UNIVERSITY OF CALIFORNIA
RIVERSIDE

Computational and Geo-Spatial Approaches to Investigate Multi-Scale Air Quality Trends in Southern California

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Chemical and Environmental Engineering

by

Khanh T. Do

June 2023

Dissertation Committee:

- Dr. Cesunica Ivey, Chairperson
- Dr. Don Collins
- Dr. David R. Cocker III

The Dissertation of Khanh T. Do is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgments

I express my sincere gratitude and appreciation to Dr. Cesunica Ivey, my advisor, for her invaluable mentorship, invaluable patience, and continuous support. She brought me into atmospheric pollution modeling research, opened doors to learn new techniques, and allowed me to pursue a master's in electrical engineering. These acquired skills ideally enriched my computational and geo-spatial air quality research.

I would like to acknowledge the dissertation committee, Dr. Don Collins and Dr. David Cocker, who provided suggestions for my advancement to candidacy and throughout my PhD work. The special topics Dr. Collins and Dr. Cocker led were my best resource to learn from air quality experimental groups.

I would like to thank Dr. Salman Asif for being my ATC committee and his help with Electrical Engineering courses. I would like to thank Arash Kashfi Yeganeh for the endless effort to compile and debug WRF and CMAQ, and the special topics students who shared their research to help me improve aerosol science.

I thank coffee for keeping me concentrating. I thank Sheba, Slug, and Wormulon for their responsiveness during my PhD work.

ABTRACT OF THE DISSERTATION

Computational and Geo-Spatial Approaches to Investigate Multi-Scale Air Quality Trends in Southern California

by

Khanh T. Do

Doctor of Philosophy, Graduate Program in Chemical and Environmental Engineering
University of California, Riverside, June 2023
Dr. Cesunica Ivey, Chairperson

Air pollution has been a significant problem in California's South Coast Air Basin for many years due to the region's unique topography, high anthropogenic emissions, weather patterns, and population density. Fine particulate matter and ozone are the two most concerning pollutants causing serious public health risks.

Chemical transport models and machine learning approaches enable an effective way to study air pollution. By looking at a large domain, scientists gain insights into the transport of precursor substances in heavily polluted areas. The methods facilitate the examination of ozone's response to emissions and meteorological factors. The main objectives are: (1) to enhance the prediction of ozone levels in the South Coast Air Basin, (2) to investigate the role of meteorology in ozone formation, and (3) to determine the most critical factors influencing ozone exceedance hours. The results showed that ozone has a strong relationship with meteorology, in which wind speed and wind direction contribute mainly to the transport and mixing of precursors, while temperature can directly contribute to ozone formation. Climate-related increases in temperature would therefore be expected to increase future ozone levels in the absence of emission changes.

Control strategies from the Air Quality Management District have improved the air quality in general. However, it did not ensure the air quality also got better for minority groups. Many

polluted areas are associated with industries, shipping activities, and warehouses that mostly affect the underrepresented communities. These communities frequently face social exclusion, yet there has been limited attention and research dedicated to understanding and addressing their specific needs and challenges.

Air pollution research often used crowdsourced data which generally reflected a higher socioeconomic status population. A data-driven approach powered by low-cost sensors showed underrepresented groups suffered higher indoor $PM_{2.5}$ due to high frequent indoor emissions from cooking, cleaning, or dusting without sufficient filtration or air exchange rate, which can be concluded that ambient and indoor PM2.5 exposures for disproportionately impacted groups cannot be generalized in population-wide. Personal exposure studies showed that people spent over 90% of their time staying indoors, emphasizing the crucial role of indoor air quality in impacting human health to a greater extent.

Table of Contents

List of Tables

List of Figures

Introduction

This dissertation seeks to address the pollution exposure disparities in the Inland Empire (Chapter 1), the role of meteorology in ozone formation (Chapter 2), and the improvement of the CMAQ computational efficiency by porting the intensive science process on the GPU (Chapter 3).

Chapter 1 utilizes low-cost sensors to study pollution exposure disparities in the Inland Empire. Ambient particulate matter (PM) has been widely studied to examine the impact of PM exposure on human health. Many air monitoring stations are operated in the U.S. to measure the trends and composition of ambient PM in support of the National Ambient Air Quality Standards (NAAQS). However, ambient PM concentrations may not reflect actual daily personal exposure (PE) (Koistinen et al., 2004). Further, the sparseness of the monitoring network leads to low spatial resolution data and necessitates gap-filling, which also affects the accuracy of PM exposure assessments that are based on ambient measurements (Yu et al., 2019). Wearable $PM_{2.5}$ sensors in real-time are used to detail a pilot-scale personal exposure campaign for five inland Southern California cities to capture the spatial and temporal variability of $PM_{2.5}$ exposures over multiple, consecutive 24-hr periods. The main objective of this pilot study was to develop and implement a high-resolution monitoring and analysis framework for characterizing $PM_{2.5}$ exposure variability for individuals from different cities of residence and subsequently different socioeconomic status (SES) neighborhoods. High concentrations of $PM_{2.5}$ can be found around industrial and shipping facilities where residential areas with low median household income are located (Houston et al., 2004; Ivey et al., 2020). These houses were built with many missing features which ensure good indoor air quality. Centering the BNSF railyard, many residential dwellings do not have an air

conditioning unit exposed to high $PM_{2.5}$ levels due to their living conditions and locations, increasing inequality and disparity.

Chapter 2 investigates the role of meteorology on ozone formation. The poor air quality in the South Coast Air Basin can be explained by the unique topography and high anthropogenic emissions. Meteorological variables and synoptic patterns greatly influence air pollution in SoCAB (Ulrickson & Mass, 1990a, 1990b). Los Angeles' temperature inversions resulting from high-pressure systems over SoCAB combined with a mountain wave-induced downslope flow create a trap that accumulates air pollutants near the ground, leading to degraded air quality (Lu & Turco, 1994, 1996). The relationship between ozone (O_3), its anthropogenic precursors, nitrogen oxides (NO_X), and volatile organic compounds (VOC) has been well studied by means of environmental chamber experiments, field studies, and air quality modeling, yet new modeling methods are still needed to better understand why rates of ozone reduction in the SoCAB have been lower than previously predicted (Baidar et al., 2015; Pusede & Cohen, 2012; Qian et al., 2019). Examining the ratio of NO_X/VOC emissions and identifying VOC and NO_X limited regimes are useful practices for creating surface ozone reduction strategies, which is one approach for developing SoCAB emission-control strategies. Chemical transport modeling is considered the most advanced approach for evaluating emission-control strategies, but is subject to uncertainties in emission rates, chemical reaction rates, and representation of meteorological influences. To further understand the quantitative relationship between ambient ozone concentrations and emission precursors, heatmaps are developed from data or modeling to capture the ozone's sensitivity due to the change in NO_x and temperature. Meteorology's role in changing ozone concentrations and ozone exceedances in SoCAB can be explored by integrating machine learning and CMAQ. The

approach in investigating meteorology-ozone sensitivity is to apply machine learning to predict Fontana (inland Southern California) ozone concentrations based on Los Angeles and Ontario meteorology. The machine learning results are analyzed against CMAQ simulations and observational data to evaluate the model performance and explore the common findings between the two approaches.

Chapter 3 presents a new approach to improving chemical transport model simulation time. Deterministic air quality models (AQMs) are designed to simulate complex physical and chemical processes taking place in the Earth's atmosphere with mathematical presentations of the atmospheric transport, diffusion, dispersion, and chemical reactions, which are solved by analytical and numerical techniques and based on the conservation of mass principle for pollutants (Lamb & Seinfeld, 1973). Handling large datasets is a challenge, given the limitation of computational efficiency and the data's complexity. Moreover, CTMs apply complex governing equations to solve for the output concentrations using the CPUs. Regarding runtime, a 12-km, two-way coupled WRF-CMAQ simulation using 34 layers of variable thickness with a domain size of 279x251 grid cells requires over 3 hours of work for 32 CPU cores per one simulated day over the five-month period (Wong et al., 2012). Simulating 12 months of ozone concentrations over Southern California using 4-km resolution with 156x102 grid cells took 20 days with 16 MPI threads. The computational efficiency of CMAQ largely suffers from solving a set of stiff differential equations when computing the gas phase chemical concentrations. For example, with the SAPRC07 chemical mechanism, the systems of photochemical reactions are calculated using Euler Backward Iteration, SMV Gear, or Rosenbrock solver (ROS) for every time step and grid cell (row x column x height) for all species in the SAPRC07 family until a specified convergence tolerance is

met. The running time is linearly proportional to the simulated domain and exponential with increased chemical species. The CMAQ simulation time can be improved by porting intensive computational processes onto GPUs. With thousands of Compute Unified Device Architecture (CUDA) cores in a single GPU, many independent arithmetic operations can be carried out simultaneously. By exploring the computer architecture, the advantages, and the disadvantages of GPU programming, a ported version of the partial derivative, decomposition, and back substitution subroutines of the ROS3 (Rosenbrock) solver to the CUDA platform reduce the computation time greatly.

References

Baidar, S., Hardesty, R. M., Kim, S. W., Langford, A. O., Oetjen, H., Senff, C. J., Trainer, M., & Volkamer, R. (2015). Weakening of the weekend ozone effect over California's South Coast Air Basin. *Geophysical Research Letters*, *42*(21). https://doi.org/10.1002/2015GL066419

Houston, D., Wu, J., Ong, P., & Winer, A. (2004). Structural disparities of urban traffic in Southern California: Implications for vehicle-related air pollution exposure in minority and high-poverty neighborhoods. *Journal of Urban Affairs*, *26*(5). https://doi.org/10.1111/j.0735-2166.2004.00215.x

Ivey, C., Gao, Z., & Do, K. (2020). *Impacts of the 2020 COVID-19 Shutdown Measures on Ozone Production in the Los Angeles Basin*. 1–10. https://doi.org/10.26434/chemrxiv.12805367.v1

Koistinen, K.J., Edwards, R.D., Mathys, P., Ruuskanen, J., Künzli, N., Jantunen, M.J., 2004a. Sources of fine particulate matter in personal exposures and residential indoor, residential outdoor and workplace microenvironments in the Helsinki phase of the EXPOLIS study. Scandinavian Journal of Work, Environment and Health 20, 36–46.

Lamb, R. G., & Seinfeld, J. H. (1973). Mathematical Modeling of Urban Air Pollution General Theory. *Environmental Science and Technology*, *7*(3). https://doi.org/10.1021/es60075a006

Lu, R., & Turco, R. P. (1994). Air Pollutant Transport in a Coastal Environment. Part I: Two-Dimensional Simulations of Sea-Breeze and Mountain Effects. *Journal of the Atmospheric Sciences*. https://doi.org/10.1175/1520-0469(1994)051<2285:aptiac>2.0.co;2

Lu, R., & Turco, R. P. (1996). Ozone distributions over the Los Angeles basin: Three-dimensional simulations with the smog model. *Atmospheric Environment*. https://doi.org/10.1016/1352-2310(96)00153-7

Pusede, S. E., & Cohen, R. C. (2012). On the observed response of ozone to NOx and VOC reactivity reductions in San Joaquin Valley California 1995-present. *Atmospheric Chemistry and Physics*, *12*(18), 8323–8339. https://doi.org/10.5194/acp-12-8323-2012

Qian, Y., Henneman, L. R. F., Mulholland, J. A., & Russell, A. G. (2019). Empirical Development of Ozone Isopleths: Applications to Los Angeles. *Environmental Science and Technology Letters*. https://doi.org/10.1021/acs.estlett.9b00160

Ulrickson, B. L., & Mass, C. F. (1990a). Numerical investigation of mesoscale circulations over the Los Angeles basin. Part I: a verification study. *Monthly Weather Review*. https://doi.org/10.1175/1520-0493(1990)118<2138:NIOMCO>2.0.CO;2

Ulrickson, B. L., & Mass, C. F. (1990b). Numerical investigation of mesoscale circulations over the Los Angeles basin. Part II: synoptic influences and pollutant transport. *Monthly Weather Review*. https://doi.org/10.1175/1520-0493(1990)118<2162:NIOMCO>2.0.CO;2

Wong, D. C., Pleim, J., Mathur, R., Binkowski, F., Otte, T., Gilliam, R., Pouliot, G., Xiu, A., Young, J. O., & Kang, D. (2012). WRF-CMAQ two-way coupled system with aerosol feedback: Software development and preliminary results. *Geoscientific Model Development*. https://doi.org/10.5194/gmd-5-299-2012

Yu, X., Stuart, A. L., Liu, Y., Ivey, C. E., Russell, A. G., Kan, H., Henneman, L. R. F., Sarnat, S. E., Hasan, S., Sadmani, A., Yang, X., & Yu, H. (2019). On the accuracy and potential of Google Maps location history data to characterize individual mobility for air pollution health studies. *Environmental Pollution*. https://doi.org/10.1016/j.envpol.2019.05.081

Chapter 1

Part 1

A data-driven approach for characterizing community scale air pollution exposure disparities in inland Southern California

The works used in this chapter were previously published in Journal of Aerosol Science.

Abstract

In 2017, Assembly Bill 617 was approved in the state of California, which mandated the allocation of resources for addressing air pollutant exposure disparities in underserved communities across the state. The bill stipulated the implementation of community scale monitoring and the development of local emissions reductions plans. The aim was to develop a streamlined, robust, and accessible $PM_{2.5}$ exposure assessment approach to support environmental justice analyses. The search is to characterize individual $PM_{2.5}$ exposure over multiple 24-hr periods in the inland Southern California region, which includes the underserved community of San Bernardino, CA. Personal sampling took place over five weeks in Spring of 2019, and personal $PM_{2.5}$ exposure was monitored for 18 adult participants for multiple, consecutive 24-hr periods. Exposure and location data were available at five-second resolution, and participant data recovery was 50.8% on average. A spatial clustering algorithm was used to classify data points as one of seven microenvironments. Mean and median personal-ambient $PM_{2.5}$ ratios were aggregated along SES lines for eligible datasets. GIS-based spatial clustering facilitated efficient microenvironment classification for more than 920,000 data points. Mean (median) personal-ambient ratios ranged from 0.02 (0.00) to 3.49 (0.55) for each microenvironment when aggregated

along SES-lines. Aggregated ratios indicated that participants from the lowest SES community experienced higher home exposures compared to participants of all other communities over consecutive 24-hr monitoring periods, despite high participant mobility and relatively low variability in ambient $PM_{2.5}$ during the study. The methods described here highlight the robust and accessible nature of the personal sampling campaign, which was specifically designed to reduce participant fatigue and engage members of the inland Southern California community who may experience barriers when engaging with the scientific community. This approach is promising for larger-scale, community-focused, personal exposure campaigns for direct and accurate analysis of environmental justice.

Introduction

Ambient particulate matter (PM) has been widely studied, and researchers have carefully examined the impact of PM exposure on human health. Many air monitoring stations are operated in the U.S. to measure the trends and composition of ambient PM in support of the National Ambient Air Quality Standards (NAAQS). However, ambient PM concentrations may not reflect actual daily personal exposure (PE) (Koistinen et al., 2004a). Further the sparseness of the monitoring network leads to low spatial resolution data and necessitates gap-filling, which affects the accuracy of PM exposure assessments that are based on ambient measurements (Yu et al., 2019).

People spend most of their time indoors (approximately 85-90%) and are most frequently exposed to indoor pollutants(Long et al., 2001). Home and workplace are the two most dominant indoor microenvironments. Indoor PM originates from cooking, smoking, cleaning products, vacuuming, and dusting; while in offices, PM is emitted from printing, mechanical grinding,

consumer products, and dusting. The Environmental Protection Agency (EPA) carried out the particulate total exposure assessment methodology (PTEAM) study on 178 non-smoking randomly selected homes in Riverside, CA. The study showed that indoor $PM_{2.5}$ (PM with an aerodynamic diameter less than or equal to 2.5 μm) levels were slightly lower than outdoor levels during the day. However, the indoor $PM_{2.5}$ levels were higher than ambient levels at night (Clayton et al., 1993; Özkaynak et al., 1996; Thomas et al., 1993a). Although ambient $PM_{2.5}$ penetrates into indoor environments, individual behaviors and living conditions are found to be the most important factors that affect indoor concentrations of PM (Kulmala et al., 1999; Long et al., 2001; Wallace, 1996).

Further, human mobility must also be taken into account for accurate exposure assessment. Yu et al. compared call detail record and home-based methods to estimate biases in exposure methods. The study showed that the home-based method both over- and under-estimates air pollutant exposure levels (Yu et al., 2018). In addition, many studies have used outputs from chemical transport models to verify the misclassification when using central monitor concentrations (CMC) to represent the exposure near the monitoring sites. Hu et al. showed that the population weighted concentrations of primary $PM_{2.5}$ of the model differ from the CMC values by -40 to +60%. The misclassification could be significant when assuming the same representative distance across central monitoring sites for multiple pollutants in a large-scale, spatial and temporal epidemiology studies (Hu et al., 2019).

Advancements in low-cost environmental sensing technologies have enabled the development of small, portable, and relatively precise PM sensors for personal exposure assessment. In a recent study by Quinn et al., filter-based, wearable, automated

microenvironmental aerosol samplers (AMAS) were used to conduct a personal exposure study with 25 high school students in Fresno, CA (Quinn et al., 2018). The wearable AMAS enabled the measurement of black carbon and oxidative potential in targeted microenvironments, but the measurements were coarsely-resolved in time. Further, low-cost optical PM sensors have very high sampling frequencies, and low-cost sensing measurements are moderately accurate (Feenstra et al., 2019). The Plantower PMS (v. 1003/3003) is a commonly used optical sensor, and has a correlation coefficient of 0.88 with the federal reference method (FRM), which reflects the viability of the sensor for exposure measurements (Kelly et al., 2017). Combined with Internet of Things (IoT) technology, the Plantower PMS can be further integrated to deliver more functionalities to end users. Data collected from a low-cost sensing device or IoT network can be uploaded to the cloud and made available in near-real-time to users. Despite all the conveniences of low-cost sensing, there are still room for improvements of PM sensor accuracy. Sensors require consistent calibration, and the measurements may require additional post-processing (Zheng et al., 2018a).

In this paper, a pilot-scale personal exposure campaign was detailed using wearable $PM_{2.5}$ sensors with real-time, remote monitoring capability. The study engaged residents of five inland Southern California cities and captured spatial and temporal variability of $PM_{2.5}$ exposures over multiple, consecutive 24-hour periods. The main objective of this pilot study was to develop and implement a high-resolution monitoring and analysis framework for characterizing $PM_{2.5}$ exposure variability for individuals from different cities of residence and subsequently different socioeconomic status (SES) neighborhoods. As Southern California historically has high ambient $PM_{2.5}$ levels, the search to understand which microenvironments posed the greatest exposure risk

in the region. The study elucidates the behavior-dependent patterns of $PM_{2.5}$ exposure in a high-traffic, industrialized region of Southern California.

Materials and Methods

Study Area

The personal exposure study was conducted in inland Southern California, better known as the Inland Empire, covering an area of approximately 200 square miles (Figure 1.1). More specifically, the study area includes the cities of Moreno Valley (2018 U.S. Census population of 209,050), Redlands (71,596), Riverside (330,063), San Bernardino (215,941), and Yucaipa (53,682), CA (Table 1.1). In 2018, median household income estimates were $63,572, $72,523, $65,313, $43,136, and $63,657; and poverty rates were 19.9%, 13.6%, 15.6%, 28.4%, and 12.3%, respectively (U.S. Census Bureau). The major routes that service these cities include interstate routes 10, 15, and 215, and U.S. highways 60, 66, 91, and 210. The major air pollution sources in inland Southern California are on-road traffic, off-road mobile sources (e.g., railyard equipment), industrial point sources (e.g., cement manufacturing and power generating facilities), and smaller point sources (e.g., auto body shops, residential combustion, and restaurants). In recent years, the logistics industry has expanded in the region, prompting the construction of large warehouses that rely on heavy-duty vehicles for goods transport.

The recently implemented California Assembly Bill 617 was designed partially to address disproportionate impacts of air pollution in environmental justice communities, and San Bernardino was selected as a Phase 1 community in 2018 (Garcia, 2017). Previous studies have highlighted health disparities in the San Bernardino community due to its proximity to a large railyard (Spencer-Hwang et al., 2016, 2015). Through the study, the search is to understand

personal exposure patterns as they relate to the unique environmental and socioeconomic characteristics of inland Southern California.

Sampling Campaign

For the sampling campaign, 18 adult participants (18 years and older; 61% males; 55% Latinx) with varied occupations (50% identified as college students) were recruited. All sampling activities and interactions with participants were pre-approved by the University of California, Riverside Institutional Review Board (protocol number: *HS 18-206*). The overall campaign took place over a five-week period from 03-10-2019 to 04-14-2019. Each week on Sunday, a PM monitoring pack to four participants was distributed, except for the first week which had two participants (Figure 1.2). Participants kept the packs for a duration of seven days, allowing the assessment of inter- and intra-day exposure variability for each individual.

Participant locations were tracked with GPS data loggers. Participants were required to carry the packs during the day, and packs were placed in their bedroom or living spaces at night. After the seven-day deployment, the packs were returned to the research facility. The GPS data from the data loggers was retrieved and removed before the next deployment for privacy. One participant's GPS data was missing, so this dataset was removed, and subsequent analyses were carried out for 17 datasets. The participant breakdown by city was the following: two from Moreno Valley, two from Redlands, five from Riverside, six from San Bernardino, and two from Yucaipa. The uncertainty introduced by the sample size and city breakdown is noticed. However, the pilot study generated useful insights that will be leveraged during the larger phase two sampling campaign.

Monitoring Equipment

Each monitoring pack (total = 4) included a battery-powered PM monitor, a GlobalSat-DG-500 (New Taipei City, Taiwan) GPS module, a Huawei Wi-Fi hotspot, Elitech temperature log, and necessary accessories. The PM monitors are developed by Applied Particle Technology (APT, St. Louis, Missouri, USA) and utilize the Plantower PMSA003 optical sensor (Figure 1.2). The monitors are commercially available, and our research team was not directly involved with monitor development. The dimensions of the PM monitors are 2 in. x 1 in. x 2.25 in. (L x W x H). The APT monitor provided four PM_1, $PM_{2.5}$, and PM_{10} measurements per minute, but only$PM_{2.5}$ measurements were analyzed due to the extensive literature and relevance of $PM_{2.5}$ exposure and health, and due to the availability of suitable reference measurements for monitor evaluation. The APT monitors also provide measurements of relative humidity and temperature, and the data are uploaded in real-time via the mobile hotspot to the vendor-hosted web interface. The size, simplicity, mobility, and accessibility of the APT device was ideal for community engagement. The sampling rate of the PM monitor was once every 15 seconds, totaling a maximum of approximately 40,320 possible measurements at the end of the seven-day sampling period, plus or minus a few hours of measurements depending on the scheduled pick-up and drop-off times. Measurements were made every 15 seconds, and data were immediately pushed to the cloud if Wi-Fi connectivity was available. If Wi-Fi connectivity was unavailable, all data are stored locally then pushed once connectivity resumed. All data were retrieved from the cloud via the vendor-managed user interface.

Data Processing

Although a uniform usage protocol was established for the study, datasets had varying degrees of availability due to the operating habits of the participants. All missing PM measurements were assigned as "-9999", then PM data were synced with the GPS data by their dates and timestamps. Since the GPS data logger's sampling rate was once every five seconds, a linear interpolation was performed on the PM data from 15 to five second intervals to obtain the highest resolution for the datasets. The resulting combined datasets provide the date, time of day, $PM_{2.5}$ concentrations, relative humidity, temperature, and the corresponding latitude and longitude. As a note, the GPS position was intermittently measured at times because the data logger stopped recording if the no movement was detected after 30 seconds. To account for the idling periods, the previous latitude and longitude were assigned to the missing timestamp if the distance between the two intervals was less than 20 meters (Figure 1.3). When the distance was greater than 20 meters and less than or equal 50 meters, linear interpolation was performed between the two points. A distance greater than 50 meters was assigned "NaN" and considered an invalid data point due to uncertainty in participant mobility during the idle period. The five-second syncing lends a maximum of approximately 120,960 possible data points for each participant.

Co-location and Adjustments

The personal PM monitors were co-located at the Mira Loma Van Buren (MLVB, AQS ID: 060658005) air monitoring site to evaluate the hourly performance of the monitors. The wearable monitors were housed in a home-built enclosure and positioned the enclosure near the site's federal equivalent method (FEM) $PM_{2.5}$ samplers (Figure 1.4). The enclosure was built using steel

mesh panels to maximize the air flow over the monitors. The monitors were kept on-site for two weeks, and the activities of each sensor were continuously monitored through the web server to ensure that each device was operating optimally. At the end of the co-location period, $PM_{2.5}$ reference data was obtained for the performance analysis. For the study, polynomial fitting was used to adjust the raw data to the FEM reference data. The measurements were determined to be uninfluenced by relative humidity and temperature, hence the polynomial fittings were solely based on two parameters: reference measurements and raw measurements (Note S1). The fitting method is well described in a paper by Zheng et al. (Zheng et al., 2018b).

Data Analysis

Microenvironments of five-second all data points were classified based on the GPS measurements. The density-based spatial clustering of applications with noise (DBSCAN, (Schubert et al., 2017)) algorithm in the QGIS (https://www.qgis.org/) open source GIS platform and DBSCAN clusters points based on a two-dimensional implementation (QGIS Development Team, 2019) were used. Each spatial cluster was defined by mandating a minimum size of 120 $PM_{2.5}$/GPS measurements within a maximum distance of 0.0005 degrees (~55 meters). This minimum size corresponds to a minimum of 120 * 5 seconds = 10 minutes of personal exposure data, and 55 meters is roughly the range of uncertainty of the GPS data logger. The clusters were manually evaluated and assigned a microenvironment class and activity by overlaying the clusters onto Google Maps. Microenvironment classes included home (H), work (or university, W), restaurant (R), retail (RE), leisure indoor (LI), leisure outdoor (LO), and transient (T); and microenvironment was classified and assigned to the cluster based on the proximity of the cluster center to labels available in Google Maps. The home and work/university clusters are identified

using the address and work/school status information provided by the participants, respectively. Restaurant and retail clusters are identified by their proximity to points of interest on Google Maps. Leisure indoor includes all clusters in proximity to non-work, non-retail indoor locations (e.g., churches and recreation centers), and leisure outdoor clusters are non-work, non-retail locations in the outdoors (e.g., parks and trails). If a participant's workplace was reported as a retail or a restaurant, then those clusters were classified as retail or restaurant. The "transient" classification indicates that the speed measurement was greater than 10 kilometers per hour, regardless of prior cluster classification. The "unclassified" classification was given to non-clustered, non-transient data points. There are no assumptions about participant mobility within the microenvironment. Further, clusters are identified and classified for each individual participant, therefore no clusters have influences from multiple participants. An example of spatial clusters can be found in Figure 1.5.

Ambient $PM_{2.5}$ Contour Fields

A $PM_{2.5}$ contour mesh over Southern California was constructed to compare the personal exposure of $PM_{2.5}$ to ambient $PM_{2.5}$. Participant mobility varied, and measurement locations were up to 100 miles away from the main study location. The input data for the ambient $PM_{2.5}$ spatial fields were accessed from the regulatory monitoring network of the South Coast Air Quality Management District. To construct hourly contour fields, cubic interpolation was performed on hourly $PM_{2.5}$ measurements from 18 monitoring stations. Participant coordinates were paired to the corresponding contour location, resulting in corresponding ambient and personal $PM_{2.5}$ data points for all participants.

Results

Personal and Ambient Data Overview

Calibration of PM monitors using the polynomial fittings resulted in good agreement between the adjusted personal measurements and reference $PM_{2.5}$ measurements. The mean bias for the four monitors ranged from -0.11 to 0.61, slopes ranged from 0.99 to 1.10, intercepts ranged from 0.012 to 0.75, and R^2 ranged from 0.41-0.45 (Note S2).

For interpolated personal measurements, data recovery is defined as the percentage of five-second data points available out of the total possible data points for each participant's sampling period (range: 0.5 – 95.6%). Mean data recovery was 50.8%, corresponding to 54,120 valid data points per participant; and median data recovery was 51.8%, corresponding to 53,921valid data points per participant (Table 1.2). Further explanation of data missingness is provided in the Discussion. In comparison to prior studies the approach was successful in collecting an exceptionally large amount of data, where valid personal data points from all 17 participants totaled 920,045 (Bekö et al., 2015; Li et al., 2017; Minet et al., 2018; Piedrahita et al., 2017; Quinn et al., 2018; Thomas et al., 1993b).

During planning for the sampling campaign, one concern was week-to-week variability in ambient $PM_{2.5}$, which may bias indoor-outdoor ratios. Ambient $PM_{2.5}$ concentrations are lowest in the spring season in southern California. Springtime ambient $PM_{2.5}$ concentrations are stable and not heavily affected by exceptional events and meteorology-induced aerosol formation. Therefore, the chosen sampling period was optimal for a multi-week pilot study. Ambient data were extracted from contours of hourly measurements from regulatory monitoring stations (Figure 1.6) and paired with the corresponding personal measurements. Median ambient $PM_{2.5}$

concentrations for each sampling week ranged from 4.4 to 10.2 μg m^{-3}, and maximum concentrations ranged from 22.3 to 28.2 μg m^{-3} (Figure 1.7). Weekly concentrations ranged from near-zero to ~30 μg m^{-3} every week, and therefore week-to-week variability was not considered a confounder in this study.

Exposure and Activity

Time series of individual personal exposure measurements identify acute $PM_{2.5}$ exposure episodes (less than one hour, > 35 μg m^{-3}), and acute exposures were highly variable for all participants. Time series of consecutive, 24-hour personal measurements at 5-seconds resolution along with the corresponding ambient hourly measurement for four participants are highlighted. Maximum acute exposures ranged from approximately 70 (Redlands) to 2500 (Moreno Valley) μg m^{-3}, further justifying the need for individual level analysis of exposure risk.

Participant 2 (San Bernardino) experienced the highest exposures in the home microenvironment in the late afternoons and early evening, as well as in an indoor residential microenvironment that was not classified as home. Participant 5 (Redlands) experienced all acute episodes in the work/university microenvironment, and the residential location university housing. Participant 5 exposures were not as severe as the other highlighted exposures.

Participant 13 (Moreno Valley) experienced frequent, extreme exposures with consistently high measurements greater than 500 μg m^{-3} in the home and leisure indoor microenvironments. High measurements were observed in short intervals in the restaurant microenvironments, specifically a popular burger and coffee chain. High measurements were also infrequently observed in the transient and work microenvironments. Based on the short duration (< 10 minutes) of the extreme exposures and the occurrence in the majority of

microenvironments, it is suspected that the participant is a smoker. Participant 15 (Riverside) experienced exposures greater than 100 μg m^{-3} in the home microenvironment, and consistently elevated $PM_{2.5}$ was observed during time spent in a restaurant microenvironment (range 20–50 μg m^{-3}). Time series for all participants can be found in Note S3 in the Supplementary Material.

Inter-City Comparative Analysis

Personal and ambient $PM_{2.5}$ data were aggregated for cities with two or more participants with 50% or greater data recovery, which was the criteria for inclusion in the inter-city analysis (Table 1.3). Results from those participants were then stratified along SES lines: Redlands/Riverside (N = 5, high SES) and San Bernardino (N = 4, low SES); there were no datasets from Moreno Valley and Yucaipa that met the aggregation criteria. Average data recovery for these participants was 73% (Redlands/Riverside) and 72% (San Bernardino). Aggregated median ambient concentrations were consistently higher than median personal concentrations, and the highest median personal concentrations were observed in home microenvironment for both SES groups. San Bernardino personal medians in the home microenvironment were higher despite having slightly lower ambient medians than Redlands/Riverside. Short-term personal exposures were higher than 20 μg m^{-3} in work, university, restaurant, retail, leisure indoor, and transient microenvironments for aggregated datasets (Figure 1.9).

For SES-aggregated datasets, mean personal-ambient (P-A) ratios for each microenvironment ranged from 0.02 to 3.49, and median ratios ranged from 0.00 to 0.55 (Table 1.4). Higher mean ratios compared to median ratios reflect the influence of the outliers in the personal measurements. Ratios less than one indicate that personal environments had lower $PM_{2.5}$ levels than those derived from ambient data. For classified microenvironment clusters, the

highest mean P-A ratios were observed in the retail 1.45 (0.60, Redlands/Riverside) and home (3.49, San Bernardino) microenvironments (Table 1.4). Redlands/Riverside had ratios greater than one for transient (1.17) and unclassified data points (2.81), while the mean home ratio was 0.76. San Bernardino retail ratio was 2.47. The highest median P-A ratios were observed in the home microenvironments for both Redlands/Riverside (0.16) and San Bernardino (0.55) for classified clusters. Wilcoxon rank sum tests indicated significant ($p < 0.05$) differences between non-outlier personal-ambient data pairs for all microenvironments and for every participant with the exception of the leisure indoor and restaurant microenvironments for Participants 5 and 8, respectively. Outlier personal data and corresponding ambient data were excluded from the Wilcoxon tests. Mean and median ratios for all participants can be found Table 1.5 and Table 1.6 in the Supplementary Material.

Discussion

The majority of data points were classified as home for the highlighted participants (mean: 65%, median: 69%) (Table 1.2). This is slightly higher, but consistent with previous personal exposure studies (Bekö et al., 2015; Hsu et al., 2020; Quinn et al., 2018). Data points were classified in these microenvironments at an average of 31% (median: 16%) of the time, therefore non-home exposures may be significant in the long-term (Table 1.2). Transient $PM_{2.5}$ measurements were within range of a previous personal exposure study conducted in California (Ham et al., 2017). Microenvironment distributions of personal and ambient measurements can be found in Note S4 in the Supplementary Material.

Calculations of time spent in each microenvironment are impacted by data recovery, and charging protocols were best adhered to in the home environments near a convenient supply of

electricity. There were compliance issues during sampling that affected data recovery, which is common in human subjects research (Chenail, 2011; Mehra, 2001). Monitor mobility and real-time data transfer of PM monitors enabled the high-resolution personal sampling of the study. However, data collection was impeded when component batteries drained, although a charging schedule was provided but not always adhered to. At times the hardware stalled, or data transfer was limited by availability of Wi-Fi signal. Participant accidents with the monitors, while rare, also interrupted sampling; minor damages to the protective casings were mended before redeployment.

The monitoring approach intuitively identifies participants that may be actively or passively exposed to cigarette or vaping smoke, as very high personal measurements (> 100 µg m^{-3}) are classified as outliers in a five-second resolution dataset (Figure 1.9) (Götschi et al., 2002; Koistinen et al., 2004b; Salmon et al., 2018; Slezakova et al., 2009). Suspected smoking events occur at relatively shorter time scales throughout the day and are easily identified in the time series and boxplots of personal measurements. Consequently, median P-A ratios derived from high temporal resolution data are useful for evaluating non-smoking related $PM_{2.5}$ exposures when smoking status is undisclosed. Therefore, when comparing the bulk (non-outliers) of personal and ambient measurements for Redlands/Riverside microenvironments, personal $PM_{2.5}$ measurements are much less than ambient $PM_{2.5}$. Conversely, the San Bernardino median home microenvironment exposure was most similar to the corresponding median ambient exposure (Table 1.3). Undisclosed smoking adds uncertainty to this study is recognized; however, the temporal resolution of the data enables the identification of potential smokers. In future efforts, smokers or individuals living in a smoking household will be pre-identified or excluded.

Considering the relatively small number of participants in the study, definitive generalizations cannot be made regarding influences of residential location. However, the large amount of measurements analyzed here provides a preliminary, yet robust, investigation of exposure disparities. San Bernardino (highest poverty rates, lowest median household income) participants with greater than 50% data recovery experienced higher home exposures compared with participants from other cities. Redlands/Riverside (second/third lowest poverty rate, highest/second-highest household incomes) participants overall had lower home personal exposures and experienced higher personal exposures outside of the home. Since most time was spent in the home microenvironment for the majority of participants, San Bernardino participants were more likely to be exposed to higher $PM_{2.5}$ concentrations, even when taking into account the high degree of mobility of participants which is reflected in the diversity of classified microenvironments.

Conclusions

The pilot study highlights the variability in community-scale exposure in a socioeconomically diverse air basin that is heavily burdened by air pollution. A novel spatial clustering approach was applied to classify the microenvironments of more than 900,000 high temporal resolution personal exposure data points. Results from the study indicate that participants from the lowest socioeconomic status community experienced overall higher personal exposures over consecutive 24-hr monitoring periods, despite high participant mobility and low variability in ambient $PM_{2.5}$ during the study. The inclusive monitoring protocol minimizes participant fatigue and is well-suited for real-time, long-term characterization of $PM_{2.5}$ exposure disparities in underserved communities. $PM_{2.5}$ serves as a useful surrogate species for many other

air pollutants that may influence disproportionate exposures. The application of the streamlined, data-driven methods in a larger-scale exposure study will further elucidate personal exposure disparities along racial and socioeconomic lines.

Data Availability

In accordance with the University of California, Riverside Institutional Review Board, personal data may only be distributed in an aggregated form to preserve participant privacy. All aggregated and anonymized data are summarized in the Supplementary Material.

Acknowledgements

We acknowledge Aaron Garcia, Saray Rodriguez, Hector Soto, Priscilla Villegas, and David Wilson for their assistance with the field data collection. We thank Chela Larios of Center for Community Action and Environmental Justice and Yassi Kavezade of Sierra Club for their help with recruiting participants. We thank Jiaxi Fang and Tandeep Chadha of Applied Particle Technology for their technical assistance, and Brandon Feenstra and Nelson Marquez for facilitating co-location at the Mira Loma Van Buren air monitoring site.

References

Bekö, G., Kjeldsen, B.U., Olsen, Y., Schipperijn, J., Wierzbicka, A., Karottki, D.G., Toftum, J., Loft, S., Clausen, G., 2015. Contribution of various microenvironments to the daily personal exposure to ultrafine particles: Personal monitoring coupled with GPS tracking. Atmospheric Environment 110, 122–129. https://doi.org/10.1016/j.atmosenv.2015.03.053

Chenail, R.J., 2011. Interviewing the investigator: Strategies for addressing instrumentation and researcher bias concerns in qualitative research. Qualitative Report.

Clayton, C.A., Perritt, R.L., Pellizzari, E.D., Thomas, K.W., Whitmore, R.W., Wallace, L.A., Ozkaynak, H., Spengler, J.D., 1993. Particle Total Exposure Assessment Methodology (PTEAM) study: distributions of aerosol and elemental concentrations in personal, indoor, and outdoor air samples in a southern California community. Journal of exposure analysis and environmental epidemiology.

Feenstra, B., Papapostolou, V., Hasheminassab, S., Zhang, H., Boghossian, B. Der, Cocker, D., Polidori, A., 2019. Performance evaluation of twelve low-cost PM2.5 sensors at an ambient air monitoring site. Atmospheric Environment 216, 116946. https://doi.org/10.1016/j.atmosenv.2019.116946

Garcia, C., 2017. Assembly Bill No. 617.

Götschi, T., Oglesby, L., Mathys, P., Monn, C., Manalis, N., Koistinen, K., Jantunen, M., Hänninen, O., Polanska, L., Künzli, N., 2002. Comparison of Black Smoke and PM 2.5 Levels in Indoor and Outdoor Environments of Four European Cities. Environmental Science & Technology 36, 1191–1197. https://doi.org/10.1021/es010079n

Ham, W., Vijayan, A., Schulte, N., Herner, J.D., 2017. Commuter exposure to PM2.5, BC, and UFP in six common transport microenvironments in Sacramento, California. Atmospheric Environment 167, 335–345. https://doi.org/10.1016/j.atmosenv.2017.08.024

Hsu, W.-T., Chen, J.-L., Candice Lung, S.-C., Chen, Y.-C., 2020. PM2.5 exposure of various microenvironments in a community: Characteristics and applications. Environmental Pollution 263, 114522. https://doi.org/10.1016/j.envpol.2020.114522

Hu, J., Ostro, B., Zhang, H., Ying, Q., Kleeman, M.J., 2019. Using Chemical Transport Model Predictions To Improve Exposure Assessment of PM 2.5 Constituents. Environmental Science & Technology Letters 6, 456–461. https://doi.org/10.1021/acs.estlett.9b00396

Kelly, K.E., Whitaker, J., Petty, A., Widmer, C., Dybwad, A., Sleeth, D., Martin, R., Butterfield, A., 2017. Ambient and laboratory evaluation of a low-cost particulate matter sensor. Environmental Pollution. https://doi.org/10.1016/j.envpol.2016.12.039

Koistinen, K.J., Edwards, R.D., Mathys, P., Ruuskanen, J., Künzli, N., Jantunen, M.J., 2004a. Sources of fine particulate matter in personal exposures and residential indoor, residential outdoor and workplace microenvironments in the Helsinki phase of the EXPOLIS study. Scandinavian Journal of Work, Environment and Health 20, 36–46.

Koistinen, K.J., Edwards, R.D., Mathys, P., Ruuskanen, J., Künzli, N., Jantunen, M.J., 2004b. Sources of fine particulate matter in personal exposures and residential indoor, residential outdoor and workplace microenvironments in the Helsinki phase of the EXPOLIS study. Scandinavian Journal of Work, Environment and Health 20, 36–46.

Kulmala, M., Asmi, A., Pirjola, L., 1999. Indoor air aerosol model: The effect of outdoor air, filtration and ventilation on indoor concentrations. Atmospheric Environment. https://doi.org/10.1016/S1352-2310(99)00070-9

Li, Z., Che, W., Frey, H.C., Lau, A.K.H., Lin, C., 2017. Characterization of PM 2.5 exposure concentration in transport microenvironments using portable monitors. Environmental Pollution 228, 433–442. https://doi.org/10.1016/j.envpol.2017.05.039

Long, C.M., Suh, H.H., Catalano, P.J., Koutrakis, P., 2001. Using time- and size-resolved particulate data to ouantify indoor penetration and deposition behavior. Environmental Science and Technology. https://doi.org/10.1021/es001477d

Mehra, B., 2001. Bias in Qualitative Research: Voices from an Online Classroom. The Qualitative Report.

Minet, L., Liu, R., Valois, M.-F., Xu, J., Weichenthal, S., Hatzopoulou, M., 2018. Development and Comparison of Air Pollution Exposure Surfaces Derived from On-Road Mobile Monitoring and Short-Term Stationary Sidewalk Measurements. Environmental Science & Technology 52, 3512–3519. https://doi.org/10.1021/acs.est.7b05059

Özkaynak, H., Xue, J., Spengler, J., Wallace, L., Pellizzari, E., Jenkins, P., 1996. Personal exposure to airborne particles and metals: Results from the particle team study in Riverside, California. Journal of Exposure Analysis and Environmental Epidemiology.

Piedrahita, R., Kanyomse, E., Coffey, E., Xie, M., Hagar, Y., Alirigia, R., Agyei, F., Wiedinmyer, C., Dickinson, K.L., Oduro, A., Hannigan, M., 2017. Exposures to and origins of carbonaceous PM2.5 in a cookstove intervention in Northern Ghana. Science of The Total Environment 576, 178–192. https://doi.org/10.1016/j.scitotenv.2016.10.069

QGIS Development Team, 2019. QGIS User Guide - Release 3.4.

Quinn, C., Miller-Lionberg, D.D., Klunder, K.J., Kwon, J., Noth, E.M., Mehaffy, J., Leith, D., Magzamen, S., Hammond, S.K., Henry, C.S., Volckens, J., 2018. Personal Exposure to PM 2.5 Black Carbon and Aerosol Oxidative Potential using an Automated Microenvironmental Aerosol Sampler (AMAS). Environmental Science & Technology 52, 11267–11275. https://doi.org/10.1021/acs.est.8b02992

Salmon, M., Milà, C., Bhogadi, S., Addanki, S., Madhira, P., Muddepaka, N., Mora, A., Sanchez, M., Kinra, S., Sreekanth, V., Doherty, A., Marshall, J.D., Tonne, C., 2018. Wearable camera-derived microenvironments in relation to personal exposure to PM2.5. Environment International 117, 300–307. https://doi.org/10.1016/j.envint.2018.05.021

Schubert, E., Sander, J., Ester, M., Kriegel, H.P., Xu, X., 2017. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. ACM Trans. Database Syst. 42, 1–21. https://doi.org/10.1145/3068335

Slezakova, K., Castro, D., Pereira, M.C., Morais, S., Delerue-Matos, C., Alvim-Ferraz, M.C., 2009. Influence of tobacco smoke on carcinogenic PAH composition in indoor PM10 and PM2.5. Atmospheric Environment 43, 6376–6382. https://doi.org/10.1016/j.atmosenv.2009.09.015

Spencer-Hwang, R., Soret, S., Knutsen, S., Shavlik, D., Ghamsary, M., Beeson, W.L., Kim, W., Montgomery, S., 2015. Respiratory Health Risks for Children Living Near a Major Railyard. Journal of Community Health 40, 1015–1023. https://doi.org/10.1007/s10900-015-0026-0

Spencer-Hwang, R., Soret, S., Valladares, J., Torres, X., Pasco-Rubio, M., Dougherty, M., Kim, W., Montgomery, S., 2016. Strategic Partnerships for Change in an Environmental Justice Community: The ENRRICH Study. Progress in Community Health Partnerships: Research, Education, and Action 10, 541–550. https://doi.org/10.1353/cpr.2016.0062

Thomas, K.W., Pellizzari, E.D., Clayton, C.A., Whitaker, D.A., Shores, R.C., Spengler, J., Ozkaynak, H., Froehlich, S.E., Wallace, L.A., 1993a. Particle Total Exposure Assessment Methodology (PTEAM) 1990 study: method performance and data quality for personal, indoor, and outdoor monitoring. Journal of exposure analysis and environmental epidemiology.

