Atmospheric Pollution Research 12 (2021) 282–291

Contents lists available at ScienceDirect

Atmospheric Pollution Research

journal homepage: www.elsevier.com/locate/apr

FLA

Assessment of air quality sensor system performance after relocation

Stefano Zauli-Sajani [a,*], Stefano Marchesi [a], Claudia Pironi [b], Carla Barbieri [c], Vanes Poluzzi [d], Annamaria Colacci [a]

[a] Regional Centre for Prevention, Environment and Health, Arpae Emilia-Romagna, Via Giardini, 472/L, 41124, Modena, Italy
[b] Environmental Prevention - West Area, Arpae Emilia-Romagna, Via Spalato, 2, 43123, Parma, Italy
[c] Environmental Prevention - Central Area, Arpae Emilia-Romagna, Viale Fontanelli, 23, 41121, Modena, Italy
[d] Regional Centre for Air Quality, Arpae Emilia-Romagna, Largo Caduti del Lavoro, 6, 40122, Bologna, Italy

ARTICLE INFO

Keywords:
Air quality
Electrochemical sensors
Field calibration
Random forest
Performance evaluation

ABSTRACT

Sensor based monitoring systems have been indicated as a promising tool to increase information on spatio-temporal distribution of air pollution but several issues have been raised about the accuracy of such monitors when used in the field. The study aimed at assessing the performance of sensor based systems after multiple relocation in different seasons and sites. The systems included electrochemical sensors to measure NO_2 and O_3 concentrations. The approach consisted in two-week field calibration of each device at a reference monitoring station and the test of the calibrated device at different reference station sites. The main specific goal was a comparison of sensor performance considering site-specific (SS) and no site-specific (no_SS) calibration, i.e. calibration and testing carried out or not at sites with similar characteristics. Calibration was performed by season using random forest (RF) models. Very good performance was found for calibrated O_3 sensors with $R^2 \geq 0.82$ regardless of seasons and sites. Mean normalized root mean square error (nRMSE) was around 7% and 6% in winter and summer tests, respectively. Very good performance of sensor systems was observed also for NO_2 during winter ($R^2 \geq 0.84$) with much better accuracy for SS compared to no_SS calibration (nRMSE equal to 6% and 17%, respectively). A marked decrease of performance was observed for NO_2 sensors during summer. Our results show a good potential of sensor based systems after SS field calibration in increasing information on the distribution of air pollution at high spatial and temporal resolution.

1. Introduction

Air pollution is associated with a range of diseases, symptoms and conditions that impair health (WHO, 2013). Despite a significant decrease in air pollutant concentrations observed in many developed countries over the last decades, air quality still represents a major public concern particularly for people living in urban and suburban areas as they are relatively more exposed (EEA, 2018).

The Air Quality Directive 2008/50/EC for European Union and similar legislation in other countries define the criteria for air quality monitoring and the reference measurement methods that States and environmental agencies shall apply when monitoring air quality. The primary role in air quality assessment is assigned to fixed-site stations equipped with reference grade monitors, but their high costs of installation and maintenance allow only for a relatively sparse monitoring, which is not anymore adequate to meet the increasing needs and demand of detailed air quality information.

In order to increase information on spatio-temporal distribution of air pollution, supplementary techniques have been proposed. In particular, during the last decade also regulatory bodies such as EU Commission and US EPA (Karagulian et al., 2019; Spinelle et al., 2013; Williams et al., 2014, 2019) have recognized the importance of new monitoring technologies based on different type of sensors. Air quality sensors have produced such high expectations to the point of making some researchers talk about "paradigm shift of air pollution monitoring" (Snyder et al., 2013). The underlying idea is that a cost-effective approach for air quality monitoring would be the implementation of mixed networks involving both reference-grade monitors as well as emerging sensor technologies (Cao et al., 2020; Mead et al., 2013). Sensor use may affect a wide range of possible applications including high resolution spatial mapping and hot-spot identification (Gulia et al., 2020), emergency intervention, near-source monitoring (Kanabkaew et al., 2019), mobile and personal monitoring (Duvall et al., 2016; Jovašević-Stojanović et al., 2015; Park et al., 2020). Epidemiological

* Corresponding author.
E-mail address: szauli@arpae.it (S. Zauli-Sajani).

https://doi.org/10.1016/j.apr.2020.11.010
Received 18 August 2020; Received in revised form 30 October 2020; Accepted 17 November 2020
Available online 24 November 2020

studies would also greatly benefit from more detailed information in terms of exposure assessment (Larkin and Hystad, 2017). The main advantages are related to the fact that sensor systems are small and portable, apparently easy to use and deploy, and often low cost or at least much less expensive than reference instruments.

However, despite the high expectations and potential pros, research and regulatory bodies have raised several issues related to the accuracy of such monitors including problems of stability, cross-sensitivity, repeatability and reproducibility (Castell et al., 2017; Morawska et al., 2018). Main current calibration solutions from the manufactures are limited to sensor testing in the laboratory under controlled conditions. This approach often provides unsatisfactory results during ambient air monitoring making the field calibration a necessary phase when using air quality sensors (Kelly et al., 2017; Mijling et al., 2018; Spinelle et al., 2017). However, also field calibration is associated with problems related to generalizability of specific calibration parameters and models.

The issue is relevant when sensors are calibrated and used at the same place because of the limited range of environmental conditions experienced during the calibration period. But the problem becomes significantly more important when sensors are calibrated in a place and used for monitoring campaigns in other places. Indeed, calibration parameters may be site-specific with an additional possible influence of sensor handling and transport. Unfortunately, while multiple relocation is a key characteristic of most the air quality sensors system applications in the real world, almost all published studies analyzed sensor performance considering calibration and testing at the same site. In this manuscript, we investigated the performance of sensor based monitoring systems after relocation. The assessment was based on multiple tests, each including field calibration of the sensor at one site using data from a reference monitoring station and the test of the performance of the calibrated sensor at other locations of the air quality monitoring network of the Emilia-Romagna region. We calibrated and tested NO_2 and O_3 sensors in sites with different characteristics and in different seasons. Main specific goal was related to compare performance of sensors considering site-specific (SS) and no site specific (no_SS) calibration, i.e. calibration carried out or not at sites with similar characteristics of those of the relocation sites.

2. Methods

2.1. Sensor system

The monitoring campaigns were carried out with three units (pods) of a commercial sensor system called AQMesh. AQMesh pods (Environmental Instruments Ltd., UK) are portable, small (22 cm 16 cm × 20 cm) and light (<2 kg) instruments consisting of one box that includes a NO_2 sensor (Alphasense NO_2-B43F), a NO sensor (Alphasense NO-B4), a O_3 sensor (Alphasense OX-B431) and three solid state sensors for temperature (T), relative humidity (RH) and atmospheric pressure (P). Electrochemical sensors provide concentration of the gases by measuring the positive or negative currents generated by reactions between the gases and an electrode. In particular, the core of electrochemical sensors consists of a "working" electrode, a "counter" electrode, and usually a "reference" electrode along with a liquid electrolyte. Gas penetrates through a membrane inside the sensor housing. Once the gas reaches the working electrode, an electrochemical oxidation or reduction (for NO_2 and O_3) reaction occurs. Oxidation causes the flow of electrons from the working electrode to the counter electrode through an external circuit while reduction is associated to the flow of electrons from the counter electrode to the working electrode. The electric current is proportional to the concentration of gas and the external circuit detect and amplifies this current.

A lithium metal battery provides power supply for about two years and a data transmission module sends the collected data to a server via a GSM connection. An IP65 enclosure guarantees protection from water and dust as well as other meteorological agents. Data from gas sensors are post-processed by a proprietary algorithm of the manufacturing company, aiming at correcting the effect of temperature, relative humidity and cross-interferences due to the presence of other gaseous pollutants. The data can then be downloaded from a dedicated website. The instruments have been widely used across the world and are the backbone of several high-resolution real time air pollution monitoring networks such as that developed within the Breathe London project (EDF Europe, 2020).

