IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 5, 2022, accepted January 17, 2022, date of publication January 25, 2022, date of current version February 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3145945

Stochastic Online Calibration of Low-Cost Gas Sensor Networks With Mobile References

GEORGI TANCEV AND FEDERICO GRASSO TORO
Swiss Federal Institute of Metrology, 3003 Wabern-Bern, Switzerland

Corresponding author: Georgi Tancev (georgi.tancev@metas.ch)

This work was supported in part by Innosuisse under Project 36779.1 IP-ENG.

ABSTRACT There has been a wide interest in high-resolution air quality monitoring with low-cost gas sensor systems in the last years. Such gas sensors, however, suffer from cross-sensitivities, interferences with environmental factors, unit-to-unit variability, aging, and concept drift. Therefore, reliability and trustworthiness of the measurements in the low parts-per-billion (ppb) range remain a concern, particularly over the course of the lifetime of a sensor network in urban environments. In this simulation study, the possibility to continuously recalibrate a wireless sensor network with mobile references and stochastic gradients, computed from encounters, is explored. By using data collected in field experiments, encounters between static and mobile nodes are modeled as a probabilistic process. Moreover, the influence of a collection of design parameters such as base calibration, initial recalibration, choice of optimization algorithm, as well as encounter frequency are analyzed and discussed. With an optimized protocol, it can be shown that long-term reliable measurements with absolute errors of about 50 ppb for CO, 3 ppb for NO_2, and 4 ppb for O_3 could be achievable with a few mobile references in urban environments.

INDEX TERMS Air quality monitoring, calibration, gas sensor, Internet of Things, low-cost, online learning, wireless sensor network.

I. INTRODUCTION

Due to the health impact of low air quality [1], [2], a lot of research with low-cost gas and particulate matter sensors for high-resolution air quality monitoring has been conducted in the last years [3]–[7]. Unit-to-unit variability [8], interferences with other gases and environmental parameters [9], [10], as well as aging [10], [11] are common problems of such sensors.

For air quality monitoring, however, data quality objectives imposed by legislators must be met [12]. Thus, researchers came up with the idea of combining an array of different sensors into so-called low-cost sensors systems with the purpose of compensating interfering effects with models obtained from machine learning algorithms [13], [14] (e.g., neural networks [4] or random forests [9]) and field data. Unfortunately, this generally leads to non-representative models followed by concept drift [15], [16]; the environmental conditions vary over time and space, so the calibration parameters need to change frequently. Hence, maintaining reliability and trustworthiness over the course of the lifetime of an air quality sensor network remains a challenge.

Traditionally, measurement instruments are recalibrated periodically against references provided by authorities to maintain trustworthiness and to assign measurement uncertainties [17]. For wireless sensor networks, however, such a workflow does not scale, and dedicated network calibration methods have been developed in the last decade. A recently published survey by Maag *et al.* [18] summarizes proposed sensor network calibration algorithms and their suitability for air quality monitoring applications.

For instance, blind calibration approaches [19]–[22] lack the (legally required) information on the measurement uncertainty and appear to not work for low-cost gas sensors [18]. Multi-hop calibration approaches with static references and mobile sensors [23]–[25] generally lead to the propagation of errors [26], but additional error sources should be avoided at all costs. Moreover, low-cost gas sensors have response times in the range of $30 - 90$ s [27]–[29], which is considerably long for mobile nodes. Alternatively, mobile reference instruments, i.e., reliable mobile devices mounted on vehicles, could continuously monitor and recalibrate static low-cost sensor nodes [30].

The associate editor coordinating the review of this manuscript and approving it for publication was Chan Hwang See.

Regarding data transmission, Saukh *et al.* [24] propose that static and mobile nodes submit measurements, indexed by time and location, to a database on the cloud at independent frequencies; they define a "rendezvous" as a time interval in which two nodes have time and space distances below certain thresholds. Note that both distances can be minimized if static nodes were placed close to stops where vehicles spend about half a minute.

This definition is particularly useful if recalibration is performed periodically in a batch so that a database can be queried for all measurements of interest. Nonetheless, this adds an additional error due to the imprecision of the indexes. In addition, the larger the allowed spatial or temporal distance for a rendezvous, the higher the additional measurement uncertainty will be. If the recalibration frequency is low, more recent measurements should receive a higher weight [31]. From a metrological point-of-view, however, a low recalibration frequency leads to less representative parameters.

While performing recalibration in batches (consisting of many encounters) is well-established, treating network calibration as an online learning (i.e., streaming data) problem [32] has not been considered so far. In a recently published conference paper [31], the idea to sequentially recalibrate nodes as part of a sensor network with mobile references and stochastic gradient descent (SGD) [33], [34] was briefly sketched. On one side, such a lightweight protocol would account for sensor aging and the rapid changes in the atmospheric conditions (i.e., concept drift) yet be robust to single anomalies (e.g., sudden artifacts in the signals).

One the other side, a traceable measurement uncertainty could be associated with each device so that measurements can be interpreted properly [17], [35]. Furthermore, due to the success of deep learning in the last decade [36], more advanced update rules like RMSProp (possibly with additional momentum of the gradient) [33], have been developed, which should be considered as well.

This work builds upon the concept and evaluates such a protocol for gas sensors using field data and simulated encounters under different base calibrations, initial recalibration, algorithms, and encounter frequencies. The paper is structured as follows. In a first step, the problem and its design parameters are presented formally, and the generalized SGD for online calibration is illustrated. Furthermore, the used experimental field data and the performed simulations are rigorously described. In a next step, the obtained results and possible limitations are presented and discussed. Finally, the paper closes with a conclusion and an outlook on future work.

II. MATERIALS AND METHODS

A. PROBLEM DEFINITION

The calibration process aims is to find the set of optimal calibration parameters $W \in \mathbb{R}^{(p+1)\times q}$ that map the sensor signal s to the reference data r, i.e., $\hat{r} = sW$. In this study, the calibration model was fixed to a linear regression with model input $s = (1, s_1, \ldots, s_p) \in \mathbb{R}^{1\times(p+1)}$ and model output $r = (r_1, \ldots, r_q) \in \mathbb{R}^{1\times q}$. Note that the model input contains a "1" for the intercept. (In principle, any other model that can be trained via SGD is also possible.)

The problem of stochastic online calibration with mobile references over a node lifetime T was modeled according to the scheme depicted in Figure 1.

