Using synthetic data to benchmark correction methods for low-cost air quality sensor networks

Joost Wesseling[1] · Derko Drukker[1] · Alicia Gressent[2] · Stijn Janssen[3] · Pascal Joassin[4] · Fabian Lenartz[4] · Sjoerd van Ratingen[1] · Vera Rodrigues[5] · Jorge Sousa[3] · Philippe Thunis[6]

Abstract

A benchmark was performed, comparing the results of three different methodologies proposed by three institutions to calibrate a network of low-cost PM2.5 sensors, on an hourly basis, using synthetically generated real concentrations and sensor measurements. The objective of the network calibrations was to correct the 2000+ sensor measurements in the Netherlands for the sensitivity to (local) environmental conditions. The option to use real measurements was dropped because the number of low-cost sensors sufficiently close to the 40 reference measurement locations was assessed to be spatially insufficient to benchmark the proposed approaches. Instead, synthetic real concentrations were generated to enable validation at all sensor locations. Hourly actual sensor and actual fixed concentrations, as well as interpolated concentration maps, were used as underlying data to generate the synthetic data sets for the period of 1 month. The synthetic sensor measurement errors were constructed by sampling from a collection of differences between actual sensor values and actual measurements. Of the three tested calibration methods, two follow a similar approach, although having differences in, e.g., outlier analyses and method of grouping sensors, leading also to comparable corrections to the raw sensor measurements. A third method uses significantly stricter rules in outlier selection, discarding considerably more sensors because of insufficient quality. Differences between the methods become most apparent when analyzing data at a smaller time scale. It is shown that two network calibration methods are better at correcting the hourly/daily bias.

Keywords Low-cost sensors · Air quality · Benchmarking · Calibration

Introduction

Air pollution is recognized as the world's biggest environmental health risk (Lelieveld et al. 2020; WHO 2021). In Europe, many cities are affected by poor air quality levels and ambient concentration levels regularly exceed both the European standards prescribed by the Ambient Air Quality Directive (AAQD) and the guidelines recommended by the World Health Organization (WHO) (EEA 2022). For fine inhalable particles (PM2.5), the EU limit value is generally met (EEA 2022), but only a few cities manage to keep concentrations below the levels recommended by the WHO (González Ortiz et al. 2020; EEA 2022; Thunis et al. 2018; Rodrigues et al. 2021). As a reference, the AAQD establishes an annual limit value of 25 μg/m³ for PM2.5 concentrations, while the WHO sets an annual guideline of 5 μg/m³. According to the latest report released by the European Environment Agency (EEA) on air quality in Europe, the WHO annual guideline level for PM2.5 was exceeded by all reporting countries, except Iceland, in 2022 (EEA 2023). According to the latest estimates, at least 238,000 premature deaths were associated with the exposure to PM2.5 (EEA 2023).

To reduce the impacts of air pollution, particularly in cities where most of the population lives (e.g., more than two-thirds of European population lived in an urban area in

✉ Joost Wesseling
joost.wesseling@rivm.nl

1   National Institute for Public Health and the Environment (RIVM), Utrecht, The Netherlands

2   INERIS, Verneuil-en-Halatte, France

3   VITO, Mol, Belgium

4   Public Service Scientific Institute (ISSeP), Liège, Belgium

5   Centre for Environmental and Marine Studies (CESAM) & Department of Environment and Planning, University of Aveiro, 3810-193 Aveiro, Portugal

6   Joint Research Centre (JRC), Ispra, Italy

2020 (EUROSTAT 2023)), it is important to define effective planning strategies for air quality improvement (Monteiro et al. 2018; Pisoni et al. 2019; Viana et al. 2020; Oliveira et al. 2022). Low-cost sensor (LCS) networks are spreading in a wide variety of designs and capabilities (Wesseling et al. 2019; Schneider et al. 2019; Volten et al. 2018; Graça et al. 2023) and may have a key role in the definition of those strategies. There are still some challenges associated with these devices, such as inconsistencies linked to the high variability in the performance of similar sensors and the variations associated with different meteorological conditions. LCS are also sensitive to developing zero drift and ageing effects, which affect the calibration and can lead to systematic errors, both for particle sensors as well as gas sensors (Ratingen et al. 2021; Woutersen et al. 2022; Wesseling et al. 2021). Nevertheless, recent progress highlights the potential of this technology (EEA 2019; WMO 2020). The European Commission (EC) submitted a proposal to revise the Ambient Air Quality Directive as part of a "zero pollution" package. The proposal includes new methods to measure concentrations or deposition levels of pollutants, such as in situ sensors (EC 2022).

The Forum for Air Quality Modeling (FAIRMODE) was launched in 2007 as a joint response initiative of the European Environment Agency and the European Commission Joint Research Centre (JRC). The forum aims to bring together air quality modelers and users to promote and support the harmonized use of models by EU Member States, with emphasis on model application under the European Air Quality Directives. LCS are very relevant for FAIRMODE, especially concerning methodologies to combine sensor networks with modeled data and official fixed measurements. There is a need to elaborate on the role of communities like FAIRMODE and AQUILA (Network of National Air Quality Reference Laboratories) in the development of data acquisition methodologies based on the use of LCS. In addition, thought is needed on how to facilitate the use of sensor data for air quality mapping. The activities of the Working Group 6 of FAIRMODE focus on understanding the strengths and weaknesses of different ways of integrating low-cost sensors into air quality assessment protocols. The main roles of the group are (i) to exchange concepts and best practices about the integration of sensor network data in air quality mapping methods and (ii) to explore how air quality modeling can contribute to the exploitation and validation of an air quality sensor network. Present experiences suggest important roles for data fusion/assimilation approaches and possibly other techniques with similar scopes.

In a first common effort, the members of this working group compare their strategies for LCS outlier detection and calibration on a PM2.5 data set combining LCS measurements from (mostly) the Sensor Community, reference measurements from the RIVM, and optionally, meteorological data from the Royal Netherlands Meteorological Institute, KNMI. In order to evaluate quantitatively the performance of the different methods, knowledge of the true pollution field is required. Since the reference values are unavailable at all sites where sensors are deployed, a synthetic data set, as representative as possible of the various LCS fluctuations, has been generated and used to benchmark three approaches.

The paper is organized as follows:

- In the "Physical and synthetic sensor data" section, the data collection methodology is described in detail, followed by a description of the approach used to generate synthetic data.
- The "Calibration methods" section describes the different approaches to perform data calibration and detection of outliers.
- The "Results" section focuses on the results, analysis, and interpretation of the results.
- Finally, in the "Discussion and concluding remarks" section, the main conclusions are discussed and summarized.

Physical and synthetic sensor data

Dutch low-cost sensor network

In the Netherlands, the Dutch National Institute for Public Health and the Environment (RIVM) runs an infrastructure allowing citizens and other parties to provide and display data from low-cost sensors for (mainly) air quality. The aim of this "Measure Together" (Dutch: "Samen Meten") program is to support citizen science for environmental parameters such as air quality, water quality, and noise (Wesseling et al. 2019). Measure Together encourages people to do these measurements together, to build communities, connect with different stakeholders, connect with RIVM and each other, exchange information, and combine data to put these in a broader context. The data is provided on the website https://sensors.rivm.nl/. For more technical information on the network, see Appendix Annex 1 Fig. 12.

Most of the sensors used in the Netherlands are of the type SDS011 from Nova Fitness Co., Ltd.; they are available in many webstores. See https://aqicn.org/sensor/sds011/ for more information about the SDS011. Recently, also other sensors, like the Sensirion SPS30 have become popular. For the present analysis, we focus on the calibration of the

SDS011; other sensors are not taken into account. The use of sensors that are less sensitive to environmental conditions (like the Sensirion SPS30) to help correct other sensors with higher sensitivity (like the SDS011) is the subject of ongoing work at RIVM and is not discussed here.

The discussed calibration methods in this paper are particularly applicable to a sensor network with a large number of sensors (network calibration). Many individual citizens participate in the "Measure Together" network; they buy sensors and manufacture their own measuring setup using instructions that are available on the Internet. Some of the participants use ready-made sensor kits. All these sensors provide data on the scale of minutes to hours, but the calibration and quality of the individual sensors are not known. There is no uniform quality control of the sensors; most sensors will work as expected, but some of these low-cost sensors will show large fluctuations in performance. Furthermore, most participants install the sensor around their house, often, but not always, at background locations (from an air quality perspective).

So, at the level of the "Measure Together" network, the challenge is to estimate particular matter concentrations from a large number (at least hundreds, up to thousands) of sensor measurements of unknown and varying quality, located at different types of locations. Given the anonymous nature of most of the participants with sensors, co-located calibration of each individual is not feasible. Therefore, a network approach was adopted for the Dutch sensors in the "Samen Meten" project (Wesseling et al. 2019). The network calibrations discussed in this paper perform a calibration every hour and do not relate that calibration to sensor specifics. Here they differ from other network-oriented calibration schemes, like for instance MOMA, where calibration is performed on a larger time scale (months) or using drift detection (Weissert et al. 2023).

A large advantage of the network-calibration methods described in this article is that no specific quality/calibration information on the individual sensors is needed. A drawback of the methods is that they only work for large enough networks (i.e., many sensors), that consist of sensor that react roughly similar to environmental conditions like relative humidity.

Creation of synthetic data

In FAIRMODE/WG6 (low-cost sensors for air quality), work is done on methods to calibrate/process the raw results of networks of low-cost sensors in such a way that the best estimates of the real/actual concentrations can be obtained. We define different concentrations, as follows:

Actual fixed concentrations/measurements	The concentrations measured with official EU reference methods (or equivalent) at fixed locations
Actual reference concentrations	The concentrations we would measure if official fixed reference measurements were available at every location
Synthetic real concentrations	The synthetic concentrations we construct to represent the real concentrations. So, after the sensor data validation and calibration, we should end up with these synthetic real concentrations
Synthetic fixed concentrations/measurements	The synthetic concentrations we construct to represent the actual fixed concentrations
Actual sensor value	The concentrations reported by physical sensors in the field at every location
Synthetic sensor value	The synthetic concentrations we construct to represent the behavior of the sensors
RIO concentrations	Concentrations calculated using the RIO model. These are an interpolation of the (background) fixed concentrations, either actual or synthetic

As we work with a large network of (2000+) low-cost (<50 Euro) sensors, we do not know how the sensor is positioned with respect to, e.g., a local street or building, and the quality of each sensor is also unknown. We assume that the large number of sensors will ensure a sufficiently good average performance of the network.

