Performance of NO, NO_2 low cost sensors and three calibration approaches within a real world application

Alessandro Bigi[1], Michael Mueller[2], Stuart K. Grange[3], Grazia Ghermandi[1], and Christoph Hueglin[2]

[1]'Enzo Ferrari' Department of Engineering, University of Modena and Reggio Emilia, Modena, Italy
[2]Empa, Swiss Federal Institute for Materials Science and Technology, Duebendorf, Switzerland
[3]Wolfson Atmospheric Chemistry Laboratory, University of York, York, United Kingdom

Correspondence to: Alessandro Bigi (alessandro.bigi@unimore.it)

Abstract. Low cost sensors for measuring atmospheric pollutants are experiencing an increase in popularity worldwide among practitioners, academia and environmental agencies, and a large amount of data by these devices is being delivered to the public notwithstanding their behaviour, performance and reliability are not yet fully investigated and understood. In the present study we investigate the medium term performance of a set of NO and NO_2 electrochemical sensors in Switzerland using 3 different regression algorithms within a field calibration approach. In order to mimic a realistic application of these devices, the sensors were initially co-located at a rural regulatory monitoring site for a 4 – month calibration period, and subsequently deployed for 4 months at two distant regulatory urban sites in traffic and urban background conditions, where the performance of the calibration algorithms was explored. The applied algorithms were Multivariate Linear Regression, Support Vector Regression and Random Forest; these were tested, along with the sensors, in terms of generalisability, selectivity, drift, uncertainty, bias, ~~precision~~ noise and suitability for spatial mapping intra-urban pollution gradients with hourly resolution. Results from the deployment at the urban sites show a better performance of the non-linear algorithms (Support Vector Regression and Random Forest) achieving RMSE < 5 ppb, R^2 between 0.74 – 0.95 and MAE between 2 – 4 ppb. The combined use of both NO and NO_2 sensor output in the estimate of each pollutant showed some contribution by NO sensor to NO_2 estimate and vice-versa. All algorithms exhibited a drift ranging between 5 – 10 ppb for Random Forest and 15 ppb for Multivariate Linear regression at the end of the deployment. The lowest concentration correctly estimated, with a 25% relative expanded uncertainty, resulted in ca. 15 – 20 ppb and it was provided by the non-linear algorithms. As an assessment for the suitability of the tested sensors for a targeted application, the probability of resolving hourly concentration difference in cities was investigated. It was found that NO concentration differences of 5 – 10 ppb (8 – 10 for NO_2) can reliably be detected (90% confidence), depending on the air pollution level. The findings of this study, although derived from a specific sensor type and sensor model, base on a flexible methodology and have a large potential to explore the performance of other low cost sensors, different in target pollutant and sensing technology.

1 Introduction

Air quality assessment for regulatory purposes is addressed by means of monitoring stations following a strict QA/QC protocol in order to deliver measurements having an uncertainty within a specific range that is appropriate for the purpose (2008/50/EC, Council of Europe, 2008). The costs associated to these monitoring sites lead to a reconfiguration of regulatory air quality networks across Europe over the last decade, ~~occasionally associated to an analysis of the redundancy and the representativeness of single sites (e.g. Righini et al., 2014; Martin et al., 2014), generally resulting in sparser and occasionally more efficient~~ resulting in improved but still spatially sparse regulatory air quality networks over the continent. Although this trend towards optimization is coherent with main regulatory needs, it is not consistent with the increasing demand for spatio-temporal air quality information in urban areas, where largest part of worldwide population live (United Nations, 2015). Up to now two of the most promising approaches to estimate air quality conditions in complex environments as urban areas are simulation models and small low cost sensors. The former approach include ~~either statistical modelling (e.g. Mueller et al., 2016) or~~ dispersion modelling (e.g. Ghermandi et al., 2015), while the latter approach consists in sensor deployment for time-resolved air quality mapping (e.g. Mueller et al., 2016), plume tracking or other tasks. Besides some devices based on the absorption in the infrared region by the target gas, most common low cost sensors for gas phase compounds are based on either metal oxide or electrochemical technology. The high expectations from these two latter types of low cost sensors were seldom met, since they often face problems of calibration (Spinelle et al., 2013), stability (Fonollosa et al., 2016), cross-sensitivity (Mead et al., 2013) and low repeatability and reproducibility (Rai et al., 2017), urging for more research and tests for their mindful use (Lewis and Edwards, 2016). Among these problems, calibration is one of the major unsolved issues, preventing a broad use of these devices: ideally a calibration should include a full description of the sensors physical/chemical working principles along with its response to all environmental conditions and with ageing. Calibration approaches should be consistent with the intended application and the resulting measuring device, made up of a sensing unit and its calibration model, should meet the performance required by the application. Indications about possible minimum requirements for air quality studies can be taken by the EU directive 2008/50/EC, requiring an expanded uncertainty of 25% for indicative measurement devices.

Main current calibration solutions involve either ~~laboratory tests~~ sensor testing in the laboratory under controlled conditions or field co-location of sensors next to a calibrated reference instrument, ~~i.e.~~ with the former being an approach based on first principles and the latter an approach based on co-location data~~; up~~. Up today the former approach provided unsatisfactory results during the model validation in the field (e.g. Spinelle et al., 2017; Fonollosa et al., 2016), making a field calibration approach more commonly and successfully applied. This latter approach however introduced issues about the generalisability of a calibration model, because of the limited and site-specific range of environmental conditions occurring during the calibration period. This holds even more true in case the calibration and the following measurements are performed at two different sites, i.e. in case of relocation, with the additional possible influence of sensor handling and transport. Nonetheless, in the common case of field calibration, the subsequent relocation is extremely likely in a realistic application of these devices, because of the sparsity of the regulatory monitoring networks and given the most straightforward applications of these sensors, i.e. the collection of time-resolved air quality data where no data is available. In the literature the effect of relocation is scarcely

described, while several studies show results from a field calibration and further deployment at the same site. For this latter case several algorithms have been tested: since field calibration consists in a data driven approach, the algorithm used has a large impact on the final results. Some studies used models from classical statistics, e.g. Multivariate Linear Regression (Mijling et al., 2017; Mueller et al., 2017), or more sophisticated methods as high-dimensional model representation (Cross et al., 2017). In other studies several machine learning algorithms have been tested, for both metal oxide and electrochemical sensors, including also laboratory calibration: different types of Artificial Neural Networks (ANN, e.g. De Vito et al., 2009; Esposito et al., 2016; Spinelle et al., 2015), Reservoir Computing (Fonollosa et al., 2015), Random Forest (Zimmerman et al., 2018) and a recent comparison of 3 algorithms fed by dynamic and static input showing promising results by Support Vector Regression (De Vito et al., 2018).

This latter literature showed how generally calibration procedures involving non-linear methods outperform those using classical statistics, and better capture the effects of environmental factors on sensor response. However the performance shown by several of the methods cited above is not taking into account the effects of relocation, which has to be expected in a realistic use of similar devices. Main notable exception is a study on SO_2 electrochemical sensors by Hagan et al. (2018), who achieved RMSE values of ~ 8 ppb and $R^2 \sim 0.88$ during a 4 months relocation, using a hybrid regression model, combining a linear with a non-linear solution. Other studies involving relocation include Esposito et al. (2018), who showed a significant degradation of NO_2 estimate by electrochemical sensors after their relocation within the urban area of Oslo (Norway), along with the one by Zimmerman et al. (2018) who showed a good performance from a Random Forest regression model on a 4-weeks relocation in the vicinity of the calibration site.

In the present study we installed a set of electrochemical sensors at a rural site exposed to highway traffic emission for calibration and subsequently deployed these same sensors in two distant urban sites in traffic and background conditions. The first aim of the study is to compare state-of-the-art calibration algorithms, using a data-driven approach, within this realistic framework. The second is to investigate the change in performance over time and after a relocation of these measuring devices, i.e. of the sensor units (the hardware) and of their individual calibration (data processing algorithm). The final aim is the quantitative assessment of the measurement uncertainty of sensor units deployed in a network and investigate whether they are suitable for mapping intra-urban pollution gradients of NO and NO_2. The results strictly apply to the type and model of sensors involved (actually extremely popular among sensor systems) and to the environmental conditions during sampling, nonetheless the flexibility of the methodology here used has a large potential for other low cost sensing instruments.

In section 2 the sensor units and the calibration methods are described. Results from the calibration and the deployment periods are found in section 3. Finally the results are discussed and main conclusions are drawn. All data processing has been performed with the software R 3.4.2 (R Core Team, 2017).

2 Materials and methods

2.1 Sensor Units

Four identical sensor units have been jointly developed with Decentlab GmbH (Dübendorf, Switzerland) and used for this study. The sensor units used are labelled SU009, SU010, SU011 and SU012. Each unit consists of one box that includes two NO_2 sensors (Alphasense NO2-B43F), two NO sensors (Alphasense NO-B4), a relative humidity (RH) and temperature (T) sensor (Sensirion STH21) and a data transmission module using GSM/GPRS connection. The system is battery powered. Two identical NO sensors and two identical NO_2 sensors are used for a better control of the data quality. NO and NO_2 sensors are housed inside the box to better protect their gas permeable membrane, and a small blower is used to draw ambient air through a teflon (PTFE) manifold to which the sensors are connected. The electrochemical (EC) sensors used employ 4 electrodes: working, reference and counter electrodes account for target gas concentration, while a fourth auxiliary electrode compensate for zero current. The former three electrodes represent an electrochemical cell where a redox reaction of the target gas occurs, generating a electric current directly proportional to the gas concentration, while the auxiliary electrode accounts for changes in baseline signal ~~(further details in Mead et al., 2013; Alphasense Ltd, 2014)~~ (further details in Baron and Saffell, 2017; Alphasense Ltd, 2014).

The signal of each sensor is ~~aggregated to one minute mean values and~~ sampled every 20 seconds. Three such values are aggregated by the SU to a 1 minute mean value. These 1 minute values are transmitted to a central ~~data base every three hours~~database every 180 minutes. Data transmission implied both an increase in energy requirement by the transmission module, causing a drop in battery level, and spikes in electrode output. A despiking procedure based upon battery level data was applied: this consisted in selecting the data associated to single drops in battery level and removing them. In case few occasional spikes remained after this first procedure, these were selectively identified and removed by the following procedure: a running median of k original readings (r_{rm}) was calculated, the standard deviation of the difference between the original readings and r_{rm} was computed (σ_{dif}), then each original reading having a difference to r_{rm} larger than s times σ_{dif} was removed. This latter procedure used the command `despike` in the `oce` package, where k and s parameters were individually set for each electrochemical sensor. One minute despiked data were subsequently averaged to 10-minutes readings and used for all following analyses, except where stated otherwise.