2.2. Measurement sites and reference instruments

The measurement campaigns were performed in the province of Parma and Modena (Fig. 1). The area is located in the southern part of the Po plain, northern Italy, and is affected by high levels of air pollution (Bigi and Ghermandi, 2016). The main urban areas of Parma and Modena have about 197,000 and 186,000 inhabitants (2018), respectively, and are at a distance of about 50 km. Sensor systems were calibrated and tested for accuracy using six fixed site monitoring stations belonging to the air quality monitoring network of the Regional Agency for Environment, Prevention and Energy of Emilia-Romagna ("Arpae"). Following European Decision 2011/850/EU, the classification of the stations was based on two indicators: "type of area" (urban, suburban, rural) and "type of station" (background, traffic, industrial). Therefore the expression "sites with similar characteristics" used throughout this paper must be interpreted as sites belonging to the same EU classification. An urban background station (UB) and an urban traffic station (UT) were available in each city. UB stations are located in large urban public parks while UT stations are located close to trafficked streets. Sensor calibration and testing included also the two suburban background stations (SB) available in each province: 1) Colorno (about 9000 inhabitants) at 13 km from Parma; and 2) Carpi (about 67,000 inhabitants) at 20 km from Modena. SBstations are located in residential areas with low traffic. The exact location of all monitoring sites are reported in Table 1S, Supplementary Information). The stations are equipped with reference-grade analysers for continous measurement, i.e. chemiluminescence for Nitrogen Oxides (EN14211:2012) and UV absorption for Ozone (EN14625:2012). The monitoring stations work under the EN 9001:2015 quality assurance protocol. The AQMesh pods were installed on the roof of the fixed site monitoring stations.

2.3. Measurement campaigns

All pods were co-located for 1 month at the same fixed site stations to assess agreement among the sensor systems. The winter co-location campaign started on 14 Dec 2018 ending on 14 Jan 2019 while the summer one started on 1 August 2019 ending on 1 September 2019. For the winter season the co-location site was next to the traffic station while for the summer season the co-location site was next to an urban background station.

The other monitoring campaigns were conducted in the period January-August 2019 following the scheme sketched out in Fig. 2. The monitoring periods are specified in Table 1. Three campaigns were carried out during the Jan-Feb period (from now on "winter season") while the other three campaigns during the Jun-Aug period (from now on "summer season"). The sequence of relocation was as follows: 1) all pods initially co-located at the UT site in Parma; 2) each pod placed at a different type of site in the province of Parma (UT, UB, SB); 3) each pod moved at the corresponding type of site in the province of Modena; 4 and 5) each pod placed again at the same type of sites in the province of Parma and Modena, respectively; 6) all pods co-located at the UB site in Parma. Each campaign lasted 14 days. The performance analysis considered only 14-days periods and therefore only half of the co-location periods.

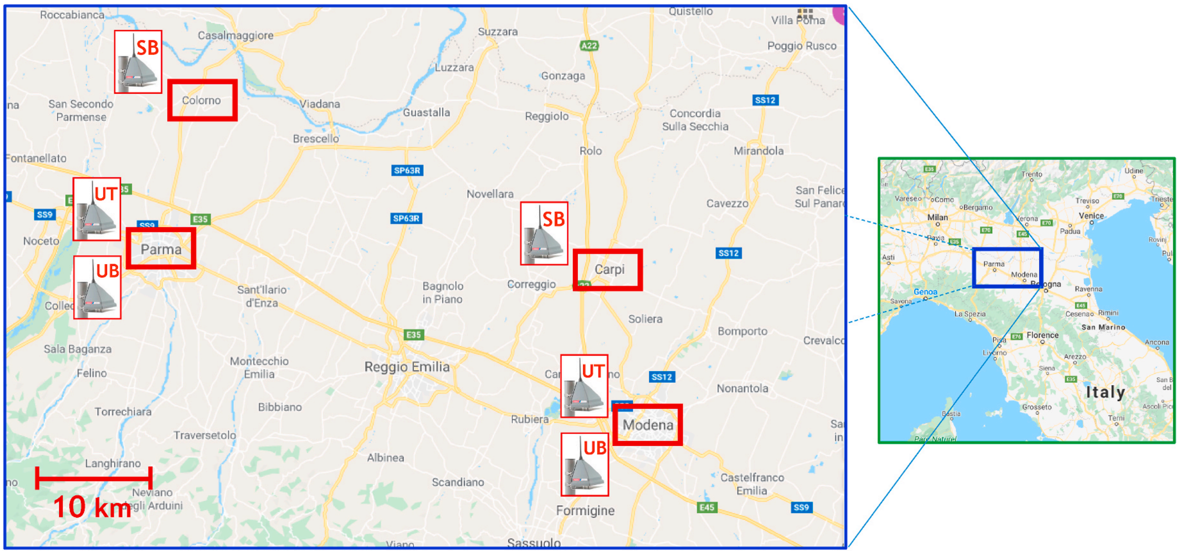

Fig. 1. Location of the monitoring sites. UB= Urban background station. UT= Urban traffic station. SB= Suburban traffic station.

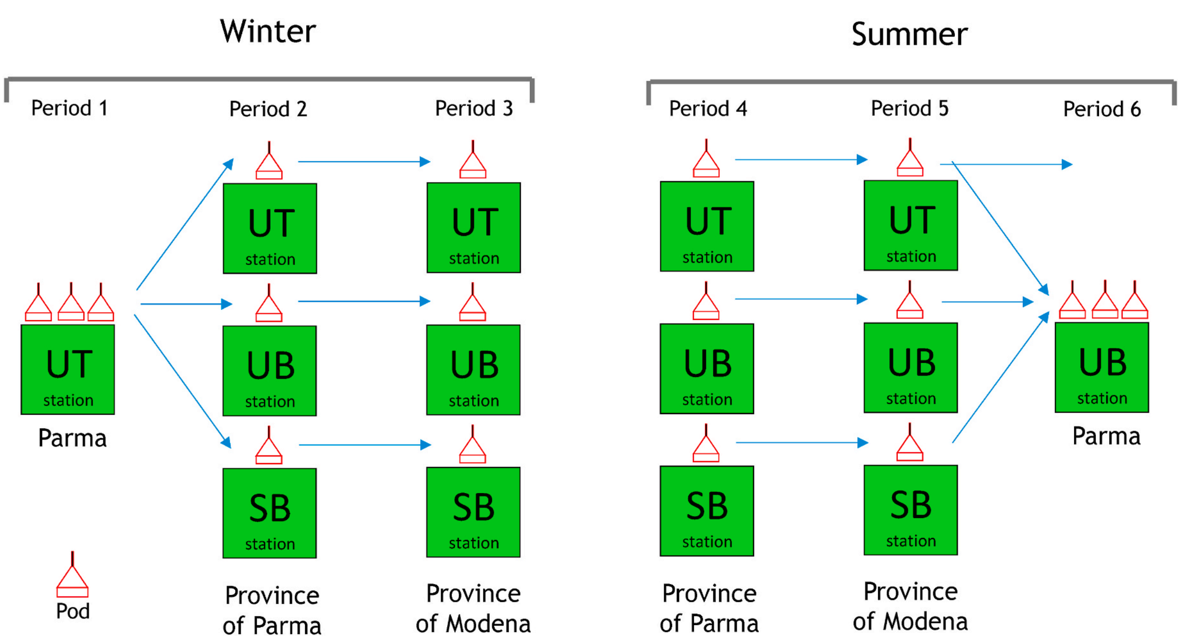

Fig. 2. Pod relocation phases for the winter and summer season.

