Measurement 200 (2022) 111601

Contents lists available at ScienceDirect

Measurement

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

A surrogate-assisted measurement correction method for accurate and low-cost monitoring of particulate matter pollutants

Marek Wojcikowski [a], Bogdan Pankiewicz [a], Adrian Bekasiewicz [a,*], Tuan-Vu Cao [b], Jean-Marie Lepioufle [b], Islen Vallejo [b], Rune Odegard [b], Hoai Phuong Ha [c]

[a] Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, 80-233 Gdansk, Poland
[b] Norwegian Institute for Air Research, Oslo, Norway
[c] University of Tromso – The Arctic University of Norway, Tromso, Norway

ARTICLE INFO

Keywords:
Air monitoring
Air quality
Kriging
IoT
Measurements correction
Particulate matter
Pollution sensor
Surrogate modeling
Temporal data
Wavelet transform

ABSTRACT

Air pollution involves multiple health and economic challenges. Its accurate and low-cost monitoring is important for developing services dedicated to reduce the exposure of living beings to the pollution. Particulate matter (PM) measurement sensors belong to the key components that support operation of these systems. In this work, a modular, mobile Internet of Things sensor for PM measurements has been proposed. Due to a limited accuracy of the PM detector, the measurement data are refined using a two-stage procedure that involves elimination of the non-physical signal spikes followed by a non-linear correction of the responses using a multiplicative surrogate model. The correction layer is derived from the sparse and non-uniform calibration data, i.e., a combination of the measurements from the PM monitoring station and the sensor obtained in the same location over a specified (relatively short) interval. The device and the method have been both demonstrated based on the data obtained during three measurement campaigns. The proposed correction scheme improves the fidelity of PM measurements by around two orders of magnitude w.r.t. the responses for which the post-processing has not been considered. Performance of the proposed surrogate-assisted technique has been favorably compared against the benchmark approaches from the literature.

1. Introduction

Air pollution is a significant environmental, economic, and social problem. Its consequences for a global economy are manifested in healthcare costs, worsened quality of life, as well as premature death rates [1,2]. Those effects are especially important in developing countries where measures oriented towards preventing excessive air pollution—resulting from rapid industrialization—are often neglected [3–5]. The short- and long-term effects of pollutants in the form of, e.g., carbon and nitrogen (mono)oxides, or particulate matter (PM) on human health, environment, and economy are well understood [6,7]. Among the mentioned contaminants, particles with diameter of up to 2.5 μm (also referred to as $PM_{2.5}$) belong to the most dangerous ones as they can penetrate natural body barriers and pass to the bloodstream resulting in, e.g., cardiovascular and/or respiratory problems [8,9]. According to a World Health Organization (WHO) [10], air pollution is the cause of around seven million premature deaths per annum [11]. Its growing costs associated with the ever increasing healthcare expenditure and negative impacts on the Earth's biosphere are far from negligible [6,7,13,14][12]. From this perspective, constant and accurate monitoring of the PM-related pollution seem to be of high importance [15,16]. Reliable PM measurements are crucial for development of early warning systems dedicated to provide information on sudden bursts, or sustaining high levels of contaminants. Availability of such data is invaluable for introduction of appropriate measures for preventing/mitigating the effects of pollution exposure (e.g., masks, air filters, stay-at-home requests, etc.). High resolution and possibly short measurement time belong to the important factors determining usefulness of the mentioned systems, especially in urban areas where accurate modeling of air quality is difficult due to its high spatial and temporal dynamics [16–18].

The accurate $PM_{2.5}$ pollution monitoring is performed using complex and expensive stationary instrumentation. The systems are normally owned by local governments and (due to a high cost) are set up in a handful of carefully chosen locations over the given area [19,20]. However, in urban areas—due to the mentioned short-term variations of

* Corresponding author.
E-mail address: bekasiewicz@ru.is (A. Bekasiewicz).

https://doi.org/10.1016/j.measurement.2022.111601
Received 30 March 2022; Received in revised form 1 June 2022; Accepted 4 July 2022
Available online 12 July 2022

air contaminants—the monitoring stations are of limited use for performing accurate and timely measurements for the purpose of e.g., research, monitoring, or development of early warning services. Mentioned challenges, manifested in the form of information gaps resulting from scarcity of the precise systems (time- and resolution-wise) [19,21], can be mitigated using low-cost sensor networks based on the Internet of Things (IoT) technology [15,22,23]. Sensor-based solutions (once appropriately calibrated/tuned) proved to be capable of providing useful data on the air quality, as well as its temporal and spatial dynamics [21,24,25].

A significant bottleneck of the existing low-cost sensors for PM measurements include their transitional unexpected behaviors manifested in the form of, e.g., non-physical signal spikes, noise, unexpected lack of pollution variations over time, or measurement drift [21,26]. These effects, often referred to as outliers [19,21,25,27,28], have to be identified and removed from the data. In [25], a simple spike identification method based on comparison of sensor responses with balanced reference data obtained from the neighboring monitoring stations has been proposed. Alternative method, where outliers detection involved a comparison of the measurements with responses of the kriging model constructed based on the sparse data from the reference stations has been proposed in [27]. Yet another approach involves analysis of the outliers based on the expected probability of the data residuals [21]. The common drawback of the mentioned methods is that they are not applicable for in-situ refinement of the responses. The main advantage of in-situ data processing is a lack of need for maintaining constant connectivity with external servers/cloud in order to provide accurate spatiotemporal information on the pollution within the given location. Consequently, the approach improves the large scale systems performance as, in the case of network failure, the data can be stored and send upon re-connection. In-situ post-processing also reduces the data-transfer demand which is due to smaller size of packets that for raw information accompanied with environmental parameters. In [27], the problem related to local data analysis has been mitigated to some extent using a procedure oriented towards identification whether the sensor response is plausible compared to the residuals of the smooth data. Apart from the proposed algorithms, detection of spikes and other signal anomalies based on a cognitive approach that involves visual inspection of the characteristics is still considered as an accurate, albeit tedious, approach [21,29]. The available body of literature indicates that the problem concerning a simple and low-cost elimination of outliers for PM measurements capable of supporting in-situ correction remains open.

Another challenge related to PM measurements using the low-cost sensors involves limited reliability of the obtained data [20,30–34]. The problem is associated with high, non-linear variations of accuracy w.r.t. changing environmental conditions such as temperature and humidity [20,30,34]. The measurement precision may fluctuate within the range of around a dozen up to a few hundred percent compared to the reference data [30]. The fluctuations of the sensor accuracy can be corrected using appropriate techniques that involve analysis of the sensor responses and environmental conditions [31–34]. In [31], inaccuracy the $PM_{2.5}$ measurements has been reduced using an analytical model based on the κ-Kohler theory that expresses the relation between the air humidity and size of the particulate matter. The model has been used to obtain up to 10% improvement of the $PM_{2.5}$ measurement accuracy compared to the reference data [31]. Another approach based on analysis of the pollutants composition in urban environments, as well as their ability to absorb water particles that deteriorates sensor-based measurements has been considered in [33]. The method involves analysis of the air contaminants composition and refines correction of [31] resulting in up to 30% improvement of the measurements quality compared to the uncorrected data. Alternative techniques involve application of regression models to reduce the $PM_{2.5}$ measurement inaccuracy by around 5% [34]. Other methods, based on the random-forest-based modeling, have also been considered for refinement of the sensor data fidelity [20,34]. Regardless of the differences between the discussed approaches, they all rely on analysis of the large amounts of data to extract the correction layers. Additionally, in [33], the tuning of the model parameters based on assumptions on the composition of the air pollutants might be required to obtain satisfactory performance. The usefulness of the discussed methods for refinement of different datasets than the ones used for extraction of the models either remains unverified or deemed unsustainable [34]. An exception is the work [20], where the model constructed based on the training dataset has been re-used to correct another while maintaining comparable data quality. The correction-related problems include limited (or lack of thereof) information on the model identification cost [20,32,34]. The latter is important when in-situ refinement of the measurements is considered. From this perspective, reliable models that do not rely on *a priori* information [32], can be identified using relatively low number of training points and at a low cost are yet to be developed.

In this work, an architecture of a compact IoT-capable platform for air pollution monitoring has been proposed. The device embeds the low-cost sensors dedicated to measurements of PM pollutants and environmental conditions, as well as the connectivity gear and internal power supply. A relatively low accuracy of the utilized PM detector is enhanced using a two-stage procedure that involves de-spiking and multiplicative correction of the PM data. The refinement process is realized by means of a kriging surrogate identified based on the training points that are automatically pre-selected from the sparse calibration data. Performance of the proposed post-processing method has been demonstrated based on the several test cases concerning measurements of the $PM_{2.5}$, as well as particulate matter with up to 1 μm and 10 μm diameters, respectively. The obtained results indicate that the presented mechanisms improve the measurements fidelity (w.r.t. the reference station data) by up to two orders of magnitude compared to the uncorrected responses. The proposed correction technique has been analyzed in terms of numerical efficiency, as well as compared against the state-of the art approaches from the literature. The surrogate-assisted bi-stage correction provides up to 3-fold improvement of the PM measurement accuracy with respect to the benchmark techniques.

2. Materials

Design of the mobile platform for accurate PM monitoring is a subject to multiple requirements concerning accuracy, repeatability of the measurements, but also modular architecture. The latter is important for straightforward extension of the system capabilities and its development oriented towards implementation of a reliable IoT-capable service. On the other hand, one has to consider constraints resulting from mobility and intended affordability of the system. These include, among others, small dimensions, relatively long battery-powered operation, as well as the use of possibly low-cost (hence, mass-produced) components. In this section, the architecture of the PM measurements platform oriented towards addressing the mentioned criteria has been considered. It should be noted that, due to the relatively low cost, the utilized commercial sensors cannot compete with high-performance measurement devices installed in stationary monitoring units. From this perspective, it is expected that the data gathered by the device will be prone to errors. To mitigate this problem, appropriate low-cost data correction mechanisms have been proposed which are explained in Section 3.

2.1. Background and assumptions

The negative effects of $PM_{2.5}$ pollution on health are well studied [8,35,36]. Even low concentration of pollutants in the air (expressed in $\mu g/m^3$) contributes to increased incidence of respiratory, and/or heart diseases [36]. It is estimated that—in the European Union alone—reduction of the $PM_{2.5}$ levels could increase of the average life expectancy of the population by around 13 months [39]. In 2005, WHO determined the acceptable average annual concentration of $PM_{2.5}$ at 10 $\mu g/m^3$ [37]. In 2021, the level was refined to 5 $\mu g/m^3$ with the tentative

targets for the industrialized areas (cities, in particular) of up to 75 μg/m^3 daily and 35 μg/m^3 yearly, respectively [38]. As already mentioned, accurate monitoring of the pollution level in urban environments is difficult due to their dynamics. For instance, in the city of Gdansk (Poland) accurate measuring stations are sparsely deployed which hinders reliable monitoring of air quality. Location of the agglomeration in a narrow valley between the sea and the hills covered by the dense forests results in notable weather changes across the area which further amplifies the PM monitoring-related challenges. From this perspective, availability of the mobile platform that provides reasonably accurate information on pollutants concentration would be invaluable.

It is expected that the monitoring system should provide: (i) modular architecture oriented around mobility, (ii) PM detector, (iii) environmental sensors, (iv) wireless connectivity, (v) computational capabilities to perform in-situ refinement of the measurements and re-calibration of the correction models. Other desirable parameters include the use of hardware that balances the performance and cost (with prospects for optimization of the latter), high measurement accuracy, and relatively long battery operation. The desirable average error between the sensor and monitoring station responses is below 5 μg/m^3.

2.2. Platform architecture

The proposed platform was designed in accordance to the requirements specified above. The schematic diagram of the device is shown in Fig. 1. It consists of independent components (modules) connected to the computing platform based on a system-on-a-chip (SoC). The latter provides a series of multi-purpose input/output (I/O) ports, but also a computational power necessary for acquisition, processing and transmission of the data. The remaining subsystems of the device include: (i) PM sensor, (ii) cellular connectivity module, (iii) environmental sensors capable of measuring temperature, humidity and pressure, (iv) data storage unit, as well as (v) energy distribution module with batteries.

The utilized SoC comprises a field programmable gate array (FPGA) with two advanced RISC machine (ARM Cortex-A9 667 MHz; note that RISC refers to reduced instruction set computer) processors and 512 MB RAM integrated on a Zynq board (ZYBO) from Digilent [40]. The platform enables integration of the device modules in terms of software and hardware. Owing to the availability of ARM processors, the system is capable of running the operating system (here, Linux). It provides a framework for interfacing with the devices, but also enables the use of high-level mechanisms for data acquisition, storage (including communication protocols, file systems, etc.), and cloud connectivity. The latter is considered important for application of the device as a component of the IoT-based air quality monitoring system which will be discussed elsewhere. Computational power of the ARM is considered sufficient for in-situ refinement of the PM measurements performed by the embedded sensor (see Section 4.5 for details). The data storage is performed on a secure digital card. The communication between the system components is realized (through a dedicated circuit board) using

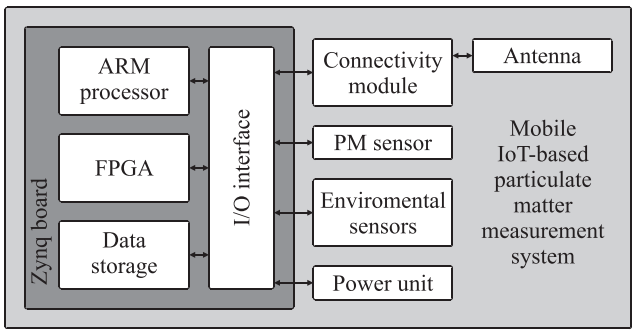

Fig. 1. An IoT-based PM measurement device – a block diagram.

serial protocols [41,42]. With relatively compact dimensions of 122 mm × 88 mm, as well as a high number of input/output ports the Zynq platform is considered suitable for development of the air-quality monitoring system.

