5-2023

Advanced Air Quality Management with Machine Learning

Cheng-Pin Kuo
University of Tennessee, Knoxville, ckuo6@vols.utk.edu

Recommended Citation

Kuo, Cheng-Pin, "Advanced Air Quality Management with Machine Learning. " PhD diss., University of Tennessee, 2023.
https://trace.tennessee.edu/utk_graddiss/8169

To the Graduate Council:

I am submitting herewith a dissertation written by Cheng-Pin Kuo entitled "Advanced Air Quality Management with Machine Learning." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Environmental Engineering.

Joshua S. Fu, Major Professor

We have read this dissertation and recommend its acceptance:

Chris Cox, Shuai Li, Russell Zaretzki

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Advanced Air Quality Management with Machine Learning

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Cheng-pin Kuo

May 2023

Dedication

I dedicate this thesis to my beloved family, my partner,
and my home country, Taiwan.

Acknowledgments

First, I would like to thank my advisor, Dr. Joshua Fu, for his help and support in my Ph.D. degree since 2018. His outstanding teaching skills and patience guided me to develop professional skills and conduct scientific research independently. His outreach network also provided me with numerous opportunities to meet and work with talented researchers from universities, research communities, and governmental agencies. Second, I would like to thank the members of my committee, Dr. Chris Cox, Dr. Shuai Li, and Dr. Russell Zaretzki for their valuable comments and suggestions on my work. I would also like to thank laboratory members, Dr. Xinyi Dong, Dr. Cheng-en Yang, Dr. Jiani Tan, Matthew Jonathan Tipton, and Rong-You Chien for their selfless help during my research works. Last but not least, I gratefully thank my beloved family for their endless support and encouragement these years.

Abstract

Air pollution has been a significant health risk factor at a regional and global scale. Although the present method can provide assessment indices like exposure risks or air pollutant concentrations for air quality management, the modeling estimations still remain non-negligible bias which could deviate from reality and limit the effectiveness of emission control strategies to reduce air pollution and derive health benefits. The current development in air quality management is still impeded by two major obstacles: (1) biased air quality concentrations from air quality models and (2) inaccurate exposure risk estimations

Inspired by more available and overwhelming data, machine learning techniques provide promising opportunities to solve the above-mentioned obstacles and bridge the gap between model results and reality. This dissertation illustrates three machine learning applications to strengthen air quality management: (1) identifying heterogeneous exposure risk to air pollutants among diverse urbanization levels, (2) correcting modeled air pollutant concentrations and quantifying the bias of sources from model inputs, and (3) examine nonlinear air pollutant responses to local emissions. This dissertation uses Taiwan as a case study, due to its well-established hospital data, emission inventory, and air quality monitoring network.

In conclusion, although ML models have become common in atmospheric and environmental health science in recent years, the modeling processes and output interpretation should rely on interdisciplinary professions and judgment. Except for meeting the basic modeling performance, future ML applications in atmospheric and environmental health science should provide interpretability and explainability in terms of human-environment interactions and interpretable physical/chemical mechanisms. Such applications are expected to feedback to traditional methods and deepen our understanding of environmental science.

Keywords:

$PM_{2.5}$, ozone, machine learning, measurement-model fusion, response surface modeling, disease burden, nonlinear response

Table of Contents

List of Tables

List of Figures

Acronyms and Abbreviations

AQMS	air quality monitoring station
AIC	Akaike information criteria
BD	burden of the disease
CFD	computational fluid dynamics
CFR	concentration-response function
CI	confidence interval
CMAQ	Community Multiscale Air Quality model
CNN	convolutional neural network
CO	carbon monoxide
cp	complexity parameter
CTM	chemical transport model
CVD	cardiovascular disease
df	degrees of freedom
DL	deep learning
ED	emergency department
EI	entropy index
EKMA	empirical kinetic modeling approach
GBM	gradient boosting model
HAP	hazardous air pollutant
HDRA	high-density residential area
HLUL	heterogeneity of land-use living pattern
IC	initial condition
IER	integrated exposure-response function
KNN	k-nearest neighbors regression
LDRA	low-density residential area
LHS	Latin hypercube sampling
LSTM	long short-term memory model
MAE	mean absolute error
MDA8	maximum daily 8-hour average
MEI	mean entropy index
ML	machine learning
MMF	measurement-model fusion
MODIS	Moderate Resolution Imaging Spectroradiometer
MSE	mean squared error
NE	normalized error
NH_3	ammonia

NOx	nitrogen oxides
NO_2	nitrogen dioxide
OR	odds ratio
O_3	ozone
PBL	planetary boundary layer
PM	particulate matter
ReLU	rectifier linear unit
RD	respiratory disease
RF	random forest
RMSE	root mean square error
RNN	recurrent neural network
RR	relative risk
RSM	response surface modeling
RT	regression tree
SOx	sulfur oxides
SO_2	sulfur dioxide
SSE	sum of square errors
TEDS	Taiwan Emission Data System
UV	ultraviolet
VIF	variance inflation factor
VOC	volatile organic compound
WHO	World Health Organization
WRF	weather research and forecasting model

Chapter 1. Introduction

1.1. Machine learning in atmospheric science

The development of atmospheric science is closely related to human health, as many air pollutants such as fine particle matter ($PM_{2.5}$, particles with aerodynamic diameters under 2.5 µm) and ozone (O_3) have been identified to significantly link with acuter or chronic adverse health effects like premature death, cardiovascular disease (CVD), and respiratory disease (RD) [1–3]. Thus, predicting air pollutant concentrations is crucial for policymakers and personal daily living patterns.

The atmospheric prediction models can be categorized into two main types: numerical models and statistical models. Numerical models such as chemical transport models (CTM), box models, Lagrangian/Eulerian dispersion models, and computational fluid dynamics (CFD) model estimate atmospheric constitutes based on scientific or empirical deterministic equations and physical and chemical mechanisms, so the applications of numerical models were convincing and popular in the past. However, the development of numerical models has been slow due to a limited understanding of the complex environment, and the high computational cost and long execution time cannot always provide timely support for policymakers. The bias between observations and predictions from numerical models also still remained significant but easily ignored.

In recent years, due to the rapid development of computational hardware, computing algorithms, and more available monitoring/measurement data, statistical models including machine learning (ML) models have aroused widespread applications in atmospheric science. The advantages of ML models including high efficiency and better nonlinear fitting capability provide an alternative option for time-limited support when numerical models cannot perform well and execute in time. The trend of ML model applications from 2000 to 2020 in atmospheric science (Figure 1-1) shows that the trend of the number of publications significantly increase after 2015, and most studies focused on particulate matter (PM) and O_3 pollution [4].

The types of ML models in atmospheric science applications depend on the task objectives, and generally, the basic algorithm of ML models can be classified into (1) linear regression (LR), (2) k-nearest neighbors' regression (KNN), (3) tree models, and (4) deep-learning (DL) structure models.

(1) Linear regression (LR)

The LR model has been developed for a long history, and the major applications of LR model were used in land-use regression [5–7], which estimates air pollutant concentrations with relatively long-term periods such as annual or monthly scale. LR models such as Multiple Linear Regression (MLR), ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic Net are based on the following basic equation:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

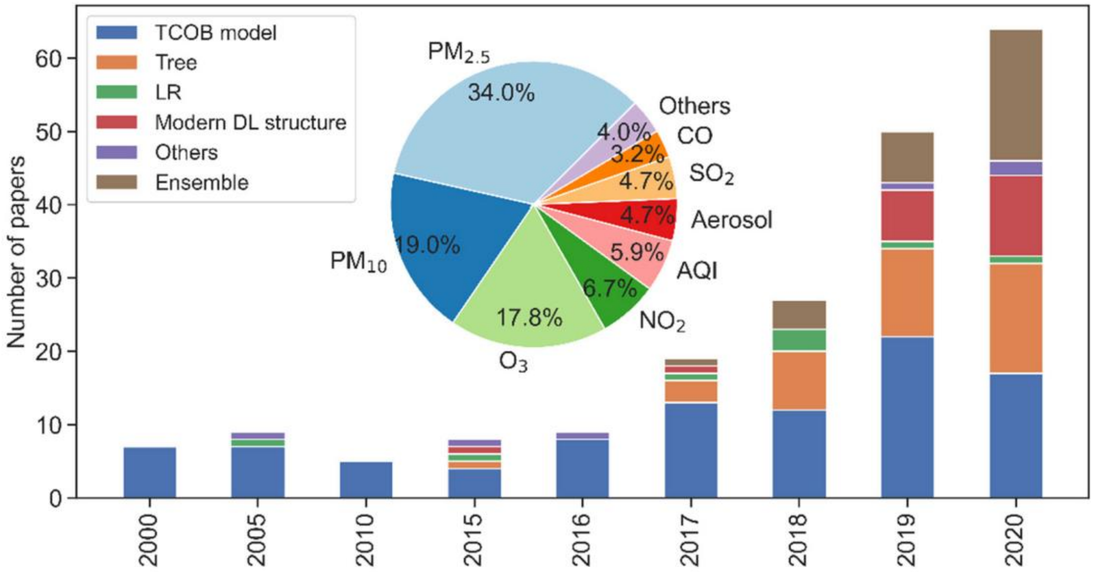


Figure 1-1. Time series of the number of literatures for ML applications in atmospheric science; The categories include traditional convex optimization-based (TCOB) models, tree models, linear regression (LR), and modern deep-learning (DL) structure models [4]

where x_i is the selected variables that are used to predict y. Although the LR model applied in land-use regression studies can have an acceptable modeling performance, LR models still hardly predict air quality over a shorter period such as daily or hourly scale due to its inferiority to deal with the nonlinear relationship between air quality and input variables.

(2) k-nearest neighbors' regression (KNN)

The KNN model is a non-parametric method developed on the assumption that similar samples exist near each other, which is suitable to predict the nonlinear response of air quality without an assumption of parametric distribution. The algorithm of KNN is to memorize the training data and predict air pollutant concentrations based on the closest samples with similar patterns of input variables. Euclidean Distance is the most common method and is calculated as follows:

$$\mathrm{D} = \sqrt{(\mathrm{a}_1 - \mathrm{b}_1)^2 + (\mathrm{a}_2 - \mathrm{b}_2)^2 + \cdots + (\mathrm{a}_\mathrm{p} - \mathrm{b}_\mathrm{p})^2}$$

where $\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_\mathrm{n}$ or $\mathrm{b}_1, \mathrm{b}_2, \ldots, \mathrm{b}_\mathrm{n}$ represent the attribute values for two points, D is the Euclidean distance between two points, and p is the number of total input variables. The user needs to select a proper number of nearest neighbors ($\mathrm{k} < \mathrm{p}$), and the prediction is the average of the values of k nearest neighbors. Theoretically, a lower k value would be sensitive to noise and may lead to overfitting, and using a higher k value would include more irrelevant data points and increases the bias [8].

(3) Tree models

Tree models including regression tree (RT) [9], random forest (RF) [10], and gradient-boosted tree models (GBM) [11] are commonly used to forecast air quality or downscale air quality concentration to a finer spatial resolution in the recent years.

The basic idea of RT is recursively partitioning the input space into binary subsets where the output becomes successively more homogeneous. The RT model divides the input variables into several non-overlapping spaces (tree construction) and optimizes the prediction with the greatest reduction in errors for each space (tree pruning). Following by RT model, the RF model fits a set of decision trees and uses averages from decision trees which are trained on a randomly selected subsample of the training data by the bagging approach [12]. The main algorithm to construct an RF regression model to predict air pollutant concentrations are based on the following equation [13]:

$$D = \{(x_m, y_m), m = 1, 2, \dots, n\}, (X, Y) \in R^i * R$$

where Y is air pollutant concentration, and X is the input variable matrix. In each tree (t_i), a random subspace D_i must be generated through a random selection, and the variables were randomly selected for prediction with the number of variables ranging from 1 to $\sqrt{p}$, where p is the total number of variables. By repeated training, an ensemble of N trees (t_i) is grown, and each tree is de-correlated because of the random selection of input variables in each tree. The predicted results are calculated from an average of N trees (h_i). RF regression is an ensemble non-linear regression model. By using the idea of a double random selection of samples and variables, resulting in RF does not intend to overfit.

GBM is an improved model based on decision trees which are grown sequentially using information from previously grown trees. The core idea of GBM uses a negative gradient of the loss function as the residual approximation during growing the trees and minimizes the loss function by reducing the residuals gradually, as shown in the following equation [14]:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^{b}(x)$$

where the prediction starts with the values $\hat{f}(x) = 0$ and the residual $r_i = y_i$, and GBM repeats updating $\hat{f}$ by adding a shrunken version of the new tree, such as $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^{b}(x)$, where λ is the shrinkage parameter that controls the learning rate of boosting until the least mean square error is the lowest. Therefore, GBM can build up consecutive trees that solve the net error of prior trees.

(4) Deep-learning (DL) structure models

DL models evolved from the development of Artificial Neural Network (ANN), which is based on a collection of parallel and interconnected neurons, and the training process uses synaptic weights to store the acquired information in each hidden layer. Modern DL models in atmospheric science applications mainly include Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). RNN is better to capture temporal information based on historic records, and advanced techniques such as Long Short-Term Memory (LSTM) network are also commonly used to forecast air quality based on historic trends.

CNN is mainly applied in the spatial prediction of air pollutants and their temporal trend [15–18]. Deep CNN consists of several neuron layers: a convolutional layer, a pooling layer, and fully connected layers, as shown in Figure 2. The convolutional layer captures different signals of the image by passing many filters over each image, which can reduce the size of the input without losing important information. Mathematically, convolution is the integral measure of the extent to which two functions overlap as one passes over the other [16,19]. The activation function, such as ReLU or softmax, embedded in the convolutional layer is used to provide nonlinear transformation for reducing input data. The pooling layer excludes features with similar attributes and can reduce the computational burden. Among several pooling operations, the max pooling operation and the average pooling operation are the most commonly used operations. The fully connected layer flattens input features into a column vector as output [16,19]. In atmospheric applications, the ReLU activation function, the max pooling operation, and the Adaptive Moment Estimation (Adam) optimizer are commonly employed to extract important features and preserve nonlinearity.

1.2. Motivation of the dissertation

In recent decades, air pollution has been intensively studied due to its significant impact on human health [20], climate [21], ecosystem [22], and regional air quality [23]. Ambient $PM_{2.5}$ and O_3 have been recognized as major air pollutants that contributed to 2.94 million and 0.47 million premature deaths, respectively, in 2017 [20]. Significant association between exposure to ambient $PM_{2.5}$ and increased risks of adverse health effects such as premature death, cardiovascular diseases (CVD), and respiratory diseases (RD) has been identified [2,24,25], while exposure to O_3 is statistically significant related to premature death, decrease of lung function, and RD [26,27].

To protect public health and prevent air quality deterioration, air quality management plans are developed to implement emission control strategies and achieve desired air quality standards such as World Health Organization (WHO) air quality guidelines [28]. The effectiveness of emission control strategies to achieve the goals must rely on the accuracy of air quality concentrations and their derived health benefits. However, current researches in air quality management are still impeded by the two major obstacles: (1) inaccurate exposure risk estimations and (2) biased air quality concentrations from modeling. Owing to more available data such as hospital records, emission inventory, real-time monitoring data, and sampling measurements, machine learning (ML) has provided promising applications to bridge the gap between model results and reality and develop more accurate and effective air quality management methods. This thesis includes three major parts to investigate the potential applications of ML in air quality management.

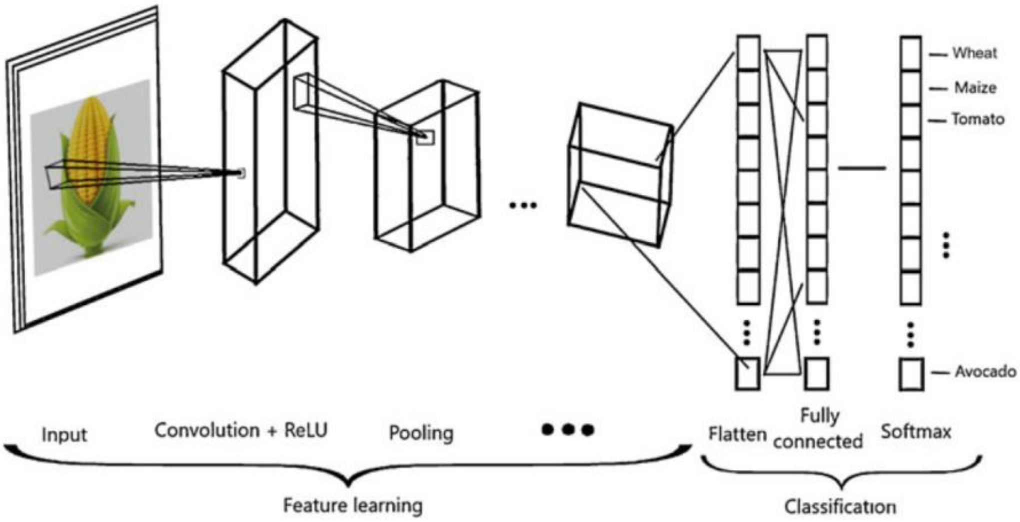


Figure 1-2. Basic structure of convolutional neural network (CNN) [19]

1.3. Aim 1: Heterogeneous exposure risks and burden of the diseases (BD)

The first part of this work presents in Chapter 2 and investigates the application of multiple datasets (e.g. hospital data) to evaluate local exposure risks to air pollutants and their spatial heterogeneities. To quantify the public health impact of air pollution or health benefits of emission control strategies, the burden of disease (BD) is a commonly used assessment tool to calculate the derived health indices such as premature deaths, emergency department (ED) visits, or hospital admissions after exposure to air pollution. Present BD estimation algorithms can provide reasonable estimations based on literature-based relative risks (RRs) and modeled/monitored $PM_{2.5}$ or O_3 concentrations, but the BD estimations still remain biased due to their algorithm assumptions. One potential concern is that most studies employed risk values from other literature or reports to estimate BD without assessing the representativeness of the study subjects or sampling bias of the reference, which could limit the application of the risk values or cause bias if the risk values were inappropriately applied. Another concern is that the spatial heterogeneity of risks among urban and rural areas was overlooked during the BD calculation. Most BD estimations at the country or regional level neglected the potential uncertainties derived from different vulnerabilities associated with living in diverse urbanization levels, for which differences in risk were related to neighboring land-use patterns and individual activity patterns [6,7,29,30] and have been identified in previous studies [5,31].

Conceptually speaking, the value of $PM_{2.5}$ exposure risks is decided by human-environment interactions, $PM_{2.5}$ toxicity, and individual factors, as shown in Figure 1-3. First, human-environment interactions refer to personal exposure duration, frequency and level in an outdoor environment, individual daily living patterns, and land-use characteristics around the home. Second, $PM_{2.5}$ toxicity relies on its size distribution and compositions. If particles contain more hazardous air pollutants (HAPs) such as diesel particles, heavy metals, or dioxins, the risks of related adverse health impacts or diseases would be higher. Third, individual factors such as demographic characteristics (gender and age), genes, habits, knowledge and awareness, social-economic status, and accessibility to the closest medical organizations can also bias the $PM_{2.5}$ exposure risk values. Thus, using a single risk value is not representative enough to illustrate people among various urbanization levels and exposure patterns. Failure to consider the spatial heterogeneity of risks among urban and rural areas could lead to a potential bias of the BD estimations, and its uncertainty should be quantified.

1.4. Aim 2: Bias correction and quantification for numerical models

The second part in Chapter 3 discusses the application of machine learning techniques in measurement-model fusion (MMF) to correct the modeling results and further quantify the bias sources. In order to assess the health impact of criteria air pollutants, either air quality monitoring stations (AQMS) or chemical transport models (CTMs) can provide air pollutant concentrations for further applications. AQMS can provide real-time measurements and alert the officials and neighboring residents in occurrence of air quality deterioration events such as haze or dust storms. However, due to their high maintenance and labor costs, most of AQMSs are usually located in populated areas or around pollutant emission sources such as industry complexes. Sparsely distributed AQMSs cannot monitor the air quality of areas without stations like suburban or rural areas, thus limiting their application in air quality management. Another challenge is the secondary formation of $PM_{2.5}$ and O_3 from chemical transformation. Monitoring data hardly link nonlinear relationships between $PM_{2.5}$ and O_3 and their precursors such as sulfur dioxide (SO_2), nitrogen oxides (NOx), ammonia (NH_3), and volatile organic compounds (VOCs).

Meanwhile, CTMs such as U.S. EPA Community Multiscale Air Quality (CMAQ) model [32]. CMAQ model is a deterministic model that can simulate the behavior of air pollutants in the atmosphere based on the known mechanisms of emission, dispersion, wet/dry deposition, and chemical and physical recreations in the atmosphere, thus CMAQ model can provide temporally and spatially varying air pollutant concentrations in a three-dimension manner. Due to its numerical basis, the CMAQ model was widely applied to forecast air quality or develop emission control strategies under presumed scenarios [13,15]. However, the bias between CMAQ-modeled estimations and observations has been underestimated or even overlooked in previous studies which may limit further applications in air quality management [33]. Multiple reasons such as inaccurate modeling inputs, accumulation of input uncertainties during modeling, and imperfect chemical and physical mechanisms of the model may contribute to considerable biases in air pollutant estimations [13,15].

In recent years, an overwhelming amount of monitoring data provide promising opportunities for machine learning techniques to develop MMF approaches, bridging the gap between modeled estimations and observations and further enhancing predicting performance [13,15]. Such techniques can be employed to apportion the biases between modeled estimations and observations, to which limited studies did not pay attention.

1.5. Aim 3: Examining nonlinear pollutant responses to local emissions

The third part showed in Chapter 4 examines the applicability of machine learning to predict nonlinear responses to precursor emissions by using ambient O_3 concentrations and NOx and VOC emission as an example.

Quantifying the air quality impact of emission sources is necessary for air quality management. The direct evaluation method is comparing scenarios with and without a target emission source, and the difference of air pollutant concentrations is the contribution of that target emission source, which method is also called the brute-force method. Nevertheless, the brute-force method can only apply to one single source and emission reduction of one pollutant. If multiple emission sources and air pollutants are needed to be evaluated, it would use a tremendous amount of computing sources and time to execute CTM modeling under different scenarios, which is a major challenge for policymakers in the before. In practical, to meet the prompt needs of emission analysis for policymakers, Response Surface Modeling (RSM) was developed to retrieve the nonlinear relationship between ambient air pollutant concentrations (e.g. O_3) and multiple precursors emissions (e.g. NOx and VOCs) from multiple emission sources based on a series of CTM simulations such as CMAQ modeling and multidimensional kriging approach [34]. RSM can construct empirical kinetic modeling approach (EKMA) diagrams, as shown in Figure 1-4, which is the USEPA developed isopleth that illustrates the response of air pollutant concentrations (e.g. O_3) to changes precursor emissions (e.g. NOx and VOC) and can be divided into NOx-limited (NOx-sensitive) and VOC-limited (VOC-sensitive) areas and used to assist policy-makers determine whether NOx or VOC emissions should be controlled preferentially in emission control strategies [35].

One typical RSM example is shown in Figure 1-5. To assess the improved O_3 of emission control strategies for industrial NOx emissions (NOx_INDUSTRY), NOx and VOC emissions from mobile sources (NOx_MOBILE and VOC_MOBILE), the Latin hypercube sampling (LHS) method would be employed to design a control matrix with different combination of emission ratios from 0.0 to 1.2 which sample size is large enough to meet the requirement of statistical power. Next, CMAQ modeling was executed based on the emission ratio settings of the control matrix. RSM can further construct the nonlinearity between ambient O_3 concentrations and individual emission sources (NOx_INDUSTRY, NOx_MOBILE, and VOC_MOBILE) based on CMAQ outputs.

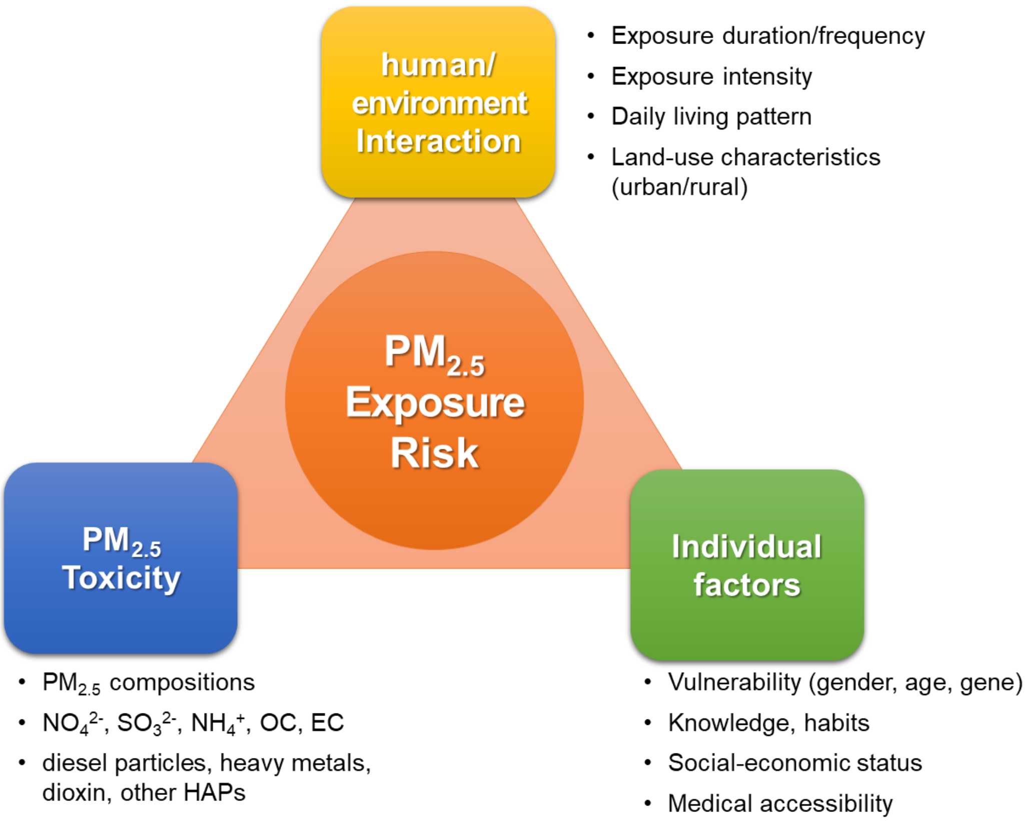

human/ environment Interaction
- Exposure duration/frequency
- Exposure intensity
- Daily living pattern
- Land-use characteristics (urban/rural)
$PM_{2.5}$ Exposure Risk
$PM_{2.5}$ Toxicity
- $PM_{2.5}$ compositions
- NO_4^{2-}, SO_3^{2-}, NH_4^+, OC, EC
- diesel particles, heavy metals, dioxin, other HAPs
Individual factors
- Vulnerability (gender, age, gene)
- Knowledge, habits
- Social-economic status
- Medical accessibility

Figure 1-3. Factors that influence $PM_{2.5}$ exposure risk value

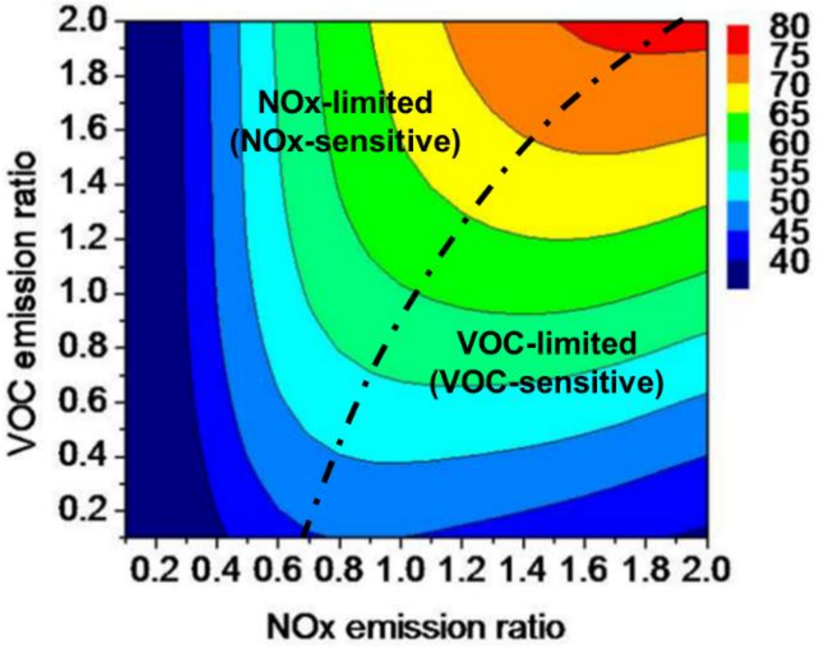


Figure 1-4. An example of typical EKMA O_3 isopleth in terms of NOx and VOC emission ratio (modified from [36])

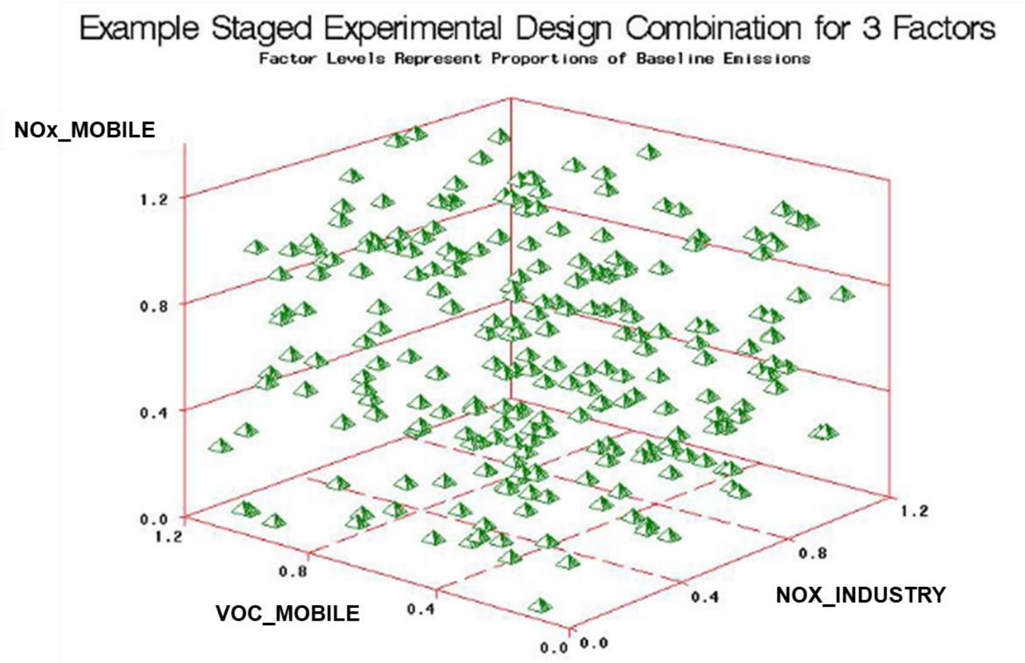


Figure 1-5. Control matrix of industrial NOx emissions (NOx_INDUSTRY), NOx and VOC emissions from mobile sources (NOx_MOBILE and VOC_MOBILE) for RSM

Finally, RSM results can retrieve O_3 concentrations instantly based on emission control strategies with any control ratio from 0% to 120%, thus providing improved O_3 concentrations to policymakers in time.

RSM technique has been successfully applied in several previous studies to optimize local emission control strategies [36,37], but these studies were still based on simulated results and neglected or underestimated the bias between the modeled estimations and observations of the benchmark case [38–40], which could largely affect the nonlinearity between pollutants and precursor emission changes. Without correction of observations in RSM, the improved air quality and derived health benefits such as avoided premature deaths may deviate from the real environment.

Inspired by more available data, the ML technique has been intensively applied and can serve as a bias corrector to adjust modeling results. Several MMF techniques [41] in post-analysis have been developed in recent years to adjust CTM results based on observations [18,42–44]. ML model can also forecast air quality based on historical observations and other auxiliary data (e.g. meteorological and land-use data) without involving CTM results and still have good performance [45–47]. However, whether ML either serves as a bias corrector or a forecaster, few ML studies examined pollutants' sensitivity to their precursor emissions based on observation-corrected results. Without correction of observations, the improved air quality and environmental benefits may deviate from the real environment.

1.6. Study region

In this thesis, Taiwan was chosen as our study region (Figure 1-6), because its well-established and high-availability hospital admission database, island geography, high-density air quality monitoring networks, and three-year-updated emission inventory can provide a supportive base to develop ML models. Also, Taiwan EPA has built up a computation-intensive RSM database for assessing diverse emission control strategies under air pollution events [37], which can also facilitate developing data-driven air quality management techniques. Furthermore, Taiwan has the 17th highest population density in the world (660/km^2), which magnifies the public health impact of air pollution exposure and illustrates the importance of fast and effective air quality policies. Also, the developed air quality management experience can be transformed to other developed and populated cities and countries.

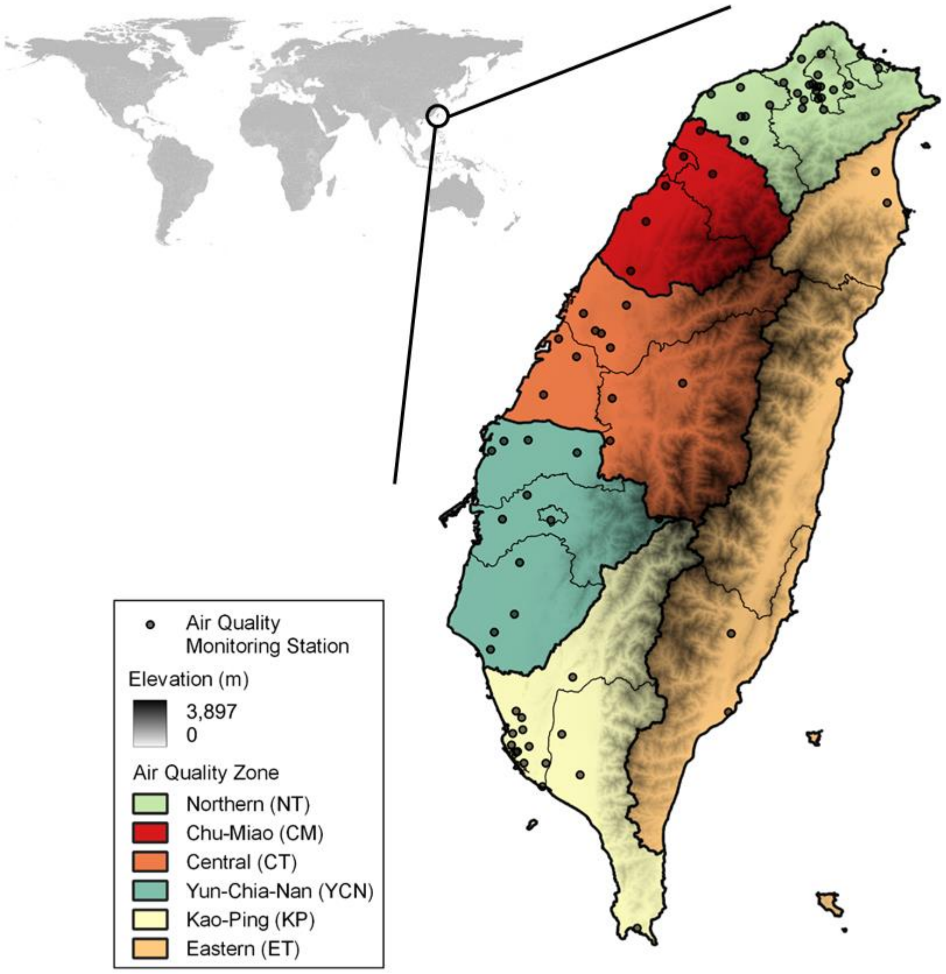


Figure 1-6. Study region of this thesis, Taiwan

Chapter 2. Spatial Heterogeneity of Exposure Risk

2.1. Abstract

The current estimations of burden of disease (BD) of $PM_{2.5}$ exposure is still potentially biased by two factors: ignorance of heterogeneous vulnerabilities at diverse urbanization levels and reliance on the risk estimates from existing literature, usually from different locations. Our objectives are (1) to build up a data fusion framework to estimate the burden of $PM_{2.5}$ exposure while evaluating local risks simultaneously and (2) to quantify their spatial heterogeneity, relationship to land-use characteristics, and derived uncertainties when calculating the disease burdens. The feature of this study is applying six local databases to extract $PM_{2.5}$ exposure risk and the BD information, including the risks of death, cardiovascular disease (CVD), and respiratory disease (RD), and their spatial heterogeneities through our data fusion framework. We applied the developed framework to Tainan City in Taiwan as a use case estimated the risks by using 2006-2016 emergency department visit data, air quality monitoring data, and land-use characteristics and further estimated the BD caused by daily $PM_{2.5}$ exposure in 2013. Our results found that the risks of CVD and RD in highly urbanized areas and death in rural areas could reach 1.20-1.57 times higher than average. Furthermore, we performed a sensitivity analysis to assess the uncertainty of BD estimations from utilizing different data sources, and the results showed that the uncertainty of the BD estimations could be contributed by different $PM_{2.5}$ exposure data (20-32%) and risk values (0-86%), especially for highly urbanized areas. In conclusion, our approach for estimating BD based on local databases has the potential to be

generalized to the developing and overpopulated countries and to support local air quality and health management plans.

2.2. Introduction

The associations between exposure to ambient $PM_{2.5}$ and increased risks of adverse health effects such as premature death, cardiovascular disease (CVD), and respiratory disease (RD) has been intensively investigated for decades [2,3,25]. The burden of disease (BD) is a commonly used assessment tool to quantify the impact of ambient $PM_{2.5}$ exposure to human health and provide references for air quality and health management. For example, the Global Burden of Disease (GBD) estimated that about 2.94 million deaths in 2017 could be attributed to particulate matter pollution [20].

Technically, the BD can be estimated by either concentration-response functions (CRFs) [48–50] or integrated exposure-response functions (IERs) [24,51]. While the applied risks in CRFs tend to be constant [48], IERs parameterize the dependence of risks on $PM_{2.5}$ concentration from meta-analysis of epidemiological studies [51]. Both of algorithms used literature-based RRs and modeled/monitored $PM_{2.5}$ concentrations to estimate disease burdens, but the BD estimations could still remain biased due to the algorithm assumptions.

One potential concern is that most studies employed risk values from other literature or reports to estimate the burden without assessing the representativeness of the study subjects or sampling bias of the reference, which could limit the application of the risk values or cause bias if the risk values were inappropriately applied. Another concern is that the spatial heterogeneity of risks among urban and rural areas were overlooked during the BD calculation. Most BD estimations at the country or regional level neglected the potential uncertainties derived from different vulnerabilities associated with living in diverse urbanization levels, for which differences in risk were related to neighboring land-use patterns and individual activity patterns [6,7,29,30] and have been identified in previous studies [5,31]. Among areas with diverse urbanization levels, $PM_{2.5}$ mass concentration and compositions adjacent to emission sources such as main roads and industrial complexes have higher $PM_{2.5}$ toxicity contributed by diesel particles, heavy metals, and oxidative potentials (Hao et al., 2020; Liu et al., 2017; Targino et al., 2016). In other words, $PM_{2.5}$ toxicity adjacent to different emission sources and land-use patterns may cause heterogeneous exposure risks to populations both near and far. Failure to consider the spatial heterogeneity of risks among urban and rural areas could lead to a potential bias of the BD estimations, and its uncertainty should be quantified.

Moreover, the data sources of $PM_{2.5}$ exposure concentration varied within studies and also contributed to the uncertainty of BD estimations. The straightforward method to estimate $PM_{2.5}$ exposure is to use measurements from the closest air quality monitoring sites, but using monitoring data alone to represent the whole region could bias short-term $PM_{2.5}$ exposure for the distant areas [31]. In addition, using more sophisticated methods such as satellite data or air quality modeling data, which need further processing before application, can identify different

$PM_{2.5}$ concentrations in urban and rural areas, but failing to validate modeled results with observations in most studies can mean overlooking important deviations from reality and increasing uncertainties of BD estimation. Even though the difference of $PM_{2.5}$ exposure estimations between monitoring data, air quality modeled data, and satellite data had been identified [5,31], the derived uncertainty for the BD calculations among urban and rural areas was still not further investigated. Nonetheless, except for the heterogeneous risks in urban and rural areas, the uncertainties between these $PM_{2.5}$ exposure assessment methods for the BD calculations need to be quantified as well.

This study applied the developed framework to Tainan City (Figure 2-2) in Taiwan as a use case, estimated the risks of death, CVD and RD by using 2006-2016 emergency department (ED) visit data, air quality monitoring data, and land-use characteristics, and further estimated the BD caused by daily $PM_{2.5}$ exposure in 2013. Land use characteristics of areas neighboring subjects' typical living activities were defined by neighboring $PM_{2.5}$ emission and Heterogeneity of Land-Use Living (HLUL) patterns. Our objectives are (1) to build up a data fusion framework to estimate the disease burden of daily $PM_{2.5}$ exposure while evaluating local risks and their spatial heterogeneities with hospital ED visit, land-use, emission, monitoring, modeling, and population data and (2) to quantify the spatial heterogeneity of risks and its relationship with land-use characteristics and derived uncertainty during BD calculation. This study systematically focused on the spatial heterogeneity of health risks at grid-level scale (1 km × 1 km) and quantified method uncertainties of the BD estimations.

2.3. Methodology

2.3.1. Data collection

(1) Hospital ED Visit Data

Study subjects were extracted from the hourly hospital ED visit database (2006-2016) maintained by Chi-Mei Hospital, a regional medical center located in Tainan City (Figure 2-2(a)). Patients with CVD (N=12,524), RD (N=18,891), and non-accidental death (N=37,846) were selected based on the International Classification of Diseases, 9th edition (ICD-9) codes (CVD: heart failure (428), cardiac dysrhythmia (426-427), cerebrovascular disease (430-437), ischemic heart disease (410-414), peripheral vascular disease (440-449); RD: chronic obstructive pulmonary disease (490-492), respiratory tract infection (464-466, 480-487)). If the subject was admitted more than once during the same month, the earlier record was used. This study was approved by the Institutional Review Board of Chi-Mei Medical Center (No. 10612-012) and was exempt from obtaining informed consent.

(2) Air Quality and Meteorological Data

Subject exposure data were collected from the nearest AQMS maintained by Taiwan EPA (Figure 2-2(a)). Hourly data of $PM_{2.5}$, PM_{10} (particulate matter ≤ 10 μm in aerodynamic diameter), NO_2, nitric oxide (NO), SO_2, carbon monoxide (CO), O_3, ambient temperature, relative humidity, and wind speed were used for analysis. The arithmetical mean

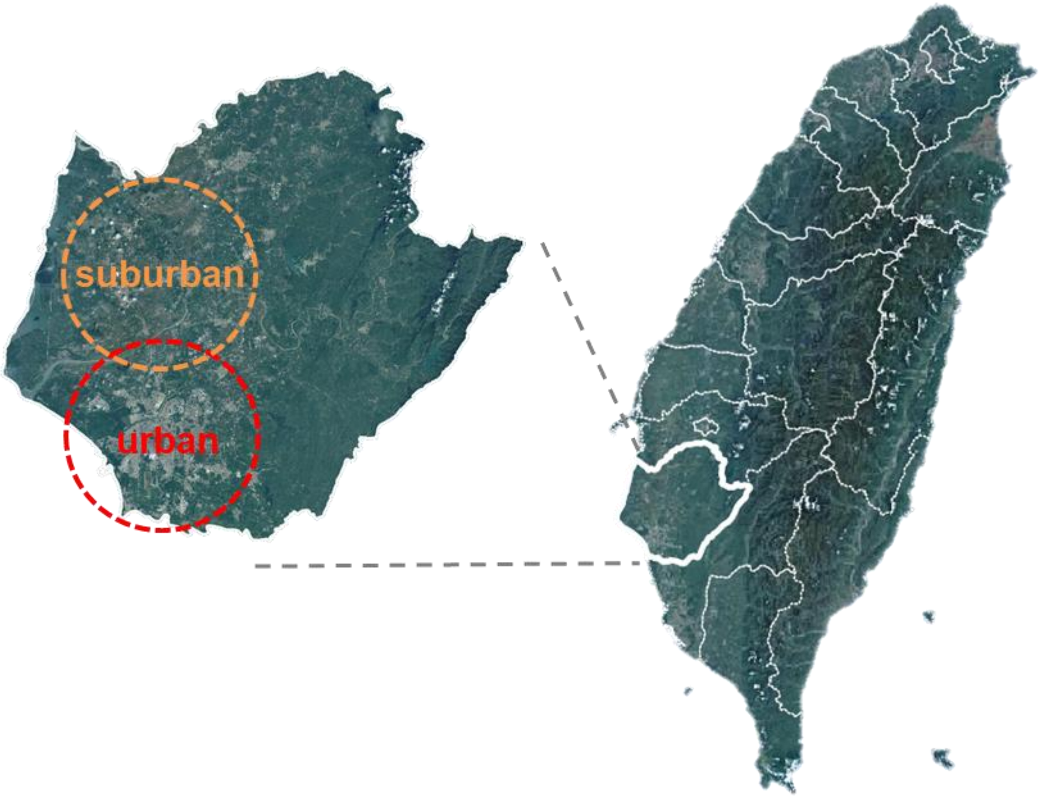


Figure 2-1. Location of Tainan City and its urban and suburban area

of the 24 hourly concentrations (n=24, 1-24 hr as an abbreviation) before the visit of each subject for each abovementioned pollutant and each abovementioned meteorological variable was calculated to represent the daily exposure of residents. Subjects with exposure to 1-24 hr $PM_{2.5}$ mean concentration ≤ 25 μg/m^3 before visit were excluded based on the daily $PM_{2.5}$ standard of the WHO [28]. Detailed analysis to test exposure threshold and lag response of $PM_{2.5}$ exposure is presented in Appendix I.

