# Virtual Sensors: Toward High-Resolution Air Pollution Monitoring using AI and IoT

Martha Arbayani Zaidan, Naser Hossein Motlagh, Brandon E. Boor, David Lu, Petteri Nurmi, Tuukka Petäjä, Aijun Ding, Markku Kulmala, Sasu Tarkoma, and Tareq Hussein

## Abstract

The present article contributes a research vision for *virtual sensing* that combines Artificial Intelligence (AI) and Internet of Things (IoT) to increase the coverage of air quality information. Virtual sensors take advantage of correlations between different pollutants to estimate the concentrations of pollutants for which no affordable sensors are available. We cover key requirements and challenges, reflecting on the current state-of-the-art and identifying key research challenges. We also demonstrate the potential and feasibility of virtual sensing through experiments conducted with data from Helsinki, Finland, which show how standard $PM_{2.5}$ and temperature measurements can be used to provide reliable estimates of $CO_2$ and black carbon concentrations. We also discuss potential applications that can benefit from the implementation of virtual air pollution sensors and establish a research roadmap for the path forward.

## Introduction

Air pollution is a major environmental health challenge impacting the respiratory and cardiovascular health of people worldwide. According to the World Health Organization (WHO), air pollution causes approximately 7 million deaths each year [1]. To monitor air pollution, professional-grade air quality monitoring stations are deployed to provide information on the concentrations and characteristics of gas- and particle-phase air pollutants in urban environments. These stations are deployed in a limited number of urban areas, thus they only provide accurate air quality information near the measurement site. Satellite-based remote sensing can provide better spatial coverage, however, often at the expense of low temporal coverage [2]. Unfortunately, both of these approaches are not effective in obtaining localized air pollution information at high spatio-temporal resolution in urban areas.

Due to low-cost sensors' (LCSs) affordability, easy installation, and limited maintenance, these sensors can be deployed at a massive scale and can offer localized and high-resolution air pollution information in urban areas [3]. However, LCSs often suffer from poor sensing accuracy, which can be corrected using various calibration techniques [4]. Another challenge with LCSs is that, due to their size, cost, power consumption, reliability, and limited sensing capability, LCSs are not always equipped with a complete package of sensors that can concurrently measure all important variables, such as meteorological conditions and numerous gas- and particle-phase species. In urban areas, the characteristics of pollutants vary across different locations and the pollutants can result from diverse sources. For example, traffic and industrial activities cause black carbon (BC) and nitrogen dioxide, whereas buildings and construction activities emit particulate matter and dust [5]. Ensuring detailed and actionable information requires novel monitoring approaches that can provide accurate and high-resolution information about a wide range of pollutants. These approaches also need to be easy and affordable to deploy and maintain as otherwise the costs of monitoring limit the scale at which information can be obtained.

The present article contributes a research vision for *virtual sensors* (also known as soft, proxy or surrogate sensors [6]) that combine Artificial Intelligence (AI) and sensor-equipped Internet of Things (IoT) devices to estimate the values of a quantity of interest for which no practical or affordable option is available. The virtual sensors concept is illustrated in Fig. 1. In the context of air quality monitoring, virtual sensors take advantage of correlations between pollutants to construct models that can estimate the values of important pollutants for which no sensors are available using the values of other pollutants [4]. We discuss the key requirements for enabling accurate virtual sensors, and present challenges and enablers for large-scale use of virtual sensors. We also reflect on the state-of-the-art to identify key research gaps and to establish a research roadmap for the path forward. We also discuss potential applications that can benefit from the availability of virtual sensors. We demonstrate the feasibility and benefits of virtual sensors through experiments conducted in Helsinki, Finland and which show how a combination of a common pollutant ($PM_{2.5}$) and environmental variables (temperature and relative humidity) can be used to obtain accurate estimates of $CO_2$ and black carbon (BC) concentrations. Being able to estimate BC is particularly significant as it is a primary air pollutant that is generated by fuel combustion and biomass burning and that is associated with severe adverse human health outcomes. Low-cost sensors for BC currently suffer from low sensing accuracy — approximately 25 percent compared to reference sensing instruments — and professional-grade BC sensors cost in excess of $10,000 USD, making them too expensive for large-scale deployments across a city [7]. Most black carbon sensors also require regular maintenance as they collect the pollutant on a filter that needs to be replaced regularly (around 1–2 weeks, depending on the extent of pollution) [8]. Indeed, as our results show, virtual sensors overcome these limitations and offer an affordable approach that overcomes these limitations and that can be used to increase the resolution at which BC information (or other pollutants) can be collected.

## Requirements for Virtual Sensing

Virtual sensors are defined as machine learning models that are integrated into LCSs and take input measurements of diverse

Martha Arbayani Zaidan is with the University of Helsinki, Finland and Nanjing University, China.