2.2 Calibration and deployment sites

All 4 units were initially installed at the Härkingen (Switzerland) monitoring site within the Swiss Federal Air Quality Monitoring Network (HAE: 47.311 N, 7.820 E, 430 m a.s.l.). SU009, SU011 and SU012 were installed on April 13, 2017, while SU010 on May 5, 2017. All boxes were removed from HAE on July 20, 2017. The HAE monitoring site encounters clean/rural air masses when northern winds blow and polluted/highway air masses when southern winds blow. This allows an exposure of sensors to a wide range of pollutant concentration (Hueglin et al., 2006). The data collected at HAE represents the *calibration*

dataset (or training dataset) and were used to develop, train and validate the three regression algorithms tested in this study. In order to estimate the performance of the sensor units within a realistic application framework, the regression models calibrated upon this latter dataset were subsequently used to estimate concentrations after deploying the units to different sites, experiencing different pollution levels and different environmental conditions. On July 28th, 2017 the units were moved to two different air quality monitoring sites: SU009, SU010 were deployed at Zurich-Kaserne, an urban background site in Switzerland (ZUE: 47.378 N, 8.530 E, 408 m a.s.l.). SU011, SU012 were deployed at an urban traffic site in Lausanne, Switzerland (LAU: 46.522 N, 6.640 E, 495 m a.s.l.). At these monitoring sites NO, NO_2, O_3, temperature (T), relative humidity (RH) were available and were used to verify the concentration estimate by the sensor units: the data collected at ZUE and LAU represents the *deployment dataset* (or testing dataset), which includes data until December 5th 2017. Table S1 of the Supplementary shows the descriptive statistics of the meteorological and pollution conditions by the regulatory network instruments at the three sites, during their respective study period. The time series of the complete dataset is shown in Figure S1, linear correlation matrix for these same data is shown in Figure S2 and the NO_2/NO_x ratio is in Figure S3. The range in NO and NO_2 levels at the calibration site is similar to the deployment sites, benefiting the data driven calibration approach used, with the calibration site showing pollution conditions more similar to ZUE than to LAU.

2.3 Regression models and explanatory variables

Three different calibration algorithms have been tested: a multivariate linear regression model (MLR), a support vector regression model (SVR) and a random forest regression model (RF). These methods were used to estimate the atmospheric concentration of NO and NO_2 using only information available by each SU, i.e. voltage output by the EC sensors, T and RH. Two identical NO and NO_2 sensors in each sensor unit allows the use of tens of different combinations of explanatory variables in the regression models, for example a set based on the mean of the net voltages of the replicate EC sensors or on the individual net signals of both.

Firstly the best set of explanatory variables was selected by comparing the performance of the algorithms in using 10 different model equations. For each tested model SVR was tuned for each pollutant and each SUs, while the same hyperparameters set was used for RF. In this task, for tuning and performance estimate, only the calibration dataset was used, consistently with the realistic framework of this study. Finally, the best performing model was selected and the regression models, tuned and calibrated upon the calibration dataset, were applied to the deployment dataset to estimate pollutant concentration. The equations of the 4 main covariate combinations that were tested are listed in Appendix A: these models are labelled *minimal* when using 1 EC sensor only (equations A1, A2), *basic* when using one NO and one NO_2 EC sensor (equations A3, A4), *single replicate* when using 2 EC sensors of the same gas (equations A5, A6) and *double replicate* when using the 4 EC sensors (equations A7, A8). All equations include ambient RH and T readings by their respective sensor within each SU.

All plots and results in the remainder of the text proceed from the model including all 4 EC sensors, i.e. the one achieving the best performance *on the calibration dataset*. However, since the redundancy in EC sensors is a feature specific to the SUs used in this study, for the sake of comparability with the literature and to verify the benefit of a redundant design, the final

performance of the SUs *at the deployment sites* using the 4 main regression models listed in Appendix A is shown in Figures S4, S5 and in Table S2.

2.3.1 Multivariate linear regression

The MLR model used in this study partly included MLR requirements of independent covariates. In a previous study Mueller et al. (2017) employed Alphasense NO_2 B42F sensors and among the explanatory variables both the weighted cumulative index of past RH changes and the change in sensor sensitivity with temperature (as observed in lab tests, Alphasense Ltd, 2017). The latter covariate was included in the 4 tested models (see Appendix A). In the present study the final regression model for NO and NO_2 followed equation 1, where $\overline{V_{NO}}$ indicates the mean net voltage produced by the replicate EC sensors for NO, $\overline{V_{NO_2}}$ indicates the net voltage produced by the replicate EC sensor for NO_2, with net voltage being the difference between the working and auxiliary electrodes. Note that this model is also listed in the Appendix in equation A7.

$$\begin{aligned} \mathrm{NO} &= \beta_0 + \beta_1 \overline{V_{NO}} + \beta_2 \overline{V_{NO_2}} + \beta_3 T + \beta_4 \mathrm{RH} + \beta_5 \overline{V_{NO}} \times T + \epsilon \\ \mathrm{NO_2} &= \beta_0 + \beta_1 \overline{V_{NO}} + \beta_2 \overline{V_{NO_2}} + \beta_3 T + \beta_4 \mathrm{RH} + \beta_5 \overline{V_{NO_2}} \times T + \epsilon \end{aligned} \tag{1}$$

2.3.2 Support Vector Regression

SVR modelling consists in a machine composed by three main steps: in the former the input data are mapped into a (high dimensional) feature space by means of a function, generally a kernel. In the second step the flattest function fitting the images of the input is found in the feature space by solving the corresponding constrained optimization equation: Support Vectors are the points corresponding to the non-null Lagrangian multipliers of this latter function. In the latter step the results are mapped back into the input space. More details on SVR modelling can be found in Smola and Schölkopf (2004). In the present study we used ϵ–SVR featured by a Gaussian radial basis kernel: the three main hyperparameters of this model are ϵ, the parameter of the insensitive-loss function, σ, the inverse kernel width, and C, the cost of constraints violation. These hyperparameters were tuned upon the calibration dataset by a 5 – fold cross-validation approach and the best performing set was selected using three different goodness-of-fit metrics, i.e. the mean of squared errors, the root-mean of squared errors and the coefficient of determination. The hyperparameters were individually tuned for each sensor unit and each pollutant.

SVR modelling and tuning were achieved using the `kernlab` and `mlr` packages for R (Karatzoglou et al., 2004; Bischl et al., 2016). Fast and optimal SVR hyperparameter tuning is an active research area within the scientific community, motivated by the hyperparameters reciprocal interaction leading to large hyperparameter spaces to be explored for an optimal result. The computing time and computing resources needed to tune the calibration dataset were significantly larger than for the other models (70-300 core-hours per sensor per pollutant on one Intel i7-6700 CPU at 3.40 GHz), moreover SVR showed a tendency to overfit the data and it often lead to similar fitting performance with different hyperparameter sets: for final optimal results, a minor manual tuning on ϵ was occasionally applied on a model bias-variance trade-off basis (Cawley and Talbot, 2010).

2.3.3 Random Forest regression

RF modelling consists in growing M randomized trees, representing the forest, where each tree is built on a random subset of the p-dimensional initial sample $\mathbb{X}^p$. A tree is grown by performing optimal cuts of each tree node (starting from the root), until the cardinality of each final cell is lower than *nodesize*. Cut optimality is estimated using the Classification And Regression Trees split criterion (CART) (Breiman et al., 1993): this algorithm compares the variance of the uncut node, with the variance of all possible cuts along *mtry* directions, where *mtry* is a random subset of sample coordinates p. The prediction is produced by averaging all tree estimates into a (pointwise) forest estimate. More details on RF regression modelling can be found in Breiman (2001).

Two main approaches exist to overcome the RF standard pointwise estimate and build an interval for model prediction, i.e. to include modelling uncertainty in the final estimate: forest-based quantile regression (QRF) and inference on RF estimates (RF-CI). Predictions by quantile regression forest results from keeping all observations in every node in every tree and estimating a weighted mean for each observation (Meinshausen, 2006). Confidence interval for RF estimates is an open research topic being tackled in different ways (e.g. Wager et al., 2014; Sexton and Laake, 2009; Mentch and Hooker, 2016). In this study, the uncertainty of point predictions was tentatively assessed by using both approaches, although still experimental. For the assessment of confidence intervals we used the approach by Athey et al. (2016), who rely the inference on asymptotically gaussian RF predictions and use the bootstrap of little bags algorithm (Sexton and Laake, 2009) to compute asymptotically valid confidence intervals. In this study standard RF modelling was performed using the `RandomForestSRC` package in R, while quantile regression and confidence interval estimate were both performed using the `grf` package in R.

Main RF hyperparameters (*mtry, nodesize, M*) were tuned upon the calibration dataset by a 5 – fold cross validation by investigating several goodness-of-fit metrics. The possible range of RF hyperparameters is narrower than SVR and RF model showed a minor sensitivity to changes in *mtry* and *nodesize*, because of the small number of covariates. Finally *nodesize* and *mtry* were set to 7 and 5 respectively, slightly larger than their recommended values, to further avoid overfitting, an unlikely event for RF models (Breiman, 2001). The number of trees was set to 1000 for standard forest and to 4000 for QRF and RF-CI forests. These hyperparameter values were used for all SUs and all pollutants. It is worth noting that small differences exists between `RandomForestSRC` and `grf`, which are mainly due to the splitting algorithm, i.e. the use of fair and unfair forests (Athey et al., 2016), besides that QRF central estimate is the forest median, while the other two RF flavours use the forest mean.

2.3.4 Features of machine learning regression models

SVR and RF modelling share the ability to build a non-linear regression model using several time series as explanatory variables and are superior to MLR in handling both autocorrelation and multicollinearity. This ability allowed to freely test any combination of the possible covariates and finally, for both SVR and RF, lead to the regression model in equation 2, where

V_{NO^A} indicates the net voltage by the NO sensor A, $V_{NO_2^A}$ indicates the net voltage by the NO_2 sensor A, and consistently V_{NO^B} and $V_{NO_2^B}$ for the respective replicate sensor B. The model in equation 2 is also listed in the Appendix as equation A8.

$$\begin{aligned} \mathrm{NO} &= \mathrm{function}(V_{NO^A}, V_{NO_2^A}, V_{NO^B}, V_{NO_2^B}, T, \mathrm{RH}) \\ \mathrm{NO_2} &= \mathrm{function}(V_{NO^A}, V_{NO_2^A}, V_{NO^B}, V_{NO_2^B}, T, \mathrm{RH}) \end{aligned} \tag{2}$$

Using a similar model structure for MLR would strongly violate the requirements for a reliable estimate of MLR errors. It is worth noting that the residuals from the SVR and RF application of equation 2 are independent, contrarily to MLR residuals from equation 1: this latter model shows autocorrelated residuals, to be expected from an ordinary linear regression on a time series, and inflated variance for its coefficients, because of the multicollinearity of the regressors. Nonetheless MLR has been included among the regression methods in this study for its wide use, also in low cost sensor calibration. A further difference among algorithms is that MLR and SVR allow to extrapolate outside the range of their input dataset, while the estimates provided by RF can only be within the bounds of the calibration space, being RF a tree-based algorithm. This worth noting feature of RF on one side implies a constraint on its application to relocated SUs, on the other side it will guarantee only positive estimates.

The role of each predictor in MLR, SVR and RF models was assessed by estimating its partial dependence, which consists in evaluating the average prediction when the covariate of interest is held constant, repeating this prediction for a set of values across the distribution of this covariate. Partial dependence plots allow to investigate the effect of each covariate on the prediction. For RF models only, it is also possible to estimate the importance of each variable by computing the increase in prediction error by randomly permuting each covariate in every tree and averaging this prediction error over the forest (Breiman, 2001): the larger the increase in prediction error, the larger is the importance of the variable for that RF model. This importance metric of a variable is the error occurring if a RF model, calibrated including that variable, is used in prediction without that same variable.