Usually, we do not know the actual reference concentrations (and certainly not the true concentrations), so we cannot test the quality of different algorithms for calibrating the sensors in a simple way. Knowing the actual reference concentrations would make it possible to compare results from different calibration methods and objectively test the effects of variations in calibration strategies. Alternatively, a synthetic data set can be generated to represent a *real* concentration field. For every hour in a test period, we take the locations of the actual monitoring stations and the physical sensors to create respectively synthetic fixed and sensor data. So, for every location where actual fixed concentrations are available, we generate synthetic fixed concentrations, and for every location of sensor measurements, we generate synthetic sensor concentrations. It is very important for the synthetic data to have (roughly) the same systematic and individual "chaotic" structure of the deviations from the actual values (noise) as actual LCS:

- Individual sensors show very non-standard distributions of hourly deviations

- Systematic deviations due to relative humidity or other environmental factors, this means local correlations between systematic deviations
- Sometimes individual sensors work well for a period of time, sometimes not, this means correlations between sensors may change over time

As analytical distributions will probably not fully describe the random behavior of LCS, the basic idea for creating realistic synthetic sensor data proposed here is to use behavior from actual/physical sensors as the basis for the synthetic behavior.

Synthetic real concentrations

To create realistic "sensor data," the starting point has to be a set of realistic "actual concentrations" at the locations where we want to create synthetic data. By using the hourly PM2.5 concentrations physically measured at official fixed monitoring locations, an interpolated concentration field can be constructed. We assume this field reproduces the *average* concentration distribution in the region/country sufficiently well. On top of this average field, we expect local variations, minima, and maxima in concentrations, reflecting the many local situations in the country.

The hourly average concentration distribution in the Netherlands is generated by the RIO model created by VITO. The RIO field is generated on an hourly basis from the official measurement data (official measurements are performed by the Dutch national network "LML," managed by RIVM). The resolution is 4×4 km^2, downscaled (interpolated) to 1×1 km^2. Land use and historical data are used to fine-tune the results of the model. Evidently, due to the relatively coarse resolution, the RIO field does not show any variations below 1 km.

We can estimate local variations in concentration from the differences between the average background concentrations estimated using the RIO model and the actual fixed measurements where these are available. These differences, derived from the actual fixed concentrations and the RIO interpolation of the actual fixed background measurements, can be used to define a distribution for the local variations at the locations of the synthetic sensors and the created synthetic real concentrations at those locations. Adding the local variations to the average field then yields the synthetic real concentrations

We constructed two methods to generate local variations in the software.

(1) The simplest is to sample from all available actual differences between RIO interpolations and actual fixed measurements in the national network in the Netherlands in the considered hour. A drawback is that potential local differences between RIO and the official

measurements are not considered by this approach; the observed differences are used for synthetic data all over the country. However, there will be circumstances when RIO will match the measured data better in some parts of the country than in other parts. For example, when there are strong spatial PM2.5 gradients in the country.

(2) To take this local spatial variability into account, we can use the differences between RIO concentrations and the actual fixed measurements located closest to the locations of the synthetic data to generate the synthetic real concentrations. This way, local correlations between concentrations are taken into account. We assume that the differences are still representative of locations if the closest actual fixed measurements are more than (roughly) 10 km away as we prefer a potentially poor local approximation over no approximation at all.

As the distribution of sensors over the country is not necessarily the same as the distribution of the reference measurement locations, the two approaches ((1) sample from all national or (2) only use local differences) will result in different sets of synthetic data. For the benchmark, we used the local approach.

Synthetic sensor concentrations

For creating the synthetic sensor data, we need to have an approximation for the deviations between the values of physical sensors in the field and the actual concentrations at those locations. The approximation for these deviations must be close enough to the real behavior of the sensors, given the local conditions. As we do not know the actual reference concentrations, we use the RIO values as an approximation to these actual reference concentrations. We assume the intrinsic random and systematic uncertainties of the sensors to be zero on average. The relation between environmental conditions and the behavior of many types of low-cost sensor is observed to be of a multiplicative character. Therefore, we assume that for sensor k at location (x,y) and hour t:

$$C_{sensor,k}(x,y,t) = \gamma_k(x,y,t) \left(C_{Bgr}(x,y,t) + C_{local}(x,y,t) + \varepsilon_k(x,y,t) \right)$$

(1)

where:

$C_{sensor,k}(x,y,t)$	actual value of a physical sensor
$C_{Bgr}(x,y,t)$	actual background concentration
$C_{local}(x,y,t)$	the local contribution to the concentration
$\varepsilon_k(x,y,t)$	the random deviation of the sensor, with an average value of zero

$\gamma_k(x,y,t)$ a factor, describing the effect of the environmental conditions

We further assume that the result of the RIO model is, on average, a sufficient approximation to the background concentration, i.e.:

$$C_{RIO}(x, y, t) \approx C_{Bgr}(x, y, t) \qquad (2)$$

with $C_{RIO}(x,y,t)$ the (background) concentrations are calculated using the RIO model. In order to create realistic synthetic sensor data, we need a sufficient approximation for the $\gamma_k(x,y,t)$. For all sensors, we determine the ratio between the RIO concentration at the sensor location and the actual sensor value. So, we use the ratios $\hat{\gamma}$ as estimates of how much each sensor deviates from the actual concentration:

$$\hat{\gamma}_k(x, y, t) = \frac{C_{sensor,k}(x, y, t)}{C_{RIO}(x, y, t)} \qquad (3)$$

It is preferred that $\hat{\gamma}$ does not deviate too much from the real value, i.e., the local contributions and random deviations in (1) are relatively small. The ratios $\hat{\gamma}$ should not be much smaller than the actual effects of the environmental conditions (γ_k), as this would lead to a behavior of the synthetic sensors that would be too smooth. As a check, we take the ratio between the γ_k that follows from (1) and the $\hat{\gamma}_k$ in (3):

$$\frac{\hat{\gamma}_k}{\gamma_k} = \frac{C_{RIO} + C_{local} + \varepsilon_k}{C_{RIO}} \approx 1 + \frac{C_{local}}{C_{RIO}} + \frac{\varepsilon_k}{C_{RIO}} \qquad (4)$$

If there are local contributions to the PM2.5 concentrations, the ratio in (4) is larger than 1, indicating that the deviations of the synthetic sensors will be at least as large as those of the real sensors. Therefore, we assume that the set of all estimated ratios $\{\hat{\gamma}_k(x,y,t)\}$ of the sensors that were actually in use during an hour is a pool of sufficiently realistic sensor ratios that can be used to generate synthetic sensor data in that specific hour. We first create a synthetic real concentration:

$$C_{synth\_real} = C_{RIO}(x, y, t) + \Delta_{local}(x, y, t) \qquad (5)$$

where Δ_{local} represents the local contributions. It is estimated by random sampling from the distribution of differences between the RIO value and the actual fixed concentrations in the proximity of the location.

We can now create a realistic synthetic sensor value ($C_{synth\_sensor,k}(x,y,t)$) at that location and time by multiplying the synthetic real concentration (C_{synth_real}) with a ratio from the pool of ratios in the neighborhood of the sensor:

$$C_{synth_{sensor},k}(x, y, t) = C_{synth\_real}(x, y, t)\, \hat{\gamma}_k(neighborhood, t) \qquad (6)$$

Taking ratios from the pool of available ratios in the neighborhood of a sensor does not take into account that there is a temporal trend/correlation in the behavior of each sensor. By taking the ratio from a sensor in the neighborhood of a location (in the order of 10 km), the synthetic sensor will show the same variations in behavior in time as a real sensor does. Optionally, the variations in γ_k can be multiplied by a random factor (around 1.0) every hour.

Test of the distributions

An important test of synthetic sensor data is to compare the concentration distribution of these synthetic values to the values of the physical sensors during the hour that we are creating synthetic data for. By using the actual behavior of sensors, there is a risk of creating artificial correlations between actual and synthetic data. A test is to calculate the correlation between the actual and synthetic data, which should be very low. On the other hand, the distributions of actual and synthetic data should be very similar, with a high correlation. Two examples of real and synthetic concentrations and distributions are shown below in Figs. 1 and 2.

The correlations between the actual concentrations are clearly very small, whereas the correlations between the concentration distributions are very high, as should be the case for a good synthetic approximation. For the benchmark, the synthetic random noise of the sensors was assumed to be 50% of the estimated random noise of the actual sensors.

Calibration methods

Calibration method INERIS

INERIS has developed a methodology to identify and eliminate outliers in the synthetic sensor dataset before performing a calibration procedure. A first cleaning of the data is performed. Negative values are removed; those values occur when the concentrations are below the limit of detection of the sensors, so they are not relevant for fine particle measurements. In addition, sensor data identified as (i) belonging to a series of "frozen" values (defined as at least three successive hourly time points for which the sensor values are constant), (ii) including a constant positive bias with respect to the reference measurements, (iii) representing very high concentration peaks (> 2 times the maximum value of the concentrations, measured at the reference stations in the study area, i.e., 240 $\mu g/m^3$) are deleted. These values correspond to 0.4% of the initial dataset.

