Cal_i) \tag{2.11}$$

RUE and BE denote respectively, the random uncertainty effect and the bias effect and are defined as follows:

$$RUE(Cal_i) = \frac{2\sqrt{\frac{RSS}{n-2} + (\lambda - (\theta_1 - 1)^2) \cdot U_{Ref_i}^2}}{Cal_i} \cdot 100\% \tag{2.12}$$

$$BE(Cal_i) = \frac{2(\theta_0 + (\theta_1 - 1)Ref_i)}{Cal_i} \cdot 100\% \tag{2.13}$$

The RUE results from uncertainties in both calibrated and reference concentrations, while the BE results from deviation of regression line represented by Eq. (2.7) from the 45° line (the line passing through the origin and making an angle of 45° with the x-axis). This line deviates from the origin when $\theta_0 \neq 0$ or $\theta_1 \neq 1$ in Eq. (2.7).

2.4 Assessing calibration schemes

To assess the feasibility of monthly and seasonal calibration schemes for different air quality applications (having different accuracy thresholds), I adopted precision and bias as accuracy measures as recommended in US EPA (Williams et al., 2014). The precision and bias estimators are calculated based on guidelines given by Camalier et al. (2007). The upper bound of the coefficient of variation (CV) of the percentage differences between the calibrated and reference concentrations is used as precision error estimator. It is calculated at 90% confidence interval using in Eq. (2.14). The bias error estimator is calculated at 95% confidence interval using Eq. (2.15).

$$CV = \sqrt{\frac{n\sum_{i=1}^{n} d_i^2 - (\sum_{i=1}^{n} d_i)^2}{n(n-1)}} \cdot \sqrt{\frac{n-1}{\chi^2_{0.1,n-1}}} \tag{2.14}$$

$$|Bias| = A + t_{0.95,n-1} \cdot \frac{B}{\sqrt{n}} \tag{2.15}$$

$$d_i = \frac{Cal_i - Ref_i}{Ref_i} \cdot 100\% \tag{2.16}$$

$$A = \frac{1}{n}\sum_{i=1}^{n} |d_i| \tag{2.17}$$

$$B = \sqrt{\frac{n\sum_{i=1}^{n}|d_i|^2 - (\sum_{i=1}^{n}|d_i|)^2}{n(n-1)}} \tag{2.18}$$

Equation(2.16)-(2.18) define the variables used in Eq. (2.14) and (2.15). $\chi^2_{0.1,n-1}$ and $t_{0.95,n-1}$ are the 10th percentile value of chi-squared distribution with $n-1$ degrees of freedom and 95th percentile value of t-distribution with $n-1$ degrees of freedom, respectively.

3 Results and discussion

Table 1 provides summary statistics of the concentration measurements reported by the reference instruments and the LCSs (LAB calibrated). The differences in the mean and standard deviation between the measurements obtained by the LCSs and the reference instruments for the four pollutants range between ca. 10 and 1300 % and ca. 30 and 2600 %, respectively. To evaluate the statistical significance of these differences, I first checked normality of all LCS and reference measurements by running Shipiro-Wilk tests, which returned p-values less than 0.05, indicating that all the LCS and reference measurements were normally distributed. Next, I performed t-tests for all the pollutants to assess the statistical significance of the difference between the mean of LCS and reference concentrations (Tiku and Akkaya, 2004). To evaluate the statistical significance of the differences in variances of the LCS and reference concentration measurements, we performed the Fligner and Killeen tests (F-K tests; Si et al. 2020) for all the cases.

Table 1: Summary statistics of concentration measurements of CO, NO_2, SO_2 and O_3 recorded by the LCS, using the laboratory (LAB) calibrations, and the reference (REF) instruments.

	CO(REF)	CO(LAB)	NO_2(REF)	NO_2(LAB)	O_3(REF)	O_3(LAB)	SO_2(REF)	SO_2(LAB)
Number of data points	111025	111025	102427	102427	94520	94520	44520	44520
Average (ppb)	470	430	18	33	18	62	1.6	27
Standard deviation (ppb)	288	262	13	26	15	39	0.98	23
Minimum (ppb)	2.87	1.79	0.26	0.03	0.002	0.07	0.001	0.17
Maximum (ppb)	3573	3110	99	1365	64	1789	96	252

The t-tests showed p-values $< 10^{-5}$ for all four pollutants, which is lower than the set significance level of 0.05, suggesting that the difference in mean values of the laboratory calibrated and reference concentrations are statistically significant. Similarly, the F-K tests returned p-values $< 10^{-10}$, indicating that the differences in variances of laboratory calibrated LCSs and reference concentration measurements for all the pollutants are statistically significant considering that a significance level of 0.05 was used in the calculations.We should note here that the difference between LCS and reference measurements was more pronounced for SO_2. This can be explained by the fact that the concentration of SO_2 at the observational site was on average less than 5ppb, which was lower than the limit of detection (LoD) of the SO_2 LCS (cf. Table 2 in the Appendix).

The statistically significant difference between the measurements recorded with the LCSs and the respective

reference instruments is not surprising considering that the laboratory calibrations carried out by the manufacturer only account for the effect of the temperature at room conditions, which does not fully capture the meteorological variabilities in ambient conditions. Moreover, the calibration equations provided by the manufacturer do not take into account the influence of other factors, such as RH which can significantly affect the performance of the LCSs (Papaconstantinou et al., 2023; Samad et al., 2020; Gonzalez et al., 2019; Pang et al., 2018).

3.1 Evaluation of ML algorithms

Figure 4 shows the correlation between the calibrated and reference concentrations for the CO, NO_2, O_3 and SO_2 for all calibration approaches we tested. We should note that for all the tests we used 80 % of the entire dataset for training and 20% for validation, whereas all the measurements had a 2-min time resolution. Evidently, we observe a good agreement between the measurements from the CO LCS and the corresponding reference measurements, even without using any of the ML calibrations (cf. Fig. 4a). The rest of the LCSs, however, exhibit moderate to low correlation with their respective reference measurements (cf. Fig. 4g, 4m and 4s). The measurements reported by the SO_2 LCS highly overestimate the SO_2 concentrations as compared to those reported by the reference instrument. Overall, we observe an improvement in the agreement between measurements reported by the LCS and the respective reference instruments after ML calibration. This improvement is more pronounced for CO, NO_2 and O_3 measurements, as reflected by the high r values that are larger than 0.9, following ML calibration. In contrast, for SO_2, despite the fact that there was a general improvement in agreement between the SO_2 LCS and reference measurements when using ML calibration, they exhibited r values that were hardly above 0.5 (cf. Fig. 4s-x). As mentioned above, this can be attributed to the low concentrations of SO_2 observed at the measuring site, which on average were far lower (i.e., 1.6 $\pm$ 1 ppb; cf. Table 1) than the LoD of the SO_2 LCS (i.e., 5 ppb; cf. Table A1 in the Appendix and Alphasense-SO2-B4 2019).

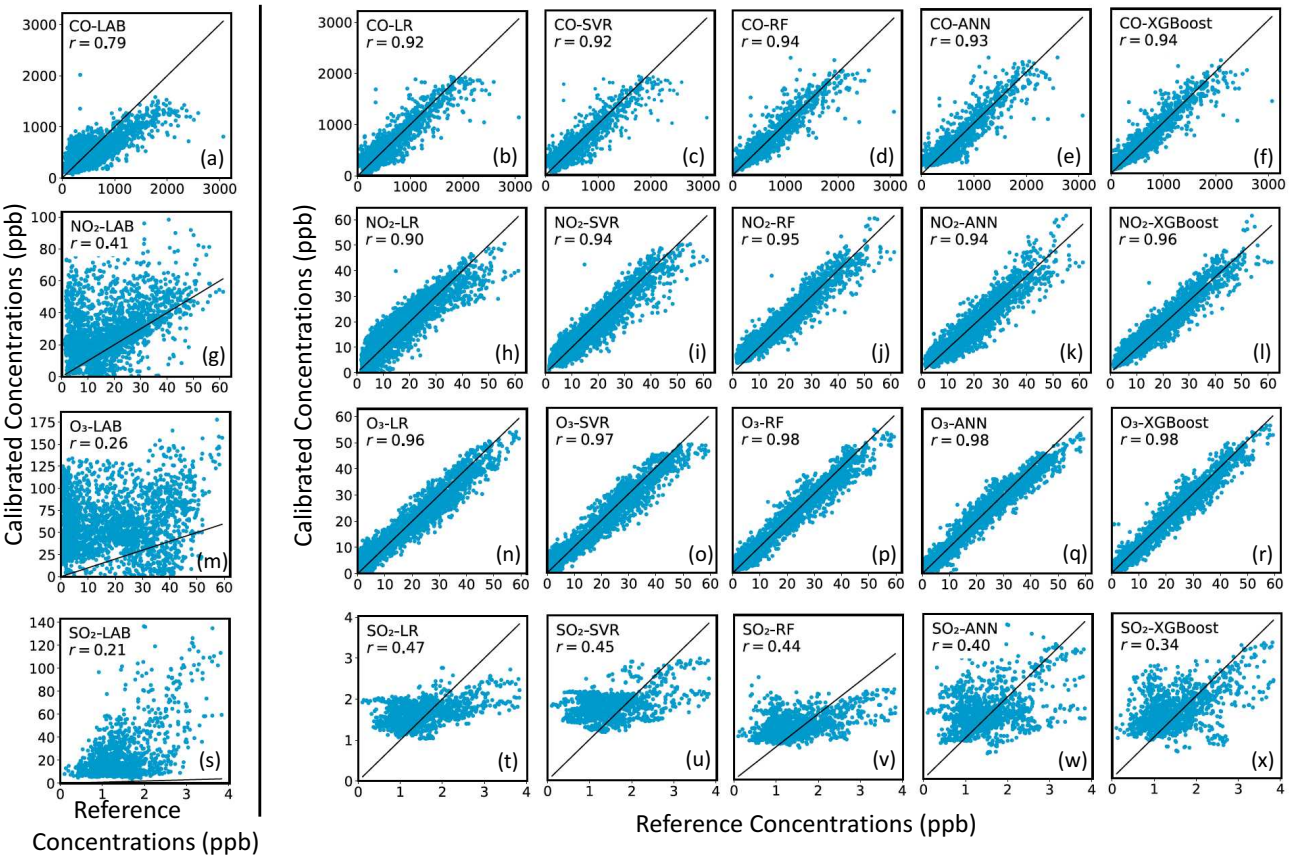

Figure 4: Correlation between calibrated LCS and reference concentration measurements for CO, NO_2, O_3 and SO_2. The black dotted correspond to 1:1 relation.

