Integrating Machine Learning and Geo-statistics for High-Resolution Surface Ozone Mapping: A Case Study in Arizona

by

R. J. Erickson

B.A. Physics, University of Santa Barbara, 2020

B.A. Geography - GIS Emphasis, University of Santa Barbara, 2020

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Art

Department of Geography

2025

Committee Members:

Guofeng Cao, Chair

Rachel Isaacs

Stefan Leyk

Erickson, R. J. (M.A., Geography)

Integrating Machine Learning and Geo-statistics for High-Resolution Surface Ozone Mapping: A Case Study in Arizona

Thesis directed by Prof. Guofeng Cao

Modeling surface ozone (O_3) concentrations at high spatial resolutions is essential for informed decision-making surrounding air pollution regulation, healthy ecologies, and biological cycles within Earth's many unique micro- and macro-biomes. Continued abundance of air pollution vastly changes atmospheric compositions based on relative O_3 concentrations via numerous ecological factors and anthropogenic trends. On the surface, O_3 varies significantly over minimal changes in urban landscapes, interacting with many oxide-related bio-geochemical cycles. Existing geospatial datasets generated through remote sensing and numerical models (e.g. GEOS-Chem, WRF-Chem) cannot capture the high resolution variations of surface O_3 concentrations. Machine learning (ML) methods have been used to improve surface O_3 mapping. However, traditional ML methods, such as random forests and gradient boosting methods, tend to ignore or simplify complex spatial patterns inherent to geographic phenomena and processes. This thesis proposes a novel and practical solution to integrate geospatial machine learning models with a geostatistical regression kriging (RK) framework to account for geospatial patterns and uncertainty. A sample of daily average maximum surface O_3 (DAMSO) concentrations in parts per billion (ppb) is presented at a 250 m resolution with RK enhancements to five common machine learning models: adaptive boost, gradient boost, extreme gradient boost, random forest, and multi-layered perceptron. The proposed method is found to be a promising solution for the incorporation of geo-statistical data into these and more advanced methods such as deep learning and artificial intelligence (AI). This methodology is readily accessible and reproducible across all areas of interest (AOIs) that have access to emerging GIS Big Data sources.

Dedication

This work is dedicated to my grandparents, Georgia and Rolland Dockham. The many road trips, life lessons, and overwhelming love for Earth inspired me to pursue Geography. May they rest in peace.

Acknowledgements

I would like to acknowledge my Brother; Jonah, Sister-In-Law; Florence, Kin; Demitri, Josh, and Halle; Advisor; Guofeng, Committee Members; Stefan and Rachel, the Department of Geography at the University of Colorado, Boulder, and its Cohort of 2021. I would also like to thank Stefan Leyk, Colleen Reid, Lori Peak, and Heather Yocum for respectively providing additional funding for this work through the CU Population Center (CUPC), Institute of Behavioral Sciences (IBS), Natural Hazards Center (NHC), and North Central Climate Adaptation Science Center (NC CASC). Without such spectacular guidance, mentorship, and friendship, I would not have succeeded.

Contents

Chapter

Appendix

Tables

Table

Figures

Figure

Chapter 1

Introduction

Historical analyses of atmospheric and ecological analyses have shown ozone (O_3) can influence the evolutionary potential of numerous systems on Earth (Manzini et al. 2024). Atmospheric chemistry and urban civilizations are significantly influenced by numerous catalytic cycles involving interfacial oxidation (Chapleski et al. 2016) from surface O_3. Greenhouse gas emissions emitted through both natural and anthropogenic sources can directly and indirectly affect the frequency of O_3 concentrations (Q. Liu et al. 2022). The ongoing regulation, monitoring, and development of air pollution control initiatives is essential to reinstate natural cycles and ensure safe exposure levels to surface O_3 (American Lung Association 2023; EPA 2020). Recent research has shown that O_3 can facilitate redox reactions in both laboratory and natural environments, serving as a product and component of several chemical pathways (Chapleski et al. 2016; Xing et al. 2016). It is a highly reactive, unstable triatomic molecule that is involved in a myriad of oxygenation processes essential to life on Earth (Stuhr, Bayer, and Von Wangelin 2022).

Anthropogenic sources have significantly affected ozone cycles by altering concentrations within various biomes in unnatural ways (Chauhan, Gupta, and Liou 2023; Flynn et al. 2014; Rovira, Domingo, and Schuhmacher 2020). As climate change continues to abet abnormal temperatures and weather patterns, the creation of high-resolution predictive models, such as those presented in this thesis, becomes increasingly critical to address the rapid rate of urbanization (Balk et al. 2018; EPA 2020; Iglesias et al. 2021). This thesis aims to provide a novel methodology that more effectively incorporates geostatistical relationships into contemporary numerical modeling ap-

proaches for urban modeling via regression kriging (RK). The predictions are applied to three of the most populous counties in Arizona for surface (O_3) due to the intricacies of its presence around urban areas.

1.1 Machine Learning/Artificial Intelligence Predictions Require Incorporation of Geo-statistical Data

Machine learning and artificial intelligence methods have revolutionized a myriad of disciplines with GIScience by enabling the analysis of massive datasets newly developed by modern Big Data (Bughin 2016). Tree-based ensembles, boost algorithms, and neural networks can effectively learn linear and nonlinear relationships for ozone modeling. This has been done with numerous predictor variables such as temperature, height of the planetary boundary layer (PBL), nitrogen dioxide (NO_2), and carbon monoxide (CO) (Y. Liu et al. 2018; Xiong, Xie, Mao, et al. 2023; Huang, C. Zhang, and Bi 2017). However, these models typically treat each pixel or grid cell as independent observations, often neglecting spatial autocorrelation, which could be accounted for to improve model performance.