The outlier detection procedure of the resulting dataset is based upon van Zoest et al. (2018), an approach that is adapted in this work to PM2.5 sensor data on a national scale. Firstly, a classification of the sensor data is performed.
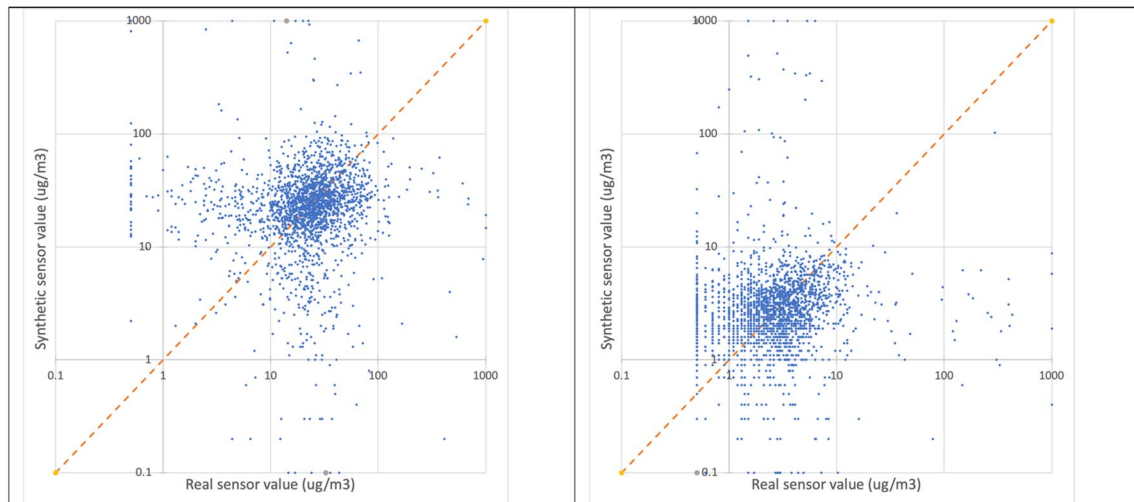
Fig. 1 Real (horizontal) and synthetic (vertical) concentrations in μg/m³. Left figure January 11th, 2022, 06:00, Right figure February 08th, 2022, 20:00
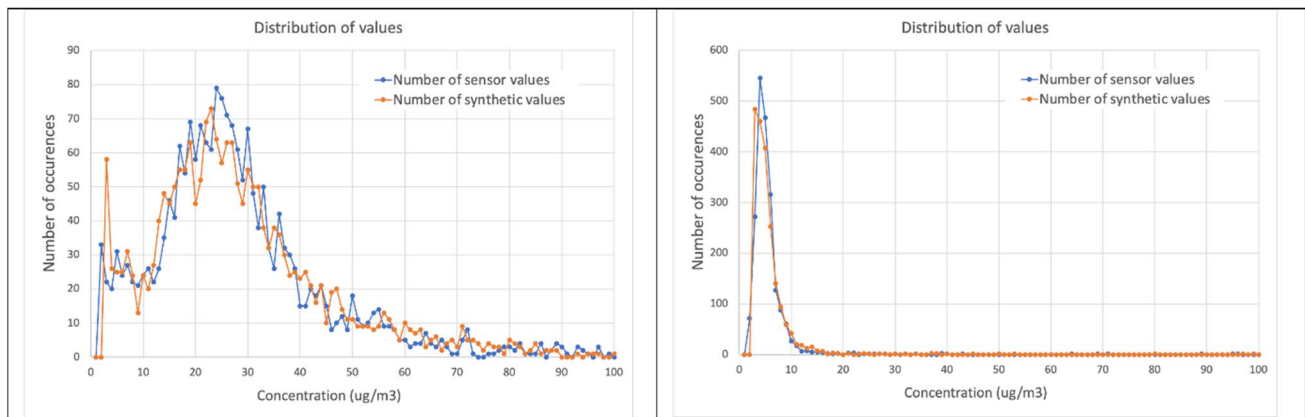

Fig. 2 Concentration distribution of real raw sensor data (blue) and synthetic sensor data (orange). Left figure January 11th, 2022, 06:00, right figure February 08th, 2022, 20:00

Each sensor is assigned to a group depending on the "type of pollution" influencing the sensor measurements, i.e., urban background, regional background, traffic, or industrial environment influence. This classification is based on information about land use, then adjusted by population density, the road network, and the type of reference stations located next to the sensors. Secondly, a clustering is made based on the distance of the sensors from each other. Therefore, groups of sensors are formed gathering the closest ones with a maximum distance of 10 km. This distance is chosen to make the link with the spatial scale (urban scale) defined in the original application of van Zoest et al. (2018). In this way, the distance clustering is independent of the existing land use groups which means that the remaining groups are an intersection of distance and land use clustering. Finally, the sensors are clustered in time into four different seasons.

In this work, only one season is considered (winter). The number of sensor groups obtained from this classification is of the order of $n_{type} * n_{dist} * n_{season}$, i.e., 257 groups as shown in Fig. 3.

Each group of sensors is considered individually to determine a confidence interval in which the sensor values must lie in order not to be eliminated. This interval is defined as described by Expression 6.

$$\mu \pm z \times \sigma \qquad (7)$$

where μ is the mean of the (log-transformed) PM$_{2.5}$ concentrations in μg/m³, σ is the standard deviation, and z is the indicator of the size of the confidence interval. Assuming independence and normality of the data, the value of z is set to 2.97 for an interval of 99.7%, which is then rounded to 3. For each group of sensors, a logarithmic transformation
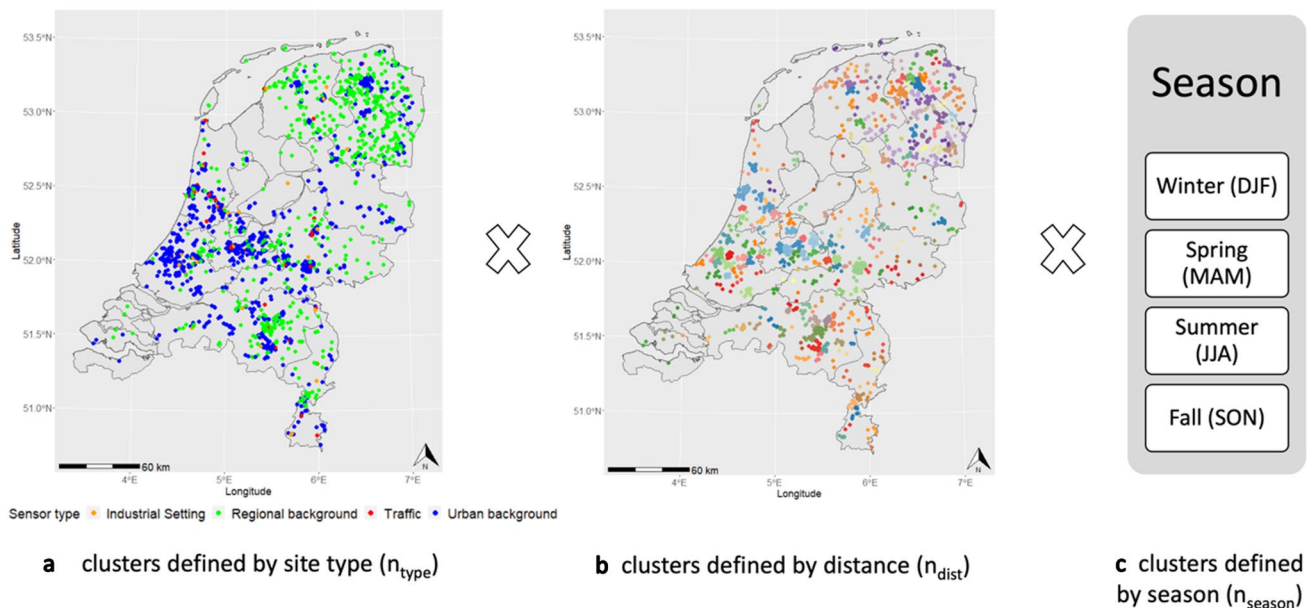
Fig. 3 The colors in (b) indicate the different clusters, of which there are many. A color scale was therefore omitted

is applied to the data to justify the use of Eq. 6. A normal distribution truncated at 1 μg/m³ is obtained and a maximum likelihood estimator is used to calculate the mean and standard deviation values of the underlying normal distribution. The confidence interval is then defined based on this estimate of the mean and standard deviation. Sensor values outside this range are identified as outliers.

Once the sensor data has been cleaned and outliers identified, a correction of the sensor network data is performed based on the method developed by the RIVM (see section "Calibration method RIVM"). The INERIS and RIVM methods differ in the data cleaning/outlier removal process but both use a correction factor based on the average cleaned sensor concentrations. At each reference station, the closest sensors are selected. In particular, sensors that are located in the buffer of the maximum representativeness of the station that varies from 50 m for stations under the influence of traffic to 25 km for stations under the influence of regional background pollution are considered. Then a correction factor is estimated at the hourly time step as described in Eq. 7.

$$F_{i\,station} = \frac{C^{ref}_{i\,station}}{C^{sensors}_{i\,station}} \tag{8}$$

with $F_{i\,station}$, the correction factor calculated at the reference station; $C^{ref}_{i\,station}$, the concentration measured at the station at the considered hourly time; and $C^{sensors}_{i\,station}$, the average concentration measured by the sensors located near the station. A spatial interpolation of the correction factors is performed by the IDW (inverse distance weighting) method on the study area. Fig. 4 shows the sensor data correction performed on 01/09/2021 at 4:00 PM.

Calibration method ISSeP

A low-cost sensor network should ideally have a high density of sensor locations but without necessarily being homogeneous in terms of spatial distribution. On the other hand, the network of official telemetric stations (fixed measurements) must be characterized by a reasonably homogeneous distribution over the low-cost sensor's deployment area.

The ISSeP calibration method is an iterative process.