(3) Land-use Data and population data

Land-use data of 2011 were acquired from National Land Survey and Mapping Center. A total of 2406 grid cells with 1 km × 1 km horizontal resolution in Tainan City was created. Considering the daily living pattern of residents, nine (k=9) land-use types were included to represent areas visited frequently by subjects. These land-use types included high-density residential area (HDRA), low-density residential area (LDRA), agriculture (including livestock farming), industrial areas, retailing sites, recreation sites (including parks, recreation centers, and gyms), schools and education institutes, road areas, and undeveloped areas (including bare lands, forest, and water bodies). Each grid was further classified into six types: urban, suburban, industrial, urban-industrial, industrial-rural and and rural area as shown in Figure 2-2(a). Basically, area of four main land-use types was calculated for each grid, including urban (including retailing sites, recreation sites and HDRA), industrial, residential (HDRA and LDRA) and rural (agricultural and other undeveloped area), and the grids and named the types of the grids by their first and second highest area. The monotype pattern was named if the second highest area is 0 or the ratio of first and second highest area is over 5, while the combination-type pattern such as urban-industrial was named if (1) the first highest area over 0.3 km^2 and the second highest area over 0.2 km^2 or (2) the ratio of first and second highest area is over 1. The “suburban” replaced the “urban-rural” pattern, and the “urban” and “rural” replaced the “urban-residential” and “rural-residential” pattern respectively.

Population data of 2013 were obtained from the Department of Household Registration within the Ministry of the Interior and further spatially distributed by total residential area (HDRA and LDRA) of each grid cell.

(4) Emission Data

Primary $PM_{2.5}$ emission data from Taiwan Emission Data System (TEDS) version 9.0 including industrial, mobile, area, and natural sources were utilized to evaluate the potential exposure of subjects to neighboring $PM_{2.5}$ emissions. $PM_{2.5}$ emission within a 1-km radius of the center of each grid cell was further summed up for analysis (Figure 2-2 (b)).

(5) Quantification of neighboring land-use characteristics

Land use characteristics of areas neighboring subjects’ typical living activities were defined by neighboring $PM_{2.5}$ emission and Heterogeneity of Land-Use Living (HLUL) patterns. The reason to use neighboring emission is that it can represent the level of primary $PM_{2.5}$ to which subjects are potentially exposed, and it does not need complicated and time-consuming modeling procedure to simulate ambient $PM_{2.5}$. In addition, HLUL patterns reflect the daily activities and exposure patterns which can enhance physical health by daily activities but also increase frequency and duration of $PM_{2.5}$ exposure. The levels of HLUL patterns were quantified by Mean Entropy Index (MEI) for all grid cells. MEI, adapted from the Entropy Index (EI) [55–57], can represent the potential mobility of residents in an area. MEI varies between 0 and 1, where a value of 1 represents the highest diversity and heterogeneity in land-use, and considers the land-use types of neighboring grid cells by using the following equation [56].

Equation 1

$$\mathbf{MEI} = -\sum_{\mathbf{x=1}}^{\mathbf{n}} \frac{\sum_{\mathbf{y=1}}^{\mathbf{k}} \frac{[\mathbf{A_{xy}} \cdot \mathbf{P_{xy}} \cdot \mathbf{ln}(\mathbf{P_{xy}})]}{\mathbf{ln}(\mathbf{k})}}{\mathbf{n}}$$

where n is the number of surrounding grid cells (n=8) of the estimated grid cell. k is the number of the total used land-use types (k=9). $\mathrm{A_{xy}}$ is the ratio of selected or surrounding grid cell to the area within a 1-km radius of each selected grid cell. $\mathrm{P_{xy}}$ is the proportion of land-use type y in the selected and surrounding xth grid cell.

2.3.2. Developed framework

Our data fusion framework is illustrated in Figure 2-3. The number of ED visits was used to evaluate the BD, and the ED visits were calculated from short-term $PM_{2.5}$ exposure risk, daily $PM_{2.5}$ concentration, and population data [58]. First, the hospital ED visit data, air quality monitoring data, emission data, and land-use data were fused by case-

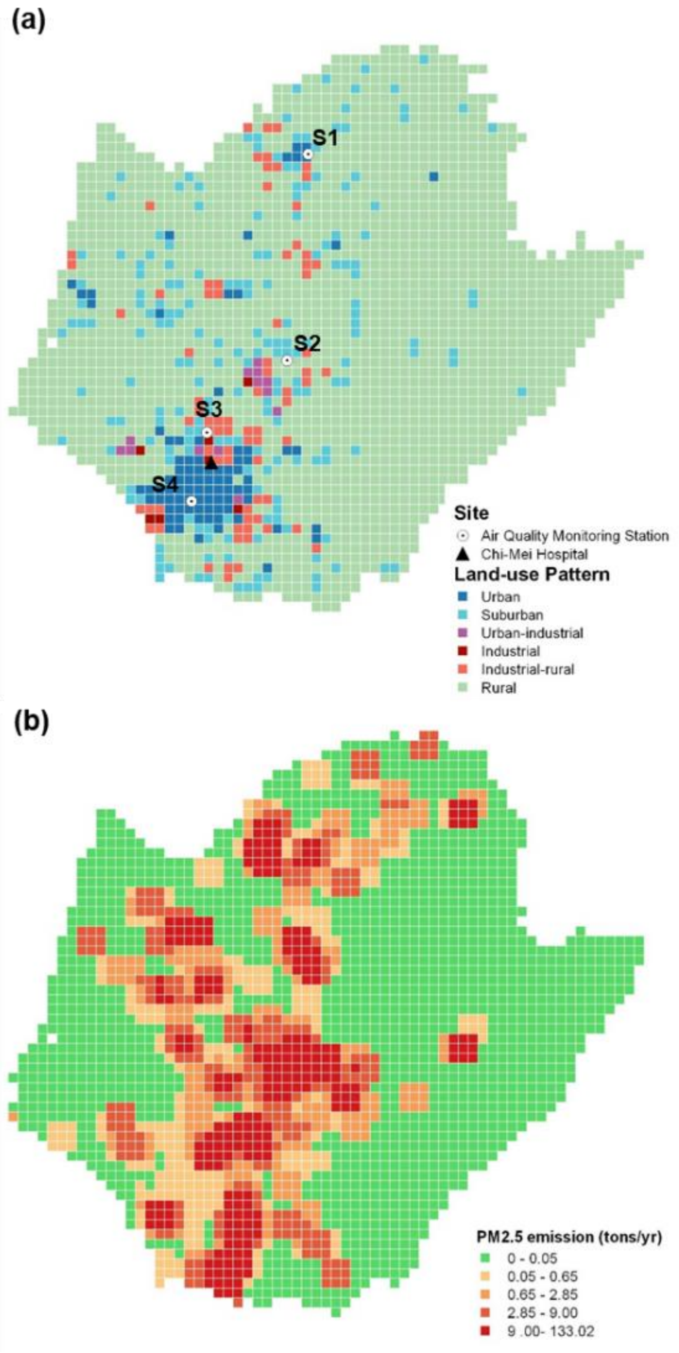


Figure 2-2. Map of Tainan City with 1 km × 1 km resolution for (a) land-use pattern (b) neighboring $PM_{2.5}$ emission (tons/year).

crossover study design and stratified analysis (Equation 2 and Equation 3) to retrieve short-term (1-24 hr) $PM_{2.5}$ exposure risk and its spatial

heterogeneity. Technically, the overall short-term $PM_{2.5}$ exposure risk was first assessed by case-crossover study design which involved hospital ED visit and air quality monitoring data, and then stratified analysis was executed to assess the heterogeneous risks for subjects living in different land-use characteristics. Second, we used daily $PM_{2.5}$ concentration data in 2013 as an example to illustrate the process to build up the BD map of CVD, RD, and death in 2013. The daily $PM_{2.5}$ concentration data were extracted from the fused estimations of modeled results and monitoring data by MMF approach [59]. The modeled results were firstly simulated by CMAQ [32] modeling and then fused with monitoring data from AQMS by applying the ratio of modeled $PM_{2.5}$ to observed $PM_{2.5}$ obtained from the nearest site (Equation 4). Overall, all parameters ($PM_{2.5}$ exposure risk, daily $PM_{2.5}$ concentration in 2013, and population) for calculating ED visits were resolved with 1 km × 1 km resolution and applied to build up the BD map in 2013 through CRFs (Equation 5). Also, a sensitivity analysis was conducted to quantify the uncertainties of the BD estimations under different scenarios.

(1) Evaluating $PM_{2.5}$ exposure risks and their spatial heterogeneity

We employed a case-crossover study design to investigate the relationship between 1-24 hr $PM_{2.5}$ exposure before the ED visit and health outcomes [60–62]. Detailed descriptions can be found in our previous publication [63]. Briefly, for each case, exposure before the visit

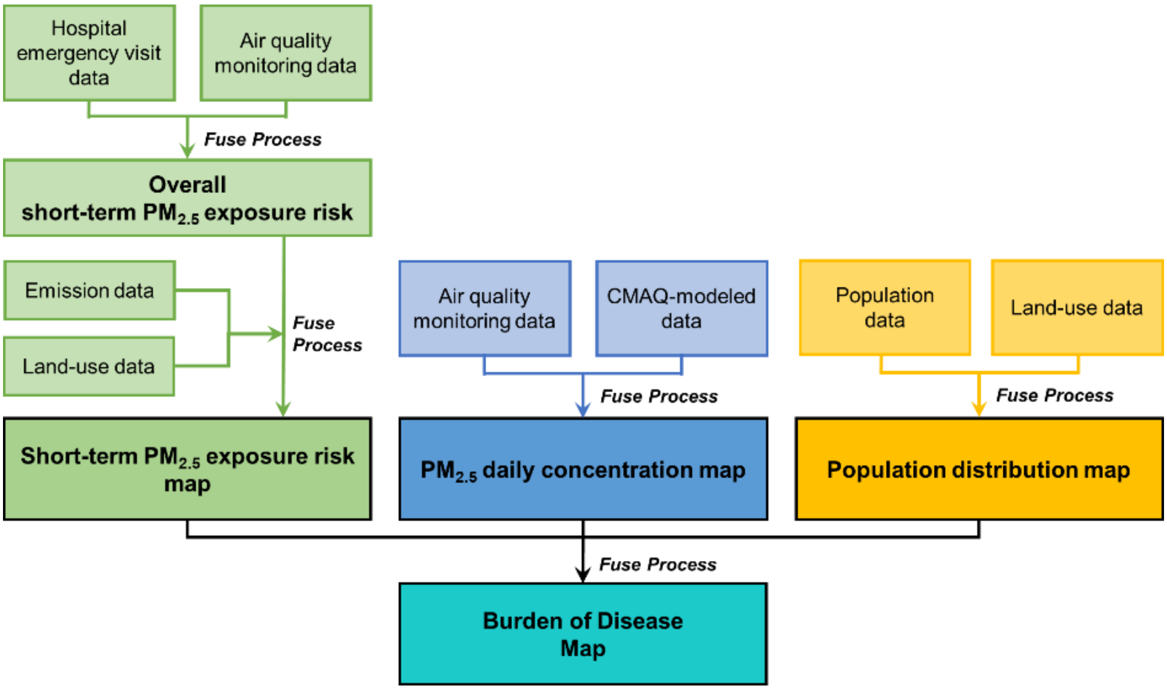

Hospital emergency visit data
Air quality monitoring data
Fuse Process
Overall short-term PM2.5 exposure risk
Emission data
Land-use data
Fuse Process
Short-term PM2.5 exposure risk map
Air quality monitoring data
CMAQ-modeled data
Fuse Process
PM2.5 daily concentration map
Population data
Land-use data
Fuse Process
Population distribution map
Fuse Process
Burden of Disease Map

Figure 2-3. Data structure of the first study

is compared with exposure at the other control periods, and the controls were selected from the same time of ED visits on the other days, on the same day of the week in the same month and year [60–62,64]. Conditional logistic regressions were conducted to estimate adjusted odds ratios (ORs) and 95% confidence intervals (CIs) for the relationship between PM2.5 exposure and health outcomes. Our multi-pollutant model was described below:

Equation 2

$$\begin{aligned}\mathrm{logit}\,\mathrm{P}(\mathrm{Y_i}) &= \ln\left(\widehat{\mathrm{OR}}_\mathrm{i}\right) \\ &= \beta_0 + \beta_1 \mathrm{PM}_{2.5_\mathrm{i}} + \beta_2 \mathrm{T_i} + \beta_3 \mathrm{RH_i} + \beta_4 \mathrm{WS_i} + \beta_5 \mathrm{P_{i1}} + \cdots + \beta_\mathrm{p} \mathrm{P_{ip}}\end{aligned}$$

where $\mathrm{P}(\mathrm{Y_i})$ is the natural logarithm of the OR for case subjects i compared with individual controls, and β_0 is the intercept. All predicting variables use 1-24 hr mean before the visit of subjects. For example, $\mathrm{PM}_{2.5_\mathrm{i}}$ is 1-24 hr PM2.5 concentration before the visit of subjects i. $\mathrm{PM}_{2.5_{\mathrm{i_i}}}$, $\mathrm{T_i}$ (temperature), $\mathrm{RH_i}$ (relative humidity), and $\mathrm{WS_i}$ (wind speed) were included as fixed variables in the model. Linear correlation between 1-24 hr PM2.5 exposure and health outcomes were assumed based on previous findings (Linares and Díaz, 2010; Yorifuji et al., 2014a, 2014b). Natural cubic splines with three degrees of freedom (df) were used in all models to adjust the potential time-variant confounders including $\mathrm{T_i}$, $\mathrm{RH_i}$, and $\mathrm{WS_i}$ [60,64,66]. $\mathrm{P_{i1}} \cdots \mathrm{P_{ip}}$ are the selected pollutants p which served as adjusting variables (covariates) and were chosen by stepwise selection. The significant level (p-value) for stepwise selection to keep or discard the variable was 0.05. For each health outcome, the final model was chosen to yield the minimum Akaike information criteria (AIC) statistic, and the collinearity of included variables was assessed by the variance inflation factor (VIF) (Table 2-1).

In advance, we performed the stratified analysis [67] by applying the same model (Equation 2) to the subjects living in the different categories of land-use characteristics to assess the spatial heterogeneity of $PM_{2.5}$ exposure risk. First, we equally categorized the subjects of death, CVD, and RD by their indices (MEI and neighboring $PM_{2.5}$ emission, respectively) with low (< 33rd percentile), medium (33rd - 66th percentile) and high (> 66th percentile) level according to their residential address. Second, for each health outcome, multi-pollutant modeling (Equation 2) was first conducted to obtain the ORs for overall subjects, and the same model was then applied to the subjects of the low-, medium- and high-level groups for MEI and neighboring $PM_{2.5}$ emission, respectively. Third, the subjects were next grouped by MEI and $PM_{2.5}$ emission categories to assess the interaction of MEI and neighboring $PM_{2.5}$ emission to the health outcomes, and a total of nine groups (three by three categories) were assessed. The same model (Equation 2) for overall subjects was also applied for each group again, and the group-specific risk could be obtained. For each group, the specific risk was calculated as:

Equation 3

$$\widehat{\mathrm{OR}}_{\mathrm{i,g}} = \exp\Big(\beta_0 + \beta_1 \mathrm{PM}_{2.5_{\mathrm{i,g}}} + \beta_2 \mathrm{T}_{\mathrm{i,g}} + \beta_3 \mathrm{RH}_{\mathrm{i,g}} + \beta_4 \mathrm{WS}_{\mathrm{i,g}} + \beta_5 \mathrm{P}_{\mathrm{i1,g}} + \cdots$$
$$+ \beta_{\mathrm{p}} \mathrm{P}_{\mathrm{ip,g}}\Big)$$

where $\widehat{\mathrm{OR}}_{\mathrm{i,g}}$ was the risk of a group g from one to nine, and the other included parameters were same as the Equation 2. By matching the group-specific risks with grid cells, the risk map can be built up.

(2) Simulating PM2.5 concentration

The daily PM2.5 concentration data in 2013 were modeled by CMAQ with 3 km × 3 km resolution and adjusted by using the monitoring data. The CMAQ modeling system was developed by the U.S. EPA and user's community. CMAQ model can provide temporally and spatially varying air pollutant concentrations by calculating physical and chemical interactions between pollutants and meteorological factors in the atmosphere [32]. The modeled daily PM2.5 concentrations were further fused with monitoring data by the MMF method. For each grid cell at each day, the modeled daily PM2.5 means were adjusted by the ratio of modeled PM2.5 to observed PM2.5 obtained from the nearest site [59], which was calculated as:

Equation 4

$$\mathrm{Fused\ \ PM_{2.5}}_{\mathrm{d,g}} = \frac{\mathrm{Obseved\ \ PM_{2.5}}_{\mathrm{d,s}}}{\mathrm{Modeled\ \ PM_{2.5}}_{\mathrm{d,s}}} \cdot \mathrm{Modeled\ \ PM_{2.5}}_{\mathrm{d,g}}$$

where $\mathrm{Fused\ PM_{2.5}}_{\mathrm{d,g}}$ and $\mathrm{Modeled\ PM_{2.5}}_{\mathrm{d,g}}$ is CMAQ-fused and CMAQ-modeled PM2.5 in grid cell g at the dth day, respectively, and $\mathrm{Obseved\ PM_{2.5}}_{\mathrm{d,s}}$ and $\mathrm{Modeled\ PM_{2.5}}_{\mathrm{d,s}}$ is observed and CMAQ-modeled PM2.5 at the dth day for the monitoring site s which is closest to the grid cell g. The 3 km × 3 km CMAQ-fused results were integrated with 1 km × 1 km population data, and each 1 km × 1 km grid cell included the closest CMAQ-fused data to calculate the number of daily excess ED visits or deaths for each grid cell. Because illustrating the different vulnerabilities in urban and rural areas and derived uncertainties from different sources is our objective, we only applied the basic MMF to fuse modeled data and monitoring data.

(3) Estimating the BD of $PM_{2.5}$ exposure

The BD of $PM_{2.5}$ exposure was calculated by CRFs, which can quantify the increased ED visits due to daily $PM_{2.5}$ exposure in 2013 [58]. The number of ED visits for each grid cell was calculated by the following equation:

Equation 5

$$\mathrm{Y} = \mathrm{E}_0 \cdot \mathrm{P} \cdot \left(1 - \mathrm{e}^{-\beta \cdot (\mathrm{C} - \mathrm{C}_0)}\right) \cdot \mathrm{A}$$

where Y is the number of daily excess ED visits or deaths caused by daily $PM_{2.5}$ exposure. E_0 is the actual morbidity or mortality rate. P is the population of each grid cell. The coefficient β is derived from RR, and RR is approximated by using OR obtained from Equation 2 or Equation 3 [64]. A is a scalar of 1/365 to convert the annual rate to daily rate. C_0 is the threshold concentration set as the WHO $PM_{2.5}$ daily standard of 25 μg/m^3. C is the daily $PM_{2.5}$ concentration in each grid cell.

2.3.3. Uncertainty analysis

We considered the uncertainty of the BD (the number of daily excess ED visits or deaths) estimations mostly originated from the value of risk coefficient (β) and daily $PM_{2.5}$ concentration (C) in Equation 5. Thus, we conducted a sensitivity analysis and compared the BD estimations from three nested scenarios, including (1) using non-local or local risk estimates to calculate the local BD. The non-local risk estimates were referred from U.S. EPA recommended values in Environmental Benefits Mapping and Analysis Program - Community Edition (BenMAP-CE, version 1.1.3) software package [68], which can be treated as the literature-based risks from other countries and regions, because U.S. EPA recommended risks were also referred from multiple well-designed studies. The local risk estimates were obtained by the conditional logistic regression (Equation 2) among overall subjects; (2) using averaged risk or heterogeneously distributed risks to calculate the burdens. The heterogeneously distributed risks were obtained by stratified analysis (Equation 3) for different land-use characteristics; (3) using monitoring data from the nearest AQMS or CMAQ-fused data to represent daily $PM_{2.5}$ exposure. SAS statistical software (SAS 9.4; SAS Institute Inc., Cary, NC, USA) was used to perform all analytical procedures.

2.4. Results and discussion

2.4.1. Descriptive analysis

Descriptive analysis of subjects (Table 2-2) showed that the number of females comprising the BD was higher than that of males. More elderly subjects visited the ED with CVD, while more young subjects visited due to RD or death. All subjects with air-quality related health conditions visited the ED more frequently during the nighttime as compared with daytime.

Table 2-1. Final models for (a) death, (b) CVD and (c) RD

(a) Death

Variables*	Unit	OR (95% CI)	p-value	VIF
$PM_{2.5}$	10 μg/m^3	1.25 (1.22, 1.27)	< 0.01	3.09
PM_{10}	10 μg/m^3	0.98 (0.97, 0.99)	< 0.01	3.39
NO_2	10 ppb	1.25 (1.17, 1.34)	< 0.01	3.88
NO	10 ppb	1.15 (1.04, 1.26)	< 0.01	1.60
CO	0.1 ppm	1.02 (1.00, 1.05)	0.04	3.00
O_3	10 ppb	1.32 (1.29, 1.35)	< 0.01	1.88

(b) Cardiovascular Disease (CVD)

Variables*	Unit	OR (95% CI)	p-value	VIF
$PM_{2.5}$	10 μg/m^3	1.27 (1.24, 1.30)	< 0.01	1.40
NO_2	10 ppb	1.35 (1.26, 1.46)	< 0.01	1.79
O_3	10 ppb	1.31 (1.26, 1.35)	< 0.01	1.46

(c) Respiratory Disease (RD)

Variable*	Unit	Odds Ratio (OR)	p-value	VIF
$PM_{2.5}$	10 μg/m^3	1.26 (1.23, 1.29)	< 0.01	1.66
NO_2	10 ppb	1.11 (1.02, 1.22)	0.02	4.87
NO	10 ppb	0.82 (0.72, 0.93)	< 0.01	1.63
SO_2	1 ppb	0.96 (0.93, 0.98)	< 0.01	1.81
CO	0.1 ppm	1.11 (1.08, 1.14)	< 0.01	3.47
O_3	10 ppb	1.22 (1.19, 1.26)	< 0.01	1.71

* Temperature, relative humidity and wind speed were included with natural cubic splines with 3 degrees of freedom (df) and not present in the tables.

During the study period, the $PM_{2.5}$ mean concentration from four AQMSs in Tainan City was 25.12±17.95 μg/m^3 (Table 2-3). For HLUL patterns, MEI was normally distributed with a mean ± standard deviation of 0.350±0.136 (unitless) (Table 2-4 and Figure 2-4). The spatial distribution of MEI in Figure 2-5 showed that the most intensely heterogeneous areas (darker color) are located in the southern city, which consists of a mix of urban, suburban, and industrial-rural land-use types. Areas adjacent to the main roads also showed high heterogeneity. In contrast, the eastern areas of the city were less heterogeneous (lighter color) due to their mountainous terrain.

2.4.2. Discussion of short-term $PM_{2.5}$ exposure risks

The odds ratios (ORs) of non-accidental death, CVD, and RD for 1-24 hr $PM_{2.5}$ exposure before the visit in this study were 1.25, 1.27, and 1.26, respectively, which are higher than those of previous studies conducted in the other developed or developing countries/regions (Table 2-5 - Table 2-7). Since the developed framework utilized local databases, it is not surprising that there is a potential difference between local risk estimations and risks from other regions or counties due to different environmental, demographic, and societal characteristics. This significant difference also implies the potential risk of using non-local risk values such as U.S. EPA recommended or literature-based risks for BD calculation.

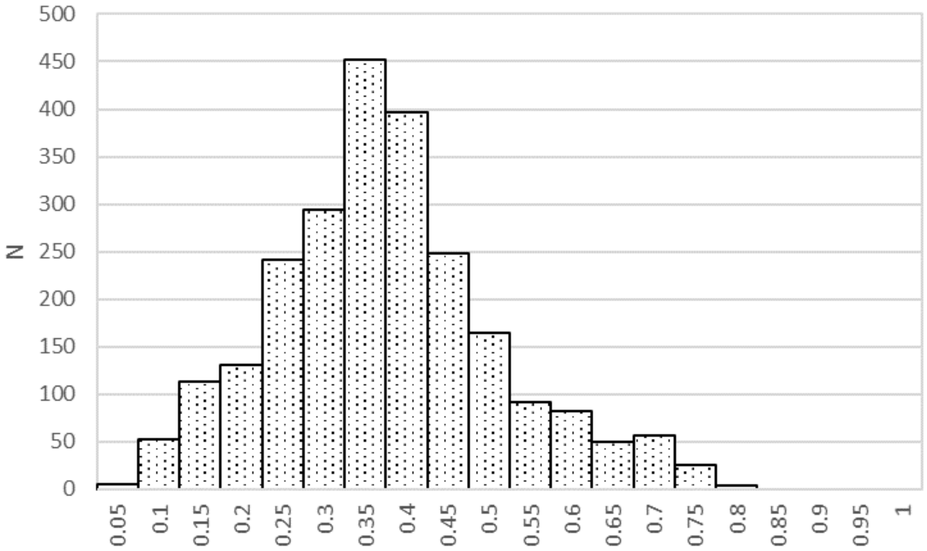


Figure 2-4. Distribution of mean entropy index (MEI) for all grids (n=2,406) in Tainan City

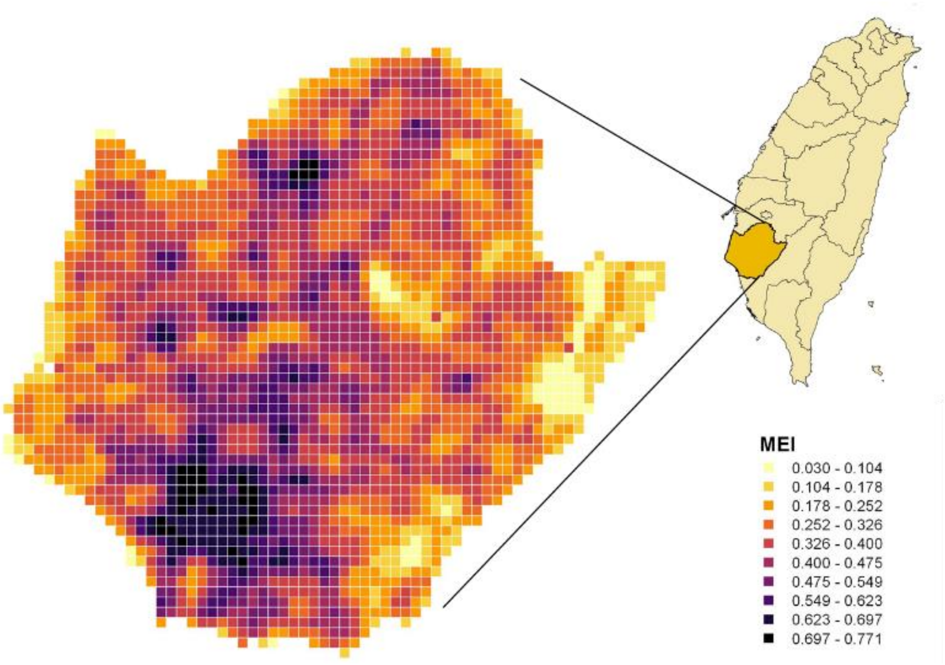


Figure 2-5. MEI map of Tainan City with 1 km × 1 km resolution

Table 2-2. Characteristics of ED visits in Chi-Mei Hospital of death, CVD and RD for the subjects residing in Tainan City, Taiwan, 2006-2016 (n=69,261)

	N (%)		
	Death	**CVD**	**RD**
Total	37,846 (100.0%)	12,524 (100.0%)	18,891 (100.0%)
Gender			
Male	17,869 (47.2%)	4,894 (39.1%)	9,047 (47.9%)
Female	19,977 (52.8%)	7,630 (60.9%)	9,844 (52.1%)
Age			
Youngers (Age<65)	23,436 (61.9%)	5,633 (45.0%)	10,901 (57.7%)
Elders (Age≥65)	14,410 (38.1%)	6,891 (55.0%)	7,990 (42.3%)
Onset time			
Daytime (8 am-7 pm)	13,073 (34.5%)	4,255 (34.0%)	8,352 (44.2%)
Nighttime (8 pm-7 am)	24,773 (65.5%)	8,269 (66.0%)	10,539 (55.8%)

Table 2-3. Statistics of air pollutants during 2006-2016 in Tainan City, Taiwan

Variable	**Mean ± SD**	**Percentile**		
		25th	**50th**	**75th**
Temperature (°C)	24.63 ± 5.20	21.00	25.83	28.50
Relative humidity (%)	75.61 ± 10.77	68.00	77.48	84.00
Wind speed (m/s)	2.33 ± 1.14	1.45	2.20	3.04
Rainfall (mm)	0.17 ± 1.42	0.00	0.00	0.00
PM_{10} (μg/m^3)	68.60 ± 41.10	39.00	60.00	89.00
$PM_{2.5}$ (μg/m^3)	25.12 ± 17.95	11.23	21.48	34.28
NO_2 (ppb)	14.40 ± 7.55	8.70	13.00	19.00
SO_2 (ppb)	3.68 ± 1.72	2.45	3.35	4.50
O_3 (ppb)	30.31 ± 21.97	13.70	24.60	41.70
NO (ppb)	3.15 ± 3.80	1.30	2.05	3.30
CO (ppm)	0.39 ± 0.17	0.27	0.37	0.49

Table 2-4. Statistics of heterogeneity indices of land-use pattern, MEI and neighboring $PM_{2.5}$ emission (tons/year) for all grids and subjects in Tainan City

	Mean	**STD**	**Median**	**33rd Percentile**	**66th Percentile**	**Minimum**	**Maximum**
All Grids (n=2406)							
MEI	0.350	0.136	0.340	0.295	0.389	0.030	0.771
$PM_{2.5}$ Emission	3.869	13.238	0.001	0.001	0.316	0.001	133.020
Death (n=37846)							
MEI	0.596	0.121	0.636	0.566	0.679	0.058	0.771
$PM_{2.5}$ Emission	14.420	24.275	2.571	0.440	8.314	0.001	130.035
CVD (n=12524)							
MEI	0.593	0.124	0.636	0.559	0.679	0.058	0.771
$PM_{2.5}$ Emission	13.268	22.999	2.475	0.439	7.734	0.001	129.933
RD (n=18891)							
MEI	0.603	0.117	0.642	0.573	0.684	0.058	0.771
$PM_{2.5}$ Emission	16.056	25.483	2.850	0.794	9.827	0.001	130.035

One explanation for our higher OR estimations could be that the composition of $PM_{2.5}$ in study region were more toxic than that of $PM_{2.5}$ measured in other countries or regions. For example, long-range transboundary air pollution transported from mainland China, which contains higher hazardous composition of $PM_{2.5}$ and by-products of combustion (e.g. sulfate), exacerbates the health impact [69,70]. The risk of exposure to $PM_{2.5}$ could be also amplified by the coexistence of NO_2 and O_3 in urban areas [67]. In this study, the significant correlation between daily $PM_{2.5}$ and daily NO_2 (Pearson's r=0.57, p-value<0.01) and O_3 (Pearson's r=0.30, p-value<0.01) magnified the health impact of $PM_{2.5}$ exposure. In addition, selection bias could also have contributed to the higher estimations. Because the subjects were selected from cases in the hospital and used as controls as well, the higher risk may have been due to subjects having been relatively more vulnerable to $PM_{2.5}$ exposure compared with the general population. Furthermore, the full-coverage and complete national health insurance among all residents and lower medical expense in Taiwan may also increase the willingness of people to visit the ED even they have non-fatal diseases or symptoms.

To exclude the possible bias from modeling, we excluded the possibility of modeling bias from multiple controls used for one case. In this study, the controls were the subjects themselves but on the other days, on the same day of the week in the same month and year. That is, in most cases, 3 controls were used for each case because the controls were selected from the same day of the week in that month. Using 1 control and 3 controls for input were both examined in this study, and both estimated

ORs were close. Meanwhile, one study pointed out that overestimation could occur when using longer lagged averages [71], but in our sensitivity analysis, a similar level of ORs was obtained for shorter (1 hr) and longer lagged period (1-48 hr and 1-96 hr).

2.4.3. Heterogeneous risks associated with HLUL patterns and $PM_{2.5}$ emission

Estimated risks of 1-24 hr $PM_{2.5}$ exposure (per 10 μg/m^3 increase) for different levels of HLUL patterns are presented in Table 2-8 and Figure 2-6(a). The average risk of CVD (OR=1.27, 95% CI: 1.24-1.30) is the highest compared with RD (OR=1.26, 95% CI: 1.23-1.29) and non-accidental death (OR=1.25, 95% CI: 1.22-1.27) (Table 5).

The risks of $PM_{2.5}$ exposure also varied with the levels of HLUL patterns. For CVD and RD, the risk increased with the increase in HLUL patterns, and high HLUL patterns increased risk up to 59% (OR=1.59, 95% CI: 1.46-1.73) and 36% (OR=1.36, 95% CI: 1.29-1.45), respectively. On the contrary, low-level HLUL patterns was correlated with a higher risk of death (OR=1.31, 95% CI: 1.25-1.37). Regarding neighboring $PM_{2.5}$ emission (Table 2-8 and Figure 2-6(b)), subjects in proximity to the medium level of $PM_{2.5}$ emission had the highest risk for all selected BD outcomes. The highest risks for CVD, death, and RD were identified in subjects adjacent to the medium-level $PM_{2.5}$ emission, for which ORs were 1.52 (95% CI: 1.40-1.66), 1.51 (95% CI: 1.41-1.61) and 1.47 (95% CI: 1.38-1.57), respectively. Furthermore, the reduced risk of death was observed for subjects in proximity to high-level $PM_{2.5}$ emission (OR=0.95, 95% CI: 0.89-1.01), although results were not statistically significant.

Table 2-5. Collected odd ratios of CVD of PM$_{2.5}$ short-term exposure from literatures

Reference	Study design	Health Outcomes	Region	Odds Ratio
Taiwan				
This study	Case-crossover design	Cardiovascular disease	Tainan, Taiwan	1.266 (1.236-1.296)
		Heart failure		1.495 (1.088-2.054)
		Arrhythmia		1.370 (1.070-1.753)
		Cerebrovascular disease		1.216 (1.168-1.265)
		Ischemic heart disease		1.372 (1.219-1.545)
[72]	Case-crossover design	Myocardial Infarction	Taipei, Taiwan	1.089 (1.062-1.121)
[73]	Case-crossover design	Cardiac arrhythmias	Taipei, Taiwan	1.094 (1.062-1.126)
[74]	Case-crossover design	Congestive heart failure	Taipei, Taiwan	1.083 (1.056-1.110)
[75]	Case-crossover design	Ischemic heart disease	Kaohsiung, Taiwan	1.140 (1.120-1.160)
		Congestive heart failure		1.140 (1.120-1.160)
		Stroke		1.140 (1.100-1.170)
		Arrhythmias		1.140 (1.100-1.180)
[76]	Case-crossover design	Ischemic Heart disease	Taipei, Taiwan	1.089 (1.073-1.099)
[77]	Case-crossover design	Ischemic heart disease	Kaohsiung, Taiwan	1.134 (1.116-1.155)
		Stroke		1.138 (1.120-1.155)
		Congestive heart failure		1.136 (1.104-1.168)
		Arrhythmias		1.140 (1.101-1.183)
[69]	Case-crossover design	Cardiovascular disease	Taiwan	2.200 (1.220-5.080)
[78]	Case-crossover design	Hypertension	Kaohsiung, Taiwan	1.141 (1.023-1.270)
Other Asia				
[62]	Case-crossover design	Cardiovascular disease	Okayama, Japan	1.019 (1.005-1.029)
[79]	Time-series analysis	Cardiovascular disease	Japan	1.004 (0.999-1.021) *
		Cerebrovascular disease		1.002 (0.994-1.011) *
[80]	Time-series analysis	Cardiovascular disease	Japan	1.013 (1.002-1.023) *
[81]	Case-crossover design	Cardiovascular disease	Beijing, China	1.005 (1.001-1.009)
[82]	Case-crossover design	Hypertension	Beijing, China	1.084 (1.028-1.139)
[83]	Case-crossover design	Ischemic stroke	Beijing, China	1.093 (1.085-1.100)
		Hemorrhagic stroke		1.084 (1.064-1.105)
[84]	Case-crossover design	Cardiovascular disease	Beijing, China	1.015 (1.002-1.027) *
[85]	Case-crossover design	Ischemic stroke	China	1.000 (1.001-1.003) *
[86]	Case-crossover design	Myocardial Infarction	Beijing, China	1.050 (1.000-1.110)
[87]	Case-crossover design	Myocardial Infarction	Tehran, Iran	1.013 (1.002-1.024)
Europe				
[88]	Time-series analysis	Cardiovascular mortality	Barcelona, Spain	1.029 (1.014-1.044) *
[89]	Case-crossover design	Cardiovascular disease	Madrid, Spain	1.025 (1.003-1.047)
[90]	Time-series analysis	Circulatory mortality	Madrid, Spain	1.022 (1.005-1.039)
		Circulatory mortality		1.025 (1.007-1.043)
[91]	Case-crossover design	Acute coronary syndrome	Rome, Italy	1.023 (1.005-1.042) *
		Heart failure		1.024 (1.003-1.045) *
[92]	Case-crossover design	Myocardial Infarction	Belgian	1.028 (1.003-1.054)
[93]	Time-series analysis	Cardiovascular mortality	French	1.051 (1.018-1.084) *
[94]	Case-crossover design	Out-of-hospital cardiac arrests	Copenhagen, Denmark	1.107 (1.020-1.199) *
[95]	Time-series analysis	Cardiovascular disease	Netherlands	1.009 (1.001-1.018)

* OR was approximated by relative risk (RR).

Table 2-5 continued. Collected odd ratios of CVD of $PM_{2.5}$ short-term exposure from literatures

Reference	Study design	Health Outcomes	Region	Odds Ratio
Americas				
[96]	Case-crossover design	Myocardial Infarction	Washington, U.S.	1.020 (0.980-1.070)
[97]	Time-series analysis	Cardiovascular disease	New-England, U.S.	1.010 (1.007-1.005) *
		Stroke		1.002 (0.999-1.006) *
[98]	Case-crossover design	Cardiovascular disease	Mid-Atlantic, U.S.	1.008 (1.001-1.001) *
[99]	Case-crossover design	Congestive heart failure	Maryland, U.S.	1.074 (0.925-1.242)
[100]	Time-series analysis	Cardiovascular disease	Connecticut and Massachusetts, U.S.	1.018 (1.004-1.031) *
[101]	Case-crossover design	Ischemic coronary disease	Utah, U.S.	1.045 (1.011-1.080) *
[102]	Case-crossover design	Congestive Heart Failure	Texas, U.S.	1.041 (1.014-1.071)
[103]	Time-series analysis	Deep vein thrombosis	U.S.	1.006 (1.000-1.013) *
[104]	Time-series analysis	Cardiac arrests	New York, U.S.	1.060 (1.020-1.100)
	Case-crossover design			1.040 (0.990-1.080)
[105]	Bayesian hierarchical modeling	Cardiovascular disease	U.S.	1.007 (1.005-1.008) *
[106]	Case-crossover design	Myocardial Infarction	Colorado, U.S.	1.020 (0.922-1.124)
		Ischemic heart disease		1.061 (1.000-1.166)
[107]	Case-crossover design	Cardiovascular disease	Massachusetts, U.S.	1.073 (1.062-1.084)
			New Jersey, U.S.	1.030 (1.022-1.037)
			New Hampshire, U.S.	1.067 (1.032-1.104)
			New York, U.S.	1.029 (1.024-1.034)
[108]	Case-crossover design	Non-cardioembolic ischemic stroke	Ontario, Canada	1.055 (0.994-1.120) *
[67]	Case-crossover design	Myocardial Infarction	Ontario, Canada	1.164 (1.084-1.254) *
[109]	Case-crossover design	Stroke	Edmonton, Canada	1.016 (0.937-1.097)
[110]	Time-series analysis	Cerebrovascular diseases	Santiago, Chile	1.021 (1.002-1.041) *
		Heart Rhythm Disturbances		1.021 (1.002-1.042) *
Australia				
[111]	Case-crossover design	Cardiovascular disease	Australia	1.027 (1.002-1.053) *
[112]	Case-crossover design	Pneumonia + acute bronchitis	Australia and New Zealand	1.023 (1.000-1.046) *

* OR was approximated by relative risk (RR).

Table 2-6. Collected odd ratios of respiratory disease of $PM_{2.5}$ short-term exposure from literatures

Reference	Study design	Health Outcomes	Region	Odds Ratio
Taiwan				
This study	Case-crossover design	Respiratory disease	Tainan, Taiwan	1.260 (1.234-1.286)
		Acute exacerbation COPD		1.275 (1.247-1.304)
		Respiratory tract infection		1.361 (1.185-1.562)
[113]	Case-crossover design	Pneumonia	Kaohsiung, Taiwan	1.142 (1.130-1.154)
		Asthma		1.120 (1.082-1.161)
		COPD		1.132 (1.108-1.157)
[114]	Case-crossover design	Pneumonia	Taipei, Taiwan	1.078 (1.062-1.089)
[115]	Case-crossover design	Asthma	Taipei, Taiwan	1.115 (1.078-1.152)
[69]	Case-crossover design	Respiratory disease	Taiwan	1.860 (1.300-2.910)
Other Asia				
[61]	Case-crossover design	Respiratory disease	Okayama, Japan	1.024 (1.005-1.043)
[79]	Time-series analysis	Respiratory disease	Japan	1.019 (1.000-1.028) *
[116]	Time-series analysis	Asthma	China	1.055 (1.043-1.069)
[84]	Case-crossover design	Respiratory disease	Beijing, China	1.020 (1.010-1.030) *
[117]	Time-series analysis	Respiratory disease	Beijing, China	1.002 (1.001-1.003) *
		Upper respiratory tract infection		1.002 (1.000-1.004) *
		Lower respiratory tract infection		1.003 (1.001-1.005) *
		Acute exacerbation COPD		1.032 (1.014-1.049) *
[118]	Case-crossover design	Respiratory disease	Hanoi, Vietnam	1.022 (1.013-1.032)
[119]	Case-crossover design	Asthma	Zonguldak, Turkey	1.150 (0.990-1.340)
		Allergic rhinitis with asthma		1.210 (1.100-1.330)
Europe				
[89]	Case-crossover design	Respiratory disease	Madrid, Spain	1.023 (1.003-1.039)
[120]	Time-series analysis	Respiratory disease	Madrid, Spain	1.028 (1.004-1.052)
[91]	Case-crossover design	Lower respiratory tract infections	Rome, Italy	1.028 (1.005-1.052) *
[95]	Time-series analysis	Respiratory disease	Netherlands	1.012 (0.997-1.014)
America				
[121]	Case-crossover design	Asthma-related hospital admission	Washington, U.S.	1.076 (1.019-1.136)
[122]	Time-series analysis	Respiratory disease	California, U.S.	1.050 (1.030-1.070)
[97]	Time-series analysis	Respiratory disease	New-England, U.S.	1.007 (1.004-1.005) *
[98]	Case-crossover design	Respiratory disease	Mid-Atlantic, U.S.	1.002 (1.002-1.003) *
[100]	Time-series analysis	Respiratory disease	Connecticut and Massachusetts, U.S.	1.007 (1.005-1.008) *
[106]	Case-crossover design	Respiratory disease	Colorado, U.S.	1.061 (1.040-1.103)
		Asthma & Wheeze		1.145 (1.082-1.210)
[123]	Case-crossover design	Asthma	Pittsburgh, U.S.	1.036 (1.001-1.073)
[124]	Bidirectional case-crossover design	Asthma	Ontario, Canada	1.097 (1.043-1.162)
	Unidirectional case-crossover design			1.011 (0.968-1.065)
	Time-series analysis			1.000 (0.968-1.043)
Australia				
[111]	Case-crossover design	Respiratory disease	Australia	1.004 (0.976-1.032) *
[112]	Case-crossover design	Pneumonia + acute bronchitis	Australia and New Zealand	1.023 (1.000-1.046) *
				1.032 (1.001-1.063) *
		Respiratory disease		1.032 (1.013-1.051) *
				1.023 (1.009-1.036) *

* OR was approximated by relative risk (RR).

Table 2-7. Collected odd ratios of death of PM$_{2.5}$ short-term exposure from literatures

Reference	Study design	Health Outcomes	Region	Odds Ratio	
This study	Case-crossover design	Non-accidental mortality	Tainan, Taiwan	1.245 (1.219-1.273)	
[88]	Time-series analysis	Cardiovascular mortality	Barcelona, Spain	1.029 (1.014-1.044)	*
[90]	Time-series analysis	Circulatory mortality	Madrid, Spain	1.022 (1.005-1.039)	
		Circulatory mortality		1.025 (1.007-1.043)	
[125]	Case-crossover design	Mortality	Italy	1.011 (1.005-1.017)	*
[93]	Time-series analysis	Non-accidental mortality	France	1.007 (0.999-1.016)	*
		Cardiovascular mortality		1.051 (1.018-1.084)	*
[95]	Time-series analysis	All-causes mortality	Netherlands	1.009 (1.004-1.014)	
		Cardiovascular mortality		1.009 (1.001-1.018)	
		Respiratory mortality		1.012 (0.997-1.014)	

* OR was approximated by relative risk (RR).

The risk estimations from different combinations of HLUL pattern levels and $PM_{2.5}$ emission categories by applying stratified analysis is shown in Table 2-9. Generally, higher risks are observed for subjects in high HLUL pattern areas and proximity to a medium-level $PM_{2.5}$ emission, and the highest risk is observed for CVD (OR=1.99, 95% CI: 1.50-2.64), followed by death (OR=1.74, 95% CI: 1.47-2.07), and RD (OR=1.52, 95% CI: 1.32-1.74). In addition, significant reduced risks (OR<1) are also observed in death subjects with low emission/medium HLUL pattern, medium emission/low HLUL pattern, and high emission/high HLUL pattern combinations.