3 Results

Several goodness-of-fit indexes were used to assess the overall performance of the 4 SUs when individually calibrated using the different described calibration approaches: these include root mean square error (RMSE), centred root mean square error (CRMSE), mean bias error (MBE), mean absolute error (MAE) and the coefficient of determination (R^2). Temporal variability of these indexes was investigated, along with an overall performance of the sensing devices.

3.1 Results for the calibration dataset

Partial plots applied to the calibration dataset of SU009 are shown in Figures 1 and 2 and of SU010, SU011 and SU012 in Figures S6 through S11. These provide insights in the role of each predictor within the model, a remedy for the widely perceived black box nature of machine learning algorithms. The most prominent result by these plots is the difference existing

among the three algorithms: MLR implies a linear response from each covariate, while SVR and RF allows non-linearity. The partial plots for EC net voltage vs its target gas show a similar pattern across all SUs and all algorithms, indicating that the final model structure generalises well across the hardware for this covariate, and that the differences existing among SUs are minor in this case. Both SVR and RF exploit the replicate EC sensors: the former algorithm shows significant response by replicate gas sensors in the estimate of their target gas (i.e. by both NO_2 EC sensors in predicting atmospheric NO_2), while RF shows large response by both replicate sensors only in case of NO by SU009 and by SU011. It is notable the similarity in the response of atmospheric variables according to SVR and RF, supporting the result also by these specific partial plots. RF correctly identifies the most informative variable (as supported by the variable importance plots in Figure 3 and S12) and it appears to be the most efficient algorithm among the three: - it shows a quasi-linear response of the EC net voltage towards its target gas, contrarily to the often non-monotonic behaviour shown by SVR - this linear behaviour is held across large part of the full range of the EC net voltage output - for RF estimating a gas, the net voltage of the EC sensor targeting that same gas has the broadest response among all covariates. The non-monotonicity in the partial response from SVR suggests that a minor overfit is still present, although this is not affecting significantly the performance during deployment.

Variable importance plots (Figures 3 and S12), possible for RF only, show how the main regression variable is the net voltage by the EC sensor of the corresponding target gas. Its importance is generally ~ 4 times larger than the second important variable, however for NO prediction by SU009 and SU011, the second most important variable is the replicate NO sensor and in this case its importance is closer to the first most important variable.

The effect of RH on sensor response is extremely low for all algorithms, consistently with results from laboratory studies (e.g. Spinelle et al., 2017). Nonetheless humidity transients are known for being responsible of spurious responses by the EC sensors (Mueller et al., 2017; Alphasense Ltd, 2017; Pang et al., 2017), but this effect was not parameterized in this study, possibly leading to a slightly degraded model estimate. However no anomalous peak was evident in the 10 – minute data, although rapid variations in atmospheric RH occurred.

Independently of the calibration algorithm, partial plots indicate a contribution by NO_2 and NO EC sensors to NO and NO_2 respectively: this might be due to the inability of the algorithms to untangle the large correlation of these pollutants in the atmosphere, and/or an existing cross-sensitivity of the EC sensors. The latter cause cannot be excluded completely ~~, since selectivity issues had been noticed for electrochemical sensors , e.g. a previous version of this~~ and was highlighted in several field deployment of EC sensors: both NO_2 sensors Alphasense NO2-B4 and Alphasense NO2-B43F exhibited a significant cross-sensitivity to CO_2 at atmospheric levels (Lewis and Edwards, 2016; Kim et al., 2018), while NO_2 ~~EC sensor (NO2B4) exhibited a significant cross sensitivity for~~ sensor Alphasense NO2-B42F was shown to have large cross-sensitivity to NO by Kim et al. (2018). Nonetheless literature studies available do not provide a clear and consistent picture about sensor selectivity and further laboratory tests are required to shed light on this topic. During this study no concurrent suitable data of atmospheric CO_2 ~~(Lewis and Edwards, 2016)~~was available, preventing an investigation of possible bias in sensor estimates of NO_2 induced by the cross-sensitivity to CO_2 in the field.

The cross-sensitivity, along with a site – and time – specific NO – NO_2 correlation, may prevent the application of a calibrated regression model over a wide spatial and temporal scale, because of a different NO/NO_2 ratio at the calibration and the

deployment site. ~~Whether their effect is present and/or large enough to result in a lack of spatial generalisation is detectable from the~~ The SU performance at LAU and ZUE (paragraph 3.2) ~~, since these are urban sites representing different~~ allows the evaluation of the effect of relocation of the sensors on the data quality, since the two sites are representing urban air pollution situations ~~compared to HAE (~~that are different from the site where the collocated measurements have been performed (HAE), see Table S1 and Figures S1, S2 and S3).

~~Variable importance plots (Figures 3 and S12), possible for RF only, show how the main regression variable is the net voltage by the EC sensor of the corresponding target gas. Its importance is generally ~ 4 times larger than the second important variable, however for prediction by SU009 and SU011, the second most important variable is the replicate sensor and in this case its importance is closer to the first most important variable.~~

In order to ~~extensively~~ further test the generalisability of the response by each covariate and hence of the proposed models, the 3 algorithms were calibrated also using the deployment dataset, in order to build partial plots at ZUE and LAU (Figures S13 through S20): note that SVR and RF were not tuned in this case, i.e. the same hyperparameter sets as for HAE were used. These latter partial plots are largely similar with those derived from the HAE dataset, particularly for MLR and RF, while SVR still exhibits some overfit. Each SU shows similar patterns between its partial plots for the calibration and the deployment dataset, including for the response of the EC sensor to their non-target gas. A minor exceptions to this latter point is the response by NO sensor B in SU010 (Figures S7 and S16), suggesting that the NO/NO_2 ratio partly influences the response of non-target gas sensors. Overall these latter partial plots also show how the main behaviour of each SU was not significantly affected by 7-months outdoor installation, notwithstanding the relocation and the change in environmental conditions.

3.2 Results for the deployment dataset

Time series of estimate from SU009, deployed at the urban background site ZUE, and from SU011, deployed at the urban traffic site LAU, are summarised in Figures 4 and 5. Summary plots for SU010 and SU012, deployed at the background and traffic site respectively, are in Figures S21 and S22. SVR and RF performed similarly and generally better than MLR, with a RMSE ranging between $2-5$ ppb for both NO and NO_2. Notwithstanding their similar goodness of fit indexes, RF showed a more regular performance than SVR across the SUs and the pollutants, and its time series predictions are more stable than the ones by SVR, which occasionally show negative spikes (e.g NO_2 by SU012 in Figure S22).

Several analyses have been performed to detail the performance of each device during deployment. Timeseries of goodness-of-fit indexes, computed with a rolling window of 1 week, indicate the change of model performance over time: in target plots (Spinelle et al., 2015) the change in performance is plotted in terms of CRMSE and MBE, both normalised by the standard deviation of the reference (σ_{ref}), and the right quadrants are used when the standard deviation of the reference is lower than the one from model predictions, and vice-versa. In target plots the distance of each target score to the origin equals RMSE normalised by σ_{ref}. Finally, a unit circumference is added to this diagram, containing model predictions having residuals with a standard deviation smaller σ_{ref}. Time-resolved target plots for the deployment dataset highlight significant variability in performance with time depending on the device, on the gas and on the algorithm. All 3 algorithms provide results within the

unit circumference for most of the deployment period, and confirm how SVR and RF results are generally better than those by MLR (Figures 6 and S23 through S25).

The timeseries of 1 week rolling RMSE in Figure 7 indicate an overall lower performance in the estimate of NO, most likely due to its larger variability, and a more steady trend for NO_2. The RMSE for MLR is, in most occasions, the largest among the three algorithms, while SVR and RF performed similarly. The lowest variation in RMSE, ranging in 2 ppb, was observed for NO estimates by RF on SU010 data, while an increase up to 6 ppb occurred in the case of NO_2 predictions by MLR on SU010 readings. In some cases the increasing trend in RMSE is evident, e.g. for NO_2 by SU009, in others the large variability hinder a clear assessment of the status of the SU, e.g. for NO_2 by SU012.

The sensing devices (i.e. the sensor units and their individual calibration) were investigated also in terms of their drift, uncertainty, bias, noise and ability of resolving spatial differences in pollution levels: for better comparison with common regulatory measurements, all these analyses were performed using 1 – hour average input data, instead of 10 – minutes as for previous ones. Nonetheless the use of 10 – minutes data delivered similar and consistent results (not shown). As a proxy for the overall drift in the estimate by sensor devices, the time series of mean daily residuals was computed: results in Figure 8 confirm the occurrence of a drift in all cases, although only occasionally with a clear trend, and among algorithms RF generally outperforms both SVR and MLR, achieving an absolute variation in the residuals between 5 – 10 ppb after 4 months of deployment. As a specific proxy for zero-drift we used the SU estimates coupled to reference instruments measurements < 0.5 ppb: this analysis, not possible for NO_2 due its low statistics of quasi-null values, confirms the better performance of the two machine learning algorithms and hints to zero-drift of ~ -10 ppb or $\sim +2$ ppb in the worst and in the best case respectively. The values of these proxies for the overall drift and the zero-drift are consistent with the results for these same EC sensors by Kim et al. (2018), who reported an absolute zero-drift (from laboratory measurements) of 2 ppb and 16 ppb for NO and NO_2 after 2.5 months of field deployment.

The uncertainty of the devices was computed as relative expanded uncertainty according to the guidelines for the data quality objective required by the directive 2008/50/EC (WG, 2010) and compared either to the expanded uncertainty of the reference instrument (EMPA, 2016), and to the 25% recommendation for indicative measurements by the same directive, as a reasonable threshold required for the detection of pollution gradients within urban areas, i.e. for a possible application of these devices. Results show some variability between the two deployment sites, with highest uncertainty for NO in Lausanne (traffic site). According to this procedure, the calculated relative expanded measurement uncertainty by SUs are within 25% for mixing ratios larger than about 15 – 20 ppb for both NO and NO_2. Calibration models based on RF have generally the lowest uncertainty among the three algorithms (Figures 9 and S26).

A further assessment of the uncertainty of these devices at the deployment sites was obtained by binning reference concentration in 1 ppb intervals and estimating for each bin the corresponding 5th – 95th quantile range of the predictions, along with the median. The quantile range was calculated only for the RF estimates using 1 – hour input data and if at least 10 values were available. Results are shown in Figure S27 and include the 1:1 line along with its 25% and 35% uncertainty intervals. In these figures the bottom shortest rug (red coloured) indicates whether the median is included in the 25% uncertainty bounds. The rug in green (blue) indicates if the 5 – 95% percentile range is included in the 35% (25%) uncertainty range. The estimate by the

sensor units is linear over a broad range of NO and NO_2, with a fairly constant 5 – 95% percentile range in most cases, besides for NO in Lausanne (traffic site), hinting to a fairly steady precision for these devices. The bias for the median is in the order of several ppb over large parts of the concentration range for both pollutants and most of the SUs.