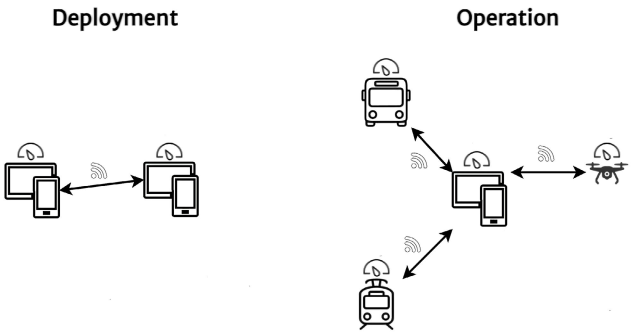


FIGURE 1. Schema of the protocol.

During a deployment phase, a low-cost sensor system with a base calibration W_b (i.e., initial calibration parameters) is synchronized with a mobile reference system for a short time period $h \ll T$ (via some wireless technology standard, treated as black box in the following) so that it adapts to its new environment.

At time point t, the tuple (s_t, r_t) of sensor and reference signals is collected. With every collected tuple, an online recalibration procedure based on SGD is performed, thereby updating the set of calibration parameters W.

Afterwards, the operational phase begins. Several reference instruments mounted on arbitrary vehicles, e.g., trams or buses, have encounters with static nodes, leading to comparison of sensor and reference values. At every encounter, the same tuple of sensor and reference signals is collected for a recalibration iteration.

In general, finding a map for continuous output variables can be achieved via least squares optimization. Since the output variables can span across different orders of magnitude (i.e., some pollutants are more abundant than others), it might be reasonable to weight their loss contributions by factors $e_1, \ldots, e_q$, stored in a matrix $E = diag(e_1, \ldots, e_q) \in \mathbb{R}^{q\times q}$. Hence, with a collection of n measurements, i.e., $S \in \mathbb{R}^{n\times(p+1)}$ and $R \in \mathbb{R}^{n\times q}$, the loss L (in matrix notation, tr refers to the trace) to be minimized is given in (1).

$$\min_W L = \frac{1}{2} tr((SW - R)^T (SW - R)E) \tag{1}$$

In case of very large data sets or streaming data, the optimal solution is best found via SGD [32], [33]. The gradient g, i.e., the derivative of the loss with respect to the model parameters, is given in (2).

$$g = \frac{dL}{dW} = S^T (SW - R)E \tag{2}$$

Algorithm 1 describes a generalized version of SGD for the stochastic online calibration of network nodes, i.e., after one encounter. ($\odot$ is the Hadamard operator.) In the deep learning literature [33], this procedure is known as RMSProp with momentum (RMSProp-m). The variable v is responsible for memorizing the momentum of the gradient; it is controlled by $\alpha \in [0, 1]$ (the momentum parameter). A high value of α accelerates the gradient and speeds up convergence when several consecutive incorrect predictions are performed [33]. The variable a stores a discounted moving average of the squared gradient and is controlled by $\beta \in [0, 1]$ (the decay rate). With a high value of β, previous (in)correct predictions are remembered for adapting the individual learning rates. Finally, $\gamma \in \mathbb{R}^+$ is the step size (the global learning rate).

With $a \leftarrow 1$, $\alpha = 0$, and $\beta = 1$, vanilla SGD is obtained (i.e., without momentum and adaptive learning rate). Setting only $\alpha = 0$ results in RMSProp (i.e., without momentum). Note that the square root operation is applied element-wise (i.e., Hadamard square root). ϵ is a stability constant and should be fixed to a value that is much smaller than the gradients. Therefore, the design parameters of the protocol are W_b, h, E, α, β, γ, and the encounter frequency.

Algorithm 1 Stochastic Online Calibration

```
procedure Calibrate($W_b, E, \alpha, \beta, \gamma$)
    $W \leftarrow W_b$
    $v \leftarrow 0$
    $a \leftarrow 0$
    $\epsilon \leftarrow 10^{-8}$
    for $t \leftarrow 1, T$ do
        if ($s_t, r_t$) collected then                #At encounter.
            $W_i \leftarrow W + \alpha v$
            $g \leftarrow s_t^T (s_t W_i - r_t) E$
            $a \leftarrow \beta a + (1 - \beta) g \odot g$
            $v \leftarrow \alpha v - \frac{\gamma}{\sqrt{a + \epsilon}} \odot g$
            $W \leftarrow W + v$
        end if
    end for
end procedure

function Measure($s_t, W$)
    $\hat{r}_t \leftarrow s_t W$
    return $\hat{r}_t$
end function
```

B. DATA

As data set, the field study conducted by Zimmerman *et al.* [9] was used. The data set was collected at an urban background site from August 2016 to February 2017 and consists of quarter-hourly measurements from 19 low-cost sensor systems. References values for carbon monoxide (CO), nitrogen dioxide (NO_2), and ozone (O_3) in the parts-per-billion (ppb) range (Table 1) are available from the second month (i.e., October). Considering that the lifetime of such sensors is roughly six to twelve months [11], the duration is sufficient to make reasonable statements about the efficacy of the protocol. There are even reports about relevant drift after one month of operation [37].

TABLE 1. Summary of the reference distribution.

	$Q_{0.05}$ ppb	$Q_{0.25}$ ppb	$Q_{0.50}$ ppb	$Q_{0.75}$ ppb	$Q_{0.95}$ ppb
$CO^{\text{ref.}}$	121	157	190	264	663
$NO_2^{\text{ref.}}$	3	6	9	14	27
$O_3^{\text{ref.}}$	4	14	21	29	43

Each low-cost sensor system contained the widely used electrochemical sensors CO-B41 [29], NO_2-B43F [28], and O_x-B431 (NO_2 and O_3 combined) [27] from Alphasense as well as sensors for temperature (T) and relative humidity (RH). According to the study authors, the sensor outputs were measured with a custom-designed electronic circuit board and optimized for signal stability. More precisely, said board comprised custom electronics to operate the device, multiple stages of filtering, and an analog-to-digital converter. In addition, the data were logged at a rate of 4 per minute but downsampled to 4 per hour by averaging.

For low-cost gas sensor systems, those are important requirements to minimize the noise. In practice, a network node would sample raw sensor signals $\tilde{s}$ at a predefined frequency and compute an average in an online fashion, thereby avoiding to store all values. With K required samples, the average $\mu_{\tilde{s}}$ with the k-th raw signal $\tilde{s}_k$ is given in (3).

$$\mu_{\tilde{s}} \leftarrow \mu_{\tilde{s}} + \frac{\tilde{s}_k}{K} \tag{3}$$

After all samples have been collected, this computation terminates; the most recent average is memorized, serving as sensor input for the next encounter with a mobile reference (i.e., $s_t \leftarrow \mu_{\tilde{s}}$, $\mu_{\tilde{s}} \leftarrow 0$).