The increased risks of living in high HLUL patterns could be related to mixed land-use patterns. Residential areas neighboring high-emission areas like industrial areas or roads have shown higher risks of health impacts from air quality [6,7,126]. In addition, subjects close to a medium-level $PM_{2.5}$ emission suffered comparable risks which is higher than proximity to a higher-level $PM_{2.5}$ emission. One explanation could be that the majority of $PM_{2.5}$ mass concentration does not only originate from local regions. Secondary $PM_{2.5}$ transported from other up-wind polluted regions may lead to subjects in these areas underestimating their exposure and reducing their awareness of air pollution. Previous studies in Taiwan found that secondary aerosol could contribute 25-60% of total $PM_{2.5}$ mass, including 8-27% of total $PM_{2.5}$ mass transported from mainland China and coastal cities in Taiwan [127–130].

2.4.4. Risk mapping building

The risk map of death (Figure 2-7(a)) illustrates that the risk of death in rural areas was 1.40 times higher than average, while the risks of CVD and RD (Figure 2-7(b) and Figure 2-7(c)) in the most urbanized areas were 1.57 and 1.20 times higher than average. The spatial heterogeneity of risks implies that even though the residents have similar $PM_{2.5}$ exposure, those who lived in areas with diverse land-use characteristics would have different vulnerabilities to $PM_{2.5}$ exposure. Based on the identified linkage between $PM_{2.5}$ exposure risk and land-use characteristics, we considered this spatial heterogeneity of vulnerability could be associated with land-use characteristics in our study region. Also, the separately quantified risks for each health outcome have similar spatial patterns, which indicates that the correlation between land-use patterns and vulnerability could exhibit a generalizable trend. Residents in the most urbanized areas have a higher vulnerability to $PM_{2.5}$ exposure for CVD and RD; residents living along main roads and highways are also less vulnerable to death and RD. In addition, residents in rural areas are at higher risk of death due to $PM_{2.5}$ exposure. While similar previous studies have ignored the land-use related heterogeneity of $PM_{2.5}$ exposure vulnerabilities and focused more on air pollutant concentrations, our results pointed out that subject vulnerability could potentially vary with the land-use characteristics of their activity spheres and those that neighbor them, and the spatial heterogeneity of those land-use characteristics and vulnerability could affect BD estimations.

Table 2-8. Adjusted ORs and 95% CIs per 10 μg/m^3 increase of PM$_{2.5}$ for different levels of neighboring heterogeneity of HLUL pattern and neighboring PM$_{2.5}$ emission

		Death	**CVD**	**RD**
All		1.25 (1.22-1.27)	1.27 (1.24-1.30)	1.26 (1.23-1.29)
HLUL	Low	1.31 (1.25-1.37)	1.10 (1.05-1.15)	1.18 (1.13-1.24)
	Medium	1.14 (1.06-1.21)	1.30 (1.21-1.40)	1.21 (1.13-1.30)
	High	1.12 (1.05-1.19)	1.59 (1.46-1.73)	1.36 (1.29-1.45)
PM$_{2.5}$ Emission	Low	1.07 (1.02-1.12)	1.07 (1.02-1.12)	1.08 (1.03-1.13)
	Medium	1.51 (1.41-1.61)	1.52 (1.40-1.66)	1.47 (1.38-1.57)
	High	0.95 (0.89-1.01)	1.18 (1.10-1.27)	1.11 (1.04-1.18)

Table 2-9. Adjusted ORs and 95% CIs per 10 μg/m^3 increase of PM$_{2.5}$ for diverse combinations of neighboring HLUL patterns and PM$_{2.5}$ emission levels

		Death			CVD			RD		
PM$_{2.5}$ Emission		Low	Medium	High	Low	Medium	High	Low	Medium	High
N										
HLUL	Low	10,100	2,844	666	3,294	1,003	182	4,958	1,376	317
	Medium	3,130	4,753	4,330	1,056	1,487	1,516	1,377	2,295	2,447
	High	55	4,917	7,051	1	1,580	2,405	137	2,644	3,340
OR (95% CI)										
HLUL	Low	1.39 (1.31-1.49)	0.48 (0.36-0.64)	-*	1.09 (1.04-1.16)	1.40 (1.10-1.79)	-*	1.17 (1.10-1.24)	1.08 (0.84-1.38)	-*
	Medium	0.49 (0.36-0.65)	1.23 (1.03-1.47)	1.60 (1.30-1.98)	0.98 (0.69-1.38)	1.28 (1.05-1.57)	1.05 (0.84-1.31)	0.93 (0.69-1.24)	1.51 (1.22-1.86)	1.02 (0.92-1.13)
	High	-**	1.74 (1.47-2.07)	0.82 (0.73-0.91)	-**	1.99 (1.50-2.64)	1.32 (1.16-1.50)	-**	1.52 (1.32-1.74)	1.20 (1.06-1.35)

* The group was combined to medium-level PM$_{2.5}$ emission and low HLUL group because of limited sample size.

** The group was combined to medium-level PM$_{2.5}$ emission and high HLUL group because of limited sample size.

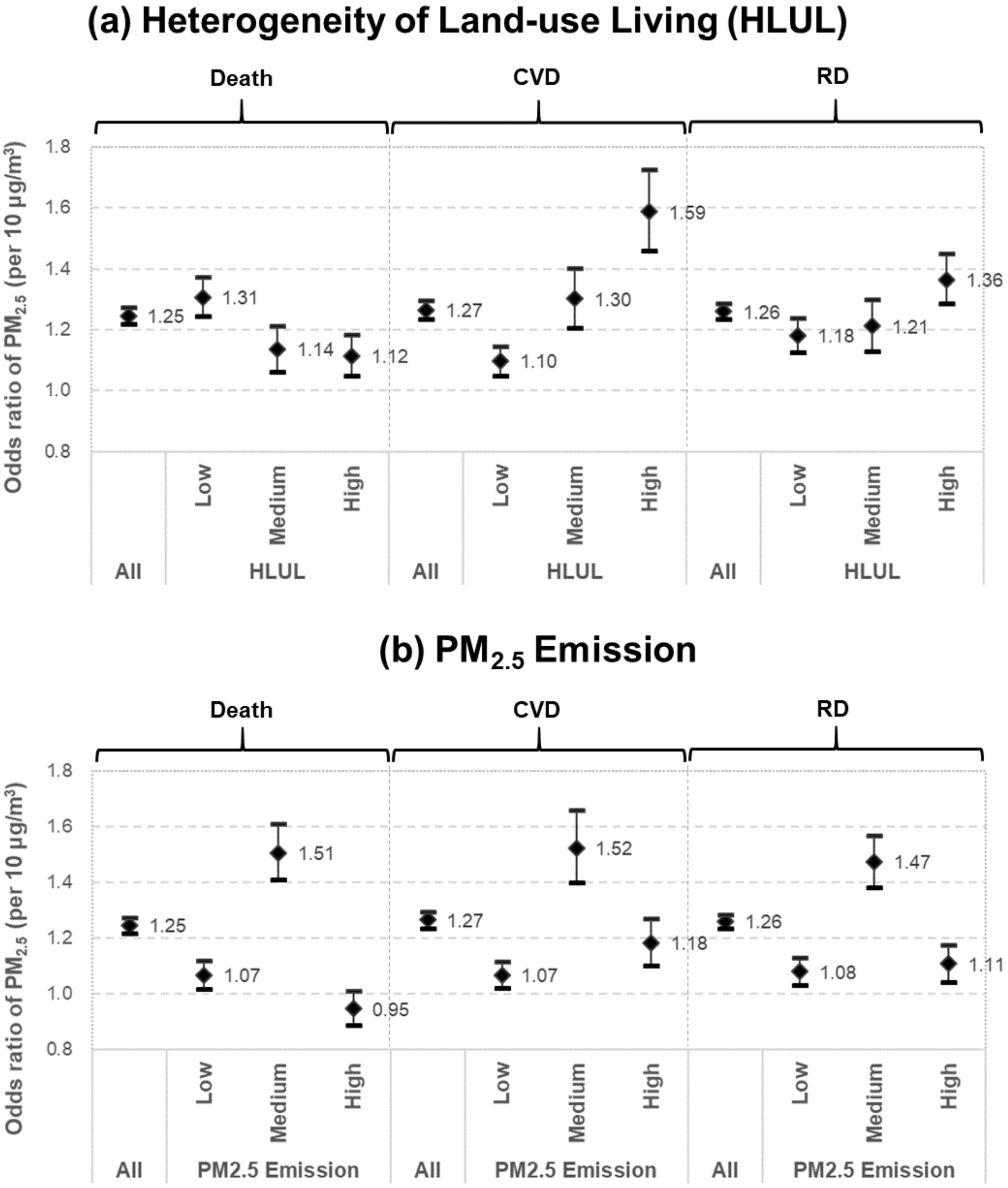


Figure 2-6. Adjusted ORs and 95% CIs per 10 μg/m^3 increase of PM$_{2.5}$ for different levels of neighboring (a) HLUL patterns and (b) PM$_{2.5}$ emission

The disparate risks among rural and urban areas might be due to the factors including human-environment interactions, individual factors, and $PM_{2.5}$ toxicity. First, concerning individual factors, one reason for the higher vulnerability in rural areas is lower accessibility of medical resources, which is attributable to longer distances to hospitals; lower abundance of medical professionals and reduced willingness to seek medical services [131,132]. Garcia et al. (2016) and Kloog et al. (2014) attributed this disparities to a “non-metropolitan penalty”, which refers to the difference in individual-level behaviors, medical accessibility, and the level of knowledge of diseases and their prevention [5,98] . Second, interaction between humans and their environment such as daily living pattern and exposure intensity, duration, and frequency could affect the level of the risk. Subjects in rural areas could have higher exposure frequency and longer exposure duration due to daily commuting and other outdoor activities [133]. Third, the difference in chemical and physical compositions between rural and urban $PM_{2.5}$ could contribute to the risk difference. Higher risks of CVD and RD in more urbanized areas could be related to higher emissions from mobile and commercial sources (e.g. restaurants, street vendors) at breathing level [134]. Additionally, it should be noted that although residing in urban areas was found to enhance the risks of CVD and RD, the risk of death in rural areas was still higher than in urban areas, which suggests that the protective factors like high transportation and medical accessibility in the urban and suburban environments could lower the risk of death, and residents in rural areas could have a lower willingness and lack medical resources to receive medical treatment in time to prevent death.

In addition, a reduced risk of $PM_{2.5}$ exposure for death was observed in suburban areas (Figure 2-7(a)), which could be associated with higher awareness to air pollution for residents in suburban areas. This finding is similar to that of Chan and Ng (2011), who speculated that vulnerable groups might be more aware of $PM_{2.5}$ events and reduce their outdoor activities [135]. Eberhardt and Pamuk (2004) also found that the health status of suburban area residents was superior to those living in both the most-rural and the most-urban areas [136].

2.4.5. disease burden uncertainty analysis

Concerning estimation uncertainties, we considered that the bias could originate from the following scenarios: (1) using non-local estimated risk for calculating the local BD, (2) using averaged risk to represent the whole study region without considering population distribution and heterogeneously distributed risks, and (3) using monitoring data to represent ambient $PM_{2.5}$ concentration among large regions including areas far from the AQMS. Thus, we designed six scenarios using three sources of risk values and two sources of $PM_{2.5}$ exposure data as shown in Table 2-10, including the estimated ED visits and the uncertainties between scenarios. Descriptive analysis of observed and CMAQ-fused $PM_{2.5}$ data are presented in Table 2-11. The validation of CMAQ-modeled results is shown in Table 2-12. Spatial distribution of the annual mean of CMAQ-fused $PM_{2.5}$ is shown in Figure 2-8. CMAQ-modeled results for all sites in this study region meet both of the performance criteria of Taiwan EPA and U.S. EPA [137].

First, risks using estimated data, based on U.S. EPA recommended values, contributed the greatest uncertainty to the results as it underestimated the local ED visits by 16-86% compared with using gridded local risks. This high uncertainty also emphasized the potential bias when using U.S. EPA recommended values or literature-based risk estimates from other countries or regions to calculate the local BD. In this study, if local gridded risk estimates were used in our study region, the estimated ED visits increased substantially by 1.19-7.29 times.

Moreover, when using monitoring data to calculate the BD, the results were lower than using CMAQ-fused data, primarily because using monitoring data could have a bias in distant areas far from the monitoring site.

Therefore, using monitoring data to represent distant areas resulted in the underestimation of ED visits by 20-32% in our study region. Although uncertainties occasionally exhibited in the monitoring data, it was still useful in the study due to higher availability and shorter time to model execution. Additionally, while using CMAQ-fused data is a more appropriate method to represent exposure more accurately for subjects in distant areas where measurements are not directly available, the modeling results still need to be validated and require more time and resources to execute. Both techniques have advantages and disadvantages, and for calculating the BD, we quantified the estimation uncertainties between these two methods to be 20-32%.

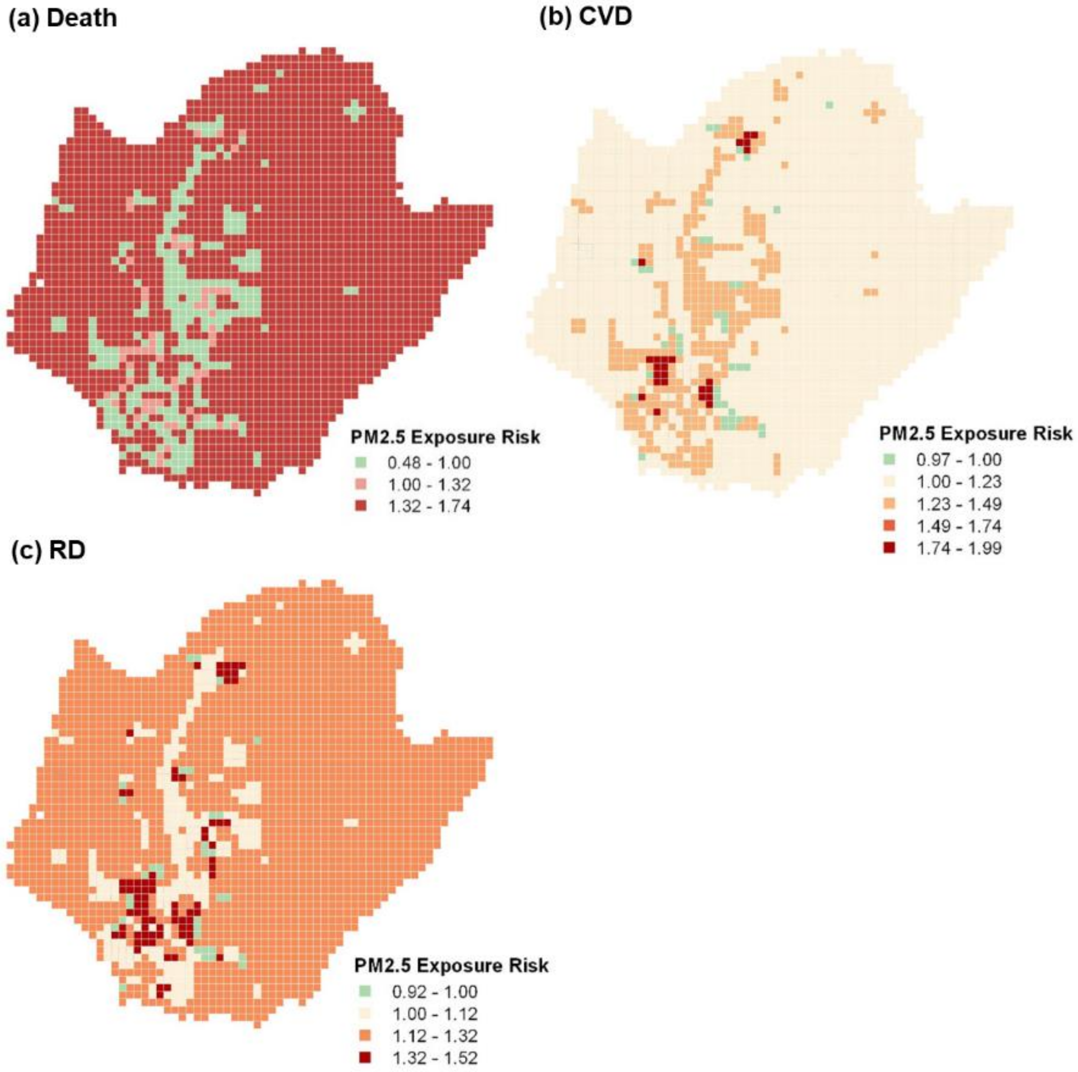


Figure 2-7. Risk map of $PM_{2.5}$ (per 10 μg/m^3 increase) for (a) death, (b) CVD and (c) RD in Tainan City.

Additionally, concerning the spatial heterogeneity of risk, when using gridded risks other than averaged risks, the number of ED visits decreases by 0-38%. This is because the disaggregated calculation considered the distribution of population and their specific risks in urban, suburban, and rural areas. Figure 2-8 shows the spatial difference between these two scenarios, where red and blue grid cells represent the overestimation and underestimation, respectively, of the averaged-risk method compared with the gridded-risk method. Using averaged risks underestimates the ED visits in urban areas for all outcomes, and overestimates risk in suburban areas for death, rural areas for CVD, and suburban and rural areas for RD.

Overall, concerning uncertainty from risk values and daily $PM_{2.5}$ concentration, using non-local risk estimates has the largest uncertainty (16-86%), followed by using monitoring data (20-32%) and applying the same averaged risks to represent the vulnerability of the population in rural and urban areas (0-38%), especially in most urbanized areas. The most biased estimation using non-local risk estimates and monitoring data underestimates the BD by 39-90% compared with using heterogeneously distributed local risks estimates and CMAQ-fused data. Thus, previous studies using nation-wide risk to estimate the BD or cost-benefit of air quality implementation could be potentially biased, because they overlooked the distribution of populations and their heterogeneous risks in areas with diverse levels of urbanization, and these risks could vary within a wide range (e.g. OR of death could be 39% higher than the average) [20,51,138]. When ignoring the spatial heterogeneity of risk, the effectiveness of air quality control implementation could be limited and remain biased.

2.5.6. Strengths of our framework

Our framework to calculate the BD has several advantages over previous calculations. First, our framework can provide cities or countries with a framework to develop their own BD estimations by using local databases, since the uncertainties arising from lack of available administrative health data in smaller regions, cities, and communities has been gradually come to be recognized [20]. For those regions or countries without local land-use or emission data, using satellite data can provide an alternative way to build up local land-use database [139], while emission data could be obtained from local government or a global emission database such as ECLIPSE that has been widely used for many international studies [140] and can be spatially distributed by population density, road lengths or other anthropogenic indices. Once regions or countries have the required databases, our developed framework can be applied in any location. Second, the risk map can identify the spatial heterogeneity of risks and identify areas with higher risks. Combined with the BD map, the maps can facilitate the allocation of public and medical resources by local governments to affected areas and reduce potential health impacts more effectively.

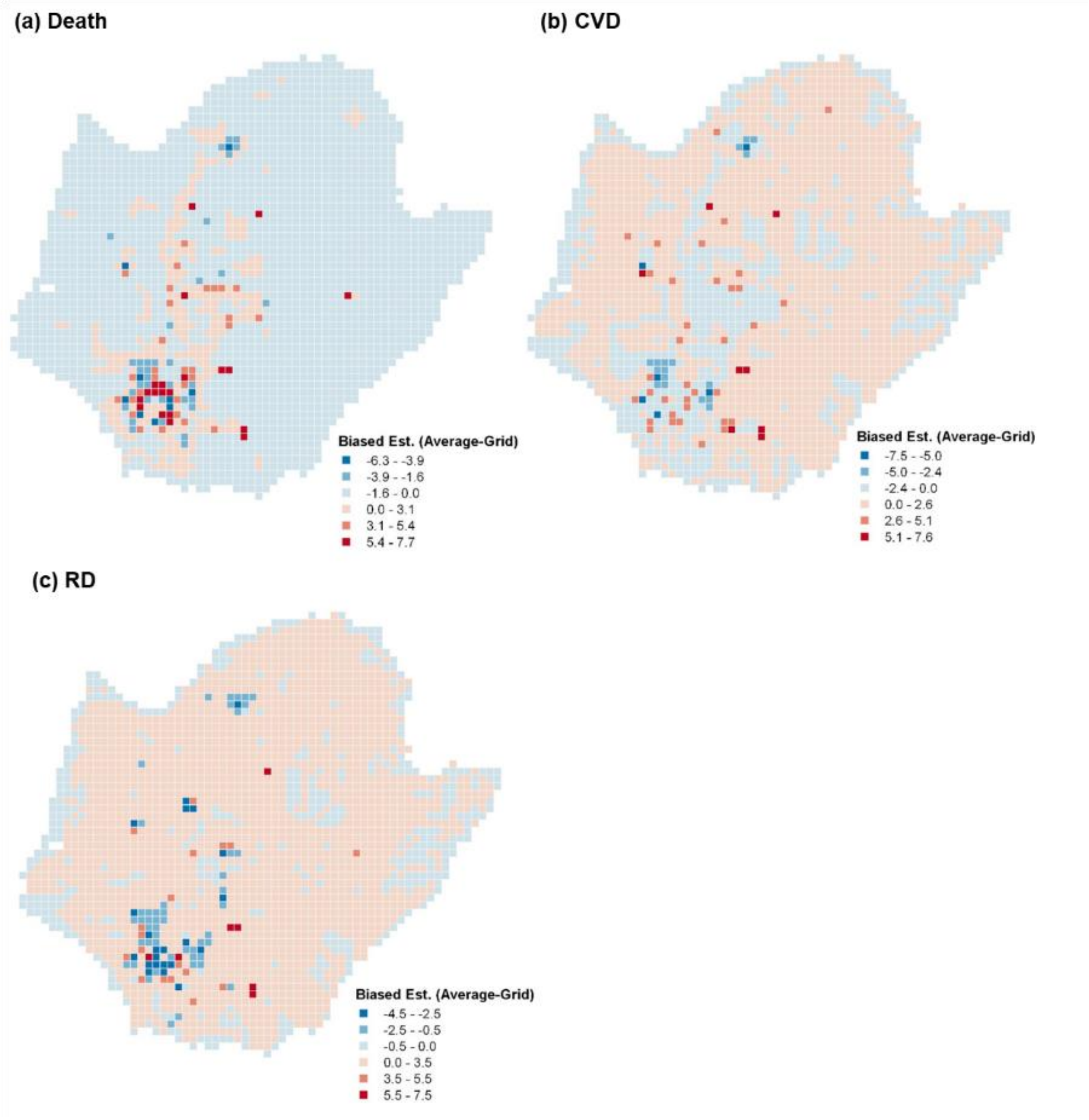


Figure 2-8. The difference of the number of ED visits (gridded-risk ED visits minus averaged-risk ED visits) for (a) death, (b) cardiovascular disease and (c) respiratory disease using monitoring data in 2013

Table 2-10. Number of increased ED visits under different scenarios in Tainan City, 2013

Scenario	Applied OR of daily $PM_{2.5}$ concentration (per 10 μg/m^3)	N (%) [1]		
		Monitoring data	CMAQ-Fused data	Difference [2]
Death				
Gridded Local Risk	1.000 [3] - 1.742	1699 (100%)	2137 (100%)	-437 (-20%)
Averaged Local Risk	1.245	1698 (100%)	2266 (106%)	-568 (-25%)
U.S. EPA Recommended Risk	1.170 [4]	1309 (77%)	1790 (84%)	-481 (-27%)
CVD				
Gridded Local Risk	1.000 [3] - 1.992	1297 (100%)	1750 (100%)	-453 (-26%)
Averaged Local Risk	1.266	1790 (138%)	2375 (136%)	-585 (-25%)
U. S. EPA Recommended Risk	1.002 [4]	178 (14%)	262 (15%)	-85 (-32%)
RD				
Gridded Local Risk	1.000 [3] - 1.518	1339 (100%)	1782 (100%)	-443 (-25%)
Averaged Local Risk	1.260	1763 (132%)	2343 (132%)	-580 (-25%)
U.S. EPA Recommended Risk	1.002 [4]	241 (18%)	355 (20%)	-113 (-32%)

[1] The percentage present the portion of ED visits divided by ED visits of gridded local risk scenario.

[2] Number of Monitoring data– Number of Fused data (% divided by fused estimation)

[3] Grids with OR<1 were replaced by OR=1 to assure non-negative number of ED visits.

[4] For each health outcome, the highest risk was used from U.S. EPA recommended risk in BenMAP-CE (Version 1.1.3), http://www.epa.gov/air/benmap.

Table 2-11. Descriptive analysis of observed and CMAQ-fused $PM_{2.5}$ in 2013

Location	Variable	Mean ± SD	Percentile		
			25th	50th	75th
Tainan City	Observed $PM_{2.5}$	26.72 ± 3.44	23.93	25.82	31.79
(2406 Grids)	CMAQ-fused $PM_{2.5}$	29.33 ± 8.37	23.29	25.82	37.52

Table 2-12. Validation of CMAQ-modeled $PM_{2.5}$ in 2013

Site #	MFB[1]	MFE[2]	Pearson r
S1	10.76%	49.27%	0.76
S2	-42.55%	59.53%	0.69
S3	-33.59%	55.50%	0.71
S4	10.76%	55.66%	0.75
All	-2.90%	13.75%	0.75

[1] $\text{Mean Fractional Bias (MFB)} = \frac{2}{\text{M}\cdot\text{N}}\sum_{k=1}^{M}\sum_{i=1}^{N}\left(\frac{P_{ik}-O_{ik}}{P_{ik}+O_{ik}}\right)$

[2] $\text{Mean Fractional Error (MFE)} = \frac{2}{\text{M}\cdot\text{N}}\sum_{k=1}^{M}\sum_{i=1}^{N}\left|\frac{P_{ik}-O_{ik}}{P_{ik}+O_{ik}}\right|$

M = number of sites

N = number of values

P_{ik} = predicted $PM_{2.5}$ of site k at day i

O_{ik} = observed $PM_{2.5}$ of site k at day i

[3] U.S. EPA performance criteria: $\text{MFB} \leq \pm 60\%$, $\text{MFE} \leq 75\%$

For example, in this study, we identified lower risks of CVD and RD but a higher risk of death in rural areas, and we suggested that one of the explanations for this tendency is the lower accessibility of residents in these areas to medical medical treatment and/or the lower willingness of residents to accept treatment. Third, with prior information about the spatial distribution of $PM_{2.5}$ exposure risk, the health officials can more accurately estimate the number of potential ED visits based on timely air pollution forecasts, so that medical and emergency care systems can be better prepared for surges in emergency department care demand.

2.5.7. Limitations

Our study still has some limitations. First, our framework only partially captures individual-level exposure pattern. For instance, although we located subjects by their residential address, it is possible that these subjects travel to another area for school or work. During working or staying indoors, the variation in intensity of individual-level exposure could also confound the impact of ambient $PM_{2.5}$. One potential strategy for solving this problem is collecting time-activity data, which would allow the duration of outdoor and indoor activities for individuals or group to be used to classify exposure level, but this method is time-consuming, costly, and its representativeness is easily limited by sample size and sampling methods. Second, we only applied the basic MMF techniques to fuse modeled data and monitoring data in this study. Advanced MMF techniques like machine learning techniques can be referred to the second study. Third, although we only applied the framework in one city, the impact of land-use characteristics on $PM_{2.5}$ exposure risk could still generally exist in the developing countries. The identified relationship between land-use characteristics and $PM_{2.5}$ exposure risk in this study region may be different from that in other regions or countries. Thus, more hospital data from other cities and regions need to be considered in the future works.

2.5. Conclusions

The finer spatial resolution of risk and BD estimations are increasingly required for regional air quality and public health management. This study built up a data fusion framework that facilitates individual development of risk and BD estimates for cities or countries based on spatial heterogeneity among different land-use characteristics and urbanization levels to enhance the accuracy of the BD estimations. In this study, we discovered that living in areas with high HLUL patterns increases the risk of CVD and RD by 59% and 35%, and that living in areas with low HLUL patterns increases the risk of death by 31%. Residents in rural areas had 1.40 times higher death risk compared with the average risk, and residents in the most urbanized areas had 1.57 and 1.20 times higher risk of CVD and RD than average. Our developed framework provides the exposure risk maps and BD map at the grid-level resolution, that visualize these risks and the BD. Such illustrations facilitate re-assessment of the potential risk of present urban planning strategies, and provide a quantified reference for air quality implementation plans and emergency episode-response plans.

Chapter 3. Quantifying Biases for Measurement-model Fusion

3.1. Abstract

Bias in chemical transport modeling (CTM) has impeded its applicability in environmental science for decades. Although emerging machine learning techniques can provide alternative and more accurate predictions, the remained bias and missing links with physical/chemical mechanisms still restrict our understanding of the real environment. The existing CTMs also hardly benefit from machine learning models.

This study proposed a machine learning-measurement model fusion (ML-MMF) framework, using the Community Multiscale Air Quality (CMAQ) model as an exemplary CTM to illustrate the ML-MMF's application in modeling improvement and bias quantification of the predicted air pollutants, $PM_{2.5}$ and O_3. The results show that the R-square of $PM_{2.5}$ and O_3 were improved from 0.41 and 0.48 to 0.86 and 0.82, respectively, in the study region. Bias quantification results showed the modeling bias is more affected by boundary conditions and local meteorology other than emission and land-use data in CMAQ modeling. The study illustrates exemplary cooperation between CTM and machine learning models, and the first developed ML-MMF framework can quantify the modeling bias structure and prioritize the improvement of CTM mechanisms and input data quality.

3.2. Introduction

Numerical models such as chemical transport models (CTMs) in environmental science have been intensively applied to simulate the environmental factors and atmospheric components based on known chemical/physical mechanisms and knowledge-based inputs for decades[141,142]. However, current numerical models or databases such as the Coupled Model Intercomparison Project Phase 6 (CMIP6) [143] or the Model Inter-Comparison Study for Asia (MICS- Asia) [141,142] still remained significant bias compared with observations due to the highly unpredictable environment. Furthermore, the demanding computational resources and time to conduct a series of numerical-based sensitivity simulations significantly slow down the improvement of modeling [39,144].

Owing to the development of environmental monitoring techniques and increasingly available environmental data in recent years, machine learning (ML) has provided effective and promising applications to prevent human and ecosystem exposure to environmental stressors [145–147]. Continental or regional forecast, or response systems based on ML algorithms to alert the public and policymakers to the near-future environmental risks and pollution such as air pollution [148,149], wildland fires [150,151], floods [152], heat waves [153], and other extreme climate events [147] have been rapidly developed and intensively implemented in practice. However, although ML techniques have shown excellent capabilities to provide more accurate estimations than traditional numerical models [43,154,155], their “black-box” modeling processes and remaining bias between predictions and observations are not much debated and insufficiently investigated [41]. Furthermore, failing to provide interpretability and explainability in terms of physical/chemical mechanisms also limits their persuasion for researchers and our understanding of the real environment [41,156].

To enhance the accuracy of CTM predictions, measurement-model fusion (MMF) techniques in CTM post-analysis have been developed to correct numerical modeling results based on observations [41]. In MMF application, ML algorithms such as regression-based model [155,157], tree-based model [43,44], and neural networks [13,154] had been utilized to optimize modeling results. Nevertheless, employing ML techniques either as a bias corrector or as a pure forecaster, few studies further analyzed the potential confounders or input components that cause bias, and the numerical model and modeling inputs still remained unfixed and hardly benefit from ML modeling except for corrected estimations.

Technically, the bias of numerical modeling is defined by the difference between modeling estimations and observations [158,159]. The biases between numerical estimations and in-situ monitoring observations are affected by modeling inputs including emission inventory, boundary conditions, meteorological factors, and land-use data [13,15]. Multiple reasons such as inaccurate modeling inputs [160,161], accumulation of input uncertainties during the modeling process, and imperfect chemical and physical mechanisms [162] in the model may contribute to considerable biases in air quality estimations. Although previous studies proved the capabilities of ML models in MMF or bias correction, the biases between modeled estimations and observations were not systematically investigated, and the bias originated from emissions, meteorology, boundary conditions, or geographical information data was not either quantified.

In this study, we used the CMAQ model [32] as an exemplary numerical model and selected Taiwan as the study region (Figure 3-1) due to its isolated geography, well-established air quality monitoring network, and routinely-updated emission inventory. The goal of this study is to illustrate the capability of ML techniques to quantify the potential sources of modeling bias from the prepared modeling inputs through the developed ML-MMF framework. The proposed ML-MMF framework was embedded with several basic ML techniques that can deal with non-linearity, including the k-nearest neighbors' regression (KNN) [163], regression tree (RT) [9], random forest (RF) [10], gradient-boosted tree models (GBM) [11], and convolutional neural network (CNN) [164]. Different modeling scenarios to include or exclude auxiliary data (emission inventory, boundary conditions, meteorological factors, and land-use data) for MMF were designed for further quantifying source-specific bias from modeling inputs. The target predictions were daily $PM_{2.5}$ average and maximum daily 8-hour ozone average (MDA8), which are the major air pollutants causing adverse health effects [20]. The objectives of this study are to improve CMAQ modeling performance through the developed ML-MMF framework, execute bias quantification for identifying the source-specific bias from CMAQ modeling inputs (emission, meteorology, boundary condition, and land-use data), and use the premature deaths as an example to emphasize the potential derived uncertainty with and without MMF.

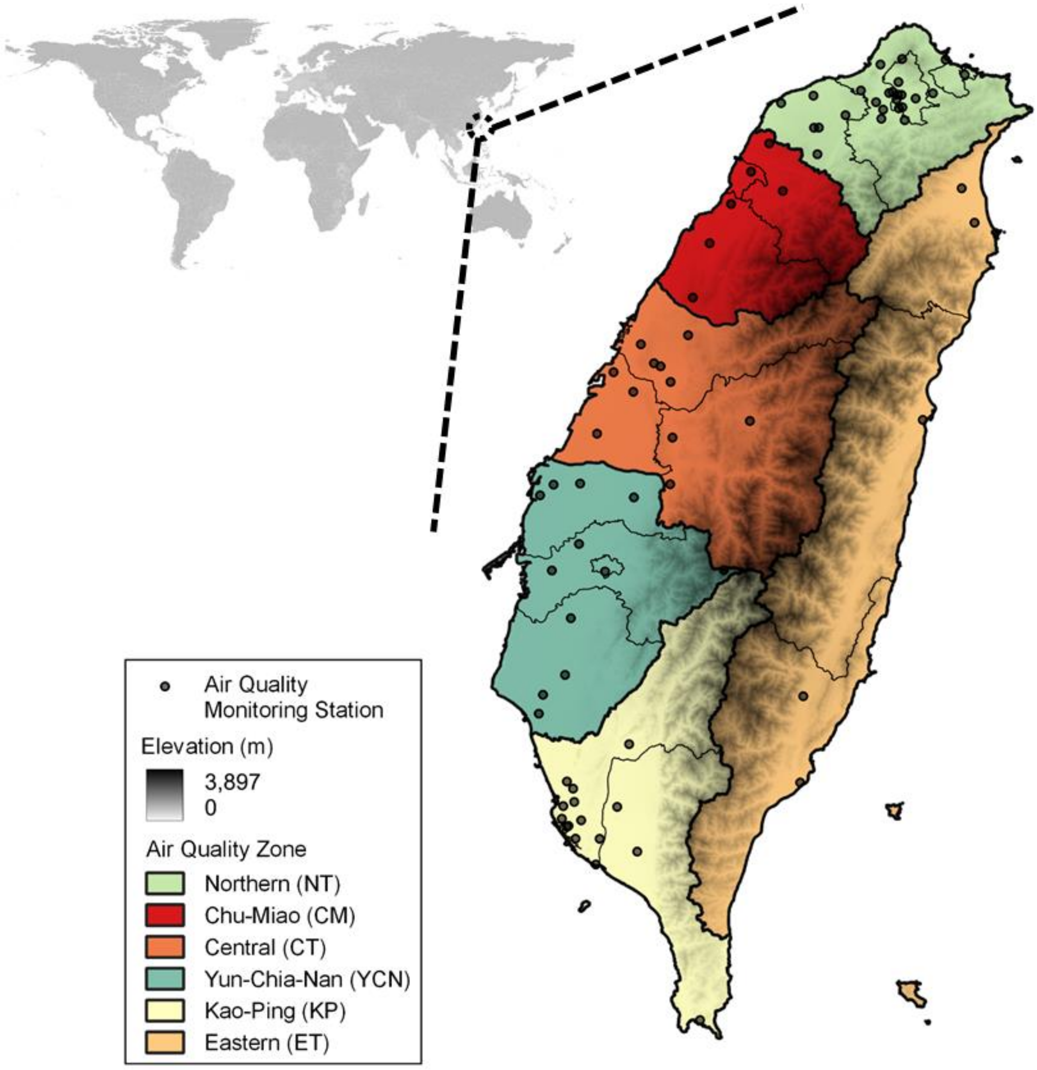


Figure 3-1. Air quality monitoring network (n=73) and air quality zones in Taiwan

3.3. Methodology

3.3.1. Dataset preparation

CMAQ model (version 5.2)[32] was used to simulate air pollutant concentrations with the Carbon Bond 6[165] and AERO6 [166] mechanisms which represent gas-phase and particulate matter chemistry, respectively. The weather research and forecasting model (WRF) [167], version 3.8, was used to simulate meteorological fields. The CMAQ and WRF modeling nested four layers from East Asia (81 km × 81 km) to Taiwan island (3 km × 3 km) which covers 90 (row) × 135 (column) horizontal grid cells. The configurations are the same as in our previous studies [129,168], as shown in Figure 3-2. Emissions are from Taiwan Emission Data System (TEDS) version 10.0 which is developed by Taiwan EPA and include industrial, mobile, area, and natural sources with 1 km × 1 km resolution.

All ML-MMF input variables are retrieved from CMAQ input data which include emission data, boundary condition data, meteorological data, and land-use data as listed in Table 3-1. Hourly observation data of $PM_{2.5}$ and O_3 in January, April, July, and October 2016 from 73 air quality monitoring stations (Figure 3-1) were used. Daily $PM_{2.5}$ and MDA8 O_3 were calculated based on the standards of WHO and Taiwan EPA [28] and used as the target datasets for ML algorithms. The chosen independent variables are related to the emission of precursors ($PM_{2.5}$, NOx, SOx, NH_3, and VOCs), the secondary formation process of $PM_{2.5}$ and O_3, and meteorological conditions that can have a significant impact on pollutant concentrations. Meteorological factors on 850 and 690 hPa layers were selected to represent the weather conditions of the mixing layer and low troposphere layer [13]. Four-day averages of boundary conditions were applied to represent the initial conditions of the modeling domain. The dominant land use category was based on MODIS classification and converted to dummy variables as inputs.

3.3.2. ML-MMF flowchart

The flowchart of the ML-MMF framework is presented in Figure 3-3. First, all inputs served as predictors including CMAQ output, emissions data, boundary condition data, meteorological data, and land-use data were compiled to the same resolution, 3 km × 3 km; Observation data from 73 air quality monitoring stations were further combined to predictor datasets, and the grid cells with observation data were used for the learning process. A random selection was employed on the complied datasets; 60% of the data set was selected as the training dataset, and the remaining 40% was used as the testing dataset. Second, five ML techniques including KNN, RT, RF, GBM, and CNN were trained with the training dataset to predict daily $PM_{2.5}$ and MDA8 O_3 the best schemes. Detailed introduction of each ML technique can be found in Appendix II. To assure modeling accuracy, for each algorithm, a 10-fold cross-validation was conducted to quantify the uncertainty of modeling performance. Finally, the testing dataset was applied to all models to validate the predictions, and the best algorithm was used for further bias quantification and assessment. Model performance was evaluated by R-square (R^2).

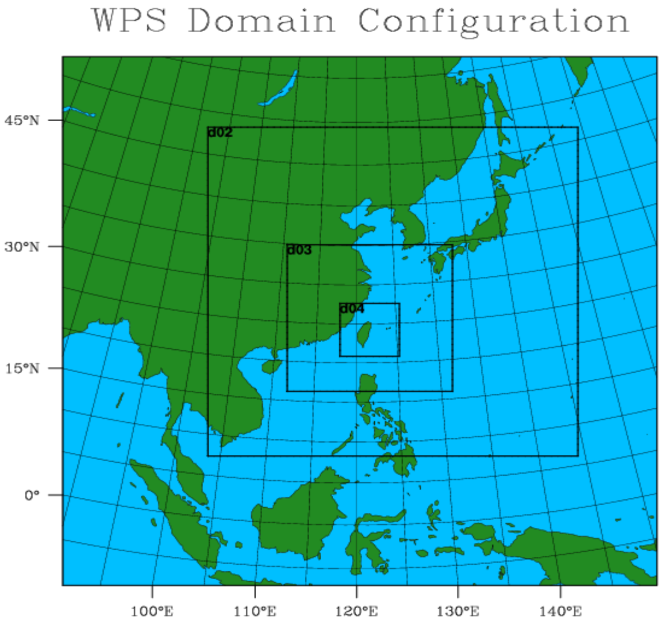


Figure 3-2. Overview of the simulated 4 nested domains from domain 1 (81 km) to domain 4 (3 km)

Table 3-1. Input variables selected to construct the machine learning (ML) model

Dataset	Source	Variables	
Observation	Monitoring stations	Daily $PM_{2.5}$ and MDA8 O_3	
Emission (n=24)	TEDS 10	$PM_{2.5}$, NOx, SOx, CO, NH_3, and VOCs emissions from point, mobile, area, and biomass sources.	
Boundary conditions (n=15)	WRF	air temperature (850 and 690 hPa), relative humidity (850 and 690 hPa), aerosol, NO_3, HNO_3, N_2O_5, NO, NO_2, sulfate, SO_2, VOCs, CO and O_3 at surface level	
Meteorological variables (n=21)	WRF	Surface (n=11)	surface pressure, PBL height, temperature at 2 m, mixing ratio at 2 m, wind speed, U wind component and V wind component, solar radiation reaching ground, precipitation, total cloud fraction, average liquid water content of cloud
		Pressure level (850 and 690 hPa) (n=10)	air temperature, potential vorticity, vertical velocity, U wind component and V wind component
Land-use (n=10)	CMAQ	evaluation, urban percent, dominant land use category based on MODIS (dummy variables)	

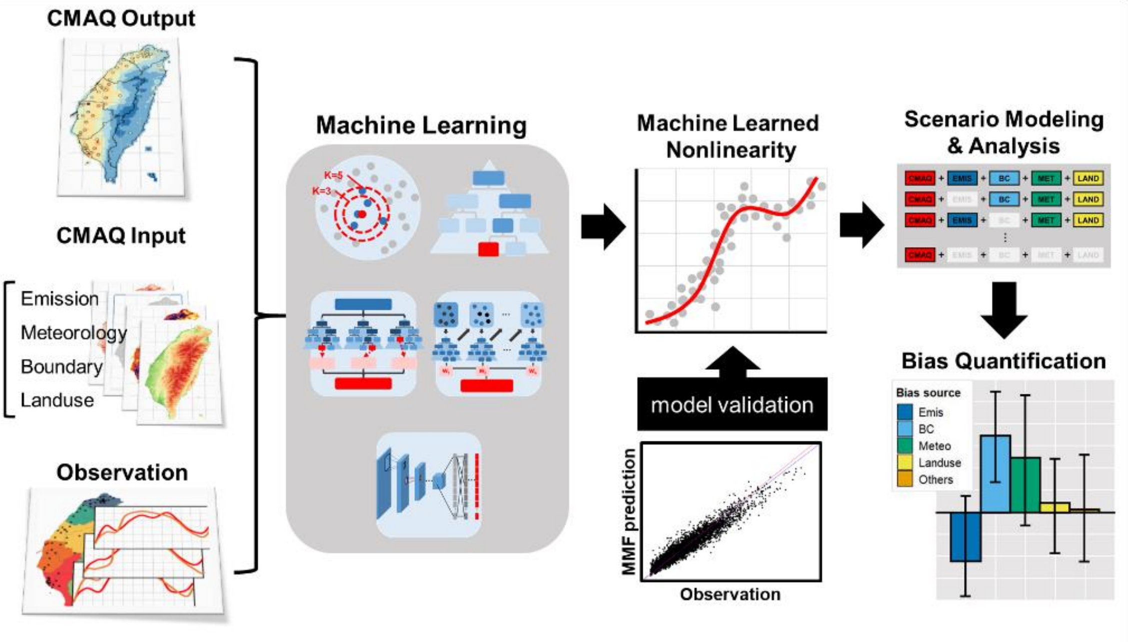


Figure 3-3. Technical flowchart of deep learning algorithm for $PM_{2.5}$ and O_3 prediction

3.3.3. Bias quantification

The bias quantification technique utilized $PM_{2.5}$ and O_3 estimations from each scenario. For each air quality zone, total bias (ΔC_{Total}) was defined by the changing population-weighted concentrations of $PM_{2.5}$ and O_3 concentrations between CMAQ raw output and S1_BASE ($\Delta C_{Total} = C_{CMAQ} - C_{S1\_BASE}$). The modeling capability of each component (emission, boundary condition, meteorology, and land-use data) was defined by either the including scenarios or the excluding scenarios. For example, the modeling capability of emission data was defined by the changing concentration between CMAQ raw output and S2_EM ($\Delta C_{1,EM} = C_{CMAQ}{-}C_{S2\_EM}$) or the changing concentration between S1_BASE and S6_nEM ($\Delta C_{2,EM} = C_{S1\_BASE} - C_{S6\_nEM}$). For all components, the calculated biases were further used to apportion their contributions to total bias through the following multiple linear regression (MLR) equation:

$$\Delta C_{Total} = \beta_0 + \beta_1 \Delta C_{i,EM} + \beta_2 \Delta C_{i,BC} + \beta_3 \Delta C_{i,MT} + \beta_4 \Delta C_{i,LU} + \varepsilon$$

where i represent the application of either including scenarios ($\Delta C_{1,EM}$, $\Delta C_{1,BC}$, $\Delta C_{1,MT}$, and $\Delta C_{1,LU}$) or excluding scenarios ($\Delta C_{2,EM}$, $\Delta C_{2,BC}$, $\Delta C_{2,MT}$, and $\Delta C_{2,LU}$) for each component; β_0 is the intercept; β_1 to β_4 represents contributed bias with a unit increase of delta $PM_{2.5}$ or O_3 concentration. The products including $\beta_1 \Delta C_{i,EM}$, $\beta_2 \Delta C_{i,BC}$, $\beta_3 \Delta C_{i,MT}$, and $\beta_4 \Delta C_{i,LU}$ are the changed concentrations from emission, boundary condition, meteorology, and land-use data respectively. ε are residuals and represent biases from other unidentified factors.