The lowest concentration correctly estimated on 90% of occurrences with a specified uncertainty is again dependent on the SU, on the site and on the pollutant: at the urban background site (Zurich) this lowest concentration is provided by SU010 and results in $\sim$ 15 ppb ($\sim$ 20 ppb) for NO (NO_2) and this is also the best result across all 4 devices. At the urban traffic site (Lausanne) the lowest concentration correctly estimated (on 90% of occurrences and with a 25% uncertainty) is $\sim$ 50 ppb (by SU012) and $\sim$ 30 ppb (by SU011) for NO and NO_2 respectively; these latter thresholds reduce to $\sim$ 15 ppb for both pollutants if a 35% uncertainty is considered.

The potential benefit of using 8 EC sensors in the same RF model was tested by combining the data of the two SUs deployed at the same site into the same RF model. Results for Zurich (combining SU009 with SU010) and Lausanne (combining SU011 with SU012) lead to figures similar to the best performing SU at the respective site, i.e. did not lead to a decrease in uncertainty, suggesting that this latter has a more fundamental constraint, either from the calibration approach or by the EC and the measurement system themselves. Nonetheless the combined use of the two SUs lead to a slight improvement in the overall goodness-of-fit indexes, with a decrease of the RMSE of $\sim$ 0.5 ppb (see Table S3).

The overall sensor noise for each bin was computed as the 2σ of the RF estimate, if at least 10 estimates were available for the bin. The median of this 2σ noise ranged in $\pm 4-7$ ppb and in $\pm 5-8$ ppb for NO and NO_2, i.e. only 1 – 2 ppb larger than the noise observed by Kim et al. (2018) under laboratory conditions on 10 s data, and half of the 2σ noise reported by the EC sensor manufacturer.

Finally, we were interested whether the tested sensor units would be appropriate for a targeted application, i.e. for resolving the intra-urban concentration gradient with hourly resolution. Assume that sensor units are deployed in the same urban environment at two distant sites A and B, where A is typically less polluted (urban background site) compared to site B (site impacted by nearby sources such as road traffic). For this assessment, the data from all four SUs have been pooled in order to account for different performance of individual sensor units, and similarly to the previous uncertainty assessments, only RF estimates using 1 – hour input data were used. Next, the concentrations measured by the reference instruments were binned in 1 ppb intervals and denoted reference bins. The corresponding sensors measurements were then linked to the reference bins. Any concentration difference between sites B and A can now been simulated by the reference bins, and the probability distribution of the concentration difference as measured by the tested sensors can be expressed by the concentration differences of the sensor measurements assigned to the corresponding reference bins. Integrating the sample probability distribution of the concentration difference over values larger than zero provided the probability that the concentration gradient between site B and A is resolved by two different SUs. This probability was computed if at least 10 estimates were available for either site A and site B. Figure 10 shows the probability that, for a given reference concentration at site A and its difference in concentration with site B, the measurements by a SU at site B are larger than measurements by a SU at site A. In Figure 10 ~~blue~~ red dots indicate the concentration difference between site B and A that can be detected with a probability of 90%. Figure 10 highlights how the possibility of resolving the gradient depends both on the gradient amount and on the concentration at site A, besides

some influence by the sample size, as evident by the lower chance of resolving differences at higher (and less frequent) levels. Generally gradients in NO above $\sim$ 5 ppb to $\sim$ 10 ppb are likely to be captured by these devices, while for NO_2 a gradient of almost 10 ppb is needed. These results were compared to the hourly gradient in a pool of European cities, including several sites in the Po valley, a NO_x hotspot for Europe. The data used proceed from 2 years of regulatory measurements at reference monitoring sites: data for years 2016 and 2017 were used for Italy and delivered directly by the local Environmental Agencies, data for years 2015 and 2016 were used for the other cities and provided by the Air Quality e-Reporting (European Environment Agency, 2017) (boxplots summarising this dataset are found in Figure S28 and S29). For each city, the intra-urban gradient was computed as the maximum hourly difference between traffic and background urban sites within the same urban area; when more than two reference sites were available, the pair of sites showing the largest concentration difference was selected. In 10 the ordinates of each city indicate its intra-urban gradient, while the abscissa expresses its median over the analysed period. As a final step, the uncertainty in RF estimates was tentatively estimated by using experimental Quantile Random Forest regression (QRF) and Confidence Interval estimates (RF-CI). Results for QRF (band including 5 th to 95 th quantiles) shows that ca. 80% of reference values are within the QRF band for both NO and NO_2. On the contrary confidence bands by RF-CI, containing ca. 20% of the predictions, appear excessively narrow, although the mean prediction still indicates a good performance for this model (Figures 11, ~~S28, S29 and S30~~S30, S31 and S32).

4 Conclusions

Four sensor units (SU) using low cost electrochemical sensors (EC) were tested with three calibration approaches. The study simulates a possible realistic application of these devices and consisted in field-calibrating the units at a single air quality monitoring site and subsequently deploy the units at two distant air quality monitoring sites. This procedure added relocation to the other well documented sources of uncertainty by low cost sensors (e.g. stability, cross-sensitivity, reproducibility), involving further possible errors generated by differences in pollution levels and environmental conditions between the calibration and deployment site and between the calibration and the deployment period. Within this realistic framework the performance of three state-of-the-art calibration algorithms were tested: Multivariate Linear Regression (MLR), Support Vector Regression (SVR) and Random Forest (RF). For each SU and for each algorithm, the overall performance and its change over time was estimated according to several metrics. Drift, uncertainty~~and bias~~, bias and noise were assessed, along with the probability to resolve spatial concentration differences by using these SUs, still within the same realistic framework.

Each unit hosted two EC sensors for each of the two monitored pollutants (NO and NO_2), resulting in several possible covariate combinations for the regression models. For all three algorithms the model fully exploiting the replicate EC sensors performed best, with RF resulting the most successful algorithm. MLR achieved the worst performance according to all goodness-of-fit indexes, along with a large drift over time, which is not surprising given the large autocorrelation in its residuals, indicating that important information from the input data were not included in the regression model. SVR overall performance is comparable, or occasionally better, than RF throughout the deployment period, however the tuning of its param-

eters is computer-intensive and the algorithm exhibited a tendency to overfit (as shown by the occasional lack of monotonicity in its partial plots), discouraging its use in a realistic production application, potentially involving several sensor units.

The lowest correctly estimated concentration resulted mainly dependent on the SU, on the pollutant and on the algorithm: best results for this study indicate 15 – 20 ppb for both NO and NO_2, if an expanded uncertainty of 25% is considered. ~~A large variability in RMSE and drift was observed, hinting to a challenging, although desirable, application of a possible QA/QC protocol for the management of this type of sensors units. Bias and precision resulted significant, although generally lower than the two previous indexes. Before using sensors for air quality measurements, it should be investigated if the selected sensors are fit for the intended purpose. We investigated whether the tested sensors are generally capable for resolving intra-urban concentration gradients on a hourly basis~~RMSE ranged between 3–7 ppb, drift resulted few ppb larger and the 2σ noise showed figures similar to RMSE. When calibrated, the sensors ~~are~~ resulted capable to detect concentration differences of about 5 – 10 ppb for NO and 8 – 10 ppb for NO_2, depending on the urban background level. ~~The data quality provided by the sensors is insufficient for cleaner cities with small intra-urban gradients, but could be considered for being used in higher polluted cities.~~

It is worth noting how the performance of the three algorithms is strongly dependent on the comparability between the calibration and the deployment space: the more similar are these spaces, the better will be the performance of the measuring device in case of field calibration. Standard RF is not able to extrapolate out-of-sample, as clearly shown e.g. by the steady NO prediction corresponding to observations larger than 100 ppb (Figure S19): notwithstanding the remarkable performance achieved by this algorithm, this feature of RF on one side represents a main limitation, on the other it allows to confine the estimates within the calibration space and to identify possible misalignments between the calibration and the deployment spaces.

Finally, although the use of a confidence band in the estimates by low cost sensors should be recommended, in the present study, confidence bands for RF resulted too experimental to be used for application studies.

On a broader view, these results recommend to investigate whether these sensors are fit for the intended purpose and the intended environment, prior to their use. Given the performance of these devices, they resulted unsuitable for cleaner urban areas (e.g. in background mountain locations) and unsuitable to reliably map small intra-urban gradients; nonetheless they also showed a large potential for time-resolved monitoring of NO and NO_2 in medium-to-high polluted areas and for quantitatively resolving intra-urban concentration gradients on a hourly basis in higher polluted and larger cities. Targeted QA/QC protocols for the management of this class of sensors and/or of a network of sensors need to be implemented for achieving the best and constant data quality during medium to longterm deployment.

Data availability. All data can be provided by the authors upon request

Competing interests. The sensor units have been jointly developed by Decentlab and Empa. The authors declare to have no other competing interests.

Acknowledgements. A.B. was supported by the Swiss National Science Foundation International Short Visit Grant (IZK0Z2-174969). S.K.G was supported by Anthony Wild with the provision of the Wild Fund Scholarship.

References

Alphasense Ltd: Alphasense 4-Electrode Individual Sensor Board (ISB), Great Notley, UK, 085-2217 edn., 2014.

Alphasense Ltd: Environmental changes: temperature, pressure, humidity, Tech. Rep. AAN 110, Great Notley, UK, http://www.alphasense.com/WEB1213/wp-content/uploads/2013/07/AAN_110.pdf, last access: 07 May 2018, 2017.

Athey, S., Tibshirani, J., and Wager, S.: Generalized Random Forests, 2016.

Baron, R. and Saffell, J.: Amperometric Gas Sensors as a Low Cost Emerging Technology Platform for Air Quality Monitoring Applications: A Review, ACS Sensors, 2, 1553–1566, https://doi.org/10.1021/acssensors.7b00620, 2017.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z. M.: mlr: Machine Learning in R, Journal of Machine Learning Research, 17, 1–5, 2016.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Breiman, L., Friedman, J., Stone, C., and Olshen, R.: Classification and Regression Trees, Chapman & Hall, 1993.

Cawley, G. C. and Talbot, N. L. C.: On over-fitting in model selection and subsequent selection bias in performance evaluation, Journal of Machine Learning Research, 11, 2079–2107, 2010.

Council of Europe: 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, Ofcial Journal of the European Union, Official Journal of the European Union, L152/1–L152/144, 2008.

Cross, E. S., Williams, L. R., Lewis, D. K., Magoon, G. R., Onasch, T. B., Kaminsky, M. L., Worsnop, D. R., and Jayne, J. T.: Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements, Atmospheric Measurement Techniques, 10, 3575–3588, https://doi.org/10.5194/amt-10-3575-2017, 2017.

De Vito, S., Piga, M., Martinotto, L., and Francia, G. D.: CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization, Sensors and Actuators B: Chemical, 143, 182 – 191, https://doi.org/10.1016/j.snb.2009.08.041, 2009.

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

EMPA: Technical report for the national monitoring network of atmospheric pollutants (NABEL), 2016 (in German), Tech. rep., EMPA, https://www.empa.ch/documents/56101/246436/Technischer+Bericht+2016/0bc321a3-f489-4f20-bcda-a323fbc4ca8a, last access: 07 May 2018, 2016.

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

Esposito, E., Salvato, M., De Vito, S., Fattoruso, G., Castell, N., Karatzas, K., and Di Francia, G.: Assessing the Relocation Robustness of on Field Calibrations for Air Quality Monitoring Devices, pp. 303–312, Springer International Publishing, https://doi.org/10.1007/978-3-319-66802-4_38, 2018.

European Environment Agency: Eionet Central Data Repository, http://cdr.eionet.europa.eu/, 2017.

