Predicting $PM_{2.5}$ in well-mixed indoor air for a large office building using regression and artificial neural network models

Brent Lagesse[†], *Shuoqi Wang*[‡], *Timothy V. Larson*[‡], *and Amy A. Kim*[*,‡]

[†]Division of Computing and Software Systems, University of Washington Bothell, Bothell, Washington 98011, United States

[‡]Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington 98195, United States

ABSTRACT

Although the exposure to $PM_{2.5}$ has serious health implications, indoor $PM_{2.5}$ monitoring is not a widely applied practice. Regulations on indoor $PM_{2.5}$ level and measurement schemes are not well-established. Compared to other indoor settings, $PM_{2.5}$ prediction models for large office buildings are particularly lacking. In response to these challenges, statistical models were developed in this paper to predict the $PM_{2.5}$ concentration in well-mixed indoor air in a commercial office building. The performance of different modeling methods, including multiple linear regression (MLR), partial least squares regression (PLS), distributed lag model (DLM), least absolute shrinkage selector operator (LASSO), simple artificial neural networks (ANN), and long-

short term memory (LSTM), were compared. Various combinations of environmental and meteorological parameters were used as predictors. The root mean square error (RMSE) of the predicted hourly $PM_{2.5}$ was 1.73 μg/m^3 for the LSTM model and in the range of 2.20~4.71 μg/m^3 for the other models when regulatory ambient $PM_{2.5}$ data were used as predictors. The LSTM models outperformed other modeling approaches across the used performance metrics by learning the predictors' temporal patterns. Even without any ambient $PM_{2.5}$ information, the developed models still demonstrated relatively high skill in predicting the $PM_{2.5}$ levels in well-mixed indoor air.

1. INTRODUCTION

The exposure to ambient fine particulate matter (PM, with an aerodynamic diameter smaller than 2.5 μm), or $PM_{2.5}$, and its health implication has been studied extensively[1-6]. The monitoring network of ambient $PM_{2.5}$ is now well established in the United States, which provides essential evidence for ambient air quality regulations. However, indoor $PM_{2.5}$ monitoring is not a widely applied practice, and regulation of indoor $PM_{2.5}$ is lacking. Previous studies have shown that people in developed countries spent up to 90% of their time indoors[7] and the American worker spends eight hours a day on average at the workplace[8]. While long term measurement is available for ambient $PM_{2.5}$ through various agencies[9], indoor $PM_{2.5}$ data are usually scarce.

To enable the indoor air quality (IAQ) assessment where direct measurement is not feasible, researchers are seeking to develop prediction models using other environmental variables that are readily available. Most recently, Wei et al.[10] conducted a review of studies using machine learning and statistical models for predicting the IAQ in various types of buildings and found that artificial neural networks (ANN) and regression were the most popular techniques. The results also showed that just five out of the 37 reviewed studies were carried out in offices, and the models of these five studies were all developed using different types of feed-forward ANNs[11-15]. Only one of the five office studies focused on predicting indoor $PM_{2.5}$ using ambient $PM_{2.5}$ measurements[11]. In that study, Challoner et al.[11] predicted indoor $PM_{2.5}$ using ambient $PM_{2.5}$ concentrations and meteorological data in a mechanically ventilated office building with ANN and reported large errors ranging from -8.09 to 4.93 μg/m^3. However, the ambient $PM_{2.5}$ concentration was calculated using a personal-exposure activity location model instead of measured directly at the building site, which might cause the large errors in the predictions[11].

Other regression models, including multiple linear regression (MLR), stepwise regression, partial least squares regression (PLS), and principal component regression (PCR), have been applied in dwellings, schools, and subway stations but not in offices[10]. It is unclear whether these regression models could be used for predicting the $PM_{2.5}$ in offices and how well they perform compared to the ANN model.

A commercial office often consists of various types of regularly occupied spaces, e.g., open workstations, conference rooms, and common areas. There is no definitive method in the placement of monitors to assess the overall $PM_{2.5}$ level in the entire space. Predicting the spatial variation of $PM_{2.5}$ in the large area is also potentially complex. Therefore, developing a prediction model for spatial-averaged $PM_{2.5}$ could be the first step. The exhaust air is a well-mixed sample of the return air from different indoor locations and could serve as a representation of the spatially averaged condition.

The objective of this paper is to develop statistical models to predict the $PM_{2.5}$ concentration in well-mixed indoor air inside a commercial office building using MLR, PLS, a simple ANN, and a specific type of ANN known as a Long Short-Term Memory (LSTM) neural network. Several recent studies have explored the use of the LSTM neural network in predicting ambient $PM_{2.5}$ concentration[16-20], given that it is better suited for long time-series predictions than simple ANN models. Yet, its use in an indoor office setting has not been investigated as a comparison to a simple ANN model. Regression models that are capable of handling time-series data and evaluating delayed effects, i.e., the distributed lag model (DLM), the least absolute shrinkage selector operator (LASSO), as well as PLS with lagged predictors (PLS-Lag), are also considered. The dependent variable in this paper was limited to indoor $PM_{2.5}$, while other pollutants, such as chemicals emitted by the occupants, were not included. The independent variables considered included meteorological variables (e.g., wind speed, wind direction, air temperature, and relative humidity), publicly available ambient $PM_{2.5}$ concentration from other locations, number of occupants at the study site, and building operational data. Various combinations of the independent variables were tested to evaluate the prediction accuracy with or without ambient $PM_{2.5}$ concentration (whether from publicly available monitoring sites or measured directly at the study site). The performance of the various models was compared using several performance indicators, including the normalized absolute error (NAE), the root mean square error (RMSE), the coefficient of determination (R^2), and the index of agreement (IA).

2. EXPERIMENTAL SECTION

2.1. Prediction Variables

The variables used to predict hourly indoor $PM_{2.5}$ (PM_E) included hourly outdoor $PM_{2.5}$, relative humidity (RH), air temperature (T), and wind speed. We also included building air intake damper opening fraction on an hourly basis as a measure of outdoor air intake. Filtration of the outdoor air is discussed in detail in Section 1.1 of the Supporting Information (SI). Occupancy level on an hourly basis was also introduced as a predictor to account for the impact of indoor human activity.

The damper opening and occupancy on an hourly basis are rarely used in other studies due to the lack of such information.

A list of all the predictor variables and relevant descriptions is given in Table 1.

Table 1. Descriptions of the prediction variables.