2.4. Calibration models and metrics for performance evaluation

Several algorithms have been tested in literature for sensor calibration spanning from univariate or multivariate linear regression models (Mijling et al., 2018) to machine learning techniques such as Artificial Neural Networks (Esposito et al., 2016; Spinelle et al., 2015), Support Vector Machine (Bigi et al., 2018; De Vito et al., 2018), Random Forest (Zimmerman et al., 2018) and Hybrid models (Malings et al., 2019). Calibration procedures involving non-linear methods generally outperform those using classical statistics, and better capture the effects of environmental factors on sensor response. Based on recent literature findings (Bigi et al., 2018; Zimmerman et al., 2018) and after some preliminary analysis, a hybrid random forest - linear regression model was adopted in this study. A RF model is a Supervised Learning algorithm which uses ensemble learning method for solving regression or classification problems (Breiman, 2001). It constructs a multitude of decision trees using a bootstrapped random sample from the training data set (in our case training data set corresponds to calibration data set). Any single tree is split into sub-nodes by considering a random subset of the input variables according to which is the strongest predictor of the response. The final output of the random forest model is the average of the prediction from each tree. The trained model is then used to make predictions on new input data.

The most critical limitation of random forest model is the difficulty in predicting new outcomes outside the domain of input variables in the training dataset. For this reason, following Zimmerman et al., (2018) the standard random forest model was modified to become a hybrid random forest - linear regression model. In particular, when air pollutant concentration in the testing dataset exceeded the 98° percentile of the concentration measured by the reference monitor during the corresponding training period, a linear model was used instead of the random forest model. The parameters of the linear model were calculated using the 30% highest concentrations of the training dataset. This approach should combine the pros of the random forest model, i.e. its ability to

Table 1
Overview of mean, minimum and maximum values of O_3 and NO_2 concentrations measured by reference stations during the six monitoring periods used to test sensor performance. Temperature (T) and relative humidity (RH) data measured by the PODs are also reported.

			NO_2 (μg/m^3)	O_3 (μg/m^3)	T (°C)	RH (%)
Period	Province	Ref. station type	Mean (Min - Max)	Mean (Min - Max)	Mean (Min - Max)	Mean (Min - Max)
Winter						
Period 1 (1-14 Jan 19)	Parma	Urban Traffic (UT)	48.4 (6 - 122)		3.4 (-3.7 - 16.1)	78.2 (31.4 - 96.7)
Period 2 (15-28 Jan 19)	Parma	Urban Traffic (UT)	49.5 (10 - 121)		3.2 (-2.7 - 11.1)	78.2 (43.1 - 92.6)
	Parma	Urban Background (UB)	34.6 (7 - 74)	13.6 (0 - 56)	3.2 (-3.2 - 15.1)	78.4 (34.8 - 94.4)
	Parma	Suburban Background (SB)	27.5 (5 - 51)	11.1 (0 - 49)	2.4 (-5.3 - 12)	82.9 (48 - 95.1)
Period 3 (6-19 Feb 19)	Modena	Urban Traffic (UT)	62.8 (14 - 176)		7.7 (0.2 - 18.8)	63.4 (32.2 - 87.3)
	Modena	Urban Background (UB)	45.8 (11 - 133)	28.3 (2 - 94)	6 (-3.1 - 19.1)	69.3 (27 - 93.3)
	Modena	Suburban Background (SB)	33.9 (6 - 84)	19.4 (4 - 99)	6.5 (-0.9 - 17.4)	69.6 (33.9 - 90.5)
Summer						
Period 4 (24 Jun - 7 Jul 19)	Parma	Urban Traffic (UT)	24.4 (6 - 107)		30.6 (20.2 - 41.6)	54.8 (33.7 - 79.8)
	Parma	Urban Background (UB)	11.3 (2 - 36)	106.3 (32 - 204)	28.6 (19.3 - 39.1)	57.2 (32.5 - 81.5)
	Parma	Suburban Background (SB)	12.3 (3 - 47)	98 (10 - 224)	29.7 (18.4 - 41.3)	56.3 (27.1 - 88.8)
Period 5 (15-28 Jul 19)	Modena	Urban Traffic (UT)	33.7 (6 - 111)		27.3 (17.2 - 41)	59.5 (34.2 - 90.5)
	Modena	Urban Background (UB)	17.5 (1 - 60)	80.7 (6 - 172)	26.5 (15.4 - 41.3)	61.4 (30.8 - 89.5)
	Modena	Suburban Background (SB)	19.4 (4 - 58)	83.5 (8 - 182)	26.8 (15.7 - 40.8)	61.8 (27.4 - 90.6)
Period 6 (1-14 Aug 19)	Parma	Urban Background (UB)	9.1 (2 - 28)	79.4 (17 - 151)	26.4 (16.1 - 36.9)	60.4 (34.2 - 81)

capture complicated nonlinear relationships between various inputs and the target output, with the ability of a simple linear model to extrapolate beyond the set of data on which the model is trained. A random forest - linear model was built for each calibration dataset including as input variables 1-h averaged raw sensor gas concentrations (NO, NO_2 and O_3), T, and RH. Calibration models were constructed only for the NO_2 and O_3 sensors, which are the most important from a regulatory point of view. The hyper-parameters were individually tuned for each sensor unit and each pollutant. To maximize utilization of the training data set, a k-fold cross-validation approach was used. A k-fold cross-validation divides the data into k equal-sized groups (where k is specified by the user), and k repeats are used to tune the model. This helps to minimize bias in training data selection when predicting new data and ensures that every point in the training window is used to build the model. 5-fold cross validation with 3-repeats was used for all calibration models. The data collected during each monitoring campaign (i.e. the data collected with each pod in each period in each site) were coupled one by one with the data from the other monitoring campaigns carried out with the same pod in other sites. Each dataset was then used both as calibration and testing dataset. This was done both to increase the number of tests but also because in real world applications is sometimes more convenient to consider calibration phase after the deployment phase (e.g. in emergency interventions). The performance of calibrated sensors were analyzed comparing the set of coupled datasets based on whether calibration and testing dataset were collected at site with similar characteristics (site-specific calibration, from now on SS calibration), or not (no_SS calibration).

All calculation related to calibration procedure and analysis of performance of calibrated pods are performed under R software (R Core Team, 2019). In particular, random forest models were run using "rf" function of the package "caret" (Kuhn, 2008) with the subset of explanatory variables randomly selected at each node ("mtry" parameter) tuned for each pollutant, calibration site and period through RMSE minimization. Metrics to quantitatively compare the calibrated data of pods with the reference monitor in the testing datasets included the coefficient of determination (R^2), the Root Mean Square Error (RMSE) and its normalized value (nRMSE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(POD_i - Ref_i)}{N}}$$

$$nRMSE = \frac{RMSE}{\sigma_{ref}}$$

where N is the number of measurements, POD_i and Ref_i are the i-th values of pod and reference measurements, and σ_{ref} is the standard deviation of reference measurements. T-test for paired data was used to assess the statistical significance of differences between sensor and reference monitor.

Target diagram was also used to show the disentangled contribution of random and systematic errors to the total nRMSE. In particular, we constructed diagrams showing the centered root mean square error (nCRMSE, which is nRMSE corrected for bias) and the normalized mean bias error (nMBE) for each calibration-testing data set. Similarly to the definition of nRMSE, normalization consisted in dividing each statistical indicator by the standard deviation of the data measured by reference monitors. Points with the same distance from the origin have equal nRMSE. The CRMSE is in the left side of the plane if model standard deviation is smaller than the standard deviation of the reference data, and vice versa.

3. Results

Table 1 provides an overview of NO_2 and O_3 concentration measured by reference stations during the monitoring campaigns. In addition, the table reports temperature and relative humidity data as measured by the pods. When the pods were co-located the indicators refer to pod 1 (differences among the pods were within 1% for T and 3% for RH).

The meteorological conditions were quite variable within and among the monitoring campaigns and taken together they were representative of the typical conditions of each season.