3.3.4. Sensitivity analysis of burden estimation

the number of premature deaths was used as a case study to illustrate the potential uncertainty if the CMAQ-modeled results were directly applied without MMF. Technically, premature deaths are calculated from concentration-response functions (CRFs) [48–50]. The number of premature deaths was calculated by the following equation:

$$\mathrm{Y} = \mathrm{E_0} \cdot \mathrm{P} \cdot \left(1 - \mathrm{e}^{-\beta \cdot (\mathrm{C} - \mathrm{C_0})}\right) \cdot \mathrm{A}$$

where Y is the number of premature deaths caused by daily $PM_{2.5}$ or O_3 exposure; $\mathrm{E_0}$ is the actual mortality rate; P is the population; The coefficient β is derived from the risk ratio from our previous study [33], The developed grid-level exposure risk values of daily $PM_{2.5}$ and 8-hour O_3 maximum are shown in Figure 3-4; A is a scalar of 1/365 to convert the annual rate to daily rate. $\mathrm{C_0}$ is the threshold concentration. The threshold concentration was set as 25 μg/m^3 for daily $PM_{2.5}$ or 60 ppb for MDA8 O_3 exposure due to the WHO [28] and Taiwan EPA.

3.4. Results and discussion

3.4.1. Improved modeling performance

The improved modeling performance by selected ML techniques and designed scenarios is shown in Table 3-2. The R^2 of CMAQ raw data is merely 0.41 and 0.48 for $PM_{2.5}$ and O_3 respectively. When including all the auxiliary data (emission, boundary condition, meteorology, and land-use data), the modeling performance of S1_BASE can be enhanced to 0.68-0.95 and 0.62-0.93 for $PM_{2.5}$ and O_3, respectively in terms of different techniques, which CNN has the highest R^2, followed by RF and GBM.

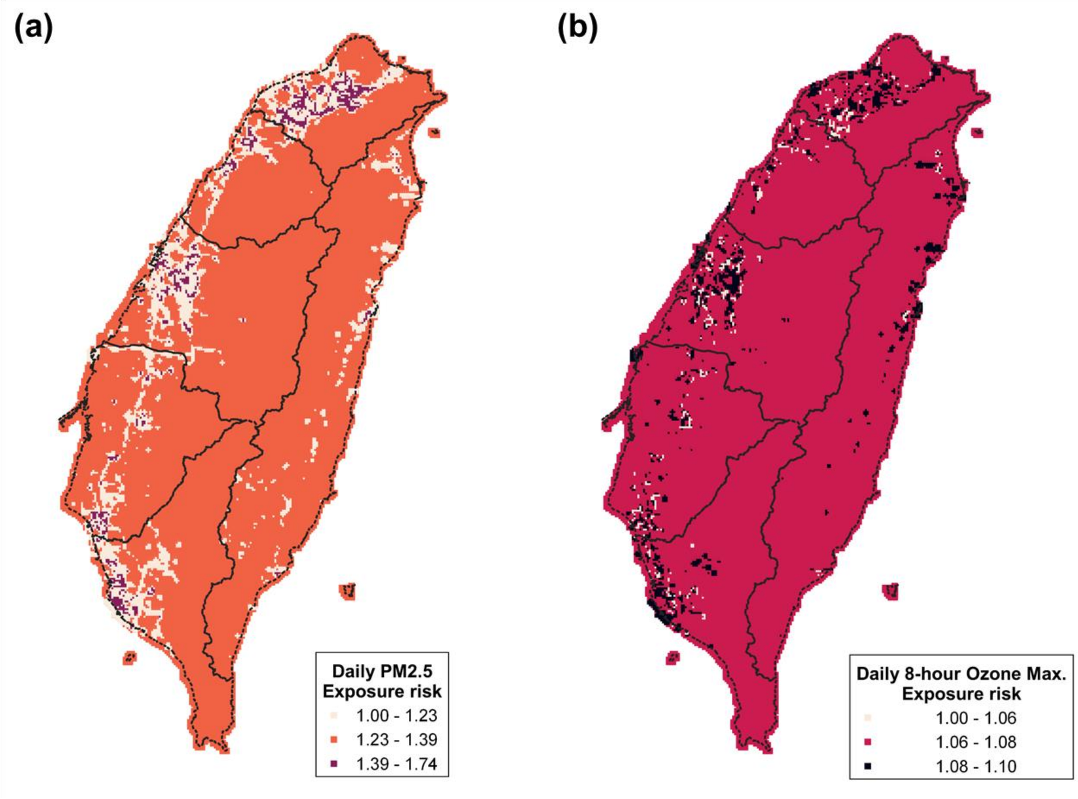


Figure 3-4. Employed risk value of each grid for premature death calculation for (a) daily $PM_{2.5}$ and (b) daily 8-hour O_3 maximum

Considering the performance of each ML technique, the R^2 (0.95 and 0.93 for $PM_{2.5}$ and O_3) of CNN for training data is much higher than the R^2 of testing data (0.83 and 0.78 for $PM_{2.5}$ and O_3), which implies its overfitting tendency when applied in MMF. Even though different split portions of training and testing data were modeled, the overfitting tendency of CNN still persisted. Thus, the results of CNN were not considered for our further analysis and application in this study. Next, both RF and GBM have comparable higher performance because of their training R^2 (RF: 0.87 and 0.84 for $PM_{2.5}$ and O_3; GBM: 0.86 and 0.82 for $PM_{2.5}$ and O_3). By comparing the spatial distribution of $PM_{2.5}$ and O_3 estimations from CMAQ, RF, and GBM outputs, as shown in Figure 3-9. $PM_{2.5}$ and O_3 estimations significantly approximate closer to observations by fusing with the other auxiliary data. After MMF, $PM_{2.5}$ concentrations are evaluated to the observed levels, especially in the CT and YCN regions, which implied using CMAQ modeled output alone could underestimate overall $PM_{2.5}$ exposure. The modeled O_3 concentrations are lowered to the observed levels, revealing the CMAQ tends to overestimate O_3 exposure, especially in western Taiwan. Additionally, by comparing CMAQ (Figure 3-9(a) and (d)) and GBM output (Figure 3-9(c) and (f)), RF (Figure 3-9(b) and (e)) showed relatively homogenous spatial patterns of $PM_{2.5}$ and O_3 concentrations and higher $PM_{2.5}$ concentrations in the central mountainous areas, which deviate from the real observations. The homogenous spatial patterns of RF imply its inferior performance in mountainous areas and could be due to its lower variable importance priorities of elevation and land-use characteristics, of which stiff terrain slopes in Taiwan could have much impact on air pollutants' concentrations. Because most monitoring stations are located in plains or basins, thus its modeling process prioritizes the other auxiliary variables other than land-use data for MMF. On the other hand, GBM presents a more reasonable spatial distribution of $PM_{2.5}$ and O_3 and is with closer concentration levels to observations. The significantly lower concentration in the central mountainous areas and the eastern valley and the higher concentration in the western plain are elaborated by GBM. In summary, GBM outputs were used for further MMF assessment and bias quantification.

3.4.2. Improved spatial distribution of air pollutants by MMF

For six air quality zones (NT, CM, CT, YCN, KP, and ET), the scatter plots between MMF-$PM_{2.5}$ and observed $PM_{2.5}$ and between MMF-MDA8 O_3 and observed MDA8 O_3 maximum at the daily scale are showed in Figure 3-5 and Figure 3-6. MMF-$PM_{2.5}$ and MMF-O_3 in each air quality zones have high correlations (R^2=0.67-0.90 and R^2=0.76-0.97) with observations, implying the MMF technique has no spatial specificity and can be generalized to any air quality zones, which results are similar to Sayeed et al. (2022) who figured out that the generalized model based on all monitoring stations can provide more stable enhancement other than on the basis of a single site [17]. For MMF-$PM_{2.5}$, The regression lines show only mild overestimations would occur in NT, YCN, and KP when $PM_{2.5}$ concentrations are higher, while $PM_{2.5}$ concentrations could be slightly underestimated in ET. For MMF-O_3, the regression lines in NT, CT, YCN, and KP imply MMF may slightly overestimate O_3 when the MDA8 O_3 maximum is over 60 ppb.

Compared with CMAQ output, MMF estimations showed different monthly spatial patterns (January, April, July, and October) for $PM_{2.5}$ and O_3, as shown in Figure 3-7 and Figure 3-8, respectively. First, MMF significantly increases the spatial heterogeneity of $PM_{2.5}$ and O_3 concentrations, which emphasizes the impact of the complexity of topography and land-use patterns on CMAQ modeling. The imperfection of physical or chemical mechanisms such as heterogeneous dust chemistry in the urbanized areas had been pointed out in the previous studies [162]. More homogenous concentrations from CMAQ output may misinterpret the real spatial distribution of $PM_{2.5}$ and O_3 and their concentrations in urbanized or populated areas. For $PM_{2.5}$, MMF elevates modeled $PM_{2.5}$ concentrations, which are closer to the observations, especially in YCN and KP in January and the whole of western Taiwan in April and July. The elevated $PM_{2.5}$ implies that directly using CMAQ output may potentially underestimate the health impact of $PM_{2.5}$ exposure in these months, especially around monitoring stations. On the contrary, $PM_{2.5}$ concentrations of CT, YCN, and KP in October are adjusted to the observed level, revealing that CMAQ tends to overestimate $PM_{2.5}$ in these regions in October. In addition, O_3 concentrations in all months and air quality regions are overestimated by CMAQ modeling, and MMF uses auxiliary data to adjust O_3 concentrations to the observed levels, especially in western Taiwan, revealing that direct using CMAQ-modeled O_3 outputs would overestimate the health impact of O_3 exposure.

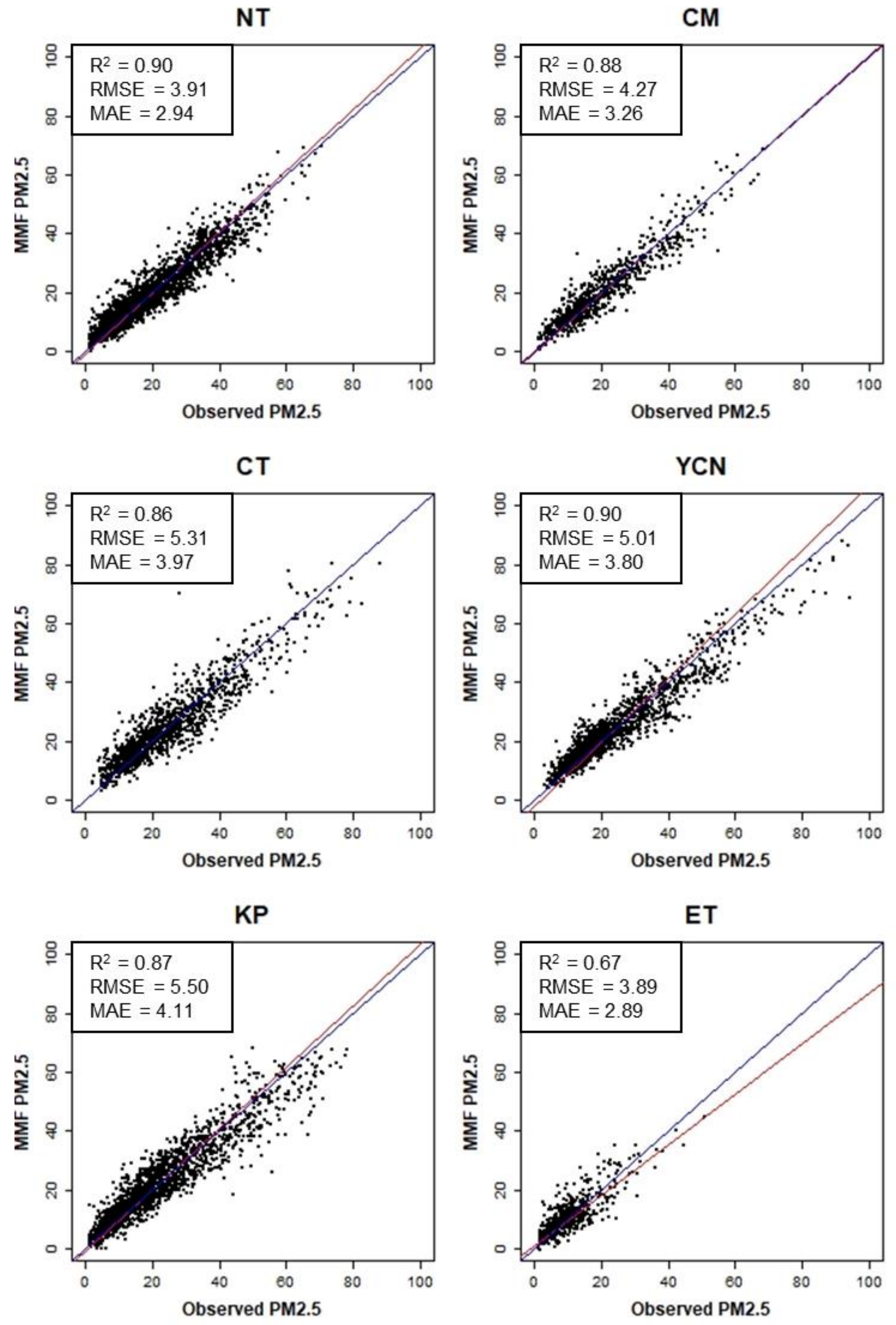


Figure 3-5. Scatter plots for six air quality zones (NT, CM, CT, YCN, KP, and ET) between MMF-daily $PM_{2.5}$ and observed daily $PM_{2.5}$

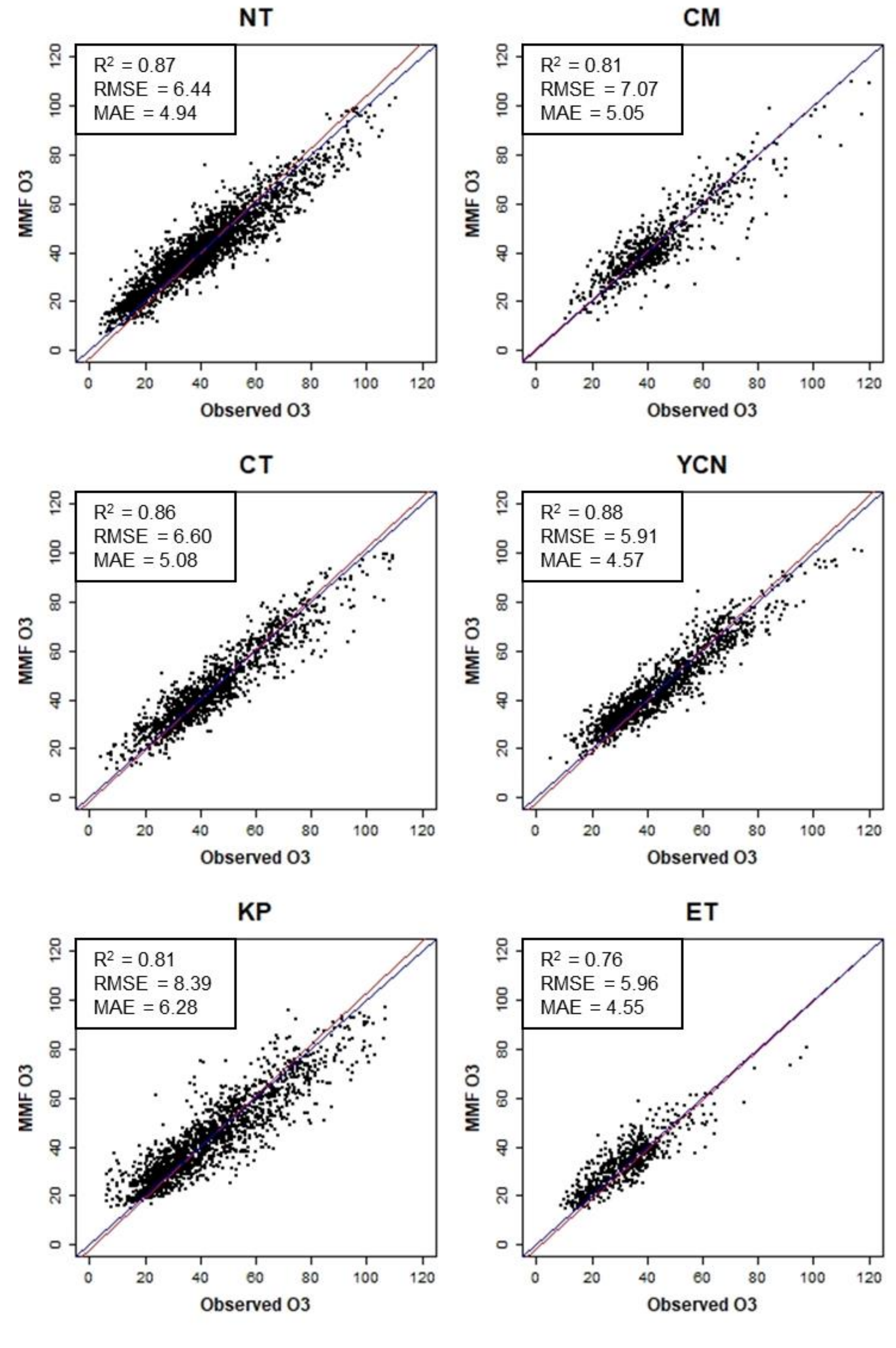


Figure 3-6. Scatter plots for six air quality zones (NT, CM, CT, YCN, KP, and ET) between MMF-MDA8 O_3 maximum and observed MDA8 O_3

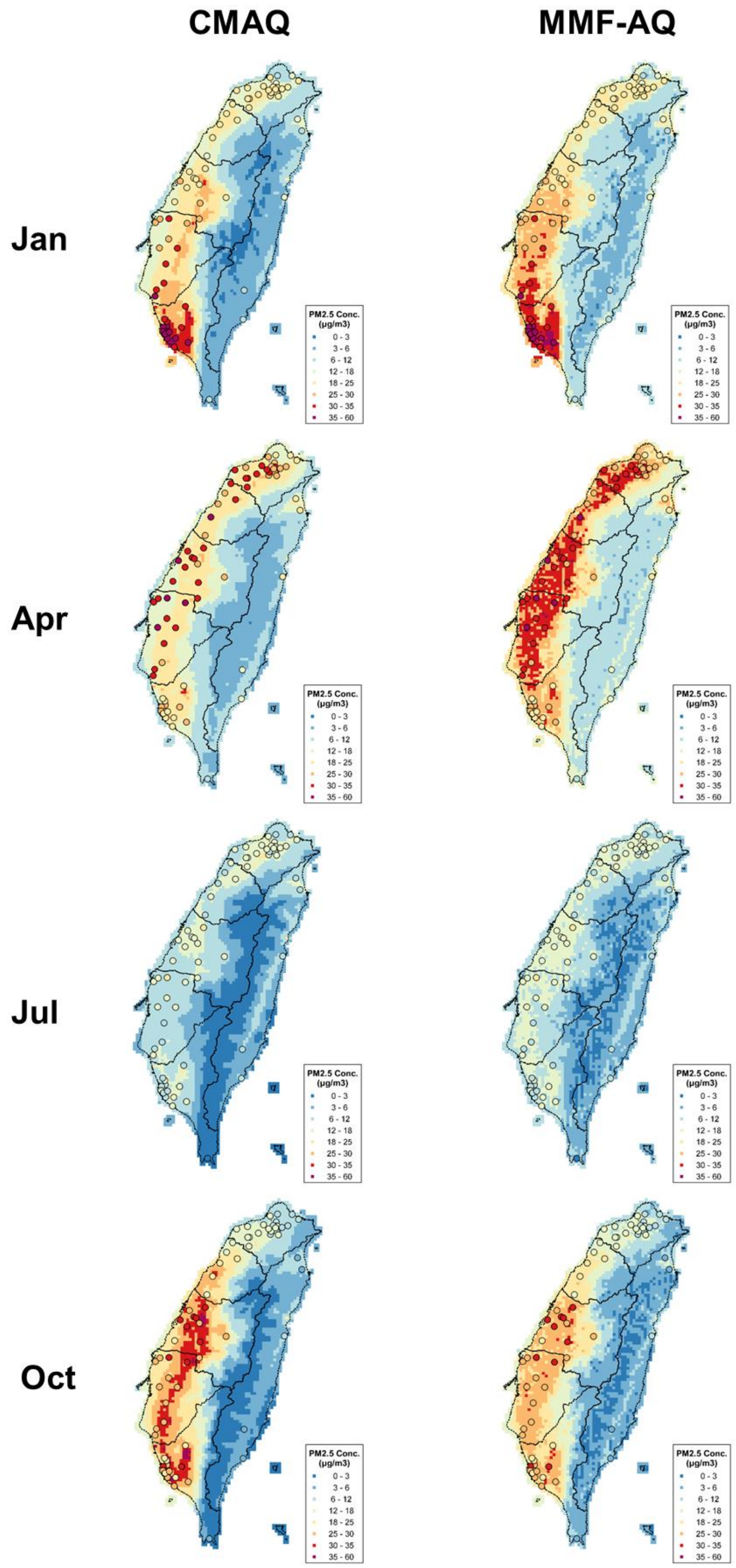


Figure 3-7. Spatial distribution of CMAQ- and MMF-PM$_{2.5}$ in January, April, July, and October, 2016

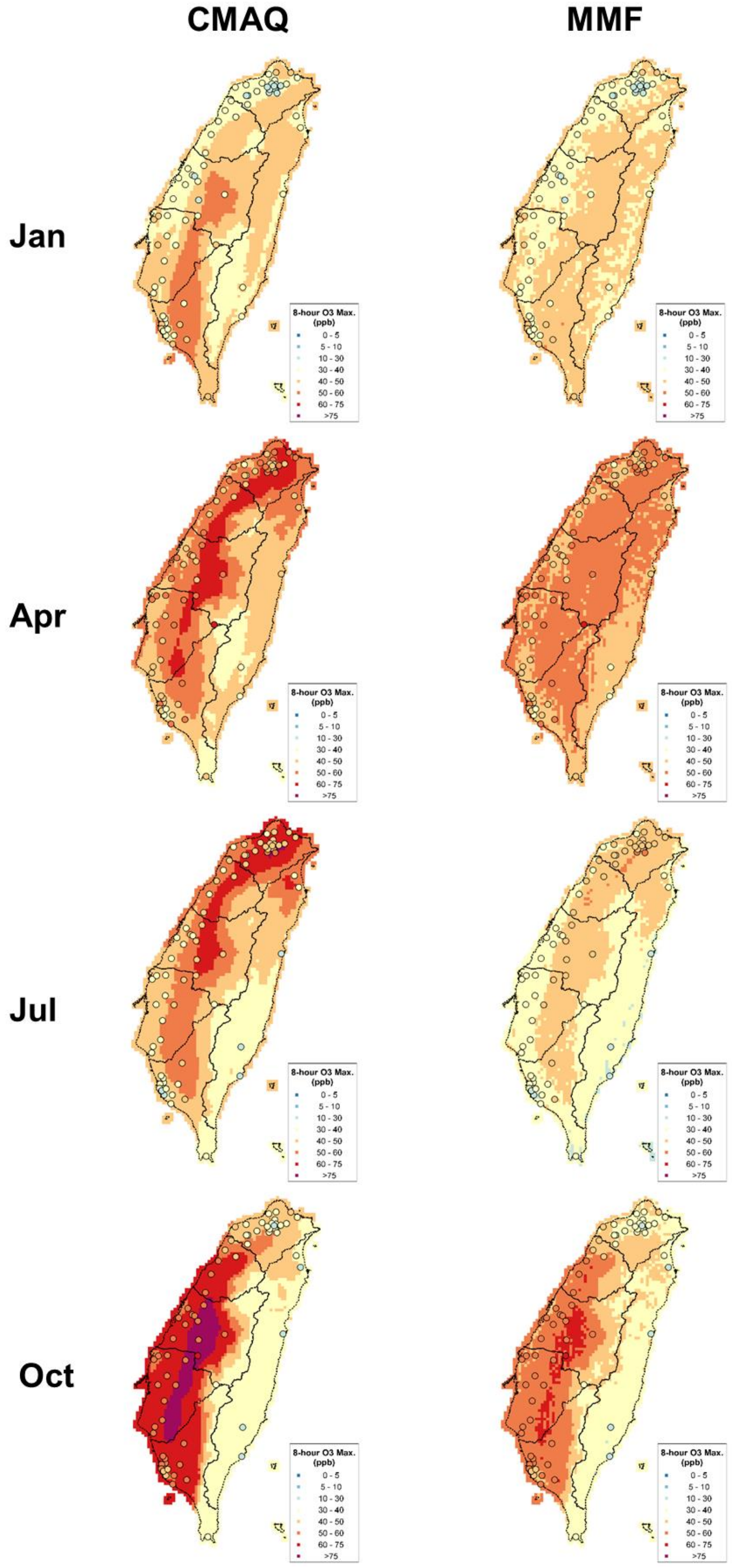


Figure 3-8. Spatial distribution of CMAQ- and MMF-O_3 in January, April, July, and October, 2016

Table 3-2. Modeling performance evaluation (R^2) of $PM_{2.5}$ and O_3 for different ML techniques and scenarios

Scenario	Input data	KNN		RT		RF		GBM		CNN	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
$PM_{2.5}$											
S1_BASE	CMAQ + Emis + BC + Met+ LU	0.68	0.68	0.76	0.77	0.87	0.87	0.86	0.86	0.95	0.83
S2_EM	CMAQ + Emis	0.51	0.54	0.47	0.48	0.59	0.59	0.58	0.59	0.62	0.57
S3_BC	CMAQ + BC	0.75	0.76	0.74	0.74	0.77	0.77	0.76	0.77	0.79	0.77
S4_MT	CMAQ + Met	0.71	0.73	0.63	0.65	0.81	0.82	0.76	0.77	0.84	0.68
S5_LU	CMAQ + LU	0.48	0.49	0.46	0.48	0.5	0.51	0.51	0.52	0.52	0.49
S6_nEM	CMAQ + BC + Met + LU	0.72	0.74	0.77	0.78	0.87	0.87	0.85	0.85	0.94	0.82
S7_nBC	CMAQ + Emis + Met + LU	0.53	0.54	0.62	0.67	0.8	0.81	0.79	0.79	0.91	0.69
S8_nMT	CMAQ + Emis + BC + LU	0.64	0.65	0.74	0.72	0.83	0.83	0.83	0.83	0.93	0.84
S9_nLU	CMAQ + Emis + BC + Met	0.7	0.7	0.77	0.77	0.87	0.87	0.86	0.86	0.94	0.83
O_3											
S1_BASE	CMAQ + Emis + BC + Met+ LU	0.62	0.63	0.73	0.73	0.84	0.85	0.82	0.81	0.93	0.78
S2_EM	CMAQ + Emis	0.51	0.49	0.49	0.48	0.53	0.52	0.56	0.55	0.58	0.51
S3_BC	CMAQ + BC	0.78	0.77	0.75	0.74	0.78	0.78	0.77	0.77	0.79	0.76
S4_MT	CMAQ + Met	0.7	0.69	0.62	0.61	0.78	0.79	0.74	0.73	0.78	0.63
S5_LU	CMAQ + LU	0.51	0.47	0.49	0.48	0.52	0.49	0.53	0.51	0.54	0.5
S6_nEM	CMAQ + BC + Met + LU	0.71	0.71	0.72	0.74	0.84	0.85	0.81	0.81	0.9	0.79
S7_nBC	CMAQ + Emis + Met + LU	0.51	0.5	0.61	0.6	0.77	0.77	0.75	0.74	0.85	0.66
S8_nMT	CMAQ + Emis + BC + LU	0.59	0.6	0.73	0.73	0.81	0.81	0.79	0.79	0.88	0.78
S9_nLU	CMAQ + Emis + BC + Met	0.67	0.68	0.71	0.72	0.84	0.85	0.82	0.81	0.89	0.78

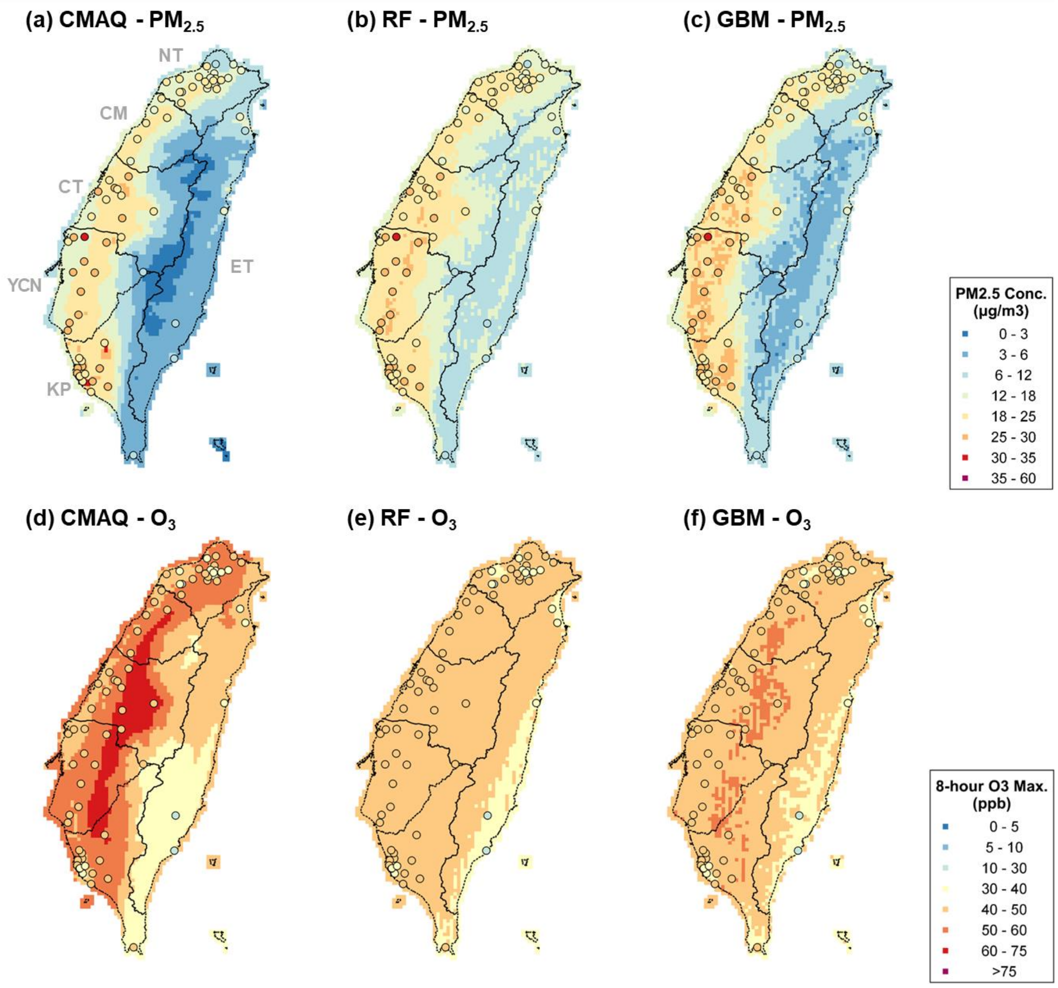


Figure 3-9. Observations (circles) and modeled estimations (grids) for $PM_{2.5}$ (a-c) and O_3 (d-f) from CMAQ, RF, and GBM outputs

3.4.3. Scenario design for prioritizing the importance of auxiliary data

Different scenarios were designed to assess the sources of bias between CMAQ raw output and observations, as shown in Table 3-2. S1_BASE is the base model that uses all inputs for prediction and serves as a baseline to compare with the other scenarios, and the following scenarios can be classified into two categories: including scenarios (S2_EM, S3_BC, S4_MT, S5_LU) and excluding scenarios (S6_nEM, S7_nBC, S8_nMT, S9_nLU). Compared with CMAQ raw output, including scenarios (S2_EM, S3_BC, S4_MT, S5_LU) assess the individual improved performance by individually including emission, boundary condition, meteorological, and land-use data individually and illustrate individual capability for MMF. Additionally, compared with the S1_BASE, excluding scenarios (S6_nEM, S7_nBC, S8_nMT, S9_nLU) exclude emission, boundary condition, meteorological, and land-use data, respectively, and the modeling results of excluding scenarios can show the decreased modeling performance and interpretability due to the lack of individual dataset in each scenario.

The modeling performance for the designed scenarios is also shown in Table 3-2. Compared with CMAQ raw output (r=0.41 for $PM_{2.5}$ and 0.48 for O_3), the R^2 suggests adding boundary conditions (S3_BC) and meteorological factors (S4_MT) for MMF would largely increase MMF modeling performance. On the other hand, when compared with S1_BASE, the lack of boundary conditions decreases most modeling capability, which is from 0.68-0.95 to 0.53-0.91 (S7_nBC) for $PM_{2.5}$ and from 0.62-0.93 to 0.51-0.85 (S7_nBC) for O_3, while excluding local meteorological factors would decrease R^2 from 0.68-0.95 to 0.64-0.93 (S8_nMT) for $PM_{2.5}$ and from 0.62-0.93 to 0.59-0.88 (S8_nMT) for O_3. In summary, the results implied that the CMAQ model could still has imperfection to well simulate $PM_{2.5}$ and O_3 with boundary conditions and local meteorology, and boundary conditions contribute to the most of explained variance for MMF modeling. One possible explanation for higher contribution from boundary conditions is frequent long-range transboundary air pollutants transported from mainland China in fall and winter [169,170], which carry primary PM and precursors of secondary PM and O_3 to Taiwan [69,70], but such hourly- and daily-scale weather conditions and air pollutant concentrations from boundary conditions are difficult to be accurately captured by global or regional emission inventory and verified by ground-level observations. Another explanation is the high sensitivity of the CMAQ model to boundary conditions around Taiwan, which suggested the CMAQ model improves the dust emission treatment for a better simulation of dust aerosol transport and deposition mechanisms over the marine boundary layer [171]. In addition, the inferior importance of local meteorology could be due to the confounding effect of its collinearity with boundary conditions. Another possible reason could be that the current meteorological models still have limitations to predict over complex terrain and under extremely stable boundary layers [28,172]. Compared with boundary conditions and local meteorology, emission inventory and land-use data only have relatively minor contributions to MMF modeling, but it does not mean emission inventory and land-use data are not essential and not sensitive for CMAQ modeling. On the contrary, it reveals the input emission and land-use data can better explain the variance of $PM_{2.5}$ and O_3 in the most of modeling periods, so their derived uncertainties are relatively lower than the uncertainty from boundary conditions and local meteorology, which dominate the modeling bias in the modeling period.

3.4.4. Bias quantification and potential impact

For both $PM_{2.5}$ and O_3, the results of bias quantification models showed that using including scenarios (R^2=0.96 for $PM_{2.5}$; R^2=0.98 for O_3) had a higher explained variance compared with using excluding scenarios (R^2=0.44 for $PM_{2.5}$; R^2=0.32 for O_3), so the further bias quantification analysis was based on including scenarios. The spatial distributions of daily $PM_{2.5}$ and MDA8 O_3 estimation biases and their apportioned biases from emission, boundary condition, local meteorology, land-use data, and other unidentified factors are shown in Figure 3-10, and the monthly averages of population-weighted $PM_{2.5}$ and O_3 estimation biases from each component are listed in Table 3-3 and Table 3-4, respectively. In Figure 3-10, the bias is defined by the subtraction of MMF estimations from CMAQ outputs, and the biases showed in the histograms are based on population-weighted concentrations, which can emphasize the exposure of the major population.

In Figure 3-10(a), compared with MMF results, the CMAQ model underestimates $PM_{2.5}$ concentrations for all air quality zones by 0.99-4.56 μg/m^3 (2-23%), by which YCN is most underestimated and KP is least underestimated. Although the CMAQ model overestimates coastal areas in KP, the population-weighted $PM_{2.5}$ concentrations were still underestimated. It is because the regions with overestimated $PM_{2.5}$ have lower population densities in coastal areas. Moreover, the spatial distribution of total bias showed that the CMAQ model tends to overestimate (red) $PM_{2.5}$ concentrations under hills and mountains and to underestimate (blue) then in plains and basins, especially around coastal areas. The monthly patterns of bias showed that April had more underestimated $PM_{2.5}$ while October's $PM_{2.5}$ concentrations in western regions (CM, CT, YCN, and KP) were overestimated. Additionally, the bias quantification results showed that boundary conditions and local meteorology are the main driving forces to cause underestimation in January and April and overestimation in October. On the contrary, land-use data contributes a positive driving force in YCN, KP, and ET and cause overestimations on the edge of hills or mountains, implying that the evaluation factors such as elevation could cause positive biases of CMAQ modeling when the pollutants accumulate under hills or mountains.

In Figure 3-10(b), the CMAQ model overestimates O_3 concentrations for all air quality zones by 5.13-10.96 ppb (17-29%), and almost O_3 concentrations in western regions are overestimated. Although CT has higher O_3 overestimation (9.81 ppb, 23%), due to its lower population density and fewer observation sites in these regions, O_3 concentrations in CT are still slightly lower than in NT (10.96 ppb, 29%). The monthly patterns showed July and October have more overestimated O_3. NT, CM, and ET regions are more overestimated in July, while CT, YCN, and KP regions are more overestimated in October. Only O_3 concentrations of KP in April are underestimated by 1.70 ppb. Similar to $PM_{2.5}$, the bias quantification results of O_3 showed that boundary conditions and local meteorology are the main driving forces causing overestimation.

Further sensitivity analysis was conducted to emphasize the impact of applying CMAQ-modeled output without MMF for calculating premature deaths. For six air quality zones, the derived premature deaths calculated from CMAQ and MMF outputs due to $PM_{2.5}$ and O_3 exposure are shown in Figure 3-11. For $PM_{2.5}$, compared with the MMF-derived total deaths (n=3641), using CMAQ output would underestimate deaths by 18% (3000 deaths). Among six air quality zones, premature deaths in NT, CT, YCN, and ET are underestimated by 6-78% with the highest underestimated premature deaths in YCN (306 deaths (35%)). Besides, premature deaths in KP are slightly overestimated by 4% when applying CMAQ-modeled $PM_{2.5}$, even though the CMAQ model underestimates $PM_{2.5}$ concentrations by 2%. The reason for this conflict is that the CMAQ model seriously overestimates $PM_{2.5}$ concentrations in KP in October but mildly underestimates $PM_{2.5}$ concentrations in the other months. For O_3, compared with the total MMF-derived deaths (n=3831), applying CMAQ output for burden calculation would highly overestimate total death by 171% (10331 deaths). The estimated premature deaths in all air quality zones are overestimated by 114-303%, and the trend of overestimation in deaths in all air quality is similar to the trend of O_3 concentrations. NT has the highest overestimated premature deaths (3016 deaths (303%)) due to its highest overestimated O_3 concentrations and dense population.

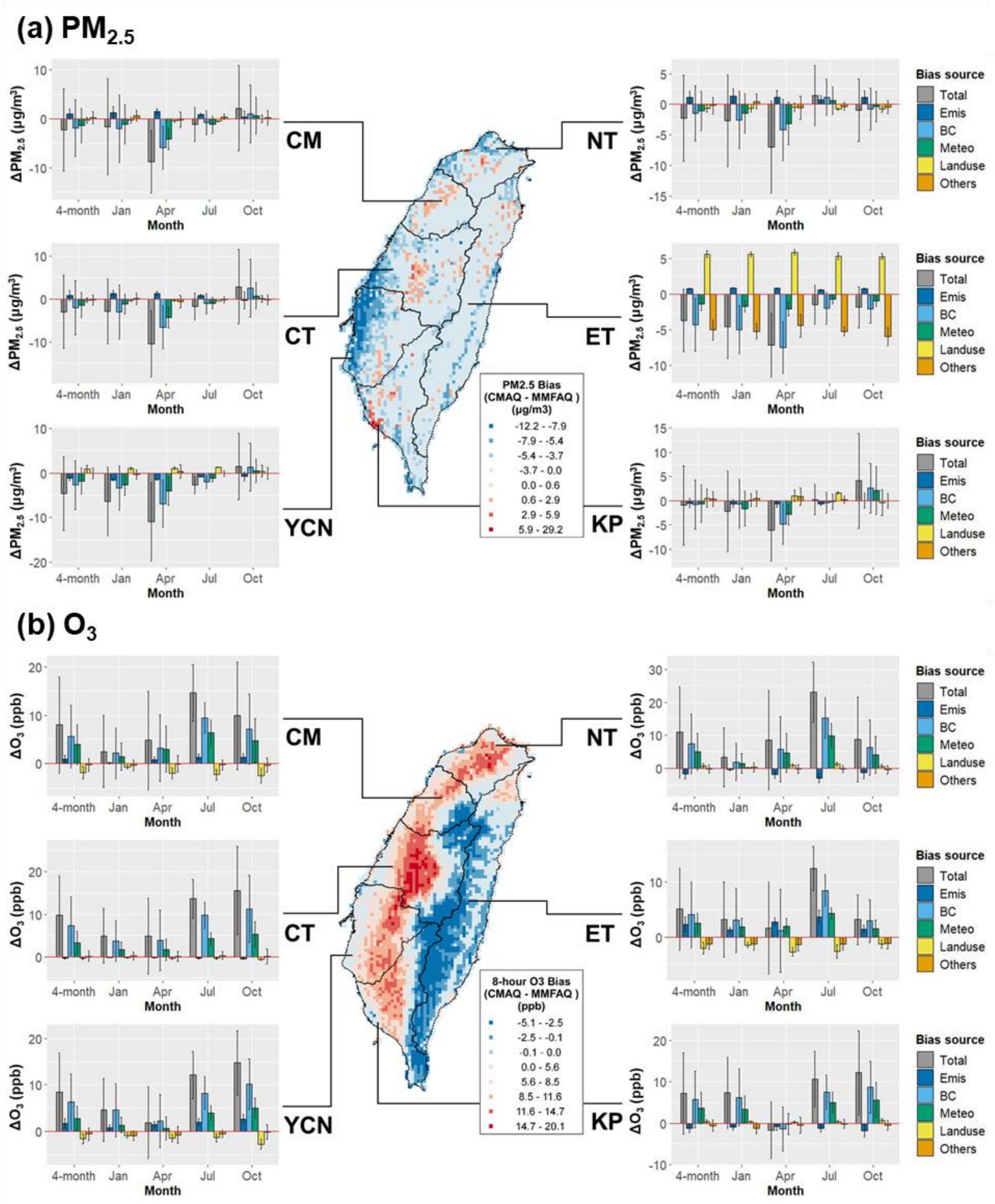


Figure 3-10. Spatial distributions of (a) daily PM$_{2.5}$ and (b) MDA8 O$_3$ maximum estimation biases and their apportioned biases from emissions, boundary conditions, local meteorology, land-use data, and other unidentified factors.

