MODELING IMPACTS OF CLIMATE CHANGE ON AIR QUALITY AND ASSOCIATED HUMAN EXPOSURES

By

TING CAI

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Environmental Sciences

Written under the direction of

Panos G. Georgopoulos

And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2019

ABSTRACT OF THE DISSERTATION

MODELING IMPACTS OF CLIMATE CHANGE ON AIR QUALITY AND ASSOCIATED HUMAN EXPOSURES

By TING CAI

Dissertation Director:

Panos G. Georgopoulos

Climate change critically affects both the atmospheric processes involved in the dynamics of air pollution systems and biogenic emissions including tree and grass pollens and fungal spores. Synergistic action of allergenic pollen with air pollutants like ozone and particulate matter has been reported as potentially exacerbating the symptoms of allergies. This dissertation investigated the spatiotemporal distributions predicted for allergenic pollen and ground-level ozone across the contiguous United States (CONUS) in 2004 and 2047 reflecting the Representative Concentration Pathways (RCP) 8.5 scenario, and estimated human exposures to those pollutants. In addition, Machine Learning (ML) methods were evaluated and applied to local-scale prediction of airborne allergenic pollen concentration.

It was estimated that ragweed pollen season will start earlier and last longer in 2047 under the RCP 8.5 scenario across the CONUS, with increasing average pollen

concentrations in most regions. The response of the oak pollen season varies across the nine climate regions of the CONUS, with the largest increase in pollen concentration occurring in the Northeast region. The oak pollen season length was estimated to shorten by 1-2 days for most regions, except for the Southeast and Southwest regions.

Analyses of observed ragweed pollen counts and ozone concentrations from 1990 to 2010 indicate that the ragweed pollen season started earlier at 76% of the monitoring stations, and the annual average number of co-occurrence of ragweed and ozone exceedances (daily maximum 8-hour average ozone $>$ 70 ppb) ranged between 0 to 17 days. Co-occurrences of ragweed pollen and ozone exceedances under climate change were investigated based on simulated ragweed pollen and ozone concentrations. Although the co-occurrence of ragweed pollen and ozone exceedances is scattered across the CONUS, it influences a remarkable fraction of the population. Inhalation exposures to ragweed pollen are higher outdoors than indoors, with significant correlation with pollen concentration. Males tend to have higher inhalation exposures to ragweed pollen and ozone than females. The inhalation exposure to ragweed pollen and ozone per unit body weight decreases with age.

Prediction of ragweed pollen concentration at the local scale, based on meteorological factors and previous ragweed pollen observations, was conducted using ML models including Support Vector Machine (SVM), Random Forest, XGBoost, Neural Network, Decision Trees, and a Bayesian Generalized Linear Model. The model parameters were optimized and the final models were evaluated using a repeated 10-fold cross-validation. Random Forest and XGBoost models outperformed other models, and pollen concentration of the previous day is the most important predictor variable for both models.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my dissertation advisor, Dr. Panos G. Georgopoulos for his tremendous support and invaluable guidance during my Ph.D. study at the Computational Chemodynamics Laboratory (CCL) at Rutgers University. His constant encouragement, unwavering pursuit of knowledge and perfection has always inspired and motivated me throughout my study.

Besides my advisor, I would like to thank all the members of my dissertation committee, Drs. Shan He (New Jersey Department of Environmental Protection), Clifford Weisel, and Tony Broccoli, for their precious time and insightful comments. I am grateful to Dr. He and Dr. Winston Hao (New York State Department of Environmental Conservation) for sharing their valuable experience in air quality modeling and helping me solve numerous problems promptly and patiently. I am thankful to Dr. Weisel, who is always available to discuss issues related to my research and progress, and who motivated me to the finishing line during difficult times. I would like to thank Dr. Broccoli for his constructive remarks on my research in climate change. I would like to thank Dr. Christopher G. Nolte (United States Environmental Protection Agency) for providing meteorological data and CMAQ output data I used in my research. I also appreciate his insightful comments on my manuscripts and generous help on CMAQ related issues. I would like to thank Dr. Leonard Bielory for providing the observed pollen counts from NAB-AAAI stations across the United States. I want to thank Drs. Qingyu Meng, Charles J. Weschler, Ann Marie Carlton, Gediminas Mainelis, Tina Fan, and Christopher Uchrin for their guidance in my research and study.

I would also like to express my gratitude to the late Dr. Paul J. Lioy, who has

been an inspiration to me in many ways. His expertise and enthusiasm in exposure science had huge impacts on my research.

I would like to thank Ms. Helene Derisi, Ms. Linda Everett and Ms. Teresa Boutillette for their great administrative and spiritual support. I would like to thank Mr. George Grindlinger (Rutgers Office of Information Technology) and Dr. Evan Patton for their tremendous work in maintaining the CCL computing cluster and moving my research forward, and for their much valued friendship.

I would also like to thank my CCL colleagues Dr. Yong Zhang, Dr. Allison Patton, Dr. Dwaipayan Mukherjee, Zhongyuan Mi, Xiang Ren, Longfei Chao, Xiaogang Tang, Steven Royce, Jocelyn Alexander, and Pamela Shade. They have helped me, encouraged me and made my study at Rutgers a wonderful experience. I would like to give special thanks to Dr. Yong Zhang, a mentor and good friend to me, who was always there to help and inspire me ever since I joined CCL. He also helped build the foundation of the CMAQ-pollen model, without which this thesis would not have been possible. I would like to thank my fellows in the Exposure Science Program and EOHSI colleagues for their help and companionship in the past years. I also want to express my sincere gratitude to my friends in the Church in Piscataway for their support to me and my family.

I could not have pursued my Ph.D. degree without the love and support from my family. I am deeply indebted to my parents, Yongzhong Cai and Chunmei He, and my parents-in-law, Liping Han and Fuquan Yuan, who have been supporting me in every possible way. I would also like to thank my brother Gen Cai and sister-in-law Biyu Hou for their great support to our family. My deepest gratitude goes to my husband, Dr. Ye Yuan, who shares my happiness, frustration and faith, gives me strength and cheers me up. Our beloved son Eric has made this journey especially meaningful.

Last but not least, I would like to acknowledge the Center for Environmental Exposures and Disease at EOHSI and the Ozone Research Center funded by the NJDEP for supporting my research.

Dedication

To my husband Ye Yuan, my son Eric Yuan

And to my parents Yongzhong Cai & Chunmei He

Table of Contents

List of Tables

List of Figures

Chapter 1
INTRODUCTION

1.1 Motivation

It has been established that climate change has broad impacts on human society and natural systems, including the economy, human health, human behavior, agriculture, water resources, ambient air quality, etc. Human activities such as burning of fossil fuels and converting forests for agriculture have played a significant role in driving climate change [4]. Ambient air quality has been substantially affected by climate change over the past decades; these effects are expected to increase in the future. Ambient ozone is an air pollutant of primary concern for public health in the United States (U.S.) and many studies predict that its concentration patterns will be impacted by climate change [5, 6, 7]. Climate change will affect atmospheric processes involved in the dynamics of air pollution systems, potentially leading to increased levels of ozone and other photochemical pollutants in certain areas.

In the U.S. Global Change Research Program (USGCRP)'s report [8], it is projected that the number and severity of wildfires in the U.S. are going to increase due to climate change, leading to increasing emissions of particulate matter and ozone precursors. Climate change is also critically affecting emissions of natural pollutants such as pollen and spores as well as biogenic gases which are components of atmospheric photochemistry reaction systems. Another key findings of the USGCRP report is that rising temperatures and changes in precipitation will also increase the levels of aeroallergens including pollen. Like wildfires, the ragweed pollen season is another

climate change indicator, and it has been observed (Figure 1.1) that the length of ragweed pollen season at various locations in the central U.S. and Canada has increased by 6 to 24 days between 1995 and 2015 [9, 1]. The Asthma and Allergy Foundation of America [2] reported that ragweed pollen production has been continuously increasing with rising global CO_2 levels since 1996 (Figure 1.2). Previous studies on nationwide observations of airborne pollen counts of selected plant species and climatic factors indicated that the start date and length of pollen season, the average peak value and annual total of daily counted airborne pollen have been affected substantially by the changing climate [3, 10, 11, 12, 13].

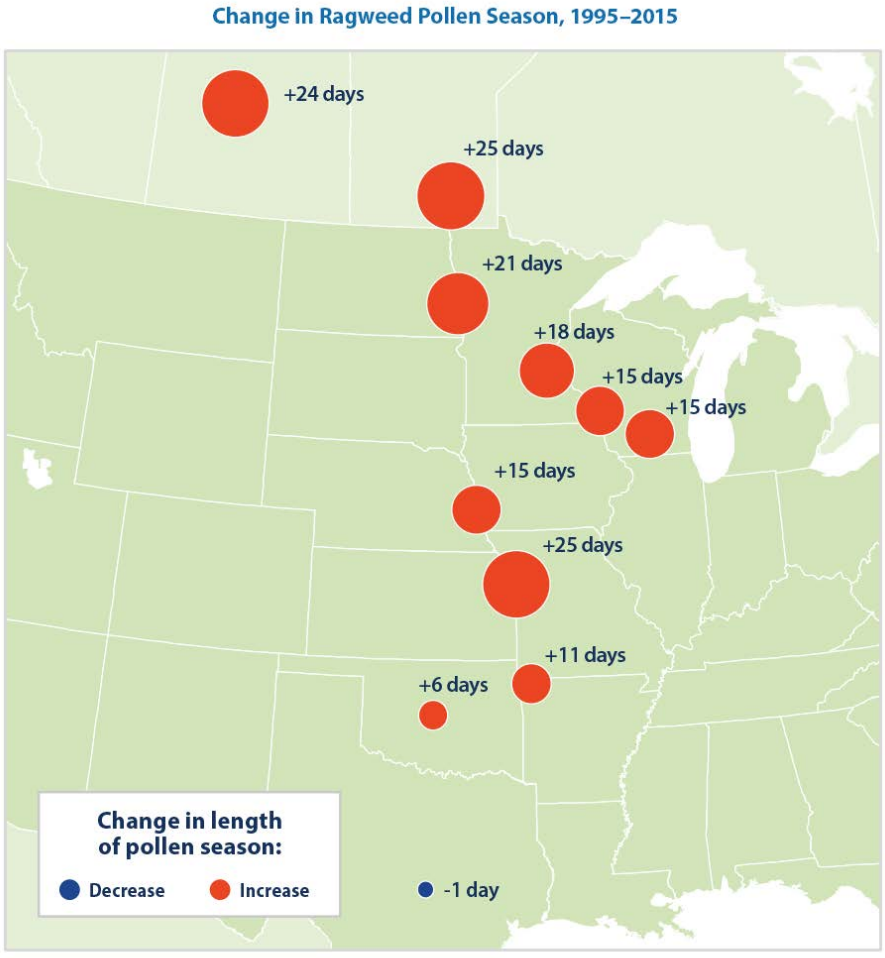


Figure 1.1: Changes in ragweed pollen season length (1995-2015) at 11 locations in the U.S. [1]

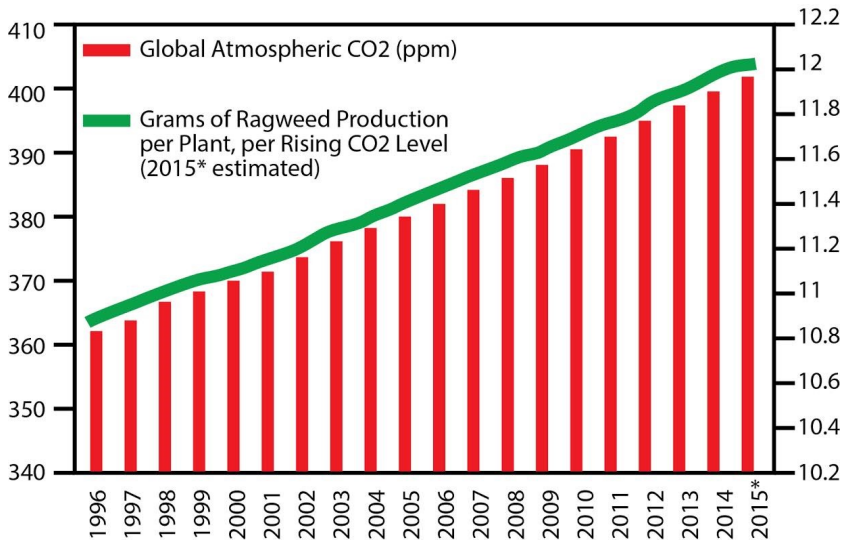


Figure 1.2: Observations of increased ragweed pollen production with rising global CO_2 level during 1996-2015 [2].

The prevalence of Allergic Airway Diseases (AAD) has grown globally in recent years resulting in increased numbers of emergency department visits and hospitalizations [14]. Clinical studies have shown that AAD can be exacerbated by the synergistic action of bioaerosols such as pollen and fungi, and atmospheric pollutants such as ozone and $PM_{2.5}$ [15]. According to the 2015 Natural Resources Defense Council (NRDC) report [16], 275 counties, where 109 million people reside, had been exposed to both ragweed pollen and unhealthy ozone levels in 2009 to 2013, and ozone concentration and ragweed pollen count are likely to increase simultaneously in some areas with climate change.

Climate change is affecting not only the ambient air pollutants, which is a major source of indoor air pollution, but also indoor air quality by directly altering the air pollutants produced indoors, including mold and volatile organic compounds [8]. Indoor exposures are critical in accessing public health since people in the U.S. spend on the average 90% of their time indoors [17]. Although extensive research has been done on the chemistry of indoor air pollutants and on impacts of climate change on public health, not many studies have focused on the impacts of climate change on indoor environments and associated public health impacts [18, 19].

Exposures to ozone can cause a wide spectrum of acute and chronic health effects, ranging from respiratory symptoms such as asthma and reduced lung function, to cardiovascular disease and even possibly cancers [20, 21, 22, 23]. Correlations between personal exposure and ambient air pollutant concentrations have been used to predict human exposures over time. Differences in outdoor ozone concentrations can lead to different ozone mortality coefficients among cities [24]. It has been reported that the personal-ambient ozone correlation coefficients are in the range 0.3-0.8. Lower correlations appear with increased time spent indoors and low indoor-outdoor air exchange rates [25]. Ambient ozone resulting from photochemical reactions and injection from the stratosphere, is transferred to indoor environment by ventilation and infiltration, and acts as the main source of indoor ozone [26, 27, 28]. Other sources of indoor ozone include photocopiers, laser printers, electrostatic air filters and electrostatic precipitators, commercial ozone generators for air purification, etc. [26], but such equipment are not commonly present in residential microenvironments. Indoor ozone concentrations are usually lower than outdoor ozone concentrations due to ozone being a highly reactive compound. However, indoor exposures to ozone often account for more than 50% of the total personal ozone exposure, since, as mentioned earlier, people spend most of their time indoors [23].

Due to its strong oxidative properties, ozone can react with many indoor organic species and generate toxic byproducts including formaldehyde, acrolein, and secondary organic aerosols (SOA), which also pose public health concerns [26, 29, 30]. Extensive studies on the reaction of limonene and ozone have been performed during the past decades [31, 32, 33, 34, 35, 36, 37]. The reaction products include gaseous species such as formaldehyde, hydrogen peroxide, and relatively short-lived reactive oxygen species (ROS), such as hydroxyl radicals, ozonides, peroxyhemiacetals, and hydroperoxides [38, 32, 33, 39, 40, 41, 42]. Although short-lived, many of these products exist long enough to be inhaled and transported into the respiratory tract. In

addition, recent studies show that ozone can also react with human skin lipids forming products that may be respiratory and skin irritants [43, 44, 45]. Squalene, the most abundant unsaturated compound in human sebum, is the major precursor for these oxidation products [46, 47, 48].

Climate change affects air quality in different ways, directly or indirectly. Meteorological factors such as temperature, cloudiness, precipitation, wind speed, etc. will influence air quality directly by affecting biogenic emissions, photochemical reactions rates, and the transport and deposition patterns of air pollutants. Climate change is also affecting human population activities, which leads to variations in emissions and meteorology, then affecting air quality indirectly. Therefore, it is challenging to model the impacts of climate change with all of the interacting variables. Researchers have to isolate the impacts on air quality caused by climate-driven changes in meteorological factors and related natural emissions by keeping other factors (such us anthropogenic emissions) unchanged. Changes of ozone concentrations under climate change and emission controls have been modeled using the Community Multiscale Air Quality (CMAQ) modeling system by many research groups. These studies differ in spatiotemporal resolutions and emission scenarios. It is known that surface ozone levels have strong seasonal cycles which vary with latitude and altitude, with the maximum usually occurring in spring to summer [49]. It is often projected that the continued rise in anthropogenic emissions, especially from developing countries, will have substantial impacts on the global surface ozone levels. Meanwhile, efforts have been made by countries to decrease the greenhouse gases (GHG) emissions to mitigate the impacts of climate change. Predicted changes in ozone concentrations will vary, depending on the emission scenarios used in the simulation.

Although a number of studies have evaluated the effects of climate change on ambient ozone concentrations, very few studies have been done for allergenic pollen, or for simultaneous evaluation of impacts of climate change on ozone and allergenic pollen and related human exposures. The co-occurrence of ozone exceedance and

allergenic pollen under changing climate has not been investigated. There also exists a major lack of assessment of population co-exposure to allergenic pollen and ozone across the contiguous United States (CONUS). Factors affecting personal exposure to allergenic pollen and ozone need to be identified.

1.2 Background

1.2.1 Modeling of future ozone concentration

Climate models are built with future scenarios of forcing agents, such as greenhouse gases and aerosols, as input to make projections of future climate changes. The Intergovernmental Panel on Climate Change (IPCC) developed a number of emission scenarios (A1, A2, B1, B2) based on different future socioeconomic conditions [50]. For example, the A1 scenario assumes very rapid economic growth in the future with global population peaks in mid-century and declines thereafter, and the rapid introduction of new and more efficient technologies. The three A1 groups are distinguished by their technological emphasis: fossil intensive (A1FI), non-fossil energy sources (A1T), or a balance across all sources (A1B) [50]. Table 1.1 lists selected modeling efforts since 2008 for predicting ozone concentration across the CONUS in mid-21st century with horizontal grid resolution of 36x36 km or higher. Using a coupled global/regional scale modeling system, Nolte *et al.* [51] predicted that the mean summertime daily maximum 8-hour average ozone (DMA8[O_3]) in Texas and parts of the eastern U.S. will increase by 2-5 ppb under the IPCC A1B greenhouse gas scenario by 2050 if the anthropogenic emissions remain at 2011 levels. In contrast, large decreases in the mean summertime DMA8[O_3] were projected for the case when anthropogenic emissions are reduced based on the A1B scenario. Tagaris *et al.* [52] also found that emission reductions and climate change together will decrease the mean summer DMA8[O_3] (-11% to -28%) in 2050 across the U.S. under the IPCC A1B emission scenario. Tagaris *et al.* [7] further investigated the health effects of

ozone under the IPCC A1B emission scenario while keeping the emissions, population, and pollution controls the same. They found that the annual mean ozone would increase for the northern U.S., but decrease for the southern part of the U.S. They also predicted that the additional annual premature deaths caused by ozone changes due to climate change would be about 300. Also under the IPCC A1B emission scenario, Lam *et al.* [53] found that by 2050 the DMA8[O_3] would decrease by 5 ppb in the eastern U.S. under the combined effect of climate change and emission reductions. Dynamical downscaling was applied in their study to improve the horizontal grid resolution from 36x36 km to 12x12 km [53]. Modeling results from a study by Gonzalez-Abraham *et al.* [54] suggested that DMA8[O_3] will increase by 2-12 ppb for most parts of the CONUS under combined effects of climate, biogenic emissions, land use and anthropogenic emissions. Penrod *et al.* [55] examined the impacts of future climate and anthropogenic emissions on ozone concentrations for 2026-2030 and found that their levels will decrease due to reduction of anthropogenic emissions. The increases in surface ozone in 2050s are largely driven by temperature, solar radiation, and cloud fraction over most of the domain [56, 57, 58].

The projections of future ozone discussed above are all based on the IPCC A1B emission scenario, which was commonly used before the introduction of the Representative Concentration Pathways (RCPs) [59]. Due to the fact that the RCP scenarios apply different emission and climate models considering future socio-economic and emission scenarios, the projections for future ozone based on them differ from those based on IPCC A1B scenarios. Fann *et al.* [60] estimated the impacts of climate change on the ozone-related health impacts in 2030 under two greenhouse gas forcing scenarios, RCP 8.5 and RCP 6.0. Simulations for both scenarios projected that the average daily maximum temperature will increase by 1-4 °C from base year 2000. The RCP 8.5 scenario showed an increase of DMA8[O_3] by 1-5 ppb all across the CONUS, while the RCP 6.0 scenario projected decreases of ozone over the Pacific Northwest and Gulf Coast areas and increases of ozone over other regions. The emissions for

future year 2030 were based on the United States Environmental Protection Agency (USEPA) estimates assuming the planned air quality policies were implemented. Gao *et al.* [61] applied the RCP 8.5 and RCP 4.5 scenarios for projection of ozone levels in 2050 and compared them with a base year of 2000. They used a coupled global and regional climate model to simulate ozone levels over the CONUS domain at 12x12 km spatial resolution. The results from both scenarios project obvious seasonal variations, with the RCP 4.5 scenario projecting a significant decrease (6-10 ppb) in ozone concentrations during summer, while the RCP 8.5 scenario shows an increase of 3-7 ppb in ozone concentrations in winter. The 2005 emissions used for base years and the future year emissions were projected based on RCP databases. Kim *et al.* [62] further studied the ozone-related health impacts in the US based on the future ozone levels simulated by Gao *et al.* [61]. They reported that the average ozone-related excess premature deaths in 2050 compared with 2000 under RCP 4.5 and RCP 8.5 scenario would be -2118 deaths/year and 1312 deaths/year, respectively. Sun *et al.* [63] estimated the future ozone levels and its impacts on mortality for 2050s under RCP 8.5 scenario over the CONUS at a 12x12 km resolution. They found that the annual mean DMA8[O_3] will increase across the Western U.S., but decrease in the Eastern U.S. Pfister *et al.* [64] simulated the summertime ozone across the U.S. in 2050 under RCP 8.5 scenario and A2 climate. The results indicated that the surface ozone will increase for most of the U.S. and the 5^{th}-95^{th} quantile of DMA8[O_3] will increase from 31-79 ppb to 30-87 ppb. Nolte *et al.* [65] reported that DMA8[O_3] is predicted to increase during summer and autumn in the central and eastern US in 2030s under three RCPs. Temperature and isoprene emissions were found to be the biggest contributors on changes of ozone concentrations.

It is generally desirable to have finer horizontal resolution when it comes to air quality modeling because it gives more information on specific areas of interest. However, fine resolution modeling requires high resolution input data, high computational

power and it is also very time-consuming for large domains. For instance, the computational requirements will increase 27 times when the horizontal resolution increases from 36x36 km to 12x12 km [66]. Coarse spatial resolution might underestimate the concentration of air pollutants in urban areas, causing biases in the estimation of health impacts due to air pollution [67]. Dynamical downscaling techniques have been used on outputs from global climate models to generate inputs for higher resolution simulations. Gao *et al.* [68] developed the Community Earth System Model (CESM) and applied the dynamical downscaling techniques on the CESM outputs for regional WRF (Weather Research and Forecast) model inputs; they were able to downscale the climate results from the 36x36 km North American domain to a 4x4 km Eastern U.S. domain.

1.2.2 Modeling of allergenic pollen concentration

Synergistic action of allergenic pollen with air pollutants like ozone and particulate matter has been reported and may exacerbate AAD symptoms [88, 15, 89, 14]. In particular, the patterns of emission and transport of allergenic pollen and common air pollutants are expected to be impacted by the changing climate [90, 61, 91, 92, 93, 12]. Studies on emission and transport of allergenic pollen from multiple taxa are needed in order to estimate their spatiotemporal distributions and the potential consequences for public health. Previous studies of emissions and transport of pollen are summarized in Table 1.2. Pollen emissions have been simulated using mostly empirical models [76, 77, 78, 72] and probabilistic models [81, 80, 79]. Pollen dispersion and transport were predicted using Gaussian model [94, 74], Lagrangian model [75, 95], Eulerian model [96] and chemical transport model such as CMAQ [3, 69, 85, 86, 87].

Kawashima and Takahashi [74] estimated the emission and dispersion of airborne cedar pollen in Japan using a Gaussian model. This model took into account various physical processes and spatial variation of flowering time using a flowering-time map. However, it could not reproduce the pollen peaks in urban areas. Using the Fifth

Generation Mesoscale Model (MM5) and the National Oceanographic and Atmospheric Administration (NOAA) Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) Model, Pasken and Pietrowicz [75] forecasted daily oak pollen concentration with 30-min increments with inconsistent performance for each day. Helbig *et al.* [76] proposed an emission model that takes into account meteorological variables such as wind stress, temperature and humidity. This emission model was incorporated into a comprehensive modeling system (COSMO) to simulate birch pollen concentration in Switzerland using COSMO-ART (Aerosols and Reactive Trace Substances) [77]. COSMO-ART was further applied for release and transport of ragweed and birch pollen in central Europe by Zink *et al.* [78, 71, 72]. Sofiev and Bergmann [80] developed a numerical model of birch pollen emission based on a double-threshold temperature sum model, ambient humidity and precipitation rate, wind speed and turbulence intensity. The emission model was evaluated by Siljamo *et al.* [79] with System for Integrated modeLling of Atmospheric coMposition (SILAM) for birch pollen dispersion in Europe. Most of the existing modeling studies have been designed to provide operational forecast of airborne pollen levels in Europe. Large scale deterministic emission and transport models need to be developed to investigate the release, transport and health effects of allergenic pollen in the U.S.

CMAQ as a chemical transport model has gained more application in this field since it can simultaneously calculate the concentrations of multiple air pollutants in gaseous or particle phase [85, 84, 78]. Efstathiou *et al.* [85] modified the parameterizations developed by Helbig *et al.* [76] to predict the emission and transport of birch and ragweed pollen over the northeastern U.S. using CMAQ and MM5. Jeon *et al.* [69] employed the emission model based on the parameterizations described by Efstathiou *et al.* [85] to simulate oak pollen emission in Southeast Texas and fed the emissions into the CMAQ model to predict the oak pollen concentration. However, the model underestimated the oak pollen concentration and was not able to capture the peaks of oak pollen concentration. A modeling framework has been developed

to incorporate WRF model and CMAQ model to study distributions of multiple allergenic pollens in southern California under climate change scenarios [86, 87]. A semi-mechanistic model has been developed for emission of airborne allergenic pollen by Zhang [3] and Cai *et al.* [97]. Based on this pollen emission model, an adapted CMAQ modeling system (CMAQ-Pollen) was used to simulate the spatiotemporal distributions of allergenic pollen including ragweed, mugwort, birch, oak and grass across the CONUS with 50x50 km grid resolution [3].

Table 1.1: Representative modeling efforts on future ozone prediction.

Base year	**Future year**	**Spatial resolution**	**Season evaluated**	**Air quality model**	**Future emission scenario**	**Reference**
1999-2003	2048-2052	36x36 km	Summer, fall	CMAQ v4.5	IPCC A1B	[51]
2001-2002	2051-2052	36x36 km	Summer	CMAQ v4.4	IPCC A1B	[56]
2001	2050	36x36 km	Summer, annual	CMAQ v4.7	IPCC A1B	[7]
1999-2001	2049-2051	36x36 km	Annual	CMAQ	IPCC A1B	[53]
2006-2010	2048-2052	36x36 km	Summer, annual	CMAQ v4.7.1	IPCC A1B	[57]
1995-1999	2048-2052	30x30 km	Summer	CMAQ v4.6	IPCC A1B	[58]
2001-2005	2026-2030	36x36 km	Summer, winter	CMAQ v5.0	IPCC A1B	[55]
1995-2004	2045-2054	36x36 km	Aunnal	CMAQ v4.7.1	IPCC A1B	[54]
1996-2008	2046-2058	36x36 km	Summer	CAM-Chem & NRCM-Chem	IPCC A2 & RCP 8.5	[64]
2001-2004	2057-2059	12x12 km	Annual	CMAQ v5.0	RCP 4.5 & RCP 8.5	[62]
2002-2004	2057-2059	12x12 km	Annual	CMAQ v5.0	RCP 8.5	[63]
1995–2005	2025–2035	36x36 km	Annual	CMAQ v5.0.2	RCP 4.5, RCP 6.0 & RCP8.5	[65]
2001-2004	2057-2059	12x12 km	Annual	CMAQ v5.0	RCP 4.5 & RCP 8.5	[61]
1995-2005	2025-2035	36x36 km	Annual	CMAQ v5.0.1	RCP 6.0 and RCP 8.5	[60]

Table 1.2: Modeling studies on pollen emission and transport.

Study location	**Spatial resolution**	**Simulation year**	**Plant taxon**	**Emission model**	**Transport model**	**Reference**
Texas, USA	4x4 km	2010	Oak	Empirical model	CMAQ	[69]
CONUS	50x50 km	2001-2004, 2047-2050	Oak, ragweed, birch, mugwort, grass	Mechanistic model	CMAQ	[3]
Europe	50x50 km	2000-2010	Ragweed	CLM v4.5	RegCM4	[70]
Central Europe	0.06°x0.06°	2012	Birch	Empirical model	COSMO-ART	[71]
France	NA	2012	Ragweed	Empirical model	COSMO-ART	[72]
Spain	50x50 km	1992-2004,2025, 2050,2075,2099	Oak	Stepwise regression	Stepwise regression	[73]
Japan	10x10 km	1990	Cedar	Meteorological parameterization	Gaussian model	[74]
CONUS	12x12 km	2000	Oak	Uniform diurnal profile	HYSPLIT	[75]
Germany	4x4 km	2000	Hazel, alder	Empirical model	DRAlS CTM	[76]
Switzerland	7x7 km	2006	Birch	Empirical model	COSMO-ART	[77]
Central Europe	7x7 km	2006	Ragweed	Empirical model	COSMO-ART	[78]
Europe	0.25°x0.25°	2006	Birch	Probabilistic model	SILAM	[79, 80]
Europe	Multiple resolution and layers	2010-2013	Birch	Probabilistic model	Multiple Ensemble Members	[81]
Finland	1x1 km	2002-2004	Birch	Aerobiology observations	SILAM	[82]
Germany	500x500 m	2000	Oak	Meteorological parameterization	METRAS	[83]
Europe	0.25°x0.25°	2005-2011	Ragweed	Empirical model	SILAM	[84]
NorthEastern USA	12x12 km	2002	Birch, ragweed	Empirical model	CMAQ	[85]
California USA	12x12 km, 4x4 km	1995–2004, 2045–2054	Multiple Taxa	STaMPS	CMAQ	[86, 87]

1.2.3 Exposures to allergenic pollen and ozone

Pollen exposure is typically calculated based on measurements at monitoring stations [98]. Peel *et al.* [99] investigated the relation between grass pollen dose and pollen concentration at pollen monitoring station and observed a median ratio of dose rate to background concentration of 0.018. Riediker *et al.* [100] reported an average indoor-outdoor ratio of 0.2 for pollen when considering personal activity patterns. However, the accuracy of exposure assessments can be limited by the availability of pollen monitoring stations and the distances between the subjects and the stations. Personal pollen samplers have been proposed to measure the actual pollen concentration people are exposed to [101, 102, 103]. Yamamoto *et al.* [102] developed the Personal Aeroallergen Sampler (PAAS), a passive sampler for aeroallergens, for personal exposure assessments of cedar and cypress pollens. They found that the seasonal peak of the personal pollen exposures was not consistent with the outdoor concentrations, which indicates insufficiency of the stationary pollen outdoor monitoring. Myszkowska *et al.* [104] compared pollen counts from personal pollen sampler and stationary pollen monitoring and found that results differed for varied conditions. The average inhalation exposure to ragweed pollen during a typical ragweed season in Baltimore was estimated to be 960 pollen grains/day for an average 11 year old child and 1536 pollen grains/day for an adult [105]. Driessen and Quanjer [106] reported that personal pollen inhalation exposure ranged from 2,500 to 20,000 pollen grains/day when pollen concentration are 250-1,000 pollen grains/m^3. Zhang [3] simulated population exposures to five taxa of pollen with observed pollen counts across the CONUS. The changes of exposures to pollen between 1990s and 2000s in the nine climate regions of the CONUS were also compared. It was found that inhalation and dermal deposition were the dominant exposure routes for allergenic pollen. The average inhalation exposure to oak pollen varied from 42 pollen grains/(day kg BW) in the Southwest region to 1,073 pollen grains/(day kg BW) in the South region. The aggregated exposure to allergenic pollen in outdoor environments was estimated to

be about three to four times of that in indoor environments.

Compared with pollen, ozone has a higher penetration rate and the indoor ozone concentration can be a significant fraction of ambient ozone concentration [27]. Therefore, ozone exposures indoors and outdoors are both important. The indoor-to-outdoor (I/O) ratios of ozone concentration range from < 0.1 to 0.9 depending on indoor environments and ventilation rates [27]. Ozone concentration at monitoring stations is a common surrogate for outdoor ozone concentration. Personal ozone monitoring is also used for ozone exposure assessment as the ozone concentration has a large variation in different micro-environments. Brauer and Brook [107] compared personal and stationary ozone monitoring with a passive sampler and found that ozone exposure based on ambient monitoring would be misclassified because time-activity patterns have significant impacts on ozone exposure. Total personal ozone exposure is affected by ozone concentration and time spent indoors and outdoors [26, 23]. Weschler [23] reported that the daily inhalation intake of ozone indoors is 25 to 60% of total daily ozone intake. The total personal daily ozone intake varied from 100 µg/day to 700 µg/day [23, 108, 109, 110, 111, 107, 112, 107]. Geyh *et al.* [109] measured personal exposure to ozone of approximately 200 school children 6-12 years old in Southern California for a year using personal samplers and ozone monitors at homes, and found that personal exposure to ozone differed by location and gender, but not by age group. Lee *et al.* [110] studied outdoor/indoor ozone exposures of 10-12 year old children in Tennessee. The results showed that personal ozone exposure has a positive correlation with time spent outdoors. Liu *et al.* [111] assessed personal exposure to ozone of 23 children in Pennsylvania using indoor, outdoor and personal ozone measurements and found that personal activity data are crucial inputs for ozone exposure.

1.2.4 Machine Learning for pollen prediction

Most of the studies on modeling and forecasting of allergenic pollen concentration under changing climate fall into two categories: numerical and statistical models. Numerical models, mostly regional-scale, can predict pollen concentrations based on mathematical equations, plant distributions, phenological, aerobiological, and meteorological data [81, 79]. Statistical models are mostly local-scale and try to predict pollen concentrations based on statistical relations between airborne pollen concentration and independent variables such as meteorological factors and previous pollen observations [113]. Machine Learning (ML) methods, a family of statistical techniques that originated from the field of artificial intelligence, have drawn increasing attention in this field due to their flexibility and capability to handle complex problems with multiple interacting elements [114, 115, 116].

ML algorithms can be divided into the following four categories according to their purpose:

- Supervised Learning algorithms deal with datasets that have both inputs and outputs. In other words, the outcomes are labeled. They can solve both classification and regression problems, and are mainly used in predicting modeling [117, 118].

- Unsupervised Learning algorithms take a dataset that contains only inputs (outcomes not labeled) and try to find structure in the data, by clustering or grouping of data points. They are generally used for descriptive modeling [119].

- Semi-supervised Learning is in-between supervised and un-supervised learning. It is important for problems in which the data are a combination of labeled and unlabeled data.

- Reinforcement Learning algorithms use observations collected from the interaction with the environment to make a specific decision. They are often used in

game theory, control theory, simulation-based optimization, etc. [120, 121].

Some of the commonly used Supervised Learning algorithms are Support Vector Machines (SVM), Decision Trees, Artificial Neural Networks (ANN), k-nearest Neighbor Algorithm, Naive Bayes, Logistic Regression, Linear Regression etc. SVM maps the input vectors to a high-dimension feature space, in which a hyperplane is constructed to separate or classify the data points [122]. Decision Trees predict the value of a target variable based on several input variables [123]. Based on the type of outcome variables, there are classification trees and regression trees. Classification and Regression Tree (CART) analysis refers to both types of Decision Trees [124]. ANN are computing systems that mimic biological Neural Networks [125]. They are a set of connected neurons organized in input, hidden and output layers. Random Forests or random decision forests are an ensemble model made of many Decision Trees constructed based on random subsets of features for classification or regression [126, 127]. They were developed to resolve the dilemma between over-fitting and achieving maximum accuracy in Decision Trees algorithm.

Several studies have applied ML methods such as Neural Networks [113, 128, 129] and Random Forest [114, 115] for pollen concentration prediction. However, the input variables and model validation methods all differ between these studies and therefore it is hard to compare these models against each other.

1.3 Main Hypotheses and Objectives

Based on the review above, the following hypotheses are proposed:

- **Hypothesis 1**: Climate change will affect the onset, duration and spatiotemporal distribution of allergenic pollen across the CONUS.
- **Hypothesis 2**: High ozone and high allergenic pollen concentrations are likely to occur at the same time, affecting human health in a synergistic manner.

- **Hypothesis 3**: Population exposures to ozone and ragweed pollen vary by age, gender and location.

- **Hypothesis 4**: Daily mean pollen concentrations can be predicted using meteorological parameters via ML methods.

The overarching goal of this dissertation is to investigate the patterns of co-occurring pollen and ground-level ozone across the CONUS under changing climate, and to estimate human exposures to those pollutants. In addition, ML methods will be evaluated for local-scale prediction of allergenic pollen concentration. The proposed hypotheses will be tested by addressing the following objectives:

- **Objective 1**:Examine spatiotemporal distributions of allergenic pollen and ozone concentrations across the CONUS in 2004 and in 2047 under the RCP 8.5 scenario.

- **Objective 2**: Examine how climate change may affect patterns of co-occurring pollen and ground-level ozone across the CONUS.

- **Objective 3**: Simulate population exposures to allergenic pollen and ozone employing results from Objective 1, human behavior characteristics, demographic, housing, and activity databases available at the Computational Chemodynamics Laboratory (CCL).

- **Objective 4**: Test different ML methods for prediction of allergenic pollen concentrations in local scales and find the best fit model.

1.4 Dissertation Overview

The overall modeling framework (Figure 1.3) of the present study is constructed with main components such as the WRF, Sparse Matrix Operator Kernel Emissions

(SMOKE), CMAQ, exposure models, and ML models. Multiple databases are available to be used as input to drive these models. The major databases used in this study are summarized in Table 1.3. Meteorology is being downscaled with WRF from the Community Earth System Model (CESM) following Representative Concentration Pathway (RCP) 8.5 scenario for the CONUS. The WRF outputs are processed with Meteorology-Chemistry Interface Processor (MCIP) to provide inputs for the CMAQ-Pollen modeling system (a customized version of CMAQ 4.7.1 extended with pollen modeling components), that has been developed by CCL [3] to model the spatiotemporal distributions of allergenic pollen. Observed daily allergenic pollen counts are collected from certified monitoring stations of the National Allergy Bureau (NAB) of the American Academy of Allergy, Asthma & Immunology (AAAAI) across the CONUS and parts of Canada. Historical ozone concentrations for the pollen monitor stations are available through the EPA air quality databases [130]. Population exposure to pollen and ozone are simulated with the Modeling Environment for Total Risk Studies (MENTOR) system developed by CCL, employing demographic, housing, and activity databases available as components of the MENTOR system [131].

In Chapter 1, the motivation and background of this study are introduced. Current status and progress of studies on impacts of climate change on ambient air quality are reviewed. Different air quality models are compared. Research on application of ML methods in air quality monitoring and forecasting are summarized. Then the hypotheses, objectives and organization of this dissertation are outlined.

Chapter 2 presents the spatiotemporal profiles of simulated allergenic pollen (oak and ragweed) concentration in a historical year (2004) and s future year (2047). The CMAQ-Pollen model was evaluated using various metrics. The changes of pollen indices (mean/maximum pollen concentration, pollen season start date/season length, pollen concentration exceedance hours) for each of the nine climate regions between historical year and future year are examined.

Chapter 3 presents analyses of the observed ragweed pollen and ozone concentrations. The changing patterns of ragweed pollen indices are investigated. The co-occurrence of ragweed pollen and ozone exceedances at selected locations are reported. The spatiotemporal distribution of simulated ragweed pollen and ozone concentrations in 2004 and 2047 are presented and compared. The co-occurrence of ragweed pollen and ozone exceedances across the CONUS and its impacts on population are then evaluated. Inhalation exposures to ragweed pollen and ozone are simulated based on CMAQ model outputs, demographic, human activity and exposure factors.

In Chapter 4, six ML methods are tested for prediction of ragweed pollen concentration at Newark, NJ. Five ML techniques are assessed for ragweed pollen level classification. The performances of each model were compared and the best model is reported. Limitations and advantages of ML techniques are presented.

In Chapter 5, the major findings of this dissertation and recommendations for future research are summarized. Applications of the research results from this study are also suggested.

Table 1.3: Information on major databases for this study.

Data	Period	Spatiotemporal resolution	Model/Observation	Domain	Note
Meteorology	2004 2047	36×36 km, hourly	WRF, MCIP	CONUS	From Dr. Christopher G. Nolte in EPA
Pollen emissions	2004 2047	36×36 km, hourly	SMOKE, pollen emission model	CONUS	Simulated
CMAQ ozone output	2004 2047	36×36 km, hourly	CMAQ	CONUS	From Dr. Christopher G. Nolte in EPA
Pollen count (oak, ragweed)	1994-2010	Daily, 86 AAAAI stations	Observation	CONUS	From Dr. Bielory
Ozone concentration	1994-2010	Hourly/Daily	Observation	CONUS	From EPA website
Observed meteorological factors	1994-2010	Hourly/Daily	Observation	CONUS	From NOAA website
Land Use Land Coverage	Established in 1998	1×1 km; Stationary	BELD3.1	CONUS	Redistributed from 1×1 km to 36×36 km
Consolidated Human Activity Database	2000	N/A	Survey	CONUS	McCurdy *et al.* (2000)
US Census Bureau Demographic Data	2000	Census tract	administrative data, survey and census data	CONUS	US Census Bureau

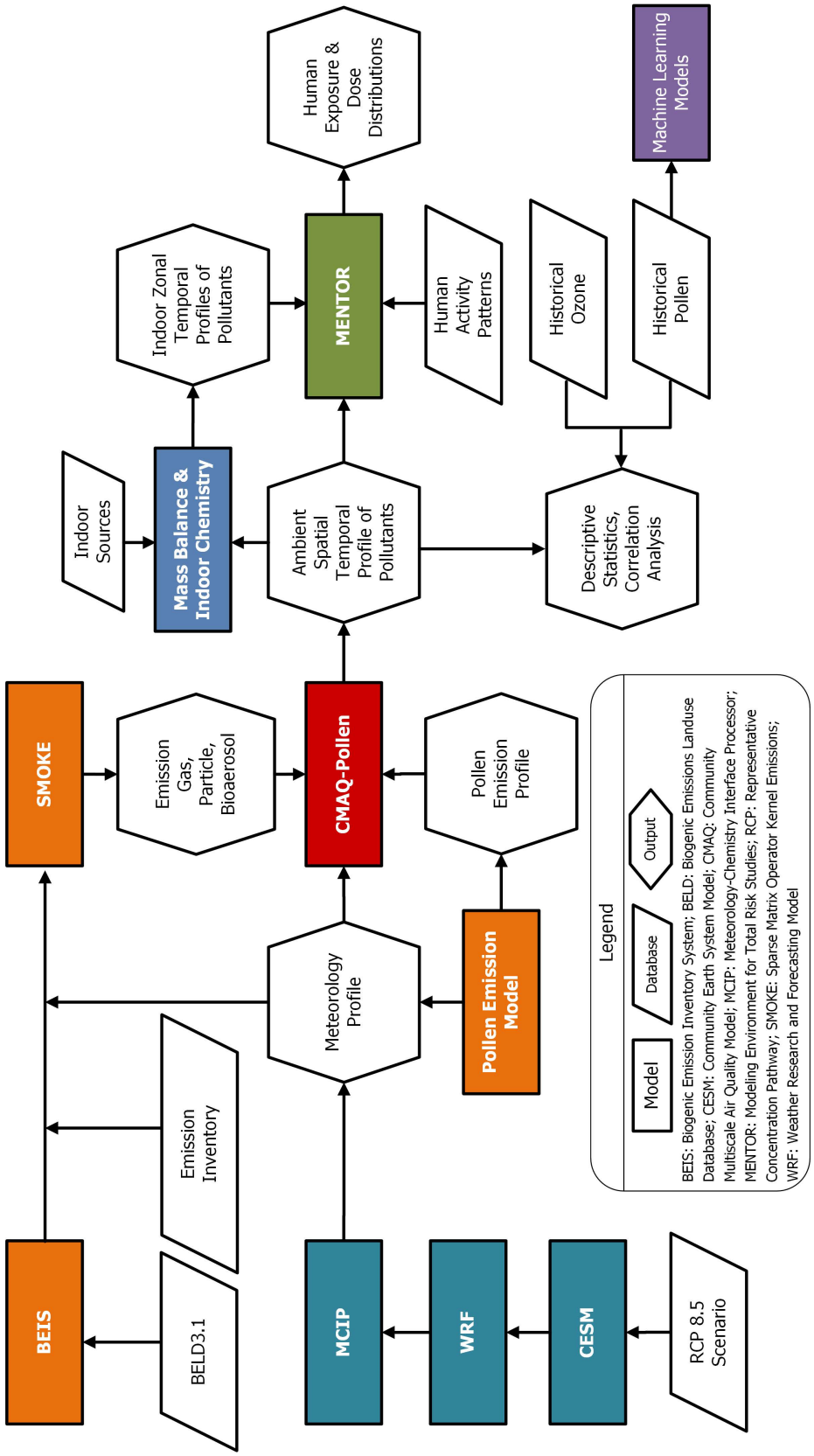


Figure 1.3: Schematic diagram of the CCL modeling system for studies of climate change effects on air quality and human exposures. (Adapted from Zhang [3])

Chapter 2

ALLERGENIC POLLEN CONCENTRATION UNDER CLIMATE CHANGE

Material in this chapter has been previously published or submitted for publication as:

Ting Cai, Yong Zhang, Xiang Ren, Leonard Bielory, Zhongyuan Mi, Christopher G. Nolte, Yang Gao, L. Ruby Leung, and Panos G. Georgopoulos. Development of a semi-mechanistic allergenic pollen emission model. *Science of The Total Environment*, 653:947957, 2019.

Ting Cai, Yong Zhang, Xiang Ren, Zhongyuan Mi, Leonard Bielory, Christopher G. Nolte, Shan He, and Panos G. Georgopoulos. Modeling spatiotemporal distributions of airborne alergenic pollen in the CMAQ-Pollen modeling system. *Science of The Total Environment* (In preparation).

2.1 Abstract

This chapter studies the impacts of climate change on the concentrations of allergenic pollen in the 2050s based on the RCP 8.5 scenario. A modeling system incorporating pollen emission and transport has been used to simulate the 2004 and 2047 spatiotemporal distributions of allergenic pollen of oak and ragweed, which are two of the top allergens in the North America. It is found that oak pollen emissions start from the Southern part of the CONUS in March and then shift gradually toward the Northern CONUS. On the other hand, ragweed pollen emissions start from the Northern CONUS in August and then shift gradually toward the Southern CONUS.

Process analysis revealed that dry deposition, emission and vertical eddy diffusion are the dominant processes determining the ambient pollen concentrations. The impact of climate change on oak pollen season varies in the nine climate regions across the CONUS. The mean and maximum hourly concentrations of oak pollen were predicted to increase in the Northeast, South and Southeast regions, but to decrease in the Northwest, East North Central, and West North Central regions. The oak pollen season was estimated to start earlier in the Central, Northeast, South and Southeast regions. The oak pollen season length was estimated to shorten by 1-2 days for most regions, except the Southeast and Southwest regions. It was estimated that ragweed pollen seasons will start earlier and last longer for all the nine climate regions. The mean and maximum hourly concentrations of ragweed pollen were predicted to increase significantly in the Northwest, Southeast, Southwest and West regions. The number of hours in which the ragweed pollen concentrations exceed the threshold value (30 pollen grains/m^3) were estimated to increase by 1.2%-34.3% in most regions, except the Central, East North Central and the West North Central regions.

2.2 Introduction

Airborne allergenic pollen from trees, weeds and grass is one of the main triggers of AAD affecting 5% to 30% of the population in industrialized countries [132, 133, 134, 80]. It has been estimated that pollen-related asthma emergency department visits across the CONUS will increase by 14% in 2090 under a high greenhouse gas emission scenario [135]. Climate change is critically affecting emissions of natural pollutants such as pollen and spores as well as biogenic gases which are components of atmospheric photochemistry reaction systems. The rising temperature and changes in precipitation will also increase the levels of aeroallergens including pollen [8]. Studies on impacts of climate change on emission and transport of allergenic pollen from multiple taxa are needed to estimate the potential consequences for public health.

Duhl *et al.* [136] studied the impact of climate change on pollen season under the IPCC SRES A1B scenario as simulated by the fifth-generation atmospheric general circulation model (ECHAM5). Meteorological inputs of current (1995–2004) and future (2045–2054) years downscaled using the WRF model were used to drive the Simulator of the Timing and Magnitude of Pollen Season (STaMPS) model to estimate the relative magnitude and timing of pollen season for six tree genera (Betula, Juglans, Morus, Olea, Platanus and Quercus) and one grass genus (Bromus) in California and Nevada. It was found that pollen season will start an average of 5-6 days earlier under predicted future climatic conditions, while the changes of pollen production varied by species. Zhang [3] used a WRF-SMOKE-CMAQ-Pollen modeling system to simulate the changes of pollen season and pollen concentration under the IPCC A2 scenario. It was estimated that the regional average pollen concentrations will decrease in the majority of climate regions during the period of 2047-2050 for ragweed, mugwort and grass. But the population will potentially have increased number of AAD attacks from oak and birch pollen in most area. The pollen season of oak was found to start earlier in recent years in Spain [73], and will continue to increase under the meteorological data forecast using the Regional Climate Model (RCM).

In RCP 8.5 scenario, future anthropogenic greenhouse gas emissions are assumed to rise continually throughout the 21st century. The goal of this study was to investigates impacts of climate change on the concentrations of allergenic pollen in the 2050s across the CONUS based on the RCP 8.5 scenario. Two of the top allergens in the North America, oak and ragweed were selected. The changes of pollen indices (Start Date, Season Length, etc.) in the nine climate region of CONUS were examined. The simulation results are evaluated using the observed pollen count from the monitor stations of the National Allergy Bureau (NAB) of the American Academy of Allergy, Asthma, and Immunology (AAAAI) across the CONUS. Process analyses are conducted to investigate the contribution of each physical process on airborne pollen concentrations.

2.3 Methods

2.3.1 Model configuration

The configurations of each component model are listed in Table 2.1. The meteorology inputs are derived from a climatological simulation of the year 2004 and 2047 by the Community Earth System Model (CESM), which was downscaled using the WRF model [65, 137]. The future meteorology in 2047 was simulated based on the RCP 8.5 scenario, which is the pathway with the highest greenhouse gas emissions [138]. The pollen emission model was developed by Zhang [3] and Cai *et al.* [97], and the details of the emission model development are provided in Appendix B. The pollen transport model CMAQ-Pollen was developed by Zhang [3] and Cai *et al.* [139]. It was adapted from the existing CMAQ modeling system (v4.7.1) [140]. Pollen grains were treated as inert coarse mode aerosol. Physical properties such as density, diameter and diameter distributions and other related information (e.g. cutoff maximum aerosol diameter) of coarse mode were adapted in relevant CMAQ modules (AERO5), so that the adapted CMAQ-Pollen model could handle the simulation of spatial and temporal distributions of pollen. The CMAQ-Pollen model was run for 2004 and 2047 covering the CONUS with 36 km horizontal grid spacing, temporal resolution of one hour, and 34 layers in the vertical direction. The simulation results for 2004 were evaluated using the observed pollen counts from the monitor stations of the NAB of the AAAAI across the CONUS.

Table 2.1: Configuration of the meteorology, emission and transport model for studying distributions of airborne allergens.

	Model	**Resolution, Layers**	**Period**	**Domain**	**Reference**
Meteorology	WRF v3.4.1	36x36 km, hourly 34	2004, 2047	CONUS	[65, 137]
Pollen emission	Semi-mechanistic model	36x36 km, hourly 1	2004, 2047	CONUS	[3, 97]
Pollen transport	Adapted CMAQ v4.7.1	36x36 km, hourly 34	2004, 2047	CONUS	[3, 139]

2.3.2 Initial and boundary conditions

For oak, the simulation for the CONUS domain was run from 00:00 of March 1st through 23:00 of April 30th of 2004 and 2047. For ragweed, the simulation was run from 00:00 of August 1st through 23:00 of September 30th of 2004 and 2047. March 1st and August 1st generally precede the earliest flowering day of oak and ragweed across the CONUS, respectively. Therefore, the simulations were initialized with no existing pollen.

For simulation on the CONUS domain, Boundary Conditions (BC) of pollen were set to zero with the eastern and western boundaries of simulation domain bordering the Atlantic and Pacific oceans, and northern and southern boundaries adjoining Canada and Mexico. To investigate the influence of BC on airborne pollen concentrations, one additional simulation on the CONUS domain was run for oak pollen between March 1st and April 30th, 2004 by prescribing the BC values for all four lateral boundaries as 10 pollen grains/m^3 at each time step of the CMAQ-Pollen model. The difference of simulated airborne pollen concentrations in the lowest layer between the simulations with the two BCs was calculated to examine the impact of BCs on airborne pollen concentrations.

2.3.3 Process analysis of pollen transport model

The physical processes governing the transport and removal of pollen grains from air include cloud processes, dry deposition, horizontal and vertical advection, and horizontal and vertical diffusion. Dry deposition process is treated in the vertical diffusion process as a flux boundary condition at the bottom of the model layer. It includes the effect of gravitational settling. Wet deposition is simulated in cloud processes, which include both in cloud and below cloud scavenging. Wet deposition depends on the precipitation rate and concentration in cloud water. Effects of convection on pollen transport are treated separately through modules of horizontal and vertical

advection.

The Process Analysis Preprocessor (PROCAN) was compiled together with an adapted CMAQ model to activate the process analysis function in the CMAQ-Pollen modeling system [140]. The process analysis was conducted to identify the contributions of each physical process on airborne pollen concentrations. The physical processes incorporated into the process analysis included cloud process, dry deposition, emission, horizontal and vertical advection, and horizontal and vertical diffusion. Process analysis was carried out using the time series of simulated hourly concentrations of allergenic oak pollen during the pollen season in 2004 in the grid cell that contains the pollen monitoring station at the Atlanta Allergy and Asthma Clinic (coordinates: 33.97°N, 84.55°W). This area has an elevation of 366 m, annual mean temperature of 16.8 °C, and annual mean precipitation of 1,286 mm.