Out of the 19 devices, three low-cost sensor systems (#4: device 1, #16: device 2, and #17: device 3) have the least values missing. In order to keep bias as low as possible, only these three systems have been considered in the analysis. Specifically, the optimal protocol was developed solely with device 1 and validated on the data from the remaining two. The measurements from devices 2 and 3 stop one month earlier (i.e., in January). Missing values between start and stop have been imputed with forward filling, since this was seen as an opportunity to simulate potential erroneous events during operation. An overview of the data set is shown in Figure 7 (Appendix A).

In their original paper [9], the authors provided two population calibrations for each sensor type (Table 2), i.e., parameters obtained from a collection of sensors that fits well on average but not necessarily for every unit; a simple (s) laboratory calibration (CO: 0-1600 ppb, NO_2: 0-50 ppb, in 3-4 points), and an extended (e) field calibration including parameters for the interferences with T and RH (calibration distribution described in Table 1).

Note that neither accounts for cross-sensitivities with other pollutants. Unfortunately, no simple calibration was made

TABLE 2. Simple (s) and extended (e) population calibrations for the three gas sensors with intercept w_0, sensitivity w_s, and corrections w_T, w_{RH}.

Sensor	Calibration	$\frac{w_0}{\text{ppb}}$	$\frac{w_s}{\text{ppb / mV}}$	$\frac{w_T}{\text{ppb / °C}}$	$\frac{w_{RH}}{\text{ppb / \%}}$
CO-B431	s	-119.0	0.8	—	—
	e	32.0	1.3	-1.1	-0.1
NO_2-B433F	s	-14.0	0.6	—	—
	e	3.9	1.2	0.1	-0.1
O_x-B43F	s	-14.0	0.6	—	—
	e	9.4	0.9	0.1	-0.2

available for the O_x-B431 sensor, so the values from the NO_2-B43F sensor had to be used instead. This can be justified by the fact that the data sheets [27], [28] suggest similar calibration parameters. In the following, both population calibrations were examined as potential base calibrations before deployment.

C. SIMULATIONS

The initial recalibration period during deployment (i.e., h) was fixed to 24 hours. Furthermore, errors for NO_2 and O_3 have been weighted by a factor of ten, since they are more relevant [2]. Where not further specified, values for the hyperparameters of the algorithms were taken from the literature (i.e., $\alpha = 0.9$, $\beta = 0.99$) [33], or determined in preliminary experiments (i.e., $\gamma = 10^{-7}$ for SGD, $\gamma = 10^{-3}$ for RMSProp and RMSProp-m).

Encounters between mobile references have been simulated via Bernoulli trials. At each time point t, there is a probability π for an encounter [31]. For $t \leqslant h$, π is equal to 1. Since the data set consists of roughly 100 data points per day, a value of $\pi = 1$ corresponds to an average encounter rate of 100 per day. Thus, different encounter rates were modeled in this manner. Since this is a stochastic process, 100 such simulations were performed to yield average results, thereby removing the influence of single encounters.

As metric for the performance of a low-cost sensor system, the absolute error for every pollutant $i \in \{CO, NO_2, O_3\}$ was computed. It is defined as the average absolute deviation of the predicted concentration from the actual reference value over the period T ignoring missing values (4).

$$\Delta_i = \frac{1}{T} \sum_{t=1}^{T} |r_{i,t} - \hat{r}_{i,t}| \tag{4}$$

III. RESULTS AND DISCUSSION

Figure 2 illustrates the performance of device 1 with the three algorithms under different encounter frequencies using a simple base calibration, whereas Figure 3 shows the same result with an extended base calibration. The dotted line is the base case error; it is the performance that is obtained if no further adjusted is made (i.e., no recalibration during deployment or operation).

The offset in the absolute error at low encounter frequency is the result from the recalibration during deployment. This observation suggests that a recalibration for a few hours during deployment is actually beneficial, as the absolute error generally decreases. This improvement has two reasons. On the one hand, the base calibration is a population calibration that does not necessarily fit for a specific unit.

On the other hand, no corrections for interferences with T and RH are included in the simple base calibration model. Specifically, the inclusion of T and RH corrections lowers the base case error (dotted line) for NO_2 and O_3 significantly in comparison with the simple calibration. In general, there is a large consensus that such compensations are necessary [3]. Figure 8 shows how such corrections are introduced during deployment. However, the sensitivities and intercepts of the extended calibration are also different (Table 2), so they might better represent these sensors.

The subplots in both Figures illustrate that RMSProp generally leads to the highest decrease in absolute error for NO_2 and O_3, whereas RMSProp-m leads to the highest decrease for CO. Further increase of the performance can only be achieved in case of several encounters per day. The reason is that either the deployment duration is too short so that calibration parameters are not yet optimal, or that they need adjustment due to aging and concept drift. For example, Figure 9 illustrates how the model parameters continuously evolve over time.

With RMSProp-m, the performance even decreases for NO_2 and O_3 at low to moderate encounter frequencies. In this case, it appears that momentum only benefits adjusting the calibration parameters of CO. Furthermore, since erroneous signals are possible (e.g., due to imputed missing values), they could also increase the absolute error at moderate encounter rates. Nonetheless, the methodology seems to be quite robust to such events, since the performance still increases at the highest encounter frequencies.

The error reduction could be probably even lower without these events; some encounters surely contain more information than others do. Hence, it might be a good idea to filter out erroneous instances before performing gradient descent. Completely faulty nodes, on the other hand, could be identified by monitoring the absolute error over time.

For CO, however, the highest concentrations are not correctly predicted, as shown in Figure 10. The explanation for the higher error of the CO measurements might be that high concentrations were not properly covered by the base calibrations; because the upper limit of the simple calibration lies higher compared to the extended calibration, it performs slightly better for CO (Figures 2 and 3). More precisely, support is lacking for high concentrations, and without these upper levels, the hyperplane might be improperly oriented, thereby resulting in predictions of low quality.