Naser Hossein Motlagh and Petteri Nurmi are with the University of Helsinki, Finland.

Brandon E. Boor is with Purdue University, USA.

David Lu is with Clarity Movements Company, USA.

Tuukka Petäjä and Markku Kulmala are with the University of Helsinki, Finland and Nanjing University, China.

Aijun Ding is with Nanjing University, China.

Sasu Tarkoma and Tareq Hussein are with the University of Helsinki, Finland.

Digital Object Identifier: 10.1109/IOTM.001.2200103

 2576-3180/23/$25.00 © 2023 IEEE 

pollutants and environmental variables. In practice, the pollutant measurements captured by LCSs are prone to errors and inaccuracies [9] and thus the measurements should be processed using machine learning-based calibration prior to using them in the virtual sensors [4]. The development block in Fig. 1 demonstrates the key requirements to develop virtual sensors, which we explain in this section.

**Reliable data sets:** Virtual sensors are developed by collecting LCS measurements together with gold-standard reference measurements, which usually requires co-locating the LCS next to a reliable reference instrument. To bolster the development of virtual sensors, there is a need for high-quality data sets that contain the necessary measurements. These data sets need to include accurate and high-resolution time-series data. They can also contain only a small amount of missing data. Reliable data sets cover diurnal cycles and seasonal variability with good spatial coverage, i.e., preferably collected from multiple locations within a city.

**Data driven models:** Establishing accurate and robust models for virtual sensors requires a machine learning (ML) pipeline that can produce such outputs [10]. ML pipelines are complex and designing them consists of setting up an architecture, defining tuning variables, applying optimization methods, and evaluating the pipelines against appropriate performance metrics. In constructing the virtual sensor pipelines, the training and testing data sets must be reliable, and efficient data processing techniques such as data harmonization and normalization, data imputations, and feature extraction need to be applied. Ensuring the developed models are accurate and robust requires a better understanding of the performance and caveats of different processing techniques, and evaluation metrics that can ensure the deployed virtual sensors will produce data that mimics physical sensors.

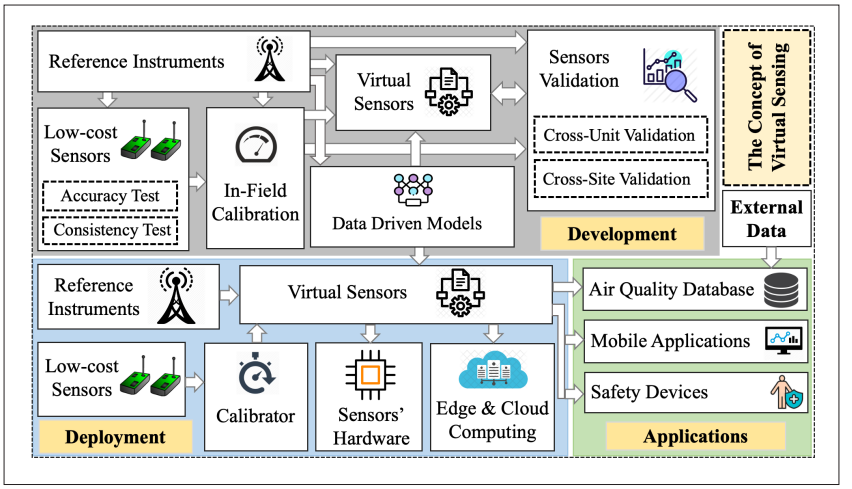


FIGURE 1. The phases and components of developing and deploying virtual air pollution sensors. The development phase uses the data generated by calibrated LCSs and reference instruments in order to establish virtual sensors based on data-driven models. The deployment phase utilizes the data from reference instruments and calibrated LCSs to be fed into the developed virtual sensors which can be embedded into LCS hardware or other computing platforms. The application phase uses the outputs of the virtual sensors in order to contribute to air quality databases and benefit diverse applications.

**Low-cost sensor validation:** While reference instruments generate ground truth data, LCSs often do not provide reliable data. Thus, LCSs should be validated for internal and external consistency. Internal consistency requires comparing the LCS units against other LCS units and ensuring the readings they produce are sufficiently similar in the same context, whereas external consistency effectively measures the accuracy, i.e., how well the measurements align with a reference instrument that provides gold-standard reference measurements. Beyond measuring consistency, there is a need for unified protocols and processes on how to perform and interpret the results of this type of measurement.