The initial results for geographic fields with ML ensembles appear sharp, yet they often fail to reflect the underlying continuity of atmospheric fields influenced by both natural and anthropogenic processes. In contrast, geostatistical methods, such as kriging, are utilized explicitly for spatial correlation. RK combines a known deterministic trend with a kriged residual to create a surface prediction. However, these trends must be related over short distances, and kriging alone cannot capture non-linear trends. When used for variables such as chemical or certain meteorological trends, there can be a large amount of complexity within a small distance. This has recently been solved with Chemical Transport Models (CTMs) which are installed to ML ensembles as linearly correlative features to represent high spatial resolutions (P. Cheng et al. 2022; Flandorfer 2019; Xiong, Xie, Mao, et al. 2023), omitting RK. CTMs are often costly in terms of development strategies, computation times, and still depicts the most error in urban areas.

Therein lies a notable methodological gap: CTMs encode the underlying physical processes

but are both expensive and coarse; ML methods can use them to capture nonlinearities for higher resolutions but overlook spatial dependencies; and RK which models spatial dependencies but cannot adequately address complex dynamics. This thesis proposed that combining the predictive power of ML models with geo-statistical relationships can be improve the predicted outcome from the established trend by considering error stemming from spatial heterogeneity. Such a concept has already shown considerable promise in modeling $PM_{2.5}$ for the contiguous United States (US) (Y. Liu et al. 2018). Applying this to surface O_3 is a testimony to the value of these models and necessity for including RK into them when modeling similar surface trends. The necessity of this stems from the lack of high resolution surface ozone representations and historical commitment to better understanding air pollution and its potential harm to human health.

1.2 Why Surface Ozone? – Relations to Urban Air Quality

In 1955, the United States enacted the Air Quality Control Act, dedicating national resources to atmospheric monitoring and addressing air pollution concerns that emerged following the scientific advances of the Space Race in the 1950s (EPA 2020). Richard Nixon, then President of the United States (US), facilitated the establishment of the Environmental Protection Agency (EPA) and the National Oceanic and Atmospheric Administration (NOAA) to urgently combat the emerging threat of air pollution (Nixon 1970). After three short decades, new environmental movements arose from the unprecedented economic growth of that era. Regulatory policies developed in the early 1990s, which initially hindered certain sectors, have ultimately demonstrated long-term health and economic benefits as noted by Ambec and Barla (2002). Geographic Information Science (GIScience) has been dedicated to effectively realizing the environmental goals established in the early 1970s.

GIScience has provided innovative methods to document the history of Earth through digital transformations of the newly available resources, addressing trends both locally and globally to solve complex questions observed on Earth. From micro- to macro-scales, geographers specialize in observations and predictions that span planetary to microbial systems, thereby weaving intricate

knowledge of space and time into comprehensive displays (Goodchild 1992). With respect to surface O_3 concentrations, depictions tend to still pose incorrect movement in and around urban areas. These can become exacerbated at any time due with significant spatial heterogeneity due to increased human densities, such as rising surface O_3 exposures due to lockdown events of the recent COVID-19 pandemic (Chauhan, Gupta, and Liou 2023; Meo et al. 2021; Staehle et al. 2022).

When large amounts of ultra-violet (UV) radiation interact with these sources, O_3 begins to form, leading to numerous effects on both organic and inorganic materials exposed to it (Chapleski et al. 2016; Brauer et al. 2024). Surface O_3 has been identified as a significant pollutant contributing to the Global Burden of Disease (GBD) (Brauer et al. 2024) but not as a full burden. Its subtle consequences can only be fully understood through careful consideration of its formation and exposure (Abdullah et al. 2017; Carvalho et al. 2022). Varying concentrations of O_3 have been found to produce numerous beneficial and adverse effects on oxide systems due to interactions with surrounding ecosystems (W. Zhang et al. 2022; EPA 2020). As the stratosphere attains its natural state, surface O_3 can be diminished due to increased UV absorption, restoring concentrations to potentially healthy levels. Studies accumulating similar data have identified several links between urbanization and populations exposed to unhealthy concentrations of surface O_3 exceeding standards established by national entities (EPA 2020; S. Liu et al. 2022).

1.3 A Tale of Two Layers

To better understand the harmful effects of surface O_3, it is essential to examine the underlying principles that influence its formation in both layers of Earth's atmosphere. Stratospheric O_3 protects the Earth's surface by absorbing UV rays, thereby maintaining appropriate levels of it (Hodzic and Madronich 2018). Due to interactions with the chemical compositions present in this layer, O_3 naturally follows seasonal cycles that correlate with the Earth's axial tilt and its distance from the sun. O_3 is highly reactive and decomposes rapidly, increasing the likelihood of reactive oxygen species (ROS) existing in any environment where it is present (EPA 2020; Place et al. 2023). At low, consistent levels, it is generally beneficial to the natural environment, serving

as a semi-catalyst for certain biological and chemical systems (Krasensky et al. 2017). Researchers have found it challenging to trace the patterns of surface O_3 (i.e., 2 m above the ground) due to the numerous interactions it has with these systems.

O_3 was first detected in the stratosphere affecting incoming light. Due to the chemical mixing within the atmosphere, the upper part of the troposphere contains O_3 as well. Since the O_3 in the stratosphere does not absorb all UV radiation (else there would be no life on Earth), residual UV rays that reflect off the Earth's surface facilitate various reactions, leading to investigations into the impacts of surface ozone on ecological processes (Claeyman et al. 2011; M. Lin et al. 2012; J. Gao et al. 2016). High concentrations of O_3 near Earth's surface tend to affect populations outside of highly industrialized zones characterized by dense vegetation, as the primary byproduct of plant processes is carbon dioxide (CO_2) (Sadiq et al. 2016). Ground-level O_3, when combined with other reactive gases, tends to correlate with the air quality experienced by human communities (Gaudel et al. 2018; Schultz et al. 2017). This presents exciting challenges for models to overcome as they become increasingly complex and widely available.