- Firstly, a time series of spatial weighted interpolations are performed to produce interpolated fields of reference values (i.e., the best guess for the concentration field).
- Secondly, the time series of each sensor is fitted by two regression models using collocated time series of the interpolated reference values: a linear model and a non-linear multi-variables model. Each regression model provides a set of parameters and a coefficient of determination. Finally, the parameters of the fits are used to calculate the weight of each sensor to be used in the interpolation process in the next step.
- The next iteration will now also include sensor data that is used in a second series of weighted spatial interpolations producing updated fields for the best estimate of the concentration fields.
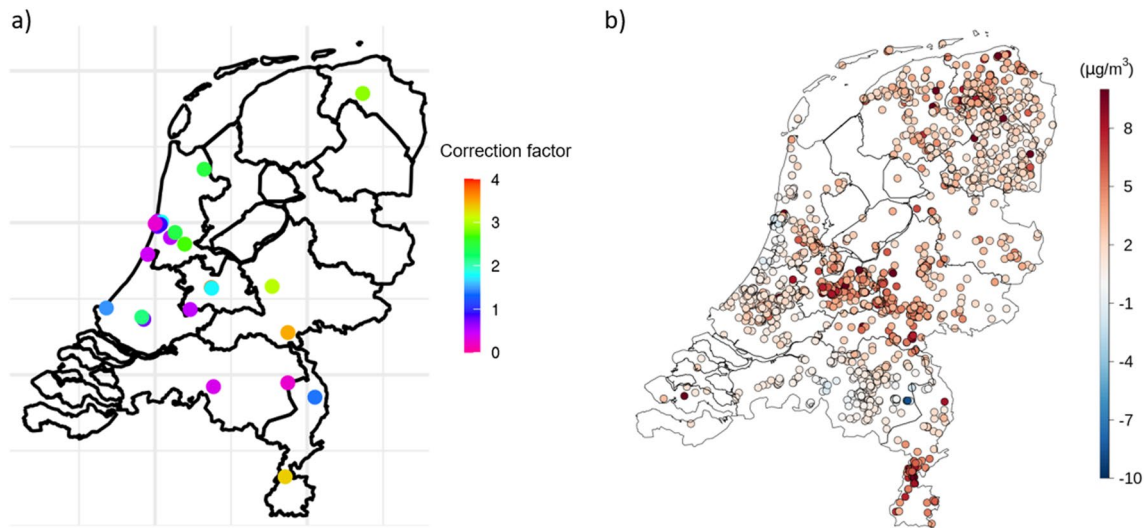
Fig. 4 Correction factors **a** calculated at reference stations and the correction (**b** the difference between the initial sensor data values and the corrected data values) after interpolation of the factors over the domain on 01/09/2022 at 4PM

- The process is repeated at least twice to obtain the final calibration results and the resulting best estimate of the concentration fields. The steps are described below in some more detail.

The interpolations that are used in the processing of the sensors are obtained using the spatial interpolation method DIVA (*Data Interpolating Variational Analysis*) (Troupin et al. 2012). This tool is developed by the GHER laboratory of the University of Liège and consists of an optimal interpolation technique based on a finite element mesh (Troupin et al. 2013). A first attempt to use DIVA to map atmospheric pollutant concentrations at various scales was made for Wallonia and the city of Antwerp (Lenartz et al. 2018).

The interpolation field is combined with a relative error field, reflecting locally the uncertainty of the interpolated reference values. The hourly value from each sensor can be used in a similar way as the reference value which is reconstructed by interpolation. The values of the fixed (official) measurements and the sensor values are distinguished by a specific weight determining their influence during the construction of the interpolation field. The weights of the official values are systematically set to a maximum value of 1, while the weights of the sensor values are initially set to 1/3 and then re-calculated using information from the regression models.

To perform the calibration of the sensors, a sampling window of about 10 days is considered to provide a sufficiently robust time series of hourly observations. For each sensor, a non-linear multi-variable regression is carried out between the sensor values time-series and the interpolated reference values at the location of the sensor. Each reference value is replaced by a couple of values obtained by multiplying the interpolated

reference value by [1-ε] and [1+ε] where ε is the interpolation relative error at the sensor location. This duplication is only considered to compute the coefficient of determination of the non-linear multi-variable regression. This regression is based on five independent variables, which are the sensor values at several different moments in its time series, in the near future and past, and at the present moment of the considered interpolated reference value. The time delays used are about one to a few hours. An exponent is applied to the sensor values, which is as close to zero as the used sensor value is more in the past or in the future, i.e., away from the present moment.

The calibrated sensor values are obtained from the interpolated concentration field at the locations of the sensors. The formulation of the used non-linear multi-variable regression is:

$$Y_m = C_a X_a^{Ea} + C_b X_b^{Eb} + C_m X_m^{1} + C_c X_c^{Ec} + C_d X_d^{Ed} + \text{intercept} \tag{9}$$

where:

Y_m is the actual reference interpolated value at the considered present moment

X_m is the actual sensor value at the considered present moment

$X_{a\text{-}b\text{-}c\text{-}d}$ are the sensor values at other moments in the sensor's time series

a is the observation index in the near past, for example 3 h before present

b is the observation index in the close past, for example 1 h before present

m is the observation index at the considered present moment

Fig. 5 Comparing the PM$_{2.5}$ concentrations in µg/m³ of the sensors within 1 km distance of official locations to those values before (left) and after (right) calibration for Jan 09th, 16:00 (top) and Jan 15th, 21:00 (below)
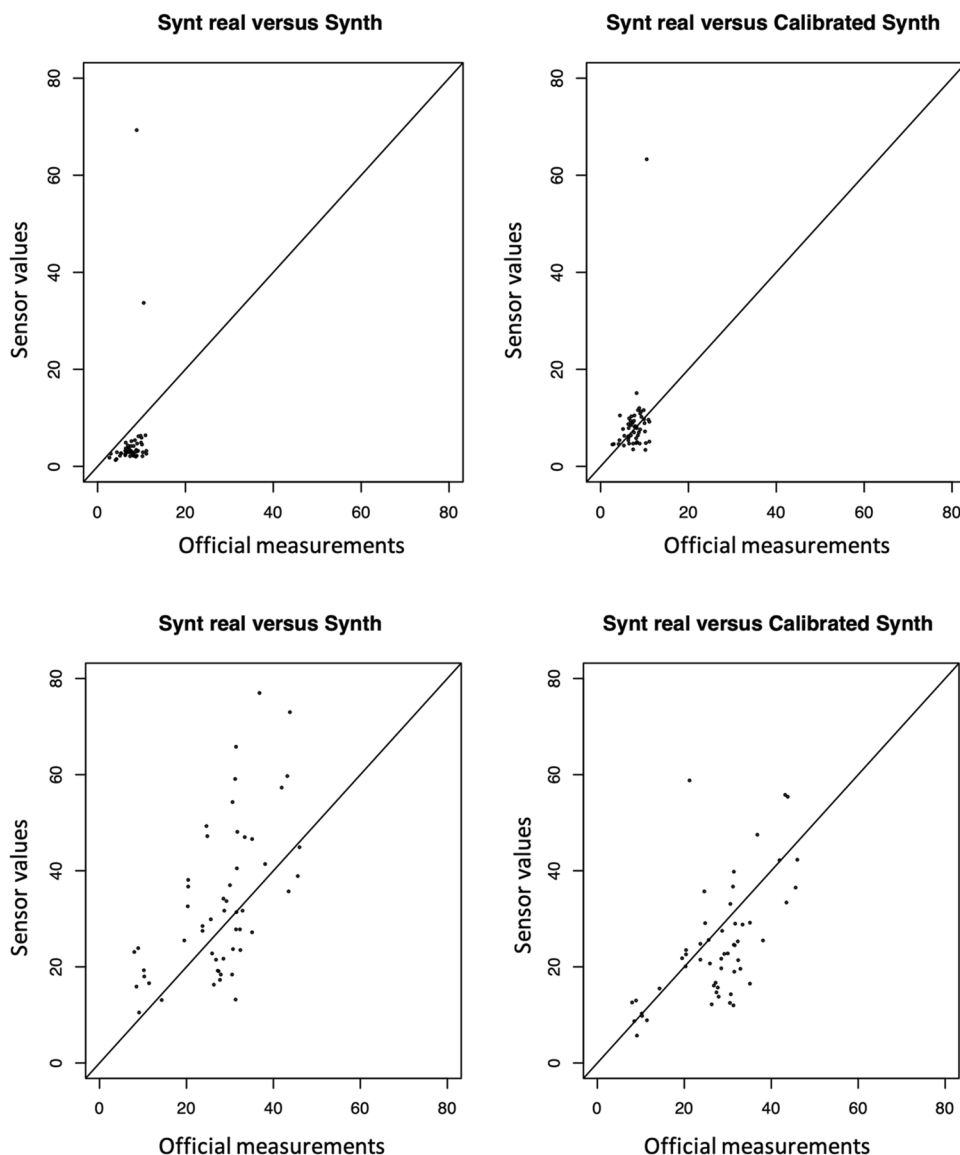
c is the observation index in the close future, for example 1 h after present

d is the observation index in the near future, for example 3 h after present

E_a and E_d are the exponents of near past/future observations, for example 0.6

E_b and E_c are the exponents of close past/future observations, for example 0.8

Coefficients C_a, C_b, C_m, C_c, C_d, and the intercept are determined by a linear least-square fitting method and compose the set of six corrective parameters specific to each sensor. A simple linear regression is also performed at this stage. It is used to evaluate the quality of the sensor through the computation of a performance index. The performance index is involved in the re-calculation of the weight of the associated sensor. During the following iteration, each sensor will provide its newly corrected and weighted values to compose the new data sets used for the construction of the new best guess of the interpolated concentration field. This will then be used again in the determination of both regression models which will produce the new parameters and calibrated values of the low-cost sensors. The sequence of iterations leads to the weights of the low-cost sensors converging towards a stable limit value. It was found that after three iterations, half of the low-cost sensors showed a weight variation of less than 10%, which was the limit applied in the presented results.

Calibration method RIVM

The calibration method used by RIVM mainly focusses on eliminating the large deviations of low-cost sensors due to environmental conditions, like humidity. Due to the large number of sensors, where individual quality and performance are

unknown, it is not feasible to use calibration functions for each sensor. Instead, we compare concentrations reported by official measurements with groups of sensors to estimate and correct the deviations. The method consists of several steps:

(1) Data selection

In practice, there are two predominant types of malfunctioning sensors: those that consistently report (1) almost zero concentrations or (2) very high concentrations. Therefore, sensors with the lowest and highest 5% of the concentrations are not included in the calibration. However, after a calibration is obtained, all sensors are calibrated and reported.