The question arises whether some better base calibrations could be obtained from experiments with orthogonal variables [15]. In field experiments, calibration ranges can not be chosen and all factors of interest are usually correlated. With calibration models obtained from field data, it is even possible to ‘‘measure’’ any pollutant with any sensor if correlations are

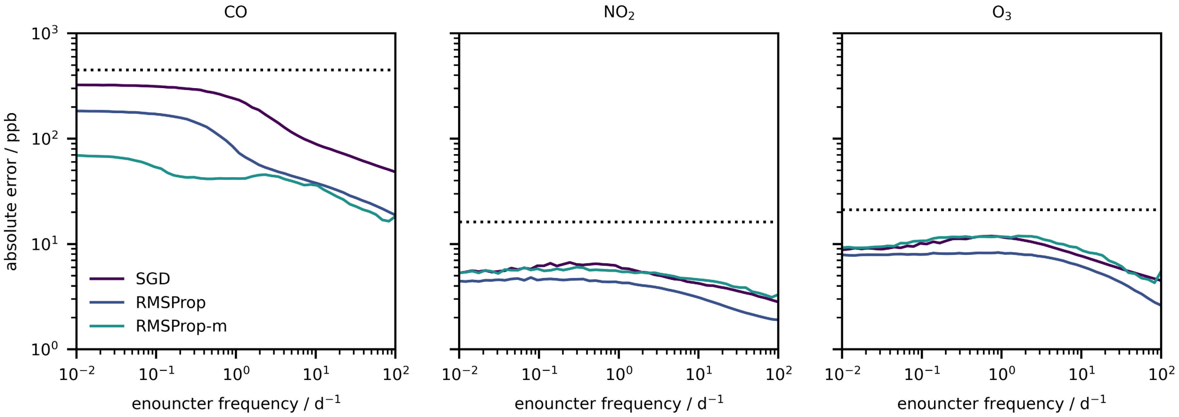


FIGURE 2. Absolute error from a simple base calibration as a function of algorithm and encounter frequency (100 simulations per encounter frequency). The dotted line refers to the absolute error without any recalibration. An offset is the result from the initial recalibration during deployment.

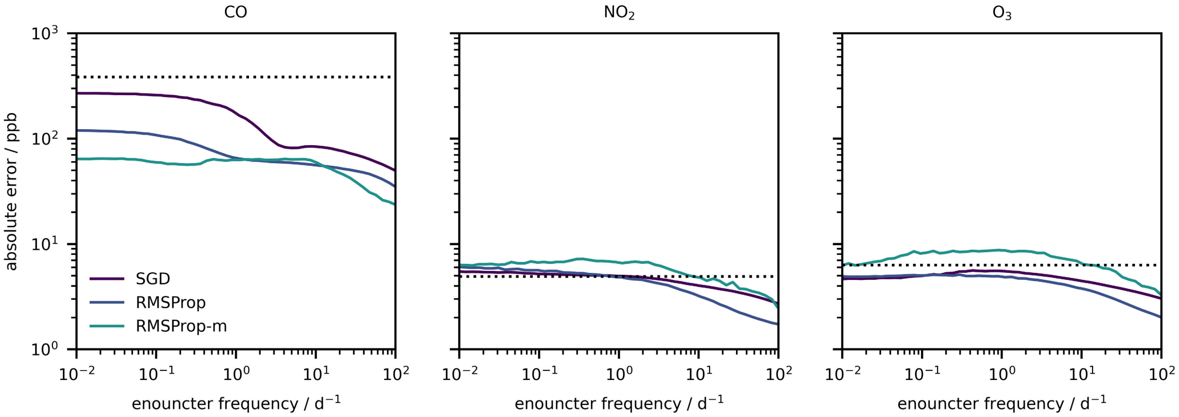


FIGURE 3. Absolute error from an extended base calibration as a function of algorithm and encounter frequency (100 simulations per encounter frequency). The dotted line refers to the absolute error without any recalibration. An offset is the result from the initial recalibration during deployment.

very strong [15]. Thus, it is challenging to find representative calibration parameters using field data.

Since SGD and RMSProp-m bring no additional benefit over RMSProp for the two relevant pollutants NO_2 and O_3, both algorithms were not considered in the further analysis. Moreover, the hyperparameters of the algorithm have not been optimized thus far. Figure 4 illustrates the error landscape (in ppb) for the three pollutants under different combinations of the hyperparameters.

It shows that the situation could be improved by implementing individual hyperparameters. By decreasing β and increasing γ, the performance can be improved for NO_2 and O_3. Moreover, predicting CO with low error requires high values of β and γ. In this case, the optimal hyperparameters should be fixed to $\beta = 0.8$ and $\gamma = 0.005$ for NO_2/O_3 as well as $\beta = 0.999$ and $\gamma = 0.01$ for CO.

Because the hyperplane might be improperly oriented initially, large consecutive gradient updates are required to reorient it, as the highest concentrations occur only rarely. Consequently, there is an interaction with the inadequate base calibration and the choice of the optimization algorithm. Alternatively, a "better" base calibration covering the full range might allow the same hyperparameters for all pollutants.

Figure 5 illustrates the agreement between sensor and reference measurements for device 1 using RMSProp with the set of optimal hyperparameters and ten encounters per day, starting with an extended calibration. The agreement is exceptionally good, considering that the measurements are coming from low-cost gas sensors.

The final absolute errors from device 1, corresponding to the standard uncertainties in metrology [35], are about 50 ppb for CO, 3 ppb for NO_2, and 4 ppb for O_3. In relation to the medians, these uncertainties are 26% for CO, 33% for NO_2, and 19% for O_3. Due to uncertainty propagation, the uncertainty from the reference instrument adds up to the one

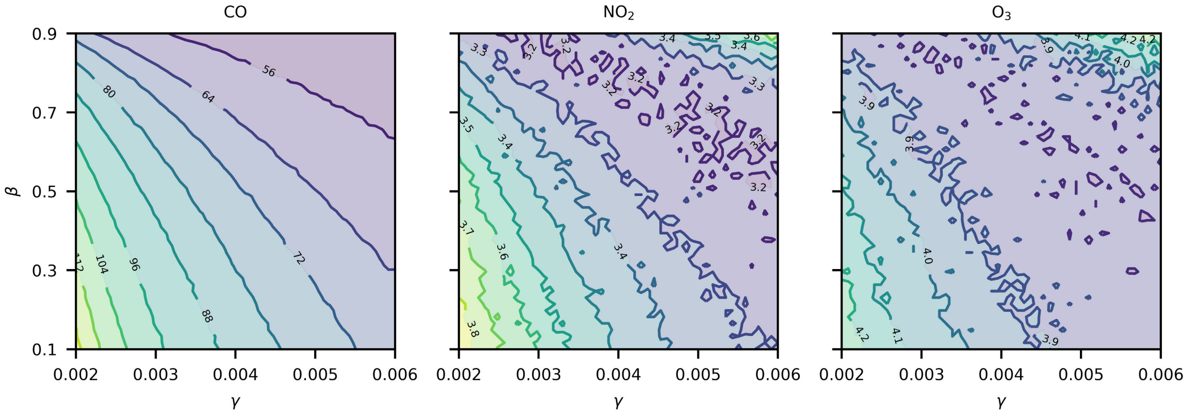


FIGURE 4. Absolute error (ppb) as a function of hyperparameters β and γ using an extended calibration and RMSProp at ten encounters per day (50 simulations per parameter combination).

