The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/0263-2772.htm

# Performing indoor $PM_{2.5}$ prediction with low-cost data and machine learning

Brent Lagesse
*Department of Computing and Software Systems, University of Washington Bothell, Bothell, Washington, USA, and*
Shuoqi Wang, Timothy V. Larson and Amy Ahim Kim
*Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington, USA*

Received 7 September 2021
Revised 13 December 2021
Accepted 5 February 2022

## Abstract

**Purpose** – The paper aims to develop a particle matter ($PM_{2.5}$) prediction model for open-plan office space using a variety of data sources. Monitoring of $PM_{2.5}$ levels is not widely applied in indoor settings. Many reliable methods of monitoring $PM_{2.5}$ require either time-consuming or expensive equipment, thus making $PM_{2.5}$ monitoring impractical for many settings. The goal of this paper is to identify possible low-cost, low-effort data sources that building managers can use in combination with machine learning (ML) models to approximate the performance of much more costly monitoring devices.

**Design/methodology/approach** – This study identified a variety of data sources, including freely available, public data, data from low-cost sensors and data from expensive, high-quality sensors. This study examined a variety of neural network architectures, including traditional artificial neural networks, generalized recurrent neural networks and long short-term memory neural networks as candidates for the prediction model. The authors trained the selected predictive model using this data and identified data sources that can be cheaply combined to approximate more expensive data sources.

**Findings** – The paper identified combinations of free data sources such as building damper percentages and weather data and low-cost sensors such as Wi-Fi-based occupancy estimator or a Plantower PMS7003 sensor that perform nearly as well as predictions made based on nephelometer data.

**Originality/value** – This work demonstrates that by combining low-cost sensors and ML, indoor $PM_{2.5}$ monitoring can be performed at a drastically reduced cost with minimal error compared to more traditional approaches.

**Keywords** Indoor air quality, Environmental sensing, Machine learning, Engineering economics, Neural networks, Air quality prediction

**Paper type** Research paper

## 1. Introduction

Fine particulate matter exposure, i.e. particle matter (PM) with an aerodynamic diameter less than 2.5 $\mu m$ or $PM_{2.5}$, has significant health implications (Burnett *et al.*, 2018; Cesaroni *et al.*, 2013; Cohen *et al.*, 2017; Li *et al.*, 2018; Pope *et al.*, 2019; Schraufnagel *et al.*, 2019). Monitoring ambient $PM_{2.5}$ has been well established and provides evidence for air quality regulations (EPA, 2011); however, indoor $PM_{2.5}$ monitoring and regulation is not widely applied in practice despite people in

This material is based upon work supported by the National Science Foundation under Grant Nos. 1852995 and 1853953.

Facilities
© Emerald Publishing Limited
0263-2772
DOI 10.1108/F-05-2021-0046

developed countries spending up to 90% of their time indoors (Klepeis *et al.*, 2001). Likewise, American workers spend approximately eight hours a day inside (Bureau of Labor Statistics, 2019).

Despite research that demonstrates the health significance of understanding indoor $PM_{2.5}$ concentrations, monitoring efforts have not followed. High-quality air quality monitoring (AQM) devices can be very expensive to deploy, operate and maintain. Likewise, these devices can only perform spot checks, so there is further added expense in either purchasing a significant number of them or identifying the optimal locations in each building. For example, the AQM device used for ground truth in this study costs approximately $6,500 and must be recalibrated yearly at the cost of $800.

*The goal of our work is to identify possible low-cost, low-effort options that building managers can use to approximate the performance of much more costly devices*, even when the data sources are not located in the same locations as the ground truth measurements. Accomplishing this will enable more building owners and building occupants, especially those with minimal resources, to have a better understanding of their indoor air quality and take measures to improve it.

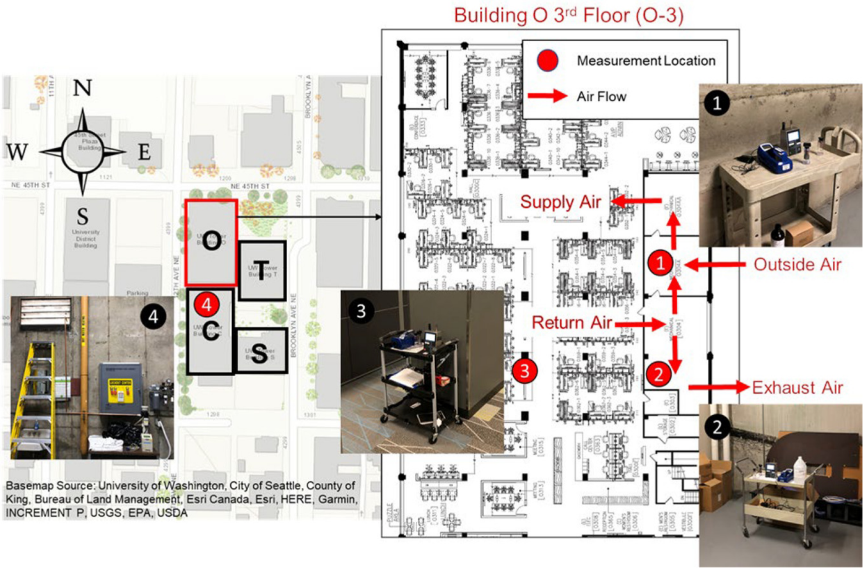
In this paper, we build predictive models based on a variety of sensors and freely available data to predict $PM_{2.5}$ mass concentration in an open-plan office space. As indoor spaces vary significantly from building to building, or even among different floors within the same building, the focus of this project is not to build the model with minimal error for our research environment, but rather to confirm that machine-learning augmented low-cost data sources can perform well in predicting indoor air quality and identifying into which principles should be explored in a wider variety of indoor environments.

We describe the following contributions in this paper that are intended to assist others in developing their own low-cost, high-quality predictive models for indoor $PM_{2.5}$:

- development of low-cost sensors in Sections 2.1.4 and 2.1.5;
- description of a variety of data sources ranging from freely available to requiring expensive hardware in Section 2.1;
- application of machine learning (ML) to augment data sources in section 3.

### *1.1 Research environment*

Research activities were mainly conducted in the University of Washington (UW) Tower complex located in the University District in Seattle, WA. The UW Tower commercial complex consists of four connected buildings, namely the O, C, T and S buildings as shown in Figure 1. The office space where indoor $PM_{2.5}$ measurements were carried out is on the third floor of O building (O-3), which is a typical open-plan office floor retrofitted in early 2018. Figure 2 gives a quick overview of the floor space. This space is occupied by the UW Information Technology group. The O-3 office space is served by a dedicated air handling unit located in the mechanical room on the floor. As shown in Figure 1, measurement Locations 1 and 2 were inside the mechanical room; Location 3 was within the office space and near the center of the floor; Location 4 was inside the penthouse on the roof of the C building where ambient air was sampled through an opening on the wall. At Location 1, the fresh outside air (not filtered) is brought in and mixed with part of the return air from the floor. The mixture is then treated with two layers of air filters, i.e. the AmAir 300X Extended Surface Pleated Panel Filter (American Air Filter Company Inc, 2018a) with a



Basemap Source: University of Washington, City of Seattle, County of King, Bureau of Land Management, Esri Canada, Esri, HERE, Garmin, INCREMENT P, USGS, EPA, USDA

**Figure 1.** UW tower complex layout and measurement locations

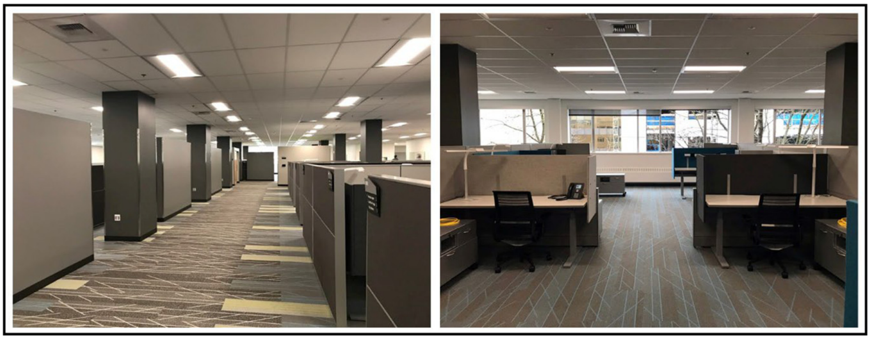


**Figure 2.** Snapshots of the O-3 office space

Minimum Efficiency Reporting Value (MERV) 8 rating and the VariCel Rigid Box Filter (American Air Filter Company Inc, 2018b) with MERV 14 rating. The filters are inspected daily and changed annually to maintain the required pressure differential. The filtered air is delivered to the floor through 48 air supply diffusers evenly spaced on the ceiling. The dampers installed on the outside air intake are controlled by the building management system (BCS) to regulate the ratio of outside air to return air based on the indoor temperature setpoint of the floor and the ambient air temperature. The return air from the O-3 floor first entered the ceiling plenum through 13 return air grilles evenly spaced on the ceiling. Part of the return air is filtered once by the variable air volume boxes installed in the plenum and then resupplied to the floor space without entering the mechanical room. The rest of the return air is delivered to the mechanical room, where the return air is split again. Some of the return air is reused and mixed with the fresh outside air (at Location 1) while the rest being exhausted. At measurement Location 2, the makeup of the room air is mainly exhaust air which is a well-mixed sample of the O-3 floor air. Therefore, the measured $PM_{2.5}$ level at Location 2 can be treated as the spatially averaged value for the entire floor.