Fonollosa, J., Sheik, S., Huerta, R., and Marco, S.: Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring, Sensors and Actuators B: Chemical, 215, 618 – 629, https://doi.org/10.1016/j.snb.2015.03.028, 2015.

Fonollosa, J., Fernández, L., Gutiérrez-Gálvez, A., Huerta, R., and Marco, S.: Calibration transfer and drift counteraction in chemical sensor arrays using Direct Standardization, Sensors and Actuators B: Chemical, 236, 1044 – 1053, https://doi.org/10.1016/j.snb.2016.05.089, 2016.

Ghermandi, G., Fabbi, S., Zaccanti, M., Bigi, A., and Teggi, S.: Micro–scale simulation of atmospheric emissions from power–plant stacks in the Po Valley, Atmospheric Pollution Research, 6, 382–388, https://doi.org/10.5094/APR.2015.042, 2015.

Hagan, D. H., Isaacman-VanWertz, G., Franklin, J. P., Wallace, L. M. M., Kocar, B. D., Heald, C. L., and Kroll, J. H.: Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments, Atmospheric Measurement Techniques, 11, 315–328, https://doi.org/10.5194/amt-11-315-2018, 2018.

Hueglin, C., Buchmann, B., and Weber, R.: Long-term observation of real-world road traffic emission factors on a motorway in Switzerland, Atmospheric Environment, 40, 3696–3709, https://doi.org/10.1016/j.atmosenv.2006.03.020, 2006.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A.: kernlab - An S4 Package for Kernel Methods in R, Journal of Statistical Software, Articles, 11, 1–20, https://doi.org/10.18637/jss.v011.i09, 2004.

Kim, J., Shusterman, A. A., Lieschke, K. J., Newman, C., and Cohen, R. C.: The BErkeley Atmospheric CO_2 Observation Network: field calibration and evaluation of low-cost air quality sensors, Atmospheric Measurement Techniques, 11, 1937–1946, https://doi.org/10.5194/amt-11-1937-2018, 2018.

Lewis, A. and Edwards, P.: Validate personal air-pollution sensors, Nature, 535, 29–31, https://doi.org/10.1038/535029a, 2016.

Martin, F., Fileni, L., Palomino, I., Vivanco, M. G., and Garrido, J. L.: Analysis of the spatial representativeness of rural background monitoring stations in Spain, Atmospheric Pollution Research, 5, 779 – 788, https://doi.org/10.5094/APR.2014.087, 2014.

Mead, M., Popoola, O., Stewart, G., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J., McLeod, M., Hodgson, T., Dicks, J., Lewis, A., Cohen, J., Baron, R., Saffell, J., and Jones, R.: The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks, Atmospheric Environment, 70, 186 – 203, https://doi.org/10.1016/j.atmosenv.2012.11.060, 2013.

Meinshausen, N.: Quantile Regression Forests, Journal of Machine Learning Research, 7, 983–999, 2006.

Mentch, L. and Hooker, G.: Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests, Journal of Machine Learning Research, 17, 1–41, 2016.

Mijling, B., Jiang, Q., de Jonge, D., and Bocconi, S.: Practical field calibration of electrochemical NO_2 sensors for urban air quality applications, Atmospheric Measurement Techniques Discussions, 2017, 1–25, https://doi.org/10.5194/amt-2017-43, 2017.

Mueller, M., Hasenfratz, D., Saukh, O., Fierz, M., and Hueglin, C.: Statistical modelling of particle number concentration in Zurich at high spatio-temporal resolution utilizing data from a mobile sensor network, Atmospheric Environment, 126, 171–181, https://doi.org/10.1016/j.atmosenv.2015.11.033, 2016.

Mueller, M., Meyer, J., and Hueglin, C.: Design of an ozone and nitrogen dioxide sensor unit and its long-term operation within a sensor network in the city of Zurich, Atmospheric Measurement Techniques, 10, 3783–3799, https://doi.org/10.5194/amt-10-3783-2017, 2017.

Pang, X., Shaw, M. D., Lewis, A. C., Carpenter, L. J., and Batchellier, T.: Electrochemical ozone sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air-quality monitoring, Sensors and Actuators B: Chemical, 240, 829 – 837, https://doi.org/10.1016/j.snb.2016.09.020, 2017.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2017.

Rai, A. C., Kumar, P., Pilla, F., Skouloudis, A. N., Sabatino, S. D., Ratti, C., Yasar, A., and Rickerby, D.: End-user perspective of low-cost sensors for outdoor air pollution monitoring, Science of The Total Environment, 607, 691 – 705, https://doi.org/10.1016/j.scitotenv.2017.06.266, 2017.

Righini, G., Cappelletti, A., Ciucci, A., Cremona, G., Piersanti, A., Vitali, L., and Ciancarella, L.: GIS based assessment of the spatial representativeness of air quality monitoring stations using pollutant emissions data, Atmospheric Environment, 97, 121 – 129, https://doi.org/10.1016/j.atmosenv.2014.08.015, 2014.

Sexton, J. and Laake, P.: Standard Errors for Bagged and Random Forest Estimators, Comput. Stat. Data Anal., 53, 801–811, https://doi.org/10.1016/j.csda.2008.08.007, 2009.

Smola, A. J. and Schölkopf, B.: A tutorial on support vector regression, Statistics and Computing, 14, 199–222, https://doi.org/10.1023/B:STCO.0000035301.49549.88, 2004.

Spinelle, L., Aleixandre, M., and Gerboles, M.: Protocol of evaluation and calibration of low-cost gas sensors for the monitoring of air pollution, Technical report EUR 26112 EN, Joint Research Centre, https://doi.org/10.2788/9916, 2013.

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

Spinelle, L., Gerboles, M., Kotsev, A., and Signorini, M.: Evaluation of low-cost sensors for air pollution monitoring, Technical report EUR 28601 EN, Joint Research Centre, https://doi.org/10.2760/548327, 2017.

United Nations: World Urbanization Prospects: The 2014 Revision, Tech. Rep. ST/ESA/SER.A/366, Department of Economic and Social Affairs, Population Division, 2015.

Wager, S., Hastie, T., and Efron, B.: Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife, Journal of Machine Learning Research, 15, 1625–1651, 2014.

WG, E.: Guide to the demonstration of equivalence of ambient air monitoring methods, Tech. rep., EC Working Group on Guidance for the Demonstration of Equivalence, http://ec.europa.eu/environment/air/quality/legislation/pdf/equivalence.pdf, last access: 7 May 2018, 2010.

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

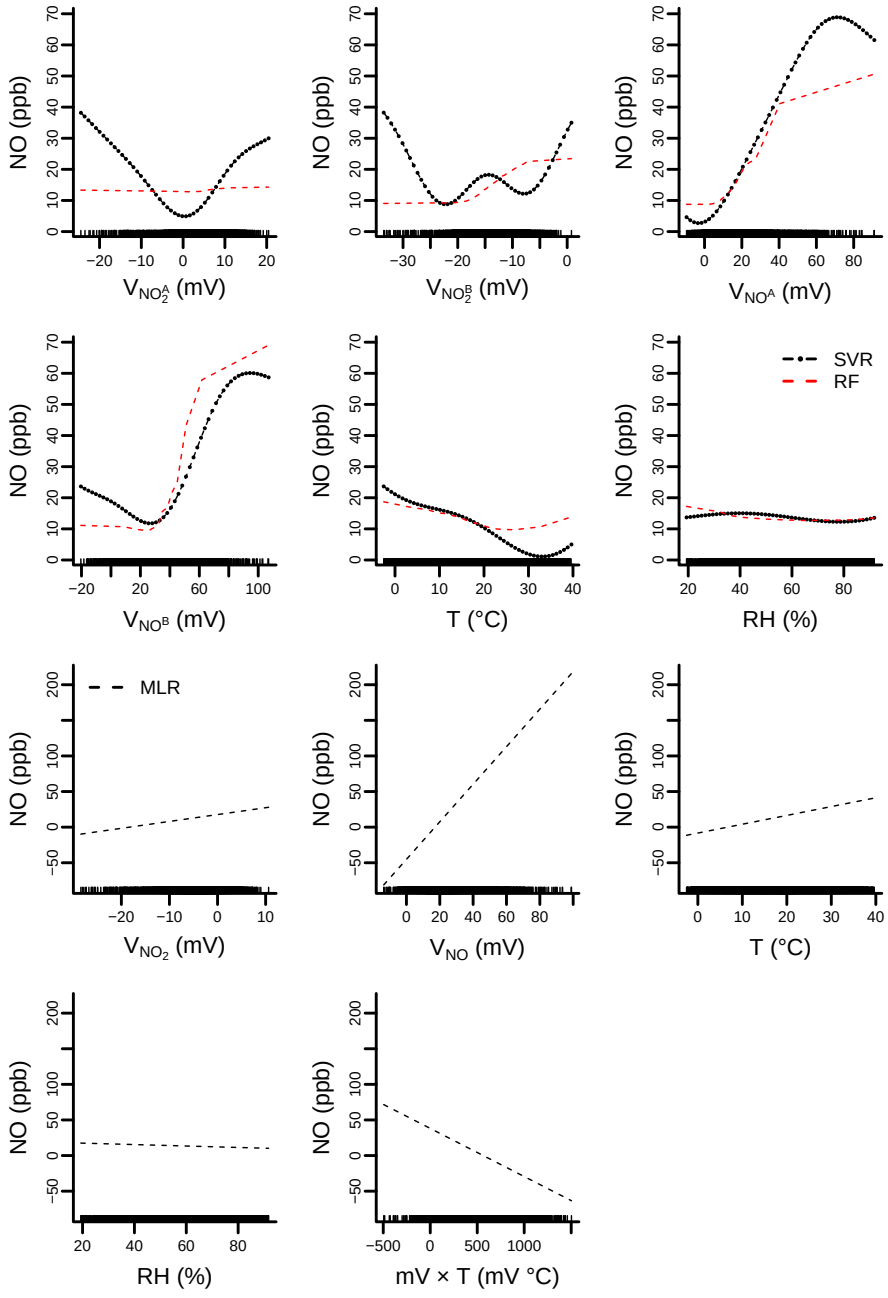

Figure 1. Partial plots for SVR, RF and MLR for the calibration dataset from SU009, NO. Rug on the abscissa indicates the range of the covariate.