Thomas, K.W., Pellizzari, E.D., Clayton, C.A., Whitaker, D.A., Shores, R.C., Spengler, J., Ozkaynak, H., Froehlich, S.E., Wallace, L.A., 1993b. Particle Total Exposure Assessment Methodology (PTEAM) 1990 study: method performance and data quality for personal, indoor, and outdoor monitoring. Journal of exposure analysis and environmental epidemiology 3, 203–226.

Wallace, L., 1996. Indoor Particles: A Review. Journal of the Air and Waste Management Association. https://doi.org/10.1080/10473289.1996.10467451

Yu, H., Russell, A., Mulholland, J., Huang, Z., 2018. Using cell phone location to assess misclassification errors in air pollution exposure estimation. Environmental Pollution 233, 261–266. https://doi.org/10.1016/j.envpol.2017.10.077

Yu, X., Stuart, A.L., Liu, Y., Ivey, C.E., Russell, A.G., Kan, H., Henneman, L.R.F., Sarnat, S.E., Hasan, S., Sadmani, A., Yang, X., Yu, H., 2019. On the accuracy and potential of Google Maps location history data to characterize individual mobility for air pollution health studies. Environmental Pollution. https://doi.org/10.1016/j.envpol.2019.05.081

Zheng, T., Bergin, M.H., Johnson, K.K., Tripathi, S.N., Shirodkar, S., Landis, M.S., Sutaria, R., Carlson, D.E., 2018a. Field evaluation of low-cost particulate matter sensors in high-and low-concentration environments. Atmospheric Measurement Techniques 11, 4823–4846. https://doi.org/10.5194/amt-11-4823-2018

Zheng, T., Bergin, M.H., Johnson, K.K., Tripathi, S.N., Shirodkar, S., Landis, M.S., Sutaria, R., Carlson, D.E., 2018b. Field evaluation of low-cost particulate matter sensors in high-and low-concentration environments. Atmospheric Measurement Techniques. https://doi.org/10.5194/amt-11-4823-2018

Tables

Table 1.1. Demographics of the cities included in the personal exposure sampling.

City	Population (2018)	Median Household Income	Poverty Rate
Moreno Valley	209,050	$63,572	19.9%
Redlands	71,596	$72,523	13.6%
Riverside	330,063	$65,313	15.6%
San Bernardino	215,941	$43,136	28.4%
Yucaipa	53,682	$63,657	12.3%

Table 1.2. Percent data recovery from participants, and the percentage of time participants spent in each microenvironment. The table header indicates the microenvironment classifications: home (H), work (or university, W), restaurant (R), retail (RE), leisure indoor (LI), leisure outdoor (LO), transient (T), and unclassified (U). The bold number indicates the microenvironment of longest time spent for each participant. Participant 12's dataset was not recovered. Star (*) indicates that the participant is a part-time or full-time student. Total valid data points = 920,045.

Participant	Valid Data Points	Recovery (%)	H (%)	W (%)	R (%)	RE (%)	LI (%)	LO (%)	T (%)	U (%)
1*	12160	11.7	31	**37**	12.6	0.0	0.0	0.0	10.9	8.5
2	95982	94.5	**69.7**	16.1	0.4	1.2	5.5	2.4	0.3	4.5
3*	85596	89.3	**72**	0.0	12.5	6.0	2.5	0.0	0.4	6.6
4	507	0.5	0	**88.8**	0.0	0.0	0.0	0.0	0.0	11.2
5*	92648	74.2	0	**97.9**	0.0	0.0	1.3	0.0	0.3	0.6
6*	88883	92.3	**75.2**	0.3	0.7	1.4	5.9	0.0	15.2	1.4
7	66098	51	**72**	0.0	0.0	22.1	2.2	0.1	0.2	3.5
8*	50847	52.7	**75.3**	18	1.2	0.6	0.7	0.4	0.7	3.1
9	56500	55.4	**95.8**	0.0	0.0	0.0	2.6	0.0	0.0	1.6
10	28378	27.7	**93.3**	0.0	0.7	0.0	0.0	2.1	0.1	3.9
11	51342	50.6	**60.3**	19.1	2.6	4.1	1.3	0.0	0.6	11.9
13*	78261	76.9	**71.2**	15.3	1.1	1.7	4.3	0.0	0.9	5.4
14*	18794	18.5	**93.5**	0.0	0.0	0.1	0.0	0.6	0.1	5.6
15*	102190	95.6	**88.5**	0.0	1.7	3.7	1.2	0.0	0.2	4.6
16	48823	38.5	**62**	0.7	3.5	12.7	4.6	2.4	0.1	14
17	33679	25.2	**76.7**	2.7	0.0	2.2	4.8	0.0	2.3	11.2
18*	9357	9.0	**97.9**	0.0	0.0	0.0	1.0	0.0	0.9	0.1
Average	54120	50.8	**66.7**	17.4	2.2	3.3	2.2	0.5	1.9	5.8
Median	53921	51.8	**72**	1.7	0.7	1.3	1.8	0.0	0.3	5.0

Table 1.3. Summary of the total number of valid data points, average data recovery, and median personal (ambient) $PM_{2.5}$ concentrations (μg m^{-3}) for Redlands and Riverside (N = 5), and San Bernardino (N = 4) participants with data recovery greater than 50%.

City	**Redlands and Riverside**		**San Bernardino**	
Number of Data Points (Average Data Recovery)	387,781 (73%)		302,305 (72%)	
	Personal (Ambient) (μg m^{-3})	% Time Spent	Personal (Ambient) (μg m^{-3})	% Time Spent
Home	1.67 (8.66)	66	5.33 (7.69)	69
Work or University	0.00 (8.91)	23	0.00 (3.85)	9
Restaurant	1.00 (9.36)	3	0.00 (4.50)	1
Retail	1.00 (8.64)	2	0.00 (7.48)	7
Leisure Indoor	0.00 (11.3)	2	2.00 (6.52)	4
Leisure Outdoor	0.00 (6.17)	0	0.00 (1.68)	1
Transient	1.00 (7.49)	0	0.00 (9.79)	4
Unclassified	1.00 (6.22)	3	0.00 (5.20)	5

Table 1.4. Mean (median) personal-ambient ratios by city of residence for Redlands and Riverside (N = 5) and San Bernardino (N = 4) participants with data recovery greater than 50%. Bold indicates higher personal $PM_{2.5}$ concentrations than the corresponding ambient concentrations.

City	**Redlands and Riverside**	**San Bernardino**
Home	0.76 (0.16)	**3.49** (0.55)
Work or University	0.30 (0.00)	0.06 (0.00)
Restaurant	0.35 (0.12)	0.48 (0.22)
Retail	1.45 (0.15)	0.09 (0.00)
Leisure Indoor	0.28 (0.00)	**2.47** (0.29)
Leisure Outdoor	0.22 (0.00)	0.02 (0.00)
Transient	**1.17** (0.08)	0.14 (0.00)
Unclassified	**2.81** (0.21)	0.23 (0.00)

Table 1.5. Mean of personal-ambient ratios for each participant. A ratio greater than one (bolded) indicates a higher personal exposure compared to exposure derived from ambient $PM_{2.5}$ concentrations. The table header indicates the microenvironment classifications: home (H), work (or university, W), restaurant (R), retail (RE), leisure indoor (LI), leisure outdoor (LO), transient (T), and unclassified (U). Participant 12's dataset was not recovered. Star (*) indicates that the participant is a part-time or full-time student. Blanks indicate that no data points were classified for the microenvironment.

Participant	City	H	W	R	RE	LI	LO	T	U
1*	Riverside	0.40	0.29	0.38	-	-	-	0.29	0.33
2	San Bernardino	**7.92**	0.00	0.00	0.00	**5.27**	0.02	0.11	0.11
3*	Riverside	0.35	-	0.24	**2.45**	0.07	-	0.67	0.61
4*	San Bernardino	-	0.00	-	-	-	-	-	0.87
5	Redlands	-	0.31	-	-	0.96	-	0.04	0.10
6*	San Bernardino	**1.18**	0.01	0.38	0.88	0.50	-	0.14	0.20
7	San Bernardino	**2.80**	-	-	0.03	0.09	0.15	0.31	0.27
8*	Riverside	**2.67**	0.23	0.94	**1.72**	0.36	0.22	**1.43**	**6.31**
9	Redlands	0.09	-	-	-	0.15	-	0.48	**1.05**
10	Yucaipa	0.31	-	**1.47**	-	-	**1.27**	0.85	**1.22**
11	San Bernardino	0.22	0.15	0.76	0.08	0.22	-	0.36	0.30
13*	Moreno Valley	**392.0**	**16.96**	**126.0**	**1.13**	**3.13**	-	**7.55**	**31.1**

Table 1.6. Median of personal-ambient ratios for each participant. A ratio greater than one (bolded) indicates a higher personal exposure compared to exposure derived from ambient $PM_{2.5}$ concentrations. The table header indicates the microenvironment classifications: home (H), work (or university, W), restaurant (R), retail (RE), leisure indoor (LI), leisure outdoor (LO), transient (T), and unclassified (U). Participant 12's dataset was not recovered. Star (*) indicates that the participant is a part-time or full-time student. Blanks indicate that no data points were classified for the microenvironment.

Participant	City	H	W	R	RE	LI	LO	T	U
1*	Riverside	0.42	0.30	0.25	-	-	-	0.31	0.30
2	San Bernardino	0.16	0.00	0.00	0.00	0.35	0.00	0.00	0.00
3*	Riverside	0.18	-	0.11	0.35	0.00	-	0.50	0.47
4*	San Bernardino	-	0.00	-	-	-	-	-	0.90
5	Redlands	-	0.00	-	-	**1.07**	-	0.00	0.00
6*	San Bernardino	0.93	0.00	0.23	0.65	0.31	-	0.00	0.09
7	San Bernardino	0.86	-	-	0.00	0.00	0.15	0.00	0.00
8*	Riverside	0.64	0.11	0.80	0.87	0.28	0.00	0.12	0.61
9	Redlands	0.00	-	-	-	0.15	-	0.45	**1.09**
10	Yucaipa	0.18	-	**1.47**	-	-	**1.18**	0.82	**1.15**
11	San Bernardino	0.00	0.00	0.57	0.00	0.13	-	0.00	0.00
13*	Moreno Valley	**5.24**	0.31	2.23	0.00	0.24	-	0.00	0.50
14*	Moreno Valley	0.21	-	-	0.00	-	0.00	0.00	0.00
15*	Riverside	0.13	-	0.17	0.00	0.00	-	0.00	0.00
16	Riverside	0.27	0.65	0.27	0.14	-	-	-	0.18
17	Moreno Valley	0.31	0.07	-	0.42	0.42	-	0.18	0.16
18*	Yucaipa	**2.84**	-	-	-	0.00	-	0.00	0.00

Figures

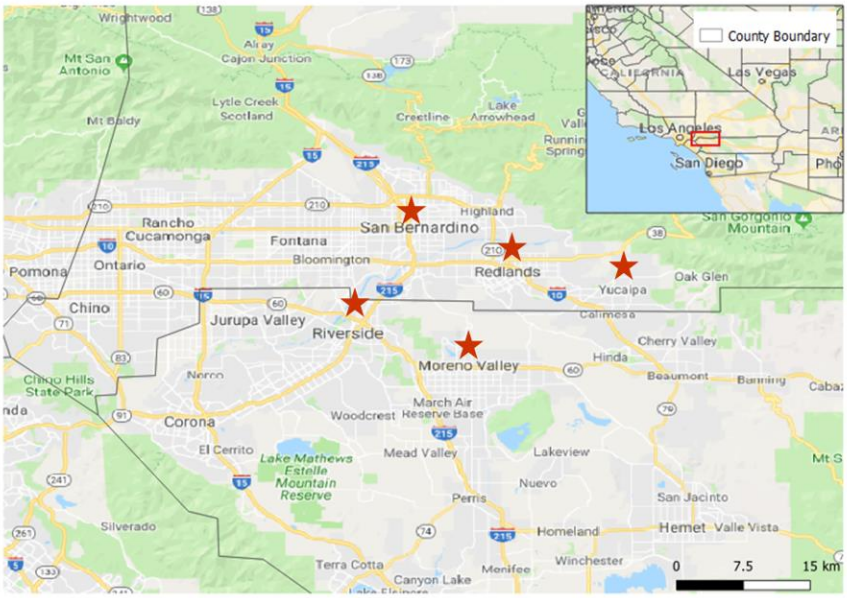


Figure 1.1. Map of the personal exposure study (Google Maps, 2019). Red stars delineate an area of approximately 200 square miles (520 square km).

Figure 1.2. (Left) Wearable particulate matter monitors from Applied Particle Technology (St. Louis, MO). Data was transmitted via Wi-Fi hotspots and was accessible online in real-time. (Right) PM sampling pack used in the personal exposure study. The monitors were clipped outside of the pack, and the Wi-Fi and GPS data loggers were housed inside of the pack.

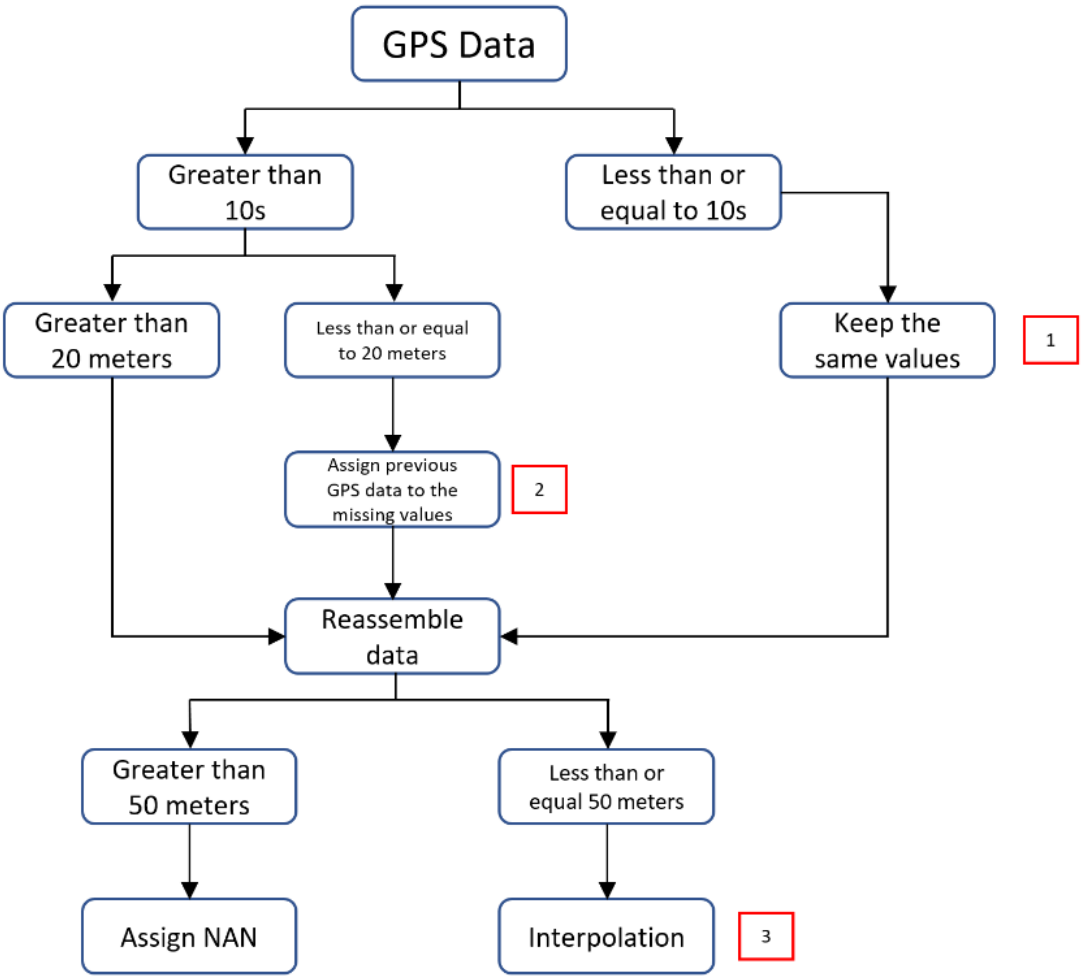


Figure 1.3. Details of the processing of GPS data obtained from the Global-Sat data loggers, which were set to sample every five seconds. GPS locations were classified as being measured more than or less than 10 seconds after the previously measured location. For locations measured more than 10 seconds after the previous location, spatial separation was also taken into account (greater than or less than 20 meters between the previous location). All locations separated by a distance of greater than 50 meters were assigned the value "NaN," and all other positions were linearly interpolated. Levels 1 to 3 indicate highest to lowest data quality, respectively.

Figure 1.4. Collocation of APT monitors at the Mira Loma Van Buren air monitoring site.

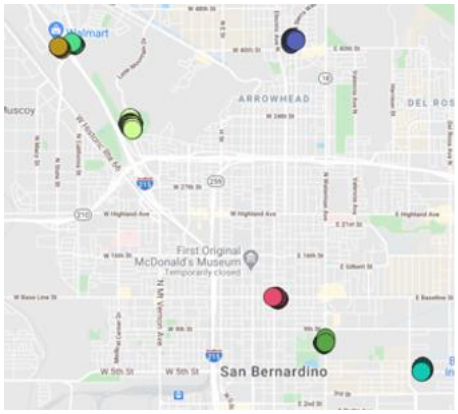


Figure 1.5. An example of spatially clustered personal measurements generated using the DBSCAN approach.

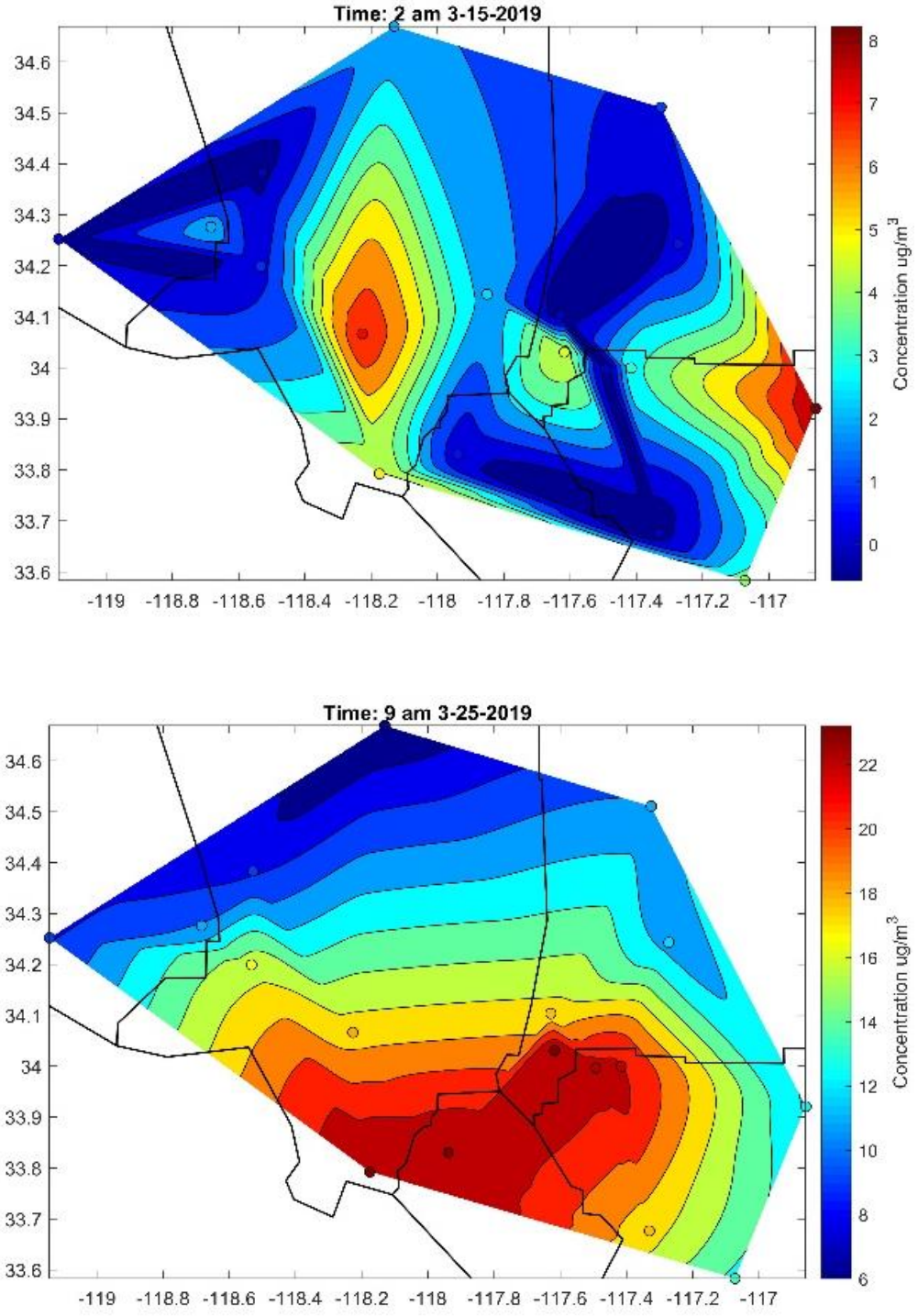


Figure 1.6. Contour spatial fields of ambient $PM_{2.5}$ overlaid with monitor values (circles) for a 2:00 AM (top) and 9:00 AM (bottom) hour during the five-week study period.

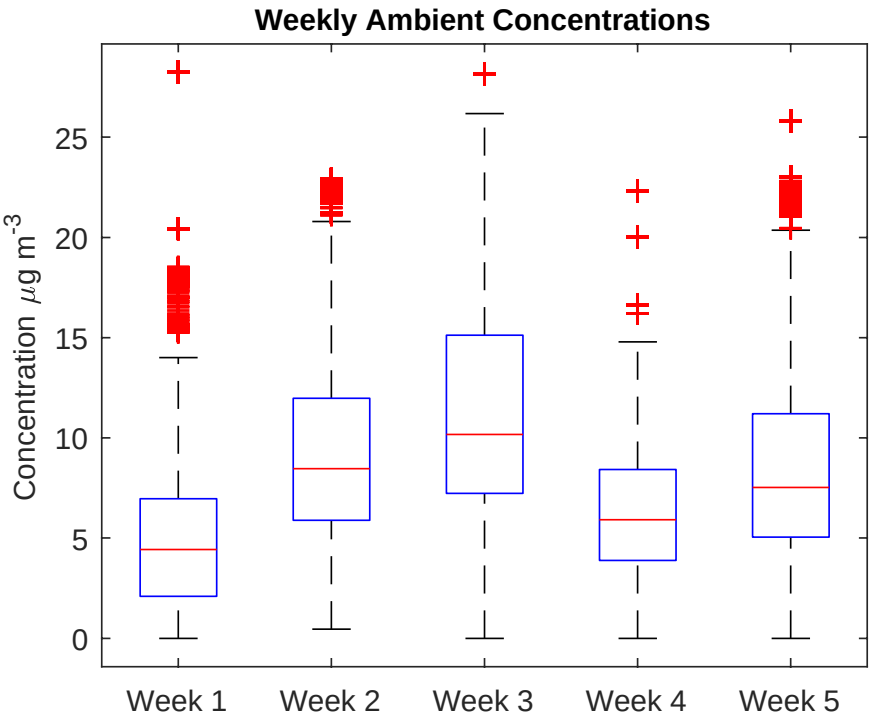


Figure 1.7. Distributions of ambient $PM_{2.5}$ concentrations (μg m^{-3}) corresponding to participant locations during each week of the study. Median concentrations were 4.4, 8.5, 10.2, 5.9, and 7.5, for weeks 1-5, respectively. All ambient data are retreived from regulatory monitoring stations.

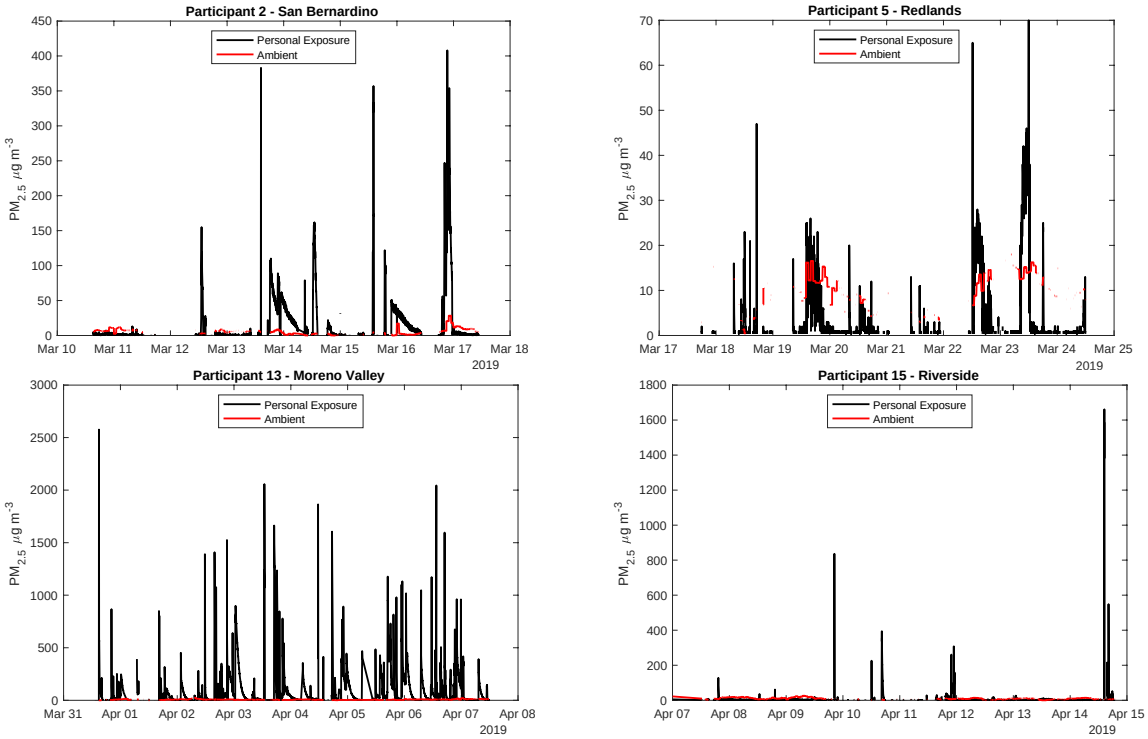


Figure 1.8. Sample time series of 5-second personal (black) and hourly ambient (red) monitoring data for four participants from San Bernardino (top-left), Redlands (top-right), Moreno Valley (bottom-left), and Riverside (bottom-right). Data are presented in log scale and maxium personal exposures are indicated in each plot.

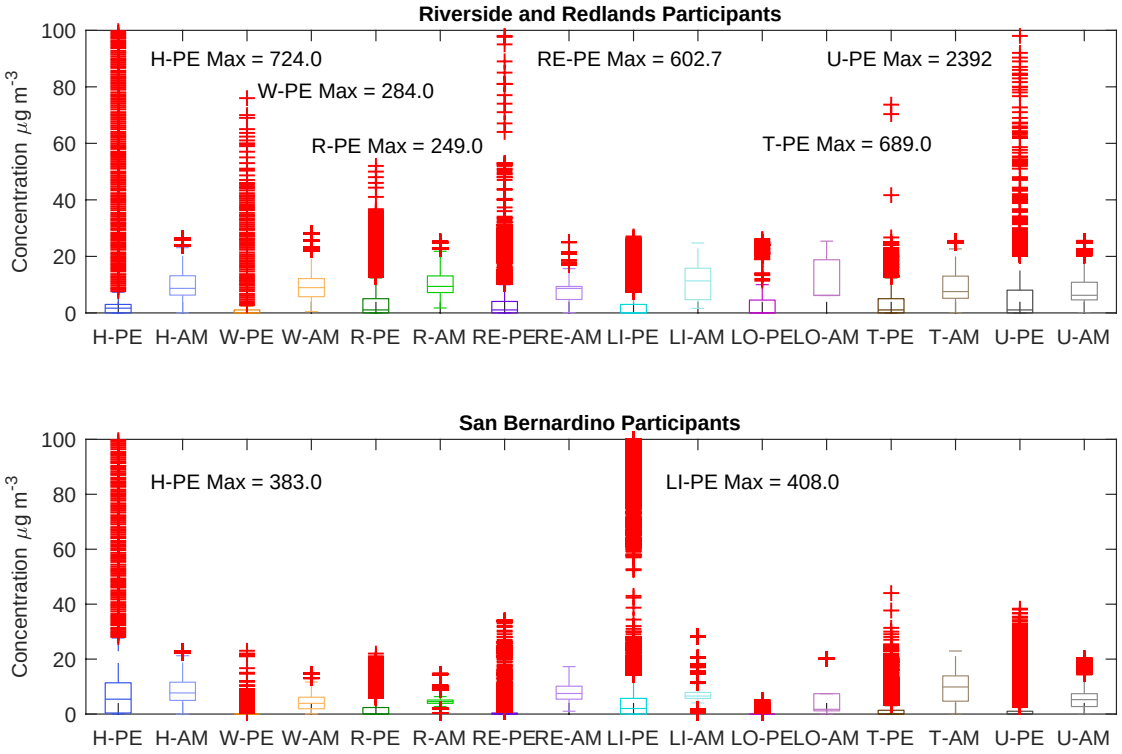


Figure 1.9. Distributions of personal and ambient $PM_{2.5}$ measurements for Redlands and Riverside (N = 5), and San Bernardino (N = 4) participants with data recovery greater than 50%. The labels indicate the microenvironment classifications: home (H), work (or university, W), restaurant (R), retail (RE), leisure indoor (LI), leisure outdoor (LO), transient (T), and unclassified (U). Personal exposure measurements are labeled "-PE," and ambient data are labeled as "-AM."

Part 2

Ambient and Indoor Influences on $PM_{2.5}$ in Rail-Impacted Homes in San Bernardino, CA

Abstract

Ambient air quality metrics are generally used to quantify the impact of air pollution on human health. However, people spend over 90% of their time in indoor environments, for which ambient air pollution may not always have the highest influence. Further, indoor and ambient air pollution analyses are rare for historically impacted communities in the U.S., and crowdsourced indoor air quality data is generally representative of higher income and less-impacted households. In this study, Indoor and ambient $PM_{2.5}$ levels were monitored for five households (ten total households for ambient monitoring) over six months using PurpleAir monitors in West San Bernardino, California, which is a state-designated community for disproportionate impacts from air pollution sources. The selected households are feet away from the Burlington Northern Santa Fe intermodal facility, which is estimated to emit air pollutants all day due to 24/7 operation. The influence of ambient $PM_{2.5}$ on indoor environments was studied using a mass balance approach. The analysis shows that household $PM_{2.5}$ levels had a higher-than-expected average infiltration factor of 0.70. Approximately 33% of the time, indoor $PM_{2.5}$ levels were greater than ambient $PM_{2.5}$ levels due to high frequent indoor emissions from cooking, cleaning, or dusting without sufficient filtration or air exchange rate. The indoor 98^{th} percentiles across the households far exceeded a healthy level at an average of 61 $\mu g/m^3$. A linear regression model confirms that the 98^{th} percentile can be reduced by increasing the air exchange rate, filtration, or reducing indoor emissions. It can be concluded that ambient and indoor $PM_{2.5}$ exposures for disproportionately impacted groups cannot be generalized in population-wide, crowdsourced applications due to lack

of availability of indoor low-cost monitoring in these communities. I recommend localized indoor monitoring efforts that are tailored to a community's needs and historical source burdens.

Introduction

Fine particulate matter (PM) is the term to describe liquid or solid particles with an aerodynamic diameter less than or equal to 2.5 microns ($PM_{2.5}$). Studies have shown that exposure to high levels of $PM_{2.5}$ can adversely affect human health, causing asthma, respiratory disease, and cardiovascular disease (Brown et al., 1950; Deng et al., 2019; Maté et al., 2010; Wang et al., 2019). In the United States, primary $PM_{2.5}$ is directly emitted from a source into the atmosphere, and sources include construction sites, smokestacks, or wildfires. $PM_{2.5}$ is also generated through complex chemical reactions in the atmosphere, known as secondary PM, which is highly correlated with urban $PM_{2.5}$ (Zawacki et al., 2018; Zhang et al., 2015). High concentrations of $PM_{2.5}$ are found in urban areas with a high volume of anthropogenic activities (Fruin et al., 2014; Gildemeister et al., 2007; Hasheminassab et al., 2014). Spatial distributions of $PM_{2.5}$ in the U.S. exhibit significant racial-ethnic disparity (Ivey, 2020; Wang et al., 2022). Specifically, highly polluted areas are often in low-income and non-white neighborhoods that are surrounded by industrial factories, shipping facilities, warehouses, and railyards (Allen, 2010a; Bluffstone and Ouderkirk, 2007; deSouza et al., 2022; Houston et al., 2004).

Additionally, people spend over 90% of the time indoors(Do et al., 2021; Long et al., 2001) and are subsequently exposed to indoor air pollutants that are generated from multiple sources. Indoor activities, such as vacuum cleaning, cooking, dusting, use of consumer products, and smoking are the primary sources of indoor $PM_{2.5}$ (Mattila et al., 2020). These activities can raise indoor $PM_{2.5}$ levels to peak concentrations in a very short period of time, approximately 10 to 30

minutes (Stephens and Siegel, 2012). An effective range hood can remove a significant amount of $PM_{2.5}$ generated during cooking activities. During high $PM_{2.5}$ episodes, air ventilation also effectively reduces indoor $PM_{2.5}$ levels by diluting with fresh outdoor air (Kang et al., 2019; Xiang et al., 2021). Further, baseline indoor $PM_{2.5}$ levels are highly influenced by the penetration of ambient $PM_{2.5}$ into the indoor environment. Although indoor air quality can be improved with proper air exchange and filtration systems, numerous studies have shown a strong relationship between indoor and ambient $PM_{2.5}$ levels (Freijer and Bloemen, 2000; Lee et al., 1997; Leung, 2015; Mousavi and Wu, 2021; Poupard et al., 2005). In particular, indoor $PM_{2.5}$ concentrations highly correlated with ambient $PM_{2.5}$ when wildfires occur (Liang et al., 2021). Closing the windows and minimizing the air exchange rate can decrease the penetration of ambient particles during such an event.

This study considers ambient and indoor $PM_{2.5}$ for a disproportionately impacted community of inland Southern California, which is located near the northern and southern borders of Riverside and San Bernardino Counties, respectively. For reference, this region is historically known for its agricultural economy and more recently for freight shipping activities and a growth of warehouses, creating a significant shift in the region's economy (Allen, 2010b; deSouza et al., 2022). The nationwide shift towards more online shopping in the United States has resulted in further expansion of freight shipping activities in the region. Roughly 45% of products imported from Asia are shipped through inland Southern California each year(Patterson, 2016) and distributed across the United States via heavy-duty diesel trucks and railway systems. The Burlington Northern Santa Fe (BNSF) intermodal facility, which is directly adjacent to residential areas within the San Bernardino community (within 200 feet of the fence line), has long been

determined as a major air pollution source and health hazard for neighboring communities (Spencer-Hwang et al., 2019, 2016, 2015, 2014). The facility's emissions are generated from diesel trucks entering and leaving the facility, equipment to load and unload containers, and locomotives (South Coast Air Quality Management District, 2019).

In this study, indoor and ambient $PM_{2.5}$ are investigated for a low-income community near the BNSF facility. Most homes were missing features that ensure good indoor air quality, such as central air conditioning. Some homes required cooling through window air conditioning units or through open windows at night to cool the indoor environment during summer, which increases ambient $PM_{2.5}$ penetration. Using a mass balance approach, the penetration, indoor emission rate, and air exchange rate, and filtration factors were estimated and compared the findings with previous work that characterized indoor air quality in California homes using crowdsourced data.

Study Area

The study was conducted in the West San Bernardino community, located in the southern region of San Bernardino County, California (inland southern California) and classified as hot-summer Mediterranean climate, with mild winters and hot, dry summers. The West San Bernardino community is bounded by a highway network of U.S. Interstates 10 to the south, 210 to the north, and 215 on the east, which are always in heavy use due to the rapid expansion of freight infrastructure. Fifteen PurpleAir (PAII) monitors were deployed in the community in ten households to assess trends in $PM_{2.5}$ over seven months (July 2023 – January 2023). Specifically, five homes were selected for the installation of indoor and ambient monitors, while the other five homes had only ambient $PM_{2.5}$ monitoring. The sample size was impacted by funding limitations, but nonetheless this effort has been of great benefit for all community members involved. The

deployment area was divided into three zones based on the distance to the BNSF facility, in which zone 1 has five PurpleAir located within 450 meters from the railyard. Zone 2 has six PurpleAir located within 1,000 meters from the railyard, and zone 3 has four PurpleAir that is further than 1000 meters from the railyard (Figure 1.10). The average household income for the participants is $50,000 (2022), which is 32% lower compared to the median household income in San Bernardino County ($72,300 in 2021).

Data and Methods

Measurements and Data Processing

Ambient PurpleAir monitors were installed in the back yard or front yard, and indoor monitors were installed in the living room (i.e., main room). The sensors were powered continuously by 120V outlets. Sampling took place over seven months, from July 2022 to January 2023. The monitors provided measurements every 120 seconds for temperature (°F), relative humidity (%), and $PM_{2.5}$ concentration ($\mu g/m^3$). 10-minute averages were used to compute indoor emission and decay rates. The data were averaged hourly to remove noise before computing statistical summaries. Hourly averages were used to evaluate data against the National Ambient Air Quality Standards (NAAQS) for 24-hour $PM_{2.5}$. A linear correction factor was applied to the raw PurpleAir $PM_{2.5}$ measurements based on recommendations by Barkjohn et al. (Eq. 1), where $PM_{2.5}$ is the corrected concentration, PA is the average raw $PM_{2.5}$ concentration from PurpleAir channels a and b, and RH is relative humidity (Barkjohn et al., 2021).

$$PM_{2.5} = 0.524PA - 0.0862RH + 5.75 \quad (1)$$

Indoor $PM_{2.5}$ Modeling

Simultaneously indoor and ambient $PM_{2.5}$ sampling enabled the derivation of a simple mass balance to estimate the loss rate constant, indoor emission rate constant, and penetration for the homes with paired monitors. The loss rate constant is the combination of the air exchange and filtration rate constant, which are responsible for the decay of indoor $PM_{2.5}$ concentrations. The indoor emission rate constant is the magnitude of indoor emissions, and the penetration rate constant represents the effectiveness of $PM_{2.5}$ transfer from the outside to the indoor environment. The mass balance applied in this study is expressed in Eq. 2:

$$\frac{dC_{in}}{dt} = aPC_{out} - (a+k)C_{in} + \left(\frac{E_{in}}{V}\right) \quad (2)$$

where C_{in} is indoor $PM_{2.5}$, C_{out} is ambient $PM_{2.5}$, a and k are the air exchange rate and filtration constant, P is the penetration factor, V is the volume of the house, and E_{in} is the indoor emissions.

Emission event: To compute indoor emission rates, the assumption was the penetration was negligible. When an emission event occurs, the rate of change in C_{in} is steep, and the penetration amount is minimal compared to indoor emissions. The solution to the ODE in Eq 2. is shown in Eq. 3, where E/V is the indoor emission rate per m^3 ($\mu g * hr^{-1} * m^{-3}$):

$$\frac{E}{V} = \frac{C_{in}(t) - C_{in}\left(t = t_{peak}\right)e^{\alpha \Delta t}}{1 - e^{\alpha \Delta t}}\alpha \quad (3)$$

For each home, multiple values were computed for α, which is $(a+k)$, and E/V based on a set of criteria (See SI).

Decay event: After an indoor emission event, zero $PM_{2.5}$ generation was assumed at the peak of C_{in} (the intersection of the green and red lines, as shown in the top panel of Figure 1.11). The decay of C_{in} only depends on the loss due to air exchange and filtration rates. At the time of peak C_{in}, the indoor $PM_{2.5}$ concentration is much higher than ambient $PM_{2.5}$. Eq. 4 can be simplified to

$$\frac{dC_{in}}{dt} = -(a + k)C_{in} \tag{4}$$

implying that right after the peak of an emission event, the change in indoor $PM_{2.5}$ depends only on the air exchange and filtration rate constants. The solution to the ODE in Eq. 4 during periods dominated by decay is Eq. 5. $C_{in}(t = peak)$ occurs when indoor $PM_{2.5}$ is maximum at the intersection of the red and green lines, as shown in Figure 1.11. Δt is the difference in time t between $C_{in}(t)$ and $C_{in}(t = peak)$.

$$\alpha = -\frac{\ln\left(\frac{C_{in}(t)}{C_{in}(t = peak)}\right)}{\Delta t} \tag{5}$$

Baseline indoor model: The indoor $PM_{2.5}$ was constructed to validate the estimated penetration and air exchange constant based on Eq. 6, where C_{model} is the modeled indoor $PM_{2.5}$ concentrations, α is the combination of air exchange rate and filtration constant, and aP is the penetration factor which is equal to C_{in}/C_{out}. Eq. 6 is valid if there are no indoor emissions and when the ambient $PM_{2.5}$ is greater than indoor $PM_{2.5}$ in the absence of indoor emission events. All ODE solution derivations can be found in the Supplemental Information.

$$C_{model}(t) = C_{model}(t-1)\mathrm{e}^{\alpha\Delta\mathrm{t}} + \frac{aPC_{out}(t)}{\alpha} \quad (6)$$

Overall, the peaks of indoor $PM_{2.5}$ were ten times greater than the indoor average, and the slopes were steep. Typically, indoor emissions were generated in 10 to 20 minutes, and the decay lasted about 10 to 50 minutes. The red lines from the bottom panel in Figure 1.11 were used to calculate average indoor emissions and decay constants.

Results and Discussion

Analysis of indoor and ambient $PM_{2.5}$: The analysis of indoor $PM_{2.5}$ was presented for the five homes where indoor and ambient pairs of PurpleAir were installed. Based on an evaluation indoor and ambient temperature (and verified by household data collected at the start of community engagement), house 3 did not use an air conditioning unit as its indoor temperature was approximately greater than the ambient temperature during summertime (Figure 1.12). The histograms in Figure 1.13 show the ratio of indoor and ambient $PM_{2.5}$ (I/O ratio). The peaks of the histogram distribution are centered around the value of one. For homes 1, 3, and 5, the mode for I/O ratio (most frequent occurrence) occurs when the indoor $PM_{2.5}$ is nearly the same as ambient $PM_{2.5}$, which contradicts previous studies, for which the distribution modes were approximately 0.62 using crowdsourced information (Liang et al., 2021).

The I/O ratios from crowdsourced data generally reflect a higher socioeconomic status population with high accessibility to indoor air quality monitoring. Further, population-based studies will likely not reflect the lived experiences of disproportionately impacted communities that have more limited access to indoor monitoring equipment. Historically, racial-ethnic minority

groups are the most sensitive and highly affected by the poor ambient air quality (Ivey, 2020; Patterson and Harley, 2019; Wang et al., 2022).

The findings also suggest that elevated ambient $PM_{2.5}$ levels directly influence indoor air quality in West San Bernardino homes (Table 1.7), which is further evidenced by the seasonal statistics (Table 1.8, Table 1.9, and Table 1.10). The consistent values across all PurpleAir monitors for the corrected 25th, 50th, and 98th percentile ambient $PM_{2.5}$ reflect good performance for ambient measurements in the West San Bernardino area. For the 50th percentile across all months, indoor $PM_{2.5}$ was less than ambient for all homes except for house 3 (no air conditioning or filtration), where indoor $PM_{2.5}$ levels were higher than ambient levels for all quartiles. Indoor mean and 98th percentile was significantly higher than corresponding ambient levels for all five houses, reflecting the influence of indoor emissions.

Seasonal variations between summer (Jul – Sep 2022) and fall (Oct 2022– Jan 2023) are provided in the Supplemental Information (Table 1.8 and Table 1.9). Summer temperatures were high, with an average of 82°F and exceeding 100°F around 5% of the time. During high-temperature periods, four out of five houses used air conditioning to regulate indoor temperatures resulting in their indoor $PM_{2.5}$ being less than ambient $PM_{2.5}$ levels (Table 1.8). This indicated that filtration systems from air conditioning units effectively reduced concentrations. The average temperature was 60 °F in the fall/winter, allowing open-window ventilation to regulate indoor environments and increasing air exchange rate and penetration. Due to increased penetration, indoor $PM_{2.5}$ baseline levels rose, leading to indoor levels exceeding ambient $PM_{2.5}$ across all quartiles (Table 1.9).

Estimated indoor emissions: Four out of five homes had an indoor 98th percentile that exceeded the 24-hour $PM_{2.5}$ NAAQS level (35 $\mu g/m^3$). High 98th percentiles resulted from high indoor emissions and poor ventilation, which can be explained by the average decay constants (Homes 1 and 5 in Table 1.11). Houses with low decay constant suffered from prolonged periods of high $PM_{2.5}$ episodes after indoor emission events (Homes 2, 3, and 4 in Table 1.11). An indoor emission event is defined as when indoor $PM_{2.5}$ levels are significantly higher than ambient $PM_{2.5}$ levels. The frequencies of indoor emissions were also estimated for the homes, considering the instances where indoor $PM_{2.5}$ concentrations peaked at levels five times higher than the average indoor $PM_{2.5}$ concentrations. Indoor emission rates per m^3 were estimated to be a minimum of 619 $\mu g * h^{-1} * m^{-3}$ and a maximum of 1190 $\mu g * h^{-1} * m^{-3}$ for houses 2 and 1, respectively.