The utilized PM sensor, SPS30 manufactured by Sensirion, performs pollution data acquisition using the laser scattering technique [43,79]. The device offers relatively long-term measurements stability, support for identification of PM_1, $PM_{2.5}$, and PM_{10} particles, as well as small volume (41 mm × 41 mm × 12 mm). According to the specification [43], the sensor is capable of detecting pollutants in a range of up to 1000 μg/m^3 with the precision of up to ± 10 μg/m^3 (the latter, however, might be lower according to the data presented in Section 4).

Communication services are provided using the module BG96 manufactured by Quectel [44]. The device supports a range of satellite navigation and cellular technologies. Monitoring of the environmental data is realized by the 24-bit barometer, hygrometer, and temperature sensor from Hoperf Electronic [45]. The accuracy of the measured temperature in a range from –20 °C to 60 °C is ± 0.3 °C.

Power for the system is delivered through an in-house circuit that embeds the STM32 microcontroller, MAX9938 voltage current monitors, and a set of four 18650 lithium-ion batteries. The module provides up to 24 h of cordless operation [46–48]. Furthermore, it allows for monitoring of the system current draw, charging state, and the battery level. The system is enclosed in a custom housing fabricated using an additive manufacturing technology (fused deposition modeling) using polyethylene terephthalene glycol-modified filament. Figs. 2 and 3 show an exploded view of the device components and its photograph. It is worth noting that, in order to ensure a decent performance of the cellular connectivity, the location of the antenna within the device has been adjusted as described in Section 2.3.

2.3. Performance-oriented connectivity tuning

High-performance wireless connectivity is an important for mobility of the proposed IoT system. Here, the data transmission is realized using a cheap, uniplanar antenna characterized by an omnidirectional radiation pattern. It should be noted that the field performance of the radiator is a function of its allocation w.r.t. the remaining components of the system. The latter ones (or rather their ground plane layers) act like reflectors for the radio-frequency signals which affect antenna

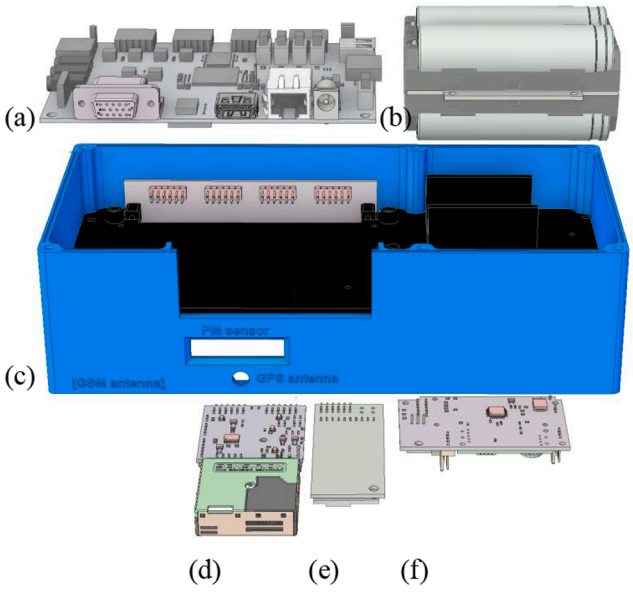

Fig. 2. Exploded view of the proposed device: (a) Zynq board [40], (b) battery pack [48], (c) housing, (d) PM sensor [43], (e) connectivity module [44], and (f) power management unit [46,47].

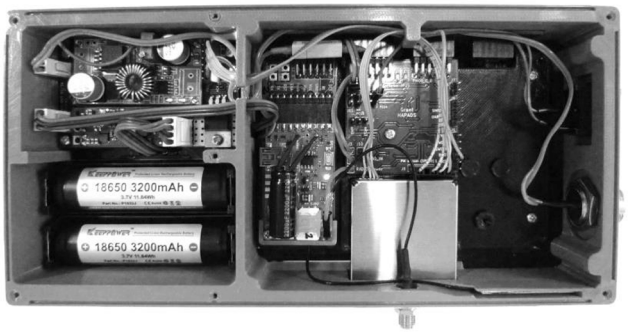

Fig. 3. The proposed mobile PM measurement system – a photograph.

characteristics. Here the environmental effects on the antenna performance have been investigated using electromagnetic (EM) simulations. The results shown in Fig. 4 indicate that location of the structure above the sensor boards results in increase of its directivity—roughly perpendicular to the boards—which is considered acceptable from the standpoint of maintaining high-performance connectivity. In this work, the radiator has been mounted in the location shown in Fig. 4(b).

3. Methods

The PM sensor of Section 2 suffers from a limited accuracy manifested in the form of a high measurement noise (spikes/peaks being up to an order of magnitude higher compared to the detection range as shown in Fig. 5), as well as non-linear variations of the pollution levels with respect to the reference data. The latter normally stem from variations of the environmental conditions such as temperature, or humidity [30,34]. On the other hand, correlation of the PM measurements with the reference station data can be improved using appropriate post-processing techniques. Here, a two-step method which implements the detection and elimination of non-physical peaks, as well as a surrogate-assisted correction of the responses has been proposed. The first step, involves identification of the signal spikes based on the PM envelope constructed from multiple measurement channels offered by the sensor. The de-spiked data is then refined using the wavelet-transform-based method [49,50]. In the next stage, the surrogate-based correction of the sensor measurements using a kriging model constructed based on the carefully selected data is performed [51,52]. The calibration set is extracted from the sparse responses obtained for both the sensor and the reference station. To make the paper self-contained, a brief discussion on the wavelet-based spikes detection, as well as kriging modeling is also provided. The methods discussed in this work have been implemented in MATLAB [76].

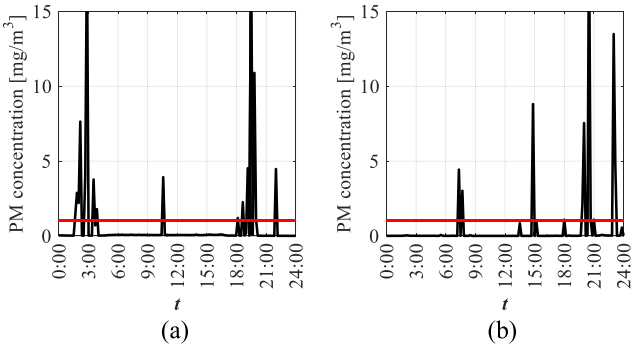

Fig. 5. The $PM_{2.5}$ pollution data obtained in Gdansk (Poland) using the SPS30 sensor on: (a) Jan 12, 2022, and (b) Apr 26, 2022. The measured peaks are up to an order of magnitude higher compared to the detection range denoted by red lines. Note that the PM concentration is in mg/m^3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1. Problem formulation

Let $\boldsymbol{R}_s(\boldsymbol{t}, \boldsymbol{k}) = [R_s(t_1, k_1) \ldots R_s(t_n, k_n) \ldots R_s(t_N, k_N)]^T$ (here, $n = 1, \ldots, N$) be the vector representing the concentration of particulate matter with up to 2.5 μm diameter recorded by the sensor of Section 2.2 over a period of time $\boldsymbol{t} = [t_1\ t_2 \ldots t_N]^T$; $\boldsymbol{k} = \boldsymbol{k}(\boldsymbol{t}) = [k_1 \ldots k_N]^T$ is a vector of temperatures recorded over $\boldsymbol{t}$. For simplicity of notation, we will often substitute $\boldsymbol{R}_s(\boldsymbol{t}, \boldsymbol{k})$ as $\boldsymbol{R}_s(\boldsymbol{t}) = [R_{s.1} \ldots R_{s.n} \ldots R_{s.N}]^T$ or simply $\boldsymbol{R}_s$ without the change of the meaning. Also, the same style of description will be used to denote other time-series sensor data.

The responses $\boldsymbol{R}_s(\boldsymbol{t})$ are distorted by the noise in the form of non-physical signal spikes (their amplitudes exceed operational range of the sensor; cf. Fig. 5) and feature variable accuracy of the PM detection as compared to the reference station data $\boldsymbol{R}_r(\boldsymbol{t})$ over the same period of time $\boldsymbol{t}$. The goal of the correction process is two-fold. First, a mapping of the form is to be applied:

$$d : \boldsymbol{R}_s(\boldsymbol{t}) \rightarrow \boldsymbol{R}_d(\boldsymbol{t}) \tag{1}$$

where $\boldsymbol{R}_d(\boldsymbol{t})$ is the refined sensor data with eliminated spikes and d denotes the function (cf. Section 3.4) that implements a threshold/envelope-based de-spiking followed by a wavelet-based peaks identification/rejection. The accuracy of the refined $\boldsymbol{R}_d$ responses is then improved using the multiplicative correction of the following form [53,54]:

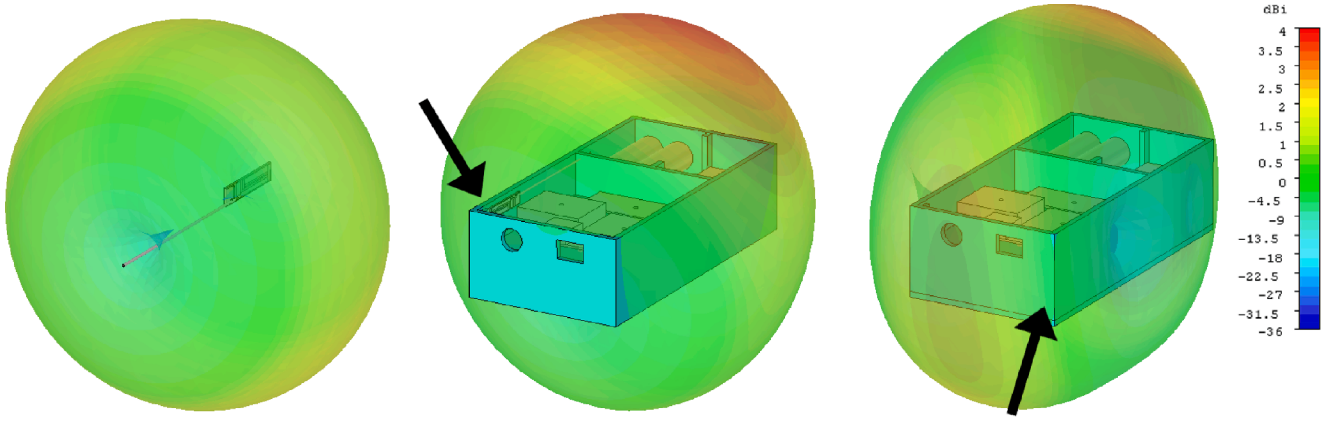

Fig. 4. The effect of the cellular antenna location within the system on its radiation characteristics (at the 830 MHz frequency): (a) sole radiator, as well as the antenna mounted in the system (arrows indicate location) in the (b) top-left corner, and (c) bottom-right corner. The omnidirectional radiation pattern of the structure is distorted by the ground plane layers of other modules of the device. Scale on the right illustrates gain of the structure.

$$\boldsymbol{R}_c(\boldsymbol{t}) = \boldsymbol{A} \circ \boldsymbol{R}_d(\boldsymbol{t}) \tag{2}$$

Here, $\boldsymbol{R}_c(\boldsymbol{t})$ represents the refined response of the sensor, "∘" denotes the component-wise multiplication, whereas $\boldsymbol{A}$ is the correction component derived from a kriging surrogate model [51]. The latter is constructed based on the calibration data derived from both the reference station $\boldsymbol{R}_r$ and de-spiked sensor responses $\boldsymbol{R}_d$. The measurements used for calibration include sensor and monitoring station responses acquired in the same location over the same period of time $\boldsymbol{t}$. The corrected response $\boldsymbol{R}_c$ is considered valid within the ranges of PM and temperatures for which the surrogate model has been identified.

3.2. Wavelet-based detection of peaks

The wavelet-based spike detection boils down to estimation of the resemblance between the signal peaks and the considered kernel function [49,50]. Here, the latter is represented by a Haar wavelet which provides reasonable shape-approximation of the peaks obtained from the sensor [55]. The method involves decomposition of the signal into a multi-scale representation (in terms of the wavelet coefficients) followed by a multi-level hypothesis testing oriented towards identification of the spike. The scales are defined as $\boldsymbol{B} = \{b_0, \ldots, b_j, \ldots, b_J\}$, where b_0/b_J are determined based on the sampling rate of the signal and the expected peaks duration [50]. The intermediate scales $\{b_1, \ldots, b_{J-1}\}$ are uniformly allocated between the b_0 and b_J, respectively.

The wavelet coefficient of the nth component $R_{s.n}$ of the $\boldsymbol{R}_s$ signal (cf. Section 3.1) at a scale b_j is given as $w(j, n) = <R_{s.n}, \psi_{j.n}>$, where $\psi_{j.n}$ is the kernel function [50]. It should be noted that the measured signal contains both information and noise. From this perspective only the signal part that carries information is used to represent the wavelet coefficient. It can be extracted from the noise (individually at each scale b_j) using the hard threshold rule [49,50]:

$$\rho(w) = \begin{cases} w, & \text{when } |w| \rangle T = \sigma\sqrt{2\ln(N)} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Here, $\rho(w)$ acts on the wavelet coefficients w obtained at the scales from $\boldsymbol{B}$, whereas the threshold T is a function of the noise coefficients, standard deviation σ, and the number of time-series samples [49,50].