**Low-cost sensor calibration:** Virtual sensors based on LCSs that are deployed should be as accurate as possible. The inputs of an LCS, which form the input to the virtual sensor, are subject to noise and errors [9] and thus there is a need to ensure the measurements are accurate. The accuracy can be improved using *calibration*. In practice, there are two types of calibrations: laboratory calibration and in-field sensor calibration. The former refers to validating and calibrating the sensors under laboratory conditions. Specifically, the LCSs and the reference instruments are placed inside a chamber where different variables, such as temperature and RH, are controlled. Then, the readings of the LCSs are compared against the data obtained from reference instruments. If the readings of LCSs do not follow the readings of reference measurements, the LCSs are adjusted or integrated with a correction function. This is feasible as long as some of the devices periodically can access a reference station as this can be used to learn a machine learning model which is then transferred to other devices [3]. This approach, however, is not feasible for regions where no reference stations are available. In this case, the model needs to be pre-calibrated in a laboratory prior to deployment and the conditions during calibration should emulate the conditions in the deployment environment as closely as possible. In many cases, laboratory calibration is not sufficient as controlled laboratory conditions do not capture the dynamics and fluctuations of field settings. This necessitates performing an in-field sensor calibration, where the LCSs are placed side-by-side with the reference instruments in-field. Then, their measurements are compared and if the readings of the LCSs do not follow the readings of the reference instruments, the LCSs are calibrated (by learning a correction function using machine learning [9]). For the in-field sensor calibration, if the field experiment is long enough, the calibration model is typically more reliable because the model captures wide ranges of environmental data, such as seasonal effects. As with other parts of the pipeline, a key requirement is to have calibration processes and measurement processes that ensure the process is carried out as accurately as possible. In-field calibration can also benefit areas with no reference stations as the model can be transferred to other areas with sufficiently similar environmental conditions (so-called calibration transfer) [9].

**Virtual sensor validation:** The validation of virtual sensors similarly should consider multiple different aspects and follow well-defined and robust protocols. First, virtual sensors developed based on individual sensors can be calibrated similarly to the inputs of the models, and this requires validating the virtual sensor against a reference instrument. Second, the virtual sensor should be validated for internal consistency using cross-unit validation, i.e., the virtual sensor model should not be sensitive to the sensor unit that runs it but should work accurately on multiple LCSs that contain the same sensors. Finally, virtual sensors should be subjected to cross-site validation where the deployed calibration and virtual sensor models are established with data from one location and tested and evaluated with another location. This validation is required since the models often do not function well when they are tested on different sites. However, the solution is to establish a *generalized calibration and virtual sensor model* which can be achieved by developing them using data from multiple sites or through transfer learning techniques.

| Components | State-of-the-Art | Key Research Challenges | Possible Solutions |
|---|---|---|---|
| **Sensing** | Monitoring stations, reference instruments, and crowd-sourcing methods | A limited number of reference stations in cities, leading to imbalanced spatio-temporal data | Increasing the number of mini-stations hosting reference instruments |
| **Data** | Ground truth data obtained from reference stations | Other sensing solutions such as LCSs which often suffer from low accuracy | Calibration models can be developed using advanced machine learning methods |
| **Models** | Existing models are developed based on statistical approaches | The deployed models work appropriately on specific environments and regions | Generalized models enable them to work in multi-sites and across different environments |
| **Computing Platforms** | Edge and cloud computing paradigms enable performing calibration and computation of virtual sensors | Sensors that are deployed in distant locations do not have access to computing platforms | Advanced communication technologies such as 5G and 6G offer access to edge and cloud platforms |
| **Performance Monitoring** | Sensor failure is identified via missing data, while anomaly and degradation are monitored by colocating them next to reference instruments | Reference instruments are not always available for in-field deployment | Anomaly and fault detection algorithms can remotely identify and diagnose the problems |

TABLE 1. The components of virtual sensors: state-of-the-art, key research challenges and possible solutions.

# Challenges

We next discuss the different components of virtual sensors, reflecting on the current state-of-the-art and identifying key research challenges. We also discuss potential enablers for these challenges within platforms needed for activating virtual sensors and monitoring that explains the performance of virtual sensors. A summary of the challenges and possible solutions are presented in Table 1.

## Platforms

As illustrated in the deployment block in Fig. 1, generally, virtual sensors can operate on different types of platforms: dedicated computing platforms that are part of fixed infrastructure, embedded LCS platforms, and network deployments that reside on the edge or in the cloud. The cost of using the underlying AI model for estimating pollution values is generally negligible and the main source of resource drain comes from updating the model.

**Fixed deployments:** Virtual sensors can operate as part of fixed deployments, e.g., as part of the urban infrastructure or as part of commercial deployments. In these cases, the virtual sensors would operate on dedicated computing platforms. Note that this could even mean integrating the virtual sensors with reference stations as cities often utilize different types of monitoring stations and their capabilities may even vary. Note that deploying the virtual sensors on fixed infrastructure does not generate data with high spatial resolution as the virtual sensors exclusively run on devices that are deployed at fixed locations. The benefit, however, is that the virtual sensor models are installed on or close to where the required input data is available. Hence, the correlations between input features will be stronger and the ground truth data can be easily accessed by the virtual sensors for computing and storing air pollutant concentrations.