Estimated decay and infiltration constants: The average decay constants, average indoor emissions per m^3, and infiltration factors for all five homes were calculated based on the mass balance (Eq. 2) and the set assumptions discussed in the Data and Methods section. Indoor activities, air exchange rates, and filtration rates were highly variable, resulting in different infiltrations values across the study period. The average infiltration values for each house also represent family habits during the community engagement period. Infiltration value ranges from zero to one, where zero represents no penetration, and one indicates the indoor $PM_{2.5}$ and ambient $PM_{2.5}$ levels. In our study, the lowest infiltration value is 0.57 the highest is 0.84 for houses 1 and 4, respectively, implying the vulnerability of indoor environments to the changes in ambient conditions (Table 1.11). The infiltration values of this study are significantly higher than those in the previous studies that rely on crowdsourced data or a test house. Stephens et al. used a mass balance, and the calculated infiltration factor was 0.34 for a test house (Utest House) (Stephens

and Siegel, 2012). Liang et al. used a similar approach and utilized the PurpleAir sensor network in California that monitored more than 1400 buildings to assess the impact of wildfire smoke on indoor air quality, and the derived average infiltration factor was 0.45 (Liang et al., 2021). The average infiltration factor in this study across the five homes is 0.70, which is relatively higher compared to previous studies, indicating a more significant impact of ambient air quality on the indoor environments of this rail-impacted community.

Baseline indoor $PM_{2.5}$ model: To evaluate the calculated infiltration and decay constant, indoor $PM_{2.5}$ concentrations were constructed using the mass balance. Here, emissions in the baseline model were not considered. Therefore, the model is only a function of decay constant, penetration, and ambient $PM_{2.5,}$ as described in Eq. 6. The model gave good predictions and captured the trend of occurrences (Figure 1.14). Although the model successfully reconstructed the distribution of indoor $PM_{2.5}$ for homes 3, 4, and 5, it did not capture the peak for house 2 and high concentrations in homes 1 and 4. The errors were caused by indoor minor emission events, which were not accounted for as long as the indoor $PM_{2.5}$ was still less than ambient $PM_{2.5}$. Minor emissions are difficult to trace with the time series without additional activity information from home occupants. Uncertainties in participants' habits, such as opening the windows, turning on the fume hood, and using air conditioning, largely contributed to the model's errors.

98th percentile regression model: Intuitively, indoor $PM_{2.5}$ levels are managed by the decay constant, $(\alpha = a + k)$ and the frequency, f. Linear regression with the two dependent variables was performed to predict the indoor 98th percentiles, for which $Indoor\ 98^{th}\ \%ile = c_1\alpha + c_2 f + c_3$, where $c_1$and c_2 are the coefficients for decay constant and frequency, respectively, and c_3 is the bias. The values for c_1, c_2, and c_3 are listed in Eq. 7, and the R^2 for the regression model is

0.84. The scatter plot for the prediction and actual indoor 98th percentile is provided in the Supplemental Information (Figure 1.15). The regression model shows that the indoor 98th percentile has a negative correlation with the decay constant and a positive correlation with indoor emission frequency.

$$Indoor\ 98^{th}\ \%ile = -11.1\alpha + 0.12f + 49 \quad (7)$$

where α is the decay constant ($\alpha = a + k$), and f is the frequency accounting for the $PM_{2.5}$ peaks, which are identified when indoor $PM_{2.5}$ is greater than five times the indoor average. Interestingly, the computed average indoor emission rates (E/V) had relatively little impact on the modeled indoor 98th percentile, for which house 1 with the highest average emission rate still had the lowest indoor 98th percentile $PM_{2.5}$.

Recommendation and model uncertainties: Our analyses show a strong effect of ambient $PM_{2.5}$ on the indoor levels for five community homes that are near the BNSF facility with an average infiltration of 0.7, a value higher than that previously published using crowdsourced data. The 98th percentile regression model implies 98th percentile concentrations are linearly correlated with the air exchange rate, filtration, and indoor emission frequency. Indoor $PM_{2.5}$ concentrations can be regulated by increasing ventilation during indoor emission events or minimizing the air exchange rate when outdoor $PM_{2.5}$ concentrations are high (during daytime peaks in fall/winter). It is strongly recommended that impacted homes near the BNSF facility have adequate air filter to minimize penetration and indoor levels. The suggestion is that PurpleAir sensors be permanently installed in impacted homes near the BNSF facility (or any large industrial source) to continuously monitor residential indoor and ambient air quality and provide real-time feedback for mitigating

indoor pollution. For instance, occupants should increase filtration and ventilation during indoor emission events when ambient $PM_{2.5}$ levels are low.

The uncertainties of estimated constants arose from the assumption that there were no emissions at the peaks (inflection points) and no penetration when indoor $PM_{2.5}$ levels were high. Infiltration uncertainty is derived from omitting minor indoor emissions from consideration, causing a slight overestimation of infiltration factors.

Supplemental Information

The Supplemental Information provides details of the mass balance derivation; statistical tables for the summer and fall periods; scatter plot for the 98th percentile model estimates and actual values; time series plot for ambient, actual indoor, and baseline model $PM_{2.5}$ concentrations; histogram plots for indoor and ambient $PM_{2.5}$ concentrations; hourly averaged $PM_{2.5}$ concentration time series plots for outdoor and indoor throughout the study; and hourly averaged ambient and indoor temperature time series throughout the study.

Acknowledgments

This paper was prepared as a result of work sponsored and paid for, in whole or in part, by the California Air Resources Board (CARB). The opinions, findings, conclusions, and recommendations are those of the authors and do not necessarily represent the views of CARB. First and foremost, we thank the community members of West San Bernardino for the collective execution of this work. We thank Janet Bernabe and the Center for Community Action and Environmental Justice (CCAEJ) for providing and coordinating the PurpleAir installation. We also

thank Ms. Jean Kayano of the People's Collective for Environmental Justice for her initial conceptualization, planning, and fundraising for the project.

References

Allen, N., 2010a. Exploring the Inland Empire: Life, Work, and Injustice in Southern California's Retail Fortress. New Labor Forum 19, 37–43. https://doi.org/10.4179/NLF.192.0000006

Allen, N., 2010b. Exploring the Inland Empire: Life, Work, and Injustice in Southern California's Retail Fortress. New Labor Forum 19, 37–43. https://doi.org/10.4179/NLF.192.0000006

Barkjohn, K.K., Gantt, B., Clements, A.L., 2021. Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor. Atmospheric Measurement Techniques 14, 4617–4637. https://doi.org/10.5194/amt-14-4617-2021

Bluffstone, R.A., Ouderkirk, B., 2007. Wearhouses, Trucks, and PM2.5: Human Health and logistics industry Growth in the Eastern Inland Empire. Contemporary Economic Policy 25, 79–91. https://doi.org/10.1111/j.1465-7287.2006.00017.x

Brown, J.H., Cook, K.M., Ney, F.G., Hatch, T., 1950. Influence of Particle Size upon the Retention of Particulate Matter in the Human Lung. American Journal of Public Health and the Nations Health. https://doi.org/10.2105/ajph.40.4.450

Deng, Q., Deng, L., Miao, Y., Guo, X., Li, Y., 2019. Particle deposition in the human lung: Health implications of particulate matter from different sources. Environmental Research. https://doi.org/10.1016/j.envres.2018.11.014

deSouza, P.N., Ballare, S., Niemeier, D.A., 2022. The environmental and traffic impacts of warehouses in southern California. Journal of Transport Geography 104, 103440. https://doi.org/10.1016/j.jtrangeo.2022.103440

Do, K., Yu, H., Velasquez, J., Grell-Brisk, M., Smith, H., Ivey, C.E., 2021. A data-driven approach for characterizing community scale air pollution exposure disparities in inland Southern California. Journal of Aerosol Science 152, 105704. https://doi.org/10.1016/j.jaerosci.2020.105704

Freijer, J.I., Bloemen, H.J.Th., 2000. Modeling Relationships between Indoor and Outdoor Air Quality. Journal of the Air & Waste Management Association 50, 292–300. https://doi.org/10.1080/10473289.2000.10464007

Fruin, S., Urman, R., Lurmann, F., McConnell, R., Gauderman, J., Rappaport, E., Franklin, M., Gilliland, F.D., Shafer, M., Gorski, P., Avol, E., 2014. Spatial variation in particulate matter components over a large urban area. Atmospheric Environment 83, 211–219. https://doi.org/10.1016/j.atmosenv.2013.10.063

Gildemeister, A.E., Hopke, P.K., Kim, E., 2007. Sources of fine urban particulate matter in Detroit, MI. Chemosphere 69, 1064–1074. https://doi.org/10.1016/j.chemosphere.2007.04.027

Hasheminassab, S., Daher, N., Saffari, A., Wang, D., Ostro, B.D., Sioutas, C., 2014. Spatial and temporal variability of sources of ambient fine particulate matter (PM2.5) in California. Atmos. Chem. Phys. 14, 12085–12097. https://doi.org/10.5194/acp-14-12085-2014

Houston, D., Wu, J., Ong, P., Winer, A., 2004. Structural Disparities of Urban Traffic in Southern California: Implications for Vehicle-Related Air Pollution Exposure in Minority and High-Poverty Neighborhoods. Journal of Urban Affairs 26, 565–592. https://doi.org/10.1111/j.0735-2166.2004.00215.x

Ivey, C., 2020. Land use predicts pandemic disparities. Nature 588, 220–220. https://doi.org/10.1038/d41586-020-03480-1

Kang, K., Kim, H., Kim, D.D., Lee, Y.G., Kim, T., 2019. Characteristics of cooking-generated PM10 and PM2.5 in residential buildings with different cooking and ventilation types. Science of The Total Environment 668, 56–66. https://doi.org/10.1016/j.scitotenv.2019.02.316

Lee, H.S., Kang, B.-W., Cheong, J.-P., Lee, S.-K., 1997. Relationships between indoor and outdoor air quality during the summer season in Korea. Atmospheric Environment 31, 1689–1693. https://doi.org/10.1016/S1352-2310(96)00275-0

Leung, D.Y.C., 2015. Outdoor-indoor air pollution in urban environment: challenges and opportunity. Front. Environ. Sci. 2. https://doi.org/10.3389/fenvs.2014.00069

Liang, Y., Sengupta, D., Campmier, M.J., Lunderberg, D.M., Apte, J.S., Goldstein, A.H., 2021. Wildfire smoke impacts on indoor air quality assessed using crowdsourced data in California. Proceedings of the National Academy of Sciences 118, e2106478118. https://doi.org/10.1073/pnas.2106478118

Long, C.M., Suh, H.H., Catalano, P.J., Koutrakis, P., 2001. Using Time- and Size-Resolved Particulate Data To Quantify Indoor Penetration and Deposition Behavior. Environmental Science & Technology 35, 2089–2099. https://doi.org/10.1021/es001477d

Maté, T., Guaita, R., Pichiule, M., Linares, C., Díaz, J., 2010. Short-term effect of fine particulate matter (PM2.5) on daily mortality due to diseases of the circulatory system in Madrid (Spain). Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2010.07.083

Mattila, J.M., Arata, C., Wang, C., Katz, E.F., Abeleira, A., Zhou, Y., Zhou, S., Goldstein, A.H., Abbatt, J.P.D., DeCarlo, P.F., Farmer, D.K., 2020. Dark Chemistry during Bleach Cleaning Enhances Oxidation of Organics and Secondary Organic Aerosol Production Indoors. Environ. Sci. Technol. Lett. 7, 795–801. https://doi.org/10.1021/acs.estlett.0c00573

Mousavi, A., Wu, J., 2021. Indoor-Generated PM2.5 During COVID-19 Shutdowns Across California: Application of the PurpleAir Indoor–Outdoor Low-Cost Sensor Network. Environ. Sci. Technol. 55, 5648–5656. https://doi.org/10.1021/acs.est.0c06937

Patterson, R.F., Harley, R.A., 2019. Effects of Freeway Rerouting and Boulevard Replacement on Air Pollution Exposure and Neighborhood Attributes. IJERPH 16, 4072. https://doi.org/10.3390/ijerph16214072

Patterson, T.C., 2016. From Acorns to Warehouses, 0 ed. Routledge. https://doi.org/10.4324/9781315428215

Poupard, O., Blondeau, P., Iordache, V., Allard, F., 2005. Statistical analysis of parameters influencing the relationship between outdoor and indoor air quality in schools. Atmospheric Environment 39, 2071–2080. https://doi.org/10.1016/j.atmosenv.2004.12.016

South Coast Air Quality Management District, 2019. Determine That Community Emissions Reduction Plan for Wilmington, Carson, West Long Beach Community Is Exempt from CEQA and Adopt Community Emissions Reduction Plan Per Assembly Bill 617 [WWW Document]. AQMD Governing Board. URL aqmd.gov/docs/default-source/Agendas/Governing-Board/2019/2019-sep6-025c.pdf

Spencer-Hwang, R., Montgomery, S., Dougherty, M., Valladares, J., Rangel, S., Gleason, P., Soret, S., 2014. Experiences of a Rail Yard Community: Life Is Hard. (cover story). Journal of Environmental Health.

Spencer-Hwang, R., Pasco-Rubio, M., Soret, S., Ghamsary, M., Sinclair, R., Alhusseini, N., Montgomery, S., 2019. Association of major California freight railyards with asthma-related pediatric emergency department hospital visits. Preventive Medicine Reports 13, 73–79. https://doi.org/10.1016/j.pmedr.2018.11.001

Spencer-Hwang, R., Soret, S., Knutsen, S., Shavlik, D., Ghamsary, M., Beeson, W.L., Kim, W., Montgomery, S., 2015. Respiratory Health Risks for Children Living Near a Major Railyard. Journal of Community Health 40, 1015–1023. https://doi.org/10.1007/s10900-015-0026-0

Spencer-Hwang, R., Soret, S., Valladares, J., Torres, X., Pasco-Rubio, M., Dougherty, M., Kim, W., Montgomery, S., 2016. Strategic partnerships for change in an environmental justice community: The ENRRICH study. Progress in Community Health Partnerships: Research, Education, and Action. https://doi.org/10.1353/cpr.2016.0062

Stephens, B., Siegel, J.A., 2012. Penetration of ambient submicron particles into single-family residences and associations with building characteristics: **Particle penetration and building characteristics**. Indoor Air 22, 501–513. https://doi.org/10.1111/j.1600-0668.2012.00779.x

Wang, C., Feng, L., Chen, K., 2019. The impact of ambient particulate matter on hospital outpatient visits for respiratory and circulatory system disease in an urban Chinese population. Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2019.02.256

Wang, Y., Apte, J.S., Hill, J.D., Ivey, C.E., Patterson, R.F., Robinson, A.L., Tessum, C.W., Marshall, J.D., 2022. Location-specific strategies for eliminating US national racial-ethnic PM2.5 exposure inequality. Proceedings of the National Academy of Sciences 119, e2205548119. https://doi.org/10.1073/pnas.2205548119

Xiang, J., Hao, J., Austin, E., Shirai, J., Seto, E., 2021. Residential cooking-related PM2.5: Spatial-temporal variations under various intervention scenarios. Building and Environment 201, 108002. https://doi.org/10.1016/j.buildenv.2021.108002

Zawacki, M., Baker, K.R., Phillips, S., Davidson, K., Wolfe, P., 2018. Mobile source contributions to ambient ozone and particulate matter in 2025. Atmospheric Environment. https://doi.org/10.1016/j.atmosenv.2018.04.057

Zhang, R., Wang, G., Guo, S., Zamora, M.L., Ying, Q., Lin, Y., Wang, W., Hu, M., Wang, Y., 2015. Formation of Urban Fine Particulate Matter. Chem. Rev. 115, 3803–3855. https://doi.org/10.1021/acs.chemrev.5b00067

Tables

Table 1.7. Statistics based on hourly averaged indoor (In) and ambient (Out) $PM_{2.5}$ concentrations (in $\mu g/m^3$) for five homes. The sampling duration is seven months (July 2022 to January 2023) spanning the summer and winter periods. The table includes the 25^{th}, 50^{th}, 75^{th}, and 98^{th} percentiles, mean, and standard deviation (STD).

	25%ile		50%ile		75%ile		98%ile		Mean		STD	
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
House 1	2.4	5.9	5.1	8.5	9.2	11.7	26.2	22.6	7.4	9.5	10.8	5.1
House 2	5.3	6.2	7.1	9.2	11.3	13.7	49.3	25.3	11.0	10.5	13.5	6.0
House 3	7.7	6.5	11.1	9.1	20.7	13.0	100	24.0	19.2	10.2	26.7	5.3
House 4	5.6	6.7	7.8	9.3	14.0	13.1	93.7	25.9	14.9	10.8	23.1	7.0
House 5	6.8	6.8	9.3	9.7	14.0	13.5	34.9	25.4	11.9	10.8	10.0	5.7

Table 1.8. Statistical summary of indoor and outdoor sensors for five houses. Sampling duration is three months from Jul 2022 to Sep 2023 spanning over the summer period. Based on 10 minute average.

Summer	25%ile		50%ile		75%ile		98%ile		Mean		STD	
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
House 1	6.2	7.1	8.1	9.1	10.7	11.2	23.9	18.9	9.3	9.6	6.0	4.2
House 2	5.4	7.1	6.9	9.1	10.2	11.2	34.5	18.0	9.9	9.4	11.7	3.3
House 3	7.3	7.3	8.9	9.1	11.7	11.3	65.9	19.1	13.1	9.7	17.0	3.8
House 4	5.0	7.4	6.7	9.2	10.5	11.2	62.4	18.5	11.8	9.7	20.7	4.1
House 5	6.4	7.6	8.1	9.6	10.1	11.8	25.0	19.3	9.3	10.1	6.8	3.8

Table 1.9. Statistical summary of indoor and outdoor sensors for five houses. Sampling duration is four months from Oct 2022 to Jan 2023 spanning over the winter period. Based on 10 minute average.

Fall	25%ile		50%ile		75%ile		98%ile		Mean		STD	
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
House 1	2.0	5.2	2.6	7.3	4.2	12.6	31.7	24.7	6.1	9.4	16.6	5.9
House 2	5.1	5.7	7.1	9.2	11.5	15.6	61.5	26.3	11.8	11.0	16.8	7.0
House 3	8.1	5.7	13.5	8.9	24.6	14.9	120.3	25.6	23.0	10.6	33.2	6.3
House 4	5.7	5.8	8.3	9.3	15.3	15.8	113.3	27.8	17.0	11.6	29.3	10.7
House 5	6.8	5.5	11.1	9.4	16.9	15.8	43.2	27.3	13.8	11.2	14.0	7.4

Table 1.10. Statistical summary of indoor and outdoor sensors for five houses using hourly average $PM_{2.5}$ concentrations. Sampling duration is six months from July 2022 to January 2023 spanning over the summer and winter periods. Based on 10 minute average.

	25%ile		50%ile		75%ile		98%ile		Mean		STD	
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
House 1	2.4	5.8	4.9	8.4	9.1	11.7	26.7	23.0	7.4	9.5	13.4	5.2
House 2	5.3	6.1	7.0	9.2	10.9	13.8	49.6	25.6	11.0	10.5	14.9	6.2
House 3	7.6	6.3	10.7	9.1	20.0	13.0	104	24.5	19.2	10.2	28.5	5.5
House 4	5.4	6.5	7.5	9.2	13.3	13.1	99.0	25.8	14.9	10.8	26.3	8.6
House 5	6.6	6.5	9.0	9.5	13.8	13.7	36.6	26.1	11.9	10.8	11.7	6.2

Table 1.11. Summary of calculated average decay constants, average indoor emissions per m^3, and infiltration factors for all five participant houses. Indoor peaks account for values greater than five times the indoor average $PM_{2.5}$.

	House 1	House 2	House 3	House 4	House 5
Indoor 98th Percentile ($\mu g/m^3$)	26	49	100	94	35
Exceed Ambient $PM_{2.5}$ %	20	27	45	36	35
Indoor Emission Peaks (frequency, f)	263	417	533	719	160
Infiltration ($F_{in} = C_{in}/C_{out}$)	0.57	0.65	0.84	0.67	0.78
Avg Decay Constant, α (hr^{-1})	4.8	2.7	2.7	3.2	3.3
Avg Indoor Emissions, E/V ($\mu g * hr^{-1} * m^{-3}$)	1190	619	663	863	779

Figures

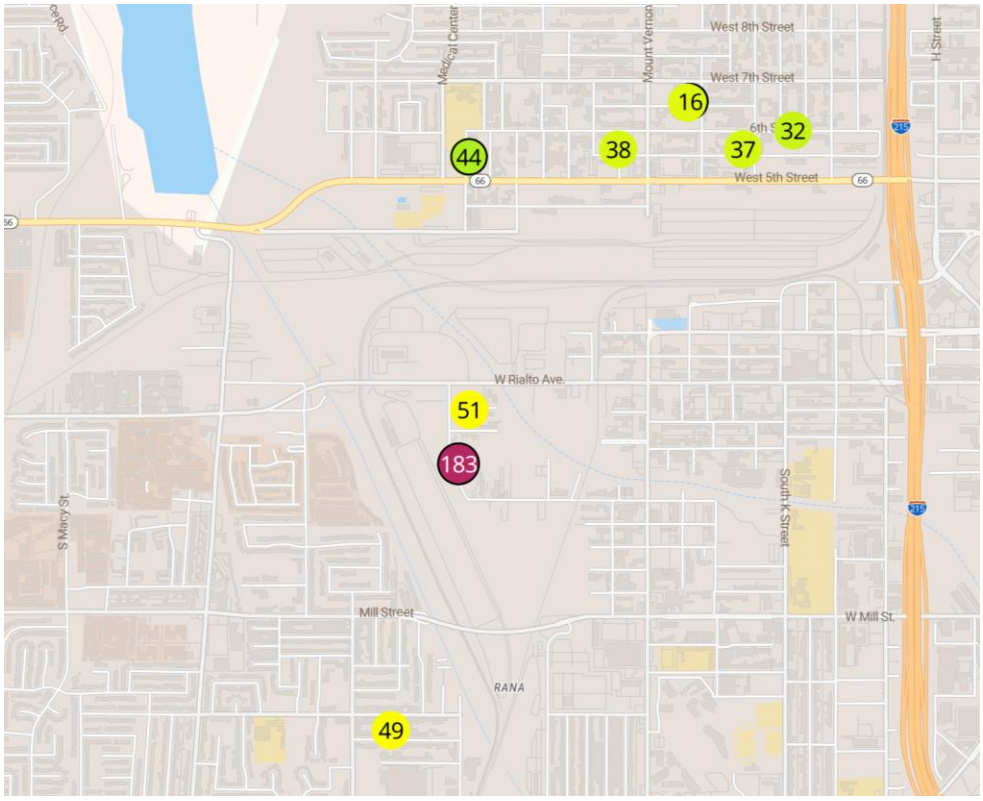


Figure 1.10. BNSF facility and household sampling locations, which are divided into three zones. Zone 1 is located within 450 meters, zone 2 is located within 1,000 meters, and zone 3 is more than 1,000 meters from the railyard. Source: map.purpleair.com

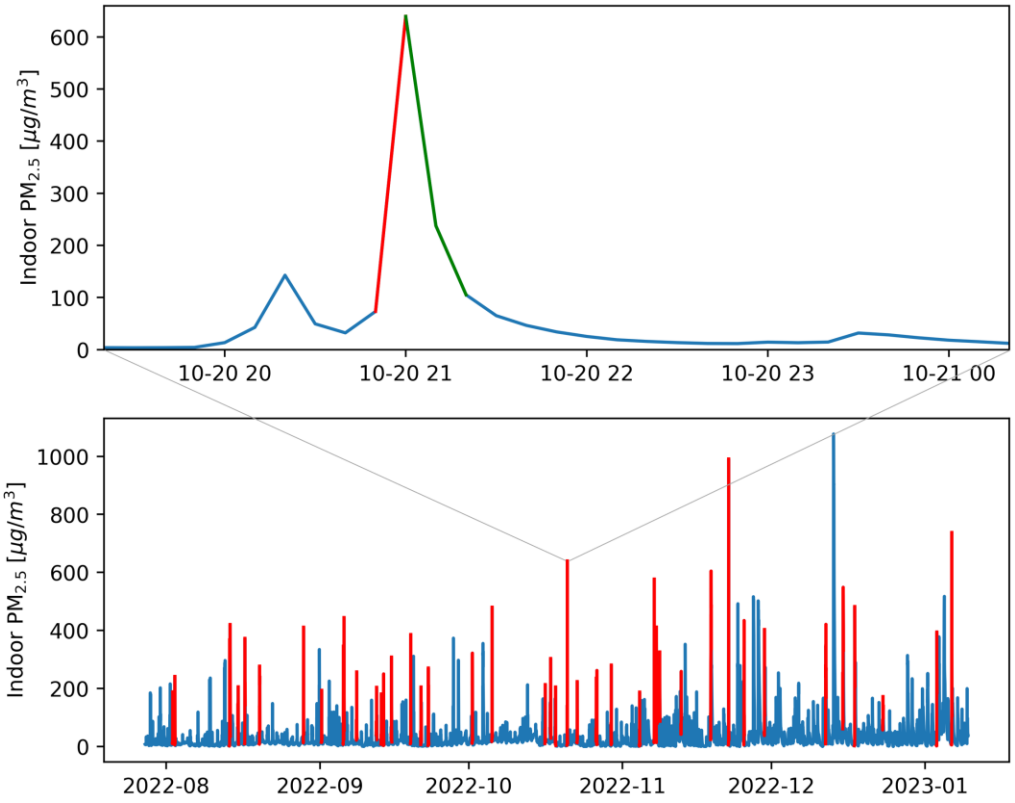


Figure 1.11. Sample time series for one home from 2022 Aug to 2023 Jan (bottom); the red lines are the data used to compute average indoor emissions. Zoom-in on the time series (top); the red line is used to calculate the indoor emissions (E/V) and green line is used to calculate the decay constant (α) based on Eqs. 3 and 5, respectively.

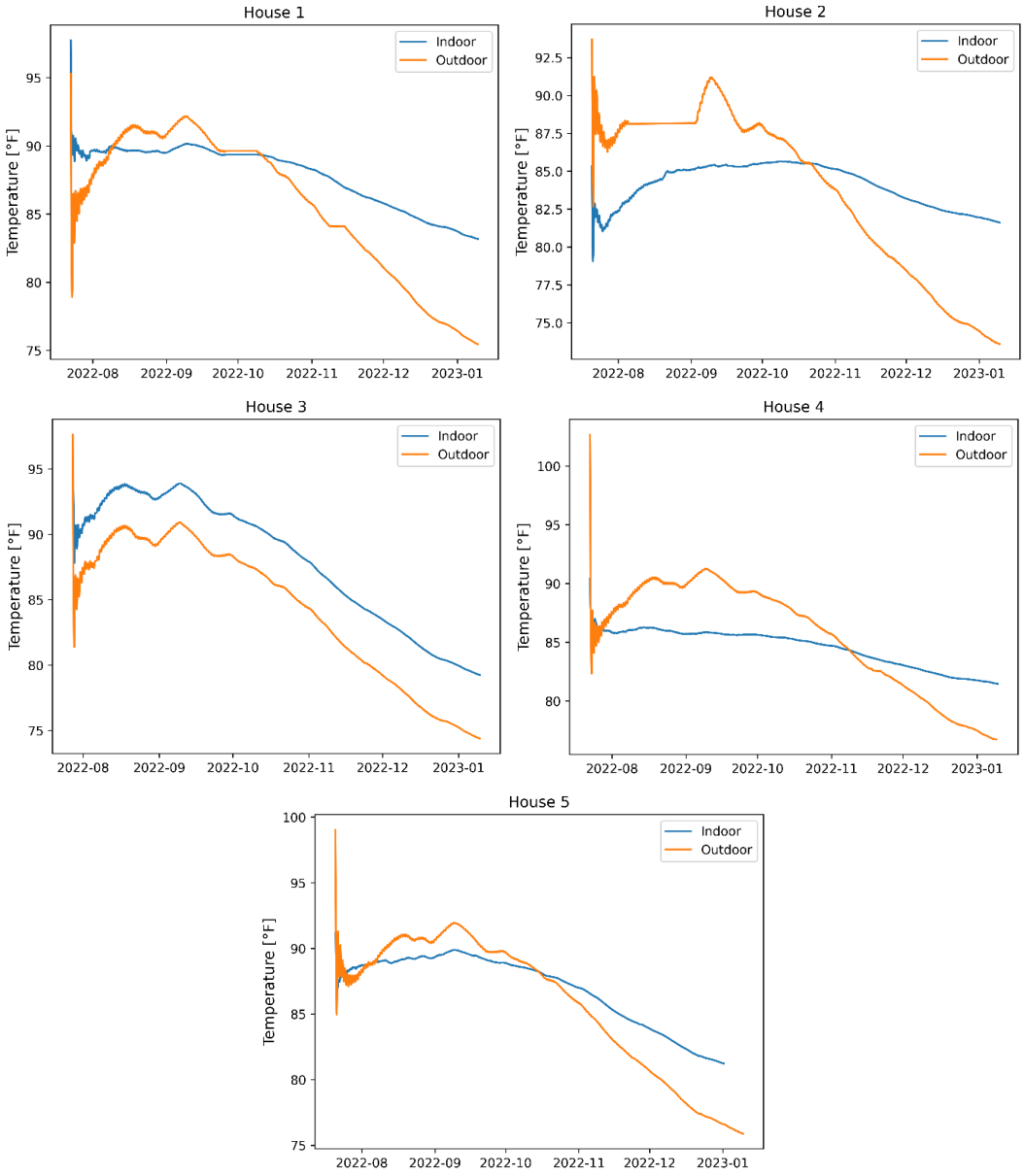


Figure 1.12. Hourly average time series plots for indoor (blue) and outdoor temperature (orange) for five participant houses. During the summertime, there were active air conditioning units to regulate indoor temperature for house 1, 2, 4, and 5. However, the indoor temperature in house 3 consistently exceeded the ambient temperature indicating there was no active air conditioning in the house.

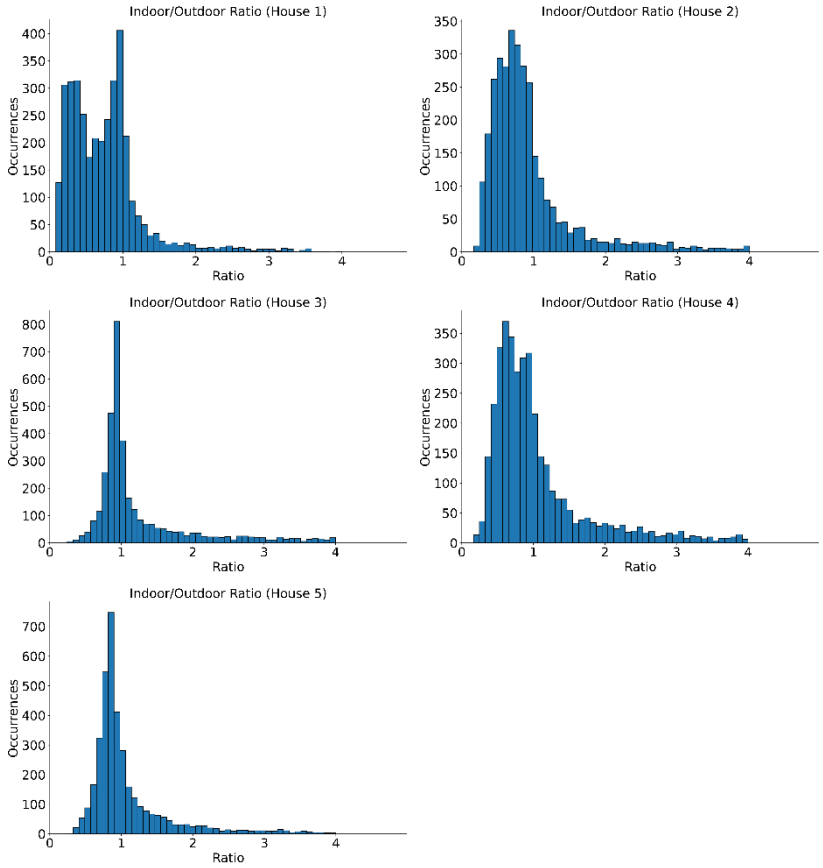


Figure 1.13. Indoor/outdoor $PM_{2.5}$ ratios for the five participant houses. The histogram was limted to 4 due to the high values when ambient concentrations were very small. Ratios are based on 10 minute average.

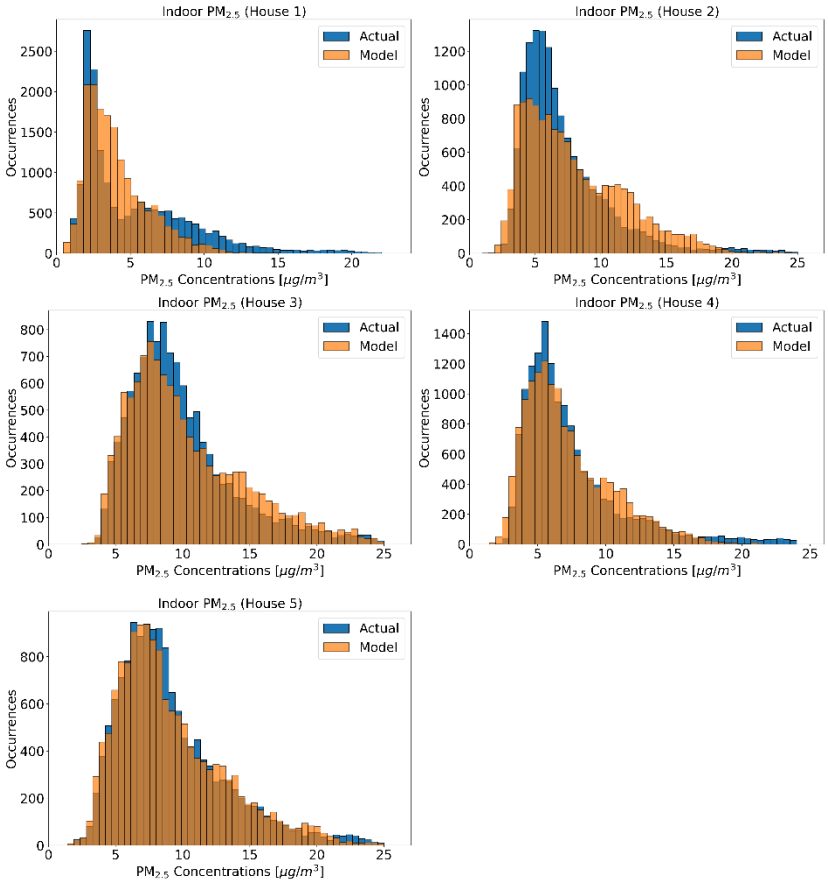


Figure 1.14. Actual indoor $PM_{2.5}$ (blue) and model $PM_{2.5}$ (orange) based on Eq. 6 based 10 minute average data. The distribution only shows the data when indoor $PM_{2.5}$ levels were less than ambient $PM_{2.5}$ levels.

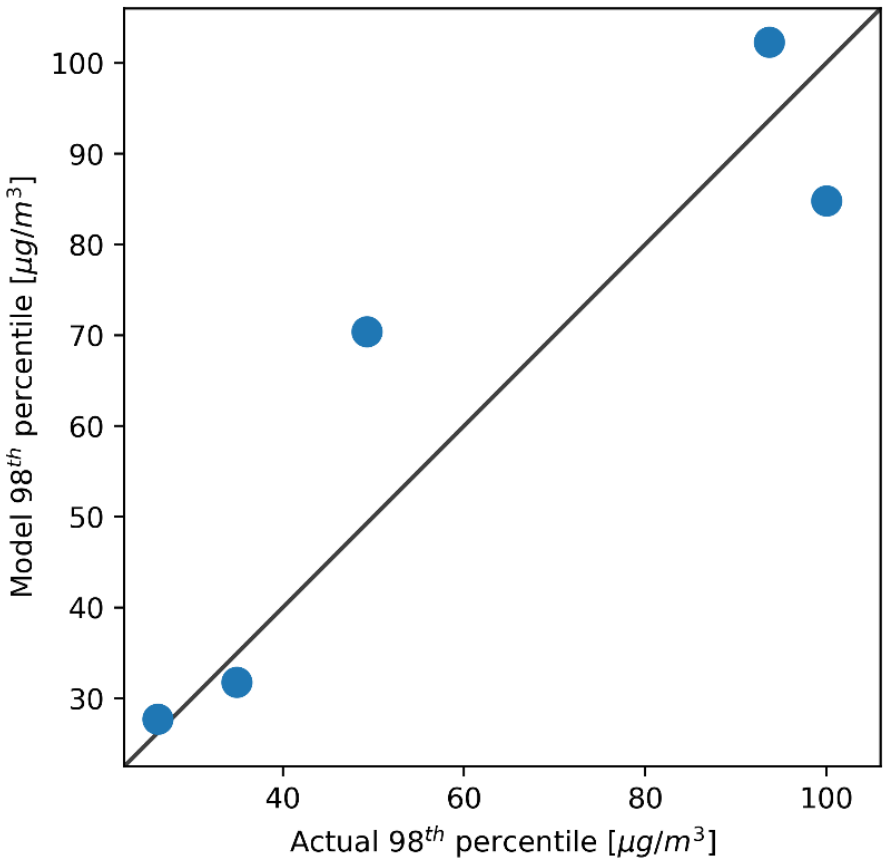


Figure 1.15. Model vs actual 98th percential of indoor $PM_{2.5}$ in five homes

Chapter 2

Part 1

A Machine Learning Approach to Quantify the Impact of Meteorology on Tropospheric Ozone in the Inland Empire, CA

The works used in this chapter were previously published in Environmental Science: Atmospheres.

Abstract

The role of meteorology in facilitating the formation and accumulation of ground-level ozone is of great theoretical and practical interest, especially due to changing global climate. In this study, with appropriate machine learning algorithms, Large meteorology and air quality datasets were analyzed to train machine learning models to (1) enhance the prediction of ozone levels in the South Coast Air Basin of California, (2) investigate the impact of recent meteorological shifts on ozone formation, and (3) determine the most critical factors influencing ozone exceedance hours. Random forest regression was used to predict historical and future trends of ozone levels, and k-nearest neighbor was used as a binary classifier for ozone exceedance prediction. The models were trained on meteorology data from Ontario and Los Angeles International Airport stations and air quality data from the Fontana, California air monitoring station, and data were collected for the 1994 to 2018 time period. Upon model evaluation, the correlation of the RFR model was 0.92, and the probability of detection for ozone exceedances using k-nearest neighbors was 0.81 for the most recent years of the analysis (2014-2018). A four km Community Multiscale Air Quality model simulation generated air pollution estimates over Southern California. As expected, ozone in Fontana was positively correlated with temperature.

The ozone exceedance hours usually occurred when the temperature was above 25 ^{o}C, and the wind direction was from 270^{o} (westerly). Ozone sensitivity as a function of temperature and NO_X was also examined. Observed troughs in hourly NO_X concentrations during midday under high temperatures suggests that most of the ambient NO_X reacted, also as expected. The results indicate that machine learning can support state implementation planning by complementing traditional air quality modeling, reducing simulation time, and exploiting large datasets for historical simulations and future air quality predictions.

Introduction

California's South Coast Air Basin (SoCAB) is well-known for its poor air quality due to its unique topography and high anthropogenic emissions. Meteorological variables and synoptic patterns greatly influence air pollution in SoCAB (Ulrickson and Mass, 1990a, 1990b). Los Angeles' temperature inversions resulting from high-pressure systems over SoCAB combined with a mountain wave-induced downslope flow creates a trap that accumulates air pollutants near the ground, leading to degraded air quality (Lu and Turco, 1995, 1994). The relationship between ozone (O_3), its anthropogenic precursors, nitrogen oxides (NO_X) and volatile organic compounds (VOC), has been well studied by means of environmental chamber experiments, field studies, and air quality modeling, yet new modeling methods are still needed to better understand why rates of ozone reduction in SoCAB have been lower than previously predicted (Baidar et al., 2015; Pusede and Cohen, 2012; Qian et al., 2019). Examining NO_X-VOC emission ratios and identifying VOC- and NO_X-limited regions are useful practices for creating surface ozone reduction strategies, thereby supporting the development of SoCAB emission-control strategies. Chemical transport modeling is generally considered the most advanced approach for evaluating emission-control

strategies, but is subject to uncertainties in emission rates, chemical reaction rates, and meteorological parameterizations. To further understand the quantitative relationship between ambient ozone concentrations and emission precursors, isopleths are developed from observed or modeled data to visualize ozone's sensitivity due to changes in NO_x and VOCs (Kinoslan, 1982; Qian et al., 2019; Sierra et al., 2013).

In recent years, SoCAB ozone has significantly decreased as a result of emissions control programs implemented by the South Coast Air Quality Management District (SCAQMD), the state of California, and the U.S. EPA. Between 1993 and 2012, NO_x and reactive organic gas (ROG) emissions in SoCAB decreased from 1,425 to 651 tons per day (tpd) of NO_x and from 1,522 to 535 tpd of ROGs (Lurmann et al., 2015). In response to the reductions, the annual average ozone from 1994 to 2011, only considering hourly concentrations between 10 am and 6 pm, decreased by 12% (64 to 57 ppb) in Riverside, CA (Lurmann et al., 2015). To achieve the 0.07 ppm 2015 National Ambient Air Quality Standard (NAAQS) for 8-hour ozone by the attainment deadline of 2037 (Figure 2.1 and Figure 2.2), SCAQMD proposed further reduction in NO_X emissions down to 250 tons per day by 2023 and 200 tons per day by 2031 by shifting from conventional fossil fuel to alternative clean fuels for mobile sources (South Coast Air Quality Management District, 2017). Since 2014, the 8-hour ozone design value for SoCAB has marginally increased despite the continuous reduction in emissions (South Coast Air Quality Management District, 2017). The emissions mitigation has unquestionably improved 8-hour ozone design value over the past several decades. However, it is conjectured that shifts in meteorology have impacted ozone improvements in recent years. Environmental researchers commonly use statistical models (i.e., multiple linear regression, generalized additive models, etc.) to predict changes in ozone

concentrations with respect to changes in meteorology and investigate the influence of synoptic and local meteorological parameters on surface ozone concentrations (Camalier et al., 2007; Gardner and Dorling, 2000; Kavassalis and Murphy, 2017; Ooka et al., 2011; Otero et al., 2016; Rao et al., 1996). The uptick in ozone concentration in recent years despite continued reductions in emissions suggests that meteorological influences should be considered when evaluating the effectiveness of control strategies in locations that are working towards NAAQS attainment. In this paper, the response of ozone to meteorology is investigated in SoCAB over a 25-year period (1994-2018) using new data-driven methods and photochemical modeling. Chemical transport models (CTMs), such as the Community Multiscale Air Quality (CMAQ) model, Goddard Earth Observing System model with atmospheric chemistry (GEOS-Chem), and the Comprehensive Air Quality Model with Extensions (CAMx), are useful tools for air quality researchers to simulate air quality trends and study the sensitivities of air pollutant levels to changes in emissions and meteorology. Although CTMs are relatively precise in representing atmospheric physics and chemical processes, handling the large datasets can be challenging given the limitation of computational efficiency and the complexity of input data. Moreover, CTM software applies complex governing equations to resolve concentrations, and most CTMs are designed for use solely with central processing units (CPUs) to carry out the simulations.

In contrast with CTMs, which solve mathematical equations to estimate the outputs, machine learning uses data to discover underlying patterns or substitute functions that mimic complex mathematical functions. CTM processes can also be optimized with modern hardware, such as graphical processing units, to reduce computation time while retaining the results' integrity (Keller et al., 2018). Presently, with many air monitoring stations across the U.S., air

quality datasets are available with high temporal resolution (hourly data). The study utilizes machine learning and air quality datasets to identify the pattern of the natural processes and explores the links between meteorology and ozone concentrations, leveraging empirical models and observational data. Previous work has been done to forecast air pollutant exceedances using supervised machine learning algorithms. For example, ozone levels have been predicted with reasonable accuracy using a feedforward neural network (Corani, 2005; Xie et al., 2009). Further, Hajek et al. (2012) presented a different approach for ozone prediction using support vector regression, which showed a significant improvement in the root mean square error (RMSE) compared to neural networks (Hájek and Olej, 2012).

The objective of this study is to explore the role of meteorology in changing ozone concentrations and ozone exceedances in SoCAB by leveraging results of machine learning and CMAQ. Meteorology-ozone sensitivity is investigated by applying machine learning to predict ozone concentrations in Fontana, California (inland Southern California), using meteorological inputs for Los Angeles and Ontario, California. The two meteorological sites represent distinct conditions due to their proximity to or distance from the Pacific Ocean. The machine learning results are analyzed against CMAQ simulations and observational data to evaluate the model performance and explore the common findings between the two approaches.

Study Location and Measurements

The California study sites include Los Angeles International Airport (LAX) and Ontario International Airport (ONT) meteorological sites and the Fontana air quality monitoring site (Figure 2.3). LAX is an upwind urban center near the coast of the Pacific Ocean, and use of LAX meteorology enables us to investigate the sensitivity of ozone concentrations at the downwind air

monitoring site with respect to upwind conditions. The temperature at LAX in the summer is lower and relative humidity is higher than the other two sites. The meteorology in ONT and Fontana is very similar because they are both in inland Southern California and are located seven miles apart, and they are approximately 50 miles from LAX. In 2018, LAX's annual average temperature was 1.0 °C lower than ONT (17.9 °C for LAX and 18.9 °C for ONT). During the 2010 to 2019 period, the 8-hour ozone design value concentration for LAX fluctuated around 80 ppb, whereas the value for Fontana was consistently above 100 ppb (Figure 2.1 and Figure 2.2).