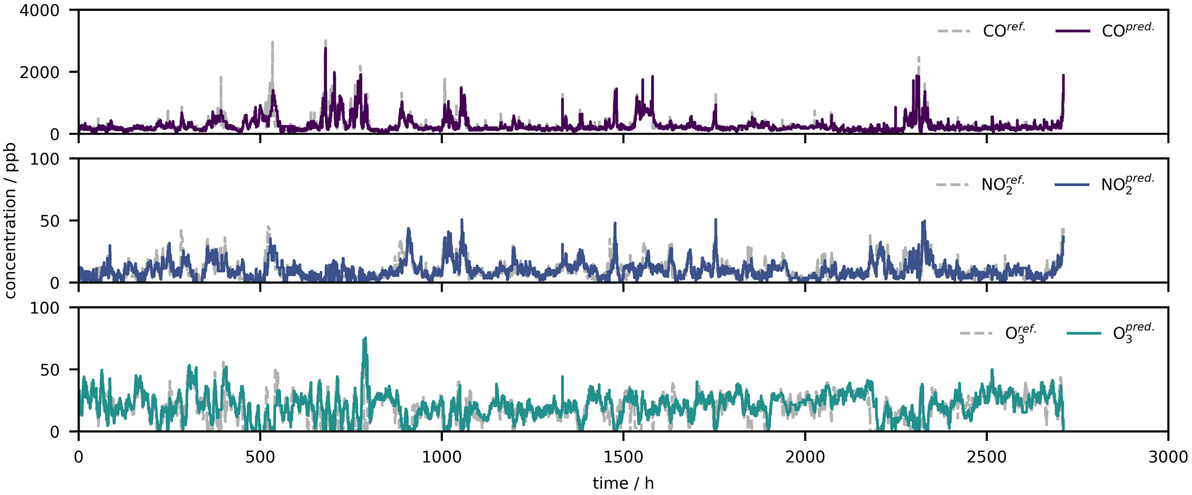


FIGURE 5. Performance of device 1 over time with optimal hyperparameters at ten encounters per day.

from the low-cost sensor system [35]; a sample calculation is given in Appendix B. To put it into legal context [12], the maximum allowed expanded measurement uncertainties are 25% for CO and NO_2 as well as 30% for O_3. The higher the pollution levels, the easier it is to meet these requirements.

Finally, since an optimal protocol has been found, it can be validated on the remaining data from devices 2 and 3. Figure 6 evaluates the absolute error and its reduction by switching from a simple base calibration without any further recalibration to an extended base calibration with an initial recalibration during deployment as well as recalibration during operation using RMSProp (with optimal hyperparameters) at ten encounters per day. It shows that this reduction is of similar magnitude for the other two devices, hence supporting the theoretical concept.

In the original study [9], the authors developed several different calibration models from the collected field data. They claimed that they could reduce the absolute errors to 8 ppb for CO and even below 1 ppb for NO_2 and O_3 with random forest models. With these calibration models, the absolute errors suddenly increased to values of about 49 ppb for CO, 5 ppb for NO_2, and 3 ppb for O_3 during an independent test at another location. Thus, their absolute errors were much higher during test time. Moreover, with purely linear models, the errors were generally higher.

On the one hand, this observation suggests that their machine learning models captured the local atmosphere of the location at which the models were developed [15], since the performance decreased at the new location. The underlying problem is that the relationships between the pollutants and/or environmental factors are different at other locations or at other time points. Therefore, less reliable measurements can be expected upon relocation of field-calibrated low-cost sensor systems [15].

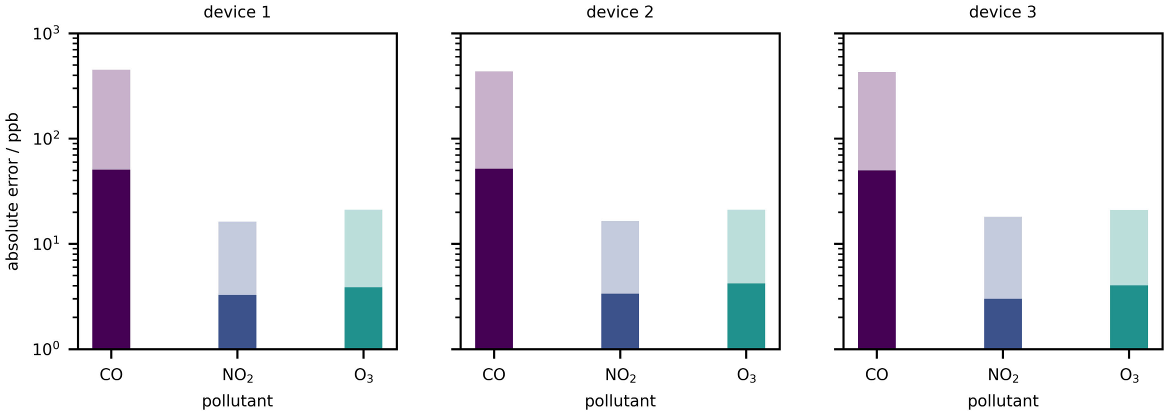


FIGURE 6. Reduction of the absolute error by switching from simple base calibration without recalibration to an optimal protocol at ten encounters per day.

On the other hand, the low error that they achieved with random forest models advocates basis expansion (i.e., the introduction of power and interaction terms). In particular, random forests (and neural networks) are great at capturing non-linear behavior without explicit basis expansion. Consequently, describing potential non-linearity should further decrease the absolute error. An example for a non-linearity would be the interaction between sensitivity and temperature that is even reported in the data sheets [27]–[29]. (This also motivated calibrating low-cost gas sensors with machine learning algorithms in the first place [3].)

With these results in mind, the question arises as to what extent several encounters per day are realistic. In the case of the city Basel in Switzerland, for example, trams circulate for about 20 hours per day and there are about eight relevant lines. A tram associated to a line has about 45 to 60 minutes from one terminal to the other.