Sand, smoke, volcanic plumes, dust, and various gaseous components, including clouds, can significantly impact the operations and complexity of O_3 cycles, as demonstrated in numerous studies (Venkanna et al. 2015). The precursors to these phenomena pose risks to human health, as highlighted by policies aimed at limiting exposure to pollutants such as particulate matter (PM), CO, and nitrogen dioxide (NO_2), as recommended by the CDC, EPA, and WHO (CDC 2024). Reducing emissions of pollutants, including carbon dioxide CO_2, formaldehyde (HCHO or CH_2O), methane (CH_4), and other nitrous oxides (NO_x), tends to decrease surface O_3 reactions. This reduction occurs by lowering the likelihood of these reactions taking place in populated areas characterized by high albedo (Wu et al. 2023).

1.3.1 Surface Ozone Exposure and Transport

While most hazardous pollutants are emitted directly, complex pollutants such as surface O_3 form as byproducts and later become constituents of other chemicals. O_3 is recognized as

a secondary pollutant (M. Lin et al. 2012; Venkanna et al. 2015), serving as both an ingredient and a product of chemical processes. Even slight changes in wind speeds, environmental conditions (ranging from coastal to arid), and ecological elements (such as available greenspace versus concrete) can influence the formation and degradation of O_3 (Meng et al. 2022; Xing et al. 2016). The data provided by these, and numerous other studies, offer insights into the historical transport of O_3 offer insights into urban, suburban, and rural development if given the proper constraints for modeling. For example, information provided by the EPA (2025) includes in situ monitor data collected since the inception of in 1970. The EPA communicates regulates air quality under the Clean Air Act by regulating five pollutants with an Air Quality Index (AQI) using the collected data (EPA 2023).

Ground level O_3 is one of the most common air pollutants with an AQI in the US, and it interacts directly with the other four pollutants: particle pollution (or particulate matter (PM_x)), CO, sulfur dioxide (SO_2), and NO_2. Furthermore, the trajectory of future emissions has a direct influence on policy decisions, which can affect various socioeconomic statuses (SES) at different rates due to the inherent interactions with vegetation (B. Li et al. 2024; Meo et al. 2021; Sadiq et al. 2016). When analyzed over long periods of time, insights from similar studies have shown O_3 to be potential indicators of environmental injustice due to the heterogeneity shown with anthropogenic sources (Adebayo-Ojo et al. 2022; Tang et al. 2024; Tian et al. 2024). Researchers use effective exposure statistics for O_3, such as Disability-Adjusted Life Years (DALYs) for public health, to assign exposures to individuals using a distance-to-nearest-monitor methodology (Rovira, Domingo, and Schuhmacher 2020). These methods illustrate how even small increases in concentrations can elevate certain negative health outcomes over an extended period. These assessments must be conducted for each study or cohort, taking into account each participant or location. A study encompassing multiple states requires the analysis of a significantly larger number of counties and subsequent data points, which are necessary for conclusive analysis.

Additionally, careful consideration of confounding factors, such as heat, hazards, and oxide-based pollutants, is crucial to properly account for the long-range transport of O_3. Many recent modifications to heat-related mortality studies now account for potential O_3 exposures after Reid

et al. (2012) revealed the complex relationship between temperature, ozone, and mortality with Directed Acyclic Graphs (DAGs). Bi-conditionally, a directed graph is considered a DAG if and only if it can be topologically ordered, the vertices are in a linear order and the resulting order is consistent with all edge directions. Utilizing DAGs incorporate fine-level quantitative information about vulnerable populations and is essential for developing effective intervention strategies for many pollutants, particularly in regard to modeling O_3 exposures over large areas. Dr. Reid established numerous DAGS for heat, O_3 exposure, and temperature and revealed that associations with and without O_3 exposure as a variable requires different analytical techniques. The initial strategy behind DAGs is reminiscent of the laws of geography used in this thesis. Chapter **??** reveals how methodology in Chapter **??** and results in Chapter **??** lay the foundations for data and techniques as inspiration from DAG analyses.

1.4 Why There? - Motivation and Concluding Insights

Proper consideration of surface O_3 exposure, particularly in relation to heat and pollutants, has demonstrated associations with several adverse health outcomes, as identified by public health researchers. Worsening respiratory symptoms, negative birth outcomes, increased overall mortality, and neurological disorders are just a few negative outcomes thought to be exacerbated by unsafe exposures to surface O_3 exposure (Carvalho et al. 2022; Turner et al. 2016; Zhao et al. 2021). Furthermore, studies conducted worldwide, including but not limited to those by Brown-Steiner and Hess (2011) and Chapleski et al. (2016) and Y. Cheng et al. (2018), have identified nitrogen oxides (NO_x) and volatile organic compounds (VoCs) as significant contributors to these health issues when O_3. Many of these interactions arise from anthropogenic sources (Chauhan, Gupta, and Liou 2023). Upon closer examination, interactions with various levels of ground-level O_3 exposure have demonstrated a lasting negative impact on human health. This includes reductions in life expectancy, hospitalizations, abrupt changes to normalized atmospheric chemistry, and alterations in related oxide-based cycles (J. Li et al. 2023; Wu et al. 2023; Barzeghar et al. 2020).

This study was developed after arduous investigation into various O_3 mechanics for exposure

analysis in a separate area of interest (AOI). The dual formation of surface O_3 involves titrations with various chemicals and complex, non-linear associations that originate closer to common pollutant sources and are later reincorporated into the environment at some distance. As a result, these O_3 reactions are challenging to identify and effectively model for decision-making and regulatory entities (Huang, C. Zhang, and Bi 2017). The many reactions taking place around O_3 make it difficult to isolate a universal feature set for prediction. Coastal cities typically exhibit stronger associations of surface O_3 with natural vegetation, which provides non-spatial heterogeneity in comparison to anthropogenic sources, reflecting conditions that were present prior to human development (Cai et al. 2019). In contrast, Phoenix and Tucson (PHOTUC) are characterized by arid climates that demonstrate significant spatial heterogeneity, with constituents linked to human development. This is influenced by dry deposition from local agricultural practices and related emissions.