Methods

The machine learning (ML) models presented in this paper were trained on Fontana air quality data with both LAX and ONT meteorological data. The models were evaluated using data from the Fontana air monitoring station. The ML models enabled the examination of the relationship between meteorology at any location (e.g., ONT/LAX) and Fontana's air quality. CMAQ simulation with 4 km horizontal spacing for the 2017 ozone season (May 1–Sep 30) in SoCAB provided a comparison dataset based on a deterministic model.

Data Processing

Meteorology and air quality datasets were obtained from the NOAA Climate Data Office and EPA Air Quality System (AQS) database, respectively. The AQS database provides air quality measurements for all valid EPA air monitoring sites in the United States. The meteorology datasets comprised multiple years of observations at LAX and ONT. Meteorological data were obtained for the years 1994 through 2018. Some AQS measurements were made using different samplers; therefore, to ensure uniformity of the data, records were selected from the same instrument whenever possible. Days where data were missing were marked as "NA." The data was temporally

synced from different locations based on their hourly, local timestamp. The data was randomly selected 80 percent from every year to create a training set, and the remaining 20 percent was used for ML model testing and evaluation.

Machine Learning Overview

Multiple regression-based ML algorithms were explored (e.g., neural network, support vector machine, k-nearest neighbors (K-NN), random forest) by training the models with processed air quality and meteorology data and evaluating predicted ozone concentrations. The RFR evaluation is mainly focused in this study, as its prediction of ozone concentrations is more accurate for SoCAB. Next, Binary classification was used to assign an ozone exceedance label when the observed and predicted hourly ozone concentrations are greater than 70 ppb. Further, Different classification methods were tested (e.g., support vector machine, logistic classification, perceptron) to choose the most suitable model for SoCAB.

The main difference between classification and regression is in the input and output. The output of classification of any input vector comes from a finite dictionary, $y \in \{1, \dots, m\}$, where y can be one of the m entries. In this study, the binary classification labels are exceedances and non-exceedances (m = 2) and the output, y can be either exceedances (labeled as 1) or non-exceedances (labeled as 0), whereas in regression, the output can be a real value number, $y \in R$. For regression, the input and output data are provided during training to build a function that correctly predicts the outputs for independent input data that were not used for training.

Random Forest Regression

Random forest regression (RFR) is a tree-based ensemble method, and each tree is trained on an independent collection of random input variables. In the study, the feature vector was

definded as $X = (X_i, \dots, X_n)^T$ where n is the number of X features. A function $f(x)$ for predicting the ozone concentration, Y can be solved. For a random forest of $\boldsymbol{J}$ trees, assuming the decision trees are split into $\boldsymbol{j}$ branches $h_1(x), \dots, h_j(x)$, the learning function computes the average from all decision trees, $f(x) = \frac{1}{J}\sum_{j=1}^{J} h_j(x)$. Thus, the final prediction is based on the average of all outputs from the regression trees. (Rodriguez-Galiano et al., 2015; Zhang and Ma, 2012)

Due to the nature of regression trees, RFR decision trees can have similarities in tree structures. Shown in Figure 2.4 is a three-node decision tree that assists RFR with predicting hourly ozone concentrations (between 12:00 noon and 5:00 PM) based on nitric oxide (NO) and nitrogen dioxide (NO_2) concentration and temperature. If a collection of trees in the RF has similar features, the model results are largely biased. To avoid a high correlation in their predictions, the RFR develops the algorithm such that predictions in their subtrees are less correlated by only allowing the trees to have access to a limited number of random samples from a pool of features (Breiman, 2001; Raschka and Mirjalili, 2006). The features used in the RFR model are temperature (T), relative humidity (RH), surface pressure (P), wind speed (WS), wind direction (WD), visibility (Vis), dewpoint temperature (DT), NO, and NO_2, with hourly ozone (O_3) as the target variable. To reduce bias, RFR selects a random number of features, and the maximum number of features is defined by the user. Since the concentration of ozone largely depends on precursor emissions and surface meteorology, ML was performed on a predetermined set of meteorological and air quality data to better capture the interactions of meteorology and emissions in an empirical model (Gao et al., 2022).

RF Algorithm Tuning (Model Descriptions)

Python *RandomForestRegressor* package from the scikit-learn 0.22 library was used. RFR was tuned with multiple configurations to choose the appropriate set of hyperparameters. Seven hyperparameters were varied to build the RFR model. A grid search was used for multiple combinations resulting in the optimal hyperparameters for the most accurate predictions. Figure 2.5 shows the mean absolute error (MAE) in ozone prediction when the RFR model is tuned with various configurations. For example, to find the best fit for the *n_estimators* parameter, Constant values were hold for the other options (e.g., *max_features* = 'auto', *max_depth* = None, *min_samples_split* = 5, and *min_samples_leaf* = 10) and vary *n_estimators* from 1 to 100. The results show the improvement of the trained model as the number of trees increases. However, when *n_estimators* approaches 16, the model shows no improvement in the overall performance. Based on the tuning exercise, the optimal values were picked for each hyperparameter that returned the lowest MAE. The optimal configurations are also informed by the scikit-learn documentation (Pedregosa et al., 2011). Further, while splitting each node during decision tree building, RFR picks the best split from the nine input features or a random subset of *max_features* (Table 2.1). Each tree is trained on a randomly drawn bootstrap sample with replacement from the original training dataset.

K-Nearest Neighbor Classifier

For any given prediction, the model needs to find the closest sample in the training dataset and assign its classification to the prediction label. There is no learned model for K-NN, and the algorithm has to search the entire training set for every test vector (Hastie et al., 2009). Figure 2.6 shows a binary classification in two dimensions with NO_2 on the x-axis and temperature on the y-

axis. The green dots are non-exceedances, the purple dots are exceedances, and the red dot is the datum needing to be classified. If k is 5, for example, K-NN searches throughout the training dataset to choose five closest data points and assigns the label by the majority vote amongst the 5 nearest neighbors. Selecting the correct nearest neighbor is crucial to train this model successfully. The model is overfitted when k is small and underfitted when k is large. By varying k from 1 to 8000 and keeping other parameters constant, the optimum k can be found that gives the best accuracy and probability of detection for specific K-NN models (Figure 2.7).

Neural Network

Other ML methods used in this study are neural network (NN) and support vector machines (SVM). NN is a multilayer perceptron, where each perceptron is a linear transformation followed by a nonlinear activation (e.g., signum, logistic, rectified linear activation function (ReLU)) (Sharma et al., 2020; Sharma and Sharma, 2017). Each perceptron can be expressed as $y_i^{[k]} = \varphi\left(w_i^{[k]T}x + b^{[k]}\right)$, where the superscript k denotes the nodes of hidden layers, $w_i^{[k]T}x + b_i^{[k]}$ is a linear combination model, subscript i is the perceptron at layer k, and φ is the nonlinear activation. NN is a deep network architecture with the depth of the network derived from the level of hidden layers (Hastie et al., 2009; Wang et al., 2020). Figure 2.8 shows the diagram for a fully connected 2-layer neural network. All the inputs x are connected to every perceptron in the hidden layer. $a_1^{[1]}$ is the perceptron 1 in hidden layer 1, which can be expressed as $a_1^{[1]} = \varphi\left(w_1^{[1]T}x + b_1^{[1]}\right)$. In this paper, ReLU was used as the activation function defined as $\varphi = \max\left(0, \left(w_1^{[1]T}x + b_1^{[1]}\right)\right)$. In terms of matrix representation, the output of the 1st hidden layer is $y^{[1]} = \varphi\left(W^{[1]}x + b^{[1]}\right)$, and the output layer is $\hat{y} = \varphi\left(W^{[2]}y^{[1]} + b^{[2]}\right)$. the weights $W^{[1]}$ and

$W^{[2]}$ that give the best prediction can be found. In general, for multi-layer neural network the output can be expressed as $\hat{y} = w^T \varphi\left(W^{[L]} \varphi\left(W^{[L-1]} \dots \varphi\left(W^{[2]} \varphi\left(W^{[1]} x\right)\right)\right)\right)$, where w is the weight of the final layer. After computing the predicted output $\hat{y}$, the loss function is used to evaluate the difference between the predictions and actual values, $L(W) = l(y, \hat{y})$. The gradient descent is used to update the weight, W to obtain better predictions for the next iterations. The process repeats with the new updated W until the loss no further substantially decreases.

Support Vector Machines

SVM is a learning algorithm that optimizes a hyperplane to maximize the margin between different data types. The property required for SVM is to find the supporting hyperplanes and maximize the gap between them (Hastie et al., 2009; Hearst et al., 1998). Figure 2.9 shows the separating hyperplane (solid line) in the center of the two supporting hyperplanes (dash lines) that maximize the margin between the black dots and white dots. The supporting hyperplanes can be expressed as $w^T x + b = c$ for black dots and $w^T x + b = -c$ for the white dots, where w is the weight, x is the input, b is the bias, and c is the arbitrary distance which can be set to 1. The distance $(\frac{2c}{||w||_2})$ between two supporting hyperplanes can be maximized by solving the optimization problem, as follows:

$$\underset{w,\,b}{minimize}\ \frac{1}{2}\left|\left|w\right|\right|_2^2$$

$$subject\ to\ y_i(w^T x_i + b) \geq 1\ for\ all\ i.$$

CMAQ Model Descriptions

CMAQ version 5.2.1 was used to carry out the simulation, as the CMAQ model is one of the EPA regulatory methods used to develop ozone attainment control strategies in SoCAB. The CMAQ simulation was carried out for the ozone season of 2017 (May-01 to Oct-01) to concurrently examine the trends that are driving NAAQS nonattainment using a first principles model alongside empirical approaches. The details of CMAQ descriptions can be found in the SI.

CMAQ Model Evaluations

Nine key monitoring stations were chosen in SoCAB to evaluate the CMAQ simulation, and the stations are located in Anaheim, Azusa, Crestline, Fontana, Los Angeles, Pasadena, Redlands, Rubidoux, and San Bernardino, California (Figure 2.10). The EPA guidelines (Usepa, 2009) and computed a set of unbiased metrics were followed to evaluate the model (Yu et al., 2006). The metrics include correlation coefficient (CC), mean bias error, mean absolute error (MAE), root mean square error (RMSE), relative root mean square error, mean normalized bias, mean normalized absolute error, normalized mean bias (NMB), normalized mean absolute error, fractional bias, fractional absolute error, model mean, and observational mean; the formulas are listed in the SI. the regression algorithms were evaluated using the intrinsic metrics of linear fit (e.g., R^2, slope, and intercept), CC, RMSE, and MAE. Different classification algorithms were evaluated based on probability of detection (PoD), accuracy, model error, and failure to predict (Eqs. 14-17 in the SI).

Results

Machine Learning Model Evaluation

Before choosing RFR as the principal regressor and K-NN as the principal classifier for the ML applications, the predictions were evaluated for a total of five different models using data for the 1994 to 2018 period (Tables 2.2 and 2.3). RFR had the best performance out of four models with the highest CC and R^2 and the lowest RMSE and MAE. For classifiers, PoD is the model's ability to detect ozone exceedances for exceedance hours only, and accuracy reflects the performance of the prediction for both exceedances and non-exceedances. Both K-NN and perceptron had reasonable evaluation results. Perceptron had a higher PoD but was less accurate and overfitted the data. K-NN was chosen as the principal classifier for the model, as model accuracy was prioritized.

A ten-fold cross-validation was carried out to further evaluate the skill of the RFR model. First, the data were shuffled randomly and split into ten groups of equal size. Nine groups were chosen to train the model, and one group was used for evaluation; this was repeated such that each group served as the evaluation group one time. The model prediction was evaluated by comparing the slope, intercept, R^2, RMSE, and MAE (Table 2.4). The ten-fold cross-validation gave consistent performance for each testing group (K) and returned the same RMSE and MAE.

The performance of the RFR model was evaluated for five-year time periods, as ozone concentrations exhibited trend changes roughly every five years (Table 2.5). The model was trained on 80% of hourly data from 12:00 noon to 5:00 PM during the period of 1994 to 2018, and the remaining 20% of the data were used to test the model in five-year increments (e.g., 1994-1998). In the three periods 1994-1998, 2004-2008, and 2009-2013, the NMB values were negative,

indicating that the model underestimates by a factor of 1.069, 1.041, and 1.029, respectively. In the other two periods from 1999-2003 and 2014-2018, NMB values were positive, and the model generally overestimated by a factor of 1.002 and 1.003. A small NMB and a consistently high CC above 0.87 suggests the high performance of the model. Therefore, it is recommended that suitable RFR evaluation metrics for ozone are $|NMB| \leq 1.07$ and a $CC \geq 0.85$.

Historical Trends with the RFR Model

Results here reflect the trained RFR model for the timespan from 12:00 PM to 5:00 PM when ozone concentrations are high. Figures 2.12 and 2.13 show the three most important variables influencing modeled ozone concentrations in Fontana based on a feature importance screening. Ozone exceedances (defined as hourly observations greater than 0.070 ppm) are associated with high temperature, moderate wind speeds, and lower observed NO_x (Figure 2.11). High temperatures accelerate ozone's photolytic cycle, while moderate wind speeds accommodate mixing and transport of precursor pollutants. Low NO_x conditions suggest that during high ozone hours, NO_x is depleted due to the rapid atmospheric turnover of NO_2. In the presence of sunlight, NO_2 is converted to NO and triplet oxygen, where the triplet oxygen reacts with O_2 to form O_3. The model performance metric R^2 was improved when the model was trained on ONT meteorology, which is most representative of Fontana meteorology, reflecting the dependence of model performance on local meteorology.

Figure 2.14 highlights the dynamic application of the RFR model for the 2014-2018 period. Since all-weather elements (e.g., temperature, RH, and surface pressure) are interdependent, varying one strongly affects others. To create the contours, not only temperature and NO_x were varied but also dynamic pressure and RH arrays were created by taking the average of the

observed RH and pressure at a certain temperature interval. A series of ozone sensitivity tests was performed by continuously feeding desired datasets to the RFR model with varied values of NO_x, temperature, temperature-dependent RH, and temperature-dependent pressure while keeping wind speed constant at 9 m/s. Figure 2.14 shows the behavior of ozone with changes in NO_x and temperature for four different wind directions (90°, 180°, 270°, and 360°). Ozone concentration reached its maximum at the mid-NO_x and high-temperature regime, as predicted by the dynamic RFR model. Ozone significantly decreased as the conditions moved orthogonally in the opposite direction of the high ozone region. Figure 2.14d shows that the exceedance usually occurred when NO_x concentration was around 10 ppb and the temperature was higher than 35 °C. More than 60% of the time, the wind direction in ONT was from the west, and 25% of the time it was between 254° and 273° (Figure 2.15), which occurred during highest concentration for the region. Ozone concentrations ranged from 0.06 ppm to 0.14 ppm, depending on wind direction. Concentrations were highest when the wind direction was 270° but reached a low for 90° wind direction. The RFR model predicted that the exceedances started to develop at a temperature greater than 30 °C. The ozone concentration gradient (i.e., the change in ozone concentration per unit change in temperature) was small at low temperatures. However, when the temperature approached 30 °C, the gradient became large, and the ozone concentration increased significantly. Ten percent of observed data were plotted on the top of the contour plots in Figure 2.16 to validate the observed likelihood of the prediction. The observational data were unevenly spread throughout the domain of the plots. However, the sparseness of observational data was found at the extremely low and high regimes of temperature.

Initially, the RFR model utilized all nine features to predict ozone concentrations. To test ozone prediction sensitivity, the model was trained with nine, eight, seven, and six features after selectively dropping features. The predicted results from each iteration were evaluated against observations using the Wilcoxon rank sum test. If the output from this test was less than or equal to 0.05, two samples were independent of one another, indicating the significance of the dropped feature(s) for the model prediction. The list Wilcoxon tests were computed. The three most important meteorological parameters were wind speed, RH, and temperature as they appeared most frequently in the significant Wilcoxon rank sum tests, suggesting that if they are absent from training features, the RFR model would likely fail to perform with similar accuracy. Also, the two-sample t-test strongly points out that if the three dropped features are RH, wind speed, and temperature or wind speed, NO, and NO_2, the model would likely have poor agreement with observations. Feature drop test shows three-feature removal combinations where the CC is less than or equal to 0.8.

Predicting the Exceedance Hours Using K-Nearest Neighbors

RFR underestimates high ozone concentrations and fails when it comes to extreme ozone levels. The k-nearest neighbor algorithm overcame this barrier when predicting exceedances and proved its accuracy for binary classification (Table 2.6). The K-NN model was evaluated for 1994-2018 in five-year intervals, similar to the procedure for the RFR model. The PoD of K-NN ranged from 0.58 for the earliest period to 0.81 for the latest period, indicating the improving model performance in later years (Table 2.6). The accuracy was above 0.71, and only 19% of the time did K-NN not detect the exceedances in the 2014 to 2018 period. Because the dataset was unbalanced due to a higher frequency of non-exceedances, the accuracy yields can be slightly misleading. Even

though the model obtained high accuracy, it failed to detect ozone exceedances for up to 42% of the time in earlier years. Figure 2.17 shows 2 x 2 confusion matrices for every five years from 1994 to 2018. The dominance of correct non-exceedance prediction (quartile I) and correct exceedance prediction (quartile IV) confirms the overall satisfactory performance of the K-NN model.

Even though NO_x and VOCs are two significant components influencing ozone formation, meteorology is also a crucial driving force. Figure 2.18 shows an oscillating pattern of temperature, alternating between winters and summers from 2014-2018 as expected. Below 22^oC, no exceedances occurred, and most exceedances occurred during the summertime. The K-NN model successfully explained the link between temperature and exceedances and accurately predicted the exceedances when the temperature is high and predicted no exceedances when the temperature fell below 22^oC. The exceedances did not usually occur for high NO_x, high RH, or low wind speed. As evident in Figures 2.16 2.23, high RH was associated with low temperatures and ozone, and lower NO_x concentrations were associated with high temperatures in this analysis. Figure 2.18 shows a strong relationship between specific meteorological regimes and exceedance hours. The marine layer penetration on foggy days might cause high RH. During these episodes, the marine layer is deep and moves farther inland with the clean air.

CMAQ Evaluation

The CMAQ simulation provides a deterministic evaluation for comparison with the ML predictions. The daily average ozone concentrations from the 2017 CMAQ simulation were extracted and evaluated against observational data at nine air monitoring sites (Table 2.7). Positive MB for all evaluation sites suggest the overall overestimates of the model with a maximum MB of 16 ppb (Fontana) and a minimum of 6 ppb (Crestline and LA). The overestimation occurred

because the model did not capture the low ozone concentrations at night (Figure 2.19), which significantly increased the CMAQ daily average ozone concentrations. Since this paper focuses on ML, the details of the comparative CMAQ evaluation can be found in the SI.

Methods Strengths and Limitations

The ML model nimbly predicts the changes of the target variable with respect to a perturbation in input features (i.e., ozone response to changing in temperature). The effect of meteorology can also be determined by examining trends over a long period of time. When using average temperature and RH from 1994 to 2018, the ML prediction minimized the effect of meteorology extremes on ozone formation (i.e., heat waves, foggy days). Figure 2.20 shows the annual 90^{th} percentile (blue), annual 98^{th} percentile (black), and the annual average (orange) ozone trends at the Fontana location for 1994 to 2018. The dashed lines are the ML prediction with the average temperature and RH, and the solid lines are the prediction with the actual features. The adjusted line shows a strong downward ozone trend from 1994 to 2010, but resisting further decrease in later years. The distance between the 98^{th} and 90^{th} ozone percentile was narrow, indicating the high frequency of high ozone concentrations in Fontana. The average meteorology had minor effects on the annual average prediction. Despite the downward trend of ozone concentrations for the 90^{th} and 98^{th} percentile, the annual average increased.

In contrast with ML, which focused on a targeted pointwise location, CMAQ simulations covered a large spatial domain (102 x 156 grid cells with 4 km spacing) over the South Coast air basin. As expected, model performance is variable when evaluated at specific locations. Despite the less-than-favorable performance at the Fontana location in terms of mean bias error (Table 2.7) compared to nine other sites, CMAQ performed better in other locations. Further, Figure 2.21

shows monthly spatial evaluations for June and July 2017 for 25 air monitoring sites in SCAQMD, which demonstrates CMAQ's utility in enabling simultaneous detailed examinations of different areas for multiple species, while covering a sizeable spatial area. This paper examines how ML vs. first principles modeling performs for a similar analysis, providing useful insight into the strengths and weaknesses of the methods for the application detailed here.

Discussion

The RFR model was the preferred regressor model for this application. However, the RFR model underestimated high ozone levels and overestimated low ozone levels due to the nature of RFR, in which the model takes the average of all the decision trees. To compensate for this limitation when predicting ozone exceedance hours, binary classifications were used. The high PoD and accuracy of the K-NN model suggested that K-NN was better suited for ozone exceedance prediction. It can be improved and fine-tuned to achieve better results by optimizing the number of neighbors, leaf size, and the algorithm to compute the nearest neighbors.

Evaluation of ML and CMAQ results showed that the temperature was the most significant contributing factor to high ozone concentrations, resulting in spikes of exceedances during the hot summer days. The relationship of temperature with ozone exceedances also varies in different topological regions. A. K. Gorai et al., showed that temperature had no uniform correlations and effects on the ozone trend in eastern Texas in May 2012 (Gorai et al., 2015). However, in SoCAB, RFR and CMAQ models strongly suggest that temperature is the primary driving force. In the VOC-limited SoCAB (Benosa et al., 2018; Heuss et al., 2003) urban area, a reduction in NO_x or increase in VOCs may increase ozone formation. During the most severe California drought years (2011-2015), isoprene decreased by more than 50% (Demetillo et al., 2019), resulting in a considerable

reduction in ozone levels. Figures 2.22 and 2.23 show the daily average emissions of biogenic isoprene and NO_x in 2012 (South Coast Air Quality Management District, 2017). From January to April, isoprene emissions slowly increase and surpass NO_x emissions in May. Emissions remain high throughout the summer and decrease after October. Vegetation emits a large amount of isoprene and other biogenic BVOCs at high temperatures (temperature-dependent isoprene emissions) (Coates et al., 2016), causing an increase in total VOC levels in the summer. NO_x emissions during 2012 were roughly constant, with the lows at 130 moles/s and the highs at 210 moles/s due to the estimated constant contributions from traffic and industrial activities throughout the year. Thus, the high summertime ozone concentration can be partially explained by increased reactions between excess BVOCs emitted from vegetation and NO_x, resulting in increased ozone levels in such a VOC-limited regime.

Wind speed and wind direction also influences ozone levels, as shown in the contour plots. Ozone precursors accumulate at low wind speeds, and high ozone levels occur when the wind speed is between 2 - 4 m/s. This is optimal wind speed and wind direction to accelerate chemical transport and mixing. More than 64% of the time, the direction of the wind in ONT is from the Los Angeles city center to the east, which transports ozone and precursors to Fontana and further contributes to ground-level ozone formation. High ozone levels occur in the summer when the temperature exceeds 25 ^{o}C, and the NO_x concentration is low due to the reaction of NO_2 with the OH radical.

Results here corroborate the previously demonstrated strong relationship of ozone with meteorology in a data-driven framework. Wind speed and wind direction contribute mainly to transport and mixing of precursors, while the temperature can be a direct contributing factor for

catalyzing ozone formation. Climate-related increases in temperature would therefore be expected to increase future ozone levels in the absence of emission changes. The time series from the RFR and CMAQ models shows spikes in temperature that correspond to ozone concentration peaks. Low RH occurs during high-temperature periods, and high RH is observed during low-temperature periods. The predicted effect of RH on ozone level as small, and when RH reached 100%, predicted ozone dramatically decreases. (Jia and Xu, 2014) RH is a significant feature for ensuring model accuracy based on significance tests.

Conclusion

Large, publicly available meteorological databases and open-source libraries (TensorFlow, scikit-learn, and PyTorch) have made ML an efficient and complementary modeling approach for studying long-term air pollution trends, compared to CTMs. CTMs and ML serve different purposes, where CTMs are useful for predicting future pollution levels in response to emission controls. This paper has shown that the RFR and K-NN models were satisfactory for ozone exceedance prediction in SoCAB during the 2017 ozone season. From significance testing and feature importance screening, meteorology data improved model prediction accuracy. In Fontana, ozone exceedances occurred at high temperatures, during periods of lower observed NO_x, wind speed above three m/s, and the RH between 10% to 50%. RFR ML models can be improved by choosing the minimum set of features spanning the tree dependency (Juszczak et al., 2009). It is of further interest to create ML models that take input from weather forecasting models to predict ozone concentrations in three dimensions. In future applications, the configurations will be tuned on multiple ML algorithms to obtain the most suitable model that accurately predicts ozone exceedances based on meteorology inputs.

Acknowledgments

The authors thank Prof. Armistead G. Russell and Dr. Sang-Mi Lee for their helpful contributions to this work. This paper was prepared as a result of work sponsored, paid for, in whole or in part, by the South Coast Air Quality Management District (SCAQMD). The opinions, findings, conclusions, and recommendations are those of the authors and do not necessarily represent the views of SCAQMD. We acknowledge Graduate Assistant in Areas of Need (GAANN) support from the University of California, Riverside Chemical and Environmental Department.

References

Baidar, S., Hardesty, R.M., Kim, S.-W., Langford, A.O., Oetjen, H., Senff, C.J., Trainer, M., Volkamer, R., 2015. Weakening of the weekend ozone effect over California's South Coast Air Basin. Geophysical Research Letters 42, 9457–9464. https://doi.org/10.1002/2015GL066419

Benosa, G., Zhu, S., Kinnon, M.M., Dabdub, D., 2018. Air quality impacts of implementing emission reduction strategies at southern California airports. Atmospheric Environment. https://doi.org/10.1016/j.atmosenv.2018.04.048

Breiman, L., 2001. Random forests. Machine Learning. https://doi.org/10.1023/A:1010933404324

Camalier, L., Cox, W., Dolwick, P., 2007. The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. Atmospheric Environment. https://doi.org/10.1016/j.atmosenv.2007.04.061

Coates, J., Mar, K.A., Ojha, N., Butler, T.M., 2016. The influence of temperature on ozone production under varying NOx conditions - A modelling study. Atmospheric Chemistry and Physics. https://doi.org/10.5194/acp-16-11601-2016

Corani, G., 2005. Air quality prediction in Milan: Feed-forward neural networks, pruned neural networks and lazy learning. Ecological Modelling. https://doi.org/10.1016/j.ecolmodel.2005.01.008

Demetillo, M.A.G., Anderson, J.F., Geddes, J.A., Yang, X., Najacht, E.Y., Herrera, S.A., Kabasares, K.M., Kotsakis, A.E., Lerdau, M.T., Pusede, S.E., 2019. Observing Severe Drought Influences on Ozone Air Pollution in California. Environmental Science and Technology. https://doi.org/10.1021/acs.est.8b04852

Gao, Z., Ivey, C.E., Blanchard, C.L., Do, K., Lee, S.-M., Russell, A.G., 2022. Separating emissions and meteorological impacts on peak ozone concentrations in Southern California using generalized additive modeling. Environmental Pollution 307, 119503. https://doi.org/10.1016/j.envpol.2022.119503

Gardner, M.W., Dorling, S.R., 2000. Statistical surface ozone models: An improved methodology to account for non-linear behaviour. Atmospheric Environment. https://doi.org/10.1016/S1352-2310(99)00359-3

Gorai, A.K., Tuluri, F., Tchounwou, P.B., Ambinakudige, S., 2015. Influence of local meteorology and NO2 conditions on ground-level ozone concentrations in the eastern part of Texas, USA. Air Quality, Atmosphere and Health. https://doi.org/10.1007/s11869-014-0276-5

Hájek, P., Olej, V., 2012. Ozone prediction on the basis of neural networks, support vector regression and methods with uncertainty. Ecological Informatics. https://doi.org/10.1016/j.ecoinf.2012.09.001

Hastie, T., Tibshirani, R., Friedman, J., 2009. Springer Series in Statistics.

Hearst, Marti.A., Scholkopf, Bernhard., Dumais, Susan., Osuna, Edgar., Platt, J., 1998. Supprot vector machines. IEEE Intelligent Systems and their Applications.

Heuss, J.M., Kahlbaum, D.F., Wolff, G.T., 2003. Weekday/weekend ozone differences: What can we learn from them? Journal of the Air and Waste Management Association. https://doi.org/10.1080/10473289.2003.10466227

Jia, L., Xu, Y., 2014. Effects of relative humidity on ozone and secondary organic aerosol formation from the photooxidation of benzene and ethylbenzene. Aerosol Science and Technology. https://doi.org/10.1080/02786826.2013.847269

Juszczak, P., Tax, D.M.J., Pękalska, E., Duin, R.P.W., 2009. Minimum spanning tree based one-class classifier. Neurocomputing. https://doi.org/10.1016/j.neucom.2008.05.003

Kavassalis, S.C., Murphy, J.G., 2017. Understanding ozone-meteorology correlations: A role for dry deposition. Geophysical Research Letters. https://doi.org/10.1002/2016GL071791

Keller, C.A., Evans, M.J., Kutz, J.N., Pawson, S., 2018. Machine learning and air quality modeling, in: Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017. https://doi.org/10.1109/BigData.2017.8258500

Kinoslan, J.R., 1982. Ozone-Precursor Relationships from EKMA Diagrams. Environmental Science and Technology. https://doi.org/10.1021/es00106a011

Lu, R., Turco, R.P., 1995. Air pollutant transport in a coastal environment-II. Three-dimensional simulations over Los Angeles basin. Atmospheric Environment. https://doi.org/10.1016/1352-2310(95)00015-Q

Lu, R., Turco, R.P., 1994. Air Pollutant Transport in a Coastal Environment. Part I: Two-Dimensional Simulations of Sea-Breeze and Mountain Effects. Journal of the Atmospheric Sciences. https://doi.org/10.1175/1520-0469(1994)051<2285:aptiac>2.0.co;2

Lurmann, F., Avol, E., Gilliland, F., 2015. Emissions reduction policies and recent trends in Southern California's ambient air quality. Journal of the Air and Waste Management Association. https://doi.org/10.1080/10962247.2014.991856

Ooka, R., Khiem, M., Hayami, H., Yoshikado, H., Huang, H., Kawamoto, Y., 2011. Influence of meteorological conditions on summer ozone levels in the central Kanto area of Japan, in: Procedia Environmental Sciences. https://doi.org/10.1016/j.proenv.2011.03.017

Otero, N., Sillmann, J., Schnell, J.L., Rust, H.W., Butler, T., 2016. Synoptic and meteorological drivers of extreme ozone concentrations over Europe. Environmental Research Letters. https://doi.org/10.1088/1748-9326/11/2/024005

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research.

Pusede, S.E., Cohen, R.C., 2012. On the observed response of ozone to NO_x and VOC reactivity reductions in San Joaquin Valley California 1995–present. Atmospheric Chemistry and Physics 12, 8323–8339. https://doi.org/10.5194/acp-12-8323-2012

Qian, Y., Henneman, L.R.F., Mulholland, J.A., Russell, A.G., 2019. Empirical Development of Ozone Isopleths: Applications to Los Angeles. Environmental Science and Technology Letters. https://doi.org/10.1021/acs.estlett.9b00160

Rao, S.T., Zurbenko, I.G., Flaum, J.B., 1996. Moderating the Influence of Meteorological Conditions on Ambient Ozone Concentrations. Journal of the Air and Waste Management Association. https://doi.org/10.1080/10473289.1996.10467439

Raschka, S., Mirjalili, V., 2006. Python Machine Learning (Second Edition).

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews. https://doi.org/10.1016/j.oregeorev.2015.01.001

Sharma, Siddharth, Sharma, Simone, 2017. Understanding Activation Functions in Neural Networks. International Journal of Engineering Applied Sciences and Technology.

Sharma, Siddharth, Sharma, Simone, Athaiya, A., 2020. ACTIVATION FUNCTIONS IN NEURAL NETWORKS. International Journal of Engineering Applied Sciences and Technology. https://doi.org/10.33564/ijeast.2020.v04i12.054

Sierra, A., Vanoye, A.Y., Mendoza, A., 2013. Ozone sensitivity to its precursor emissions in northeastern Mexico for a summer air pollution episode. Journal of the Air and Waste Management Association. https://doi.org/10.1080/10962247.2013.813875

South Coast Air Quality Management District, 2017. Final 2016 Air Quality Management Plan.

Ulrickson, B.L., Mass, C.F., 1990a. Numerical investigation of mesoscale circulations over the Los Angeles basin. Part I: a verification study. Monthly Weather Review. https://doi.org/10.1175/1520-0493(1990)118<2138:NIOMCO>2.0.CO;2

Ulrickson, B.L., Mass, C.F., 1990b. Numerical investigation of mesoscale circulations over the Los Angeles basin. Part II: synoptic influences and pollutant transport. Monthly Weather Review. https://doi.org/10.1175/1520-0493(1990)118<2162:NIOMCO>2.0.CO;2

Usepa, 2009. Guidance on the development, evaluation, and application of environmental models. USEPA Publication.

Wang, X., Kumar, A., Shelton, C.R., Wong, B.M., 2020. Harnessing deep neural networks to solve inverse problems in quantum dynamics: Machine-learned predictions of time-dependent optimal control fields. Physical Chemistry Chemical Physics. https://doi.org/10.1039/d0cp03694c

Xie, H., Ma, F., Bai, Q., 2009. Prediction of indoor air quality using artificial neural networks, in: 5th International Conference on Natural Computation, ICNC 2009. https://doi.org/10.1109/ICNC.2009.502

Yu, S., Eder, B., Dennis, R., Chu, S.-H., Schwartz, S.E., 2006. New unbiased symmetric metrics for evaluation of air quality models. Atmospheric Science Letters. https://doi.org/10.1002/asl.125

Zhang, C., Ma, Y., 2012. Ensemble machine learning: Methods and applications, Ensemble Machine Learning: Methods and Applications. https://doi.org/10.1007/9781441993267

Tables

Table 2.1. Final configurations for RFR model

Hyperparameter	**Description**
n_estimators = 16	The number of trees in the forest.
max_features = 'auto'	The number of features to consider when looking for the best split.
max_depth = None	The maximum depth of the tree.
min_samples_split = 5	The minimum number of samples required to split an internal node.
min_samples_leaf = 30	The minimum number of samples required to be at a leaf node.
min_weight_fraction_leaf = 0	The minimum weighted fraction of the sum total of weights required to be at a leaf node.
max_leaf_nodes = None	Best nodes are defined as relative reduction in impurity.
n_jobs = 8	The number of jobs to run in parallel.

Table 2.2. Ozone prediction evaluation metrics for four regression models (random forest, neural network, support vector machine, and K-nearest neighbors). The models were trained on nine features from 1994 to 2018. The models were constructed using 80% of data and evaluated using 20% of the data from the 2014-2018 period. The evaluations are in the unit ppm.

Regressor	**CC**	**Slope**	**Intercept**	**R^2**	**RMSE**	**MAE**
RFR	0.927	0.875	0.00605	0.861	0.009	0.006
Neural Network	0.860	0.807	0.00703	0.689	0.014	0.011
SVR	0.787	1.03	0.0194	0.619	0.028	0.024
K-NN	0.921	0.869	0.00702	0.848	0.010	0.007

Table 2.3. Ozone prediction evaluation metrics for four classifier models (support vector machine, neural network, k-nearest neighbors, and perceptron). The models were trained on nine features from 1994 to 2018. The models were constructed using 80% of data and evaluated using 20% of the data from the 2014-2018 period. The evaluations are in the unit ppm.

Classifier	**PoD**	**Accuracy**	**Failure to Predict**
SVM	0.07	0.83	0.93
Neural Network	0.76	0.71	0.24
K-NN	0.81	0.71	0.19
Perceptron	0.83	0.69	0.17

Table 2.4. Ten-fold cross-validation evaluation metrics for the RFR model for the period from 1994 to 2018.

Metrics	K1	K2	K3	K4	K5	K6	K7	K8	K8	K10
Slope	0.798	0.798	0.786	0.794	0.786	0.789	0.788	0.787	0.791	0.789
Intercept	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007
R^2	0.768	0.769	0.767	0.763	0.773	0.759	0.765	0.765	0.763	0.764
RMSE	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013
MAE	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008

Table 2.5. Five-year summary statistics for the RFR model vs. observational data from the Fontana air quality monitoring station. The differences between the model and observational means were minimal. Biases and errors are in units of ppm.

Year	CC	MB	MAE	RMSE	MNB	MNAE	NMB	NMAE	FB	FAE	$\bar{M}$	$\bar{O}$
1994-1998	0.88	-0.004	0.013	0.018	0.035	0.263	-0.069	0.214	-0.039	0.234	0.055	0.059
1999-2003	0.882	0	0.01	0.014	0.101	0.276	0.002	0.203	0.028	0.235	0.048	0.048
2004-2008	0.884	-0.002	0.009	0.013	0.011	0.19	-0.041	0.165	-0.025	0.182	0.052	0.054
2009-2013	0.884	-0.002	0.008	0.011	-0.009	0.147	-0.029	0.142	-0.024	0.15	0.055	0.057
2014-2018	0.927	0	0.006	0.011	0.036	0.151	0.003	0.142	0.017	0.143	0.058	0.058

Table 2.6. Summary statistics for K-NN exceedance hour predictions. Exceedance hours occurred when ozone concentrations were greater than 70 ppb. The K-NN model was evaluated using 20% of the data from 1994-2018. The probability of detection was calculated as the number of correct exceedance predictions divided by the total actual exceedances. Failure to predict is 1 – PoD. Accuracy is the correct predictions for non-exceedance and exceedance hours divided by the total hourly observations.

Year	PoD	Accuracy	Failure to Predict
1994-1998	0.58	0.84	0.42
1999-2003	0.69	0.87	0.31
2004-2008	0.69	0.86	0.31
2009-2013	0.74	0.87	0.26
2014-2018	0.81	0.71	0.19

Table 2.7. CMAQ benchmarking statistical summary of ozone simulation for nine SCAQMD air monitoring stations. Units are in ppm.

Stations	Max	Min	Mean	Median	Q1	Q3	RMSE	rRMSE	MAE	MBE
Anaheim	0.107	0	0.037	0.041	0.032	0.053	0.014	0.391	0.012	0.009
Azusa	0.114	0	0.042	0.046	0.033	0.060	0.017	0.402	0.014	0.008
Crestline	0.12	0.004	0.060	0.062	0.055	0.069	0.023	0.499	0.020	0.006
Fontana	0.116	0	0.045	0.050	0.036	0.065	0.022	0.567	0.019	0.016
LA	0.101	0	0.033	0.037	0.023	0.051	0.015	0.757	0.013	0.006
Pasadena	0.111	0	0.043	0.046	0.035	0.058	0.018	0.433	0.016	0.012
Redlands	00.108	0	0.053	0.056	0.046	0.067	0.021	0.413	0.018	0.008
Rubidoux	0.118	0	0.045	0.049	0.035	0.065	0.018	0.762	0.015	0.009
SB	0.119	0	0.047	0.052	0.038	0.066	0.020	0.449	0.017	0.007

Figures

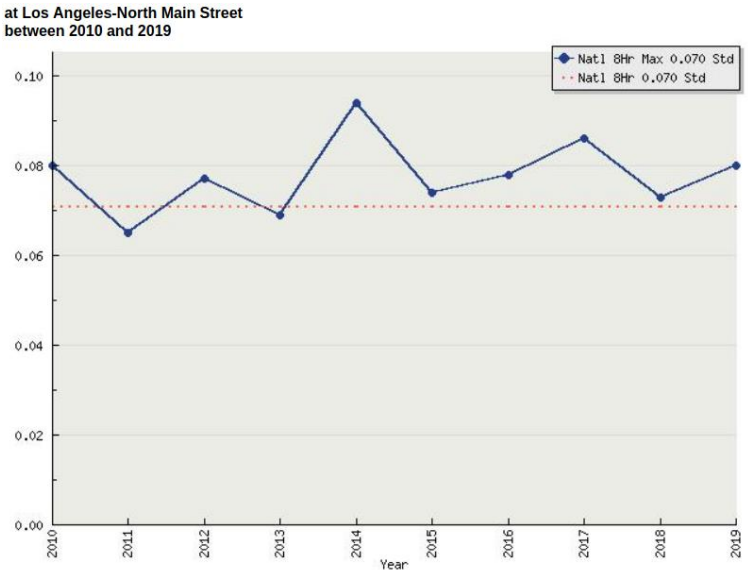


Figure 2.1. 8-hour ozone design value concentrations in Los Angeles. The red dash line is the National Ambient Air Quality Standard (NAAQS) for 8-hour ozone (0.070 ppm, 2015). Source: California Air Resources Board.

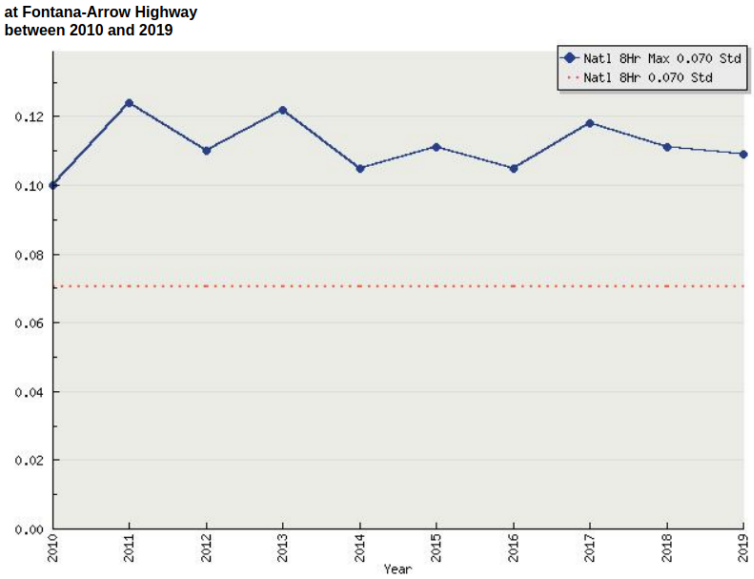


Figure 2.2. 8-hour ozone design value concentrations in Fontana. The red dash line is the National Ambient Air Quality Standard (NAAQS) for 8-hour ozone (0.070 ppm, 2015). Source: California Air Resources Board.

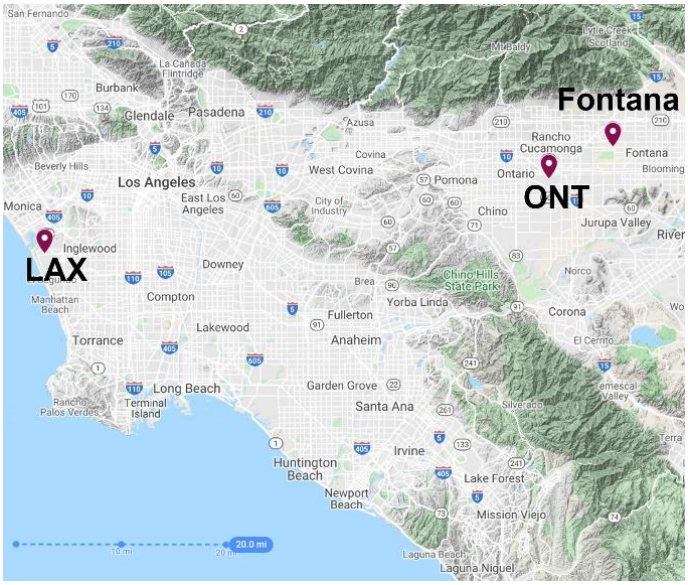


Figure 2.3. Site location map highlighting the Los Angeles (LAX) and Ontario (ONT) International Airports and the Fontana air monitoring site.

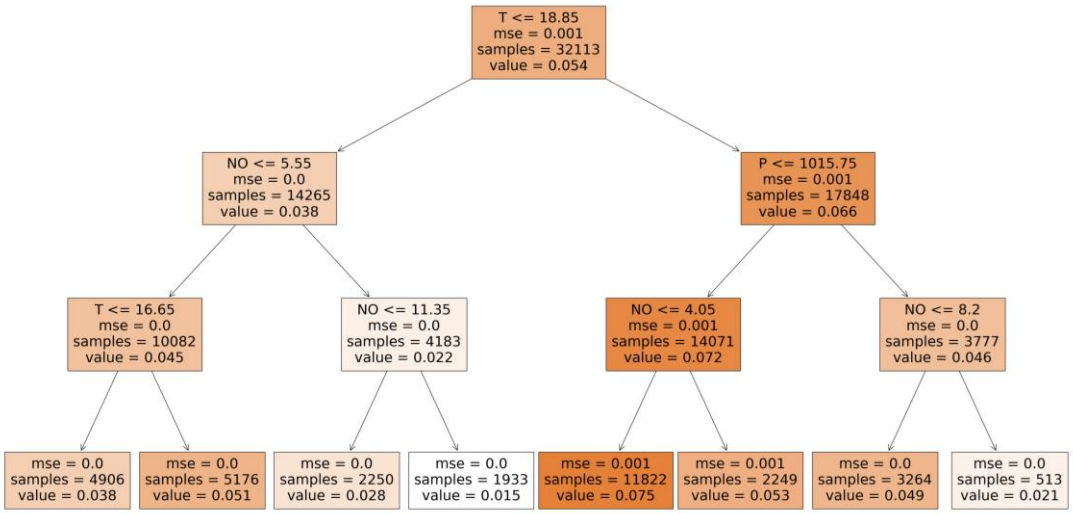


Figure 2.4. Three node decision trees based on air quality and meteorological input from 2014-2018. The meteorology data is from LAX, and air quality data is from Fontana. The predictions were made based on 12:00 noon to 5:00 PM training data.