The missing data due to sensor failure were a tiny percentage and no data were removed as anomalous from the dataset.

3.1. Intercomparison between pod raw data and reference measurements

The two one-month pod co-location data were used to inter-compare raw data (i.e. pod measurement data before field calibration) and compare them with the data collected by the reference measurements. Very high correlation among pods and with reference monitor were found for O_3 during the summer co-location campaign ($R \geq 0.97$). Very high correlations ($R \geq 0.97$) were also observed between NO_2 sensors for both winter and summer co-location. Very high correlations with reference values were found for all NO_2 sensors during winter ($0.92 \leq R \leq 0.95$) while lower correlations were observed during summer ($0.49 \leq R \leq 0.54$). A further analysis was carried out in terms of mean differences between pod raw data and reference concentrations. Large differences in concentrations were found for both O_3 and NO_2. These mean differences were highly significant ($p < 0.01$) among the pods as well as comparing the pods with reference data. Fig. 3 (panel A and C) shows the differences in mean concentrations for NO_2 (winter co-location) and O_3 (summer co-location). Particularly evident is the difference between the O_3 mean values measured by pod 3 and the other two pods. This behaviour was not associated with some kind of instrument malfunction, as confirmed by the very high levels of correlation with the other pods and by the stability over time of their differences (Fig. 3, panel B). The stability over time of the differences between pods and reference measurements was assessed calculating RMSE over 14-day time periods. Fig. 3 (panel B and D) shows for O_3 (during summer co-location) and NO_2 (during winter co-location) the little variability of RMSE.

3.2. Performance of calibrated O_3 sensors after relocation

Calibrated O_3 sensors showed good performance in all tests (Fig. 4). During summer R^2 values calculated between pods and reference monitors were always higher than 0.91 (Table 2). In all tests nRMSE were less than 7.5% in winter and less than 6.5% in summer. No statistically significant differences in R^2 and RMSE were found between SS and no_SS tests during summer. On the contrary, significant differences were found during winter between SS and no_SS in relation to correlation with mean R^2 increasing from 0.91 to 0.97 while only a small decrease was observed for RMSE (Fig. 5). RMSE were higher in summer than in winter and showed much higher variability across the tests. A small contribution of systematic errors (nMBE) compared to the random errors (nCRMSE) was observed during summer while more similar contribution emerged during winter (Fig. 6).

3.3. Performance of calibrated NO_2 sensors after relocation

The performance scores for NO_2 were different in relation to season. Excellent agreement between calibrated pod data and reference measurements were found during winter (Fig. 7). R^2 values across tests were always higher than 0.84 with mean R^2 increasing from 0.90 for no_SS to 0.93 for SS tests (Fig. 8). A large, highly significant ($p < 0.01$) drop (−70%) was observed for RMSE (Fig. 8 and Table 2) which decreased from 13 μg/m^3 (no_SS tests) to 5 μg/m^3 (SS tests). A marked improvement of performance was also found considering nRMSE which decreased from 17% (no_SS tests) to 6% (SS tests). The target diagram of Fig. 9 provides a visual representation of this marked decrease of nRMSE and highlight that most of this decrease of nRMSE was due to a decrease of normalized MBE, i.e. a decrease of the contribution of systematic errors to the total nRMSE.

A marked decrease of performance was observed for NO_2 sensors during summer. Average R^2 was 0.38 for no_SS tests and 0.50 for SS tests. nRMSE values were much higher than during the winter season (27.7% for no_SS tests and 18.7% for SS tests). The target diagram in Fig. 7 shows the dispersion of data points and the balanced contribution between systematic and random errors.

4. Discussion

4.1. Comparison of results with published studies

We found only a few scientific papers assessing the performance of sensor systems after relocation. Two interesting studies considered

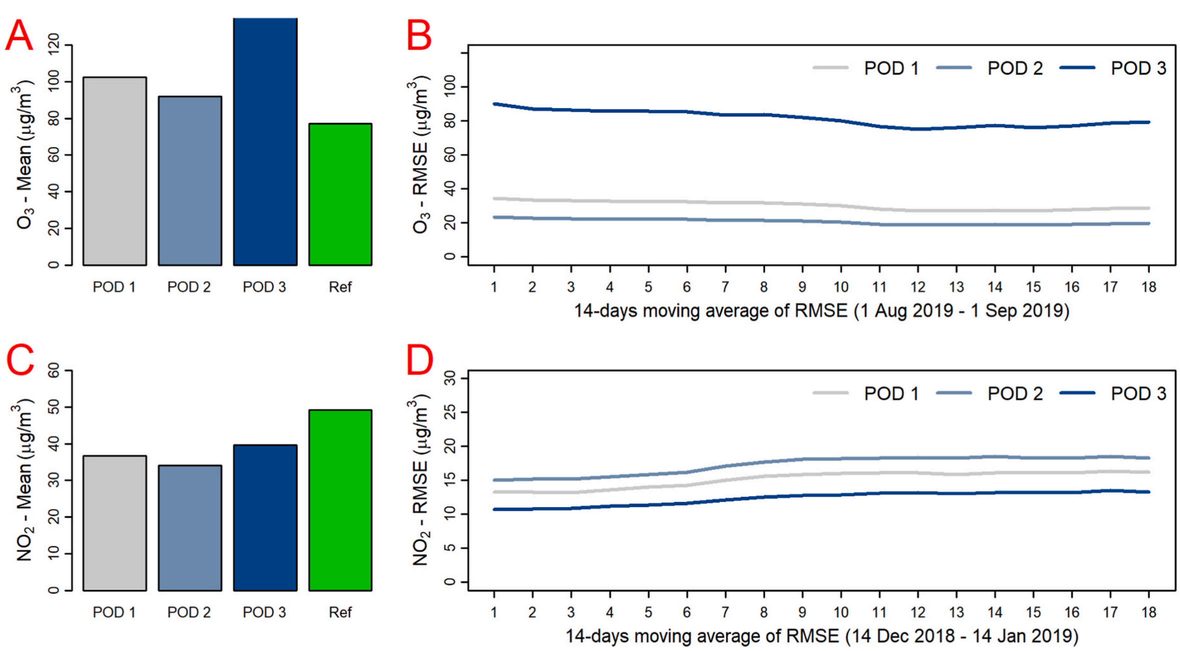

Fig. 3. Mean values of O_3 concentrations (panel A) and 14-days moving average of RMSE (panel B) for the one-month summer co-location monitoring campaign. The same graphs are reported in panel C and D for the one-month NO_2 winter co-location campaign.