Existing surface O_3 products are often generated at coarse resolutions (1–10 kilometers (km)), which obscures intra-urban heterogeneity driven by land use, meteorology, and emissions (M. Lin et al. 2012) often seen throughout AOIs like PHOTUC. ML methods often fail to highlight critical heterogeneity from anthropogenic sources and interactions with natural disasters at finer spatial scales (Nawaz et al. 2023). These models result in either too broad or strict of an established trend for used over a large area. While Chemical Transport Models (CTMs) embed process-based knowledge for finer resolutions, they tend to be computationally intensive and typically generate larger near-surface errors for highly reactive species. This thesis addresses these gaps by combining (i) physically informed features derived from satellite data with (ii) geo-statistical regression kriging (RK) to re aggregate the prediction to 250 meters (m). The overall goal is to achieve a robust, high-resolution mapping technique that emphasizes spatial heterogeneity and quantifies predictive uncertainty. PHOTUC offers a small AOI with unique challenges to overcome for generic mapping and processing of surface O_3. Completion of the work done here can be expanded to many other AOIs that have more access to private data and prominent training features for surface O_3.

1.5 Structure of Thesis

A statistical model and residual Kriging methodology are proposed to estimate surface O_3 with a spatial resolution of 250 m, which is suitable for urban analysis (Y. Wang et al. 2023). This thesis aims to compare the performance of five ML/AI methods: adaptive boost, gradient boost, extreme gradient boost, random forest, and multi-layered perceptron, with enhancements provided by RK forecasted on O_3 uncertainty to further develop a Python-based air pollution modeling library. Beginning with a comprehensive literature review on the formation of O_3, numerous aspects of the reaction can affect local ecologies. These factors are discussed in detail to provide a robust foundation for feature creation and model tuning based on scientific evidence. A section dedicated to data sources and materials offers further justification for the features best suited for ML/AI ensembles and their integration with the RK method. The comprehensive combination of these techniques generates daily high spatial resolution rasters for the urban areas of Phoenix and Tucson in Arizona. These cities are predominantly located in two counties, with a third county located between them. The methodology employed in this thesis is characterized as a case study that utilizes the scientific method, geo-statistical analysis, and applications of the three main laws of geography to contemporary ML/AI modeling methodologies.

Chapter 2

Literature Review

A literature database was utilized in this thesis to provide a general background on the formation of O_3, supporting the need for development of surface O_3 models in urban environments. Figure 2.1 illustrates the increasing interest in this area of research since the establishment of the Environmental Protection Agency (EPA) in 1970. Observations from earlier work investigating

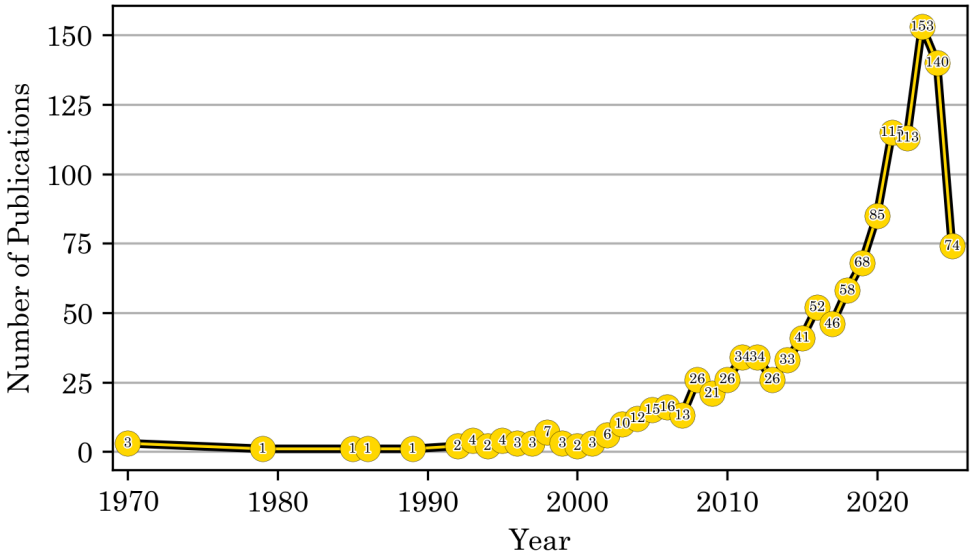


Figure 2.1: There is a notable increase of publications for this area of research since the establishment of the Environmental Protection Agency (EPA) in 1970. Work from 2019 and on were focused on with some literature published before 2000 (Farkas 1979; Goodchild 1992; S. He and Carmichael 1999) being kept primarily due to their methodologies. Their work remains relevant and can be found in numerous disciplines related to surface O_3.

stratospheric O_3 have also shown that chlorinated compounds can evolve into reactive radicals (Cl, ClO) that subsequently catalyze the destruction of stratospheric O_3. In the 1980s, research

unveiled catalytic tropospheric O_3-destroying cycles persist until the initial activation reservoirs are depleted (Farkas 1979). Overtime, it was found to be the inevitable cause of the stratospheric O_3 depletion event in the Arctic (Nadzir et al. 2018). Recently, these stratospheric O_3 mixing ratios were observed to be restored from a minimum of 5 ppb to 10 to 13 ppb, showing the benefits of continued monitoring and policies which reduce the effects of climate change.