(2) Group sensor data, determine corrections

For each available official fixed measurement location at a particular hour, the surrounding sensors are grouped in clusters with a typical maximum distance of 5 km. The sensors just across the border (Germany and Belgium) are also used in this grouping while, for the synthetic sensor data, only locations of Dutch sensors were used. Depending on the number of sensors in a group, the highest and lowest values may be excluded from the calibration; i.e., with more than 20 sensors in a group, the three highest and lowest sensors are not used in the calibration. With 10–20 sensors in a group, the highest and lowest two are not used. With less than 10 sensors, the highest and lowest sensors are not used. For each group, the ratio between the average of the included sensors and the official measurement is calculated, yielding a map with local calibration factors. The effects of the local calibrations are shown in Fig. 5, where we compare the official measured fixed concen-

Fig. 6 Locations of the (synthetic) sensors. The color scale indicates the average synthetic real concentrations during the benchmark period in mg/m^3
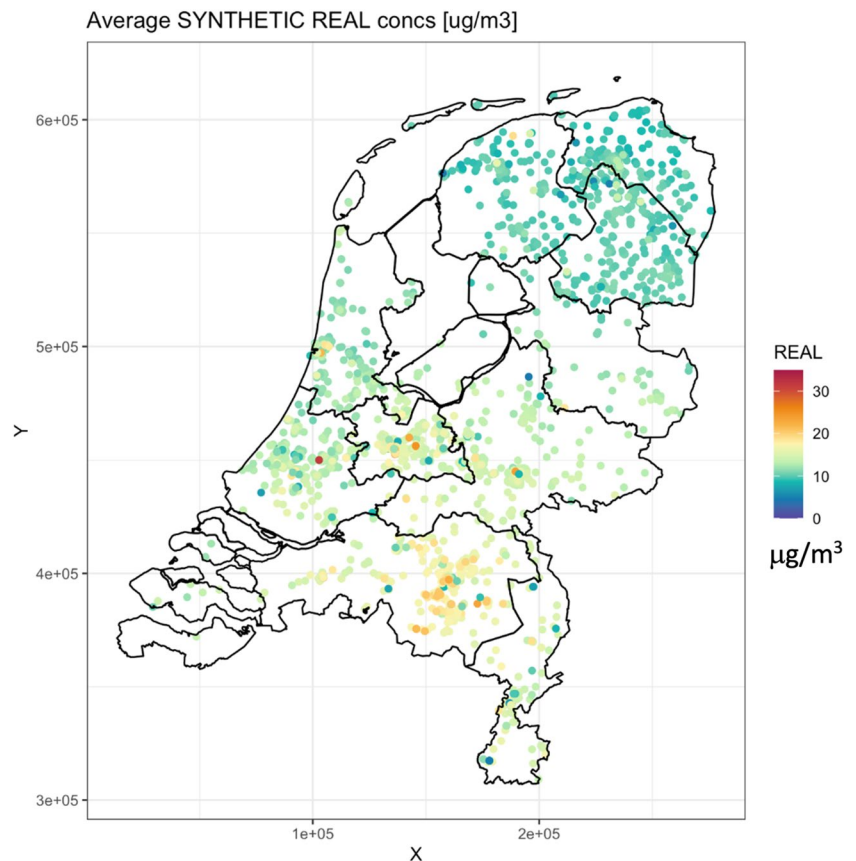

Table 1 Median value of R and RMSE of individual sensors. Mean absolute value and standard deviation of the daily bias.

Calibration	median R		median RMSE µg/m^3		mean abs day bias µg/m^3		stdev day bias µg/m^3	
	Raw	Cal	Raw	Cal	Raw	Cal	Raw	Cal
RIVM	0.84	0.89	7.68	6.02	2.53	1.07	3.14	0.58
INERIS	0.84	0.88	7.65	6.15	2.56	1.15	3.15	0.97
ISSEP	0.84	0.87	7.26	5.43	2.53	2.21	3.17	2.73

Fig. 7 Biases of the raw sensor data (vertical) versus the calibrated sensor data (horizontal). All concentrations in μg/m³. Shown are results from RIVM (top, left), INERIS (top, right), and ISSeP (bottom, left)
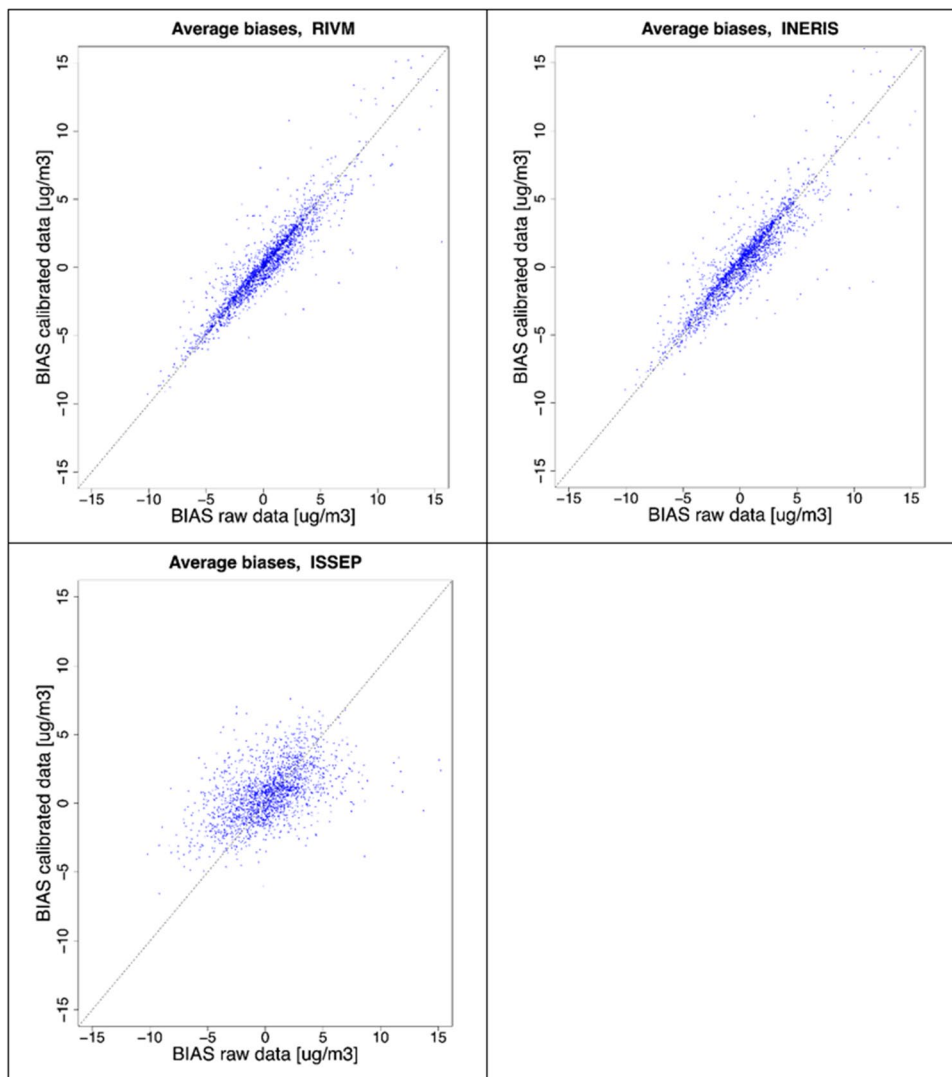
trations and the nearby sensors (<1 km). On January 09th, the concentrations were low, and meteorological conditions resulted in sensors underestimating the official fixed measurements. After the calibration, the average values are in better agreement; however, few sensors with an already high value are also scaled up. On January 15th, the mean and the variance of the concentrations are much higher. Furthermore, in some regions, the average official fixed measurements and sensors are comparable whereas in other parts, the sensors are substantially larger than the official fixed measurements. After the calibration, there is a better overall agreement in the whole country.

(3)   Interpolate corrections

It is assumed that the local calibration factors are sufficiently representative of the surrounding areas. All the available local calibration factors are then used to inter-

polate a calibration factor for the whole country. For simplicity, we applied the inverse distance weighting (IDW) with the modified Shepard's method, using only the nearest neighbors. To reduce the typical circular patterns in IDW, the maximum distance of interpolations is varied to (almost) always include several neighboring local calibration factors in the interpolation.

(4)   Calibrate all sensors

All sensors (including the outliers excluded from the calibration step) are calibrated based on the interpolated calibration field. Quality parameters are available for each sensor, indicating if the sensor was used in the calibration and, if not, why.

For use in a subsequent data fusion process using the calibrated sensors, the whole calibration process is repeated many times in a bootstrap procedure varying the available official measurements and the available sensors to estimate the 95% confidence interval of the sensor values.