Therefore, it can be expected that the encounter frequency should be up to 20 per day if one tram per line would be equipped with a mobile reference. Thus, operating in this regime with eight such references would be not too far-fetched, though the question arises what kind of reference instruments could possibly be used, as such devices should be affordable yet reliable [30].

In contrast to using only mobile reference instruments mounted on trams or buses, the temporal resolution with additional static nodes can be much higher. In this manner, down-sampling can be performed when a high data acquisition rate is available, which essentially lowers the noise in the measurements. In addition, multiple data points per hour can be made available.

The presented outcomes demonstrate that stochastic online calibration with reliable mobile references bears the potential for long-term accurate measurements from low-cost gas sensors, since unit-to-unit variability, aging, and concept drift can be continuously compensated. Although the algorithms were only applied to low-cost gas sensors system within one data set, it can be expected that the concept generalizes to arbitrary sensor systems and networks, since SGD was successfully applied in several distinct online learning scenarios before [32].

Despite these promising results, it is also important to point out some of their shortcomings. In particular, several additional error sources can be expected in real-world scenarios. For instance, the response times of the measurement instruments have been assumed to be zero. Due to response times of the reference instruments, a spatial carryover equals to $s_l = v_m t_r$ (with vehicle velocity v_m and response time t_r) would result in a real-world scenario. Hence, a short response time can be seen as one requirement for candidate reference devices. A complete list of requirements would help to identify an existing product or guide the development of a new one.

Although every reader intuitively understands the term ‘‘encounter’’, it is not absolutely defined. During an encounter, the distance first decreases and then increases again. In practice, an encounter between nodes requires machine-to-machine communication and the range depends on the chosen wireless technology standard (e.g., Bluetooth 5 has ranges up to 200 m [38]). Hence, there is a also time window for data exchange. The larger the allowed distance, the less representative a communicated reference measurement is.

In the presented schema (Figure 1), every encounter triggers a computation, thereby consuming a small amount of energy, which might not always be available. Alternatively, before updating the parameters, the loss in (1) could be computed to assess whether a gradient update would be even required, e.g., by setting a minimum error threshold. (Intuitively, no gradient step is necessary if predictions and reference values coincide.)

Yet another option would be that the stationary node transmits the sensor data together with the current calibration parameters so that the mobile reference can perform all

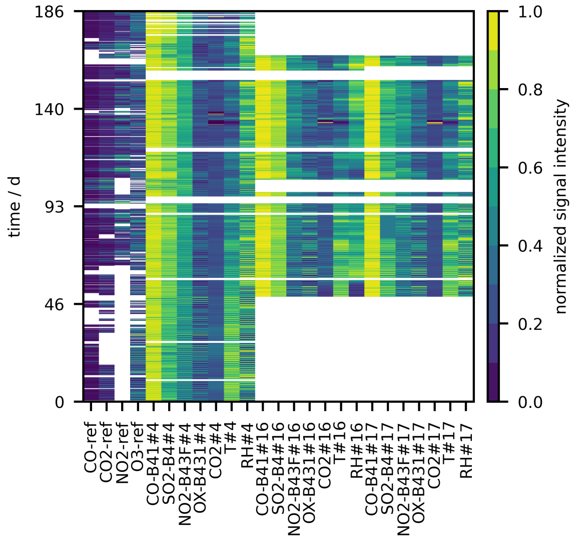


FIGURE 7. Overview of the data set. White regions correspond to missing measurements.

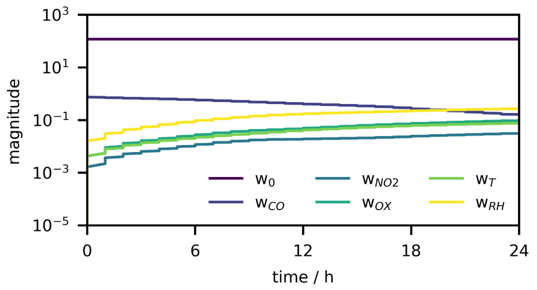


FIGURE 8. Adaptation of the calibration parameters (absolute values) from simple base calibration for CO during deployment (i.e., within the first 24 hours) with SGD.

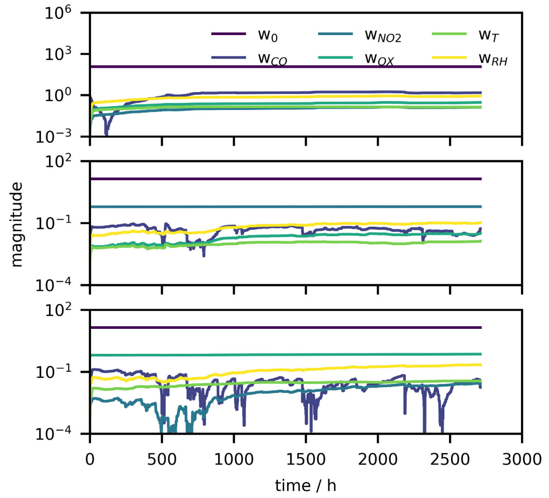


FIGURE 9. Adaptation of the calibration parameters (absolute values) from simple base calibration for CO, NO_2 and O_3 (from top to bottom) during operation (SGD at ten encounters per day).

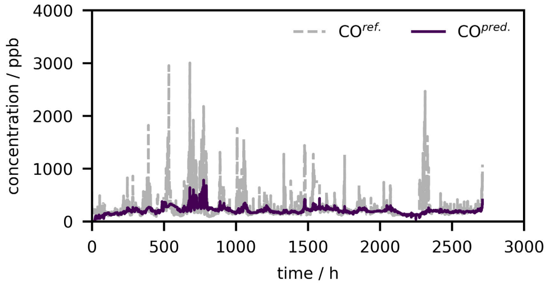


FIGURE 10. Agreement between reference and prediction of CO from device 1 over time (RMSProp with $\beta = 0.99$ and $\gamma = 0.001$) at ten encounters per day.

computations. Nonetheless, transmitting data also requires energy. For an upcoming real-world solution, other aspects such as price and availability of references, network overhead, energy consumption of nodes, or costs of computations on the cloud need to be considered as well.

IV. CONCLUSION AND OUTLOOK

This work expanded the concept of sequentially recalibrating nodes of a low-cost gas sensor network for air quality monitoring with mobile references and stochastic gradients. With such mobile references and the proposed algorithm, network nodes could be monitored and continuously recalibrated, thereby aiding to maintain trust in the measurements.