Table 3-3. Monthly average of population-weighted $PM_{2.5}$ concentration biases from emission data, boundary conditions, local meteorology, land-use data, and other unidentified factors

Air quality zone	Period	$PM_{2.5}$ Bias Source (μg/m^3)					
		Total	Emis	BC	Met	Landuse	Others
NT	4-month	-2.25	1.11	-1.50	-1.07	-0.66	-0.21
	Jan	-2.68	1.38	-2.59	-1.48	-0.64	0.44
	Apr	-6.98	1.14	-4.16	-3.13	-0.45	-0.55
	Jul	1.46	0.77	1.17	0.64	-0.80	-0.32
	Oct	-0.95	1.19	-0.73	-0.37	-0.76	-0.42
CM	4-month	-2.31	0.94	-1.91	-1.45	-0.30	0.25
	Jan	-1.60	1.18	-2.07	-1.15	-0.28	0.63
	Apr	-8.82	1.43	-5.91	-4.16	-0.36	-0.24
	Jul	-1.17	0.84	-0.81	-1.20	-0.52	0.40
	Oct	2.14	0.30	0.98	0.63	-0.05	0.23
CT	4-month	-2.99	0.77	-2.00	-1.40	-0.24	-0.21
	Jan	-2.88	1.30	-3.02	-1.13	-0.28	0.07
	Apr	-10.45	1.25	-6.58	-4.29	-0.33	-0.60
	Jul	-1.76	0.82	-1.04	-0.94	-0.44	-0.16
	Oct	2.91	-0.30	2.59	0.74	0.11	-0.13
YCN	4-month	-4.56	-1.15	-2.71	-1.78	0.93	-0.02
	Jan	-6.36	-1.52	-3.31	-2.71	0.99	-0.31
	Apr	-10.89	-1.48	-6.88	-3.99	1.11	0.36
	Jul	-2.65	-0.85	-1.99	-1.13	1.27	-0.08
	Oct	1.46	-0.75	1.36	0.57	0.39	-0.06
KP	4-month	-0.99	-0.50	-0.77	-0.62	0.52	0.34
	Jan	-2.23	-0.62	-0.67	-1.63	0.15	0.43
	Apr	-6.18	-0.70	-4.83	-2.86	1.03	0.82
	Jul	0.22	-0.69	-0.41	-0.21	1.58	0.18
	Oct	4.09	0.01	2.61	2.11	-0.44	-0.05
ET	4-month	-3.72	0.77	-4.30	-1.37	5.60	-5.03
	Jan	-4.53	0.87	-4.98	-1.73	5.59	-5.23
	Apr	-7.15	0.81	-7.49	-2.08	5.87	-4.41
	Jul	-1.45	0.61	-2.02	-0.70	5.29	-5.23
	Oct	-1.86	0.77	-2.10	-0.99	5.26	-5.95

Table 3-4. Monthly average of population-weighted O_3 concentration biases from emission data, boundary conditions, local meteorology, land-use data, and other unidentified factors

Air quality zone	Period	O_3 Bias Source (ppb)					
		Total	Emis	BC	Met	Landuse	Others
NT	4-month	10.96	-1.71	7.47	5.06	0.74	-0.17
	Jan	3.30	-0.38	1.92	1.43	0.22	0.27
	Apr	8.56	-1.91	5.80	4.65	0.89	-0.23
	Jul	23.10	-3.05	15.28	9.91	1.23	-0.27
	Oct	8.79	-1.43	6.33	4.06	0.59	-0.47
CM	4-month	7.99	0.90	5.61	3.92	-1.88	-0.26
	Jan	2.50	0.14	2.14	1.44	-0.71	-0.33
	Apr	4.81	0.84	3.22	2.99	-2.02	-0.06
	Jul	14.62	1.36	9.47	6.35	-2.28	-0.28
	Oct	9.90	1.25	7.17	4.69	-2.51	-0.34
CT	4-month	9.81	-0.28	7.28	3.33	-0.38	0.18
	Jan	4.94	-0.10	3.77	1.71	-0.17	0.20
	Apr	4.85	-0.24	3.91	1.67	-0.37	0.06
	Jul	13.66	-0.33	9.73	4.31	-0.37	0.31
	Oct	15.65	-0.44	11.28	5.27	-0.60	0.13
YCN	4-month	8.40	1.70	6.39	2.76	-1.66	-0.61
	Jan	4.59	0.82	4.56	1.36	-0.94	-0.96
	Apr	1.87	1.37	2.25	0.60	-1.50	-0.77
	Jul	12.15	1.96	8.23	3.92	-1.41	-0.55
	Oct	14.77	2.58	10.16	4.98	-2.75	-0.20
KP	4-month	7.21	-1.24	5.74	3.60	0.43	-0.59
	Jan	7.36	-0.98	6.24	3.38	0.34	-1.22
	Apr	-1.70	-0.86	-1.35	-0.14	0.32	-0.50
	Jul	10.69	-1.22	7.50	5.00	0.34	-0.15
	Oct	12.20	-1.83	8.72	5.65	0.71	-0.50
ET	4-month	5.13	2.28	4.03	2.47	-2.02	-1.27
	Jan	3.16	1.24	3.04	1.81	-1.45	-1.22
	Apr	1.62	2.72	1.15	2.01	-2.69	-1.39
	Jul	12.48	3.62	8.49	4.26	-2.59	-1.30
	Oct	3.14	1.45	2.94	1.54	-1.30	-1.18

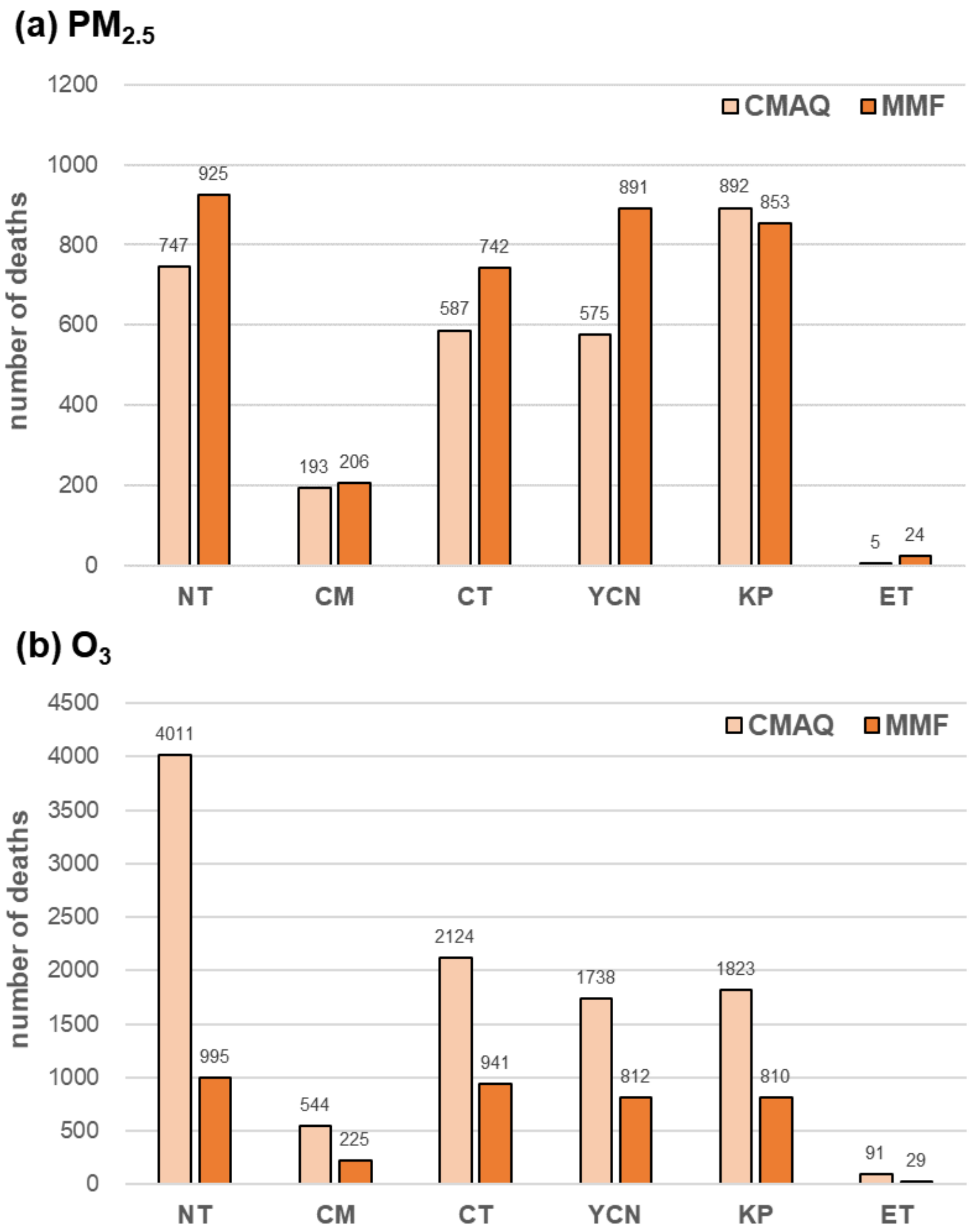


Figure 3-11. Premature deaths for six air quality zones calculated from CMAQ and MMF output due to (a) daily $PM_{2.5}$ and (b) MDA8 O_3 maximum exposure

3.4.5. Discussion

The proposed ML-MMF framework has several merits. First, the framework can serve as a post-processing procedure to improve CMAQ modeling performance, and the users can select the optimal ML algorithms as the MMF technique based on the improved performance and spatial distribution. In this study, the overall modeling performance of daily $PM_{2.5}$ and MDA8 O_3 was improved from 0.41 and 0.48 to 0.86 and 0.82, respectively, based on the GBM algorithm, which showed more reasonable $PM_{2.5}$ and O_3 concentrations and spatial distributions compared with CMAQ raw output and reflected the impact of complex topography on $PM_{2.5}$ and O_3 concentrations in the finer scale.

Second, the developed ML-based bias quantification technique provides details about the bias sources of numerical estimations. By employing the bias quantification technique, the model developer can have priority to optimize algorithms in the numerical model, or the users can know the bias structure and have a priority to improve their modeling input data quality in their study region. In our study region, the biases mainly originated from boundary conditions and local meteorology, and most of the overestimations occur under hills and mountains, implying the CMAQ model may have inferiority to model the air pollutants trapped in valleys or foothills of mountains.

Third, this study highlights the potential concern if the modeled results from numerical models were directly applied to air quality management and public health policies without correcting or fusing with observations that are commonly used for the current policymaking. In this study, we used the number of premature deaths as an example to illustrate the deviations with and without MMF by ML techniques. For all air quality zones, compared with MMF results, the CMAQ model underestimates population-weighted $PM_{2.5}$ concentrations by 0.99-4.56 μg/m^3 (2-23%) and overestimates population-weighted O_3 concentrations for all air quality zones by 5.13-10.96 ppb (17-29%). The biased $PM_{2.5}$ and O_3 estimation would cause the underestimation of premature death due to $PM_{2.5}$ exposure by 18% and the overestimation of death due to O_3 exposure by 171%, which may mislead the emission control strategies and air quality policies.

This study still has some limitations. First, the applicability of the ML model highly depends on geological characteristics around monitoring stations and study regions. In this study, because most monitoring stations are located in populated areas such as coastal, basins, and plains, and there are very little monitoring data in mountainous areas, some ML techniques such as RF cannot properly utilize land-use variables for MMF modeling. Even though satellite data may serve as an alternative source of observations, satellite data are still easily biased by clouds and columns of atmospheric layers. Second, although the proposed bias quantification can quantify the bias contributions from each input component (emissions, boundary conditions, meteorological variables, and land-use data), the bias of each component is still the combined uncertainty of inaccuracy of input data and imperfect mechanisms in the model, which cannot be easily differentiated. For example, in this study, the meteorology-contributed bias could be from the inaccurate estimations of WRF modeling or imperfect physical and chemical mechanisms in the CMAQ model.

3.5. Conclusion

Bias in numerical models has impeded their applicability for decades. Although ML may serve as an alternative tool to forecast environmental trends and prevent human and ecosystem exposure to environmental stressors and pollutions, its “black-box” modeling processes, modeling bias, and missing links with physical/chemical mechanisms limit its persuasion and our understanding of the real environment. The existing numerical models also hardly benefit from ML models except for use for bias correction.

This study proposed an ML-MMF framework, using the CMAQ model as an exemplary CTM to illustrate its application in model improvement and further attribute the bias to the prepared model inputs. the R^2 of daily $PM_{2.5}$ and O_3 exposure were significantly improved from 0.41 and 0.48 to 0.86 and 0.82 by learning from auxiliary data including emission, boundary condition, meteorology, and land-use data. The proposed ML-MMF framework can not only adjust $PM_{2.5}$ and O_3 concentrations to the observed levels but also improve the spatial heterogeneity of $PM_{2.5}$ and O_3 concentrations, which emphasizes the impact of the complexity of topography and land-use patterns on CMAQ modeling. Bias quantification results showed that the bias is more affected by boundary conditions and local meteorology than other inputs in the study region, implying that the CMAQ model still has imperfect mechanisms to well simulate $PM_{2.5}$ and O_3 with boundary conditions and local meteorology.

The study illustrates exemplary cooperation between CTM and machine learning methods. The firstly developed bias quantification technique can provide a bias structure for numerical modeling and serve as an assessment tool to improve embedded algorithms and input data quality.

Chapter 4. Examining Ozone Response Modeling to NOx and VOC Emissions

4.1. Abstract

Current machine learning (ML) applications in atmospheric science focus on forecasting and bias correction for numerical modeling estimations, but few studies examined the nonlinear response of their predictions to precursor emissions. This study uses ground-level maximum daily 8-hour ozone average (MDA8 O_3) as an example to examine O_3 responses to local anthropogenic NOx and VOC emissions in Taiwan by Response Surface Modeling (RSM). Three different datasets for RSM were examined, including the Community Multiscale Air Quality (CMAQ) model data, ML-measurement-model fusion (ML-MMF) data, and ML data, which respectively represent direct numerical model predictions, numerical predictions adjusted by observations and other auxiliary data, and ML predictions based on observations and other auxiliary data.

The results show that both ML-MMF (R=0.93-0.94) and ML predictions (R=0.89-0.94) present significantly improved performance in the benchmark case compared with CMAQ predictions (R=0.41-0.80). While ML-MMF isopleths exhibit O_3 nonlinearity close to actual responses due to their numerical base and observation-based correction, ML isopleths present biased predictions concerning their different controlled ranges of O_3 and distorted O_3 responses to NOx and VOC emission ratios compared with ML-MMF isopleths, which implies that using data without support from CMAQ modeling to predict the air quality could mislead the controlled targets and future trends. Meanwhile, the observation-corrected ML-MMF isopleths also emphasize the impact of transboundary pollution from mainland China on the regional O_3 sensitivity to local NOx

and VOC emissions, which transboundary NOx would make all air quality regions in April more sensitive to local VOC emissions and limit the potential effort by reducing local emissions.

Future ML applications in atmospheric science like forecasting or bias correction should provide interpretability and explainability, except for meeting statistical performance and providing variable importance. Assessment with interpretable physical and chemical mechanisms and constructing a statistically robust ML model should be equally important.

4.2. Introduction

Air pollution has gained great attention owing to its adverse effects on human health [3,33], climate [170], agriculture (Tai and Martin, 2017), ecosystems [174], and visibility [175]. In regional air quality management, controlling local anthropogenic emissions is a common way to improve regional air quality. Predicting air quality under designed emission control scenarios by using chemical transport models (CTMs) like the Community Multiscale Air Quality (CMAQ) model has been much studied (Arnold and Dennis, 2006; Che et al., 2011).

Moreover, to meet the prompt and various needs of policymakers, response surface modeling (RSM) was developed to assess the improved or changed air quality based on designed emission control strategies without extra CTM simulations. That is, RSM can retrieve the nonlinear equation between ambient air pollutant concentrations (e.g. O_3) and multiple precursors emissions (e.g. NOx and VOCs) from multiple emission sources based on a series of CTM simulations [178–180], and the user can apply the retrieved nonlinear equation to timely estimate the changed air pollutant concentrations based on input emission ratios and support emission control strategies. For example, RSM has been intensively applied to assess the NOx and VOC emission control strategies of O_3 pollution [40,180,181]. RSM can identify ambient O_3 sensitivity to NOx and VOC emission by O_3 isopleth and further classify the isopleth regimes into two chemical regimes, NOx-limited and VOC-limited. In the NOx-limited regimes, O_3 increases with increased NOx emissions and exhibits limited response to VOC emissions, and vice versa. Classification of O_3 formation regime in the isopleth can assist policymakers determine whether NOx or VOC emissions should be controlled preferentially in emission control strategies [35]. However, although RSM was employed to improve regional air quality in several previous studies, these studies were still based on simulated results and neglected the bias between the modeled estimations and observations in the benchmark case [38–40], which could largely affect the nonlinearity between pollutants and precursor emission changes.

To forecast air quality and support air quality policies, machine learning (ML) or machine intelligence has been rapidly developed and intensively implemented in environmental science and air quality management [4,145]. Technically, ML is driven by monitoring and/or measurement data, so it is relatively easy to execute and can provide more accurate predictions compared with CTMs, which still need complicated data-preparing processes and computationally-intensive time and resources, and have a larger modeling bias. However, ML is also debated and remains low persuasiveness due to its black-box modeling process and failure to provide interpretability and explainability concerning physical/chemical mechanisms [4,41].

In previous applications, ML can serve as a bias corrector to adjust modeling results. Several measurement-model fusion (MMF) [41] techniques in post-analysis have been developed in recent years to adjust CTM results based on observations [18,42–44]. Air pollutant estimations without correction by observations could underestimate or overestimate improved air quality and derived environmental benefits [41]. In addition, ML can also forecast air quality based on historical observations and other auxiliary data (e.g. meteorological and land-use data) without involving CTM results and still have good performance (Ausati and Amanollahi, 2016; Song et al., 2015; Zhou et al., 2019). However, whether ML either serves as a bias corrector or a forecaster, few ML studies examined pollutants’ sensitivity to their precursor emissions based on observation-corrected results.

In this study, we used the maximum daily 8-hour O_3 average (MDA8 O_3) as the target index and selected Taiwan as the study region due to its island geography, high-density air quality monitoring stations (n=73), and three-year-updated emission inventory. The goal of this study is to (1) verify the capability of ML to correct CMAQ modeling results (denoted as ML-MMF data) and predict MDA O_3 without CMAQ-modeled results (denoted as ML data) in the benchmark-case modeling performance and (2) examine O_3 nonlinear responses to all anthropogenic

NOx and VOC emission ratios by O_3 isopleths by using CMAQ, ML-MMF, and ML modeling results.

4.3. Methodology

The technical work is illustrated in Figure 4-1. Three types of data (CMAQ, ML-MMF, and ML) were prepared for RSM; CMAQ data were direct outputs from CMAQ-modeled results; ML-MMF data are the corrected estimates by observation and CMAQ inputs (emissions, boundary conditions, meteorology, and land-use), and the constructed model was employed to predict O_3 concerning the changed CMAQ outputs, the other CMAQ inputs and changed emissions (kg/day) under different emission scenarios; ML outputs are predictions constructed by using O_3 observations and CMAQ inputs, and the constructed model was utilized to predict O_3 concerning the changed emissions (kg/day) and the other CMAQ inputs under different emission scenarios. Second, RSM was executed for each dataset to predict O_3 under different NOx and VOC emission ratios in proportion to the baseline emissions (ratio=1) in the benchmark case. O_3 isopleths were finally constructed for each dataset and were validated by out-of-samples and observations.

4.3.1. Data preparation

Air quality zones in this study were categorized into six regions: Northern (NT), Chu-Miao (CM), Central (CT), Yun-Chia-Nan (YCN), Kao-Ping (KP), and Eastern (ET). A total of 73 air quality monitoring stations with hourly O_3 measurements were included (Figure 4-2). Modeling period included January, April, July, and October 2016 to represent different seasons. Meteorological fields were firstly simulated by WRF (version 3.8), and hourly O_3 concentrations were simulated by the CMAQ model (version 5.2) with the gas-phase chemistry module, Carbon Bond 6, [165] and aerosol module, AERO6 [166] mechanisms. The configurations of the simulation domain nested four layers from East Asia (81 km × 81 km) to Taiwan island (3 km × 3 km) which covers 90 (row) × 135 (column) horizontal grid cells (Lai and Lin, 2020). Daily emission data from Taiwan Emission Data System (TEDS) version 10.0 with 3 km × 3 km resolution developed by Taiwan EPA including industrial, mobile, fugitive, and biogenic sources were used. Before CMAQ modeling, emission distribution and speciation were processed by the USEPA SMOKE program [182], which can apply temporal and spatial allocation and chemical speciation for industrial, mobile, and fugitive sources from TEDS. The modeling performance assessment of meteorology factors and CAMQ-modeled O_3 are shown in Table 4-1 and Table 4-2, respectively.

The details of developed ML-MMF and ML framework were illustrated in Chapter 3, and the major difference between ML-MMF and ML framework is predicting O_3 with and without CMAQ direct outputs for prediction, as shown in the following conceptual equations:

$$\mathrm{MLMMF\ \ O_3\ \ predictions = f(CMAQ, Emis, BC, Met, Land)}$$

$$\mathrm{ML\ \ O_3\ \ predictions = f(Emis, BC, Met, Land)}$$

where CMAQ is CAMQ output; Emis are emission data (kg/day); BC are boundary conditions; Met are meteorological variables; Land are land-use variables. Basically, Both of ML-MMF and ML models employed five techniques including the k-nearest neighbors' regression (KNN) [163], regression tree (RT) [9], random forest (RF) [10], gradient boosted tree models (GBM) (Natekin and Knoll, 2013), and convolutional neural network (CNN) [164], which can construct the nonlinearity between input variables (emissions, boundary conditions, meteorology, and land-use data) and observed O_3 concentrations in the benchmark case. The best model with higher accuracy, reasonable spatial predictions, and no overfitting tendency was selected to predict O_3 concentrations (Appendix I); 60% and 40% of the data set were selected as the training dataset and the testing dataset, respectively; The 10-fold cross-validation was also conducted to optimize hyperparameters based on the uncertainty of modeling performance. The selected model for ML-MMF and ML would be further utilized to predict O_3 based on CMAQ outputs (not for the ML model), the changed emissions (kg/day), and the other fixed auxiliary data (boundary conditions, meteorology, and land-use data) under different emission scenarios.

The variables selected for ML-MMF and ML were related to emissions of precursor species including NOx and VOCs, boundary conditions that affect the background level of O_3, NOx, and VOCs, meteorological factors involved with photochemical reactions and transport fluxes of air, and time-independent land-use geographical information (Table 4-3). Meteorological factors on 850 hPa and 690 hPa represent the weather conditions of the mixing layer and lower troposphere layer [13]. MDA8 O_3 would be the dependent variable because O_3 concentrations usually are higher during the daytime due to the existence of ultraviolet energy [183]. For each month, the modeling performance was evaluated by correlation coefficient (r), mean absolute error (MAE), and root mean square error (RMSE) as shown in the following equation.

$$\mathrm{r} = \sqrt{\frac{\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\left(M_{ij} - \bar{M}\right)\left(O_{ij} - \bar{O}\right)\right]}{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(M_{ij} - \bar{M}\right)^2 \sum_{i=1}^{n}\sum_{j=1}^{m}\left(O_{ij} - \bar{O}\right)^2}}$$

$$\mathrm{MAE} = \frac{\left|M_{ij} - O_{ij}\right|}{N \times M}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(M_{ij} - O_{ij}\right)^2}{N \times M}}$$

where M_{ij} is the modeled O_3 for ith station in jth day; O_{ij} the observed O_3 for ith station in jth day; $\bar{M}$ is the average of modeled O_3 for all stations; $\bar{M}$ is the average of observed O_3 for all stations; N is the number of stations; M is the number of days.

4.3.2. Response surface modeling (RSM)

RSM can retrieve non-linear O_3 responses to anthropogenic NOx and VOC emission ratios, which are changed ratios of emission compared with baseline emission in the benchmark case (emission ratio=1). First, to generate the control matrix of anthropogenic NOx and VOC emission ratios for each air quality region, the Latin hypercube sampling (LHS) method was utilized to estimate enough sample size of the number of CMAQ simulations for RSM while providing enough statistical power and saving computing resources.

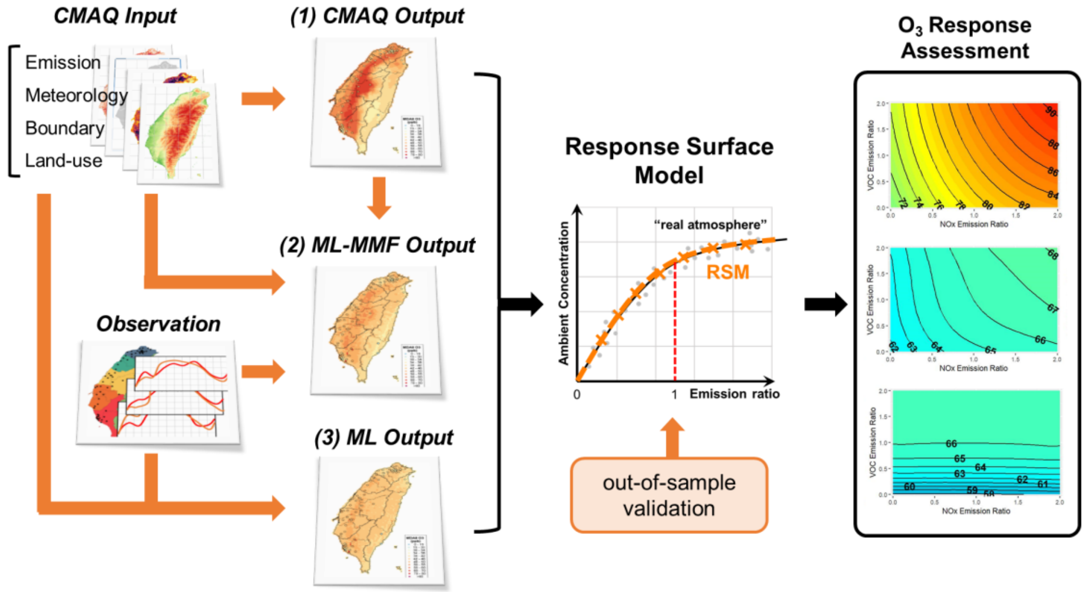


Figure 4-1. Technical flowchart of this study

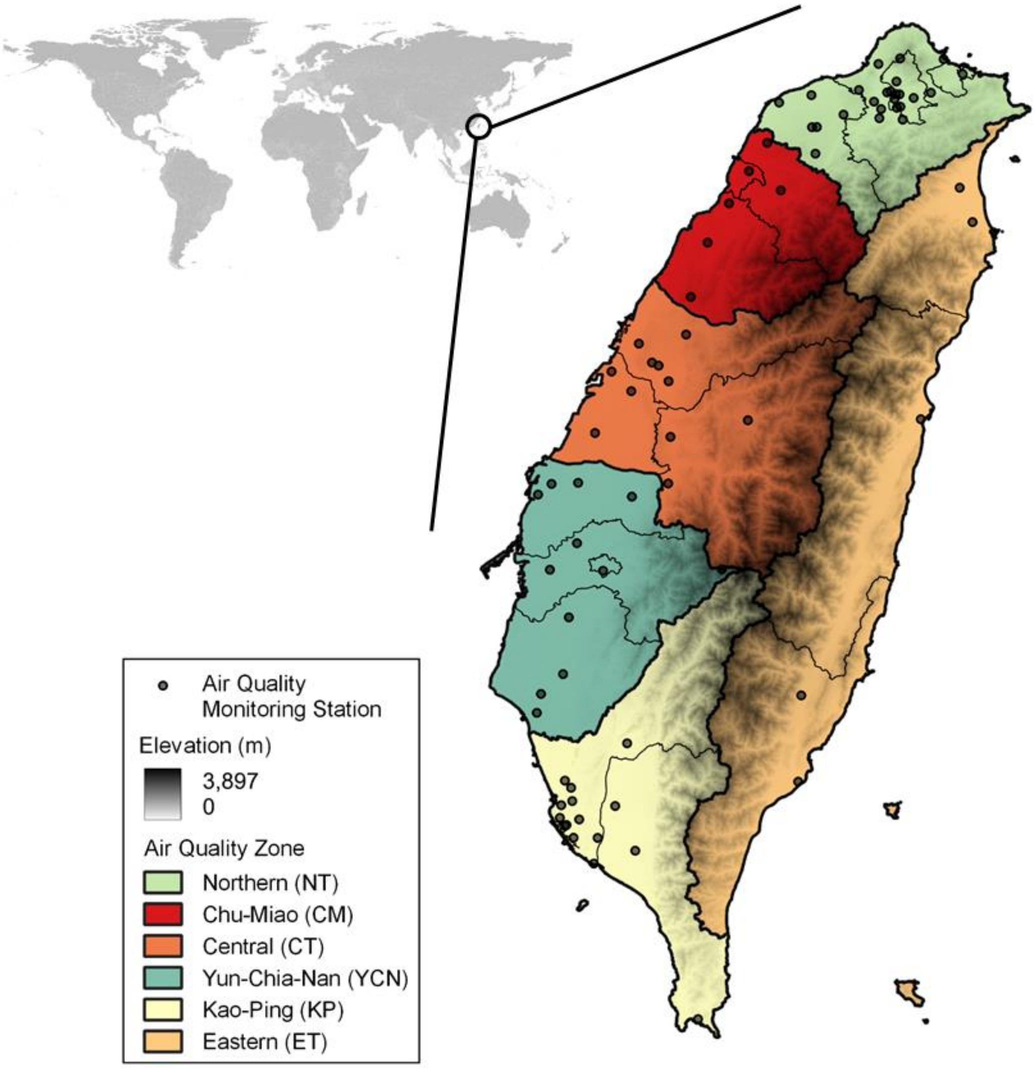


Figure 4-2. Air quality monitoring stations (n=73) and six air quality zones in Taiwan

Table 4-1. Modeling performance assessment of temperature, wind speed, and wind direction in 2016

Region/Month		Temperature		Wind speed		Wind direction	
		MB[1]	**MAGE[2]**	**MB[1]**	**RMSE[3]**	**NMB[4]**	**NME[5]**
Northen/ Chu-Miao/ Yilan	January	0.17	0.64	0.40	1.19	1%	16%
	April	0.30	0.89	0.43	1.09	-1%	18%
	July	0.38	0.84	0.54	1.28	1%	17%
	October	0.01	0.65	0.48	1.42	-2%	15%
Central	January	0.50	0.78	0.36	1.03	2%	7%
	April	0.91	1.53	0.24	1.04	3%	13%
	July	0.30	1.20	0.23	0.84	5%	12%
	October	0.22	0.89	0.24	0.87	0%	13%
Yun-Chia	January	0.36	1.22	0.27	0.86	-3%	6%
	April	1.07	1.76	0.31	1.13	-2%	15%
	July	0.18	0.83	0.01	1.01	2%	14%
	October	-0.31	0.97	0.27	0.82	2%	14%
Southern (Tainan/Kao-Ping)	January	0.32	1.09	-0.14	1.06	-2%	7%
	April	0.42	1.14	0.12	0.96	-3%	12%
	July	0.12	0.98	0.19	1.08	-3%	12%
	October	-0.08	1.01	-0.04	0.98	-2%	11%
Hua-Tung	January	0.28	0.64	0.41	1.09	7%	13%
	April	0.06	0.58	0.31	1.09	4%	18%
	July	-0.26	0.56	0.33	1.15	-1%	14%
	October	0.04	0.52	0.27	0.87	4%	13%

[1] Mean Bias, MB

$$\mathrm{MB} = \frac{1}{M \times N}\sum_{k=1}^{M}\sum_{i=1}^{N}\left(P_{i,k} - O_{i,k}\right)$$

[2] Mean Absolute Gross Error, MAGE

$$\mathrm{MB} = \frac{1}{M \times N}\sum_{k=1}^{M}\sum_{i=1}^{N}\left|P_{i,k} - O_{i,k}\right|$$

[3] Root Mean Square Error, RMSE

$$\mathrm{MB} = \left[\frac{1}{M \times N}\sum_{k=1}^{M}\sum_{i=1}^{N}\left(P_{i,k} - O_{i,k}\right)^2\right]^{1/2}$$

[4] Normalized Mean Bias, NMB

$$\mathrm{NMB} = \frac{\sum_{k=1}^{M}\sum_{i=1}^{N}\left(P_{i,k} - O_{i,k}\right)}{M \times N \times 360°} \times 100\%$$

[5] Normalized Mean Error, NME

$$\mathrm{NME} = \frac{\sum_{k=1}^{M}\sum_{i=1}^{N}\left|P_{i,k} - O_{i,k}\right|}{M \times N \times 360°} \times 100\%$$

where $P_{i,k}$ is the modeled temperature or wind speed of ith hour for k stations; $O_{i,k}$ is the observed temperature or wind speed of ith hour for k stations; N is the number of total hours; M is the number of monitoring sites.

* Taiwan EPA and U.S. EPA performance criteria:

MB (K) ±1.5, RMSE (K) <3.0, MB (m/s) ±1.5, RMSE (m/s) <3, NMB (%) ±10, NME(%) <30

Table 4-2. Modeling performance assessment of CMAQ-modeled O3 in 2016

Air Quality Region	**O_3**			
	MB[1]	**MNB[2]**	**MNE[3]**	**r[4]**
Northern	-9%	-11%	11%	0.38
Chu-Miao	-4%	-9%	11%	0.64
Central	7%	1%	9%	0.79
Yun-Chia-Nan	-1%	-5%	9%	0.76
Kao-Ping	4%	-1%	10%	0.78
Yilan	-4%	-6%	6%	0.50
Hua-Tung	2%	-3%	6%	0.38
All regions	-36%	3%	4%	0.77

[1] Mean bias, MB

$$\mathrm{MB} = \frac{1}{M \times N}\sum_{k=1}^{M}\sum_{j=1}^{N}\left(\frac{Max_{i=1}^{24}(P_{i,j,k}) - Max_{i=1}^{24}(O_{i,j,k})}{Max_{i=1}^{24}(O_{i,j,k})}\right)$$

where $P_{i,j,k}$ is the modeled estimates of ith hour in the jth day for k stations; $O_{i,j,k}$ is the observations of ith hour in the j day for k stations; N is the number of total hours (or days); M is the number of monitoring sites.

[2] Mean Normalized Bias, MNB

$$\mathrm{MNB} = \frac{1}{M \times N}\sum_{k=1}^{M}\sum_{i=1}^{N}\left(\frac{P_{i,k} - O_{i,k}}{O_{i,k}}\right)$$

[3] Mean Normalized Error, MNE

$$\mathrm{MNE} = \frac{1}{M \times N}\sum_{k=1}^{M}\sum_{i=1}^{N}\left|\frac{P_{i,k} - O_{i,k}}{O_{i,k}}\right|$$

where $P_{i,k}$ is the modeled estimates of jth day for k station; $O_{i,k}$ is the observations of jth day for k station.

[4] Correlation coefficient, r

$$\mathrm{r} = \frac{1}{M \times N}\sum_{k=1}^{M}\sum_{j=1}^{N}\left[\frac{(P_{i,k} - \bar{P})(O_{i,k} - \bar{O})}{S_P S_O}\right]$$

where $\bar{P}$ is the averaged predictions of all sites in the modeling region; $\bar{O}$ is the averaged observations of all sites in the modeling region; S_P is the standard deviation of hourly predictions from all sites in the modeling region; S_O is the standard deviation of hourly observations from all sites in the modeling region

* Taiwan EPA and U.S. EPA performance criteria: MB ± 10%, MFB ± 15%, MFE ± 35%, r > 0.45

Table 4-3. Selected variables for ML-MMF and ML model

Dataset	Source	Variables	
Emission	TEDS 10	NOx and VOCs emissions from point, mobile, fugitive, and biogenic sources.	
Boundary conditions	WRF	air temperature (850 and 690 hPa), relative humidity (850 and 690 hPa), NO_3, HNO_3, N_2O_5, NO, NO_2, VOCs and O_3 at surface level	
Meteorological variables	WRF	Surface	surface pressure, PBL height, temperature at 2 m, mixing ratio at 2 m, wind speed, U wind component and V wind component, solar radiation reaching ground, precipitation, total cloud fraction, average liquid water content of cloud
		Pressure level (850 and 690 hPa)	air temperature, potential vorticity, vertical velocity, U wind component and V wind component
Land-use	CMAQ	evaluation, urban percent	

The LHS design can provide flexibility, which selects a number of runs based on limited computing resources [34], and also can capture the nonlinearity of ozone. The number of CMAQ runs for RSM was decided by the emission ratio range (0%-200%), the accuracy of O_3 response to NOx and VOC emissions, and the available computing resources. Typically, the number of CMAQ runs for RSM around 50 simulations is enough for constructing a statistically robust model [40]. Extra simulations for individual air quality regions were also conducted due to higher O_3 concentrations in spring, fall, and winter. Emission ratios of NOx and VOC are designed from 0% to 200%, and the number of CMAQ runs for six air quality regions is shown in Table 4-4.

In each air quality region, RSM involves multiple precursor emissions (NOx and VOC) and anthropogenic emission sectors (industrial, mobile, and fugitive sources) from multiple cities and counties, as shown in Figure 4-3 and Table 4-5. Multiple city/county-level NOx and VOC emission sectors (Table 4-5) were used for each air quality region. Self-adaptive RSM (SA-RSM) based on regression method and stepwise selection was employed to predict O_3 based on designed emission ratios of NOx and VOC, and the followed multidimensional kriging method was used to illustrate O_3 isopleths to show O_3 nonlinear response to NOx and VOC emission ratios [34,40,180,184]. In each air quality region, the averages of grid cells with daily MDA8 O_3 higher than 60 ppb were used for RSM based on WHO and Taiwan EPA standards [28]. Also, higher O_3 concentrations were identified to be more sensitive to anthropogenic emissions of NOx and VOC [36]. In each region, the following equation linking the O_3 concentration ($\Delta\mathrm{Conc}$) in each grid to city/county-level NOx or VOC emission ratios of the emission sectors and stepwise selection that can automatically select polynomial variables for all grid cells with 0.15 of the entering level and the leaving level were utilized [40]:

$$\Delta\mathrm{Conc}(m,n)=\sum_{i=1}^{k}X_i\cdot(\Delta E_i)^{a_i}+\sum_{i=1}^{k}\sum_{j=1}^{k-1}X_{ij}\cdot(\Delta E_i)^{b_i}\left(\Delta E_j\right)^{c_j}$$

where $\Delta\mathrm{Conc}$ is the changed daily MDA8 O_3 concentration (response) from the baseline scenario (the benchmark case, emission ratio=1) in the grid(m,n); ΔE_i is the changed emission ratios from 1 of k city/county-level NOx or VOC emission sectors; ΔE_j is the other changed emission ratios from 1 of emission sector other than ΔE_i; a_i, b_i, and c_j are the nonnegative integer powers of ΔE_i and interaction terms of ΔE_i and ΔE_j. We set a_i from 1 to 3 and b_i and c_j from 1 to 2, respectively. A total of 2782 CMAQ simulations were used for RSM. To assure the performance of RSM, a total of 107 CMAQ simulations were used for out-of-sample validation (Table 4-4). The RSM performance was evaluated by correlation coefficient (r), Mean normalized error (Mean NE), and Maximum NE (Max NE) [180] as shown in the following equations.

$$\mathrm{r}=\sqrt{\frac{[\sum_{i=1}^{N}(M_i-\bar{M})(O_i-\bar{O})]}{\sum_{i=1}^{N}(M_i-\bar{M})^2\sum_{i=1}^{N}(O_i-\bar{O})^2}}$$

$$\mathrm{Mean\ \ NE}=\frac{1}{N}\sum\nolimits_{i=1}^{N}\frac{|M_i-O_i|}{O_i}$$

$$\mathrm{Max\ \ NE}=\max\left(\frac{|M_i-O_i|}{O_i}\right)$$

where M_i and O_i are the average of grid cells for RSM predictions and CAMQ, ML-MMF, or ML predictions in the ith simulation over different emission scenarios ; N is the number of scenarios; $\bar{M}$ and $\bar{O}$ are the average of the RSM predictions and CAMQ, ML-MMF, or ML predictions in the ith simulation over different emission scenarios. Finally, the response of O_3 concentrations to changes in anthropogenic NOx and VOC emission ratios from all sources and all cities and counties in each air quality region was illustrated by O_3 isopleths, which can be divided into NOx-limited (NOx-sensitive and VOC-rich) and VOC-limited (VOC-sensitive and NOx-rich) regimes and used to help determine whether NOx or VOC emissions should be controlled more aggressively in strategies to alleviate ground-level O_3 concentrations [35].

4.4. Results and discussion

4.4.1. Benchmark case modeling performance

The performance of CMAQ, ML-MMF, and ML modeling compared with observations is shown in Table 4-6Table 4-6. Compared with ML-MMF and ML predictions, the CMAQ predictions have lower performance for all months (R=0.41-0.80, RMSE=13.45-21.19). After CMAQ data were adjusted by the auxiliary data (emission, meteorology, boundary condition, and land-use data) and corrected by observations, ML-MMF modeling presents significantly better performance for all months (R=0.93-0.94, RMSE=4.49-7.43). The improved performance highlights the benefits of adding auxiliary data in ML-MMF applications. Even though excluding CMAQ output and purely using auxiliary data for ML modeling, the ML model still maintains comparable modeling performance (R=0.89-0.94, RMSE=4.62-8.94).

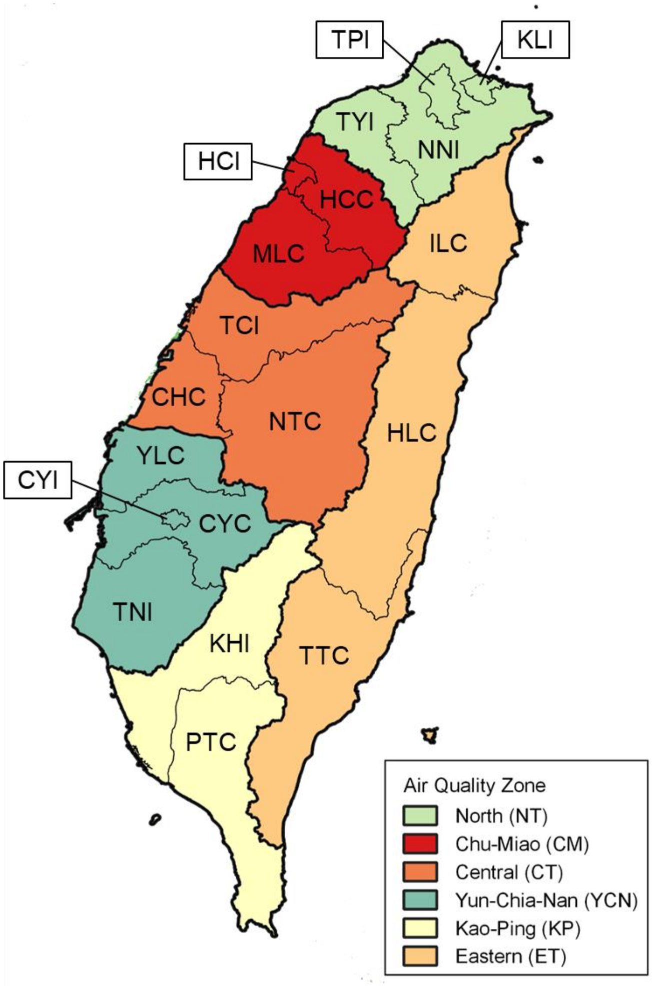


Figure 4-3. City-level and county-level subcategories in air quality regions.

Table 4-4. Finished CMAQ simulations for six air quality zones

Zone	Training samples				Out of samples	Total
	Jan	April	July	October		
NT	200	200	62	200	21	683
CM	150	150	47	150	21	518
CT	150	47	47	47	15	306
YCN	200	62	62	62	15	401
KP	120	98	62	62	14	356
ET	150	150	47	150	21	518
Total	970	707	327	671	107	2782

Table 4-5 Types and Number of emission sectors in six air quality regions

Region	City/County	Species	Source	# Emission sector
Northern (NT)	KLI	NOx	industrial	24
	TPI	VOC	mobile	
	NNI		fugitive	
	TYI			
Chu-Miao (CM)	HCC	NOx	industrial	18
	HCI	VOC	mobile	
	MLC		fugitive	
Central (CT)	TCI	NOx	industrial	18
	CHC	VOC	mobile	
	NTC		fugitive	
Yun-Chia-Nan (YCN)	YLC	NOx	industrial	24
	CYI	VOC	mobile	
	CYC		fugitive	
	TNI			
Kao-Ping (KP)	KHI	NOx	industrial	12
	PTC	VOC	mobile	
			fugitive	
Eastern (ET)	ILC	NOx	industrial	24
	HLC	VOC	mobile	
	TTC		fugitive	

Both ML-MMF and ML have no overfitting tendency considering their similar performance in training and testing data. Averages of observed and modeled MDA8 O_3 for CMAQ, ML-MMF, and ML in selected months are presented in Figure 4-4. Figure 4-4(a) shows obvious overestimations of CMAQ compared with observations, especially in the regions along with the west side of the mountains and the central basin. Figure 4-4(b)(c) shows the ML-MMF and ML estimations are much close to observations and have lower modeling bias (RMSE and MAE) compared with CMAQ.

Monthly concentrations of CMAQ, MMF, and ML are presented in Figure 4-5. Among all selected months, Figure 4-5(a) shows CMAQ remain overestimated in all months and has spatial specificity in each month, in which CT, YCN, and KP region are overestimated in January, NT, CM, and CT regions are overestimated in April, and whole west regions overestimated in July and October. On the other hand, Figure 4-5(b)(c) shows the ML-MMF model and ML model present similar spatial distributions for all months except for the ET region in July and October. The different estimations in the ET region are due to fewer air quality monitoring stations in the region, so the ML model needs to learn the data from the monitoring sites in the western regions.

4.4.2. Adjusted O_3 seasonal patterns

In Taiwan, frequent long-range transboundary pollution events from mainland China with higher air pollutant concentrations occur in spring, fall, and winter when northeast monsoon prevails [169,170]. Our method to classify event and non-event days is illustrated in Appendix III. In the modeling period, there were 14, 13, and 3 event days in January, April, and October, respectively, and no event days occurred in July. Figure 4-6 presents the monthly MDA8 O_3 average of event and non-event days based on ML-MMF estimations. In January, transboundary plumes mostly contained higher PM, NOx, and CO, so there is no significant O_3 concentration difference between the event and non-event days. In April, the MDA8 O_3 averages of 13 event days show a significant impact on the whole island. Compared with event days, O_3 exceedances on non-event days mostly occur in central mountainous areas, which may result from higher biogenic VOC emissions in spring [185]. In July, higher O_3 level in northern Taiwan is due to the prevalence of southwest and south monsoon, and NOx, VOC, and formatted O_3 transport from southern regions accumulated in northern regions. In October, both event days and non-event days show similar distribution, but event days present higher O_3 concentration in western areas, especially in the NT region.