Figure 5 provides heat maps that show the performance of LAB and ML calibrations based on R^2 and NRMSE. With the exception of the SO_2 LCS, all the ML calibrations performed reasonably well in terms of R^2 as indicated in Figure 5a. Overall, the RF model exhibited the highest R^2 values for all pollutants, followed by the ANN and the XGBoost models. The good performance of the ML calibrations was also reflected by the relatively low NRMSE values (cf. Fig. 5b), which exhibit a significant improvement compared to the values of their corresponding LAB calibration. Another key point to note here is that, even though the CO LCS exhibited a good performance without ML calibration, it is evident that performance of ML calibrations for NO_2 and O_3 measurements are generally better than that of the CO. This is explained by the fact that, for the NO_2 and O_3 ML calibrations we accounted for cross-sensitivities by including O_3 and NO_2 reference concentrations, respectively, as in input variables.

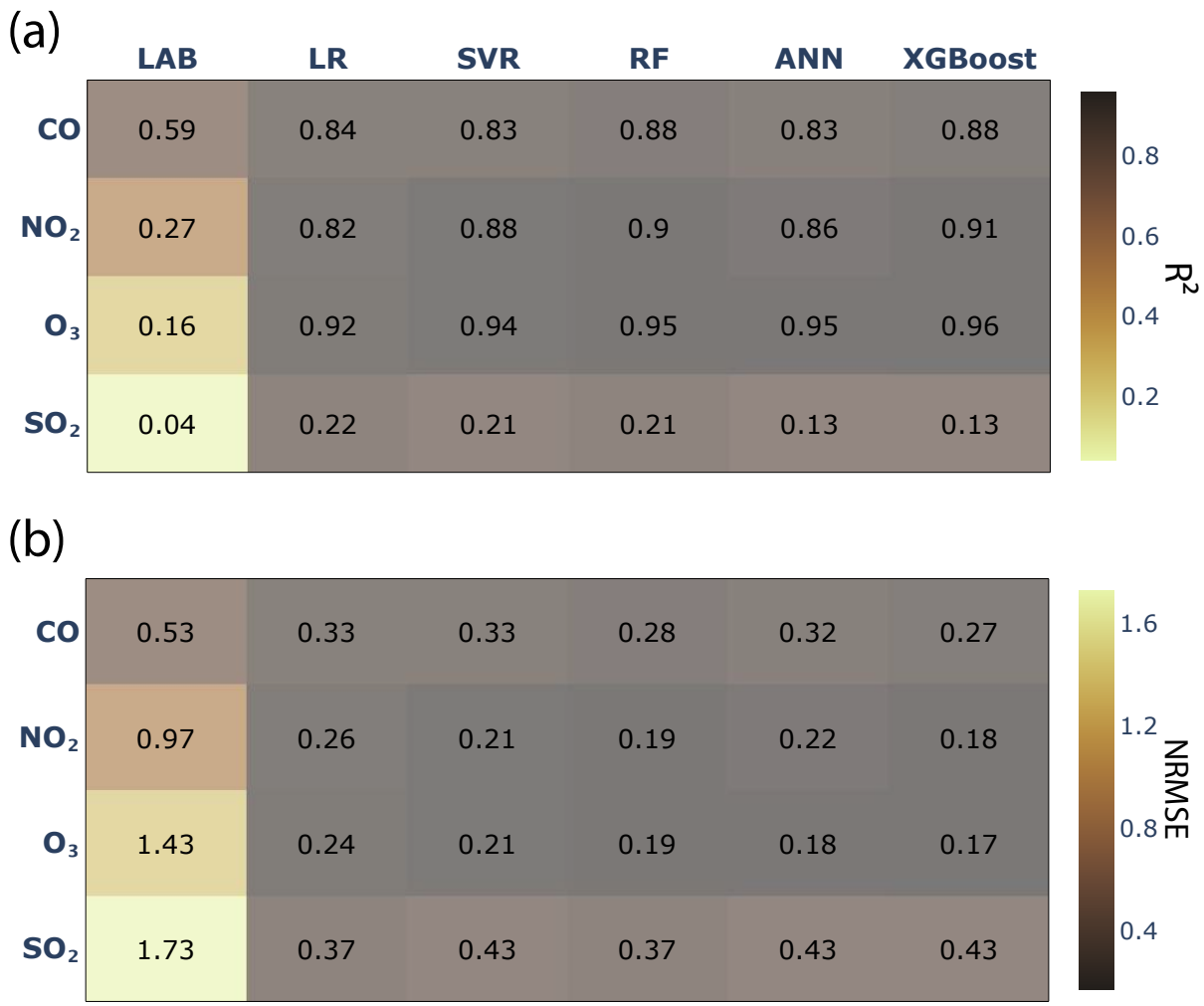

Figure 5: Heat maps showing the performance of the LCSs in terms of (a) coefficient of determination (R^2) and (b) normalized root mean squared error (NRMSE) after calibration using the laboratory tests carried out by the manufacturer (LAB) and the five ML algorithms employed in this work.

Figure 6 shows target diagrams that help to further assess the performance of the ML models in terms of the bias and variance, and to compare the standard deviations of their predictions with those of their respective reference concentrations. In each diagram the normalised MBE is plotted against the normalized CRMSE. The MBE and CRMSE were normalised by the standard deviation of their corresponding reference concentrations to facilitate comparison of the calibration performance of the models across different pollutants. Consequently, the distance of each point from the centre of the diagrams corresponds to the RMSE normalized by the standard deviation of the respective reference concentrations. We will denote this with nRMSE to distinguish it from the one described in Eq. (2.3). As shown in the diagrams, the ML calibrations generally exhibited lower nRMSE values compared to corresponding laboratory calibrations. The RF, ANN and XGBoost models generally performed better in terms of nRMSE when compared to the LR and SVR models.