2.3.4 Evaluation of model performance

Due to the fact that the meteorology data used in the study are downscaled from a global climate model without assimilating weather observations, the day-to-day weather variability cannot be represented [65]. Therefore, it is hard to compare the simulation results with the pollen observation on a daily basis. Instead, the evaluation is focused on seasonal and monthly temporal scales. Correlation analysis of the observed seasonal mean pollen concentrations at pollen monitoring stations with the corresponding simulated seasonal mean pollen concentrations was conducted with normalized pollen data (mean zero and unit standard deviation). The observed pollen concentrations at a monitor station are paired with the simulated pollen concentrations in a grid cell that contains the corresponding pollen monitoring station. The simulated pollen concentrations are derived from the simulated hourly concentrations in the model's lowest layer (i.e., layer 1) because observations of pollen counts are generally made near the surface. The model's lowest layer on average extends from 0 to 60 m above the ground. Fractional bias (FB) of simulated seasonal pollen counts

(sum of daily pollen concentration) are reported:

$$FB_i = 2\frac{SC_{Sim,i} - SC_{Obs,i}}{SC_{Sim,i} + SC_{Obs,i}} \tag{2.1}$$

where FB_i is the fractional bias of simulated seasonal pollen count at station i, $SC_{Sim,i}$ is the simulated seasonal pollen count of station i, $SC_{Obs,i}$ is the observed seasonal pollen count of station i. Hit and false rates are common indexes to evaluate the simulated daily pollen concentration. Procedures from the literature are followed to calculate the hit and false rates at three different concentration levels [84, 78], which are 10, 50 and 100 pollen grains/m^3, respectively. The details of the calculations are presented in Appendix B.

2.3.5 Uncertainty analysis

Uncertainties generally pervade the entire modeling process [141]. They may result from different components and modules of the modeling framework [142]. In the current study, we made qualitative judgment on the relevance of uncertainty sources for each modeling component, and the general procedures to diagnose and reduce uncertainty.

2.3.6 Impacts of climate change on spatiotemporal distribution of allergenic pollen

To examine the impact of climate change on pollen season and pollen levels across the CONUS, five metrics were evaluated: the mean hourly concentration, maximum hourly concentration, Start Date (SD) of pollen season, pollen Season Length (SL), and the number of hours exceeding the pollen threshold concentrations during pollen season. Figure 2.1 shows the procedures for calculating the pollen indices for the nine climate regions (Figure 2.2). The climate regions are classified according to the long term observed temperature and precipitation based on the database of the National

Climatic Data Center of the National Oceanic and Atmospheric Administration [143]. The threshold concentrations for calculating the number of exceedance hours are 13 pollen grains/m^3 for oak and 30 pollen grains/m^3 for ragweed. These threshold values were chosen based on the clinical symptoms of allergic disease in sensitive patients [80, 144, 145].

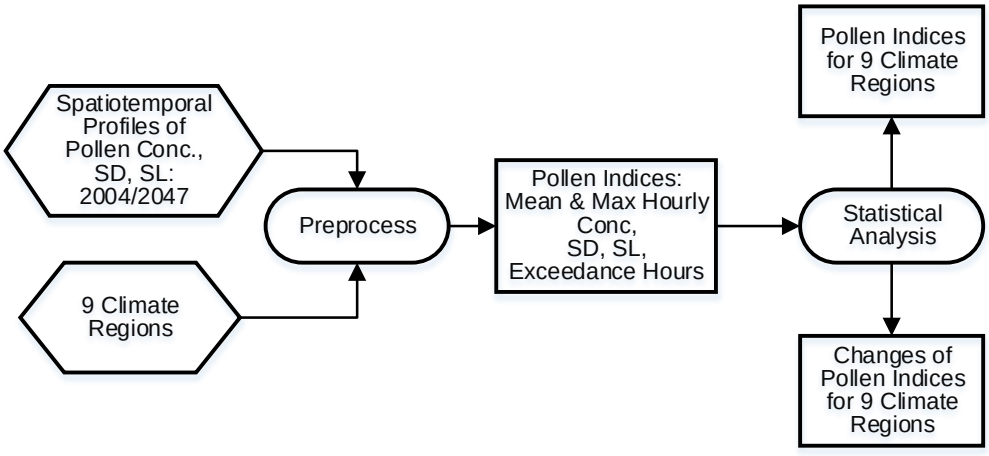


Figure 2.1: Calculation of pollen indices to assess climate change impacts on allergenic pollen.

Only the simulated hourly pollen concentrations in the first layer of the modeling grid (up to 60 meters above the ground) were analyzed because people are mainly exposed to allergenic pollen is this layer. The mean hourly pollen concentrations ($C_{hr,Mn}(i,j)$), the maximum hourly pollen concentrations ($C_{hr,Mx}(i,j)$), and the number of exceedance hours ($N_{Exd}(i,j)$) in each cell of the modeling grid were calculated according to Equation 2.2,

$$\begin{cases} C_{hr,Mn}(i,j) = \frac{\sum_{hr} C(hr,i,j)}{N_{hr}} \\ C_{hr,Mx}(i,j) = \max_{hr} C(hr,i,j) \\ N_{Exd}(i,j) = \sum_{hr} \mathbb{1}_{C(hr,i,j) \geq C_{Thr}} \end{cases} \tag{2.2}$$

where N_{hr} is the number of simulation hours in each grid. $\mathbb{1}$ is the indicator function; it takes 1 as its value when the hourly concentration $C(hr,i,j)$ is greater or equal to

the threshold concentration C_{Thr}, otherwise takes 0 as its value. The pollen indices for 2004 and 2047 were calculated, and then the changes of each index between 2047 and 2004 were calculated using Equation 2.3,

$$\begin{cases} \Delta C_{hr,Mn}(i,j)/C^{2004}_{hr,Mn}(i,j) = \frac{C^{2047}_{hr,Mn}(i,j) - C^{2004}_{hr,Mn}(i,j)}{C^{2004}_{hr,Mn}(i,j)} \\ \Delta C_{hr,Mx}(i,j)/C^{2004}_{hr,Mx}(i,j) = \frac{C^{2047}_{hr,Mx}(i,j) - C^{2004}_{hr,Mx}(i,j)}{C^{2004}_{hr,Mx}(i,j)} \\ \Delta SD(i,j) = SD^{2047}(i,j) - SD^{2004}(i,j) \\ \Delta SL(i,j) = SL^{2047}(i,j) - SL^{2004}(i,j) \\ \Delta N_{Exd}(i,j)/N^{2004}_{Exd}(i,j) = \frac{N^{2047}_{Exd}(i,j) - N^{2004}_{Exd}(i,j)}{N^{2004}_{Exd}(i,j)} \end{cases} \tag{2.3}$$

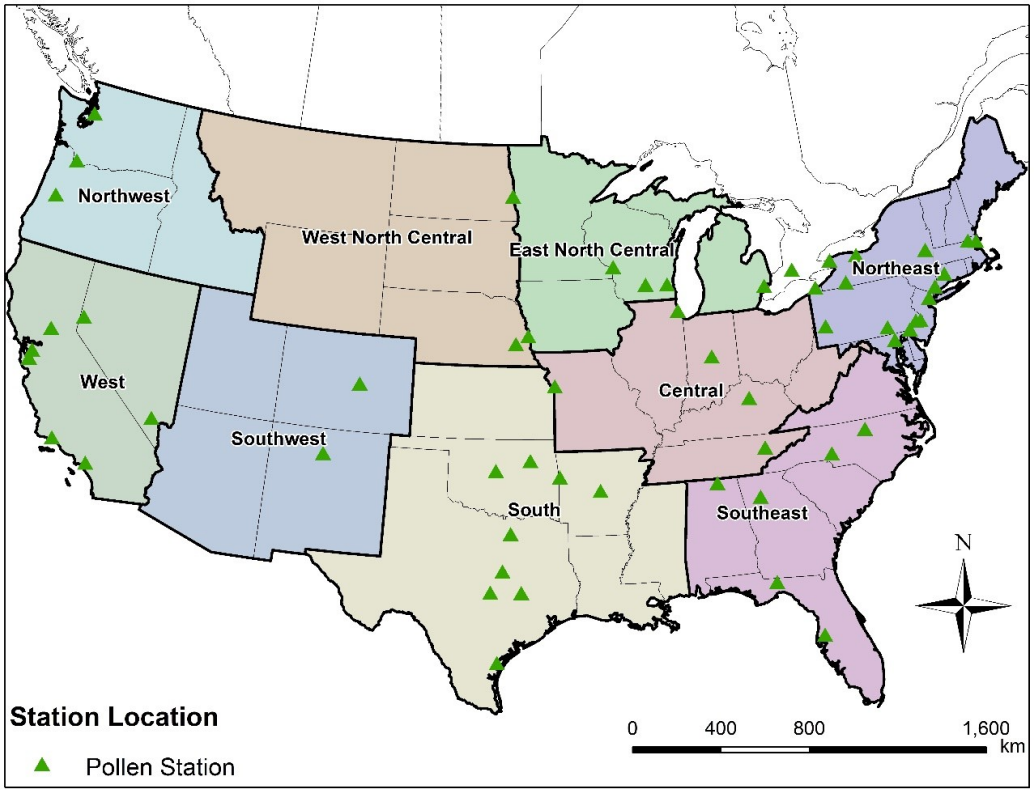


Figure 2.2: Distribution of the 58 studied pollen stations across the nine climate regions in the contiguous US.

The mean and standard deviation of the change of a pollen index(e.g, *SD*) in a climate region were the average value of the pollen index in the grids of that region. As

an example, Equation 2.4 shows how to calculate the mean and standard deviations of changes in SD in climate region k:

$$\begin{cases} \overline{\Delta SD}_k = \frac{\sum_{(i,j)\in \text{Region}k} \Delta SD(i,j)}{N_k} \\ \Delta SD_{k,Std} = \frac{\sum_{(i,j)\in \text{Region}k} (\Delta SD(i,j) - \overline{\Delta SD}_k)^2}{N_k} \end{cases} \tag{2.4}$$

where N_k is the number of grid cells in climate region k. The mean and standard deviation of other indices in other climate regions were calculated similarly.

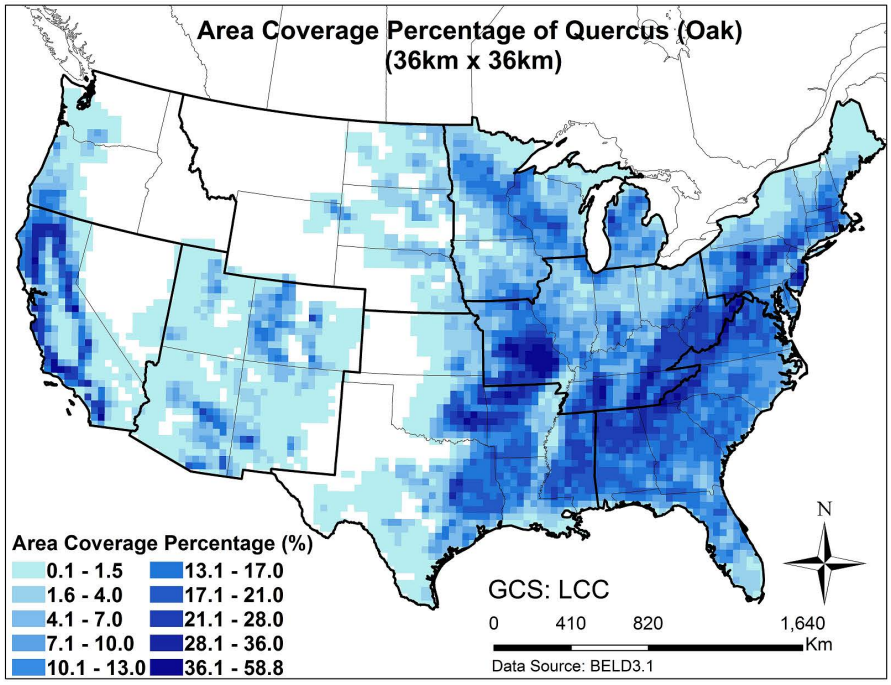
2.4 Results and Discussion

2.4.1 Vegetation coverage

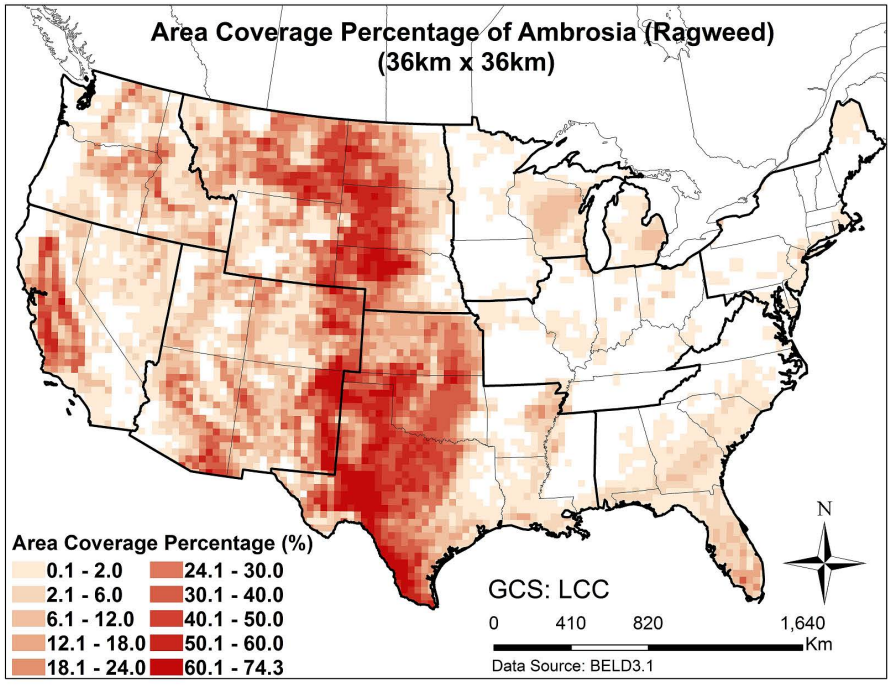
Figure 2.3 presents the percentage of the area occupied by oak and ragweed in each cell of the modeling grid covering the CONUS. The distributions of oak and ragweed are kept as constant for 2004 and 2047. Oak trees are distributed mostly across eight of the nine climate regions of the CONUS with the highest area coverages (36%-58.8%) in the West, South, Central and Southeast climate regions. Ragweed is mainly distributed in the western US, with the highest area coverages (60.1%-74.3%) in the South and the West North Central climate regions. The classification of the nine climate regions across the CONUS is illustrated in Figure 2.2. These vegetation coverage maps are important inputs to the pollen emission model to calculate the pollen emission fluxes in each cell of the modeling grid covering the CONUS.

2.4.2 Spatiotemporal distribution of airborne pollen concentration

To examine the temporal distribution patterns of the simulated airborne pollen, the monthly mean oak/ragweed pollen concentrations at ground level during their early and late flowering season are plotted in Figure 2.4. The overall patterns are consistent with their emission patterns shown in B.1. Oak pollen only appeared in the Southern



(a)



(b)

Figure 2.3: Area coverage of: (a) oak and (b) ragweed with 36-km horizontal grid spacing over the CONUS.

CONUS in March, then occurred in the Northern CONUS in April. The maximum mean oak pollen concentration is 4,500 pollen grains/m^3. Ragweed pollen appeared

first in the Northern CONUS in August and then shifted toward the Southern CONUS in September. The distribution pattern also follows its emission pattern. The mean ragweed pollen can reach up to $2x10^4$ pollen grains/m^3 during its peak season. Figure 2.5 displays simulated average oak and ragweed pollen concentrations for different hours of the day. The oak pollen concentration in each cell of the modeling grid at 11:00 UTC is higher than that at 18:00 UTC (averaged over April 21-April 30, 2004), and the ragweed pollen concentration in each cell of the modeling grid at 14:00 UTC is higher than that at 18:00 UTC (averaged over September 21-September 30, 2004).

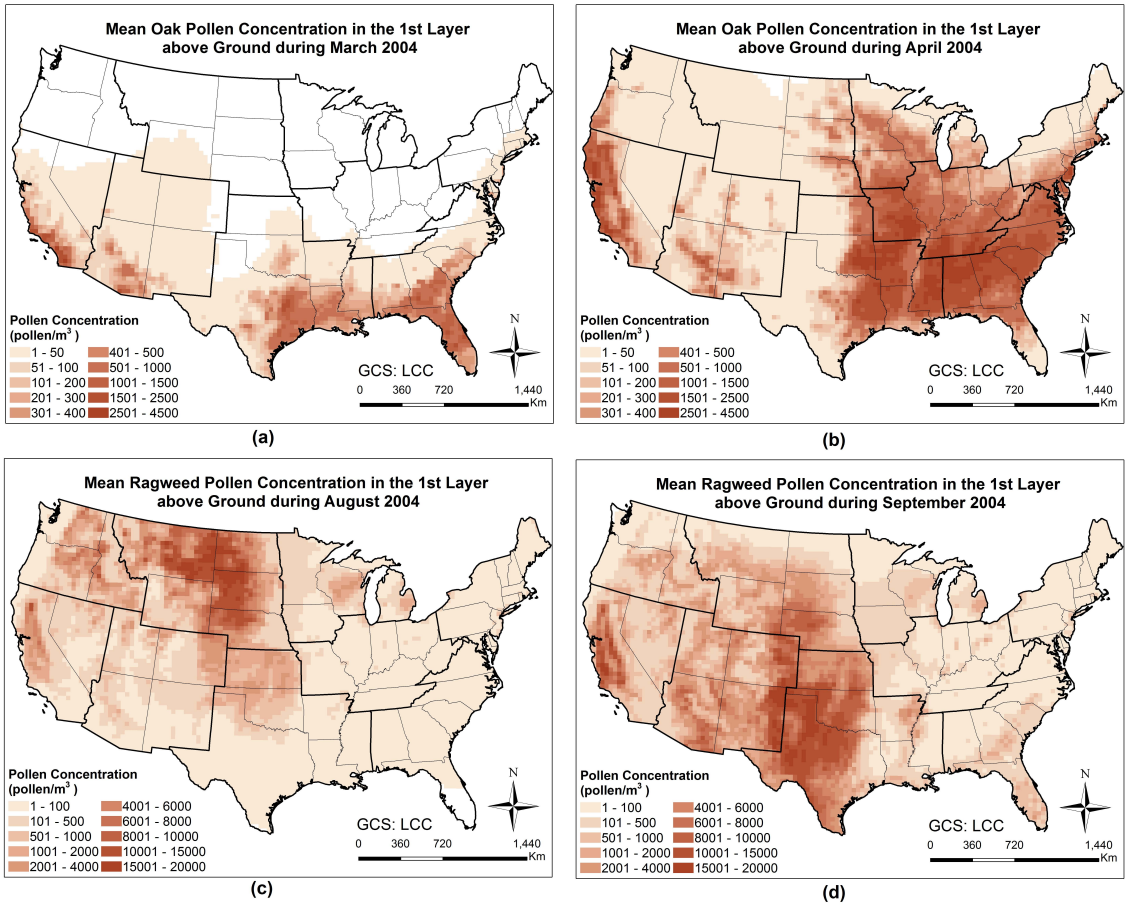


Figure 2.4: Spatial patterns of mean concentration of (a) oak pollen in March 2004; (b) oak pollen in April 2004; (c) ragweed pollen in August 2004; (d) ragweed pollen in September 2004.

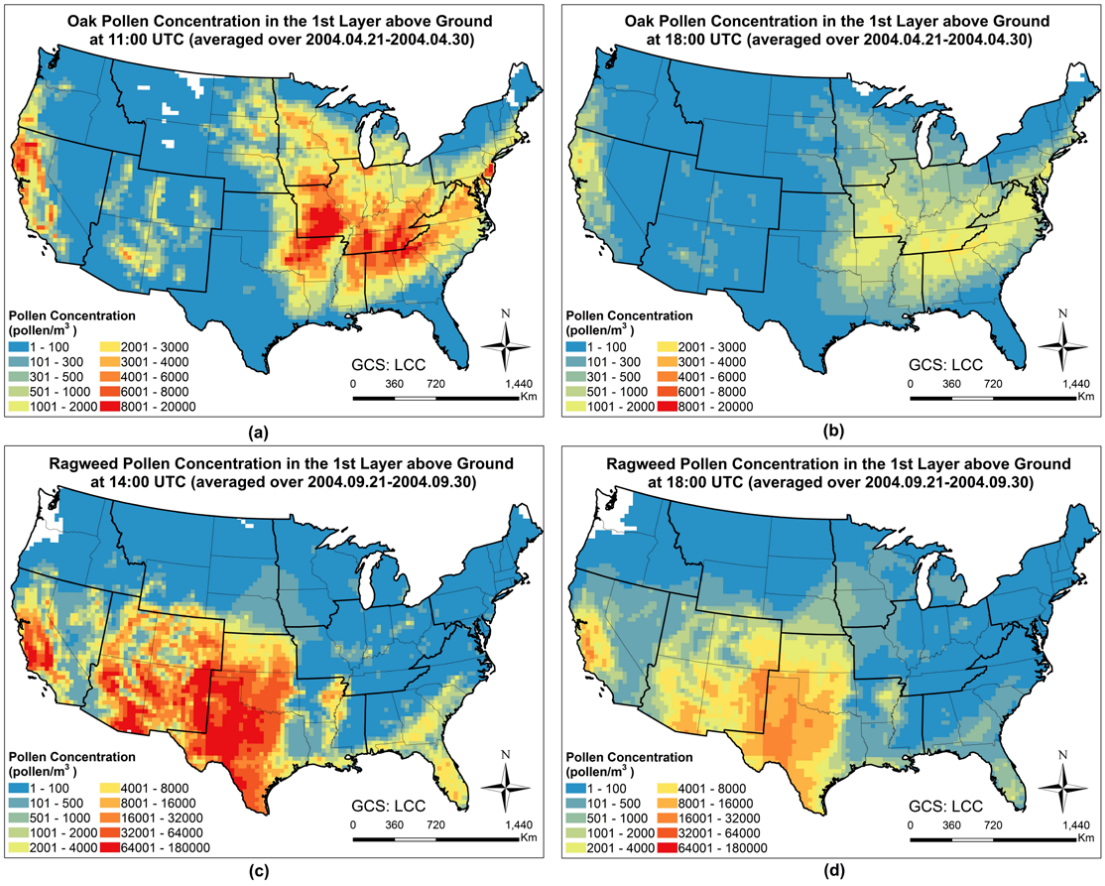


Figure 2.5: Time slices of spatiotemporal concentration profiles of (a) oak pollen at 11:00 UTC (averaged over April 21-April 30, 2004); (b) oak pollen at 18:00 UTC (averaged over April 21-April 30, 2004); (c) ragweed pollen at 14:00 UTC (averaged over September 21-September 30, 2004).

2.4.3 Evaluation of model performance

As shown in Figure 2.6, there is statistically significant correlation between observed seasonal mean concentrations with the corresponding simulated seasonal mean concentrations for both oak and ragweed pollen. The Pearson correlation coefficient is 0.345 (p-value $0.0252 < 0.05$) for oak pollen based on data from 42 monitoring stations, and 0.399 (p-value $0.0055 < 0.05$) for ragweed pollen based on data from 47 monitoring stations. The data points for oak pollen are evenly distributed around the 45-degree line. Three ragweed monitoring stations have larger deviations from other stations that our model was not able to capture. The statistical distributions

of the daily simulated pollen concentrations at each pollen monitoring station during pollen season compared with corresponding observation data are shown in Figure 2.8. For each pollen monitoring station, similar distribution between simulation and observation data indicate good model performance. Our model was able to capture the distribution of observed pollen concentration for most of the stations. For oak pollen, the model was also able to simulate the extreme data points at the stations with high concentration outliers.

Figure 2.7 shows the fractional bias of simulated seasonal pollen count. The fractional bias for seasonal oak pollen count was mostly greater than 0, indicating overestimation of the pollen concentration. The fractional bias for ragweed pollen concentration was also greater than 0 for most stations, suggesting overestimation in the model performance. However, the model was able to capture the variation of the pollen observation as shown in the correlation analysis in Figure 2.6. Hit and false rates are metrics to check whether the simulated and observed exceedances are consistent and co-located. Figure B.8 and Figure B.9 present the hit rates and false rates for predicted and observed daily oak and ragweed pollen concentrations for three pollen levels at the studied stations during 2004 across the CONUS. The hit rates for airborne oak and ragweed pollen levels of 10, 50 and 100 pollen grains/m^3 were all between 70% and 100% for most of the studied stations. This indicates that the observed exceedances of three thresholds were mostly correctly predicted by the modeling system of pollen emission and transport. The false rates for airborne oak pollen level of 10 pollen grains/m^3 were between 0 and 30% for most of the studied stations, but the false rate increased at levels of 50 and 100 pollen grains/m^3. The false rates were over 30% for most stations, which indicates that the CMAQ-Pollen model overestimated the ragweed pollen concentration.

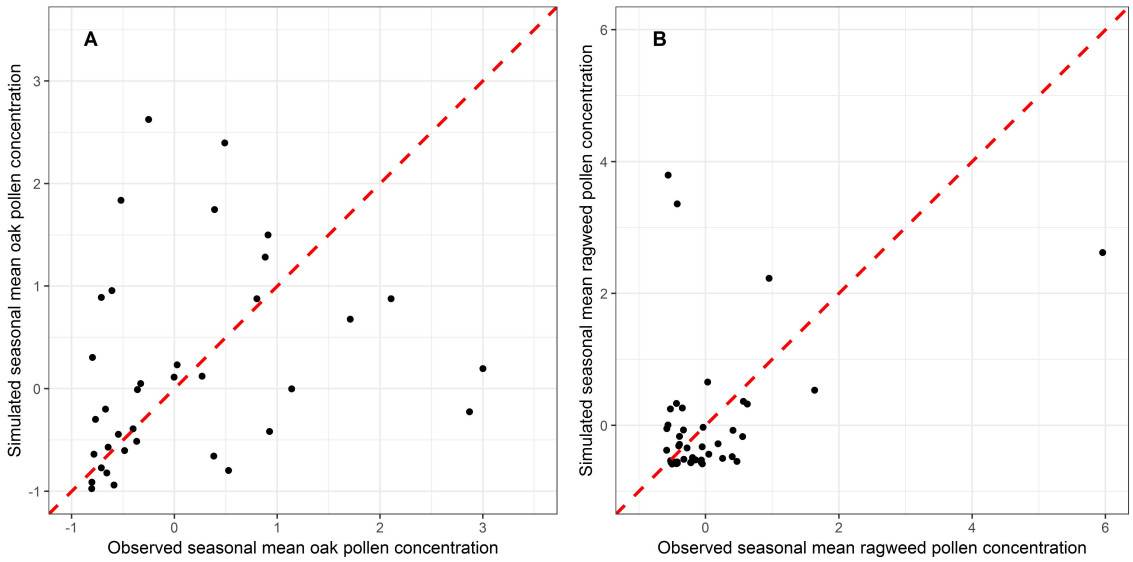


Figure 2.6: Scatterplots of normalized observed seasonal mean concentrations and simulated seasonal mean concentrations in 2004 for oak and ragweed pollen at selected pollen monitoring stations with 45-degree line.

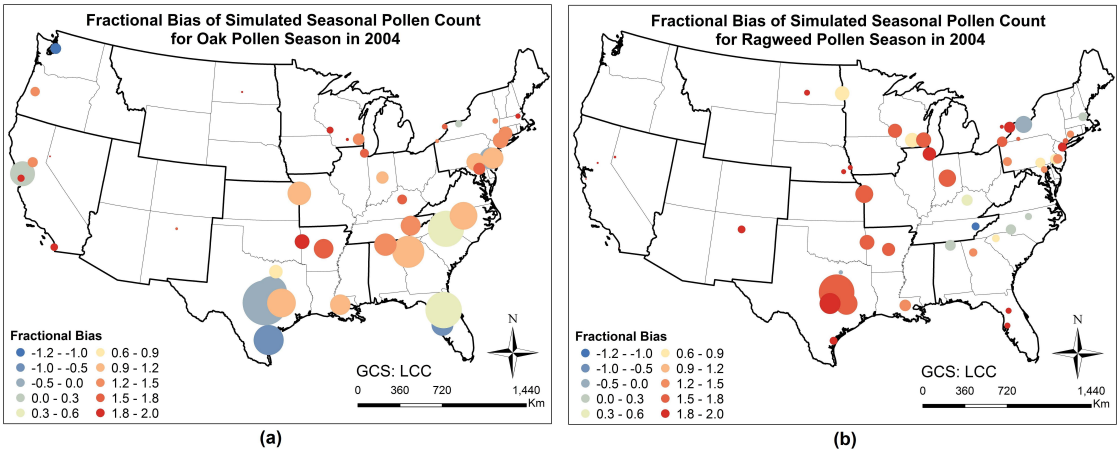


Figure 2.7: Fractional biases of predicted pollen concentration during 2004 across the CONUS. (a) Fractional bias of seasonal oak pollen counts; (b) Fractional bias of seasonal ragweed pollen counts.

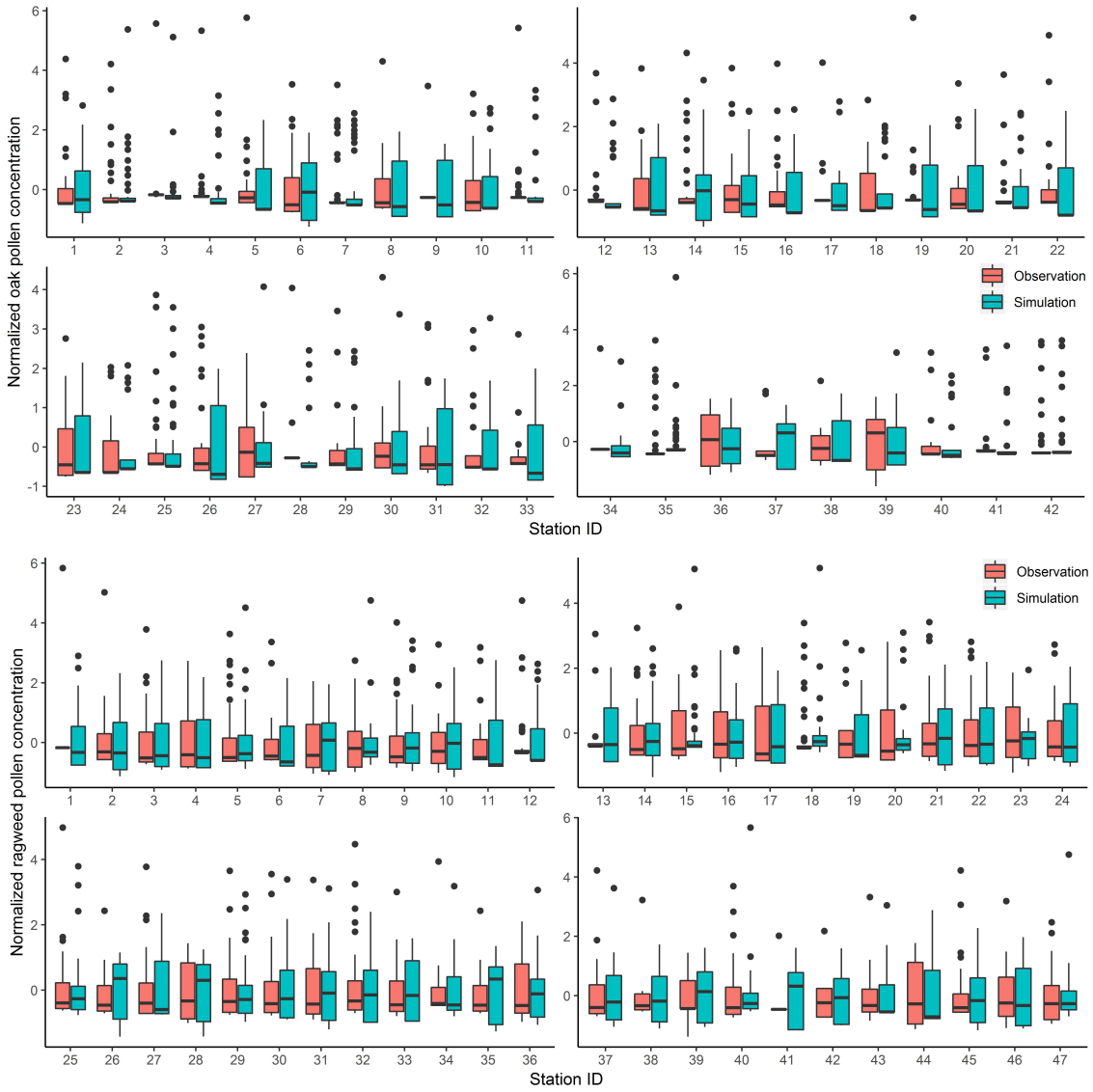


Figure 2.8: Seasonal box plots of normalized simulated daily concentrations of oak pollen (top) and ragweed pollen (bottom) compared against observed pollen concentrations in 2004 at pollen monitoring stations. Boxes range from the 25th to 75th percentiles with the dark line denoting the median and the dark dots denoting the outliers.

2.4.4 Process analysis

Figure 2.9 presents the contributions of advection, diffusion, dry deposition, emission and cloud processes on the hourly oak pollen concentrations between local time 00:00 UTC April 1 to 23:00 UTC April 10, 2004 in Atlanta, Georgia. Dry deposition, emission and vertical eddy diffusion were the dominant processes determining ambient concentrations of oak pollen. The emission process continuously released pollen grains into the air following a regular diurnal pattern. The two emission and concentration peaks at around local time EST 5 AM and 5 PM reflect the diurnal profiles of oak pollen emission due to different meteorological conditions in early morning and late afternoon. The majority of ambient pollen grains were removed from the air through dry deposition. Generally, the pollen concentration near the surface represents a balance between emission and dry deposition. However, vertical diffusion may dominate the transport of ambient pollen grains when there is strong turbulent atmospheric movement. For example, pollen grains might be lifted up by turbulence at neighboring locations (e.g., during a frontal passage) and subsequently transported horizontally to Atlanta where vertical diffusion could bring them down to the lowest layer under special weather conditions, and increase the pollen concentrations in the lowest layer. The cloud process also played an important role through in-cloud and below-cloud scavenging during rainy time (red line between April 6 and April 8 in Figure 2.9). Hence, depending on the meteorological conditions and emissions, both varying diurnally and seasonally, different processes can play important roles in determining the pollen concentrations near the surface, and influence the model prediction skill.

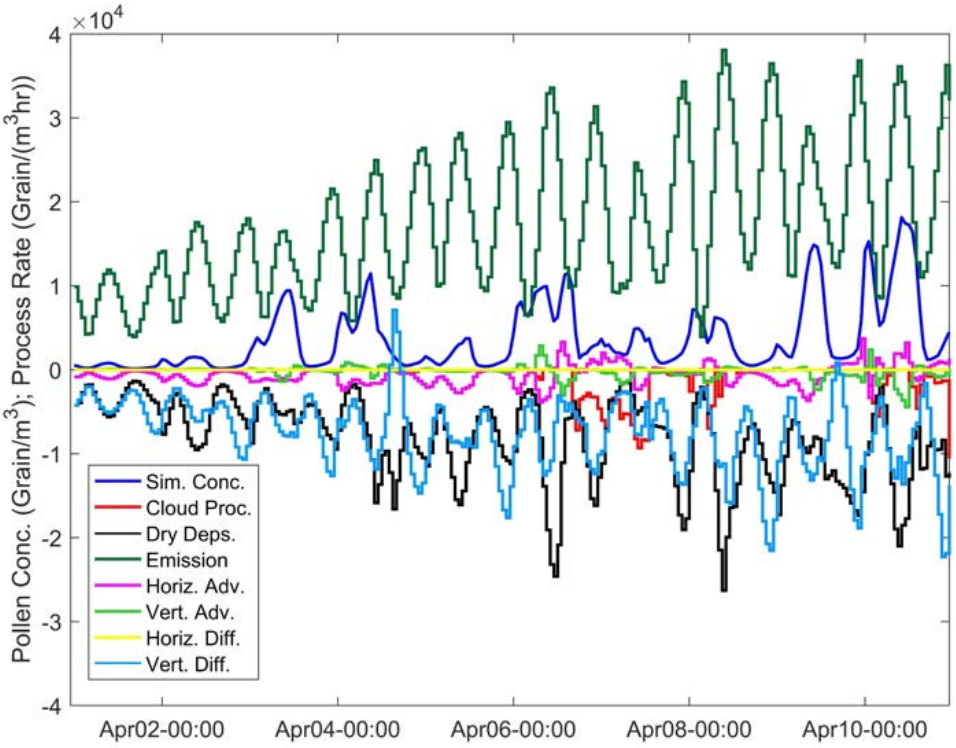


Figure 2.9: Contributions of advection, diffusion, dry deposition and cloud process on the hourly oak pollen concentrations between 00:00 UTC April 1 to 23:00 UTC April 10, 2004 in Atlanta, Georgia. (Sim. Conc.: Simulated Concentration, Cloud Proc.: Cloud Processes, Dry Deps.: Dry Deposition, Horiz. Adv.: Horizontal Advection, Vert. Adv.: Vertical Advection, Horiz. Diff.: Horizontal Diffusion, Vert. Diff.: Vertical Diffusion)

2.4.5 Influence of boundary conditions

Figure 2.10 displays the difference in ambient oak pollen concentration due to different boundary conditions. In the majority of the grid cells, the mean hourly pollen concentrations seem not to be remarkably influenced by the boundary conditions. In some area of the west coast, such as California, Nevada, Arizona, Utah, Oregon and Washington, the mean hourly concentrations increased by 1-2 pollen grains/m^3 because of the changes in boundary conditions. These areas appear to have relatively low oak area coverage and low emissions, so transport of pollen from the boundaries may be a dominant source of oak pollen in those regions. In the current study, the

simulated pollen concentrations were only mapped within the CONUS boundaries, which are different from the model boundaries. An early study in some regions of California has reported that dynamic BCs had barely improved the model performances, and that perturbations in emissions significantly influenced the simulated pollen concentrations [87]. Further investigations are needed to identify the impact of BC on simulated pollen concentrations.

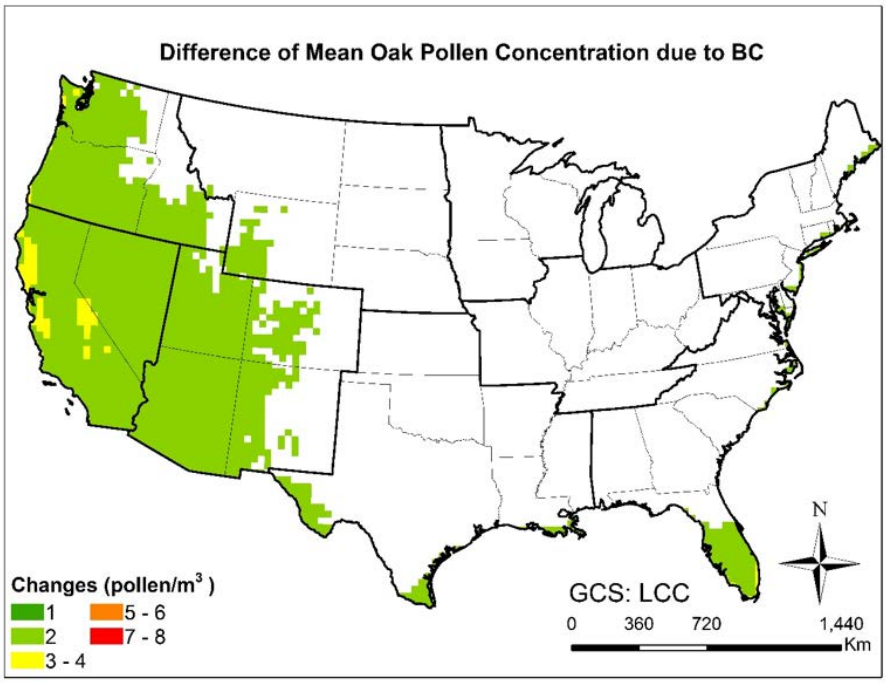


Figure 2.10: The difference in mean hourly concentrations of oak pollen between two different boundary conditions (BC). The default BC was set as 0 pollen grains/m^3, and the other BC was set as 10 pollen grains/m^3.

2.4.6 Impact of climate change on allergenic pollen

Distributions of allergenic pollen during 2004 and 2047

Figure 2.11 and Figure 2.12 show the spatial distribution of the mean and maximum concentrations of oak and ragweed pollen in 2004 and 2047, which were calculated based on Equation 2.2. The distribution patterns of oak and ragweed pollen both follow the patterns of their area coverage as shown in Figure 2.3.

The mean and maximum concentrations vary across the nine climate regions. For oak pollen, the highest mean and maximum hourly concentrations occurred in the Central, Southeast and South regions. For ragweed pollen, the highest mean and maximum hourly concentrations occurred in the West North Central, South, and Southwest regions. The mean hourly concentrations of oak pollen went up to 2,442 pollen grains/m^3 and the maximum hourly concentrations of oak pollen can reach up to 29,175 pollen grains/m^3. The mean hourly concentrations of ragweed pollen ranged from 1-12,187 pollen grains/m^3, and the maximum hourly concentrations of ragweed pollen varied from 23-3x10^6 pollen grains/m^3.

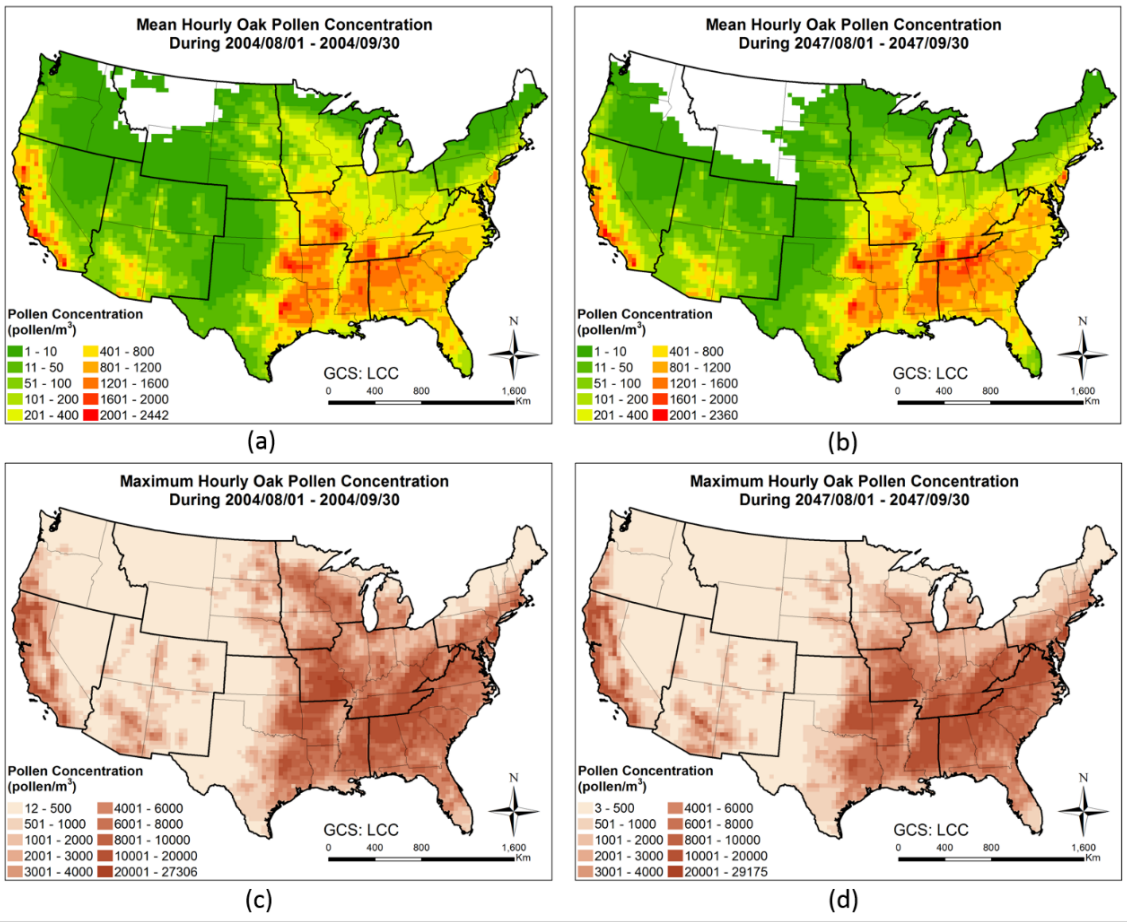


Figure 2.11: Mean (Fig. a and b) and maximum (Fig. c and d) simulated hourly concentrations of oak pollen in 2004 and 2047.

Figure 2.13 and Figure 2.14 present the simulated start date and season length of

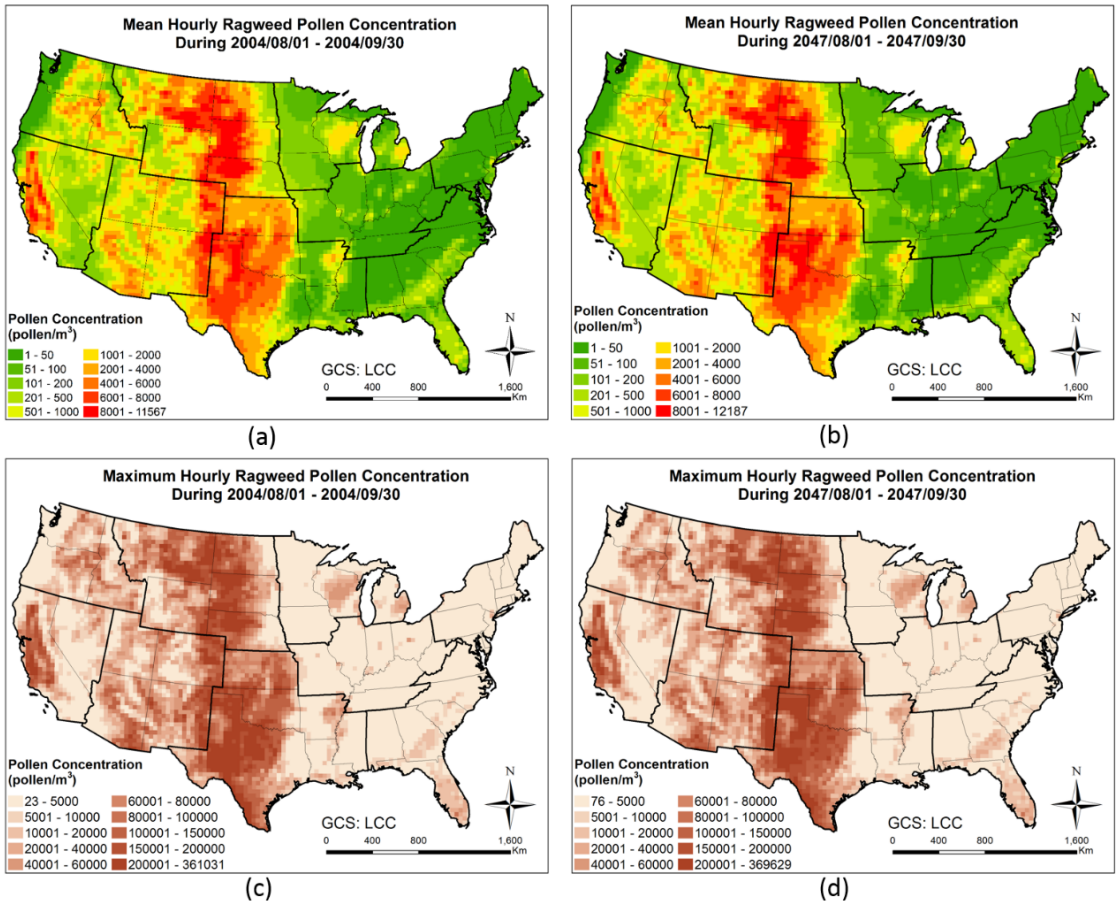


Figure 2.12: Mean (Fig. a and b) and maximum (Fig. c and d) simulated hourly concentrations of ragweed pollen in 2004 and 2047.

oak and ragweed pollen season in 2004 and 2047. The oak pollen season in 2004 and 2047 started around March in the Southern US, around April in the Northern US. While the ragweed pollen started from the Northern US in August, and then shifted toward the Southern US in September. The oak pollen season length ranged from 10 to 46 days, and the ragweed pollen season length varied between 30 to 58 days across the CONUS.

Figure 2.15 and Figure 2.16 display the number of hours in which oak and ragweed pollen concentration exceeded the threshold values (13 pollen grains/m^3 for oak and 30 pollen grains/m^3 for ragweed) in 2004 and 2047 across CONUS. The exceedances were calculated based on Equation 2.2. The oak pollen exceedances ranged from 0 to 1,462 hours, with the highest numbers appeared in the South, Southeast, Southwest

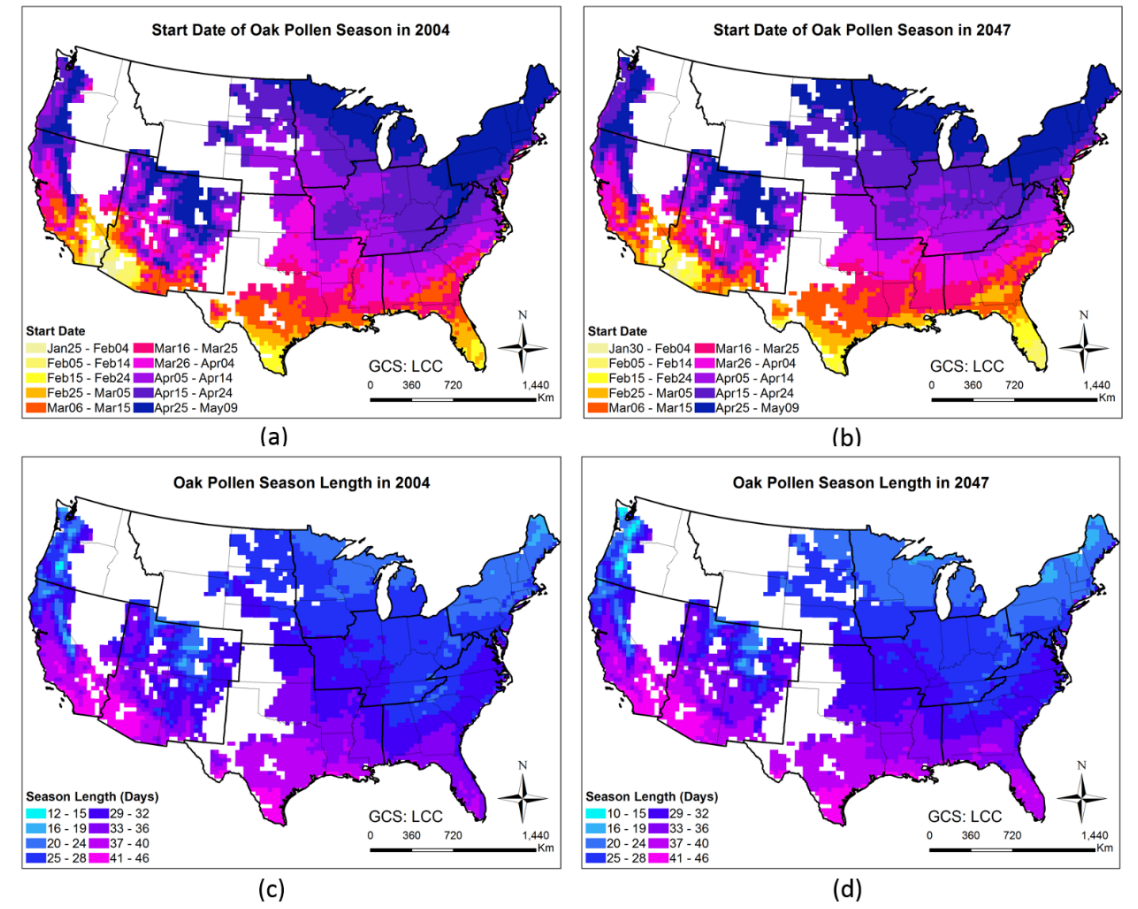


Figure 2.13: Start date (Fig. a and b) and season length (Fig. c and d)) of oak pollen season in 2004 and 2047.

and the West climate regions. The ragweed pollen exceedances ranged from 0 to 1,234 hours, with the highest numbers in the West North Central, the Southwest, West, and the Northwest climate regions.

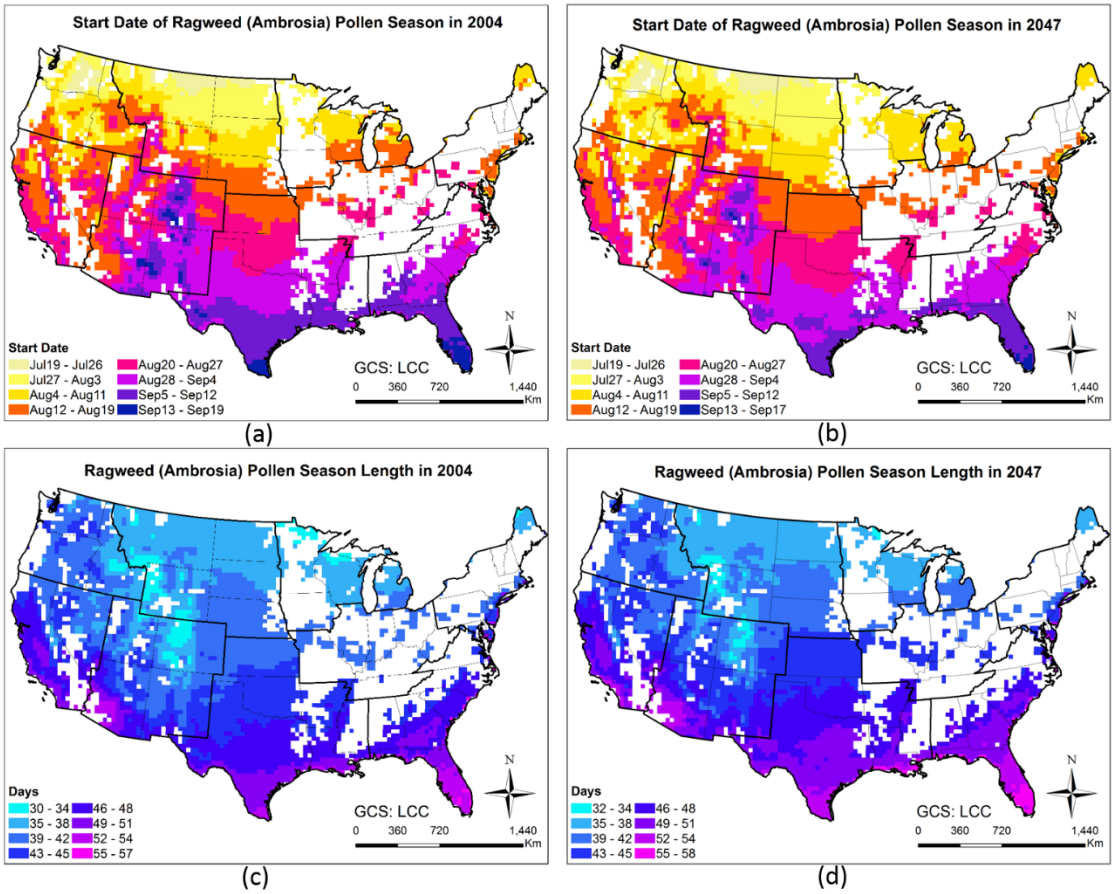


Figure 2.14: Start date (Fig. a and b) and season length (Fig. c and d) of ragweed pollen season in 2004 and 2047.