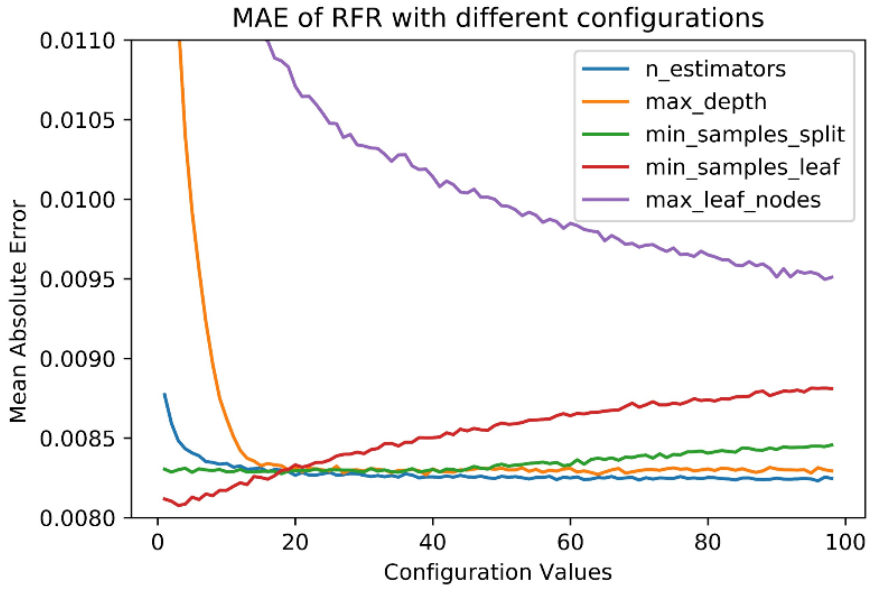


Figure 2.5. RFR mean absolute error (MAE) with different hyperparameter values. The value of the tuning parameter was varied from 1 to 100 while keeping others constant. MAE is in units of ppm.

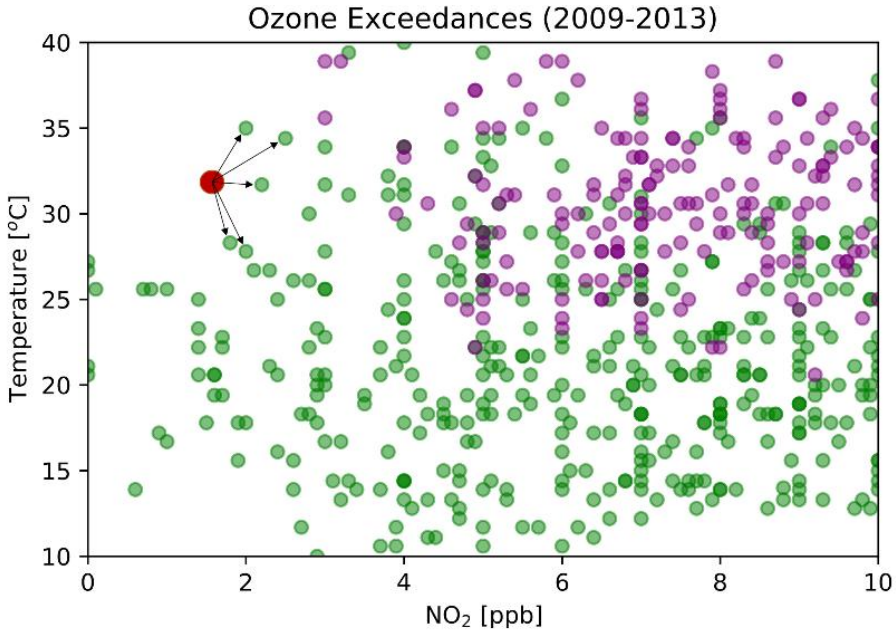


Figure 2.6. Classification in two dimensions, coded as a binary variable (green = non-exceedances, purple = exceedances). The predicted class of the red point is chosen by the majority vote amongst the 5 nearest neighbors.

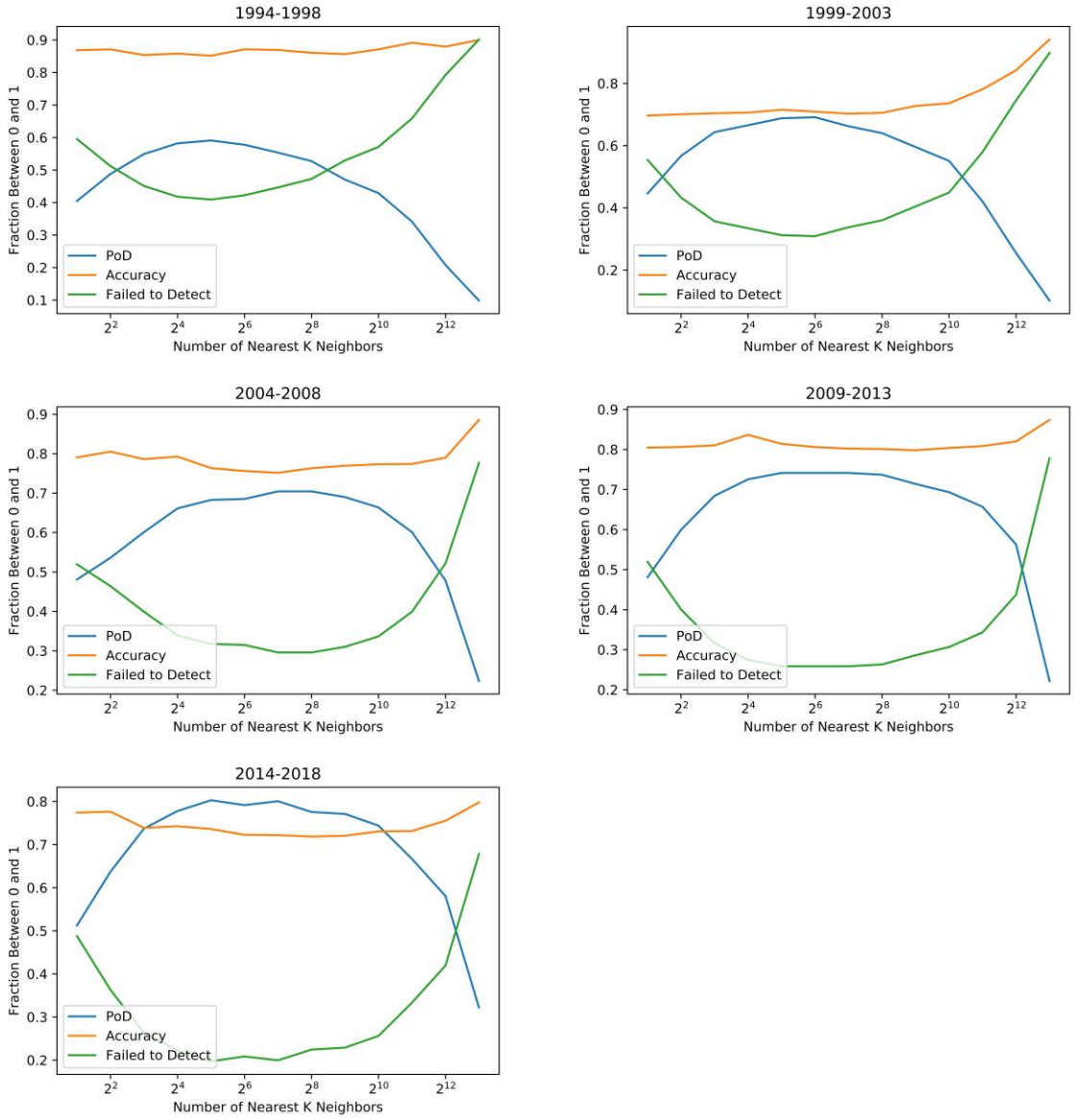


Figure 2.7. Testing the performance of K-NN by varying the number of nearest neighbors while keeping other parameters constant.

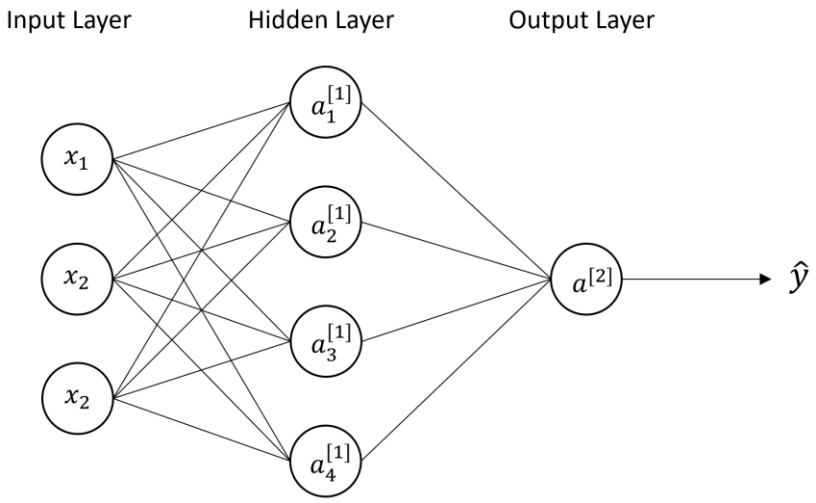


Figure 2.8. A fully connected 2-layer neural network diagram with inputs x, four perceptrons in the hidden layer, and one in the output layer.

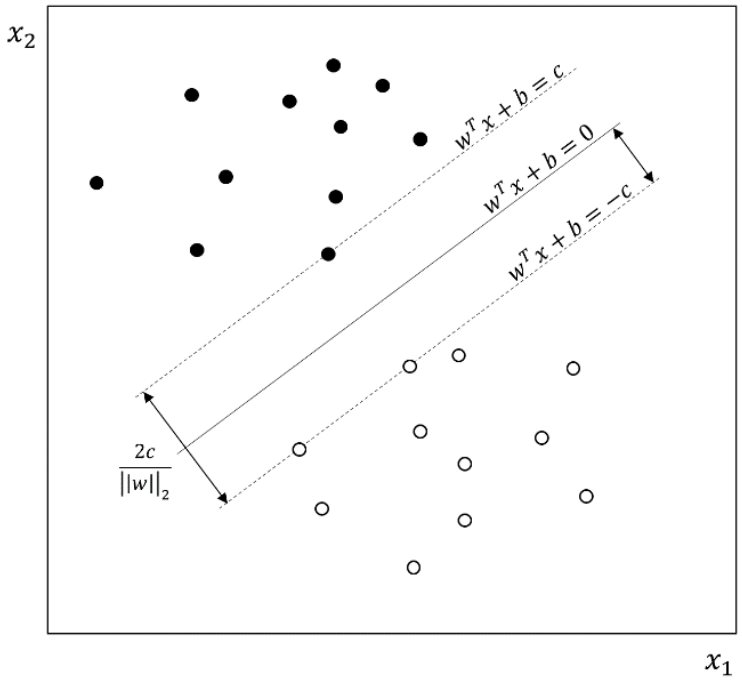


Figure 2.9. Support vector machine separating black dots and white dots. The separating hyperplane (solid line) is in the center of the two supporting hyperplanes for which the margin is maximized.

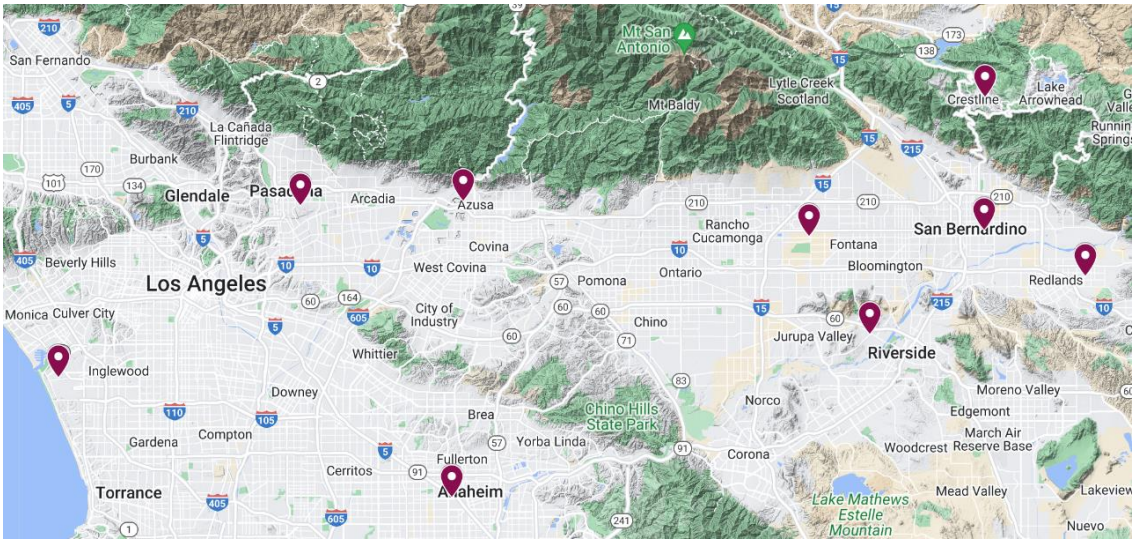


Figure 2.10. Nine evaluation sites from SCAQMD. From left to right, LAX, Pasadena, Anaheim, Azusa, Fontana, Riverside, San Bernardino, Crestline, and Redlands

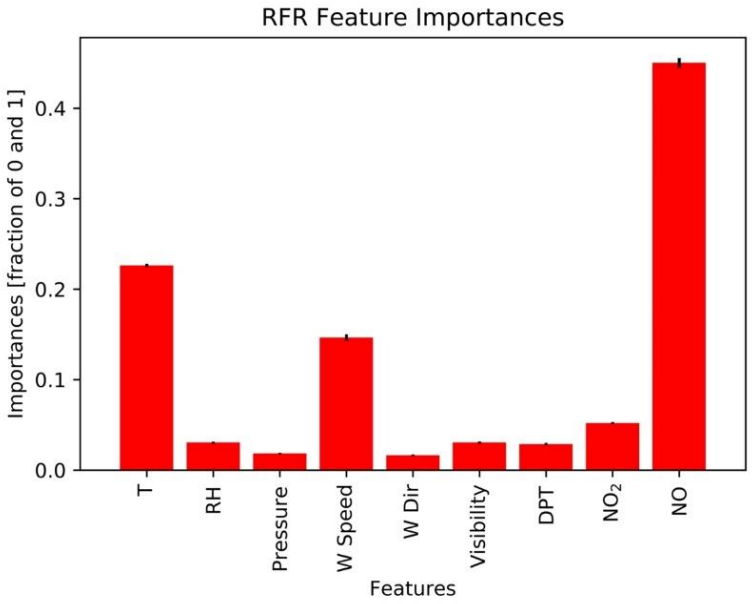


Figure 2.11. Feature importance generated from the RFR model. NO, T, and wind speed are the three most important features.

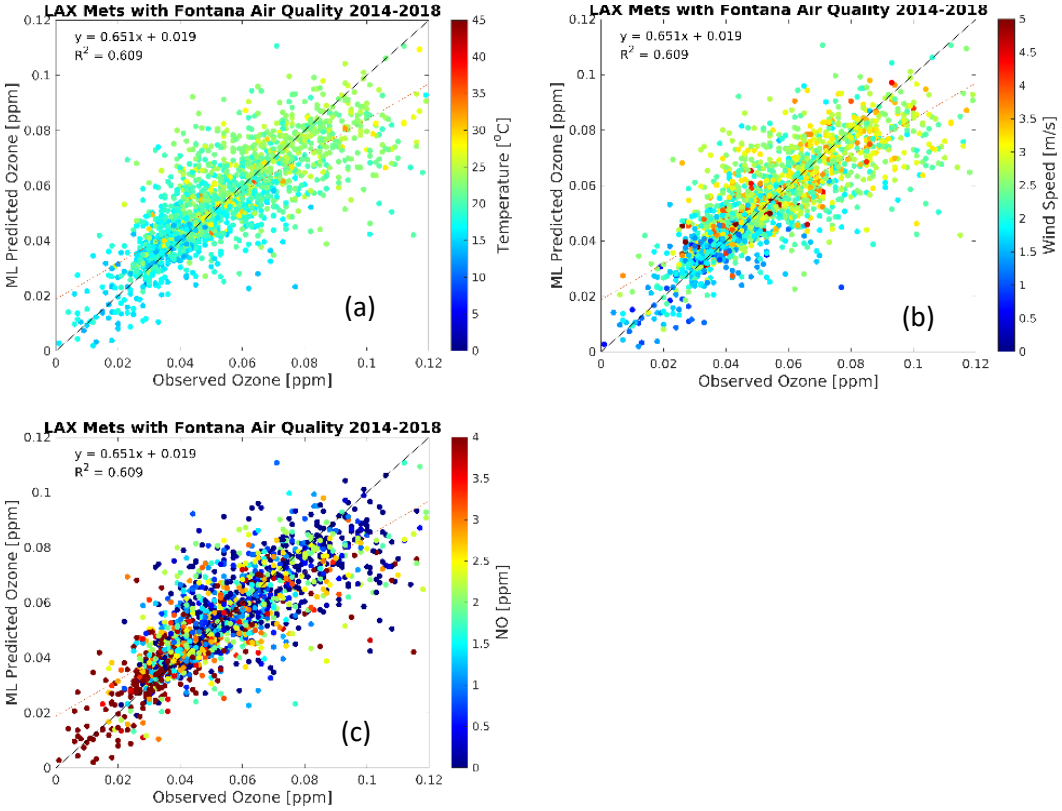


Figure 2.12. Observational O_3 (x-axis) and RFR predictions (y-axis) for Fontana air quality and meteorology from the LAX international airport monitoring station. The plots are for the most recent five-year increment from 2014-2018. The color bars show temperature (a), wind speed (b), and NO (c). Plots for other periods are provided in the SI.

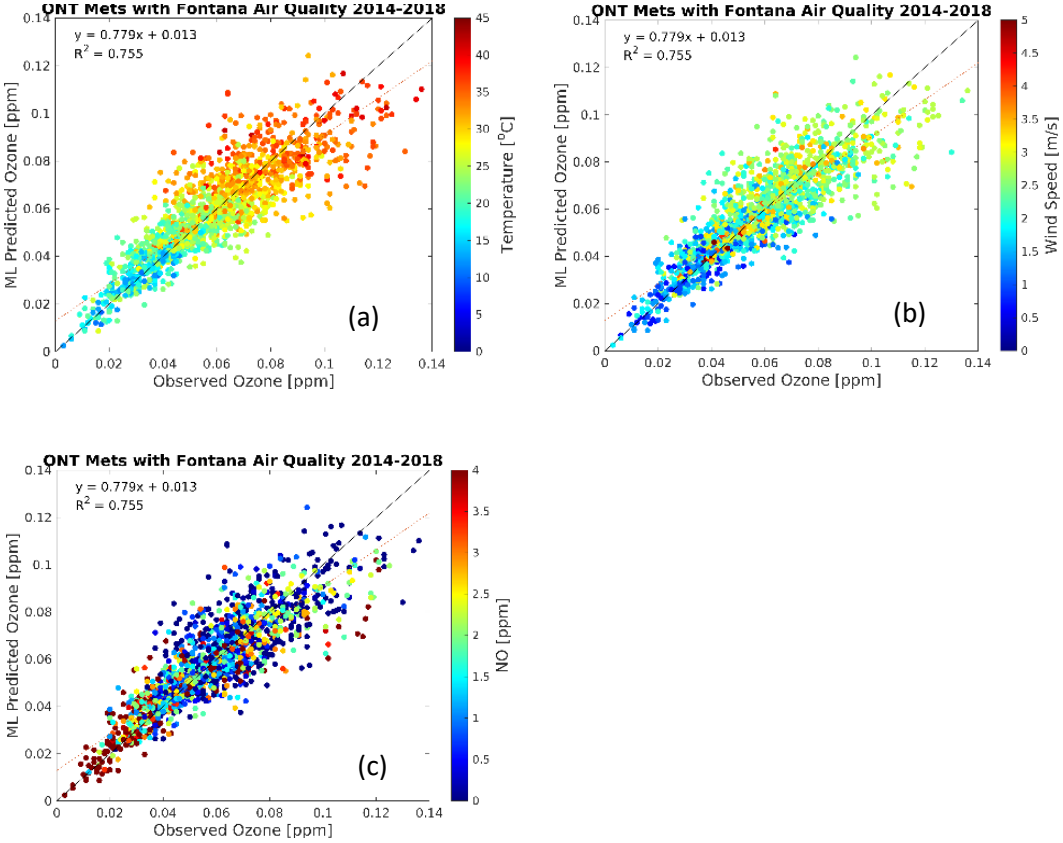


Figure 2.13. Observational O_3 (x-axis) and RFR predictions (y-axis) for Fontana air quality and meteorology from the ONT international airport monitoring station. The plots are for the most recent five-year increment from 2014-2018. The color bars show temperature (a), wind speed (b), and NO (c). Plots for other periods are provided in the SI.

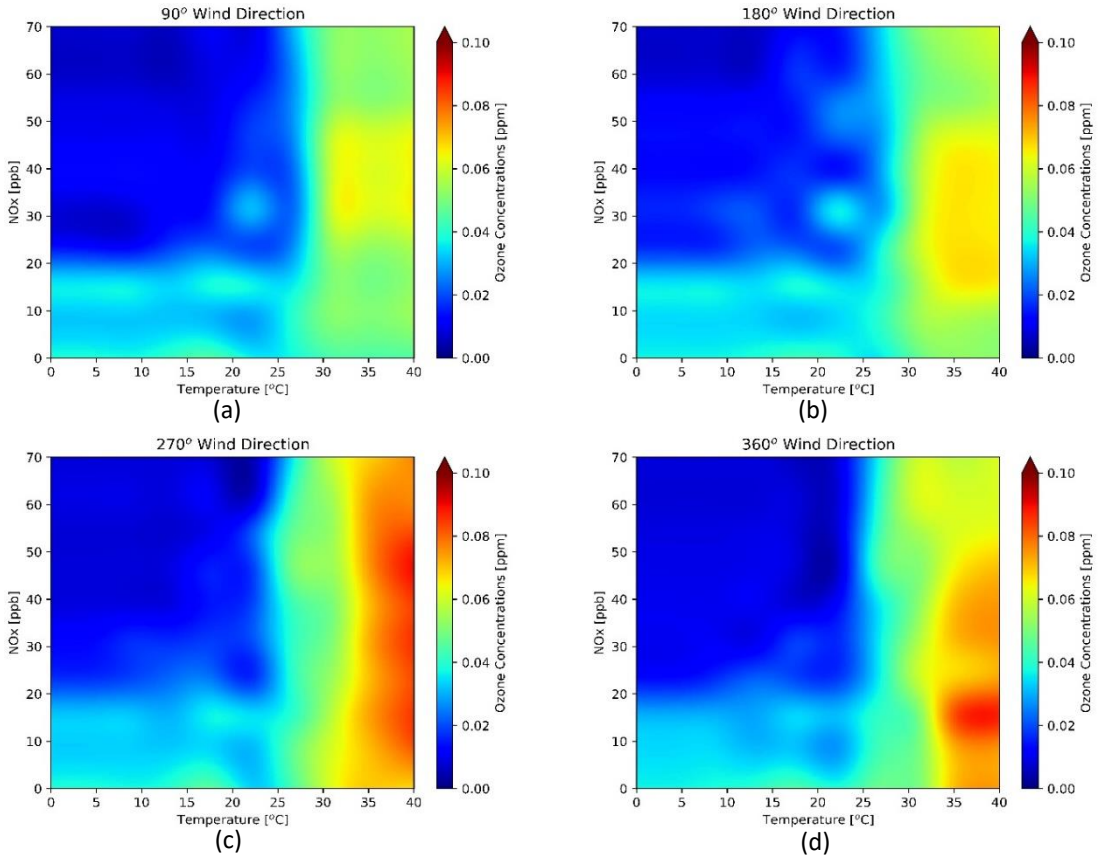


Figure 2.14. Contour plots generated by the RFR model trained on ONT meteorology and Fontana air quality at constant wind speed (9 m/s), visibility (16000 m), dynamic pressure, dynamic relative humidity, and for four discrete wind directions: (a) 90°, (b) 180°, (c) 270°, and (d) 360°.

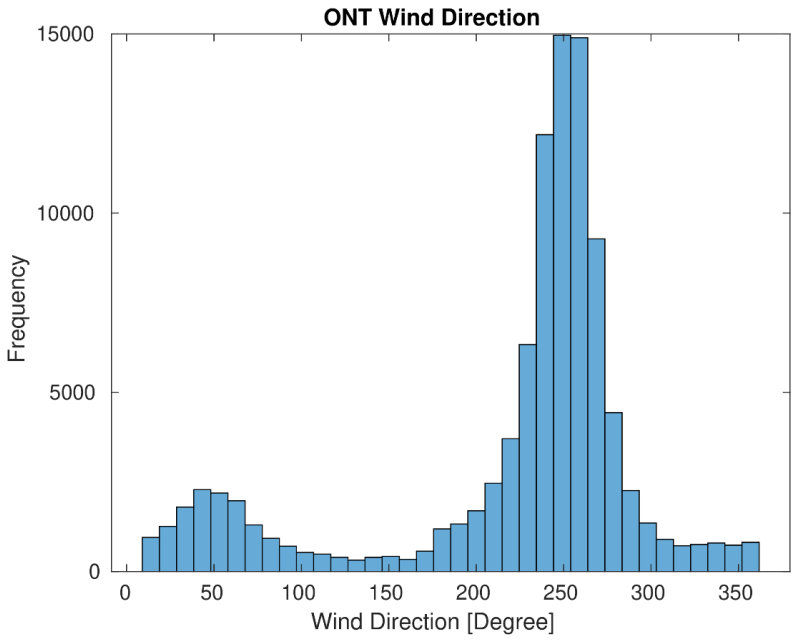


Figure 2.15. Wind direction in Ontario international airport. 25% of wind directions are from 254-273 degrees, and 64% of wind directions are from 225-273 degrees.

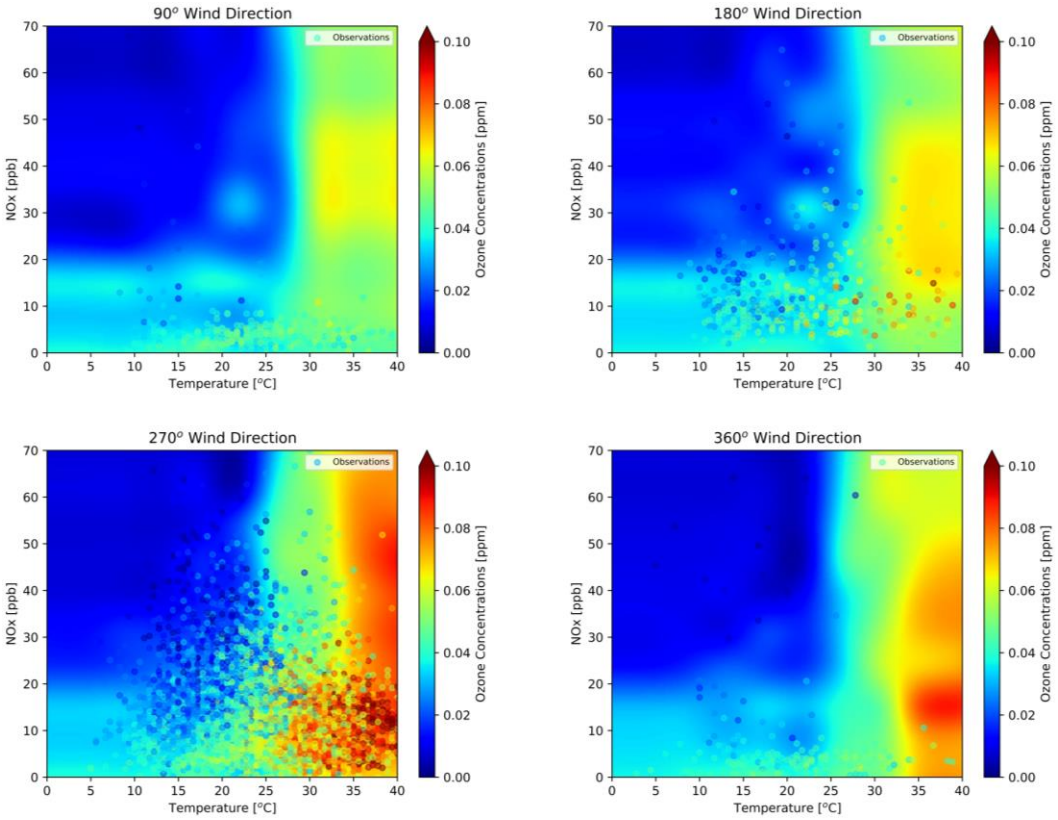


Figure 2.16. Contour plots generated by the RFR model trained on ONT meteorology and Fontana air quality at constant wind speed (9 m/s), visibility (16000 m), dynamic pressure, dynamic relative humidity, and at four discrete wind direction levels (90, 180, 270, 360). The dots are observational data plotted on the top of the contours.

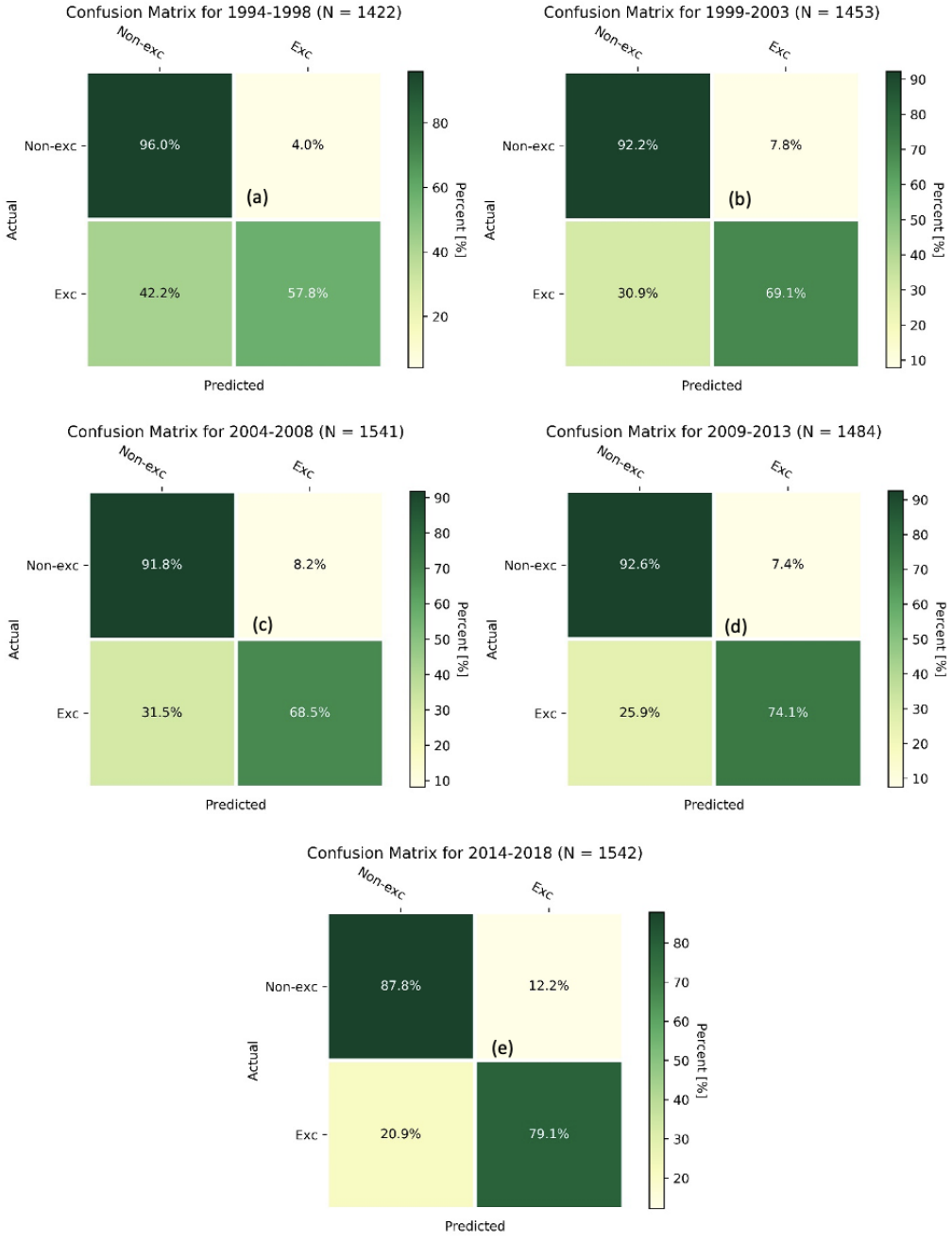


Figure 2.17. Confusion matrices for ozone exceedances evaluated for the K-NN model for the periods (a) 1994-1998, (b) 1999-2003, (c) 2004-2008, (d) 2009-2013, and (e) 2014-2018. N is the total number of valid data points.

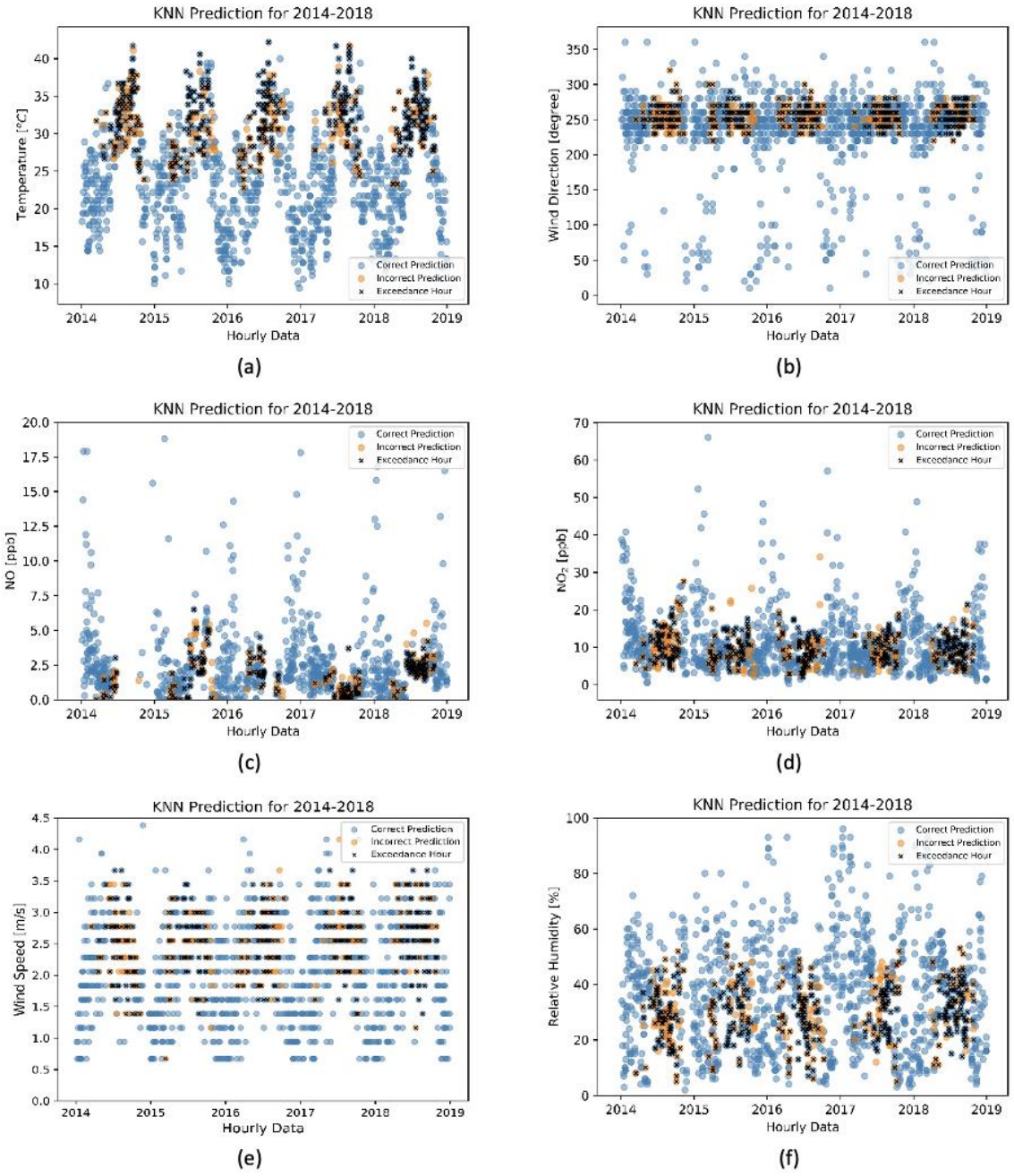


Figure 2.18. Non-exceedance and exceedance hours for observed input variables: (a) temperature in ˚C, (b) wind direction in degrees, (c) NO in ppb, (d) NO_2 in ppb, € wind speed in m/s, and relative humidity in %. Predictions were made using K-NN for the years 2014-2018 for Fontana using ONT meteorology. Hourly data from 12 pm to 5 pm are highlighted to reflect the peak ozone period.

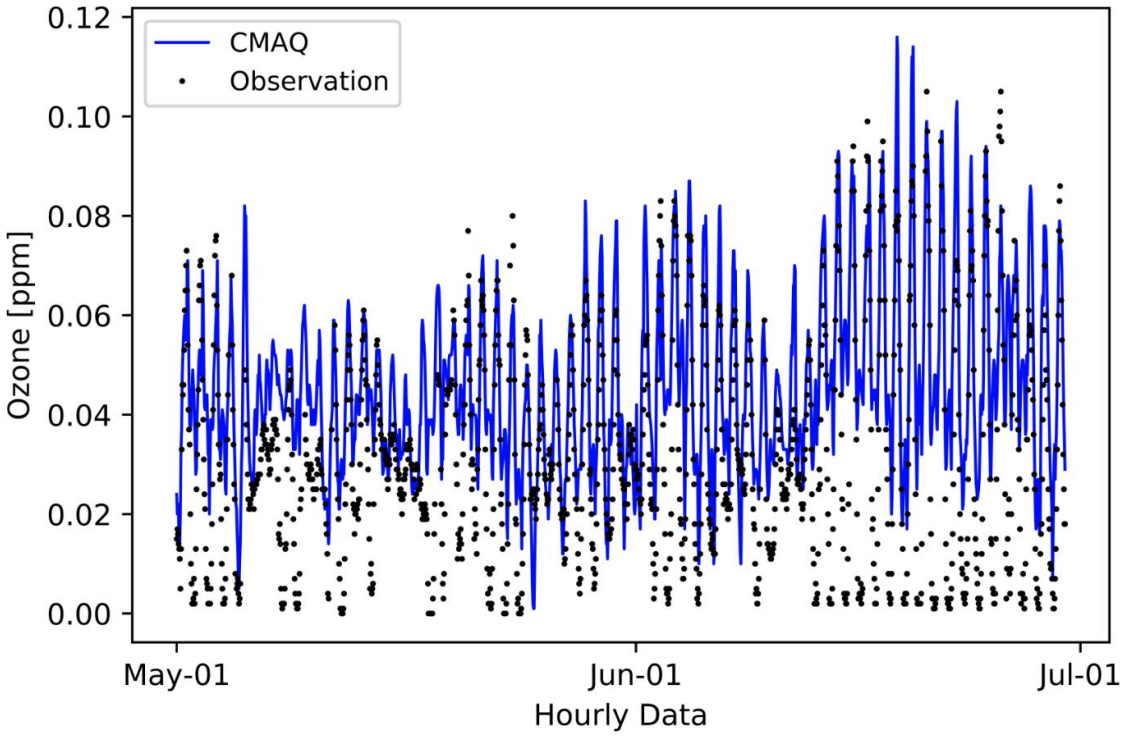


Figure 2.19. Time series of ozone concentration in Fontana, CA. The blue line is CMAQ simulation results, and the dots are observational data.

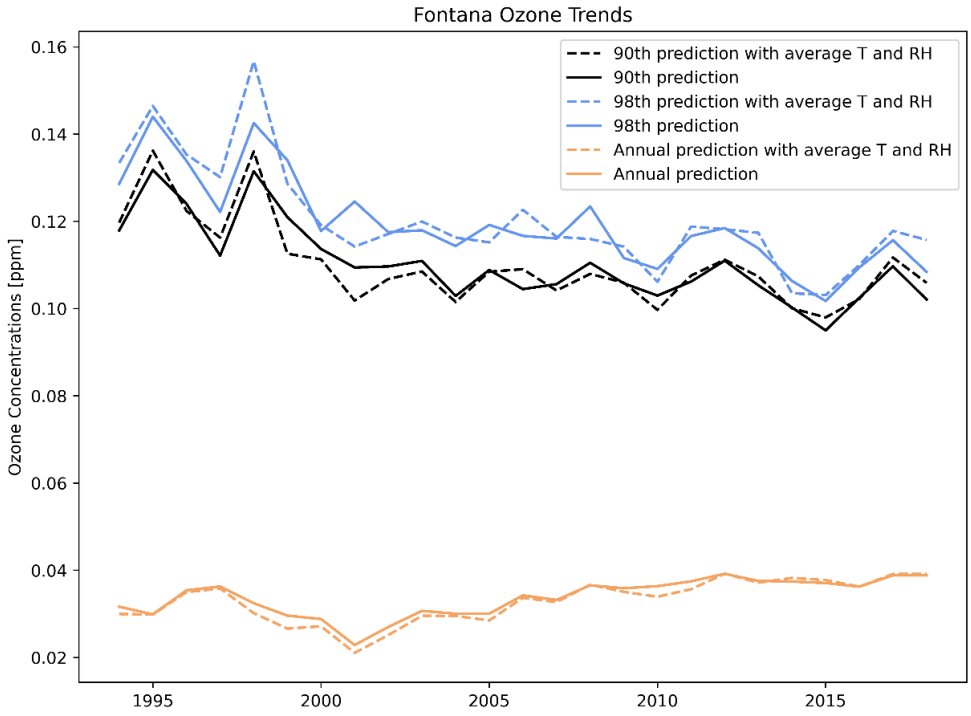


Figure 2.20. Fontana trends for 90th (black), 98th (blue) percentile, and annual average (orange) ozone concentration. The dashed lines were predicted with hourly average temperature and RH from 1994 to 2018. The solid lines were predicted with actual values.

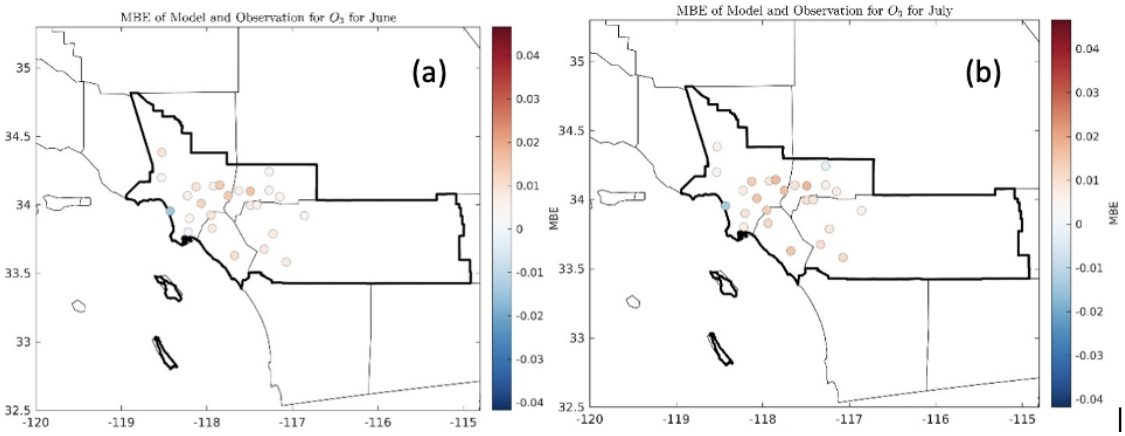


Figure 2.21. Monthly mean bias error for ozone for 25 air monitoring sites in SoCAB; (a) June 2017, (b) July 2017.