**LCS embedded systems:** Advances in embedded systems and computing technologies have made it possible to embed complex models, including ML pipelines, directly on embedded devices. Installing virtual sensor models directly on LCS platforms increases the resolution of information but comes with the caveat that updating the models becomes challenging. Virtual sensors can also affect the operation of the LCS as the storage and computing requirements drain memory and energy from other operations. This means the duty cycles of the LCS devices need to be carefully optimized to ensure the virtual sensors do not significantly hamper the other operations of the device, e.g., data collection needs to be limited to the most relevant periods of time to avoid overusing the limited storage of the LCS platforms. Deploying the virtual sensors on the LCS nevertheless has the advantage of deploying them independently without a need for data communication, which is particularly well suited for urban areas that are not covered by extensive smart city infrastructure.

**Edge and cloud platforms:** The resource drain on the LCS platforms can be reduced by running the virtual sensor model on an edge or a cloud platform and generating virtual data on them [11]. Alternatively, the model can be installed on the embedded LCS device, and model updates can be performed on-the-fly while data is transmitted to edge or cloud platforms (i.e., federated learning). The main challenges with this approach are ensuring continuous communication links and guaranteeing data security. Nevertheless, edge deployments bring the computing platforms closer to the sensors where the data is generated. Reliable communication links also enable monitoring virtual sensor models continuously and calibrating and updating conveniently. Another benefit comes from the potential to support virtual sensors for variables that are not available on every LCS platform. Naturally, this option requires the availability of edge (or cloud) computing infrastructure and economically feasible models for harnessing the resources. In lower-density areas, the models can run directly on the LCS and be supplemented by a dedicated sensor network that monitors areas close to known pollution sources (e.g., industrial locations or transportation).

## Monitoring Sensor Performance

In practice, the operation of the virtual sensors would encounter several challenges as addressed in this subsection.

**Physical sensor degradation:** LCSs deployed in the field are subject to wear and tear, and can even malfunction due to changes in environmental factors during their service life. The degradation naturally affects the quality and reliability of the data they produce. In many cases the only way to overcome the issue is to perform regular maintenance, e.g., sensor air flows can become clogged and need to be cleaned or the battery of the device may lose a significant part of its capacity. Handling maintenance in practice is challenging for massive amounts of sensors, particularly if they are deployed widely in city infrastructure and carried around by citizens. This requires smart condition-based maintenance to automatize the estimation needed for maintenance and then optimize maintenance schedules in a resource-efficient way.

**Calibration and virtual sensor model drift:** The performance of calibration and virtual sensor models might also drift if environmental conditions change during the operational time of the models. The challenge is that virtual sensors which are developed using air pollutant data collected during one period

of time may not work properly once the environmental situation changes. For example, the traffic volume in a city district may change over time, impacting the virtual sensors' model accuracy. This requires re-training the models or developing adaptive models that can re-train themselves. Examples of adaptive models include transfer learning and federated learning — both of which require networking capabilities on the sensors. An alternative solution to mitigate model drifts is to use robust models, such as physics-based models that are resistant to variations in environmental factors.

**Physical, calibrator, and virtual sensor monitoring:** Reference instruments in cities are in fixed locations whereas LCSs are distributed unevenly and in large numbers within the city. The LCSs also are not guaranteed to be co-located to a reference instrument frequently. This makes continuous validation a challenging task. Instead of requiring the sensors to be co-located, an alternative solution is to employ *drift monitoring* [12] which enables LCSs hardware to self-monitor and detect model drifts in calibrators and virtual sensors. Drift monitoring enables cross-checking and identifying drift in the sensors, whereas concept drift detection can further be used to detect model degradation in the calibrators and the virtual sensors [13]. Sensors with drift can then be scheduled for re-calibration and taken close to a reference station or a transfer learning mechanism can be applied.

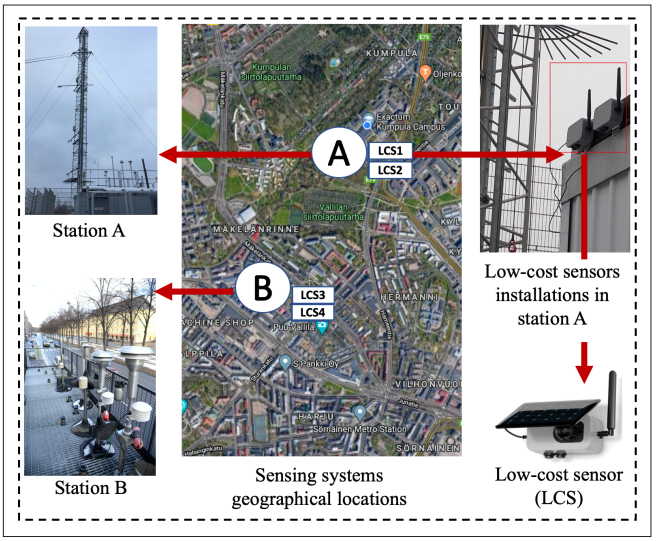


FIGURE 2. Reference stations (**A** and **B**) are located approximately one kilometer apart from each other at two different environmental profiles, enabling the development and deployment of generalized virtual sensors.