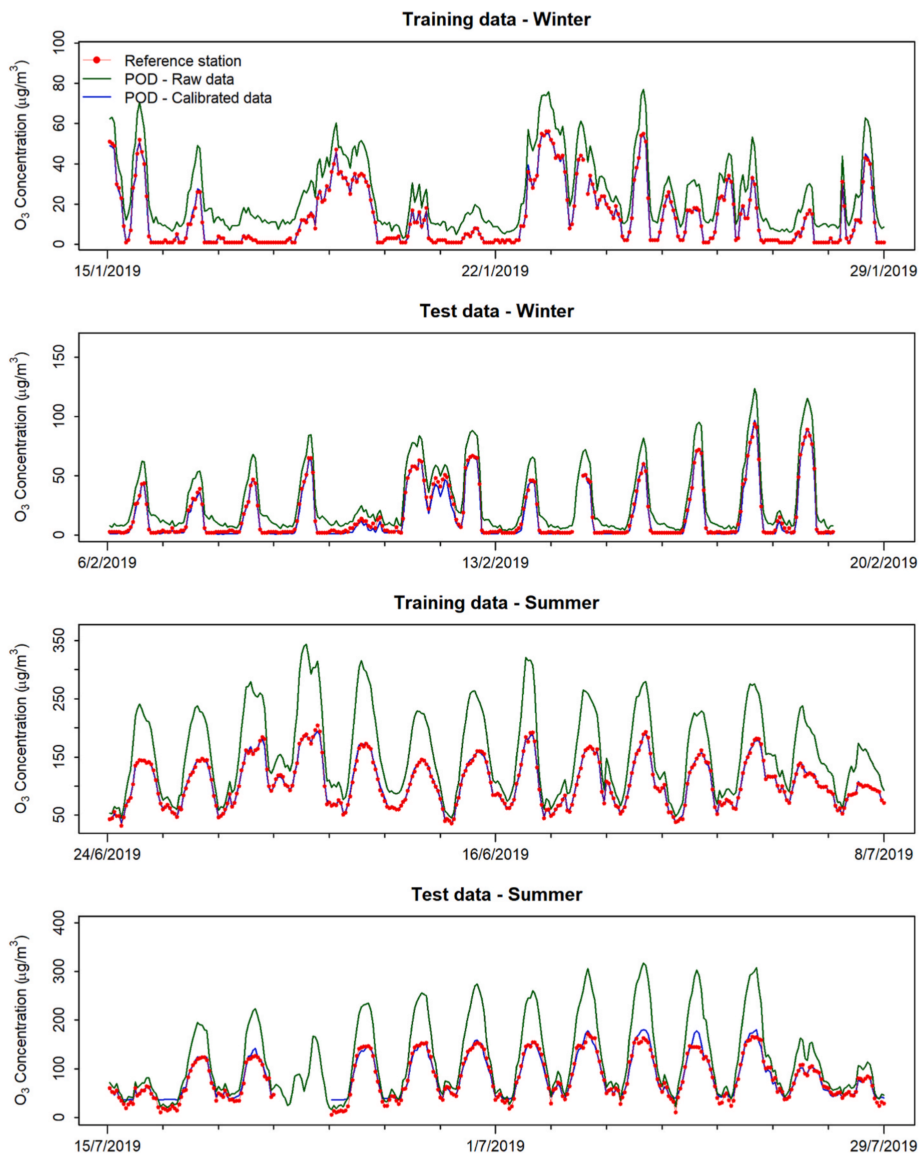

Fig. 4. Examples of comparison of O_3 reference measurements with POD raw and calibrated data. Training data were collected at the urban background station in Parma (Period 2) while the testing data refers to the urban background station in Modena (Period 3)..

Table 2
Overview of mean, minimum and maximum values of R^2, RMSE and nRMSE calculated comparing O_3 and NO_2 measurements from calibrated sensor and reference monitors. SS= site-specific calibration tests. no_SS= no site-specific calibration tests.

–			R^2			RMSE ($\mu g/m^3$)			nRMSE (%)		
			Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
O_3											
	Winter										
		no_SS	0.91	0.82	0.99	5.8	2.2	6.3	7.4	2.4	11.1
		SS	0.97	0.96	0.99	4.5	2.4	6.7	6.8	2.6	12.0
	Summer										
		no_SS	0.95	0.93	0.96	10.6	8.2	16.1	6.4	5.2	7.5
		SS	0.95	0.94	0.97	9.9	6.5	17.0	5.9	4.8	8.0
NO_2											
	Winter										
		no_SS	0.90	0.84	0.95	14.1	9.8	20.2	17.0	10.1	23.7
		SS	0.93	0.84	0.98	5.6	3.7	11.5	6.0	2.7	12.3
	Summer										
		no_SS	0.38	0.19	0.49	9.1	4.5	15.9	27.7	12.2	61.2
		SS	0.50	0.36	0.68	11.1	5.8	21.3	18.7	10.3	31.7

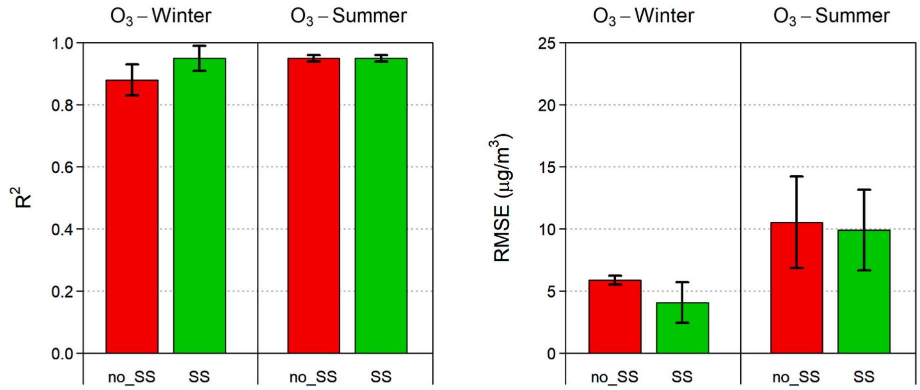

Fig. 5. Performance metrics (R^2, left panel; RMSE, right panel) for O_3 sensor tests. Mean and standard deviation of individual R^2 and RMSE tests are shown.

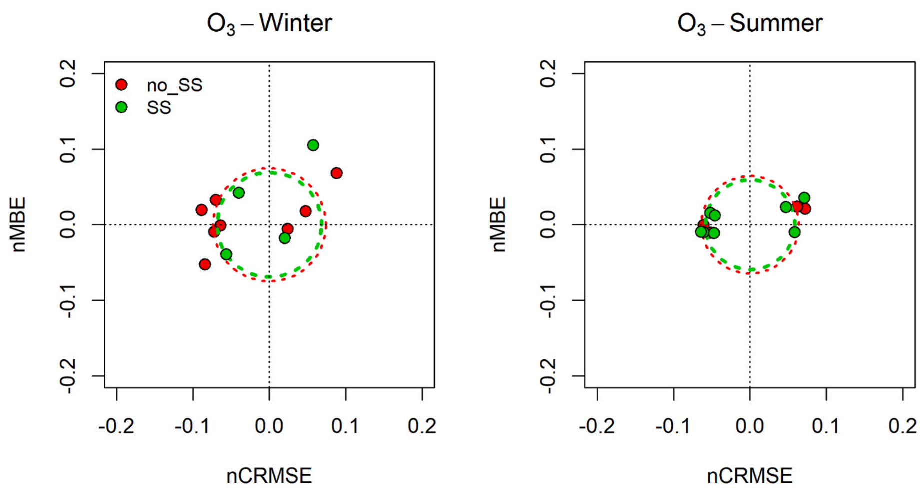

Fig. 6. Target diagrams for O_3 for all testing periods grouped for season, and no_SS and SS tests.

similar statistical approaches for calibration and included relocation tests even though not in a systematic way and without a season specific performance analysis.

Zimmermann et al. (2018) used RF model to calibrate a sensor system at an urban background site (from August 2016 to February 2017) and then moved the system to another site characterized by increased traffic volume. The testing period didn't include summer months (from February to May 2017). They found correlation levels for O_3 equal to 0.92, a bit lower than our findings. For NO_2 the authors reported moderate R^2 values ($R^2 = 0.75$), i.e. lower than our winter findings and significantly higher than our summer ones. The authors reported accuracy in terms of mean absolute error (MAE) showing a value of 6.4 μg/m^3 for O_3 and 9.4 μg/m^3 for NO_2. We calculated this performance indicator with reference to our SS tests obtaining 3.4 μg/m^3 and 9.0 μg/m^3 for O3 in the winter and summer season, respectively, and 4.8 μg/m^3 and 7.4 μg/m^3 for NO_2.

Bigi et al. (2018) analyzed the performance of four systems including NO_2 sensors. The systems were initially deployed for calibration at a suburban site (from April to July 2017) and then relocated at an urban background site (two units) and a traffic site (the other two units). The testing dataset included data from August to early December. R^2 for RF models ranged from 0.79 to 0.91 at the urban background site and from 0.82 to 0.85 at the traffic site. RMSE values ranged from 5.4 μg/m^3 to 9.6 μg/m^3 with slightly worse performance at the traffic site.