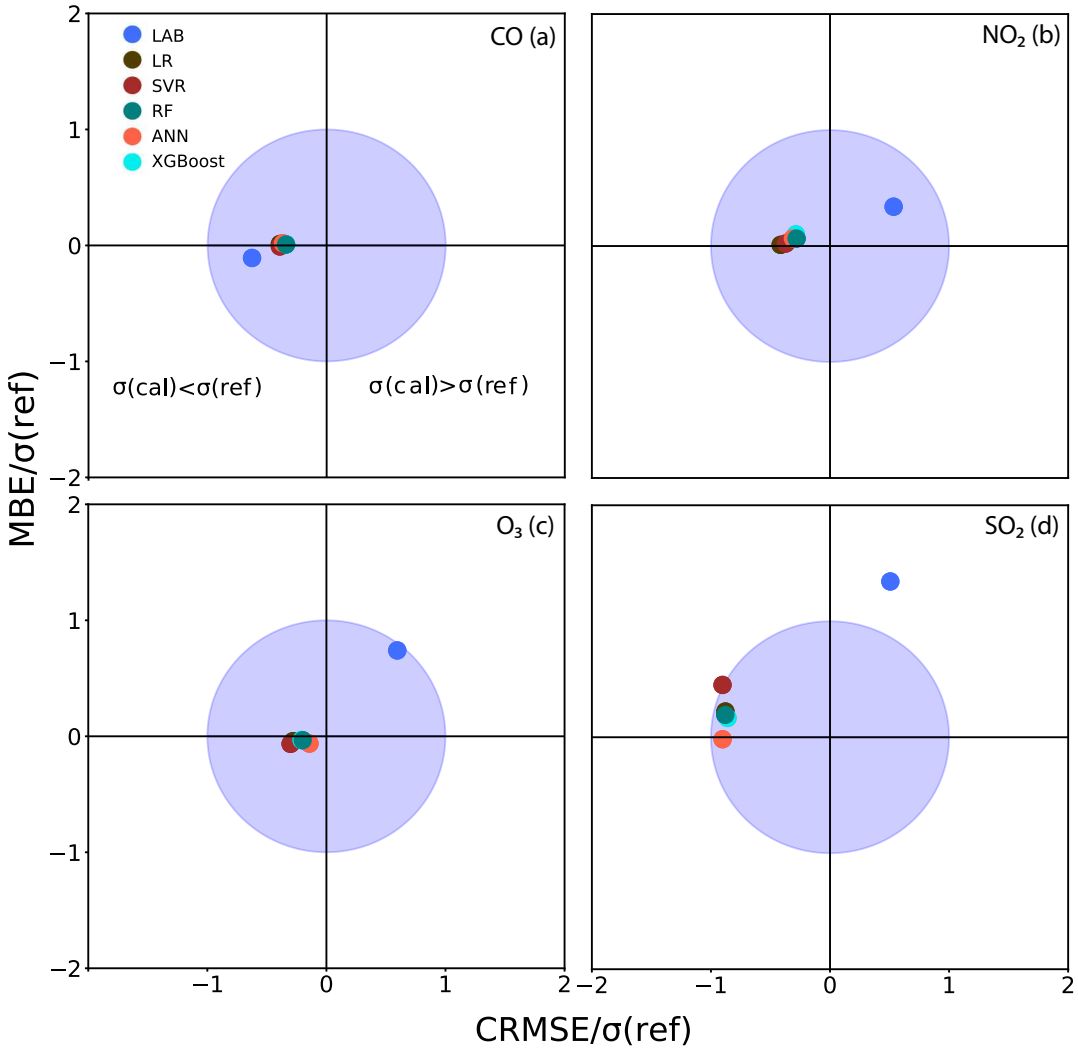

Figure 6: Target diagrams showing the overall performance of the laboratory and ML learning calibrations of the data from the LCS tested in this work. The light blue circles show the region where the NRMSE is less than unity and the variance of residuals of the calibrated concentrations are lower than those of their corresponding reference concentrations.

The horizontal lines passing through the centers of the circles in Figure 6 correspond to cases with zero bias. Any point above or below these lines indicate that the corresponding algorithm overestimates or underestimates the values reported by the reference instruments, respectively. The CO LAB calibration exhibit a small bias and thus corroborating the good correlation between those and the respective reference measurements as shown in Figure 4. This is not the case for the other sensors. As indicated in Figure 6b-d, the LAB calibrations for the measurements with the NO_2, O_3 and SO_2 LCSs overestimate the respective reference measurements. With the exception of the SO_2 measurements, the calibrations by the ML models have a bias close to zero, indicating a significant improvement in the agreement between the LCS and reference concentration measurements.

When the standard deviation of the calibrated LCS data is lower than that of their respective reference measurements, the point associated with the model used for calibration would be illustrated on the left quadrant of the respective target diagram and vice versa. Therefore, the standard deviations of LAB calibrated measurements

from all the LCSs, except the one for CO, are higher than those of their respective reference concentrations, whereas ML calibrated concentrations generally have standard deviations which are lower than those of the corresponding reference concentrations. All the points corresponding to ML models generally lie inside the unit circles, indicating that the variance of residuals of their calibrated concentrations are lower than those of the respective reference concentrations and thus that they do not over-fit on the training data.

3.2 Data quality assessments

According to the EU directive ,the data from LCSs is primarily used as indicative measurements (i.e measurements with less stringent uncertainty requirements) and in this case, the directive sets out the uncertainty limits at 25% for CO, NO_2 and SO_2, and 30% for O_3. Essentially, these uncertainty requirements have to be met at limit values for CO, NO_2 and SO_2, and at the target value for O_3, in order to achieve compliance with the directive's DQOs. The limit values are specific concentrations determined based on scientific knowledge, with the aim of preventing harmful effects on human health and the environment. On the other hand, target values are levels fixed in order to avoid long-term effects on human heath and the environment (Equivalence, 2010). The limit or target values are for specific averaging periods for each pollutant. Based on hourly averaging, the limit values for NO_2 and SO_2 are 200 ppb and 350 ppb, respectively. The limit value for CO and the target value for O_3, which are based on 8 hour averaging period are 10 ppm and 125 ppb, respectively (EU-directive, 2008).

To assess compliance with EU directive DQOs, the two effects RR (Eq. (2.12)) and RB (Eq. (2.13)) that account for variation in REUs were calculated for the calibrated concentrations and plotted in target diagrams as shown in Figure 7. Note that from the color-bar legends, the maximum concentrations observed were lower than the limit or target values set by EU directive for all the pollutants. Hence, our analysis focus on assessing whether the calibrated concentrations can still meet the EU DQOs even at concentrations lower than the limit values or target values. The uncertainties for all the LAB calibrations were higher than 75% and therefore are not included in the target diagrams in Figure 7.

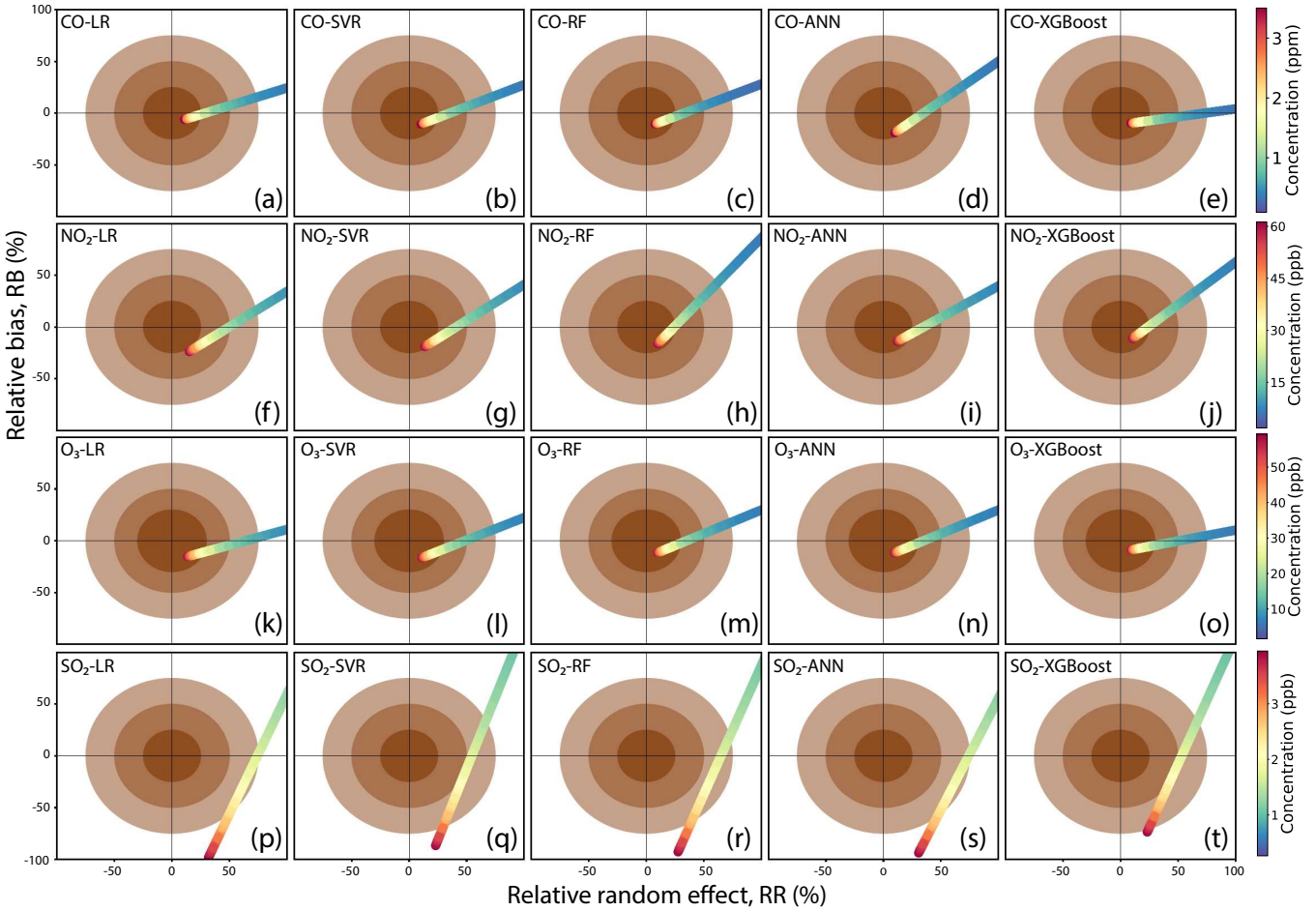

Figure 7: Target diagrams to assess the level of uncertainty of ML calibrated concentrations. The distance from each point on the coloured lines to the centre of the corresponding diagram represent its relative expanded uncertainty. The coloured circles in each diagrams represent the uncertainty targets. The inner circles are the 25% uncertainty targets for CO, NO_2 and SO_2 and 30% uncertainty targets O_3. The middle and outer circles in all the diagrams are 50% and 75% uncertainty targets, respectively. The large green point at the end of each line is the scatter plot RR and RB corresponding to the maximum concentration.

For CO, NO_2 and O_3, the DQO criteria (i.e., uncertainty below 25% for CO and NO_2, and 30% for O_3) are met by the RF, ANN and XGBoost calibrated data at concentrations above approximately 1.5 ppm for CO, 45 ppb for NO_2 and 30 ppb for O_3, which are significantly lower compared to the limit or target values in the current EU directive (i.e., 10 ppm for CO, 200 ppb for NO_2 and 125 ppb for O_3). On the other hand, although LR and SVR calibrations improved the data quality of the CO measurements (cf. Figure 7a-b), the calibrations by these models did not generally meet EU DQOs for the measurements from the other LCSs.