## Feasibility Study

We demonstrate the benefits of virtual sensors through a feasibility study that demonstrates how virtual sensors can be used to increase the spatiotemporal resolution of obtaining carbon dioxide ($CO_2$) and black carbon (BC). We focus on these two pollutants as they are significant health risks and as they are difficult to capture on LCS platforms. As inputs for the virtual sensors, we consider variables that are most commonly available on LCS platforms: $PM_{2.5}$, temperature, and relative humidity (RH).

### Overview of the Field Experiment

We collected air quality measurements from 13 March 2018 to 18 June 2019 using four LCS units by co-locating them at two different reference stations in Helsinki, Finland. The two locations had different urban characteristics, thus affecting also the pollutant distributions, and the sensors were located at different altitudes relative to the ground to ensure the locations were as different as possible. The locations are shown in Fig. 2. Two of the LCSs were installed on the top of the container at about 4 meters above ground level at an official urban air quality monitoring station (**A**). The LCS units we use cost less than $250 USD per unit [14]. The reference station is located in a sparsely populated urban area and is separated from the nearest busy road by approximately 150 m bands of deciduous forest and also surroundings by buildings, parking lots, and small vegetation. Hence, the pollution profile at the reference station is similar to the urban background. The other two LCSs were installed on a second reference station (**B**) at about 2 meters from ground level. The reference station was located in a street canyon within a busy street segment (approximately 28,100 vehicles per workday) [15], reflecting higher pollution concentrations. The reference instruments in these stations measure a large number of atmospheric variables including concentrations of aerosols, trace gases, solar radiation, and meteorological variables.

The locations where the LCSs are installed onto the reference stations are shown in Fig. 2. The LCSs used in our experiment are developed by Clarity Movements Company, based in Berkeley, CA, USA. These sensors are capable of measuring meteorological variables including temperature via band-gap technology and RH via capacitive technology. The sensors also measure particulate matter ($PM_{2.5}$) and $CO_2$ via laser light scattering technology and metal oxide semiconductor technology, respectively. The sensors have been calibrated by the manufacturer in both laboratory and field environments. While the frequency of measurements varies around 16–23 minutes per data point, the sensors that are equipped with an LTE-4G communication module send their measurements to a cloud platform provided by Clarity. The data is further stored in the cloud and downloadable anytime by request.

To develop the virtual sensor models, we followed the steps proposed earlier. We validated the LCSs by *consistency* and *accuracy* tests to ensure they function well when they are deployed in the field. In this step, we performed in-field calibration for the LCSs and ensured their readings follow the readings of the reference instruments. Then, we collected ground truth data from the reference instruments at sites (**A**) and (**B**). We used the variables temperature, RH, and $PM_{2.5}$ from the reference instruments and the inputs from the LCS measurements to develop the virtual sensor model.

For the model, we use nonlinear autoregressive with exogenous inputs network (NARX) which has been found to outperform other state-of-the-art models. The optimal NARX hyperparameters which consist of one hidden layer containing fifty neurons are obtained through grid search. Bayesian regularization backpropagation is also used for NARX parameters' estimation to ensure the model generalization and avoid over-fitting [4]. We further conducted *cross-unit validation* and *cross-site validation* for both the calibrated LCSs and virtual sensors model.

### Results

Figure 3 shows the box plots of the reference instruments (blue boxes) and virtual sensors (red boxes) for BC and $CO_2$ concentrations. While Fig. 3a illustrates the daily diurnal cycle of BC concentrations, Fig. 3b shows the $CO_2$ concentration in monthly aggregation. The blue dashed lines inside the plots represent the mean of BC and $CO_2$ concentrations obtained from the reference measurements. The red dashed lines represent the mean of BC and $CO_2$ concentrations obtained from the outputs of the virtual sensors. In all box-plots, the lines inside each box indicate the median. The bottom edges of the boxes show the 25th percentiles and the top edges show the 75th percentiles. The whiskers present the most extreme data points.