No.	Variable	Unit	Description
1	PM_{LYN}	μg/m^3	Ambient $PM_{2.5}$ measured at the PSCAA Lynnwood site.
2	PM_B	μg/m^3	Ambient $PM_{2.5}$ measured at the PSCAA Bellevue site.
3	PM_W	μg/m^3	Ambient $PM_{2.5}$ measured at the PSCAA TW site.
4	PM_{LFP}	μg/m^3	Ambient $PM_{2.5}$ measured at the PSCAA LFP site.
5	PM_D	μg/m^3	Ambient $PM_{2.5}$ measured at the PSCAA Duwamish site.
6	PM_O	μg/m^3	Ambient $PM_{2.5}$ measured at location 4.
7	T_S	°C	Air temperature of the supply air (location 1).
8	T_E	°C	Air temperature of the exhaust air (location 2).
9	T_F	°C	Air temperature of the floor air (location 3).
10	T_O	°C	Air temperature of the ambient air logged by the BCS.
11	RH_S	%	Relative humidity of the supply air (location 1).
12	RH_E	%	Relative humidity of the exhaust air (location 2).
13	RH_F	%	Relative humidity of the floor air (location 3).
14	O_F	-	Relative occupancy of the floor (see SI).
15	D	-	Air intake damper opening fraction logged by the BCS.
16	WD	°	Wind direction recorded on the ATG rooftop.
17	WS	m/s	Wind speed recorded on the ATG rooftop.

2.2. Ambient $PM_{2.5}$ Measurements

The hourly average ambient $PM_{2.5}$ concentrations were collected from five monitoring sites managed by the Puget Sound Clean Air Agency (PSCAA)[21] as part of Washington's air monitoring network[22]. The selected five sites were all within 16 km of the UW Tower building. A general description of the environment near each site is given in Table S1. The wind speed and wind direction records were obtained from a weather station on the rooftop of the Atmospheric Sciences-Geophysics (ATG) Building[23] on the UW campus, approximately 0.9 km from the UW Tower. Figure S1 shows the location of UW Tower in relation to the PSCAA sites and the weather station.

2.3. Indoor Measurements

The indoor measurements were recorded every five minutes on floor O-3 in the University of Washington (UW) Tower building in Seattle. The building schematic and floor plan of the selected O-3 office space is shown in Figure S2. A detailed description of the ventilation system operation can also be found in Section 1.1 of the SI. The building control system (BCS) manages the building ventilation and logged the air intake damper opening fraction and ambient air temperature every five minutes. Figure 1 shows the configuration of the air handling unit and variable air volume boxes on floor O-3. Since the supply air is a mixture of outside air and return air, the air intake damper opening is a better indicator of the amount of outdoor air brought into the space than ventilation rate. The RH and T were measured at three locations on floor O-3, as shown in Figure S2, using three units of Particles Plus 7302-AQM air quality monitors (AQM)[24]. The $PM_{2.5}$ concentration in the exhaust air was measured at location 2. The exhaust air was considered a well-mixed air sample representing the spatial average of the indoor air on the O-3 floor. A Radiance M903 nephelometer[25] was used at location 4 outdoors to record the concentration of ambient $PM_{2.5}$ adjacent to the building. The particle mass concentration calibration process is detailed in Section 1.2 of the SI.

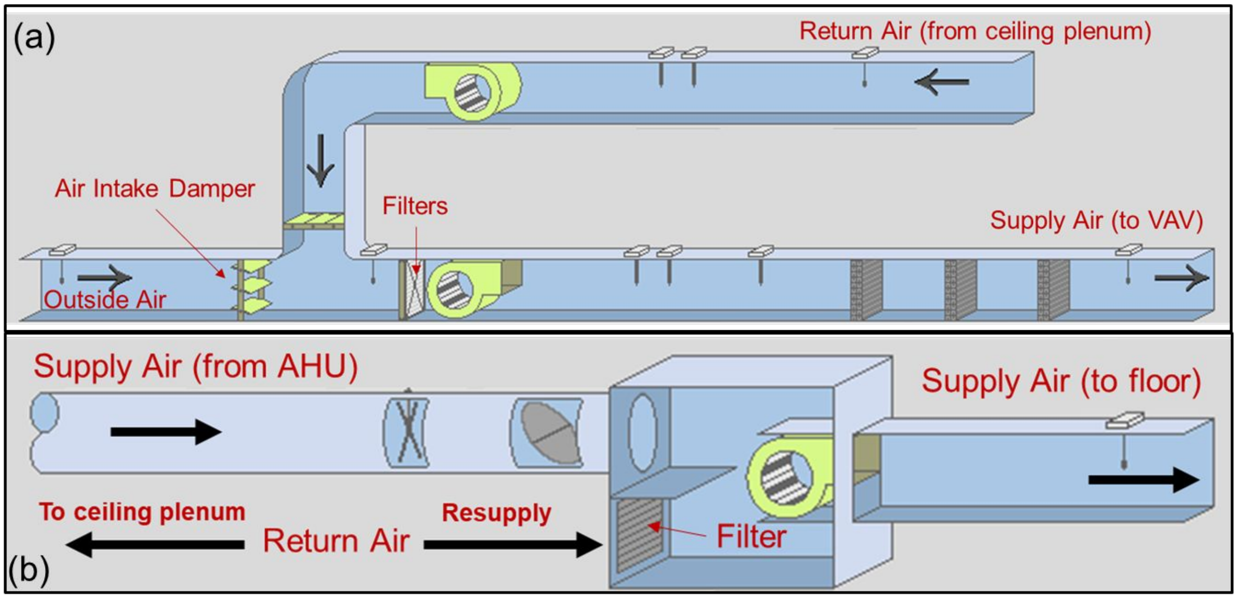

Figure 1. Configuration of the ventilation system on floor O-3. (a) air handling unit (AHU) located in the mechanical room; (b) one of the variable air volume (VAV) boxes located in the ceiling plenum space on the floor.

In addition to the AQM, an occupancy sensor was installed at location 3 in the open office space on floor O-3 (see Figure S2). It estimated relative occupancy by counting the number of Media Access Control (MAC) addresses that communicate during the five-minute sampling period. The working theory of the sensor is explained in Section 1.3 of the SI.

2.4. Data Processing

The measurements were conducted from August 2nd through November 13th in 2019. Due to various technical issues (e.g., repair of power supply, failure to start logging), missing values and measurement gaps existed for some of the variables. The duration of the measurements for each variable is shown in Figure S3. Only a subset of the observations was used to construct the analysis dataset. The collected data were transformed into 1-hour averages and merged to create a time series matrix. The PM_E data were also checked for outliers. The first quantile ($Q1$), third quantile

($Q3$), and the interquartile range (IQR) were calculated, and the data points above the upper bound (defined as $Q3 + 1.5IQR$) or below the lower bound (defined as $Q1 - 1.5IQR$) were labeled as outliers. In total, 21 outliers were identified for PM_E out of 1,535 observations (1.4%) and removed. After removing missing observations from all the other variables, the remaining matrix with complete data contained 670 observations (hourly averaged values) for a total of 17 variables.

3. MODELS

3.1. Multiple Linear Regression (MLR)

MLR models are the most commonly used for IAQ predictions as summarized in Wei et al.[10] The MLR approach enables the use of various independent variables (e.g., ambient $PM_{2.5}$, meteorological, occupancy) to predict the outcome (i.e., the $PM_{2.5}$ in well-mixed indoor air). The coefficient of each independent variable reflects the effect of the variable in predicting the outcome. All 17 independent variables were included at the beginning and a stepwise selection was conducted to find the final model that has the lowest Akaike Information Criterion (AIC) value. Multicollinearity is a known issue with MLR models when high correlations exist between independent variables[26, 27]. Therefore, the variance inflation factors (VIF) of the independent variables in the final model were examined. Variables with large VIF values were excluded to ensure low correlations among all the predictors in the final model [27].