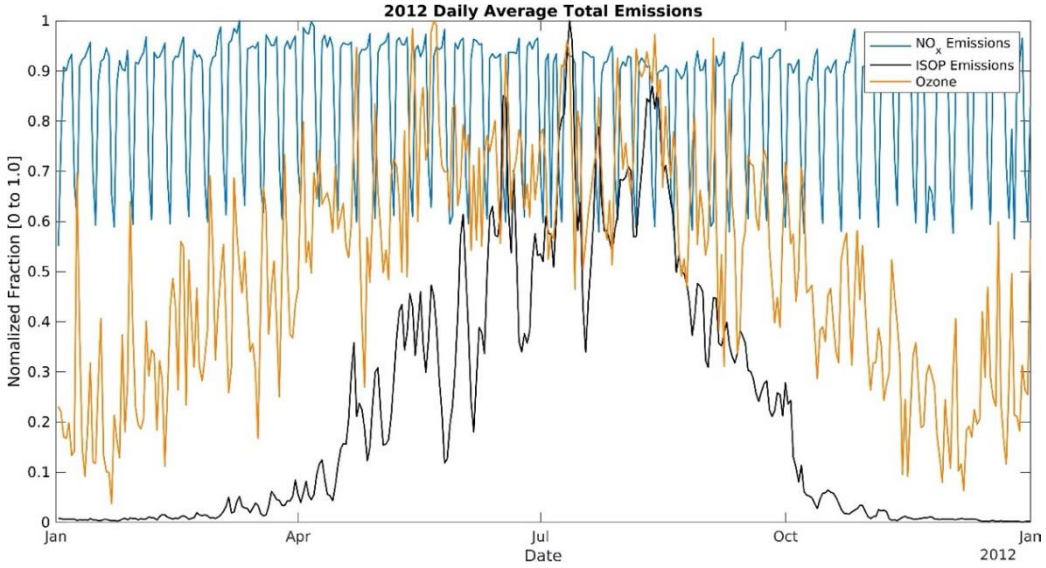


Figure 2.22. Daily average NO_x and isoprene (ISOP) emissions over the model domain normalized by the maximum value in the domain. The periodic oscillation of NO_x emissions (blue line) is due to weekday/weekend behavior. The black line is the biogenic isoprene emissions in the entire domain. NO_x and ISOP emissions were extracted from gridded SCAQMD emissions. Twenty-four-hour ozone averages were sampled from the Fontana air monitoring station.

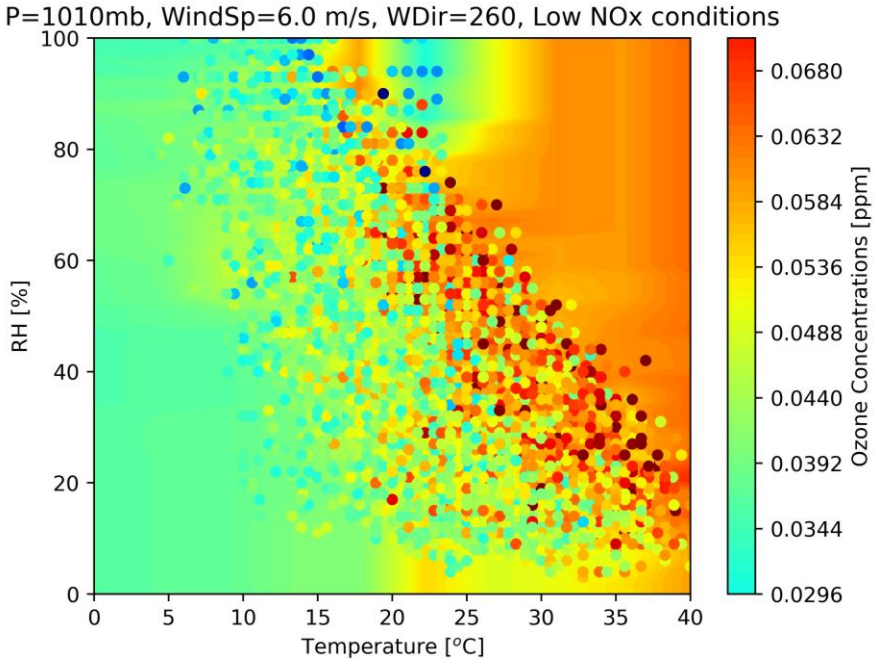


Figure 2.23. Contour plots generated by the RFR model trained on ONT meteorology and Fontana air quality at constant wind speed (6.0 m/s), visibility (16000 m), wind direction from 260 degree, and 1010 mb pressure. The dots are observational data plotted on the top of the contours.

Part 2

Machine Learning with Spatial Interpolation Techniques for Constructing 2-Dimensional Ozone Concentrations in Southern California during the COVID-19 Shutdown

Abstract

In this study, machine learning and geospatial interpolations is combined to create a two-dimensional high-resolution ozone concentration field over the South Coast Air Basin for the entire year of 2020. Three spatial interpolation methods (bicubic, IDW, and ordinary kriging) were employed. The predicted ozone concentration fields were constructed using 15 building sites, and random forest regression was employed to test predictability of 2020 data based on input data from past years. Spatially interpolated ozone concentrations were evaluated at twelve sites that were independent of the actual spatial interpolations to find the most suitable method for SoCAB. Ordinary kriging interpolation had the best performance overall for 2020: concentrations were overestimated for Anaheim, Compton, LA North Main Street, LAX, Rubidoux, and San Gabriel sites and underestimated for Banning, Glendora, Lake Elsinore, and Mira Loma sites. The model performance improved from the West to the East, exhibiting better predictions for inland sites. The model is best at interpolating ozone concentrations inside the sampling region (bounded by the building sites), with R^2 ranging from 0.56 to 0.85 for those sites, as prediction deficiencies occurred at the periphery of the sampling region, with the lowest R^2 of 0.39 for Winchester. All the interpolation methods poorly predicted and underestimated ozone concentrations in Crestline during summer (up to 19 ppb). Poor performance for Crestline indicates that the site has a distribution air pollution level independent from all other sites. Therefore, historical data from coastal and inland sites should not be used to predict ozone in Crestline using data-driven spatial

interpolation approaches. The study demonstrates the utility of machine learning and geospatial techniques for evaluating air pollution levels during anomalous periods.

Introduction

In the atmosphere, the non-linear relationship between nitrogen oxides (NO_X), volatile organic compounds (VOCs), and ozone is complex. In the United States, the COVID-19 pandemic and the ensuing shutdown presented an unintentionally optimal period to observe, revise, and improve our existing air quality models and observe the sensitivity of the NO_X-VOC-ozone relationship in real time. In California, the pandemic shutdown began on March 16, 2020, resulting in a significant drop in traffic volume. In Los Angeles and Ventura Counties, there was approximately a 30% decrease in vehicle miles traveled (VMT) on weekdays and up to a 40% decrease on weekends in 2020 (Caltrans, 2023). This unusual event temporarily changed the conventional distribution of primary and secondary air pollutants in the South Coast Air Basin (SoCAB). NOx and VOC emissions declined with the reduction in traffic flow (Jiang et al., 2021). As a result, a drop in ozone concentrations was expected in Southern California. Several studies were published regarding the pandemic that investigated the effects of the COVID-19 shutdown on air pollutants. Jiang et al., used WRF-Chem to simulate the major air pollutants under two scenarios (i.e., before lockdown and during lockdown) and found an increase in ozone in urban areas due to emission reduction during the lockdown (Jiang et al., 2021). The COVID-19 shutdown provided an estimation of the impacts of future large-scale emission reductions strategies on ozone formation in SoCAB (Ivey et al., 2020).

Over the past several decades, ozone levels in Southern California significantly decreased as a result of emissions control programs implemented by the South Coast Air Quality

Management District (SCAQMD), thereby reducing emissions from mobile sources and shifting to renewable energy sources (Lurmann et al., 2015; South Coast Air Quality Management District, 2017). However, during the past decade, ozone concentrations in the SoCAB have slightly plateaued despite further emissions reductions (Figure 2.24) (Do et al., 2023).

This paper focuses on the performance of deterministic and statistical models under rapid changes in emissions and meteorological conditions. Chemical transport models (CTM) are conventionally used for air quality research and regulatory purposes. The Community Multiscale Air Quality (CMAQ) modeling system, developed by the U.S. Environmental Protection Agency (EPA), is well-known for multi-day air quality simulations to estimate air pollutant concentrations with prescribed emissions and meteorology inputs (Ooka et al., 2011; Rao et al., 1996; Wong et al., 2012). From the model outputs, scientists and regulators can better predict the interactions between future emissions, meteorology, and air pollutants to strengthen recommendations for emissions control programs. Zhu et al. used CMAQ to investigate the sensitivity of ozone and particulate matter less than 2.5 microns ($PM_{2.5}$) to incremental changes in volatile organic compounds (VOC) by updating the VOC emissions from recent literature, and simulated maximum daily 8-hour ozone concentrations increased by 17.4 ppb and 15.6 ppb in summer and winter, respectively (Zhu et al., 2019). With a similar approach, Karamchandani et al., found that near-recent regulatory modeling for SoCAB generally underestimated the response of ozone design values to the changes in precursor emissions (Karamchandani et al., 2017).

Recently, machine learning (ML) as an alternative modeling approach has attracted more attention from air quality researchers. Although ML and chemical transport models have a similar goal to accurately predict air pollution, ML heavily depends on the quality and quantity of data

available. Conversely, CTMs are based on first principles equations and are initiated with interpolated observation data, hence avoiding most obstacles introduced by data missingness in observations. In contrast with CTMs, which produce larger scale, spatially resolved outputs, ML only provides predictions strictly at trained locations when used for ambient air quality applications. SCAQMD operates 38 air monitoring stations in Southern California over an area of approximately 10,743 square miles, including SoCAB, portions of the Salton Sea Air Basin, and Mojave Desert Air Basin, with an average of 283 square miles per monitoring station (Miyasato et al., 2016; South Coast Air Quality Management District, 2017). Due to the relative sparseness of monitoring stations and locality of air pollutants, using air monitoring stations to represent spatially-varying air quality over a large area may result in incorrect information (Apte et al., 2017). To overcome this limitation when high-resolution measurements are not available, researchers opt to use spatial interpolation methods (e.g., nearest neighbors, linear or polynomial interpolation, continuous natural neighbor interpolation, etc.) (Joseph et al., 2013). Yu et al., evaluated 14 unique spatial modeling methods for eight air pollutants in Atlanta, Georgia for developing spatiotemporal air pollutant concentrations fields (Yu et al., 2018). Wong et al., assessed four spatial interpolation methods (spatial averaging, nearest neighbor, inverse distance weighting (IDW), and kriging) to estimate ozone and PM_{10} air concentrations (Wong et al., 2004). In this paper, three spatial interpolation techniques are compared to the CMAQ model and evaluated biases related to COVID-19 lockdown anomalies.

Study Area and Datasets

This study targeted the Southern California region, including Los Angeles, Orange County, Riverside, and San Bernardino counties. The region has been historically challenged with poor air

quality, with especially higher ozone concentrations than the rest of the United States. The coastal areas tend to have higher relative humidity (RH) and lower temperatures than inland Southern California. Since the turn of the century, SoCAB has been designated as a nonattainment area for the 1997 8-hour ozone standard (80 ppb), with design values for ozone well above the 2015 standard of 70 ppb (Figure 2.24). In 2019, the maximum daily 8-hour average (MDA8) ozone concentration in SoCAB 108 ppb at the design value location with a classification of "extreme" (Redlands, California) (California Air Resources Board, 2023).

Model Input Data

The input meteorological data for the CMAQ simulation were generated using the Weather Research and Forecasting (WRF) model. WRF was initiated using data from the North American Mesoscale (NAM) Forecast System integrated with high-resolution sea surface temperature (SST) from the Group for High Resolution Sea Surface Temperature. The WRF Objective Analysis program was used to improve the meteorological simulation, and this step blends observed surface and upper air observations with background WRF fields. The surface and upper air observations are sourced from ds461 and ds351 datasets via the National Center for Atmospheric Research's Research Data Archive, respectively (Wang et al., 2017).

Gridded 4 km emissions were projected from 2019 for the year 2020 using a two-step adjustment to account for changes due to the COVID-19 (Zhu et al., 2023). In the first step, a linear projection factor (Eq. 1) was applied to 2019 gridded emissions based on SCAQMD basin-wide, total annual emissions spanning from 2012 to 2034, where the District's future projections began at year 2020). The correction factor was calculated for seven air pollutant groups (total organic gases, reactive organic gases, CO, NO_X, SO_X, NH_3, PM).

$$Linear\ projection\ factor = \frac{2020\ emis - 2019\ emis}{2019\ emis} \quad (1)$$

The second step accounted for traffic reductions due to the COVID-19 lockdown, and reductions were highest from March to May 2020 then slowly but not fully rebounding to pre-lockdown levels toward the end of 2020 (Caltrans, 2023). SCAQMD basin-wide projections understandably did not reflect the decrease in mobile source emissions due to traffic reductions. Moreover, weekly traffic metrics in 2020 were acquired for the total flow, flow change, and speed change at 2991 locations in Southern California (Tanvir et al., 2023). Since the traffic data were not evenly distributed over the study domain, k-nearest neighbors (k-NN) was used to obtain the traffic data for grid cells (locations) that had no more than five reported data points (k value ≤ 5). For the grid cells with more than five reported data points, traffic volume was normalized and then averaged the normalized data.

Machine Learning Inputs

Meteorological and air quality data from 15 air monitoring sites in SoCAB were used (Figure 2.25). Hourly meteorological and air quality data used for ML training and validation were obtained from the Air Quality System (AQS) data mart (https://aqs.epa.gov/aqsweb/airdata/download_files.html#Raw, last access Jan 19th, 2023). Data was checked to ensure the hourly data was available for all training features. If there was a missing data point for one of the features, the invalid hour and all corresponding features were removed. The date range of the model training data was 2009-2010 and 2016-2019 for all 15 sites (Figure 2.25). The period from 2011-2015 was not included in the models due to the limited availability of wind direction and wind speed at the sites. The 2020 data was used for model testing and evaluation (Table 2.8).

Methods

CMAQ Modeling

In this study, the performance of both CMAQ and ML was compared with spatial interpolations of ozone concentrations in SoCAB for the year 2020. The CMAQ simulation covered three distinct periods to study the impact of COVID-19 lockdown on air pollutant concentrations: pre-lockdown (Jan 1st to Mar 18th), lockdown (Mar 16th to May 15th), and post-lockdown (after May 16th) periods. Meteorological modeling was carried out using the Weather Research and Forecasting (WRF) model version 3.9 with 4 km horizontal grid spacing, 11 vertical layers for the finest domain (10 layers near the surface), and 156 x 102 grid cells (Figure 2.26). There were two parent domains with coarser horizontal grid spacing (36 km and 12 km for domain 1 and domain 2, respectively). WRF configurations were optimized for SoCAB, and they included use of USGS land use, the thermal diffusion surface layer scheme, and the Yonsei University planetary boundary layer scheme (Hong et al., 2006; Huang et al., 2014). The CMAQ simulation used the modified 2020 emissions and previously described WRF simulations as inputs. The choice of chemical mechanism was SAPRC07tc_ae6_aq, i.e., SAPRC07tc photochemical mechanism, aerosol module 6, and aqueous chemistry (Byun & Schere, 2006; Carter, 2010).

Machine Learning

In a preceding study, multiple ML algorithms were tested to obtain a better method that resulted in the highest prediction accuracy for ozone concentrations in the SoCAB. Those included neural network, support vector machine, k-nearest neighbors, and random forest (Do et al., 2023). Here, random forest regression (RFR) was selected, as RFR is the most suitable ML algorithm for predicting ozone concentrations in SoCAB (Do et al., 2023). For reference, RFR is a supervised

learning algorithm with a tree-based ensemble method, i.e., a combination of multiple decision trees trained on an independent collection of input variables. In this application, RFR selected a random collection of features from the six input features for each decision tree to reduce bias, and the output of RFR is the average result from all decision trees (Rodriguez-Galiano et al., 2015; Zhang and Ma, 2012).

In this study, six training features were selected to predict ozone concentrations, which included two air quality features (NO and NO_2) and four meteorological features (temperature, relative humidity, wind speed, and wind direction). The two air quality features are directly related to ozone formation in the troposphere. Ozone undergoes the photolytic cycle during the day and is removed by NOx during nighttime (Brune, 2001; Liu et al., 1980; Trousdell et al., 2019). The four meteorological features were well studied in our previous work and were shown as the most important features to capture the variability in annual ozone, especially in SoCAB (Camalier et al., 2007; Gao et al., 2022; Jaffe, 2020).

The scikit-learn 0.22 library supported by the Python programming language was used to train the RFR model. Again, the input features are NO_2, NO, temperature, relative humidity, wind speed, and wind direction, and the label is ozone. The algorithm was tuned by varying the number of decision trees, the depth of the tree, sample split, and the sample leaf to obtain the best prediction accuracy. The same model tuning approached described in Do et al.was used (Table 2.9) (Do et al., 2023).

Spatial Interpolation

To generate a two-dimensional ozone concentration map, the RFR model was carried out to obtain the ozone concentrations at each air monitoring location (15 sites), which served as the

model building sites. In other words, a pointwise ML algorithm was applied to predict ozone concentration at each trained location. Next, the output was spatially interpolated over the target Southern California region. Three different spatial interpolation methods (ordinary kriging, inverse distance weighting (IDW), and bicubic interpolation) were applied and comparatively evaluated the performance of each method. Each interpolation approach is described below.

Ordinary kriging was applied to interpolate the ozone concentration at the 10 km resolution over the study area. Ordinary kriging is a well-known spatial interpolation method developed by Danie G. Krige. Generally, kriging predicts the values for unknown locations by performing a series of linear combinations of values at known locations. Equation 1 expresses the generic form of the estimator to predict the optimum value $Z^*$of an unknown location by combining the known values Z_i with their weights λ_i (Oliver and Webster, 1990). The variance σ^2 can be written as an optimization problem (Eq. 2) that can be solved using the Lagrange multiplier μ (Eq. 3).

$$Z^*(u) = \sum_{i=1}^{n} \lambda_i Z(u_i) \tag{1}$$

$$\sigma^2(u) = Var[Z(u) - Z^*(u)] = -\sum_{j=1}^{n}\sum_{i=1}^{n} \lambda_j \lambda_i \gamma(u_i - u_j) + 2\sum_{i=1}^{n} \lambda_i \gamma(u_i - u) \tag{2}$$

$$\sum_{j=1}^{n} \lambda_j (u_i - u_j) + \mu = \gamma(u_i - u) \tag{3}$$

and

$$\sum_{j=1}^{n} \lambda_j = 1 \tag{4}$$

μ is the Lagrange multiplier, u_i and u_j are the distance of known locations from unknown locations u, γ is the variogram, and $i = 1, \dots, n$. Equations 1 and 2 are called the kriging system, and λ is the kriging weight. The values for λ_i and the optimum value Z^* are obtained by solving the kriging system and Equation 1 (Yamamoto, 2000).

Bicubic interpolation is another method for interpolating data points on a two-dimensional grid. The interpolated surface can be written in terms of two variables (Eq. 5). The polynomial p consists of sixteen coefficients a_{ij} that are solved with sixteen boundary conditions (i.e., $(x = 0, y = 0), (x = 1, y = 0), (x = 0, y = 1), (x = 1, y = 1)$) and its derivatives with respect to x, y, and xy (Seiler and Seiler, 1989).

$$p(x, y) = \sum_{i=0}^{3} \sum_{j=0}^{3} a_{ij} x^i y^j \tag{5}$$

The IDW interpolation method accounts for the distances between the interpolated points and the measured locations. The assumption for IDW is that points close to each other are more alike and have more significant influence than those farther apart. Thus, the nearest measured values have greater weights assigned. Equation 6 shows that the predicted value $Z(x)$ is inversely proportional to the distance between the measured and interpolated points $d(x, x_i)$.

$$Z(x) = \frac{\sum_{i=1}^{n} \frac{Z_i}{d(x, x_i)^p}}{\sum_{i=1}^{n} \frac{1}{d(x, x_i)^p}} \tag{6}$$

where $Z(x)$ is the predicted value, d is the distance, x is the unknown point, x_i is the known location, Z_i is the value of a known location, and p is the power (Bartier and Keller, 1996).

Model Evaluation and Discussion

Figure 2.27 shows a snapshot of the ozone concentrations over the interpolation region at 4:00 PM on June 22, 2020 (highest ozone episode of the day), using ordinary kriging. The colored dots with a white border are the actual values at the evaluation sites, and those without a white border are the RFR predicted values for training sites. The model successfully reconstructed the spatial trends in the region where the lowest ozone levels were in the southwest (coastal) and the highest were in the east (inland), and there was good agreement with the actual ozone concentrations. Figure 2.28 and Figure 2.29 show the heatmap for bicubic and IDW interpolation for the same timestamp. Although all interpolation methods predicted the lowest ozone concentrations in the Southwest, the highest ozone concentrations were predicted in the Northeast of the study region for bicubic and in the North for IDW. The concentration gradient increased from south to north for bicubic and IDW, but from west to east for ordinary kriging.

The performance of the models was evaluated based on commonly-used statistical metrics: mean bias (MB), correlation coefficient, root mean square error, and R^2 (equations listed in SI). The models were evaluated based on data from 27 air monitoring stations in SoCAB, of which 15 sites were used to evaluate the training sites and the other 12 sites were used to evaluate the performance of the three interpolation methods at non-training sites. Table 2.10 and Table 2.11 highlight R^2 for daily average ozone for the bicubic, IDW, and ordinary kriging interpolations, as well as R^2 for the CMAQ comparison. the entire year was used to evaluate the interpolation methods, but the five highest ozone months from May to September were used for the CMAQ evaluation.

The bicubic R^2 indicates the poorest performance of the three interpolation methods. The lowest R^2 values for the 12 evaluation sites were 0.15 and 0.29, Mission Viejo and West LA, respectively (Table 2.11). The poor performance resulted from the method used to calculate the coefficients a_{ij} (Eq. 5), for which the values of coefficients did not depend on the distance between interpolating points but were dependent on the formation of a smooth curve. Bicubic is best for evenly distributed points, such as interpolating image pixels. IDW showed a significant improvement compared to bicubic interpolation. The lowest R^2 was 0.36 for Mission Viejo, and the highest R^2 was 0.83 for Pomona. Since IDW accounts for the distances between the interpolation points and the data points, farther data points have less influence on the interpolation points. Ordinary kriging resulted in the best interpolation method, with the lowest R^2 of 0.39 for Winchester and the highest R^2 of 0.84 for Pomona. Kriging not only accounts for the distance between building points and interpolated data by assigning larger weight λ_i to the near neighbors, but it also considers the variability of data by considering the variance of input data, σ^2. The basis of the variogram function represents the spatial variability of data. The variance depends not on observation values but on the variogram model and geometry (Kebaili Bargaoui and Chebbi, 2009) (Eq. 2).

ML with interpolation gave a poor performance for Crestline and Winchester locations. Crestline is located in the mountains and to the northeast of SoCAB , which is elevated terrain associated with upper air and a different air mass at times. Crestline ozone was not well-correlated with coastal or inland sites. Thus, interpolated Crestline ozone based on coastal or inland data points will likely yield poor results. The Winchester air monitoring site is located near the Skinner Reservoir (Figure 2.30), far away from other data points (Lake Elsinore and Banning). Low R^2 for Winchester can be explained by the influence of the lake and local meteorology and

air quality. The ordinary kriging model performed well for locations bounded by data points with R^2 above 0.56. However, poor interpolation results occurred for peripheral locations in SoCAB (Crestline, Mission Viejo, and Winchester). LAX ozone levels were not well correlated with meteorology, and training the ML model with fewer meteorological features did not affect the performance of the LAX location. Overall, model performance increased from the West to the East, with better prediction for inland sites.

The distribution of the monthly mean bias (MB) for ordinary kriging interpolation centered around zero with the range between +9 ppb for Compton (August) and -11 ppb for Glendora (October). Eleven building sites have a net positive monthly MB, and four have a net negative monthly MB (Figure 2.31). The results from the CMAQ simulation overestimated the ozone levels. CMAQ's best performance was from May to October when the MBs were the smallest. In general, ozone concentrations in the SoCAB are highest during the summer and lowest in the winter, corresponding with the temperature. Even though the CMAQ simulation captures diurnal variation, the seasonal variation is not as well-represented (Figure 2.32, Figure 2.33, Figure 2.34, and Figure 2.35). Lower performing CMAQ results could come from uncertainties in emissions estimates. CMAQ generally overestimated ozone concentrations because the simulated nighttime ozone concentrations were higher than those observed, potentially due to underestimated nighttime NO_X emissions (Zhu et al., 2023). In other words, that there was not enough NOx emitted in the model during the daytime for ozone formation and at night for ozone removal (Awang and Ramli, 2017; Brown et al., 2004).

Training features can be varied to study the sensitivity to modeled ozone response. For example, the temperature, RH, or emissions values can be perturbed to examine the ozone levels corresponding to the change in the features. However, because the formation of ozone results

from a complex combination of chemical reactions, resulting impacts are nonlinear and interdependent. Therefore, when using ML to test for sensitivity to a feature, one should consider feature dependencies. For example, in testing temperature impacts on ozone concentration, both how temperature impacts photolysis rates (NO_2 degradation) as well as simultaneous correlations/anticorrelation with other meteorological variables, such as RH or wind speed must be considered.

The reduction in traffic volumes during the lockdown from March to May led to a decrease in observed CO and NO_X (Ivey et al., 2020; Tanvir et al., 2023). As a result, an overall reduction in ozone levels was expected over the SoCAB region. The average diurnal ozone concentrations before the lockdown (Jan - Feb) in 2020 were noticeably greater than the average from 2016 – 2019 for all 15 building sites. Figure 2.36 shows the averaged diurnal profiles of three 2020 periods for inland sites, Rubidoux and Fontana: pre-lockdown (a, d), lockdown (b, e), and post-lockdown (c, f) periods. During lockdown, observed 2020 ozone levels (red line) marginally dropped in Rubidoux but still were higher than Fontana's four-year average (blue line). Post-lockdown differences compared to the four-year average were not significant across the 15 sites. The RFR model captured ozone trends throughout 2020, although slightly lower during and despite the observed reduction in NOx, suggesting that besides the air quality features (NO and NO_2), meteorology would play an important role in predicting ozone levels during anomalous episodes. Actual and modeled discrepancies also indicate anomalous ozone behavior during lockdown. For instance, several sites in the SoCAB showed an increase in ozone levels based on the diurnal profile implying that the urban locations in the SoCAB were VOCs-limited regimes, where reduction in NOx reduction-initiated ozone enhancement (Parker et al., 2020).

Conclusion

This study highlights the advantages of spatial interpolation methods for ozone predictions during anomalous environmental events. With modern processor architectures (e.g., AMD Zen 3 or Intel Alder Lake), training RFR model and performing high-resolution interpolation over the SoCAB region for one prediction year took less than five minutes of walltime with a 16-core processor. In contrast, CMAQ walltime was 16 days for a year-long simulation for the SoCAB region. Further, ozone modeling for 2020 was challenging because of expected emissions conditions from March to September, during which traffic volume significantly decreased (up to 40% reduction in some locations). The hypothesis was that mid-2020 ozone levels would decrease semi-proportionally to the decline in traffic volume. However, the changes in ozone levels in the SoCAB was small in magnitude, but directionally the changes were informative for future emissions reductions planning (increased ozone indicates VOC limitations). Ordinary kriging interpolation using ML building provided daily data, addressed data missingness, and captured 2020 ozone trends with fairly low bias despite the sudden change in emissions.

Acknowledgment

This paper was prepared as a result of work sponsored, paid for in part by, the South Coast Air Quality Management District (SCAQMD). The opinions, findings, conclusions, and recommendations are those of the authors and do not necessarily represent the views of SCAQMD. We acknowledge Graduate Assistant in Areas of Need (GAANN) support from the University of California, Riverside Chemical and Environmental Department. The authors thank the South Coast Air Quality Management District for providing emissions data and Dwaraknath Ravichandran and Prof. Shams Tanvir, who provided 2020 traffic data for emission correction. The

authors acknowledge partial support from the University of California Institute for Transportation Studies for work conducted for this manuscript. The authors thank Prof. Armistead G. Russell and Dr. Charles L. Blanchard for their guidance on this work.

References

Apte, J.S., Messier, K.P., Gani, S., Brauer, M., Kirchstetter, T.W., Lunden, M.M., Marshall, J.D., Portier, C.J., Vermeulen, R.C.H., Hamburg, S.P., 2017. High-Resolution Air Pollution Mapping with Google Street View Cars: Exploiting Big Data. Environmental Science and Technology 51, 6999–7008. https://doi.org/10.1021/acs.est.7b00891

Awang, N.R., Ramli, N.A., 2017. Preliminary Study of Ground Level Ozone Nighttime Removal Process in an Urban Area. Journal of Tropical Resources and Sustainable Science (JTRSS) 5. https://doi.org/10.47253/jtrss.v5i2.595

Bartier, P.M., Keller, C.P., 1996. Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). Computers & Geosciences 22, 795–799. https://doi.org/10.1016/0098-3004(96)00021-0

Brown, S.S., Dibb, J.E., Stark, H., Aldener, M., Vozella, M., Whitlow, S., Williams, E.J., Lerner, B.M., Jakoubek, R., Middlebrook, A.M., DeGouw, J.A., Warneke, C., Goldan, P.D., Kuster, W.C., Angevine, W.M., Sueper, D.T., Quinn, P.K., Bates, T.S., Meagher, J.F., Fehsenfeld, F.C., Ravishankara, A.R., 2004. Nighttime removal of NOx in the summer marine boundary layer. Geophysical Research Letters 31. https://doi.org/10.1029/2004GL019412

Brune, W.H., 2001. Introduction to Atmospheric Chemistry: Daniel J. Jacob; Princeton University Press, Princeton, NJ, 1999, 266pp., ISBN 0-691-00185-5. Atmospheric Environment 35, 1715. https://doi.org/10.1016/S1352-2310(00)00432-5

California Air Resources Board, 2023. Trends Summary [WWW Document]. URL https://www.arb.ca.gov/adam/trends/trends1.php

Caltrans, 2023. Caltrans PeMS [WWW Document]. URL https://dot.ca.gov/programs/traffic-operations/census/mvmt

Camalier, L., Cox, W., Dolwick, P., 2007. The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. Atmospheric Environment 41, 7127–7137. https://doi.org/10.1016/j.atmosenv.2007.04.061

Do, K., Manasi, M., Kashfi Yeganeh, A., Gao, Z., Blanchard, C.L., Ivey, C.E., 2023. A Machine Learning Approach to Quantify the Impact of Meteorology on Tropospheric Ozone in the Inland Empire, CA. Accepted in Environmental Science: Atmospheres.

Gao, Z., Ivey, C.E., Blanchard, C.L., Do, K., Lee, S.-M., Russell, A.G., 2022. Separating emissions and meteorological impacts on peak ozone concentrations in Southern California using generalized additive modeling. Environmental Pollution 307, 119503. https://doi.org/10.1016/j.envpol.2022.119503

Ivey, C., Gao, Z., Do, K., Kashfi Yeganeh, A., Russell, A., Blanchard, C.L., Lee, S.-M., 2020. Impacts of the 2020 COVID-19 Shutdown Measures on Ozone Production in the Los Angeles Basin (preprint). https://doi.org/10.26434/chemrxiv.12805367.v1

Jaffe, D., 2020. Role of Meteorology, Emissions and Smoke on Ozone in the South Coast Air Basin, Final project report for CRC Project A-118. Coordinating Research Council, Alpharetta, GA.

Jiang, Z., Shi, H., Zhao, B., Gu, Y., Zhu, Y., Miyazaki, K., Lu, X., Zhang, Y., Bowman, K.W., Sekiya, T., Liou, K.-N., 2021. Modeling the impact of COVID-19 on air quality in southern California: implications for future control policies. Atmos. Chem. Phys. 21, 8693–8708. https://doi.org/10.5194/acp-21-8693-2021

Joseph, J., Sharif, H.O., Sunil, T., Alamgir, H., 2013. Application of validation data for assessing spatial interpolation methods for 8-h ozone or other sparsely monitored constituents. Environmental Pollution 178. https://doi.org/10.1016/j.envpol.2013.03.035

Karamchandani, P., Morris, R., Wentland, A., Shah, T., Reid, S., Lester, J., 2017. Dynamic Evaluation of Photochemical Grid Model Response to Emission Changes in the South Coast Air Basin in California. Atmosphere 8, 145. https://doi.org/10.3390/atmos8080145

Kebaili Bargaoui, Z., Chebbi, A., 2009. Comparison of two kriging interpolation methods applied to spatiotemporal rainfall. Journal of Hydrology 365, 56–73. https://doi.org/10.1016/j.jhydrol.2008.11.025

Liu, S.C., Kley, D., McFarland, M., Mahlman, J.D., Levy, H., 1980. On the origin of tropospheric ozone. Journal of Geophysical Research. https://doi.org/10.1029/jc085ic12p07546

Lurmann, F., Avol, E., Gilliland, F., 2015. Emissions reduction policies and recent trends in Southern California's ambient air quality. Journal of the Air & Waste Management Association 65, 324–335. https://doi.org/10.1080/10962247.2014.991856

Miyasato, M., Tisopulos, L., Low, J., Bermudez, R., Vlasich, B., 2016. Annual Air Quality Monitoring Network Plan.

Oliver, M.A., Webster, R., 1990. Kriging: A method of interpolation for geographical information systems. International Journal of Geographical Information Systems 4. https://doi.org/10.1080/02693799008941549

Parker, H.A., Hasheminassab, S., Crounse, J.D., Roehl, C.M., Wennberg, P.O., 2020. Impacts of Traffic Reductions Associated With COVID-19 on Southern California Air Quality. Geophys. Res. Lett. 47. https://doi.org/10.1029/2020GL090164

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews. https://doi.org/10.1016/j.oregeorev.2015.01.001

Seiler, M.C., Seiler, F.A., 1989. Numerical Recipes in C: The Art of Scientific Computing. Risk Analysis 9. https://doi.org/10.1111/j.1539-6924.1989.tb01007.x

South Coast Air Quality Management District, 2017. Final 2016 Air Quality Management Plan.

Tanvir, S., Ravichandran, D., Ivey, C., Barth, M., Boriboonsomsin, K., 2023. Traffic, Air Quality, and Environmental Justice in the South Coast Air Basin During California's COVID-19 Shutdown, in: Loukaitou-Sideris, A., Bayen, A.M., Circella, G., Jayakrishnan, R. (Eds.), Pandemic in the Metropolis. Springer International Publishing, Cham, pp. 131–148. https://doi.org/10.1007/978-3-031-00148-2_9

Trousdell, J.F., Caputi, D., Smoot, J., Conley, S.A., Faloona, I.C., 2019. Photochemical production of ozone and emissions of NOx and CH4 in the San Joaquin Valley. Atmospheric Chemistry and Physics. https://doi.org/10.5194/acp-19-10697-2019

Wang, W., Bruyere, C., Duda, M., Dudhia, J., Gill, D., Kavulich, M., Keene, K., Chen, M., Lin, H.-C., Michalakes, J., Rizvi, S., Zhang, X., Berner, J., Ha, S., Fossell, K., 2017. WRF Version 3.9 User's Guide.

Wong, D.W., Yuan, L., Perlin, S.A., 2004. Comparison of spatial interpolation methods for the estimation of air quality data. Journal of Exposure Analysis and Environmental Epidemiology. https://doi.org/10.1038/sj.jea.7500338

Yu, H., Russell, A., Mulholland, J., Odman, T., Hu, Y., Chang, H.H., Kumar, N., 2018. Cross-comparison and evaluation of air pollution field estimation methods. Atmospheric Environment. https://doi.org/10.1016/j.atmosenv.2018.01.045

Zhang, C., Ma, Y., 2012. Ensemble machine learning: Methods and applications, Ensemble Machine Learning: Methods and Applications. https://doi.org/10.1007/9781441993267

Zhu, S., Horne, J.R., Mac Kinnon, M., Samuelsen, G.S., Dabdub, D., 2019. Comprehensively assessing the drivers of future air quality in California. Environment International 125, 386–398. https://doi.org/10.1016/j.envint.2019.02.007

Zhu, Z., Do, K., Ibarra Gomez, D., Ivey, C.E., Collins, D., 2023. Assessing CMAQ Model Discrepancies in Vertical Ozone Profiles in a Heavily-Polluted Air Basin using UAV Measurements. In preparation.

Tables

Table 2.8. Data summary for machine learning modeling.

Ground Monitoring Locations	Anaheim, Azusa, Banning, Compton, Fontana, Glendora, Lake Elsinore, LAX, LA North Main Street, Mira Loma, Rubidoux, San Gabriel, Santa Clarita, San Bernardino, Upland
Features	NO_x, NO, temperature, relative humidity, wind speed, wind direction
Label	Ozone
Data sources	EPA AQS data mart, CARB AQMIS
Training years	2009, 2010, 2016, 2017, 2018, 2019
Evaluation year	2020

Table 2.9. Optimal RFR configurations for the study

Hyperparameter	**Description**
n_estimators = 16	The number of trees in the forest.
max_features = 'auto'	The number of features to consider when looking for the best split.
max_depth=None	The maximum depth of the tree.
min_samples_split=5	The minimum number of samples required to split an internal node.
min_samples_leaf=30	The minimum number of samples required to be at a leaf node.
min_weight_fraction_leaf=0	The minimum weighted fraction of the sum total of weights required to be at a leaf node.
max_leaf_nodes=None	Best nodes are defined as relative reduction in impurity.
n_jobs=8	The number of jobs to run in parallel.

Table 2.10. Daily average R^2 at the 15 building sites for three interpolation methods for the year 2020. R^2 for CMAQ was computed using the five highest ozone months May - September of 2020.

Sites	Bicubic R^2	IDW R^2	Ordinary Kriging R^2	CMAQ R^2
Anaheim	0.66	0.67	0.74	0.41
Azusa	0.52	0.64	0.77	0.59
Banning	0.17	0.46	0.73	0.26
Compton	0.65	0.67	0.77	0.48
Fontana	0.88	0.89	0.87	0.59
Glendora	0.46	0.53	0.72	0.52
Lake Elsinore	0.52	0.70	0.79	0.56
LA North Main ST	0.36	0.67	0.78	0.48
LAX	0.31	0.48	0.65	0.25
Mira Loma	0.56	0.71	0.86	0.67
Rubidoux	0.46	0.65	0.86	0.68
San Bernardino	0.68	0.85	0.86	0.67
San Gabriel	0.53	0.77	0.81	0.62
Santa Clarita	0.27	0.72	0.84	0.61
Upland	0.76	0.80	0.86	0.61

Table 2.11. Daily average R^2 at 12 evaluation sites, and these were not used spatial interpolation. R^2 for CMAQ was computed using the five highest ozone months, May - September of 2020.

Sites	Bicubic R^2	IDW R^2	Ordinary Kriging R^2	CMAQ R^2
Crestline	0.35	0.42	0.42	0.23
La Habra	0.75	0.80	0.77	0.44
Long Beach	0.46	0.60	0.56	0.30
Mission Viejo	0.15	0.36	0.49	0.39
North Hollywood	0.67	0.67	0.79	0.59
Pasadena	0.55	0.71	0.78	0.57
Perris	0.55	0.72	0.80	0.56
Pomona	0.71	0.83	0.84	0.68
Redlands	0.60	0.74	0.71	0.57
Reseda	0.63	0.63	0.71	0.01
West LA	0.29	0.56	0.60	0.28
Winchester	0.37	0.40	0.39	0.45

Figures

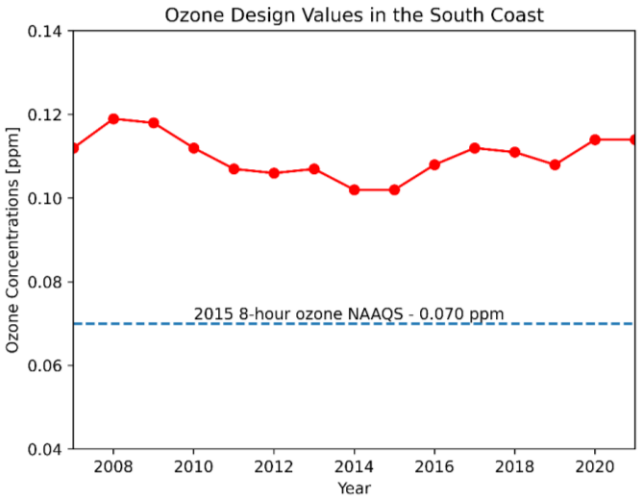


Figure 2.24. Ozone design values for the South Coast Air Basin from 2006 to 2020 (https://www.epa.gov/air-trends/air-quality-design-values).

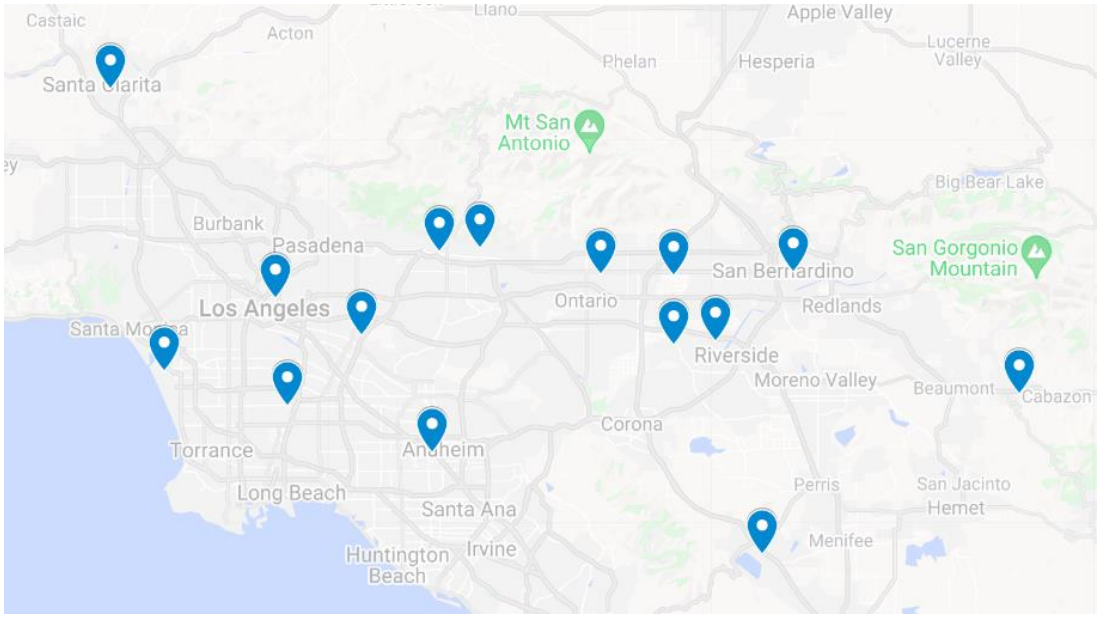


Figure 2.25. Data from 15 air monitoring stations (Anaheim, Azusa, Banning, Compton, Fontana, Glendora, Lake Elsinore, LAX, LA North Main Street, Mira Loma, Rubidoux, San Gabriel, Santa Clarita, San Bernardino, Upland) were used for ML model predictions of ozone concentrations.

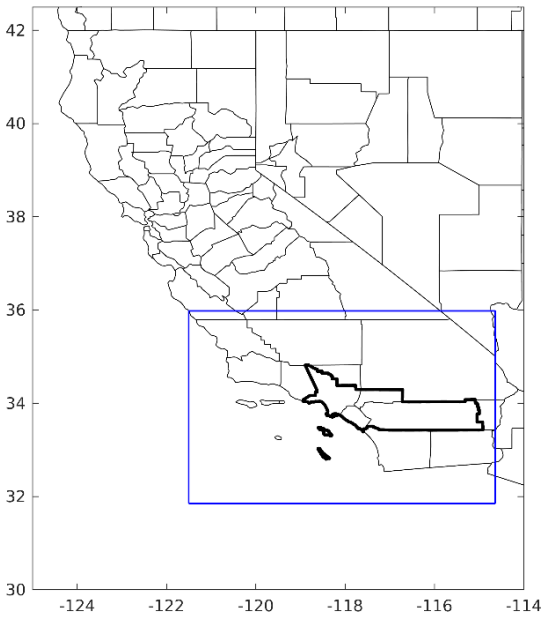


Figure 2.26. The third and inner-most domain (blue boundary) with 4 km. Horizontal grid spacing covered the entire SCAQMD region (thick black lines).

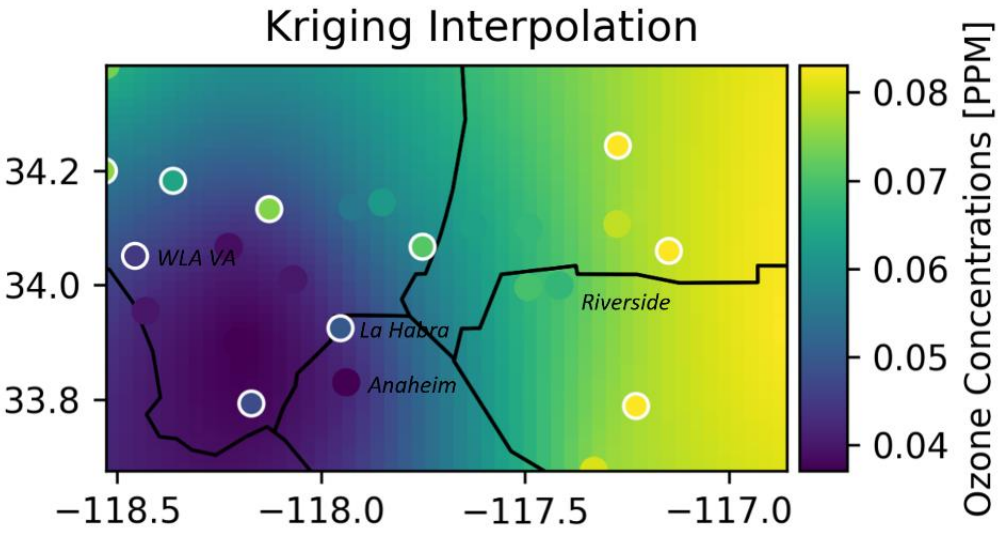


Figure 2.27. Hourly ozone heatmap (16:00 on June 22, 2020) using ordinary kriging. The dots with white borders are the evaluation sites, and dots without borders are the training sites.