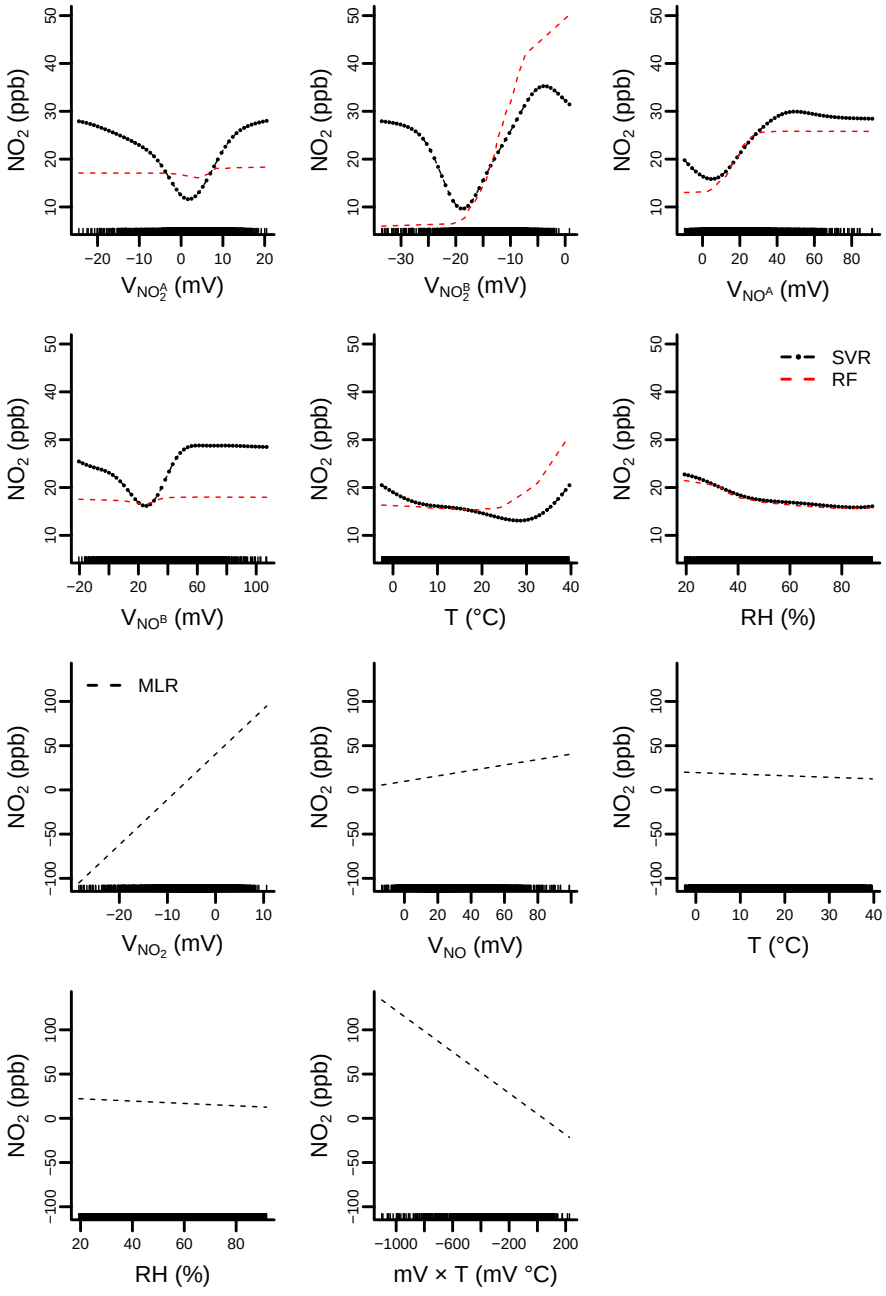

Figure 2. Partial plots for SVR, RF and MLR for the calibration dataset from SU009, NO_2. Rug on the abscissa indicates the range of the covariate.

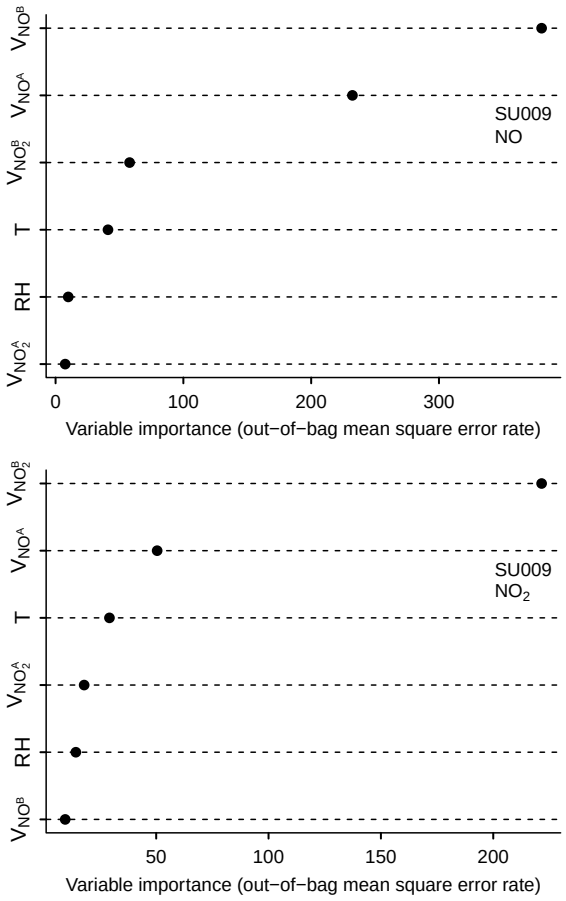

Figure 3. Variable importance plot for the prediction by SU009 of NO (top) and $\mathrm{NO_2}$ (bottom).

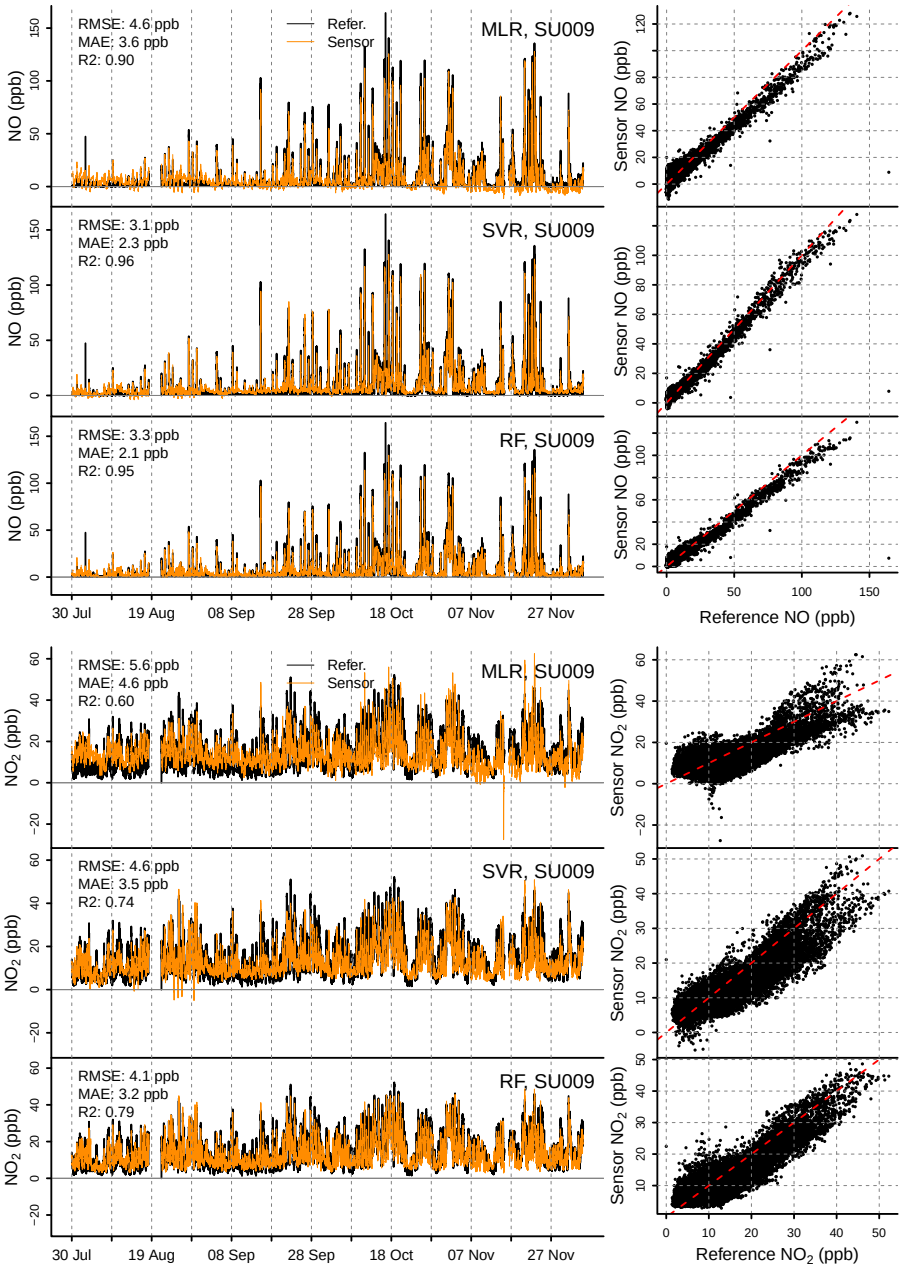

Figure 4. Comparison of NO (top) and NO_2 (bottom) estimates by SU009 with observations by reference instruments. 1:1 red dashed line is added in the scatterplots.

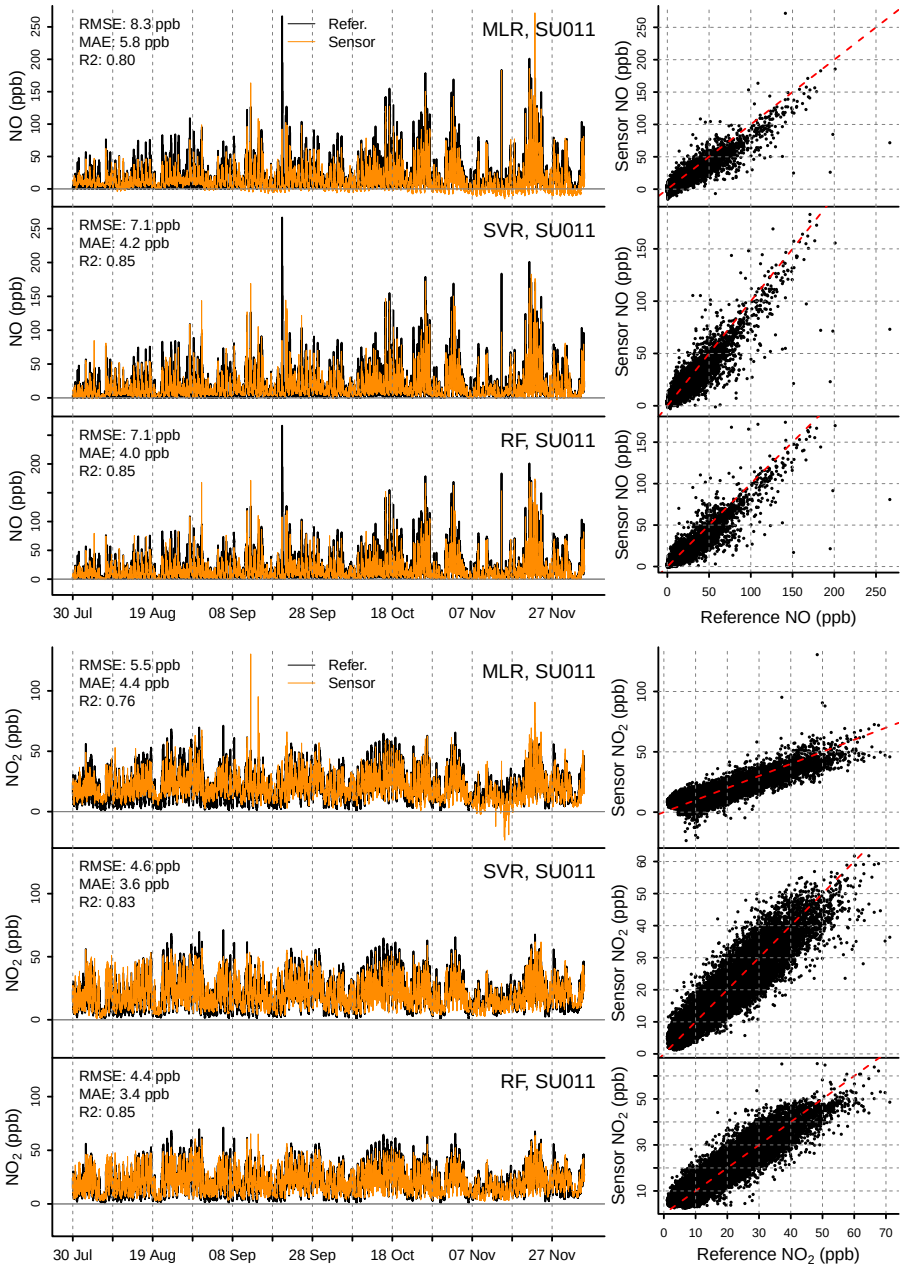

Figure 5. Comparison of NO (top) and NO_2 (bottom) estimates by SU011 with observations by reference instruments. 1:1 red dashed line is added in the scatterplots.