3.2. Partial Least Squares (PLS)

PLS is a regression method with dimension reduction capability. For building a prediction model with many potentially correlated predictors, applying the PLS technique allows the user to transform the predictors into a reduced set of orthogonal latent variables (or components), which are a linear combination of the original predictors[28]. Compared to MLR, the PLS method has shown its capability in building robust $PM_{2.5}$ prediction models by coping with the multicollinearity issue present with a large number of predictors[29-32].

3.3. Artificial Neural Network (ANN)

An ANN is a collection of algorithms used to learn patterns from data and then use those patterns for predicting or classifying new data that has not previously been seen. This paper utilized a simple ANN (additional background information is included in Section 1.4 of the SI), and a type of recurrent neural network (RNN) called LSTM. An RNN is a neural network where some of the connections propagate backward in addition to forward. RNNs are commonly used in applications that involve time-series data since the feedback connections help it learn temporal sequences. LSTMs are a refinement on RNNs that mitigate the vanishing gradient problem where the model unintentionally learns to ignore the feedback connections. Handling missing data is an active area of research in the machine learning community[33]. The models described in this paper ignore observations with missing features and the LSTM uses the two most recent observations, even if they are not the two preceding hours, but other approaches from emerging research could be examined in future work.

The term “ANN” is used hereafter to only refer to the simple ANN model. Hyperparameters for each model were chosen through a grid search approach after eliminating hyperparameters that never resulted in competitive models. The hyperparameters used for the ANN and LSTM can be found in Table S2, while Table S3 lists the options for hyperparameters that were selected.

3.4. Time Series Regression

Given LSTM’s capability in handling time-series data and learning temporal patterns, two additional regression techniques, i.e., DLM and LASSO, were employed as a way to control for autocorrelation. Both DLM and LASSO can model the delayed effects of the independent time-series variables on the dependent time-series variable including different time lags. The PLS model can also be modified to include lagged independent variables as predictors (PLS-Lag). To facilitate a fair comparison with LSTM, additional ANN models (ANN-Lag) were evaluated where the two most recent observations were used in the learning process similar to LSTM.

3.5. Model Training and Testing

The dataset of 670 observations with 17 columns of the independent variables and one column of the dependent variable was separated into a training set and a testing set. The time sequence structure of the dataset was maintained. The first 546 observations (81%) were kept in the training set and the rest in the testing set. This split was done due to the fact that a large time gap existed between observations 546 and 547 because of missing data. The training set was further split into a training subset and validation subset for the implementation of cross-validation (CV). A rolling forecasting origin technique as discussed by Hyndman and Athanasopoulos[34] was used to create 10 resamples of the training and validation subsets while maintaining time sequence. The details of the CV scheme are illustrated in Figure S4. During CV, the final model was selected based on RMSE.

3.6. Model Implementation

A consistent three-phase model framework was implemented across the four modeling approaches, i.e., MLR, PLS, ANN, and LSTM:

- During Phase 1, for PLS, ANN, and LSTM models, all 17 predictors were included. For the MLR model, because some predictors could be highly correlated, a bidirectional stepwise regression was run to determine the best subset of predictors for the full model. Remaining predictors in the subset with high VIF values were removed.
- During Phase 2, the Phase 1 models were re-evaluated after removing predictor PM_O.

- During Phase 3, the Phase 2 models were re-evaluated after removing all other $PM_{2.5}$ independent variables.

The three phases were designed in accordance with the potential difficulties of obtaining $PM_{2.5}$ measurements. On-site ambient $PM_{2.5}$ measurement requires high-grade instruments to provide accurate readings regardless of the ambient weather condition, which may not be feasible for some buildings due to economic or operational constraints. Ambient $PM_{2.5}$ records from government agencies are publicly available, but some labor costs may be involved to acquire and process the data. In addition, regulatory monitoring sites may not exist in the target city or even nearby cities. Therefore, by evaluating models without certain $PM_{2.5}$ predictors, users are given the option to choose the best approach based on each building's unique condition.

3.7. Model Evaluation

Several indicators, i.e., NAE, RMSE, R^2, and IA, are used to compare the performance of the predictive models. The use of these indicators was also demonstrated in Elbayoumi et al[26]. The NAE and RMSE are smaller-the-better metrics that measure the existing error of the model, while the R^2 and IA are larger-the-better metrics that measure the accuracy of the model. Calculation of each indicator is given in Equations (1) through (4):

$$NAE = \frac{\sum_{i=1}^{N} |P_i - O_i|}{\sum_{i=1}^{N} O_i} \tag{1}$$

$$RMSE = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(P_i - O_i)^2} \tag{2}$$

$$R^2 = \left(\frac{\sum_{i=1}^{N}(P_i - \overline{P})(O_i - \overline{O})}{N \cdot S_p \cdot S_o}\right)^2 \tag{3}$$

$$IA = 1 - \left[\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(|P_i - \overline{O}| + |O_i - \overline{O}|)^2}\right] \tag{4}$$

where N is the number of observations; P_i and O_i are the predicted and observed values of the i^{th} observation; $\overline{P}$ and $\overline{O}$ are the averages of the predicted and observed values; S_p and S_o are the standard deviations of the predicted and observed values.

4. RESULTS

4.1. Descriptive Statistics

A summary of the descriptive statistics of the dataset is provided in Table S4. The dataset contained observations from October 10$^{\text{th}}$ through November 13$^{\text{th}}$ in 2019. As shown in Table S4, the mean hourly averaged indoor $PM_{2.5}$ was 5.68 μg/m^3 while the mean hourly ambient $PM_{2.5}$ measured at the UW Tower was 4.07 μg/m^3. The ambient $PM_{2.5}$ at the five PSCAA monitoring sites were also within acceptable range per the National Primary and Secondary Ambient Air Quality Standards[35] most of the time (see Figure S5). Pearson's correlation coefficients of the predictors are summarized in Table S5.

4.2. Model Performance

Using the three-phase implementation framework and the three modeling approaches, several models were evaluated for their performance in predicting indoor $PM_{2.5}$. Considering that regulatory ambient $PM_{2.5}$ records from nearby monitoring sites should be relatively easy to obtain for most commercial office buildings located in urban centers, only the Phase 2 model results are presented here to be succinct. Results for Phase 1 and Phase 3 models are included in the SI (See

Figure S6, Table S6, and Table S7). The values of calculated Phase 2 model performance indicators are summarized in Table 2. Figures 2 and 3 show the predicted values versus observations for different models using the testing dataset without and with the temporal information considered. Details of the results obtained for each modeling approach are discussed in the following sections.

Table 2. Performance indicators of the Phase 2 models.