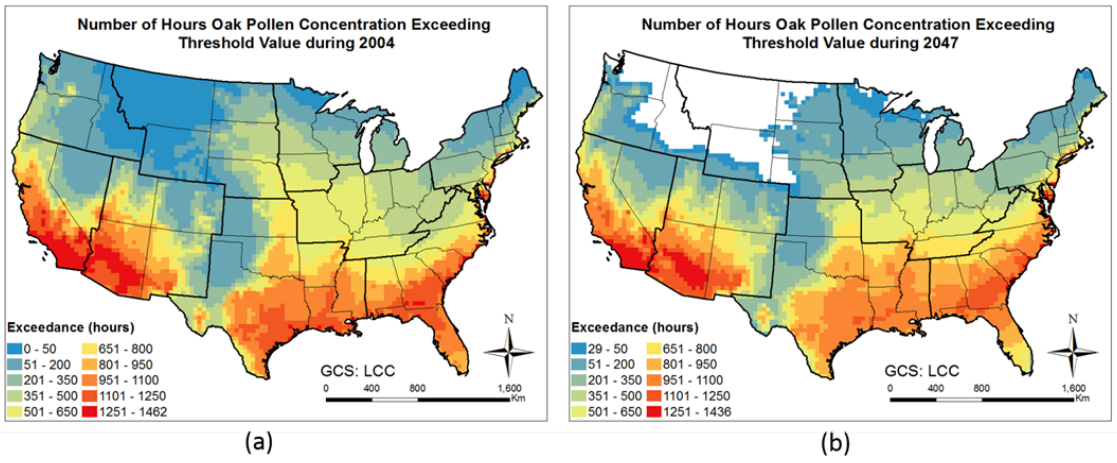


Figure 2.15: Number of hours in which oak pollen concentration exceeds 13 pollen grains/m^3 during 2004 and 2047.

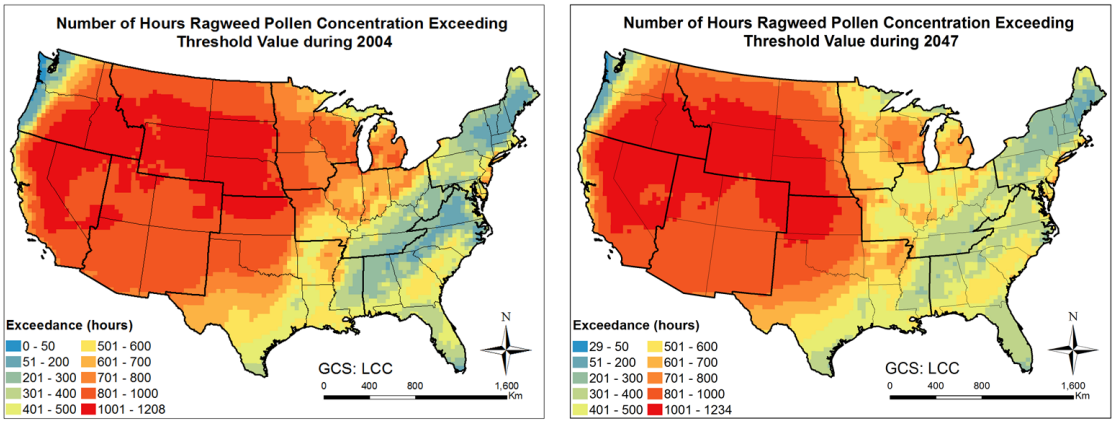


Figure 2.16: Number of hours in which ragweed pollen concentration exceeds 30 pollen grains/m^3 during 2004 and 2047.

Changes of allergenic pollen season between 2004 and 2047

Figure 2.17 presents the changes of the five pollen indices for oak: mean hourly concentrations, maximum hourly concentrations, start date, season length, and exceedance hours, which were calculated based on Equation 2.3. Table 2.2 summarized the regional mean and standard deviation of the changes in each pollen index for oak pollen. As shown in Figure 2.17 and Table 2.2, the impact of climate change on oak pollen season varies in the nine climate regions. The mean and maximum hourly concentrations of oak pollen were predicted to increase in the Northeast, South and Southeast regions, but to decrease in the Northwest, East North Central, and West North Central regions. The Northeast region was estimated to experience the highest increase in mean and maximum hourly concentrations on average for oak pollen. The oak pollen season was estimated to start earlier in the Central, Northeast, South and Southeast regions. The oak pollen season length was estimated to shorten by 1-2 days for most regions, except the Southeast and Southwest regions. The number of hours in which the oak pollen concentrations exceed the threshold value (13 pollen grains/m^3) were estimated to increase most in the Northeast region by 31.6%.

Similarly, Figure 2.18 presents the changes of the five pollen indices for ragweed:

mean hourly concentrations, maximum hourly concentrations, start date, season length, and exceedance hours. Table 2.3 summarized the regional mean and standard deviation of the changes in each pollen index for ragweed pollen. As shown in Figure 2.18 and Table 2.3, the response of ragweed pollen season to future climate in 2047 varies in the nine climate regions. The mean and maximum hourly concentrations of ragweed pollen were predicted to increase significantly in the Northwest, Southeast, Southwest and West regions. The ragweed pollen season was estimated to start 1-3 days earlier in all the climate regions with longer pollen season. The number of hours in which the ragweed pollen concentrations exceed the threshold value (30 pollen grains/m^3) were estimated to increase by 1.2%-34.3% in six regions, while there was a decrease in the Central, East North Central and the West North Central regions.

Table 2.2: Regional average and standard deviation of the changes in mean and maximum hourly concentrations, start date, season length and exceedance hours for oak pollen. (mean ± standard deviation).

Climate Region	Mean Hourly (%)	Max Hourly (%)	Start Date (day)	Season Length (day)	Exceedance Hours (%)
Central	5.0 ± 43.4	-7.8 ± 23.9	-1.6 ± 4.0	-0.4 ± 1.0	2.2 ± 17.3
East North Central	-63.2 ± 24.7	-38.8 ± 73.9	3.1 ± 3.8	-1.5 ± 1.0	-32.7 ± 48.4
Northeast	89.7 ± 117.2	85.0 ± 193.6	-2.2 ± 1.2	-0.6 ± 1.2	31.6 ± 71.7
Northwest	-59.3 ± 39.4	-30.2 ± 54.9	2.9 ± 2.2	-2.1 ± 2.0	-60.7 ± 33.1
South	3.0 ± 28	18.6 ± 39.6	-1.4 ± 3.4	-0.2 ± 1.2	1.1 ± 24.9
Southeast	13.2 ± 37.7	5.2 ± 24.6	-7.7 ± 2.5	1.0 ± 1.4	9.4 ± 21.3
Southwest	5.9 ± 31.3	-0.2 ± 35.7	0.5 ± 3.3	1.4 ± 1.2	7.0 ± 30.1
West	-5.6 ± 31.7	3.0 ± 40.1	4.1 ± 2.8	-0.2 ± 1.1	-0.3 ± 22.9
West North Central	-52.4 ± 41.8	-45.8 ± 32.0	6.8 ± 2.8	-1.4 ± 1.1	-66.6 ± 30.1

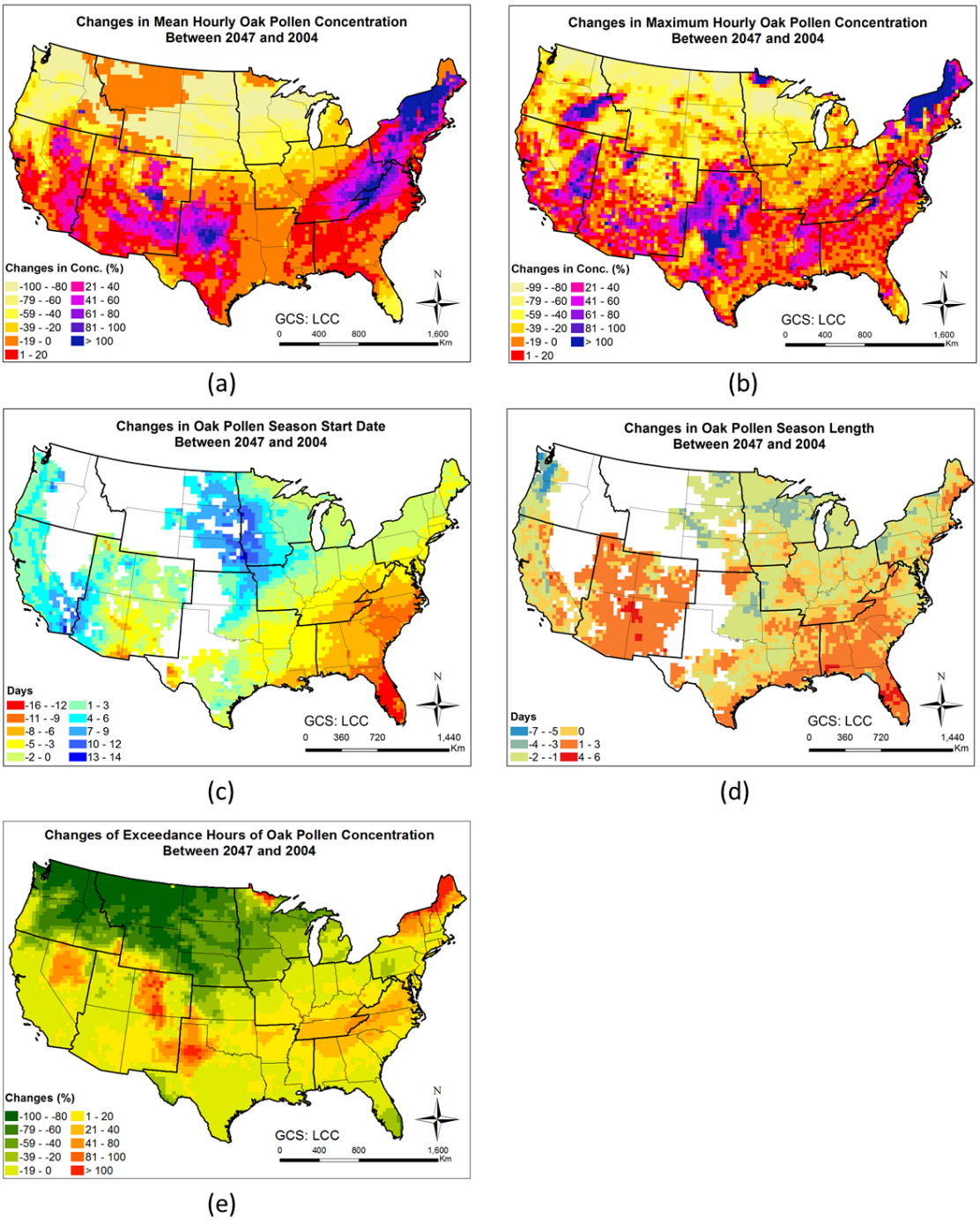


Figure 2.17: Changes in oak pollen season between 2004 and 2047. (a) Mean hourly concentrations, (b) Maximum hourly concentrations, (c) Start date, (d) Season length, and (e) Exceedance hours

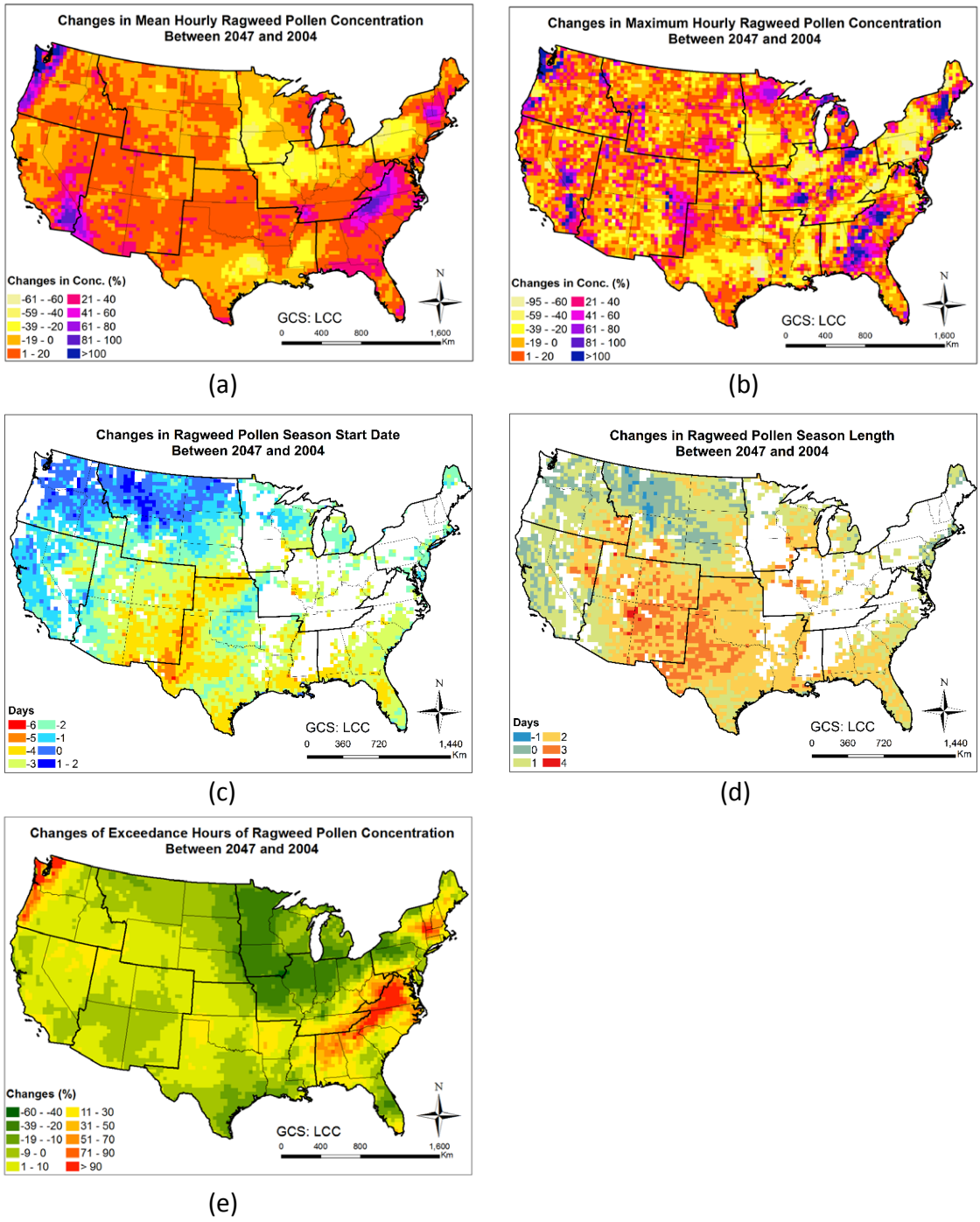


Figure 2.18: Changes in ragweed pollen season between 2004 and 2047. (a) Mean hourly concentrations, (b) Maximum hourly concentrations, (c) Start date, (d) Season Length, and (e) Exceedance hours

Table 2.3: Regional average and standard deviation of the changes in mean and maximum hourly concentrations, start date, season length and exceedance hours for ragweed pollen. (mean ± standard deviation).

Climate Region	Mean Hourly (%)	Max Hourly (%)	Start Date (day)	Season Length (day)	Exceedance Hours (%)
Central	-2.1 ± 23.3	-1.2 ± 38.1	-2.8 ± 0.7	1.8 ± 0.6	-0.6 ± 28.9
East North Central	-12.5 ± 18.5	2.3 ± 32.3	-2.0 ± 0.8	1.4 ± 0.7	-19.7 ± 9.7
Northeast	-0.6 ± 25.1	0.1 ± 53.5	-1.8 ± 0.7	0.7 ± 0.5	11.9 ± 26.9
Northwest	19.3 ± 40.7	18.6 ± 50.7	-0.7 ± 0.9	0.9 ± 0.8	34.1 ± 195.1
South	0.5 ± 14.3	-5.3 ± 24.1	-3.3 ± 1.0	2.0 ± 0.6	1.8 ± 9.7
Southeast	22.7 ± 21.0	12.4 ± 43.2	-2.9 ± 0.9	1.4 ± 0.6	34.3 ± 36.5
Southwest	10.7 ± 14.6	3.7 ± 33.7	-3.1 ± 0.9	2.1 ± 0.7	1.2 ± 4.8
West	11.4 ± 23.2	9.0 ± 33.6	-1.5 ± 0.8	1.0 ± 0.8	3.9 ± 6.6
West North Central	-2.5 ± 12.0	2.7 ± 25.2	-1.1 ± 1.2	0.8 ± 0.8	-3.8 ± 8.3

2.4.7 Uncertainty analysis

There are substantial uncertainties in each of the components and modules of the developed modeling system of pollen emission and transport. For each of the model components and its modules, the uncertainty has mainly resulted from the model formulations, parameters and the input data. In the current study, great efforts have been made to identify and reduce the uncertainties in each of the model components and modules based on different methods.

For the meteorology simulations from the WRF model, quality control measures have been applied by a modeling group at USEPA to evaluate the quality of the archived meteorology data [65, 137]. The pollen counts themselves may also be subject to large uncertainties [146]. The observed pollen counts from NAB-AAAAI stations were examined carefully according to quality control measures to ensure data quality [10, 12].

For the developed pollen emission model, global sensitivity analysis was conducted to identify sensitive and interactive input parameters based on Morris' design [147, 97]. The values of highly sensitive and interactive parameters were carefully chosen from the literature or parameterized using literature data. Many iterations of the

emission model have been tried to ensure the consistency and quality of the simulated pollen emission data. For the pollen transport model, process analysis has been conducted to identify the contributions of each physical process on the airborne pollen concentrations.

2.5 Summary

A modeling system incorporating pollen emission and transport has been applied to simulate the spatiotemporal distributions of allergenic pollen of oak and ragweed under impacts of climate change. The simulation domain covers the CONUS with 36 km horizontal grid spacing, temporal resolution of one hour, and 34 layers in the vertical direction. The results were evaluated with correlation analysis, fractional bias, hit and false rates. This model was able to capture the distribution of observed pollen concentration. Process analyses indicate that dry deposition, emission and vertical eddy diffusion were the dominant processes determining the ambient pollen concentrations. Uncertainty analyses were conducted to identify the sources of uncertainties in each of the components of the pollen emission and transport system. The boundary condition of airborne pollen concentration exerted remarkable influence on mean pollen concentrations at locations with small emission sources.

Five pollen indices (mean hourly concentrations, maximum hourly concentrations, start date, season length, and exceedance hours) were reported and compared for each pollen species across the nine climate regions. It was estimated that ragweed pollen season will start earlier and last longer under the RCP 8.5 scenario for all the nine climate regions, with increasing average pollen concentrations in most regions. The response of oak pollen season varies across the nine climate regions, with the largest increase in pollen concentration in the Northeast region.

Chapter 3

CO-OCCURRENCE OF ALLERGENIC POLLEN AND OZONE EXCEEDANCES UNDER CLIMATE CHANGE

3.1 Abstract

Prevalence of Allergic Airway Disease (AAD) is growing globally, resulting in increased numbers of emergency department visits and hospitalizations. Clinical studies show that AAD can be exacerbated by the synergistic action of aeroallergens such as pollen and spores, and atmospheric pollutants such as ozone. The present study investigates changes of ragweed pollen indices (start date, season length, etc.) and ozone indices (standard exceedances) during 1994-2010 in the nine climate regions of the CONUS. Analyses of observed pollen counts and ozone concentrations at the locations of 58 pollen monitor stations were conducted. The simultaneous impacts of climate change on ragweed pollen and ozone concentration were studied. The co-occurrence of ragweed pollen and ozone standard exceedances during historical and future periods were investigated. It is predicted that the ragweed pollen and ozone concentrations in 2047 under RCP 8.5 scenario will simultaneously increase for the Southwest and West regions, but decrease for the Central, East North Central and West North Central regions. The co-occurrence of ragweed pollen and ozone exceedances will affect a remarkable fraction of population. Inhalation exposures to ragweed pollen are higher in outdoor environments compared with indoor environments. Male and younger population groups tend to have higher exposures to ragweed pollen and ozone.

3.2 Introduction

Ambient air quality has been substantially impacted by climate change over the past few decades. In fact, climate change critically affects both the atmospheric processes involved in the dynamics of air pollution systems and the dynamics of biogenic emissions, including tree and grass pollens and fungal spores. Synergism of allergenic pollen with air pollutants like ozone and particulate matter has been reported and can exacerbate the AAD of allergy sufferers [88, 134, 89, 14]. In particular, the patterns of emission and transport of allergenic pollen and common air pollutants are expected to be impacted by the changing climate [90, 61, 91, 92, 93, 12, 148].

The interaction between airborne pollen and gaseous pollutants, such as ozone, has been investigated by some studies to evaluate the impacts of exposures to these pollutants on pollen [149, 150, 151, 152, 153]. High ambient ozone levels have been found to be a critical factor for enhanced allergenicity of birch pollen [154, 155]. It was revealed that ozone exposure triggers changes in both the chemotactic and the immune modulatory potential of the pollen. Increased ozone exposure generated enhanced allergen content and skin prick test reactivity. With increasing ozone levels, the symptoms of pollen allergic patients will also increase. Experiments demonstrated that elevated ozone exposure during the growth phase of plants will lead to increased protein and allergen content of rye grass pollen [156]. Similar results were found with another allergenic species, Arizona cypress, by skin tests and in vitro tests [157]. These findings indicate that with expected increase in air pollutants under climate change, the airborne allergens and associated allergies are likely to increase. It was estimated that 77 million people in Europe will have sensitization to ragweed pollen by 2041–2060 under RCP 4.5 and 8.5 scenarios, which is double the current number [148, 158].

The goal of this study was to investigate the changes of ragweed pollen indices (start date, season length, etc.) and an ozone index (standard exceedances) during

1994-2010 in the nine climate regions of the CONUS. Analyses of observed pollen counts and ozone concentrations at the locations of 58 pollen monitor stations were conducted. The spatiotemporal distribution of ozone concentration in 2047 under RCP 8.5 scenario is compared with predicted ragweed pollen concentration under the same climate scenario. The simulated co-occurrence of ragweed pollen and ozone exceedances in 2004 and 2047 were also compared. In addition, population exposures to ragweed pollen and ozone were simulated with a virtual population. Factors affecting people's exposure to pollen and ozone were identified.

3.3 Methods

3.3.1 Analysis of historical ragweed pollen observation and ozone exceedances

Data sources

Observed daily airborne ragweed pollen counts were obtained from all available monitoring stations of the National Allergy Bureau (NAB) at the American Academy of Allergy, Asthma and Immunology (AAAAI) during the period of 1994-2010 across the CONUS (Figure 2.2). Fifty-seven NAB-AAAI stations were selected, because they recorded valid data for at least four years, for performing further analyses and modeling parameterization. The main climate characteristics and geographical locations of the studied stations are listed in Table C.1. Observed daily temperatures, precipitation and other climatic factors were obtained from the National Oceanic and Atmospheric Administration (NOAA) meteorology stations nearest to the corresponding NAB-AAAAI pollen stations.

Observed daily maximum 8-hour average ozone (DMA8[O_3]) data were downloaded from EPA website [130] and matched with the ragweed pollen data according to the longitude and latitude of the pollen monitoring stations and observation dates. The closest ozone monitoring station to each pollen monitoring station was selected.

Pollen indices

Four pollen indices were examined in this study: Start Date (SD), Season Length (SL), Peak Value (PV), and Annual total Production (AP) of daily counted airborne pollen. The pollen season start date and end date (days from January 1st of the year) are defined as the days when the cumulative pollen count reaches 5% and 95%, respectively, of the annual total pollen count. Season length (days) is defined as the duration between the start date and end date. Peak value (pollen grains/m^3) is the maximum daily pollen count recorded during the pollen season. Annual production (pollen grains/year) is the sum of daily pollen counts of a pollen season. Pollen data with SL of less than 7 or greater than 80 days are assumed to be unreasonable and excluded from analyses.

Changes of mean pollen indices between two periods: 1994-2000 and 2001-2010

To investigate the changes in ragweed pollen indices in the past two decades, the pollen data were divided into two periods: 1994-2000 and 2001-2010. There are limitations in data availability as most of the pollen monitoring stations do not have a full record of data during the two periods. Stations that have at least three years of data in each period were selected for further analyses. Student's t tests were performed to check the significance of changes in pollen indices during the periods of 1994-2000 and 2001-2010.

Changes in pollen indices at station i were calculated using Equation 3.1,

$$\begin{cases} \Delta\overline{SD}_i = \overline{SD}_{i,P2} - \overline{SD}_{i,P1} \\ \Delta\overline{SL}_i = \overline{SL}_{i,P2} - \overline{SL}_{i,P1} \\ \Delta\overline{AP}_i/\overline{AP}_{i,P1} = (\overline{AP}_{i,P2} - \overline{AP}_{i,P1})/\overline{AP}_{i,P1} \\ \Delta\overline{PV}_i/\overline{PV}_{i,P1} = (\overline{PV}_{i,P2} - \overline{PV}_{i,P1})/\overline{PV}_{i,P1} \end{cases} \tag{3.1}$$

where $\overline{SD}_{i,P1}$, $\overline{SL}_{i,P1}$, $\overline{AP}_{i,P1}$ and $\overline{PV}_{i,P1}$ are the mean SD, SL, AP and PV, respectively, during the period of 1994-2000 at station i; and $\overline{SD}_{i,P2}$, $\overline{SL}_{i,P2}$, $\overline{AP}_{i,P2}$ and $\overline{PV}_{i,P2}$ are the mean SD, SL, AP and PV, respectively, during the period of 2001-2010. All the statistical tests are performed using R version 3.5.2 [159].

Co-occurrence of ragweed pollen and ozone exceedances in 1994-2010

After the observed ozone and ragweed data had been collected for each pollen monitoring station and matched according to observation date, the number of days when both ragweed pollen concentration is greater than or equal to 1 and ozone exceedances (DMA8[O_3]>70 ppb) occur, are summed for each station and each year during 1994-2010. Then the annual average number of days for co-occurrence of ragweed pollen and ozone exceedances was calculated for each station. The pattern of this co-occurrence is analyzed in the following.

3.3.2 Spatiotemporal distribution of ozone concentration under climate change

The spatiotemporal distributions of ozone concentration in 2004 and 2047 were simulated by our collaborator Dr. C. Nolte's group at the USEPA. The meteorology inputs are the same as those used for the ragweed pollen concentration simulation, which were downscaled from the CESM using WRF model. Only the meteorological conditions and the methodologically dependent emissions are changed between 2004 and 2047 for the CMAQ simulations. All other inputs, such as the anthropogenic emissions, chemical lateral boundary condisions, and land use and land cover classifications etc. were kept the same for the two simulations. The simulation domain covers the CONUS with 36 km horizontal grid spacing, temporal resolution of one hour, and 34 layers in the vertical direction.

3.3.3 Spatiotemporal distribution of ragweed pollen concentration under climate change

The simulations of spatiotemporal distributions of ragweed pollen concentration in 2004 and 2047 were discussed in Chapter 2. The monthly mean ragweed pollen concentrations are presented in Figure 2.4. The seasonal mean and maximum ragweed pollen concentrations are displayed in Figure 2.12. The results will be analyzed in this chapter to compare with ozone concentration under climate change.

3.3.4 Co-occurrence of ragweed pollen and ozone exceedances under climate change

Based on the simulated ragweed pollen concentration and ozone concentration in 2004 and 2047, the regional changes of mean ragweed pollen concentration and DMA8[O_3] during August and September between 2004 and 2047 in each of the nine climate regions were calculated. The number of days when both ragweed and ozone exceedance occur during the pollen season (August and September) in 2004 and 2047 were summed in each cell of the modeling grid and then plotted. The difference of the number of days between 2004 and 2047 were calculated. These changes are calculated based on Equation 3.2,

$$\begin{cases} \Delta C_{hr,Mn}(i,j)/C^{2004}_{hr,Mn}(i,j) = \frac{C^{2047}_{hr,Mn}(i,j) - C^{2004}_{hr,Mn}(i,j)}{C^{2004}_{hr,Mn}(i,j)} \\ \Delta DMA8[O_3]/DMA8[O_3]^{2004}(i,j) = \frac{DMA8[O_3]^{2047}(i,j) - DMA8[O_3]^{2004}(i,j)}{DMA8[O_3]^{2004}(i,j)} \\ \Delta N_{Co-oc}(i,j) = N^{2047}_{Co-oc}(i,j) - N^{2004}_{Co-oc}(i,j) \end{cases} \tag{3.2}$$

where $C^{2004}_{hr,Mn}(i,j)$ and $C^{2047}_{hr,Mn}(i,j)$ are the mean ragweed pollen concentration (pollen grains/m^3) for cell (i,j) of the modeling grid during the pollen season in 2004 and 2047, $DMA8[O_3]^{2004}(i,j)$ and $DMA8[O_3]^{2047}(i,j)$ are the average daily maximum 8-hour average ozone for cell (i,j) of the modeling grid during August and September in 2004 and 2047, $N^{2004}_{Co-oc}(i,j)$ and $N^{2047}_{Co-oc}(i,j)$ are the number of days when both

ragweed and ozone exceedance occur in cell (i, j) during the pollen season (August and September) in 2004 and 2047.

3.3.5 Exposures to ragweed pollen and ozone

The exposures to ragweed pollen and ozone are estimated with a probabilistic model based on the simulated ragweed pollen and ozone concentration in 2004, the demographic data and human activity patterns. Figure 3.1 presents the diagram of the exposure modeling system. Because allergenic airway disease is caused by inhalation exposure to allergens, only inhalation exposures to ragweed pollen and ozone are considered.

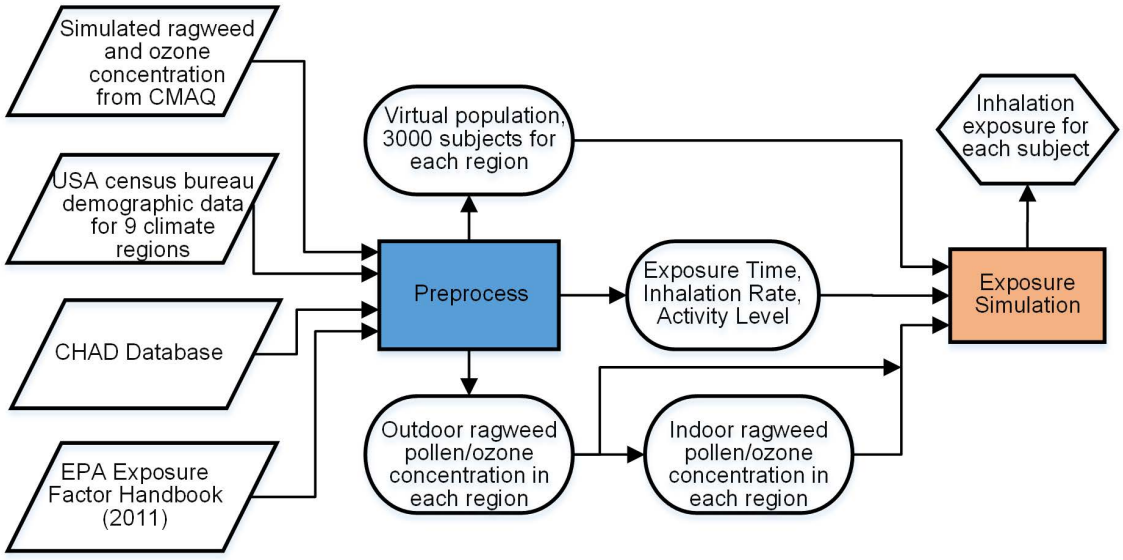


Figure 3.1: The schematic illustration of exposure modeling system.

The exposures were simulated for 3000 "virtual subjects" in each of the nine climate regions. The subjects were sampled from the demographic data for 2000 in each region from the US Census Bureau [160] so that they can represent the population in the region. Each subject was chosen randomly and assigned an age and gender.

The ragweed pollen/ozone concentrations are derived from the CMAQ simulation results in 2004. Outdoor daily ragweed pollen/ozone concentrations ($C_{out}(i, j)$) for a "virtual subject" on day i in climate region j are sampled from the simulated

daily pollen/ozone concentration on day i in climate region j. The indoor daily ragweed pollen/ozone concentrations ($C_{in}(i,j)$) for a "virtual subject" on day i in climate region j is a ratio of the outdoor daily ragweed pollen/ozone concentrations ($C_{out}(i,j)$), and the ratio is assumed to be uniformly distributed within a range as reported in the literature [24, 27, 26, 161, 3].

The exposure time indoor and outdoor for each subject varies with his/her age, gender, climate region and day of the year. The outdoor exposure time $t_{out}(i,j)$ (hours) for a "virtual subjects" on day i in climate region j was sampled from the observed outdoor exposure times based on his/her age and gender, which are retrieved from the Consolidated Human Activity Database (CHAD) [162].

Exposure factors, including inhalation rates, are derived from the EPA Exposure Factor Handbook [163]. The inhalation rate indoors, $IR_{in}(a,g)$ and outdoors, $IR_{out}(a,g)$, depend on the subject's age a, gender g and activity level as shown in Equation 3.3 and 3.4,

$$\begin{aligned} IR_{in}(a,g) = IR(a,g,L_P)f_{P,in} + IR(a,g,L_L)f_{L,in} + \\ IR(a,g,L_M)f_{M,in} + IR(a,g,L_H)f_{H,in} \end{aligned} \tag{3.3}$$

$$\begin{aligned} IR_{out}(a,g) = IR(a,g,L_P)f_{P,out} + IR(a,g,L_L)f_{L,out} + \\ IR(a,g,L_M)f_{M,out} + IR(a,g,L_H)f_{H,out} \end{aligned} \tag{3.4}$$

where L_P, L_L, L_M and L_H indicate passive, low, moderate and high activity levels, respectively; and $f_{P,in}$, $f_{L,in}$, $f_{M,in}$ and $f_{H,in}$ are fractions of time spent in indoor environments at passive, low, moderate and high activity levels, respectively, $f_{P,out}$, $f_{L,out}$, $f_{M,out}$ and $f_{H,out}$ are fractions of time spent in outdoor environments at passive, low, moderate and high activity levels, respectively. The inhalation rates for different ages, genders and activity levels, and the fractions of time spent at the corresponding activity level are derived from data in the EPA Exposure Factors Handbook [163].

The inhalation exposures ($E_{inha}(i,j,a,g)$) to ragweed pollen/ozone for a "virtual

subject" with age a and gender g on day i in climate region j is calculated based on Equation 3.5,

$$E_{inha}(i,j,a,g) = IR_{in}(a,g)C_{in}(i,j)t_{in}(i,j) + IR_{out}(a,g)C_{out}(i,j)t_{out}(i,j) \quad (3.5)$$

3.4 Results and Discussion

3.4.1 Historical ragweed pollen observation and ozone exceedances

Mean ragweed pollen indices across latitude

In order to investigate the spatial patterns of ragweed pollen across the CONUS, the mean pollen indices for each pollen monitoring station during 1994 to 2010 are calculated and plotted against the latitude of each station. As shown in Figure 3.2, the ragweed pollen season starts earlier in areas with higher latitude and then shifts to lower latitude areas. But the pollen season lasts longer at lower latitudes. The peak pollen concentration and annual production both decrease as latitude increases. Statistical tests shows that there is significant correlation (p-value < 0.05) between start date/season length/annual production and latitude as shown in Figure 3.2.

Correlation between pollen indices and meteorological factors

Figure 3.3 presents the Pearson correlation heat map for the ragweed annual pollen indices during 1994 to 2010 and the following factors: mean precipitation (PRCP), mean wind speed (WDSP), minimum temperature (TMin), maximum temperature (TMax), mean temperature (Temp) during pollen season, and the elevation and latitude of the monitoring stations. Dots indicate significant correlation (p-value < 0.05) between two variables, while an "x" indicates that the correlation was not statistically significant. Start date has significant correlation with all the factors except mean temperature. Season length has significant correlation with all the factors except mean wind speed. Peak value is mainly affected by mean temperature, while

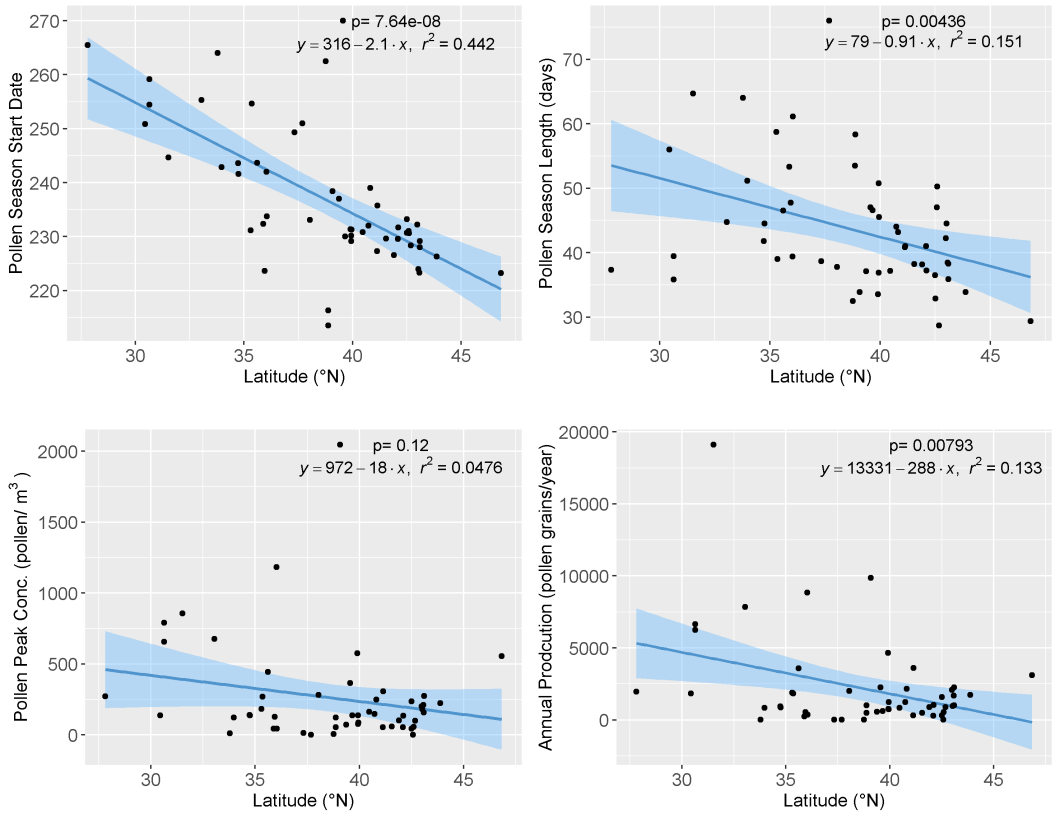


Figure 3.2: The mean pollen indices (1994-2010) for each station across latitudes. The pollen season start dates are represented as the number of days from January 1st of the year.

annual production is influenced by mean temperature, mean wind speed, maximum temperature and latitude. Multiple linear regression was performed for each pollen index and the factors above, and the coefficients are shown in Table 3.1. It is shown that maximum temperature and elevation contribute significantly to start date in the multiple regression model; maximum temperature, minimum temperature, elevation and latitude are major contributors to Season Length; wind speed and latitude are the significant contributors to Annual Production in the model.

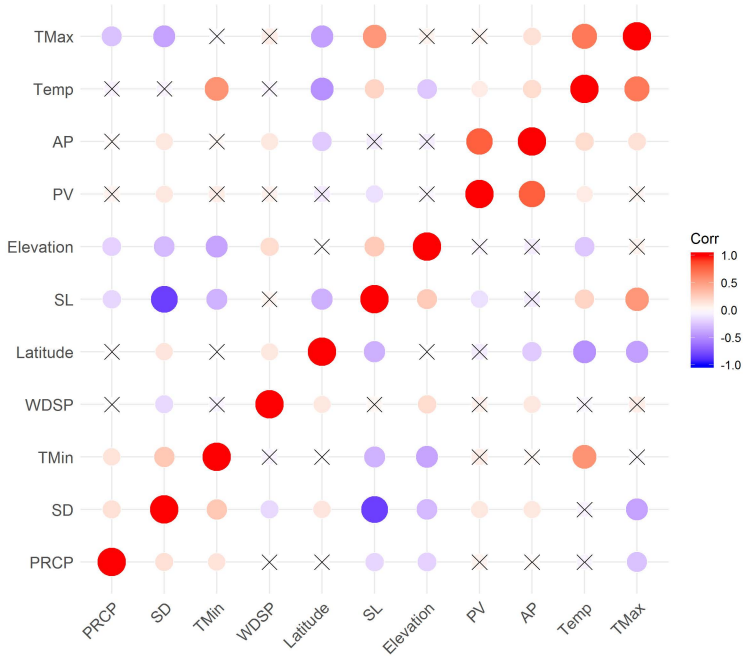


Figure 3.3: Pearson correlation heat map (with hierarchical clustering) for the pollen indices and meteorology indices.

Table 3.1: Coefficients for the multiple linear regression between pollen indices and meteorological and geological factors. Asterisk (*) indicates significant estimate (p<0.05). Temp: mean temperature; PRCP: mean precipitation; WDSP: mean wind speed; TMax: maximum temperature; TMin: minimum temperature.

Pollen Index	Temp	PRCP	WDSP	Tmax	Tmin	Elevation	Latitude
Start Date	1.75	-7.87	-2.21	-3.74*	0.65	-0.02*	0.28
Season Length	0.95	9.18	-0.31	2.90*	-1.86*	0.02*	-1.31*
Peak Value	4.08	332.96	30.26	0.11	4.19	-0.02	-8.47
Annual Production	80.92	917.28	357.59*	-7.00	-29.15	-0.71	-145.93*

Changes of mean pollen indices between periods 2001-2010 and 1994-2000

To analyze the changes in mean pollen indices for each station, there must be at least three years of ragweed pollen data at each monitoring station in both periods: 2001-2010 and 1994-2000. Therefore, only 17 stations are qualified. Table 3.2 displays the differences of mean pollen indices for each station between periods 2001-2010

and 1994-2000. The changes in start date and season length are visualized on the maps in Figure 3.4 and Figure 3.5. The ragweed pollen season tends to start earlier at most of the monitoring stations (13 of the 17 stations). The pollen monitoring stations in the Central climate region all experienced earlier onset of pollen season with the biggest change of 30 days. Changes in season length vary across latitudes and climate regions (-8 days to +8 days). Nine of the seventeen stations experienced a longer ragweed pollen season. Changes in peak value and annual production both have very big variation. Changes in peak value vary from -84% to 258.9%, while changes in annual production range between -89.8% to 148%. When examining the changes in four pollen indices across latitude, it is found that the changes in peak value and annual production decreased with latitude, while the changes in start date and season length increased with latitude. This might be caused by the changes in meteorological factors such as temperature, precipitation, humidity etc.

Table 3.2: Changes of mean pollen indices for each station between periods 2001-2010 and 1994-2000. Asterisk (*) indicates significant changes ($p<0.05$).

Pollen Station ID	**Start Date (days)**	**Season Length (days)**	**Peak Value (%)**	**Annual Production (%)**	**Latitude (ºN)**	**Longitude (ºW)**
25	1.5	-2.9	2.4	9.1	34.7	86.6
26	5.3	4.5	121.3	148.0*	34.8	92.4
28	-0.3	-0.5	20.6	1.3	35.3	80.8
32	-30.3*	-6.5	258.9*	67.9	35.9	84.0
42	-9.2	6.7	-84.0*	-89.8*	38.0	84.5
47	3.3	-2.7	61.9*	24.1	39.4	76.5
50	-1.0	-2.8	-20.9	-48.8*	39.9	86.2
51	-4.6	7.6	-11.4	-44.0	39.9	74.9
53	-7.2*	7.1	42.7	-25.4	40.0	75.2
58	-2.5	-3.3	-59.3	-55.2*	40.7	74.2
62	-2.5	-8.3*	26.6	24.2	41.5	73.1
65	1.1	2.3	-57.2*	-66.7*	42.1	78.4
66	-1.9	1.4	-21.0	-13.6	42.1	80.1
67	-8.4	4.5	-38.9*	-33.7*	42.5	70.9
68	-3.0	-0.3	-4.6	8.2	42.5	82.9
70	-12.2	6.5	9.3	-17.8	42.6	71.4
76	-1.7	7.7*	19.2	81.6*	43.1	77.6

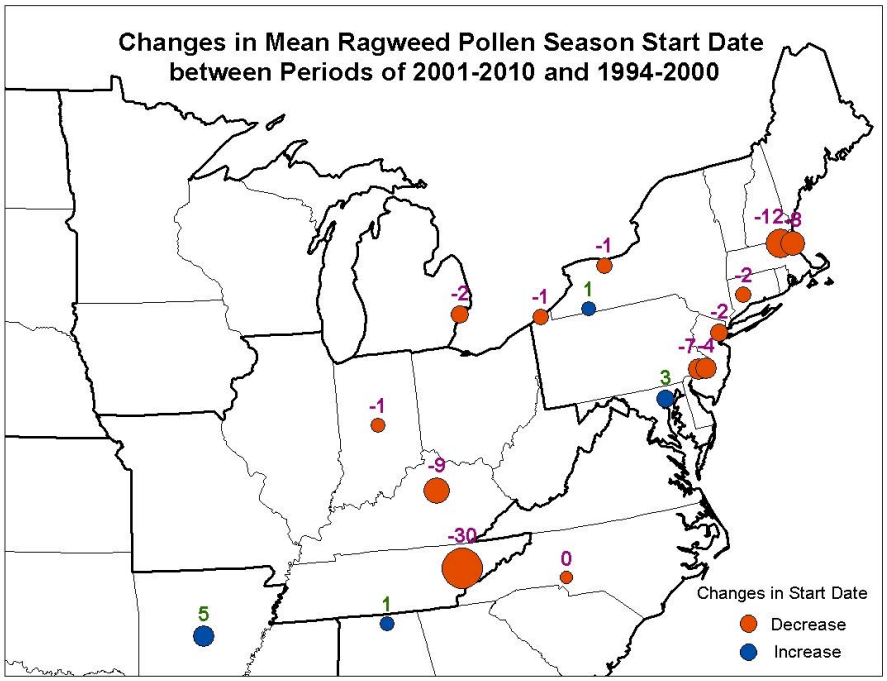


Figure 3.4: Changes in mean ragweed pollen season start date between periods of 2001-2010 and 1994-2000.

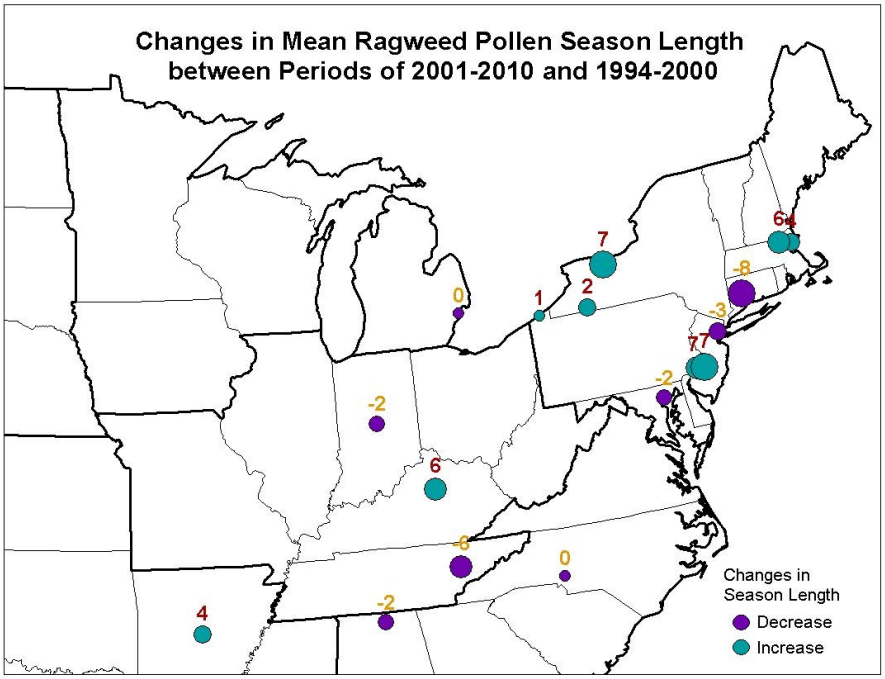


Figure 3.5: Changes in mean ragweed pollen season length between periods of 2001-2010 and 1994-2000.

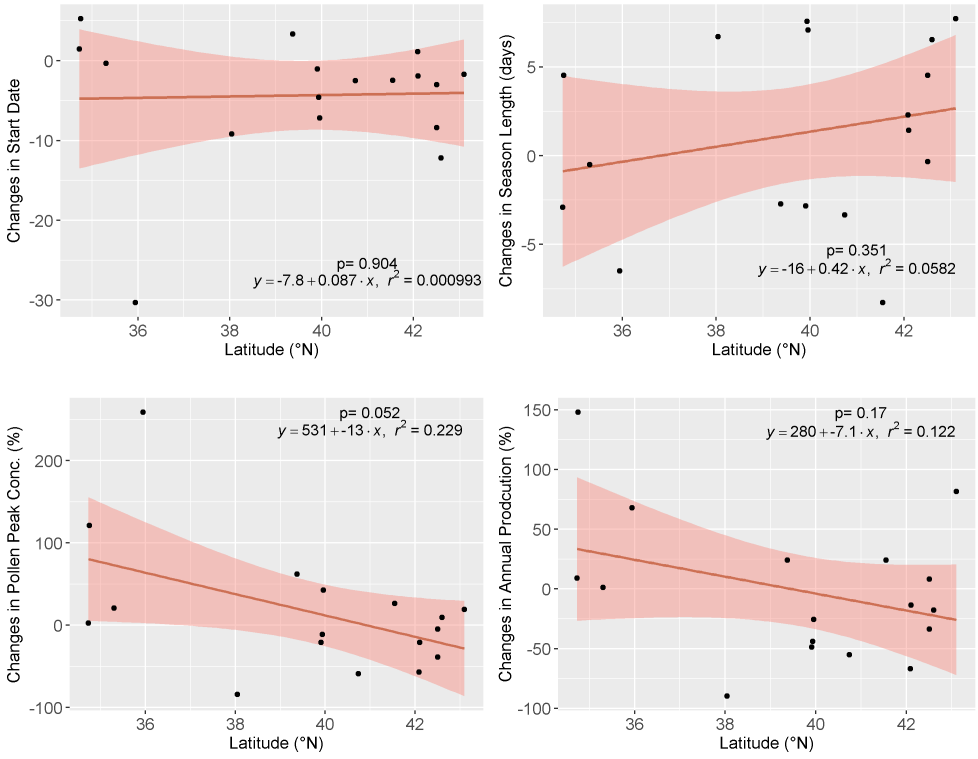


Figure 3.6: Changes in mean ragweed pollen indices between periods of 2001-2010 and 1994-2000 across latitudes.

Co-occurrence of ragweed pollen and ozone exceedances during 1994-2010

Figure 3.7 depicts the average number of days when ragweed is greater or equal to 1 and DMA8[O_3] is greater than 70 ppb for the location of each pollen monitoring station during 1994-2010, except 2001, 2002 and 2009 when there are no ragweed pollen observation data. The number of co-occurrences ranged from 0 to 17 days. The largest number of co-occurrences appeared in the West, Southwest, Southeast and South regions. The changes in ragweed pollen season length and ozone exceedance days across the nine climate regions between 2001-2010 and 1994-2000 are shown in Figure 3.8. The median ragweed pollen season lengths in each climate region for the two periods are displayed. The size of the the dots indicates the fraction of ozone exceedance days during the ragweed pollen season. The pollen season length has increased after 2000 for the East North Central, South, Southeast, Southwest, and West North Central regions. The fraction of ozone exceedance days during ragweed pollen season also increased for the Southwest and Northeast regions after 2000. The Southwest region has experienced increase in both ragweed pollen season length and ozone exceedances.

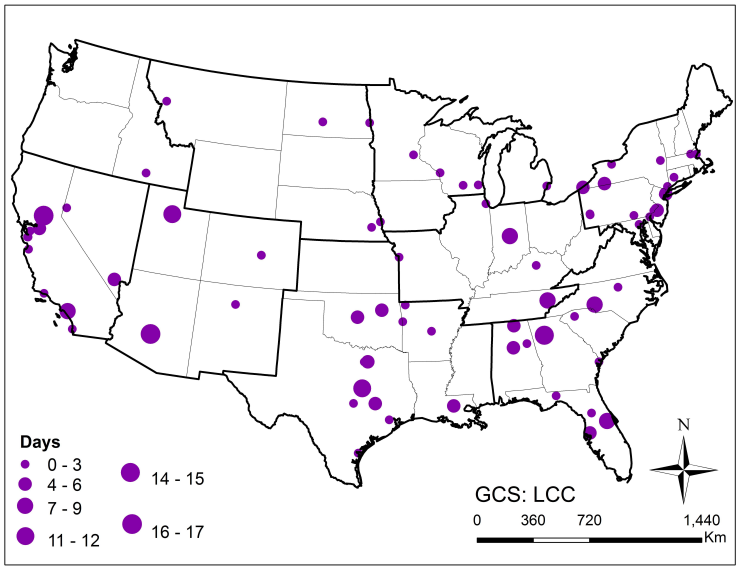


Figure 3.7: Annual average number of days when both ragweed pollen ≥1 and ozone exceedances (DMA8[O_3]>70 ppb) occur for 58 pollen stations during 1994-2010 (except 2001, 2002, and 2009).

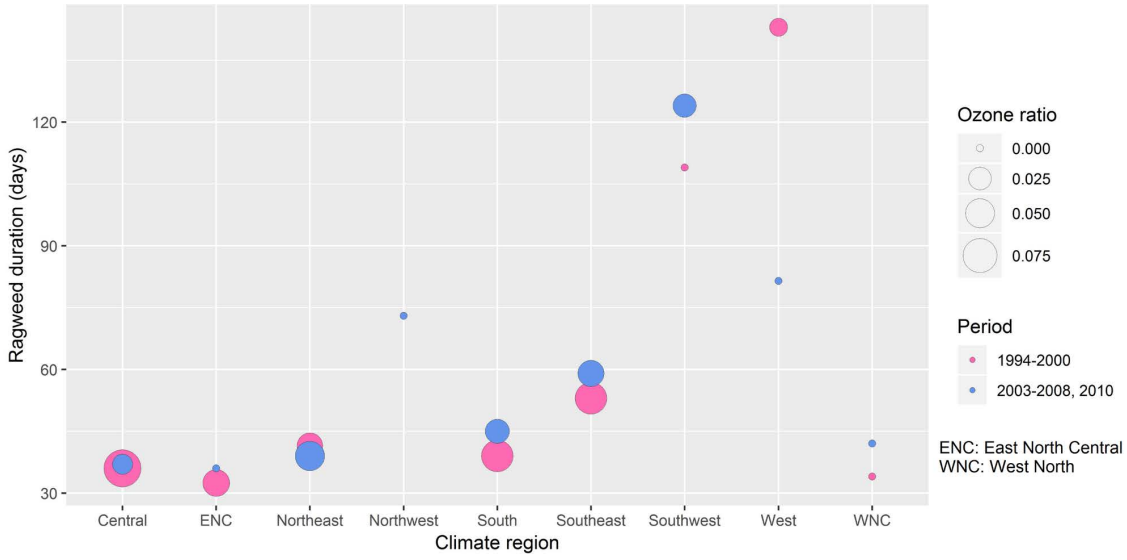


Figure 3.8: Changes in ragweed pollen season length and ozone exceedance days across the nine climate regions.(Ozone ratio: the fraction of ozone exceedance days during ragweed pollen season)

3.4.2 Distributions of ozone and ragweed pollen concentrations during 2004 and 2047

The simulated mean DMA8[O_3] across the CONUS during August to September in 2004 and 2047 is displayed in Figure 3.9. The mean DMA8[O_3] in 2004 ranged from 22 ppb to over 70 ppb. It is relatively higher in the West, Southwest, South and West North Central regions, and relatively lower in the Northwest and the Southeast regions in 2004. The main pattern of mean DMA8[O_3] did not change significantly in 2047. Figure 3.10 shows the changes in mean DMA8[O_3] during August to September between 2047 and 2004. The mean DMA8[O_3] will increase in most areas of the West, Southwest, South and Northeast regions, but decrease in most parts of the Southeast, Central and East North Central regions. When looking at the average changes across each climate region in mean DMA8[O_3] and mean ragweed pollen concentration in Figure 3.11, there is simultaneous increase in both of them for the Southwest and the West regions, while the Central, East North Central and West North Central regions will experience decreases in both mean DMA8[O_3] and mean ragweed pollen concentration. The Southeast region has the highest increase in mean ragweed pollen concentration in 2047, followed by the Northwest region, but both of the two regions will expect decrease in mean DMA8[O_3] in 2047.