For SO_2, although the quality of the data improved after ML calibration, the improvement generally did not attain the EU directive DQOs at any concentration level (cf. Fig. 7p-t). This was expected considering that all the statistical indicators shown in Table 5 were not improved substantially for SO_2 even after using ML models for calibration. This highlights the fact that that in areas with SO_2 concentrations lower than limit of detection, both the laboratory and field calibration of SO_2 LCS data might hardly attain the EU DQOs.

According to US EPA, the DQOs that the LCS data should meet for use in regulatory purposes are more stringent compared to that of non-regulatory applications. The agency set out 10, 25, 30 and 50% error thresholds as the DQOs that the LCS data must meet for use in assessing respectively regulatory compliance, spatial gradient studies, intervention studies and hotspot determination as well as citizen science projects (Schäfer et al., 2021). To assess whether the quality of calibrated LCS data meets these DQOs, the precision and bias errors for LAB and ML calibrations were calculated and represented in spider plots shown in Figure 8.

The level of precision and bias errors is generally higher for LAB calibrations compared to ML calibrations. Nevertheless, LAB calibration for the CO LCS (cf. Figure 8a) appears to have precision and bias errors falling in the range 30-50%, suggesting that even without any further re-calibration, the measurements from CO LCS can still be used for hotspot identification and carrying out citizen science projects. For the NO_2, O_3 and SO_2 LCS (cf. Figure 8b-d) however, their measurements corresponding to LAB calibrations have bias errors exceeding 50% and therefore cannot be used to serve any of the purposes mentioned in the previous paragraph unless they are further recalibrated. As shown in Figure 8a-c, the level of precision and bias errors for ML calibrations is generally lower than 30%, with the calibrations using RF, ANN and XGBoost algorithms exhibiting even a more lower precision and bias errors (i.e lower than 25%). This is an indication that whenever ML models are used to calibrate the measurements from CO, NO_2 and O_3 LCSs, the quality of their data greatly improves to a level that can be used to serve non-regulatory LCS applications ranging from spatial gradient studies to hotspot analysis as discuss above.

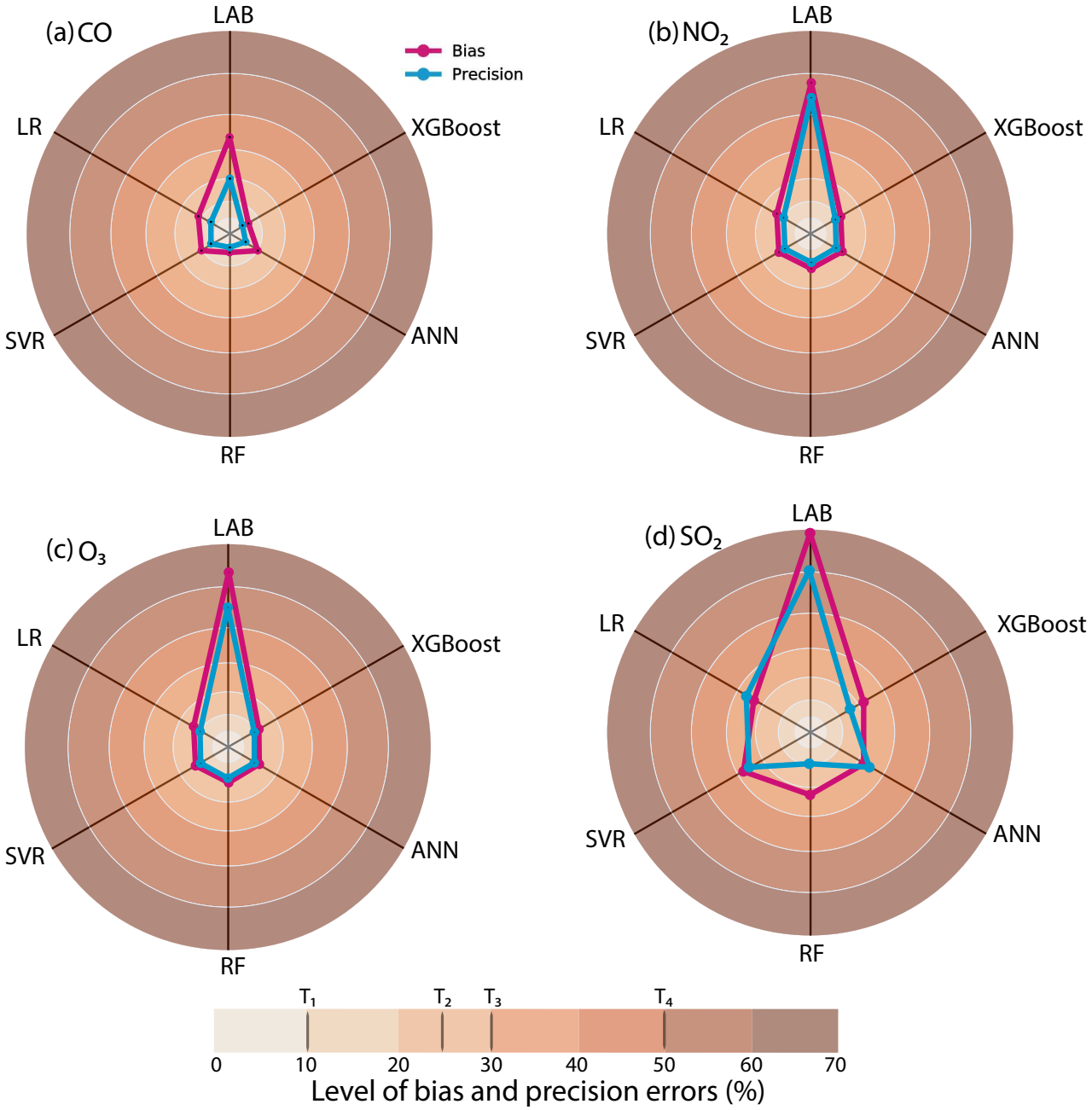

Figure 8: Spider plots showing the level of precision and bias errors for both LAB and ML calibrated CO, NO_2, O_3 and SO_2 LCS data. T_1 to T_4 in the color-bar are thresholds on the level of precision and bias errors set out by US EPA for different LCS applications. T_1 is the threshold for regulatory compliance, T_2 the threshold for spatial gradient studies, T_3 the threshold for Intervation studies and T_4 is the threshold for hotspot determination & citizen science projects.

For the SO_2 LCS, the quality of its measurements improved only slightly calibration by ML algorithms. As shown in Figure 8d, all the ML calibration attained precision and bias errors between 30 and 45%, making it appropriate only for determining SO_2 hotspots and in carrying out citizen science projects.

So far we have seen that ML calibrations have precision and bias errors meeting the thresholds for non-regulatory LCS applications. However, the 10% threshold for regulatory compliance is hardly met by both LAB and ML

calibrations in all the LCSs. This highlight the fact that the data from LCSs, both calibrated and non-calibrated can only be used as indicative measurements to serve non-regulatory air quality assessments.

3.3 Feature importance

To quantify the influence of a number of features that affect the performance of the CO, NO_2, O_3 and SO_2 LCSs, we carried out sensitivity analysis using the RF model, which exhibited the best performance as discussed in the previous sections. To do that, the model is trained with the entire 6-month dataset by using reference concentrations, temperature, relative humidity, month, day of week and hour as input/independent variables, and the NSS as dependent variables. As stated earlier, the NO_2 LCS is highly cross-sensitive to O_3 and vice versa (ANN803-05, 2019). To assess the effect of cross-sensitivity on model performance, O_3 and NO_2 reference concentrations were included as additional input variables in the training of the RF model for the NO_2 and the O_3 LCSs, respectively. The influence of a given input variable to model performance was then calculated by permuting its values and determining the change in the performance. For each variable, 20 random permutations were carried out and the average change in R^2 was calculated and expressed as a percentage of the total contribution from all the variables.