Temporal Information	Model No.	Method	Predictors	NAE	RMSE	R^2	IA
Not considered	M1	MLR	10	0.37	3.07	0.60	0.82
	P1	PLS	16	0.35	2.89	0.60	0.83
	A1	ANN	16	0.26	2.38	0.67	0.88
Considered	DL1	DLM	10	0.24	2.20	0.71	0.91
	LA1	LASSO	20	0.33	2.65	0.65	0.85
	PL1	PLS-Lag	30	0.54	4.71	0.02	0.50
	AL1	ANN-Lag	16	0.29	2.63	0.60	0.89
	L1	LSTM	16	0.18	1.73	0.83	0.94

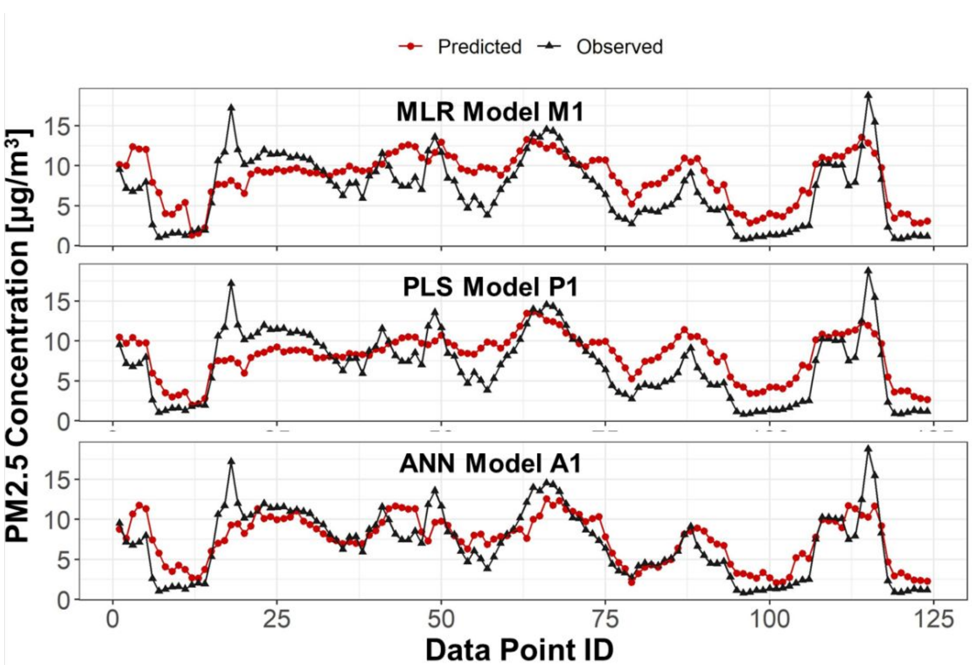

Figure 2. Plots of predicted and observed values for Phase 2 MLR, PLS, and ANN models using the testing set.

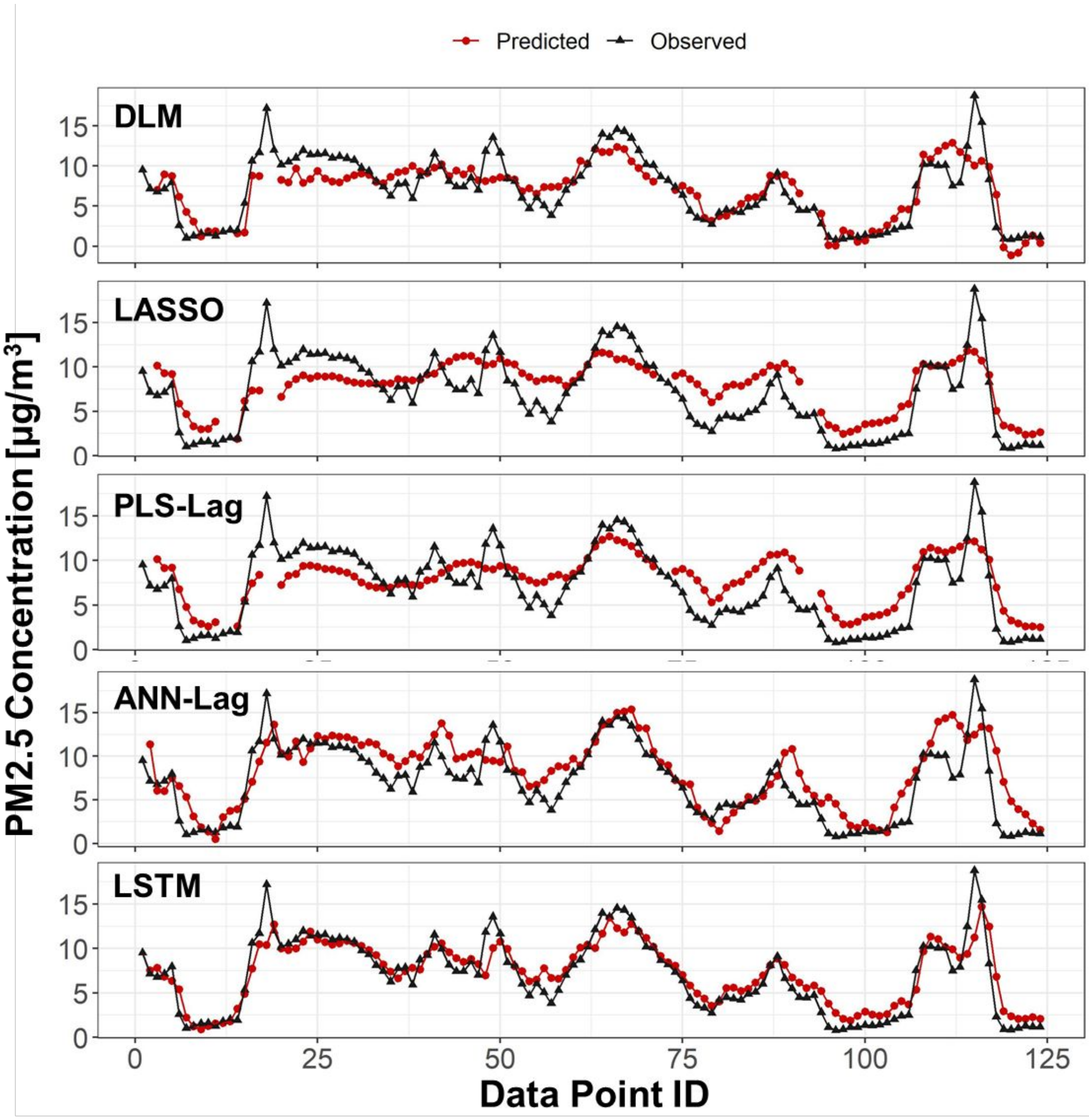

Figure 3. Time series plots of predicted and observed values for DLM, LASSO, PLS-Lag, ANN-Lag and LSTM models.