Fig. 8 Target plots of the raw monthly averaged sensor data (red) versus the calibrated sensor data (blue). All concentrations in μg/m³. Shown are results from RIVM (top, left), INERIS (top, right) and ISSeP (bottom, left)
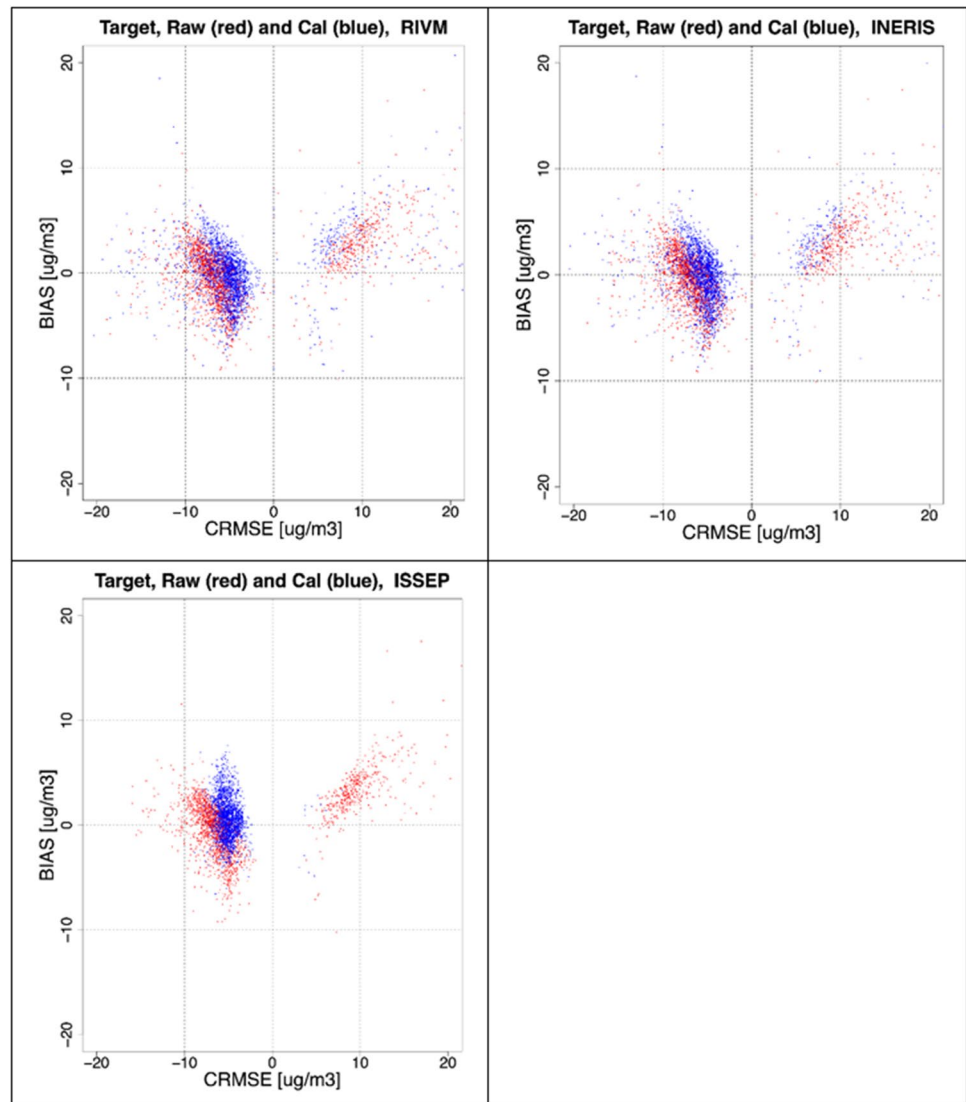
## Short description of differences between calibration methods

The data processing methods of INERIS and RIVM are similar. The biggest difference is the extensive clustering of the sensor data performed by INERIS, based on a number of parameters. For all clusters, a separate outlier detection using log-transformed concentrations is performed. Contrary to this, RIVM bases the outlier selection on the highest and lowest percentiles of the data, followed by an outlier selection in groups of sensors surrounding fixed measurement locations. The method employed by ISSeP is different, as it compares the sensors to interpolated values of the concentrations at the official fixed locations. The sensor values are compared

to the interpolated values, and the differences are used to estimate a correction that is to be applied to the sensors. The methods of RIVM and INERIS method thus first compute divergence ratios between official stations and sensor measurements and then spatially interpolate them whereas the ISSeP method first spatially interpolates official stations combined with weighted sensor measurements and then computes divergence ratios.

Results

The three different approaches to calibrating the sensors resulted in just over 1.4 million data points from INERIS, just below 1.4 million from RIVM, and 1.2 million from

Fig. 9 Scatter plots of the daily average of all sensors for raw data (red) versus the calibrated data (blue). All concentrations in µg/m³. Shown are results from RIVM (top, left), INERIS (top, right), and ISSeP (bottom, left)
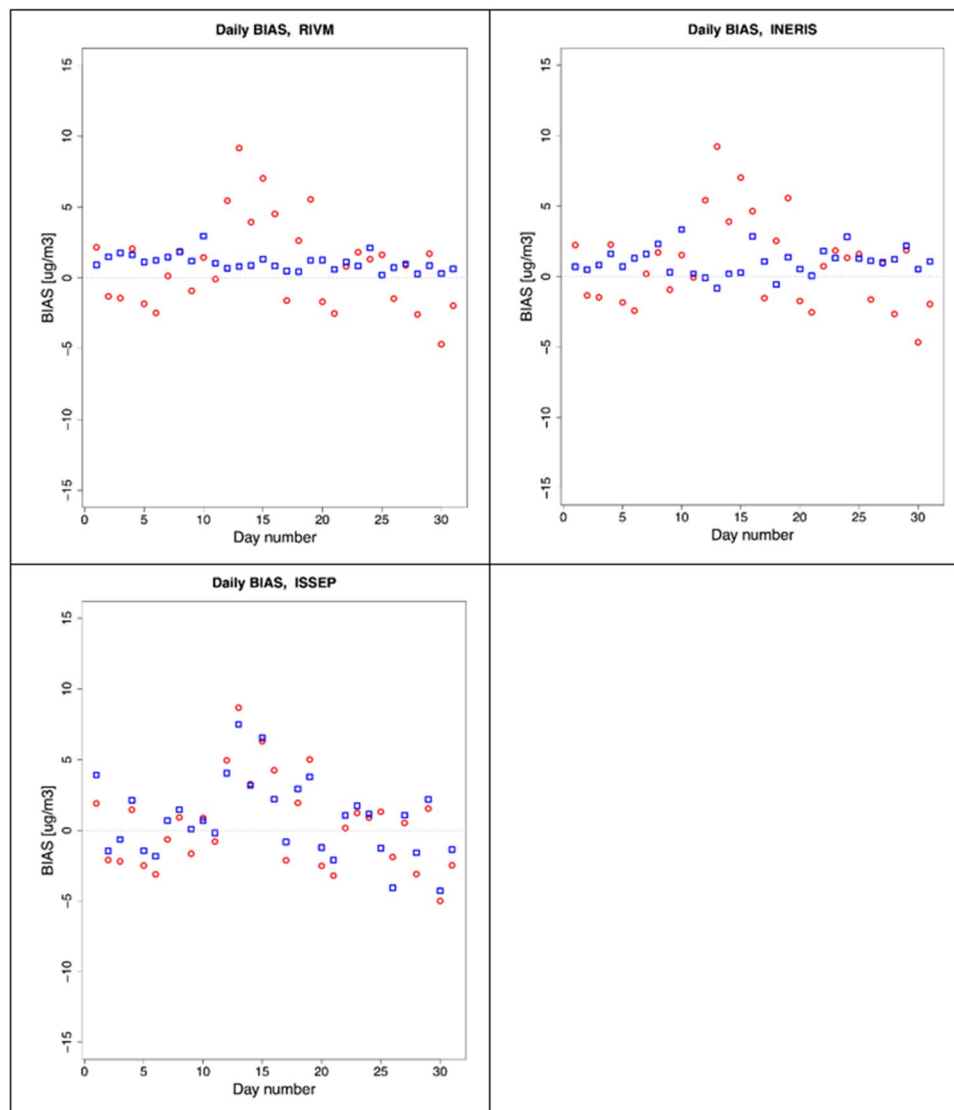

ISSeP. As each method has different schemes to select valid data, the number of calibrated sensors that are reported depends on the method. For a sensible comparison, only the hourly results for sensors with raw (uncalibrated) values below 250 mg/m³ were compared. Apart from the first hours of a new year, this is a concentration level that is not expected to occur. For the selected sensors and hours, the raw, uncalibrated values and the calibrated values are available. Figure 6 depicts the locations of the sensors during the benchmark period.

To make a first comparison between the three different calibration methods as described in the previous section, a set of statistical parameters was calculated. The Pearson correlation and the root mean square error were calculated for all raw and calibrated sensors individually during the entire measurement period. The median values of these individual metrics are

shown in Table 1. Additionally, the average difference, taken over all sensors, between sensors and synthetic real concentrations was calculated for each day in the measurement period. The mean absolute value and standard deviation of these daily biases are also reported in Table 1. The parameters in the table for the raw (uncalibrated) sensor data can still differ between the calibration methods because of the differences in data points that have been analyzed. All parameters improve when comparing calibrated with raw sensor data.

A simple next test is to compare the average values of the raw and calibrated sensor values to the synthetic real values. Alternatively, the raw and calibrated biases can be compared, i.e., the average differences between the raw sensor values and the synthetic real values compared to the average differences between the calibrated sensor values and the synthetic real values. The results for the methods are shown below in Fig. 7.

Fig. 10 Average daily biases of the raw sensors (red) and the daily average relative humidity (gray), in percent, as a function of the day in the month of January 2022
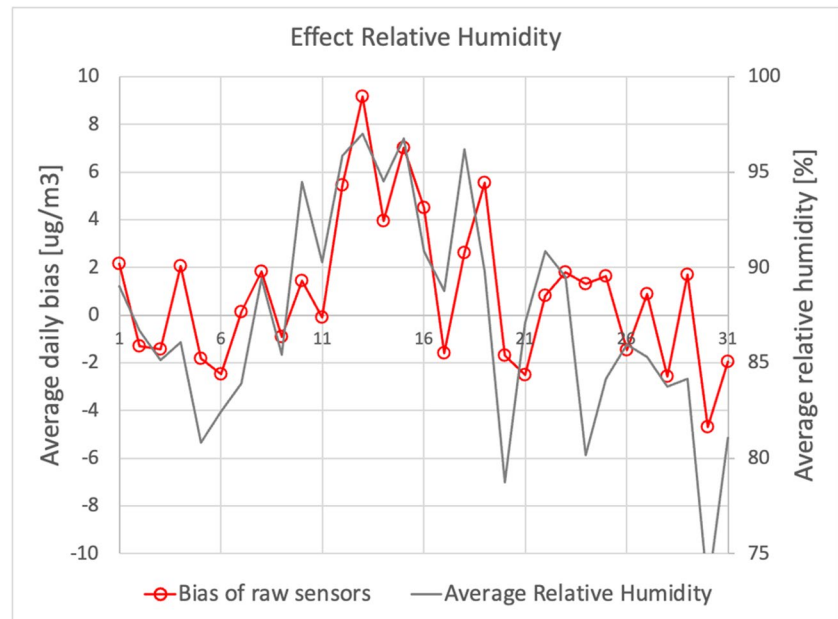
The raw and calibrated biases are similar in the case of the INERIS and RIVM methods, as if the calibration process has only a limited effect on the sensors. However, the method of ISSeP produces biases that are different before and after calibration.