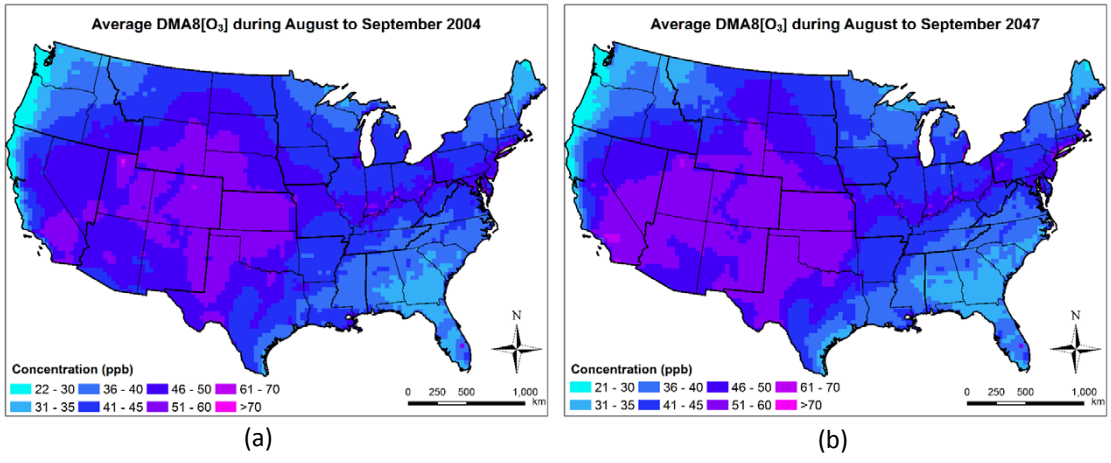


Figure 3.9: Average DMA8[O_3] during August to September in 2004 and 2047.

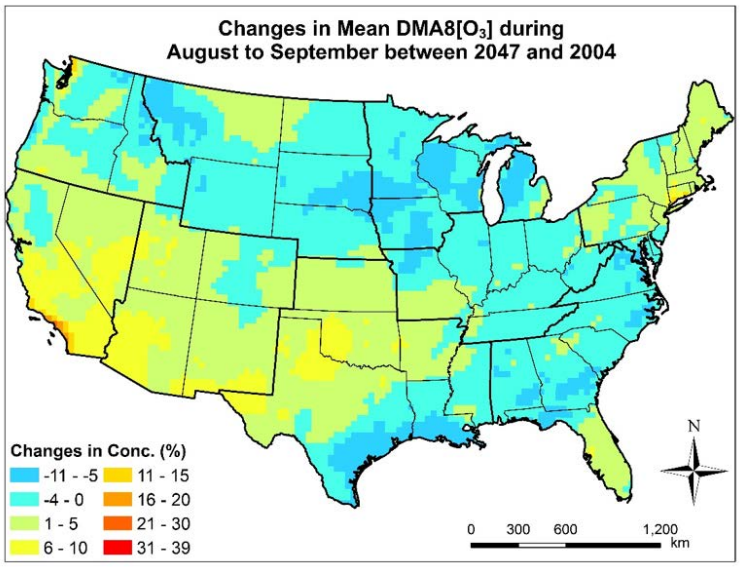


Figure 3.10: Changes in DMA8[O_3] between 2004 and 2047.

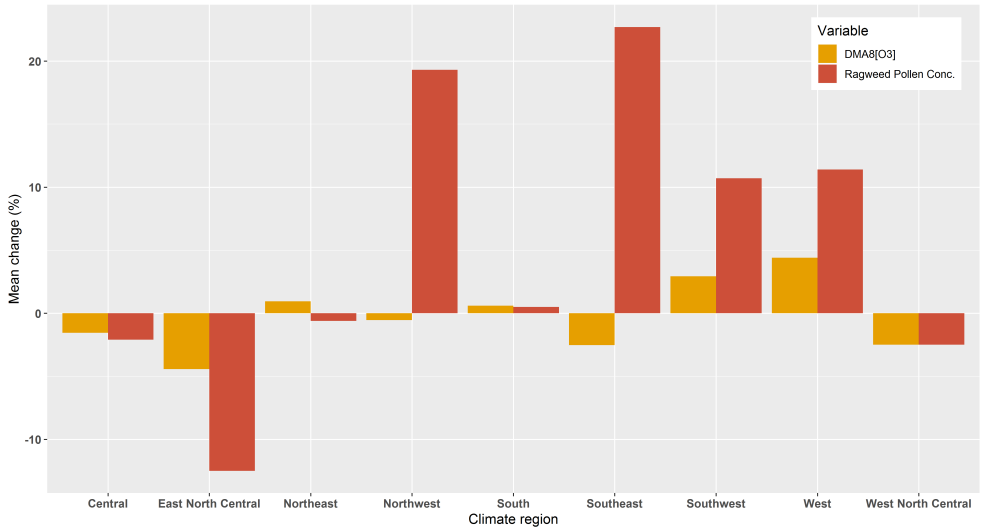


Figure 3.11: Average changes in DMA8[O_3] and ragweed pollen for the nine climate regions between 2047 and 2004.

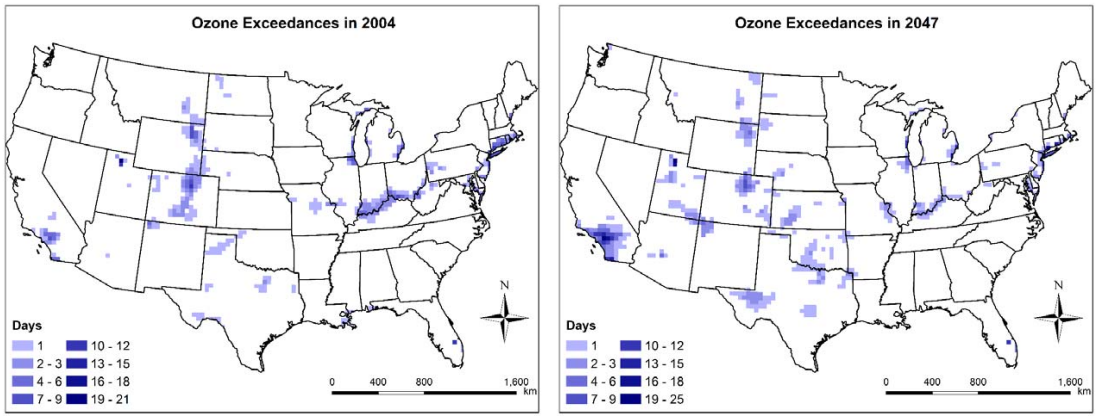
3.4.3 Co-occurrence of ragweed pollen and ozone exceedances

The simulated ozone standard exceedances (DMA8[O_3]>70 ppb) during August to September in 2004 and 2047 are shown in Figure 3.12. The ozone exceedances mainly occur in the West, West North Central, Southwest and Central regions in both 2004

and 2047. The South region is expected to have more ozone exceedances in 2047 compared with 2004. The number of ozone exceedances ranged from 0 to 21 days in 2004, and 0 to 25 days in 2047. Figure 3.13 displays the co-occurrence of ragweed pollen and ozone exceedances in 2004 and 2047. The pattern of co-occurrence is similar to that of ozone exceedances, which implies that the co-occurrence is predominantly influenced by ozone concentration. The changes in co-occurrence of ragweed pollen and ozone exceedances between 2047 and 2004 (Figure 3.14) vary from -9 to 14 days. The increase in co-occurrence mainly appears in the West, Southwest and South regions, while the co-occurrence will decrese in the Central region.

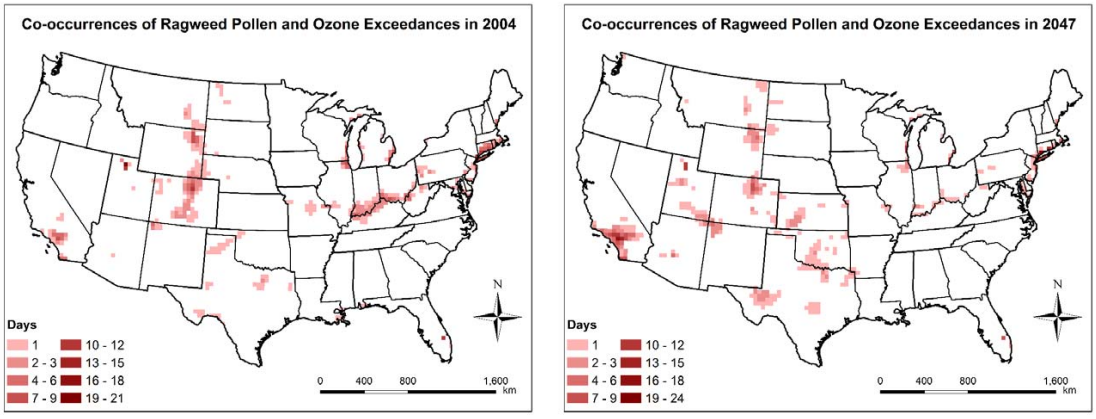
Although the co-occurrence of ragweed pollen and ozone exceedance scatters over the CONUS, it influences a remarkable fraction of population, as shown in Figure 3.15. Most of the heavily populated areas are predicted to have ragweed pollen and ozone exceedances in 2004 and 2047. Among them, five of the top 10 largest cities by population are included in 2004: New York, NY, Los Angeles, CA, Chicago, IL, Dallas, TX, and Detroit, MI. San Diego, CA will be added to the list in 2047.

Figure 3.17 illustrates the trend of DMA8[O_3] and daily average ragweed pollen concentration for two representative cities: Los Angeles (LA) and New York City (NYC), which are the top two largest U.S. cities by population. The co-occurrence of ozone exceedance and ragweed pollen is highlighted with red dots in the figure. Both LA and NYC will experience increased number of co-occurrence in 2047 compared with 2004. It is predicted that LA will have 17 days of co-occurrence in 2047, which is 10 days more than 2004, while NYC will have 20 days of co-occurrence, compared with 9 days in 2004. Some of the co-occurrences will happen consecutively for four or five days as shown in Figure 3.17.



(a) (b)

Figure 3.12: Simulated ozone exceedances during August to September in 2004 and 2047.



(a) (b)

Figure 3.13: Co-occurrence of ragweed pollen and ozone exceedances in 2004 and 2047.

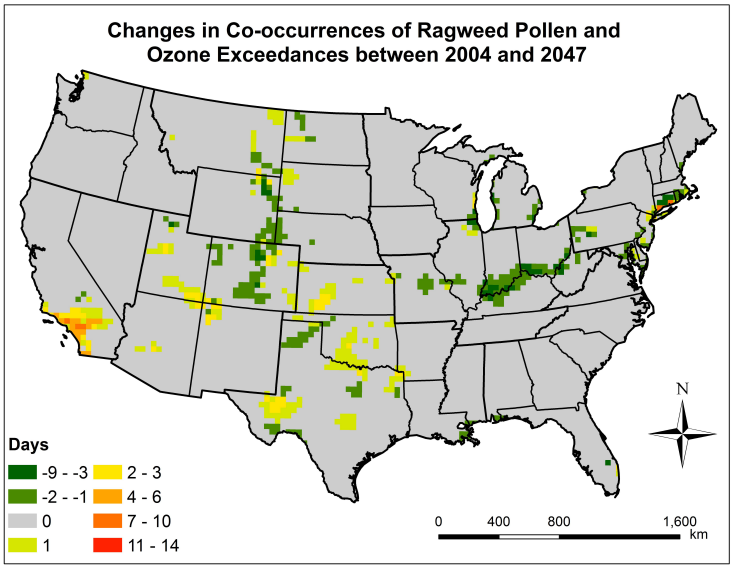


Figure 3.14: Changes in co-occurrences of ragweed pollen and ozone exceedances between 2047 and 2004.

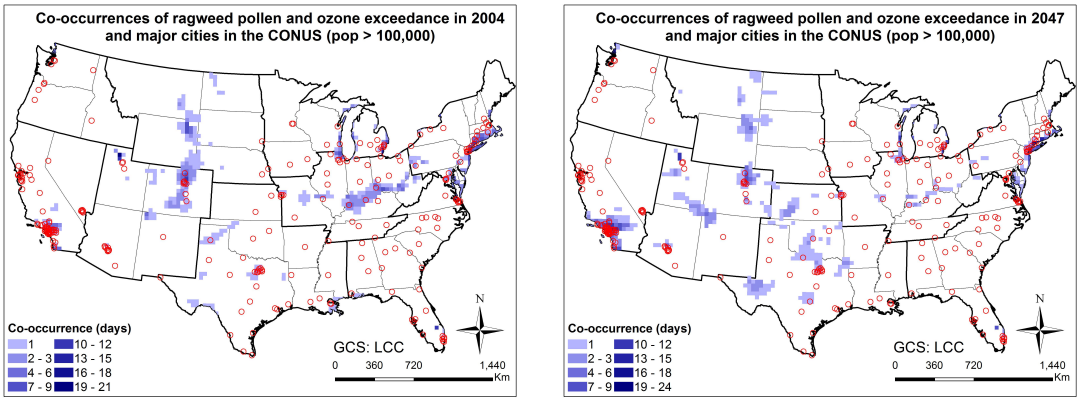


Figure 3.15: Co-occurrences of ragweed pollen and ozone exceedance in 2004 and 2047 and cities with population larger than 100,000 which are represented with red circles in the figure.

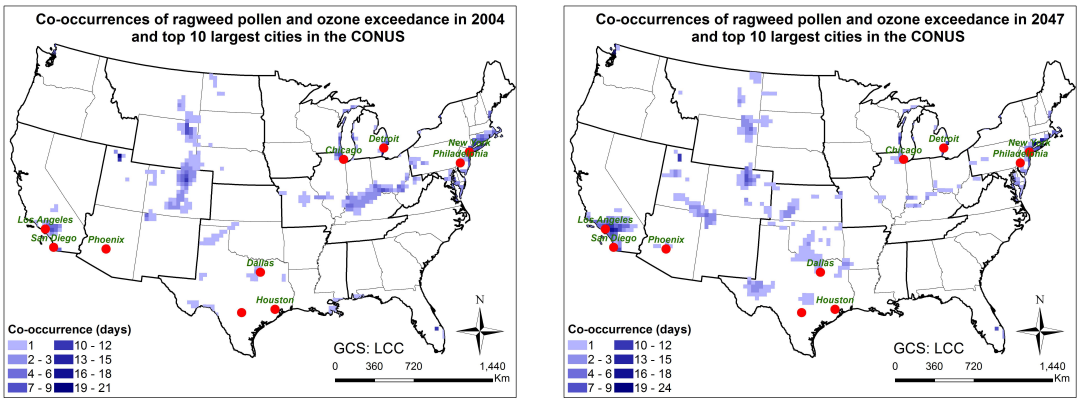


Figure 3.16: Co-occurrences of ragweed pollen and ozone exceedance in 2004 and 2047 and top 10 largest cities in the U.S.

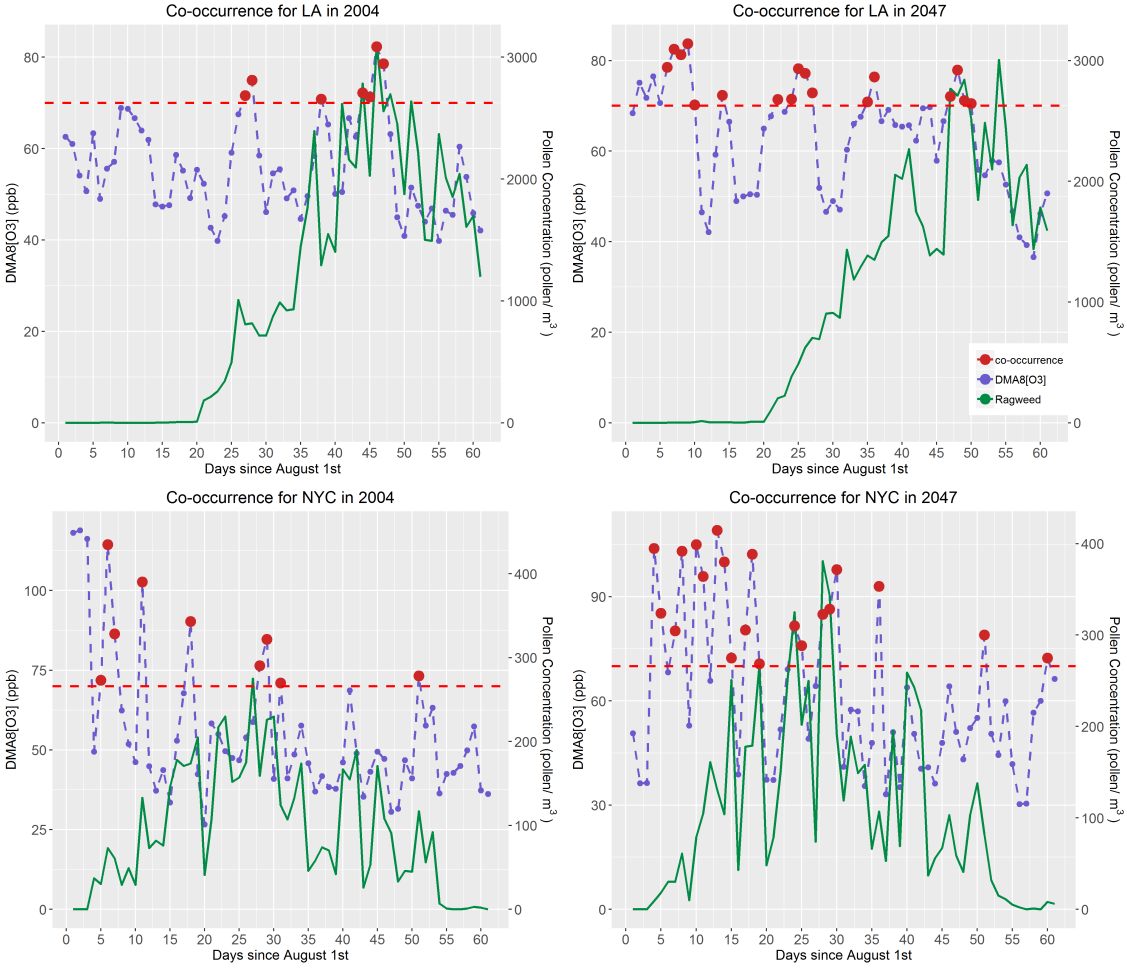


Figure 3.17: Time series plot of DMA8[O_3] and ragweed pollen concentration during August and September in 2004 and 2047 for Los Angeles (LA) and New York City (NYC). The red dots in the figures indicate the co-occurrence of ozone exceedance and ragweed pollen. The red dashed line indicates the ozone standard of 70 ppb.

3.4.4 Simulated exposures to ragweed pollen and ozone

The mean daily inhalation exposures to ragweed pollen and ozone in the nine climate regions have been compared by microenvironment (indoor/outdoor, Figure 3.18), gender (Figure 3.19) and age group (Figure 3.20). In general, ragweed pollen inhalation exposures outdoors are higher than its indoor inhalation exposures across the nine climate regions, which is the opposite of what happens for ozone exposures. Although people spent most of their time indoors, the pollen concentration is much higher outdoors. The ratios of indoor to outdoor ragweed pollen concentration and ozone concentration vary significantly. In this study, the ratios of indoor to outdoor ozone concentrations are assumed to vary from 0.05 to 0.7, whereas the ratios of indoor to outdoor ragweed pollen concentrations were bounded between 0.006 to 0.2, based on data from various studies [24, 27, 26, 161]. Due to lack of data on the ratios of indoor to outdoor ozone across the nine climate regions, the same ratios were applied to all regions. The inhalation exposure to ragweed has a large variation across different regions mainly due to ragweed coverage (Figure 2.3) and ragweed pollen concentration (Figure 2.12). The West North Central region has the highest population inhalation exposure to ragweed pollen (total of about 250 pollen grains/(day kg BW)), followed by the South, Southwest and West regions. The inhalation exposure to ozone has less variation across the CONUS compared with ragweed, and the ozone exposures indoors are higher than those outdoors. The mean calculated total daily ozone inhalation exposure is about 10 µg/(day kg BW), which is comparable with that reported by Weschler [23]. The outdoor ozone inhalation exposure is about 1/2 to 2/3 of indoor ozone inhalation exposure.

The inhalation exposures to ragweed pollen and ozone for males are higher than females due to the fact that males in general have higher inhalation rates [163] and spend more time outdoors (Figure 3.23). The variations between the nine climate regions are similar to those observed in Figure 3.18. The West North Central region has the highest mean inhalation exposure to ragweed pollen for males, which is about

260 pollen grains/(day kg BW). The population in the South West region have the highest mean daily inhalation exposures to ozone, which are 11.2 µg/(day kg BW) for males and 10.5 µg/(day kg BW) for females.

The Physiological Daily Inhalation Rates (PDIRs) per Unit of Body Weight (m^3/(hr kg BW)) decrease by age (Figure 3.21). Therefore, the inhalation exposures to ragweed and ozone also decrease from younger populations to older populations as shown in Figure 3.20. The most susceptible population is the 1-4 years old population group, especially for the West North Central region, where the mean daily inhalation exposure to ragweed pollen could reach up to 750 pollen grains/(day kg BW). The patterns of mean time spent outdoors for each age group are similar between different regions (Figure 3.22). Even though the 1-4 years old population group spends less time outdoors, their high inhalation rate per unit of body weight results in the highest inhalation exposure to ragweed. The 1-4 years old population group also has the highest mean inhalation exposure to ozone among the whole population across the nine climate regions. The population in the Southwest region has slightly higher exposure to ozone than people in other regions due to its high ozone concentration (Figure 3.9).

To show the changes of inhalation exposure to ragweed pollen and ozone with time and related factors, Figure 3.24 and Figure 3.25 illustrates the profiles of simulated daily ragweed pollen/ozone concentration, exposure time, inhalation rate and inhalation exposure to ragweed pollen/ozone for a simulated "virtual subject", who is a 3 years old male in the West region. The simulation period is from August 1st to September 30th in 2004. The inhalation exposure to ragweed pollen for the subject follows the patterns of ragweed pollen concentration, which starts to increase in late August and peaks in mid September. The inhalation exposure to ozone indoors and outdoors changes with ozone concentration and exposure time. The inhalation exposure to ozone indoors overall is higher than that outdoors.

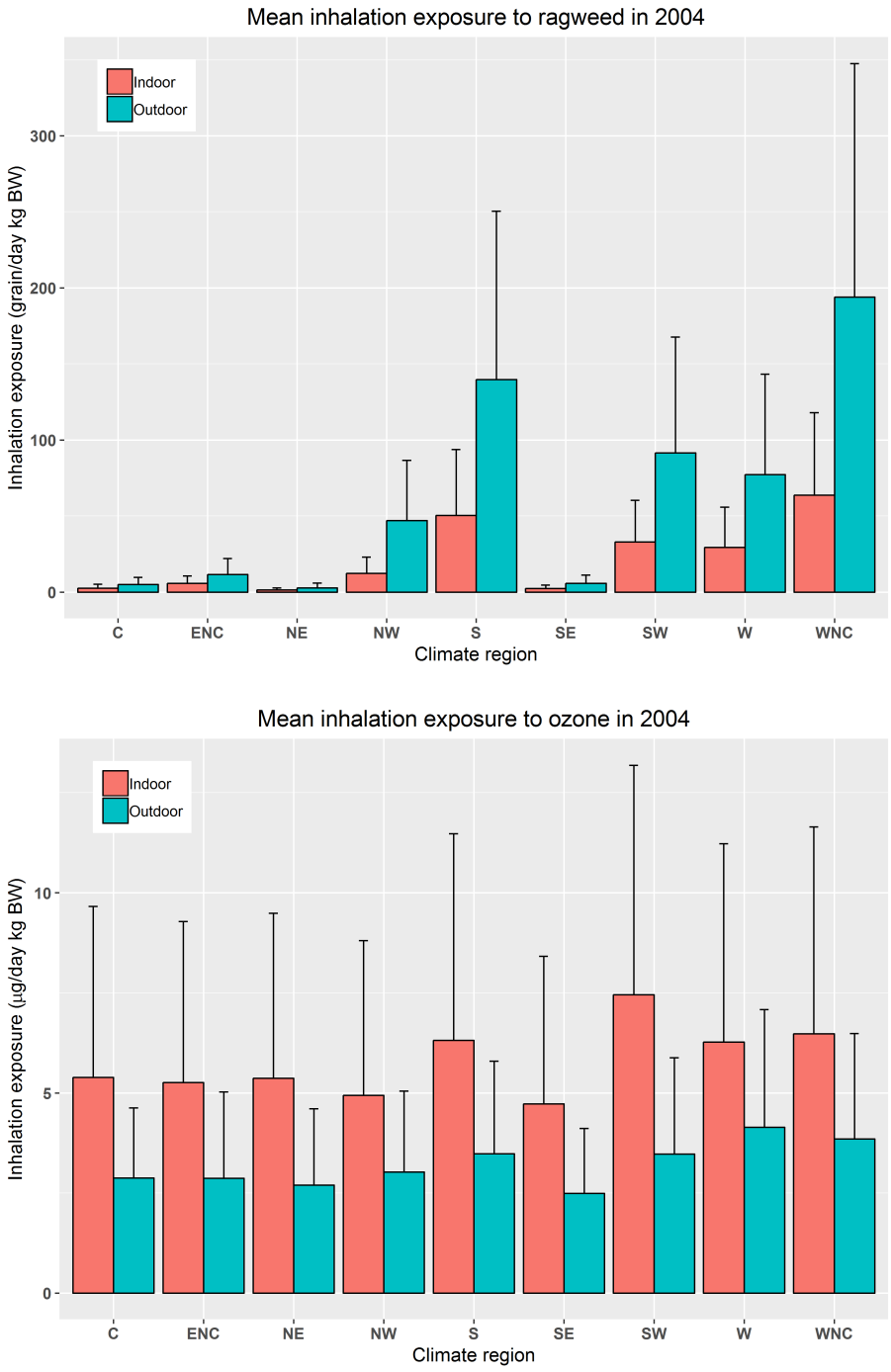


Figure 3.18: Mean daily inhalation intakes of ragweed pollen (top figure) and ozone (bottom figure) in indoor and outdoor environments in the nine climate regions in August 2004 and September 2004.

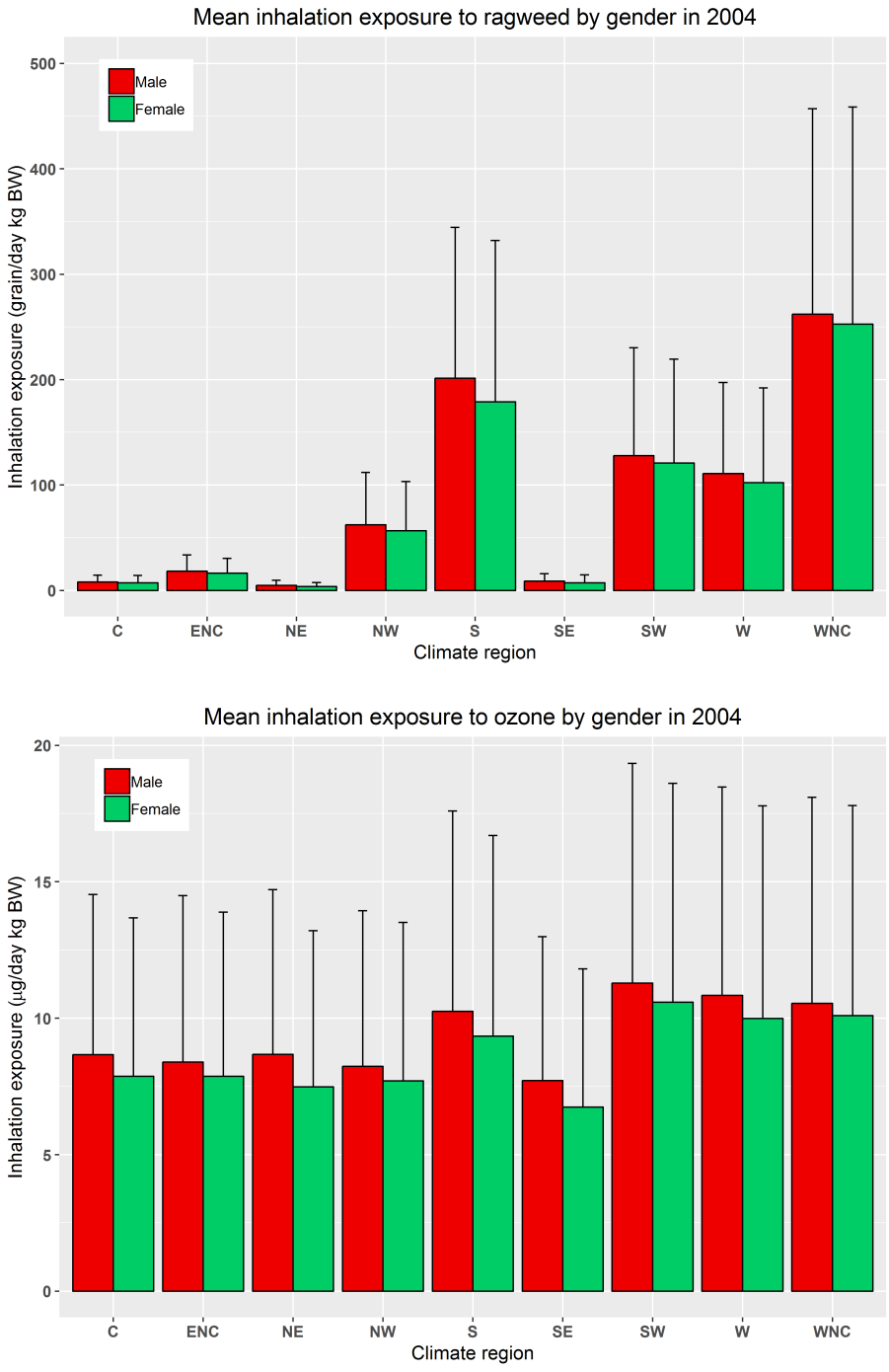


Figure 3.19: Mean daily inhalation intakes of ragweed pollen (top figure) and ozone (bottom figure) by gender in 2004.

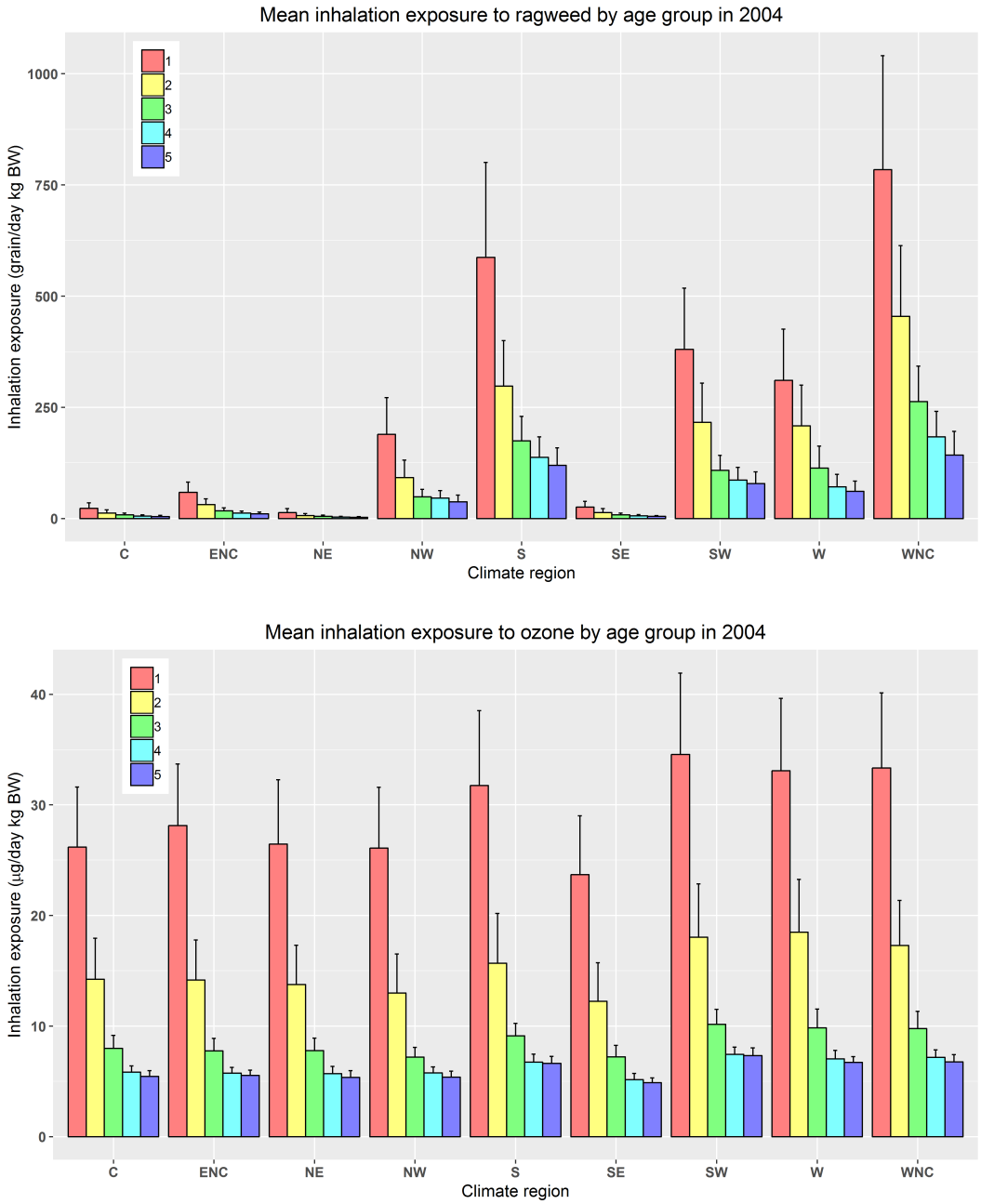


Figure 3.20: Mean daily inhalation intakes of ragweed pollen (top figure) and ozone (bottom figure) by age group in 2004. Age group 1: 1-4 years old, Age group 2: 5-11 years old, Age group 3: 12-17 years old, Age group 4: 18-64 years old, Age group 5: >64 years old.

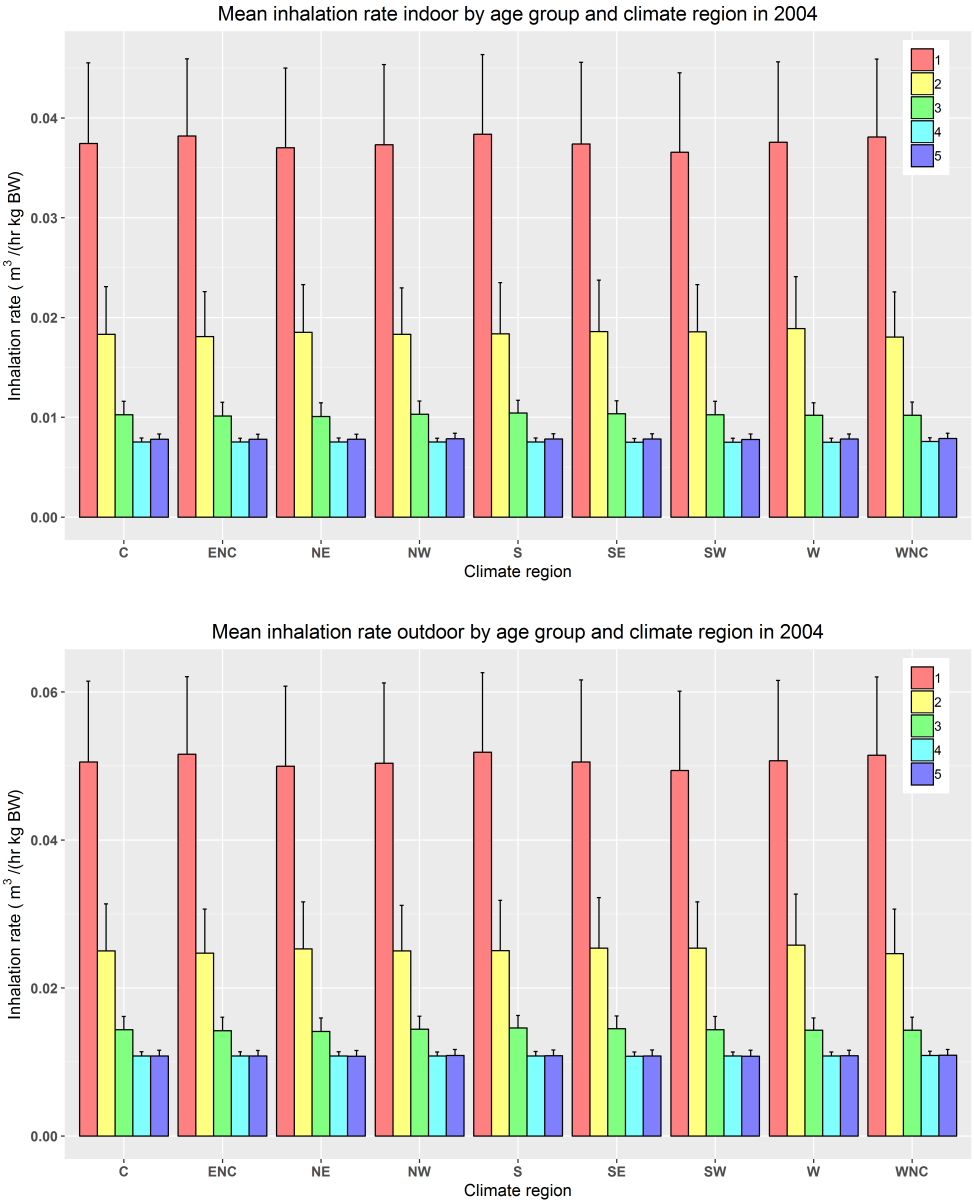


Figure 3.21: Mean inhalation rate indoors and outdoors by age group.

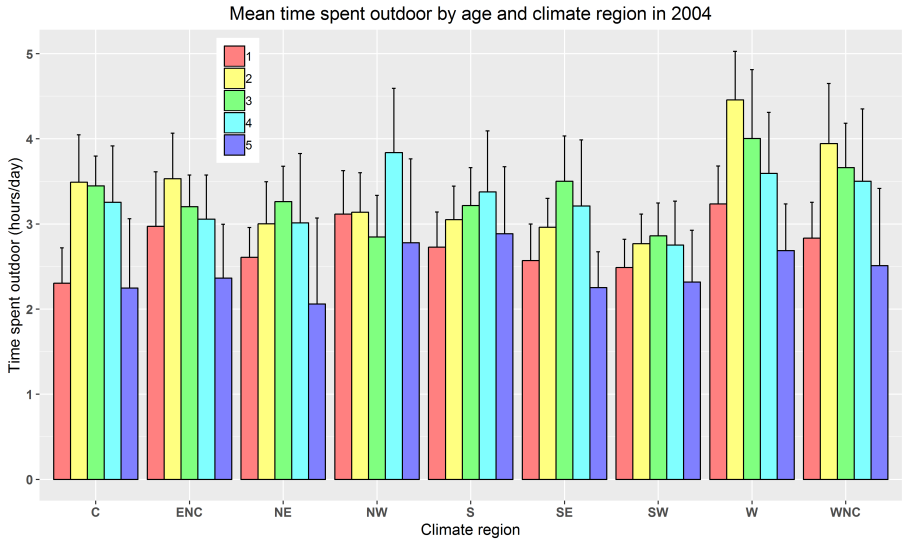


Figure 3.22: Mean time spent outdoors by age group.

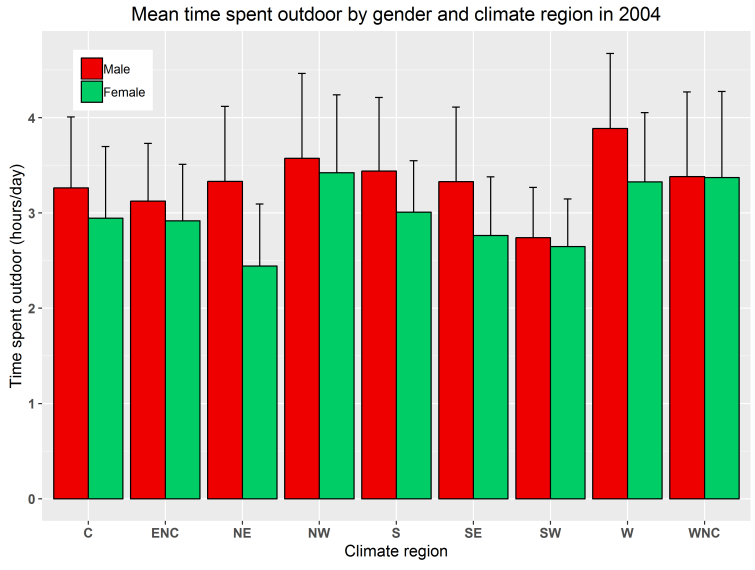


Figure 3.23: Mean time spent outdoors by gender.

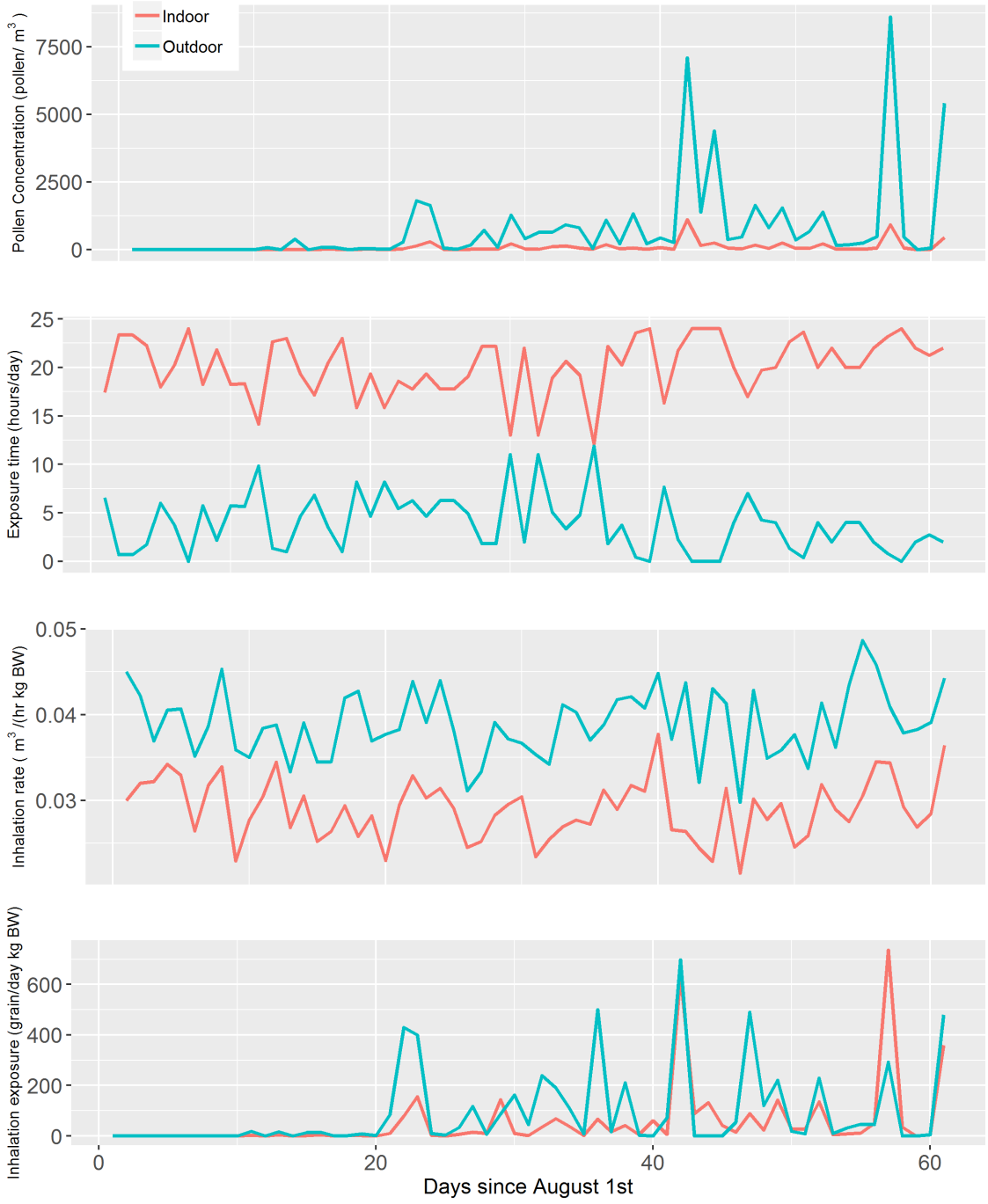


Figure 3.24: Time series of daily ragweed pollen concentration, exposure time, inhalation rate, and inhalation intakes of ragweed pollen. The simulated virtual subject is a 3 years old male in the West region.

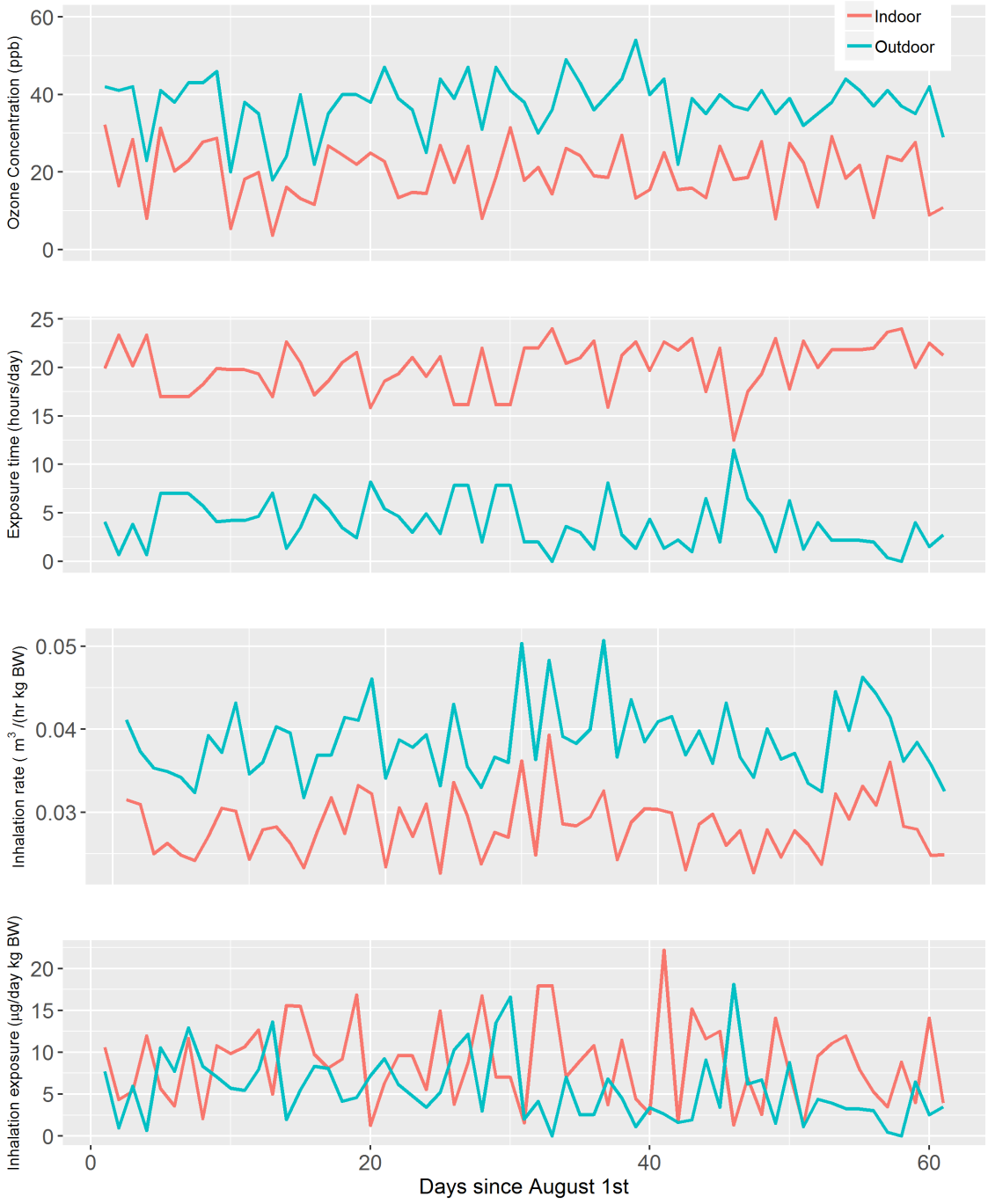


Figure 3.25: Time series of daily ozone concentration, exposure time, inhalation rate, and inhalation intakes of ozone. The simulated virtual subject is a 3 years old male in the West region.

3.5 Summary

The changes in ragweed pollen indices during 1994 to 2010 were examined based on observation data. Ragweed pollen season tends to start earlier for the Northern CONUS and shifts to the Southern CONUS. Ragweed pollen season lasts longer in the Southern CONUS. The peak pollen concentration and annual pollen production decrease from lower latitudes to higher latitudes. Correlation analysis reveals that start date is significantly affected by climate factors such as maximum temperature, precipitation, wind speed, while season length is mainly affect by precipitation, mean temperature, maximum temperature and minimum temperature.

The simulated ragweed pollen and ozone concentrations in 2047 and 2004 were compared and it is predicted that the Southwest and West regions will experience simultaneous increases in ragweed pollen and ozone concentrations in 2047, while the Central, East North Central and West North Central regions will have decreased ragweed pollen and ozone concentrations. Although the co-occurrence of ragweed pollen and ozone exceedance scatters across the CONUS, it influences a remarkable fraction of population. Five of the 10 largest cities by population are affected by the co-occurrence of ragweed pollen and ozone exceedance in 2004, and one more city will be added to the list in 2047.

The simulated exposures to ragweed pollen and ozone have distinct patterns by microenvironment, gender and population age group. Exposures to ragweed pollen outdoors is higher than indoors, with significant correlation with pollen concentration. Males tend to have higher inhalation exposures to ragweed pollen and ozone than females. The inhalation exposure to ragweed pollen and ozone per unit body weight decreases with age. Individual exposures to both pollutants follow the main patterns of the pollutant concentration during the whole simulation period.

Chapter 4

PREDICTING RAGWEED POLLEN CONCENTRATION USING MACHINE LEARNING METHODS

4.1 Abstract

Prediction of ragweed pollen concentration based on meteorological factors and previous ragweed pollen observation was conducted using Machine Learning (ML) models that included Support Vector Machine, Random Forest, XGBoost, Neural Network, Decision Trees, etc. Two types of prediction models were applied: regression models and classification models, which predict ragweed pollen concentration (pollen grains/m^3) and pollen levels (low, medium, high), respectively. The model parameters have been optimized and the final models were validated using a repeated 10-fold cross-validation. The performances of regression models were evaluated based on coefficient of determination (R^2), Root Mean Square Error ($RMSE$) and Mean Absolute Error (MAE), while classification models were evaluated based on accuracy and $F1$ score. The variable importance of predictors in the best models was calculated and compared. Random Forest and XGBoost outperformed other models for both regression and classification. The XGBoost model had the lowest $RMSE$ (22.99 ± 7.48) and shared the same R^2 (0.58 ± 0.12) with Random Forest model in regression. The accuracy and $F1$ score of the XGBoost model were 0.77 ± 0.05 and 0.90 ± 0.03 in classification, which are similar to Random Forest. Pollen concentration of the previous day is the most important predictor variable in both Random Forest and XGBoost models.

4.2 Introduction

Modeling and forecasting of pollen concentration a few days ahead could help individuals avoid exposures or take preventive medicine to reduce the adverse health effects of allergenic pollen [164]. Numerical and statistical forecasting models are the two most popular models [81, 165, 84, 76]. Numerical models are capable of predicting pollen concentrations for large areas based on mathematical equations, plant distributions, phenological, aerobiological, and meteorological data [81, 79]. Statistical models are mostly local-scale and do not depend on the knowledge of physical processes of pollen emission and dispersion, but aim to predict pollen concentrations based on numerical relations between pollen and independent variables [113]. The relations can be built using various statistical methods that include Multiple Regression [116, 166], Discriminant Linear Analysis [129], Random Forest [114, 115], Artificial Neural Networks [128, 113], gaussian, gamma and logistic distribution models [167], etc. There are quite a few requirements regarding the analyzed data when applying some of these statistical models. For example, (muti-)linear regressions assume linear relationships between the outcome variable and the independent variables, residuals are normally distributed, and the independent variables are not highly correlated with each other [168]. Log-transform of pollen concentrations is usually needed to meet the normal distribution requirements [169, 170].

ML techniques have drawn increasing interest in pollen concentration prediction in recent years due to their powerful algorithms and fewer restrictions regarding the input data. ML methods are able to find patterns in nonlinear and high dimensional data, and make predictions with relatively high accuracy [171]. ML models work well in both regression and classification problems. Some of the representative ML methods include nonlinear models such as Support Vector Machine (SVM), Artificial Neural Networks (ANN), K-Nearest neighbors, and tree-based models such as Random Forest, boosting, bagged trees [172]. Studies of pollen concentration prediction using

ML methods have been conducted for some pollen species. Castellano-Mendez *et al.* [113] used Neural Networks to forecast birch (Betula) pollen levels in a city of northwest Spain based on three input variables: daily pollen concentration, daily rainfall, and daily mean temperature. Puc [128] also applied Neural Networks to model daily birch pollen concentrations with meteorological data in Szczecin (Poland). Nowosad [114] predicted pollen levels of Corylus, Alnus, and Betula in Poland using Random Forest based on gridded meteorological data. Nowosad *et al.* [115] compared several modeling techniques in linear regression models, nonlinear regression models, and regression trees and rule-based models for prediction of Corylus, Alnus, and Betula pollen concentrations in nine cities in Poland. Mesa *et al.* [129] found that Neural Networks performed better than linear models for forecasting the severity of the Poaceae pollen season in a region with a typical Mediterranean climate. However, there is no study focusing on a comparison of different ML techniques for ragweed (Ambrosia) pollen prediction in the CONUS.