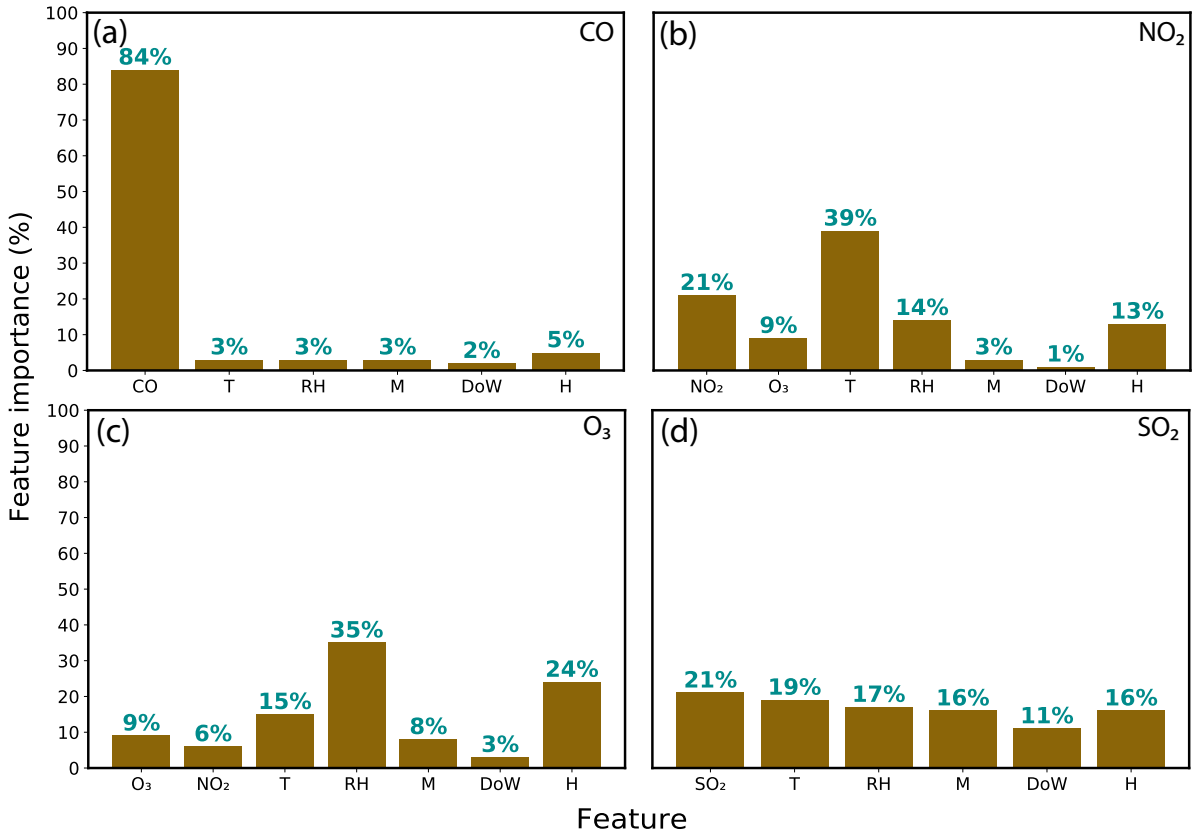

Figure 9: Estimated importance of each input feature in determining the variability of the signals reported by CO, NO_2, O_3 and SO_2 LCSs, determined for each sensor using the RF model. For CO and SO_2 the model is trained using respective reference concentrations of the target gases, temperature, relative humidity, month, day of week and hour as input variables, and the NSS as dependent variable. For NO_2 and O_3, the training of the RF algorithm also includes respectively, the O_3 and NO_2 reference concentrations as input variables in order to assess cross-sensitivity.

For the CO (cf. Fig. 9a), the target gas (i.e., the CO reference concentration) has the highest degree of importance in the performance of the model relative to the importance of other variables. This implies that the CO sensor is not affected much by the other variables used as input to the model, and thus responds well to the fluctuations of the CO concentration in ambient air. This is not the case for NO_2 and O_3. For these two cases, the influence of target gases on the performance of the model seems to be low, with temperature and RH having strong influence. By accounting for cross-sensitivity, the performance of the models further improved by 6-9 % (cf. Fig. 9c and 9d). For the SO_2, the performance of the model is almost equally affected by the target gas, temperature and RH as well as some of the temporal parameters (cf. Fig. 9d). This is expected considering that the SO_2 LoD is higher compared to what is observed.

3.4 Assessing calibration practices

For the evaluation of the models described in subsection 3.1, we have used 80 % of the dataset for training the models. While it is a good practice in ML to use higher fraction of data (e.g., >70 %) for training, it is not

practical for calibrating the data from the LCSs as that will require collocating LCSs with corresponding reference instruments for long periods of time, adding significantly to the cost of the measurements. Using RF model, here we investigate how the amount of training data can be minimised, while not significantly sacrificing their quality, by varying practical parameters such as (1) the temporal resolution of the training data, (2) how the training is sampled, and (3) the frequency of calibration.

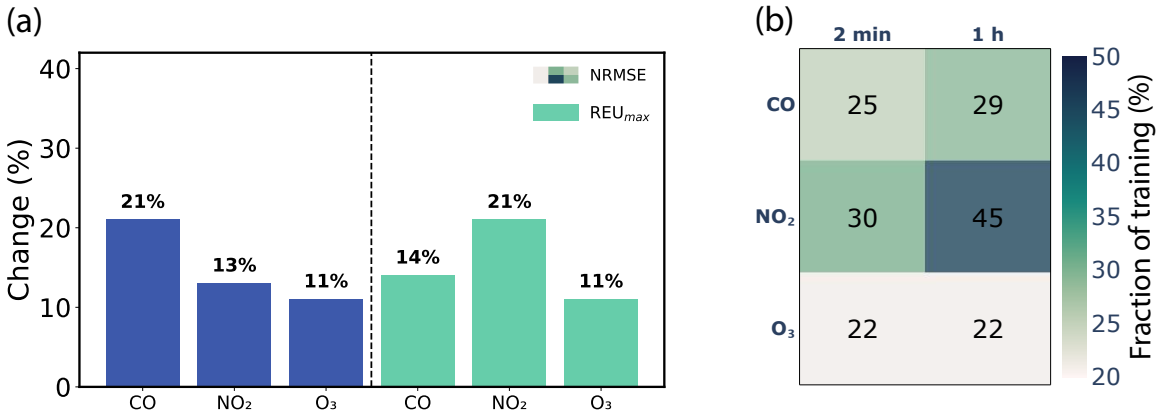

Figure 10: (a) Percentage change in the normalized root mean squared error (NRMSE) and the average relative expanded uncertainty calculated at maximum concentration observed (REU_{max}) when the resolution of the data used to train the RF model is increase from 1 h to 2 min, and (b) the minimum fraction for 2 min and 1 h measurements required to train the RF model that will guarantee its calibrations to meet the data quality objective for indicative measurements defined by the EU directive.

Figure 10 show the percentage change in (1) the normalized root mean squared error (ΔNRMSE), and (2) the relative expanded uncertainty calculated at maximum concentration observed (ΔREU_{max}), when the resolution of the data used to train the RF models is increase from 1 h to 2 min. Calculations of ΔNRMSE and ΔREU_{max} were done using (A.1) and (A.2) shown in the Appendix. The results shows that by increasing the temporal resolution of the training data from 1 h to 2 min (which is the base case here), the performance of the RF model in terms of NRMSE and REU_{max} improves by 11-21 %. As a results, the minimum fraction of data required for training the RF models that will qualify their calibrations for indicative measurements as defined by EU directive reduces by an average of 6.3 % (cf Fig. 10b)

To investigate the effect of how the training data is sampled and calibration frequency on the amount of data required for training, we performed calibrations using different fractions of the dataset every 1, 3 and 6 months. For 1 month calibration, the training data is obtained through continuously sampling, i.e, training data sampled continuously at the start of each month. For 3 and 6 month calibrations, in addition to continuous data sampling, we also considered interceptive data sampling, where the training datasets were sampled from all the months. Note that for the later case, the post-calibration of the LCSs by applying the trained ML algorithm and the evaluation of the algorithm are conducted retrospectively at the end each period (i.e., every 3 or 6 months for the 3- and 6- month periods, respectively).

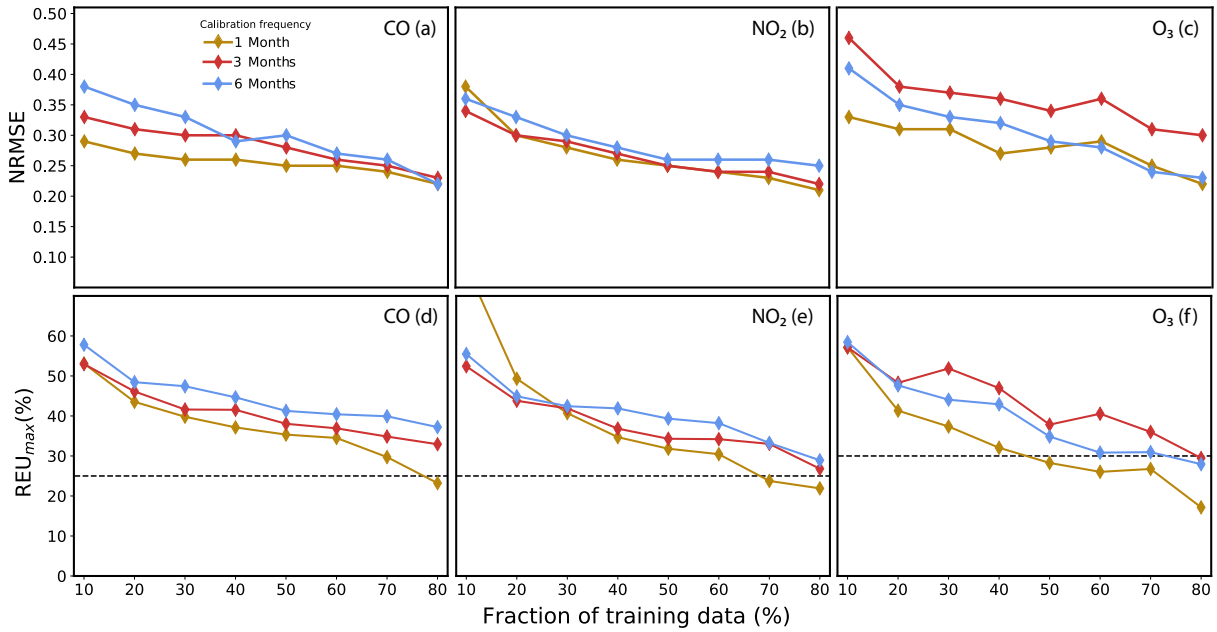

Figure 11: Average normalized root mean squared error (NRMSE; a-c), and the average relative expanded uncertainty calculated at maximum concentration observed (REU_{max}: d-f) for 1, 3 and 6 calibrations obtained by training RF following continuous data sampling (i.e., continuously at the beginning of each period). The dotted lines in the bottom show the EU DQOs for indicative measurements, which are set to 25 % for CO and NO_2, and 30 % for O_3.