In the Target diagram, developed in the FAIRMODE project and implemented in the Delta Tool (Janssen and Thunis 2020), two sets of data can be compared, models and observations or other variables. The *X* and *Y* axes of the Target diagram correspond to the BIAS and the CRMSE (centered root mean square error). Depending on whether the CRMSE is dominated by the correlation or by the spread in values, the value is plotted as either negative or positive, see the manual of the Delta Tool for more information. In the standard Target diagram, the BIAS and CRMSE are normalized by the measurement uncertainty and parametrized as part of the FAIRMODE work. For the present analysis, we have chosen not to normalize the BIAS and CRMSE.

The Target diagrams (Fig. 8) for the three methods show more differences between the raw and calibrated data than the comparison of biases. The calibrations of INERIS, ISSeP, and RIVM all result in slightly smaller (absolute) CRMSE's than are observed in the Target plots of the raw sensor data. The BIAS of both INERIS and RIVM is, on average, practically zero, whereas the BIAS of ISSep is slightly positive. In the Target diagrams, it is obvious that the method of ISSeP produces a smaller number of calibrated sensor values than the other methods. After the calibration, there are practically no results with positive values of the CRMSE.

The differences between the raw and calibrated sensor values become clear when looking at the spatial average values of the sensors on a daily basis; see Fig. 9. The raw sensors show substantial positive deviations from the synthetic real concentrations during the 10 days between January 10th and 20 h, probably due to relatively high humidity in that period. On the other hand, the raw sensors underestimate the synthetic real concentrations during the other days of the month. Although the calibrations of INERIS and RIVM have significant effects on hours and days with large biases, there also are enough days in the benchmark period with limited corrections, such that the average effect of the calibration over the whole month is limited.

The sensors in the benchmark, the SDS011, are known to be very sensitive to moisture (VACUUMS 2021). To see if the relative humidity is driving the biases of the raw sensors, we have calculated the daily average relative humidity of all stations of the Dutch met-office (KNMI). Figure 10 shows the daily biases of the uncalibrated sensors in red and the daily average relative humidity. The curves show similar behavior over the days of the month. During January 10th–19th, the average relative humidity is clearly higher than during the other days.

The effects of the three types of calibrations become more pronounced when looking at the hourly average concentrations of the raw and calibrated sensor data. In Fig. 11, the calibrated sensor data (blue dots) seems to slightly but systematically overestimate the synthetic real concentrations. This is the result of including part of the sensors that are not functioning properly, resulting in high values, both raw and in the analysis of the benchmark

Fig. 11 Scatter plots of the hourly average of all sensors for raw data (red) and the calibrated data (blue) versus the synthetic real concentrations. All concentrations in µg/m³. Shown are results from RIVM (top, left), INERIS (top, right), and ISSeP (bottom, left)
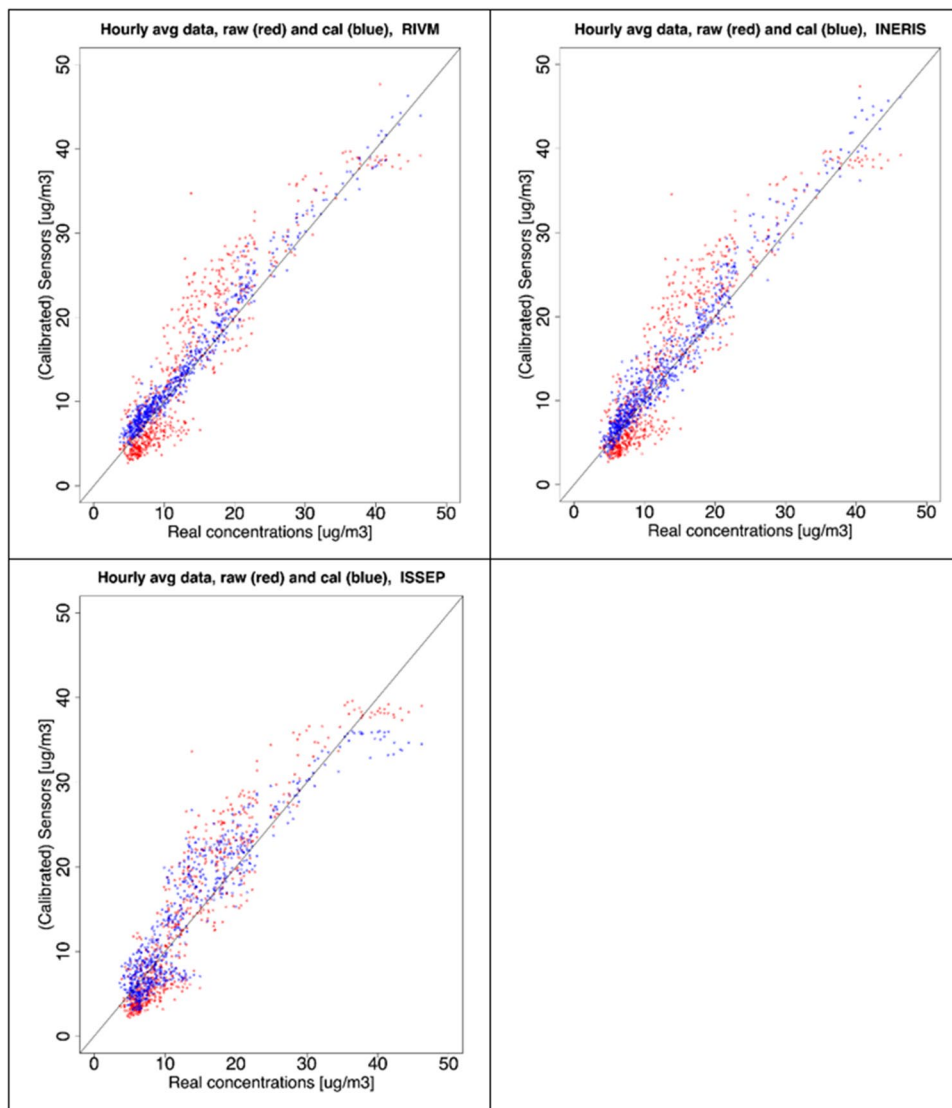

Table 2 Correlation, RMSE, offset, and slope of a fit to the hourly raw and calibrated data for the results of INERIS, ISSeP, and RIVM, shown in Fig. 11

	Correlation R		RMSE		Offset linear fit		Slope linear fit	
	Raw	Cal	Raw	Cal	Raw	Cal	Raw	Cal
RIVM	0.93	0.99	4.0	1.7	2.03	−0.99	0.79	0.99
INERIS	0.93	0.98	4.0	2.1	2.01	−0.31	0.79	0.95
ISSEP	0.92	0.92	3.9	3.3	2.71	0.65	0.77	0.90

results. Removing these high value data in the analysis eliminates the systematic overestimation but may also result in eliminating actually occurring high concentrations. Therefore, we have chosen to include sensors with high concentrations in the analysis.

After the calibration using the method of RIVM, the results follow the diagonal in a relatively small band. The results using the method of INERIS show a similar pattern. The calibration method of ISSeP leads to more scatter along the diagonal, in a band that is practically as wide as that of the raw sensor data. The different results are reflected in the correlation and RMSE, as well as the offset and slope of a fit to the raw and calibrated data in Fig. 11, shown in Table 2.

The table shows slightly different values for the raw sensor data for RIVM, INERIS, and ISSeP. This is due to the fact that the different methods made different selections of outliers that were returned in the processing.

Discussion and concluding remarks

Based on the tests performed and described in this article, several conclusions can be drawn.

First of all, benchmarking is an important and useful process to study the effects of different approaches in the calibration of data from large networks of low-cost sensors.

For large sensor networks, it is not possible to compare the results of raw/calibrated sensors to the results of official fixed measurements at the sensor locations as the official data is simply not available. However, sufficiently realistic synthetic real concentrations and synthetic sensor data can be constructed, and these are valuable for an objective benchmark of different sensor network processing algorithms.

The importance of data cleaning, handling of uncertainty, interpolation, and calibration of low-cost sensors is demonstrated and investigated. The algorithms applied in the benchmark for network calibration can substantially correct the influence of environmental conditions on the performance of the SDS011 PM2.5 sensors. The results obtained by INERIS and RIVM are comparable. The results obtained by ISSeP, based on spatial averages for each hour, show less improvement after the calibration. Looking at the extent of the BIAS, the calibrated ISSeP data is better centered around zero. This is likely due to the fact that the ISSeP method is discarding more data points which is an inherent feature of the method.

The methods employed by RIVM-INERIS are suited for a calibration approach looking for a robust good mean calibration, with tolerance for a few "bad" corrected sensors, whereas the ISSeP method is suited for calibrations with low tolerance for badly corrected sensors.

The SDS011 sensor, used as a basis for the synthetic data, has a large random uncertainty that cannot be corrected by network calibration. As a result, there remains a substantial scatter in the results of the sensors which limits individual use. Combining the calibrated PM2.5-sensor data with existing air quality maps in a data fusion approach is expected to improve the level of detail and the quality of the air quality maps, especially when zooming in spatially and in time. This will be part of the next phase of sensor processing in FAIRMODE Working Group 6; see https://fairmode.jrc.ec.europa.eu/activity/ct6.

Appendix

Annex 1. Data collection for the benchmark

As described before, the data used in the benchmark is based on citizen data in The Netherlands. This annex gives a short technical description of the data collection and processing.