### *1.2 Related work*

The research on enhancing indoor $PM_{2.5}$ monitoring has been conducted in two directions, i.e. creating low-cost sensors and developing prediction models.

As pointed out by Morawska *et al.* (2018) and Stamp *et al.* (2020), one important reason for the limited application of continuous indoor $PM_{2.5}$ monitoring, either in research or practice, was that the required instruments were traditionally expensive, labor-intensive and intrusive. Therefore, one direction is to develop reliable, accurate and easy-to-use low-cost $PM_{2.5}$ sensors. Several recent studies (Jovašević-Stojanović *et al.*, 2015; Lowther *et al.*, 2019; Zusman *et al.*, 2020) have reviewed the status of the existing low-cost sensors and evaluated their performance against reference instruments, and a majority of these sensors have shown consistent and accurate readings. However, most of the performance validations were conducted in the outdoor environment and reports on the sensor's indoor use in commercial buildings were limited.

As an alternative to on-site monitoring, indoor $PM_{2.5}$ prediction models have been studied by many researchers, including physics-based models and statistical models. Mechanistic models (Chen *et al.*, 2006; Goyal and Khare, 2011; Schneider *et al.*, 2004; Tran *et al.*, 2017; Hussein and Kulmala, 2008) and building simulations (Feng *et al.*, 2012; NIST, 2019) are both commonly used physics-based approaches that require detailed input information such as outdoor pollutant level, indoor emission sources, building envelope and ventilation system configuration to correctly model the target pollutant concentration. These models are often preferred during the design phase of new constructions for evaluating competing design strategies that affect indoor pollutants. However, the complex nature of the inputs makes it challenging to implement in natural indoor environments when human activities are involved, especially for large buildings.

In contrast, ML models and regression models are the most widely used statistical models for predicting indoor $PM_{2.5}$ in existing buildings, as recently reviewed by Wei *et al.* (2019). Inputs for both ML and regression models consist of indoor environmental and outdoor meteorological variables such as previous hour $PM_{2.5}$, indoor air temperature, indoor relative humidity (RH), indoor $CO_2$, outdoor $PM_{2.5}$, outdoor temperature and wind speed. In general, the ML models appeared to generate prediction models that have lower root mean square error (RMSE) values during training compared to the regression models considering that many of the inputs were highly correlated (Wei *et al.*, 2019). Among the reviewed studies using ML models, only one was conducted in commercial office buildings and a feed-forward neural network model was developed (Challoner *et al.*, 2015), whereas others were focused on residential housing, school buildings and subway stations (Wei *et al.*, 2019). Compared to the ML models for ambient $PM_{2.5}$ prediction (Joharestani *et al.*, 2019; Pak *et al.*, 2020; Xiao *et al.*, 2018; Sun and Sun, 2017; Chen *et al.*, 2020), the indoor models are not as diverse. Several recent studies have demonstrated the potential of long short-term memory (LSTM) neural network in predicting ambient $PM_{2.5}$ (Bai *et al.*, 2019; Kim *et al.*, 2019; Li *et al.*, 2017; Qi *et al.*, 2019) but its use in commercial office buildings has not been assessed. This paper aims to evaluate the performance of LSTM in predicting $PM_{2.5}$ in a commercial office space. In addition, the cost of input variables is considered to test the feasibility of a reduced model with low-cost data only.

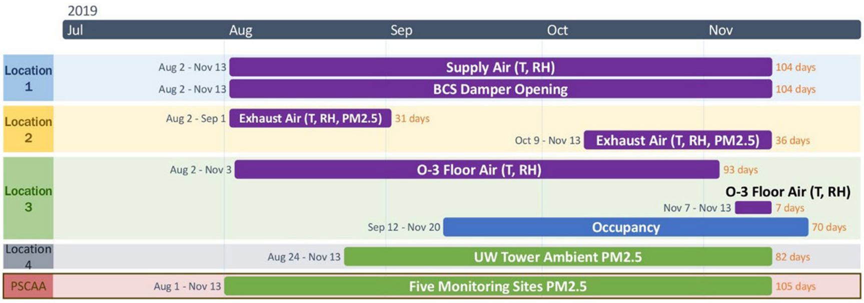
## 2. Materials and methods

Section 2 describes the collected data sets, the sensors used for data collection and the neural networks used to perform prediction.

### *2.1 Data sets used*

This study used data collected into seven different data sets described in this section. An overview of when the data was collected for each data set is depicted in Figure 3. The relative cost for obtaining each data set is summarized in Table 1. Details of each data set are provided in the following subsections. Except for the occupancy and damper data sets, all other five data sets are related to indoor or outdoor environmental conditions and are commonly used in $PM_{2.5}$ prediction models. The occupancy data set accounted for the impact of human activities on $PM_{2.5}$ concentration. In a typical office setting, human activities such as walking on the carpet could cause particle resuspension, which in turn contributes to the change in $PM_{2.5}$ concentration (Qian *et al.*, 2014; Tian *et al.*, 2014). In addition, particles could also detach from clothing when people engage in various types of physical activities (McDonagh and Byrne, 2014b; McDonagh and Byrne, 2014a). The damper data set, on the other hand, provided a direct measure of the amount of outdoor air intake. Inclusion of the occupancy and damper data sets, which has not been seen in many other studies, was expected to improve the prediction accuracy.

*2.1.1 Regional ambient $PM_{2.5}$ data.* There are five ambient $PM_{2.5}$ monitoring sites in the Puget Sound region surrounding the City of Seattle, as shown in Figure 4. These five sites are all within 16 km of the UW Tower building. The surrounding environment of each site is summarized in Table 2. Three of the sites, i.e. Lynnwood, Lake Forest Park (LFP) and Duwamish are owned and operated by the Puget Sound Clean Air Agency (PSCAA) (PSCAA, 2021), while the other two sites, i.e. 10th and Weller (TW) and Bellevue, are owned and operated by the Washington State Department of Ecology (WADOE) (WADOE, 2021).



**Note:** The location numberings are as shown in Figure 1

**Figure 3.** Data collection timeline

**Table 1.** Relative cost of data sources

| Measurement | Data Source | Cost |
|---|---|---|
| UW Tower Ambient $PM_{2.5}$ | Radiance Research M903 Nephelometer | High |
| Air temperature, RH, indoor $PM_{2.5}$ | Particles Plus 7302-AQM | High |
| Indoor $PM_{2.5}$ | Plantower PMS7003 | Low |
| Relative occupancy | Occupancy sensor | Low |
| Meteorological | UW Weather Station | Free |
| Regional Ambient $PM_{2.5}$ | Government agency monitoring sites | Free |
| Outside air intake damper opening | UW Tower BCS | Free |