The detection is performed (separately for each scale b_j) by testing the null hypothesis that the signal (here, understood as the spike) is not present and its alternative that the combination of the signal and noise does exist [56]. Owing to the transient nature of peaks, the alternative hypothesis holds only within the specific time intervals [49]. Finally, the decisions resulting from the hypothesis testing performed at all scales are combined and utilized to estimate the arrival time of the peaks. For the ith peak the time of its occurrence is calculated as [49,50]:

$$t_i = \|B_{H.i}\|^{-1} \sum_{b_j \in B_{H.i}} t_{i.j} \tag{4}$$

where $B_{H.i} = \{b_j \in B: |w(j,n)| > \Theta_j, n \in C_{H.i}\}$ with $C_{H.i}$ being the subset of the basis function translations for which the hypothesis related to the presence of the signal and noise holds over all scales; Θ_j is the hypothesis acceptance threshold at the b_j scale [49]. The parameter $t_{i.j}$ represents estimated location of the ith peak at b_j and is defined as [49,50]:

$$t_{i.j} = \arg\max_{n \in C_{H.i}} \left\{ |w(j,n)| : |w(j,n)| \rangle \Theta_j \right\} \tag{5}$$

The identified peaks candidates can be further removed using appropriate post-processing [50,57–59].

3.3. Kriging-based modeling

Kriging surrogate is used to provide a correction layer dedicated to refine the PM sensor measurements. The model provides a smooth approximation of the data based on the interpolation of all the training points used for its construction. The formulation of kriging follows the description of [52,60]. Let $\boldsymbol{X} = [\boldsymbol{x}_1 \ \ldots \ \boldsymbol{x}_m \ \ldots \ \boldsymbol{x}_M]^T$ ($m = 1, \ldots, M$) represent the vectors of training points and $\boldsymbol{Y} = [y_1 \ \ldots \ y_m \ \ldots \ y_M]^T$ their responses. Now, let $\boldsymbol{x}_l$ be the lth point—specified within the bounds determined by the training data—with unknown response. The surrogate model response $y_l = \boldsymbol{R}_{KR}(\boldsymbol{x}_l)$ can be obtained as [60]:

$$\boldsymbol{R}_{KR}(\boldsymbol{x}_l) = \mu + \upsilon^T \boldsymbol{\Psi}^{-1}(\boldsymbol{Y} - \boldsymbol{1}\mu) \tag{6}$$

Here, $\boldsymbol{1}$ denotes the vector of ones and $\boldsymbol{\upsilon} = [\mathrm{cor}(y_1, y_l) \ \ldots \ \mathrm{cor}(y_M, y_l)]^T$ is the vector of correlations between the training data and the y_l prediction. The correlation matrix and a mean base term are given as [60]:

$$\boldsymbol{\Psi} = \begin{bmatrix} \mathrm{cor}(y_1, y_1) & \cdots & \mathrm{cor}(y_1, y_M) \\ \vdots & \ddots & \vdots \\ \mathrm{cor}(y_M, y_1) & \cdots & \mathrm{cor}(y_M, y_M) \end{bmatrix} \tag{7}$$

$$\mu = \frac{\boldsymbol{1}^T \boldsymbol{\Psi}^{-1} \boldsymbol{Y}}{\boldsymbol{1}^T \boldsymbol{\Psi}^{-1} \boldsymbol{1}} \tag{8}$$

For the pair of selected G-dimensional designs $\boldsymbol{x}_i = [x_{i.1} \ \ldots \ x_{i.g} \ \ldots \ x_{i.G}]^T$, $\boldsymbol{x}_j = [x_{j.1} \ \ldots \ x_{j.g} \ \ldots \ x_{j.G}]^T$ and their responses y_i, y_j, the correlation is expressed using a basis function of the form [60]:

$$\mathrm{cor}(y_i, y_j) = exp\left(- \sum_{g=1}^{G} \theta_g \left| x_{i.g} - x_{j.g} \right|^{p_j} \right) \tag{9}$$

The parameter vectors $\boldsymbol{\theta} = [\theta_1 \ \ldots \ \theta_k \ \ldots \ \theta_K]^T$ and $\boldsymbol{p} = [p_1 \ \ldots \ p_k \ \ldots \ p_K]^T$ can be estimated through maximization of the ln-likelihood function realized using a suitable numerical optimization algorithm [51]. For more detailed discussion on kriging and the implementation used in this work, see [51,52,60].

3.4. Spikes detection and elimination

As already mentioned, the peaks detection and elimination procedure are conducted in two stages. First, the hard-threshold and envelope-based PM data refinement is performed. It should be noted that—even though the main focus of the work is identification of $PM_{2.5}$ pollution—the sensor is also capable of estimating the PM_{10} and PM_1 pollutants concentration. Owing to the correlation of PM_1, $PM_{2.5}$, and PM_{10} pollution levels resulting from growing granularity of the measurements (e.g., particles with up to 2.5 μm diameter also account for PM_1 pollutants) [72–75], the amount of data required for accurate de-spiking can be increased.

Let $\boldsymbol{R}(\boldsymbol{t}) = [\boldsymbol{R}_1(\boldsymbol{t}) \ \boldsymbol{R}_2(\boldsymbol{t}) \ \boldsymbol{R}_3(\boldsymbol{t})]^T$ where $\boldsymbol{R}_j(\boldsymbol{t}) = \boldsymbol{R}_j(\boldsymbol{t}, \boldsymbol{k})$, $j = 1, 2, 3$, represent the concentration of PM_{10}, $PM_{2.5}$ (note that $\boldsymbol{R}_2(\boldsymbol{t}) = \boldsymbol{R}_s(\boldsymbol{t})$; cf. Section 3.1), and PM_1 particles over a period $\boldsymbol{t}$ obtained from the sensor of Section 2. From the $\boldsymbol{R}(\boldsymbol{t})$ matrix, non-physical measurements in the jth row and the nth time instance (column) can be identified and removed as follows:

$$R_{e.j.n} = \begin{cases} R_{j.n}, & \text{when } R_{j.n} \leqslant \delta_1 \\ 0.5\left(R_{j.n-l_1} + R_{j.n+l_2}\right), & \text{otherwise} \end{cases} \tag{10}$$

where $n - l_1$ and $n + l_2$ represent the nearest indices around nth element of $\boldsymbol{R}_j$ for which the responses are below the δ_1 threshold. Here, $\delta_1 = 375$ μg/m^3 is used, which corresponds to a five-fold violation of the (tentative) maximum acceptable 24-hour exposure limit to the $PM_{2.5}$ pollution specified by the WHO [38]. Note that δ_1 is user-defined and its particular value can be determined based on analysis of the historical data available for the given area. The refined matrix is of the form:

$$\boldsymbol{R}_e = [\boldsymbol{R}_{e.1} \quad \boldsymbol{R}_{e.2} \quad \boldsymbol{R}_{e.3}]^T = \begin{bmatrix} R_{e.1.1} & R_{e.2.1} & R_{e.3.1} \\ \vdots & \vdots & \vdots \\ R_{e.1.N} & R_{e.2.N} & R_{e.3.N} \end{bmatrix}^T \tag{11}$$

In the next step, the corrected responses are used for construction of the lower-bound PM envelope $\boldsymbol{E} = [E_1 \ \ldots \ E_N]^T$. The latter represents the minimum value among the available PM measurements with different levels of granularity. The component of $\boldsymbol{E}$ in nth time instance is

extracted as:

$$E_n = \min\left(\left[R_{e.1.n} \quad R_{e.2.n} \quad R_{e.3.n}\right]^T\right) \tag{12}$$

The envelope is utilized to eliminate the peaks with amplitudes lower than the hard-defined threshold δ_1. The process is performed based on analysis of the relative variation between the $\boldsymbol{R}_{e.j}$ (the selected j depends on the PM of interest) and the envelope. In other words, the nth component of the $\boldsymbol{R}_p = [R_{p.1} \ldots R_{p.n} \ldots R_{p.N}]$ vector is constructed from $\boldsymbol{E}$ and $\boldsymbol{R}_{e.j}$ as follows:

$$R_{p.n} = \begin{cases} E_n, & \text{when } \xi_n \geqslant \delta_2 \\ R_{e.j.n}, & \text{otherwise} \end{cases} \tag{13}$$

Here, $\xi_n = (R_{e.j.n} - E_n)/\max(|\boldsymbol{R}_{e.j} - \boldsymbol{E}|)$, whereas the user-defined threshold is $\delta_2 = 0.05$. It should be noted that substitution of the identified peak with the envelope response introduces some inaccuracy to the $\boldsymbol{R}_p$ data. On the other hand, the typical discrepancy for the considered test cases is around an order of magnitude lower compared to the one resulting from the presence of the spike. Consequently, the process results in overall improvement of the PM measurements fidelity.

In the second stage of the refinement process, the remaining peaks are identified using the method of Section 3.2. The approach detects a set of time points $\boldsymbol{t}_p \in \boldsymbol{t}$, based on (5), for which the spikes are present. For the identified time instances—similarly as in (10)—the signal is reconstructed as the average of the nearest $PM_{2.5}$ response before and after the peak. Owing to low computational cost, the process has the potential for in-situ operation, albeit at the expense of a certain time-delay resulting from (10). An important remark is that the denominator of the coefficient ξ_n, i.e., $\max(|\boldsymbol{R}_{e.j} - \boldsymbol{E}|)$, has to be "stabilized" over a set of measurements. This can be achieved (at a low cost) through its calculation as $\xi_n = (R_{e.j.n} - E_n)/E_{\text{hist}}$, where E_{hist} represents the history of the sensor responses given as:

$$E_{hist} = \begin{cases} \left|R_{e.j.n} - -E_n\right|, & \text{when } n = 1 \\ \max\left(\left[E_{hist}, \ \left|R_{e.j.n} - -E_n\right|\right]\right), & \text{otherwise} \end{cases} \tag{14}$$

The example $PM_{2.5}$ measurements at each stage of the de-spiking process are shown in Fig. 6. The generality of the proposed approach might be affected by capability of the sensor in terms of the number and span of PM granularity levels that can be identified [72]. For instance, usefulness of PM_{10} for correction of PM_1 might be limited (depending on the environment) due to notable (up to an order of magnitude) difference between the dimensions of accounted particles [72]. From this perspective, the envelope (12) should be constructed using a combination of the PM measurements with similar granularity to the target parameter (here, $PM_{2.5}$).

3.5. Identification of the kriging-based data correction model

The sensor-based PM measurements are characterized by fluctuations of accuracy w.r.t. the high-end monitoring stations. In this work, the discrepancies are accounted for using a kriging model [60]. Upon identification, the surrogate is used as the multiplicative correction layer (2) for the $\boldsymbol{R}_d$ responses [54]. The goal of the process is to provide the mapping of the discrepancies between the $\boldsymbol{R}_d(\boldsymbol{t})$ and $\boldsymbol{R}_r(\boldsymbol{t})$ responses, while ensuring that the model (6) provides a representation of the calibration data over the available (possibly large) ranges of the environmental conditions and PM measurements. This is ensured using an appropriate data treatment scheme oriented towards correction of the signal rather than the local noise.

Due to the sparse and non-uniform distribution of measurement samples as a function of e.g., temperature and/or PM-levels (see Fig. 7), their direct use for construction of the surrogate would restrict usefulness of the interpolation model to a small fraction of the data. Here, the problem is addressed through a linear transformation and confinement of the PM measurements oriented towards optimization of the number/location of representative training points that can be used for model

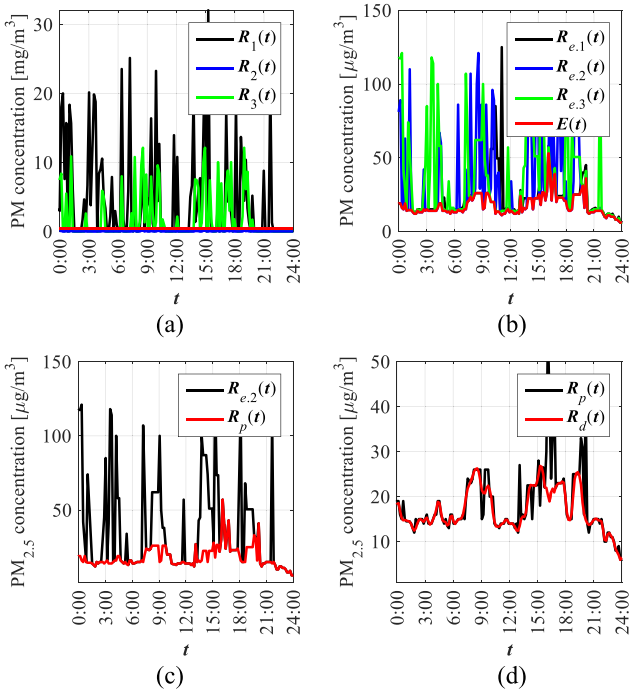

Fig. 6. The $PM_{2.5}$ data de-spiking procedure: (a) trimming of the peaks above δ_1 (red line), (b) determination of the envelope (red) based on PM_1–PM_{10} responses, (c) envelope-based spikes rejection, and (d) wavelet-based elimination of peaks. The measurement data has been acquired on Jan 16, 2022 in Gdansk (Poland). Note that, in (a), the PM concentration scale is in mg/m^3, whereas in (b)-(d) it is in μg/m^3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

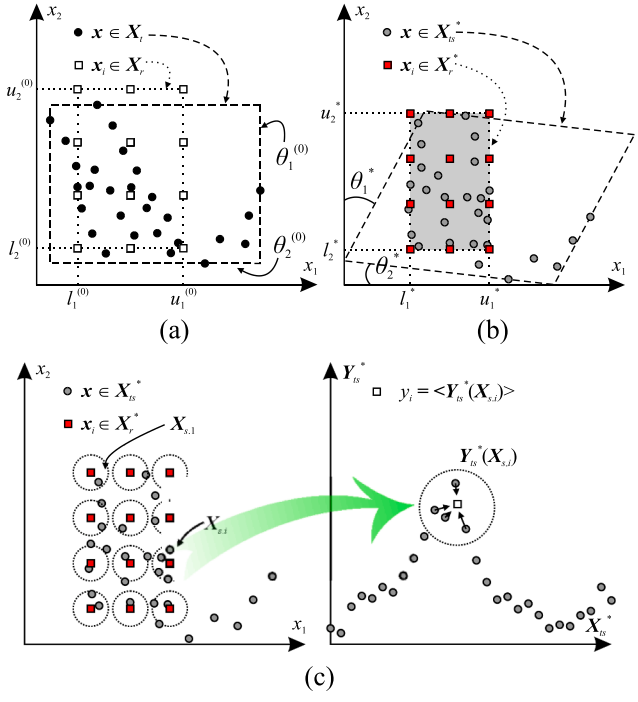

Fig. 7. Conceptual illustration of the sensor data processing steps. Visualization of the input data points: (a) before and (b) after minimization of (15), as well as (c) identification of the designs located within the user-defined distance λ around the reference designs (left) and aggregation of their associated responses (right; visualized using a two-dimensional projection) to generate the set of responses $\boldsymbol{Y}_r^* = \{y_i\}_{i=1,\ldots,J}$ corresponding to the $\boldsymbol{X}_r^*$ designs.

identification. The conceptual illustration of the process is shown in Fig. 7.