4.2.1. MLR Modeling Results

The MLR was conducted in R[36] using the "caret"[37] package for cross-validation and the "MASS"[38] package for bidirectional stepwise regression. As discussed in Mansfield and Helms[39], multicollinearity is not a problem if the VIFs are not unusually larger than 1.0. The VIFs of the ten predictors in Model M1 are in the range of 1.22-2.92, and the air intake damper opening D appears to have a significant effect on the indoor $PM_{2.5}$, as shown in Table 3. Similar results can be observed for the Phase 1 and Phase 3 models (see Section 1.5.1 and Table S8 in the SI). The exclusion of on-site ambient $PM_{2.5}$ predictor PM_O led to a slight increase of RMSE (12%) in Model M1 compared to Model M2 while the ambient $PM_{2.5}$ from other locations were kept in the model. Removal of all the ambient $PM_{2.5}$ predictors, led to an increase of RMSE of 18% in Model M3 compared to Model M2. A modified version of Model M2 was evaluated by swapping PM_W, PM_{LFP}, and PM_D with on-site PM_O data, and the results were similar (RMSE decreased by 9%). It shows that the inclusion of some ambient $PM_{2.5}$ information, not necessarily measured on-site, could improve the prediction accuracy of the model.

Table 3. Summary of Model M1 results.

	Intercept	PM_W	PM_{LFP}	PM_D	T_S	T_F	RH_E	O_F	D	WD	WS
VIF	-	2.71	2.21	2.92	1.84	1.44	1.32	1.26	1.43	1.22	1.49
Coef	-7.28	0.08	0.14	0.17	-0.10	0.17	0.17	0.01	-6.19	0.004	-0.96

4.2.2. PLS Modeling Results

The PLS regression was conducted in R[36] using the "pls"[40] package and the kernel algorithm[41]. The optimal number of components in the model was determined using a randomization test approach[42], which checked whether the squared prediction errors of the models with fewer components were significantly larger than in the reference model and selected the smallest model not significantly worse than the reference model. Figure S7 shows the cross-validation plots and the determined number of components for each model.

The percentage of variance explained by each component of the PLS model for both the predictors and outcome is summarized in Table S9. Component 1 of all three PLS models appears to make the most contribution (49.54%~62.01%) in explaining the variance of the outcome variable. For each component of Model P1, the loading value of each predictor is shown in Figure 4. It can be seen that all of the ambient $PM_{2.5}$ predictors carried large positive loading values in Component 1 when they were included in the model. A similar effect can be observed for Models P2 and P3 (see Figure S8). Additional discussion regarding other predictors can be found in Section 1.5.2 of SI.

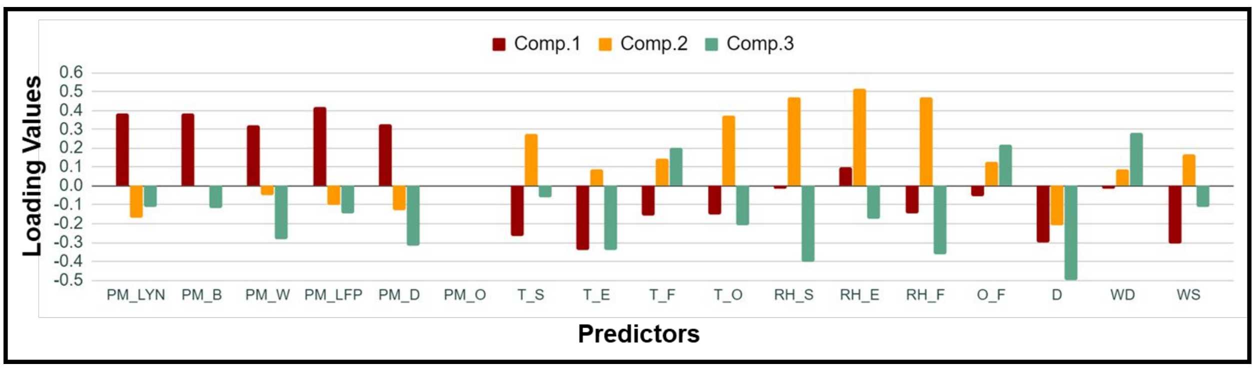

Figure 4. Loading values of each variable for each component of Model P1.

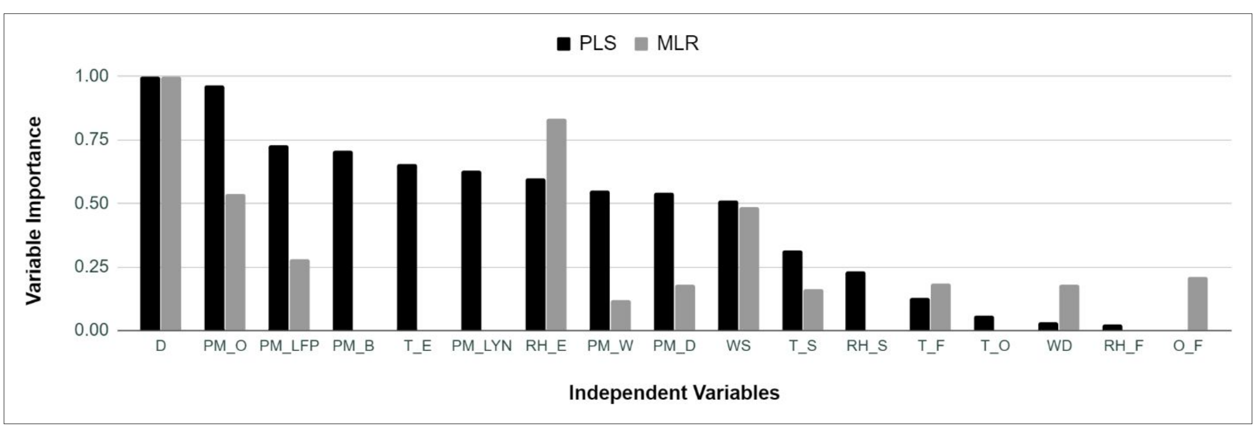

Figure 5. The variable importance of all the predictors using PLS and MLR approaches.

The variable importance (VIP) was calculated using the "caret"[37] package in R for both the MLR and PLS Phase 1 models which included all the predictors, as shown in Figure 5. It can be seen that the critical predictors are relatively consistent in either approach. The damper opening, ambient $PM_{2.5}$ level (on-site as well as from other locations), well-mixed air temperature and RH, and wind speed appear to be the variables that affect the prediction to a greater extent.

4.2.3. ANN and LSTM Modeling Results

The results reported in Table 2 represent the average values for each metric across all folds of the 10-fold cross-validation run using that model (see Table S6 for Phase 1 and Phase 3 models). The model was selected based on the best average RMSE value. Table S10 shows the hyperparameters that were used in each of these models. Note that many of the models built with alternate hyperparameter configurations also performed nearly as well in RMSE and some performed better in the other metrics.

The ANN was tested both with and without temporal information included in the training. The model performance did not always improve when the previous two hours' data were considered in the training (RMSE degraded by 0.89 and 0.25 μg/m^3 for Phase 1 and Phase 2 models, and improved by 0.52 μg/m^3 for the Phase 3 model). The LSTM model, which learns temporal patterns, was also trained and tested with inputs from the previous two hours in addition to the current values for every variable excluding the dependent variable, PM_E. As shown in Table 2, the LSTM model outperformed the other models across the four metrics.