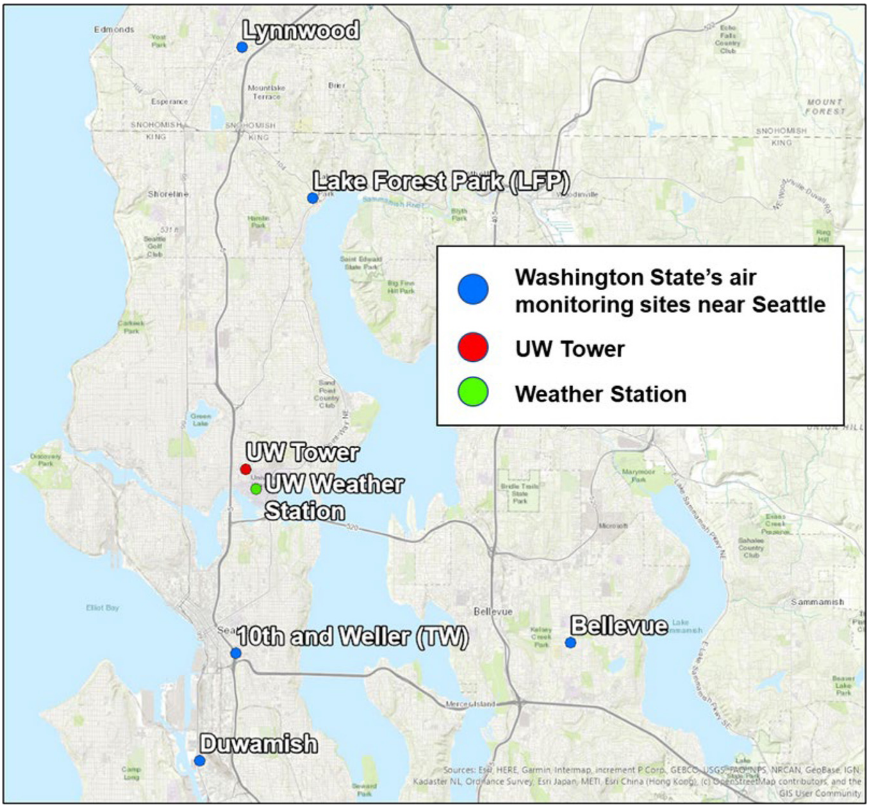

Lynnwood
Lake Forest Park (LFP)
UW Tower
UW Weather Station
10th and Weller (TW)
Bellevue
Duwamish
● Washington State's air monitoring sites near Seattle
● UW Tower
● Weather Station
Sources: Esri, HERE, Garmin, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), (c) OpenStreetMap contributors, and the GIS User Community

**Figure 4.**
Locations of ambient $PM_{2.5}$ monitoring sites and UW weather station

**Table 2.**
Descriptions of the surrounding environment of each ambient $PM_{2.5}$ monitoring site

| | Location | | | Activity | | |
|---|---|---|---|---|---|---|
| Station | Urban Center | Suburban | Rural | Commercial | Industrial | Residential |
| Lynnwood | | ✓ | | ✓ | | |
| LFP | | ✓ | | ✓ | | ✓ |
| TW | ✓ | | | | | |
| Duwamish | ✓ | | | | ✓ | |
| Bellevue | ✓ | | | ✓ | | |

PSCAA is one of the seven local clean air agencies established in the State of Washington. These agencies and WADOE regional offices manage the air quality of different areas of the state. PSCAA and WADOE are both the state partners of the United States Environmental Protection Agency (EPA) and work together on data quality assurance to ensure that data generated by the monitoring sites are comparable to those generated by the filter-based federal reference method defined by the EPA (PSCAA, 2020). At the TW site, where the $PM_{2.5}$ is measured using a Met One BAM 1020 beta attenuation monitor (Met One Instruments, 2021), only hourly averaged data is available. At the other sites, the Radiance Research M903 nephelometers (NOAA, 2021) are being used, which can provide $PM_{2.5}$

measurements in one-minute intervals. We obtained hourly averaged $\mathrm{PM}_{2.5}$ data from all five sites, which is publicly available online (PSCAA, 2021; WADOE, 2021). In addition, we obtained 1-minute $\mathrm{PM}_{2.5}$ measurements for the three sites (i.e. Lynnwood, LFP and Duwamish), which could be requested from PSCAA for free.

Cost – This data is freely available.

*2.1.2 University of Washington tower ambient $PM_{2.5}$ data.* A Radiance Research M903 nephelometer (NOAA, 2021) was set up at measurement Location 4, as shown in Figure 1 to provide the ambient $\mathrm{PM}_{2.5}$ measurements at the UW Tower site. This nephelometer is identical to the instrument used by PSCAA and WADOE for some of the monitoring sites. It measures the particle backscattering extinction coefficient of light (bscat) by aerosols. The bscat is not a direct measurement of $\mathrm{PM}_{2.5}$. However, PSCAA and WADOE implements the US EPA guidance for mathematically relating bscat to $\mathrm{PM}_{2.5}$ mass concentrations from a Federal Reference or Equivalent Method (FRM/FEM) $\mathrm{PM}_{2.5}$ instrument via site-specific relationships. Because no FRM/FEM instrument was available at the UW Tower, the coefficients of the correlation between the nephelometer bscat and FRM/FEM for the Lynnwood site was obtained from PSCAA and used to convert the bscat value at measurement Location 4 to $\mathrm{PM}_{2.5}$ mass concentration $M$ [see equation (1)]. The Lynnwood site is selected because its surrounding environment is similar to the UW Tower's.

$$M = 0.6 + 22.2 \times bscat \tag{1}$$

Cost – While prices are not typically available online, the estimated cost of this device is approximately $10,000.

*2.1.3 Indoor $PM_{2.5}$ data.* As discussed in Section 1.1, the $\mathrm{PM}_{2.5}$ level measured at Location 2 is treated as the ground truth for the entire O-3 floor, considering that it is a well-mixed return air sample. A Particles Plus 7302-AQM (AQM) (Particles Plus Inc, 2021) was used to measure the $\mathrm{PM}_{2.5}$ level at Location 2. The AQM uses long-life diode technology to detect particles in the range of 0.3 to 25 $\mu m$ and was calibrated by the manufacturer before deployment. The AQM was programmed to count the particles in a two-minute air sample at a flow rate of 2.83 liters per minute every five minutes, and the particles were divided into six bins based on diameters, i.e. 0.3–0.5 $\mu m$, 0.5–1 $\mu m$, 1–2.5 $\mu m$, 2.5–5 $\mu m$, 5–10 $\mu m$ and 10–25 $\mu m$. The $\mathrm{PM}_{2.5}$ particle counts were the sum of the first three bins and were converted into mass concentration using equation (2) (Chan and Noris, 2011):

$$M = \sum_i \frac{\pi}{6} \rho \bar{d}_i^3 N_i \tag{2}$$

The particle density $\rho$ was set to $2.2\,g/cm^3$ as reported by Hasheminassab *et al.* (2014). The equivalent particle diameter ($\mu m$) for each size bin $i$ is denoted by $\bar{d}_i$ and $N_i$ is the particle count. The calculation of $\bar{d}_i$ is given in equation (3) where $d_{i,a}$ and $d_{i,b}$ are the lower and upper diameters of each size bin:

$$\bar{d}_i = \left[\frac{d_{i,b}^4 - d_{i,a}^4}{4(d_{i,b} - d_{i,a})}\right]^{\frac{1}{3}} \tag{3}$$

In addition to particle counts, the AQM also provided air temperature and RH measurements at Location 2 through an add-on sensor. Two additional AQMs were placed at measurement Locations 1 and 3 to record air temperature and RH.

Cost – The AQM has an upfront purchase price of around $6,500 and requires annual calibration that costs $800 each time. The add-on temperature and RH sensor for the AQM could be purchased separately for $295. However, many other stand-alone low-cost temperature and RH sensors could be used, given the maturity of the market.

*2.1.4 Occupancy data.* The occupancy sensor runs on a Raspberry Pi 4 and estimates relative occupancy by counting the number of media access control (MAC) addresses that communicate during a given time frame. This information was acquired by using a USB Wi-Fi networking card and entering it into monitor mode. Typically, a network card will drop any network packets that are not intended for it, but in monitor mode, it will accept any packets and pass them to a packet handler. The software reads the sender MAC addresses of these packets that have a received signal strength indicator greater than $-70$ dB and saves them for five minutes. After each five-minute time step, the software reports the count to the server and throws away the MAC addresses. During these five minutes, the software hops through the list of available 2.4 GHz Wi-Fi frequency channels to ensure that it is sampling networking devices on each of the frequency bands used for Wi-Fi. This provided a relative estimate of the occupancy of the floor throughout the experimental time frame while still preserving the privacy of the occupants. The occupancy sensor is placed at Location 3 to ensure a good signal coverage of the entire floor.

Cost – The cost of this device was $60 for everything needed to run the occupancy sensor, including the Raspberry Pi; however, this cost could be reduced by running the occupancy sensor on existing or surplus computing hardware.

*2.1.5 Low-cost indoor $PM_{2.5}$ data.* As a low-cost alternative to lab-grade particle counters such as the AQM and the nephelometer, we also deployed a Plantower PMS7003 sensor (Plantower, 2015) that was attached to the Raspberry Pi occupancy sensor at Location 3. The Plantower PMS7003 is a laser particle sensor that can provide particle counts in three particle size ranges, i.e. 0.3–1 $\mu m$, 1–2.5 $\mu m$ and 2.5–10 $\mu m$. This sensor was queried every five minutes, and the minimum, maximum and mean values of $PM_1$, $PM_{2.5}$ and $PM_{10}$ for the five-minute interval were recorded.