The main goals of this study were to compare selected ML modeling techniques for regression analysis of ragweed pollen concentration and classification of ragweed pollen levels and to assess the variable importance for the best models. Six ML techniques were used to predict pollen concentration and five were used in pollen level classification.

4.3 Methods

4.3.1 Study area and predictor variables

The pollen monitoring station located in Newark, New Jersey (coordinates: 40.74°N, 74.19°W) was chosen for this study. There are in total 12 years of observed ragweed pollen data from the NAB. This area has an elevation of 43 m, annual mean temperature of 13.0 °C, and annual mean precipitation of 1,213 mm. Twelve meteorological parameters from the same day as pollen concentration and the pollen

concentration/level of previous day were used as independent variables (Table 4.1), and the daily pollen concentration was used as the dependent variable in the regression model, while the daily pollen level was used as the dependent variable in the classification model. The meteorological factors (except cumulative temperature and cumulative precipitation) are retrieved from the nearest meteorology station (coordinates: 40.68°N, 74.17°W) from NOAA [173]. The distance between the pollen monitoring station and the meteorology station is 6.4 km. The pollen concentration on the previous day will have lag effects on the pollen concentration on the next day, therefore, it is added into the model as an input variable.

Table 4.1: Input variables for the ML models.

Predictor variable name	Abbreviation	Unit
Mean temperature for the day	TEMP	ºC
Mean dew point for the day	DEWP	ºC
Mean station pressure for the day	STP	Millibars
Mean visibility for the day	VISIB	Miles
Mean wind speed for the day	WDSP	Knots
Maximum sustained wind speed for the day	MXSPD	Knots
Maximum temperature during the day	MaxTemp	ºC
Minimum temperature during the day	MinTemp	ºC
Total precipitation during the day	PRCP	Inches
Mean relative humidity for the day	RH	%
Cumulative temperature since the first day of year	CumTemp	ºC
Cumulative precipitation since first day of year	CumPRCP	Inches
Mean pollen concentration/level on previous day	PollenDay_1	Pollen grains/m^3

In the classification models, the dependent variable is pollen level of the day. In most of the ambient air quality monitoring systems, pollen levels are reported as absent, low, moderate, high or very high [174, 175]. In this study, the ragweed pollen

levels are classified into four categories based on threshold values as below,

$$Pollen\ Level = \begin{cases} 1\ (\text{Low}),\ \text{if}\ C < 10 \\ 2\ (\text{Medium}),\ \text{if}\ 10 \leq C < 30 \\ 3\ (\text{High}),\ \text{if}\ C \geq 30 \end{cases} \tag{4.1}$$

where C is the daily ragweed pollen concentration (pollen grain/m^3). Ragweed pollen concentration below 10 pollen grain/m^3 is considered low level by NAB [174]. The threshold concentration, 30 pollen grain/m^3 was chosen based on the health effects of ragweed pollen in sensitive patients, at which concentration they started to have allergy symptoms [80, 144, 145].

4.3.2 ML models

Six ML techniques were used for prediction of ragweed pollen concentrations:

- Support Vector Machine (SVM) [176]
- Random Forest [177]
- eXtreme Gradient Boosting (XGBoost) [178, 179]
- Bayesian Generalized Linear Model (BayesGLM) [180]
- Neural Networks [181]
- Classification and Regression Tree (CART) [124]

Five of these methods (except BayesGLM) are also used in the pollen level classification models. All the statistical analyses are conducted using R version 3.5.2 [159] and R packages caret, ggplot2, e1071, doSNOW, etc. [182, 183, 184, 185, 186]. Models were built using the following methods in caret: Support Vector Machines with Linear Kernel (method = 'svmLinear') [187], Random Forest (method = 'rf') [188], eXtreme Gradient Boosting (method = 'xgbTree') [189], Bayesian Generalized

Linear Model (method = 'bayesglm') [190], Neural Network (method = 'nnet') [191], CART (method = 'rpart') [192].

4.3.3 Workflow of modeling tasks

The main goal of this study is to evaluate ML methods for the prediction of ragweed pollen concentration/level in Newark, NJ. The workflow is as following:

(1) Ten meteorological variables are retrieved from the nearest meteorology station. In addition, cumulative temperature and cumulative precipitation since January 1st to the observation date in each year are calculated.

(2) Pollen concentrations on each day are transformed into pollen levels based on Equation 4.1 for classification models. The dependent variable (pollen concentration/pollen level) and independent variables are combined into one dataset.

(3) Different ML techniques are applied to predict ragweed pollen concentration/level. Models are validated using a repeated (three times) 10-fold cross validation procedure [172].

(4) The ML models are compared based on predictive performance. Regression models are evaluated using the coefficient of determination (R^2), Root Mean Square Error ($RMSE$) and Mean Absolute Error (MAE), while classification models will be evaluated based on accuracy and $F1$ score.

Figure 4.1 presents the processes for ML.

An R^2 value is the squared correlation coefficient between the observed and predicted value. It ranges between 0 and 1. $RMSE$ is the standard deviation of the prediction errors, namely residuals. $RMSE$ is a measure of how spread out these residuals are and tells us how concentrated the data are around the line of best fit [193]. MAE is the average of the difference between the original values and the

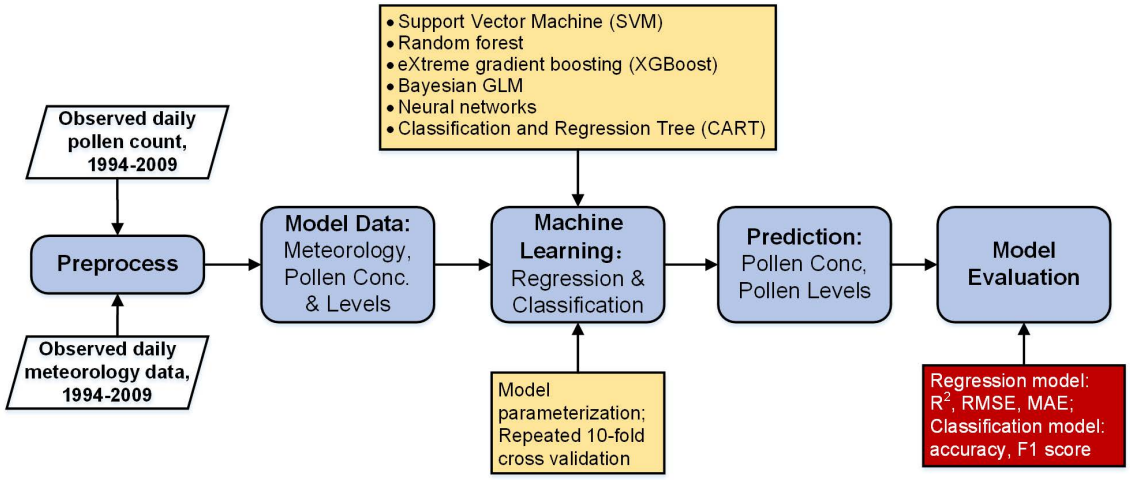


Figure 4.1: The schematic illustration of ML modeling system.

predicted values as represented in Equation 4.2.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{4.2}$$

The definition of accuracy, precision P, recall R and $F1$ score is based on confusion matrix (Table 4.2) [194, 195]:

Table 4.2: Confusion matrix for pollen level classification models.

Level *i*		Actual level	
		Positive	Negative
Predicted level	Positive	True Positive (TP_i)	False Negative (FN_i)
	Negative	False Positive (FP_i)	True Negative (TN_i)

$$\begin{cases} Accuracy = \frac{\sum_{i=1}^{|C|} TP_i + TN_i}{\sum_{i=1}^{|C|} TP_i + FP_i + TP_i + FN_i} \\ P = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{TP_i}{TP_i + FP_i} \\ R = \frac{1}{|C|}\sum_{i=1}^{|C|}\frac{TP_i}{TP_i + FN_i} \\ F1 = \frac{2PR}{P+R} \end{cases} \tag{4.3}$$

where $|C| = 3$ is the number of classes of pollen levels.

4.4 Results and Discussion

4.4.1 Description of observed ragweed pollen concentration

Figure 4.2 presents the distribution of ragweed pollen concentration in Newark between 1994 to 2009 except 1998, 1999, 2000, and 2001. The pollen concentration (pollen grain/m^3) are mostly between 0-100. The pollen concentrations in 1994 and 1995 are higher than the other years. The cumulative pollen count for each year is displayed in Figure 4.3. The pollen season start date is defined as the the day when the cumulative pollen count reaches 5% of annual total count, and the end date is the day when the cumulative pollen count reaches 95% of annual total count. The pollen season started around the 230th day (green dots in Figure 4.3) in each year with a standard deviation of 2.4 days, and ended around the 272th day (red dots in Figure 4.3) every year with a standard deviation of 6.4 days. The pollen season length, which is the duration between start date and end date, has an average of 41 days in Newark with a standard deviation of 5.6 days.

Ragweed pollen concentration is transformed into pollen levels and displayed in Figure 4.4. Pollen levels are distributed predominantly at low level for years 2002, 2003, 2005 and 2008, while in other years, the pollen levels are almost evenly distributed at low, medium and high levels.

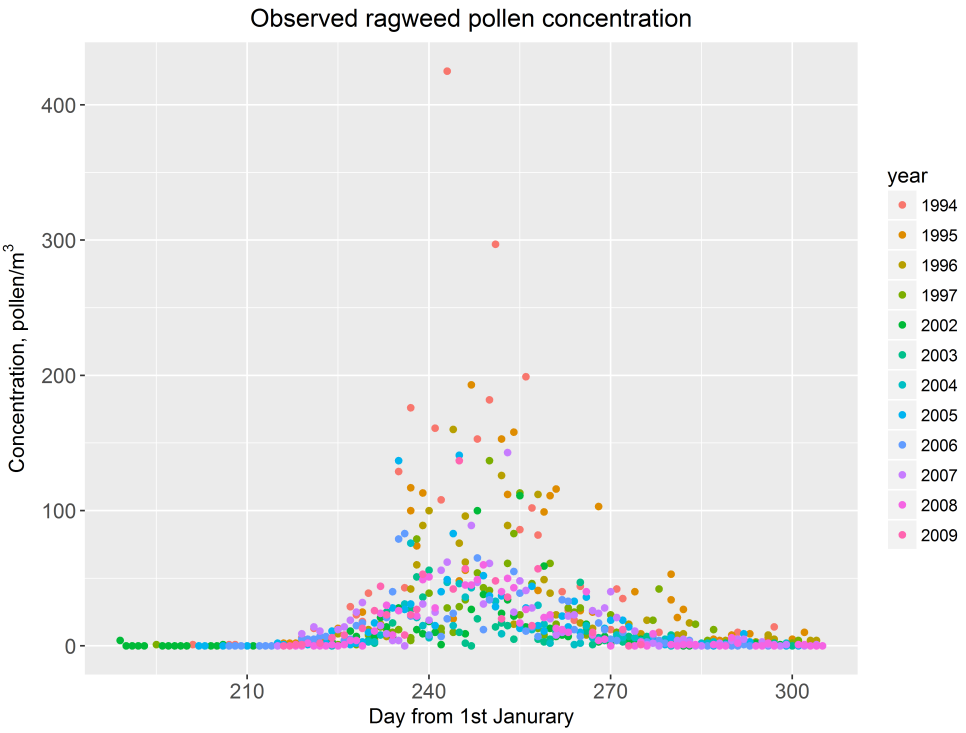


Figure 4.2: Observed ragweed pollen concentration in Newark, NJ during 1994-2009.

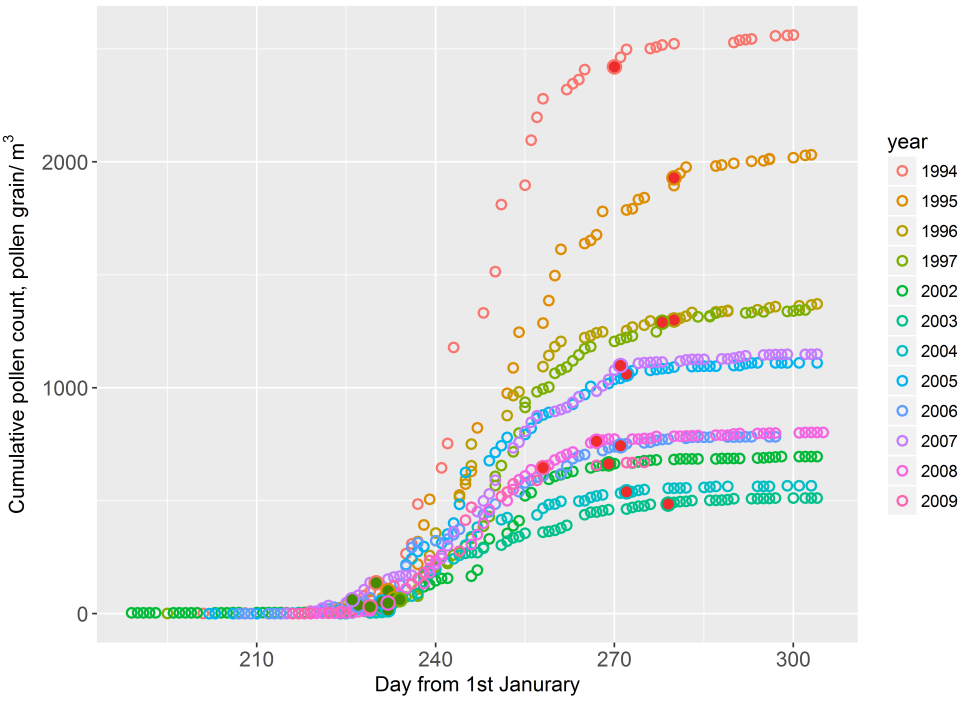


Figure 4.3: Cumulative observed ragweed pollen concentration in Newark, NJ during 1994-2009. The green dots indicate the start dates of ragweed pollen season, the red dots indicates the end dates.

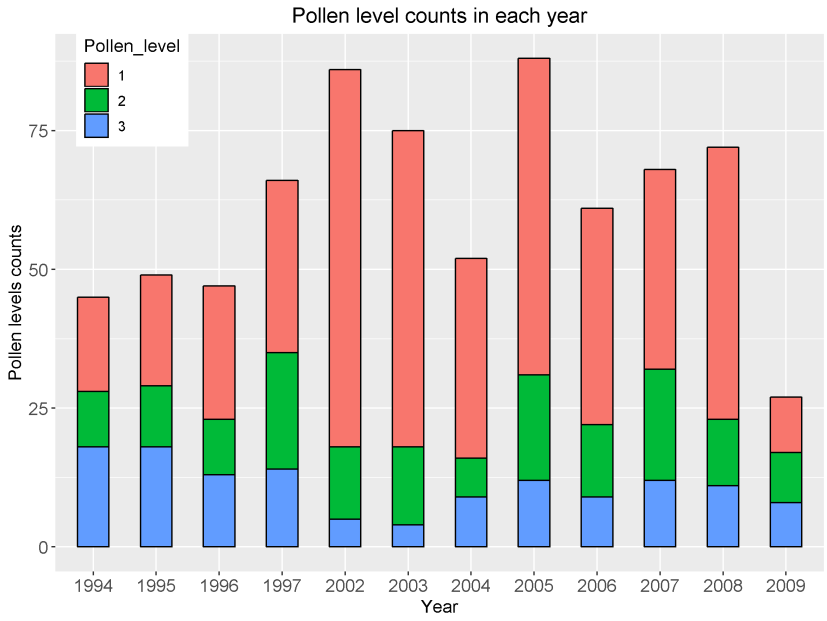


Figure 4.4: The cumulative number of days with pollen level in each year.

4.4.2 Correlation between pollen concentration and predictor variables

The correlation between pollen concentration and meteorological factors is illustrated in Figure 4.5. The pollen concentration is significantly (p-value < 0.05) correlated with mean temperature, mean dew point, mean visibility, maximum temperature, minimum temperature, relative humidity and cumulative temperature of the day.

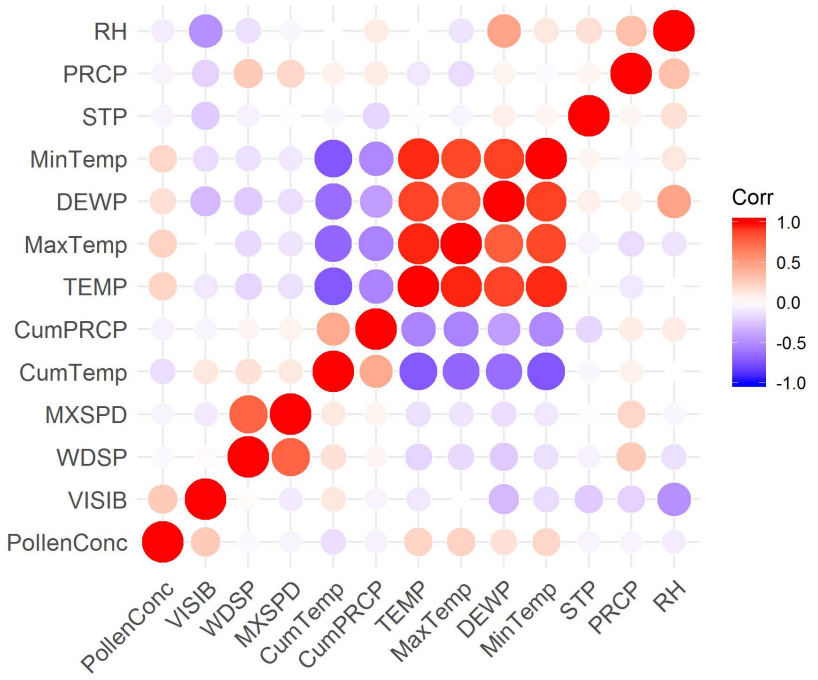


Figure 4.5: Pearson correlation heat map (with hierarchical clustering) for ragweed pollen concentration in Newark, NJ and 12 meteorological factors.

4.4.3 Performance of regression models

The performances of six regression models for estimation of ragweed pollen concentration were compared using $RMSE$, R^2 and MAE in Figure 4.6. The XGBoost model had the lowest $RMSE$ and MAE. The Random Forest model gave similar R^2 (0.58 ± 0.11) to the XGBoost model, which is the highest among the six models. Figure 4.7 displays scatterplots of observed and predicted daily ragweed pollen concentration during 1994-2009 in Newark, NJ for the six regression models. The Spearman coefficients (Table D.2) ranged from 0.62 with Neural Network model to 0.86 with Random Forest. Some extremely high ragweed pollen concentrations are under-predicted, but they might be outliers in the observation data. Overall, all the models capture the distribution of observed ragweed pollen concentration in Newark, NJ.

All the models were cross-validated and their parameters were optimized as shown in Figure 4.8 to Figure 4.12. The performances of the models are affected significantly by the setting of parameters. The Random Forest model gave the lowest $RMSE$ when there are three randomly selected predictors in the model. The number of boosting iterations (nrounds) has significant effects on the performance of the XGBoost model. The XGBoost model had best fit when the maximum tree depth is 4 and the shrinkage (eta) is 0.025. The structure of the boosting tree and the decision tree (CART) are shown in Figure 4.13 and Figure 4.14.

Both Random Forest model and XGBoost model perform better than the other models based on R^2 and $RMSE$. Therefore, these two models were selected as the best two models for further analysis. The variable importance for each of them was estimated and presented in Figure 4.15. Pollen concentration of the previous day is the most important variable in both models. The other variables have distinctly lower importance. In the Random Forest model, cumulative temperature is the second most important variable, followed by cumulative precipitation and mean daily visibility. Cumulative precipitation is the second important variable in the XGBoost model. Mean station pressure and mean daily precipitation for the day are the least important variables in both models. The variables with higher importance part a bigger prediction power in the model. Nowosad *et al.* [115] also found that Random Forest model had the best overall performance for all the pollen species they tested. But the cumulative growing degree days (GDD) and daily maximum temperature were the most important predictor variables in their model.

Figure 4.16 shows a comparison between observed and predicted time series of daily pollen concentrations during ragweed pollen season in 1995 and 1996 in Newark, New Jersey using Random Forest and XGBoost models. Both models captured the trend of observed ragweed pollen very well, with a few days of underestimation.

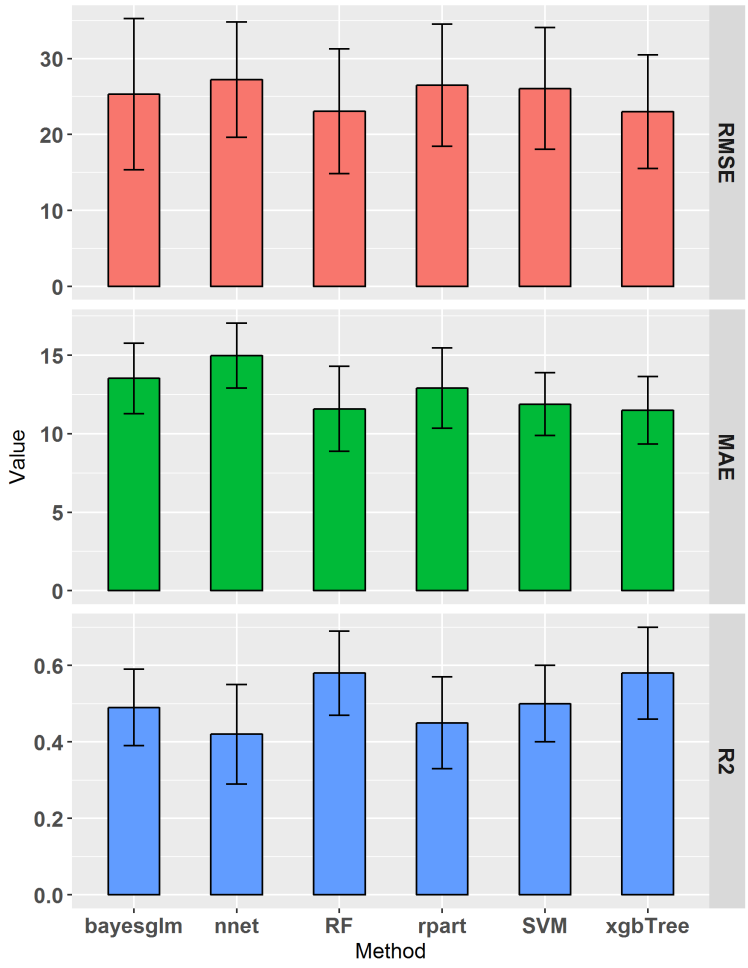


Figure 4.6: The performance metrics (mean ± standard deviation) of ML models on estimates of daily ragweed pollen concentration.

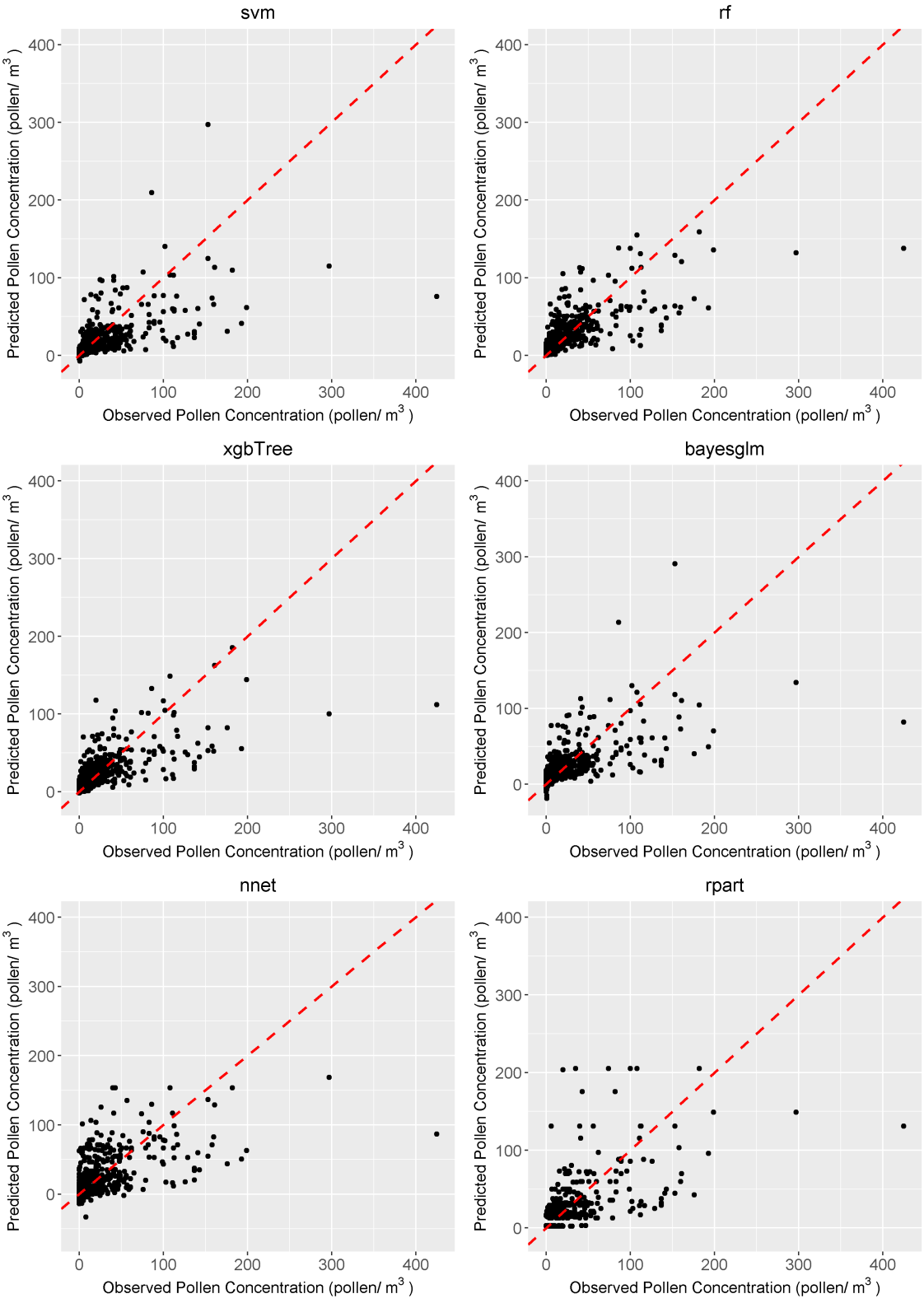


Figure 4.7: Scatterplots of observed and predicted daily ragweed pollen concentration with 45-degree line using six models: SVM, Random Forest, XGBoost, BayesGLM, Neural Network and CART.

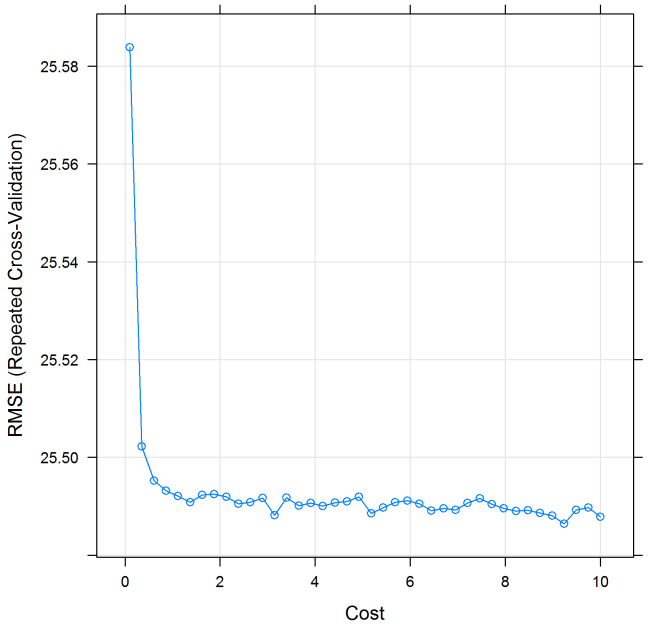


Figure 4.8: Cross-validated $RMSE$ profile for the SVM model. The optimal model parameter is Cost = 9.24.

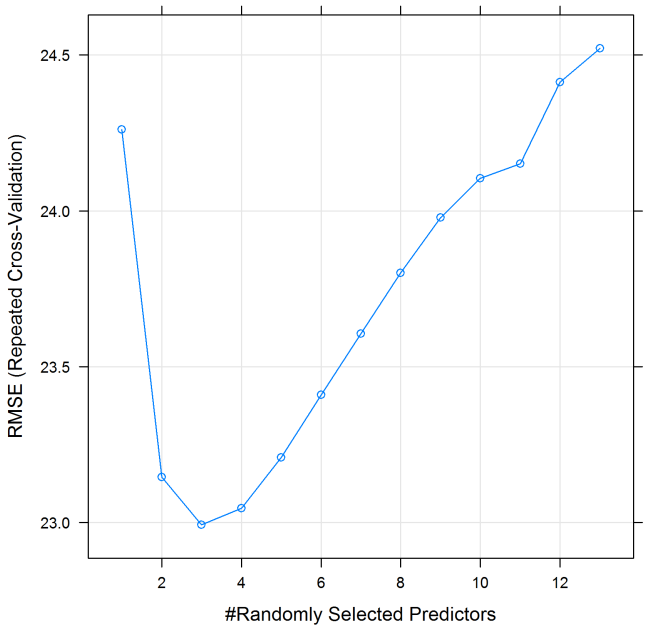


Figure 4.9: Cross-validated $RMSE$ profile for the Random Forest model. The final model was fit with mtry = 3.

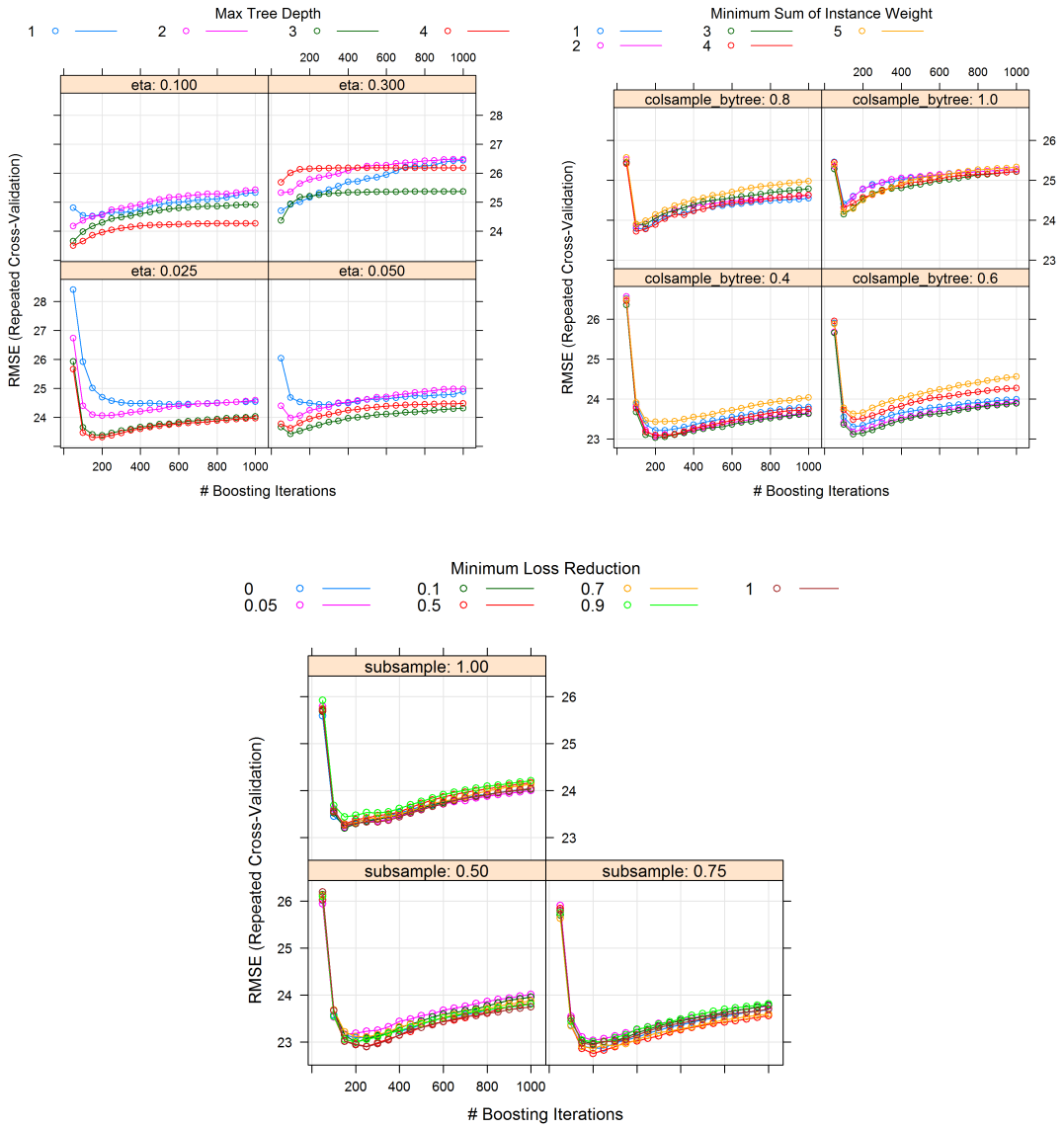


Figure 4.10: Cross-validated $RMSE$ profile for the XGBoost model. The final model parameters are listed in Table D.1.

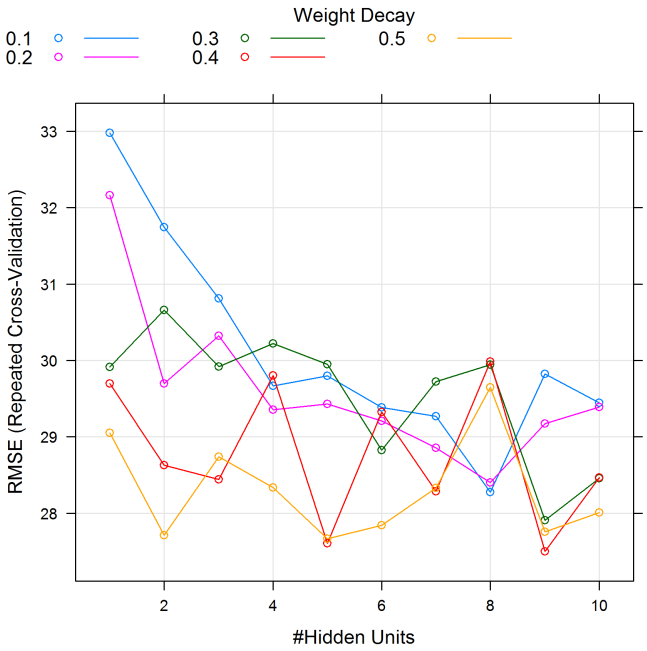


Figure 4.11: Cross-validated $RMSE$ profile for the Neural Network model. The optimal model used decay of 0.5 and 9 hidden units.

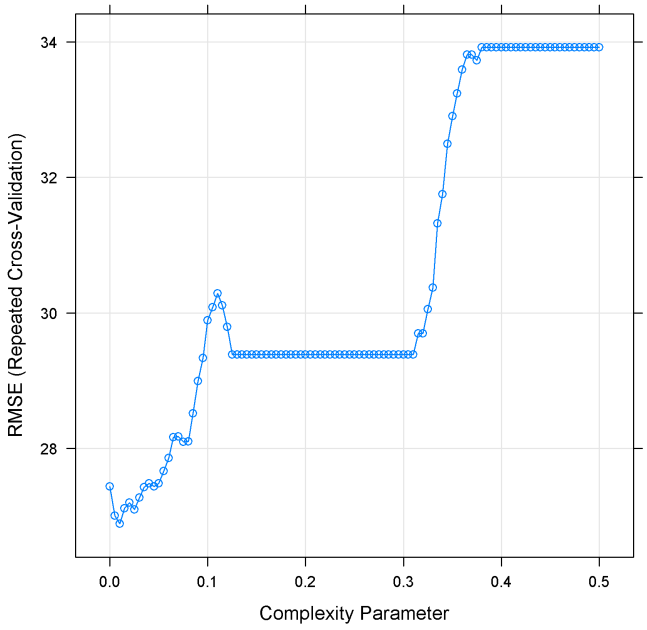


Figure 4.12: Cross-validated $RMSE$ profile for the CART model. The complexity parameter (cp) of the final model is 0.01.

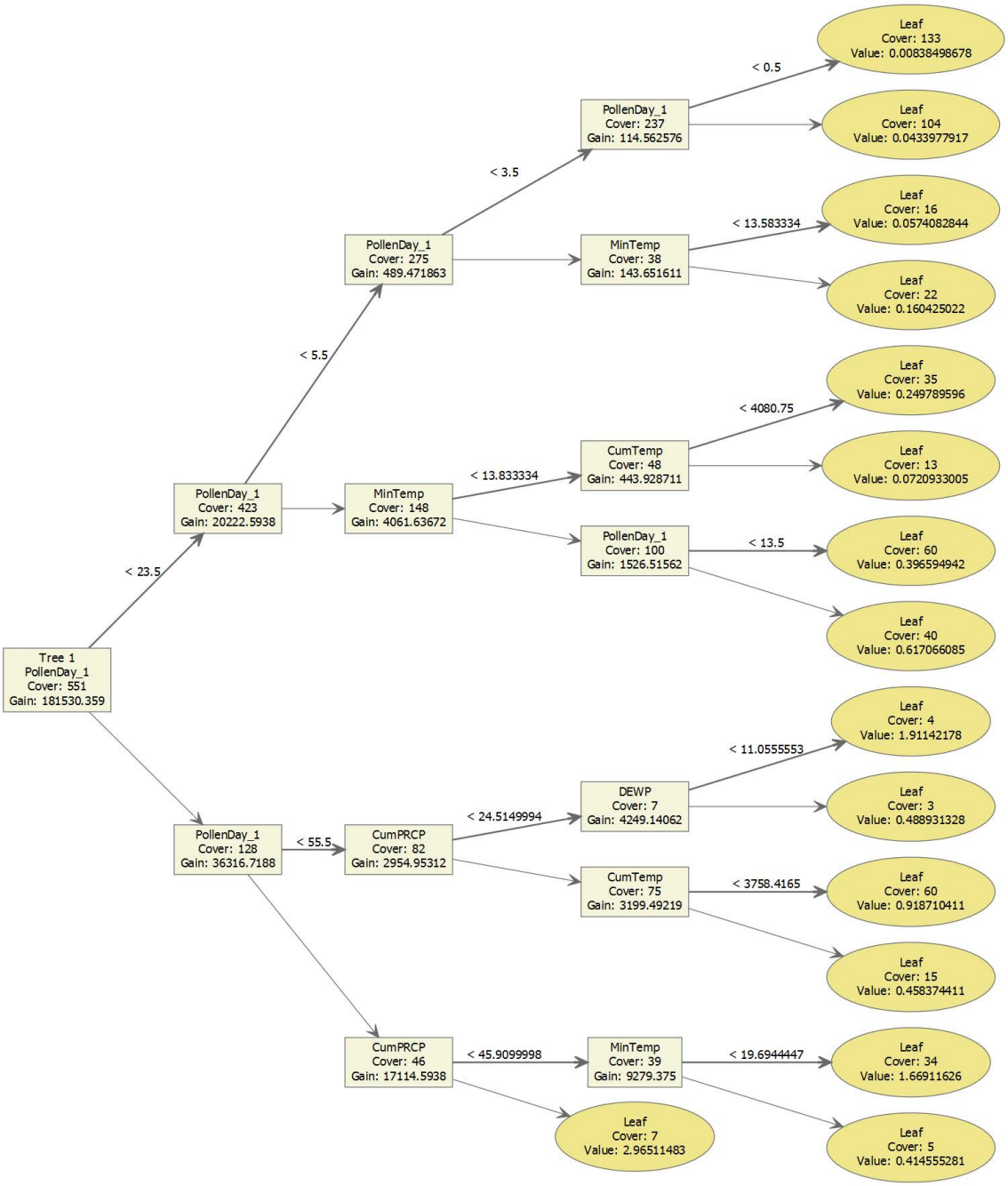


Figure 4.13: Structure of the boosting tree for prediction of ragweed pollen concentrations. The nodes represent the conditions and input variables, the leaves represent ragweed pollen concentrations.

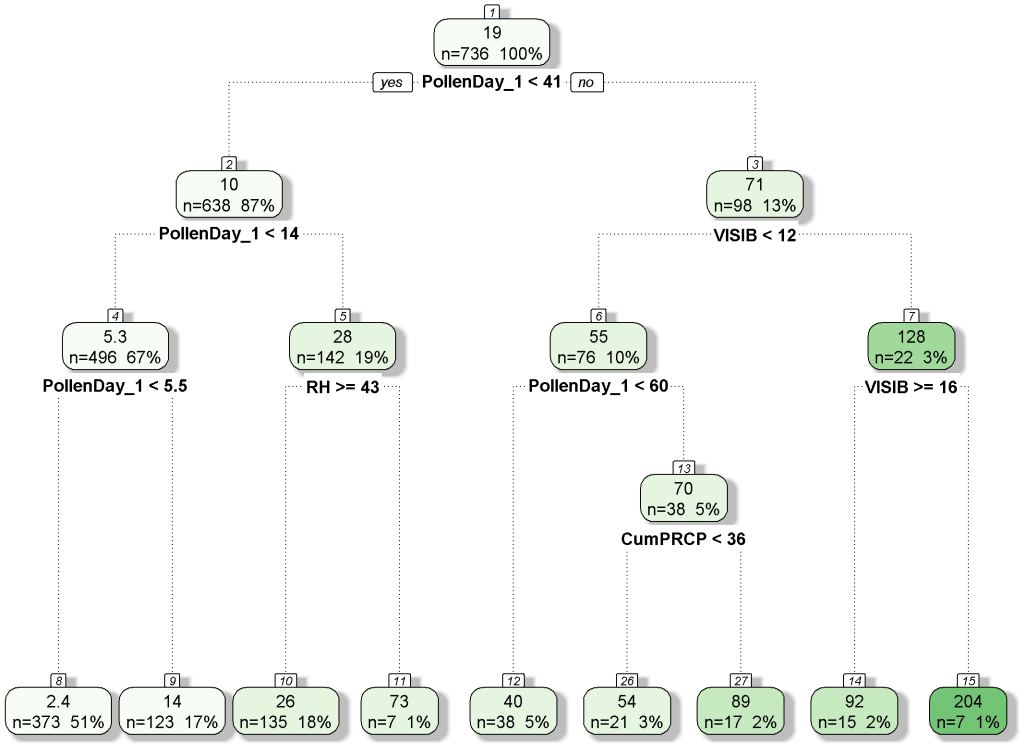


Figure 4.14: Structure of the Decision Tree for prediction of ragweed pollen concentrations. The nodes represent the conditions and input variables, the leaves represent ragweed pollen concentrations.

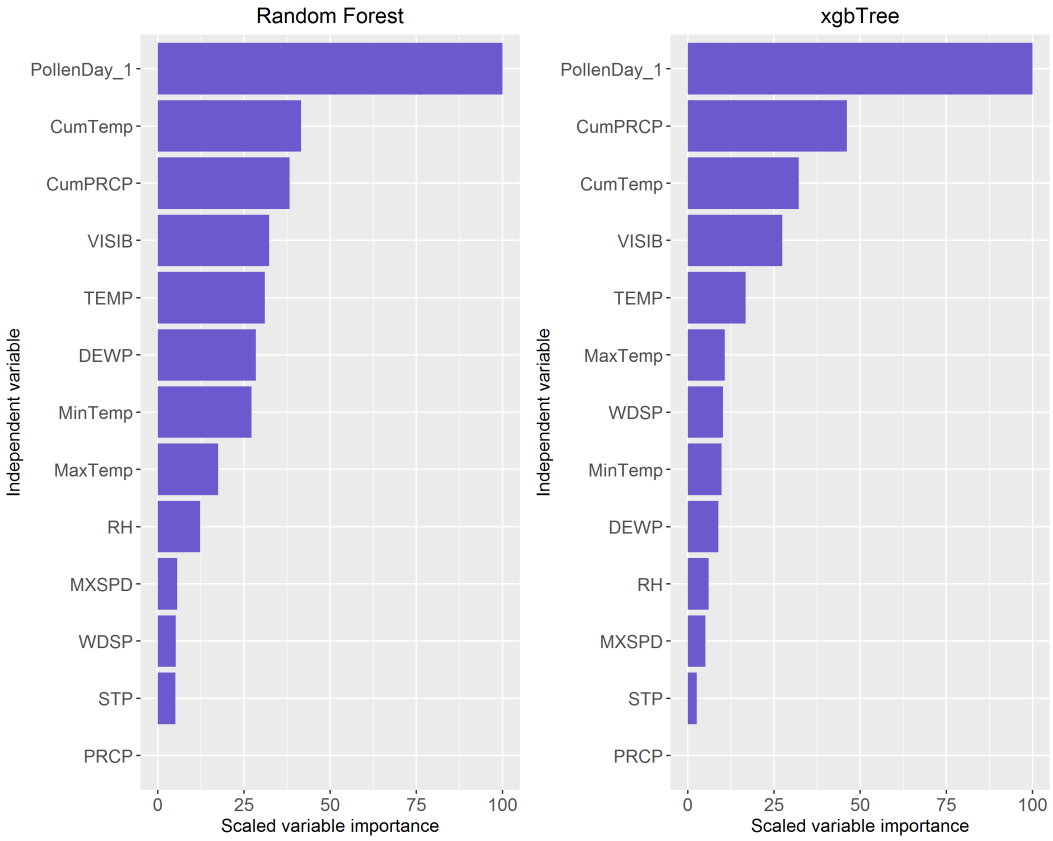


Figure 4.15: Importance of each predictor variable in the Random Forest model and the XGBoost model for ragweed pollen concentration estimation in Newark, NJ.

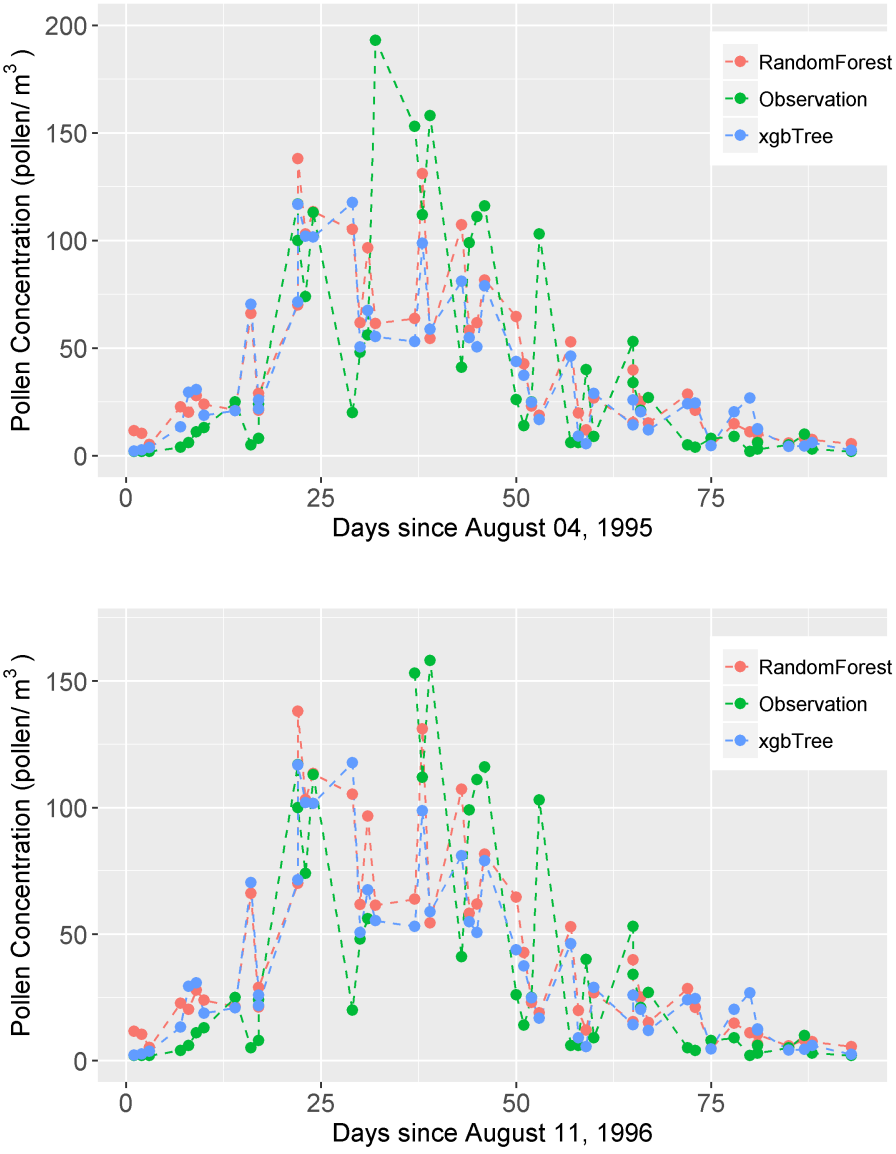


Figure 4.16: Simulated and observed time series plots of pollen concentration in Newark, NJ in 1995 and 1996.

4.4.4 Performance of classification models

Five ML models including SVM, Random Forest, XGBoost, Neural Network and CART were tested for classification of pollen levels and their performances are summarized in Figure 4.4.4. The XGBoost model gave the best performance based on both accuracy (0.7740) and $F1$ score (0.8973). The Random Forest model has very close performance to the XGBoost model. SVM and CART had similar accuracy and $F1$ score. The Neural Network did not perform well for this case. All the models have been optimized and the profiles of the cross-validated accuracy are shown in Figure 4.18 to Figure 4.22. The method used in the XGBoost model is xgbTree in R and it has the highest number of parameters and its tuning is a little more complicated than for the other models. First, Boosting Iterations (nrounds), Max Tree Depth (max_depth) and Shrinkage (eta) were tuned with the other parameters fixed and the optimal values were obtained. Then the three parameters were fixed with their optimal values and Subsample Ratio of Columns (colsample_bytree) and Minimum Sum of Instance Weight (min_child_weight) were tuned. Lastly, Minimum Loss Reduction (gamma) and Subsample Percentage (subsample) were optimized. All the final parameters for each model is listed in Table D.3.

Figure 4.23 presents the model structure of a Neural Network for prediction of ragweed pollen levels. The first layer is the input layer consisting of all the thirteen input variables in Table 4.1. The second layer is the hidden layer of 9 units, and the third layer is the output layer consisting of three pollen levels. The model structure for CART is shown in Figure 4.24.

To find out which variables contribute most to the XGBoost model and the Random Forest model, the scaled variable importance is compared for each variable in Figure 4.25. Pollen concentration on previous day is the predominant contributor in both models, followed by cumulative temperature. Maximum wind speed and precipitation is negligible in the XGBoost model, while mean wind speed and cumulative precipitation had least impacts in the Random Forest model.

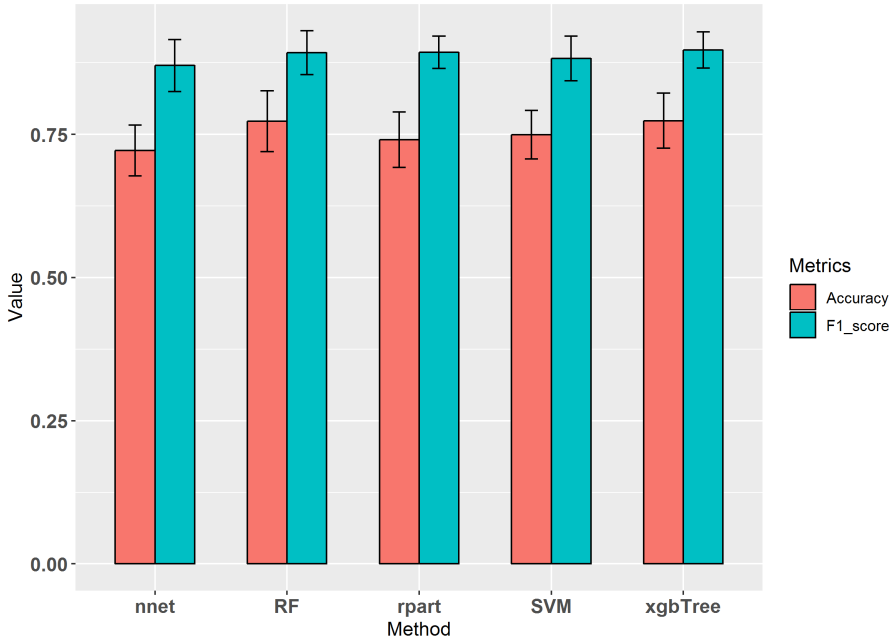


Figure 4.17: The performance metrics (mean ± standard deviation) of ML models on estimates of daily ragweed pollen levels.

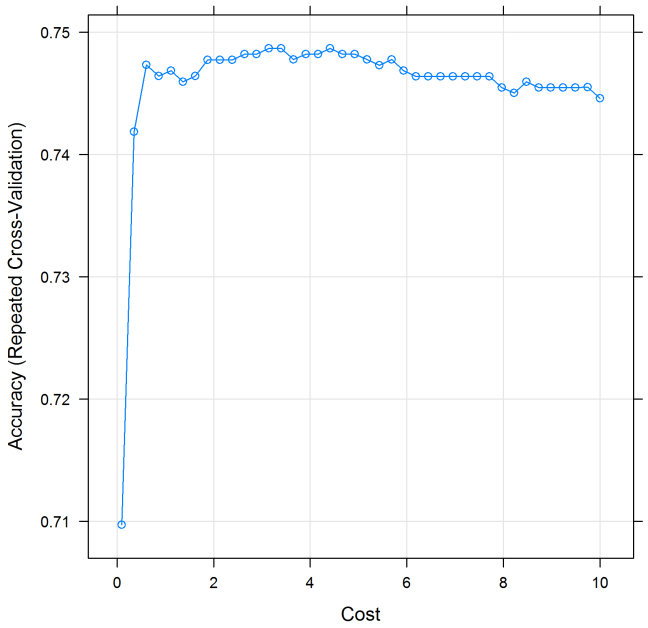


Figure 4.18: Cross-validated accuracy profile for the SVM model. The optimal model parameter is Cost = 4.4.

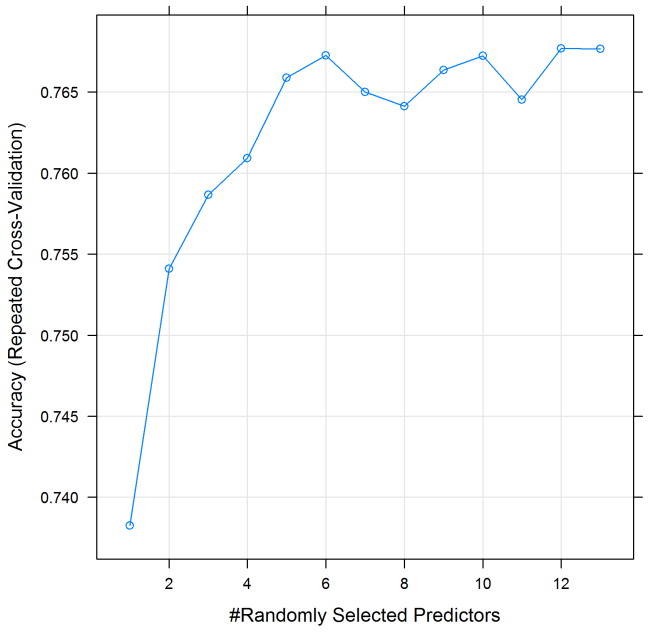


Figure 4.19: Cross-validated accuracy profile for the Random Forest model. The final model was fit with 12 predictors.