Figure 11 shows the average NRMSE and REU_{max} for 1, 3 and 6 months calibrations for CO, NO_2 and O_3 obtained through continuous data sampling and varying the training data from 10 to 80 %. The NRMSE and the REU_{max} decreases with increase in the fraction of training data. The results also shows that under the continuous data sampling scheme, the monthly calibration performing better in terms of NRMSE and REU_{max} values compared to calibration after 3- or 6-month. This is explained by the fact that the RF algorithm lacks the power to extrapolate on training data, and thus will perform better in monthly calibration where the seasonal variabilities of the training and testing datasets are lower compared to calibration after 3 or 6 month. It is also evident that under the continuous data sampling, a larger fraction of data (at least 70 % for CO and NO_2, and 50 % for O_3) is needed to train the models for the quality of calibrated data to meet the EU directive DQOs for indicative measurements (cf. Fig. 11d-f).

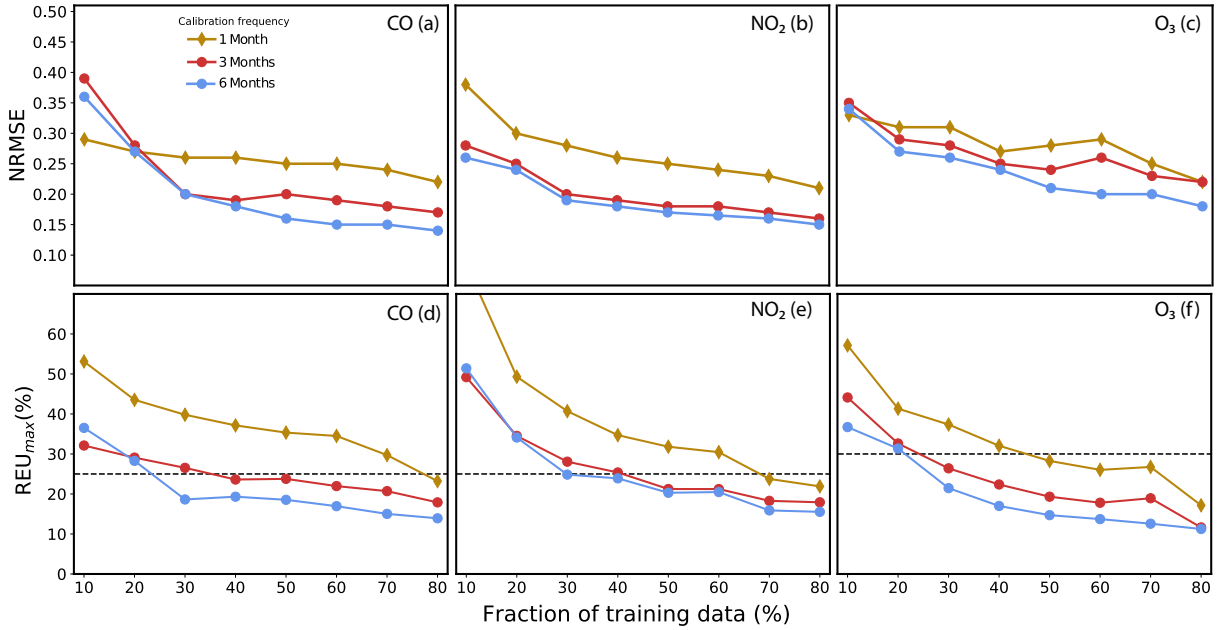

Figure 12: Average normalized root mean squared error (NRMSE: a-c), and the average relative expanded uncertainty calculated at maximum concentration observed (REU_{max}: d-f) for 1, 3 and 6 month calibration periods obtained by training the RF following interceptive data sampling for 3 and 6 month calibrations and comparing with monthly calibration carried out through continuous data sampling. The dotted lines in the bottom show the EU DQOs for indicative measurements, which are set to 25 % for CO and NO_2, and 30 % for O_3.

Figure 12 shows the average NRMSE and REU_{max} for CO, NO_2 and O_3 obtained by performing 3 and 6 months through interceptive sampling. Note that the results of monthly calibrations are added here for comparison. As shown in the figure, performing calibration after 3 or 6 months using interceptive data sampling scheme requires less training data for achieving a data quality meeting the requirements of EU directive DQOs. For instance, performing 6 month calibration under this schemes will reduce the fraction of training data to as low as 22 % without deteriorating the quality of the data to levels below those that will qualify them for indicative measurements. This reduces substantially the cost of these measurements, as significantly less amount of time (from 70-80 % down to 22 %) is needed for collocated measurements with expensive reference grade instruments.

4 Conclusion

I have evaluated the performance of a number of ML algorithms for calibrating data from CO, NO_2, O_3 and SO_2 LCSs, and identified that the Random Forest model provides the best results. This model is then used (1) to investigate how different input variables affect the calibration performance, and (2) to investigate the extend to which practical parameters such as the temporal resolution of the measurements, how the training data is sampled, and the frequency of calibration reduces the fraction of data needed for training, while keeping the quality of calibrated data within the accepted levels. My results show that the Alphasense CO LCS respond well

to the target gas and is not affected much by the other variables such as temperature and RH. This is not the case for NO_2 and O_3 Alphasense LCSs. For these two cases, the influence of target gases on the performance of the model seems to be low, with temperature and RH having strong influence on their performance. By increasing the temporal resolution of the training data from 1 h to 2 min, the minimum fraction of data required for training the RF models that will qualify their calibrations for indicative measurements as defined by EU directive reduces by an average of 6.3 %. The results also suggest that the minimum fraction of data required for training the ML models depends on the frequency of carrying out collocated measurements with reference instruments and using the resulting datasets for training the calibration model. If the calibrations are carried out on a monthly basis, ca. 50 % of the period is needed for collecting data to train the RF algorithm and qualify the LCSs for indicative measurements as defined by the EU directive (2008/50/EC). If the training is carried out every 3 or 6 months by following continuous data sampling, then ca. 60 % of the measuring period is required for collecting training data. In those cases, if the sampling of the training data is made interpretively over specific periods every month, but the entire training dataset is used to calibrate the measurements over 3 or 6 months, then the amount of data required for qualifying the LCSs for indicative measurements can significantly reduce down to 22 %. This reduces substantially the cost of these measurements, as the amount of time needed for collocated measurements with expensive reference grade instruments significantly reduce from 70-80 % down to 22 %). This contributes not only in improving the data quality of measurements obtained by LCS networks but achieves it in the most cost-effective manner, maintaining the cost-efficient orientation/goal of employing LCS networks.

References

Alphasense-SO2-B4. So2-b4 sulfur dioxide sensor technical specification 4-electrode. https://www.alphasense.com/wp-content/uploads/2019/09/SO2-B4.pdf, 2019.

ANN803-05. Correcting for background currents in four electrode toxic gas sensors (2019). alphasense application note aan 803-05. 2019.

P. Arroyo, J. Gómez-Suárez, J. I. Suárez, and J. Lozano. Low-cost air quality measurement system based on electrochemical and pm sensors with cloud connection. *Sensors*, 21(18):6228, 2021.

J. Baranwal, B. Barse, G. Gatto, G. Broncova, and A. Kumar. Electrochemical sensors and their applications: A review. *Chemosensors*, 10(9):363, 2022.

L. Camalier, S. Eberly, J. Miller, and M. Papp. Guideline on the meaning and use of precision and bias data required by 40 cfr part 58, appendix a. *EPA-454/B-07-001*, 2007.

C.-C. Chen, C.-T. Kuo, S.-Y. Chen, C.-H. Lin, J.-J. Chue, Y.-J. Hsieh, C.-W. Cheng, C.-M. Wu, and C.-M. Huang. Calibration of low-cost particle sensors by using machine-learning method. In *2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, pages 111–114. IEEE, 2018.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Equivalence. Guide to the demonstration of equivalence of ambient air monitoring methods. 2010.

EU-directive. Directive 2008/50/ec of the european parliament and of the council of 21 may 2008 on ambient air quality and cleaner air for europe. *Official Journal of the European Union*, 2008.

P. Ferrer-Cid, J. M. Barcelo-Ordinas, J. Garcia-Vidal, A. Ripoll, and M. Viana. Multisensor data fusion calibration in iot air pollution platforms. *IEEE Internet of Things Journal*, 7(4):3124–3132, 2020.

A. Gonzalez, A. Boies, J. Swason, and D. Kittelson. Field calibration of low-cost air pollution sensors. *Atmospheric Measurement Techniques Discussions*, pages 1–17, 2019.

N. Isaac, I. Pikaar, and G. Biskos. Metal oxide semiconducting nanomaterials for air quality gas sensors: operating principles, performance, and synthesis techniques. *Microchimica Acta*, 189(5):196, 2022.

J. K. Jolliff, J. C. Kindle, I. Shulman, B. Penta, M. A. Friedrichs, R. Helber, and R. A. Arnone. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. *Journal of Marine Systems*, 76(1-2): 64–82, 2009.

P. Kumar, L. Morawska, C. Martani, G. Biskos, M. Neophytou, S. Di Sabatino, M. Bell, L. Norford, and R. Britter. The rise of low-cost sensing for managing air pollution in cities. *Environment international*, 75:199–205, 2015.

V. Kumar and M. Sahu. Evaluation of nine machine learning regression algorithms for calibration of low-cost pm2. 5 sensor. *Journal of Aerosol Science*, 157:105809, 2021.