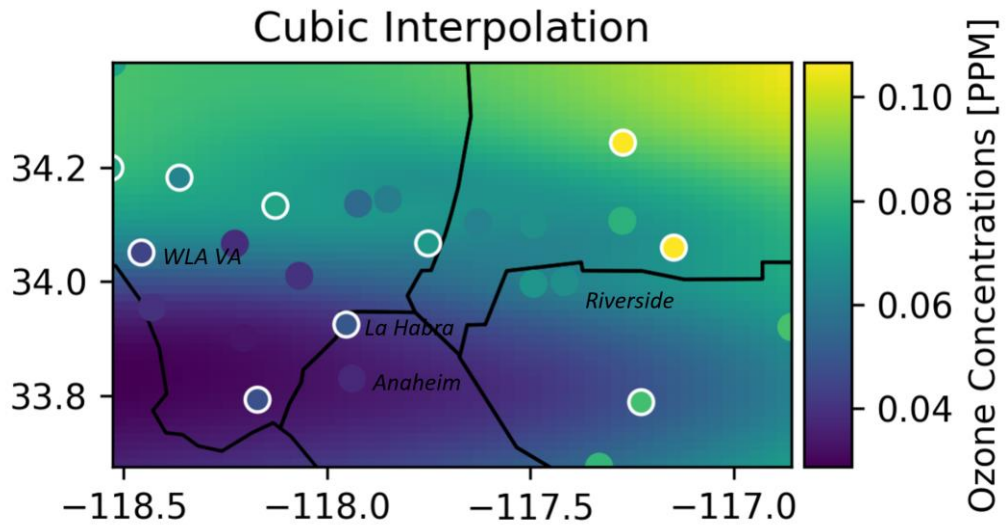


Figure 2.28. Hourly ozone heatmap (@16pm June 22, 2020) using cubic interpolation.

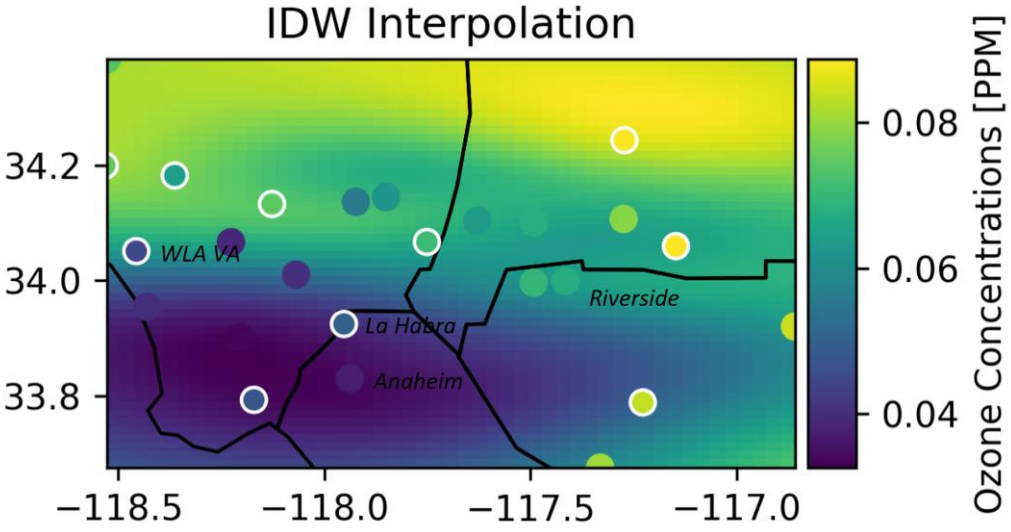


Figure 2.29. Hourly ozone heatmap (@16pm June 22, 2020) using IDW interpolation.

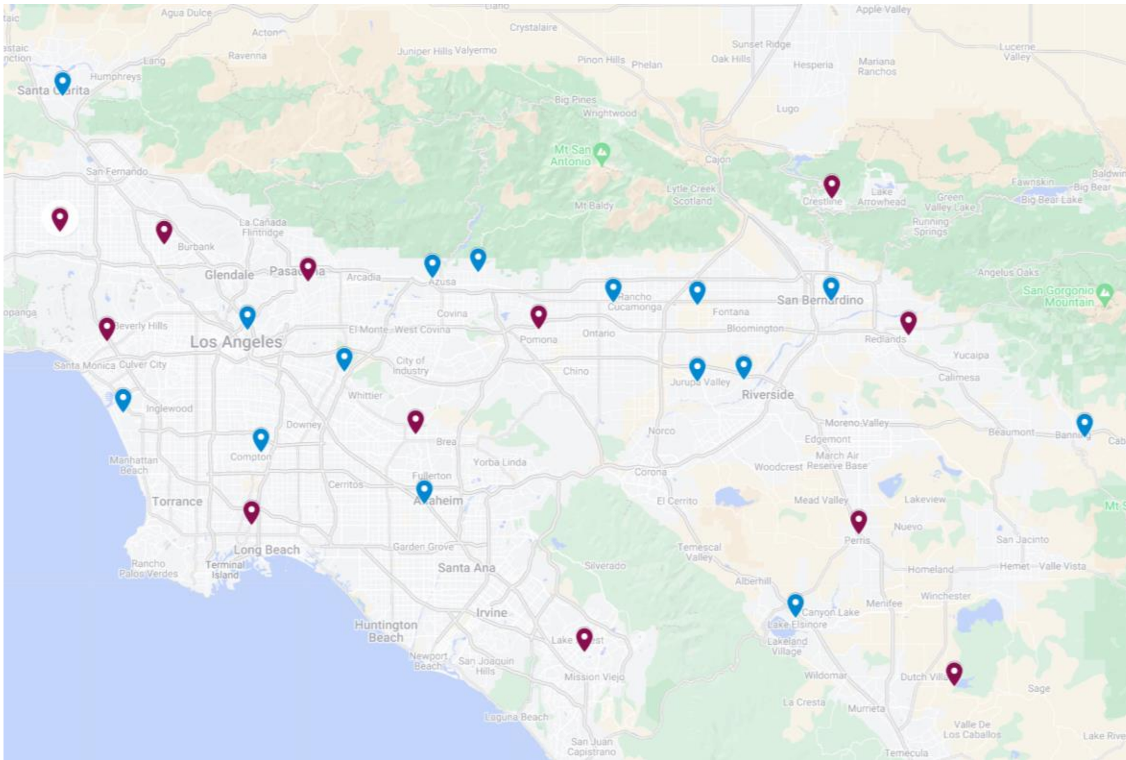


Figure 2.30. The map shows 27 air monitoring sites in the SoCAB. Blue labels were used for interpolation points, and red labels were used for interpolation performance evaluation.

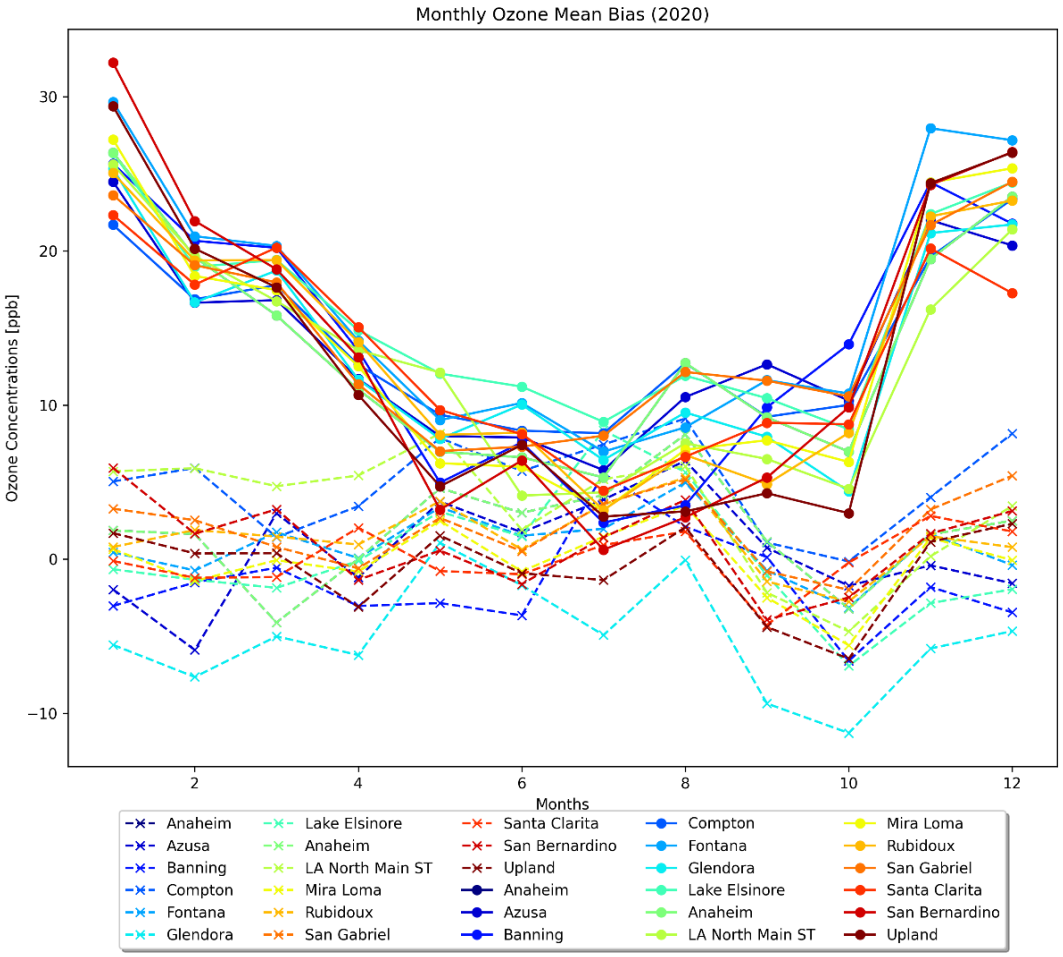


Figure 2.31. Monthly mean bias of the ordinary kriging application (dashed lines) and CMAQ simulation (solid lines).

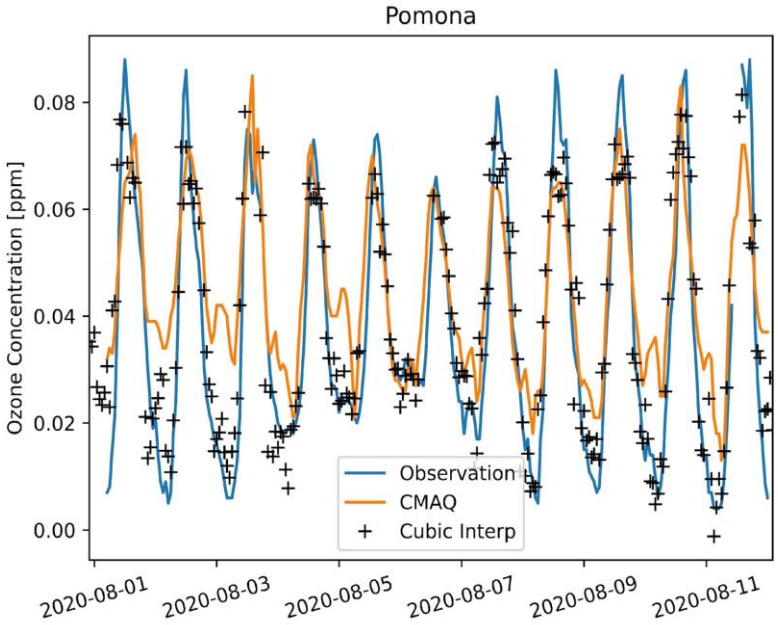


Figure 2.32. Time series plotting ozone concentrations for CMAQ model, cubic interpolation, and observation

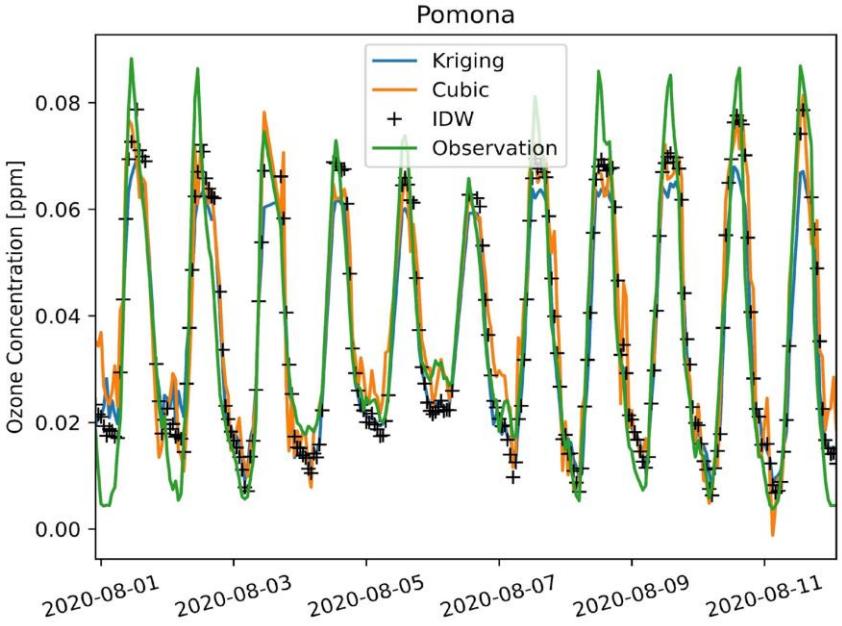


Figure 2.33. Time series plotting ozone concentrations for three different interpolation methods (kriging, cubic, and IDW) with observation

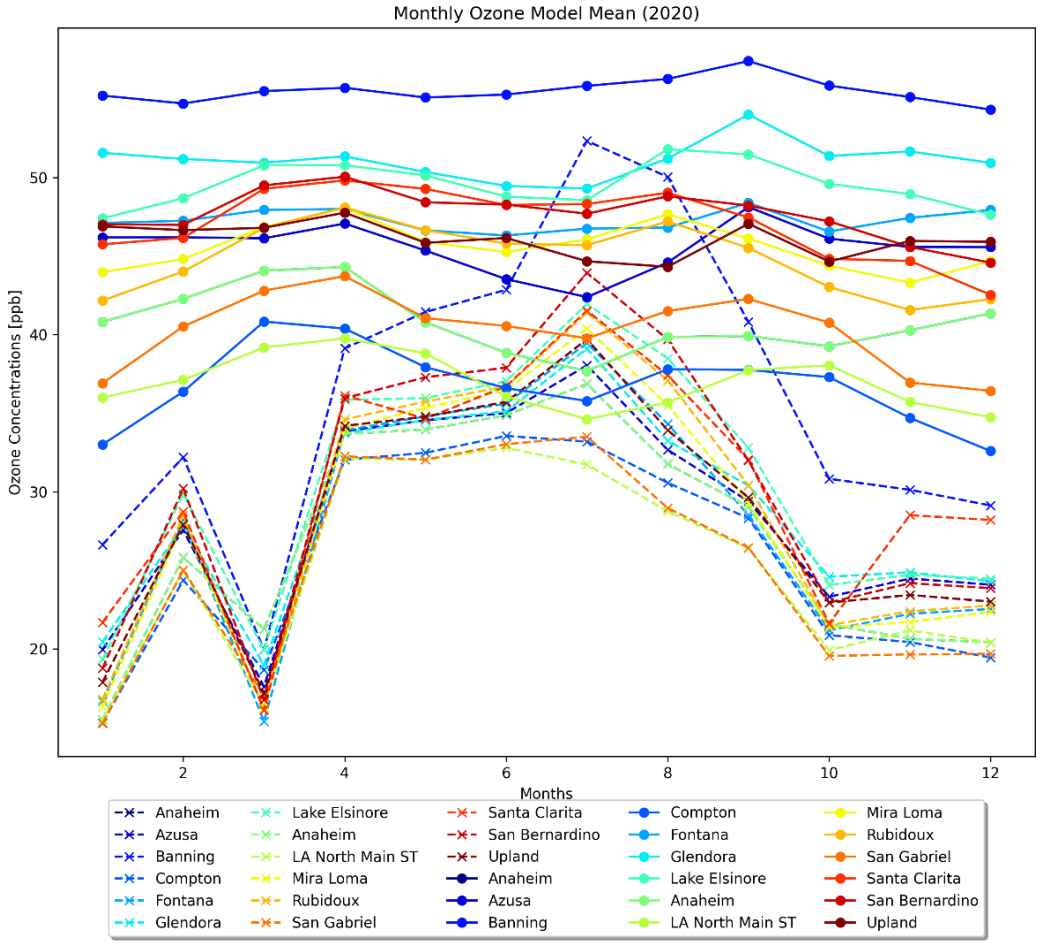


Figure 2.34. CMAQ (solid lines) vs. ML building sites (dash lines) model mean.

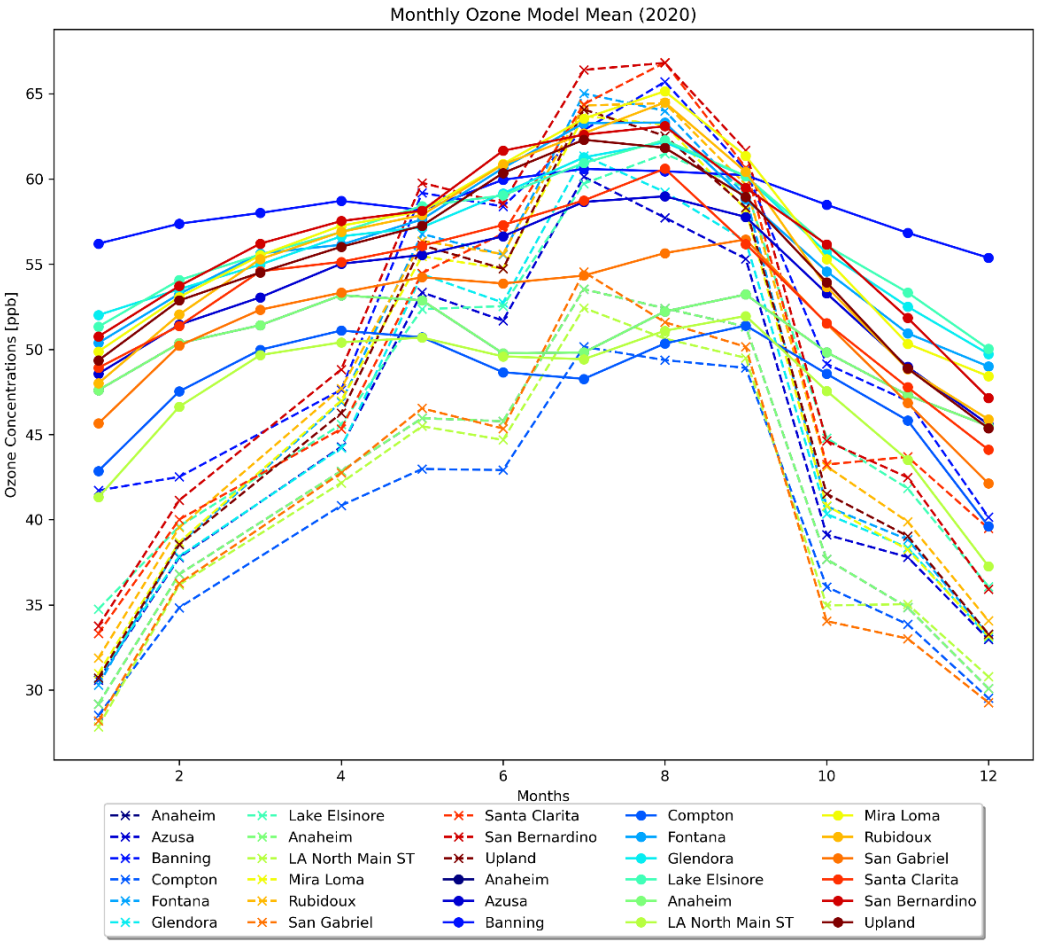


Figure 2.35. CMAQ (solid lines) vs. ML building sites (dash lines) from 9AM to 4Pm.

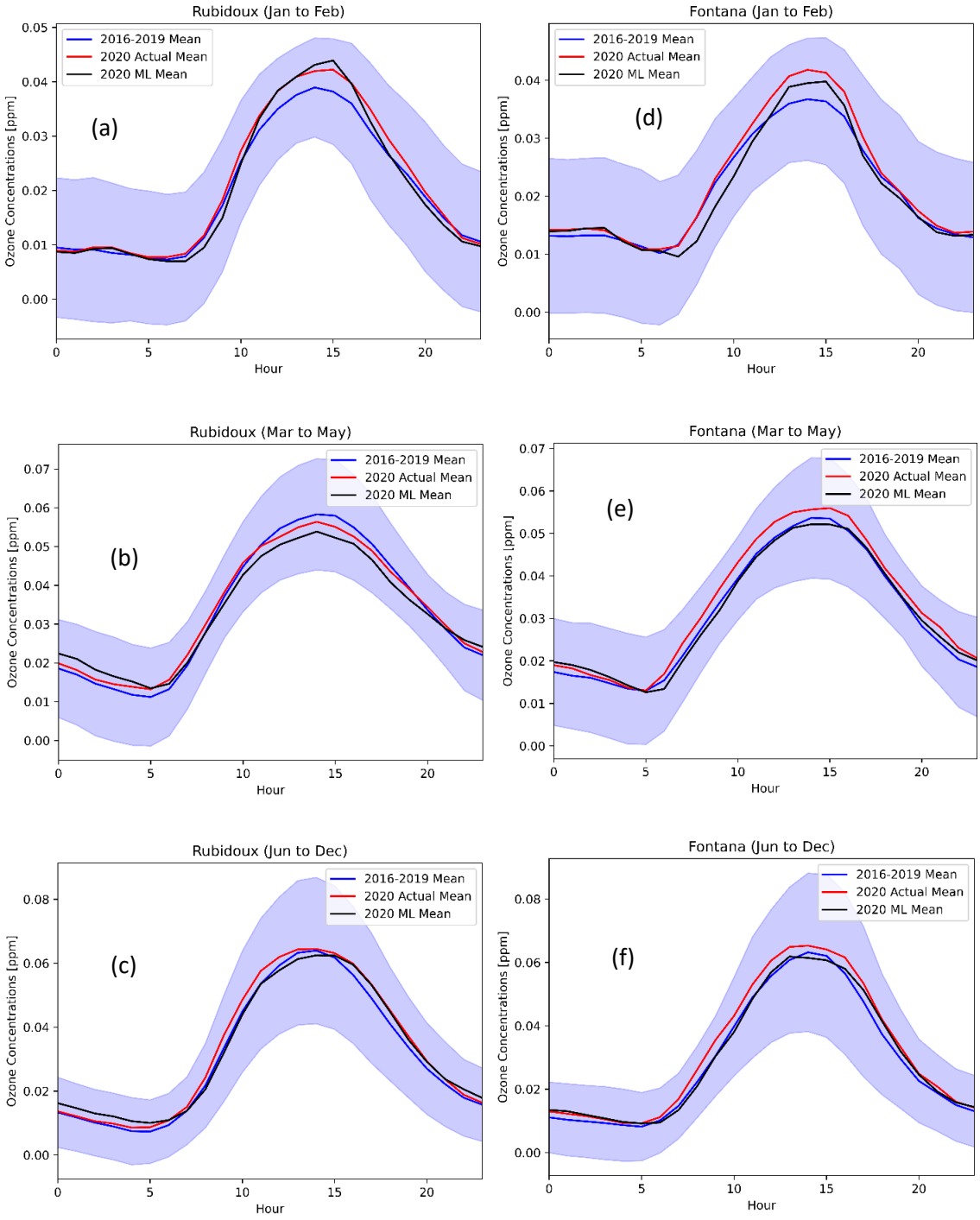


Figure 2.36. Averaged diurnal profiles of 2016 - 2019 (blue), actual 2020 (red), and ML predicted 2020 (black) ozone concentrations (ppm) at Rubidoux (a, b, c) and Fontana (d, e, f) for three different periods: (a,d) pre-lockdown (Jan to Feb), (b,e) lockdown (Mar to May), and (c,f) post-lockdown (after May). The shaded area is the standard deviation of the 2016 - 2019 measurements.

Chapter 3

GPU-Assisted Computation for a Gas-Phase Chemical Solver in CMAQ

Abstract

The Earth's atmosphere is extremely complex. Mimicking the atmosphere involves many scientific processes, such as dispersion, diffusion, deposition, and chemical reactions. Researchers improve the predictability of air quality models by integrating more scientific processes with increasing the number of chemical reactions by adding a significant number of species to the mechanisms, which degrades the computational efficiency for the most comprehensive modeling applications. The disadvantage of the method is the worsening of simulation time. Offline chemical transport models spend much time simulating large atmospheric domains, with the most time solving for the gas-phase chemistry. To improve the simulation time while maintaining the integrity of the models, graphics processing units (GPU) were utilized to replace the central processing units (CPU) for computing the most extensive science process. The gas-phase chemistry solver has been successfully ported onto a GPU to reduce computational time. The actual kernel computing time for the solver is twice as fast as the CPU with the BLKSIZE of 8,000. However, the GPU solver suffers from moving data back and forth between the system memory to the GPU memory. This paper focusesthe details of (1) the compilation of the Community Multiscale Air Quality (CMAQ) model with CUDA kernels, (2) porting the gas-phase chemistry solver onto the GPU, and (3) optimizing the solver to improve GPU computational efficiency. The good results from the ported solver show the promising future for intensive parallel computing applications benefiting researchers in reducing the simulation time and accelerating the research.

Introduction

Deterministic air quality models (AQMs) are designed to simulate complex physical and chemical processes taking place in the Earth's atmosphere with mathematical presentations of the atmospheric transport, diffusion, dispersion, and chemical reactions, which are solved by analytical and numerical techniques and based on the conservation of mass principle for pollutants (Lamb & Seinfeld, 1973). Recently, with the rapid growth of machine learning, AQM can be fully based on empirical statistical relationships between historical data. However, ML models have not been ready to involve in U.S. regulatory decision-making but for research purposes. AQMs do not have all the features of a real system. Depending on the research and regulatory purposes, AQMs were designed to focus on a set of interesting substances, which were used for decision-making on environmental problems and predicting future pollution levels in response to emission controls.

AQMs operate on a set of input data, including meteorology (e.g., wind information, temperature, relative humidity, and turbulent coefficients), emissions of harmful air pollutants and their precursors, topography, and land uses. The outputs usually are mass concentrations of air pollutants, such as ozone, particulate matter (PM), simulated NO_x, and VOCs (Haurie et al., 2004), closely matching measurements from air monitoring stations over the entire simulation domain. AQMs are often used to predict the trends of criteria pollutants for environmental compliance and attainment purposes, which were established by the United States Environmental Protection Agency for six common air pollutants (i.e., PM, O_3, CO, SO_2, NO_2, and Pb).

There are two common types of AQMs. (1) Atmospheric quality dispersion models are dispersion models defined as mathematical descriptions of transport and dispersion using

emission sources and meteorology data. Dispersion models are used to estimate the downwind concentrations of air pollutants and operate on emissions, meteorology, and topography (Hennig et al., 2016; Jerrett et al., 2005). In general, dispersion models are designed for atmospheric dispersion. A few dispersion models account for chemical reactions during the dispersion process. However, most dispersion models consider only inactive species. (2) Atmospheric chemical transport models (CTM) are deterministic models that simulate the atmosphere. CTMs account for space and time in the simulation domain using three-dimensional numerical models to simulate the change of air pollutant concentrations by solving a set of mass balance equations.

CMAQ is one of the Environmental Protection Agency (EPA) regulatory methods used to develop attainment control strategies for criteria air pollutants (Foley et al., 2015; Kim et al., 2010; Yu et al., 2008) for evaluating pollution control, new science processes, and the sources of air pollution. The model is the most widely used in air quality modeling systems in recent years (Simon et al., 2012). The advances in chemical mechanisms and transport have improved the accuracy of the air quality model and the ability to reproduce atmospheric air pollution concentrations. The advances in the science processes in CMAQ successfully mimic the processes in the atmosphere. Zhang et al. carried out a seven-year CMAQ simulation in which the simulated concentrations were within the model performance criteria based on the EPA criteria (Zhang et al., 2014).

CTMs include multiple science processes which accurately simulate the states of air pollutants in the atmosphere, such as transport, photolysis, radiation, multiphase chemistry, and cloud formation, in which photolysis provides the energy from the sun that is sufficient for many chemical reactions, multiphase chemistry used data from laboratory experiments to obtain reaction rates and chemical mechanisms to predict the products from chemical reactions (Al-

Abadleh, 2022; Byun et al., 1998; Dütsch, 1971; Mebust et al., 2003; Søvde et al., 2012). The CMAQ model has two popular chemical mechanisms: Carbon Bond (CB) and the Statewide Air Pollution Research Center (SAPRC). SAPRC was developed to tackle the reactions of emitted volatile organic compounds (VOCs) in the presence of NO_x to form O_3 and other secondary air pollutants. The updated versions of the SAPRC mechanism in the later year give better predictions for secondary pollutants by adding a significant number of species. The numbers of model species for SAPRC-99, SAPRC-07, and SAPRC-18 are 82, 126, and 516, respectively. With the increase in the model species, chemical reactions exponentially increase, with 211 reactions for SAPRC-99, 569 reactions for SAPRC-07, and 1772 reactions for SAPRC-18 (Carter, 2023). Undoubtedly, the simulation time of CMTs is proportional to the chemical reactions. CMAQ is precise in representing atmospheric physics and chemical processes.

Handling large datasets is a challenge, given the limitation of computational efficiency and the data's complexity. Moreover, CTMs apply complex governing equations to solve for the output concentrations using the CPUs. Regarding runtime, a 12-km, two-way coupled WRF-CMAQ simulation using 34 layers of variable thickness with a domain size of 279x251 grid cells requires over 3 hours of work for 32 CPU cores per one simulated day over the five-month period (Wong et al., 2012). From our previous work, 2020 ozone concentrations were simulated over Southern California using 4-km resolution with 156x102 grid cells took 20 days to simulate 12 months with 16 MPI threads.

The computational efficiency of CMAQ largely suffers from solving a set of stiff differential equations when computing the gas phase chemical concentrations. For example, with the SAPRC07 chemical mechanism, the systems of photochemical reactions are calculated using Euler

Backward Iteration, SMV Gear, or Rosenbrock solver (ROS) for every time step and grid cell (row x column x height) for all species in the SAPRC07 family until a specified convergence tolerance is met. The running time is linearly proportional to the simulated domain and exponential with increased chemical species.

In this paper, the CMAQ simulation time is improved by porting intensive computational processes onto GPUs. With thousands of Compute Unified Device Architecture (CUDA) cores in a single GPU, many independent arithmetic operations can be carried out simultaneously. The computer architecture, the advantages, and the disadvantages of GPU programming are closely examined. The partial derivative, decomposition, and back substitution subroutines of the ROS3 (Rosenbrock) solver were successfully converted to the CUDA platform. Our CPU-GPU version of the CMAQ model is tested with SAPRC07 for a simulation over Southern California with 102 x 156 x 11 grid cells.

CMAQ on CPU using MPI: All air quality models were designed to utilize CPUs to calculate the science processes. With a set of instructions, scheduler, and high clock speed, CPUs are superior for arithmetic operations, and all programming languages are naturally compiled and executed by CPUs. CMAQ is a grid model in which the 3-dimensional space is subdivided into 3-dimensional grid cells, and the resolution of the simulation domain is defined by user inputs. CMAQ uses Message Passing Interface (MPI) to parallel its simulation process, where multiple grids are executed simultaneously by the number of available CPU cores, local memory, and data must be explicitly shared by passing messages between processes (Clarke et al., 1994; Gropp et al., 1996; Lusk et al., 2009). Even though CPUs are fast for arithmetic operations and CMAQ simulation time is linearly inversely proportional to the number of CPU cores, CPU threads are

limited and expensive. For a modern consumer CPU lineup, the most core CPU consists of 24 cores for Intel (Core i9-13900K) and 16 for AMD (Ryzen 9 7950X). The best CPU for server platforms is AMD EPYC 7773X with 64 cores and costs $8,800 (as of April 2023). To accelerate the simulation, researchers can purchase expensive clusters or pay the premium for cloud services (e.g., AWS, Azure). Somehow, this seems less feasible for small research and consultant groups with less funding in the United States or developing countries.

GPU Advantages: Modern GPUs are capable of computing vector operations with floating-point arithmetic. New-generation GPUs can handle double-precision floating point numbers to improve model accuracy (Nvidia et al., 2009; Whitehead, 2011). Despite low clock speed and naïve scheduler (Guevara et al., 2009), GPUs still outperform the CPUs for multithreaded applications. Ada Lovelace architecture GPUs from NVIDIA have up to 16,384 CUDA cores with a 2.52 GHz boost clock and 24 GB of GPU memory (NVIDIA RTX 4090). Many CUDA cores enable a GPU to perform thousands of arithmetic operations simultaneously, significantly benefiting from parallelized computing and handling extensive data. Multiple GPU streams perform simple operations; all processes in all GPU threads are identical. GPU computing is more affordable and scalable than high-performance computers with numerous nodes. The cost of an RTX 4090 is $1599 (as of April 2023), and installing a GPU into an existing computer is quite simple.

Related work: GPUs were traditionally developed to accelerate graphic rendering for output to display devices. The rendering executes a single set of instructions on multiple GPU cores, emphasizing parallel processing on one specific task. In 2006, NVIDIA launched CUDA, the first commercial solution for general-purpose computing on GPUs (GPGPU). CUDA has provided unique frameworks that allow developers to integrate GPU computing across different

programming languages. CUDA has become an effective tool for training deep learning and machine learning models. In recent years, with the popularity of GPGPU and the improvement of CUDA from NVIDIA, air quality researchers have sought alternative solutions to accelerate chemical transport models. Delic has shown that porting a selected loop of CMAQ to a GPU was feasible (Delic, 2010). In 2023, Kai Cao et al. successfully ported the horizontal advection process (HAVDPPM) from CAMx onto a GPU. The gain for GPU-HAVDPPM was substantial compared to the original HAVDPPM on the CPU (Cao et al., 2023). However, the port of HAVDPPM to the GPU was incomplete and not native to the Fortran programing language. Kai Cao translated HAVDPPM to C programming languages before using CUDA C to execute the process on the GPU. This method reduced the overall efficiency of the computing time due to the heterogeneity in programming languages between Fortran and C.

The work introduces the port of the CMAQ chemical solver onto the GPU using CUDA Fortran, the native programing language of the model. The method (1) simplifies the compilation process, (2) improves the overall model simulation time, and (3) increases readability for future development. One step compilation of the CMAQ's CCTM was developed by embedding CUDA Fortran into the Makefile, in which the original CMAQ modules were compiled with GNU or Intel compiler, and CUDA subroutines were compiled with CUDA Fortran.

CMAQ Model Descriptions

CMAQ describes the dynamics of the atmosphere with a set of governing equations on grid cells in terms of the column, row, and layer. Figure 3.1 shows the simplified overview of CMAQ's modules, in which the science processes operate in series. The science process module calls vertical diffusion (vdiff), horizontal advection (hadv), vertical advection (zadv), horizontal diffusion

(hdiff), cloud process (cldproc), and chem (chem) (Byun et al., 1999; EPA, 2019). The first four processes imitate the transport of the model using meteorological data inputs from WRF or other meteorology models. The cloud process computes the concentration changes in the cloud due to aqueous chemistry, scavenging, and wet deposition. The chem process calculates the gas phase concentrations based on the chemical mechanism provided by user input (SAPRC or CB).

Three ordinary differential equation (ODE) solvers built-in CMAQ are Euler Backward Iteration (EBI), Rosenbrock (ROS), and Sparse Matrix Vectorized Gear (smvgear) (EPA, 2019). The default solver method in CMAQ is EBI due to its superior computing time. However, EBI is prone to inaccurate results and convergence errors with small time steps for steep differential equations. On the other hand, smvgear has the most accurate results (*CMAQv5.2 Operational Guidance Document*, 2017). However, the prolonged computing time of the gear method is a big disadvantage. Figure 3.2Figure 2.32 shows the overall CMAQ simulation time for three solvers with different MPI threads. EBI is 2.5x and 2.1x faster than smvgear and ROS3 solver, respectively. Increasing the number of MPI threads improve the overall simulation time; however, the time differences between the solvers remain the same.

Descriptions of ODE solvers: The EBI solver is the default CMAQ method due to its computational efficiency. However, considering accuracy and data structure, the EBI method is unsuitable for GPU computation. The general differential equations for a chemical system can be expressed in Equation 1, in which the change in concentration of specie i equals the difference between the production and the loss of specie i.

$$\frac{dc_i}{dt} = P_i - L_i c_i \; and \; i = 1, \dots, s \tag{1}$$

where P_i is the production, L_i is the loss term of specie i, and s is the number of chemical species (Jacobson, 2005). The numerical solution using EBI approximation is shown as $c_i^{n+1} = c_i^n + P_i^{n+1}\Delta t - L_i^{n+1}\Delta t c_i^{n+1}$, and can be written in the form:

$$c_i^{n+1} = \frac{c_i^n + P_i^{n+1}\Delta t}{1 - L_i^{n+1}\Delta t} \tag{2}$$

The solution using the EBI method is just a simple linear combination, which makes the method effective in solving ODEs. To solve the concentrations in the gas phase, CMAQ iterates Equation 2 through all the gas species until the criteria are met. Because the correct concentrations depend on the order of the species, Equation 2 must be carried out in order to obtain the correct solution. The EBI method is configured in series means porting EBI to the GPU is not beneficial.

The Rosenbrock and SMVGEAR solvers were designed based on the code originally developed by M. Jacobson by adding a sparse-matrix package and vectorized loops about the grid-cell dimension to improve computational burden (Z. Jacobson & Turco, 1994). For a set of ODEs as $\frac{d\boldsymbol{c}}{dt} = f(t, \boldsymbol{c})$, the prediction matrix is $\boldsymbol{P} \cong I - h\beta_o J$, where $\boldsymbol{c}$ is the concentration vector, I is the identity matrix, h is the time step, β is the scalers, and $J = \frac{\partial \boldsymbol{c}}{\partial t}$ is the Jacobian matrix. The Gear method uses decomposition with back-substitution to solve for the concentrations (Jacobson, 2005; Z. Jacobson & Turco, 1994). The calculation of the Gear method can be carried out independently with matrix and vector operations. The method has a high degree of parallelization and favors GPU computation.

Methods

Determine the slowest science process: The science process modules in CMAQ have different computational times and require various hardware resources. The time for one simulation day of five science processes was measured, as shown in Figure 3.3. The gas phase chemistry was the most time-consuming step, taking five times longer than VDIFF and HADV and significantly impacting the overall CMAQ simulation duration. Successfully porting the CHEM module onto the GPU essentially improves the model performance.

Compilation: CMAQ was written in Fortran and compiled using Intel Fortran or the open-source GNU Fortran compiler. The CMAQ compilation process is straightforward, with appropriate pre-install libraries. The port of CMAQ subroutines onto GPU requires heterogenous compilers, in which Fortran compiler is used to compile traditional CMAQ modules (.F files), and nvfortran is used to compile CUDA subroutines (.cuf files). The data flow between .F and .cuf subroutines is strictly enforced, in which .cuf subroutines cannot directly inherit variables from CMAQ data modules.

Parallelization independent loops: Parallelizing dependent loops result in inaccurate outputs. In gas-phase chemistry, concentrations of several species must be computed in priority. For example, in the EBI method, the concentrations of NO_2, NO, O_3, and O_3P need to calculate first before computing HO, HO_2, HONO, and HNO_4. In the ROS method, the order of the decomposition loop needs to be executed in series. Therefore, performing loop-dependent analysis before porting to the GPU is essential to maintain the model's integrity. Bernstein's conditions were used to test for statements or operations that can be interchanged without altering the model's

outputs. The conditions state that if neither Equation 3 to Equation 5 holds, the statements can be interchangeable.

$$OUT_1 \cap IN_2 = \emptyset \tag{3}$$

$$IN_1 \cap OUT_2 = \emptyset \tag{4}$$

$$OUT_1 \cap OUT_2 = \emptyset \tag{5}$$

where IN and OUT are the inputs and outputs of task 1 and task 2.

Effects of BLKSIZE: BLKSIZE parameter can be set before compiling CMAQ for the smvgear and ROS solvers. The default BLKSIZE is 50. BLKSIZE influences the way CMAQ handles the grid cells. Large BLKSIZE results in fewer function-call for a given domain, but the solver effectively deals with big matrices and data promoting GPU computing. In the CMAQ configurations, the grid dimension is 102 x 156 x 11 (row x column x layer), resulting in a total number of grid cells of 175,032. If the BLKSIZE is 50, CMAQ calls the gas-chem solver 3,500 times. When the BLKSIZE is set to 2,000, the gas-chem solver is called 88 times. However, the concentration matrix is 40 times larger than the BLKSIZE 50. Figure 3.4 shows the effect of the BLKSIZE on CMAQ simulation time. The gas-chem module and other science process modules were timed for one iteration timestep. Larger BLKSIZE degraded CMAQ performance; with 10,000 BLKSIZE, CMAQ is about 3.5 times slower than the default BLKSIZE. Increasing the BLKSIZE would not influence CMAQ simulation time in an ideal condition. However, because of the limitations of hardware, big concentration matrices due to large BLKSIZE overflow CPU cache (Matam et al., 2012; Ristov et al., 2014; Sulatycke & Ghose, 1998), and when the CPU performs matrix operations, the CPU has to retrieve

the data from the system memory which is much slower compared to CPU cache explaining increasing the BLKSIZE has a negative effect of the overall CMAQ simulation time.

Figure 3.4 shows the BLKSIZE parameter impacts the gas-phase chemistry process. The simulation time of the science processes increased solely due to the computation time of the gas-phase chemistry, and other science processes' computing time, such as diffusion, advection, or aerosol, is unaffected by the BLKSIZE.

GPU Computing: A GPU can be seen as a computing unit with its own instruction sets, arithmetic-logic units (ALUs), GPU cache, and GPU memory. The GPU (device) communicates with the CPU (host) through a peripheral component interconnect express (PCIe) (Figure 3.5). Carrying out an operation on a GPU has to follow three steps: (1) send a copy of the data from the host to the device, (2) launch a CUDA kernel for instruction to compute on the GPU, and (3) send the GPU computed data (results) from the device to the host. Each step adds to the overall computing time of the system. The general limitation of GPU computing is the bottleneck of the PCIe bandwidth for data transferring between the host and the device.

System Configurations: The simulation from CMAQ and CMAQ-CUDA was carried out on a consumer desktop computer with an Intel Core i5 8400, 16GB of system memory, and NVIDIA RTX 3090 GPU. The low-end computer was used to ensure the performance of CMAQ on a wide range of devices, including high-performance computers and regular desktops with upgraded graphics cards. The CMAQ test domain over Southern California with 4 x 4km resolution consists of 156 x 102 x 11 grid cells. The input meteorological data for CMAQ simulation were the North American Mesoscale Forecast System (NAM) integrated with NOAA high-resolution sea surface temperature (SST). OBSGRID was used to improve meteorological analyses, incorporating the

observed surface and upper air to correct the NAM data corresponding to the ds461 and ds351 datasets, respectively (Wang et al., 2017).

Results and Discussion

Figure 3.6 shows the timing breakdown of the GEAR subroutines for one timestep over the computational domain. The slowest subroutines are the calculation for decomposition (DECOMP) and partial derivative (PDERIVE), in which the concentration matrix is computed a partial derivative with respect to the species and decomposed into lower and upper matrices. The GPU porting of the solver prioritizes the slow subroutines/loops to optimize the model's performance.

Figure 3.7 shows a block diagram of the CUDA Rosenbrock solver for CMAQ-CUDAv1.0. The codes from blue blocks are executed using the CPU. The CUDA kernels are the red blocks that are executed using the GPU. Ideally, more ported GPU subroutines significantly improve the computing time. However, only selected subroutines can be parallelized due to the data dependence. Because .cuf subroutines (compiled with nvfortran) cannot understand .F (compiled with Fortran compiler), an intermediate.F was introduced (compiled with nvfortran), which can communicate with both subroutines. The intermediate.F is the bridge between the host and the device where data and variables must pass through intermediate.F. The calculation of rbfeval.F, rbjacob.F, rbdecomp.F, and rbsolve.F is repeated until tolerance is met.

Table 3.1 summarizes the computing time for the four subroutines from the Rosenbrock solver performed on conventional CMAQ and CMAQ-CUDAv1.0 for 2,100 BLKSIZE and 10,000 BLKSIZE. The actual GPU computing time (kernel time) is much faster than CPU time with the same arithmetic operations. With 10,000 BLKSIZE, GPU computing is about 91% faster for the RBFEVAL

subroutine, 86% for the RBJACOB subroutine, 93% for RBDECOMP, and 92% for RBSOLVE. Even though the computation carried out on the GPU outperformed CPU performance, the overall CMAQ simulation time severely suffered from data transferring between the host and the device. The gain from the kernel was offset by the data allocation, increasing CMAQ simulation time.

CMAQ-CUDAv1.0 implemented a naïve version of the Rosenbrock solver on the GPU, in which every subroutine on the solver has its kernel and requires multiple data transfer through a PCIe (Figure 3.7). When the dimension of concentration matrices is larger with large BLKSIZE, GPU computing experiences loss due to the bottleneck of the PCIe bandwidth.