Cost – The Plantower PMS7003 sensor costs approximately $30 and was attached to the Raspberry Pi that was used for the occupancy sensor. We have reduced this cost even further by building a custom embedded system that does not need the Raspberry Pi.

*2.1.6 Outside air intake damper opening data.* The outside air intake dampers at Location 1 are controlled by the UW Tower BCS. The fraction of damper opening, as well as the ambient air temperature, were logged in five-minute intervals by the BCS and the records were exported.

Cost – This data is freely available.

*2.1.7 Weather data.* Weather data, including wind speed, wind direction, solar radiation, ambient pressure and ambient RH, was acquired from a pre-existing weather station on the rooftop of the Atmospheric Sciences-Geophysics Building on the UW campus (University of Washington, 2021). The location is shown by the green dot in Figure 4.

### *2.2 Prediction*

For this work, we predicted the $PM_{2.5}$ levels of well-mixed indoor air in the mechanical room at Location 2, as shown in Figure 1. To do this, we used the data sets listed in Table 3 as input variables (or predictors) and trained a neural network to make predictions. The estimated outcome variable from the neural network, i.e. $PM_{2.5}$ mass concentration at Location 2, was then compared to the actual measurement from the AQM to evaluate model performance.

**Table 3.** Description of data sets used in experiments

| Data set *al*ias | Unit | Description |
|---|---|---|
| PSCAA_Hourly | $\mu g/m^3$ | Hourly ambient $PM_{2.5}$ measured at the Lynnwood site |
| | | Hourly ambient $PM_{2.5}$ measured at the Bellevue site |
| | | Hourly ambient $PM_{2.5}$ measured at the TW site |
| | | Hourly ambient $PM_{2.5}$ measured at the LFP site |
| | | Hourly ambient $PM_{2.5}$ measured at the Duwamish site |
| PSCAA | $\mu g/m^3$ | One-minute ambient $PM_{2.5}$ measured at the Lynnwood site |
| | | One-minute ambient $PM_{2.5}$ measured at the LFP site |
| | | One-minute ambient $PM_{2.5}$ measured at the Duwamish site |
| Nephelometer | $\mu g/m^3$ | UW Tower ambient $PM_{2.5}$ measured at location 4 |
| AQM | °C | Air temperature of the supply air (location 1) |
| | | Air temperature of the exhaust air (location 2) |
| | | Air temperature of the floor air (location 3) |
| | | Air temperature of the ambient air logged by the UW Tower BCS |
| | % | RH of the supply air (location 1) |
| | | RH of the exhaust air (location 2) |
| | | RH of the floor air (location 3) |
| Particle | $\mu g/m^3$ | Five-minute $PM_{2.5}$ from Plantower PMS7003 (location 3) |
| Occupancy | – | Relative occupancy of the floor |
| Damper | – | Air intake damper opening fraction logged by the UW Tower BCS |
| Weather | Degree | Wind direction recorded on the ATG rooftop |
| | m/s | Wind speed recorded on the ATG rooftop |
| | Watts/$m^2$ | Solar radiation recorded on the ATG rooftop |
| | mbar | Ambient pressure recorded on the ATG rooftop |
| | % | Ambient RH recorded on the ATG rooftop |

*2.2.1 Pre-processing.* The data that was collected came from a variety of sources with differing collection intervals and formats. First, all of the timestamps had to be normalized to the same format and time zone. During this process, any data with errors or blank cells were thrown out. Once the timestamps were normalized, we merged the data sets that were to be used for each experiment. Additionally, we examined both five-minute time bins and one-hour time bins. As the sampling intervals of each data set differed, we collected all data within a time bin and took the average over that time. After the time bins of data were collected, we examined all the data sets in the experiment and identified all of the data points where there was a data point from every data set for the same time (within a tolerance of 10 min) and merged those into a single data set. This final data set was used for training and testing of the models.

*2.2.2 Training.* We examined a variety of neural network architectures, including traditional artificial neural networks, generalized recurrent neural networks and LSTM neural networks. We also examined many traditional ML models, the best of which had a normalized RMSE of 0.11 which is nearly twice the median error of our neural network models, so those are omitted from this report. The neural networks were built using TensorFlow 2.2. We then ran preliminary tests to identify a subset of hyperparameters to explore by eliminating those that consistently produced poor results since grid search time grows rapidly due to the curse of dimensionality. We then performed a hyperparameter grid search based on the hyperparameters identified in Table 4 for the neural network and validated the results using RMSE. The best results were chosen although many of the models performed similarly. We note that an exhaustive grid search is not necessary to obtain a highly effective model. Our grid search took about 1 h to run on an Nvidia RTX 2080 GPU and only needs to be run once. The traditional ANN using a densely connected input layer (sized to the number of features), an intermediate densely connected layer of 500 nodes and an output layer of a single densely connected node performed best, so we focus on that particular architecture in the remainder of this paper. This search was performed for every combination of input data, so each model had different hyperparameters. The performance was similar for a wide variety of hyperparameters, but for reference, we have included the hyperparameters for the best overall model in Table 5.

We trained with every combination of data sets to produce the results in Section 3. The $PM_{2.5}$ concentration of the well-mixed return air measured by the AQM at Location 2 was used as the ground truth for training.

## 3. Results

The data sets collected and processed from Table 3 were used to train and evaluate the model. A model was built from each possible combination of data sets and trained to predict the well-mixed air at Location 2 in Figure 1. The results in this section are presented as the average values of 10-fold cross-validation. Data was split 80% training, 10% testing and 10% validation. The results presented in this section are normalized values that have been scaled because depending on the data sets that were merged for a particular experiment, the underlying statistics can vary significantly, so an absolute RMSE value can be misleading. As a result, when we use RMSE in the remainder of this paper, it refers to the normalized RMSE produced by the MinMaxScaler from sklearn [1] which scales the input values linearly in a range of 0 to 1. Note that the predicted value sometimes exceeds the maximum testing value, so this is when the figure goes beyond 1.0.

In these experiments, we summarize the effectiveness of the sensors and data sources described in Section 2.1 and demonstrate that low-cost sensors plus ML can often be as useful as expensive sensors in predicting the $PM_{2.5}$ concentration indoors. We have evaluated the results at the five-minute granularity and one-hour granularity.

### *3.1 Five-minute granularity results*

As many of our data sources had reported at a five-minute or higher period, we conducted analysis at a five-minute granularity in addition to the hourly granularity.

**Table 4.**
Hyperparameter search space

| Hyperparameter | Options |
|---|---|
| Activation | Linear, Relu |
| Loss | Mean Squared Error, Huber Loss, Mean Absolute Error |
| Optimizer | Stochastic Gradient Descent, Nadam, Adam, RMSProp |
| Dropout | 0.1, 0.3, 0.5 |
| Training Epochs | 10, 50, 80, 100, 500 |
| Neurons | 100, 500, 1000 |

**Table 5.**
Hyperparameter selection for neural network

| Hyperparameter | Value |
|---|---|
| Activation | Linear |
| Loss | Mean Squared Error |
| Optimizer | Stochastic Gradient Descent (*lr* = 0.01, momentum = 0.9) |
| Dropout | 0.3 |
| Training Epochs | 80 |
| Neurons | 500 |

Using the methodology described in Section 2.1, we were able to produce enough data to make predictions on approximately 7,000–9,000 data points for many of our combined data sets.

*3.1.1 Discussion.* Figures 5–8 demonstrate the predictions of several models trained on the five-minute data. Figure 5 is the model that has the best overall performance with an RMSE of 0.024. Figures 6 and 7 are the two best performing model that uses free data and low-cost sensors. Figure 8 is the best-performing model that uses only free data.

The use of the nephelometer improves performance in nearly every case in this experiment; however, the predictions that include data from the low-cost sensors, even the occupancy sensor, provide results that are close to the best performing model. When only the freely available data is used, the quality of the results degrades significantly.

### *3.2 One-hour granularity results*

We also examined the results at a one-hour granularity. This was accomplished by averaging the results over each hour when the results were reported at a more frequent than once per hour frequency. Using the methodology described in Section 2.2.1, we were able to produce enough data to make predictions on approximately 600–800 data points for many of our combined data sets. In general, the predictions were worse in the one-hour granularity than they were in the five-minute granularity. This appears to be the result of the averaging process losing information that was more useful to prediction in the five-minute granularity.