The problem of robustness to relocation of calibrated air quality multisensors devices has been also faced in a recent paper by De Vito et al. (2020). Based on data collected with AQMesh pods in different locations and seasons, they found that performance degradation was associated to difference between testing and calibration conditions in terms of probability distribution of relevant sensors drivers (target gas, non target interferents and environmental variables). Their findings are therefore in line with the lower performance observed in out study for no_SS compared to SS tests. It is worth to note that in the study by De Vito et al. (2020) relocation is limited to the urban area of Oslo while the present performance assessment considered relocation at distance up to 50 km.

The role of the type of monitoring location, or land use, in sensor calibration has been further emphasized by recent studies that have investigated the possibility of remote sensor calibration. These techniques do not rely on sensor co-location with regulatory grade instruments but on the use of regulatory grade instruments as proxies based on broad land use categories. (Miskell et al., 2018; Weissert et al., 2020). Remote calibration, while surely promising and more cost-effective than calibration based on co-location, seems at present not guarantee the same level of performance.

4.2. Calibration approach

The calibration approach adopted in the study deserves some discussion. A first point is related to the choice of considering season specific calibration. The comparison with the above mentioned published studies showed that mean performance values found in literature for NO_2 are often lower than our winter findings and higher than our summer ones. Starting from our data and literature findings we can conclude that season specific calibration may represent a pragmatic and effective way to calibrate and deploy sensor systems without carrying

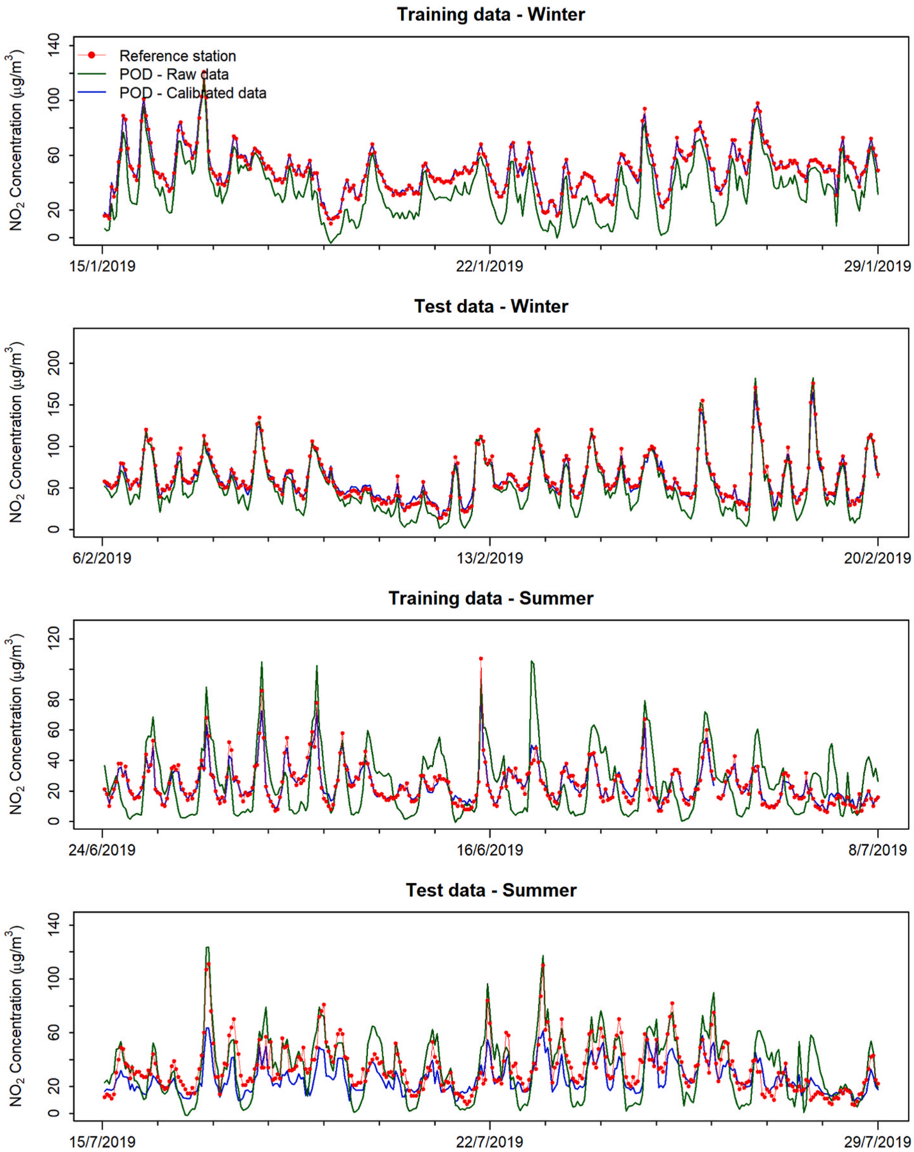

Fig. 7. Examples of comparison of NO_2 reference measurements with POD raw and calibrated data. Training data were collected at the urban traffic station in Parma (Period 4) while the testing data refers to the urban traffic station in Modena (Period 5).

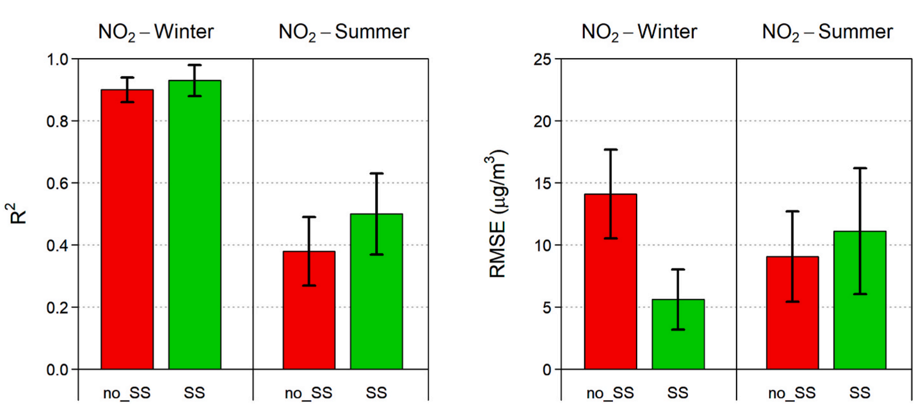

Fig. 8. Performance metrics (R^2, left panel; RMSE, right panel) for NO_2 sensor tests. Mean and standard deviation of individual R^2 and RMSE tests are shown.

out long monitoring campaigns required to train complex non-linear models such as RF or other machine learning models. After all, most of the sensor system applications in the real world are related to short term monitoring campaigns. Spatial mapping (which is one of the most important application of sensor systems) is usually carried out with short term measurement campaigns in particular when associated to Land Use

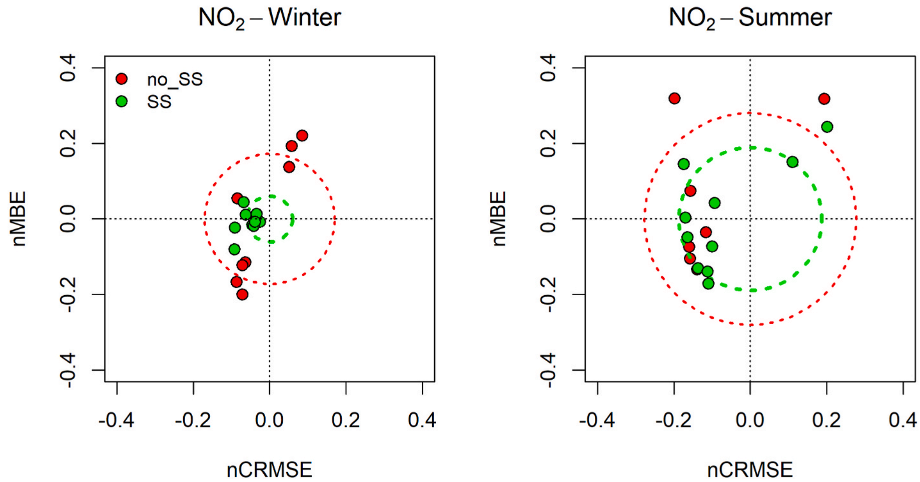

Fig. 9. Target diagrams for NO_2 for all testing periods grouped for season, and no_SS and SS tests.