4.4.3. RSM performance

Because higher O_3 concentrations are more sensitive to anthropogenic NOx and VOC emissions [36], we performed RSM for O_3 exceedance days (MDA8 O_3 > 60ppb) for April, July, and October and excluded January due to no O_3 exceedance days. RSM performances by using CMAQ, ML-MMF, and ML data for six air quality regions (NT, CM, CT, YCN, KP, and ET) of the event and non-event days in April, July, and October are illustrated in Table 4-7. Most of the RSM results meet the statistical requirement of mean NE < 3% and max NE <10% considering r (0.855-1.000), mean NE (0.02%-2.62%), and max NE (0.08%-9.88%), except for the CM region's October (max NE=11.48%) and the YCN

Table 4-6. Modeling performance of CMAQ, ML-MMF, and ML of benchmark case in selected months in 2016

Data source		Index	Month			
			January	April	July	October
CMAQ		R	0.41	0.53	0.79	0.80
		RMSE	13.45	17.87	21.19	19.09
		MAE	9.72	13.61	17.33	14.54
ML-MMF	Train	R	0.93	0.93	0.93	0.93
		RMSE	4.54	6.83	5.94	7.43
		MAE	3.36	5.14	4.26	5.34
	Test	R	0.93	0.93	0.93	0.94
		RMSE	4.49	6.74	6.19	7.19
		MAE	3.41	5.08	4.51	5.11
ML	Train	R	0.92	0.92	0.90	0.90
		RMSE	4.66	7.18	7.04	8.94
		MAE	3.49	5.35	5.14	6.58
	Test	R	0.93	0.92	0.89	0.90
		RMSE	4.62	7.10	7.35	8.94
		MAE	3.49	5.41	5.41	6.40

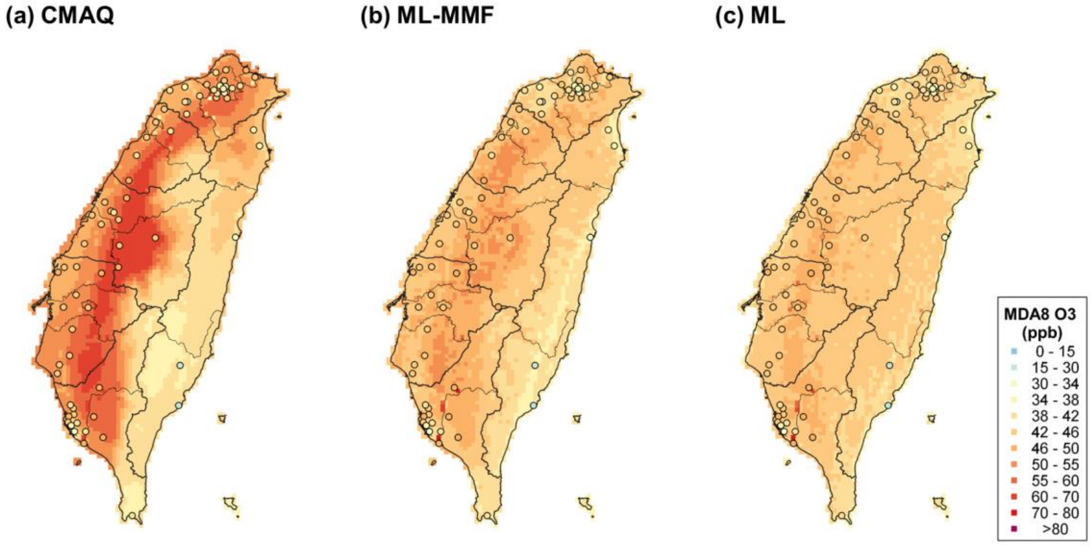


Figure 4-4. Observed MDA8 O_3 (circles) and modeled estimations of (a) CMAQ, (b) ML-MMF, and (c) ML

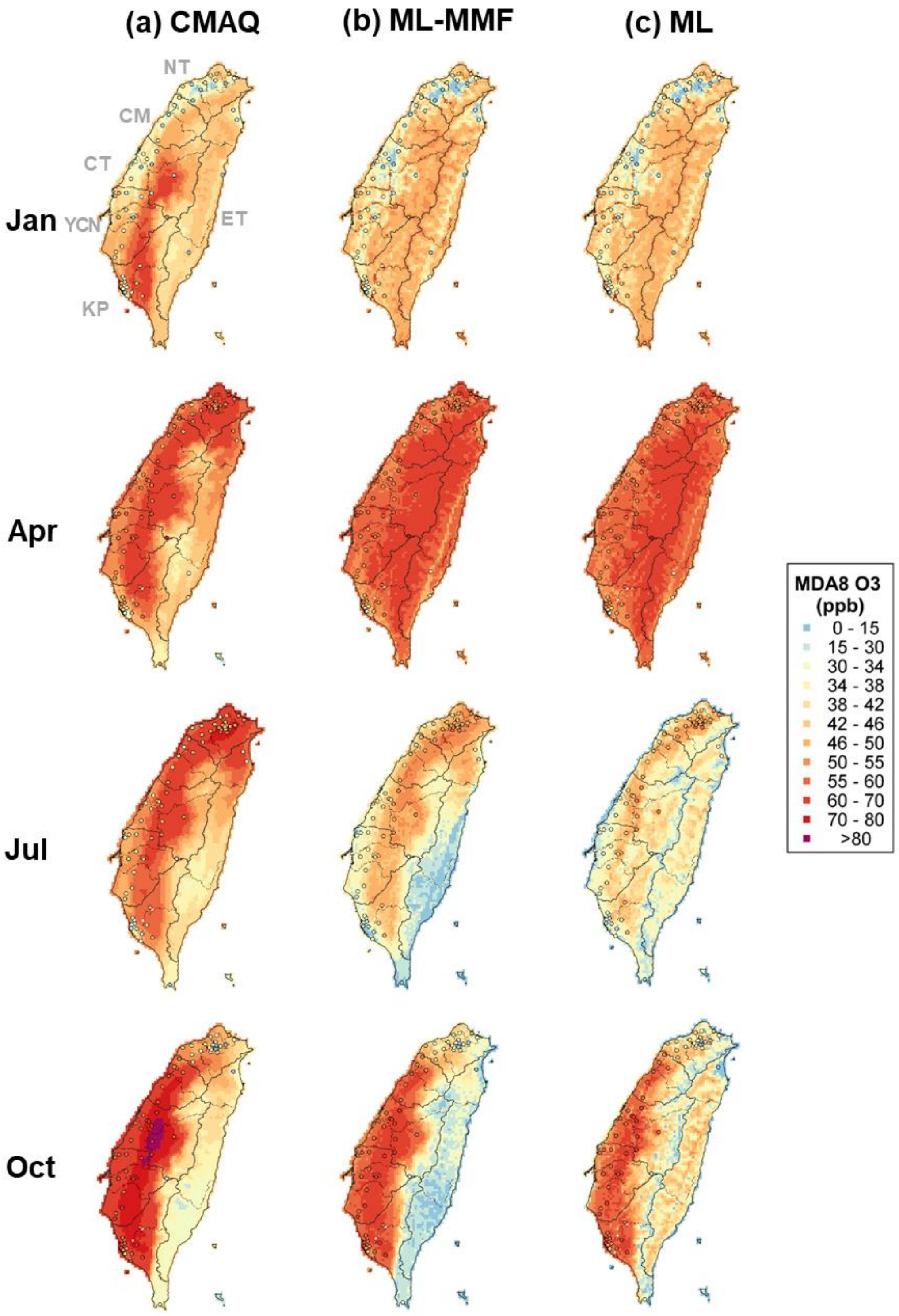


Figure 4-5. Monthly observed MDA8 O_3 (circles) and modeled estimations of (a) CMAQ, (b) ML-MMF, and (c) ML

region’s April (max NE=11.38%) and October (max NE=12.04%) of CMAQ data and KP’s April (R=0.759) of ML-MMF data. Good RSM performance represents that RSM can reproduce CMAQ, ML-MMF, or ML outputs by only using county/city-level NOx and VOC emission ratios without extra CMAQ simulations or ML-MMF/ML modeling. The higher max NE may be due to unstable RSM performance when NOx and VOC emission ratios are very low (0%) or very high (200%) [186]. Although the limited performance could be improved by adding more simulations under these extreme emission ratios, these scenarios are still impractical and hardly verified by observations. For example, it is almost impossible to measure pollutant concentrations without any anthropogenic emissions in urban areas. Furthermore, the validation of the baseline emission ratio (emission ratio=1, Figure 4-7) compared with observations shows a satisfactory performance (Mean NE=2%-6%). Second, the higher bias may imply the high impact of non-local emissions, which means air pollutants or their precursors from upwind countries or other air quality zones dominate downwind local air quality, so local emissions only have a limited impact on local air quality and hardly explain the temporal variation of observed concentrations. As previously mentioned, in April and October, long-range transboundary air pollution transported from mainland China and plumes from upwind regions frequently accumulate and deteriorate air quality in southern air quality regions.

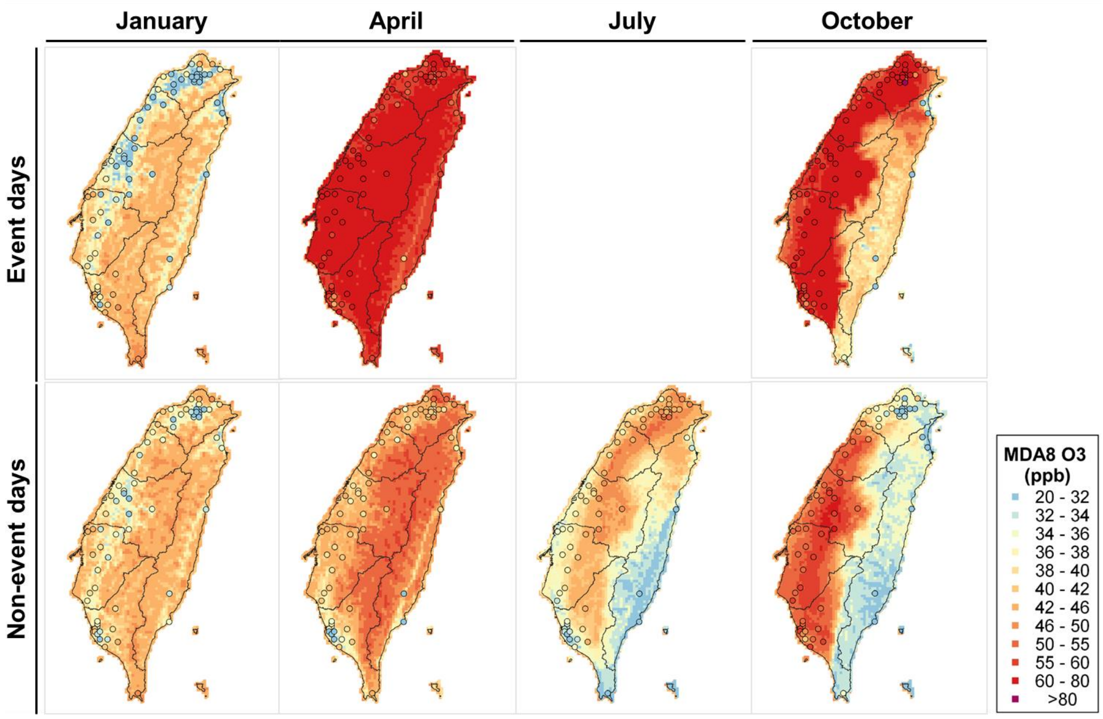


Figure 4-6. Monthly MDA8 O_3 concentrations for event and non-event days in selected months by ML-MMF data. No event days occurred in July.

4.4.4. O_3 sensitivity to local NOx/VOC emissions

Monthly combined O_3 isopleths of all air quality regions in April, July, and October for the event and non-event days by using CMAQ, ML-MMF, and ML data are presented in Figure 4-8, and the O_3 isopleths for the individual region are shown in Figure 4-9 - Figure 4-13. All isopleths for each air quality region show the averages of grid cells having MDA8 O_3 concentrations higher than 60 ppb in the region, and the x-axis and y-axis represent the NOx and VOCs emission ratios for all sources in all cities and counties in the region.

First, ML-MMF isopleths are closer to the real O_3 nonlinear responses to anthropogenic NOx and VOC emissions due to its observation-based adjustment, and ML-MMF isopleths show the real improved effort by reducing local emissions should be less than what CMAQ model simulated. For example, the CMAQ isopleth in July ranges from 72 ppb to 90 ppb concerning different emission ratios, but the ML-MMF isopleth presents a narrower range, which is from 62 ppb to 68 ppb. The reduced range of isopleths emphasizes the importance of adjustment by observations, and CMAQ-based RSM results may bias the improved air quality by emission control strategies.

Second, the ML-MMF O_3 isopleths show changed O_3 sensitivity after fusing with observations. For example, CMAQ isopleths in April with a combined regime of NOx-limited and VOC-limited trends (Figure 4-8 (a)) are corrected to VOC-limited in ML-MMF isopleths (Figure 4-8 (b)) for the event and non-event days. If further comparing CMAQ modeled NOx and VOCs with observations (Figure 4-17), CMAQ modeled estimations in the

Table 4-7. RSM performance of event and non-event days in selected months by CMAQ, ML-MMF, and ML data for six air quality regions

Region	Month	CMAQ			ML-MMF			ML		
		R	Mean NE	Max NE	R	Mean NE	Max NE	R	Mean NE	Max NE
Event days										
NT	April	0.999	0.65%	2.98%	0.999	0.34%	3.82%	1.000	0.27%	1.11%
	October	0.994	1.94%	6.74%	0.998	0.93%	2.88%	0.992	0.46%	1.01%
CM	April	0.999	0.48%	2.41%	1.000	0.17%	0.72%	1.000	0.11%	0.35%
	October	0.949	2.20%	8.14%	0.962	1.18%	3.60%	0.929	1.29%	4.79%
CT	April	0.997	0.36%	1.87%	1.000	0.25%	1.07%	1.000	0.16%	0.33%
	October	0.997	0.65%	1.26%	0.995	0.62%	0.98%	0.891	0.83%	1.33%
YCN	April	0.990	1.15%	7.01%	0.999	0.57%	4.19%	1.000	0.17%	0.52%
	October	0.991	2.24%	4.88%	0.997	0.56%	1.75%	0.962	0.51%	1.52%
KP	April	0.984	2.42%	8.79%	0.993	1.22%	9.53%	1.000	0.06%	0.24%
	October	0.996	0.49%	0.82%	0.984	0.91%	1.72%	0.938	0.76%	1.66%
ET	April	1.000	0.20%	3.58%	1.000	0.17%	1.79%	1.000	0.07%	0.20%
	October	0.999	0.34%	2.59%	0.992	1.04%	7.43%	0.981	0.85%	1.82%
Non-event days										
NT	April	1.000	0.43%	1.85%	0.984	0.29%	2.76%	0.993	0.19%	0.73%
	July	0.982	1.61%	9.21%	0.950	1.38%	9.88%	0.855	1.29%	6.26%
	October	0.997	1.28%	4.62%	0.993	1.01%	4.92%	0.987	0.53%	2.88%
CM	April	1.000	0.30%	3.11%	0.958	0.21%	1.69%	1.000	0.03%	0.24%
	July	0.999	0.29%	1.07%	0.975	0.58%	4.71%	0.979	0.68%	1.97%
	October	0.990	1.53%	11.48%	0.997	0.85%	3.45%	0.999	0.17%	0.75%
CT	April	1.000	0.12%	0.62%	0.867	0.30%	2.24%	1.000	0.05%	0.16%
	July	0.998	0.45%	2.18%	0.919	1.11%	7.82%	0.940	1.06%	2.60%
	October	0.998	0.61%	3.12%	0.992	0.81%	5.04%	0.999	0.20%	0.71%
YCN	April	0.991	1.52%	11.38%	0.997	0.36%	0.99%	1.000	0.02%	0.08%
	July	0.990	1.70%	8.15%	0.930	1.61%	9.31%	0.964	0.22%	0.49%
	October	0.987	2.62%	12.04%	0.992	0.97%	8.60%	1.000	0.11%	0.40%
KP	April	0.996	2.19%	8.73%	0.712	1.19%	4.15%	1.000	0.03%	0.19%
	July	0.998	0.48%	3.04%	0.924	0.84%	2.57%	-*	-*	-*
	October	0.991	0.52%	5.57%	0.991	0.80%	5.62%	0.999	0.20%	1.55%
ET	April	1.000	0.11%	0.98%	0.885	0.40%	5.05%	1.000	0.05%	0.22%
	July	0.999	0.42%	1.32%	0.950	0.52%	3.36%	-*	-*	-*
	October	0.992	0.81%	2.66%	0.999	0.64%	1.75%	0.928	0.75%	2.28%

* No exceedance (>60 ppb) occurred based on the used data sources.

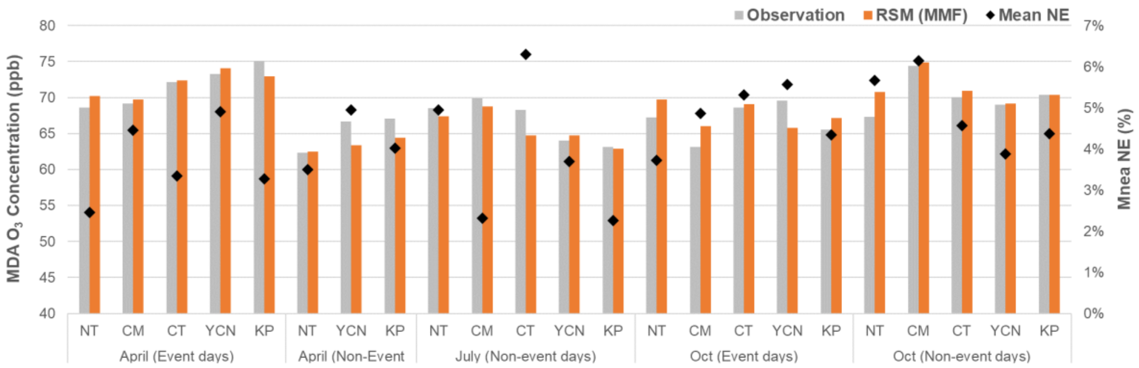


Figure 4-7. Comparison between RSM prediction based on ML-MMF data and observations and mean normalize errors (Mean NE) in each region for the event day and non-event days in selected months

benchmark case were found to overestimate NOx in the NT region and underestimate VOCs compared with observations for all regions, which means the real O_3-NOx-VOC system should be more VOC-limited and NOx-rich.

Third, ML O_3 isopleths present diverse O_3 ranges and distorted NOx and VOC regimes compared with ML-MMF O_3 isopleths, which implies ML without the support from CMAQ output could provide biased air quality predictions concerning its different O_3 response to emissions, although ML model performances well in the benchmark case. For example, for the non-event days of the NT region in April (Figure 4-9), the ML isopleth identifies an obvious VOC-limited trend, but the ML-MMF isopleth shows a combined NOx-limited and VOC-limited trend. For the other months, the ML O_3 isopleths also present a lower O_3 concentration level and narrower range, even though the trend may be similar to the ML-MMF O_3 isopleths. Even different ML techniques were performed, the ML-O_3 response to emissions still remained distorted (Figure 4-16).

The disparity between ML-MMF and ML O_3 isopleths implies that previous air quality forecasting studies [45–47] without involving CTM results and only using historical observations or other auxiliary data to predict future air quality may be potentially biased, even though the ML models met statistical requirements. Because the ML model could deviate the pollutant’s responses to its precursor emissions from the real responses of air pollutant concentrations. One major explanation is the lower variable importance priority of emission variables in the ML model. The variable importance priority of emission variables is relatively lower than other variables like meteorological variables (Figure 4-14), so the model missed the link between O_3 responses with NOx and VOC emissions. The lower variable importance priority is because most emission inventories are constructed by top-down methodology with routine monthly and weekly variation, the regular temporal variation of emission data has limited capability to explain dramatically varied pollutants. On the contrary, CMAQ-modeled O_3 has a higher importance priority in the ML-MMF model (Figure 4-14), so its nonlinear information to NOx and VOC emissions still remains in ML-MMF outputs. Another potential reason is that the ML model only learns grid-level information without considering the air pollutants transported and reacted between grids, so the ML predictions are unable to reflect the accumulated pollutants from the upwind grids or air quality zones, even adding zonal emissions as independent variables cannot improve the performance. Moreover, the limited sample size from observations could be not large enough for the ML model, because O_3 exceedance (>60 ppb) does not always occur around the monitoring stations within limited exceedance days (Figure 4-15). Therefore, the ML model under certain seasons (e.g. non-event days in April) has limited data to construct a robust model. In summary, we suggest not using ML O_3 isopleths for emission control strategies.

Fourth, transboundary pollution from upwind may change O_3 sensitivity to local NOx and VOC emissions. In the NT region, which is the first region suffering from transboundary pollution from mainland China, the ML-MMF isopleths (Figure 4-9) of non-event days in April remain a combined regime of NOx-limited and VOC-limited, but for event days in April, the isopleth changes to serious VOC-limited. Because the northeast monsoon carries high NOx concentrations from the upwind, the NT region turns to a NOx-rich atmosphere and local O_3 response becomes more sensitive to VOC emissions. Even though there is no local NOx emission (emission ratio=0), transboundary NOx still can support local O_3 formation. Moreover, if taking out boundary conditions and local meteorology during the fusion process (Figure 4-18), boundary conditions and local meteorology play a significant role that changing O_3 sensitivity, which also proves the significant impact of transboundary pollution. The changed O_3 sensitivity affected by outside pollution from the upwind was also reported in California, where wildfires in the late summer would emit large amounts of VOCs that can be transported to urban areas and significantly change the urban O_3 sensitivity [187].

4.4.5. Suggested emission control preference

Monthly O_3 isopleths for the event and non-event days based on ML-MMF data are shown in Figure 4-19. The O_3 isopleths in April show VOC-limited trends for both the event and non-event days for all western regions, but the range reveals a limited improved effort by reducing VOC emission, which is around 0.25 ppb to 4.5 ppb. Transboundary NOx could also accumulate for days and affect O_3 sensitivity to local emissions on non-event days. In July, controlling NOx emissions is a more effective way to lower O_3 concentrations in the CM, CT, YCN, and ET regions, and the NT and KP regions present a combined NOx-limited and VOC-limited

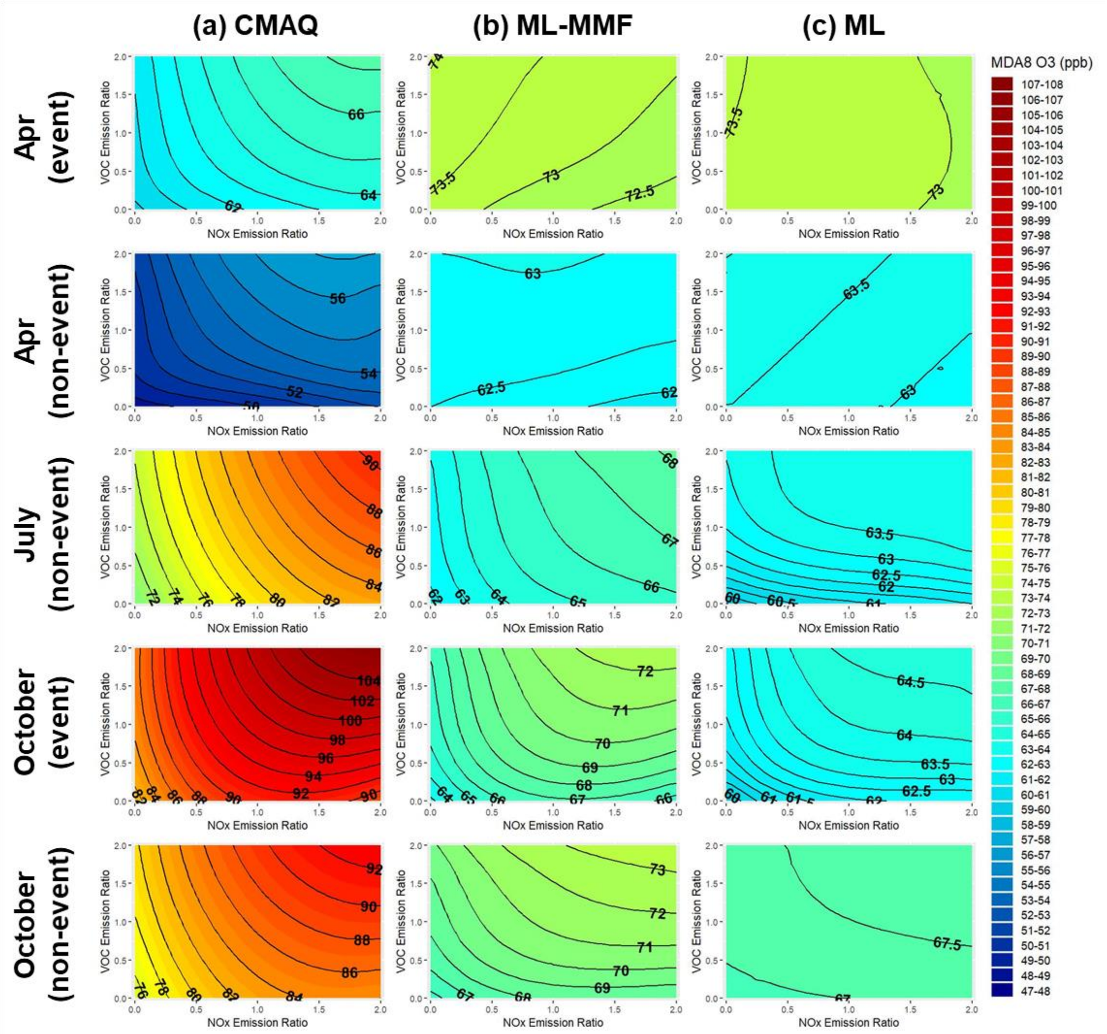


Figure 4-8. Monthly combined O_3 isopleths from all air quality regions (NT, CM, CT, YCN, KP, and ET) for the event and non-event days in selected months by using CMAQ, ML-MMF, and ML data

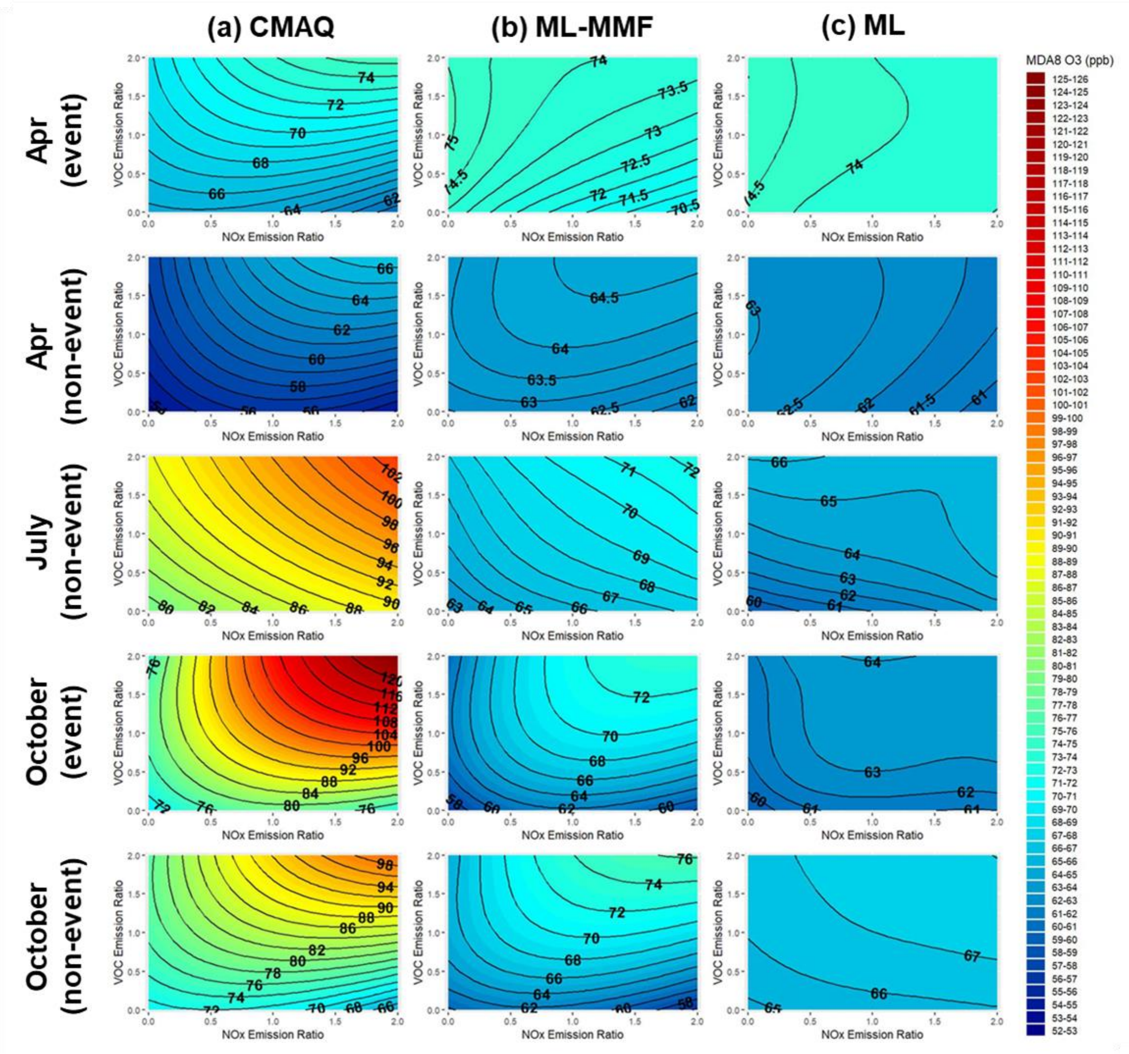


Figure 4-9. Monthly combined O_3 isopleths from the NT region for the event and non-event days in selected months by using CMAQ, ML-MMF, and ML data

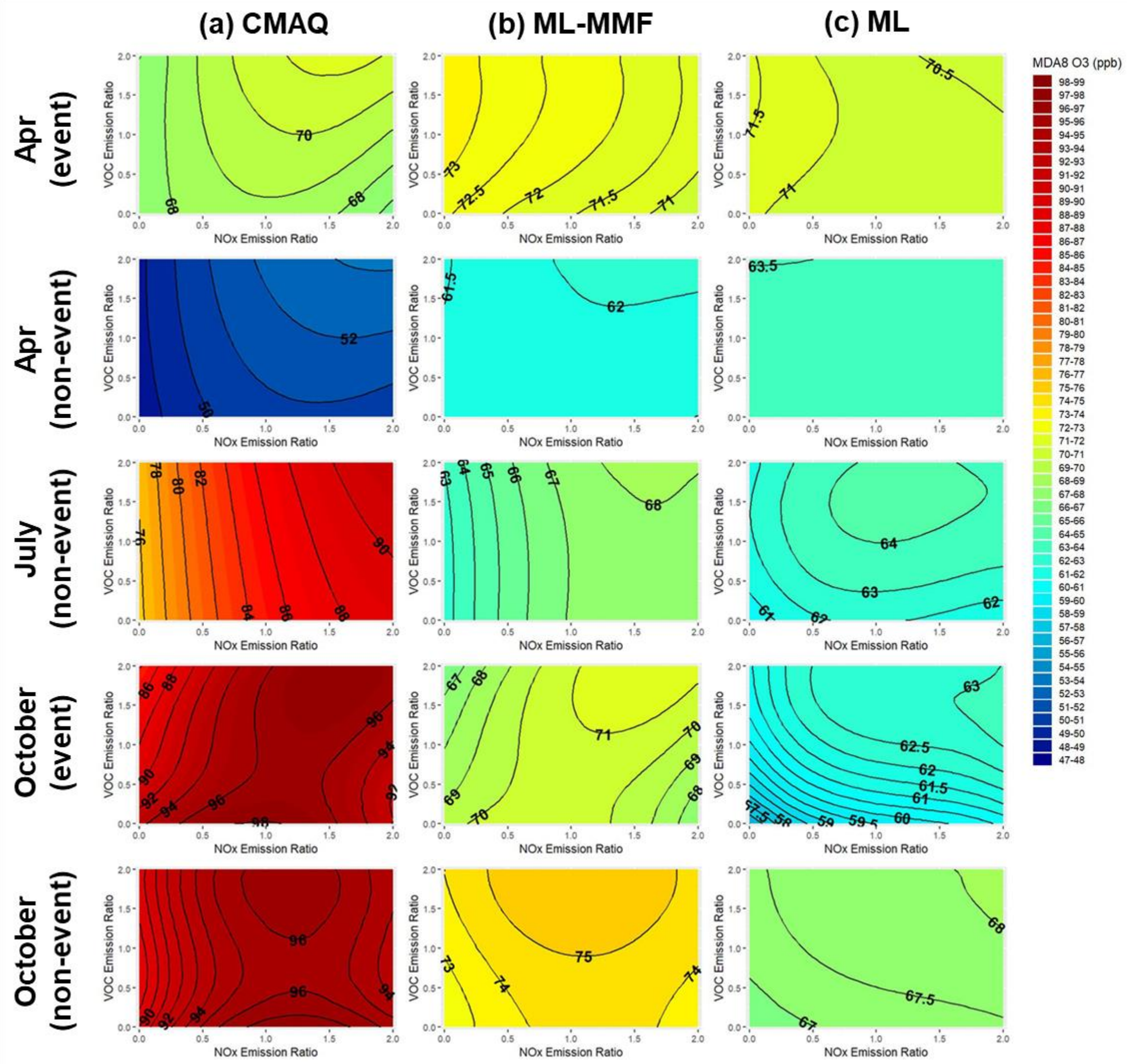


Figure 4-10. Monthly combined O_3 isopleths from the CM region for the event and non-event days in selected months by using CMAQ, ML-MMF, and ML data

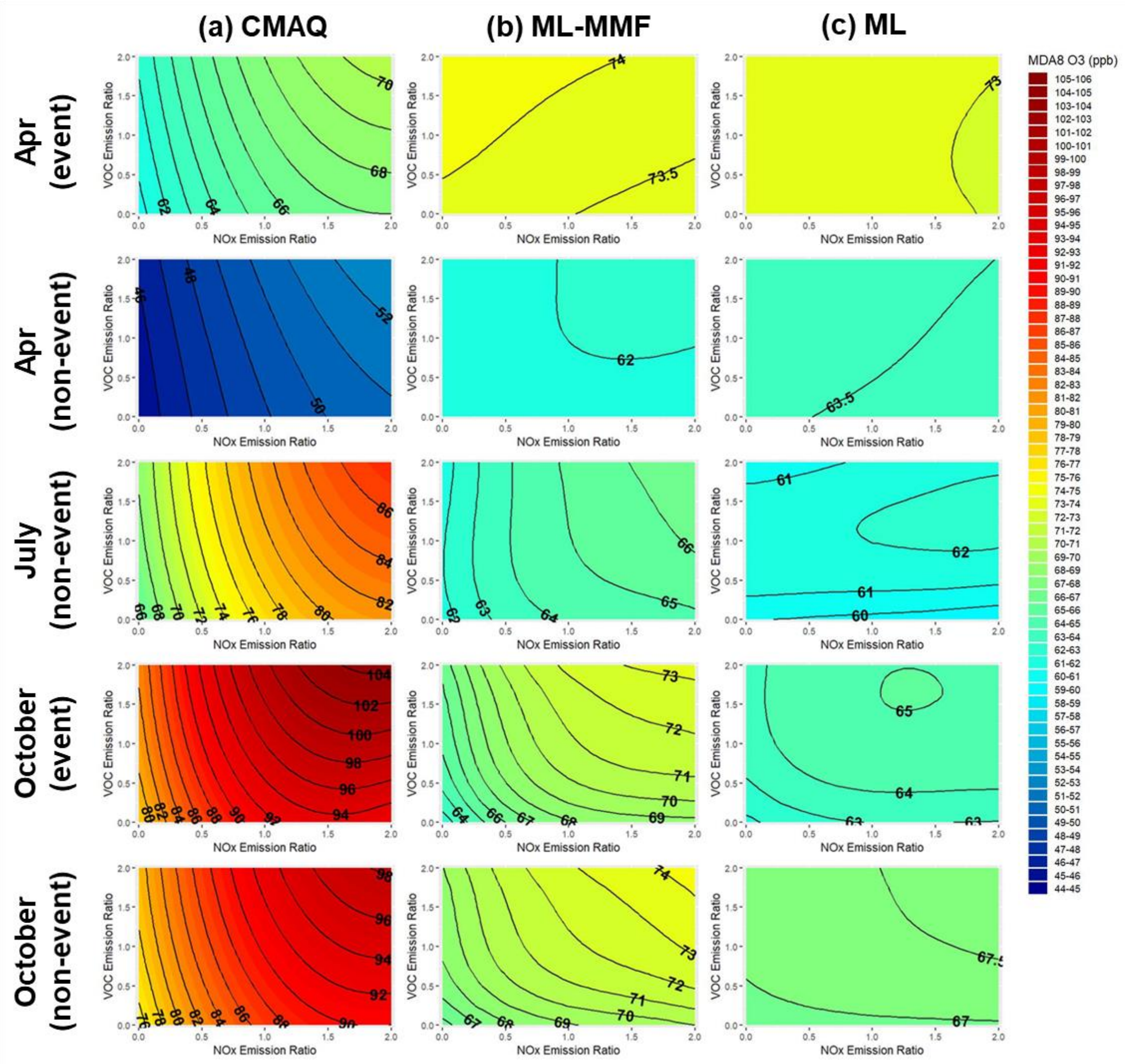


Figure 4-11. Monthly combined O_3 isopleths from the CT region for the event and non-event days in selected months by using CMAQ, ML-MMF, and ML data

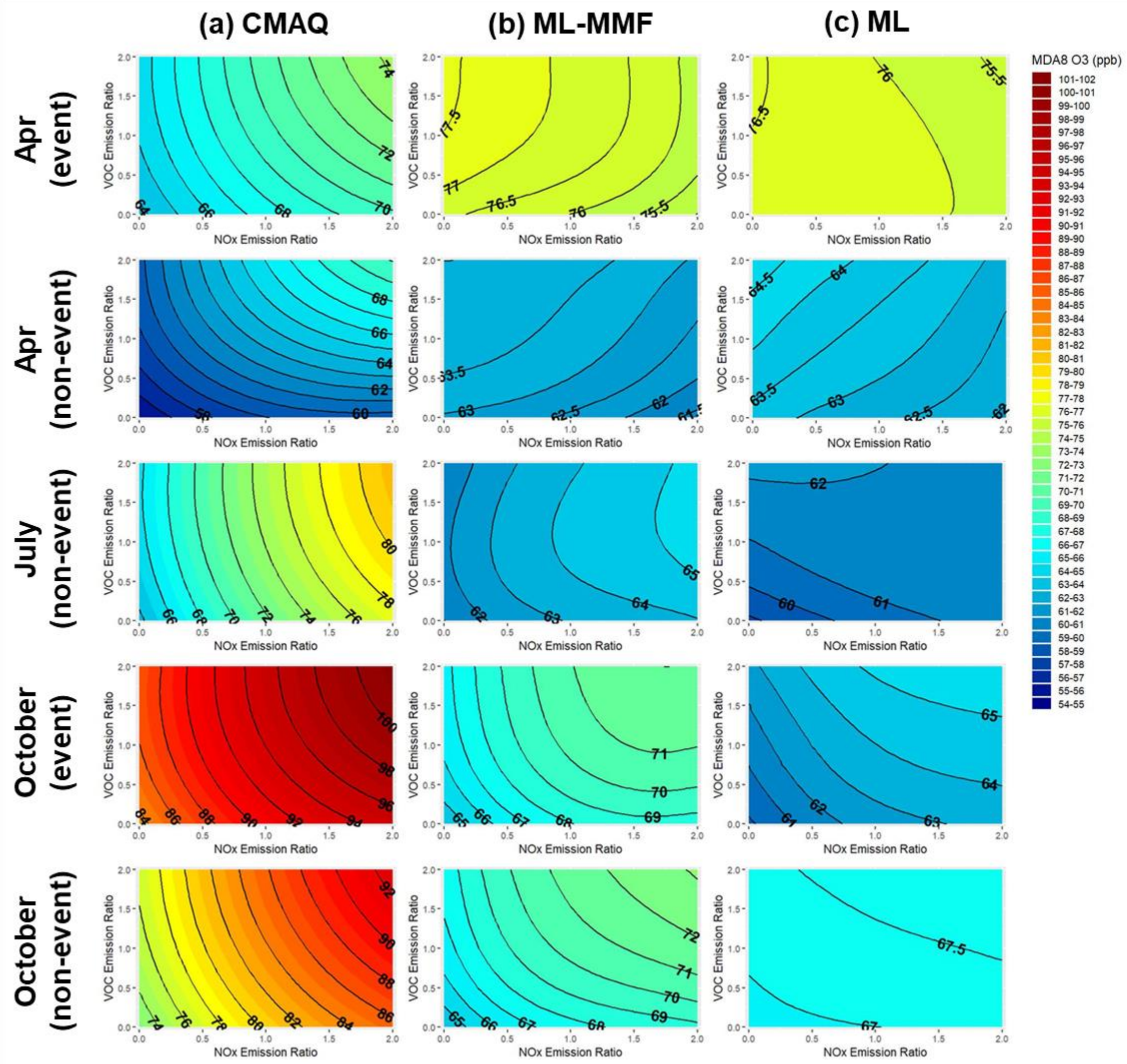


Figure 4-12. Monthly combined O_3 isopleths from the YCN region for the event and non-event days in selected months by using CMAQ, ML-MMF, and ML data

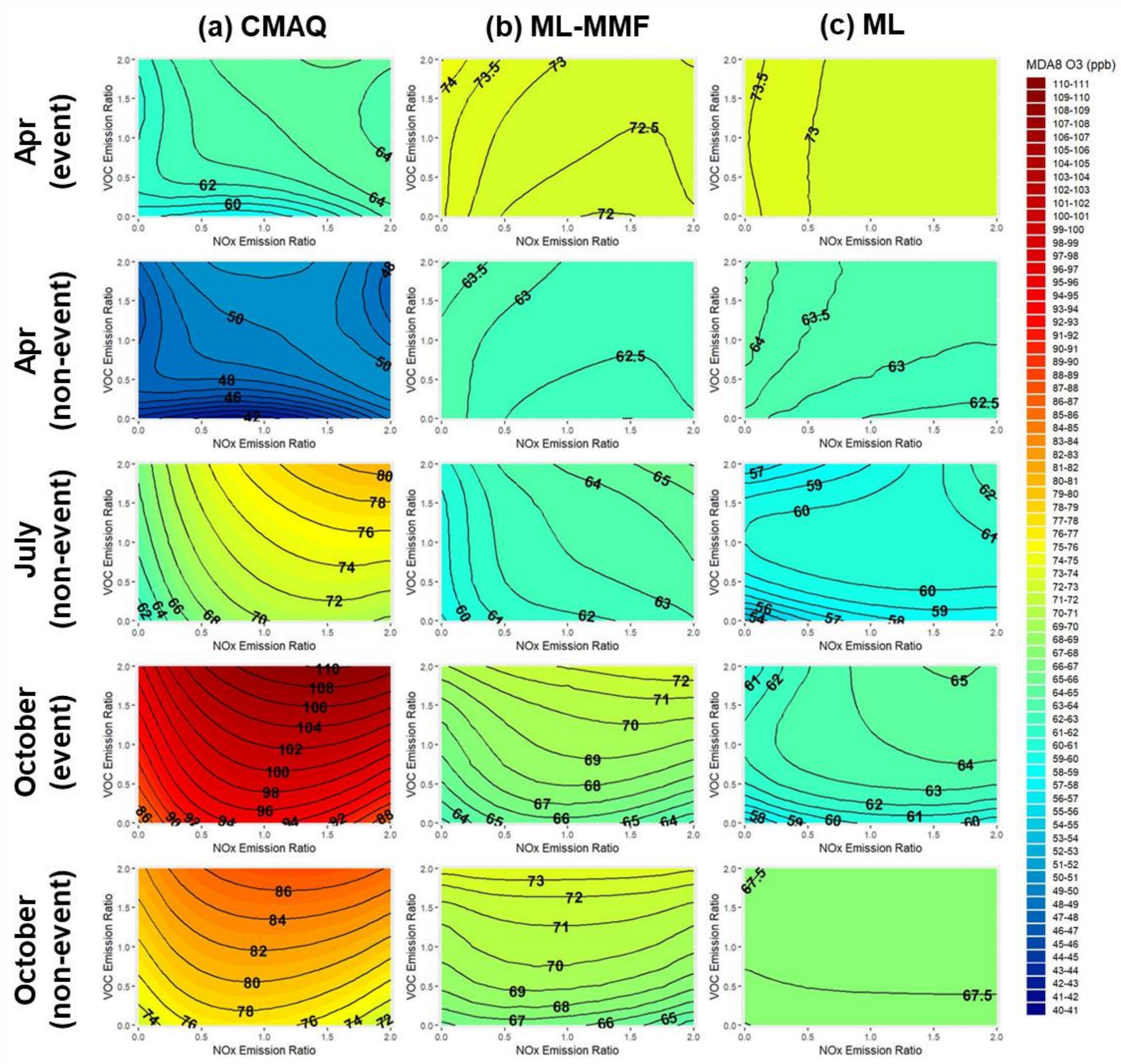


Figure 4-13. Monthly combined O3 isopleths from the KP region for the event and non-event days in selected months by using CMAQ, ML-MMF, and ML data

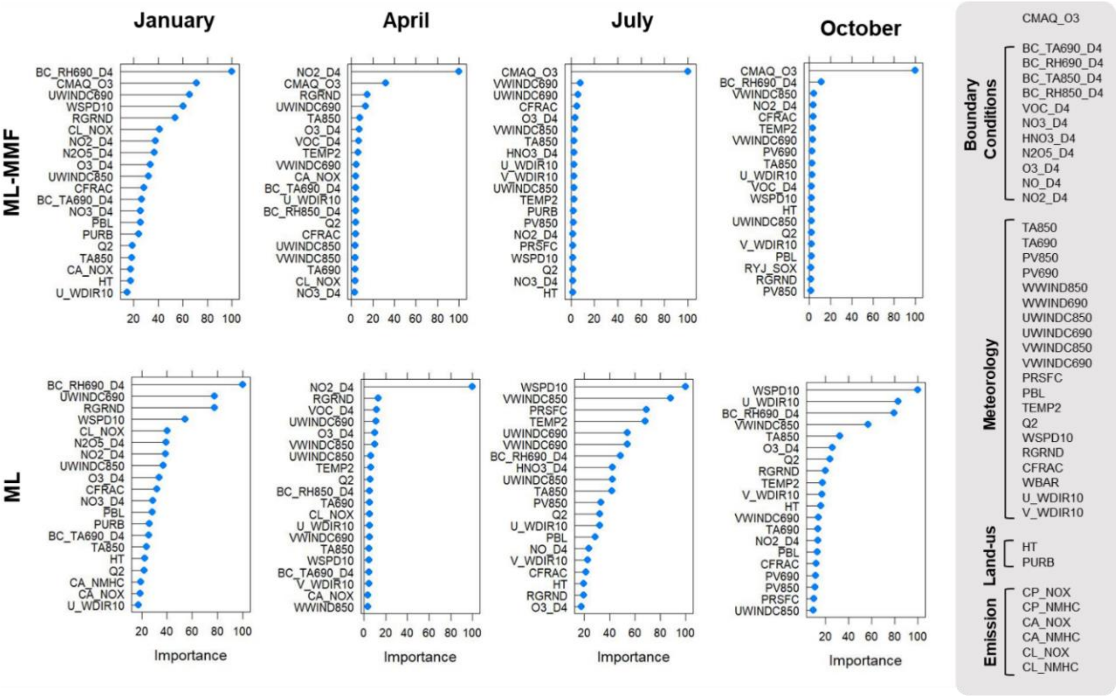


Figure 4-14. Variable importance plots of ML-MMF and ML model for January, April, July, and October in 2016.