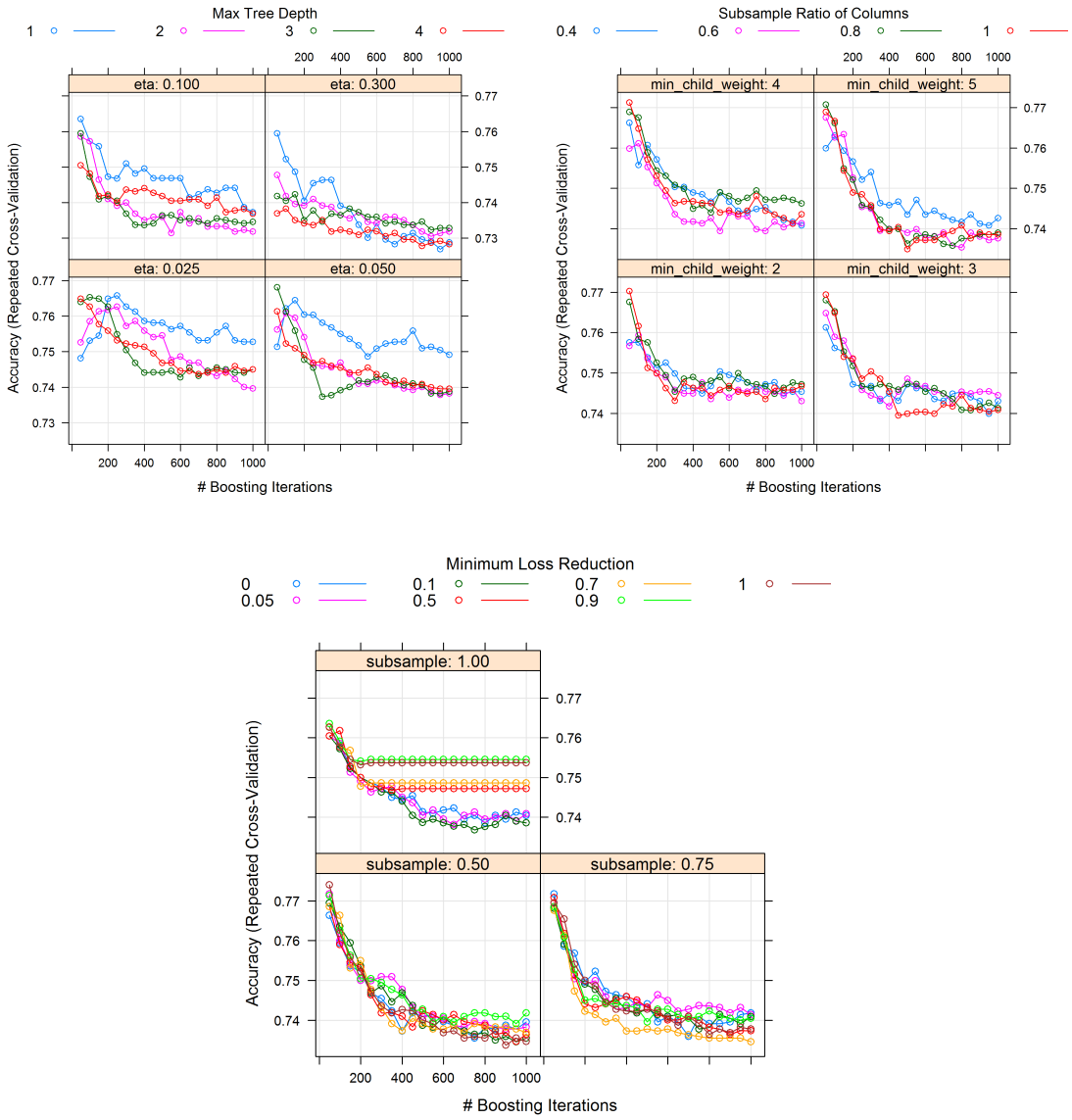


Figure 4.20: Cross-validated accuracy profile for the Boosting model. The final model parameters are listed in Table D.3.

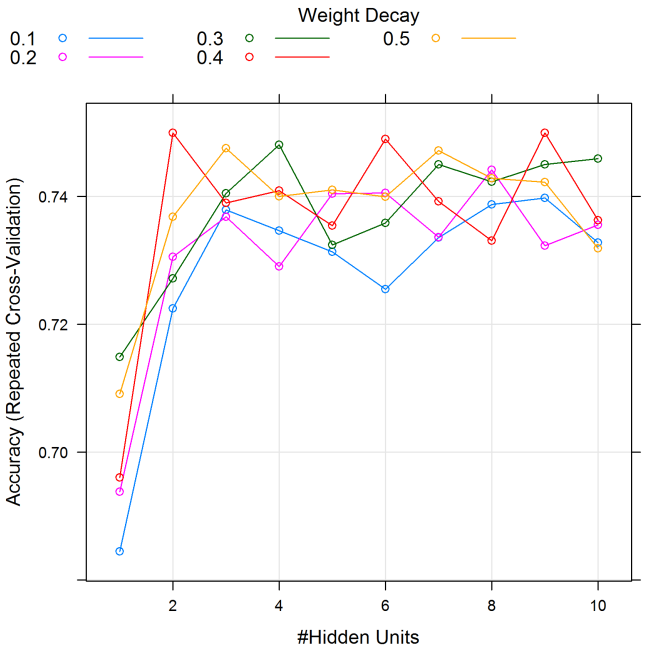


Figure 4.21: Cross-validated accuracy profile for the Neural Network model. The optimal model used decay of 0.5 and 9 hidden units.

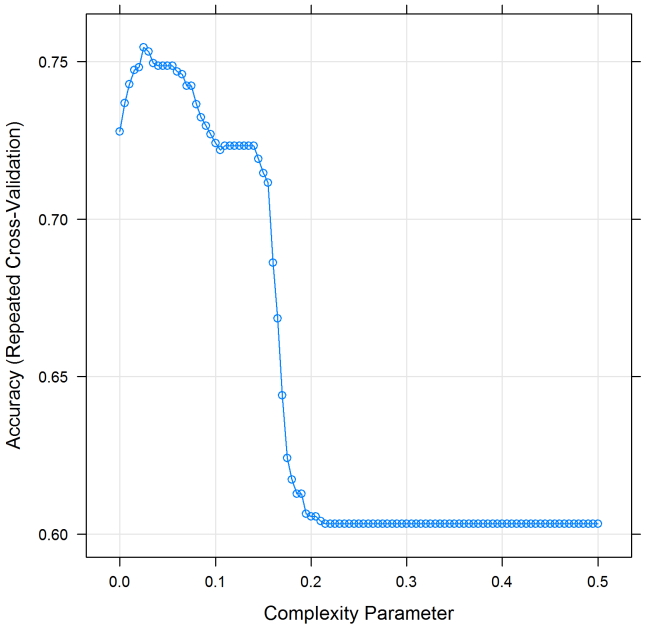


Figure 4.22: Cross-validated accuracy profile for the CART model. The complexity parameter (cp) of the final model is 0.025.

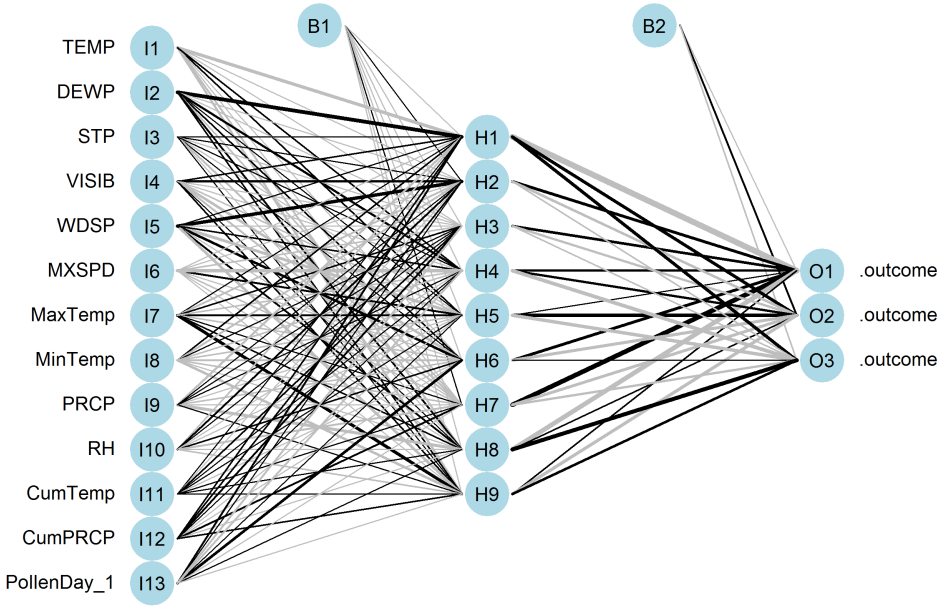


Figure 4.23: The structure of Neural Network model for prediction of ragweed pollen levels.

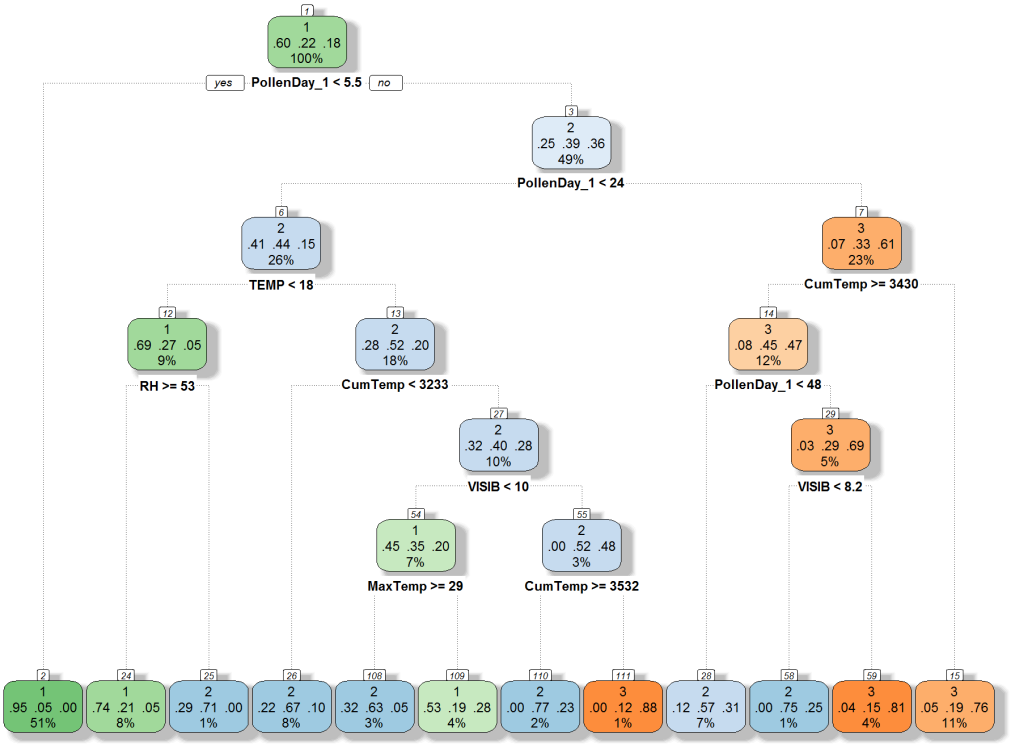


Figure 4.24: The structure of the Decision Tree for prediction of ragweed pollen levels. The nodes represent the conditions and input variables, the leaves represent ragweed pollen levels.

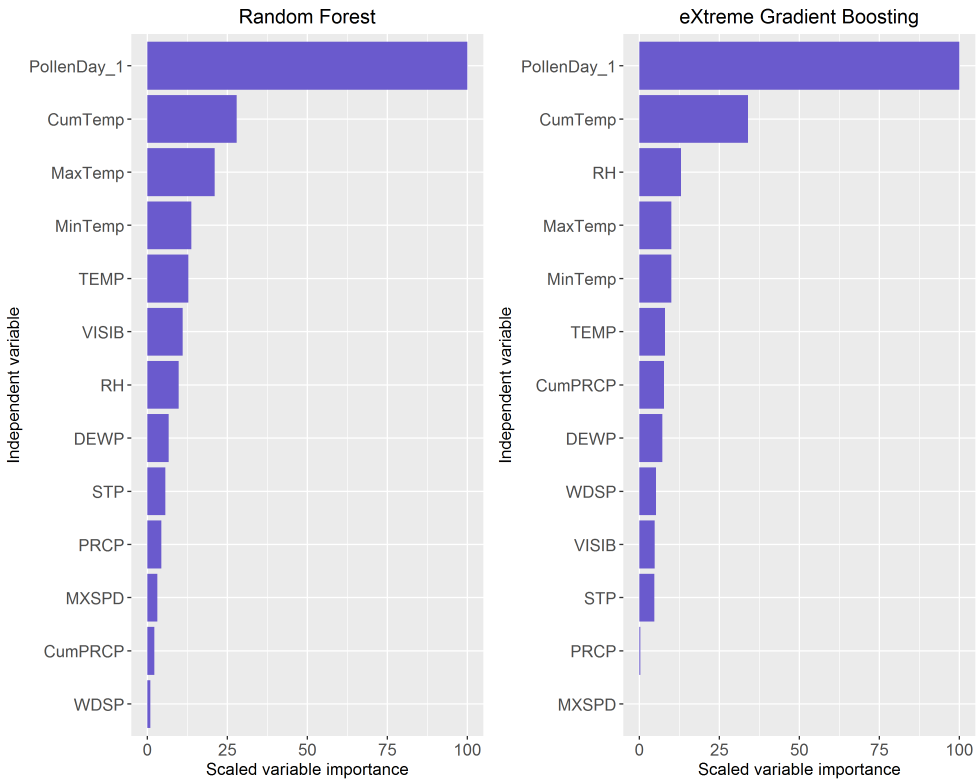


Figure 4.25: Importance of each predictor variable in the Random Forest model and the XGBoost model of pollen levels.

4.5 Summary

This chapter explored different ML techniques for prediction of ragweed pollen concentrations/levels with 12 meteorological factors at a local scale. Newark, NJ was selected as the sample location. The ML model performances were compared based on various metrics. All the models were cross-validated and the optimal parameters were obtained. The results showed that Random Forest and XGBoost models performed better than other models in both regression and classification problems. Direct comparison of the observed and predicted ragweed pollen concentration confirmed the good performance of the models obtained and the ability to recreate most of the variation. Only some extreme values of pollen counts were underestimated. Random Forest and XGBoost models predicted 77% of the pollen levels correctly with $F1$ score of nearly 0.9. The ML methods can be used in local-scale estimation of pollen concentration/level; however, the models need to be trained and optimized for best performance at different locations. Forecasting of pollen concentrations a few days in advance can help reduce the adverse health effects of allergenic pollen with pre-emptive medication and behavioral adaptation.

Chapter 5
CONCLUSIONS AND RECOMMENDATIONS

5.1 Main Findings

This dissertation investigated the impacts of climate change on air quality and associated human exposures based on mechanistic modeling, statistical analysis and data-driven ML modeling. The key findings and conclusions are summarized below:

(i) The CMAQ-Pollen modeling system was able to simulate the spatiotemporal distributions of oak and ragweed pollen concentration in 2004 across the CONUS with relatively good performance based on results from correlation analysis, fractional bias, hit and false rates. This regional modeling system has 36 km horizontal grid resolution, which is the highest spatial resolution employed across the CONUS using the CMAQ-Pollen modeling system. Dry deposition, emission and vertical eddy diffusion are the dominant processes determining ambient pollen concentrations. The boundary conditions of airborne pollen concentration exert remarkable influence on mean pollen concentrations at locations with small emission sources.

(ii) Considering solely the single impact of meteorological conditions simulated under the RCP 8.5 scenario, the ragweed pollen season in 2047 will start earlier and last longer for all the nine CONUS climate regions, with increasing average pollen concentrations in most regions. The mean and maximum hourly concentrations of oak pollen were predicted to increase in the Northeast, South and Southeast regions, but to decrease in the Northwest, East North Central,

and West North Central regions. The oak pollen season was estimated to start earlier in the Central, Northeast, South and Southeast regions. The oak pollen season length was estimated to shorten by 1-2 days for most regions, except the Southeast and Southwest regions.

(iii) Statistical analysis of observed pollen observation from 1990 to 2010 at 56 pollen monitoring stations across the CONUS indicates that the ragweed pollen season starts from higher latitudes and then shifts to lower latitudes, while the pollen season lasts longer at lower latitudes. The peak pollen concentration and annual production both decrease as latitude increases. Comparisons of mean pollen indices between periods 2001-2010 and 1994-2000 showed that the mean ragweed pollen season from 2001 to 2010 tends to start earlier at 76% of the monitoring stations, with the biggest shift being 30 days earlier. Changes in season length do not have significant correlation with latitude or climate regions. The annual average number of days when both ragweed pollen $\geq$1 and ozone exceedances (DMA8[O_3]>70 ppb) occur for 58 pollen stations during 1994-2010 ranged from 0 to 17 days, with the the largest number of co-occurrence appearing in the West, Southwest, Southeast and South region.

(iv) Co-occurrence of ragweed pollen and ozone exceedances under climate change were investigated based on simulated ragweed pollen and ozone concentrations. The co-occurrence of ragweed pollen and ozone exceedances ranged from 0 to 21 days in 2004, and is predicted to be 0 to 24 days in 2047. The spatial distribution pattern of the co-occurrence is similar to that of ozone exceedances. Although the co-occurrence of ragweed pollen and ozone exceedances does not appear across a large area of the CONUS, it influences a remarkable fraction of the population, including five of the 10 largest cities by population in 2004 and six of the 10 largest cities in 2047.

(v) Population inhalation exposures to airborne ragweed pollen and ozone were simulated with a probabilistic model based on CMAQ estimated ragweed pollen and ozone concentrations in 2004. Inhalation exposure to ragweed pollen is higher outdoors than indoors across the CONUS. On the contrary, people in general have higher inhalation exposures to ozone from time spent indoors than outdoors. There is distinct variation in inhalation exposures to ragweed pollen among the nine CONUS climate regions due to the large differences in ragweed coverage and ragweed pollen concentrations. The inhalation exposures to ragweed pollen and ozone also vary by age and gender. Males have higher inhalation exposures to ragweed and pollen because they have higher mean inhalation rates and generally spend more time outdoors. The inhalation exposure to ragweed pollen and ozone per unit body weight decreases with age due to decreased physiological daily inhalation rates and increased weight.

(vi) Local scale statistical modeling of ragweed pollen was conducted using ML methods. The models only need as inputs meteorological variables and previous day pollen concentration. Compared with regional deterministic models, such as the CMAQ-Pollen modeling system, statistical models have much less demand in computational power and input data. Six regression models and five classification models were tested for prediction of pollen concentration (pollen grains/m^3) and pollen levels (low, medium, high), respectively. Random Forest and XGBoost outperformed other models for both regression and classification problems with the lowest $RMSE$ and the highest $F1$ score. Pollen concentration of the previous day was found to be the most important predictor variable in both the Random Forest and XGBoost models.

5.2 Future Research Directions

There are certain limitations in this study that can be addressed with the following recommendations. Focus of future research work is also suggested.

(i) Improve the performance of the CMAQ-Pollen modeling system using high quality and high resolution meteorological inputs. The meteorology data used in this dissertation were downscaled from a global climate model without assimilating local weather observations, therefore, the day-to-day weather variability could not be represented, which leads to discrepancies in daily observed and simulated pollen concentrations for 2004. The CMAQ-Pollen modeling system has been proven to be applicable at 50 km and 36 km horizontal grid resolutions, and this can be expanded to 12 km, 4 km or even 2 km, if the required meteorology data become available.

(ii) The performance of the CMAQ-Pollen modeling system is also partly determined by the accuracy of the pollen emission model. Vegetation coverage is an important input to the pollen emission model. In this study, ragweed coverage, which was not available in the BELD3.1 database, was estimated using ragweed pollen counts and vegetation coverage information from BELD3.1. This model can be further improved with updated information from satellite data, local observations, and knowledge of ragweed ecology. The emission model can also be parameterized for other pollen species to expand the application of the CMAQ-Pollen modeling system.

(iii) Due to year-to-year variations in meteorological conditions, the spatiotemporal distributions of allergenic pollen predicted in the 2050s can not be well represented by a single year. Multiple years of simulation need to be conducted to get reasonable estimates of average metrics of pollen season and pollen concentration in the future. Due to limitation in meteorology data, it was not possible

to complete this task in this study. Future research work on this topic should take this into consideration.

(iv) The health impacts of co-occurrences of ragweed pollen and ozone exceedances could be evaluated with data on asthma from the National Health and Nutrition Examination Surveys (NHANES), National Asthma Survey (NAS), National Survey of Children's Health (NSCH) etc. Statistical relationships between asthma and the co-occurring allergenic pollen and air pollutants could help predict the potential consequences of climate change for public health.

(v) The exposure model can also be improved to better assess exposures to ozone and allergenic pollen. The indoor-to-outdoor (I/O) ratio of air pollutant is a critical parameter in our exposure model. It varies with indoor environments and ventilation rate. So far, to our knowledge, there is no statistical distribution of I/O ratio of ozone available for each climate region in the CONUS. Therefore, the same uniform distribution was assumed for I/O ratios for each CONUS climate region in this study. It is expected that people will adapt to climate change with new behavioral and building design strategies. All the exposure factors need to be adjusted or re-evaluated to reflect exposures expected in the future.

(vi) The rapid development of ML methods is bringing more powerful data-driven techniques for prediction of air pollutants. This study only explored several ML methods for local scale pollen prediction with meteorological factors. With growing databases, including remote sensing data, fine-scale geographic information systems (GIS) data, satellite image data, mobile phone big data, etc., the inputs to the ML models could be substantially expanded to achieve prediction with improved spatiotemporal resolution and high accuracy.

Bibliography

[1] L. Ziska, K. Knowlton, C. Rogers, National Allergy Bureau, and Canada. Aerobiology Research Laboratories. Update to data originally published in: Ziska, l., et al. 2011. Recent warming by latitude associated with increased length of ragweed pollen season in Central North America. *Proceedings of the National Academy of Sciences of the United States of America*, page 4251, 2016.

[2] Asthma and Allergy Foundation of America. Extreme allergies and global warming. *www.aafa.org*, 2015.

[3] Y. Zhang. Climate change and airborne allergens. 2015.

[4] A. J. McMichael, R. E. Woodruff, and S. Hales. Climate change and human health: present and future risks. *Lancet*, 367(9513):859–869, 2006.

[5] M. L. Bell, R. Goldberg, C. Hogrefe, P. L. Kinney, K. Knowlton, B. Lynn, J. Rosenthal, C. Rosenzweig, and J. A. Patz. Climate change, ambient ozone, and health in 50 US cities. *Climatic Change*, 82(1-2):61–76, 2007.

[6] D. J. Jacob and D. A. Winner. Effect of climate change on air quality. *Atmospheric Environment*, 43(1):51–63, 2009.

[7] E. Tagaris, K. J. Liao, A. J. Delucia, L. Deck, P. Amar, and A. G. Russell. Potential impact of climate change on air pollution-related human health effects. *Environmental Science & Technology*, 43(13):4979–4988, 2009.

[8] USGCRP. The impacts of climate change on human health in the United States: A scientific assessment. *U.S. Global Change Research Program, Washington, DC,*, page 312 pp., 2016.

[9] USEPA. Climate change indicators in the United States, 2016. fourth edition. *EPA 430-R-16-004*, 2016.

[10] Y. Zhang, L. Bielory, T. Cai, Z. Mi, and P. Georgopoulos. Predicting onset and duration of airborne allergenic pollen season in the United States. *Atmospheric Environment*, 103:297–306, 2015.

[11] Y. Zhang, P. G. Georgopoulos, and L. Bielory. Climate change effects on ragweed (*Ambrosia*) and mugwort (*Artemisia*) pollen seasons in USA. *Proceedings of the Air and Waste Management Association's Annual Conference and Exhibition, AWMA*, 4:2983–2993, 2013.

[12] Yong Zhang, Leonard Bielory, Zhongyuan Mi, Ting Cai, Alan Robock, and Panos Georgopoulos. Allergenic pollen season variations in the past two decades under changing climate in the United States. *Global Change Biology*, 21(4): 1581–1589, 2015. doi: 10.1111/gcb.12755.

[13] Yong Zhang, Leonard Bielory, and Panos Georgopoulos. Climate change effect on *Betula* (birch) and *Quercus* (oak) pollen seasons in the United States. *International Journal of Biometeorology*, 58(5):909–919, 2013.

[14] Ki-Hyun Kim, Shamin Ara Jahan, and Ehsanul Kabir. A review on human health perspective of air pollution with respect to allergies and asthma. *Environment International*, 59:41–52, 2013.

[15] Sabit Cakmak, Robert E. Dales, and Frances Coates. Does air pollution increase the effect of aeroallergens on hospitalization for asthma? *Journal of Allergy and Clinical Immunology*, 129(1):228–231, 2012.

[16] Juan Declet-Barreto, Sean Alcorn, and Natural Resources Defense Council. NRDC: Sneezing and wheezing - how climate change could increase ragweed allergies, air pollution, and asthma. 2015.

[17] N. E. Klepeis, W. C. Nelson, W. R. Ott, J. P. Robinson, A. M. Tsang, P. Switzer, J. V. Behar, S. C. Hern, and W. H. Engelmann. The national human activity pattern survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Analysis and Environmental Epidemiology*, 11(3):231–252, 2001.

[18] Institute of Medicine. *Climate Change, the Indoor Environment, and Health*. 2011. ISBN 978-0-309-20941-0. doi: 10.17226/13115.

[19] C. J. Weschler. Chemistry in indoor environments: 20 years of research. *Indoor Air-International Journal of Indoor Air Quality and Climate*, 21(3):205–218, 2011.

[20] Douglas W. Dockery, C. Arden Pope, Xiping Xu, John D. Spengler, James H. Ware, Martha E. Fay, Benjamin G. Ferris, and Frank E. Speizer. An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine*, 329(24):1753–1759, 1993.

[21] B. J. Hubbell, A. Hallberg, D. R. McCubbin, and E. Post. Health-related benefits of attaining the 8-hr ozone standard. *Environmental Health Perspectives*, 113(1):73–82, 2005.

[22] A. J. Krupnick, W. Harrington, and B. Ostro. Ambient ozone and acute health-effects - evidence from daily data. *Journal of Environmental Economics and Management*, 18(1):1–18, 1990.

[23] Charles J. Weschler. Ozone's impact on public health: Contributions from indoor exposures to ozone and products of ozone-initiated chemistry. *Environmental Health Perspectives*, 114(10):1489–1496, 2006. doi: 10.1289/ehp.9256.

[24] Chun Chen, Bin Zhao, and Charles J Weschler. Indoor exposure to "outdoor PM_{10}": assessing its influence on the relationship between PM_{10} and short-term mortality in U.S. cities. *Epidemiology*, 23 6:870–8, 2012.

[25] USEPA. 2013 final report: Integrated science assessment for ozone and related photochemical oxidants. *U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-10/076F*, 2013.

[26] Charles J. Weschler and Helen C. Shields. The influence of ventilation on reactions among indoor pollutants: Modeling and experimental observations. *Indoor Air*, 10(2):92–100, 2000.

[27] Charles J. Weschler, Helen C. Shields, and Datta V. Naik. Indoor ozone exposures. *JAPCA*, 39(12):1562–1568, 1989.

[28] C. J. Weschler. Ozone in indoor environments: Concentration and chemistry. *Indoor Air-International Journal of Indoor Air Quality and Climate*, 10(4): 269–288, 2000.

[29] C. J. Weschler. New directions: Ozone-initiated reaction products indoors may be more harmful than ozone itself. *Atmospheric Environment*, 38(33):5715–5716, 2004.

[30] C. J. Weschler, H. C. Shields, and D. V. Nalk. Indoor chemistry involving O_3, NO, and NO_2 as evidenced by 14 months of measurements at a site in Southern California. *Environmental Science & Technology*, 28(12):2120–2132, 1994.

[31] X. Chen and P. K. Hopke. A chamber study of secondary organic aerosol formation by limonene ozonolysis. *Indoor Air-International Journal of Indoor Air Quality and Climate*, 20(4):320–328, 2010.

[32] Z. H. Fan, P. Lioy, C. Weschler, N. Fiedler, H. Kipen, and J. F. Zhang. Ozone-initiated reactions with mixtures of volatile organic compounds under simulated indoor conditions. *Environmental Science and Technology*, 37(9):1811–1821, 2003.

[33] Z. H. Fan, C. J. Weschler, I. K. Han, and J. F. Zhang. Co-formation of hydroperoxides and ultra-fine particles during the reactions of ozone with a complex voc mixture under simulated indoor conditions. *Atmospheric Environment*, 39(28): 5171–5182, 2005.

[34] G. Sarwar and R. Corsi. The effects of ozone/limonene reactions on indoor secondary organic aerosols. *Atmospheric Environment*, 41(5):959–973, 2007.

[35] T. Wainman, J. F. Zhang, C. J. Weschler, and P. J. Lioy. Ozone and limonene in indoor air: A source of submicron particle exposure. *Environmental Health Perspectives*, 108(12):1139–1145, 2000.

[36] C. J. Weschler and H. C. Shields. Experiments probing the influence of air exchange rates on secondary organic aerosols derived from indoor chemistry. *Atmospheric Environment*, 37(39-40):5621–5631, 2003.

[37] C. J. Weschler and H. C. Shields. Indoor ozone/terpene reactions as a source of indoor particles. *Atmospheric Environment*, 33(15):2301–2312, 1999.

[38] K. S. Docherty, W. Wu, Y. B. Lim, and P. J. Ziemann. Contributions of organic peroxides to secondary aerosol formed from reactions of monoterpenes with O_3. *Environmental Science & Technology*, 39(11):4049–4059, 2005.

[39] Barbara J. Finlayson-Pitts and Jr. James N. Pitts. *Chemistry of the Upper and Lower Atmosphere*. 2000.

[40] A. W. Norgaard, J. K. Nojgaard, K. Larsen, S. Sporring, C. K. Wilkins, P. A. Clausen, and P. Wolkoff. Secondary limonene endo-ozonide: A major product from gas-phase ozonolysis of *R*-(+)-limonene at ambient temperature. *Atmospheric Environment*, 40(19):3460–3466, 2006.

[41] Annette C Rohr. The health significance of gas-and particle-phase terpene oxidation products: A review. *Environment International*, 60:145–162, 2013.

[42] Michael S Waring, J Raymond Wells, and Jeffrey A Siegel. Secondary organic aerosol formation from ozone reactions with single terpenoids and terpenoid mixtures. *Atmospheric Environment*, 45(25):4235–4242, 2011.

[43] C. J. Weschler, A. Wisthaler, S. Cowlin, G. Tamas, P. Strom-Tejsen, A. T. Hodgson, H. Destaillats, J. Herrington, J. J. Zhang, and W. W. Nazaroff. Ozone-initiated chemistry in an occupied simulated aircraft cabin. *Environmental Science & Technology*, 41(17):6177–6184, 2007.

[44] A. Wisthaler, G. Tamas, D. P. Wyon, P. Strom-Tejsen, D. Space, J. Beauchamp, A. Hansel, T. D. Mark, and C. J. Weschler. Products of ozone-initiated chemistry in a simulated aircraft environment. *Environmental Science & Technology*, 39(13):4823–4832, 2005.

[45] A. Wisthaler and C. J. Weschler. Reactions of ozone with human skin lipids: Sources of carbonyls, dicarbonyls, and hydroxycarbonyls in indoor air. *Proceedings of the National Academy of Sciences of the United States of America*, 107 (15):6568–6575, 2010.

[46] P. Fruekilde, J. Hjorth, N. R. Jensen, D. Kotzias, and B. Larsen. Ozonolysis at vegetation surfaces: A source of acetone, 4-oxopentanal, 6-methyl-5-hepten-2-one, and geranyl acetone in the troposphere. *Atmospheric Environment*, 32 (11):1893–1902, 1998.

[47] N. Nicolaides. Skin lipids: their biochemical uniqueness. *Science*, 186(4158): 19–26, 1974.

[48] T. Nikkari. Comparative chemistry of sebum. *J Invest Dermatol*, 62(3):257–67, 1974.

[49] R. Vingarzan. A review of surface ozone background levels and trends. *Atmospheric Environment*, 38(21):3431–3442, 2004.

[50] IPCC. Special report on emissions scenarios, tech. rep. *Intergovernmental Panel on Clim. Change, New York*, 2000.

[51] C. G. Nolte, A. B. Gilliland, C. Hogrefe, and L. J. Mickley. Linking global to regional models to assess future climate impacts on surface ozone levels in the United States. *Journal of Geophysical Research-Atmospheres*, 113(D14), 2008.

[52] E. Tagaris, K. Manomaiphiboon, K. J. Liao, L. R. Leung, J. H. Woo, S. He, P. Amar, and A. G. Russell. Impacts of global climate change and emissions on regional ozone and fine particulate matter concentrations over the United States. *Journal of Geophysical Research-Atmospheres*, 112(D14), 2007.

[53] Y. F. Lam, J. S. Fu, S. Wu, and L. J. Mickley. Impacts of future climate change and effects of biogenic emissions on surface ozone and particulate matter concentrations in the United States. *Atmospheric Chemistry and Physics*, 11 (10):4789–4806, 2011.

[54] R. Gonzalez-Abraham, S. H. Chung, J. Avise, B. Lamb, E. P. Salathé Jr, C. G. Nolte, D. Loughlin, A. Guenther, C. Wiedinmyer, T. Duhl, Y. Zhang, and D. G. Streets. The effects of global change upon United States air quality. *Atmos. Chem. Phys.*, 15(21):12645–12665, 2015.

[55] A. Penrod, Y. Zhang, K. Wang, S. Y. Wu, and L. R. Leung. Impacts of future climate and emission changes on US air quality. *Atmospheric Environment*, 89: 533–547, 2014.

[56] Y. Zhang, X. M. Hu, L. R. Leung, and W. I. Gustafson. Impacts of regional climate change on biogenic emissions and air quality. *Journal of Geophysical Research-Atmospheres*, 113(D18), 2008.

[57] M. Trail, A. P. Tsimpidi, P. Liu, K. Tsigaridis, J. Rudokas, P. Miller, A. Nenes, Y. Hu, and A. G. Russell. Sensitivity of air quality to potential future climate change and emissions in the United States and major cities. *Atmospheric Environment*, 94:552–563, 2014.

[58] Hao He, Xin-Zhong Liang, Hang Lei, and Donald J. Wuebbles. Future U.S. ozone projections dependence on regional emissions, climate change, long-range transport and differences in modeling design. *Atmospheric Environment*, 128: 124–133, 2016.

[59] Detlef P Van Vuuren, Jae Edmonds, Mikiko Kainuma, Keywan Riahi, Allison Thomson, Kathy Hibbard, George C Hurtt, Tom Kram, Volker Krey, and Jean-Francois Lamarque. The representative concentration pathways: an overview. *Climatic Change*, 109(1-2):5–31, 2011.

[60] Neal Fann, Christopher G. Nolte, Patrick Dolwick, Tanya L. Spero, Amanda Curry Brown, Sharon Phillips, and Susan Anenberg. The geographic distribution and economic value of climate change-related ozone health impacts in the United States in 2030. *Journal of the Air & Waste Management Association (Taylor & Francis Ltd)*, 65(5):570–580, 2015.

[61] Yang Gao, Joshua S Fu, John B Drake, J-F Lamarque, and Yang Liu. The impact of emission and climate change on ozone in the United States under representative concentration pathways (RCPs). *Atmospheric Chemistry and Physics*, 13(18):9607–9621, 2013.

[62] Y. M. Kim, Y. Zhou, Y. Gao, J. S. Fu, B. A. Johnson, C. Huang, and Y. Liu. Spatially resolved estimation of ozone-related mortality in the United States under two representative concentration pathways (RCPs) and their uncertainty. *Climatic Change*, 128(1-2):71–84, 2015.

[63] Jian Sun, Joshua S. Fu, Kan Huang, and Yang Gao. Estimation of future $PM_{2.5}$- and ozone-related mortality over the continental United States in a changing climate: An application of high-resolution dynamical downscaling technique. *Journal of the Air & Waste Management Association*, 65(5):611–623, 2015.

[64] G. G. Pfister, S. Walters, J. F. Lamarque, J. Fast, M. C. Barth, J. Wong, J. Done, G. Holland, and C. L. Bruyère. Projections of future summertime ozone over the U.S. *Journal of Geophysical Research: Atmospheres*, 119(9): 5559–5582, 2014.

[65] C. G. Nolte, T. L. Spero, J. H. Bowden, M. S. Mallard, and P. D. Dolwick. The potential effects of climate change on air quality across the conterminous US at 2030 under three representative concentration pathways. *Atmos. Chem. Phys.*, 18(20):15471–15489, 2018.

[66] C. W. Tessum, J. D. Hill, and J. D. Marshall. Twelve-month, 12km resolution north american WRF-CHEM v3.4 air quality simulation: performance evaluation. *Geoscientific Model Development*, 8(4):957–973, 2015.

[67] Y. Li, D. K. Henze, D. Jack, and P. L. Kinney. The influence of air quality model resolution on health impact assessment for fine particulate matter and its components. *Air Quality Atmosphere and Health*, 9(1):51–68, 2016.

[68] Yang Gao, Joshua S. Fu, John B. Drake, and Yun-Fat Lam. Regional climate downscaling study in Eastern United States. *10th Annual CMAS Conference*, Chapel Hill, NC(October 24-26, 2011), 2011.

[69] Wonbae Jeon, Yunsoo Choi, Anirban Roy, Shuai Pan, Daniel Price, Mi-Kyoung Hwang, Kyu Rang Kim, and Inbo Oh. Investigation of primary factors affecting the variation of modeled oak pollen concentrations: A case study for southeast Texas in 2010. *Asia-Pacific Journal of Atmospheric Sciences*, 54(1):33–41, 2018.

[70] L. Liu, F. Solmon, R. Vautard, L. Hamaoui-Laguel, C. Z. Torma, and F. Giorgi. Ragweed pollen production and dispersion modelling within a regional climate system, calibration and application over Europe. *Biogeosciences*, 13(9):2769–2786, 2016.

[71] K. Zink, A. Pauling, M. W. Rotach, H. Vogel, P. Kaufmann, and B. Clot. EMPOL 1.0: a new parameterization of pollen emission in numerical weather prediction models. *Geoscientific Model Development*, 6(6):1961–1975, 2013.

[72] K. Zink, P. Kaufmann, B. Petitpierre, O. Broennimann, A. Guisan, E. Gentilini, and M. Rotach. Numerical ragweed pollen forecasts using different source maps: a comparison for France. *International Journal of Biometeorology*, 61(1):23–33, 2017.

[73] H. Garcia-Mozo, Carmen Galán, Victoria Jato, Jordina Belmonte, Consuelo Diaz de la Guardia, Delia Fernandez, Gutiérrez-Bustillo Adela M, M. Jesus Aira, Joan M. Roure, Luisa Ruiz, M. Mar Trigo, and Eugenio Dominguez-Vilches. *Quercus pollen season dynamics in the Iberian Peninsula: Response to meteorological parameters and possible consequences of climate change*, volume 13. 2006.

[74] S. Kawashima and Y. Takahashi. An improved simulation of mesoscale dispersion of airborne cedar pollen using a flowering-time map. *Grana*, 38(5):316–324, 1999.

[75] Robert Pasken and Joseph A. Pietrowicz. Using dispersion and mesoscale meteorological models to forecast pollen concentrations. *Atmospheric Environment*, 39(40):7689–7701, 2005.

[76] Nora Helbig, Bernhard Vogel, Heike Vogel, and Franz Fiedler. Numerical modelling of pollen dispersion on the regional scale. *Aerobiologia*, 20(1):3–19, 2004. doi: 10.1023/b:aero.0000022984.51588.30.

[77] Heike Vogel, Andreas Pauling, and Bernhard Vogel. Numerical simulation of birch pollen dispersion with an operational weather forecast system. *International Journal of Biometeorology*, 52(8):805–814, 2008.

[78] Katrin Zink, Heike Vogel, Bernhard Vogel, Donát Magyar, and Christoph Kottmeier. Modeling the dispersion of *Ambrosiaartemisiifolia* L. pollen with the model system COSMO-ART. *International Journal of Biometeorology*, 56 (4):669–680, 2012.

[79] P. Siljamo, M. Sofiev, E. Filatova, L. Grewling, S. Jager, E. Khoreva, T. Linkosalo, S. O. Jimenez, H. Ranta, A. Rantio-Lehtimaki, A. Svetlov, L. Veriankaite, E. Yakovleva, and J. Kukkonen. A numerical model of birch pollen emission and dispersion in the atmosphere. model evaluation and sensitivity analysis. *International Journal of Biometeorology*, 57(1):125–136, 2013.

[80] Mikhail Sofiev and Karl-Christian Bergmann. Allergenic pollen: a review of the production, release, distribution and health impacts. 2013.

[81] M. Sofiev, U. Berger, M. Prank, J. Vira, J. Arteta, J. Belmonte, K.-C. Bergmann, F. Chéroux, H. Elbern, E. Friese, C. Galan, R. Gehrig, D. Khvorostyanov, R. Kranenburg, U. Kumar, V. Marécal, F. Meleux, L. Menut, A.-M. Pessi, L. Robertson, O. Ritenberga, V. Rodinkova, A. Saarto, A. Segers, E. Severova, I. Sauliene, P. Siljamo, B. M. Steensen, E. Teinemaa, M. Thibaudon, and V.-H. Peuch. MACC regional multi-model ensemble simulations of birch pollen dispersion in Europe. *Atmospheric Chemistry and Physics*, 15(14):8115–8130, 2015.

[82] M. Sofiev, P. Siljamo, H. Ranta, and A. Rantio-Lehtimäki. Towards numerical forecasting of long-range air transport of birch pollen: theoretical considerations and a feasibility study. *International Journal of Biometeorology*, 50(6):392–402, 2006.

[83] Silvio Schueler and Katharina Schlünzen. Modeling of oak pollen dispersal on the landscape level with a mesoscale atmospheric model. *Environmental Modeling and Assessment*, 11(3):179–194, 2006.

[84] Marje Prank, Daniel S. Chapman, James M. Bullock, Jordina Belmonte, Uwe Berger, Aslog Dahl, Siegfried Jäger, Irina Kovtunenko, Donát Magyar, Sami Niemelä, Auli Rantio-Lehtimäki, Viktoria Rodinkova, Ingrida Sauliene, Elena Severova, Branko Sikoparija, and Mikhail Sofiev. An operational model for forecasting ragweed pollen release and dispersion in Europe. *Agricultural and Forest Meteorology*, 182–183(0):43–53, 2013.

[85] Christos Efstathiou, Sastry Isukapalli, and Panos Georgopoulos. A mechanistic modeling system for estimating large-scale emissions and transport of pollen and co-allergens. *Atmospheric Environment*, 45(13):2260–2276, 2011.

[86] TR Duhl, R Zhang, A Guenther, SH Chung, MT Salam, JM House, RC Flagan, EL Avol, FD Gilliland, and BK Lamb. The simulator of the timing and magnitude of pollen season (STAMPS) model: a pollen production model for regional emission and transport modeling. *Geoscientific Model Development Discussions*, 6(2):2325–2368, 2013.

[87] R Zhang, T Duhl, MT Salam, JM House, RC Flagan, EL Avol, FD Gilliland, A Guenther, SH Chung, and BK Lamb. Development of a regional-scale pollen emission and transport modeling framework for investigating the impact of

climate change on allergic airway disease. *Biogeosciences*, 11(6):1461–1478, 2014.

[88] Atin Adhikari, Tiina Reponen, Sergey A. Grinshpun, Dainius Martuzevicius, and Grace LeMasters. Correlation of ambient inhalable bioaerosols with particulate matter and ozone: A two-year study. *Environmental Pollution*, 140(1): 16–28, 2006.

[89] Robert E. Dales, Sabit Cakmak, Stan Judek, Tom Dann, Frances Coates, Jeffrey R. Brook, and Richard T. Burnett. Influence of outdoor aeroallergens on hospitalization for asthma in Canada. *Journal of Allergy and Clinical Immunology*, 113(2):303–306, 2004.

[90] Thomas Frei and Ewald Gassner. Climate change and its impact on birch pollen quantities and the start of the pollen season an example from Switzerland for the period 1969–2006. *International Journal of Biometeorology*, 52(7):667–674, 2008.

[91] Herminia García-Mozo, C. Galán, P. Alcázar, C. de la Guardia, D. Nieto-Lugilde, M. Recio, P. Hidalgo, F. Gónzalez-Minero, L. Ruiz, and E. Domínguez-Vilches. Trends in grass pollen season in southern Spain. *Aerobiologia*, 26(2): 157–169, 2010.

[92] P. L. Kinney. Climate change, air quality, and human health. *American Journal of Preventive Medicine*, 35(5):459–467, 2008.

[93] C.G. Nolte, P.D. Dolwick, N. Fann, L.W. Horowitz, V. Naik, R.W. Pinder, T.L. Spero, D.A. Winner, and L.H. Ziska. Air quality. In impacts, risks, and adaptation in the United States: Fourth National Climate Assessment, Volume II. *U.S. Global Change Research Program, Washington, DC, USA*, pages 512–538, 2018.

[94] W. Pfender, K. Graw, W. Bradley, M. Carney, and L. Maxwell. Emission rates, survival, and modeled dispersal of viable pollen of creeping bentgrass. *Crop Science*, 47(6):2529–2539, 2007.

[95] Donald E. Aylor. Quantifying maize pollen movement in a maize canopy. *Agricultural and Forest Meteorology*, 131(3–4):247–256, 2005.

[96] S. Dupont, Y. Brunet, and N. Jarosz. Eulerian modelling of pollen dispersal over heterogeneous vegetation canopies. *Agricultural and Forest Meteorology*, 141(2-4):82–104, 2006.

[97] Ting Cai, Yong Zhang, Xiang Ren, Leonard Bielory, Zhongyuan Mi, Christopher G. Nolte, Yang Gao, L. Ruby Leung, and Panos G. Georgopoulos. Development of a semi-mechanistic allergenic pollen emission model. *Science of The Total Environment*, 653:947–957, 2019.

[98] B. J. Green, T. O'Meara, J. K. Sercombe, and E. R. Tovey. Measurement of personal exposure to outdoor aeromycota in Northern New South Wales, Australia. *Annals of Agricultural and Environmental Medicine*, 13(2):225–234, 2006.

[99] Robert G. Peel, Ole Hertel, Matt Smith, and Roy Kennedy. Personal exposure to grass pollen: relating inhaled dose to background concentration. *Annals of Allergy, Asthma & Immunology*, 111(6):548–554, 2013.

[100] M. Riediker, S. Keller, B. Wuthrich, T. Koller, and C. Monn. Personal pollen exposure compared to stationary measurements. *Journal of Investigational Allergology & Clinical Immunology*, 10(4):200–203, 2000.

[101] Uwe Berger, Maximilian Kmenta, and Katharina Bastl. Individual pollen exposure measurements: are they feasible? *Current Opinion in Allergy and Clinical Immunology*, 14(3):200–205, 2014.

[102] N. Yamamoto, H. Matsuki, and Y. Yanagisawa. Application of the personal aeroallergen sampler to assess personal exposures to Japanese cedar and cypress pollens. *Journal of Exposure Science and Environmental Epidemiology*, 17(7): 637–643, 2007.

[103] T. Sehlinger, K. Boehm, F. Goergen, and K. C. Bergmann. Measuring individual pollen exposure. *Journal of Allergy and Clinical Immunology*, 131(2): Ab79–Ab79, 2013.

[104] D. Myszkowska, B. Bilo, D. Stepalska, J. Wotek, and K. Obtutowicz. Personal and stationary pollen monitoring with regard to pollen allergy symptoms. *Allergy & Clinical Immunology International-Journal of the World Allergy Organization*, 19(3):108–114, 2007.

[105] Gary L. Rosenberg, Richard R. Rosenthal, and Philip S. Norman. Inhalation challenge with ragweed pollen in ragweed-sensitive asthmatics. *Journal of Allergy and Clinical Immunology*, 71(3):302–310, 1983.

[106] M. N. Driessen and P. H. Quanjer. Pollen deposition in intrathoracic airways. *European Respiratory Journal*, 4(3):359, 1991.

[107] M. Brauer and J. R. Brook. Personal and fixed-site ozone measurements with a passive sampler. *Journal of the Air & Waste Management Association*, 45(7): 529–537, 1995.

[108] J. A. Sarnat, K. W. Brown, J. Schwartz, B. A. Coull, and P. Koutrakis. Ambient gas concentrations and personal particulate matter exposures - implications for studying the health effects of particles. *Epidemiology*, 16(3):385–395, 2005.

[109] A. S. Geyh, J. P. Xue, H. Ozkaynak, and J. D. Spengler. The Harvard Southern California chronic ozone exposure study: Assessing ozone exposure of grade-school-age children in two Southern California communities. *Environmental Health Perspectives*, 108(3):265–270, 2000.

[110] K. Lee, W. J. Parkhurst, J. P. Xue, A. H. Ozkaynak, D. Neuberg, and J. D. Spengler. Outdoor/indoor/personal ozone exposures of children in Nashville, Tennessee. *Journal of the Air & Waste Management Association*, 54(3):352–359, 2004.

[111] L. J. S. Liu, P. Koutrakis, H. H. Suh, J. D. Mulik, and R. M. Burton. Use of personal measurements for ozone exposure assessment - a pilot-study. *Environmental Health Perspectives*, 101(4):318–324, 1993.

[112] W. S. Linn, D. A. Shamoo, K. R. Anderson, R. C. Peng, E. L. Avol, J. D. Hackney, and H. Gong. Short-term air pollution exposures and responses in Los Angeles area schoolchildren. *Journal of Exposure Analysis and Environmental Epidemiology*, 6(4):449–472, 1996.

[113] M. Castellano-Mendez, M. J. Aira, I. Iglesias, V. Jato, and W. Gonzalez-Manteiga. Artificial neural networks as a useful tool to predict the risk level of betula pollen in the air. *International Journal of Biometeorology*, 49(5): 310–316, 2005. 922rv Times Cited:33 Cited References Count:50.

[114] J. Nowosad. Spatiotemporal models for predicting high pollen concentration level of Corylus, Alnus, and Betula. *International Journal of Biometeorology*, 60(6):843–855, 2016.

[115] J. Nowosad, A. Stach, I. Kasprzyk, K. Chlopek, K. Dabrowska-Zapart, L. Grewling, M. Latalowa, A. Pedziszewska, B. Majkowska-Wojciechowska, D. Myszkowska, K. Piotrowska-Weryszko, E. Weryszko-Chmielewska, M. Puc, P. Rapiejko, and T. Stosik. Statistical techniques for modeling of Corylus, Alnus, and Betula pollen concentration in the air. *Aerobiologia*, 34(3):301–313, 2018.

[116] Olga Ritenberga, Mikhail Sofiev, Victoria Kirillova, Laimdota Kalnina, and Eugene Genikhovich. Statistical modelling of non-stationary processes of atmospheric pollution from natural sources: example of birch pollen. *Agricultural and Forest Meteorology*, 226-227:96 – 107, 2016. ISSN 0168-1923.

[117] Stuart J. Russell and Peter Norvig. *Artificial Intelligence - A Modern Approach (3. internat. ed.)*. Pearson Education, 2010. ISBN 978-0-13-207148-2.

[118] Ethem Alpaydin. *Introduction to Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2014. ISBN 978-0-262-02818-9.

[119] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *Acm Computing Surveys*, 31(3):264–323, 1999.

[120] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. 2000. ISBN 1886529094.

[121] Martijn van Otterlo and Marco Wiering. *Reinforcement Learning and Markov Decision Processes*, pages 3–42. 2012. ISBN 978-3-642-27645-3. doi: 10.1007/978-3-642-27645-3_1.

[122] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

[123] A. Holzinger. Data mining with decision trees: Theory and applications. *Online Information Review*, 39(3):437–438, 2015. doi: 10.1108/Oir-04-2015-0121.

[124] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984. ISBN 0-534-98053-8.

[125] R. M. Golden. Neural networks: A comprehensive foundation - haykin,s. *Journal of Mathematical Psychology*, 41(3):287–292, 1997.

[126] T. K. Ho. The random subspace method for constructing decision forests. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[127] T. K. Ho. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1:278–282 vol.1, 1995. doi: 10.1109/ICDAR.1995.598994.

[128] Małgorzata Puc. Artificial neural network model of the relationship between betula pollen and meteorological factors in Szczecin (Poland). *International Journal of Biometeorology*, 56(2):395–401, Mar 2012. ISSN 1432-1254.

[129] J. A. S. Mesa, C. Galan, and C. Hervas. The use of discriminant analysis and neural networks to forecast the severity of the poaceae pollen season in a region with a typical mediterranean climate. *International Journal of Biometeorology*, 49(6):355–362, 2005.

[130] USEPA. URL `https://aqs.epa.gov/aqsweb/airdata/download_files.html`.

[131] P. G. Georgopoulos and P. J. Lioy. From a theoretical framework of human exposure and dose assessment to computational system implementation: The modeling environment for total risk studies (MENTOR). *Journal of Toxicology and Environmental Health-Part B-Critical Reviews*, 9(6):457–483, 2006.

[132] Leonard Bielory, Kevin Lyons, and Robert Goldberg. Climate change and allergic disease. *Current Allergy and Asthma Reports*, 12(6):485–494, 2012. doi: 10.1007/s11882-012-0314-z.

[133] Marie-Claude Breton, Michelle Garneau, Isabel Fortier, Frédéric Guay, and Jacques Louis. Relationship between climate, pollen concentrations of *Ambrosia* and medical consultations for allergic rhinitis in Montreal, 1994–2002. *Science of The Total Environment*, 370(1):39–50, 2006.

[134] Sabit Cakmak, Robert E. Dales, Richard T. Burnett, Stan Judek, Frances Coates, and Jeffrey R. Brook. Effect of airborne allergens on emergency visits by children for conjunctivitis and rhinitis. *The Lancet*, 359(9310):947–948, 2002.

[135] James E. Neumann, Susan Anenberg, Kate R. Weinberger, Meredith Amend, Sahil Gulati, Allison Crimmins, Henry Roman, Neal Fann, and Patrick L. Kinney. Estimates of present and future asthma emergency department visits associated with exposure to oak, birch, and grass pollen in the United States. *GeoHealth*, 0(ja), 2018. doi: 10.1029/2018GH000153.

[136] Tiffany Duhl, Ronson Zhang, Alex Guenther, Serena Chung, Muhammad Salam, James House, R. C Flagan, Ed Avol, F. D Gilliland, Brian Lamb, Timothy Vanreken, Yuchao Zhang, and Eric Salathé. *The Simulator of the Timing and Magnitude of Pollen Season (STaMPS) model: a pollen production model for regional emission and transport modeling*, volume 6. 2013. doi: 10.5194/gmdd-6-2325-2013.