Let $\boldsymbol{X}_t = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m, \ldots \boldsymbol{x}_M\}$ and $\boldsymbol{Y}_t = \{y_1, \ldots, y_m, \ldots, y_M\}$ represent the sets of sparsely located test points and their corresponding responses acquired over a certain period of time (i.e., the calibration data). Here, $\boldsymbol{x}_m = [x_{1.m}\ x_{2.m}]^T = [R_d(t_m, k_m)\ k_m(t_m)]^T$ and $y_m = R_r(t_m, k_m)/R_d(t_m, k_m)$, respectively. Then, let.

$\boldsymbol{X}_r = \{\boldsymbol{x}_i\}_{i = 1, \ldots, J}$ be the set of standard basis points located on a grid and scaled w.r.t. the unknown lower/upper $\boldsymbol{l}/\boldsymbol{u}$ bounds that results from the distribution of the available test points [61]. The dataset transformation and confinement are realized by solving the minimization problem of the following form:

$$\boldsymbol{z}^* = \text{argmin}(-U_1(\boldsymbol{z}) + U_2(\boldsymbol{z})) \tag{15}$$

where the vector $\boldsymbol{z} = [\boldsymbol{l}\ \boldsymbol{u}\ \boldsymbol{\theta}]^T = [l_1\ l_2\ u_1\ u_2\ \theta_1\ \theta_2]^T$ represents the lower/upper bounds for the $\boldsymbol{X}_r$ set and slant angles for $\boldsymbol{x} \in \boldsymbol{X}_t$ (see Fig. 7) used for determination of the transformed dataset $\boldsymbol{X}_{ts}$:

$$\boldsymbol{X}_{ts} = \{\boldsymbol{x}\boldsymbol{S}(\boldsymbol{\theta}) : \boldsymbol{x} \in \boldsymbol{X}_t\} \tag{16}$$

where

$$\boldsymbol{S}(\boldsymbol{\theta}) = \begin{bmatrix} 1 & \tan(\theta_1) \\ \tan(\theta_2) & 1 \end{bmatrix} \tag{17}$$

The objective function $U_1(\boldsymbol{z}) = |\boldsymbol{X}_c|/|\boldsymbol{X}_t|$ where $\boldsymbol{X}_c = \{\boldsymbol{x}: \boldsymbol{x} \in \boldsymbol{X}_{ts} \wedge \boldsymbol{l} \leq \boldsymbol{x} \leq \boldsymbol{u}\}$ (note that $|\bullet|$ denotes the cardinality of the set) involves maximization of the number of data points confined within $\boldsymbol{l}/\boldsymbol{u}$ bounds w.r.t. $\boldsymbol{X}_t$, whereas $U_2(\boldsymbol{z}) = <[d_1 \ldots d_i \ldots d_J]^T>$ with d_i being a minimum distance between the ith reference point $\boldsymbol{x}_i \in \boldsymbol{X}_r$ and the elements from the transformed set of training data (note that the symbol $<\bullet>$ denotes average). The final design $\boldsymbol{z}^*$—found through solving (15)—is used to generate the dataset $\boldsymbol{X}_{ts}{}^*$ and its associated responses $\boldsymbol{Y}_{ts}{}^* = \boldsymbol{Y}_t$ (note that change of $\boldsymbol{X}_{ts}$ basis does not alter PM measurements obtained from the sensor), as well as $\boldsymbol{X}_r{}^*$. The latter is the $\boldsymbol{X}_r$ set rescaled w.r.t. the bounds obtained from the optimized $\boldsymbol{z}^*$ vector.

Once the transformed set $\boldsymbol{X}_{ts}{}^*$ and its associated responses $\boldsymbol{Y}_{ts}{}^*$ are found, a total of J sequences $\{\boldsymbol{X}_{s.1}, \ldots, \boldsymbol{X}_{s.i}, \ldots, \boldsymbol{X}_{s.J}\}$, where $\boldsymbol{X}_{s.i} = \{\boldsymbol{x}: \boldsymbol{x} \in \boldsymbol{X}_{ts}{}^*, \boldsymbol{x}_i \in \boldsymbol{X}_r{}^* \wedge ||\boldsymbol{x} - \boldsymbol{x}_i|| \leq \lambda\}$ are constructed where λ is the user-defined parameter that exhibits search radius around the ith design from $\boldsymbol{X}_r{}^*$ (here, $\lambda = 3$). The elements of $\boldsymbol{X}_{s.i}$ are ordered from the least to most distant from $\boldsymbol{x}_i$. For each sequence, the responses of the first k designs (here, $k = 5$) used to construct the response using the median $y_i = M(\boldsymbol{Y}_{ts}{}^*(\boldsymbol{X}_{s.i.1\text{-}k}))$. Finally, the set of reference points $\boldsymbol{X}_r{}^*$ and their responses $\boldsymbol{Y}_r{}^* = \{y_i\}_{i = 1, \ldots, J}$ are used for construction of the kriging interpolation model $\boldsymbol{R}_{KR}$ as described in Section 3.3. The example functional landscapes obtained from $\boldsymbol{X}_r{}^*$ and $\boldsymbol{Y}_r{}^*$ are shown in Fig. 8. The identification of the kriging-based correction model can be summarized as follows:

1. Obtain $\boldsymbol{R}_d(\boldsymbol{t}, \boldsymbol{k})$ and $\boldsymbol{R}_r(\boldsymbol{t}, \boldsymbol{k})$ and $\boldsymbol{k}(\boldsymbol{t})$ data over a selected time period $\boldsymbol{t}$ (cf. Section 3.1);
2. Define the reference set $\boldsymbol{X}_r$;
3. Set $\boldsymbol{z}^{(0)}$ and obtain $\boldsymbol{z}^*$ by solving (15); use the optimized parameters to generate $\boldsymbol{X}_{ts}{}^*$;
4. Extract the set of $\boldsymbol{Y}_{ts}{}^*$ responses that corresponds to the designs from $\boldsymbol{X}_{ts}{}^*$;
5. Obtain $\boldsymbol{X}_r{}^*$ by rescaling $\boldsymbol{X}_r$ w.r.t. the optimized $\boldsymbol{l}/\boldsymbol{u}$ bounds (see Fig. 7 (a)-(b));
6. Generate the set of averaged responses $\boldsymbol{Y}_r{}^* = \{y_i\}_{i = 1, \ldots, J}$ from the elements of $\boldsymbol{Y}_{ts}{}^*$ that correspond to the ordered elements of $\boldsymbol{X}_{ts}{}^*$ located in the vicinity to the reference designs from $\boldsymbol{X}_r{}^*$ (see Fig. 7 (c));
7. Use the $\boldsymbol{X}_r{}^*$, $\boldsymbol{Y}_r{}^*$ for identification of the kriging model $\boldsymbol{R}_{KR}$ (cf. Section 3.3).

The identified surrogate model can be used to refine the data preprocessed as described in Section 3.4. The multiplicative correction vector for the N-point measurement is of the form $\boldsymbol{A} = [a_1 \ldots a_j \ldots a_N]^T$, where:

$$a_j = \boldsymbol{R}_{KR}([R_d(t_j, k(t_j))\quad k(t_j)]\boldsymbol{S}(\boldsymbol{\theta}^*)) \tag{18}$$

It is worth reiterating that identification of the correction coefficient for jth sample involves transformation of the sensor data using matrix (17) where the slant angles are determined from solving (15). The latter is performed only in the course of model identification process. The correction layer can be applied both to vector, or scalar responses. Consequently, once the model is established, it has the potential to support in-situ refinement of the sensor-based PM measurements.

It should be noted that grid-based design of experiments used to generate $\boldsymbol{X}_r$ is not mandatory. It has been selected for simplicity, however for multi-dimensional input data utilization other methods such as Latin hypercube, or orthogonal sampling may be of interest as well [62,63]. Bearing sufficient amount of training samples obtained from the sensor and the weather station, the model features reasonable accuracy within the optimized $\boldsymbol{l}^*/\boldsymbol{u}^*$ bounds.

3.6. Summary of the data-correction method

The correction of the $\boldsymbol{R}_s(\boldsymbol{t}, \boldsymbol{k})$ samples acquired by the sensor involves de-spiking of the data followed by its transformation oriented towards determination of the multiplicative correction and, finally, determination of the refined response from (2). Assuming that the kriging surrogate model (6) is identified, the PM measurement correction of the N-point data vector can be summarized as follows (see Fig. 9 for a block diagram):

1. Perform spikes elimination on $\boldsymbol{R}_s$ using (10) and obtain the refined matrix of sensor responses $\boldsymbol{R}_e$;

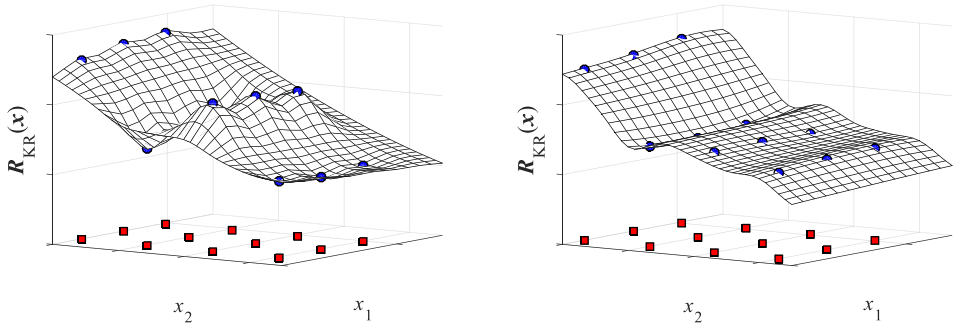

Fig. 8. Example functional landscapes of the kriging-based correction layer extracted from different $\boldsymbol{X}_r{}^*$ (red rectangles) and their corresponding $\boldsymbol{Y}_r{}^*$ responses (blue circles). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

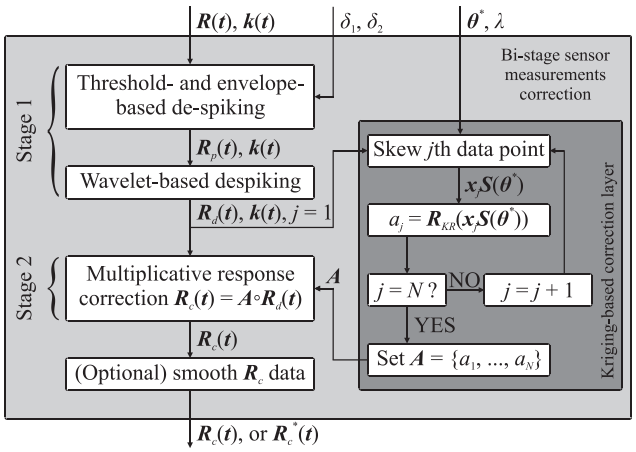

Fig. 9. A two-step post-processing of the PM measurements that involves data de-spiking (stage 1) and kriging-based multiplicative correction (stage 2). Note that δ_1, δ_2, and λ are user-defined whereas $\boldsymbol{\theta}^*$ is obtained from minimization of (15).

2. Generate the envelope $\boldsymbol{E}$ from (12) and obtain $\boldsymbol{R}_p$ using de-spiking method (cf. Section 3.4);
3. Perform wavelet-based peaks elimination (cf. Section 3.2) on $\boldsymbol{R}_p$ to generate $\boldsymbol{R}_d$.
4. Set $j = 1$;
5. Calculate $a_j = \boldsymbol{R}_{KR}(\boldsymbol{x}_j\boldsymbol{S}(\boldsymbol{\theta}^*))$ where $\boldsymbol{x}_j = [\boldsymbol{R}_d(t_j,k(t_j))\ k(t_j)]$ as in (17);
6. If $j = N$ set $\boldsymbol{A} = \{a_j\}_{j\,=\,1,\,\ldots,\,N}$ go to step 7; otherwise set $j = j + 1$ and go to step 5;
7. Calculate $\boldsymbol{R}_c$ from (2);
8. (Optional) Generate $\boldsymbol{R}_c^*$ by smoothing the obtained $\boldsymbol{R}_c$ data.