*3.2.1 Discussion.* Figures 9–11 demonstrate the predictions of several models trained on the one-hour data. Figure 9 is the model that has the best overall performance with an RMSE

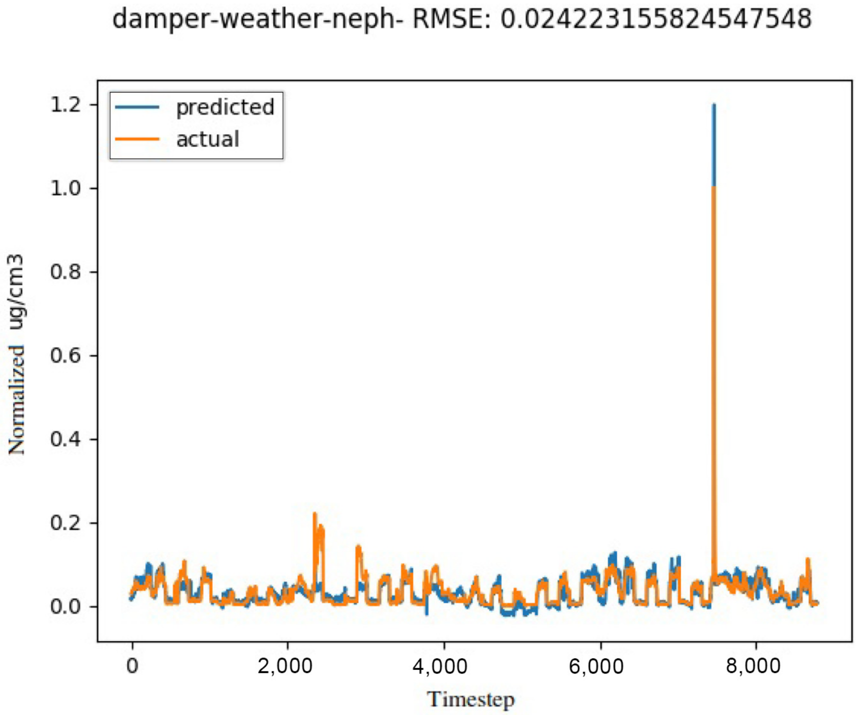


**Figure 5.**
Five-minutes: damper, weather and nephelometer data sets

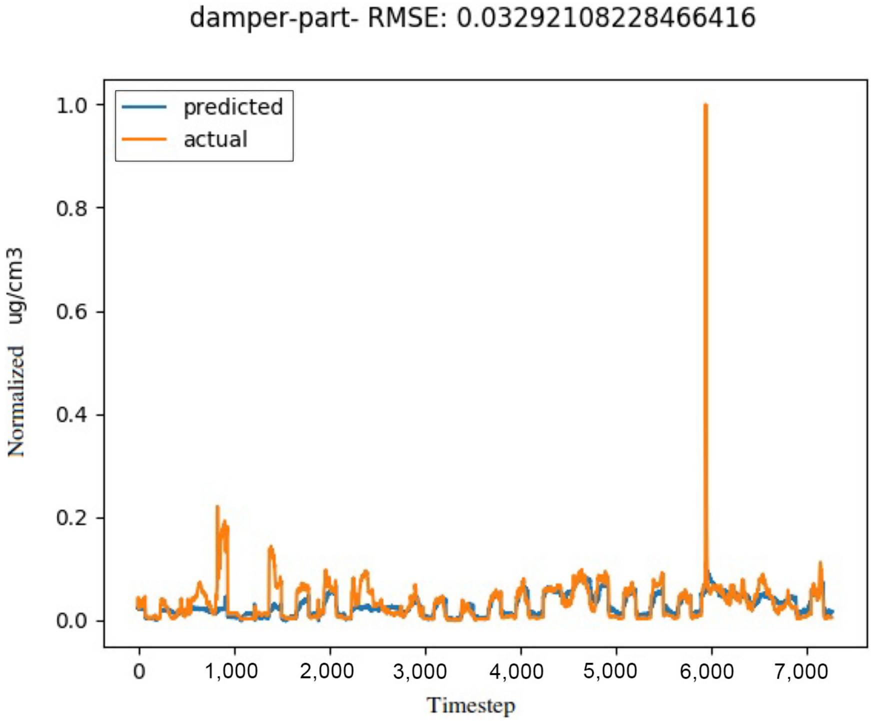


**Figure 6.**
Five-minutes: damper and particle data sets

of 0.053. Figure 10 is the best-performing model that uses free data and low-cost sensors. Figure 11 is the best-performing model that uses only free data.

In this section, we see similar trends in the one-hour experiments to the five-minute experiments. Once again, the best model does include the most expensive hardware, the best low-cost model performs well, but not as well and the best free data model has degraded results compared to the other two classifications of models.

### *3.3 Result comparisons*

As focusing on only the best model for our building would only have limited value to the research community, we also examined how models built with specific data sets performed. Table 6 presents these results. In this case, there is definitely a difference based on data set selection in performance in models that perform in the top 10, which will be discussed in the next subsection.

We also examined the performance of single data set models for both the five-minute and one-hour data sets. These results are presented in Table 7. This table shows that in both the five-minute and 1-hour case, adding the right data to a model can improve upon just using a single data source to train the model as combining the Damper and Particle data sets in Figure 6 nearly performs as well as the data set collected from the very expensive nephelometer. Overall, the Nephelometer, Occupancy Sensor and Particle Sensor performed well as individual data sets, but each was improved by adding additional data to the model.

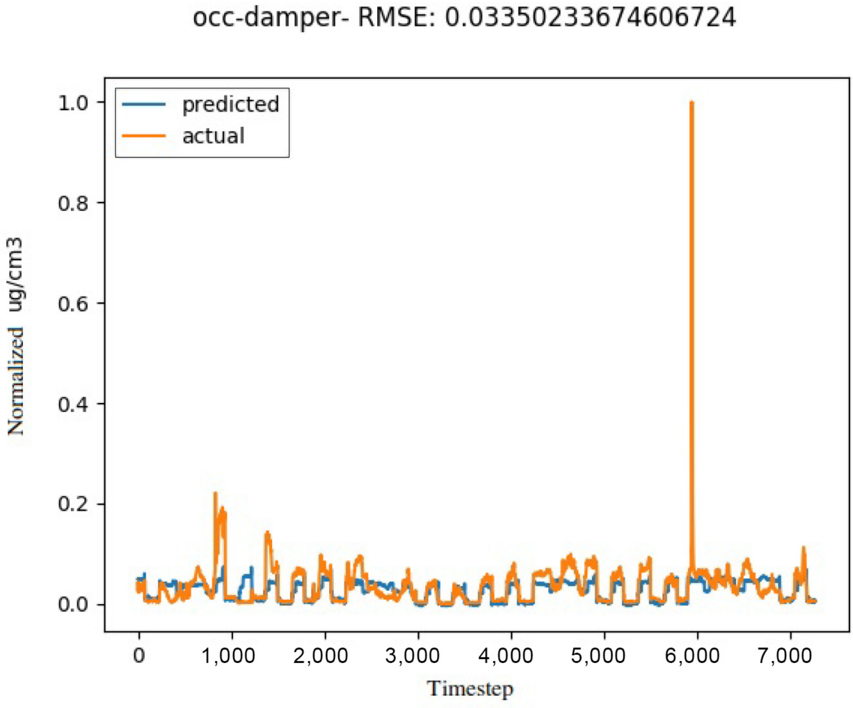


**Figure 7.**
Five-minutes: occupancy and damper data sets

*3.3.1 Discussion.* The key takeaways from this table are as follows:

- Sampling at five-minute intervals vs one-hour intervals makes a significant difference in which data sets were most useful. The PSCAA data was not very useful when sampling at five-minute intervals but was of typical usefulness for hourly sampling.
- At five-minute intervals, the nephelometer dominated the data set usage for top-performing models, but at one-hour intervals, the weather and damper settings dominated the top-performing models.
- In both cases, the more models we examined, the more the usage of each data set in training the models approached the expected value.
- The performance of a model based on a single data set correlates with its performance in combined data set models, but it does not mean that the data sets used in poorly performing models do not have value.

Expanding on Takeaway 1, the more that the data are averaged together (or if the median value for the timestep is taken), the more information is lost. This can be valuable when hardware produces anomalies or outliers since then the outlier is mitigated or removed from the data set, depending on which technique is used to determine the value for a particular timeset. Additionally, this loss of information means that the predictive power of models based on individual data sources is less, so inclusion in the top *N* models approaches the mean.