A proper base calibration is crucial for the success of the approach. Thus, characterizing a small population of low-cost gas sensor systems to obtain an adequate base calibration might be an effort worth taken. As of now, this is mostly achieved in lengthy field campaigns, but establishing it as an efficient and inexpensive service in a laboratory setting with orthogonal variables is planned [15].

It could be shown that an initial recalibration during deployment is beneficial for a low-cost gas sensor system, as every device needs unit-specific calibration parameters. Moreover, a calibration model should also include compensations for interfering variables and cover the range of interest.

If the calibration parameters change over time due to aging and concept drift, they are adjusted accordingly. Moreover, it could be demonstrated that the choice of gradient update rule matters, since RMSProp performed better than vanilla SGD. With optimal hyperparameters and an encounter frequency of up to 20 per day, the absolute error can be reduced to about 50 ppb for CO, 3 ppb for NO_2, and 4 ppb for O_3 by performing gradient descent updates after encounters.

Finally, future work should focus on filtering out erroneous instances to further increase the performance as well as defining requirements for adequate mobile reference systems. If no commercially available devices meet the requirements, novel ones could be developed. Once potential references have been determined, the proposed protocol needs to be validated in field studies.

APPENDIX A
SUPPLEMENTARY FIGURES

See Figs. 7–10.

APPENDIX B
UNCERTAINTY PROPAGATION

With K standard uncertainty sources u_j, $j \in \{1, \ldots, K\}$, the expanded uncertainty U of a measurement instrument is defined in (5) [35].

$$U = 2 \times \sqrt{\sum_{j=1}^{K} u_j^2} \tag{5}$$

Typically, references for air quality (e.g., NO_2) have standard uncertainties of 5% [39]. With 30% of standard uncertainty coming from a low-cost sensor, the resulting expanded uncertainty is $U = 2 \times \sqrt{(30\%)^2 + (5\%)^2} = 61\%$.

AUTHOR CONTRIBUTIONS

Georgi Tancev performed the analysis and wrote the manuscript. Federico Grasso Toro reviewed the manuscript.

ACKNOWLEDGMENT

The authors would like to thank Zimmerman *et al.* for providing the raw data.

CODE AVAILABILITY

The code can be found on GitHub (*gtancev/n-count-r*).

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

REFERENCES

[1] S. Khomenko, M. Cirach, E. Pereira-Barboza, N. Mueller, J. Barrera-Gómez, D. Rojas-Rueda, K. de Hoogh, G. Hoek, and M. Nieuwenhuijsen, "Premature mortality due to air pollution in European cities: A health impact assessment," *Lancet Planet. Health*, vol. 5196, no. 20, pp. e121–e134, 2021.

[2] J. R. Balmes and M. D. Eisner, "Indoor and outdoor air pollution," in *Murray and Nadel's Textbook of Respiratory Medicine*, 6th ed. Amsterdam, The Netherlands: Elsevier, 2016, pp. 1331–1342. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/B9781455733835000749

[3] F. Karagulian, M. Barbiere, A. Kotsev, L. Spinelle, M. Gerboles, F. Lagler, N. Redon, S. Crunaire, and A. Borowiak, "Review of the performance of low-cost sensors for air quality monitoring," *Atmosphere*, vol. 10, no. 9, p. 506, Aug. 2019.

[4] S. De Vito, M. Piga, L. Martinotto, and G. Di Francia, "CO, NO_2 and NO_x urban pollution monitoring with on-field calibrated electronic nose by automatic Bayesian regularization," *Sens. Actuators B, Chem.*, vol. 143, no. 1, pp. 182–191, 2009.

[5] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. Bonavitacola, "Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide," *Sens. Actuators B, Chem.*, vol. 215, pp. 249–257, Aug. 2015, doi: 10.1016/j.snb.2015.03.031.

[6] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. Bonavitacola, "Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO_2," *Sens. Actuators B, Chem.*, vol. 238, pp. 706–715, Jan. 2017, doi: 10.1016/j.snb.2016.07.036.

[7] A. Bigi, M. Mueller, S. K. Grange, G. Ghermandi, and C. Hueglin, "Performance of NO, NO_2 low cost sensors and three calibration approaches within a real world application," *Atmos. Meas. Techn.*, vol. 11, no. 6, pp. 3717–3735, 2018.

[8] S. Feinberg, R. Williams, G. S. W. Hagler, J. Rickard, R. Brown, D. Garver, G. Harshfield, P. Stauffer, E. Mattson, R. Judge, and S. Garvey, "Long-term evaluation of air sensor technology under ambient conditions in Denver, Colorado," *Atmos. Meas. Techn.*, vol. 11, no. 8, pp. 4605–4615, Aug. 2018.

[9] N. Zimmerman, A. A. Presto, S. P. Kumar, J. Gu, A. Hauryliuk, E. S. Robinson, A. L. Robinson, and R. Subramanian, "A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring," *Atmos. Meas. Techn.*, vol. 11, no. 1, pp. 291–313, 2018.

[10] J. Tryner, J. Mehaffy, D. Miller-Lionberg, and J. Volckens, "Effects of aerosol type and simulated aging on performance of low-cost PM sensors," *J. Aerosol Sci.*, vol. 150, Dec. 2020, Art. no. 105654, doi: 10.1016/j.jaerosci.2020.105654.

[11] J. Li, A. Hauryliuk, C. Malings, S. R. Eilenberg, R. Subramanian, and A. A. Presto, "Characterizing the aging of alphasense NO_2 sensors in long-term field deployments," *ACS Sensors*, vol. 6, no. 8, pp. 2952–2959, 2021.

[12] EUR-Lex. *Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe*. Accessed: Dec. 15, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A32008L0050

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*, 2nd ed. New York, NY, USA: Springer, 2007.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[15] G. Tancev and C. Pascale, "The relocation problem of field calibrated low-cost sensor systems in air quality monitoring: A sampling bias," *Sensors*, vol. 20, no. 21, p. 6198, Oct. 2020.

[16] S. De Vito, E. Esposito, N. Castell, P. Schneider, and A. Bartonova, "On the robustness of field calibration for smart air quality monitors," *Sens. Actuators B, Chem.*, vol. 310, May 2020, Art. no. 127869, doi: 10.1016/j.snb.2020.127869.

[17] T. J. Quinn and J. Kovalevsky, "Measurement and society," *Comp. Rendus Phys.*, vol. 5, no. 8, pp. 791–797, 2004.

[18] B. Maag, Z. Zhou, and L. Thiele, "A survey on sensor calibration in air pollution monitoring deployments," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4857–4870, Dec. 2018.