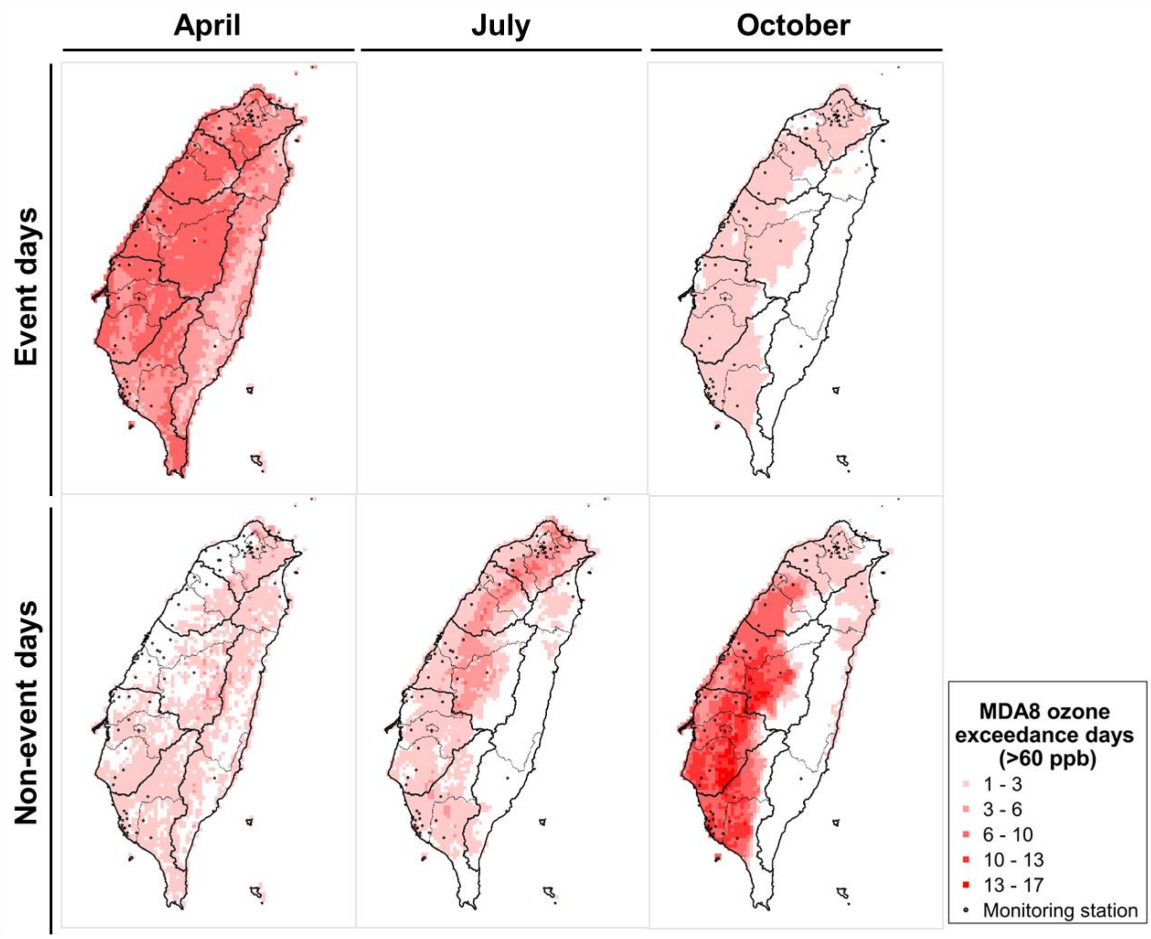


Figure 4-15. Number of MDA8 O_3 exceedance days (>60 ppb) in the selected months

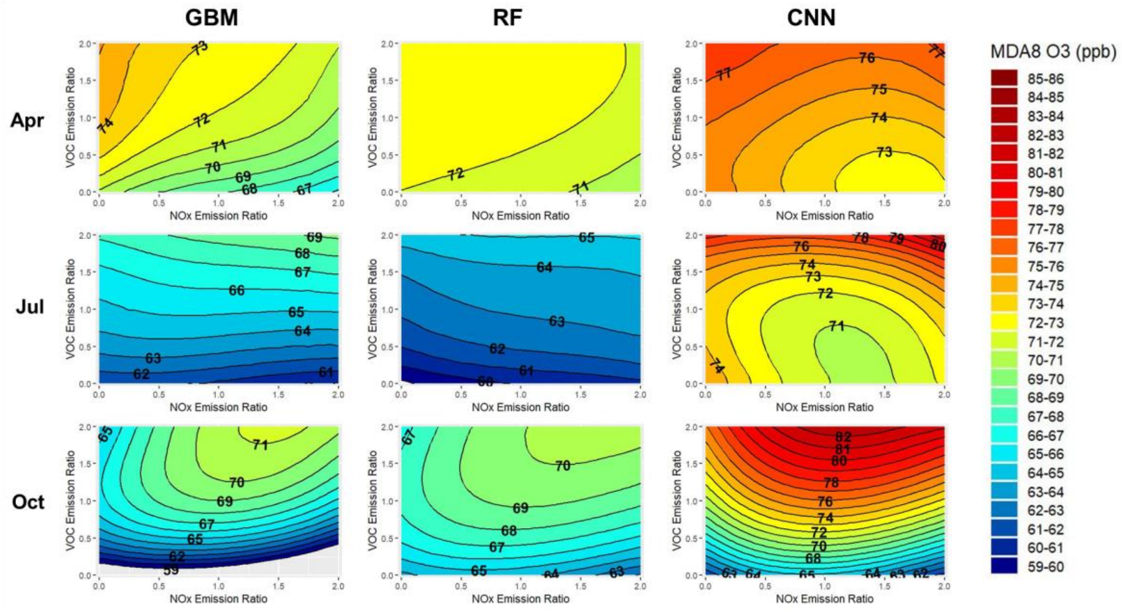


Figure 4-16. O_3 isopleths in NT region of selected months for different ML techniques (GBM, RF, and CNN)

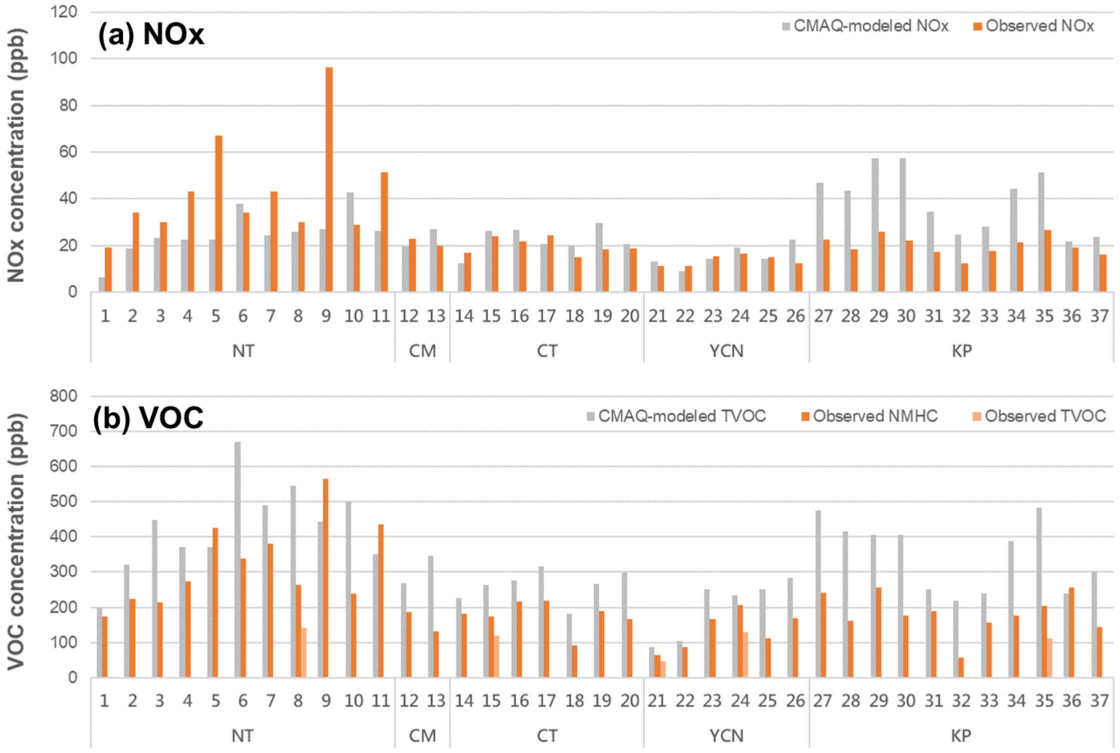


Figure 4-17. Comparison between CAMQ modeled estimations and observations of (a) NOx and (b) VOCs

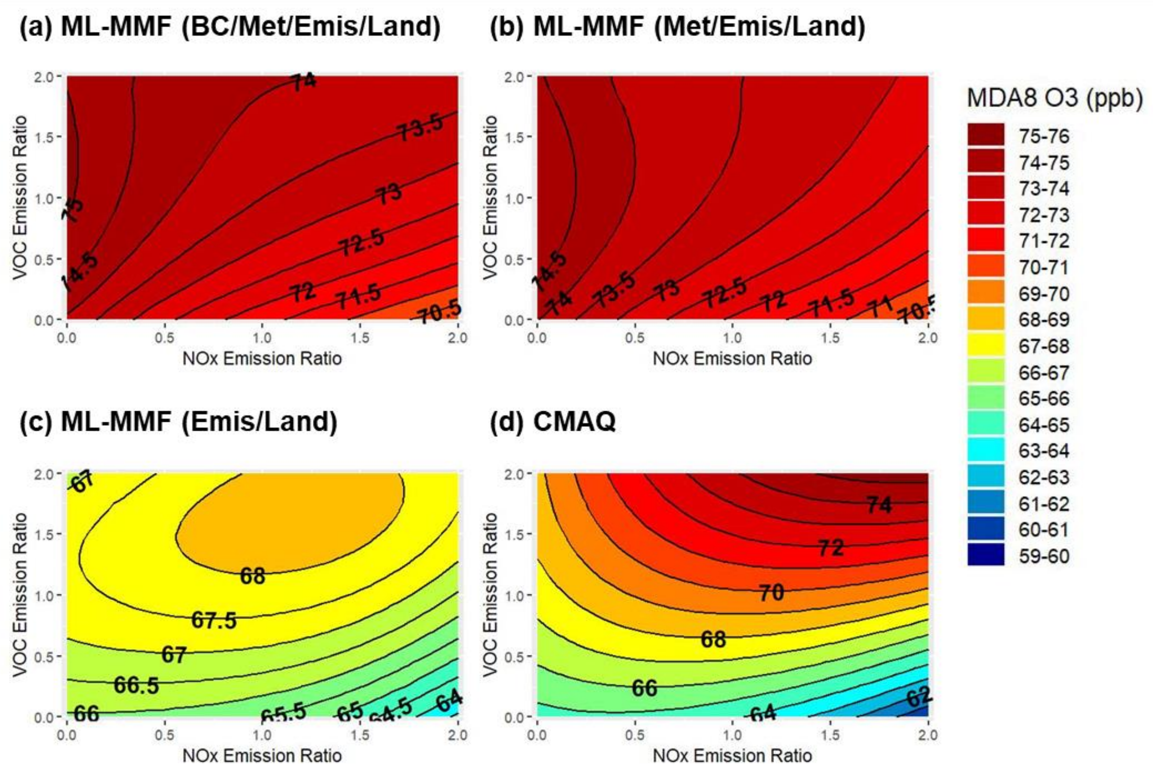


Figure 4-18. O_3 isopleths for different inputs: (a) ML-MMF (boundary conditions, meteorology, emission, and land-use data); (b) ML-MMF (meteorology, emission, and land-use data); (c) ML-MMF (emission and land-use data); (d) CMAQ data

regimes. The more VOC-limited trend in the NT and KP regions may be due to more population and vehicle emissions in these regions.

In October, O_3 isopleths are similar between the event and non-event days, because transboundary pollution in October mostly carries PM and SO_2, which are not related to the formation of O_3. NT, CT, and YCN regions have a combined NOx-limited and VOC-limited trend, and the potential improvement can be significant (58-76 ppb) in the NT region. The more VOC-limited trend in the CM and KP regions could be due to the impact of local geography. Because of many plateaus and hills in the CM and KP region, which higher terrain roughness would slow down the wind and cause accumulation of pollutants like NOx from the upwind regions. In the ET region, the O_3 isopleths suggest an obvious NOx-limited trend for all months. In summary, most of the regions are VOC-limited in April and October but NOx-limited in July. The suggestion of controlling VOC emissions in fall and winter is also coherent with the other study in Taiwan [188].

3.6. Limitations and future works

This study still has some limitations. First, the ML-MMF model needs to rely on enough monitoring stations (sample size) and good-quality of emission inventory to build a robust model. Multiple monitoring stations can explain the variance of space-related variables like emissions, meteorology, and land uses between sites and provide enough statistical power, and the emission inventory should reflect the various emissions around the monitoring stations. In our case, the employed TEDS emission inventory is updated every three years with finer to 1 km spatial resolution, thus it can reflect emissions around stations within space and time. For those regions or countries without enough monitoring stations, using remote sensing data from satellites may be an alternative way to obtain observations, but satellite data could be still biased by cloud and vertical column densities [187]. For regions or countries without proper emission inventory, using a global emission database like ECLIPSE [140] could be a potential solution to provide emission data around stations, but the spatial distribution method from national emissions needs careful assessment.

Second, the ML-MMF RSM results or O_3 isopleths are hardly validated by observations. Even though the ML-MMF O_3 isopleths with baseline emission ratio (emission ratio=1) were validated by observations and have a satisfactory performance (Figure 4-7, Mean NE=2%-6%), O_3 responses under extreme emission scenarios are hardly validated by observations. Because there are no observations to validate the estimations under extreme emission scenarios like zero (emission ratio=0) or double (emission ratio=2) anthropogenic emissions. Thus, the margins of O_3 isopleths still remained potential uncertainty. Further study to assess the uncertainty of O_3 isopleths should be investigated in the future.

4.5. Conclusion

ML models become mainstream in atmospheric science because of more available real-time monitoring data and measurements. But its black-box modeling process, failure to involve physical and chemical mechanisms, and weak cooperation with the existing numerical models lower its stringency and persuasion. Except for developing more

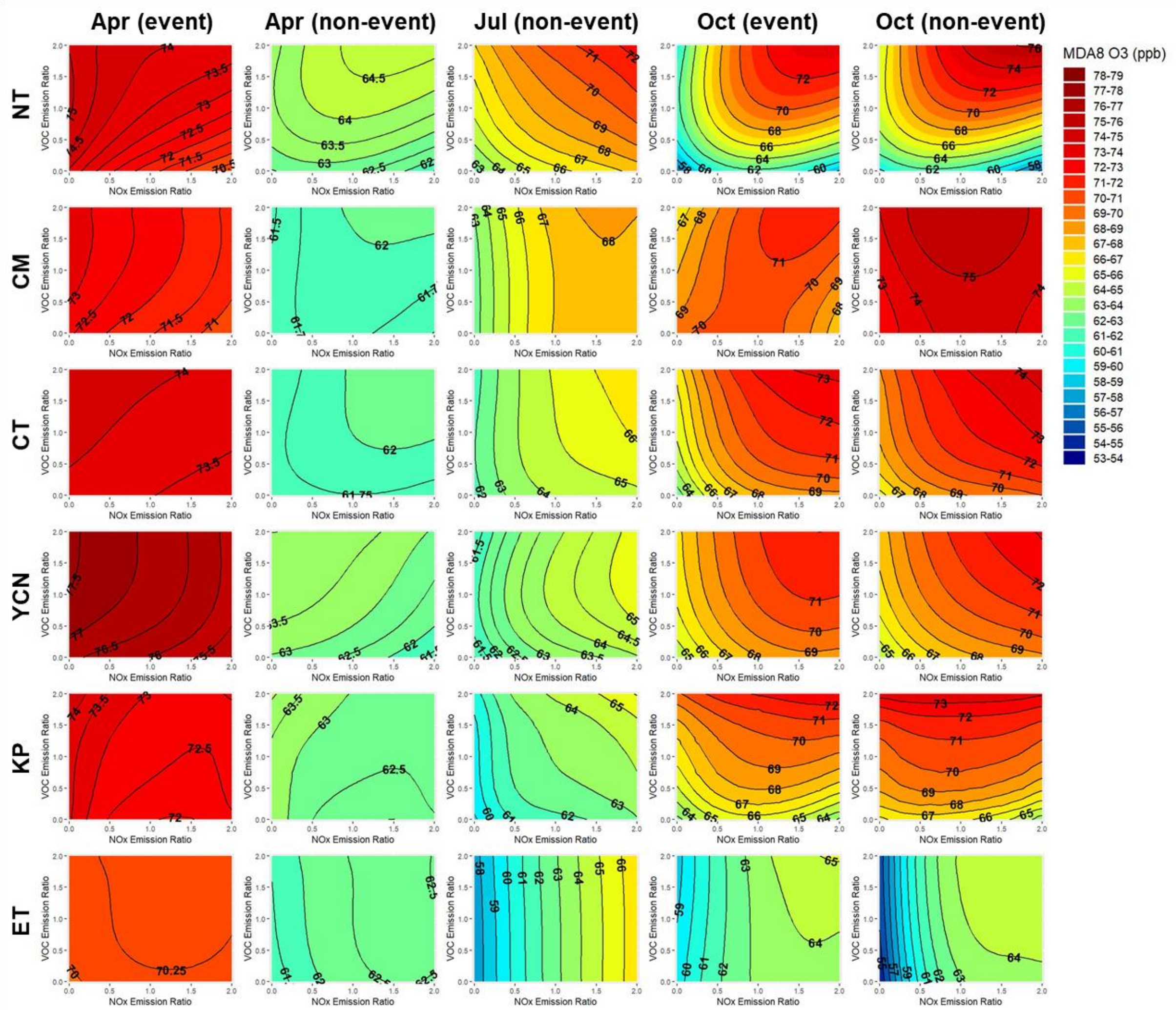


Figure 4-19. Monthly O_3 isopleths for event and non-event days in selected months based on ML-MMF data

sophisticated model designs to increase predicting accuracy, the interpretability and explainability of predictions should be evaluated as well. In this study, the capability of the ML model to serve as a bias corrector (ML-MMF output) based on CMAQ modeled results and a forecaster (ML output) was examined and applied to predict O_3 nonlinear responses to anthropogenic NOx and VOC emissions. Three types of data were examined for constructing O_3 isopleths by RSM: CMAQ data, ML-MMF data (a bias corrector), and ML data (a forecaster).

Compared with CMAQ predictions (r=0.41-0.80), both ML-MMF (r=0.93-0.94) and ML predictions (r=0.89-0.94) showed significantly improved performance in the benchmark case. While ML-MMF isopleths exhibit O_3 nonlinearity close to actual responses due to their numerical base and observation-based correction, ML isopleths present different O_3 ranges and distorted NOx and VOC-limited regimes compared with ML-MMF O_3 isopleths even though the ML model meets the statistical requirement in the benchmark case,. Without involving CMAQ results, the ML model presents biased predictions concerning its different O_3 responses to NOx and VOC emission ratios. It also implies that only using historical observations or other auxiliary data without support from CMAQ or other CTMs to forecast the air quality could mislead the future trend. Meanwhile, after being corrected by observations, ML-MMF data present changed O_3 sensitivity compared with CMAQ data. The corrected O_3 isopleths emphasize the impact of transboundary pollution from mainland China on the local O_3 sensitivity, which transboundary NOx in April would make all air quality regions in Taiwan more sensitive to local VOC emissions and limit the potential effort by reducing local VOC emissions.

It is advisable for future ML applications in atmospheric science like forecasting or bias correction to provide interpretability and explainability while requiring modeling performance. Failing to interpret the interaction between predicted air quality, emissions, and environmental factors may mislead controlled targets and air quality policies. Assessment with interpretable physical and chemical mechanisms and constructing a statistically robust ML model should be equally important.

Chapter 5. Conclusions and Future Prospects

Owing to the increasing population, more developed countries, and arising public health perceptions, community-level air quality management and exposure risk assessment are increasingly required in the near future. Inspired by more available and overwhelming data, machine learning techniques provide promising opportunities to develop more accurate and effective air quality management methods. This thesis attempts to resolve two major obstacles in air quality management: (1) inaccurate exposure risk estimations and (2) biased air quality concentrations from air quality models and illustrates three machine learning solutions, trying to bridge the present gaps between models and the real environment.

The first study identified the local exposure risk of $PM_{2.5}$ exposure and the spatial heterogeneity among different urbanization levels. Residents in rural areas had 1.40 times higher death risk compared with the average risk, and residents in the most urbanized areas had 1.57 and 1.20 times higher risk of CVD and RD than average. The imbalance exposure risk would also significantly affect the estimations of the burden of the diseases, contributing 0-86% of uncertainty, especially for highly urbanized areas.

The second study illustrated the capability of machine learning for MMF and bias quantification. The MMF results showed the R^2 of daily $PM_{2.5}$ and O_3 concentrations were significantly improved from 0.41 and 0.48 to 0.86 and 0.82 and also emphasized the impact of the complexity of topography and land-use patterns on CMAQ modeling. Bias quantification results revealed that the bias is more affected by boundary conditions and local meteorology than emission and land-use data, implying that the CMAQ model still has imperfect mechanisms to well simulate $PM_{2.5}$ and O_3 with boundary conditions and local meteorology, especially under hills or mountains where pollutants easily accumulate.

The third study examined the capability of the ML model to serve as a bias corrector and a forecaster to predict O_3 nonlinear responses to NOx and VOC emissions. Although the ML predictions (R=0.89-0.94) showed significantly improved performance in the benchmark case, the ML predictions still present different O_3 ranges and distorted NOx/VOC-limited regimes compared with the ML-MMF predictions, implying that using historical observations or other auxiliary data without support from CMAQ to forecast the air quality could mislead the future air quality trend and derived health benefits concerning diverse NOx and VOC emission control ratios. In addition, after being corrected by observations, ML-MMF predictions emphasize the significant impact of transboundary pollution from mainland China on the local O_3 sensitivity, which transboundary NOx would make Taiwan more sensitive to local VOC emissions and limit the potential effort by reducing local VOC emissions.

In the recent years, ML models have become mainstream in atmospheric and environmental health science, but model design, parameter settings, data input pretreatment, variable selection, and output interpretation still need interdisciplinary professions. While meeting the basic statistical modeling requirements, future ML applications in atmospheric and environmental health science should provide interpretability and explainability in terms of human-environment interactions and interpretable physical and chemical mechanisms. Such applications are expected to feedback to traditional methods, construct finer resolution of exposure scenarios, and deepen our understanding of environmental science.

References

[1] Wang C, Tu Y, Yu Z, Lu R. $PM_{2.5}$ and Cardiovascular Diseases in the Elderly : An Overview. Int J Environ Res Public Health 2015;12:8187–97. https://doi.org/10.3390/ijerph120708187.

[2] Xing Y, Xu Y, Shi M, Lian Y. The impact of $PM_{2.5}$ on the human respiratory system. J Thorac Dis 2016;8:69–74. https://doi.org/10.3978/j.issn.2072-1439.2016.01.19.

[3] Apte JS, Brauer M, Cohen AJ, Ezzati M, Pope CA. Ambient $PM_{2.5}$ Reduces Global and Regional Life Expectancy. Environ Sci Technol Lett 2018;5:546–51. https://doi.org/10.1021/acs.estlett.8b00360.

[4] Zheng L, Lin R, Wang X, Chen W. The Development and Application of Machine Learning in Atmospheric Environment Studies. Remote Sens 2021;13. https://doi.org/10.3390/rs13234839.

[5] Garcia CA, Yap PS, Park HY, Weller BL. Association of long-term $PM_{2.5}$ exposure with mortality using different air pollution exposure models: Impacts in rural and urban California. Int J Environ Health Res 2016;26:145–57. https://doi.org/10.1080/09603123.2015.1061113.

[6] Ebisu K, Holford TR, Belanger KD, Leaderer BP, Bell ML. Urban land-use and respiratory symptoms in infants. Environ Res 2011;111:677–84. https://doi.org/10.1016/j.envres.2011.04.004.

[7] Son JY, Kim H, Bell ML. Does urban land-use increase risk of asthma symptoms? Environ Res 2015;142:309–18. https://doi.org/10.1016/j.envres.2015.06.042.

[8] Kumar V, Sahu M. Evaluation of nine machine learning regression algorithms for calibration of low-cost $PM_{2.5}$ sensor. J Aerosol Sci 2021;157:105809. https://doi.org/10.1016/j.jaerosci.2021.105809.

[9] Loh W. Classification and Regression Tree Methods. Encycl Stat Qual Reliab 2008;1:315–23. https://doi.org/10.5860/choice.45-6515.

[10] Deng W, Huang Z, Zhang J, Xu J. Random forests. Elem. Stat. Learn., New York, NY: Springer; 2001, p. 587–604. https://doi.org/10.1109/ICCECE51280.2021.9342376.

[11] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7. https://doi.org/10.3389/fnbot.2013.00021.

[12] Zamani Joharestan M, Cao C, Ni X, Bashir B, Talebiesfandarani S. $PM_{2.5}$ Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data Mehdi 2013:6425–32.

[13] Lu H, Xie M, Liu X, Liu B, Jiang M, Gao Y, et al. Adjusting prediction of ozone concentration based on CMAQ model and machine learning methods in Sichuan-Chongqing region, China. Atmos Pollut Res 2021;12:101066. https://doi.org/10.1016/j.apr.2021.101066.

[14] Zhang T, He W, Zheng H, Cui Y, Song H, Fu S. Satellite-based ground $PM_{2.5}$ estimation using a gradient boosting decision tree. Chemosphere 2021;268:128801. https://doi.org/10.1016/j.chemosphere.2020.128801.

[15] Sayeed A, Choi Y, Eslami E, Jung J, Lops Y, Salman AK, et al. A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone

concentrations 14 days in advance. Sci Rep 2021;11:10891. https://doi.org/10.1038/s41598-021-90446-6.

[16] Sayeed A, Choi Y, Eslami E, Lops Y, Roy A, Jung J. Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance. Neural Networks 2020;121:396–408. https://doi.org/10.1016/j.neunet.2019.09.033.

[17] Sayeed A, Eslami E, Lops Y, Choi Y. CMAQ-CNN: A new-generation of post-processing techniques for chemical transport models using deep neural networks. Atmos Environ 2022;273:118961. https://doi.org/10.1016/j.atmosenv.2022.118961.

[18] Sayeed A, Lops Y, Choi Y, Jung J, Salman AK. Bias correcting and extending the PM forecast by CMAQ up to 7 days using deep convolutional neural networks. Atmos Environ 2021;253:118376. https://doi.org/10.1016/j.atmosenv.2021.118376.

[19] Montesinos López OA, Montesinos López A, Crossa J. Multivariate Statistical Machine Learning Methods for Genomic Prediction. Multivar Stat Mach Learn Methods Genomic Predict 2022. https://doi.org/10.1007/978-3-030-89010-0.

[20] GBD 2017 Risk Factor Collaborators. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Stu. Lancet 2018;392:1923–94. https://doi.org/10.1016/S0140-6736(18)32225-6.

[21] IPCC. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. 2021.

[22] Fuhrer J, Val Martin M, Mills G, Heald CL, Harmens H, Hayes F, et al. Current and future ozone risks to global terrestrial biodiversity and ecosystem processes. Ecol Evol 2016;6:8785–99. https://doi.org/10.1002/ece3.2568.

[23] Hung WT, Lu CH (Sarah), Wang SH, Chen SP, Tsai F, Chou CCK. Investigation of long-range transported $PM_{2.5}$ events over Northern Taiwan during 2005–2015 winter seasons. Atmos Environ 2019;217:116920. https://doi.org/10.1016/j.atmosenv.2019.116920.

[24] Apte JS, Marshall JD, Cohen AJ, Brauer M. Addressing Global Mortality from Ambient $PM_{2.5}$. Environ Sci Technol 2015;49:8057–66. https://doi.org/10.1021/acs.est.5b01236.

[25] Wang X, Kindzierski W, Kaul P, Sun Q. Air pollution and acute myocardial infarction hospital admission in Alberta, Canada: A three-step procedure case-crossover study. PLoS One 2015;10:1–15. https://doi.org/10.1371/journal.pone.0132769.

[26] Kim SY, Kim E, Kim WJ. Health effects of ozone on respiratory diseases. Tuberc Respir Dis (Seoul) 2020;83:S6–11. https://doi.org/10.4046/TRD.2020.0154.

[27] Chen TM, Gokhale J, Shofer S, Kuschner WG. Outdoor air pollution: Ozone health effects. Am J Med Sci 2007;333:244–8. https://doi.org/10.1097/MAJ.0b013e31803b8e8c.

[28] World Health Organization. WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: Global update 2005 2005:1–21. https://doi.org/10.1016/0004-6981(88)90109-6.

[29] Barton H. Land use planning and health and well-being. Land Use Policy 2009;26:115–23. https://doi.org/10.1016/j.landusepol.2009.09.008.

[30] Wernham A. Health impact assessments are needed in decision making about environmental and land-use policy. Health Aff 2011;30:947–56. https://doi.org/10.1377/hlthaff.2011.0050.

[31] McGuinn LA, Ward-Caviness C, Neas LM, Schneider A, Di Q, Chudnovsky A, et al. Fine particulate matter and cardiovascular disease: Comparison of assessment methods for long-term exposure. Environ Res 2017;159:16–23. https://doi.org/10.1016/j.envres.2017.07.041.

[32] U.S. EPA Office of Research and Development. CMAQ 2017. https://doi.org/10.5281/ZENODO.3585898.

[33] Kuo CP, Fu JS, Wu PC, Cheng TJ, Chiu TY, Huang CS, et al. Quantifying spatial heterogeneity of vulnerability to short-term $PM_{2.5}$ exposure with data fusion framework. Environ Pollut 2021;285:117266. https://doi.org/10.1016/j.envpol.2021.117266.

[34] USEPA. Technical Support Document for the Proposed PM NAAQS Rule 2006:48.

[35] Gipson GL, Freas WP, Kelly RF, Meyer EL. Guideline for use of city-specific EKMA in preparing ozone SIPS. Draft report. United States: 1980.

[36] Xing J, Wang SX, Jang C, Zhu Y, Hao JM. Nonlinear response of ozone to precursor emission changes in China: A modeling study using response surface methodology. Atmos Chem Phys 2011;11:5027–44. https://doi.org/10.5194/acp-11-5027-2011.

[37] Lai HC, Hsiao MC, Liou JL, Lai LW, Wu PC, Fu JS. Using costs and health benefits to estimate the priority of air pollution control action plan: A case study in Taiwan. Appl Sci 2020;10:1–16. https://doi.org/10.3390/app10175970.

[38] Xing J, Zheng S, Li S, Huang L, Wang X, Kelly JT, et al. Mimicking atmospheric photochemical modeling with a deep neural network. Atmos Res 2022;265:105919. https://doi.org/10.1016/j.atmosres.2021.105919.

[39] Kelly JT, Jang C, Zhu Y, Long S, Xing J, Wang S, et al. Predicting the nonlinear response of $PM_{2.5}$ and ozone to precursor emission changes with a response surface model. Atmosphere (Basel) 2021;12. https://doi.org/10.3390/atmos12081044.

[40] Li J, Dai Y, Zhu Y, Tang X, Wang S, Xing J, et al. Improvements of response surface modeling with self-adaptive machine learning method for $PM_{2.5}$ and O_3 predictions. J Environ Manage 2022;303:114210. https://doi.org/10.1016/j.jenvman.2021.114210.

[41] Fu JS, Carmichael GR, Dentener F, Aas W, Andersson C, Barrie LA, et al. Improving Estimates of Sulfur, Nitrogen, and Ozone Total Deposition through Multi-Model and Measurement-Model Fusion Approaches. Environ Sci Technol 2022;56:2134–42. https://doi.org/10.1021/acs.est.1c05929.

[42] Lu X, Yuan D, Chen Y, Fung JCH, Li W, Lau AKH. Estimations of Long-Term nss-SO42-and NO$_3$-Wet Depositions over East Asia by Use of Ensemble Machine-Learning Method. Environ Sci Technol 2020;54:11118–26. https://doi.org/10.1021/acs.est.0c01068.

[43] Geng G, Meng X, He K, Liu Y. Random forest models for PM$_{2.5}$ speciation concentrations using MISR fractional AODs. Environ Res Lett 2020;15. https://doi.org/10.1088/1748-9326/ab76df.

[44] Requia WJ, Di Q, Silvern R, Kelly JT, Koutrakis P, Mickley LJ, et al. An Ensemble Learning Approach for Estimating High Spatiotemporal Resolution of Ground-Level Ozone in the Contiguous United States. Environ Sci Technol 2020;54:11037–47. https://doi.org/10.1021/acs.est.0c01791.

[45] Ausati S, Amanollahi J. Assessing the accuracy of ANFIS, EEMD-GRNN, PCR, and MLR models in predicting PM$_{2.5}$. Atmos Environ 2016;142:465–74. https://doi.org/10.1016/j.atmosenv.2016.08.007.

[46] Song Y, Qin S, Qu J, Liu F. The forecasting research of early warning systems for atmospheric pollutants: A case in Yangtze River Delta region. Atmos Environ 2015;118:58–69. https://doi.org/10.1016/j.atmosenv.2015.06.032.

[47] Zhou Y, Chang FJ, Chang LC, Kao IF, Wang YS. Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. J Clean Prod 2019;209:134–45. https://doi.org/10.1016/j.jclepro.2018.10.243.

[48] Ding D, Zhu Y, Jang C, Lin C, Wang S, Fu J, et al. Evaluation of health benefit using BenMAP-CE with an integrated scheme of model and monitor data during Guangzhou Asian Games. J Environ Sci 2016;42:9–18. https://doi.org/10.1016/j.jes.2015.06.003.

[49] Qiu X, Zhu Y, Jang C, Lin C, Wang S, Fu J, et al. Development of an integrated policy making tool for assessing air quality and human health benefits of air pollution control 2015;9:1056–65. https://doi.org/10.1007/s11783-015-0796-8.

[50] Wang J, Wang S, Voorhees AS, Zhao B, Jang C, Jiang J, et al. Assessment of short-term $PM_{2.5}$-related mortality due to different emission sources in the Yangtze River Delta, China. Atmos Environ 2015;123:440–8. https://doi.org/10.1016/j.atmosenv.2015.05.060.

[51] Burnett RT, Arden Pope C, Ezzati M, Olives C, Lim SS, Mehta S, et al. An integrated risk function for estimating the global burden of disease attributable to ambient fine particulate matter exposure. Environ Health Perspect 2014;122:397–403. https://doi.org/10.1289/ehp.1307049.

[52] Liu B, Wu J, Zhang J, Wang L, Yang J, Liang D, et al. Characterization and source apportionment of $PM_{2.5}$ based on error estimation from EPA PMF 5.0 model at a medium city in China. Environ Pollut 2017;222:10–22. https://doi.org/10.1016/j.envpol.2017.01.005.

[53] Targino AC, Gibson MD, Krecl P, Rodrigues MVC, dosSantos MM, dePaula Corrêa M. Hotspots of black carbon and $PM_{2.5}$ in an urban area and relationships to traffic characteristics. Environ Pollut 2016;218:475–86. https://doi.org/10.1016/j.envpol.2016.07.027.

[54] Hao Y, Meng X, Yu X, Lei M, Li W, Yang W, et al. Quantification of primary and secondary sources to $PM_{2.5}$ using an improved source regional apportionment method in an industrial city, China. Sci Total Environ 2020;706:135715. https://doi.org/10.1016/j.scitotenv.2019.135715.

[55] Hess PM, Moudon AV, Logsdon MG. Measuring land use patterns for transportation research. Transp Res Rec 2001:17–24. https://doi.org/10.3141/1780-03.

[56] Kockelman KM. Travel behavior as function of accessibility, land use mixing, and land use balance: Evidence from San Francisco Bay Area. Transp Res Rec 1997:116–25. https://doi.org/10.3141/1607-16.

[57] Song Y, Merlin L, Rodriguez D. Comparing measures of urban land use mix. Comput Environ Urban Syst 2013;42:1–13. https://doi.org/10.1016/j.compenvurbsys.2013.08.001.

[58] Davidson K, Hallberg A, McCubbin D, Hubbell B. Analysis of $PM_{2.5}$ using the environmental benefits mapping and analysis program (BenMAP). J Toxicol Environ Heal - Part A Curr Issues 2007;70:332–46. https://doi.org/10.1080/15287390600884982.

[59] Zhang Y, Foley KM, Schwede DB, Bash JO, Pinto JP, Dennis RL. A Measurement-Model Fusion Approach for Improved Wet Deposition

Maps and Trends. J Geophys Res Atmos 2019;124:4237–51. https://doi.org/10.1029/2018JD029051.

[60] Li S, Guo Y, Williams G. Acute impact of hourly ambient air pollution on preterm birth. Environ Health Perspect 2016;124:1623–9. https://doi.org/10.1289/EHP200.

[61] Yorifuji T, Suzuki E, Kashima S. Hourly differences in air pollution and risk of respiratory disease in the elderly: A time-stratified case-crossover study. Environ Heal A Glob Access Sci Source 2014;13:1–11. https://doi.org/10.1186/1476-069X-13-67.

[62] Yorifuji T, Suzuki E, Kashima S. Cardiovascular emergency hospital visits and hourly changes in air pollution. Stroke 2014;45:1264–8. https://doi.org/10.1161/STROKEAHA.114.005227.

[63] Wu P-C, Cheng T-J, Kuo C-P, Fu JS, Lai H-C, Chiu T-Y, et al. Transient risk of ambient fine particulate matter on hourly cardiovascular events in Tainan City, Taiwan. PLoS One 2020;15:e0238082. https://doi.org/10.1371/journal.pone.0238082.

[64] Bhaskaran K, Armstrong B, Hajat S, Haines A, Wilkinson P, Smeeth L. Heat and risk of myocardial infarction: Hourly level case-crossover analysis of MINAP database. BMJ 2013;346:1–14. https://doi.org/10.1136/bmj.e8050.

[65] Linares C, Díaz J. Short-term effect of concentrations of fine particulate matter on hospital admissions due to cardiovascular and respiratory causes among the over-75 age group in Madrid, Spain. Public Health 2010;124:28–36. https://doi.org/10.1016/j.puhe.2009.11.007.

[66] Rich DQ, Utell MJ, Croft DP, Thurston SW, Thevenet-Morrison K, Evans KA, et al. Daily land use regression estimated woodsmoke and traffic pollution concentrations and the triggering of ST-elevation myocardial infarction: a case-crossover study. Air Qual Atmos Heal 2018;11:239–44. https://doi.org/10.1007/s11869-017-0537-1.

[67] Weichenthal S, Lavigne E, Evans G, Pollitt K, Burnett RT. Ambient $PM_{2.5}$ and risk of emergency room visits for myocardial infarction: Impact of regional $PM_{2.5}$ oxidative potential: A case-crossover study. Environ Heal A Glob Access Sci Source 2016;15:1–9. https://doi.org/10.1186/s12940-016-0129-9.

[68] U.S. EPA. Environmental Benefits Mapping and Analysis Program - Community Edition (Version 1.1.3) 2014. www.epa.gov/benmap (accessed April25, 2020).

[69] Liu ST, Liao CY, Kuo CY, Kuo HW. The effects of $PM_{2.5}$ from asian dust storms on emergency room visits for cardiovascular and respiratory diseases. Int J Environ Res Public Health 2017;14:1–10. https://doi.org/10.3390/ijerph14040428.

[70] Tsai JH, Huang KL, Lin NH, Chen SJ, Lin TC, Chen SC, et al. Influence of an Asian dust storm and Southeast Asian biomass burning on the characteristics of seashore atmospheric aerosols in Southern Taiwan. Aerosol Air Qual Res 2012;12:1105–15. https://doi.org/10.4209/aaqr.2012.07.0201.

[71] Guo Y, Barnett AG, Pan X, Yu W, Tong S. The impact of temperature on mortality in Tianjin, china: A case-crossover design with a

distributed lag nonlinear model. Environ Health Perspect 2011;119:1719–25. https://doi.org/10.1289/ehp.1103598.

[72] Chang CC, Kuo CC, Liou SH, Yang CY. Fine particulate air pollution and hospital admissions for myocardial infarction in a subtropical city: Taipei, Taiwan. J Toxicol Environ Heal - Part A Curr Issues 2013;76:440–8. https://doi.org/10.1080/15287394.2013.771559.

[73] Chiu HF, Tsai SS, Weng HH, Yang CY. Short-term effects of fine particulate air pollution on emergency room visits for cardiac arrhythmias: A case-crossover study in Taipei. J Toxicol Environ Heal - Part A Curr Issues 2013;76:614–23. https://doi.org/10.1080/15287394.2013.801763.

[74] Hsieh YL, Tsai SS, Yang CY. Fine particulate air pollution and hospital admissions for congestive heart failure: A case-crossover study in Taipei. Inhal Toxicol 2013;25:455–60. https://doi.org/10.3109/08958378.2013.804609.

[75] Chen YC, Weng YH, Chiu YW, Yang CY. Short-Term Effects of Coarse Particulate Matter on Hospital Admissions for Cardiovascular Diseases: A Case-Crossover Study in a Tropical City. J Toxicol Environ Heal - Part A Curr Issues 2015;78:1241–53. https://doi.org/10.1080/15287394.2015.1083520.

[76] Chiu HF, Peng CY, Wu TN, Yang CY. Short-term effects of fine particulate air pollution on ischemic heart disease hospitalizations in Taipei: A case-crossover study. Aerosol Air Qual Res 2013;13:1563–9. https://doi.org/10.4209/aaqr.2013.01.0013.

[77] Chang CC, Chen PS, Yang CY. Short-term effects of fine particulate air pollution on hospital admissions for cardiovascular diseases: A case-crossover study in a tropical city. J Toxicol Environ Heal - Part A Curr Issues 2015;78:267–77. https://doi.org/10.1080/15287394.2014.960044.

[78] Tsai SS, Tsai CY, Yang CY. Fine particulate air pollution associated with increased risk of hospital admissions for hypertension in a tropical city, Kaohsiung, Taiwan. J Toxicol Environ Heal - Part A Curr Issues 2018;81:567–75. https://doi.org/10.1080/15287394.2018.1460788.

[79] Phung VLH, Ueda K, Kasaoka S, Seposo X, Tasmin S, Yonemochi S, et al. Acute effects of ambient $PM_{2.5}$ on all-cause and cause-specific emergency ambulance dispatches in Japan. Int J Environ Res Public Health 2018;15:1–12. https://doi.org/10.3390/ijerph15020307.

[80] Ueda K, Nitta H, Ono M. Effects of fine particulate matter on daily mortality for specific heart diseases in Japan. Circ J 2009;73:1248–54. https://doi.org/10.1253/circj.CJ-08-1149.

[81] Guo Y, Jia Y, Pan X, Liu L, Wichmann HE. The association between fine particulate air pollution and hospital emergency room visits for cardiovascular diseases in Beijing, China. Sci Total Environ 2009;407:4826–30. https://doi.org/10.1016/j.scitotenv.2009.05.022.

[82] Guo Y, Tong S, Zhang Y, Barnett AG, Jia Y, Pan X. The relationship between particulate air pollution and emergency hospital visits for hypertension in Beijing, China. Sci Total Environ 2010;408:4446–50. https://doi.org/10.1016/j.scitotenv.2010.06.042.

[83] Huang F, Luo Y, Guo Y, Tao L, Xu Q, Wang C, et al. Particulate Matter and Hospital Admissions for Stroke in Beijing, China: Modification Effects by Ambient Temperature. J Am Heart Assoc 2016;5:1–12. https://doi.org/10.1161/JAHA.116.003437.

[84] Liang F, Tian L, Guo Q, Westerdahl D, Liu Y, Jin X, et al. Associations of $PM_{2.5}$ and black carbon with hospital emergency room visits during heavy haze events: A case study in Beijing, China. Int J Environ Res Public Health 2017;14. https://doi.org/10.3390/ijerph14070725.

[85] Liu H, Tian Y, Xu Y, Zhang J. Ambient Particulate Matter Concentrations and Hospitalization for Stroke in 26 Chinese Cities: A Case-Crossover Study. Stroke 2017;48:2052–9. https://doi.org/10.1161/STROKEAHA.116.016482.

[86] Zhang Q, Qi W, Yao W, Wang M, Chen Y, Zhou Y. Ambient particulate matter ($PM_{2.5}$/PM_{10}) exposure and emergency department visits for acute myocardial infarction in Chaoyang District, Beijing, China during 2014: A case-crossover study. J Epidemiol 2016;26:538–45. https://doi.org/10.2188/jea.JE20150209.

[87] Akbarzadeh MA, Khaheshi I, Sharifi A, Yousefi N, Naderian M, Namazi MH, et al. The association between exposure to air pollutants including PM_{10}, $PM_{2.5}$, ozone, carbon monoxide, sulfur dioxide, and nitrogen dioxide concentration and the relative risk of developing STEMI: A case-crossover design. Environ Res 2018;161:299–303. https://doi.org/10.1016/j.envres.2017.11.020.

[88] Ostro B, Tobias A, Querol X, Alastuey A, Amato F, Pey J, et al. The effects of particulate matter sources on daily mortality: A case-crossover study of Barcelona, Spain. Environ Health Perspect 2011;119:1781–7. https://doi.org/10.1289/ehp.1103618.

[89] Reyes M, Díaz J, Tobias A, Montero JC, Linares C. Impact of Saharan dust particles on hospital admissions in Madrid (Spain). Int J Environ Health Res 2014;24:63–72. https://doi.org/10.1080/09603123.2013.782604.

[90] Maté T, Guaita R, Pichiule M, Linares C, Díaz J. Short-term effect of fine particulate matter ($PM_{2.5}$) on daily mortality due to diseases of the circulatory system in Madrid (Spain). Sci Total Environ 2010;408:5750–7. https://doi.org/10.1016/j.scitotenv.2010.07.083.