The results in both Fig. 3a and Fig. 3b. demonstrate that the developed virtual sensors generate data similar to the measurements of the reference instruments. This is illustrated by the dashed lines which show that the mean concentrations of BC and $CO_2$ concentrations follow similar patterns. Likewise, the median values and the box sizes (both blue and red boxes) follow similar patterns. These results demonstrate the high per-

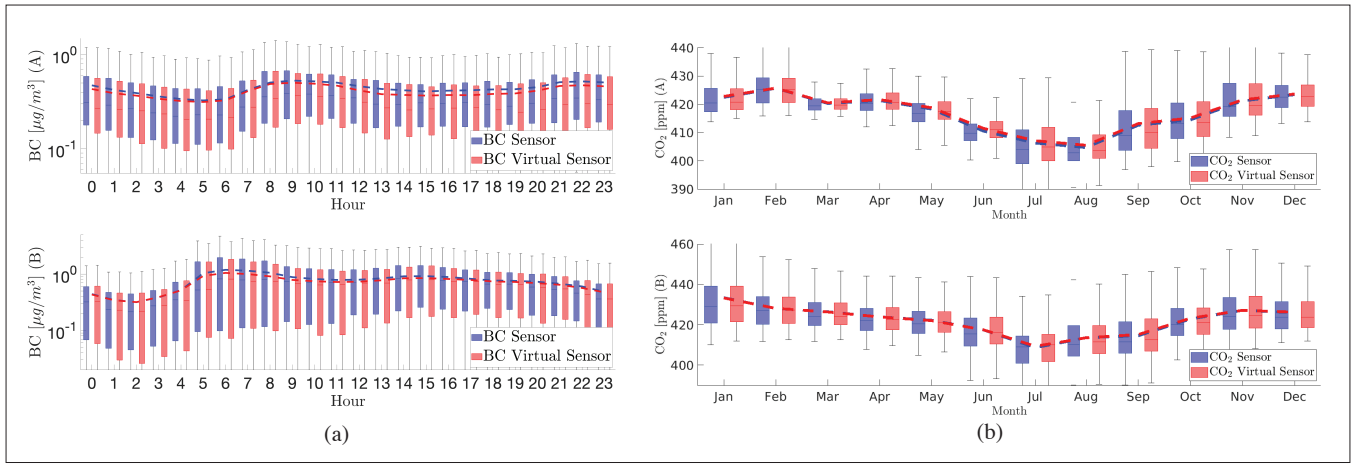


FIGURE 3. Box plot visualization of reference instrument measurements and virtual sensors: the similar length of box plots between sensor measurements (blue) and virtual sensors (red) indicate that the virtual sensors estimate the values close to the physical sensor measurements: a) diurnal cycles of BC concentrations at sites **A** (top) and **B** (bottom); b) $CO_2$ monthly concentrations at sites **A** (top) and **B** (bottom).

formance of the developed virtual sensors which function accurately to emulate the measurements of two different pollutants at two different measurement sites.

Massive deployment of BC and $CO_2$ virtual sensors is beneficial to generate high-resolution spatio-temporal maps which can complement sparse measurements of reference instruments and LCSs. The virtual sensor measurements lead to an improved understanding of pollutant concentrations and sources in urban areas. As shown in Fig. 3a, investigating diurnal cycles assist in understanding the pattern of BC concentrations around the measurement area. The BC concentration increases during rush hours (about 5–8 AM in the morning and about 3–5 PM in the afternoon). This result is expected as site **A** is located on a street that is one of the busiest roads in Helsinki. In addition, the increase in BC concentrations also takes place in the morning and afternoon, but with about 1 hour delay. The reason is that site **B** is located in an urban background (surrounded by buildings and plants) and BC is usually transported from nearby urban areas.

In addition, the results in Fig. 3b show the pattern of carbon concentration in the northern Hemisphere by illustrating more $CO_2$ in the winter than in the summer. The result implies a healthy environment in the region as the existing leaves on trees absorb more $CO_2$ in the summer than in the winter. These results demonstrate that the developed virtual sensors in this article perform well by virtually generating data points that follow the measurements of physical reference instruments. Indeed, deploying BC and $CO_2$ virtual sensors can offer dense measurements, reduce maintenance costs, and collect large amounts of data. Thus, analyzing the data enables an understanding of the environmental impacts of air pollution mitigation strategies.

## Discussion: The Impact of Virtual Sensors

As illustrated in the application block of Fig. 1, virtual sensors improve air quality information and enable the development of various applications as described below.

**Virtual Sensors and Deployments:** The experiments focused on demonstrating the benefits of virtual sensors and in practical deployments different benefits and costs need to be assessed. Naturally when data quality is the main consideration then high-cost precision instruments should be used but this results in limited spatiotemporal resolution of the data. The comparison between virtual sensors and low-cost sensors in turn is more complex. Virtual sensors can provide better accuracy than dedicated low-cost sensors as they can use multiple inputs to reduce data uncertainty, but they require sufficient quantities of sensors providing the necessary inputs. This may be more costly than deploying dedicated low-cost sensors for the target variable. For example, low-cost sensors for black carbon require regular maintenance as they rely on replaceable filters. As virtual sensors do not require maintenance, they are best suited for long-term monitoring to augment the available information whenever sufficient sensor deployments are already available.