Arctic O_3 loss serves as an indicator of short-range tropospheric warming due to stratospheric cooling as its presence wanes (Minghu Ding et al. 2020). This cooling can be linked to broader climate feedback mechanisms within the troposphere, driven by both natural and anthropogenic factors, such as the effects of ice albedo and greenhouse warming, respectively. The latter process, in excess, has been associated with ice melt and sea level rise (Girach et al. 2023). O_3 primarily absorbs UV radiation and plays a complex role in atmospheric chemistry by modulating the concentrations of other trace gases, including reactive chlorine and nitrogen species. This activity positions O_3 as a catalytic agent within stratospheric cycles (Hodzic and Madronich 2018). Its simplicity, characterized by three oxygen molecules, presents numerous complexities to the subject matter at hand.

Modeling both short- and long-term trends requires intuitive knowledge of these processes occurring at satisfactory resolutions. This literature review aims to leverage a wide range of resources, including professional CUB guidance, coursework, and Big Data sources, to enhance understanding of overall surface O_3 formation and interactions. Given the vast potential of computer science, this project adheres to two informal principles from the fields of computer science and military doctrine: Keep It Stupid Simple (K.I.S.S.) and Proper Planning and Preparation Prevents Piss Poor Performance (the 7P's). Streamlined access to the data enables the utilization of a multitude of resources for project implementation across various fields. A thorough yet concise analysis of contemporary O_3 reaction-based models from diverse disciplines will be conducted to develop a deep understanding of surface O_3 reactions, independent of their occurrence.

Literature was also gathered from colleagues and coursework completed during the writing of this thesis. The models employed are expected to be consistent across most cases of complex

adaptive systems (CAS). This literature review aims to utilize a wide range of resources, including professional CUB guidance, coursework, and Big Data sources, to develop a comprehensive understanding of overall surface O_3 formation and interactions within similar CAS. A thorough yet concise analysis of contemporary O_3 reaction models from diverse disciplines will be conducted to foster a deeper understanding of surface reaction formation, irrespective of their specific occurrence. These systems typically exhibit numerous predictive features that are similar, albeit lacking an identifiable unifying principle. Consequently, significant multicollinearity is anticipated among the key features associated with O_3 (Alvarez-Mendoza, Teodoro, and Ramirez-Cando 2019). To address this issue, methods were developed in this section to combine certain features; however, not all were utilized in the final model. This will be further elaborated upon in Chapter **??**.

2.1 Search Methods And Literature Database Creation

The synthesized literature employed the extensive academic resources of the University of Colorado, Boulder (UCB) to conduct a comprehensive investigation into O_3 mechanisms and processes associated with air pollutants. A Python script accessed the API of each database using the categorization of key terms in Table 2.1. With access to a substantial body of academic literature through the CUB library, EBSCOhost, and Web of Science, various documents were selected based on keywords identified during literature reviews conducted throughout coursework. Individ-

Topic	Category	Search Terms	EBSCO	WoS
1	O_3	surface ozone, ground ozone, O3, ozone	112,673	177,569
2	Models	boosting, machine learn, deep learn	5,140	8,833
3	Ecology	public health, pollution, chemistry	49,081	68,666
4	Human	mortality, injur*, illness*, hospital*	8,716	16,200
5	Risk	dispropo*, vulner*, risk*, burden*	21,125	48,387
6	Prediction	predict*, air qual*, air chem*, model	51,508	90,932
7	Transport	transport*, trajector*, circulat*, advection*	17,520	31,633
Total Unfiltered Count			265,763	442,220

Table 2.1: Initially, all categories served as a single search parameter to yield an idea of the vast information available for this work. Combinations of these terms in SQL, like (T1) OR (T2), allowed for concise, precise, and accurate catagorization of literature for this thesis.

ual counts from each search are presented in Table 2.1. Citations gathered from these searches were imported into Python and Zotero for proper formatting and de-duplication. Many works utilized throughout this thesis, including the introduction, were sourced through this process, while others were obtained from informational meetings and coursework.

The combinations of terms that follow topic 1 were employed to establish patterns within the abstracts and titles of the literature, thereby facilitating the categorization of the collected material. This categorization is further analyzed in this chapter, providing a formal foundation for the development of ozone terminology and atmospheric modeling as a comprehensive discipline. The final grouping of key terms can be found in 2.2 with 160 documents being selected for this thesis. All reviewed literature was accessed using the associated DOI with CUB credentials. For

Set Combination	**EBSCO**	**WoS**
All sets	29	196
T1, T2, T6, T7	436	964
T1, T2, T6	2,365	4,750
T1, T3, T4, T5, T7	350	1,238
Literature used in thesis		160

Table 2.2: Over 1,000 unique documents related to O_3 were identified across both databases. The database was reduced to 179 sources for use in this review and thesis.

this thesis, over 1,000 unique documents related to O_3 were identified across both databases. The reviewed literature encompassed various disciplines, including Environmental Sciences, Ecology, Meteorology, Atmospheric Sciences, and Public Health. Each of these fields concerns surface O_3 reactions and has employed relevant techniques that are beneficial to this thesis, such as Abdullah et al. (2017), Manzini et al. (2024), L. He et al. (2024), Turner et al. (2016), and Xiao Lu et al. (2019).

Upon review of the the database, this thesis identified that many sources utilized various numerical models and CTM to simulate, forecast, or analyze O_3 behavior through sensitivity analyses (e.g., Kleinert, Leufen, and Schultz (2020), Emmons et al. (2010), and S. Ma et al. (2021)). The results of the search also indicates a significant portion of the literature emphasizes the quantification and analysis of O_3 through various exposure and chemical transport statistics. Figure

2.2 illustrates a Python package called **wordcloud** where commonly used terms such as ozone, concentration, surface, pollutant, model, data and study are prominent throughout the collected literature. This illustration is based on the titles, abstracts, and keyword terminology of each source in the database before reduction to the final count of around 180 sources. The word cloud highlights common factors such as VOCs, precursors, and temperature, emphasizing frequently referenced O_3 precursors and meteorological conditions noted earlier in Chapter 1.2. Terms like year, period, daily, summer, trend, and long term indicate potential temporal resolutions required for this study. Furthermore, terms such as simulation, analysis, method, measurement, and trend