It should be reiterated that, owing to the low cost, the proposed data de-spiking and correction methods can be implemented in the hardware of Section 2. The goal of the optional $\boldsymbol{R}_c$ data smoothing is to mitigate the effects of local noise (resulting from a limited accuracy of the $\boldsymbol{R}_{KR}$ model) on quality of the responses obtained from the PM sensor. Although in Section 4 the above algorithm has been demonstrated using vector data, it can also be executed on the individual time-series data samples.

4. Results and discussion

Validation of the proposed device and data correction methods have been performed based on a series of measurements obtained in the city of Gdansk (Poland). The particulate matter data has been acquired for a total of 48 days during three campaigns between: (i) January 12th, 2022 and January 18th, 2022, (ii) February 2nd, 2022 and February 16th, 2022, as well as (iii) April 6th, 2022 and May 4th, 2022, respectively—in the same location as the reference monitoring station in order to obtain information required for identification of the correction model, as well as its further verification of the models quality [64]. It should be noted that, even though the proposed data correction methodology is dedicated for improving the $PM_{2.5}$ detection accuracy, the usefulness of the proposed method has also been demonstrated for particles of different diameters. Finally, the presented device and data-correction method have been benchmarked against solutions from the literature in terms of the performance of the measurements. Discussion on cost of the device has also been provided. The quantitative comparison between the sensor and the reference station data is expressed in terms of a root-mean square error (RMSE):

$$e_{\mathrm{RMSE}} = \frac{1}{\sqrt{N}}\sqrt{\sum_{j=1}^{N}\left(\boldsymbol{R}(t_j) - \boldsymbol{R}_r(t_j)\right)^2} \tag{19}$$

Moreover, a coefficient of determination is used to evaluate capability of the corrected responses in terms of fitting the reference station data:

$$r^2 = 1 - \frac{\sum_{j=1}^{N}\left(\boldsymbol{R}(t_j) - \boldsymbol{R}_r(t_j)\right)^2}{\sum_{j=1}^{N}\left(\boldsymbol{R}(t_j) - \langle\boldsymbol{R}_r(t)\rangle\right)^2} \tag{20}$$

In the above equations, the parameter $\boldsymbol{R}$ (when appropriate) refers to the uncorrected, de-spiked, or corrected PM measurements, respectively, whereas N denotes the number of data points in the dataset. It is worth emphasizing that $r^2 = 0$ means that predictions are as bad as random guess, while $r^2 = 1$ refers to perfect match between the sensor and reference station responses. RMSE is expressed in $\mu g/m^3$.

4.1. Identification and validation of the data-correction model

The data-correction model for the $PM_{2.5}$ pollution has been extracted as described in Section 3.5. Here, the first dataset has been used. Fig. 10 shows a time-series comparison of the sensor and reference station measurements [64]. A large discrepancy between the responses ($e_{\mathrm{RMSE}} = 1694$) renders a significant part of the $\boldsymbol{R}_s(\boldsymbol{t})$ data (around 16% of samples is characterized by the relative error above 100%) useless for accurate prediction of the particulate matter concentration. For the same reasons, the responses cannot be directly used for construction of the multiplicative correction model. Consequently, de-spiking of the measurements is mandatory before $\boldsymbol{R}_{KR}$ identification. The RMSE and coefficient of determination for the de-spiked measurements $\boldsymbol{R}_d(\boldsymbol{t})$ are 7.88 and 0.73, respectively, which represents a two orders of magnitude improvement compared to $\boldsymbol{R}_s(\boldsymbol{t})$. The resulting data has been used for construction of the correction model of Section 3.3.

The reference set $\boldsymbol{X}_t$ consists of 12 designs located on a 3×4 grid (cf. Section 3.5). The initial parameters for $\boldsymbol{X}_{ts}$ adjustment are $\boldsymbol{x}^{(0)} = [10\ 1\ 14\ 60\ 0.3\ 0]^T$. The optimized vector $\boldsymbol{x}^* = [10.16\ 0.8\ 13.86\ 61.1\ 0.49\ -0.03]^T$ used for the determination of $\boldsymbol{X}_{ts}^*$, $\boldsymbol{X}_r^*$ and $\boldsymbol{\theta}^*$ has been found through minimization of (15). It should be emphasized that, due to transformation of the already available measurement data, the cost of numerical optimization is low (cf. Section 4.5). Next, medians of the responses that correspond to the sets of transformed measurements located within the distance λ around the reference designs from the $\boldsymbol{X}_r^*$ set have been calculated to obtain $\boldsymbol{Y}_r^*$. Finally, the kriging correction layer $\boldsymbol{R}_{KR}$ has been identified using $\boldsymbol{X}_r^*$ and $\boldsymbol{Y}_r^*$ data (cf. Section 3.5). It should be emphasized that the training set contains only 12 points, whereas $\boldsymbol{Y}_r$ for the considered dataset comprises around 1100 samples. Furthermore, a total of 50 training samples (roughly 4 points around each $\boldsymbol{x}_i \in \boldsymbol{X}_r^*$) have been used for construction of $\boldsymbol{R}_{KR}$, which represent only 4.5% of the data contained in the $\boldsymbol{R}_s(\boldsymbol{t})$ and $\boldsymbol{R}_r(\boldsymbol{t})$ datasets.

The kriging surrogate $\boldsymbol{R}_{KR}$ has been used for refinement of the $\boldsymbol{R}_d$

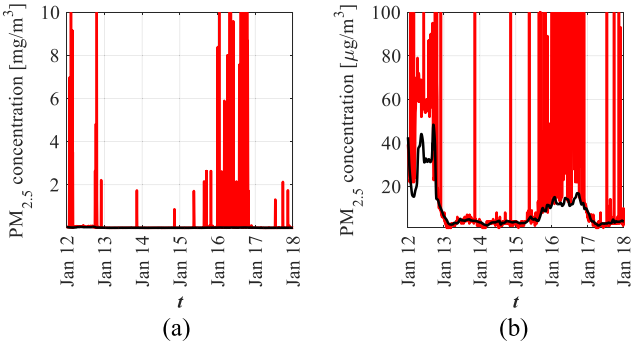

Fig. 10. $PM_{2.5}$ concentration obtained from the sensor (red) and the reference station (black) in: (a) original scale and (b) narrowed down range. High spikes present in large part of the measurements render them useless for the air quality estimation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

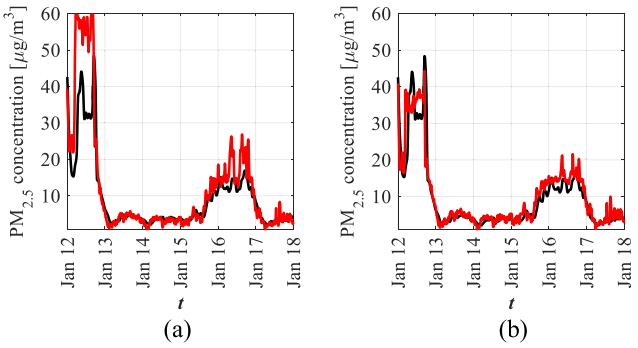

Fig. 11. Comparison of the particulate matter pollution measurements obtained from the reference station (black) and the proposed system (red) after correction using: (a) de-spiking method of Section 3.4 ($e_{RMSE} = 7.88$) and (b) de-spiking followed by multiplication-based correction of Section 3.5 ($e_{RMSE} = 2.98$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

samples form the first dataset. The results shown in Fig. 11 indicate a notable improvement of the $\boldsymbol{R}_c$ responses compared to the $\boldsymbol{R}_d$ data. The coefficient of determination calculated for the corrected PM measurements is 0.87, which represents around 16% improvement w.r.t. the de-spiked responses. At the same time, the RMSE of the corrected measurements is only 2.98, which represents over 2-fold improvement compared to $\boldsymbol{R}_d(\boldsymbol{t})$. Fig. 12 shows $\boldsymbol{R}_d(\boldsymbol{t})$ and $\boldsymbol{R}_c(\boldsymbol{t})$ responses as functions of $\boldsymbol{R}_r(\boldsymbol{t})$ characteristics. The results indicate a significant improvement of the correlation between the reference data and corrected responses compared to the de-spiked measurements. A comparison of the $\boldsymbol{R}_c$ characteristics with the ones obtained after (optional) data smoothing $\boldsymbol{R}_c$*—shown in Fig. 13(a)—suggests that the latter may be useful for mitigating the effects of local correction inaccuracy resulting from the increased noise and/or locally worsened performance of the $\boldsymbol{R}_{KR}$ model. Although the response changes resulting from smoothing are not substantial, they improve r^2 by another 2% to 0.89. The RMSE of the $\boldsymbol{R}_c$* is 2.82 which also represents a slight improvement compared to $\boldsymbol{R}_c$. Evolution of the r^2 factor for consecutive steps of the sensor responses refinement is illustrated in Fig. 13(b).

4.2. Refinememt of the inaccurate PM measurements

The kriging correction model of Section 4.1 has been used for refinement of the PM measurements gathered in the second dataset. As before a large portion of the high-amplitude peaks renders the $\boldsymbol{R}_s$ data useless for evaluation of the air quality. Application of the de-spiking

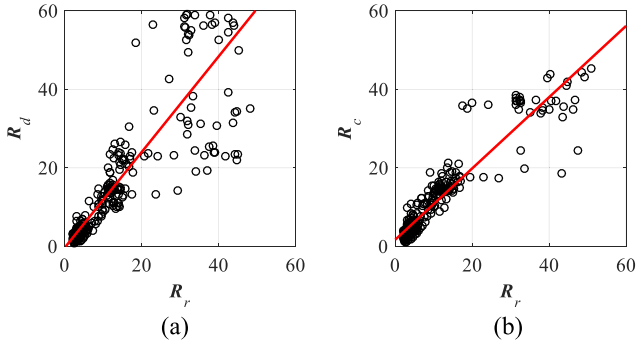

Fig. 12. Corrected sensor measurements as a function of the reference station data: (a) data with eliminated spikes ($r^2 = 0.73$) and (b) responses with removed spikes and corrected using a multiplicative layer ($r^2 = 0.87$). The red line represents linear interpolation of the data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

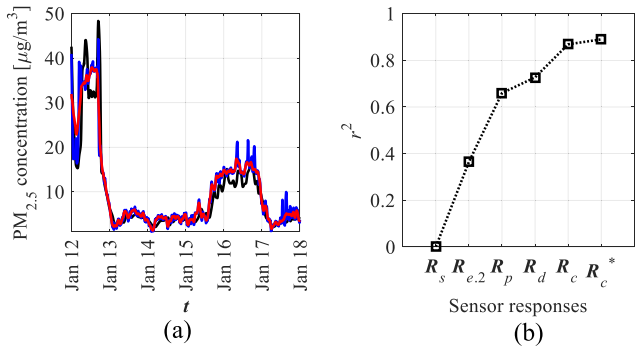

Fig. 13. Correction of sensor-based measurements: (a) comparison of $\boldsymbol{R}_r$ (black), $\boldsymbol{R}_c$ (blue), and $\boldsymbol{R}_c$* (red), as well as (b) "evolution" of the r^2 for various correction steps of the of the PM measurement. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

procedure results in improvement of the e_{RMSE} from 1685 to 5.63 (two orders of magnitude). The multiplicative correction of the $\boldsymbol{R}_d$ data results in further reduction of the RMSE to 3.13 ($r^2 = 0.84$) which contributes to over 44% improvement w.r.t. the de-spiked data. It should be stressed out that the correction has been performed using the $\boldsymbol{R}_{KR}$ model constructed based on the measurements contained in the first dataset. Comparisons of the responses before and after multiplicative correction with the reference data is shown in Fig. 14, whereas visualization of the $\boldsymbol{R}_c$ as a function of the $\boldsymbol{R}_r$ is given in Fig. 15.

To validate the effect of increasing the number of calibration data points (and their number) on the $\boldsymbol{R}_c$ accuracy, the correction model $\boldsymbol{R}_{KR}$ has been re-set using a combination of the measurements from the first two datasets. The surrogate identification procedure follows the discussion from Sections 3.5 and 4.1. Again, the reference set $\boldsymbol{X}_r$ is in the form of a 3 × 4 grid. The optimized vector for adjustment of $\boldsymbol{X}_{ts}$ transformation is $\boldsymbol{x}^* = [9.92\ 0.69\ 14.12\ 59.97\ 0.16\ -0.02]^T$. The new surrogate has been constructed using a total of 60 samples that amount to only 2% of all the available data points. The model has been used for correction of the PM measurements from both datasets. As expected, the results shown in Figs. 16 and 17 imply that increasing the data density improves the predictive performance of the model. The RMSE calculated for the refined sets is 2.64 (model $r^2 = 0.92$) which indicates over 2-fold improvement compared to the data processed only using the de-spiking algorithm. It should be noted that, for the given size of the training set $\boldsymbol{X}_t$, increasing the number of training points beyond certain level results in saturation. Consequently, performance of the correction (understood as improvement of correlation between $\boldsymbol{R}_c$ and $\boldsymbol{R}_r$) depends more on the distribution of samples rather than their density (see Fig. 7).