There are no a priori requirements for parties to deliver their data. RIVM receives Citizen Science Air Quality data from a variety of data streams such as the data from the https://sensor.community/en/. The key elements in the hard- and software configuration for handling and storing this data are a Node-Red-server, an InfluxDB-server, and a PostgresDB-server. The Node-Red receives the data and forwards it into an InfluxDB-timeseries database. This InfluxDB-server is also used to receive data from (mostly older) sensor initiatives that use WiFi as a communication medium.

Apart from these push mechanisms, RIVM also pulls data from several external APIs and stores these in InfluxDB. The data is then transferred and harmonized to a PostgresDB-server where it can be combined with other data like sensor types, location data, and official data from Reference Networks.

Fig. 12 Schematic overview of the data flow
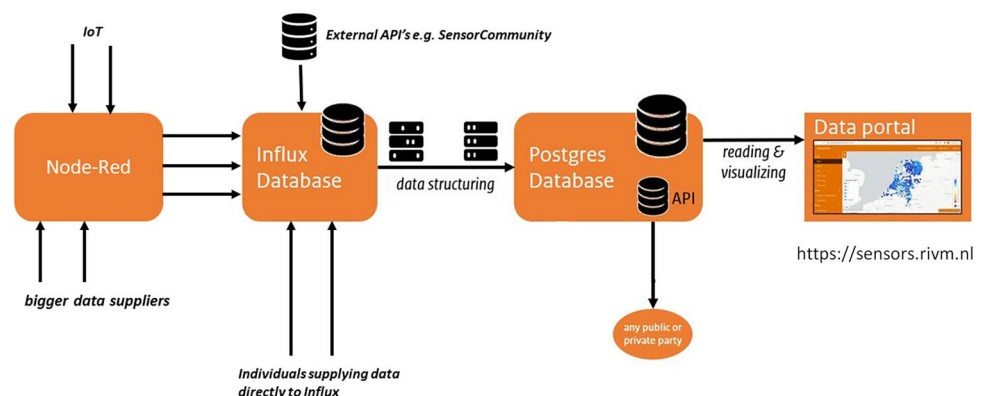
The PostgresDB-server is the source for public services like the web-based data portal and a "SamenMeten" API, https://api-samenmeten.rivm.nl/v1.0. For the FAIRMODE benchmark, the data was also provided in a compressed format for all interested parties to develop and test their calibration methods.

Author contributions All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by AG, PJ, and JW. The first draft of the manuscript was written by all authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval Not applicable, no human or animal subjects were involved in the study.

Consent for publication Not applicable, no human subjects were involved in the study.

Conflict of interest The authors declare no competing interests.

References

EC (2022) Proposal for a directive of the European parliament and of the council on ambient air quality and cleaner air for Europe (recast), COM/2022/542 final, Document 52022PC0542. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2022:542:FIN. Accessed Dec 2023

EEA (2019) Assessing air quality through citizen science. Technical Report EEA Report No 19/2019. European Environment Agency. https://www.eea.europa.eu/ds_resolveuid/5eb208f66c6f40ae88f53b8e663b0e6f. Accessed Dec 2022

EEA (2022) Europe's air quality status 2022. Publications Office of the European Union. https://www.eea.europa.eu/ds_resolveuid/7367c9142c3e4ed9bd77c6f98dee66d9. Accessed Dec 2023

EEA (2023) Europe's air quality status 2023. Technical Report. European Environment Agency. https://www.eea.europa.eu/ds_resolveuid/de4a1ea81f2442a690989e036d83d189. Accessed Dec 2023

EUROSTAT (2023) Key figures on the EU in the world – 2023 Edition. Technical Report. Publications Office of the European Union. https://doi.org/10.2785/57156

González Ortiz A, Guerreiro C, Soares J (2020) Air quality in Europe: 2020 report. EEA Publications Office. https://doi.org/10.2800/786656

Graça D, Reis J, Gama C, Monteiro A, Rodrigues V, Rebelo M, Borrego C, Lopes M, Miranda AI (2023) Sensors network as an added value for the characterization of spatial and temporal air quality patterns at the urban Scale. Sensors 23:1859. https://doi.org/10.3390/s23041859

Janssen S, Thunis P (2020) FAIRMODE guidance document on modelling quality objectives and benchmarking, EUR 30264 EN. Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/30226

Lelieveld J, Pozzer A, Pöschl U, Fnais M, Haines A, Münzel T (2020) Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective. Cardiovasc Res 116:1910–1917. https://doi.org/10.1093/cvr/cvaa025

Lenartz F, Troupin C, Lefebvre W (2018) Data interpolating variational analysis for the generation of atmospheric pollution maps at various scales. In: Mensink C, Kallos G (eds) Air Pollution Modeling and its Application XXV. Springer International Publishing, Cham, pp 231–235

Monteiro A, Sá E, Fernandes A, Gama C, Sorte S, Borrego C, Lopes M, Russo MA (2018) How healthy will be the air quality in 2050? Air Qual Atmos Health 11:353–362. https://doi.org/10.1007/s11869-017-0466-z

Oliveira K, Rodrigues V, Slingerland S, Vanherle K, Soares J, Rafael S, Trozzi C, Bouman EA, Ferreira J, Kewo A, Nielsen PS, Diafas I, Monteiro A, Miranda AI, Lopes M, Hayes E (2022) Assessing the impacts of citizen-led policies on emissions, air quality and health. J Environ Manage 302:114047. https://doi.org/10.1016/j.jenvman.2021.114047

Pisoni E, Christidis P, Thunis P, Trombetti M (2019) Evaluating the impact of "sustainable urban mobility plans" on urban background air quality. J Environ Manage 231:249–255. https://doi.org/10.1016/j.jenvman.2018.10.039

Ratingen Sv, Vonk J, Blokhuis C, Wesseling J, Tielemans E, Weijers E (2021) Seasonal influence on the performance of low-cost no2 sensor calibrations. Sensors 21:7919. https://doi.org/10.3390/s21237919

Rodrigues V, Gama C, Ascenso A, Oliveira K, Coelho S, Monteiro A, Hayes E, Lopes M (2021) Assessing air pollution in European cities to support a citizen centered approach to air quality management. Sci Total Environ 799:149311. https://doi.org/10.1016/j.scitotenv.2021.149311

Schneider P, Bartonova A, Castell N, Dauge FR, Gerboles M, Hagler GSW, Hüglin C, Jones RL, Khan S, Lewis AC, Mijling B, Müller M, Penza M, Spinelle L, Stacey B, Vogt M, Wesseling J, Williams RW (2019) Toward a unified terminology of processing levels for low-cost air-quality sensors. Environ Sci Technol 53:8485–8487. https://doi.org/10.1021/acs.est.9b03950

Thunis P, Degraeuwe B, Pisoni E, Trombetti M, Peduzzi E, Belis CA, Wilson J, Clappier A, Vignati E (2018) PM2.5 source allocation in European cities: a SHERPA modelling study. Atmos Environ 187:93–106. https://doi.org/10.1016/j.atmosenv.2018.05.062

Troupin C, Barth A, Sirjacobs D, Ouberdous M, Brankart J-M, Brasseur P, Rixen M, Alvera-Azcárate A, Belounis M, Capet A, Lenartz F, Toussaint M-E, Beckers J-M (2012) Generation of analysis and consistent error fields using the Data Interpolating Variational Analysis (DIVA). Ocean Modelling 52–53:90-101. https://doi.org/10.1016/j.ocemod.2012.05.002

Troupin C, Ouberdous M, Sirjacobs D, Alvera-Azcárate A, Barth A, Toussaint ME, Watelet S, Beckers JM (2013) Diva User Guide

VACUUMS (2021) Performance evaluation of six low-cost particulate matter sensors in the field, VACUUMS project. https://vaquums.eu/test-results. Accessed Dec 2023

Viana M, de Leeuw F, Bartonova A, Castell N, Ozturk E, González Ortiz A (2020) Air quality mitigation in European cities: status and challenges ahead. Environ Int 143:105907. https://doi.org/10.1016/j.envint.2020.105907

Volten H, Devilee J, Apituley A, Carton L, Grothe M, Keller C, Kresin F, Land-Zandstra A, Noordijk E, Putten EM, Rietjens J, Snik F, Tielemans E, Vonk, J, Voogt M, Wesseling J, (2018) Chapter in book "Enhancing national environmental monitoring through local citizen science". pp. 337–352 https://www.ucl.ac.uk/ucl-press/browse-books/citizen-science.

Weissert LF, Henshaw GS, Williams DE, Feenstra B, Lam R, Collier-Oxandale A, Papapostolou V, Polidori A (2023) Performance evaluation of MOMA (MOment MAtching) - a remote network calibration technique for PM2.5 and PM10 sensors, Atmospheric. Meas Tech 16:4709–4722 https://amt.copernicus.org/articles/16/4709/2023/

Wesseling J, Hendricx W, de Ruiter H, van Ratingen S, Drukker D, Huitema M, Schouwenaar C, Janssen G, van Aken S, Smeenk JW et al (2021) Assessment of PM2.5 exposure during cycle trips in The Netherlands using low-cost sensors. Int J Environ Res Public Health 18:6007. https://doi.org/10.3390/ijerph18116007

Wesseling J, de Ruiter H, Blokhuis C, Drukker D, Weijers E, Volten H, Vonk J, Gast L, Voogt M, Zandveld P et al (2019) Development and implementation of a platform for public information on air quality, sensor measurements, and citizen Science. Atmosphere 10:445. https://doi.org/10.3390/atmos10080445

WHO (2021) WHO global air quality guidelines: particulate matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization

WMO (2020) An update on low-cost sensors for the measurement of atmospheric composition. technical report WMO report no. 1215. World Meteorological Organization

Woutersen A, de Ruiter H, Wesseling, J, Hendricx W, Blokhuis C, van Ratingen S, Vegt K, Voogt M (2022) Farmers and local residents collaborate: application of a participatory citizen science approach to characterising air quality in a rural area in The Netherlands. Sensors 22:8053. https://doi.org/10.3390/s22208053

van Zoest VM, Stein A, Hoek G (2018) Outlier detection in urban air quality sensor networks. Water Air Soil Pollut 229:111. https://doi.org/10.1007/s11270-018-3756-7

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.