A. Lewis, W. R. Peltier, and E. von Schneidemesser. Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications. 2018.

B. Maag, O. Saukh, D. Hasenfratz, and L. Thiele. Pre-deployment testing, augmentation and calibration of cross-sensitive sensors. In *EWSN*, pages 169–180, 2016.

A. Masic, D. Bibic, B. Pikula, and F. Razic. New approach of measuring toxic gases concentrations: Principle of operation. *Annals of DAAAM & Proceedings*, 29, 2018.

P. Nowack, L. Konstantinovskiy, H. Gardiner, and J. Cant. Machine learning calibration of low-cost no 2 and pm 10 sensors: non-linear algorithms and their impact on site transferability. *Atmospheric Measurement Techniques*, 14(8):5637–5655, 2021.

N. U. Okafor and D. T. Delaney. Application of machine learning techniques for the calibration of low-cost iot sensors in environmental monitoring networks. In *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, pages 1–3. IEEE, 2020.

X. Pang, M. D. Shaw, A. C. Lewis, L. J. Carpenter, and T. Batchellier. Electrochemical ozone sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air-quality monitoring. *Sensors and Actuators B: Chemical*, 240:829–837, 2017.

X. Pang, M. D. Shaw, S. Gillot, and A. C. Lewis. The impacts of water vapour and co-pollutants on the performance of electrochemical gas sensors used for air quality monitoring. *Sensors and Actuators B: Chemical*, 266:674–684, 2018.

R. Papaconstantinou, M. Demosthenous, S. Bezantakos, N. Hadjigeorgiou, M. Costi, M. Stylianou, E. Symeou, C. Savvides, and G. Biskos. Field evaluation of low-cost electrochemical air quality gas sensors under extreme temperature and relative humidity conditions. *Atmospheric Measurement Techniques*, 16(12):3313–3329, 2023.

S. S. Patra, R. Ramsisaria, R. Du, T. Wu, and B. E. Boor. A machine learning field calibration method for improving the performance of low-cost particle sensors. *Building and Environment*, 190:107457, 2021.

A. Samad, D. R. Obando Nuñez, G. C. Solis Castillo, B. Laquai, and U. Vogt. Effect of relative humidity and air temperature on the results obtained from low-cost gas sensors for ambient air quality measurements. *Sensors*, 20(18):5175, 2020.

K. Schäfer, K. Lande, H. Grimm, G. Jenniskens, R. Gijsbers, V. Ziegler, M. Hank, and M. Budde. High-resolution assessment of air quality in urban areas?a business model perspective. *Atmosphere*, 12(5):595, 2021.

M. Si, X. Ying, S. Du, and K. Du. Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine learning technologies 2. 2020.

J. Song, K. Han, and M. E. Stettler. Deep-maps: Machine-learning-based mobile air pollution sensing. *IEEE Internet of Things Journal*, 8(9):7649–7660, 2020.

L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. Bonavitacola. Field calibration of a cluster of low-cost available sensors for air quality monitoring. part a: Ozone and nitrogen dioxide. *Sensors and Actuators B: Chemical*, 215:249–257, 2015.

P. Thunis, A. Pederzoli, and D. Pernigotti. Performance criteria to evaluate air quality modeling applications. *Atmospheric Environment*, 59:476–482, 2012.

M. L. Tiku and A. D. Akkaya. *Robust estimation and hypothesis testing*. New Age International, 2004.

I. Vajs, D. Drajic, N. Gligoric, I. Radovanovic, and I. Popovic. Developing relative humidity and temperature corrections for low-cost sensors using machine learning. *Sensors*, 21(10):3338, 2021.

S.-E. Walker and P. Schneider. A study of the relative expanded uncertainty formula for comparing low-cost sensor and reference measurements. *NILU rapport*, 2020.

C. Wang, Q. Wu, M. Weimer, and E. Zhu. Flaml: A fast and lightweight automl library. *Proceedings of Machine Learning and Systems*, 3:434–447, 2021.

R. Williams, V. Kilaru, E. Snyder, A. Kaufman, T. Dye, A. Rutter, A. Russell, and H. Hafner. Air sensor guidebook. *US Environmental Protection Agency*, 2014.

L. Yu, K. K. Lai, S. Wang, and W. Huang. A bias-variance-complexity trade-off framework for complex system modeling. In *International Conference on Computational Science and Its Applications*, pages 518–527. Springer, 2006.

N. Zimmerman. Tutorial: Guidelines for implementing low-cost sensor networks for aerosol monitoring. *Journal of Aerosol Science*, 159:105872, 2022.

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

C. Zuidema, C. S. Schumacher, E. Austin, G. Carvlin, T. V. Larson, E. W. Spalt, M. Zusman, A. J. Gassett, E. Seto, J. D. Kaufman, et al. Deployment, calibration, and cross-validation of low-cost electrochemical sensors for carbon monoxide, nitrogen oxides, and ozone for an epidemiological study. *Sensors*, 21(12):4214, 2021.

Table 2: Low cost sensor specifications. The superscript [1] stands for the lowest concentration difference that can be distinguished by the sensor while [2] stands for the lowest detectable reading that can be reliably measured and displayed. It is an absolute quantity of the gas that can be detected. ISB stands for individual sensor board

	CO-B41	**NO_2-B43F**	**OX-B431**	**SO_2-B4**
Detection range (ppm)	1000	20	20	100
Temperature range (°C)	-30 to 50	-30 to 40	-30 to 40	-30 to 50
Humidity range (%RH)	15 to 90	15 to 85	15 to 85	15 to 19
Response time (s)	$<$ 25 (From 0 to 10 ppm)	$<$60 (From 0 to 2 ppm)	$<$ 60 (From 0 to 1ppm)	$<$ 60 (From 0 to 2ppm)
Resolution[1] $\pm$ 2 standard deviation (ppb)	4	15	15	5
Major cross-sensitivity gas(es)	H_2S	O_3, Cl_2	NO, NO_2, Cl_2	NO, CO, H_2S
Sensitivity[2] from 20°C to 50°C / zero current (nA)	-50 to -200	-10 to 50	5 to 100	10 to 30
Operation until the loss of 50% of the original signal (months)	$>$36 (24 warranted)	$>$24 (24 warranted)	$>$24 (24 warranted)	$>$36 (24 warranted)
Dimensions (mm)	32 x 16.5	32 x 16.5	32 x 16.5	32 x 16.5
Weight (g)	$<$13	$<$13	$<$13	$<$13
Cost with ISB (US $)	153	158	164	153

A

Table 3: Specifications of reference instruments used by Cyprus Department of Labour and Inspection for air quality monitoring

	CO	**NO_2**	**O_3**	**SO_2**
Model	Ecotech Serinus 30	Ecotech Serinus 40	Thermo Scientific 49i	Ecotech Serinus 50
ISO_EN Standard	CYS EN	CYS EN	CYS EN	CYS EN
Number	14626:2012	14211:2012	14625:2012	14212:2012
Method	Infrared Spectroscopy	Chemiluminescence	Ultraviolet Photometry	Ultraviolet Fluorescence
Sample flow rate	1.0 slpm	0.3 slpm (0.6 slpm total flow for the NO and NOX flow path)	1–3 lpm	0.750 slpm
Precision	20 ppb or 0.1 % of the reading, whichever is greater	0.4 ppb or 0.5% of the reading, whichever is greater		0.5 ppb or 0.15 % of reading, whichever
Response time	60 seconds to 95%	15 seconds to 90%	20 seconds (10 seconds lag time)	60 seconds to 95 %
Lower detectable limit (ppb)	40	0.4	1	0.3

Table 4: Optimal values of the main hyper-parameters for SVR, RF, ANN and XGBoost ML algorithms

SVR-CO	**RF − CO**	**ANN − CO**	**XGBoost − CO**
kernel='rbf'	n_estimators=236	number of hidden layers=3	n_estimators=970
C=1	max_leaf_nodes=744	number of neurons in the 1st and 2nd hidden layers=128	max_depth=0
ϵ= 1	max_features=0.51	number of neurons in the 3rd hidden layer=100	eta=0.092
		activation function in the input and output layers='linear'	subsample=0.76
		activation function in the in the hidden layers='ReLU'	colsample_bytree=1
		optimizer='Adam', learning rate=0.01	alpha=0.091
		batch size=100 , number of epochs=200	
SVR-NO_2	**RF − NO_2**	**ANN − NO_2**	**XGBoost − NO_2**
kernel='rbf'	n_estimators=194	number of hidden layers=3	n_estimators=499
C=1	max_leaf_nodes=1564	number of neurons in the 1st and 2nd hidden layers=128	max_depth=0
ϵ= 1	max_features=0.31	number of neurons in the 3rd hidden layer=100	eta=0.038
		activation function in the input and output layers='linear'	subsample=0.515
		activation function in the in the hidden layers='ReLU'	colsample_bytree=0.85
		optimizer='Adam', learning rate=0.01	alpha=0.0017
		batch size=100 , number of epochs=200	
SVR-O_3	**RF − O_3**	**ANN − O_3**	**XGBoost − O_3**
kernel='rbf'	n_estimators=418	number of hidden layers=3	n_estimators=1260
C=1	max_leaf_nodes=1836	number of neurons in the 1st and 2nd hidden layers=128	max_depth=0
ϵ=1	max_features=0.72	number of neurons in the 3rd hidden layer=100	eta=0.099
		activation function in the input and output layers='linear'	subsample=0.9
		activation function in the in the hidden layers='ReLU'	colsample_bytree=0.38
		optimizer='Adam', learning rate=0.01	alpha=0.9
		batch size=100 , number of epochs=200	
SVR-SO_2	**RF − SO_2**	**ANN − SO_2**	**XGBoost − SO_2**
kernel='rbf'	n_estimators=474	number of hidden layers=3	n_estimators=924
C=1	max_leaf_nodes=557	number of neurons in the 1st and 2nd hidden layers=128	max_depth=0
ϵ= 1	max_features=0.32	number of neurons in the 3rd hidden layer=100	eta=0.024
		activation function in the input and output layers='linear'	subsample=0.56
		activation function in the in the hidden layers='ReLU'	colsample_bytree=0.89
		optimizer='Adam', learning rate=0.01	alpha=0.045
		batch size=100 , number of epochs=200	