4.2.4. Time Series Regression Results

The DLM regression was conducted in R[36] using the "dLagM"[43] package. The built-in autoregressive distributed lag (ARDL) bounds testing function was used to compute the optimal lag structure for the independent variables of Model DL1. The maximum lag period considered was two hours (same as in LSTM) and only ambient $PM_{2.5}$, occupancy, and damper opening variables were included in the lag structure. The results suggested a 2-hour lag for PM_D, 1-hour lag for damper opening and zero lags were used for all the other variables (as listed in Table S7). By using the lagged variables, the delayed effect of these predictors was included in the DL1 model.

The LASSO regression was ran using the "glmnet"[44] package in R. The optimal tuning parameter λ which controls the overall strength of the penalty was selected by rolling forecast origin CV as discussed in Section 3.5. Both 1-hour and 2-hour lagged PM_W, PM_{LFP}, PM_D, O, and D were included in LASSO as well as the unlagged versions. Other variables were included without any lags. The PLS-Lag model was based on Model P1. Similar to Model LA1, both 1-hour and 2-hour lagged PM_{LYN}, PM_B, PM_W, PM_{LFP}, PM_D, O and D were included in PLS-Lag as well as their unlagged versions. Other variables were kept without any lags.

The results show better performance of Model DL1 compared to LA1 and PL1 as listed in Table 2. However, from Figure 3, the three models are outperformed by LSTM. Notice that gaps exist in the predicted time series of DL1, LA1, and PL1 models, due to missing data of in some of the predictors. As mentioned in Section 3.3, LSTM was able to use the most recent two observations regardless of missing data, no gaps exist for the LSTM predictions.

5. DISCUSSION

The development and comparison of the various predictive models have shown that the indoor $PM_{2.5}$ in the well-mixed air in this office space could be estimated by using readily available variables. In general, when temporal information is not considered, the performance of the models developed using the MLR, PLS, and ANN methods were comparable in terms of their NAE, RMSE, R^2, and IA, as shown in Table 2 and Table S6. With temporal information included in the model, the LSTM method outperformed the DLM, LASSO, PLS-Lag, and ANN-Lag presumably because the LSTM took into consideration the lagged effect as well as the rate of change in the predictor variables. For example, the LSTM model may learn that the value of a variable from 2 hours ago has an effect on PM_E at the current time. In addition, it may also learn that the rate of change of a variable over the past 2 hours has an effect on PM_E at the current time. The models with a large number of independent variables appear to provide a marginally better prediction for regression models (M2 and P2), but not for the neural network models (A2, AL2, and L2). The reduced models with fewer predictors are still capable of making accurate predictions. The results suggest that although on-site measurement of ambient $PM_{2.5}$ could aid in predicting the indoor level, using measurements from other publicly available monitors instead has minimal impact on the model performance. By including some form of ambient $PM_{2.5}$ measurements (not necessarily on-site), there is a significant improvement in the model results.

The RMSE values of the regression and ANN models developed in this paper for the office space are in the range of 2.05~4.71 μg/m^3 while the values of the LSTM models are in the range of 1.73~1.93 μg/m^3. In comparison, as summarized in Wei et al.[10], the RMSE values for regression models developed for indoor $PM_{2.5}$ in schools[26] and private dwellings[45-47] were in the range of 0.45~1.7 μg/m^3. The ANN type models have been used to predict indoor $PM_{2.5}$ in subway stations[48, 49], dwellings[50], and schools[26]. Similarly to the regression models, the reported error values were small for dwellings and schools (1~3 μg/m^3 RMSE) but large for the subway stations (RMSE over 10 μg/m^3). This paper shows that regression and simple ANN models are quite capable of predicting indoor $PM_{2.5}$ in offices. Using an LSTM to account for time trends in the predictor variables further provides a significant improvement of the prediction performance over time series regression methods.

Unlike schools, dwellings, and ambient air, a mechanically ventilated office has a relatively consistent indoor environment controlled by the ventilation system. With the air filtration in place, the correlation of indoor and outdoor $PM_{2.5}$ exists but is not as high as in a naturally ventilated building. Therefore, in addition to the usual meteorological variables, i.e., air temperature, RH, and wind speed, other building-related variables, i.e., damper opening and occupancy, also appeared to be useful. Due to the difficulties in obtaining these building-related variables, few studies have included them in the prediction models. As shown in Figure 5, the damper opening was the most important variable in both MLR and PLS models. Since the studied office space did not have any operable windows, the ventilation system was the main route through which ambient PM entered the indoor environment while infiltration and tracking remained secondary routes. The high variable importance of the damper opening and ambient PM variables in Figure 5 show that the ambient condition has a major influence on the indoor environment. In addition, the air temperature and relative humidity of the well-mixed air, as well as the outdoor wind speed, also appear to carry large weights in the prediction model. As discussed in Gundel and Destaillats[51], the ambient particles go through a phase change when entering the building via ventilation or infiltration due to the change of temperature and relative humidity conditions.

The results presented in this paper have some practical implications. First, a rooftop ambient $PM_{2.5}$ monitor is not always necessary to produce a fairly good prediction of well-mixed indoor $PM_{2.5}$ unless nearby regulatory monitors do not exist. From a building management perspective, this is encouraging as there is always cost associated with conducting on-site $PM_{2.5}$ measurements, including material and labor costs for sensor procurement, installation, and data analysis. The other predictors, such as indoor air temperature, relative humidity, and wind speed, are all commonly monitored environmental parameters in existing commercial buildings or could be obtained from nearby weather stations. The air intake damper opening could be difficult to obtain if the building control system lacks the capability of continuous monitoring of the damper position. In this case, substitute parameters, e.g., outdoor airflow rate, could be used if such monitoring is easier to implement. Building occupancy monitoring also requires a specialized sensor for data collection. Nevertheless, the importance of occupancy in the models is relatively low compared to the other predictors, as shown in Figure 5.

Some limitations exist in this paper. The analysis dataset only contained measurements from October and November in 2019 when the outdoor weather was relatively mild in the Seattle area. Extensive data collection is needed to evaluate the model performance in different seasons. The ambient $PM_{2.5}$ in the Seattle area was maintained at a healthy level during the measurement. It is unknown whether the degradation of ambient air quality (e.g., during wildfire events) could affect the predictive capability of the models. The outcome $PM_{2.5}$ variable was measured in the exhaust/well-mixed air, and it may not be the same as the $PM_{2.5}$ measured at other locations in the building. How well these models predict the $PM_{2.5}$ level at other indoor locations is not in the scope of this paper. It is also recognized that the models were trained and tested using data from one floor in a single building. Their applicability at other building sites with different ambient air condition and building characteristics is rather limited. Buildings with natural ventilation and operable windows could have very different set of significant predictors than discovered in this paper which in turn would affect the model performance. This issue has also been raised in Challoner et al.[11] and Wei et al.[10] Future field studies covering various climate regions and building types could validate and improve the results obtained from small-scale investigations. As the low-cost PM sensors become more reliable and widely used, it would require less effort to conduct these large-scale investigations to obtain more generalized findings. IAQ simulation tools such as CONTAM[52] could also serve as another avenue for validating the predictive models from a physical and mechanical perspective. When the simulation is coupled with commercial reference buildings[53], the results could be applicable to the most common commercial buildings.