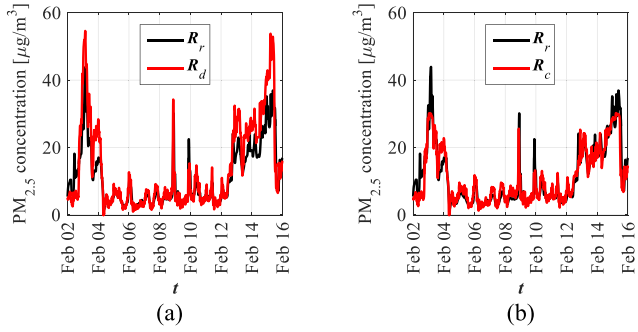

Fig. 14. Comparison of the PM measurements obtained from the reference station (black) and the proposed sensor (red) after correction using: (a) the de-spiking method ($e_{RMSE} = 5.63$) and (b) the combination of de-spiking and multiplication-based correction ($e_{RMSE} = 3.13$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

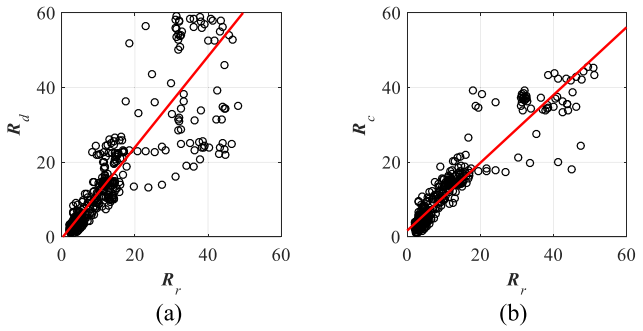

Fig. 15. Refined response of the sensor versus reference station measurements obtained for: (a) de-spiked data ($r^2 = 0.8$) and (b) de-spiked data corrected using multiplicative layer constructed based on the training set from Section 4.1 ($r^2 = 0.84$).

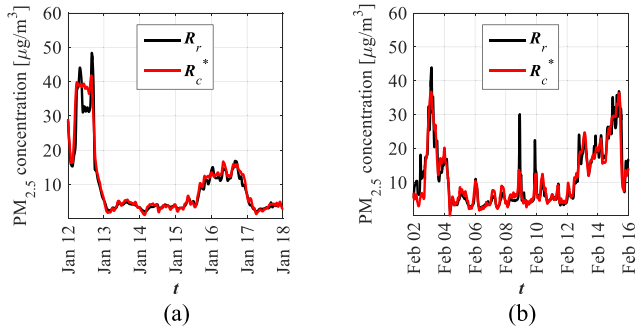

Fig. 16. Comparison of the $PM_{2.5}$ from the reference station (black) and the sensor responses corrected using the model constructed based on the data both datasets (red): (a) the first, as well as (b) the second. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

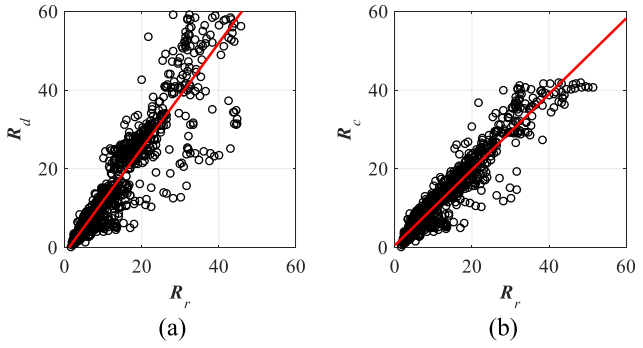

Fig. 17. Combined datasets versus reference data: (a) de-spiked model – $r^2 = 0.81$ and (b) de-spiked model with correction layer extracted from both datasets – $r^2 = 0.93$.

Notwithstanding, adjusting the number of training points $\times \in \boldsymbol{X}_{ts}$* around the individual reference designs (controlled by λ; cf. Section 3.5) may improve performance of the surrogate in terms of its immunity to the local noise.

4.3. Correction of the PM_{10} and PM_1 data

Although the main focus of the presented correction method is improvement of the $PM_{2.5}$ detection accuracy, the approach is also applicable to correction of the particulate matter pollutants concentration characterized by different diameters. Here, the refinement of PM_1 and PM_{10} is considered. The data on pollutants concentration from the sensor of Section 2 and the reference stations have been acquired simultaneously with the $PM_{2.5}$ measurements from both datasets [64]. Figs. 18 and 19 show comparison of the de-spiked and corrected responses for the PM_1 and PM_{10} detections using the calibration data extracted from the first and second measurement campaigns. Note that separate kriging models $\boldsymbol{R}_{KR}$ have been constructed for PM_1 and PM_{10} data. The RMSE factors calculated for the combined datasets before and after correction are 6.38 and 3.05 ($r^2 = 0.89$) for PM_1, as well as 6.32 and 4.07 ($r^2 = 0.88$) for PM_{10}, respectively. It should be noted that the uncorrected PM measurements are characterized by e_{RMSE} of 17.8 for PM_1 and 3437 for PM_{10}. The obtained results indicate that—for the considered datasets—the proposed correction methods provide substantial improvement of PM measurements accuracy when using the low-cost sensor of Section 2 [43].

4.4. The effects of inter-season measurements on the correction performance

The last case study involves performance analysis of the multiplicative correction layer applied to the data obtained during different seasons of the year. It should be reiterated that the PM measurements from datasets (i) and (ii) have been obtained in the winter, whereas the third set has been acquired during the spring. The de-spiking procedure of Section 3.4 has been used to reduce the RMSE by two-orders of magnitude, i.e., from 1302 ($r^2 = 0.0018$) for $\boldsymbol{R}_c$ to 3.93 ($r^2 = 0.84$) for $\boldsymbol{R}_d$, respectively. In the next step, the correction layer from Section 4.2 has been used to further adjust the PM measurements. However, multiplicative modification of the response results in deterioration of the $\boldsymbol{R}_c$ quality manifested by increase of the e_{RMSE} to 4.62 and decrease of r^2 to 0.67. This unintended effect stems from calibration of the correction model using the winter data which do not coincide with the environmental conditions pertinent to the spring season. Consequently, for the considered test case the winter-based model is unsuitable for correction of the $\boldsymbol{R}_d$ response. The problem has been mitigated through re-set of the $\boldsymbol{R}_{KR}$ layer based on the combination of all available datasets. The RMSE calculated for the PM data refined using the new model is 3.84 ($r^2 = 0.85$), which represents a slight improvement compared to the $\boldsymbol{R}_d$.

For the sake of comparison, another $\boldsymbol{R}_{KR}$ correction layer (based only on the calibration data obtained during the spring season) has been constructed and used for refinement of the low-fidelity PM measurements. The resulting $\boldsymbol{R}_c$ response is characterized by $e_{\text{RMSE}} = 3.79$ and $r^2 = 0.85$, respectively. Comparisons of the characteristics generated using both $\boldsymbol{R}_{KR}$ models that result in improvement of the PM measurements quality are shown in Fig. 20. It should be noted that only slight differences between both considered $\boldsymbol{R}_c$ responses can be noticed. A more detailed discussion concerning identification of the correction layer for different environmental conditions is provided in Section 4.6.

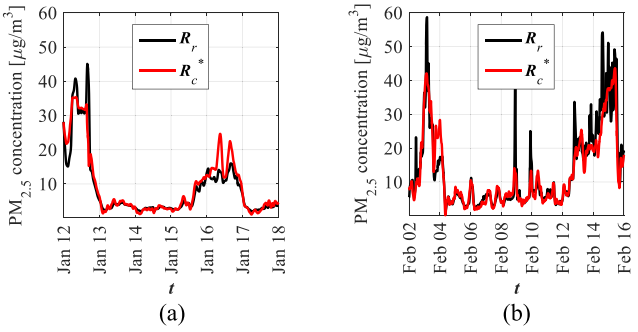

Fig. 18. Comparison of the PM_1 measurements from the reference station (black) and the sensor after correction (red): (a) the first, as well as (b) the second dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

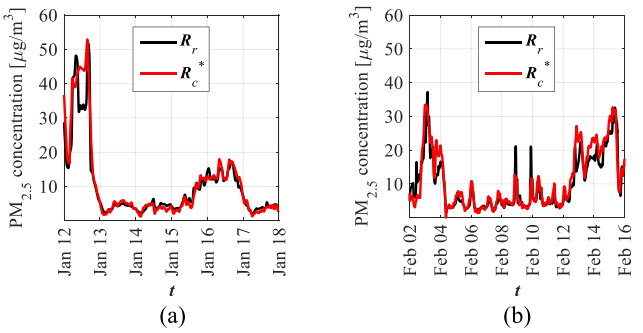

Fig. 19. Comparison of the PM_{10} measurements from the reference station (black) and the sensor after correction (red): (a) the first dataset, as well as (b) the second dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

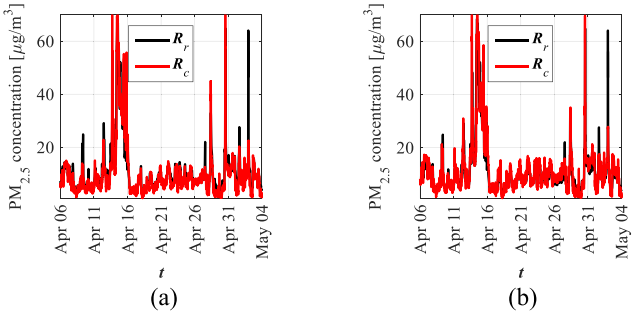

Fig. 20. Comparison of the $PM_{2.5}$ measurements performed in the spring season from the reference station (black) and the sensor after correction (red) using the $\boldsymbol{R}_{KR}$ model constructed based on: (a) all combined datasets, as well as (b) only the third (spring) dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.5. Discussion and comparisons

The presented device and correction methods have been demonstrated based on a total of seven test cases compiled from the measurements acquired during three data-gathering campaigns (cf. Section 4). The tests considered above include: (a) $PM_{2.5}$ correction based on the first dataset, (b) refinement of the $PM_{2.5}$ responses from the second dataset using the surrogate model from (a), as well as improvement of the data from two measurement campaigns using correction layer constructed based on the combination of available data for (c) $PM_{2.5}$, (d) PM_1, (e) PM_{10}. The remaining two test cases concern: (f) refinement of the spring-based $PM_{2.5}$ data using the $\boldsymbol{R}_{KR}$ model identified from the datasets obtained in the winter season, and (g) adjustment of the spring-based PM2.5 data using the kriging model from the same season. A summary of the metrics expressing performance of the de-spiking/correction is gathered in Table 1. It should be reiterated that, for all of the considered measurements, the excessive noise rendered the uncorrected data of little to no use for reliable assessment of the particulate matter concentration in the air. The obtained results clearly show that the de-spiking procedure is mandatory to make use of the gathered data. Moreover, utilization of the proposed correction layer improves the averaged r^2 (i.e., calculated for all of the considered experiments) by around 7% (from 0.80 to 0.87), but also provides nearly 2-fold reduction of the average RMSE (from 5.8 μg/m^3 to 3.3 μg/m^3). For the selected test cases, smoothing of the responses provide further, albeit small, improvement of the PM measurements fidelity. It is worth noting that, with the average RMSE for the corrected measurements of around 3 μg/m^3, the assumption concerning maximum permitted discrepancy between the reference and sensor measurements of 5 μg/m^3 is fulfilled (cf. Section 2.1).

Performance of the proposed data de-spiking and correction methods has been compared against the approaches from the literature [21,31,33,34] based on the $PM_{2.5}$ measurements from the datasets obtained in the winter season (cf. Section 4). A total of eight test cases have been considered that include the smooth-enhanced peaks elimination method (I; Step 1) from [21] combined with (Step 2): (i) random-forest [34], (ii) κ-Kohler theory [31], (iii) humidity correction [33], and (iv) the proposed kriging-based refinement methods, as well as the proposed bi-stage de-spiking method (II; Step 1) coupled with the mentioned response correction approaches (Step 2). For the sake of fair comparison, the random-forest model was constructed using the same number of reference points as the proposed multiplicative correction layer. Due to stochastic nature, performance of the former has been estimated as an average of 30 independent model identification steps. The results gathered in Table 2 indicate that the proposed methods are characterized by noticeably better performance compared to the reference techniques. For the considered test case, application of the presented de-spiking technique results in improvement of the RMSE and r^2 by 64% and 15%, respectively. Similarly, the multiplicative refinement provides the average improvement of the RMSE and r^2 factors by 53% and 13% when peaks elimination has been performed using I, as well as 44% and 7% for the de-spiking using II (proposed) approach. It should be emphasized that, among the considered test cases, the introduced correction mechanism produced both the lowest errors and the highest coefficients of determination.

Table 2
Benchmark of the proposed data refinement methods.

	Method	RMSE [μg/m^3]	r^2	Method	RMSE [μg/m^3]	r^2
Step 1	I	10.36	0.66	II (this work)	6.32	0.81
Step 2	I-(i)*	6.92	0.71	II-(i)*	5.13	0.82
	I-(ii)	7.21	0.73	II-(ii)	3.91	0.89
	I-(iii)	7.86	0.71	II-(iii)	4.39	0.87
	I-(iv)	3.43	0.85	II-(iv) (this work)	2.50	0.93

* Due to stochastic nature of the method, the correction performance has been estimated as an average from 30 instances of the identified model.

Table 1
Summary of the sensor-based PM measurements correction.