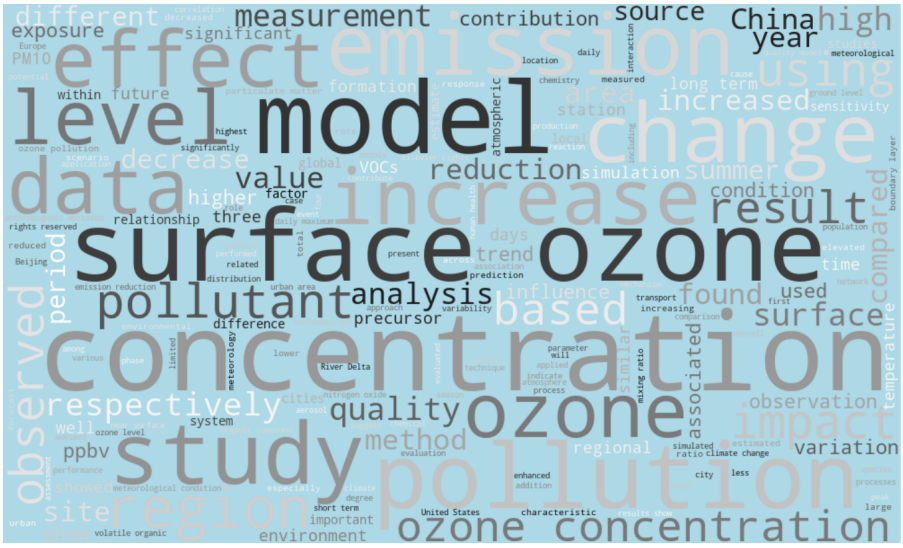


Figure 2.2: Abstracts, titles, subjects, and **wordcloud** in Python were combined to produce this out of interest to the researcher. Modeling strategies came from terms highlighted in this word cloud. Terms like year, period, daily, summer, trend, and long term indicate potential temporal resolutions required for this study. Words such as simulation, analysis, method, measurement, and trend underscore the reliance on data-driven models, as well as statistical and simulation techniques for reporting O_3 trends.

underscore the reliance on data-driven models, as well as statistical and simulation techniques for reporting O_3 trends. Geographical terms like China, city, urban, station, area and United States pinpoint key regions and study sites, particularly in East Asia and urban settings. Each literature source gathered was further generalized into transportation mechanics, modeling techniques, and the impact of O_3 to guide the researcher through a comprehensive journey of historical air pollution

modeling.

Modeling mainly comprised of papers that included keywords from topics 1 and 2 in Figure 2.2. Transportation relates to topics 1 and 7. The remaining categories were associated with health, featuring overlapping terms such as effect, impact, change, quality, emissions, exposure, and climate. These terms reflect O_3's relevance to air quality, human health, and climate interactions, bolstering the categories they were assigned to. Further analysis of the literature indicated a recent increase in interest regarding O_3 production at the surface during the SARS COVID-19 pandemic (Chauhan, Gupta, and Liou 2023; Meo et al. 2021; Y. Pan et al. 2024; Staehle et al. 2022). A majority of these factors are noted to exist in the area of interest (AOI) within this study. The shared terminology was utilized in Figure 2.3 to gauge the availability of this work in modern academia. The various

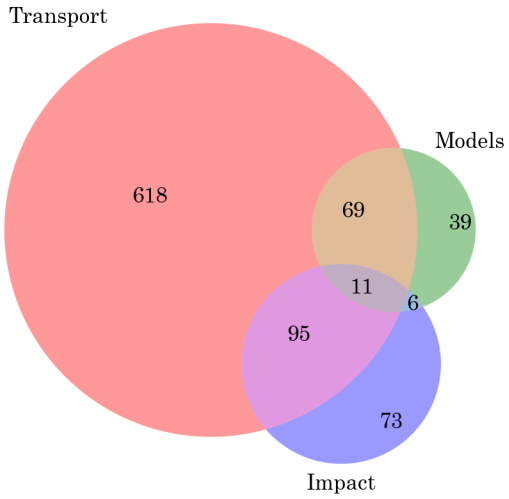


Figure 2.3: The Venn-Diagram highlights the categorization of literature from the collected material. Using **sklearn** in Python, individual categories Transport, Models, and Impact were created to seperate literature into the core reasoning, methods, and overall necessity of these models, respectively. The literature in-between combinations of the three main categories were mainly used throughout the thesis.

interactions that ozone has with both organic and inorganic materials constitute complex processes on surfaces in urban environments, significantly influencing these interactions (Stuhr, Bayer, and Von Wangelin 2022). Like most naturally occurring atmospheric gasses, surface O_3 displays cyclical

increases resulting from interactions within anthropogenic spaces (Gaudel et al. 2018). The summaries of O_3 models, exposure methods, and ecological interactions are crucial to understanding which models are necessary for consideration for the final outcome.

2.2 Surface Ozone Formation

Unlike most pollutants, surface O_3 tends to non-uniformly increase at locations distant from both non-maintained and excessively maintained spaces (Choi et al. 2012). Literature interpretation also reveals that the rates of O_3 reactions exhibit both seasonal and anthropogenic patterns over very short distances. Recent studies have demonstrated that surface O_3 formation follows these seasonal patterns, with concentration fluctuations that consistently increase in urban areas over time due to interactions with UV radiation and volatile organic compounds (VOCs) (Napi et al. 2021). The constituents and drivers of O_3 significantly impact urban layouts, as geophysical systems contribute to elevated temperatures in densely populated areas.