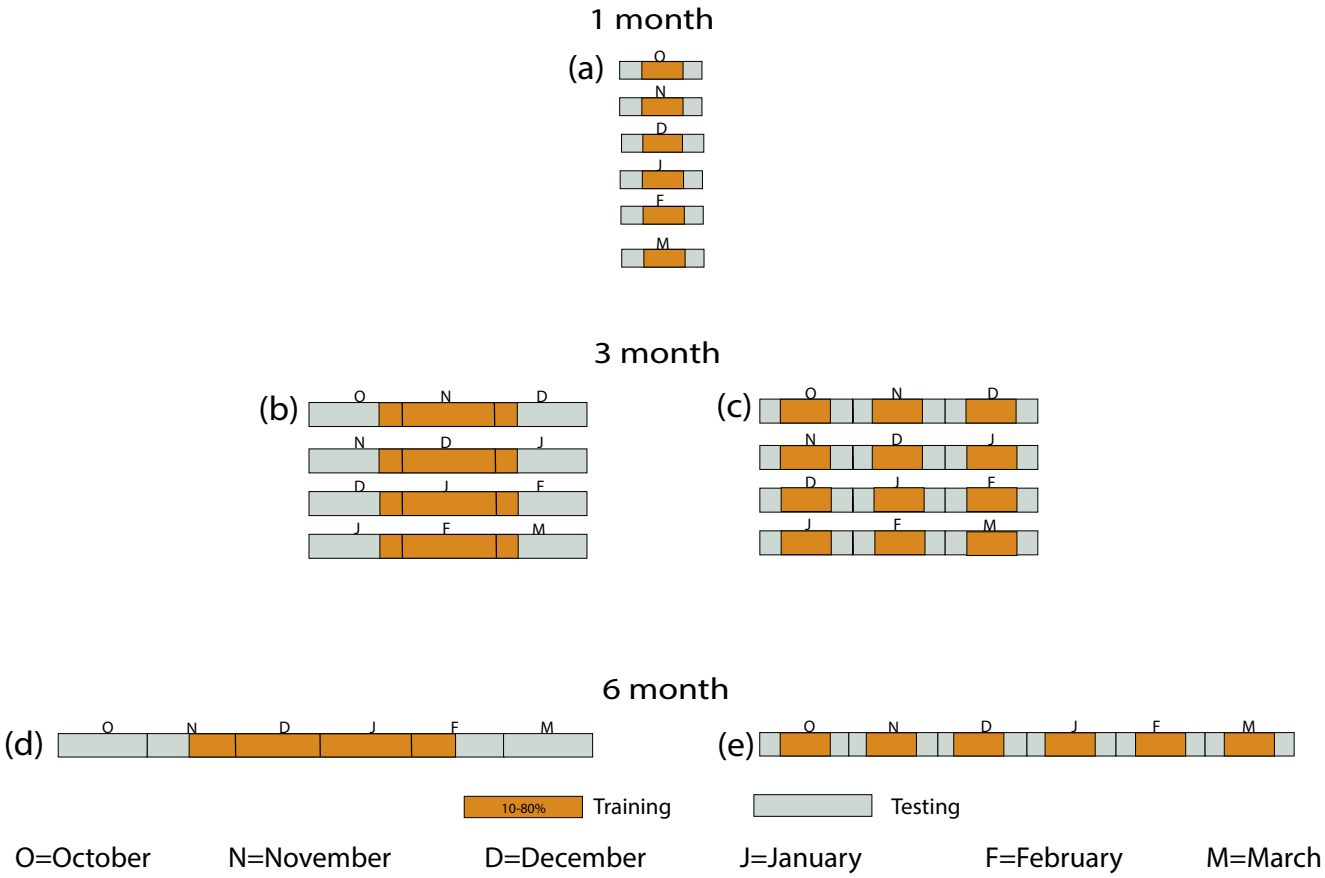

Figure 13: Data splitting schemes for 1, 3 and 6 month calibrations. For 1 month calibration the training data is obtained continuously at the middle of each month. For 3 and 6 month calibration, two data splitting cases are considered; the first case is where the training data is obtained continuously at the middle each calibration period while the second case is where the training data is sampled at the middle of month.

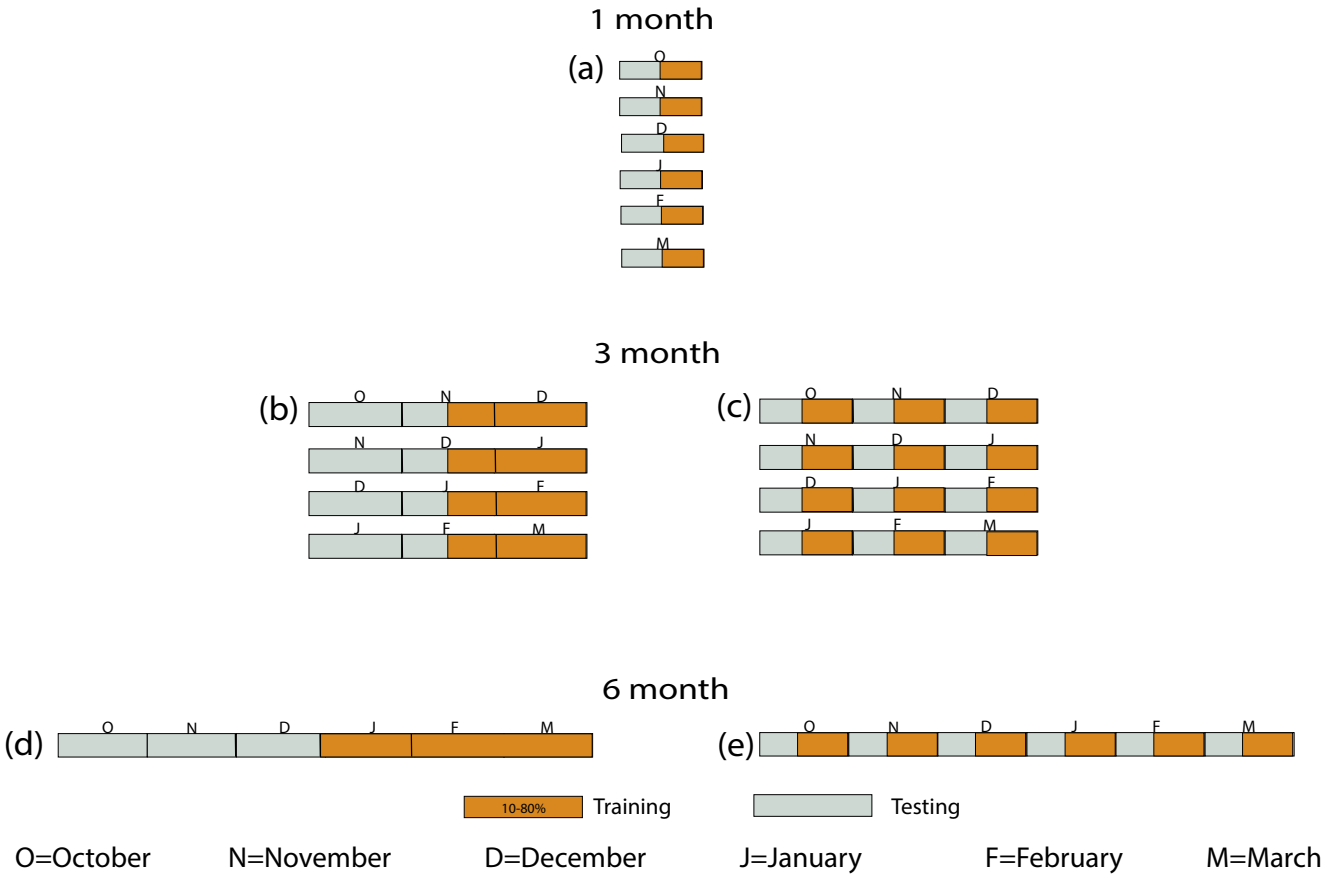

Figure 14: Data splitting schemes for 1, 3 and 6 month calibrations. For 1 month calibration the training data is obtained continuously at the end each month. For 3 and 6 month calibration, two data splitting cases are considered; the first case is where the training data is obtained continuously at the end of each calibration period while the second case is where the training data is sampled at the end of each month.

$$\Delta\text{NRMSE} = \frac{|\text{NRMSE(2min)} - \text{NRMSE(1h)}|}{\text{NRMSE(1h)}} * 100\% \tag{A.1}$$

$$\Delta\text{REU}_{\text{max}} = \frac{|\text{REU}_{\text{max}}\text{(2min)} - \text{REU}_{\text{max}}\text{(1h)}|}{\text{REU}_{\text{max}}\text{(1h)}} * 100\% \tag{A.2}$$