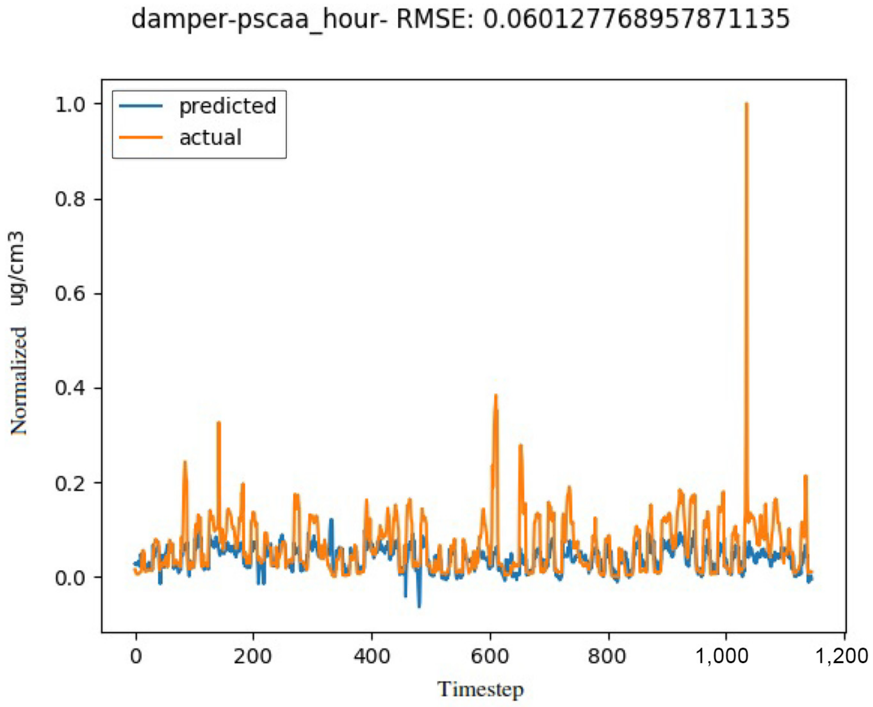


**Figure 8.**
Five-minutes: damper and PSCAA hourly data sets

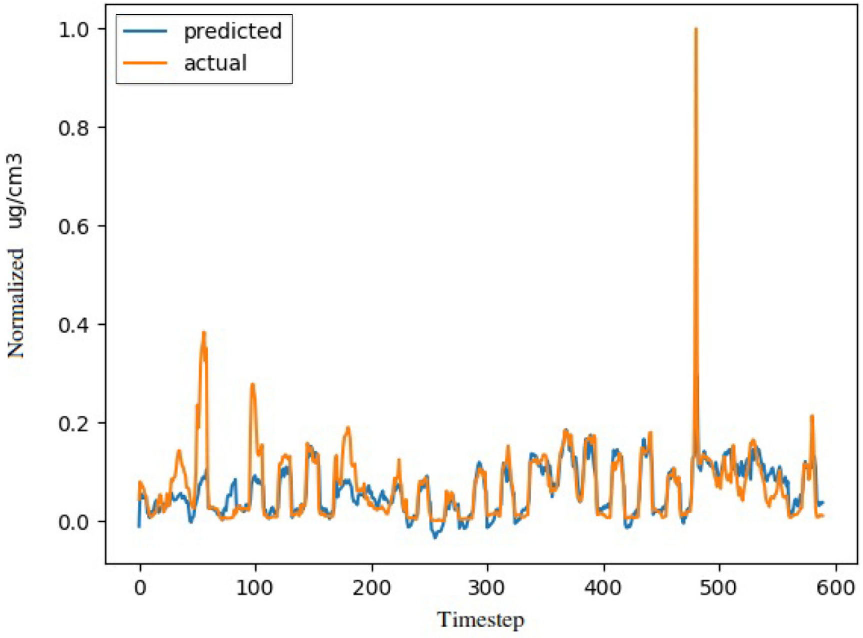


**Figure 9.**
One-hour: occupancy, damper, weather and nephelometer data sets

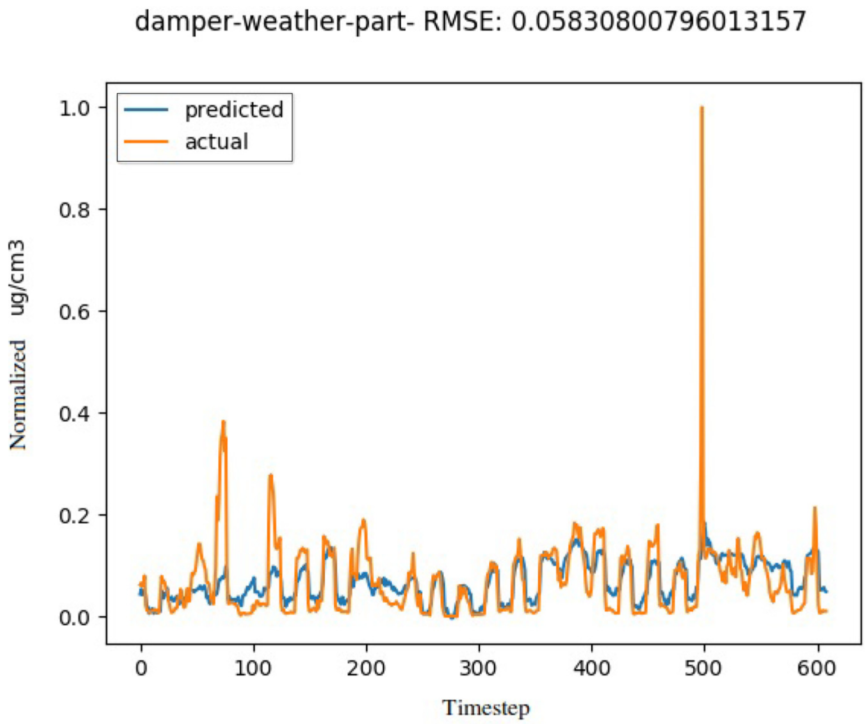


**Figure 10.**
One-Hour: damper, weather and particle data sets

Expanding on Takeaway 2, this is a topic that needs further experimentation and will be addressed in future work. At this point, we have identified starting points for developing best-practice predictive models for indoor air quality, but more work is needed to determine if this is specific to our experimental environment or if there is a greater trend that can be learned.

Expanding on Takeaway 3, there are definitely differences in how the data sets that are included affect the results in the best performing models for both cases. This means that the data sets that are selected are very important to this process and need to be combinatorially tested for a given environment.

Expanding on Takeaway 4, the combination of data sets for training a model typically improved the overall performance relative to just a single data set, as expected; however, the best models were not necessarily just the ones that used the top-performing single data sets. Just because a data set performs poorly in isolation does not mean that it may not be useful to a model in combination with other data sets. Even though the worst-performing single model using a single data set in isolation was the one-hour PSCAA Hourly data set, it was still part of the best overall performing free data model.

Although it was proved feasible to train a neural network model to predict indoor $PM_{2.5}$ levels with fairly good accuracy, the applicability of the developed models at other building sites is uncertain. The purpose of this paper is to serve as a proof of concept that by adopting ML algorithms, free and low-cost data from a few locations could be used to estimate the average $PM_{2.5}$ level in an office space. Given the current rapid development of low-cost sensors and increased public awareness on indoor air quality issues, future studies with large-scale sensor deployment covering multiple buildings in different climate conditions