Test case	Particulate matter	Uncorrected		De-spiking		Correction layer		Smoothing	
		RMSE [μg/m^3]	r^2	RMSE [μg/m^3]	r^2	RMSE [μg/m^3]	r^2	RMSE [μg/m^3]	r^2
(a)	$PM_{2.5}$	1694	$9\bullet10^{-4}$	7.88	0.73	2.98	0.87	2.82	0.89
(b)	$PM_{2.5}$	1685	10^{-3}	5.63	0.80	3.13	0.84	3.03	0.85
(c)	$PM_{2.5}$	1688	10^{-3}	6.32	0.81	2.64	0.92	2.50	0.93
(d)	PM_1	17.76	0.37	6.38	0.79	3.05	0.89	3.05	0.89
(e)	PM_{10}	3437	$5\bullet10^{-4}$	6.32	0.81	4.07	0.88	3.72	0.89
(f)	$PM_{2.5}$	1302	$2\bullet10^{-3}$	3.93	0.84	3.84	0.85	3.87	0.80
(g)	$PM_{2.5}$	1302	$2\bullet10^{-3}$	3.93	0.84	3.79	0.85	3.59	0.82
Average	N/A	1589	$5\bullet10^{-2}$	5.77	0.80	3.35	0.87	3.26	0.87

It should be re-iterated that the bi-stage correction procedure presented in this work is implemented in MATLAB environment [76]. Consequently, the developed methods and algorithms need to be rewritten in programming languages suitable for embedded devices in order to enable in-situ post-processing of the PM measurements. Usefulness of the method for on-device data refinement can be justified by comparing computational performance of the personal computer (PC) and the developed device. The machine used for post-processing is based on a 10-core Intel Xeon e5-2650 v4 processor (clock speed – 2.2 GHz) and 16 GB RAM. The mobile device, is equipped with a double core ARM Cortex-A9 unit with 667 MHz clock. An important remark is that MATLAB is a high-level programming language and thus it is not well optimized for performance. The correction is performed using only one of the ten available Xeon cores (typical for MATLAB-based programs). The average (per core) performance of the considered processors—evaluated using 7zip package and expressed in the million instructions per second (MIPS) [77]—amount to 3456 and 511 for Intel and ARM, respectively. Consequently, the PC is around 7-fold faster compared to the developed device. Table 3 summarizes the numerical costs (averaged over 100 independent runs) associated with each stage of the PM measurements correction, as well as identification of the kriging model. The presented numerical results are obtained based on the dataset considered in Section 4.4. For the considered test case, the estimated combined costs of de-spiking and multiplicative correction correspond to 0.2 s (PC) and 1.33 s (mobile device), respectively. The approximated cost of in-situ kriging model re-set amounts to 20 min which is acceptable having in mind infrequent execution of the process (i.e., only when the device is located near the reference station), as well as availability of the second ARM core that can ensure the device stability. It should be emphasized that the typical PM measurement intervals are in the order of minutes to dozens of minutes. Therefore, the CPU would be in the idle state for most of the time (important for the sake of energy conservation). Another important remark is that the post-processing cost has been estimated based on a relatively large number of 4235 samples corrected at once. It is expected that, in the real-world scenario, the de-spiking and correction procedures will be performed on the sets comprising at most a few dozen of PM samples. Finally, the provided estimations do not account for a poor optimization of MATLAB-based programs w.r.t. to, e.g., algorithms implemented using C/C++ languages (very popular for embedded devices due to the efficient use of available resources) [78]. The mentioned factors indicate that the data of Table 3 represent the worst-case estimate to the correction of PM measurements.

It is worth reiterating that, owing to the low cost of the proposed spike elimination and multiplicative based correction/identification, refinement of the sensor measurements and enhancements of the surrogate-based correction layer could be implemented directly on the device of Section 2. It seems to be an important advantage over more complex correction schemes such as the ones based e.g., on neural networks, which require substantial computational power in order to extract the necessary information for sensor calibration [65–67].

The cost-breakdown of the proposed device, provided in Table 4, indicates that the price of the utilized sensors amounts to only around 18% of the system price, whereas the data-processing, power, and connectivity units contribute to almost 70% of the overall cost. The results show a potential of the presented platform for further cost optimization, which could be achieved, e.g., through maintaining a balance between the performance and computational power of the system, as well as optimization of the energy consumption. It should be noted, however, that more detailed discussion on the topic is beyond the scope of this work.

Table 3
Estimated CPU-time of the PM data correction.

CPU			Intel e5-2650 v4	ARM Cortex-A9
MIPS/core [77]			3456	511
Relative speed/core			1	0.15
Test	De-spiking	CPU cost	0.04 s	0.27 s*
	Correction		0.16 s	1.06 s*
—	Model re-set		180 s	1200 s*

* Estimated based on the relative speed of the ARM-based CPU w.r.t. Intel Xeon (per single core).

Table 4
Cost-breakdown of the proposed IoT mobile platform for PM measurements.

Category	Item	Price [EUR]	Fraction of cost [%]
Sensors	Particulate matter	40	17.8
	Environmental	84	
Connectivity	GSM module	123	17.7
CPU and power	Zybo Boards with FPGA	233	33.4
	Power supply	129	18.5
Miscellaneous	Interconnections	48	6.87
	Mechanical parts	40	5.73
	Cost per device	697	100

It is worth emphasizing that the direct price-wise, comparison of the system with other solutions from the literature is challenging. One of the main reasons is that, to the best knowledge of the authors, the detailed costs of components used for construction of the competitive in-house mobile sensors are not provided. Furthermore, the mobility aspect of the other solutions is often reduced to focus on the dimensions rather than implementation of the detector as a part of the standalone platform capable not only for acquiring of the PM measurements, but also supporting cloud connectivity, data processing, and cordless operation. The solutions from the literature either involve the discussion of the sensors that could be accommodated to work as components interconnected with mobile communication devices [68–70], or rather stationary devices with capability to be re-set in different location [70,71]. The mentioned cases either lack the necessary mobility aspects (e.g., cordless operation, wireless connectivity), or do not consider the challenges related to deviations of data accuracy.

4.6. Limitations of the method and recommendations

As already mentioned in Sections 3.4 and 3.5 the presented de-spiking and correction techniques are based on a few assumptions that might not hold for in general. Possible bottlenecks of the approach include: (i) limited PM granularity ranges supported by the low-cost sensor, (ii) low (or lack of thereof) correlation between various PM levels, and/or (iii) lack of environmental/reference data that would justify the application of the kriging-based correction model.

The first two challenges are partially related, as correlation between the PM levels changes with the measurement conditions, as well as the difference between physical size of particles accounted for within the given granularity [72–75]. In other words, for the selected PM level, the correlation between the mass concentration of particles might weaken when the physical discrepancy (i.e., size) between them increase. From this perspective, construction of the de-spiking envelope based on the measurements characterized by similar (order-wise) granularity is recommended. Consequently, the low-cost sensor is supposed to support PM measurements at multiple levels around the one of interest for air quality monitoring. Another important aspect associated with the problem involves the environmental effects on the PM-levels correlation. As shown in [72], in closed areas (buildings, transportation means, etc.) the resemblance between the mass concentration of particles characterized by different sizes is noticeably better compared to open space. The effect stems from dynamics of the particles movement induced by the external factors such as wind, temperature changes, etc. However, even when the correlation between the PM samples is limited, the main role of the envelope is to replace the identified non-physical PM measurements. The response is further refined using wavelet-based method and the kriging correction layer.

As explained in Section 3.5, the kriging-based correction model is set-up within the ranges specified by the optimized lower and upper bounds that correspond to the environmental conditions for which PM measurements from the reference station and sensor are available. Consequently, the model is useful for correction only within the mentioned ranges or slightly beyond them (due to the ability to indicate trends between the environmental factors and the discrepancy between the low-fidelity and reference measurements). In other words, one cannot expect that the kriging model identified in the winter season will be suitable for correction of the sensor data gathered during the spring (cf. Section 4.4). The problem can be mitigated by application of the multiplicative refinement only to the fraction of the low-fidelity PM data which is within the ranges of model validity. On the other hand, given availability of the high-fidelity measurements (e.g., due to temporary location of the mobile device in the vicinity of the reference station), re-set/extension of the correction layer can be performed to improve generalization capability of the model. Furthermore, the device can be configured so as to switch between the models identified for various conditions (e.g., seasons; cf. Section 4.4). These concepts will be investigated as a part of future work. An important remark is that, even when the kriging model is unsuitable for correction due to the out-of-bounds environmental conditions, the de-spiking procedure offers a substantial improvement of the detected PM quality with respect to the reference data (at least for the considered test cases; cf. Section 4.5).

5. Conclusions

In this work, a mobile IoT-capable platform for measurements of $PM_{2.5}$ air pollutants has been presented. The device features a modular architecture that incorporates the FPGA with integrated ARM processor, PM detector, connectivity module, as well as environmental sensors. The used low-cost PM detector is characterized by a limited measurement accuracy manifested in the form of non-physical spikes in concentration of the measured pollutants, as well as non-linear drift of the PM measurements compared to the reference data. The measurements fidelity has been improved using the proposed data de-spiking and correction algorithms. The first is implemented as a combination of the lower bound envelope extracted from the multi-level PM measurements followed by the wavelet-based data post-processing. The second is performed using a multiplicative layer in the form of the kriging surrogate identified using the calibration set. The latter is carefully extracted from the PM sensor and reference station measurements performed in the same time interval and location. The device and the correction methods have been both demonstrated based on the measurements obtained during three data-gathering campaigns. The obtained results show that the proposed correction algorithms improve the sensor-based PM measurements fidelity by around two orders of magnitude compared to the responses for which post-processing has not been considered. Furthermore, the methods offer up to over 3-fold reduction of the measurement error as compared to the benchmark approaches from the literature. Although the presented methods and algorithms have been developed with correction of $PM_{2.5}$ pollution in mind, their usefulness for refinement of PM_1 and PM_{10} measurements has also been demonstrated.

It should be emphasized that identification of the correction model, as well as data de-spiking procedure proposed in this work are numerically cheap. Due to availability of the ARM processor, as well as connectivity module, the proposed device has the potential for performing in-situ re-set of the multiplicative correction model based on the data obtained from the nearby located reference stations (when available). The mentioned concept will be the focus of the future work. Fusion of the measurements from multiple sensors oriented towards providing temporal/spatial monitoring of the PM pollution in the given area, cost-related optimization of the platform, as well as validation of the presented measurement correction methods on different air pollutants will also be investigated. Finally, the forthcoming research will concentrate on the integration of multiple measurement platforms in the form of the IoT-based, cloud-connected system for reliable air quality monitoring within the dynamically changing environments.

CRediT authorship contribution statement

Marek Wojcikowski: Investigation, Software, Data curation, Funding acquisition, Project administration. **Bogdan Pankiewicz:** Software, Resources, Validation. **Adrian Bekasiewicz:** Conceptualization, Methodology, Software, Supervision, Validation, Formal analysis, Visualization, Writing – original draft. **Tuan-Vu Cao:** Funding acquisition. **Jean-Marie Lepioufle:** . **Islen Vallejo:** . **Rune Odegard:** . **Hoai Phuong Ha:** .

Declaration of Competing Interest

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

Acknowledgements

This work was supported in part by the National Centre for Research and Development Grant NOR/POLNOR/HAPADS/0049/2019-00.

The authors would like to thank the Agency of Regional Atmospheric Monitoring Gdansk-Gdynia-Sopot for providing (free of charge) the data from the reference measurements stations that were mandatory for construction and analysis of our models.

References

[1] J. Lelieveld, J.S. Evans, M. Fnais, D. Giannadaki, A. Pozzer, The contribution of outdoor air pollution sources to premature mortality on a global scale, Nature 525 (7569) (2015) 367–371.
[2] M. Kampa, E. Castanas, Human health effects of air pollution, Environ. Pollut. 151 (2) (2008) 362–367.
[3] N. Manojkumar, B. Srimuruganandam, Health effects of particulate matter in major Indian cities, Int. J. Environ. Health. Res. 31 (3) (2021) 258–270.
[4] J. Wang, B. Zhao, S. Wang, F. Yang, J. Xing, L. Morawska, A. Ding, M. Kulmala, V.-M. Kerminen, J. Kujansuu, Z. Wang, D. Ding, X. Zhang, H. Wang, M.i. Tian, T. Petäjä, J. Jiang, J. Hao, Particulate matter pollution over China and the effects of control policies, Sci. Total Environ. 584-585 (2017) 426–447.
[5] Y. Awe, et al., *Air Quality Management – Poland*, Report no, The World Bank, 2017. AUS0000585.
[6] C.A. Pope, R.T. Burnett, M.J. Thun, E.E. Calle, D. Krewski, K. Ito, G.D. Thurston, Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution, J. Am. Med. Assoc. 287 (2002) 1132–1141.
[7] C.A. Pope, D.W. Dockery, Health effects of fine particulate air pollution: Lines that connect, J. Air Waste Manag. Assoc. 56 (6) (2006) 709–742.
[8] L. Yang, C. Li, X. Tang, The impact of PM2.5 on the host defense of respiratory system, Front. Cell Dev. Biol., vol. 8, art no. 91, 2020.
[9] T. Li, R. Hu, Z.i. Chen, Q. Li, S. Huang, Z. Zhu, L.-F. Zhou, Fine particulate matter (PM2.5): The culprit for chronic lung diseases in China, Chronic Dis. Transl. Med. 4 (3) (2018) 176–186.
[10] World Health Organization, Av. Appia 20, 1211 Geneve, Switzerland.
[11] H. Orru, K.L. Ebi, B. Forsberg, The interplay of climate change and air pollution on health, Curr. Envir. Health Rpt. 4 (4) (2017) 504–513.
[12] V. Ramanathan, P.J. Crutzen, J.T. Kiehl, D. Rosenfeld, Atmosphere: Aerosols, climate, and the hydrological cycle, Science 294 (2001) 2119-2124.
[13] Oecd, The Cost of Air Pollution: Health Impacts of Road Transport, OECD Publishing, Paris, 2014.
[14] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, E. Bezirtzoglou, Environmental and Health Impacts of Air Pollution: A Review, Front. Public Health, vol. 8, 2020.
[15] S. Xie, J.R. Meeker, L. Perez, W. Eriksen, A. Localio, H. Park, A. Jen, M. Goldstein, A.F. Temeng, S.M. Morales, C. Christie, R.E. Greenblatt, F.K. Barg, A.J. Apter, B. E. Himes, Feasibility and acceptability of monitoring personal air pollution exposure with sensors for asthma self-management, Asthma Res Pract. 7 (1) (2021), https://doi.org/10.1186/s40733-021-00079-9.
[16] A. Gressent, L. Malherbe, A. Colette, H. Rollin, R. Scimia, Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value, Environ. Int., vol. 143, art. no. 105965, 2020.
[17] H. Guo, G. Dai, J. Fan, Y. Wu, F. Shen, Y. Hu, A mobile sensing system for urban $P_{M2.5}$ monitoring with adaptive resolution, J. Sensors, art. no. 7901245, 2016.
[18] P. deSouza, A. Anjomshoaa, F. Duarte, R. Kahn, P. Kumar, C. Ratti, Air quality monitoring using mobile low-cost sensors mounted on trash-trucks: Methods development and lessons learned, Sustain. Cities Soc. vol. 60, art. no. 102239, 2020.