When the gap between urban density and available greenspace is substantial, surface O_3 reactions tend to occur in regions characterized by high biological activity or those vulnerable to natural hazards associated with heat and aerosol movement, such as storms, heat waves, and volcanic eruptions (Brown-Steiner and Hess 2011). Although the latter may not be present in most areas of the US, communities located downwind of golf courses, urban parks, and similar anthropogenic constructions may experience elevated concentrations of O_3 due to these transport mechanisms (Girach et al. 2023). These works show naturally high concentrations of O_3 in summer and lower concentrations in winter.

Moreover, additional sources suggest that these concentrations have gradually increased over time in urban areas, primarily due to the influences of solar radiation and VoCs (Place et al. 2023). This pattern is particularly notable during milder winters, influenced by the overall geography of the AOI. The formation of surface O_3 can be induced by stable tropospheric O_3 cycles during sustained high temperatures and oxide-based constituents, as noted by Cai et al. (2019). As an oxidizing agent, O_3 has been found to participate in redox reactions while simultaneously undergoing

reduction due to its molecular instability (e.g. S. He and Carmichael (1999) and Krasensky et al. (2017) and Stuhr, Bayer, and Von Wangelin (2022)).

2.2.1 Modeling Complexities of Urban Areas

The ideal distance for accurately representing surface ozone patterns in urban areas has yet to be formally established (L. Wang et al. 2023), largely due to the intricate molecular interactions of O_3. Large populations can significantly contribute to O_3 formation and are particularly vulnerable to complex variations due to disparities in access to greenspaces and industrialized areas (Y. Pan et al. 2024; Meo et al. 2021). Chemists have identified a specific reaction that occurs when the ozone molecule splits upon absorbing photons; photolysis. Low temperatures, combined with heavily polluted conditions and reduced photolysis, typically result in decreased ozone (O_3) concentrations (S. He and Carmichael 1999; Manzini et al. 2024).

Photolysis is the chemical representation of the light-driven oxidation event known as photosynthesis, or flora. The distribution of urban flora does not conform to standard non-spatial patterns; therefore, pollutants that exhibit spatial non-heterogeneity over short distances must be appropriately correlated with their sources. This thesis notes the many different computational approaches used to describe O_3 to build a complex set of functions for model. These approaches allow for the integration of temporal changes in atmospheric conditions and surface interactions, enhancing the model's accuracy and reliability.

By utilizing various model types, this thesis more effectively represents the complexities of surface O_3 values under changing conditions by considering various mechanics. In addition, an analysis on the impact of various constituents on the performance of these models is conducted. This ensures that the results are comprehensive and applicable across the selected region. The methodology section (Chapter **??**) presents detailed descriptions of data collection, processing techniques, and model validation processes as a result of this literature.

2.3 Chemical Transport Models

Chemical Transport Models (CTMs) are process-based, mechanistic models that provide data for ML ensembles. Recent studies demonstrate that these approaches can be effectively combined to enhance surface ozone (O_3) forecasting accuracy at coarse resolutions, which can subsequently be aggregated to finer resolutions (J.-T. Lin et al. 2012). Typically, CTMs and ML ensembles are situated at opposite ends of the modeling spectrum; however, they can complement one another in the study of pollutants such as surface O_3 (Yu et al. 2018). While ML ensembles do not inherently account for heterological spatial data due to the collinearity it poses, they can be integrated with CTMs. CTMs operate within a mechanistic three-dimensional (3D) Eulerian framework and primarily account for large spatial variation through linear associations over a major trend predicted by the ensemble (Travis and Jacob 2019). They can be integrated with CTMs, which operate within a mechanistic three-dimensional (3D) Eulerian framework and subsequently only accounts for large spatial variation through linear associations over a major trend predicted by the ensemble (Travis and Jacob 2019). This integration has shown promise in estimating surface O_3 concentrations using satellite data (Kang et al. 2021); however, further improvements can be attained through the incorporation of geospatial uncertainty within the machine learning model.

Most CTMs were found to have approximately 8 to 13 parts-per-billion (ppb), with a 5 ppb RMSE associated with their predictions, likely attributed to the modifiable unit area problem (MAUP) encountered when resampling outputs to achieve the desired higher resolution (Travis and Jacob 2019) with complex trends. CTMs separate emissions and transport mechanisms to model the relationships between surface measurements and satellite detections for consideration by the selected trend. They typically incorporate valuable spatial information into the overall error of the associated model by analyzing emissions and localized precursors from specialized databases (J. Gao et al. 2016; S. Liu et al. 2022; Nawaz et al. 2023). This makes CTMs computationally and temporally costly. Most models require extensive access to Big Data systems and developments in technology for accurate depictions of surface trends in a timely manner, spurring the development

of low cost emission sensors (Cavaliere et al. 2023). CTMs that incorporate ML and AI methods for transport can be further enhanced by these monitors by effectively integrating geo-spatially regressed uncertainty from monitored data through RK.

Integrating the RK approach into an ML ensemble with CTM-based data enhances the reliability of these systems by increasing accessibility to relevant monitoring data separate from the emissions database. The transport representations in CTMs are based on continuity equations, encompassing advection, convection, emissions, and detailed interactions in both gas and aqueous phases, along with relevant atmospheric equations and chemistry (Flandorfer 2019). Proper incorporation of RK into these numerical models is essential, as uncertainties in general aerosol models significantly increase when aggregated to higher resolutions (Sofen et al. 2015). CTMs typically utilize meteorological data obtained from remote sensing and employ selected chemical mechanisms to simulate concentrations within approximately 10 kilometers (km) in regions such as China (Dou et al. 2024), India (Chauhan, Gupta, and Liou 2023), and the US (Flynn et al. 2014). These physically coherent fields are suitable for investigating chemical lifetimes, transport, and stratosphere–troposphere cycles of specific chemicals; however, highly reactive molecules, such as O_3, often yield the highest errors. The general laws of physics discussed in Chapter **??** derived from the mechanisms found within CTMs, are utilized to combine linear and non-linear features of O_3 into a single predictor variable.