A new algorithm was developed to minimize data transfer and optimize parallel computing in CMAQ-GPUv2.0. In the second version, the four subroutines of the solver were vectorized and combined to enhance the data transfer. The initial subroutine prepares all required data used by the solver and sends the data to the GPU memory. After the kernel calculation, the RBSOLVE returns the results to the host (Figure 3.8). This version still requires sending the results from the GPU back to the host for convergence check. This optimization requires only a one-time data transfer between the host and the device. The disadvantage of the method is that the size of each copy of data is large and requires greater GPU memory. For CMAQ-CUDAv1.0, each copy of data is 3.5GB and will be cleared after the kernel is finished. For CMAQ-CUDAv2.0, the size of the data is 6.7GB per copy. If launching three kernels in parallel, the amount of data exceeds 20GB when dealing with large BLKSIZE. NVIDIA's latest GPUs for data centers offer an impressive 48GB of memory at a substantial cost.

Figure 3.9 shows the average time of the Rosenbrock solver for one iteration for CMAQ, CMAQ-CUDAv1.0, and CMAQ-CUDAv2.0 for 8,000 BLKSIZE. The actual computation time (orange

bar) is faster with CMAQ-CUDA than with traditional CMAQ. The gain from the kernel for CMAQ-CUDAv1.0 was offset by the transferring time to the device (blue bars) and the host (yellow bars). The implementation of CMAQ-CUDAv2.0 optimized the transferring time and showed a significant improvement. However, the kernel of CMAQ-CUDAv2.0 experienced longer computation time due to extended arithmetic operations added to the kernel. Because more codes from the RBSOLVER were ported to the kernel of CMAQ-CUDAv2.0, the kernel had to compute a large number of series operations, in which the slow clock speed from the GPU harmed the overall computation time. CMAQ-CUDAv2 allocated a large amount of data on the kernel, causing the GPU memory management to be less efficient and adding extra time for searching through the memory to acquire the data (Fauzia et al., 2015; Winter et al., 2021). CMAQ-CUDAv2.0 also performed better with large BLKSIZE. The total time for one iteration of the Rosenbrock solver, including data transferring and computation time, is shown in Figure 3.10. The traditional CMAQ gave a good performance with a small data size (BLKSIZE), but when the BLKSIZE was greater than 3700, CMAQ-CUDAv2.0 had better computation time.

The accuracy of CMAQ-CUDA was evaluated by carrying out the simulation for 24 hours and comparing the outputs with the original CMAQ version. Figure 3.11 shows the simulated concentrations of the most common species for CMAQ and CMAQ-CUDA. The left panels are the output concentrations from CMAQ, the middle panels are CMAQ-CUDA simulation, and the right panels are the concentration differences between CMAQ and CMAQ-CUDA. The two CMAQ versions produced very similar concentrations across all the species, including the maximum and minimum values of the entire simulated domain. The mean bias over the domain is small, and the maximum errors between the two values are about 0% for SO_2 and CO, 5% for O_3, and 7% for NO.

The errors between CMAQ and CMAQ-CUDA came from the differences in hardware architecture. Numerical discrepancies between CPU and GPU are merging double-precision multiplication and addition into double-precision fused multiply-add (FMA) architectures to improve accuracy by reducing rounding and preventing subtractive cancellation (Blanchard et al., 2020; Quinnell et al., 2008, 2015; Zhang et al., 2019). Calculation errors from the differences between GPU and CPU architectures will not be significant. However, dealing with extremely small numbers for concentrations and small timesteps for solving stiff ODEs will magnify the errors. Another uncertainty contributing to the errors could be the definition and allocation of double precision floating point for GCC compiler and CUDA Fortran platform.

Future work: Typically, the solution for the ODEs must go back and forth in the Rosenbrock solver more than six times until the numerical solution meets a set of criteria. With CMAQ-CUDAv2.0, the solution from the kernel has to copy to the host for convergence check and upload to the device again for the next iteration. In future work, CMAQ-CUDAv3.0 will be introduced to overcome this limitation. Implementing CMAQ-CUDAv3.0 will port the convergence check of the solver onto the kernel for further optimizing the transferring time (Figure 3.12), and the outputs from the kernel are the final solution to reduce the transferring time by a factor of five. This version will optimize the number of selected variables that need for the kernel computation. The constant data, such as reaction rates, will be stored on the GPU memory permanently.

Hardware Optimization: The CMAQ-GPU is currently limited to computer hardware. The major disadvantage of GPU computing is the transferring time between the host and the device. Even though the GPU kernel performs much faster than the CPU for parallel applications, the

moving and reallocation of data can offset any gain from the GPU kernel. The data moving time can sometimes be even greater than the actual computing time. The bottleneck comes from the PCIe bandwidth, system memory speed, and GPU memory speed. In the future, the release of newer hardware will lift the bottleneck limitation for GPU computing. The release of PCIe 7.0 (Table 3.2) and DDR6 in 2025 is twice faster as DDR5 and four times faster than DDR4, and the PCIe bandwidth is sixteen times faster than PCIe 4.0 in the test system. With the better-supported hardware framework, CMAQ-GPU will be about four times faster than the current version.

Conclusion

Results from CMAQ-CUDAv2.0 show a promising future of GPU computing for CTMs. An optimized ported kernel significantly reduces the computation time for large data sizes. The actual computation time from the GPU kernel is much faster than on the CPU. However, the time for moving the data between the host and the device added a significant amount of time to the overall results. The CMAQ-CUDAv3.0 algorithm are proposed to minimize the transfer time by porting the convergence check loop onto the kernel. The current limitations of the test system are the PCIe bandwidth and memory speed bottleneck. In the future, with the newer generation of PCIe and DDR, remarkable improvements can be obtained in GPU computing for scientific applications. The development of CMAQ-CUDA provided the framework for GPU computing, in which newly developed highly parallel computing for science process modules can be easily written for GPUs and compiled using the one-step compilation method for Fortran modules and CUDA kernels. Scientists can turn on and off GPU computing options with a flag in the build scripts.

Acknowledgments

The authors thank Prithviraj Yuvaraj for his contribution to the research. This paper was prepared as a result of work sponsored and paid for, in whole or in part, by the Computational and Data-Enabled Science and Engineering from the National Science Foundation (NSF CDS&E). The opinions, findings, conclusions, and recommendations are those of the authors and do not necessarily represent the views of NSF CDS&E.

References

Al-Abadleh, H. A. (2022). Atmospheric aerosol chemistry: State of the science. In *Atmospheric Aerosol Chemistry: State of the Science*. https://doi.org/10.1515/9781501519376

Blanchard, P., Higham, N. J., Lopez, F., Mary, T., & Pranesh, S. (2020). Mixed precision block fused multiply-add: Error analysis and application to GPU tensor cores. *SIAM Journal on Scientific Computing*, *42*(3). https://doi.org/10.1137/19M1289546

Byun, D. W., Ching, J. K. S., Novak, J., & Young, J. (1998). Development and Implementation of the EPA's Models-3 Initial Operating Version: Community Multi-Scale Air Quality (CMAQ) Model. In *Air Pollution Modeling and Its Application XII*. https://doi.org/10.1007/978-1-4757-9128-0_37

Byun, D. W., Young, J., & Talat Odman, M. (1999). Ch06: Governing Equations and Computational Structure of the Community Multiscale Air Quality (CMAQ) Chemical Transport Model. *Science Algorithms of the EPA Models-3 Community Multiscale Air Quality (CMAQ) Modeling System*.

Cao, K., Wu, Q., Wang, L., Wang, N., Cheng, H., Li, D., & Wang, L. (2023). GPU-HADVPPM V1.0: high-efficient parallel GPU design of the Piecewise Parabolic Method (PPM) for horizontal advection in 2 air quality model (CAMx V6.10) 3. *EGUsphere*. https://doi.org/10.5194/egusphere-2023-410

Carter, W. P. L. (2023). *DOCUMENTATION OF THE SAPRC-22 MECHANISM*. https://intra.cert.ucr.edu/~carter/SAPRC/22.

Clarke, L., Glendinning, I., & Hempel, R. (1994). The MPI Message Passing Interface Standard. In *Programming Environments for Massively Parallel Distributed Systems*. https://doi.org/10.1007/978-3-0348-8534-8_21

CMAQv5.2 Operational Guidance Document. (2017).

Delic, G. (2010). *DEVELOPING CMAQ FOR MANY-CORE AND GPGPU PROCESSORS*.

Dütsch, H. U. (1971). Photochemistry of atmospheric ozone. *Advances in Geophysics*, *15*(C). https://doi.org/10.1016/S0065-2687(08)60303-9

EPA. (2019). *CMAQv5.3 User Manual*.

Fauzia, N., Pouchet, L. N., & Sadayappan, P. (2015). Characterizing and enhancing global memory data coalescing on GPUs. *Proceedings of the 2015 IEEE/ACM International Symposium on Code Generation and Optimization, CGO 2015*. https://doi.org/10.1109/CGO.2015.7054183

Foley, K. M., Dolwick, P., Hogrefe, C., Simon, H., Timin, B., & Possiel, N. (2015). Dynamic evaluation of CMAQ part II: Evaluation of relative response factor metrics for ozone attainment demonstrations. *Atmospheric Environment*, *103*. https://doi.org/10.1016/j.atmosenv.2014.12.039

Gropp, W., Lusk, E., Doss, N., & Skjellum, A. (1996). A high-performance, portable implementation of the MPI message passing interface standard. *Parallel Computing*, *22*(6). https://doi.org/10.1016/0167-8191(96)00024-5

Haurie, A., Kübler, J. J. E., Clappier, A., & Van Den Bergh, H. (2004). A metamodeling approach for integrated assessment of air quality policies. *Environmental Modeling and Assessment*, *9*(1). https://doi.org/10.1023/B:ENMO.0000020886.39231.52

Hennig, F., Sugiri, D., Tzivian, L., Fuks, K., Moebus, S., Jöckel, K. H., Vienneau, D., Kuhlbusch, T. A. J., de Hoogh, K., Memmesheimer, M., Jakobs, H., Quass, U., & Hoffmann, B. (2016). Comparison of land-use regression modeling with dispersion and chemistry transport modeling to assign air pollution concentrations within the Ruhr area. *Atmosphere*, *7*(3). https://doi.org/10.3390/atmos7030048

Jacobson, M. Z. (2005). Fundamentals of atmospheric modeling second edition. In *Fundamentals of Atmospheric Modeling Second Edition* (Vol. 9780521839709). https://doi.org/10.1017/CBO9781139165389

Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., & Giovis, C. (2005). A review and evaluation of intraurban air pollution exposure models. In *Journal of Exposure Analysis and Environmental Epidemiology* (Vol. 15, Issue 2). https://doi.org/10.1038/sj.jea.7500388

Kim, Y., Fu, J. S., & Miller, T. L. (2010). Improving ozone modeling in complex terrain at a fine grid resolution - Part II: Influence of schemes in MM5 on daily maximum 8-h ozone concentrations and RRFs (Relative Reduction Factors) for SIPs in the non-attainment areas. *Atmospheric Environment*, *44*(17). https://doi.org/10.1016/j.atmosenv.2010.02.038

Lamb, R. G., & Seinfeld, J. H. (1973). Mathematical Modeling of Urban Air Pollution General Theory. *Environmental Science and Technology*, *7*(3). https://doi.org/10.1021/es60075a006

Lusk, E., Huss, S., Saphir, B., & Snir, M. (2009). MPI: A message-passing interface standard Version 3.0. *International Journal of Supercomputer Applications*, *8*(3/4).

Matam, K., Krishna Bharadwaj Indarapu, S. R., & Kothapalli, K. (2012). Sparse matrix-matrix multiplication on modern architectures. *2012 19th International Conference on High Performance Computing, HiPC 2012*. https://doi.org/10.1109/HiPC.2012.6507483

Mebust, M. R., Eder, B. K., Binkowski, F. S., & Roselle, S. J. (2003). Model-3 Community Multiscale Air Quality (CMAQ) model aerosol component 2. Model evaluation. *Journal of Geophysical Research D: Atmospheres*, *108*(6). https://doi.org/10.1029/2001jd001410

Nvidia, W., Generation, N., & Compute, C. (2009). Whitepaper NVIDIA's Next Generation CUDA Compute Architecture. *ReVision*, *23*(6).

Quinnell, E., Swartzlander, E. E., & Lemonds, C. (2008). Bridge floating-point fused multiply-add design. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *16*(12). https://doi.org/10.1109/TVLSI.2008.2001944

Quinnell, E., Swartzlander, E. E., & Lemonds, C. (2015). Floating-point fused multiply-add architectures. In *Computer Arithmetic: Volume III*. https://doi.org/10.1142/9789814651141

Ristov, S., Gusev, M., & Velkoski, G. (2014). Optimal block size for matrix multiplication using blocking. *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014 - Proceedings*. https://doi.org/10.1109/MIPRO.2014.6859580

Simon, H., Baker, K. R., & Phillips, S. (2012). Compilation and interpretation of photochemical model performance statistics published between 2006 and 2012. In *Atmospheric Environment* (Vol. 61). https://doi.org/10.1016/j.atmosenv.2012.07.012

Søvde, O. A., Prather, M. J., Isaksen, I. S. A., Berntsen, T. K., Stordal, F., Zhu, X., Holmes, C. D., & Hsu, J. (2012). The chemical transport model Oslo CTM3. *Geoscientific Model Development*, *5*(6). https://doi.org/10.5194/gmd-5-1441-2012

Sulatycke, P. D., & Ghose, K. (1998). Caching-efficient multithreaded fast multiplication of sparse matrices. *Proceedings of the 1st Merged International Parallel Processing Symposium and Symposium on Parallel and Distributed Processing, IPPS/SPDP 1998*, *1998-March*. https://doi.org/10.1109/IPPS.1998.669899

Wang, W., Bruyere, C., Duda, M., Dudhia, J., Gill, D., Kavulich, M., Keene, K., Chen, M., Lin, H.-C., Michalakes, J., Rizvi, S., Zhang, X., Berner, J., Ha, S., & Fossell, K. (2017). *WRF Version 3.9 User's Guide*. https://www2.mmm.ucar.edu/wrf/users/docs/user_guide_V3/user_guide_V3.9/ARWUsersGuideV3.9.pdf

Whitehead, N. (2011). *Precision & Performance: Floating Point and IEEE 754 Compliance for NVIDIA GPUs*.

Winter, M., Parger, M., Mlakar, D., & Steinberger, M. (2021). Are dynamic memory managers on GPUs slow?: A survey and benchmarks. *Proceedings of the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPOPP*. https://doi.org/10.1145/3437801.3441612

Wong, D. C., Pleim, J., Mathur, R., Binkowski, F., Otte, T., Gilliam, R., Pouliot, G., Xiu, A., Young, J. O., & Kang, D. (2012). WRF-CMAQ two-way coupled system with aerosol feedback: Software development and preliminary results. *Geoscientific Model Development*. https://doi.org/10.5194/gmd-5-299-2012

Yu, S., Mathur, R., Schere, K., Kang, D., Pleim, J., Young, J., Tong, D., Pouliot, G., McKeen, S. A., & Rao, S. T. (2008). Evaluation of real-time PM2.5 forecasts and process analysis for PM2.5 formation over the eastern United States using the Eta-CMAQ forecast model during the 2004 ICARTT study. *Journal of Geophysical Research Atmospheres*, *113*(6). https://doi.org/10.1029/2007JD009226

Z. Jacobson, M., & Turco, R. P. (1994). SMVGEAR: A sparse-matrix, vectorized gear code for atmospheric models. *Atmospheric Environment*, *28*(2). https://doi.org/10.1016/1352-2310(94)90102-3

Zhang, H., Chen, D., & Ko, S. B. (2019). Efficient Multiple-Precision Floating-Point Fused Multiply-Add with Mixed-Precision Support. *IEEE Transactions on Computers*, *68*(7). https://doi.org/10.1109/TC.2019.2895031

Zhang, H., Chen, G., Hu, J., Chen, S. H., Wiedinmyer, C., Kleeman, M., & Ying, Q. (2014). Evaluation of a seven-year air quality simulation using the Weather Research and Forecasting (WRF)/Community Multiscale Air Quality (CMAQ) models in the eastern United States. *Science of the Total Environment*, *473–474*. https://doi.org/10.1016/j.scitotenv.2013.11.121

Tables

Table 3.1. Computing time for RBFEVAL, RBJACOB, RBCECOMP, and RBSOLVE subroutines for CMAQ and CMAQ-CUDAv1.0. The time was measured in seconds and was the average of a CMAQ timestep for 2,100 and 10,000 BLKSIZE. CTD (copy to device) is the time for transferring data from host to device. CTH (copy to host) is the time for copying the results from the device to the host. KER (kernel) is the GPU computing time.

	RBFEVAL (s)			RBJACOB (s)			RBDECOMP (s)			RBSOLVE (s)		
	CTD	KER	CTH	CTD	KER	CTH	CTD	KER	CTH	CTD	KER	CTH
CMAQCUDAv1.0 2,100 BLKS	3.8 $*10^{-3}$	2.5 $*10^{-3}$	3.9 $*10^{-4}$	1.2 $*10^{-2}$	1.4 $*10^{-2}$	1.3 $*10^{-2}$	1.1 $*10^{-2}$	1.1 $*10^{-2}$	9.8 $*10^{-3}$	9.9 $*10^{-3}$	1.6 $*10^{-3}$	1.1 $*10^{-2}$
CMAQ 2,100 BLKS	0	4.0 $*10^{-3}$	0	0	1.2 $*10^{-2}$	0	0	3.8 $*10^{-2}$	0	0	4.1 $*10^{-3}$	0
CMAQCUDAv1.0 10,000 BLKS	2.4 $*10^{-2}$	1.9 $*10^{-3}$	1.7 $*10^{-3}$	5.0 $*10^{-2}$	7.6 $*10^{-3}$	5.8 $*10^{-2}$	5.3 $*10^{-2}$	1.3 $*10^{-2}$	4.5 $*10^{-2}$	4.6 $*10^{-2}$	1.7 $*10^{-3}$	5.3 $*10^{-2}$
CMAQ 10,000 BLKS	0	2.3 $*10^{-2}$	0	0	5.6 $*10^{-2}$	0	0	1.9 $*10^{-1}$	0	0	2.1 $*10^{-2}$	0

Table 3.2. Improvement of data transfer rate over PCIe generations.

Generation	Year of Release	Data Transfer Rate	Bandwidth x1	Bandwidth x16
PCIe 3.0	2010	8.0 GT/s	1.0 GB/s	16 GB/s
PCIe 4.0	2017	16.0 GT/s	2.0 GB/s	32 GB/s
PCIe 5.0	2019	32.0 GT/s	4.0 GB/s	64 GB/s
PCIe 6.0	2022	64.0 GT/s	8.0 GB/s	128 GB/s
PCIe 7.0	2025	128.0 GT/s	16.0 GB/s	256 GB/s

Figures

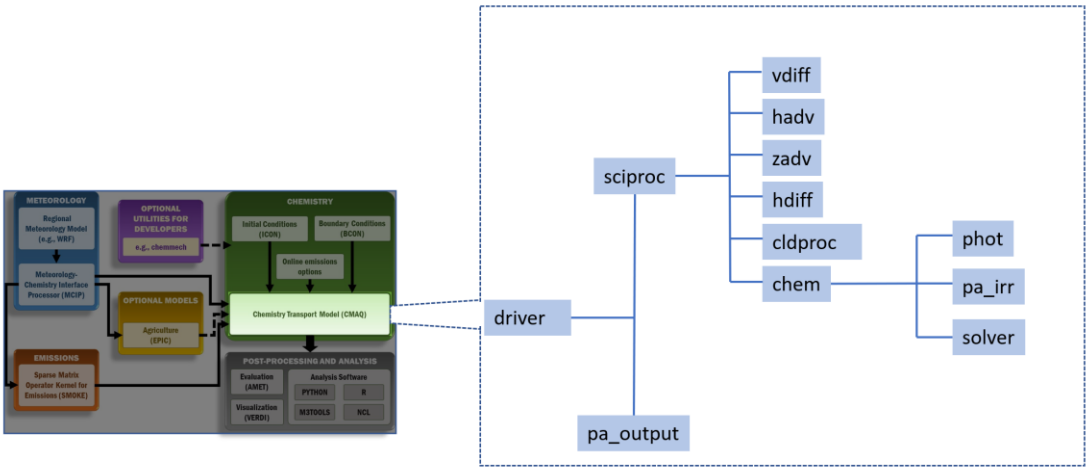


Figure 3.1. CMAQ's CCTM science process modules.

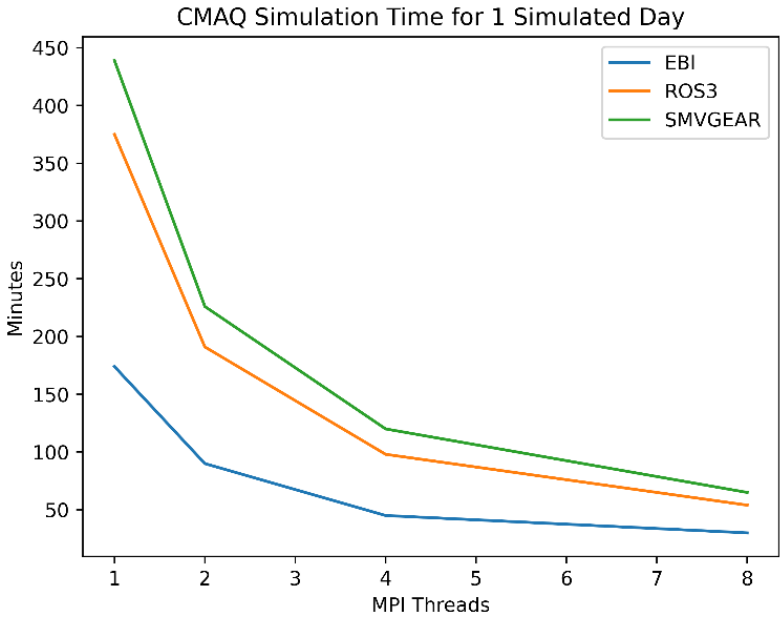


Figure 3.2. CMAQ simulation time for 1 simulated date were carried out using EBI (blue), ROS3 (orange), and SMVGEAR (green) solver with different number of MPI threads.

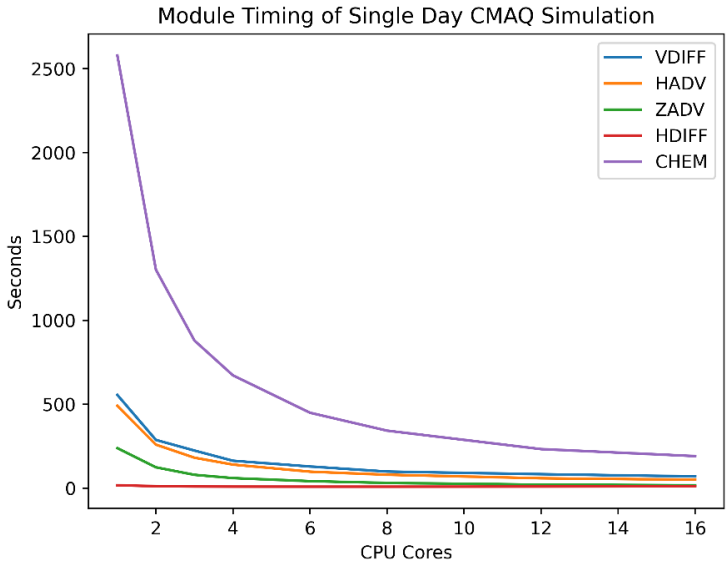


Figure 3.3. Module timing of a single simulated day for five science modules in CMAQ. Gas phase chemistry (purple line) is the slowest module across all CPU cores.

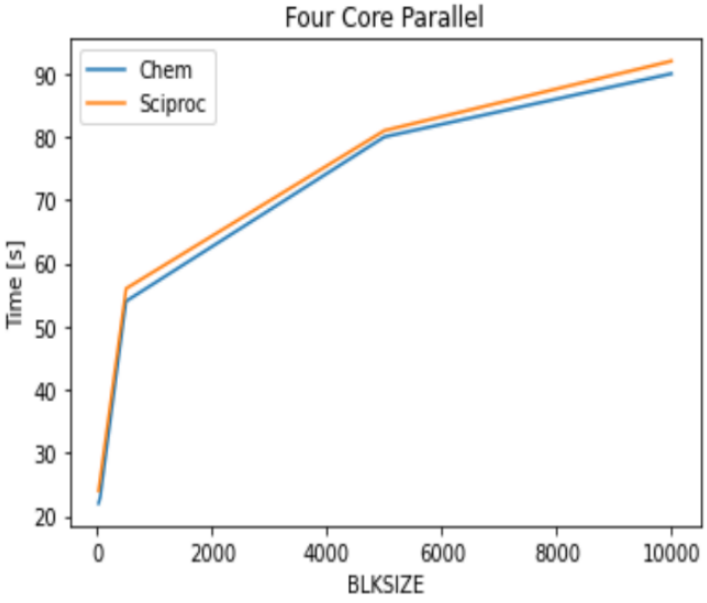


Figure 3.4. Effect of BLKSIZE on CMAQ's science processes per simulation timestep. The blue line is gas-phase chemistry process, and the orange line is the entire science processes

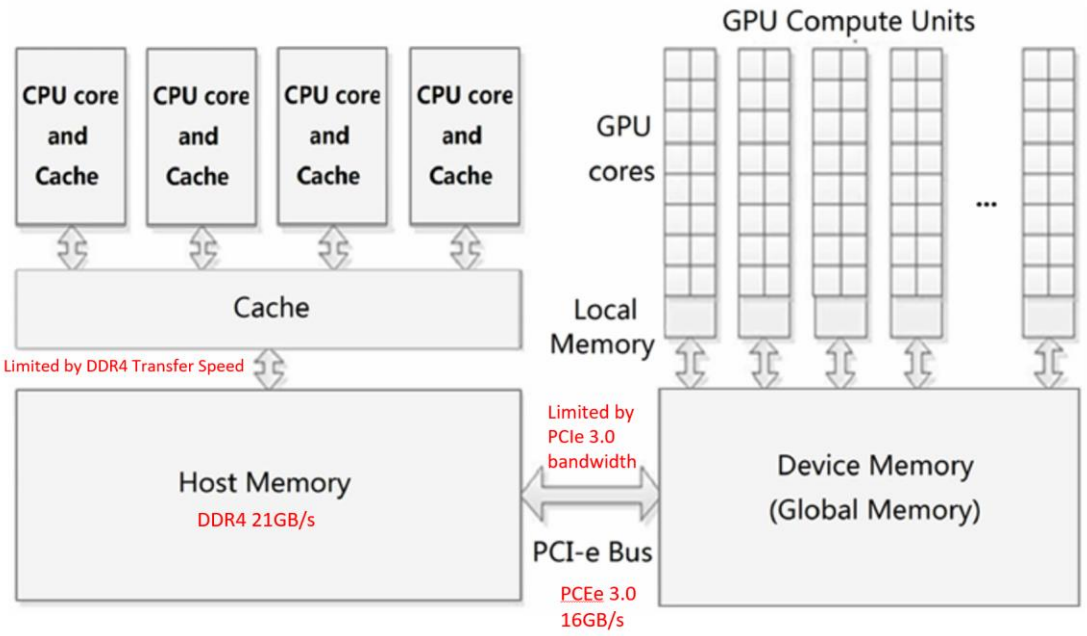


Figure 3.5. Scheme of a computer with a GPU.

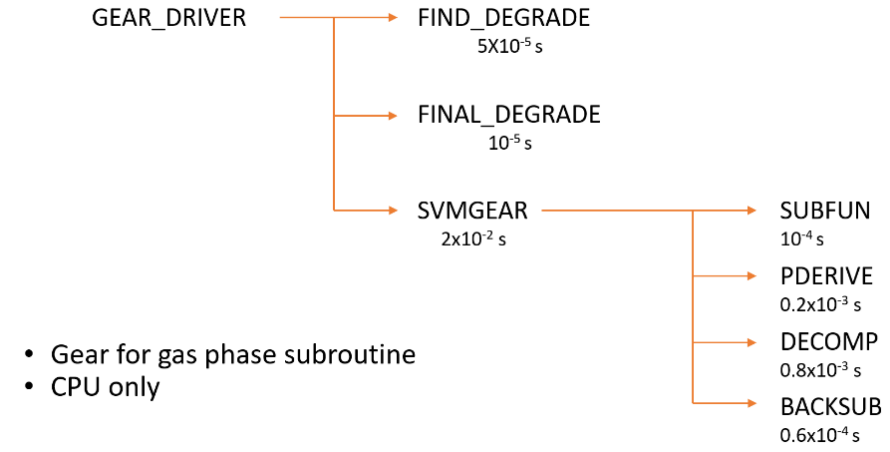


Figure 3.6. Timing of GEAR subroutines.

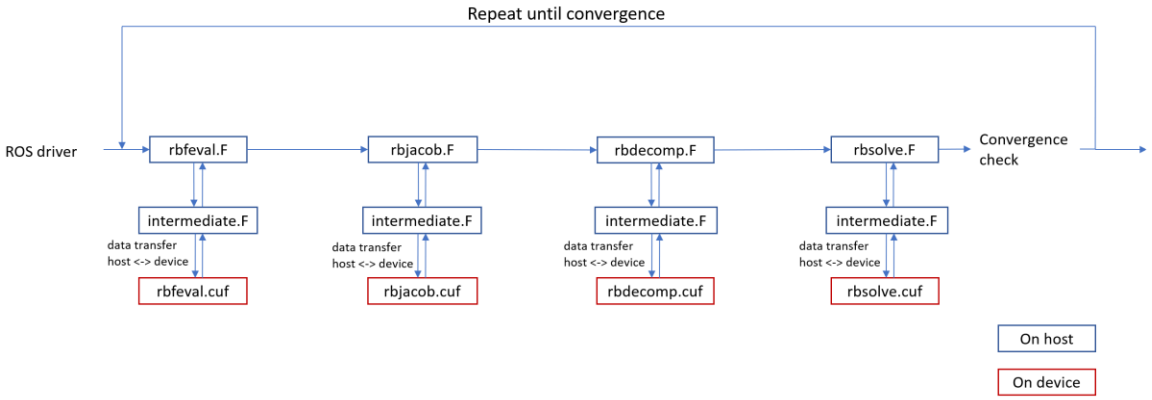


Figure 3.7. CUDA Rosenbrock solver block diagram for CMAQ-CUDAv1.0. The blue blocks are executed using the CPU (host), and the red blocks are executed using GPU (device). Because of the different compilers, the .cur and .F can communicate through intermediate.F.

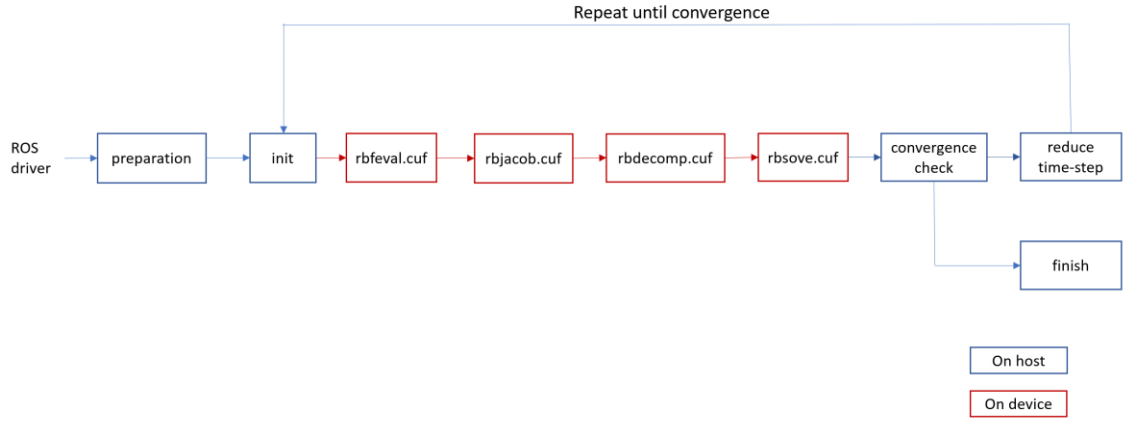


Figure 3.8. CUDA Rosenbrock solver block diagram for CMAQ-CUDAv2.0. The blue blocks are executed using the CPU (host), and the red blocks are executed using GPU (device). The four subroutines (red blocks) operate on the GPU without requiring data transfer between each subroutine.

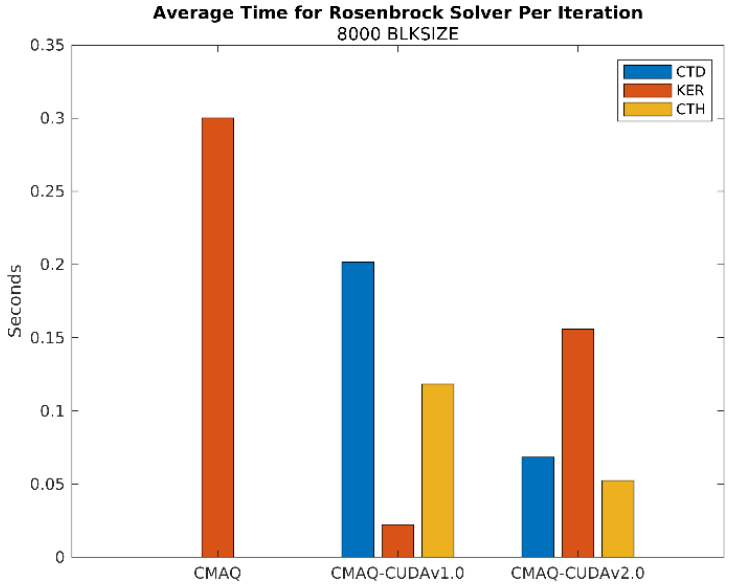


Figure 3.9. Average time for Rosenbrock solver for one iteration step for CMAQ-CUDAv1.0, CMAQ-CUDAv2.0, and CMAQ. Blue bars are the time to copy data to the device (CTD), orange bars are the actual computing time (KER), and yellow bars are the time for copying data from the device to the host (CTH).

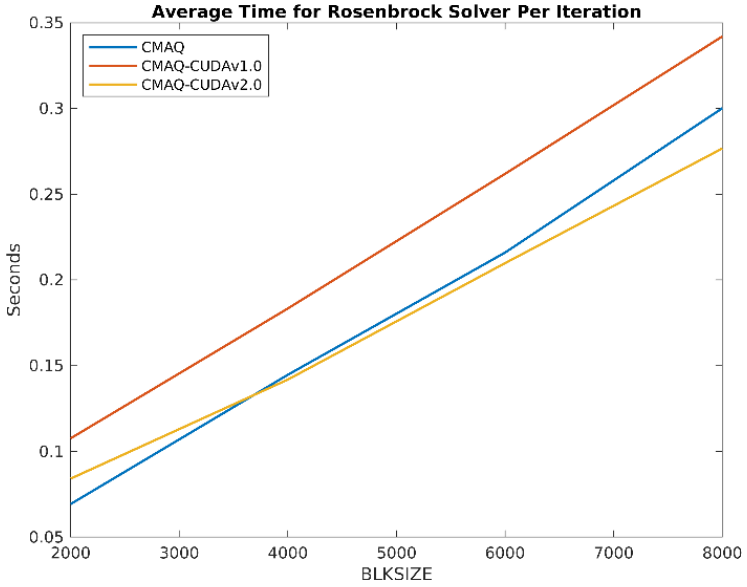
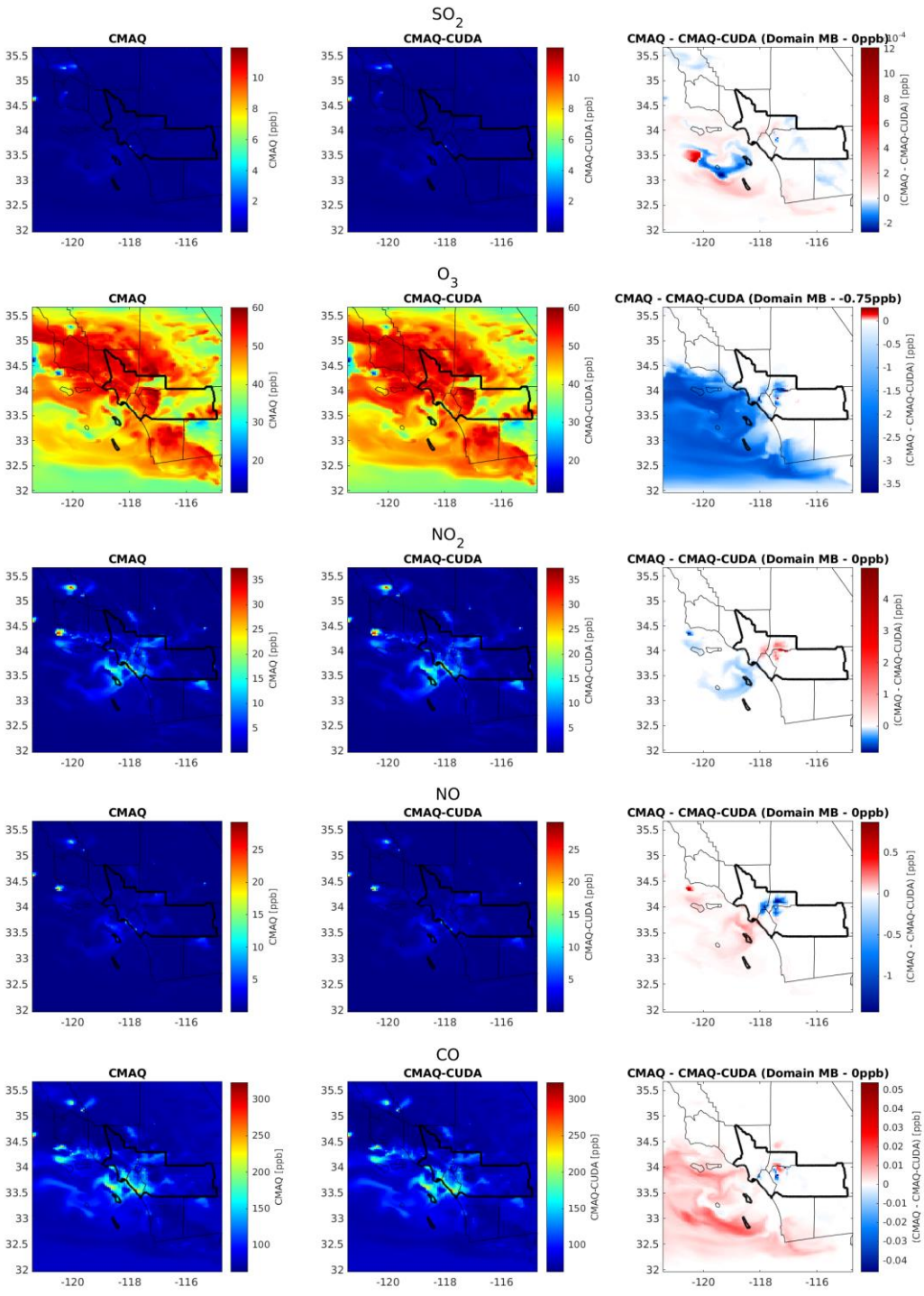


Figure 3.10. Average time for Rosenbrock solver for one iteration with different BLKSIZEs. Blue line is convention CMAQ, orange line is CMAQ-CUDAv1.0, and yellow line is CMAQ-CUDAv2.0.


SO2
CMAQ
CMAQ-CUDA
CMAQ - CMAQ-CUDA (Domain MB - 0ppb)
CMAQ [ppb]
CMAQ-CUDA [ppb]
(CMAQ - CMAQ-CUDA) [ppb]
O3
CMAQ - CMAQ-CUDA (Domain MB - -0.75ppb)
NO2
NO
CO

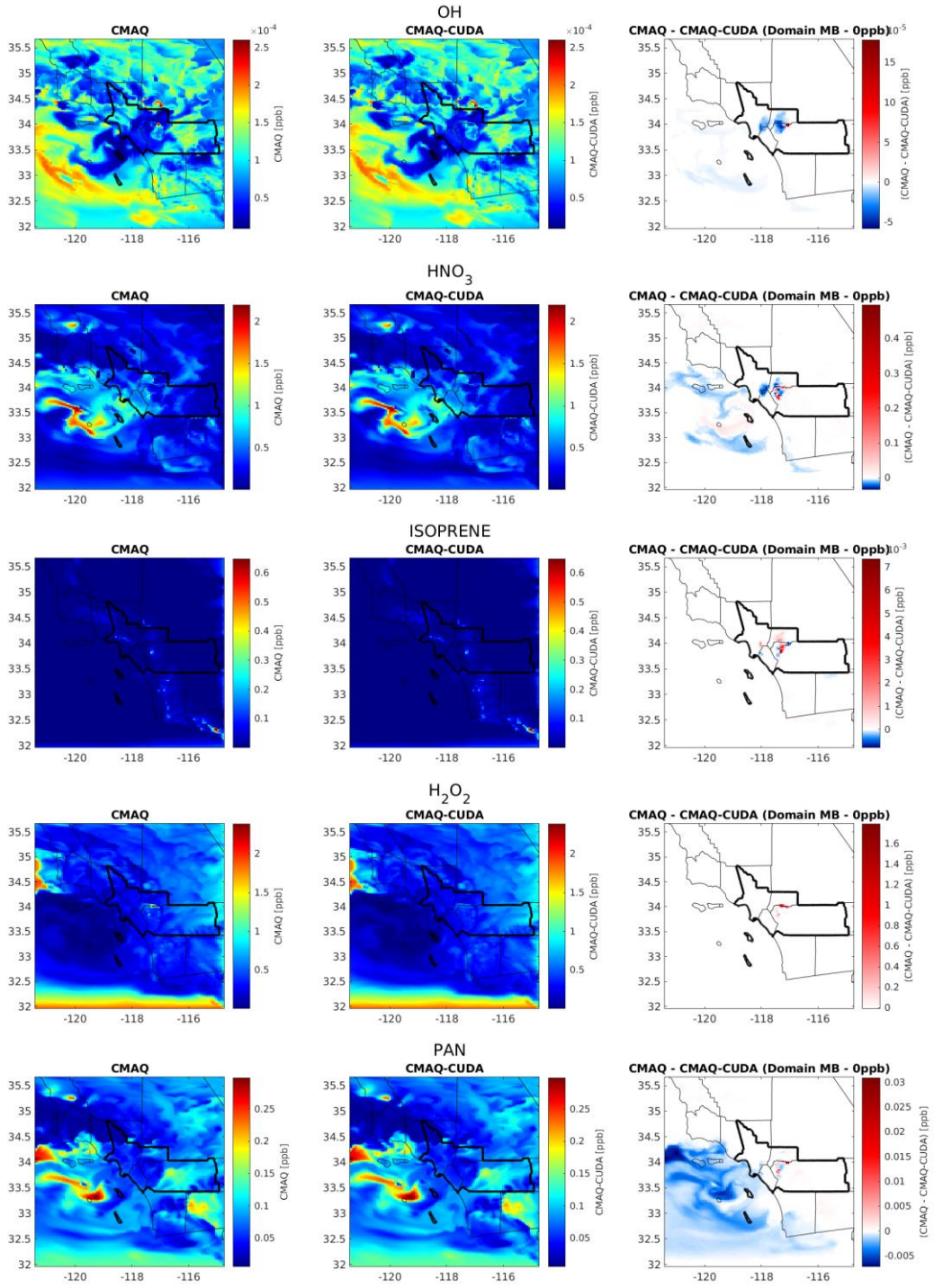


Figure 3.11. The outputs after 24-hour simulation of CMAQ-CUDA and CMAQ. The results were compared among common species (SO_2, O_3, NO_2, NO, CO, OH, HNO_3, Isoprene, H_2O_2, and PAN). The left panels are CMAQ simulation, the middle panels are CMAQ-CUDA simulation, and the right panels are the differences between CMAQ and CMAQ-CUDA in ppb.

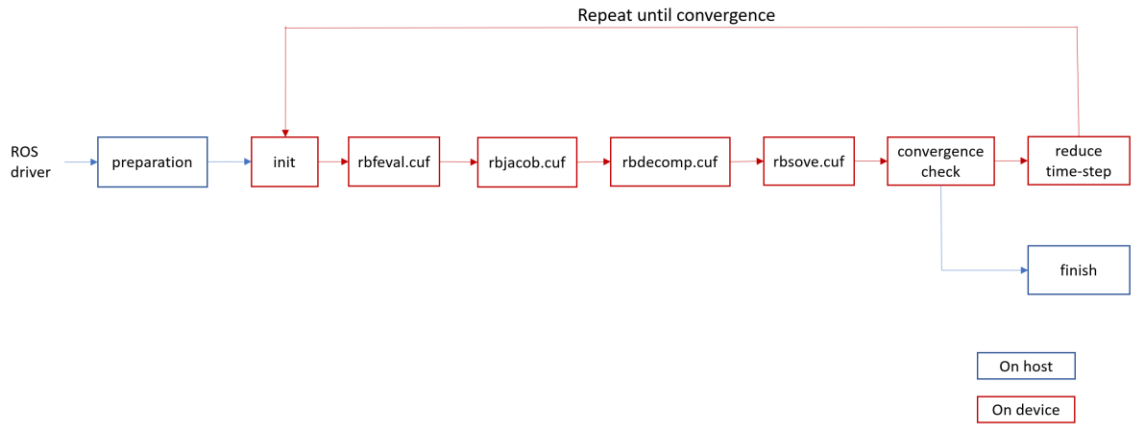


Figure 3.12. CUDA Rosenbrock solver block diagram for CMAQ-CUDAv3.0. The blue blocks are executed using the CPU (host), and the red blocks are executed using GPU (device). The convergence check is ported to the kernel to optimize the transferring time. The outputs from the kernel are the final solutions.