[137] T. L. Spero, C. G. Nolte, J. H. Bowden, M. S. Mallard, and J. A. Herwehe. The impact of incongruous lake temperatures on regional climate extremes downscaled from the CMIP5 archive using the WRF model. *Journal of Climate*, 29 (2):839–853, 2016.

[138] Keywan Riahi, Shilpa Rao, Volker Krey, Cheolhung Cho, Vadim Chirkov, Guenther Fischer, Georg Kindermann, Nebojsa Nakicenovic, and Peter Rafaj. RCP 8.5-a scenario of comparatively high greenhouse gas emissions. *Climatic Change*, 109(1):33, Aug 2011. ISSN 1573-1480.

[139] Ting Cai, Yong Zhang, Xiang Ren, Zhongyuan Mi, Leonard Bielory, Christopher G. Nolte, Shan He, and Panos G. Georgopoulos. Modeling spatiotemporal distributions of airborne alergenic pollen in the CMAQ-Pollen modeling system. *Science of The Total Environment*, In preparation.

[140] Daewon Byun and Kenneth L. Schere. Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (CMAQ) modeling system. *Applied Mechanics Reviews*, 59(2):51–77, 2006.

[141] Colin M Beale and Jack J Lennon. Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1586):247–258, 2012.

[142] Helen M Regan, Mark Colyvan, and Mark A Burgman. A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications*, 12(2):618–628, 2002.

[143] T. Karl and W. J. Koss. Regional and national monthly, seasonal, and annual temperature weighted by area, 1895–1983. *National Climatic Data Center, Asheville, NC, USA*, 1984.

[144] M. Thibaudon. The pollen-associated allergic risk in France. *Eur Ann Allergy Clin Immunol*, 35(5):170–2, 2003.

[145] P. Rapiejko, W. Stanlaewicz, K. Szczygielski, and D. Jurkiewicz. Threshold pollen count necessary to evoke allergic symptoms. *Otolaryngol Pol*, 61(4): 591–594, 2007.

[146] T. R. Cotos-Yáñez, F. J. Rodríguez-Rajo, A. Pérez-González, M. J. Aira, and V. Jato. Quality control in aerobiology: comparison different slide reading methods. *Aerobiologia*, 29(1):1–11, Mar 2013. ISSN 1573-3025.

[147] Max D Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991.

[148] I. R. Lake, N. R. Jones, M. Agnew, C. M. Goodess, F. Giorgi, L. Hamaoui-Laguel, M. A. Semenov, F. Solomon, J. Storkey, R. Vautard, and M. M. Epstein. Climate change and future pollen allergy in Europe. *Environmental Health Perspectives*, 125(3):385–391, 2017.

[149] B. Ferreira, H. Ribeiro, M. S. Pereira, A. Cruz, and I. Abreu. Effects of ozone in *Plantagolanceolata* and *Salixatrocinerea* pollen. *Aerobiologia*, 32(3):421–430, 2016.

[150] G. Chichiricco and P. Picozzi. Reversible inhibition of the pollen germination and the stigma penetration in *Crocusvernus* ssp *vernus* (iridaceae) following fumigations with NO_2, CO, and O_3 gases. *Plant Biology*, 9(6):730–735, 2007.

[151] S. Pasqualini, E. Tedeschini, G. Frenguelli, N. Wopfner, F. Ferreira, G. D'Amato, and L. Ederli. Ozone affects pollen viability and NAD(P)H oxidase release from *Ambrosiaartemisiifolia* pollen. *Environmental Pollution*, 159(10):2823–2830, 2011.

[152] H. Ribeiro, L. Duque, R. Sousa, A. Cruz, C. Gomes, J. E. da Silva, and I. Abreu. Changes in the IgE-reacting protein profiles of Acer negundo, Platanus x acerifolia and Quercus robur pollen in response to ozone treatment. *International Journal of Environmental Health Research*, 24(6):515–527, 2014.

[153] H. Ribeiro, L. Duque, R. Sousa, and I. Abreu. Ozone effects on soluble protein content of *Acernegundo*, *Quercusrobur* and *Platanus* spp. pollen. *Aerobiologia*, 29(3):443–447, 2013.

[154] Isabelle Beck, Susanne Jochner, Stefanie Gilles, Mareike McIntyre, Jeroen T. M. Buters, Carsten Schmidt-Weber, Heidrun Behrendt, Johannes Ring, Annette Menzel, and Claudia Traidl-Hoffmann. High environmental ozone levels lead to enhanced allergenicity of birch pollen. *PLOS ONE*, 8(11):e80147, 2013. doi: 10.1371/journal.pone.0080147.

[155] L.G. Cuinica, A. Cruz, I. Abreu, and J.C.G.E. da Silva. Effects of atmospheric pollutants (CO, O_3, SO_2) on the allergenicity of $Betulapendula, Ostryacarpinifolia$, and $Carpinusbetulus$ pollen. *International Journal of Environmental Health Research*, 25(3):312–321, 2015.

[156] J. Eckl-Dorna, B. Klein, T. G. Reichenauer, V. Niederberger, and R. Valenta. Exposure of rye (*Secalecereale*) cultivars to elevated ozone levels increases the allergen content in pollen. *Journal of Allergy and Clinical Immunology*, 126(6): 1315–1317, 2010. doi: 10.1016/j.jaci.2010.06.012.

[157] I. Cortegano, E. Civantos, E. Aceituno, A. del Moral, E. Lopez, M. Lombardero, V. del Pozo, and C. Lahoz. Cloning and expression of a major allergen from *Cupressusarizonica* pollen, Cup a 3, a PR-5 protein expressed under polluted environment. *Allergy*, 59(5):485–490, 2004.

[158] J. Storkey, P. Stratonovitch, D. S. Chapman, F. Vidotto, and M. A. Semenov. A process-based approach to predicting the effect of climate change on the distribution of an invasive allergenic plant in Europe. *Plos One*, 9(2), 2014.

[159] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

[160] USCB. United States Census 2010. http://www.census.gov/, 2012. URL `http://www.census.gov/`. Accessed March 28, 2014.

[161] Junfeng Zhang and Paul J. Lioy. Ozone in residential air: Concentrations, I/O ratios, indoor chemistry, and exposures. *Indoor Air*, 4(2):95–105, 1994.

[162] T. McCurdy, G. Glen, L. Smith, and Y. Lakkadi. The National Exposure Research Laboratory's consolidated human activity database. *Journal of Exposure Analysis and Environmental Epidemiology*, 10(6):566–578, Nov-Dec 2000.

[163] U.S. EPA. *Exposure Factors Handbook 2011 Edition*. U.S. Environmental Protection Agency, Washington, DC,, 2011 edition, 2011.

[164] M. Huynen, B. Menne, H. Behrendt, R. Bertollini, S. Bonini, R. Brandao, C. Brown-Fahrländer, B. Clot, C. d' Ambrosio, P. De Nuntiis, K.L. Ebi, J. Emberlin, E. Erdei Orbanne, C. Galán, S. Jäger, S. Kovats, P. Mandrioli, P. Martens, A. Menzel, B. Nyenzi, A. Rantio-Lehtimäki, J. Ring, O. Rybnicek, C. Traidl-Hoffmann, A.J.H. van Vliet, T. Voigt, S. Weiland, and M. Wickman. *Phenology and human health: allergic disorders*. Number 1 in Health and global environmental change. Health and global environmental change, 2003.

[165] V. Marecal, V. H. Peuch, C. Andersson, S. Andersson, J. Arteta, M. Beekmann, A. Benedictow, R. Bergstrom, B. Bessagnet, A. Cansado, F. Cheroux, A. Colette, A. Coman, R. L. Curier, H. A. C. D. van der Gon, A. Drouin, H. Elbern, E. Emili, R. J. Engelen, H. J. Eskes, G. Foret, E. Friese, M. Gauss, C. Giannaros, J. Guth, M. Joly, E. Jaumouille, B. Josse, N. Kadygrov, J. W. Kaiser, K. Krajsek, J. Kuenen, U. Kumar, N. Liora, E. Lopez, L. Malherbe, I. Martinez, D. Melas, F. Meleux, L. Menut, P. Moinat, T. Morales, J. Parmentier, A. Piacentini, M. Plu, A. Poupkou, S. Queguiner, L. Robertson, L. Rouil, M. Schaap, A. Segers, M. Sofiev, L. Tarasson, M. Thomas, R. Timmermans, A. Valdebenito, P. van Velthoven, R. van Versendaal, J. Vira, and A. Ung. A regional air quality forecasting system over Europe: the MACC-II daily ensemble production. *Geoscientific Model Development*, 8(9):2777–2813, 2015.

[166] M. Inatsu, S. Kobayashi, S. Takeuchi, and A. Ohmori. Statistical analysis on daily variations of birch pollen amount with climatic variables in Sapporo. *Sola*, 10:172–175, 2014.

[167] I. Kasprzyk and A. Walanus. Gamma, gaussian and logistic distribution models for airborne pollen grains and fungal spore season dynamics. *Aerobiologia*, 30 (4):369–383, 2014.

[168] L. Jantschi, D. Balint, and S. D. Bolboaca. Multiple linear regressions by maximizing the likelihood under assumption of generalized gauss-laplace distribution of the error. *Computational and Mathematical Methods in Medicine*, 2016.

[169] Francisco Javier Toro, Marta Recio, María del Mar Trigo, and Baltasar Cabezudo. Predictive models in aerobiology: data transformation. *Aerobiologia*, 14(2):179, Sep 1998. ISSN 1573-3025.

[170] E. Limpert, J. Burke, C. Galan, M. D. Trigo, J. S. West, and W. A. Stahel. Data, not only in aerobiology: how normal is the normal distribution? *Aerobiologia*, 24(3):121–124, 2008.

[171] Friedrich Recknagel. Applications of machine learning to ecological modelling. *Ecological Modelling*, 146(1):303 – 310, 2001. ISSN 0304-3800.

[172] Max Kuhn and Kjell Johnson. Applied predictive modeling. New York, Springer. 2013.

[173] The NCDC Climate Services Branch (CSB). *Global surface summary of day data*. URL `https://www7.ncdc.noaa.gov/CDO/cdoselect.cmd?datasetabbv=GSOD&countryabbv=&georegionabbv=&resolution=40`.

[174] National Allergy Bureau (NAB). *NAB pollen count charts*. URL `https://www.aaaai.org/global/nab-pollen-counts/reading-the-charts`.

[175] STARx Allergy and LLC Asthma Center. *New York and New Jersey daily pollen monitoring*. URL `http://www.nynjpollen.com/`.

[176] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex J. Smola, and Vladimir Vapnik. Support vector regression machines. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press, 1997.

[177] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

[178] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. URL <GotoISI>://WOS:000173361700001.

[179] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting - rejoinder. *Annals of Statistics*, 28(2):400–407, 2000. URL <GotoISI>://WOS:000089669700007.

[180] A. Gelman, A. Jakulin, M. G. Pittau, and Y. S. Su. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(4):1360–1383, 2008. URL <GotoISI>://WOS:000262731100012.

[181] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., 1995. ISBN 0198538642.

[182] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*, 2018. URL https://CRAN.R-project.org/package=caret. R package version 6.0-81.

[183] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL http://ggplot2.org.

[184] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2019. URL https://CRAN.R-project.org/package=e1071. R package version 1.7-0.1.

[185] Microsoft Corporation and Stephen Weston. *doSNOW: Foreach Parallel Adaptor for the 'snow' Package*, 2017. URL https://CRAN.R-project.org/package=doSNOW. R package version 1.0.16.

[186] Matt Dowle and Arun Srinivasan. *data.table: Extension of 'data.frame'*, 2019. URL https://CRAN.R-project.org/package=data.table. R package version 1.12.0.

[187] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9): 1–20, 2004. URL http://www.jstatsoft.org/v11/i09/.

[188] Andy Liaw and Matthew Wiener. Classification and regression by random forest. *R News*, 2(3):18–22, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

[189] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, and Yutian Li. *xgboost: Extreme Gradient Boosting*, 2019. URL https://CRAN.R-project.org/package=xgboost. R package version 0.81.0.1.

[190] Andrew Gelman and Yu-Sung Su. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2018. URL https://CRAN.R-project.org/package=arm. R package version 1.10-1.

[191] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4.

[192] Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2018. URL https://CRAN.R-project.org/package=rpart. R package version 4.1-13.

[193] Anthony G. Barnston. Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Weather and Forecasting*, 7, 12 1992.

[194] Yutaka Sasaki. *The truth of the F-measure*. 2007.

[195] David Powers. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*, volume 2. 2008.

[196] E. Kinnee, C. Geron, and T. Pierce. United States land use inventory for estimating biogenic ozone precursor emissions. *Ecological Applications*, 7(1): 46–58, 1997.

[197] Alison M Eyth and K Habisak. The mims spatial allocator: a tool for generating emission surrogates without a geographic information system. *Proceedings, 12th International Emission Inventory Conference, San Diego*, 2003.

[198] Andrea Saltelli, Karen Chan, and E. M. Scott. *Sensitivity analysis*, volume 134. 2000.

[199] P. Brose, D. Van Lear, and R. Cooper. Using shelterwood harvests and prescribed fire to regenerate oak stands on productive upland sites. *Forest Ecology and Management*, 113(2-3):125–141, 1999.

[200] M. Chamecki, C. Meneveau, and M. B. Parlange. Large eddy simulation of pollen transport in the atmospheric boundary layer. *Journal of Aerosol Science*, 40(3):241–255, 2009.

[201] M. M. Foster, P. M. Vitousek, and P. A. Randolph. Effects of ragweed ($Ambrosia-artemisiifolia$ L.) on nutrient cycling in a 1st-year old-field. *American Midland Naturalist*, 103(1):106–113, 1980.

[202] B. Fumanal, B. Chauvel, and F. Bretagnolle. Estimation of pollen and seed production of common ragweed in France. *Annals of Agricultural and Environmental Medicine*, 14(2):233–236, 2007.

[203] Y. S. Wang, D. R. Miller, J. M. Welles, and G. M. Heisler. Spatial variability of canopy foliage in an oak forest estimated with fisheye sensors. *Forest Science*, 38(4):854–865, 1992.

[204] W. Deen, C. J. Swanton, and L. A. Hunt. A mechanistic growth and development model of common ragweed. *Weed Science*, 49(6):723–731, 2001.

[205] Michael D. Martin, Marcelo Chamecki, and Grace S. Brush. Anthesis synchronization and floral morphology determine diurnal patterns of ragweed pollen dispersal. *Agricultural and Forest Meteorology*, 150(9):1307–1317, 2010.

[206] John H. Seinfeld and Spyros N. Pandis. *Atmospheric chemistry and physics: from air pollution to climate change*. 1997. ISBN 0471178152.

[207] Margaret B. Davis and Linda B. Brubaker. Differential sedimentation of pollen grains in lakes. *Limnology and Oceanography*, 18(4):635–646, 1973.

[208] L. Ziska, K. Knowlton, C. Rogers, D. Dalan, N. Tierney, M. A. Elder, W. Filley, J. Shropshire, L. B. Ford, C. Hedberg, P. Fleetwood, K. T. Hovanky, T. Kavanaugh, G. Fulford, R. F. Vrtis, J. A. Patz, J. Portnoy, F. Coates, L. Bielory, and D. Frenz. Recent warming by latitude associated with increased length of ragweed pollen season in Central North America. *Proceedings of the National Academy of Sciences of the United States of America*, 108(10):4248–4251, 2011.

[209] K. A. Stinson, J. M. Albertine, L. M. S. Hancock, T. G. Seidler, and C. A. Rogers. Northern ragweed ecotypes flower earlier and longer in response to elevated CO_2: what are you sneezing at? *Oecologia*, 182(2):587–594, 2016.

[210] B. Šikoparija, G. Mimić, M. Panić, O. Marko, P. Radišić, T. Pejak-Šikoparija, and A. Pauling. High temporal resolution of airborne ambrosia pollen measurements above the source reveals emission characteristics. *Atmospheric Environment*, 192:13–23, 2018.

[211] J. Zhang and Y. Shao. A new parameterization of particle dry deposition over rough surfaces. *Atmospheric Chemistry and Physics*, 14(22):12429–12440, 2014.

[212] L. Zhang and Z. He. Technical note: An empirical algorithm estimating dry deposition velocity of fine, coarse and giant particles. *Atmospheric Chemistry and Physics*, 14(7):3729–3737, 2014.

Appendix A
LIST OF ACRONYMS

AAAAI	American Academy of Allergy, Asthma & Immunology
AAD	Allergic Airway Diseases
ANN	Artificial Neural Networks
BayesGLM	Bayesian Generalized Linear Model
BC	Boundary Condition
BELD3.1	Biogenic Emissions Landuse Database, version 3.1
BEIS	Biogenic Emission Inventory System
CART	Classification and Regression Tree
CESM	Community Earth System Model
CCTM	CMAQ Chemical Transport Model
CHAD	Consolidated Human Activity Database
CMAQ	Community Multiscale Air Quality
CONUS	CONtiguous United States
DMA8[O_3]	Daily Maximum 8-hour Average Ozone
GCM	General Circulation Model
GDD	Growing Degree Days
GHG	GreenHouse Gases
IC	Initial Condition
IPCC	Intergovernmental Panel on Climate Change
LULC	Land Use and Land Coverage
MCIP	Meteorology-Chemistry Interface Processor
MENTOR	Modeling Environment for Total Risk studies

MM5	The Fifth Generation Mesoscale Model
MAE	Mean Absolute Error
NAB	National Allergy Bureau
NJDEP	New Jersey Department of Environmental Protection
NOAA	National Oceanic and Atmospheric Administration
PDIRs	Physiological Daily Inhalation Rates
PM	Particulate Matter
RCPs	Representative Concentration Pathways
RMSE	Root Mean Square Error
SD	Start Date
SL	Season Length
SMOKE	Sparse Matrix Operator Kernel Emissions
SVM	Support Vector Machines
WRF	Weather Research and Forecasting model
USEPA	US Environmental Protection Agency
XGBoost	eXtreme Gradient Boosting

Appendix B
SUPPLEMENT DATA FOR CHAPTER 2

B.1 Pollen Emission Model

The pollen emission flux in a modeling grid with area of S_g was calculated through Equation B.1,

$$F_g = F_e S_g P_c \tag{B.1}$$

where P_c (%) is the percentage of area coverage of allergenic plants in the corresponding modeling grid. The upward emission flux F_e on a unit surface is derived using mass balance of pollen grain fluxes in the near surroundings of allergenic plants [97]. It is formulated using Equation B.2,

$$F_e = \frac{q_p L_d L_h (K_e LAI + C_r K_r (1 + LAI))}{1 + v_d (1 + LAI)/u_*} \tag{B.2}$$

where q_p is the annual total emission flux. L_d and L_h are the daily and hourly flowering likelihood, respectively.K_e and K_r (dimensionless) are the lumped meteorology adjustment factors for direct emission and resuspension fluxes, respectively. LAI is leaf area index. C_r is a proportional factor to relate the resuspension to direct emission flux. u_* and v_d are characteristic velocity and deposition velocity, respectively. All these terms on the right side of Equation B.2, can either be measured, or parameterized and approximated through measurable factors. Details of the derivation and parameterization of the emission model are presented in Cai *et al.* [97]. The emission fluxes occur only in the first model layer, up to 60 meters above the ground.

B.1.1 Coverage of oak and ragweed

The coverages in 2004 and 2047 were considered the same for each of oak and ragweed. The area coverage for oak was derived from the Biogenic Emissions Land Use Database, version 3.1 (BELD3.1) [196]. The area coverage for oak across the CONUS was generated using Spatial Allocator to redistribute the BELD3.1 data with 1-km grid resolution into 36-km grid resolution [197]. The area coverage for ragweed was generated using an algorithm developed on the basis of observed ragweed pollen counts and vegetation coverage information from BELD3.1 [97]. It was found that the mean annual production of ragweed pollen was mainly relevant to area coverages of grass land, crop grass land, and savanna land. The estimation of ragweed plant coverage in a cell of the modeling grid was generated using:

$$P_R = b_G P_G + b_{CG} P_{CG} + b_{Sa} P_{Sa} \tag{B.3}$$

where P_G, P_{CG} and P_{Sa} (%) are the area coverage of grass land, crop grass land, and savanna land, respectively. b_G, b_{CG} and b_{Sa} (dimensionless) are the corresponding coefficients (Table B.1). The coefficient b_G represents roughly the fraction of grass land area occupied by ragweed plants, likewise for other coefficients.

Table B.1: Coefficients used to calculate the area coverage of ragweed.

Land Class **Coefficient**	**Grass** b_G **(unitless)**	**Crop Grass** b_{CG} **(unitless)**	**Savanna** b_{Sa} **(unitless)**
Ragweed	0.7684	0.5000	0.7497

B.1.2 Sensitivity analysis

Global sensitivity analyses were performed to test the sensitivity of the pollen emission model to multiple inputs and parameters based on Morris's design [147]. This design estimates the main effect of a parameter by computing a number of local sensitivities at random points of the parameter space. The mean of these randomized

local sensitivities indicates the overall influence of a given parameter on the output metric, while the corresponding standard deviation indicates the effects of interaction and nonlinearity [198].

The regional mean hourly emission (F_{hrMn}) was selected as a metric for testing the emission model's sensitivity to multiple inputs and parameters. The definition of this metric is presented in Equation B.4,

$$F_{hrMn} = \frac{\sum_{i,j,t} F_g(i,j,t)}{N_i N_j N_t} \tag{B.4}$$

where N_i and N_j are the values of spatial indices i and j, respectively.

In the current study, each of the 23 parameters (Table B.2) was sampled 6,000 times according to Morris' method from 250 random trajectories (each has 24 steps) in the parameter space [147, 198]. Each of the parameters was perturbed between 50% and 150% of its base value or distribution while keeping other parameters unchanged. Equation B.5 was used to calculate the Normalized Sensitivity Coefficient (NSC) for regional hourly mean emission at a local point:

$$NSC_{hrMn} = \frac{\Delta F_{hrMn}/F_{hrMn}}{\Delta P/P} \tag{B.5}$$

where F_{hrMn} and P are the regional mean hourly emission flux and the input parameter, respectively; and ΔF_{hrMn} and ΔP are the perturbations in the emission flux and input parameters, respectively.

The global NSC of a parameter, NSC_g, is defined as the mean of the corresponding local sensitivities. The average absolute global NSC, $\overline{|NSC_g|}$, for each parameter and pollen taxon can be derived based on means of the absolute NSC_g. Similarly, the standard deviations averaged over each parameter and pollen taxon ($\overline{STD}$) can be obtained to evaluate the interaction and nonlinearity effect of input parameters on modeling output.

Table B.2: Parameters for pollen emission model. These parameters were derived from the literature, and also used for global sensitivity analysis.

Parameter and ID	**Oak**	**Ragweed**
1 H_c, plant height (m)	30 [a]	0.69 [b]
2 C_r, proportional factor (unitless)	0.7 [c]	0.7 [c]
3 q_p, annual emission flux (pollen grain /(m^2 yr))	1.0×10^9 [d]	2.8×10^9 [e, f]
4 LAI, leaf area index (m^2/m^2)	3.4 [h]	1.2 [i]
5 u_*, friction velocity (m/s)	WRF data [j]	WRF data [j]
6 c_{Te}, c_{Ue}, c_{Ve}, correction factor for direct emission (unitless)	1 [c]	1 [c]
7 c_{Ur}, c_{Vr}, correction factor for resuspension (unitless)	1 [c]	1 [c]
8 T_{te}, threshold temperature for direct emission (ºC)	10 [c]	0 [k]
9 U_{te}, threshold relative humidity for direct emission (%)	60 [c]	60 [c]
10 V_{te}, threshold velocity for direct emission (m/s)	2.65 [c]	2.9 [k]
11 U_{tr}, threshold relative humidity for resuspension (%)	85[c]	85 [c]
12 V_{tr}, threshold velocity for resuspension (m/s)	0.9 [c]	0.9 [c]
13 r_a, aerodynamic resistance (hr/m)	calculated [l]	calculated [l]
14 r_b, quasi-laminar resistance (hr/m)	calculated [l]	calculated [l]
15 v_s, settling velocity (m/s)	calculated [l]	calculated [l]
16 P_c, percentage of area coverage (%)	BELD 3.1	calculated [l]
17 L_d, daily flowering likelihood (%)	calculated [l]	calculated [l]
18 L_h, hourly flowering likelihood (%)	calculated [l]	Literature [m]
19 u_{*t}, threshold friction velocity (m/s)	calculated [l]	calculated [l]
20 z_0, surface roughness (m)	10 [n]	0.1 [n]
21 d_p, diameter of pollen grain (μm)	28 [p]	18 [p]
22 ρ_p, density of pollen grain (kg/m^3)	1200 [p]	1280 [p]
23 C_c, slip correction factor (unitless)	1.008 [p]	1.008 [p]

[a]Brose *et al.* [199], [b]Chamecki *et al.* [200], [c]Helbig *et al.* [76], [d]Schueler and Schlünzen [83],
[e]Foster *et al.* [201], [f][202], [h]Wang *et al.* [203], [i]Deen *et al.* [204],
[j]Nolte *et al.* [65], Spero *et al.* [137], [k]Zink *et al.* [78], [l]Cai *et al.* [97],
[m]Martin *et al.* [205], [n]Seinfeld and Pandis [206], [p]Davis and Brubaker [207]

B.2 Spatiotemporal Distribution of Pollen Emission

To examine the temporal pattern of pollen emissions, the mean hourly emission fluxes over the early and late flowering period for oak and ragweed are plotted in Figure B.1. Oak pollen emissions started from the Southern CONUS in March and then shifted gradually toward the Northern CONUS in April. In contrast to oak, ragweed

pollen emissions started from the Northern CONUS in August and then shifted gradually toward the Southern CONUS in September. This pattern is consistent with long term observations [10, 208], and is simulated for the first time in this study. It has been identified that summer-flowering ragweed has earlier flower initiation at high latitudes than lower latitudes, which is a typical adaption to the environment to maximize reproductive success [209]. The time slices of pollen emissions in Figure B.2 display the diurnal variation of pollen emission. In general, the oak pollen emission flux in each cell of the modeling grid at 11:00 UTC is higher than that at 18:00 UTC (averaged over Apr 21-Apr 30, 2004), and the ragweed pollen emission flux in each cell of the modeling grid at 14:00 UTC is higher than that at 18:00 UTC (averaged over Sept 21-Sept 30, 2004), which is mainly caused by variation of hourly flowering likelihood. The emission model developed in this study demonstrated the daily and hourly variation of pollen emission flux due to the intrinsic physiological characteristics of the plants and influences of meteorological factors such as temperature, wind velocity and relative humidity.

Figure B.3 depicts the spatial patterns of oak pollen emissions during the entire pollen season in 2004. The spatial patterns were examined for four metrics: mean, maximum, seasonal total, and standard deviation of hourly emissions at each cell of the modeling grid covering the CONUS. These four metrics were calculated based on the simulated hourly emissions of oak pollen between 1 March 2004 and 30 April 2004. The spatial patterns of mean, maximum, seasonal and standard deviation of hourly emission flux all roughly follow the pattern of area coverage of oak trees shown in Figure 2.3. The oak pollen emissions varied substantially in different climate regions. The seasonal total oak pollen emissions in the Southeast, South and Central climate regions were higher than those in other regions in the CONUS. The mean hourly oak emission flux can increase up to $9\text{x}10^5$ pollen/(m^2 h). The maximum hourly oak pollen emission flux was $5.8\text{x}10^6$ pollen/(m^2 h). As shown in Figure 7, the spatial patterns of ragweed emission flux also follow the patterns of ragweed area coverage.

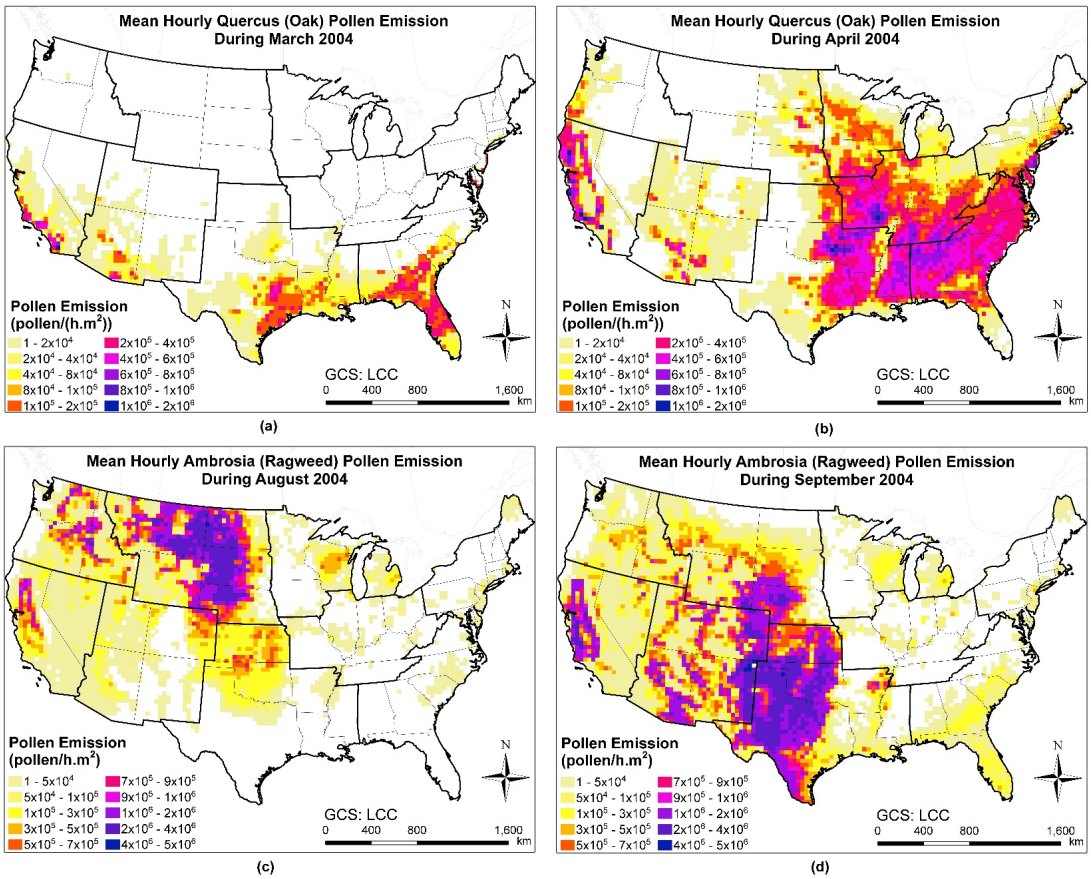


Figure B.1: Spatial patterns of mean hourly emission of (a) oak pollen in March 2004; (b) oak pollen in April 2004; (c) ragweed pollen in August 2004; and (d) ragweed pollen in September 2004.

The South and West North Central climate regions had the highest ragweed seasonal total emission of 4x10^9 pollen/m^2. The standard deviations of the hourly pollen emission flux throughout the season were caused by variations in daily and hourly flowering likelihood, and the meteorological factors contributing to pollen emission. The mean hourly ragweed pollen emission can reach up to 2.4x10^6 pollen/(m^2 h). The magnitude of the simulated pollen emission flux is comparable to that reported by Šikoparija *et al.* [210], which were obtained through an empirical study. The observed average hourly ragweed pollen emission flux were derived from direct measurements of airborne pollen concentrations in the field and ranged from 2.7x10^6 pollen/(m^2 h) to 3.2x10^8 pollen/(m^2 h).

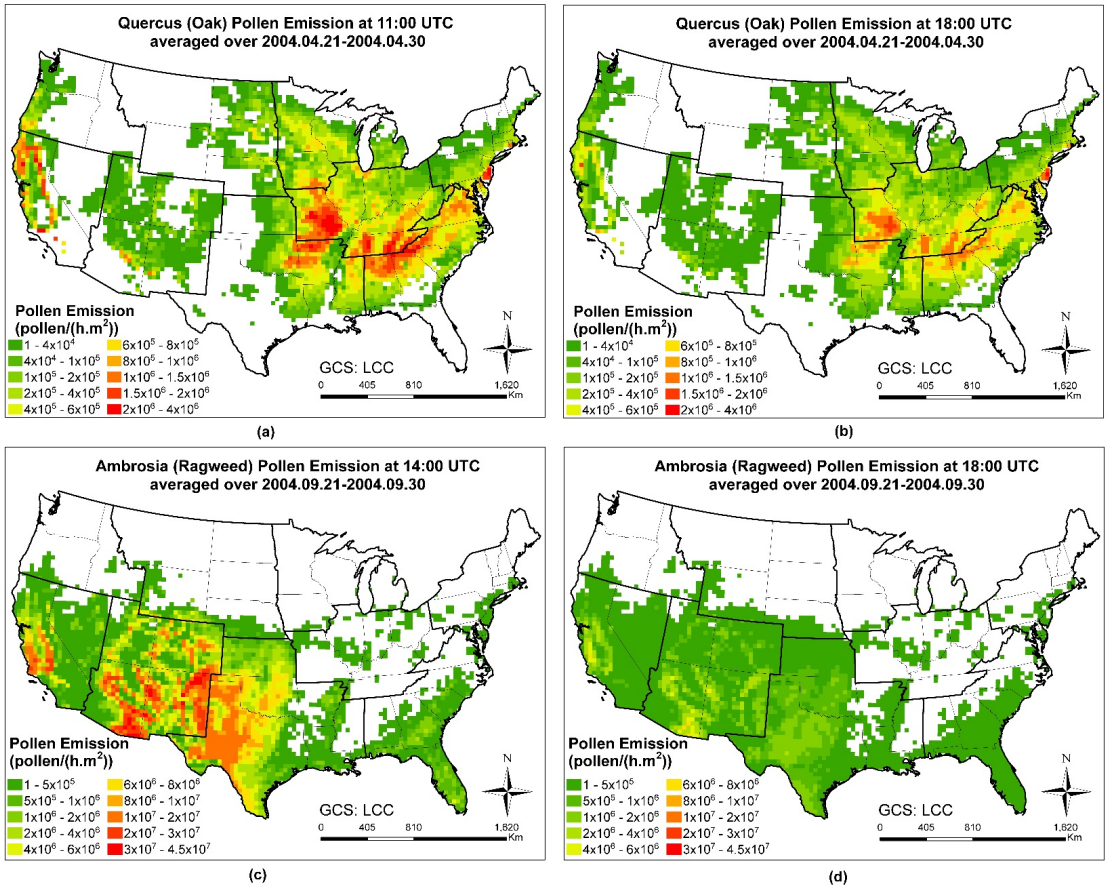


Figure B.2: Time slices of spatiotemporal emission profiles of (a) oak pollen at 11:00 UTC (averaged over April 21-April 30, 2004); (b) oak pollen at 18:00 UTC (averaged over April 21-April 30, 2004); (c) ragweed pollen at 14:00 UTC (averaged over September 21-September 30, 2004); and (d) ragweed pollen at 18:00 UTC (averaged over September 21-September 30, 2004).

B.2.1 Sensitivity Analysis of the emission model

The global sensitivity of the simulated regional mean hourly pollen emissions to different parameters is presented in Figure B.5. The global NSC of all parameters, except the density of oak pollen grain (ρ_p), varied within -0.1 and 0.1 for pollen emissions from oak. The average absolute global NSC of density of oak pollen grain (ρ_p) is 0.1311. The ragweed pollen emission model is also robust to 22 of the 23 parameters (-0.1 $<$ global NSC $<$ 0.1), but more sensitive to diameter of pollen grain (d_p), with an average absolute global NSC of 0.1438.

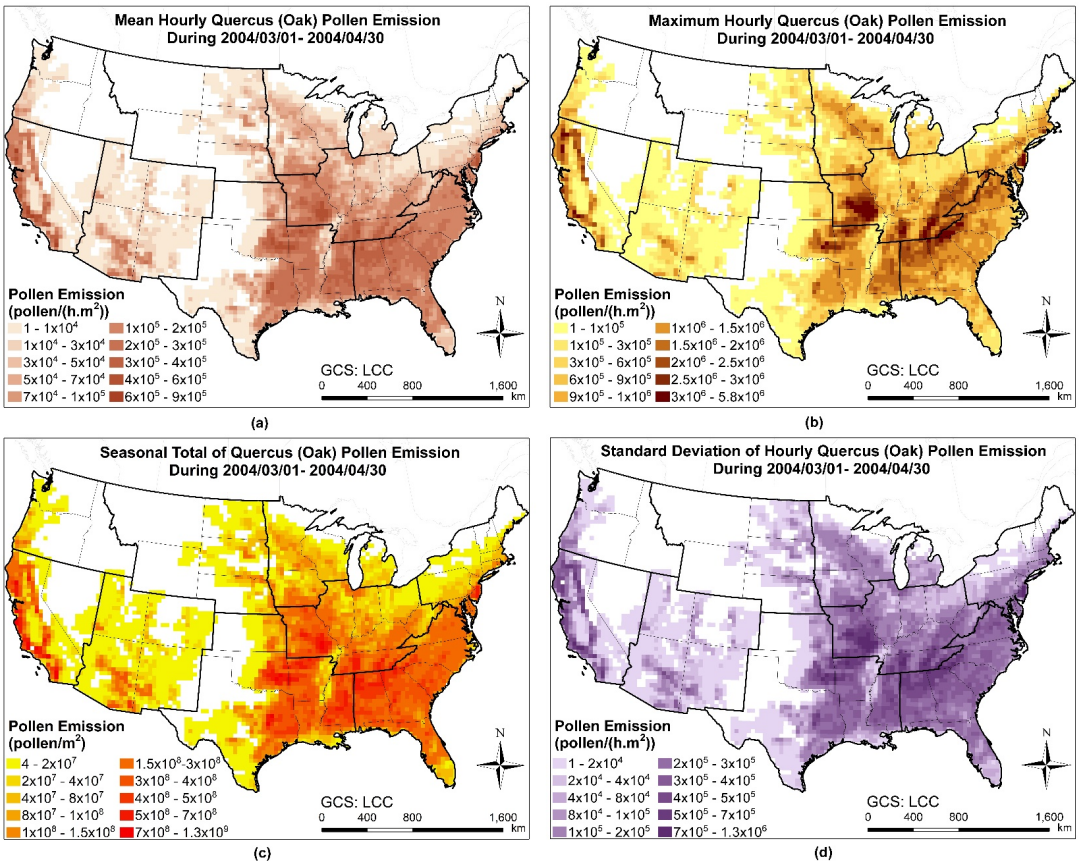


Figure B.3: Spatial patterns of mean, maximum, seasonal total and standard deviation of hourly emission of oak pollen. (a) Hourly mean, (b) Hourly maximum, (c) Seasonal total, and (d) Standard deviation.

The standard deviations of $NSCs$ for pollen emissions of oak and ragweed were between 0.4626 and 0.6836. This indicated low interaction and nonlinearity effects among parameters for pollen emissions of oak and ragweed.

Uncertainties in sensitive and interactive input parameters can result to large deviations of model predictions. In particular, we acknowledge that there are substantial uncertainties in dry deposition velocity, plant area coverage, flowering likelihood, and our assumptions of quasi-steady state and quasi-equilibrium balance of pollen fluxes. A new parameterization of deposition velocity has been developed similar to that in Zhang and Shao [211] and Zhang and He [212]. Particularly, deposition velocity for large particles such as pollen still has substantial uncertainties, and will need to be

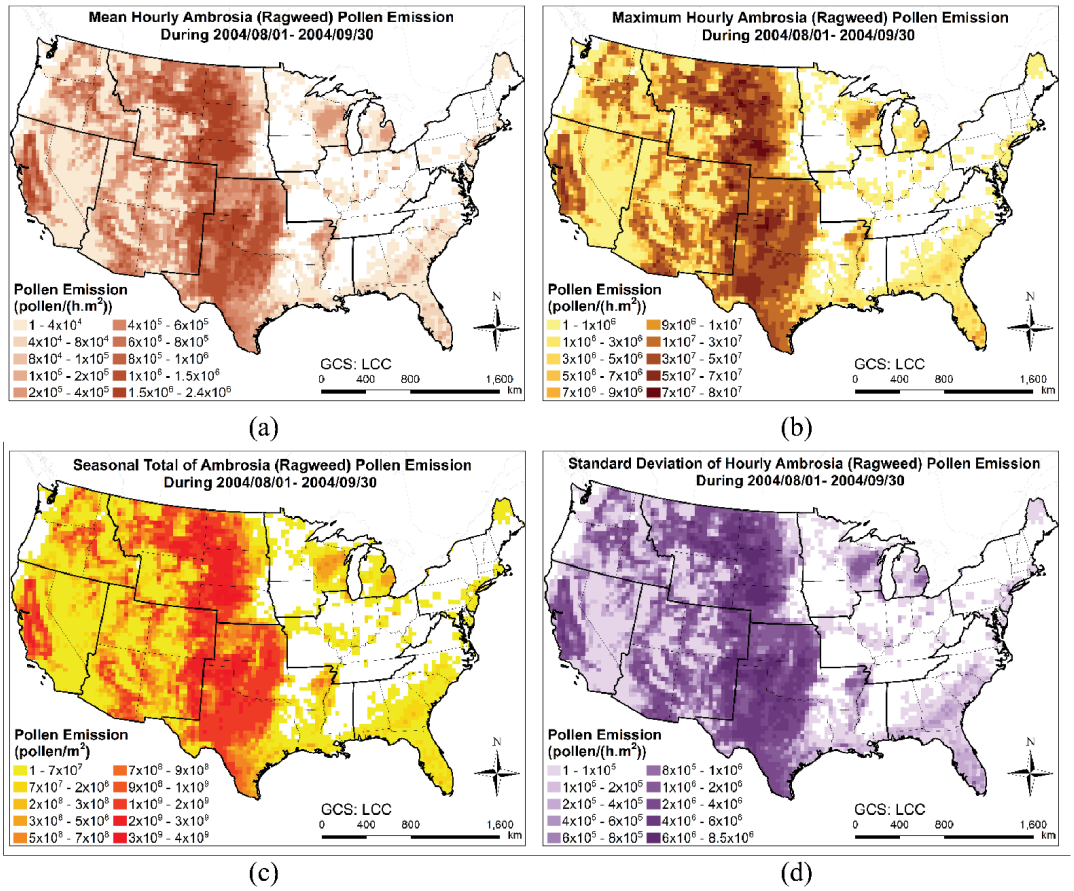


Figure B.4: Spatial patterns of mean, maximum, seasonal total and standard deviation of hourly emission of ragweed pollen. (a) Hourly mean, (b) Hourly maximum, (c) Seasonal total, and (d) Standard deviation.

further investigated. Plant area coverages calculated using land use data in combination with annual pollen counts seem more reasonable than those calculated based on plant inventory or ecological model [72].

B.2.2 Evaluation of the emission model

To investigate the accuracy of the simulated pollen emissions with respect to plant coverage, scatterplots of normalized coverage and seasonal total emission for oak and ragweed pollen are shown in Figure B.6. The Pearson correlation coefficients for oak is 0.813 (p-value < 0.0001), and 0.892 (p-value < 0.0001) for ragweed. The data points should fall on or near the 45-degree line in Figure B.6 if there is linear

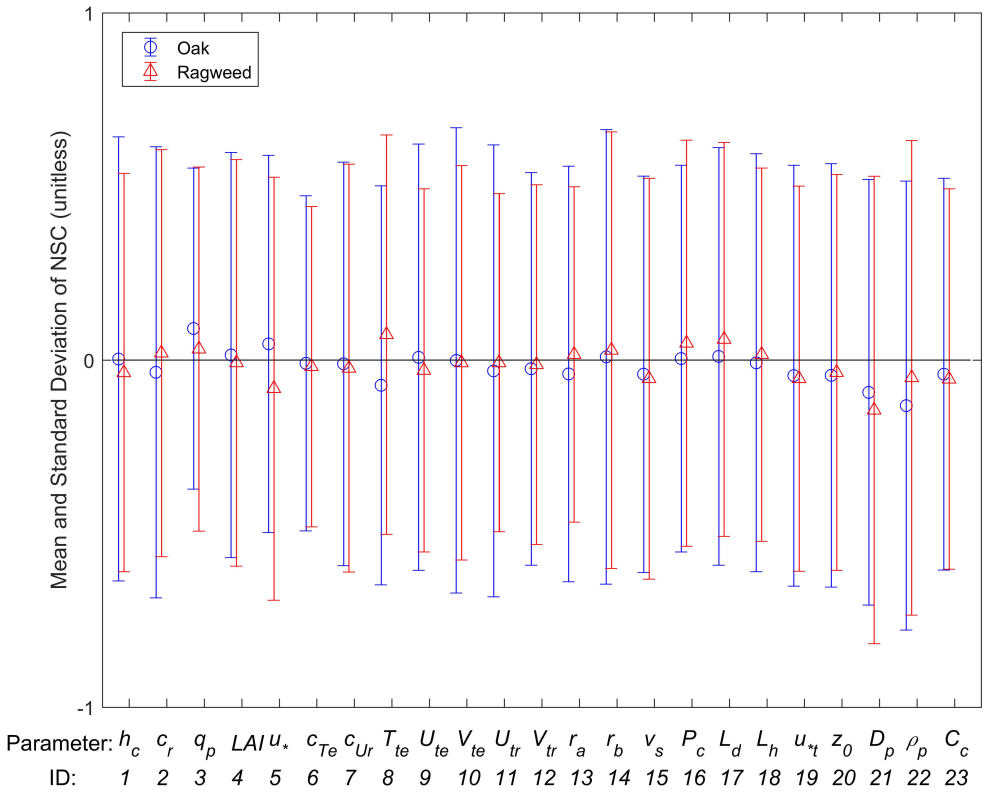


Figure B.5: Mean and standard deviation of Normalized Sensitivity Coefficient (NSC) for each parameter for the pollen emission model of oak and ragweed. All parameters are described in Table B.2.

relationship between the two normalized variables. The deviations from the diagonal line are caused by various factors such as temperature, humidity, wind speed, etc. that were fully considered in the emission model. The correlation of observed annual total pollen counts at selected pollen monitoring stations with the corresponding simulated seasonal total emission is illustrated in Figure B.7. Due to the incompleteness of pollen observations at the monitoring stations in 2004, only stations with more than 20 days of valid observations were selected. Data from 36 monitoring stations for oak show a Pearson correlation coefficient of 0.421 (p-value is 0.0105), while data from 34 monitoring stations for ragweed result in a Pearson correlation coefficient of 0.787 (p-value < 0.0001). Pollen counts at monitoring stations are affected not only by pollen emission, but also by atmospheric transport of pollen, which is incorporated

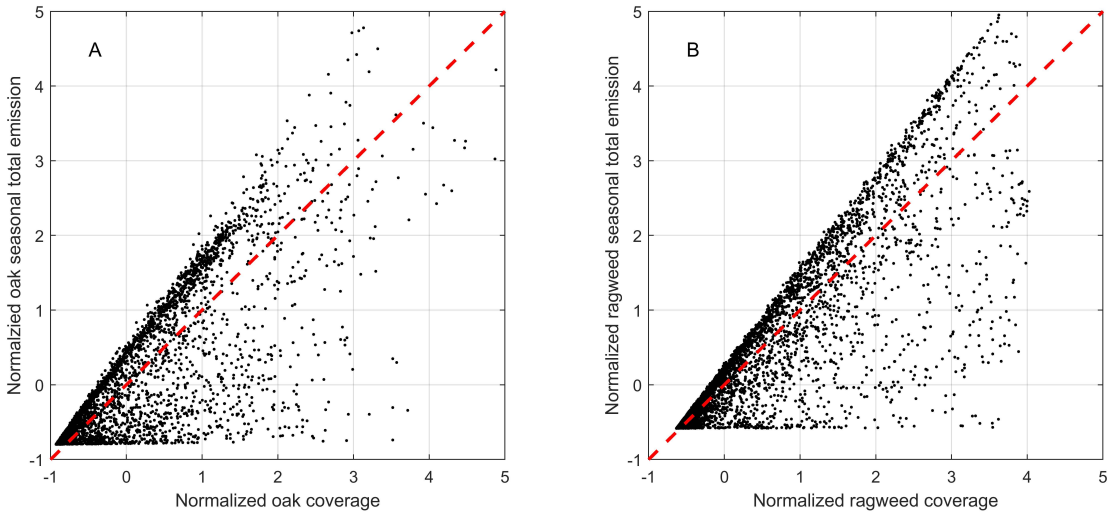


Figure B.6: Scatterplot of normalized coverage and normalized seasonal total emission in 2004 for oak and ragweed pollen with 45-degree line.

and simulated in the new pollen transport model that was developed and tested by our group.

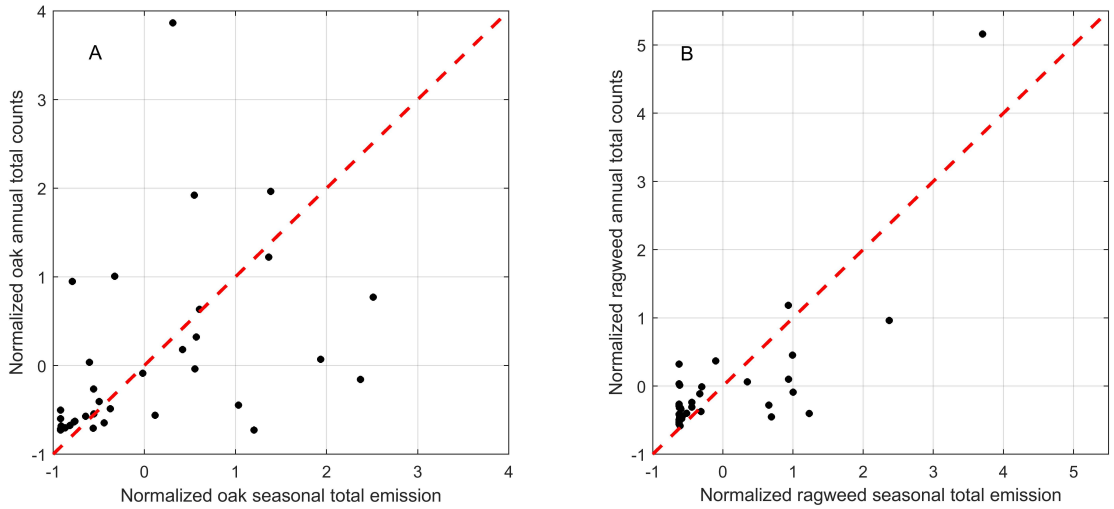


Figure B.7: Scatterplot of normalized annual pollen counts observation and normalized seasonal total emission in 2004 for oak and ragweed pollen with 45-degree line.

B.3 Pollen Transport Model

B.3.1 Calculation of hit and false rates

The simulated daily pollen concentrations, and sum of daily pollen concentrations during the pollen season (hereafter referred as seasonal count) in 2004 were first paired with corresponding observations. For example, the observed daily pollen concentrations at a monitor station was paired with the simulated daily pollen concentrations in a grid cell that contains the corresponding pollen monitoring station; similar pairings were performed for seasonal count. As shown in Equation B.6,

$$C(Day, i, j) = \frac{\sum_{hr \in Day} C(hr, 1, i, j)}{24} \tag{B.6}$$

the simulated daily pollen concentration at a given day in a grid cell (i, j), $C(Day, i, j)$ is defined as the daily average concentrations derived from the simulated hourly concentrations on the model's lowest layer (i.e., layer 1) because observations of pollen counts are generally made near the surface. The model's lowest layer on average extends from 0 to 60 m above the ground. Hit and false rates were calculated for

Table B.3: The confusion matrix for calculating hit rate and false rate.

Greater or equal to concentration level i (i.e., $\geq C_{Li}$)		Observation	
		True	False
Prediction	True	TP_i	FP_i
	False	FN_i	TN_i

evaluation of the simulated daily pollen concentration. Procedures from the literature were followed to calculate the hit and false rates at three different concentration levels [84, 78], which are 10, 50 and 100 pollen grains/m^3, respectively. Table B.3 lists the confusion matrix representing the number of cases of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) of observed and predicted pollen concentrations for a given concentration level. The hit rate (Hi) and false rate (Fi) for a given pollen concentration level C_{Li} are defined using Equation B.7 ,

$$\begin{cases} H_i = \frac{TP_i}{TP_i+FN_i} \\ F_i = \frac{FP_i}{TP_i+FP_i} \end{cases} \tag{B.7}$$

Both hit and false rates are defined based on observations and simulations day by day. If both the observation and simulation for a given day at a given station are equal to or higher than threshold value level i (i.e., C_{Li}), it is counted as a hit (i.e. True Positive) for level i at this station. If the observation is below C_{Li} and the simulation is equal to or greater than C_{Li}, it is counted as a false (i.e., False Positive). The hit rate indicates among the observed airborne concentrations greater than or equal to C_{Li}, how many are correctly predicted. The false rate indicates among the predicted airborne concentrations greater than or equal to C_{Li}, how many are falsely predicted.

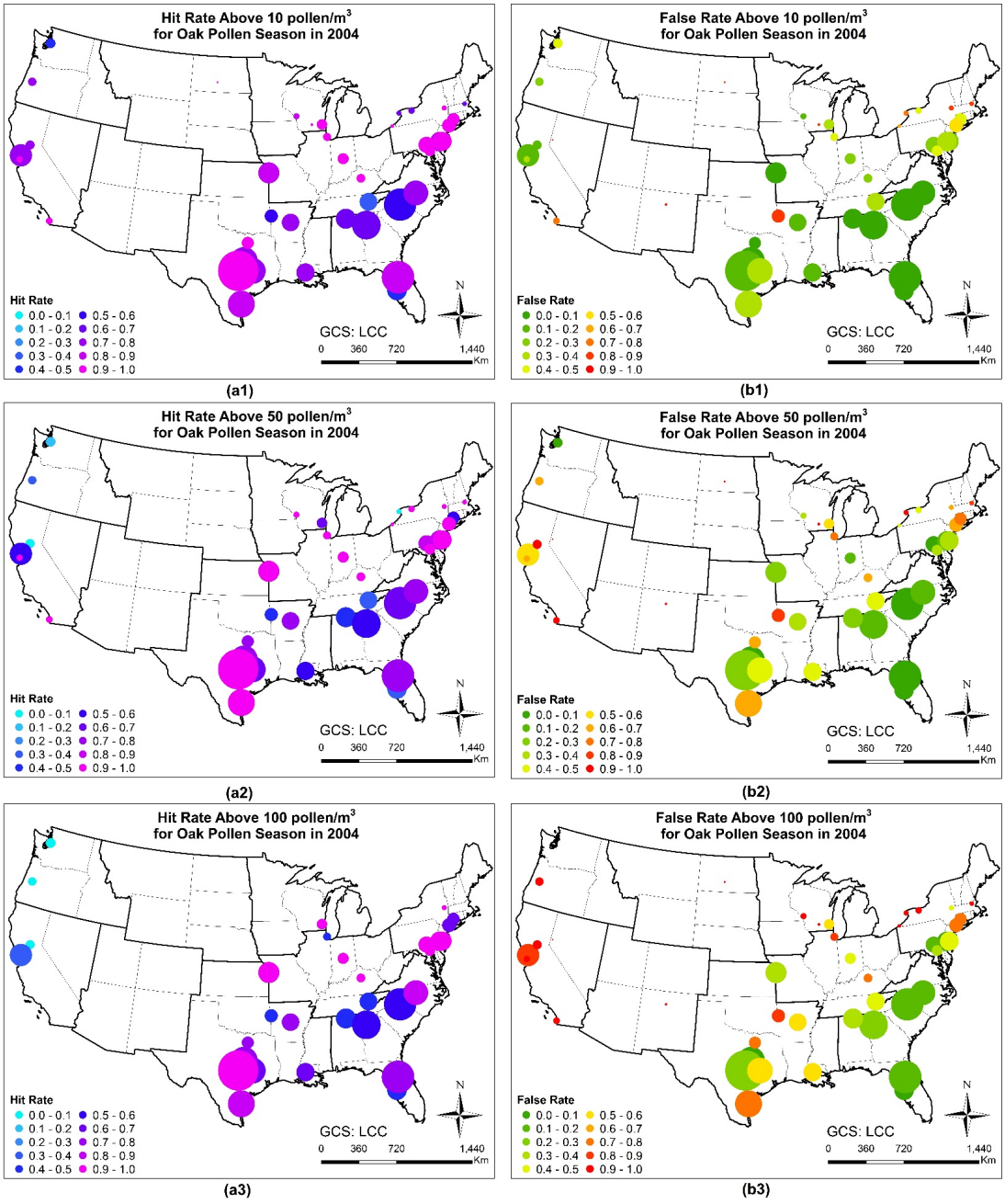


Figure B.8: Hit and false rates for predicted and observed daily oak pollen concentration during 2004 across the CONUS. The size of the circle indicates the oak pollen abundance at that station. (a1-a3): Hit rates for 10, 50 and 100 pollen grains/m^3, respectively; (b1-b3): False rates for 10, 50 and 100 pollen grains/m^3, respectively.