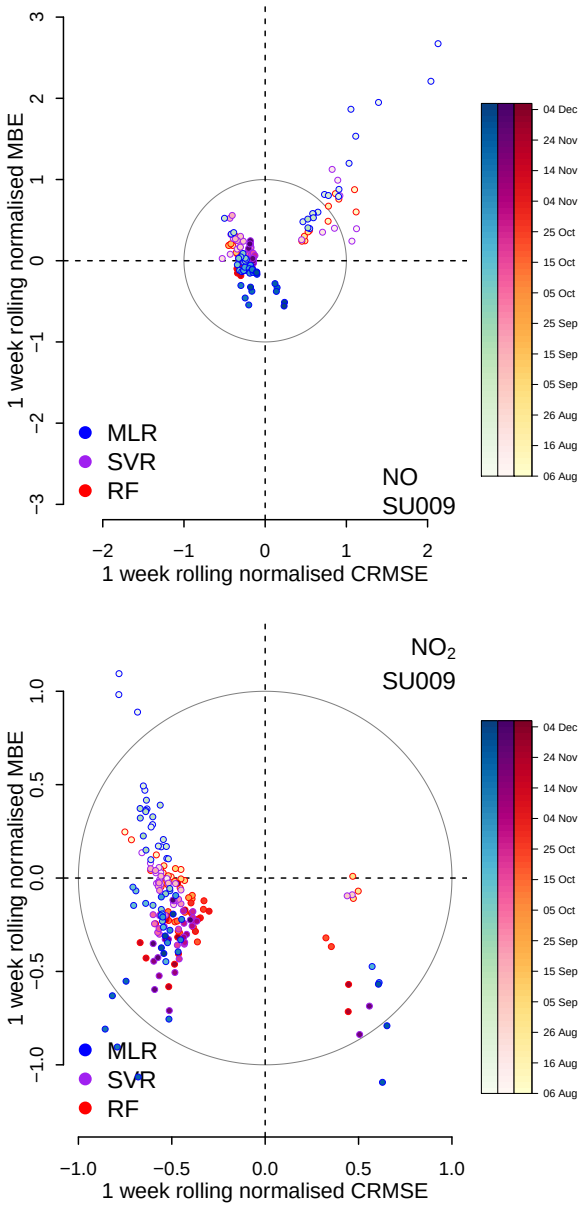

Figure 6. Target plot for timeseries of 1 week rolling goodness-of-fit indexes of NO (top) and NO_2 (bottom) estimate by SU009, in Zurich urban background.

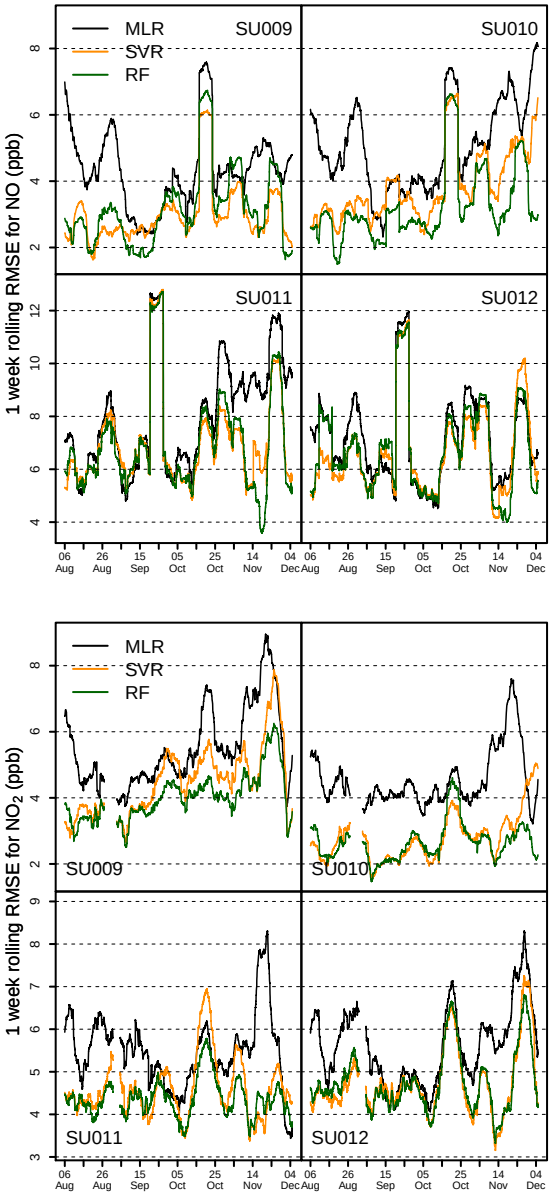

Figure 7. Timeseries of 1 week rolling RMSE for 10 – minute data of NO (top) and NO_2 (bottom) at the deployment sites.

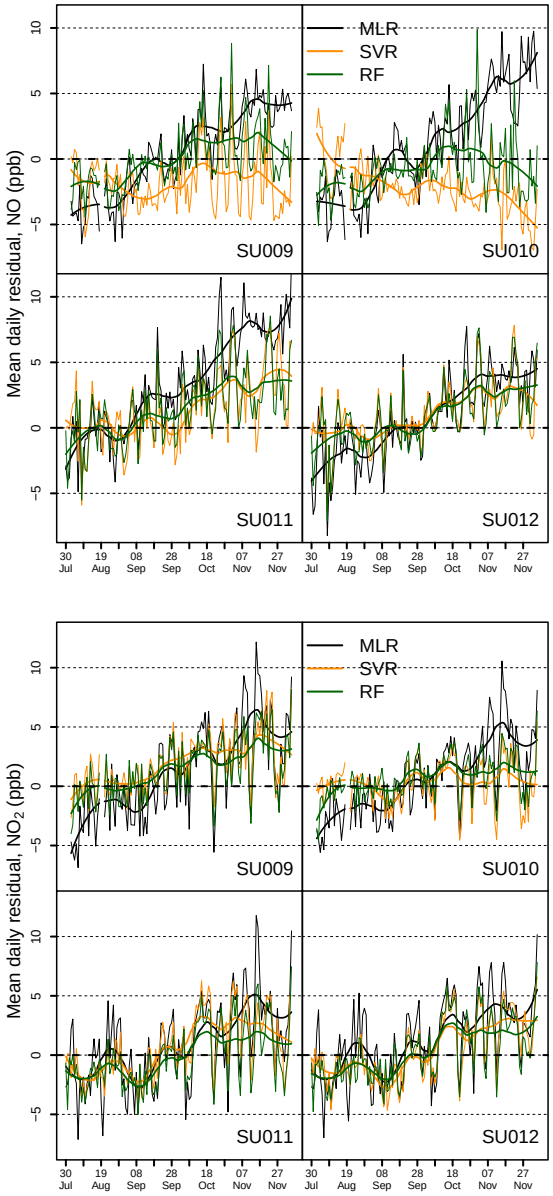

Figure 8. Time series of mean daily residuals for NO and NO_2 estimates, from 1 hour average data. Smooth lines from locally weighted polynomial regression, by `loess` function in R, were added.

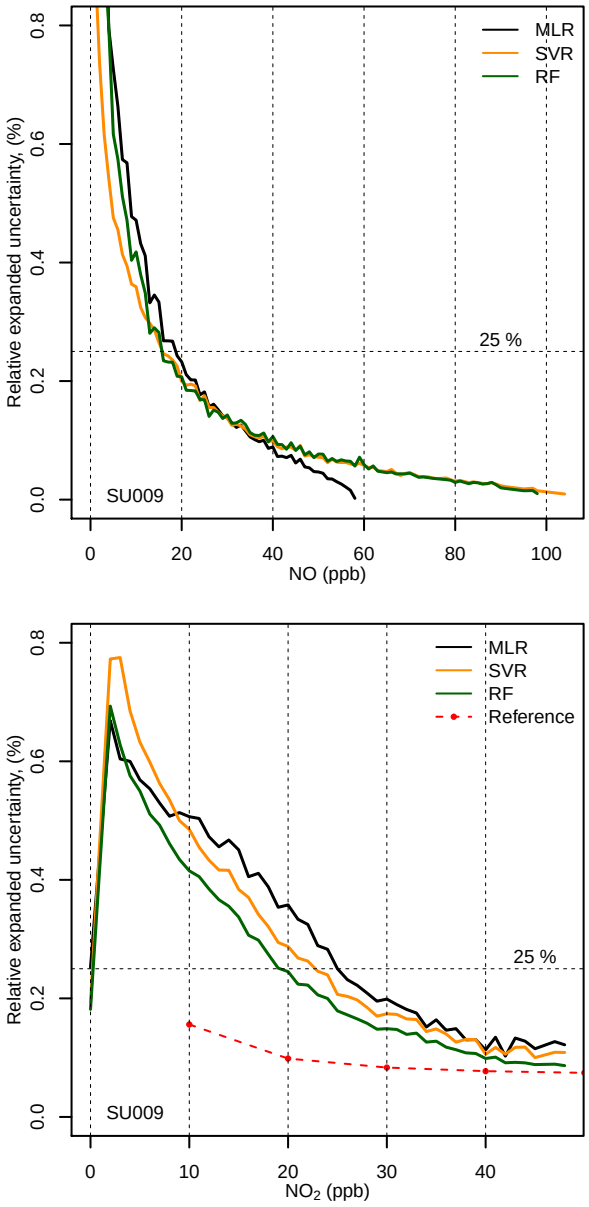

Figure 9. Comparison of expanded relative uncertainty and reference NO and NO_2 concentration for the SU009, as deployed at the urban background site Zurich-Kaserne, using 1 hour average data.