[19] L. Balzano and R. Nowak, "Blind calibration of networks of sensors: Theory and algorithms," in *Networked Sensing Information and Control*, 1st ed. Springer, 2008, ch. 1, pp. 9–37.

[20] M. S. Stanković, S. S. Stanković, and K. H. Johansson, "Distributed blind calibration in lossy sensor networks via output synchronization," *IEEE Trans. Autom. Control*, vol. 60, no. 12, pp. 3257–3262, Dec. 2015.

[21] Y. Wang, A. Yang, Z. Li, X. Chen, P. Wang, and H. Yang, "Blind drift calibration of sensor networks using sparse Bayesian learning," *IEEE Sensors J.*, vol. 16, no. 16, pp. 6249–6260, Aug. 2016.

[22] Y. Wang, A. Yang, X. Chen, P. Wang, Y. Wang, and H. Yang, "A deep learning approach for blind drift calibration of sensor networks," *IEEE Sensors J.*, vol. 17, no. 13, pp. 4158–4171, Jul. 2017.

[23] E. Miluzzo, N. D. Lane, A. T. Campbell, and R. Olfati-Saber, "CaliBree: A self-calibration system for mobile sensor networks," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Lecture Notes in Computer Science), vol. 5067. Berlin, Germany: Springer, 2008, pp. 314–331.

[24] O. Saukh, D. Hasenfratz, and L. Thiele, "Reducing multi-hop calibration errors in large-scale mobile sensor networks," in *Proc. 14th Int. Conf. Inf. Process. Sensor Netw.*, Apr. 2015, pp. 274–285.

[25] L.-J. Sun, Y.-Q. Su, S. Shen, R.-C. Wang, and W.-J. Li, "Cooperative calibration scheme for mobile wireless sensor network," in *Proc. 15th Int. Conf. Mobile Ad-Hoc Sensor Netw. (MSN)*, Dec. 2019, pp. 143–150.

[26] D. Hasenfratz, O. Saukh, and L. Thiele, "On-the-fly calibration of low-cost gas sensors," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Lecture Notes in Computer Science), vol. 7158. Berlin, Germany: Springer, 2012, pp. 228–244.

[27] Alphasense. *Technical Specification OX-B431*. Accessed: Feb. 1, 2020. [Online]. Available: http://www.alphasense.com/WEB1213/wp-content/uploads/2019/09/OX-B431.pdf

[28] *Technical Specification NO2-B43F*. Accessed: Feb. 1, 2020. [Online]. Available: http://www.alphasense.com/WEB1213/wp-content/uploads/2019/09/NO2-B43F.pdf

[29] *Technical Specification CO-B4*. Accessed: Feb. 1, 2020. [Online]. Available: http://www.alphasense.com/WEB1213/wp-content/uploads/2019/09/CO-B4.pdf

[30] J. S. Apte, K. P. Messier, S. Gani, M. Brauer, T. W. Kirchstetter, M. M. Lunden, J. D. Marshall, C. J. Portier, R. C. Vermeulen, and S. P. Hamburg, "High-resolution air pollution mapping with Google street view cars: Exploiting big data," *Environ. Sci. Technol.*, vol. 51, no. 12, pp. 6999–7008, 2017.

[31] G. Tancev and F. G. Toro, "Sequential recalibration of wireless sensor networks with (stochastic) gradient descent and mobile references," *Meas.: Sensors*, vol. 18, Dec. 2021, Art. no. 100115.

[32] A. A. Benczúr, L. Kocsis, and R. Pálovics, "Online machine learning algorithms over data streams," in *Encyclopedia of Big Data Technologies*. Cham, Switzerland: Springer, 2019, pp. 1207–1218.

[33] Y. Bengio, I. Goodfellow, and A. Courville, *Deep Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[34] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," 2019, *arXiv:1904.09237*.

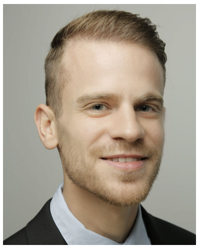
[35] *Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement*, Standard JCGM 100:2008, International Organization for Standardization, Geneva, Switzerland, 2008. [Online]. Available: http://www.bipm.org/en/publications/guides/gum.html

[36] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, Feb. 2015.

[37] M. Mueller, J. Meyer, and C. Hueglin, "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. Techn.*, vol. 10, no. 10, pp. 3783–3799, 2017.

[38] M. Collotta, G. Pau, T. Talty, and O. K. Tonguz, "Bluetooth 5: A concrete step forward toward the IoT," *IEEE Commun. Mag.*, vol. 56, no. 7, pp. 125–131, Jul. 2018.

[39] D. Leuenberger, M. Guillevic, C. Pascale, G. Baur, and A. Ackermann, "Concepts and challenges in the dynamic generation of SI-traceable nitrogen dioxide reference gas mixtures at ambient air amount fractions," in *Proc. EGU Gen. Assem. Conf. Abstr.*, vol. 20, 2018, p. 14334.



GEORGI TANCEV was born in Strumica, North Macedonia, in 1990, and moved to Switzerland, in 1995. He received the B.S. and M.S. degrees in chemical and bioengineering from the Swiss Federal Institute of Technology (ETH) Zurich, in 2015 and 2017, respectively.

He is currently a Research Scientist with the Swiss Federal Institute of Metrology (METAS), Bern, focusing on trustworthy and reliable air quality monitoring with low-cost sensors. Previously, he worked on process modeling and optimization at Novartis, and in medical device development and clinical data science with the University Children's Hospital Basel. His research interests include mathematical modeling, numerical simulation, optimization, and machine learning.

FEDERICO GRASSO TORO received the M.S. degree in electronic engineering (major in automation and robotics) from the Universidad Nacional de San Juan, Argentina, in 2011, and the Ph.D. degree in mechanical engineering (focus on artificial intelligence for global satellite navigation-based localization) from Technische Universität Braunschweig, Germany, in 2014.

From 2014 to 2016, he worked as a Project Manager at iQST, Braunschweig, and from 2016 to 2018, as a Postdoctoral Researcher with the Physikalisch-Technische Bundesanstalt (PTB), Berlin. Since 2019, he has been the Project Leader in research and development with the Swiss Federal Institute of Metrology (METAS), Bern, working on digitalization of metrological services.

Dr. Grasso Toro is a member of the Technical Committee for Digitalization (TC6) at the International Measurement Confederation (IMEKO), aiming to develop, organize, and disseminate fundamental concepts of measurement science that relate to digitalization and digital transformation in science, industry, and society.