[91] Belleudi V, Faustini A, Stafoggia M, Cattani G, Marconi A, Perucci CA, et al. Impact of fine and ultrafine particles on emergency hospital admissions for cardiac and respiratory diseases. Epidemiology 2010;21:414–23. https://doi.org/10.1097/EDE.0b013e3181d5c021.

[92] Argacha JF, Collart P, Wauters A, Kayaert P, Lochy S, Schoors D, et al. Air pollution and ST-elevation myocardial infarction: A case-crossover study of the Belgian STEMI registry 2009–2013. Int J Cardiol 2016;223:300–5. https://doi.org/10.1016/j.ijcard.2016.07.191.

[93] Pascal M, Falq G, Wagner V, Chatignoux E, Corso M, Blanchard M, et al. Short-term impacts of particulate matter (PM_{10}, PM_{10}-2.5, $PM_{2.5}$) on mortality in nine French cities. Atmos Environ 2014;95:175–84. https://doi.org/10.1016/j.atmosenv.2014.06.030.

[94] Wichmann J, Folke F, Torp-Pedersen C, Lippert F, Ketzel M, Ellermann T, et al. Out-of-Hospital Cardiac Arrests and Outdoor Air Pollution Exposure in Copenhagen, Denmark. PLoS One 2013;8:2–11. https://doi.org/10.1371/journal.pone.0053684.

[95] Janssen NAH, Fischer P, Marra M, Ameling C, Cassee FR. Short-term effects of $PM_{2.5}$, PM_{10} and $PM_{2.5-10}$ on daily mortality in the Netherlands. Sci Total Environ 2013;463–464:20–6. https://doi.org/10.1016/j.scitotenv.2013.05.062.

[96] Sullivan J, Sheppard L, Schreuder A, Ishikawa N, Siscovick D, Kaufman J. Relation between short-term fine-particulate matter exposure and onset of myocardial infarction. Epidemiology 2005;16:41–8. https://doi.org/10.1097/01.ede.0000147116.34813.56.

[97] Kloog I, Coull BA, Zanobetti A, Koutrakis P, Schwartz JD. Acute and chronic effects of particles on hospital admissions in New-England. PLoS One 2012;7:2–9. https://doi.org/10.1371/journal.pone.0034664.

[98] Kloog I, Nordio F, Zanobetti A, Coull BA, Koutrakis P, Schwartz JD. Short term effects of particle exposure on hospital admissions in the mid-atlantic states: A population estimate. PLoS One 2014;9:1–7. https://doi.org/10.1371/journal.pone.0088578.

[99] Symons JM, Wang L, Guallar E, Howell E, Dominici F, Schwab M, et al. A case-crossover study of fine particulate matter air pollution and onset of congestive heart failure symptom exacerbation leading to hospitalization. Am J Epidemiol 2006;164:421–33. https://doi.org/10.1093/aje/kwj206.

[100] Bell ML, Ebisu K, Leaderer BP, Gent JF, Lee HJ, Koutrakis P, et al. Associations of $PM_{2.5}$ constituents and sources with hospital admissions: Analysis of four counties in connecticut and Massachusetts (USA) for persons ≥ 65 years of age. Environ Health Perspect 2014;122:138–44. https://doi.org/10.1289/ehp.1306656.

[101] Pope CA, Muhlestein JB, May HT, Renlund DG, Anderson JL, Horne BD. Ischemic heart disease events triggered by short-term exposure to fine particulate air pollution. Circulation 2006;114:2443–8. https://doi.org/10.1161/CIRCULATIONAHA.106.636977.

[102] Grineski SE, Herrera JM, Bulathsinhala P, Staniswalis JG. Is there a Hispanic Health Paradox in sensitivity to air pollution? Hospital admissions for asthma, chronic obstructive pulmonary disease and congestive heart failure associated with NO_2 and $PM_{2.5}$ in El Paso, TX, 2005-2010. Atmos Environ 2015;119:314–21. https://doi.org/10.1016/j.atmosenv.2015.08.027.

[103] Kloog I, Zanobetti A, Nordio F, Coull BA, Baccarelli AA, Schwartz J. Effects of airborne fine particles ($PM_{2.5}$) on deep vein thrombosis admissions in the northeastern United States. J Thromb Haemost 2015;13:768–74. https://doi.org/10.1111/jth.12873.

[104] Silverman RA, Ito K, Freese J, Kaufman BJ, DeClaro D, Braun J, et al. Association of ambient fine particles with out-of-hospital cardiac arrests in New York city. Am J Epidemiol 2010;172:917–23. https://doi.org/10.1093/aje/kwq217.

[105] Bell ML, Son JY, Peng RD, Wang Y, Dominici F. Brief Report: Ambient $PM_{2.5}$ and Risk of Hospital Admissions: Do Risks Differ for Men and Women? Epidemiology 2015;26:575–9. https://doi.org/10.1097/EDE.0000000000000310.

[106] Alman BL, Pfister G, Hao H, Stowell J, Hu X, Liu Y, et al. The association of wildfire smoke with respiratory and cardiovascular emergency department visits in Colorado in 2012: A case crossover study. Environ Heal A Glob Access Sci Source 2016;15:1–9. https://doi.org/10.1186/s12940-016-0146-8.

[107] Talbott EO, Rager JR, Benson S, Ann Brink L, Bilonick RA, Wu C. A case-crossover analysis of the impact of $PM_{2.5}$ on cardiovascular disease hospitalizations for selected CDC tracking states. Environ Res 2014;134:455–65. https://doi.org/10.1016/j.envres.2014.06.018.

[108] O’Donnell MJ, Fang J, Mittleman MA, Kapral MK, Wellenius GA. Fine Particulate Air Pollution ($PM_{2.5}$) and the Risk of Acute Ischemic Stroke. Epidemiology 2008;23:1–7. https://doi.org/10.1038/jid.2014.371.

[109] Villeneuve PJ, Chen L, Stieb D, Rowe BH. Associations between outdoor air pollution and emergency department visits for stroke in Edmonton, Canada. Eur J Epidemiol 2006;21:689–700. https://doi.org/10.1007/s10654-006-9050-9.

[110] Vera J, Cifuentes L. Association of Hospital Admissions for Cardiovascular Causes and Air Pollution (PM_{10}, PM25 and O_3) in Santiago, Chile. Epidemiology 2008;19.

[111] Hansen A, Bi P, Nitschke M, Pisaniello D, Ryan P, Sullivan T, et al. Particulate air pollution and cardiorespiratory hospital admissions in a temperate Australian city: A case-crossover analysis. Sci Total Environ 2012;416:48–52. https://doi.org/10.1016/j.scitotenv.2011.09.027.

[112] Barnett AG, Williams GM, Schwartz J, Neller AH, Best TL, Petroeschevsky AL, et al. Air pollution and child respiratory health: A case-crossover study in Australia and New Zealand. Am J Respir Crit Care Med 2005;171:1272–8. https://doi.org/10.1164/rccm.200411-1586OC.

[113] Tsai SS, Chiu HF, Liou SH, Yang CY. Short-term effects of fine particulate air pollution on hospital admissions for respiratory diseases: A case-crossover study in a tropical city. J Toxicol Environ Heal - Part A Curr Issues 2014;77:1091–101. https://doi.org/10.1080/15287394.2014.922388.

[114] Tsai SS, Yang CY. Fine particulate air pollution and hospital admissions for pneumonia in a subtropical city: Taipei, Taiwan. J Toxicol Environ Heal - Part A Curr Issues 2014;77:192–201. https://doi.org/10.1080/15287394.2013.853337.

[115] Cheng MH, Chen CC, Chiu HF, Yang CY. Fine particulate air pollution and hospital admissions for asthma: A case-crossover study in Taipei. J Toxicol Environ Heal - Part A Curr Issues 2014;77:1075–83. https://doi.org/10.1080/15287394.2014.922387.

[116] Hua J, Yin Y, Peng L, Du L, Geng F, Zhu L. Acute effects of black carbon and $PM_{2.5}$ on children asthma admissions: A time-series study

in a Chinese city. Sci Total Environ 2014;481:433–8. https://doi.org/10.1016/j.scitotenv.2014.02.070.

[117] Xu Q, Li X, Wang S, Wang C, Huang F, Gao Q, et al. Fine particulate air pollution and hospital emergency room visits for respiratory disease in urban areas in Beijing, China, in 2013. PLoS One 2016;11:1–17. https://doi.org/10.1371/journal.pone.0153099.

[118] Luong LMT, Phung D, Sly PD, Morawska L, Thai PK. The association between particulate air pollution and respiratory admissions among young children in Hanoi, Vietnam. Sci Total Environ 2017;578:249–55. https://doi.org/10.1016/j.scitotenv.2016.08.012.

[119] Tecer LH, Alagha O, Karaca F, Tuncel G, Eldes N. Particulate matter ($PM_{2.5}$, $PM_{10\text{-}2.5}$, and PM 10) and children's hospital admissions for asthma and respiratory diseases: A bidirectional case-crossover study. J Toxicol Environ Heal - Part A Curr Issues 2008;71:512–20. https://doi.org/10.1080/15287390801907459.

[120] Guaita R, Pichiule M, Mate T, Linares C, Diaz J. Short-term impact of particulate matter ($PM_{2.5}$) on respiratory mortality in Madrid. Int J Environ Health Res 2011;21:260–74. https://doi.org/10.1080/09603123.2010.544033.

[121] Gan RW, Ford B, Lassman W, Pfister G, Vaidyanathan A, Fischer E, et al. Comparison of wildfire smoke estimation methods and associations with cardiopulmonary-related hospital admissions. GeoHealth 2017;1:122–36. https://doi.org/10.1002/2017GH000073.

[122] Yap PS, Gilbreath S, Garcia C, Jareen N, Goodrich B. The influence of socioeconomic markers on the association between fine particulate matter and hospital admissions for respiratory conditions among children. Am J Public Health 2013;103:695–702. https://doi.org/10.2105/AJPH.2012.300945.

[123] Glad JA, Brink LL, Talbott EO, Lee PC, Xu X, Saul M, et al. The relationship of ambient ozone and $PM_{2.5}$ levels and asthma emergency department visits: Possible influence of gender and ethnicity. Arch Environ Occup Heal 2012;67:103–8. https://doi.org/10.1080/19338244.2011.598888.

[124] Lin M, Chen Y, Burnett RT, Villeneuve PJ, Krewski D. The influence of ambient coarse particulate matter on asthma hospitalization in children: Case-crossover and time-series analyses. Environ Health Perspect 2002;110:575–81. https://doi.org/10.1289/ehp.02110575.

[125] Alessandrini ER, Stafoggia M, Faustini A, Berti G, Canova C, Togni ADe, et al. Association between short-term exposure to $PM_{2.5}$ and PM_{10} and mortality in susceptible subgroups: A multisite case-crossover analysis of individual effect modifiers. Am J Epidemiol 2016;184:744–54. https://doi.org/10.1093/aje/kww078.

[126] Cambra K, Martínez-Rueda T, Alonso-Fustel E, Cirarda FB, Ibáñez B, Esnaola S, et al. Mortality in small geographical areas and proximity to air polluting industries in the Basque Country (Spain). Occup Environ Med 2011;68:140–7. https://doi.org/10.1136/oem.2009.048215.

[127] Lu HY, Lin SL, Mwangi JK, Wang LC, Lin HY. Characteristics and source apportionment of atmospheric $PM_{2.5}$ at a coastal city in Southern Taiwan. Aerosol Air Qual Res 2016;16:1022–34. https://doi.org/10.4209/aaqr.2016.01.0008.

[128] Liao HT, Chou CCK, Chow JC, Watson JG, Hopke PK, Wu CF. Source and risk apportionment of selected VOCs and $PM_{2.5}$ species using partially constrained receptor models with multiple time resolution data. Environ Pollut 2015;205:121–30. https://doi.org/10.1016/j.envpol.2015.05.035.

[129] Kuo CP, Liao HT, Chou CCK, Wu CF. Source apportionment of particulate matter and selected volatile organic compounds with multiple time resolution data. Sci Total Environ 2014;472:880–7. https://doi.org/10.1016/j.scitotenv.2013.11.114.

[130] Hsu CY, Chiang HC, Chen MJ, Chuang CY, Tsen CM, Fang GC, et al. Ambient $PM_{2.5}$ in the residential area near industrial complexes: Spatiotemporal variation, source apportionment, and health impact. Sci Total Environ 2017;590–591:204–14. https://doi.org/10.1016/j.scitotenv.2017.02.212.

[131] De-Miguel-Balsa E, Latour-Pérez J, Baeza-Román A, Llamas-Álvarez A, Ruiz-Ruiz J, Fuset-Cabanes MP. Accessibility to Reperfusion Therapy among Women with Acute Myocardial Infarction: Impact on Hospital Mortality. J Women's Heal 2015;24:882–8. https://doi.org/10.1089/jwh.2014.5011.

[132] Simões PP, Almeida RMVR. Geographic accessibility to obstetric care and maternal mortality in a large metropolitan area of Brazil. Int J Gynecol Obstet 2011;112:25–9. https://doi.org/10.1016/j.ijgo.2010.07.031.

[133] Chuang KJ, Lin LY, Ho KF, Su CT. Traffic-related $PM_{2.5}$ exposure and its cardiovascular effects among healthy commuters in Taipei, Taiwan. Atmos Environ X 2020;7:100084. https://doi.org/10.1016/j.aeaoa.2020.100084.

[134] Du Y, Xu X, Chu M, Guo Y, Wang J. Air particulate matter and cardiovascular disease: The epidemiological, biomedical and clinical evidence. J Thorac Dis 2016;8:E8–19. https://doi.org/10.3978/j.issn.2072-1439.2015.11.37.

[135] Chan CC, Ng HC. A case-crossover analysis of Asian dust storms and mortality in the downwind areas using 14-year data in Taipei. Sci Total Environ 2011;410–411:47–52. https://doi.org/10.1016/j.scitotenv.2011.09.031.

[136] Eberhardt MS, Pamuk ER. The importance of place of residence: Examining health in rural and nonrural areas. Am J Public Health 2004;94:1682–6. https://doi.org/10.2105/AJPH.94.10.1682.

[137] U.S. EPA. Guidance on the Use of Models and Other Air Quality Goals for Ozone, $PM_{2.5}$, and Regional Haze. 2007.

[138] Chen L, Shi M, Gao S, Li S, Mao J, Zhang H, et al. Assessment of population exposure to $PM_{2.5}$ for mortality in China and its public health

benefit based on BenMAP. Environ Pollut 2017;221:311–7. https://doi.org/10.1016/j.envpol.2016.11.080.

[139] Talukdar S, Singha P, Mahato S, Shahfahad, Pal S, Liou YA, et al. Land-use land-cover classification by machine learning classifiers for satellite observations-A review. Remote Sens 2020;12. https://doi.org/10.3390/rs12071135.

[140] Stohl A, Aamaas B, Amann M, Baker LH, Bellouin N, Berntsen TK, et al. Evaluating the climate and air quality impacts of short-lived pollutants. Atmos Chem Phys 2015;15:10529–66. https://doi.org/10.5194/acp-15-10529-2015.

[141] Tan J, Fu JS, Dentener F, Emmons L. Multi-model study of HTAP II on sulfur and nitrogen deposition. Atmos Chem Phys 2018;18:6847–66.

[142] Tan J, Fu JS, Carmichael GR, Itahashi S, Tao Z, Huang K, et al. Why do models perform differently on particulate matter over East Asia? A multi-model intercomparison study for MICS-Asia III. Atmos Chem Phys 2020;20:7393–410.

[143] Li X, Liu Y, Wang M, Jiang Y, Dong X. Assessment of the Coupled Model Intercomparison Project phase 6 (CMIP6) Model performance in simulating the spatial-temporal variation of aerosol optical depth over Eastern Central China. Atmos Res 2021;261:105747. https://doi.org/10.1016/J.ATMOSRES.2021.105747.

[144] Xing J, Zheng S, Ding D, Kelly JT, Wang S, Li S, et al. Deep Learning for Prediction of the Air Quality Response to Emission Changes.

Environ Sci Technol 2020;54:8589–600. https://doi.org/10.1021/acs.est.0c02923.

[145] Kang GK, Gao JZ, Chiao S, Lu S, Xie G. Air Quality Prediction: Big Data and Machine Learning Approaches. Int J Environ Sci Dev 2018;9:8–16. https://doi.org/10.18178/ijesd.2018.9.1.1066.

[146] Haupt SE, Cowie J, Linden S, McCandless T, Kosovic B, Alessandrini S. Machine learning for applied weather prediction. Proc - IEEE 14th Int Conf EScience, e-Science 2018 2018:276–7. https://doi.org/10.1109/eScience.2018.00047.

[147] O'Gorman PA, Dwyer JG. Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. J Adv Model Earth Syst 2018;10:2548–63. https://doi.org/10.1029/2018ms001351.

[148] Eslami E, Khan Salman A, Choi Y, Sayeed A, Lops Y. A data ensemble approach for real-time air quality forecasting using extremely randomized trees and deep neural networks. Neural Comput Appl 2020;32:7563–79. https://doi.org/10.1007/s00521-019-04287-6.

[149] Du S, Li T, Member S, Yang Y, Horng S-J. Deep Air Quality Forecasting Using Hybrid Deep Learning Framework. IEEE Trans Knowl Data Eng 2021;33:2412–24. https://doi.org/10.1109/TKDE.2019.2954510.

[150] Mandel J, Vejmelka M, Kochanski A, Farguell A, Haley J, Mallia D, et al. An interactive data-driven HPC system for forecasting weather, wildland fire, and smoke. Proc Urgent 2019 1st Int Work HPC Urgent

Decis Mak - Held Conjunction with SC 2019 Int Conf High Perform Comput Networking, Storage Anal 2019:35–44. https://doi.org/10.1109/UrgentHPC49580.2019.00010.

[151] Liu Y, Kochanski A, Baker KR, Mell W, Linn R, Paugam R, et al. Fire behaviour and smoke modelling: Model improvement and measurement needs for next-generation smoke research and forecasting systems. Int J Wildl Fire 2019;28:570–88. https://doi.org/10.1071/WF18204.

[152] Speight LJ, Cranston MD, White CJ, Kelly L. Operational and emerging capabilities for surface water flood forecasting. Wiley Interdiscip Rev Water 2021;8:1–24. https://doi.org/10.1002/wat2.1517.

[153] Asadollah SBHS, Khan N, Sharafati A, Shahid S, Chung E-S, Wang X-J. Prediction of heat waves using meteorological variables in diverse regions of Iran with advanced machine learning models. Stoch Environ Res Risk Assess 2022;36:1959–74. https://doi.org/10.1007/s00477-021-02103-z.

[154] Lightstone SD, Moshary F, Gross B. Comparing CMAQ forecasts with a neural network forecast model for $PM_{2.5}$ in New York. Atmosphere (Basel) 2017;8. https://doi.org/10.3390/atmos8090161.

[155] Oh I, Hwang MK, Bang JH, Yang W, Kim S, Lee K, et al. Comparison of different hybrid modeling methods to estimate intraurban NO_2 concentrations. Atmos Environ 2021;244:117907. https://doi.org/10.1016/j.atmosenv.2020.117907.

[156] Lin W-Y, Hsiao M-C, Wu P-C, Fu JS, Lai L-W, Lai H-C. Analysis of air quality and health Co-benefits regarding electric vehicle promotion coupled with power plant emissions. J Clean Prod 2019:119152.

[157] Cheng FY, Feng CY, Yang ZM, Hsu CH, Chan KW, Lee CY, et al. Evaluation of real-time $PM_{2.5}$ forecasts with the WRF-CMAQ modeling system and weather-pattern-dependent bias-adjusted $PM_{2.5}$ forecasts in Taiwan. Atmos Environ 2021;244:117909. https://doi.org/10.1016/j.atmosenv.2020.117909.

[158] Dennis R, Fox T, Fuentes M, Gilliland A, Hanna S, Hogrefe C, et al. A framework for evaluating regional-scale numerical photochemical modeling systems. Environ Fluid Mech 2010;10:471–89. https://doi.org/10.1007/s10652-009-9163-2.

[159] USEPA. Meteorological Monitoring Guidance for Regulatory Modeling Applications. 2000.

[160] Mao Q, Gautney LL, Cook TM, Jacobs ME, Smith SN, Kelsoe JJ. Numerical experiments on MM5-CMAQ sensitivity to various PBL schemes. Atmos Environ 2006;40:3092–110. https://doi.org/10.1016/j.atmosenv.2005.12.055.

[161] Kim HC, Kim E, Bae C, Hoon Cho J, Kim BU, Kim S. Regional contributions to particulate matter concentration in the Seoul metropolitan area, South Korea: Seasonal variation and sensitivity to meteorology and emissions inventory. Atmos Chem Phys 2017;17:10315–32. https://doi.org/10.5194/acp-17-10315-2017.

[162] Dong X, Fu JS, Huang K, Tong D, Zhuang G. Model development of dust emission and heterogeneous chemistry within the Community Multiscale Air Quality modeling system and its application over East Asia. Atmos Chem Phys 2016;16:8157–80. https://doi.org/10.5194/acp-16-8157-2016.

[163] Kramer O. Dimensionality Reduction with Unsupervised Nearest Neighbors. vol. 51. 2013. https://doi.org/10.1007/978-3-642-38652-7.

[164] Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. 2017 Int. Conf. Eng. Technol., 2017, p. 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186.

[165] Sarwar G, Luecken D, Yarwood G, Whitten GZ, Carter WPL. Impact of an updated carbon bond mechanism on predictions from the CMAQ modeling system: Preliminary assessment. J Appl Meteorol Climatol 2008;47:3–14. https://doi.org/10.1175/2007JAMC1393.1.

[166] Appel KW, Pouliot GA, Simon H, Sarwar G, Pye HOT, Napelenok SL, et al. Evaluation of dust and trace metal estimates from the Community Multiscale Air Quality (CMAQ) model version 5.0. Geosci Model Dev 2013;6:883–99. https://doi.org/10.5194/gmd-6-883-2013.

[167] Powers JG, Klemp JB, Skamarock WC, Davis CA, Dudhia J, Gill DO, et al. The weather research and forecasting model: Overview, system efforts, and future directions. Bull Am Meteorol Soc 2017;98:1717–37. https://doi.org/10.1175/BAMS-D-15-00308.1.

[168] Lai HC, Lin MC. Characteristics of the upstream flow patterns during $PM_{2.5}$ pollution events over a complex island topography. Atmos

Environ 2020;227:117418. https://doi.org/10.1016/j.atmosenv.2020.117418.

[169] Huang K, Fu JS, Lin NH, Wang SH, Dong X, Wang G. Superposition of Gobi Dust and Southeast Asian Biomass Burning: The Effect of Multisource Long-Range Transport on Aerosol Optical Properties and Regional Meteorology Modification. J Geophys Res Atmos 2019;124:9464–83. https://doi.org/10.1029/2018JD030241.

[170] Dong X, Fu JS, Huang K, Zhu Q, Tipton M. Regional Climate Effects of Biomass Burning and Dust in East Asia: Evidence From Modeling and Observation. Geophys Res Lett 2019;46:11490–9. https://doi.org/10.1029/2019GL083894.

[171] Kong SSK, Fu JS, Dong X, Chuang MT, Ooi MCG, Huang WS, et al. Sensitivity analysis of the dust emission treatment in CMAQv5.2.1 and its application to long-range transport over East Asia. Atmos Environ 2021;257:118441. https://doi.org/10.1016/j.atmosenv.2021.118441.

[172] Serafin S, Adler B, Cuxart J, J De Wekker SF, Gohm A, Grisogono B, et al. Exchange Processes in the Atmospheric Boundary Layer Over Mountainous Terrain. Atmosphere (Basel) 2018;9:102. https://doi.org/10.3390/atmos9030102.

[173] Tai APK, Martin MV. Impacts of ozone air pollution and temperature extremes on crop yields: Spatial variability, adaptation and implications for future food security. Atmos Environ 2017;169:11–21. https://doi.org/10.1016/j.atmosenv.2017.09.002.

[174] Zarnetske PL, Gurevitch J, Franklin J, Groffman PM, Harrison CS, Hellmann JJ, et al. Potential ecological impacts of climate intervention by reflecting sunlight to cool Earth. Proc Natl Acad Sci U S A 2021;118:1–11. https://doi.org/10.1073/pnas.1921854118.

[175] Huang K, Fu JS, Gao Y, Dong X, Zhuang G, Lin Y. Role of sectoral and multi-pollutant emission control strategies in improving atmospheric visibility in the Yangtze River Delta, China. Environ Pollut 2014;184:426–34. https://doi.org/10.1016/j.envpol.2013.09.029.

[176] Arnold JR, Dennis RL. Testing CMAQ chemistry sensitivities in base case and emissions control runs at SEARCH and SOS99 surface sites in the southeastern US. Atmos Environ 2006;40:5027–40. https://doi.org/10.1016/j.atmosenv.2005.05.055.

[177] Che W, Zheng J, Wang S, Zhong L, Lau A. Assessment of motor vehicle emission control policies using Model-3/CMAQ model for the Pearl River Delta region, China. Atmos Environ 2011;45:1740–51. https://doi.org/10.1016/j.atmosenv.2010.12.050.

[178] Wang S, Xing J, Jang C, Zhu Y, Fu JS, Hao J. Impact assessment of ammonia emissions on inorganic aerosols in East China using response surface modeling technique. Environ Sci Technol 2011;45:9293–300. https://doi.org/10.1021/es2022347.

[179] Zhu Y, Lao Y, Jang C, Lin C-J, Xing J, Wang S, et al. Development and case study of a science-based software platform to support policy making on air quality. J Environ Sci 2015;27:97–107.

[180] Xing J, Ding D, Wang S, Zhao B, Jang C, Wu W, et al. Quantification of the enhanced effectiveness of NOx control from simultaneous reductions of VOC and NH3 for reducing air pollution in the Beijing-Tianjin-Hebei region, China. Atmos Chem Phys 2018;18:7799–814. https://doi.org/10.5194/acp-18-7799-2018.

[181] Zhao B, Wang SX, Xing J, Fu K, Fu JS, Jang C, et al. Assessing the nonlinear response of fine particles to precursor emissions: Development and application of an extended response surface modeling technique v1.0. Geosci Model Dev 2015;8:115–28. https://doi.org/10.5194/gmd-8-115-2015.

[182] U.S. EPA. SMOKE v4. 6 User Manual. 2018.

[183] Ware LB, Zhao Z, Koyama T, May AK, Matthay MA, Lurmann FW, et al. Long-Term Ozone Exposure Increases the Risk of Developing the Acute Respiratory Distress Syndrome. Am J Respir Crit Care Med 2016;193:1143–50. https://doi.org/10.1164/rccm.201507-1418OC.

[184] Xing J, Ding D, Wang S, Dong Z, Kelly JT, Jang C, et al. Development and application of observable response indicators for design of an effective ozone and fine-particle pollution control strategy in China.pdf. Atmos Chem Phys 2019;19:13627–46. https://doi.org/10.5194/acp-19-13627-2019.

[185] Chang KH, Chen TF, Huang HC. Estimation of biogenic volatile organic compounds emissions in subtropical island - Taiwan. Sci Total Environ 2005;346:184–99. https://doi.org/10.1016/j.scitotenv.2004.11.022.

[186] Luo H, Zhao K, Yuan Z, Yang L, Zheng J, Huang Z, et al. Emission source-based ozone isopleth and isosurface diagrams and their significance in ozone pollution control strategies. J Environ Sci (China) 2021;105:138–49. https://doi.org/10.1016/j.jes.2020.12.033.

[187] Wu S, Lee HJ, Anderson A, Liu S, Kuwayama T, Seinfeld JH, et al. Direct measurements of ozone response to emissions perturbations in California. Atmos Chem Phys 2022;22:4929–49. https://doi.org/10.5194/acp-22-4929-2022.

[188] Chen T-F, Tsai C-Y, Chen C-H, Chang K-H. Effect of Long-range Transport from Changing Emission on Ozone-NOx-VOC Sensitivity: Implication of Control. J Innov Technol 2021;3:39–49.

[189] Shang Z, Deng T, He J, Duan X. A novel model for hourly $PM_{2.5}$ concentration prediction based on CART and EELM. Sci Total Environ 2019;651:3043–52. https://doi.org/10.1016/j.scitotenv.2018.10.193.

[190] Bartz-Beielstein T, Markon S. Tuning search algorithms for real-world applications: A regression tree based approach. Proc 2004 Congr Evol Comput CEC2004 2004;1:1111–8. https://doi.org/10.1109/cec.2004.1330986.

[191] Ture M, Kurt Omurlu I. Determining of complexity parameter for recursive partitioning trees by simulation of survival data and an application on breast cancer data. J Stat Manag Syst 2018;21:125–38. https://doi.org/10.1080/09720510.2017.1386878.

[192] Chollet F. keras 2015.

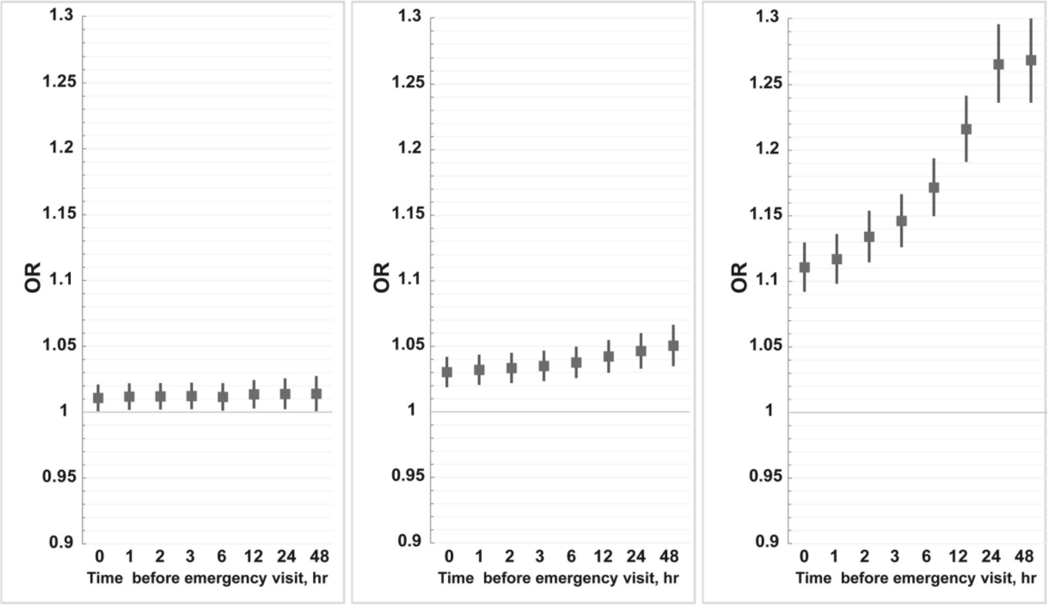
[193] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems 2016.

Appendix

I. Threshold and lag response analysis of $PM_{2.5}$ exposure risk

The odds ratio (OR) of vascular disease events and $PM_{2.5}$ was first assessed by a single-pollutant model (Figure 1). ORs of cardiovascular emergency visits with 10 μg/m^3 increases in $PM_{2.5}$ at 0 to 48 hours before disease onset were all statistically significant when including all cases in the model.

The lag-response relationship increased slightly with cumulative exposure before the case event from 0 to 48 hours in all cases. When I only include cases occurring with $PM_{2.5}$ >10 μg/m^3 and $PM_{2.5}$ >25 μg/m^3, very significant ORs could be observed for 10 μg/m^3 increases in $PM_{2.5}$ at 0 and 1 hours, implying fine particulate exposure could promptly trigger vascular disease events. Moreover, a very clear increased risk level could be observed with cumulative exposure from 0 to 48 hours, especially in those cases identified when $PM_{2.5}$ >25 μg/m^3.



(A) All cases

(B) $PM_{2.5}$>10 µg/m^3

(C) $PM_{2.5}$>25 µg/m^3

Figure 1. Odds ratio (OR) of cardiovascular emergency visits for 10 µg/m^3 increase in $PM_{2.5}$ exposure

II. Employed machine learning methods

(1) k-nearest neighbors (KNN) regression

KNN model is a non-parametric method developed on the assumption that similar samples exist near each other. That is, KNN memorizes the training data and predicts the $PM_{2.5}$ or O_3 concentrations based on the closest samples with similar patterns of input variables. Technically, KNN includes the following steps [8]:

Step 1: Compute the distance from other data points to the desired point and sort the points in increasing order of distance. Euclidean Distance is the most common method and is calculated as follows:

$$\mathrm{D} = \sqrt{(\mathrm{a}_1 - \mathrm{b}_1)^2 + (\mathrm{a}_2 - \mathrm{b}_2)^2 + \cdots + \left(\mathrm{a}_\mathrm{p} - \mathrm{b}_\mathrm{p}\right)^2}$$

where $\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_\mathrm{n}$ or $\mathrm{b}_1, \mathrm{b}_2, \ldots, \mathrm{b}_\mathrm{n}$ represent the attribute values for two points, D is the Euclidean distance between two points, and p is the number of total input variables.

Step 2: Select the number of nearest neighbors ($\mathrm{k} < \mathrm{p}$). The $PM_{2.5}$ or O_3 prediction is the average of the values of k nearest neighbors. Since a lower k value would be sensitive to noise and may lead to overfitting, while using a higher k value would include more irrelevant data points and increases the bias [8], the number of k ranging from 1 to 25 was tested to obtain the best-tuned KNN model.

(2) regression tree (RT)

The basic idea of RT is recursively partitioning the input space into binary subsets where the output becomes successively more homogeneous. RT model divides the input variables into several non-overlapping spaces (tree construction) and optimizes the prediction with the greatest reduction in errors for each space (tree pruning). The modeling steps are detailed as follows:

Step 1: RT begins with the root node, which represents the whole input space and contains all training samples. First, we select the predictor $\mathrm{X_j}$ and the cutpoint s such that splitting the predictor space into region $\mathrm{R_1(j,s) = \{X|X_j < s\}}$ and $\mathrm{R_2(j,s) = \{X|X_j \geq s\}}$ leads to minimizing the expected sum of square errors (SSE) for two subsets. That is [189]:

$$\min\left[\sum_{\mathrm{i:x_i \in R_1(j,s)}} \left(\mathrm{y_i} - \hat{\mathrm{y}}_{\mathrm{R_1}}\right)^2 + \sum_{\mathrm{i:x_i \in R_2(j,s)}} \left(\mathrm{y_i} - \hat{\mathrm{y}}_{\mathrm{R_2}}\right)^2\right]$$

Next, RT repeats the process to find the best predictor and best cutpoint in order to split the data so as to minimize the SSE.

Step 2: After repeating the splitting process, considering a very large tree $\mathrm{T_0}$ with no splits, we prune it back in order to obtain the optimal subtree. The pruning approach is based on the error-complexity measure, which considers the accuracy of subtrees and that of complexity (given by the number of terminal nodes of a tree). The error-complexity measure R is defined for any node t and its branch $\mathrm{T_t}$ as [190,191]:

$$\mathrm{R_\alpha(T_t) = R(T_t) + \alpha|\overline{T}_t|}$$

where $\mathrm{|\overline{T}_t|}$ is the number of the terminal nodes (or leaves) or complexity of T, α is the threshold complexity parameter (cp) which is used for the control of tree growth and served as a penalty term for each new added split in the tree. The idea is to prune the branch $\mathrm{T_t}$ if its error-complexity measure is not lower than the error-complexity measure of its root t. To obtain the optimized subtree, we tuned the complexity parameter from 10^{-5} to 10^{0}.

(3) random forest (RF) regression

RF model fits a set of decision trees and uses averages from decision trees which is trained on a randomly selected subsample of the training data by the bagging approach. Due to its higher robustness to noise and less tendency of overfitting, RF has been applied in various fields of ML [12]. The main steps to construct an RF regression model for predicting $PM_{2.5}$ or O_3 concentration are described as follows [13]:

Step 1: The basis feature data can be formulated as follows:

$$D = \{(x_m, y_m), m = 1, 2, \dots, n\}, (X, Y) \in R^i * R$$

where Y is $PM_{2.5}$ or O_3 observations, and X is input variable matrix.

Step 2: To grow each tree (t_i), a random subspace D_i must be generated through a random selection with replacement from D, among which the variables were randomly selected for prediction with the number of variables ranging from 1 to $\sqrt{p}$, where p is the total number of variables. By repeated training, an ensemble of N trees (t_i) is grown, and each tree are de-correlated because of the random selection of input variables in each tree.

Step 3: The predicted results are calculated from an average of N trees (h_i). RF regression is an ensemble non-linear regression model. By using the idea of a double random selection of samples and variables, resulting in RF does not intend to overfit.

(4) gradient boosting model (GBM)

GBM is an improved model based on decision trees which are grown sequentially using information from previously grown trees. The core idea of GBM uses a negative gradient of the loss function as the residual approximation during growing the trees and minimizes the loss function by reducing the residuals gradually. The details steps include [14]:

Step 1: Set the predicted values $\hat{f}(x) = 0$ and the residual $r_i = y_i$ for all samples i in the training data.

Step 2: For each number of tree $b = 1, 2, \ldots, B$, repeat updating $\hat{f}$ by adding a shrunken version of the new tree, such as $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$, where λ is the shrinkage parameter that controls the learning rate of boosting, and updating the residuals like $r_i \leftarrow r_i - \lambda \hat{f}^b(x)$, until the least mean square error is the lowest. The final output of the model is:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

Therefore, GBM will build up consecutive trees that solve the net error of prior trees. We tuned the number of trees (B) from 0 to 2500 with a learning rate (λ) of 0.01 or 0.001.

(5) convolutional neural network (CNN)

CNN is generally applied for image classification and has been intensively used in air quality forecasting [15–18]. Deep CNN consists of several neuron layers: a convolutional layer, a pooling layer, and fully connected layers, as shown in Figure 2. The convolutional layer captures different signals of the image by passing many filters over each image, which can reduce the size of the input without losing important information. Mathematically, convolution is the integral measure of the extent to which two functions overlap as one passes over the other [16,19]. The activation function, such as ReLU or softmax, embedded in the convolutional layer is used to provide nonlinear transformation for reducing input data. The rectifier linear unit (ReLU) is one of the common activation functions, which is defined by:

$$\mathrm{f(x)} = \max(0, \mathrm{x})$$

The ReLU activates a node only if the input is higher than a threshold, as shown in Figure 3. The pooling layer excludes features with similar attributes and can reduce the computational burden. Among several pooling operations, the max pooling operation and the average pooling operation are the most commonly used operations. The fully connected layer flattens input features into a column vector as output [16,19].

We implemented two-layer CNN in the Keras environment [192] with the TensorFlow [193] backend with the Adaptive Moment Estimation (Adam) optimizer. Each layer consists of 40 filters and is convolved through a kernel of size 2×1. The ReLU activation function and the max pooling operation were employed to extract important features and preserve nonlinearity. The loss function used in this study is based on mean squared error (MSE).

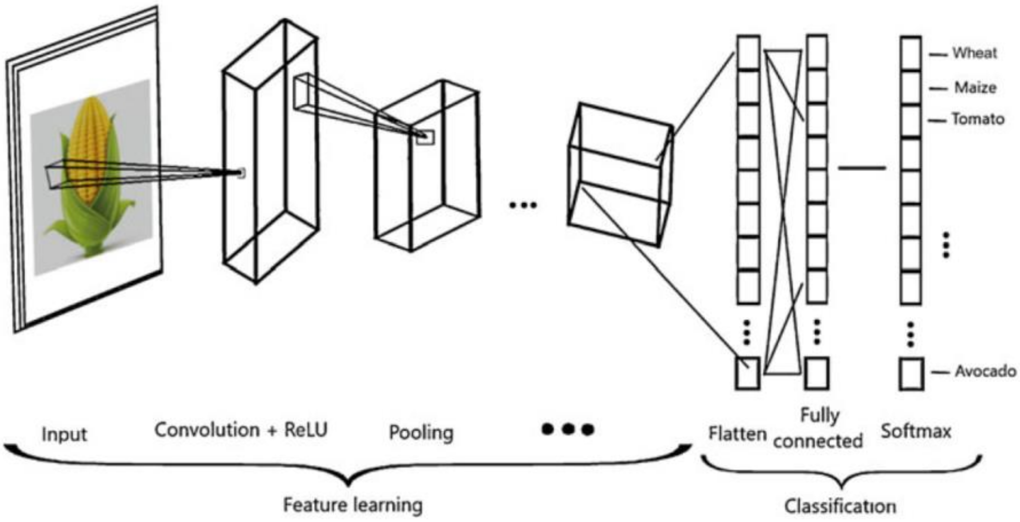


Figure 2. Basic structure of convolutional neural network (CNN) [19]

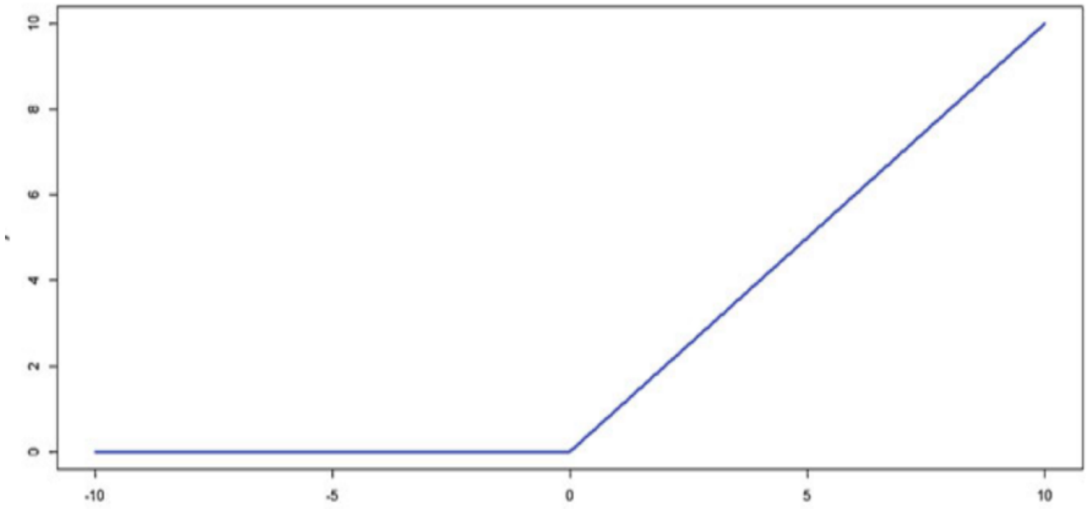


Figure 3. Example of the ReLU activation function [19]

III. Definition of long-range transboundary pollution event day

The days with long-range transboundary pollution were defined by the following two criteria:

(1) For each site, the exceedance threshold of the criteria pollutants (NOx, SO_2, CO, $PM_{2.5}$, PM_{10}, and NMHC) was defined by the daily concentration higher than the 75th quantile of daily concentrations of each pollutant in 2016.

(2) Transboundary day is defined as the day having over 50% of monitoring stations received exceedance on the same day for any of three criteria pollutants.

IV. Academic records

Table 1. Milestones for publications and conferences related to this proposal

Time	Title	Conferences / Journals	Note
01/2023	Ozone Response Modeling to NOx and VOC Emissions: Examining Machine Learning Models	(submitted)	1st author
12/2022	Apply Learning Intelligence to Quantify Biases for Measurement-model Fusion of Environmental Pollution	(in manuscript)	1st author
11/2021	Uncertainty Analysis of CMAQ-derived Burden Estimation of $PM_{2.5}$ Exposure with Bias Correction Technique	2021 CMAS	Oral
04/2021	Quantifying spatial heterogeneity of vulnerability to short-term $PM_{2.5}$ exposure with data fusion framework.	Environmental Pollution	1st author
10/2020	Assessing heterogeneity of the burden of disease of $PM_{2.5}$ exposure at diverse urbanization levels with CMAQ-fused data	2020 CMAS	Oral
08/2020	Transient risk of ambient fine particulate matter on hourly cardiovascular events in Tainan City, Taiwan	Plos One	Coauthor
06/2020	Quantifying potential uncertainty of burden of disease with spatial heterogeneity of $PM_{2.5}$ exposure risk	2020 A&WMA	2nd Place Doctoral Student Poster Competition
12/2019	Assessing relationship between heterogeneity of land-use pattern and vulnerability of residents exposing to $PM_{2.5}$: A Case Study in Taiwan	2019 AGU	Poster

Table 2. Publications and conferences for the other research topics

Time	Title	Conferences / Journals	Note
04/2022	Localized energy burden, concentrated disadvantage, and the feminization of energy poverty	Iscience	Coauthor
01/2020	Evaluating the impact of mobility on COVID-19 pandemic with machine learning hybrid predictions. Science of The Total Environment	Science of The Total Environment	1st author
06/2020	An artificial intelligence framework to forecast air quality.	2020 A&WMA	Coauthor
04/2021	How limitations in energy access, poverty, and socioeconomic disparities compromise health interventions for outbreaks in urban settings	Iscience	Coauthor
06/2021	Predicting the near future county-level COVID-19 pandemic trend under lockdown and reopen scenarios	2021 A&WMA	Poster
06/2021	Present and Future Wildfire Impacts on Dryness in a Changing Climate	2021 A&WMA	Oral
07/2021	A Hybrid Machine Learning Framework to Identify Driving Forces at Early Stage of COVID-19 Pandemic	2021 GAW Symposium	Oral
11/2021	Projections of Wildfire Impacts on Air Toxics in the Western US.	2021 CMAS	Oral
12/2021	Trend of Wildfire Impacts in Present and Future Climate	2021 A&WMA Climate Change	Oral

Vita

Cheng-pin Kuo was born in Taoyuan, Taiwan in 1989. He received his bachelor’s and master’s degree in the College of Public Health at the National Taiwan University in 2011 and 2013, respectively. He served one year of mandatory military service as a Corporal in Tri-Service General Hospital Song-shan Branch, Taipei in 2014. After a four-year experience in the environmental consultant company in Taiwan, he pursued his Ph.D. degree in Civil and Environmental Engineering at the University of Tennessee in 2018 fall. In May 2023, he graduated with the Ph.D. degree in Environmental Engineering and a minor in Computational Science.