**Comprehensive air quality database:** Virtual sensors are beneficial for supplementing existing air quality databases, which usually contain data sets obtained from air quality models, satellite remote sensing and official air pollution monitoring stations. As virtual sensors are integrated with LCSs, they can estimate additional air pollutants which are not measured via physical sensors to increase the resolution and the spatial and temporal coverage.

**New air quality index (AQI):** The current AQI is limited to a few pollutants, such as particulate matter, and carbon monoxide (CO). However, human exposure cannot be accurately assessed merely from a limited number of pollutants [16]. Thus, a new AQI requires the inclusion of additional variables such as BC, lung deposited surface area (LDSA), and particle number concentration (PNC). Virtual sensors help to estimate these pollutants for an improved AQI.

**Applications and Beneficiaries:** Virtual sensors enable the development of various applications. For example, they can be integrated with personalized health devices which are carried by individuals to report their exposure [17]. This integration also enables estimation of exposure to other pollutants such as BC and LDSA which are not measured by portable sensors, and allows monitoring of traffic pollution or showing green route maps for commuters. This can be beneficial particularly in residential areas that are exposed to pollutants yet do not have sufficient economic incentives to deploy professional-grade sensors. For example, our ongoing research explores virtual sensors for BC in a residential district where the residents commonly burn wood [4]. Virtual sensors can also be used as safety devices by estimating variables that LCSs are incapable of measuring. For example, in indoor environments, the measurements of PNC and poisonous gas concentrations enable the detection of fire and leakage of harmful gases (e.g. CO), respectively [18]. The potential beneficiaries of these applications include citizens, hospitals, NGOs, and other organizations, industries, and policymakers.

## Conclusions

Virtual sensors offer a potential way to increase the scale of air quality monitoring by harnessing correlations and dependencies between different pollutants and environmental variables to estimate the concentrations of pollutants that are otherwise difficult to capture. This allows harnessing other types of sensors

that are easier and more affordable to deploy, instead of requiring deployments of dedicated sensor platforms. We offered a research vision for the use of virtual sensors, reflecting back on current technology and state-of-the-art solutions for air quality monitoring, and identifying key research gaps. These gaps relate to algorithmic challenges for dealing with data imbalance and sparsity, generality of the models and the data that is available for training, and ways to increase resilience to having only intermittent access to reference instruments or networking capability. We also demonstrated the feasibility and benefits of virtual sensors through experiments that used three common sensors (particulate matter, temperature, and relative humidity) to estimate black carbon and carbon dioxide concentrations. The results show that virtual sensors provide data that closely aligns with reference instruments and contains expected seasonal and diurnal patterns. Overall, our work shows that virtual sensors are a highly promising solution for air quality monitoring, especially for pollutants that are costly or difficult to capture with a high spatial and temporal resolution. Our ongoing work is exploring the potential of virtual sensors further through deployments of low-cost sensors in the city of Helsinki. These efforts seek to provide a better understanding of how different environments, sensor types, and environmental factors affect the performance of virtual sensors.

## Acknowledgment

This work is supported by Helsinki Institute for Information Technology (HIIT) with grant number 75233229. This work is also supported by Nokia Center of Advance Research (NCAR) and by the Academy of Finland with grant numbers 335934, 345008, and 339614, and in part by EMME-CARE with grant number 4100199.