In summary, this paper shows that it is feasible to develop a relatively accurate indoor $PM_{2.5}$ prediction model for well-mixed air in a mechanically ventilated office space using some readily available meteorological and building-related variables. A straightforward indoor $PM_{2.5}$ prediction model could provide the building owner, facility manager, and occupants insight into the average air quality of the space and empower the stakeholders to make informed decisions related to the management of the indoor environment.

In the future, researchers should continue to explore not just prediction models, but also how to optimize the cost and accessibility of prediction relative to accuracy. The quantity, quality, and placement of sensors augmented by external information and machine learning models are critical to widespread access to such systems. Furthermore, research in this field should progress from prediction to active management, where predictive models such as those presented in this paper are used to actively improve the efficiency of building operations and the quality of life of the building's inhabitants.

ASSOCIATED CONTENT

Supporting Information (SI)

The Supporting Information is available free of charge at http://pubs.acs.org.

Texts: building characteristics and sampling locations; particle mass concentration calculation; occupancy sensor background; ANN background; additional model results.

Figures: location of the UW Tower and other monitoring sites; overview of the study site and sampling locations; measurement duration for all of the variables; data splitting scheme for the cross-validation; time series plot of the indoor and outdoor $PM_{2.5}$ measurements; plots of predicted and observed values for Phase 1 and 3 models; cross-validation plots of the PLS models; loading values of each variable in Models P2 and P3; boxplots of $PM_{2.5}$ level at different hours.

Tables: descriptions of the surrounding environment of each ambient $PM_{2.5}$ monitoring site; list of hyperparameters used for the ANN and LSTM; hyperparameter search space; descriptive statistics of the variables; Pearson's correlation coefficients between independent variables; performance indicators of Phase 1 and 3 models; independent variables included in Phase 1 and 3 models; summary of MLR models M2 and M3; percentage of variance explained by each component of the PLS models; hyperparameters of the best ANN and LSTM models; recursive model results.

AUTHOR INFORMATION

Corresponding Author

*Phone: +1 (206) 685-0228; E-mail: amyakim@uw.edu.

Author Contributions

The contributions of each author are: conceptualization, B.L., T.L., and A.K.; methodology, B.L., S.W., and T.L.; data curation, B.L. and S.W.; writing – orginal draft preparation, B.L., S.W. and A.K.; writing – review and editing, T.L.; visualization, S.W.; and supervision, A.K.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors would like to thank Troy Swanson, UW Tower Facility Manager; Kim Lokan, UW-IT Facilities Services Manager. This material is based upon work supported by the National Science Foundation under Grant No. 1852995.

REFERENCES

1. 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.; Spadaro, J. V., Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. *Proceedings of the National Academy of Sciences* **2018,** *115* (38), 9592-9597.
2. Cesaroni, G.; Badaloni, C.; Gariazzo, C.; Stafoggia, M.; Sozzi, R.; Davoli, M.; Forastiere, F., Long-Term Exposure to Urban Air Pollution and Mortality in a Cohort of More than a Million Adults in Rome. *Environmental Health Perspectives* **2013,** *121* (3), 324-331.
3. Cohen, A. J.; Brauer, M.; Burnett, R.; Anderson, 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.; Forouzanfar, M. H., 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* **2017,** *389* (10082), 1907-1918.
4. Li, T.; Zhang, Y.; Wang, J.; Xu, D.; Yin, Z.; Chen, H.; Lv, Y.; Luo, J.; Zeng, Y.; Liu, Y.; Kinney, P. L.; Shi, X., All-cause mortality risk associated with long-term exposure to ambient $PM_{2.5}$ in China: a cohort study. *The Lancet Public Health* **2018,** *3* (10), e470-e477.
5. Pope III, C. A.; 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.; Burnett, R. T., Mortality Risk and Fine Particulate Air Pollution in a Large, Representative Cohort of U.S. Adults. *Environmental Health Perspectives* **2019,** *127* (7), 077007.
6. 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.; Wuebbles, D. J., Air Pollution and Noncommunicable Diseases: A Review by the Forum of International Respiratory Societies' Environmental Committee, Part 1: The Damaging Effects of Air Pollution. *CHEST* **2019,** *155* (2), 409-416.
7. Klepeis, N. E.; Nelson, W. C.; Ott, W. R.; Robinson, J. P.; Tsang, A. M.; Switzer, P.; Behar, J. V.; Hern, S. C.; Engelmann, W. H., The National Human Activity Pattern Survey

(NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Science & Environmental Epidemiology* **2001,** *11*, 231-252.

8. Bureau of Labor Statistics Graphics for Economic News Releases. https://www.bls.gov/charts/american-time-use/activity-by-sex.htm# (accessed November 29, 2019).

9. EPA Air Quality System (AQS). https://www.epa.gov/aqs (accessed January 27, 2020).

10. Wei, W.; Ramalho, O.; Malingre, L.; Sivanantham, S.; Little, J. C.; Mandin, C., Machine learning and statistical models for predicting indoor air quality. *Indoor Air* **2019,** *29* (5), 704-726.

11. Challoner, A.; Pilla, F.; Gill, L., 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* **2015,** *12* (12), 15233-15253.

12. Putra, J. C. P.; Safrilah, S.; Ihsan, M. In *The prediction of indoor air quality in office room using artificial neural network*, Proceedings of the 4th International Conference on Engineering, Technology, and Industrial Application, Surakarta, Indonesia, AIP Conference Proceedings: Surakarta, Indonesia, 2018; p 020040.

13. Sofuoglu, S. C., Application of artificial neural networks to predict prevalence of building-related symptoms in office buildings. *Building and Environment* **2008,** *43* (6), 1121-1126.

14. Xie, H.; Ma, F.; Bai, Q. In *Prediction of Indoor Air Quality Using Artificial Neural Networks*, 2009 Fifth International Conference on Natural Computation, Tianjin, China, Aug 14-16; Wang, H.; Low, K. S.; Wei, K.; Sun, J., Eds. IEEE: Tianjin, China, 2009; pp 414-418.

15. Chen, X.; Zheng, Y.; Chen, Y.; Jin, Q.; Sun, W.; Chang, E.; Ma, W.-Y. In *Indoor air quality monitoring system for smart buildings*, UbiComp '14: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, WA, USA, Association for Computing Machinery: Seattle, WA, USA, 2014; pp 471-475.

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

17. Kim, H. S.; Park, I.; Song, C. H.; Lee, K.; Yun, J. W.; Kim, H. K.; Jeon, M.; Lee, J.; Han, K. M., 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* **2019,** *19* (20), 12935-12951.

18. Li, X.; Peng, L.; Yao, X.; Cui, S.; Hu, Y.; You, C.; Chi, T., Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution* **2017,** *231* (Part 1), 997-1004.

19. Qi, Y.; Li, Q.; Karimian, H.; Liu, D., 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* **2019,** *664*, 1-10.

20. Zhao, J.; Deng, F.; Cai, Y.; Chen, J., Long short-term memory - Fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. *Chemosphere* **2019,** *220*, 486-492.