Regression (LUR) model construction and testing (Kashima et al., 2018; Vizcaino and Lavalle, 2018).

Another point to be discussed is the choice of a two week window for calibration. We verified in some preliminary analysis and also after the conclusion of the monitoring campaigns that calibration models generally converge to quite stable performance values after a few days. For our performance assessment we made an a priori choice of a two-week period because we considered this as a reasonable duration for both calibration and testing in relation to the possible occurrence of anomalous weather conditions not representative of the season. However, from our data and in particular for NO_2 during winter and O_3 during summer, it seems possible to achieve very good performance in testing sensor systems also considering shorter calibration periods.

A final important point that deserves a discussion is related to the time stability of sensor performance and calibration settings. Our study supports the hypothesis of a good stability of the sensors during the same season. Excellent results were obtained with a specific analysis, consisting in the two one-month co-location campaign. If sensor stability would be rigorously proved over longer time periods (e.g. three/four months) this would allow to plan a large number of monitoring campaigns without sensor re-calibration during the same season. While some recent studies using higher-class sensor systems such that used in this study seem to be promising in this regard (Mueller et al., 2017; Zimmerman et al., 2018; Liu et al., 2020), others showed relevant drifts or performance degradation (Mueller et al., 2017), possibly due to both the change over time of response of sensor exposed to the same environmental conditions and to different sensor response in changing environmental conditions.

4.3. *Seasonal variability of sensor performance*

The decline of performances of NO_2 sensor during summer warrants a specific discussion. It is well known that electrochemical sensors may suffer from interference due to changing temperature, relative humidity or other gases that may affect the oxidation-reductions occurring at the working electrode (Mueller et al., 2017). AQMesh pods apply some algorithms to compensate electrochemical sensor response due to temperature, relative humidity and ozone. These algorithms are primarily based on some guidance provided by the Alphasense sensor manufacturing company (Sensor Technology House, UK) that was shown to perform well at temperature under 20 °C but may not lead to gas concentration values of acceptable accuracy (Cross et al., 2017) above 25-30 °C. The non linear RF calibration model adopted in our study, while largely superior to a simple linear model (data not shown) resulted not able to provide appropriate additional corrections. It is worth noting that ambient conditions during our summer campaigns were particularly critical with ozone concentration values up to 204 $\mu g/m^3$ and air temperature up to 41.6 °C, i.e. conditions rarely experienced in previous studies.

It is also important to note that the summer co-location campaign of the pods showed highly correlated time trends among pods, but much less correlated trends between each pod and the reference station. We can therefore conclude that the sensors showed deterministic responses to ambient conditions but low precision and accuracy. This supports the hypothesis that more sophisticated calibration model or longer training dataset may lead to more effective calibration procedures. On the contrary, the use of an array of sensors to measure the same pollutant in each pod, as tested by some authors (Bigi et al., 2018; Williams, 2019), may be not effective in improving sensor system performance. As a matter of fact, the high correlations between different NO_2 sensors found in our study lead to the conclusion that multiple arrays of NO_2 sensors may be not very effective in improving sensor system performance.

5. Conclusions

The study aimed at assessing the performance of sensor based systems after multiple relocation in different seasons and sites. The approach consisted in two-week field calibration of these devices at some reference monitoring stations and the test of the calibrated devices at different sites. Excellent performance was observed for O_3 in all season while NO_2 sensors showed high accuracy and precision during winter but a marked decrease of performance during the warm season. An improvement of sensor performance was found when sensors were calibrated and deployed in sites with similar characteristics (SS calibration). This improvement was especially evident for NO_2 during winter. In conclusion, our results showed a great potential of sensor based systems after SS field calibration to increase the spatial density of air quality monitoring at intra-urban up to regional scale and support exposure assessment studies.

CRediT authorship contribution statement

Stefano Zauli-Sajani: Conceptualization, Methodology, Investigation, Formal analysis, Writing - original draft. **Stefano Marchesi:** Conceptualization, Methodology, Investigation, Writing - review & editing. **Claudia Pironi:** Investigation. **Carla Barbieri:** Investigation. **Vanes Poluzzi:** Methodology, Writing - review & editing. **Annamaria Colacci:** Project administration, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was conducted as part of the AWAIR project (*Environmental integrated, multilevel knowledge and approaches to counteract critical air pollution events, improving vulnerable citizens quality of life in Central Europe functional urban areas*), which was supported and financed by the European Union within the Interreg Central Europe Programme (CE1226).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apr.2020.11.010.