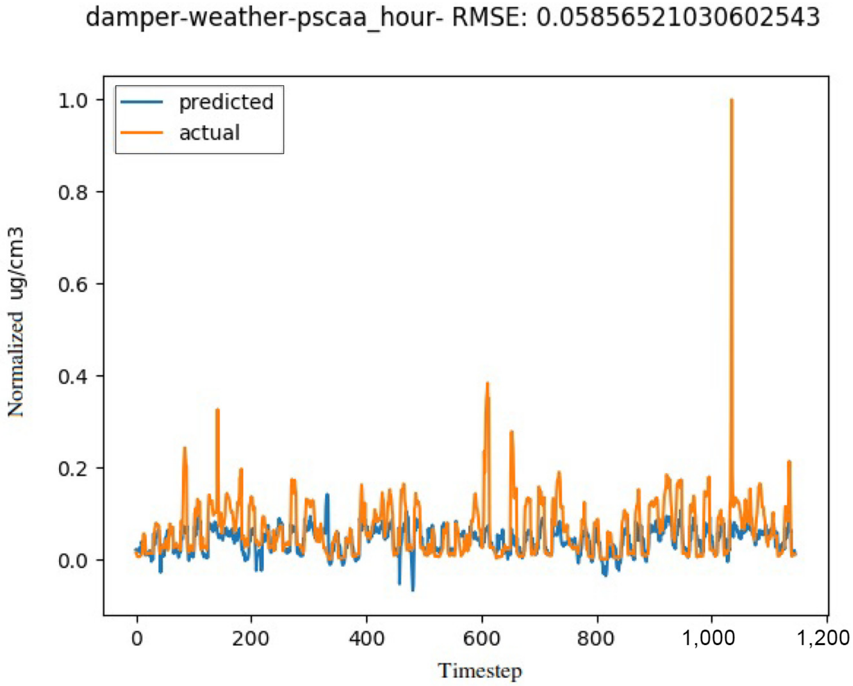


**Figure 11.**
One-Hour: damper, weather and PSCAA hourly data sets

**Table 6.**
Comparison of data sets based on relative performance rankings

| Sample | Top | Mean RSME | Occupancy | Damper | Weather | Nephelometer | Particle | PSCAA | PSCAA Hourly |
|---|---|---|---|---|---|---|---|---|---|
| Five-minute | 10 | 0.029 | 5 | 6 | 6 | 10 | 5 | 0 | 0 |
| | 25 | 0.033 | 13 | 13 | 11 | 15 | 15 | 1 | 1 |
| | 50 | 0.041 | 29 | 30 | 25 | 27 | 28 | 23 | 23 |
| One-hour | 10 | 0.058 | 5 | 9 | 10 | 6 | 5 | 5 | 5 |
| | 25 | 0.060 | 14 | 17 | 19 | 13 | 12 | 14 | 14 |
| | 50 | 0.065 | 29 | 28 | 28 | 26 | 28 | 26 | 26 |

**Table 7.**
Comparison of RMSE for individual data source models

| Data Source | Five-minute | One-hour |
|---|---|---|
| Nephelometer | 0.032 | 0.061 |
| Occupancy | 0.035 | 0.067 |
| Particle | 0.041 | 0.077 |
| Damper | 0.061 | 0.120 |
| PSCAA Hourly | 0.105 | 0.276 |
| Weather | 0.106 | 0.235 |
| PSCAA | 0.299 | 0.246 |

are possible. In that case, a more generalizable model could be developed. Nevertheless, this paper has shown that when the resource is limited and large-scale sensor deployment in an indoor office environment is not feasible, the limited amount of data could still be useful in predicting the average indoor $PM_{2.5}$ level.

## 4. Conclusion

In this study, we collected a variety of data from different sources ranging in cost from free to very expensive. From this data, we were able to train a neural network to make predictions about the $PM_{2.5}$ mass concentration in well-mixed air in our research environment. After analyzing the results, we were able to demonstrate that free and low-cost data combined with ML can effectively predict air quality even if they are not located in the sampling location. Furthermore, we identified trends in the performance of predictive models for our research environment that need to be explored as possible guiding principles for air quality prediction.

In the future, we intend to expand this study to examine more indoor spaces in a variety of locations and usages. Furthermore, we are working to identify additional data that would be useful to incorporate that is easy and cheap to acquire, including developing our own low-cost sensors to assist with the process. We also intend to see how more advanced ML models can be used with additional data collection.

## Note

[1] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

## References

American Air Filter Company Inc (2018a), "AmAir 300 – AAF international", available at: www.aafintl.com/en-gb/commercial/browse-products/commercial/panel-filters/amair-300-gt (accessed 1 December 2019).

American Air Filter Company Inc (2018b), "VariCel – AAF international", available at: www.aafintl.com/en/commercial/browse-products/commercial/box-filters/varicel (accessed 1 December 2019).

Bai, Y., Zeng, B., Li, C. and Zhang, J. (2019), "An ensemble long short-term memory neural network for hourly $PM_{2.5}$ concentration forecasting", *Chemosphere*, Vol. 222, pp. 286-294.

Bureau of Labor Statistics (2019), "Graphics for economic news releases", available at: www.bls.gov/charts/american-time-use/activity-by-sex.htm# (accessed 29 November 2019).

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C.A., Apte, J.S., Brauer, M., Cohen, A., Weichenthal, S., Coggins, J., Di, Q., Brunekreef, B., Frostad, J., Lim, S.S., Kan, H., Walker, K.D., Thurston, G.D., Hayes, R.B., Lim, C.C., Turner, M.C., Jerrett, M., Krewski, D., Gapstur, S.M., Diver, W.R., Ostro, B., Goldberg, D., Crouse, D.L., Martin, R.V., Peters, P., Pinault, L., Tjepkema, M., van Donkelaar, A., Villeneuve, P.J., Miller, A.B., Yin, P., Zhou, M., Wang, L., Janssen, N.A.H., Marra, M., Atkinson, R.W., Tsang, H., Quoc Thach, T., Cannon, J.B., Allen, R.T., Hart, J.E., Laden, F., Cesaroni, G., Forastiere, F., Weinmayr, G., Jaensch, A., Nagel, G., Concin, H. and Spadaro, J.V. (2018), "Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter", *Proceedings of the National Academy of Sciences*, Vol. 115 No. 38, pp. 9592-9597.

Cesaroni, G., Badaloni, C., Gariazzo, C., Stafoggia, M., Sozzi, R., Davoli, M. and Forastiere, F. (2013), "Long-Term exposure to urban air pollution and mortality in a cohort of more than a million adults in Rome", *Environmental Health Perspectives*, Vol. 121 No. 3, pp. 324-331.

Challoner, A., Pilla, F. and Gill, L. (2015), "Prediction of indoor air exposure from outdoor air quality using an artificial neural network model for inner city commercial buildings", *International Journal of Environmental Research and Public Health*, Vol. 12 No. 12, pp. 15233-15253.

Chan, W.R. and Noris, F. (2011), "Side-by-side comparison of particle count and mass concentration measurements in a residence", Lawrence Berkeley National Laboratory, Berkeley, CA, available at:

https://eta-publications.lbl.gov/sites/default/files/side-by-side-comparison.pdf (accessed 21 June 2021).

Chen, F., Yu, S.C.M. and Lai, A.C.K. (2006), "Modeling particle distribution and deposition in indoor environments with a new drift–flux model", *Atmospheric Environment*, Vol. 40 No. 2, pp. 357-367.

Chen, Y.C., Lei, T.C., Yao, S. and Wang, H.P. (2020), "PM2.5 prediction model based on combinational hammerstein recurrent neural networks", *Mathematics*, Vol. 8 No. 12, p. 2178.

Cohen, A.J., Brauer, M., Burnett, R. and erson, H.R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., Pope, C.A., III, Shin, H., Straif, K., Shaddick, G., Thomas, M., van Dingenen, R., van Donkelaar, A., Vos, T., Murray, C.J.L. and Forouzanfar, M.H. (2017), "Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015", *The Lancet*, Vol. 389 No. 10082, pp. 1907-1918.

EPA (2011), "Exposure factors handbook 2011 edition (final report)", U.S. Environmental Protection Agency, Washington, DC, available at: https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=236252 (accessed 29 September 2019).

Feng, W., Grunewald, J., Nicolai, A., Zhang, C. and Zhang, J.S. (2012), "CHAMPS-multi zone – a combined heat, air, moisture and pollutant simulation environment for whole-building performance analysis", *HVAC and R Research*, Vol. 18, pp. 233-251.

Goyal, R. and Khare, M. (2011), "Indoor air quality modeling for PM10, PM2.5 and PM1.0 in naturally ventilated classrooms of an urban Indian school building", *Environmental Monitoring and Assessment*, Vol. 176 Nos 1/4, pp. 501-516.

Hasheminassab, S., Pakbin, P., Delfino, R.J., Schauer, J.J. and Sioutas, C. (2014), "Diurnal and seasonal trends in the apparent density of ambient fine and coarse particles in los Angeles", *Environmental Pollution*, Vol. 187, pp. 1-9.

Hussein, T. and Kulmala, M. (2008), "Indoor aerosol modeling: basic principles and practical applications", *Water, Air and Soil Pollution: Focus*, Vol. 8 No. 1, pp. 23-34.

Joharestani, M.Z., Cao, C., Ni, X., Bashir, B. and Talebiesfandarani, S. (2019), "PM2.5 prediction based on random forest, XG boost and deep learning using multisource remote sensing data", *Atmosphere*, Vol. 10 No. 7, p. 373.