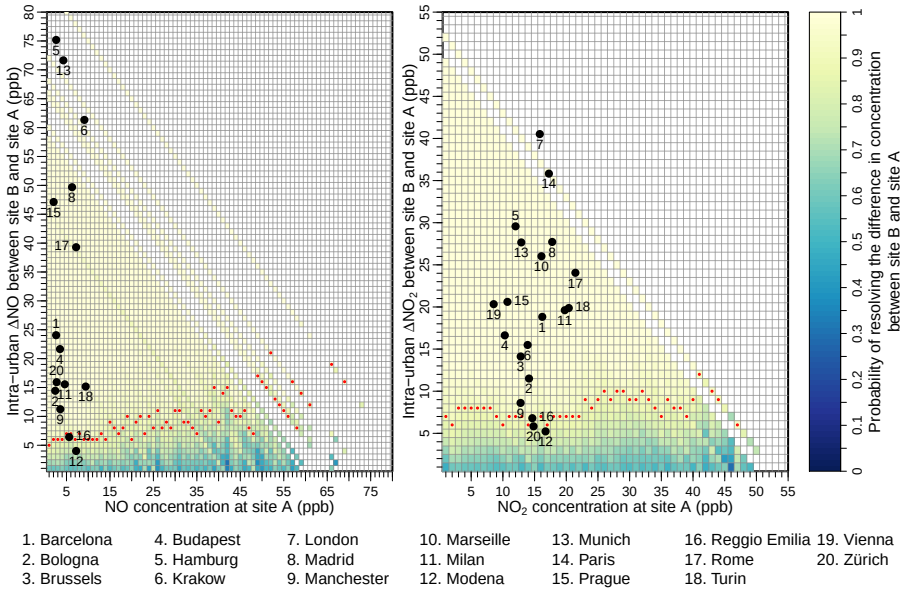

Figure 10. Probability of resolving spatial intra-urban difference in NO and NO_2 between site B and site A, with the latter exposed to lower concentrations. ~~Blue~~ Red dots indicate the concentration difference between site B and A that can be detected with a probability of 90%. Numbered dot coordinates indicate pollution condition for 20 European cities: x coordinate is the urban median concentration, y coordinate is the median intra-urban gradient for hourly concentration data by the air quality monitoring sites within that same urban area.

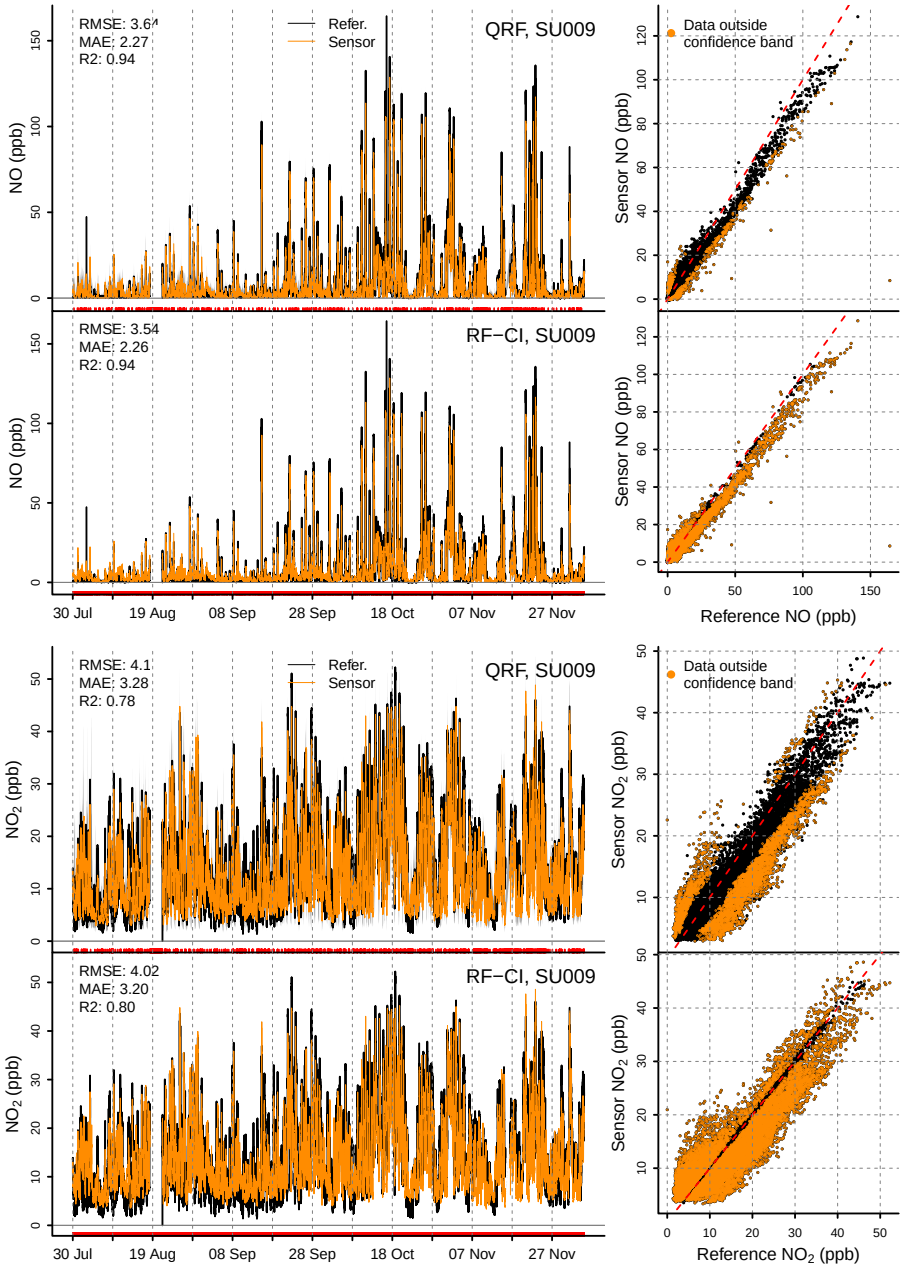

Figure 11. Comparison of QRF and RF-CI estimates of NO (top) and NO_2 (bottom) by SU009 with observations by reference instruments. The grey shaded area indicates either the 5 – 95% quantiles band (QRF case) or the 95% confidence interval (RF-CI case). 1:1 red dashed line is added in the scatterplots.

Appendix A

Equation of the *minimal* model for Multivariate Linear Regression: only EC sensor A for the target pollutant is used.

$$\begin{aligned}\text{NO} &= \beta_0 + \beta_1 V_{\text{NO}^{\text{A}}} + \beta_2 T + \beta_3 \text{RH} + \beta_4 V_{\text{NO}^{\text{A}}} \times T + \epsilon \\ \text{NO}_2 &= \beta_0 + \beta_1 V_{\text{NO}_2^{\text{A}}} + \beta_2 T + \beta_3 \text{RH} + \beta_4 V_{\text{NO}_2^{\text{A}}} \times T + \epsilon\end{aligned} \tag{A1}$$

Equation of the *minimal* model for Support Vector Regression and Random Forest: only EC sensor A for the target pollutant is used.

$$\begin{aligned}\text{NO} &= \text{function}(V_{\text{NO}^{\text{A}}}, T, \text{RH}) \\ \text{NO}_2 &= \text{function}(V_{\text{NO}_2^{\text{A}}}, T, \text{RH})\end{aligned} \tag{A2}$$

Equation of the *basic* model for Multivariate Linear Regression: EC sensors A both for NO and NO_2 are used.

$$\begin{aligned}\text{NO} &= \beta_0 + \beta_1 V_{\text{NO}^{\text{A}}} + \beta_2 V_{\text{NO}_2^{\text{A}}} + \beta_3 T + \beta_4 \text{RH} + \beta_5 V_{\text{NO}^{\text{A}}} \times T + \epsilon \\ \text{NO}_2 &= \beta_0 + \beta_1 V_{\text{NO}^{\text{A}}} + \beta_2 V_{\text{NO}_2^{\text{A}}} + \beta_3 T + \beta_4 \text{RH} + \beta_5 V_{\text{NO}_2^{\text{A}}} \times T + \epsilon\end{aligned} \tag{A3}$$

Equation of the *basic* model for Support Vector Regression and Random Forest: EC sensors A both for NO and NO_2 are used.

$$\begin{aligned}\text{NO} &= \text{function}(V_{\text{NO}^{\text{A}}}, V_{\text{NO}_2^{\text{A}}}, T, \text{RH}) \\ \text{NO}_2 &= \text{function}(V_{\text{NO}^{\text{A}}}, V_{\text{NO}_2^{\text{A}}}, T, \text{RH})\end{aligned} \tag{A4}$$

Equation of the *single replicate* model for Multivariate Linear Regression: $\overline{V_{\text{NO}}}$ indicates the mean net voltage produced by the twin EC sensors for NO, $\overline{V_{\text{NO}_2}}$ indicates the net voltage produced by the two EC sensor for NO_2.

$$\begin{aligned}\text{NO} &= \beta_0 + \beta_1 \overline{V_{\text{NO}}} + \beta_2 T + \beta_3 \text{RH} + \beta_4 \overline{V_{\text{NO}}} \times T + \epsilon \\ \text{NO}_2 &= \beta_0 + \beta_1 \overline{V_{\text{NO}_2}} + \beta_2 T + \beta_3 \text{RH} + \beta_4 \overline{V_{\text{NO}_2}} \times T + \epsilon\end{aligned} \tag{A5}$$

Equation of the *single replicate* model for Support Vector Regression and Random Forest: either EC sensor A for NO and EC sensor A for NO_2 are used.

$$\begin{aligned}\text{NO} &= \text{function}(V_{\text{NO}^{\text{A}}}, V_{\text{NO}_2^{\text{A}}}, T, \text{RH}) \\ \text{NO}_2 &= \text{function}(V_{\text{NO}^{\text{A}}}, V_{\text{NO}_2^{\text{A}}}, T, \text{RH})\end{aligned} \tag{A6}$$

Equation of the *double replicate* and final model for Multivariate Linear Regression: $\overline{V_{\mathrm{NO}}}$ indicates the mean net voltage produced by the twin EC sensors for NO, $\overline{V_{\mathrm{NO_2}}}$ indicates the net voltage produced by the two EC sensor for $\mathrm{NO_2}$.

$$\begin{aligned}\mathrm{NO} &= \beta_0 + \beta_1\overline{V_{\mathrm{NO}}} + \beta_2\overline{V_{\mathrm{NO_2}}} + \beta_3 T + \beta_4 \mathrm{RH} + \beta_5\overline{V_{\mathrm{NO}}} \times T + \epsilon \\ \mathrm{NO_2} &= \beta_0 + \beta_1\overline{V_{\mathrm{NO}}} + \beta_2\overline{V_{\mathrm{NO_2}}} + \beta_3 T + \beta_4 \mathrm{RH} + \beta_5\overline{V_{\mathrm{NO_2}}} \times T + \epsilon\end{aligned} \tag{A7}$$

Equation of the *double replicate* and final model for Support Vector Regression and Random Forest: either EC sensor A for NO and EC sensor A for $\mathrm{NO_2}$ are used.

$$\begin{aligned}\mathrm{NO} &= \mathrm{function}(V_{\mathrm{NO^A}}, V_{\mathrm{NO^B}}, V_{\mathrm{NO_2^A}}, V_{\mathrm{NO_2^B}}, T, \mathrm{RH}) \\ \mathrm{NO_2} &= \mathrm{function}(V_{\mathrm{NO^A}}, V_{\mathrm{NO^B}}, V_{\mathrm{NO_2^A}}, V_{\mathrm{NO_2^B}}, T, \mathrm{RH})\end{aligned} \tag{A8}$$