References

Bigi, A., Ghermandi, G., 2016. Trends and variability of atmospheric $PM_{2.5}$ and $PM_{10\text{-}2.5}$ concentration in the Po valley, Italy. Atmos. Chem. Phys. 16, 15777–15788.
Bigi, A., Mueller, M., Grange, S.K., Ghermandi, G., Hueglin, C., 2018. Performance of NO, NO_2 low cost sensors and three calibration approaches within a real world application. Atmos. Meas. Tech. 11, 3717–3735.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Cao, R., Li, B., Wang, Z., Peng, Z.R., Lou, S., 2020. Using a distributed air sensor network to investigate the spatiotemporal patterns of PM2.5 concentrations. Environ. Pollut. 264, 114549. https://doi.org/10.1016/j.envpol.2020.114549.
Castell, N., Dauge, F.R., Schneider, P., Vogt, M., Lerner, U., Fishbain, B., Broday, D., Bartonova, A., 2017. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environ. Int. 99, 293–302.
Cross, E.S., Williams, L.R., Lewis, D.K., Magoon, G.R., Onasch, T.B., Kaminsky, M.L., Worsnop, D.R., Jayne, J.T., 2017. Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements. Atmos. Meas. Tech. 10, 3575–3588.
De Vito, S., Esposito, E., Salvato, M., Popoola, O., Formisano, F., Jones, R., Di Francia, G., 2018. Calibrating chemical multisensory devices for real world applications: an in-depth comparison of quantitative machine learning approaches. Sensor. Actuator. B Chem. 255, 1191–1210.
De Vito, S., Esposito, E., Castell, N., Schneider, P., Bartonova, A., 2020. On the robustness of field calibration for smat air quality monitors. Sensor. Actuator. B Chem. 310, 127869.
Decision 2011 850 EU, 2020. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32011D0850&from=EN. (Accessed 14 August 2020).
Duvall, R.M., Long, R.W., Beaver, M.R., Kronmiller, K.G., Wheeler, M.L., Szykman, J.J., 2016. Performance evaluation and community application of low-cost sensors for ozone and nitrogen dioxide. Sensors 16, 1698.
EDF Europe. Breathe London project. https://www.breathelondon.org/. (Accessed 14 August 2020).
EEA (European Environment Agency, 2018. Air Quality in Europe — 2018 Report, ISBN 978-92-9213-989-6. https://doi.org/10.2800/777411.
EN 14211, 2012. Ambient Air - Standard Method for the Measurement of the Concentration of Nitrogen Dioxide and Nitrogen Monoxide by Chemiluminescence. European Committee for Standardization.
EN 14625, 2012. Ambient Air - Standard Method for the Measurement of the Concentration of Ozone by Ultraviolet Photometry. European Committee for Standardization.
EN 9001, 2015. Quality Management Systems - Requirements. European Committee for Standardization.
Esposito, E., De Vito, S., Salvato, M., Bright, V., Jones, R.L., Popool, O., 2016. Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems. Sensor. Actuator. B Chem. 231, 701–713.
Gulia, S., Prasad, P., Goyal, S.K., Kumar, R., 2020. Sensor-based Wireless Air Quality Monitoring Network (SWAQMN) - a smart tool for urban air quality management. Atmos. Pollut. Res. 11, 1588–1597.
Jovašević-Stojanović, M., Bartonova, A., Topalović, D., Lazović, I., Ristovski, Z., 2015. On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter. Environ. Pollut. 206, 696–704.
Kanabkaew, T., Mekbungwan, P., Raksakietisak, S., Kanchanasut, K., 2019. Detection of PM2.5 plume movement from IoT ground level monitoring data. Environ. Pollut. 252, 543–552.
Karagulian, F., Gerboles, M., Barbiere, M., Kotsev, A., Lagler, F., Borowiak, A., 2019. Review of Sensors for Air Quality Monitoring. EUR 29826 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-09255-1, p. JRC116534. https://doi.org/10.2760/568261.
Kashima, S., Yorifuji, T., Sawada, N., Nakaya, T., Eboshida, A., 2018. Comparison of land use regression models for NO2 based on routine and campaign monitoring data from an urban area of Japan. Sci. Total Environ. 631-632, 1029–1037.
Kelly, K.E., Whitaker, J., Petty, A., Widmer, C., Butterfield, A., 2017. Ambient and laboratory evaluation of a low-cost particulate matter sensor. Environ. Pollut. 221, 491–500.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Software 28, 1–26.
Larkin, A., Hystad, P., 2017. Towards personal exposures: how technology is changing air pollution and health research. Curr. Environ. Health Rep. 4, 463–471.
Liu, X., Jayaratne, R., Thai, P., Kuhn, T., Zing, I., Christensen, B., Lamont, R., Dunbabin, M., Zhu, S., Gao, J., Wainwright, D., Neale, D., Kan, R., Kirkwood, J., Morawska, L., 2020. Low-cost sensors as an alternative for long-term air quality monitoring. Environ. Res. 185, 109438.
Malings, C., Tanzer, R., Hauryliuk, A., Kumar, S.P.N., Zimmerman, N., Kara, L.B., Presto, A.A., Subramanian, R., 2019. Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. Atmos. Meas. Tech. 12, 903–920.
Mead, I., Popoola, O., Stewart, G., Landshoff, P., Calleja, M., Hayes, M., Baldovi, J., McLeod, M.W., Hodgson, T., Dicks, J., Lewis, A.C., Cohen, J., Baron, R., Saffell, J., Jones, R., 2013. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. Atmos. Environ. 70, 186–203.
Mijling, B., Jiang, Q., De Jonge, D., Bocconi, S., 2018. Field calibration of electrochemical NO_2 sensors in a citizen science context. Atmos. Meas. Tech. 11, 1297–1312.
Miskell, G., Salmond, J.A., Williams, D.E., 2018. Solution to the problem of calibration of low-cost air quality measurement sensors in networks. ACS Sens. 3, 832–843.
Morawska, L., Thai, P., Liu, X., Asumadu-Sakyia, A., Ayoko, G., Bartonova, A., Bedini, A., Chai, F., Christensen, B., Dunbabin, M., Gao, J., Hagler, G.S.W., Jayaratne, R., Kumar, P., Lau, A.K.H., Louie, P.K.K., Mazaheri, M., Ning, Z., Motta, N., Mullins, B., Rahman, M.M., Ristovski, Z., Shafiei, M., Tjondronegoro, D., Westerdahl, D., Williams, R., 2018. Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: how far have they gone? Environ. Int. 116, 286–299.
Mueller, M., Meyer, J., Hueglin, C., 2017. Design of an ozone and nitrogen dioxide sensor unit and its long-term operation within a sensor network in the city of Zurich. Atmos. Meas. Tech. 10, 3783–3799.
Park, J., Ryu, H., Kim, E., Choe, Y., Heo, J., Lee, J., Cho, S.-H., Sung, K., Cho, M., Yang, W., 2020. Assessment of PM2.5 population exposure of a community using sensor-based air monitoring instruments and similar time-activity groups. Atmos. Pollut. Res. https://doi.org/10.1016/j.apr.2020.08.010.
R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Snyder, E.G., Watkins, T.H., Solomon, P.A., Thoma, E.D., Williams, R.W., Hagler, G.S., Shelow, D., Hindin, D.A., Kilaru, V.J., Preuss, P.W., 2013. The changing paradigm of air pollution monitoring. Environ. Sci. Technol. 47, 11369–11377.
Spinelle, L., Aleixandre, M., Gerboles, M., 2013. Protocol of Evaluation and Calibration of Low-Cost Gas Sensors for the Monitoring of Air Pollution. European Commission EUR 26112 - Joint Research Centre - Institute for Environment and Sustainability. Publications Office of the European Union - Scientific and Technical Research series, Luxembourg. https://doi.org/10.2788/9916. ISSN 1831-9424 (online), ISBN 978-92-79-32691-2 (pdf).
Spinelle, L., Gerboles, M., Villani, M.G., Aleixandre, M., Bonavitacola, F., 2015. Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: ozone and nitrogen dioxide. Sensor. Actuator. B Chem. 215, 249–257.
Spinelle, L., Gerboles, M., Kotsev, A., Signorini, M., 2017. Evaluation of Low-Cost Sensors for Air Pollution Monitoring: Effect of Gaseous Interfering Compounds and Meteorological Conditions. Publications Office of the European Union. https://doi.org/10.2760/548327.
Vizcaino, P., Lavalle, C., 2018. Development of European NO_2 Land Use Regression Model for present and future exposure assessment: implications for policy analysis. Environ. Pollut. 240, 140–154.
Weissert, L., Miles, E., Miskell, G., Alberti, K., Feenstra, B., Henshaw, G.S., Papapostolou, V., Patel, H., Polidori, A., Salmond, J.A., Williams, D.E., 2020. Hierarchical network design for nitrogen dioxide measurement in urban environments. Atmos. Environ. 228, 117428.
WHO Regional Office for Europe, 2013. Review of Evidence on Health Aspects of Air Pollution: REVIHAAP Project. Copenhagen. http://www.euro.who.int/_data/assets/pdf_file/0004/193108/REVIHAAP-Final-technical-report-final-version.pdf?ua=1. (Accessed 17 July 2018).
Williams, D.E., 2019. Low cost sensor networks: how do we know the data are reliable? ACS Sens. 4, 2558–2565.
Williams, R., Kilaru, V., Snyder, E., Kaufman, A., Dye, T., Rutter, A., Russell, A., Hafner, H., 2014. Air Sensor Guidebook. U.S. Environmental Protection Agency, Washington, DC. EPA/600/R-14/159 (NTIS PB2015-100610).
Williams, R., Duvall, R., Kilaru, V., Hagler, G., Hassinger, L., Benedict, K., Rice, J., Kaufman, A., Judge, R., Pierce, G., Allen, G., Bergin, M., Cohen, R.C., Fransioli, P., Gerboles, M., Habre, R., Hannigan, M., Jack, D., Louie, P., Martin, N.A., Penza, M., Polidori, A., Subramanian, R., Ray, K., Schauer, J., Seto, E., Thurston, G., Turner, J., Wexler, A.S., Ning, Z., 2019. Deliberating performance targets workshop: potential paths for emerging $PM_{2.5}$ and O_3 air sensor progress. Atmos. Environ. X. 2, 100031.
Zimmerman, N., Presto, A.A., Kumar, S.P.N., Gu, J., Hauryliuk, A., Robinson, E.S., Robinson, A.L., Subramanian, R., 2018. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. Atmos. Meas. Tech. 11, 291–313.