[19] S. Araki, H. Shimadera, K. Yamamoto, A. Kondo, Effect of spatial outliers on the regression modelling of air pollutant concentrations: A case study in Japan, Atmos. Env. 153 (2017) 83–93.
[20] J.-M. Lepioufle, L. Marsteen, M. Johnsrud, Error prediction of air quality at monitoring stations using random forest in a total error framework, Sensors 21, art. no. 2160, 2021.
[21] H. Wu, X. Tang, Z. Wang, et al., Probabilistic automatic outlier detection for surface air quality measurements from the China national environmental monitoring network, Adv. Atmos. Sci. 35 (2018) 1522–1532.
[22] K.K. Johnson, M.H. Bergin, A.G. Russell, G.S.W. Hagler, Field test of several low-cost particulate matter sensors in high and low concentration urban environments, Aerosol Air Qual. Res. 18 (3) (2018) 565–578.
[23] J. Yun, J. Woo, IoT-enabled particulate matter monitoring and forecasting method based on cluster analysis, IEEE Internet Things J. 8 (9) (2021) 7380–7393.
[24] C. Báthory, Z. Dobó, A. Garami, Á. Palotás, P. Tóth, Low-cost monitoring of atmospheric PM—development and testing, J. Environ. Manage. 304(art no. 114158) (2022).
[25] O. Kracht, M. Gerboles, H.I. Reuter, First evaluation of a novel screening tool for outlier detection in large scale ambient air quality datasets, Int. J. Environmental Pollution 55 (1/2/3/4) (2014) 120, https://doi.org/10.1504/IJEP.2014.065912.
[26] Y. Yao, A. Sharma, L. Golubchik, R. Govindan, Online anomaly detection for sensor systems: A simple and efficient approach, Perform. Eval. 67 (2010) 1059–1075.
[27] M. Bobbia, M. Misiti, Y. Misiti, J.-M. Poggi, B. Portier, Spatial outlier detection in the PM monitoring network of Normandy (France), Atmos. Pollut. Res. 6 (2015) 476–483.
[28] M. Čampulová, P. Veselík, J. Michálek, Control chart and Six sigma based algorithms for identification of outliers in experimental data, with an application to particulate matter PM_{10}, Atmos. Pollut. Res. 8 (4) (2017) 700–708.
[29] J. Mroczka, The cognitive process in metrology, Measurement 46 (2013) 2896–2907.
[30] M. Badura, P. Batog, A. Drzeniecka-Osiadcz, P. Modzel, "Evaluation of low-cost sensors for ambient PM2.5 monitoring, J. Sensors, art no. 5096540, 2018.
[31] L.R. Crilley, M. Shaw, R. Pound, L.J. Kramer, et al., Evaluation of a low-cost optical particle counter (Alphasense OPC-N2) for ambient air monitoring, Atmos. Meas. Tech. 11 (2018) 709–720.
[32] C. Lin, N. Masey, H. Wu, M. Jackson, et al., Practical field calibration of portable monitors for mobile measurements of multiple air pollutants, Atmosphere 8(art. no. 231) (2017).
[33] A. Di Antonio, O.A.M. Popoola, B. Ouyang, J. Saffell, R.L. Jones, Developing a relative humidity correction for low-cost sensors measuring ambient particulate matter, Sensors 18, art. no. 2790, 2018.
[34] H.-Y. Liu, P. Schneider, R. Haugen, M. Vogt. Performance assessment of a low-cost $PM_{2.5}$ sensor for a near four-month period in Oslo, Norway, Atmosphere 10, art. no. 41, 2019.
[35] Y.F. Xing, Y.H. Xu, M.H. Shi, Y.X. Lian, The impact of PM2.5 on the human respiratory system, J. Thorac. Dis. 8 (1) (2016) E69–E74.
[36] S. Feng, D. Gao, F. Liao, F. Zhou, X. Wang, The health effects of ambient PM2.5 and potential mechanisms, Ecotoxicol. Environ. Saf. 128 (2016) 67–74.
[37] R.H. Anderson, et al., WHO air quality guidelines global update 2005, World health Organization, Bonn, Germany, 2005.
[38] H. Adair-Rohani, et al., WHO global air quality guidelines: Particulate matter (PM and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, World health Organization, Geneva, Switzerland, 2021.
[39] A. Rodriguez-Alvarez, Air pollution and life expectancy in Europe: Does investment in renewable energy matter? Sci. Total Environ. 792 (2021) 148480.
[40] Zybo Zynq-7000, Digilent Inc., 1300 NE Henley Ct. Suite 3, Pullman, WA 99163, USA.
[41] I^2C-bus specification and user manual, UM10204, NXP Semiconductors, High Tech Campus 60, 5656 AG Eindhoven, The Netherlands.
[42] Serial Peripheral Interface (SPI) User Guide, SPRUGP2A, Texas Instruments, 12500 TI Blvd., Dallas, TX 75243, USA.
[43] SPS30 - $PM_{2.5}$ Sensor for HVAC and air quality applications (datasheet), Sensirion AG, Laubisruetistrasse 50, 8712 Stafa, Switzerland.
[44] BG96, Quectel, Shanghai Business Park Phase III (Area B), No.1016 Tianlin Road, Minhang District, 200233 Shanghai, China.
[45] HP206C, Hoperf, Shenzhen Hope Microelectronics Co., Ltd., 30th floor of 8th Building, C Zone, Vanke Cloud City, Xili Sub-district, Nanshan, Shenzhen, China.
[46] STM32L031, STMicroelectronics, 39 Chemin du Champ-des-Filles, Geneve, Switzerland.
[47] MAX9938, Maxim Integrated, 160 Rio Robles, San Jose, CA 95134 USA.
[48] P1832J 18650 Cell, Keeppower Technology Ltd., 5F, Bldg 4, FenMenao Industrial Park, Gangtou, Bantian, Long Gang District, Shenzhen 518129, China.
[49] Z. Nenadic, J.W. Burdick, Spike detection using the continuous wavelet transform, IEEE Trans. Biomed. Eng. 52 (1) (2005) 74–87.
[50] Y.X. Xia, Y.-Q. Ni, A wavelet-based despiking algorithm for large data of structural health monitoring, Int. J. Distrib. Sensor Networks, 14(12) (2018).
[51] A.I. Forrester, A. Sobester, A.J. Keane, Multi-fidelity optimization via surrogate modelling, Roy. Soc. Proc. Royal Soc. 463 (2007) 3251–3269.
[52] T.W. Simpson, J.D. Pelplinski, P.N. Koch, J.K. Allen, Metamodels for computer-based engineering design: survey and recommendations, Eng. Comput. 17 (2001) 129–150.
[53] S. Koziel, L. Leifsson, Multi-point response correction for reduced-cost EM-simulation-driven design of antenna structures, Microwave Opt. Tech. Lett. 55 (9) (2013) 2070–2074.
[54] S. Koziel, Q.S. Cheng, J.W. Bandler, Space mapping, IEEE Microwave Mag. 9 (6) (2008) 105–122.
[55] X. Yang, S.A. Shamma, A totally automated system for the detection and classification of neural spikes, IEEE Trans. Biomed. Eng. 35 (10) (1988) 806–816.
[56] S.M. Kay, Fundamentals of Statistical Signal Processing: Detection Theory, Englewood Cliffs, Prentice-Hall, New Jersey, 1998.
[57] Y. Tian, K.S. Burch, Automatic spike removal algorithm for Raman spectra, Appl. Spectrosc. 70 (11) (2016) 1861–1871.
[58] A.X. Patel, P. Kundu, M. Rubinov, P.S. Jones, et al., A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series, Neuroimage. 95 (100) (2014) 287–304.
[59] D.B. Percival, A.T. Walden, *Wavelet Methods for Time Series Analysis*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cabridge, 2000.
[60] S.N. Lophaven, H.B. Nielsen, J. Sondergaard, "dace,, a Matlab kriging toolbox", Technical University of Denmark, 2002.
[61] L.N. Trefethen, D. Bau, Numerical Linear Algebra, Society for Industrial and Applied Mathematics, Philadelphia, 1997.
[62] B. Beachkofski, R. Grandhi, "Improved distributed hypercube sampling," American Institute of Aeronautics and Astronautics, AIAA 2002-1274, 2002.
[63] R. Davis, P. John, Application of taguchi-based design of experiments for industrial chemical processes, in: V. Silva [Ed.], Statistical Approaches With Emphasis on Design of Experiments Applied to Chemical Processes, IntechOpen, London, 2018.
[64] AM3 measurement station, Agency of Regional Atmospheric Monitoring Gdansk-Gdynia-Sopot, Brzozowa 15 A, 80-243 Gdansk, Poland.
[65] Y.-C. Chen, T.-C. Lei, S. Yao, H.-P. Wang, „PM2.5 prediction model based on combinational hammerstein recurrent neural networks," Mathematics, vol. 8, art. no. 2178, 2020.
[66] S. Chae, J. Shin, S. Kwon, et al., PM10 and PM2.5 real-time prediction models using an interpolated convolutional neural network, Sci Rep., vol. 11, art. no. 11952, 2021.
[67] N. Onyeuwaoma, D. Okoh, B. Okere, „A neural network-based method for modeling PM 2.5 measurements obtained from the surface particulate matter network, Environ. Monit. Assess. 193(5) (2021) art. no. 261.
[68] H.H. Hsu, G. Adamkiewicz, E.A. Houseman, J.D. Spengler, J.I. Levy, Using mobile monitoring to characterize roadway and aircraft contributions to ultrafine particle concentrations near a mid-sized airport, Atmos. Environ. 89 (2014) 688–695.
[69] M. Nyarku, M. Mazaheri, R. Jayaratne, M. Dunbabin, et al., Mobile phones as monitors of personal exposure to air pollution: Is this the future?" PLoS ONE, 13(2) (2018) art. no. e0193150.
[70] F. Gozzi, G. Della Ventura, A. Marcelli, "Mobile monitoring of particulate matter: State of art and perspectives", *Atmospheric*, Pollution Research 7 (2) (2016) 228–234.
[71] C. Bathory, Z. Dobo, A. Garami, A. Palotas, P. Toth, "Low-cost monitoring of atmospheric PM—development and testing," J. Environ. Manage. 304, art no. 114158, 2022.
[72] Z. Qiu, J. Song, X. Xu, Y. Luo, R. Zhao, W. Zhou, B. Xiang, Y. Hao, Commuter exposure to particulate matter for different transportation modes in Xi'an, China, Atmos. Pollution Res. 8 (5) (2017) 940–948.
[73] M. Yang, Y.-M. Guob, M.S. Bloom, S.C. Dharmagee, et al., Is PM1 similar to PM2.5? A new insight into the association of PM1 and PM2.5 with children's lung function, Env. Int. 145, art no. 106092, 2020.
[74] X. Zhou, Z. Cao, Y. Ma, L. Wang, et al., Concentrations, correlations and chemical species of PM2.5/PM10 based on published data in China: Potential implications for the revised particulate standard, Chemosphere 144 (2016) 518–526.
[75] H. Fan, C. Zhao, Y. Yang, X. Yang, Spatio-temporal variations of the PM2.5/PM10 ratios and its application to air pollution type classification in China, Front. Env. Sci. 9 (2021).
[76] MathWorks MATLAB, v. 2013a, MathWorks, Inc., 3 Apple Hill Drive, Natick, 01760 MA, USA.
[77] J.W. Smith, I. Sommerville, Workload classification & software energy measurement for efficient scheduling on private cloud platforms, arXiv, art no. 1105.2584, 2011.
[78] T. Andrews, Computation time comparison between MATLAB and C++ using launch windows, California Polytechnik State University, San Luis, 93407 CA, 2012.
[79] H. Grimm, D.J. Eatough, Aerosol measurement: the use of optical light scattering for the determination of particulate size distribution, and particulate mass, including the semi-volatile fraction, J. Air & Waste Manag. Assoc. 59 (1) (2009) 101–107.