Jovašević-Stojanović, M., Bartonova, A., Topalović, D., Lazović, I., Pokrić, B. and Ristovski, Z. (2015), "On the use of small and cheaper sensors and devices for indicative citizen-based monitoring of respirable particulate matter", *Environmental Pollution*, Vol. 206, pp. 696-704.

Kim, H.S., Park, I., Song, C.H., Lee, K., Yun, J.W., Kim, H.K., Jeon, M., Lee, J. and Han, K.M. (2019), "Development of a daily $PM_{10}$ and $PM_{2.5}$ prediction system using a deep long short-term memory neural network model", *Atmospheric Chemistry and Physics*, Vol. 19 No. 20, pp. 12935-12951.

Klepeis, N.E., Nelson, W.C., Ott, W.R., Robinson, J.P., Tsang, A.M., Switzer, P., Behar, J.V., Hern, S.C. and Engelmann, W.H. (2001), "The national human activity pattern survey (NHAPS): a resource for assessing exposure to environmental pollutants", *Journal of Exposure Science and Environmental Epidemiology*, Vol. 11 No. 3, pp. 231-252.

Li, T., Zhang, Y., Wang, J., Xu, D., Yin, Z., Chen, H., Lv, Y., Luo, J., Zeng, Y., Liu, Y., Kinney, P.L. and Shi, X. (2018), "All-cause mortality risk associated with long-term exposure to ambient $PM_{2.5}$ in China: a cohort study", *The Lancet Public Health*, Vol. 3 No. 10, pp. e470-e477.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C. and Chi, T. (2017), "Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation", *Environmental Pollution*, Vol. 231 No. Part 1, pp. 997-1004.

Lowther, S.D., Jones, K.C., Wang, X., Whyatt, J.D., Wild, O. and Booker, D. (2019), "Particulate matter measurement indoors: a review of metrics, sensors, needs and applications", *Environmental Science and Technology*, Vol. 53 No. 20, pp. 11644-11656.

McDonagh, A. and Byrne, M.A. (2014a), "The influence of human physical activity and contaminated clothing type on particle resuspension", *Journal of Environmental Radioactivity*, Vol. 127, pp. 119-126.

McDonagh, A. and Byrne, M.A. (2014b), "A study of the size distribution of aerosol particles resuspended from clothing surfaces", *Journal of Aerosol Science*, Vol. 75, pp. 94-103.

Met One Instruments (2021), "Continuous particulate monitor BAM 1020", Met One Instruments, Grants Pass, OR, available at: https://metone.com/products/bam-1020/ (accessed June 19 2021).

Morawska, L., Thai, P.K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., Bedini, A., Chai, F., Christensen, B., Dunbabin, M., Gao, J., Hagler, G.S.W., Jayaratne, R., Kumar, P., Lau, A.K.H., Louie, P.K.K., Mazaheri, M., Ning, Z., Motta, N., Mullins, B., Rahman, M.M., Ristovski, Z., Shafiei, M., Tjondronegoro, D., Westerdahl, D. and Williams, R. (2018), "Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: how far have they gone?", *Environment International*, Vol. 116, pp. 286-299.

NIST (2019), "CONTAM", National Institute of Standards and Technology, Gaithersburg, MD, available at: www.nist.gov/services-resources/software/contam (accessed 2 August 2020).

NOAA (2021), "Radiance research nephelometer", available at: www.esrl.noaa.gov/gmd/aero/instrumentation/RR_neph.html (accessed 10 June 2021).

Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K. and Pak, C. (2020), "Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: a case study of Beijing, China", *Science of the Total Environment*, Vol. 699, p. 133561.

Particles Plus Inc (2021), "7301-AQM and 7302-AQM remote air quality and environmental monitor", Particles Plus, Stoughton, MA, available at: https://particlesplus.com/7301-iaq-remote-particle-counter/ (accessed 21 June 2021).

Plantower (2015), "PMS 7003-PM2.5-Plantower technology", Beijing Plantower Co., Beijing, available at: www.plantower.com/en/content/?110.html (accessed 22 June 2021).

Pope, C.A., III, Lefler, J.S., Ezzati, M., Higbee, J.D., Marshall, J.D., Kim, S.-Y., Bechle, M., Gilliat, K.S., Vernon, S.E., Robinson, A.L. and Burnett, R.T. (2019), "Mortality risk and fine particulate air pollution in a large", *Environmental Health Perspectives*, Vol. 127 No. 7, p. 77007.

PSCAA (2020), "2019 Air quality data summary", Puget Sound Clean Air Agency, Seattle, WA, available at: https://pscleanair.gov/DocumentCenter/View/4164/Air-Quality-Data-Summary-2019 (accessed 19 June 2021).

PSCAA (2021), "Network map", Puget Sound Clean Air Agency, Seattle, WA, available at: https://secure.pscleanair.org/AirQuality/NetworkMap (accessed 19 June 2021).

Qi, Y., Li, Q., Karimian, H. and Liu, D. (2019), "A hybrid model for spatiotemporal forecasting of $PM_{2.5}$ based on graph convolutional neural network and long short-term memory", *Science of the Total Environment*, Vol. 664, pp. 1-10.

Qian, J., Peccia, J. and Ferro, A.R. (2014), "Walking-induced particle resuspension in indoor environments", *Atmospheric Environment*, Vol. 89, pp. 464-481.

Schneider, T., Alstrup Jensen, K., Clausen, P.A., Afshari, A., Gunnarsen, L., Wåhlin, P., Glasius, M., Palmgren, F., Nielsen, O.J. and Fogh, C.L. (2004), "Prediction of indoor concentration of 0.5–4 $\mu$m particles of outdoor origin in an uninhabited apartment", *Atmospheric Environment*, Vol. 38 No. 37, pp. 6349-6359.

Schraufnagel, D.E., Balmes, J.R., Cowl, C.T., De Matteis, S., Jung, S.-H., Mortimer, K., Perez-Padilla, R., Rice, M.B., Riojas-Rodriguez, H., Sood, A., Thurston, G.D., To, T., Vanker, A. and Wuebbles, D.J. (2019), "Air pollution and non-communicable diseases: a review by the forum of international respiratory societies' environmental committee, part 1: the damaging effects of air pollution", *Chest*, Vol. 155 No. 2, pp. 409-416.

Stamp, S., Burman, E., Shrubsole, C., Chatzidiadou, L., Mumovic, D. and Davies, M. (2020), "Long-term, continuous air quality monitoring in a cross-sectional study of three UK non-domestic buildings", *Building and Environment*, Vol. 180, p. 107071.

Sun, W. and Sun, J. (2017), "Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm", *Journal of Environmental Management*, Vol. 188, pp. 144-152.

Tian, Y., Sul, K., Qian, J., Mondal, S. and Ferro, A.R. (2014), "A comparative study of walking-induced dust resuspension using a consistent test mechanism", *Indoor Air*, Vol. 24 No. 6, pp. 592-603.

Tran, D.T., Alleman, L.Y., Coddeville, P. and Galloo, J.-C. (2017), "Indoor particle dynamics in schools: determination of air exchange rate, size-resolved particle deposition rate and penetration factor in real-life conditions", *Indoor and Built Environment*, Vol. 26 No. 10, pp. 1335-1350.

University of Washington (2021), "Rooftop Observations – ATG building UW", University of Washington, Seattle, WA, available at: https://a.atmos.washington.edu/cgi-bin/list_uw.cgi (accessed 22 June 2021).

WADOE (2021), "Washington's air monitoring network", Washington State Department of Ecology, Olympia, WA, available at: https://fortress.wa.gov/ecy/enviwa/ (accessed 19 June 2021).

Wei, W., Ramalho, O., Malingre, L., Sivanantham, S., Little, J.C. and Mandin, C. (2019), "Machine learning and statistical models for predicting indoor air quality", *Indoor Air*, Vol. 29 No. 5, pp. 704-726.

Xiao, Q., Chang, H.H., Geng, G. and Liu, Y. (2018), "An ensemble Machine-Learning model to predict historical PM2.5 concentrations in China from satellite data", *Environmental Science and Technology*, Vol. 52 No. 22, pp. 13260-13269.

Zusman, M., Schumacher, C.S., Gassett, A.J., Spalt, E.W., Austin, E., Larson, T.V., Carvlin, G., Seto, E., Kaufman, J.D. and Sheppard, L. (2020), "Calibration of low-cost particulate matter sensors: model development for a multi-city epidemiological study", *Environment International*, Vol. 134, pp. 105329.

**Corresponding author**

Brent Lagesse can be contacted at: lagesse@uw.edu