21. Puget Sound Clean Air Agency Air Quality. https://www.pscleanair.gov/27/Air-Quality (accessed 2020 July 29).

22. Washington State Department of Ecology Washington's Air Monitoring Network. https://fortress.wa.gov/ecy/enviwa/ (accessed November 30, 2019).

23. University of Washington Rooftop Observations - ATG building UW. https://a.atmos.washington.edu/cgi-bin/list_uw.cgi (accessed 2020 July 29).

24. Particles Plus Inc. 7301-AQM and 7302-AQM Remote Air Quality and Environmental Monitor. https://particlesplus.com/7301-iaq-remote-particle-counter/ (accessed December 1, 2019).

25. NOAA Radiance Research Nephelometer. https://www.esrl.noaa.gov/gmd/aero/instrumentation/RR_neph.html (accessed December 1, 2019).

26. Elbayoumi, M.; Ramli, N. A.; Yusof, N. F. F. M., Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor $PM_{2.5–10}$ and $PM_{2.5}$ concentrations in naturally ventilated schools. *Atmospheric Pollution Research* **2015,** *6* (6), 1013-1023.

27. Elbayoumi, M.; Ramli, N. A.; Yusof, N. F. F. M.; Yahaya, A. S. B.; Al Madhoun, W.; Ul-Saufie, A. Z., Multivariate methods for indoor PM10 and PM2.5 modelling in naturally ventilated schools buildings. *Atmospheric Environment* **2014,** *94*, 11-21.

28. Marengo, E.; Bobba, M.; Robotti, E.; Liparota, M. C., Modeling of the Polluting Emissions from a Cement Production Plant by Partial Least-Squares, Principal Component Regression, and Artificial Neural Networks. *Environmental Science and Technology* **2006,** *40* (1), 272-280.

29. Huang, T.; Yu, Y.; Wei, Y.; Wang, H.; Huang, W.; Chen, X., Spatial–seasonal characteristics and critical impact factors of PM2.5 concentration in the Beijing–Tianjin–Hebei urban agglomeration. *PLoS ONE* **2018,** *13* (9), e0201364.

30. Kim, M.; SankaraRao, B.; Kang, O.; Kim, J.; Yoo, C., Monitoring and prediction of indoor air quality (IAQ) in subway or metro systems using season dependent models. *Energy and Buildings* **2012,** *46*, 48-55.

31. Lee, S.; Kim, M. J.; Kim, J. T.; Yoo, C. K., In search for modeling predictive control of indoor air quality and ventilation energy demand in subway station. *Energy and Buildings* **2015,** *98*, 56-65.

32. Lim, J.; Kim, Y.; Oh, T.; Kim, M.; Kang, O.; Kim, J. T.; Kim, I.-W.; Kim, J.-C.; Jeon, J.-S.; Yoo, C., Analysis and prediction of indoor air pollutants in a subway station using a new key variable selection method. *Korean Journal of Chemical Engineering* **2012,** *29* (8), 994-1003.

33. Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; Liu, Y., Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports* **2018,** *8*, 6085.

34. Hyndman, R. J.; Athanasopoulos, G., *Forecasting: principles and practice*. OTexts: Melbourne, Australia, 2018.

35. EPA, National Primary and Secondary Ambient Air Quality Standards. U.S. Environmental Protection Agency, Ed. 2016; Vol. 40 CFR Part 50.

36. R Core Team *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing: Vienna, Austria, 2019.

37. Kuhn, M. *caret: Classification and Regression Training*, R Foundation for Statistical Computing: Vienna, Austria, 2019.

38. Ripley, B. *MASS: Support Functions and Datasets for Venables and Ripley's MASS*, R Foundation for Statistical Computing: Vienna, Austria, 2019.

39. Mansfield, E. R.; Helms, B. P., Detecting Multicollinearity. *The American Statistician* **1982,** *36* (3 Part 1), 158-160.

40. Mevik, B.-H.; Wehrens, R.; Liland, K. H. *pls: Partial Least Squares and Principal Component Regression*, R Foundation for Statistical Computing: Vienna, Austria, 2019.

41. Dayal, B. S.; MacGregor, J. F., Improved PLS algorithms. *Journal of Chemometrics* **1997,** *11* (1), 73-85.
42. van der Voet, H., Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and Intelligent Laboratory Systems* **1994,** *25* (2), 313-323.
43. Demirhan, H. *dLagM: Time Series Regression Models with Distributed Lag Models*, R Foundation for Statistical Computing: Vienna, Austria, 2020.
44. Friedman, J.; Hastie, T.; Tibshirani, R.; Narasimhan, B.; Tay, K.; Simon, N.; Qian, J. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*, R Foundation for Statistical Computing: Vienna, Austria, 2020.
45. Jafta, N.; Barregard, L.; Jeena, P. M.; Naidoo, R. N., Indoor air quality of low and middle income urban households in Durban, South Africa. *Environmental Research* **2017,** *156*, 47-56.
46. Yuchi, W. Modelling Fine Particulate Matter Concentrations inside the Homes of Pregnant Women in Ulaanbaatar, Mongolia. Master of Science, Simon Fraser University, Burnaby, British Columbia, Canada, 2017.
47. Yuchi, W.; Gombojav, E.; Boldbaatar, B.; Galsuren, J.; Enkhmaa, S.; Beejin, B.; Naidan, G.; Ochir, C.; Legtseg, B.; Byambaa, T.; Barn, P.; Henderson, S. B.; Janes, C. R.; Lanphear, B. P.; McCandless, L. C.; Takaro, T. K.; Venners, S. A.; Webster, G. M.; Allen, R. W., Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environmental Pollution* **2019,** *245*, 746-753.
48. Kim, M.; Kim, Y.; Sung, S.; Yoo, C. In *Data-driven prediction model of indoor air quality by the preprocessed recurrent neural networks*, Proceedings of the ICROS-SICE International Joint Conference 2009, Fukuoka City, Japan, IEEE: Fukuoka City, Japan, 2009; pp 1688-1692.
49. Loy-Benitez, J.; Vilela, P.; Li, Q.; Yoo, C., Sequential prediction of quantitative health risk assessment for the fine particulate matter in an underground facility using deep recurrent neural networks. *Ecotoxicology and Environmental Safety* **2019,** *169*, 316-324.
50. Das, P.; Shrubsole, C.; Jones, B.; Hamilton, I.; Chalabi, Z.; Davies, M.; Mavrogianni, A.; Taylor, J., Using probabilistic sampling-based sensitivity analyses for indoor air quality modelling. *Building and Environment* **2014,** *78*, 171-182.
51. Gundel, L. A.; Destaillats, H., Aerosol Chemistry and Physics - An Indoor Perspective. In *Aerosols handbook : measurement, dosimetry, and health effects*, 2nd ed.; Ruzer, L. S.; Harley, N. H., Eds. CRC Press: Boca Raton, FL, 2013.
52. NIST CONTAM. https://www.nist.gov/services-resources/software/contam (accessed August 2).
53. DOE Commercial Reference Buildings. https://www.energy.gov/eere/buildings/commercial-reference-buildings (accessed August 2).