## References

[1] World Health Organization, "World Health Statistics 2019: Monitoring Health for the SDGs, Sustainable Development Goals," 2019.
[2] M. Sorek-Hamer, R. Chatfield, and Y. Liu, "Strategies for Using Satellite-Based Products in Modeling PM2.5 and Short-Term Pollution Episodes," *Environment International*, vol. 144, 2020, p. 106057.
[3] N. H. Motlagh *et al.*, "Toward Massive Scale Air Quality Monitoring," *IEEE Commun. Mag.*, vol. 58, no. 2, 2020, pp. 54–59.
[4] M. A. Zaidan *et al.*, "Intelligent Calibration and Virtual Sensing for Integrated Low-Cost Air Quality Sensors," *IEEE Sensors J.*, 2020, pp. 1–1.
[5] P. L. Kinney, "Interactions of Climate Change, Air Pollution, and Human Health," *Current Environmental Health Reports*, vol. 5, no. 1, 2018, pp. 179–86.
[6] L. Liu, S. M. Kuo, and M. Zhou, "Virtual Sensing Techniques and Their Applications," *2009 Int'l. Conf. Networking, Sensing and Control*, 2009, pp. 31–36.
[7] J. J. Caubel, T. E. Cados, and T. W. Kirchstetter, "A New Black Carbon Sensor for Dense Air Quality Monitoring Networks," *Sensors*, vol. 18, no. 3, 2018, p. 738.
[8] J. J. Caubel *et al.*, "A Distributed Network of 100 Black Carbon Sensors for 100 Days of Air Quality Monitoring in West Oakland, California," *Environmental Science & Technology*, vol. 53, no. 13, 2019, pp. 7564–73.
[9] F. Concas *et al.*, "Low-Cost Outdoor Air Quality Monitoring and Sensor Calibration: A Survey and Critical Analysis," *ACM Trans, Sensor Networks (TOSN)*, vol. 17, no. 2, 2021, pp. 1–44.
[10] J. Song, K. Han, and M. E. Stettler, "Deep-Maps: Machine-Learning-Based Mobile Air Pollution Sensing," *IEEE Internet of Things J.*, vol. 8, no. 9, 2020, pp. 7649–60.
[11] X. Su *et al.*, "Intelligent and Scalable Air Quality Monitoring with 5G Edge," *IEEE Internet Computing*, vol. 25, no. 2, 2021, pp. 35–44.
[12] M. A. Zaidan *et al.*, "Intelligent Air Pollution Sensors Calibration for Extreme Events and Drifts Monitoring," *IEEE Trans. Industrial Informatics*, vol. 19, no. 2, 2023, pp. 1366–79.
[13] J. Gama *et al.*, "A Survey on Concept Drift Adaptation," *ACM Computing Surveys (CSUR)*, vol. 46, no. 4, 2014, pp. 1–37.
[14] E. Lagerspetz *et al.*, "Megasense: Feasibility of Low-Cost Sensors for Pollution Hot-Spot Detection," *2019 IEEE 17th Int'l. Conf. Industrial Informatics (INDIN)*, vol. 1, 2019, pp. 1083–90.
[15] P. L. Fung *et al.*, "Evaluation of White-Box Versus Black-Box Machine Learning Models in Estimating Ambient Black Carbon Concentration," *J. Aerosol Science*, vol. 152, 2021, p. 105694.
[16] A. Monteiro *et al.*, "Towards an Improved Air Quality Index," *Air Quality*, Atmosphere & Health, vol. 10, no. 4, 2017, pp. 447–55.
[17] P. W. Oluwasanya *et al.*, "Portable Multi-Sensor Air Quality Monitoring Platform for Personal Exposure Studies," *IEEE Instrumentation & Measurement Mag.*, vol. 22, no. 5, 2019, pp. 36–44.
[18] A. Bozek *et al.*, "The Use of Combustible Gas Detection in Hazardous Locations: Additional Safety Precautions Around Flammable Gas or Vapors," *IEEE Industry Applications Mag.*, vol. 24, no. 3, 2018, pp. 64–74.

## Biographies

Martha Arbayani Zaidan (martha.zaidan@helsinki.fi) is a Senior Researcher at the Department of Computer Science and INAR, University of Helsinki, Finland. He completed his Ph.D. from the University of Sheffield, UK. His research interests include AI and environmental sciences.

Naser Hossein Motlagh (naser.motlagh@helsinki.fi) is a Researcher at the Department of Computer Science, University of Helsinki. He completed his D.Sc. at Aalto University, Finland in 2018. His research interests include the Internet of Things, wireless sensor networks, unmanned aerial vehicles, and autonomous underwater vehicles.

Brandon E. Boor (bboor@purdue.edu) is an Associate Professor at Purdue University, USA. He received his Ph.D. from The University of Texas, the USA in 2015. His research interests include low-cost air quality monitoring and human exposure assessment.

David Lu (david.lu@clarity.io) is a Co-Founder and the CEO of Clarity Movement Co. in California, USA. He received his B.Sc. degree from the University of California at Berkeley, USA. His research interests include air quality measurements and monitoring.

Petteri Nurmi (petteri.nurmi@helsinki.fi) is an Associate Professor at the University of Helsinki. He received his Ph.D. from the University of Helsinki in 2009. His research interests include distributed systems, pervasive data science, and sensing systems.

Tuukka Petäjä (tuukka.petaja@helsinki.fi) is a Professor at the University of Helsinki, Finland. He completed his Ph.D. in Physics at the University of Helsinki. His research interests include atmospheric aerosol particles and their role in climate change and air quality.

Aijun Ding (dingaj@nju.edu.cn) is a Professor at Nanjing University, China. He received his Ph.D. in Meteorology from Nanjing University. His research interests include air pollution-weather/climate interactions and Lagrangian dispersion modeling.

Markku Kulmala (markku.kulmala@helsinki.fi) is an Academy Professor at the University of Helsinki, Finland. He received his Ph.D. in from the University of Helsinki. His research interests include atmospheric aerosol nucleation and biosphere-aerosol-cloud-climate interactions.

Sasu Tarkoma (sasu.tarkoma@helsinki.fi) is a Professor at the University of Helsinki. He received his Ph.D. from the University of Helsinki in 2006. His research interests include mobile computing and AI.

Tareq Hussein (tareq.hussein@helsinki.fi) is a Professor at the University of Helsinki, Finland. He received his Ph.D. from the University of Helsinki. His research interests include urban and indoor air quality and exposure.