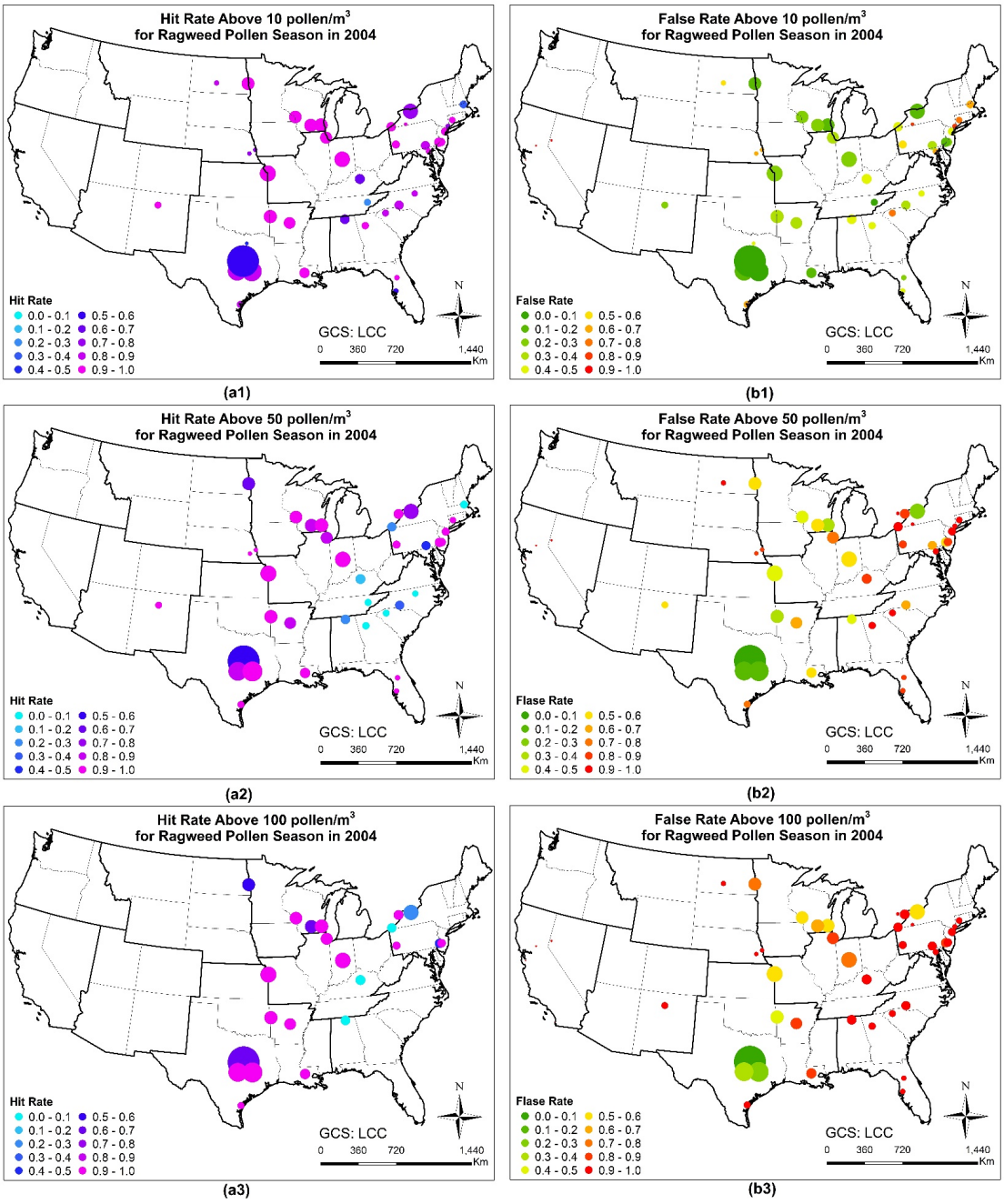


Figure B.9: Hit and false rates for predicted and observed daily ragweed pollen concentration during 2004 across the CONUS. The size of the circle indicates the ragweed pollen abundance at that station. (a1-a3): Hit rates for 10, 50 and 100 pollen grains/m^3, respectively; (b1-b3): False rates for 10, 50 and 100 pollen grains/m^3, respectively.

Appendix C

SUPPLEMENTARY DATA FOR CHAPTER 3

Table C.1: Coordinates, elevations, main climate characteristics and years of data for the pollen stations in this study.

ID	Station Name	Latitude (°N)	Longitude (°W)	Elevation (m)	Mean Temp. (°C)	Annual Precip. (mm)	Years of Data (Yrs.)	
							Oak	Ragweed
1	Seattle, WA	47.66	122.29	20	11.9	603	13	-
2	Fargo, ND	46.84	96.87	277	5.9	569	11	12
3	Vancouver, WA	45.62	122.50	89	12.3	960	4	-
4	Eugene, OR	44.04	123.09	129	11.3	1065	13	-
5	LaCrosse, WI	43.88	91.19	216	9.0	905	10	9
6	Rochester, NY	43.10	77.58	148	9.3	878	14	14
7	Niagara Falls, ON , CA	43.09	79.09	188	9.3	893	-	4
8	Madison, WI	43.08	89.43	263	8.7	909	7	7
9	Waukesha, WI	43.02	88.24	270	9.6	557	6	6
10	London, ON, CA	42.99	81.25	250	8.3	476	4	4
11	Albany, NY	42.68	73.77	72	9.4	992	5	-
12	Chelmsford, MA	42.60	71.35	37	10.0	814	9	8
13	St. Clair Shores, MI	42.51	82.9	180	9.8	863	6	7
14	Salem, MA	42.50	70.92	42	10.9	1082	10	10
15	Erie, PA	42.10	80.13	215	10.1	1002	9	13
16	Olean, NY	42.09	78.43	433	7.3	974	8	14
17	Chicago, IL	41.91	87.77	189	11.0	617	7	7
18	Waterbury, CT	41.55	73.07	140	11.8	665	10	10
19	Omaha, NE	41.14	95.97	305	10.9	854	12	13
20	Armonk, NY	41.13	73.73	187	11.1	865	7	7
21	Lincoln, NE	40.82	96.64	371	11.0	699	4	5
22	Newark, NJ	40.74	74.19	43	13.0	1213	10	12
23	Pittsburgh, PA	40.47	79.95	287	11.2	858	5	7
24	Philadelphia, PA	39.96	75.16	12	13.5	1106	11	10
25	York, PA	39.94	76.71	195	13.0	948	6	7
26	Cherry Hill, NJ	39.94	74.91	13	12.7	550	13	14
27	Indianapolis, IN	39.91	86.2	254	12.0	1095	11	11
28	New Castle, DE	39.66	75.57	3	13.5	1106	4	5
29	Reno, NV	39.56	119.77	1382	12.1	195	6	-
30	Baltimore, MD	39.37	76.47	36	13.3	1117	6	10
31	Kansas City, MO	39.08	94.58	288	13.9	750	8	8
32	Colorado Springs 1, CO	38.87	104.82	1867	9.8	346	5	4
33	Roseville, CA	38.76	121.27	57	17.0	637	10	-
34	Lexington, KY	38.04	84.5	299	13.1	1225	8	9
35	Pleasanton, CA	37.69	121.91	100	14.2	256	10	-
36	San Jose 1, CA	37.33	121.94	35	15.7	234	10	-
37	San Jose 2, CA	37.31	121.97	47	15.7	234	6	-
38	Durham, NC	36.05	78.9	110	15.7	1160	9	8
39	Tulsa 1, OK	36.03	95.87	207	16.2	1072	4	5
40	Knoxville, TN	35.95	84.01	305	15.0	1285	13	12
41	Los Alamos, NM	35.88	106.32	2227	11.8	323	6	-
42	Oklahoma City, OK	35.61	97.6	340	15.9	886	7	6
43	Fort Smith, AR	35.35	94.39	186	16.5	1149	4	-
44	Charlotte, NC	35.3	80.75	229	16.0	1097	8	7
45	Little Rock, AR	34.75	92.39	115	17.3	1198	8	8
46	Huntsville, AL	34.73	86.59	191	16.3	1325	12	13
47	Santa Barbara, CA	34.44	119.76	57	14.9	354	7	-
48	Atlanta, GA	33.97	84.55	366	16.8	1286	14	-
49	Orange, CA	33.78	117.86	53	17.9	170	4	-
50	Dallas, TX	33.04	96.83	207	19.3	912	7	7
51	Waco, TX	31.51	97.2	185	19.4	945	4	-
52	Georgetown, TX	30.64	97.76	269	20.3	1009	7	7
53	College Station, TX	30.64	96.31	91	19.5	509	10	10
54	Tallahassee, FL	30.44	84.28	62	19.7	1478	6	6
55	Tampa, FL	28.06	82.43	12	22.7	1101	7	-
56	Corpus Christi, TX	27.80	97.4	2	22.2	794	7	6

Appendix D

SUPPLEMENTARY DATA FOR CHAPTER 4

Table D.1: The performance metrics (mean ± standard deviation) and model parameters of ML models used to develop estimates of daily ragweed pollen concentration.

Model	R method	RMSE	R^2	MAE	Parameters
Support Vector Machine	"svm"	26.06 ± 7.99	0.50 ± 0.10	11.88 ± 2.00	C= 9.24
Random Forest	"rf"	23.06 ± 8.19	0.58 ± 0.11	11.59 ± 2.70	mtry=3
XGBoost	"xgbTree"	22.99 ± 7.48	0.58 ± 0.12	11.49 ± 2.15	eta=0.025, nrouds=150, max_depth=4, min_child_weight=3, colsample_bytree=0.4, gamma=0.5, subsample=0.75
BayeGLM	"bayesglm"	25.30 ± 9.93	0.49 ± 0.10	13.53 ± 2.24	n/a
Neural Network	"nnet"	27.21 ± 7.58	0.42 ± 0.13	14.98 ± 1.74	size=9, decay=0.5
CART	"rpart"	26.49 ± 8.02	0.45 ± 0.12	12.92 ± 2.56	cp=0.01

Table D.2: The Spearman coefficient of ML models on estimates of daily ragweed pollen concentration.

Model	Spearman coefficient
Support Vector Machine	0.83
Random Forest	0.86
XGBoost	0.84
Bayesian Generalized Linear Model	0.78
Neural Network	0.62
Classification and regression tree	0.74

Table D.3: The performance metrics (mean ± standard deviation) and model parameters of ML models on estimates of pollen level.

Model	R method	Accuracy	F1 score	Parameters
Support Vector Machine	"svm"	0.7494 ± 0.0423	0.8826 ± 0.0388	C = 4.4154
Random Forest	"rf"	0.7731 ± 0.0530	0.8926 ± 0.0382	mtry = 12
XGBoost	"xgbTree"	0.7740 ± 0.0479	0.8973 ± 0.0317	eta=0.05, nrouds=50, max_depth=3, min_child_weight=4, colsample_bytree=1, gamma=1, subsample=0.5
Neural Network	"nnet"	0.7219 ± 0.0447	0.8703 ± 0.0453	size=9, decay=0.5
CART	"rpart"	0.7410 ± 0.0484	0.8934 ± 0.0281	cp=0.025

Appendix E

R CODES FOR CHAPTER 4

```
#NOTES: #### indicates section, ### indicates subsection, # indicates notes and comments
# Scripts used to build regression models for pollen concentration prediction in Newark, NJ
setwd('C://Users/Ting Cai/Documents/PhD dissertation/MachineLearning/figures')
library(ggplot2)
#package to use SVM and SVR
library('e1071');

#### Step1. load input data processed from ML_data_preprocessing.R
model_data=read.csv('C://Users/Ting Cai/Documents/PhD dissertation/MachineLearning/model_input_variables.csv',header=TRUE)
# Subset data to features we wish to keep/use.
features <- c("PollenConc", "TEMP", "DEWP", "STP", "VISIB",
              "WDSP", "MXSPD", "MaxTemp", "MinTemp","PRCP","RH","CumTemp","CumPRCP",
              "PollenDay_1")
data_input <- model_data[, features]
#summary(data_input);
#I'm doing 10-fold cross validation in each method
set.seed(148000515)
library(caret)
# to report the F1 score
ctrl <- trainControl(method = "repeatedcv",
                     number = 10,
```

```
                      repeats = 3,
                      savePredictions = TRUE)

ctrl_cv10 <- trainControl(method = "cv",
                      number = 10,
                      savePredictions = TRUE)

#### Step2. fit SVM model
### base model
modelSVM <- train(PollenConc ~ ., data = data_input,
                  method = "svmLinear",
                  trControl = ctrl)
print(modelSVM)
modelSVM$finalModel

### TUNE SVM
modelLookup('svmLinear')
set.seed(148000515)
mygrid <- expand.grid(C = seq(0.1, 10, length = 40))
modelSVM<- train(PollenConc ~ ., data = data_input,
                 method = "svmLinear", tuneGrid=mygrid, trControl =
                    ctrl,preProcess = c("center","scale"))
modelSVM$finalModel
(SVM_results=modelSVM$results)
png(filename = "SVM_tuning_regression.png",width = 15, height = 15,
    units= "cm", res = 300)
plot(modelSVM)
dev.off()

### Final model: to get performance metrics
modelSVM$bestTune
mygrid <- expand.grid(C = 9.238462)  ## best C
modelSVM <- train(PollenConc ~ ., data = data_input,
                  method = "svmLinear", tuneGrid=mygrid,
```

```
54                    trControl = ctrl_cv10)
55 (SVM_results=modelSVM$results)
56
57 ConcPredictSVM=modelSVM$pred
58 # report Spearman coefficient
59 (corSpearmanSVM=cor(cbind(ConcPredictSVM$pred,ConcPredictSVM$obs),
       method='spearman')[2,1]);
60
61 # scatterplot SVM
62 gg_svm <- ggplot(ConcPredictSVM, aes(x=obs, y=pred)) +
63   geom_point()+
64   geom_abline(linetype ='dashed', slope=1, intercept=0,color="red",
         size=1)+
65   ylim(-10, 400)+
66   labs(x=expression('Observed Pollen Concentration (pollen/'~m^3~')'
         ),
67        y=expression('Predicted Pollen Concentration (pollen/'~m^3~')
             '),
68        title="svm") +
69   theme(plot.title = element_text(hjust = 0.5),text = element_text(
         size=12),axis.text = element_text(size=12))
70
71
72 #### Step3. random forest
73 ### base model
74 modelRF <- train(PollenConc ~ ., data = data_input,
75                    method = "rf",
76                    trControl = ctrl)
77 print(modelRF)
78 modelRF$finalModel
79
80 ### Tuning
81 modelLookup('rf')
82 set.seed(148000515)
```

```
mygrid <- expand.grid(.mtry=c(1:13))
modelRF<- train(PollenConc ~ ., data = data_input,
                 method = "rf", tuneGrid=mygrid, trControl = ctrl)
png(filename = "RF_tuning_regression.png",width = 15, height = 15,
    units= "cm", res = 300)
plot(modelRF)
dev.off()

### Final model: to get performance metrics
modelRF$bestTune

mygrid <- expand.grid(.mtry=3)
modelRF <- train(PollenConc ~ ., data = data_input,
                  method = "rf", tuneGrid=mygrid, importance=TRUE,
                  trControl = ctrl_cv10)
(RF_results=modelRF$results)

ConcPredictRF=modelRF$pred
# report Spearman coefficient
(corSpearman=cor(cbind(ConcPredictRF$pred,ConcPredictRF$obs),method=
    'spearman')[2,1]);

# scatterplot RF
gg_rf <- ggplot(ConcPredictRF, aes(x=obs, y=pred)) +
  geom_point()+
  geom_abline(linetype ='dashed', slope=1, intercept=0,color="red",
      size=1)+
  ylim(-10, 400)+
  labs(x=expression('Observed Pollen Concentration (pollen/'~m^3~')'
      ),
       y=expression('Predicted Pollen Concentration (pollen/'~m^3~')
           '),
       title="rf") +
```

```
111   theme(plot.title = element_text(hjust = 0.5),text = element_text(
          size=12),axis.text = element_text(size=12))
112
113
114 ### variable importance
115 Imp_RF=varImp(modelRF, scale = TRUE)$importance
116 Imp_RF$variable=row.names(Imp_RF)
117 Imp_RF=Imp_RF[order(Imp_RF$Overall, decreasing = TRUE),]
118 plot(varImp(modelRF, scale = TRUE))
119
120 Imp_RF_gg <- ggplot(Imp_RF, aes(x = reorder(Imp_RF$variable, Overall
      ), y = Overall)) +
121   geom_bar(position=position_dodge(), stat="identity", fill="
          slateblue") +
122   coord_flip() +
123   labs(x=expression('Independent variable'),
124        y=expression('Scaled variable importance'),
125        title="Random Forest") +
126   theme(plot.title = element_text(hjust = 0.5),text = element_text(
          size=12),axis.text = element_text(size=12))
127
128
129 #### Step4. xgbTree
130 ### base model
131 modelxgbTree <- train(PollenConc ~ ., data = data_input,
132                       method = "xgbTree",importance=TRUE,
133                       trControl = ctrl)
134 print(modelxgbTree)
135 modelxgbTree$finalModel
136 (xgbTree_results=modelxgbTree$results)
137
138 xgbTree_results[xgbTree_results$nrounds==50&xgbTree_results$max_
      depth==1&xgbTree_results$eta==0.3&
```

```
139                   xgbTree_results$gamma==0&xgbTree_results$colsample
                         _bytree==0.6&xgbTree_results$min_child_weight
                         ==1&
140                   xgbTree_results$subsample==0.5,]
141 ### tuning xgbtree
142 library(doSNOW)
143 cl <- makeCluster(6, type = "SOCK")  # I have 8 in total on my PC
144
145 # Register cluster so that caret will know to train in parallel.
146 registerDoSNOW(cl)
147 modelLookup(model='xgbTree')
148
149 # tune eta, nrouds, max_depth first
150 tune.gridxgb <- expand.grid(eta = c(0.025, 0.05, 0.1, 0.3), # 0.025
151                             nrounds = seq(from = 50, to = 1000, by =
                                  50),  # 150
152                             max_depth = 1:4,  # 4
153                             min_child_weight = 1,
154                             colsample_bytree = c(0.6),
155                             gamma = 0,
156                             subsample = c(0.75))
157 modelxgbTree <- train(PollenConc ~ ., data = data_input,
158                       method = "xgbTree",importance=TRUE, tuneGrid=
                             tune.gridxgb,
159                       trControl = ctrl)
160 png(filename = "xgbTree_tuning_regression_part1.png",width = 15,
       height = 15,  units= "cm", res = 300)
161 plot(modelxgbTree)
162 dev.off()
163
164 print(modelxgbTree)
165 # tune the min_child_weight and colsample_bytree
166 tune.gridxgb <- expand.grid(eta = 0.025, # best 0.025
```

```
167                             nrounds = seq(from = 50, to = 1000, by =
                                    50),  # best 150
168                             max_depth = 4,  # best 4
169                             min_child_weight =c(1, 2, 3,4,5),
170                             colsample_bytree = c(0.4,0.6,0.8,1),
171                             gamma = 0,
172                             subsample = c(0.75))
173 modelxgbTree <- train(PollenConc ~ ., data = data_input,
174                        method = "xgbTree",importance=TRUE, tuneGrid=
                              tune.gridxgb,
175                        trControl = ctrl)
176 png(filename = "xgbTree_tuning_regression_part2.png",width = 15,
    height = 15,  units= "cm", res = 300)
177 plot(modelxgbTree)
178 dev.off()
179
180 print(modelxgbTree)
181 # tune gama and subsumbple
182 tune.gridxgb <- expand.grid(eta = 0.025, # best 0.05
183                             nrounds = seq(from = 50, to = 1000, by =
                                    50),  # best 50
184                             max_depth = 4,  # best 3
185                             min_child_weight =3,  #best 3
186                             colsample_bytree = 0.4, #best 0.4
187                             gamma = c(0, 0.05, 0.1, 0.5, 0.7, 0.9,
                                  1.0), # best is 1
188                             subsample = c(0.5, 0.75, 1.0))  # best
                                  is 0.5
189
190 modelxgbTree <- train(PollenConc ~ ., data = data_input,
191                        method = "xgbTree",importance=TRUE, tuneGrid=
                              tune.gridxgb,
192                        trControl = ctrl)
193 print(modelxgbTree)
```

```
194 png(filename = "xgbTree_tuning_regression_part3.png",width = 15,
        height = 15,  units= "cm", res = 300)
195 plot(modelxgbTree)
196 dev.off()
197
198 ### Final model: to get accuracy and its SD and F1 score and SD
199 modelxgbTree$bestTune
200 finalGrid <- expand.grid(eta = 0.025, # best 0.025
201                          nrounds = 150,  # best 150
202                          max_depth = 4,  # best 4
203                          min_child_weight =3,  #best 3
204                          colsample_bytree = 0.4, #best .4
205                          gamma = 0.5, # best is .5
206                          subsample = 0.75)  # best is 0.75
207
208 modelxgbTree <- train(PollenConc ~ ., data = data_input,
209                       method = "xgbTree",importance=TRUE, tuneGrid=
                             finalGrid,
210                       trControl = ctrl_cv10)
211 modelxgbTree$results
212 stopCluster(cl)
213
214 (xgbTree_results=modelxgbTree$results)
215 ConcPredictxgbTree = modelxgbTree$pred
216 # report Spearman coefficient
217 (corSpearmanxgbTree=cor(cbind(ConcPredictxgbTree$pred,
        ConcPredictxgbTree$obs),method='spearman')[2,1]);
218
219 # scatterplot
220 gg_xgbTree <- ggplot(ConcPredictxgbTree, aes(x=obs, y=pred)) +
221   geom_point()+
222   geom_abline(linetype ='dashed', slope=1, intercept=0,color="red",
          size=1)+
223   ylim(-10, 400)+
```

```
224   labs(x=expression('Observed Pollen Concentration (pollen/'~m^3~')'
          ),
225        y=expression('Predicted Pollen Concentration (pollen/'~m^3~')
              '),
226        title="xgbTree") +
227   theme(plot.title = element_text(hjust = 0.5),text = element_text(
          size=12),axis.text = element_text(size=12))
228
229
230 ### plot xgbtree
231 library(xgboost)
232 #xgb.plot.multi.trees(modelXgbtree$finalModel, feature_names =
       modelXgbtree$finalModel$feature_names, features_keep = 10)
233 xgb.importance(modelxgbTree$finalModel$feature_names, modelxgbTree$
       finalModel)
234 #install.packages('DiagrammeR')
235 #install.packages('rsvg')
236 library(DiagrammeR)
237 library(rsvg)
238 gr <- xgb.plot.tree(feature_names = modelxgbTree$finalModel$feature_
       names, model = modelxgbTree$finalModel,
239                     trees=1,render=FALSE)  #must add "render=FALSE"
                          for export_graph to work!!!!!!!!!!
240 export_graph(gr, 'xgbtree_regression.png', width=1500)
241
242
243 ### variable importance
244 Imp_xgbTree=varImp(modelxgbTree, scale = TRUE)$importance
245 Imp_xgbTree$variable=row.names(Imp_xgbTree)
246 Imp_xgbTree=Imp_xgbTree[order(Imp_xgbTree$Overall, decreasing = TRUE
       ),]
247 #plot(varImp(Imp_xgbTree, scale = TRUE))
248
```

```
249 Imp_xgbTree_gg <- ggplot(Imp_xgbTree, aes(x = reorder(Imp_xgbTree$
      variable, Overall), y = Overall)) +
250   geom_bar(position=position_dodge(), stat="identity", fill="
        slateblue") +
251   coord_flip() +
252   labs(x=expression('Independent variable'),
253        y=expression('Scaled variable importance'),
254        title="xgbTree") +
255   theme(plot.title = element_text(hjust = 0.5),text = element_text(
        size=12),axis.text = element_text(size=12))
256
257 library(gridExtra)
258 g <- grid.arrange(Imp_RF_gg,Imp_xgbTree_gg, nrow = 1,ncol=2)
259
260 ggsave("variable_importance_RF_xgbTree_regression.png", g, device =
      png(), path = 'C://Users/Ting Cai/Documents/PhD dissertation/
      MachineLearning/figures',
261        width = 25, height = 20, units = "cm", dpi = 300)
262
263
264 #### Step5. bayesglm
265 ### base model
266 modelbayesglm <- train(PollenConc ~ ., data = data_input,
267                   method = "bayesglm",
268                   trControl = ctrl)
269 print(modelbayesglm)
270 modelbayesglm$finalModel
271 modelbayesglm$results
272 modelbayesglm$bestTune
273
274 modelbayesglm<- train(PollenConc ~ ., data = data_input,
275                       method = "bayesglm",
276                       trControl = ctrl_cv10)
277 print(modelbayesglm)
```

```
ConcPredictbayesglm=modelbayesglm$pred
modelbayesglm$finalModel
(bayesglm_results=modelbayesglm$results)

# report Spearman coefficient
(corSpearmanbayesglm=cor(cbind(ConcPredictbayesglm$pred,
    ConcPredictbayesglm$obs),method='spearman')[2,1]); # 0.7765936
# scatterplot
gg_bayesglm <- ggplot(ConcPredictbayesglm, aes(x=obs, y=pred)) +
  geom_point()+
  geom_abline(linetype ='dashed', slope=1, intercept=0,color="red",
      size=1)+
  ylim(-20, 400)+
  labs(x=expression('Observed Pollen Concentration (pollen/'~m^3~')'
      ),
       y=expression('Predicted Pollen Concentration (pollen/'~m^3~')
           '),
       title="bayesglm") +
  theme(plot.title = element_text(hjust = 0.5),text = element_text(
      size=12),axis.text = element_text(size=12))


#### Step6. nnet
### base model
modelnnet <- train(PollenConc ~ ., data = data_input,
                   method = "nnet",
                   trControl = ctrl)
print(modelnnet)
modelnnet$finalModel

### TUNing
modelLookup('nnet')
set.seed(148000515)
nnetGrid <-  expand.grid(size = seq(from = 1, to = 10, by = 1),
```

```
307                          decay = seq(from = 0.1, to = 0.5, by = 0.1)
                                )
308
309 modelnnet<- train(PollenConc ~ ., data = data_input,
310                       method = "nnet",maxit=1000,linout = TRUE,
311                       tuneGrid=nnetGrid, trControl = ctrl)
312 png(filename = "nnet_tuning_regression.png",width = 15, height = 15,
        units= "cm", res = 300)
313 plot(modelnnet)
314 dev.off()
315 modelnnet$results
316
317 ### Final model: to get performance metrics
318 modelnnet$bestTune
319 mygrid <- expand.grid(size = 9, decay = 0.4)
320 modelnnet <- train(PollenConc ~ ., data = data_input,
321                    method = "nnet", tuneGrid=mygrid,
322                    trControl = ctrl_cv10)
323 (nnet_results=modelnnet$results)
324 ConcPredictnnet=modelnnet$pred
325 ConcPredictnnet=ConcPredictnnet[ConcPredictnnet$decay==0.2 &
      ConcPredictnnet$size==9,][c(1:736),]
326 modelnnet$finalModel
327 (nnet_results=modelnnet$results)
328 # Spearman coefficient
329 (corSpearmannnet=cor(cbind(ConcPredictnnet$pred,ConcPredictnnet$obs)
      ,method='spearman')[2,1]); #0.6216421
330 # scatterplot
331 gg_nnet <- ggplot(ConcPredictnnet, aes(x=obs, y=pred)) +
332   geom_point()+
333   geom_abline(linetype ='dashed', slope=1, intercept=0,color="red",
        size=1)+
334   ylim(-60, 400)+
```

```
335   labs(x=expression('Observed Pollen Concentration (pollen/'~m^3~')'
          ),
336        y=expression('Predicted Pollen Concentration (pollen/'~m^3~')
               '),
337        title="nnet") +
338   theme(plot.title = element_text(hjust = 0.5),text = element_text(
          size=12),axis.text = element_text(size=12))
339
340
341
342 #### Step7. rpart (CART)
343 ### base model
344 modelrpart <- train(PollenConc ~ ., data = data_input,
345                     method = "rpart",
346                     trControl = ctrl)
347 print(modelrpart)
348
349 ### TUNE rpart
350 modelLookup(model='rpart')
351 set.seed(148000515)
352 mygrid <- expand.grid(cp=seq(0, 0.5, 0.005))
353 modelrpart<- train(PollenConc ~ ., data = data_input,
354                    method = "rpart", tuneGrid=mygrid, trControl =
                          ctrl)
355 print(modelrpart)
356 modelrpart$finalModel
357 png(filename = "rpart_tuning_regression.png",width = 15, height =
       15,  units= "cm", res = 300)
358 plot(modelrpart)
359 dev.off()
360
361 ### Final model:
362 mygrid <- expand.grid(cp=0.01)  ## best C
363 modelrpart <- train(PollenConc ~ ., data = data_input,
```

```
364                     method = "rpart", tuneGrid=mygrid,
365                     trControl = ctrl)
366 (rpart_results=modelrpart$results)
367 ConcPredictrpart=modelrpart$pred
368 ConcPredictrpart=ConcPredictrpart[ConcPredictrpart$cp==0.01,][c
    (1:736),]
369 modelnnet$finalModel
370 (nnet_results=modelnnet$results)
371 # Spearman coefficient
372 (corSpearmanrpart=cor(cbind(ConcPredictrpart$pred,ConcPredictrpart$
    obs),method='spearman')[2,1]); #0.7386159
373
374 #install.packages('rattle')
375 library(rattle)
376 png(filename = "rpart_tree_regression.png",width = 20, height = 20,
     units= "cm", res = 300)
377 fancyRpartPlot(modelrpart$finalModel,caption = "")
378 dev.off()
379
380 # scatterplot
381 gg_rpart <- ggplot(ConcPredictrpart, aes(x=obs, y=pred)) +
382   geom_point()+
383   geom_abline(linetype ='dashed', slope=1, intercept=0,color="red",
      size=1)+
384   ylim(-10, 400)+
385   labs(x=expression('Observed Pollen Concentration (pollen/'~m^3~')'
      ),
386        y=expression('Predicted Pollen Concentration (pollen/'~m^3~')
          '),
387        title="rpart") +
388   theme(plot.title = element_text(hjust = 0.5),text = element_text(
      size=12),axis.text = element_text(size=12))
389
390
```

```
391 # combined all the scatterplots
392 library(gridExtra)
393 g <- grid.arrange(gg_svm, gg_rf,gg_xgbTree,gg_bayesglm,gg_nnet,gg_
    rpart, nrow = 3,ncol=2)
394 ggsave("scatterplot_ML_regression.png", g, device = png(), path = 'C
    ://Users/Ting Cai/Documents/PhD dissertation/MachineLearning/
    figures',
395        width = 25, height = 35, units = "cm", dpi = 300)
396
397
398 #### plot results (RMSE, R2, MAE):
399 # save prvious results in csv file first.
400 regression_metrics=read.csv('C://Users/Ting Cai/Documents/PhD
    dissertation/MachineLearning/ML_regression_results.csv',header=
    TRUE)
401 library(data.table)
402 regression_metrics$Spearman.coefficient.SD=0
403 regression_metrics$Method_short=c("SVM","RF","xgbTree","bayesglm","
    nnet","rpart")
404
405 library(ggplot2)
406
407 ### three plots in one figure
408 regression_metrics_long=melt(regression_metrics[,c(2,6,4,10)],
    variable.name="Method_short")
409 regression_metrics_long[c(13:18),2]= as.factor(as.character("R2"))
410 regression_metrics_long2=melt(regression_metrics[,c(3,7,5,10)],
    variable.name="Method_short")
411
412 regression_metrics_new=cbind(regression_metrics_long,regression_
    metrics_long2)
413 regression_metrics_new=regression_metrics_new[,-4]
414 colnames(regression_metrics_new) <- c("Method","Metrics","Metrics_
    value","Metrics_SD","SD_value")
```

```
415
416 gg <- ggplot(regression_metrics_new,aes(x=Method,y=Metrics_value,
        fill=Metrics)) +
417   geom_bar(width = 0.5,position=position_dodge(), stat="identity",
        color="black") +
418   geom_errorbar(aes(ymin=Metrics_value-SD_value, ymax=Metrics_value+
        SD_value), width=.2,position=position_dodge(.9)) +
419   labs(y = "Value", x = "Method",
420        title = "")+
421   theme(legend.position = "none")+
422   #theme(legend.position = c(0.9, 0.9))+
423   theme(plot.title = element_text(hjust = 0.5),text = element_text(
        size=12),axis.text = element_text(size=12,face="bold"))+
424   theme(strip.text = element_text(face="bold", size=12))+
425   facet_grid(Metrics ~ ., scales = "free_y")
426
427
428 ggsave("ML_regression_results.png", gg, device = png(), path = 'C://
        Users/Ting Cai/Documents/PhD dissertation/MachineLearning/figures
        ',
429              width = 15, height = 20, units = "cm", dpi = 300)
430
431
432 #### Time series for predicted and observed pollen concentration
433 ConcPredictxgbTree = ConcPredictxgbTree[order(ConcPredictxgbTree$
        rowIndex),]
434 ConcPredictRF = ConcPredictRF[order(ConcPredictRF$rowIndex),]
435 data_time_series = cbind(model_data$YEARMODA, ConcPredictRF[,c
        (1,2,3)],ConcPredictxgbTree[,1])
436 data_time_series = data_time_series[,-4]
437 colnames(data_time_series) <- c("YEARMODA","RandomForest","
        Observation","xgbTree")
438 data_time_series$YEARMODA
439 data_time_series$PollenConc
```

```
440 data_time_series$PollenAllPredictRF
441
442 library(data.table)
443 data_time_series_long <- melt(data_time_series[46:94,], id="YEARMODA
    ")  # convert to long format, year 1995
444 #data_time_series_long <- melt(data_time_series[95:141,], id="
    YEARMODA")  # convert to long format, year 1996
445 data_time_series_long$day=as.numeric(as.Date(data_time_series_long$
    YEARMODA)-as.Date("1995-08-04")+1)
446 #data_time_series_long$day=as.numeric(as.Date(data_time_series_long$
    YEARMODA)-as.Date("1996-08-11")+1)
447
448 g1995 <- ggplot(data_time_series_long, aes(x=day, y=value, colour=
    variable)) +
449   geom_line(size=0.5,linetype = "dashed")+
450   geom_point(size=2)+
451   ylim(0, 200)+
452   theme(legend.position = c(0.9, 0.8))+
453   theme(legend.title=element_blank())+
454   #geom_smooth(method=lm, fill = "salmon", color='salmon3')+
455   labs(x=expression('Days since August 04, 1995'),
456        y=expression('Pollen Concentration (pollen/' ~m^3~')'),
457        title="") +
458   theme(plot.title = element_text(hjust = 0.5),text = element_text(
      size=12),axis.text = element_text(size=12))
459
460 g1996 <- ggplot(data_time_series_long, aes(x=day, y=value, colour=
    variable)) +
461   geom_line(size=0.5,linetype = "dashed")+
462   geom_point(size=2)+
463   ylim(0, 170)+
464   theme(legend.position = c(0.9, 0.8))+
465   theme(legend.title=element_blank())+
466   #geom_smooth(method=lm, fill = "salmon", color='salmon3')+
```

```
467   labs(x=expression('Days since August 11, 1996'),
468        y=expression('Pollen Concentration (pollen/' ~m^3~')'),
469        title="") +
470   theme(plot.title = element_text(hjust = 0.5),text = element_text(
          size=12),axis.text = element_text(size=12))
471
472 library(gridExtra)
473 g <- grid.arrange(g1995, g1996, nrow = 2,ncol=1)
474
475 ggsave("ML_time_series_RF_xgbTree.png", g, device = png(), path = 'C
      ://Users/Ting Cai/Documents/PhD dissertation/MachineLearning/
      figures',
476        width = 15, height = 20, units = "cm", dpi = 400)
```

```
1  #NOTES: #### indicates section, ### indicates subsection, #
       indicates notes and comments
2  # Scripts used to build classification models for pollen level
       prediction in Newark, NJ
3
4  #package to use SVM and SVR
5  library('e1071');
6  #package to provide commonly used function such as
       creatDataPartition for regression and classification
7  library('caret');
8  #load ggplot for visualization
9  #library('ggplot2');
10 setwd('C://Users/Ting Cai/Documents/PhD dissertation/MachineLearning
       /')
11
12 #### load the input data
13 model_data=read.csv('C://Users/Ting Cai/Documents/PhD dissertation/
       MachineLearning/model_input_variables.csv',header=TRUE)
14
15 #### Step 1: preprocessing the data
```

```
16 model_data$Pollen_level=NA
17
18 # divide the pollen conc into three levels: <10, (10,30),(30,100)
       ,>100
19 library(data.table)
20 model_data$Pollen_level[model_data$PollenConc < 10]=1
21 model_data$Pollen_level[model_data$PollenConc >= 10 & model_data$
       PollenConc < 30]=2
22 model_data$Pollen_level[model_data$PollenConc >= 30]=3
23
24 table(model_data$Pollen_level)
25 YR=substring(model_data$YEARMODA,1,4)
26 model_data$Year=as.factor(YR)
27 model_data$Pollen_level=as.factor(model_data$Pollen_level)
28
29 #### ggplot of pollen levels for each year
30
31 gg <- ggplot(model_data, aes(x = Year, fill = Pollen_level)) +
32   #theme_bw() +
33   geom_bar(width = 0.5, color = "black") +
34   theme(legend.position = c(0.1, 0.9))+
35   labs(y = "Pollen levels counts", x = "Year",
36        title = "Pollen level counts in each year")+
37   theme(plot.title = element_text(hjust = 0.5),text = element_text(
       size=12),axis.text = element_text(size=12))
38
39 ggsave("Pollen_level_counts_by_year.png", gg, device = png(), path =
       'C://Users/Ting Cai/Documents/PhD dissertation/MachineLearning/
      figures',
40        width = 20, height = 15, units = "cm", dpi = 300)
41
42 # Subset data to features we wish to keep/use.
43 features <- c("Pollen_level", "TEMP", "DEWP", "STP", "VISIB",
```

```
44                "WDSP", "MXSPD", "MaxTemp", "MinTemp","PRCP","RH","
                     CumTemp","CumPRCP",
45                "PollenDay_1")
46 data_input<- model_data[, features]
47
48 # I'm doing 10-fold cross validation with 3 repeats in each method
49 set.seed(148000515)
50 library(caret)
51
52 ## to report the F1 score
53 library(MLmetrics)
54 f1 <- function(data, lev = NULL, model = NULL) {
55   f1_val <- F1_Score(y_pred = data$pred, y_true = data$obs, positive
         = lev[1])
56   c(F1 = f1_val)
57 }
58
59 ctrl <- trainControl(method = "repeatedcv",
60                      number = 10,
61                      repeats = 3,
62                      savePredictions = TRUE)
63
64 ctrl_F1 <- trainControl(method = "repeatedcv",
65                         number = 10,
66                         repeats = 3,
67                         summaryFunction = f1,
68                         savePredictions = TRUE)
69
70 #### Step2. SVM
71 ### base model
72 modelSVM <- train(Pollen_level ~ ., data = data_input,
73                   method = "svmLinear",
74                   trControl = ctrl)
75 print(modelSVM)
```

```
modelSVM$finalModel
m <- svm(Pollen_level~., data = data_input)

plot(m, data_input, CumTemp ~ CumPRCP)
confusionMatrix(ConcPredictSVM$pred, ConcPredictSVM$obs)
modelSVM$finalModel
(SVM_results=modelSVM$results)  ## gives F1 and its SD


### TUNE SVM
set.seed(148000515)
mygrid <- expand.grid(C = seq(0.1, 10, length = 40))
modelSVM<- train(Pollen_level ~ ., data = data_input,
                 method = "svmLinear", tuneGrid=mygrid, trControl =
                     ctrl,preProcess = c("center","scale"))
modelSVM$finalModel
(SVM_results=modelSVM$results)

png(filename = "SVM_tuning_classification.png",width = 15, height =
   15,  units= "cm", res = 300)
plot(modelSVM)
dev.off()

### Final model: to get accuracy and its SD and F1 score and SD
mygrid <- expand.grid(C = 4.4154)  ## best C
modelSVM <- train(Pollen_level ~ ., data = data_input,
                  method = "svmLinear",
                  trControl = ctrl)
(SVM_results=modelSVM$results)

modelSVM<- train(Pollen_level ~ ., data = data_input,metric = "F1",
                 method = "svmLinear", tuneGrid=mygrid, trControl =
                    ctrl_F1,preProcess = c("center","scale"))
modelSVM$finalModel
```

```
107 (SVM_results=modelSVM$results)
108
109
110 #### Step 3. random forest
111 ### base model
112 modelRF <- train(Pollen_level ~ ., data = data_input,
113                  method = "rf",
114                  trControl = ctrl)
115
116 print(modelRF)
117 modelRF$finalModel
118
119 ### TUNE RF
120 modelLookup(model='rf')
121 set.seed(148000515)
122 mygrid <- expand.grid(.mtry=c(1:13))
123 modelRF<- train(Pollen_level ~ ., data = data_input,
124                 method = "rf", tuneGrid=mygrid, trControl = ctrl)
125 print(modelRF)
126 modelRF$finalModel
127 (RF_results=modelRF$results)
128
129 png(filename = "RF_tuning_classification.png",width = 15, height =
    15,  units= "cm", res = 300)
130 plot(modelRF)
131 dev.off()
132
133 ### Final model: to get accuracy and its SD and F1 score and SD
134 mygrid <- expand.grid(.mtry=12)  ## best C
135 modelRF <- train(Pollen_level ~ ., data = data_input,
136                  method = "rf", tuneGrid=mygrid,
137                  trControl = ctrl)
138 (RF_results=modelRF$results)
139 modelRF<- train(Pollen_level ~ ., data = data_input,metric = "F1",
```

```
140                   method = "rf", tuneGrid=mygrid, trControl = ctrl_F1
                          )
141 modelRF$finalModel
142 (RF_results=modelRF$results)
143
144
145 ### variable importance
146 Imp_RF=varImp(modelRF, scale = TRUE)$importance
147 Imp_RF$variable=row.names(Imp_RF)
148 Imp_RF = Imp_RF[,c(3,4)]
149 colnames(Imp_RF) = c("Overall", "variable")
150 Imp_RF=Imp_RF[order(Imp_RF$Overall, decreasing = TRUE),]
151 plot(varImp(modelRF, scale = TRUE))
152
153 Imp_RF_gg <- ggplot(Imp_RF, aes(x = reorder(Imp_RF$variable, Overall
       ), y = Overall)) +
154   geom_bar(position=position_dodge(), stat="identity", fill="
         slateblue") +
155   coord_flip() +
156   labs(x=expression('Independent variable'),
157        y=expression('Scaled variable importance'),
158        title="Random Forest") +
159   theme(plot.title = element_text(hjust = 0.5),text = element_text(
         size=12),axis.text = element_text(size=12))
160
161 #### Step 4. xgbTree
162 ### base model
163
164 modelxgbTree <- train(Pollen_level ~ ., data = data_input,
165                       method = "xgbTree",importance=TRUE,
166                       trControl = ctrl)
167
168 print(modelxgbTree)
169 modelxgbTree$finalModel
```

```
170 (xgbTree_results=modelxgbTree$results)
171
172 xgbTree_results[xgbTree_results$nrounds==50&xgbTree_results$max_
    depth==1&xgbTree_results$eta==0.3&
173                  xgbTree_results$gamma==0&xgbTree_results$colsample
                        _bytree==0.6&xgbTree_results$min_child_weight
                        ==1&
174                  xgbTree_results$subsample==0.5,]
175
176 ### tuning xgbtree
177 library(doSNOW)
178 cl <- makeCluster(6, type = "SOCK")  # I have 8 in total on my PC
179
180 # Register cluster so that caret will know to train in parallel.
181 registerDoSNOW(cl)
182 modelLookup(model='xgbTree')
183
184 ## tune eta, nrouds, max_depth first
185 tune.gridxgb <- expand.grid(eta = c(0.025, 0.05, 0.1, 0.3), # 0.05
186                             nrounds = seq(from = 50, to = 1000, by =
                                    50),  # 50
187                             max_depth = 1:4,  # 3
188                             min_child_weight = 1,
189                             colsample_bytree = c(0.8),
190                             gamma = 0,
191                             subsample = c(0.75))
192 png(filename = "xgbTree_tuning_classification_part1.png",width = 15,
     height = 15,  units= "cm", res = 300)
193 plot(modelxgbTree)
194 dev.off()
195
196 ## tune the min_child_weight and colsample_bytree
197 tune.gridxgb <- expand.grid(eta = 0.05, # best 0.05
```

```
198                             nrounds = seq(from = 50, to = 1000, by =
                                    50),  # best 50
199                             max_depth = 3,  # best 3
200                             min_child_weight =c(2, 3,4,5),  #best 4
201                             colsample_bytree = c(0.4,0.6,0.8,1), #
                                   best 1
202                             gamma = 0,
203                             subsample = c(0.75))
204 png(filename = "xgbTree_tuning_classification_part2.png",width = 15,
        height = 15,  units= "cm", res = 300)
205 plot(modelxgbTree)
206 dev.off()
207 ## tune gama and subsumbple
208 tune.gridxgb <- expand.grid(eta = 0.05, # best 0.05
209                             nrounds = seq(from = 50, to = 1000, by =
                                    50),  # best 50
210                             max_depth = 3,  # best 3
211                             min_child_weight =4,  #best 4
212                             colsample_bytree = 1, #best 1
213                             gamma = c(0, 0.05, 0.1, 0.5, 0.7, 0.9,
                                   1.0), # best is 1
214                             subsample = c(0.5, 0.75, 1.0))  # best
                                   is 0.5
215
216 modelxgbTree <- train(Pollen_level ~ ., data = data_input,
217                       method = "xgbTree",importance=TRUE, tuneGrid=
                             tune.gridxgb, #metric = "F1",
218                       trControl = ctrl)
219
220 print(modelxgbTree)
221 png(filename = "xgbTree_tuning_classification_part3.png",width = 15,
        height = 15,  units= "cm", res = 300)
222 plot(modelxgbTree)
223 dev.off()
```

```
224
225
226 ### Final model: to get accuracy and its SD and F1 score and SD
227 finalGrid <- expand.grid(eta = 0.05, # best 0.05
228                          nrounds = 50,  # best 50
229                          max_depth = 3,  # best 3
230                          min_child_weight =4,  #best 4
231                          colsample_bytree = 1, #best 1
232                          gamma = 1, # best is 1
233                          subsample = 0.5)  # best is 0.5
234
235 modelxgbTree <- train(Pollen_level ~ ., data = data_input,
236                       method = "xgbTree",importance=TRUE, tuneGrid=
                             finalGrid,
237                       trControl = ctrl)
238 modelxgbTree$results
239
240 modelxgbTree <- train(Pollen_level ~ ., data = data_input,
241                       method = "xgbTree",importance=TRUE, tuneGrid=
                             finalGrid, metric = "F1",
242                       trControl = ctrl_F1)
243
244 stopCluster(cl)
245
246 print(modelxgbTree)
247 modelxgbTree$finalModel
248 (xgbTree_results=modelxgbTree$results)
249 plot(modelxgbTree)
250
251 ### variable importance
252 Imp_xgbTree=varImp(modelxgbTree, scale = TRUE)$importance
253 Imp_xgbTree$variable=row.names(Imp_xgbTree)
254 Imp_xgbTree=Imp_xgbTree[order(Imp_xgbTree$Overall, decreasing = TRUE
    ),]
```

```
255 #plot(varImp(Imp_xgbTree, scale = TRUE))
256
257 Imp_xgbTree_gg <- ggplot(Imp_xgbTree, aes(x = reorder(Imp_xgbTree$
      variable, Overall), y = Overall)) +
258   geom_bar(position=position_dodge(), stat="identity", fill="
        slateblue") +
259   coord_flip() +
260   labs(x=expression('Independent variable'),
261        y=expression('Scaled variable importance'),
262        title="eXtreme Gradient Boosting") +
263   theme(plot.title = element_text(hjust = 0.5),text = element_text(
        size=12),axis.text = element_text(size=12))
264
265 library(gridExtra)
266 g <- grid.arrange(Imp_RF_gg,Imp_xgbTree_gg, nrow = 1,ncol=2)
267
268 ggsave("variable_importance_RF_xgbTree_classification.png", g,
      device = png(), path = 'C://Users/Ting Cai/Documents/PhD
      dissertation/MachineLearning/figures',
269        width = 25, height = 20, units = "cm", dpi = 300)
270
271
272 #### Step5. nnet
273 ### base model
274 modelnnet<- train(Pollen_level ~ ., data = data_input,
275                   method = "nnet",trControl = ctrl)
276 print(modelnnet)
277 modelnnet$finalModel
278 (nnet_results=modelnnet$results)
279
280 ### tuning nnet
281 library(doSNOW)
282 cl <- makeCluster(6, type = "SOCK")  # I have 8 in total on my PC
283 # Register cluster so that caret will know to train in parallel.
```

```
284 registerDoSNOW(cl)
285
286 modelLookup(model='nnet')
287 nnetGrid <-  expand.grid(size = seq(from = 1, to = 10, by = 1),
288                          decay = seq(from = 0.1, to = 0.5, by = 0.1)
                             )
289
290 modelnnet<- train(Pollen_level ~ ., data = data_input,
291                   method = "nnet",maxit=1000,linout = TRUE,
292                   tuneGrid=nnetGrid, trControl = ctrl)
293
294 print(modelnnet)
295 plot(modelnnet)
296 modelnnet$finalModel
297 (nnet_results=modelnnet$results)
298
299 png(filename = "nnet_tuning_classification.png",width = 15, height =
      15,  units= "cm", res = 300)
300 plot(modelnnet)
301 dev.off()
302
303 library(devtools)
304 source_url('https://gist.githubusercontent.com/fawda123/7471137/raw/
    466c1474d0a505ff044412703516c34f1a4684a5/nnet_plot_update.r')
305 png(filename = "ModelStructure_neuralnetwork.png",width = 25, height
      = 20,  units= "cm", res = 300)
306 plot.nnet(modelnnet);
307 dev.off()
308
309 ### Final model: to get accuracy and its SD and F1 score and SD
310 finalGrid <- expand.grid(.decay=0.4, .size=9)
311
312 modelnnet <- train(Pollen_level ~ ., data = data_input,
```

```
313                       method = "nnet",importance=TRUE, tuneGrid=
                             finalGrid,
314                       trControl = ctrl)
315 modelnnet$results
316
317 modelnnet <- train(Pollen_level ~ ., data = data_input,
318                       method = "nnet",importance=TRUE, tuneGrid=
                             finalGrid, metric = "F1",
319                       trControl = ctrl_F1)
320 modelnnet$results
321 stopCluster(cl)
322
323
324 #### Step6. rpart (CART)
325 ### base model
326 modelrpart <- train(Pollen_level ~ ., data = data_input,
327                     method = "rpart",
328                     trControl = ctrl)
329 print(modelrpart)
330
331 ### TUNE rpart
332 modelLookup(model='xgbTree')
333 set.seed(148000515)
334 mygrid <- expand.grid(cp=seq(0, 0.5, 0.005))
335 modelrpart<- train(Pollen_level ~ ., data = data_input,
336                    method = "rpart", tuneGrid=mygrid, trControl = ctrl)
337 print(modelrpart)
338 modelrpart$finalModel
339
340 png(filename = "rpart_tuning_classification.png",width = 15, height
    = 15,  units= "cm", res = 300)
341 plot(modelrpart)
342 dev.off()
343
```

```
344 ### Final model: to get accuracy and its SD and F1 score and SD
345 mygrid <- expand.grid(cp=0.025)  ## best C
346 modelrpart <- train(Pollen_level ~ ., data = data_input,
347                  method = "rpart", tuneGrid=mygrid,
348                  trControl = ctrl)
349 (rpart_results=modelrpart$results)
350
351 modelrpart<- train(Pollen_level ~ ., data = data_input,metric = "F1"
    ,
352                method = "rpart", tuneGrid=mygrid, trControl = ctrl_
                      F1)
353 modelrpart$finalModel
354 (rpartresults=modelrpart$results)
355
356 #install.packages('rattle')
357 library(rattle)
358 png(filename = "rpart_tree_classification.png",width = 20, height =
    20,  units= "cm", res = 300)
359 fancyRpartPlot(modelrpart$finalModel,caption = "")
360 dev.off()
361
362 #### Step7. PLOT THE RESULTS
363 # save results from previous section first
364 classification_metrics=read.csv('C://Users/Ting Cai/Documents/PhD
    dissertation/MachineLearning/ML_classification_results.csv',
    header=TRUE)
365 library(data.table)
366
367 library(ggplot2)
368
369 classification_metrics_long=melt(classification_metrics[,c(2,4,1)],
    variable.name="Method")
370 classification_metrics_long2=melt(classification_metrics[,c(3,5,1)],
     variable.name="Method")
```

```
371
372 classification_metrics_new=cbind(classification_metrics_long,
        classification_metrics_long2)
373 classification_metrics_new=classification_metrics_new[,-4]
374 colnames(classification_metrics_new) <- c("Method","Metrics","
        Metrics_value","Metrics_SD","SD_value")
375
376 gg <- ggplot(classification_metrics_new,aes(x=Method,y=Metrics_value
        , fill=Metrics)) +
377   geom_bar(width = 0.6,position=position_dodge(), stat="identity",
        color="black") +
378   geom_errorbar(aes(ymin=Metrics_value-SD_value, ymax=Metrics_value+
        SD_value), width=.2,position=position_dodge(.6)) +
379   labs(y = "Value", x = "Method",
380        title = "")+
381   theme(plot.title = element_text(hjust = 0.5),text = element_text(
        size=12),axis.text = element_text(size=12,face="bold"))
382
383
384 ggsave("ML_classification_results.png", gg, device = png(), path = '
        C://Users/Ting Cai/Documents/PhD dissertation/MachineLearning/
        figures',
385        width = 20, height = 15, units = "cm", dpi = 300)
```