2.4 Statistical Regression

CTMs have used linear regression to model associations with anthropogenic sources. Linear regression is a well-established statistical method for regression and prediction. Its simplicity, interpretability, and efficiency make it a valuable tool for geographic predictions, especially in binary classification scenarios such as land cover changes, habitat presence, or the classification of environmental hazards. Multi-linear regression has been used for models comparing O_3 production

by season (Napi et al. 2021). Linear regression is typically expressed using the following notation:

$$f(x) = m(x) + b + \varepsilon \tag{2.1}$$

where $m(x)$ is an indirect rise-over-run correlation between the independent variable and a specific feature, b represents the y-intercept, and ε indicates the residual error that each point x deviates from the mean trend. Extending this equation to multiple covariates leads to multi-linear regression:

$$y(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon \tag{2.2}$$

where each β_i represents a weighted value attributed to $y(x)$. There are minimal interpretations to consider, and linear methods typically require little tuning unless additional weighting techniques are employed. However, surface O_3 entails greater complexity. CTMs offer complex relationships that contribute to linear correlations, facilitating effective modeling strategies (Napi et al. 2021). Although linear regression is often applied alongside Principal Component Analysis (PCA) and Artificial Neural Networks (ANN) (Kleinert, Leufen, and Schultz 2020; Kang et al. 2021), O_3 trends cannot be modeled at high resolutions with methods that do not account for complex systems stemming from urban drivers. This thesis seeks to utilize linear scopes for feature analysis, but found that non-linear approaches have seen better results in the recent decade (Cavaliere et al. 2023; Q. Li et al. 2024; Venkanna et al. 2015). Machine Learning (ML) and Artificial Intelligence (AI) methods operate with increased complexity, learning weights by aggregating samples into statistically related bins—a process commonly referred to as boosting.

2.5 Machine Learning

Modern CTM models employ a combination of statistical trends and weighted boosting methods to offer linear combinations of meteorological tendencies for an ML trend. While CTMs provide accurate depictions of atmospheric physics and can utilize ML methods, they do not inherently learn from data. Instead, they serve as features in an ML model and carry their own spatial uncertainties, necessitating corrections from the researcher (Yu et al. 2018; Travis and Jacob 2019; Xiong,

Xie, Huang, et al. 2024). CTMs come at a high computational cost, often resulting in hundreds to thousands of CPU-hours for resolutions of 500 m to 1 km. They are also dependent on emission inventories and parameterizations, which introduce additional uncertainties (C. Chen et al. 2021; Gilliland et al. 2008). Utilizing these to represent resolutions from, 100–300 m is unsuitable due to the uncertainty in model error at these fine resolutions. CTMs such as MOZART-TS1 (Emmons et al. 2010), the Community Earth System Model (CESM (Kan Yi et al. 2016)), and the Community Multiscale Air Quality modeling system (CMAQ (Place et al. 2023)) are grounded in standard representations of tropospheric–stratospheric chemistry.

Most CTMs are tree-based algorithms, as they implement complex trends to create linear associations. Literature comparing regression and tree-based correction ensembles highlights the effectiveness of these learners in ML models, the impact of improper tuning on error rates, and the necessity for adequate database management (Wen et al. 2021; Xiong, Xie, Huang, et al. 2024). Initially, these models struggled to identify appropriate trends for urban areas; however, further refinements made within CTM based features have led to improved performance in these contexts (Djalalova et al. 2010; Staehle et al. 2022; Xiong, Xie, Huang, et al. 2024). All statistical concepts within boosting ensembles are analogous; they involve sorting binned data using predetermined constraints into trees, which are then employed collectively in an ensemble to generate a predictive algorithm. This algorithm is selected based on an outcome that minimizes a loss function, such as the Mean Absolute Error (MAE):

$$\sum_{i=1}^{D} |x_i - y_i| \tag{2.3}$$

where x_i and y_i are the predicted and observed measurements, respectively. Replacing the absolute function with a squared function yields the Mean Squared Error(MSE):

$$\sum_{i=1}^{D} (x_i - y_i)^2 \tag{2.4}$$

Following the same trend, applying a square root function after squaring the values yields the

Root-Mean Squared Error (RMSE):

$$\sum_{i=1}^{D} \sqrt{(x_i - y_i)^2} \tag{2.5}$$

Given their role in model creation, these methods will also be employed in model comparison, as discussed later in Chapter **??**. Although complex in nature, these representations provide powerful insights into systems that extend beyond linear frameworks (R. Cao et al. 2024; W. Wang et al. 2022).

The number of branches and decisions varies within each ensemble and can be adjusted to represent different data distributions (Ko, Cho, and Rao 2022). The arrangement of trees facilitates a range of complex algorithms known as machine learning, with each tree reflecting trends identified during the binning process (Q. Pan, Harrou, and Sun 2023). While normality and non-stochastic conditions are generally preferred, these ensembles incorporate parameters that can adapt to varying factors. This underscores the necessity of mastering these methods and incorporating them into this thesis. Boosting algorithms can be categorized into two primary types: Sequential and Parallel. Both techniques function analogously to series and parallel circuits in electrodynamics; the former learns through iterative data binning, while the latter processes all data simultaneously to identify trends within generated statistical subsets.

2.5.1 Sequential Boosting

Sequential boosting (or bagging) involves training base learners one after another, where each subsequent model tries to correct the errors made by its predecessor. The model assigns higher weights to data points that were previously misclassified, ensuring that future learners focus on the harder cases. This dependency creates a feedback loop that allows the ensemble to iteratively improve its performance. While sequential bagging generally achieves high accuracy, it can be more sensitive to noise and overfitting due to its focus on hard-to-classify examples. To mitigate these issues, linear regularization methods—specifically L1 and L2 regularization techniques—can be integrated to adjust the weights assigned to errors during training. These methods in equations 2.6 and 2.7 are referred to as the least absolute shrinkage and selection operation (LASSO) and Ridge

Regression, respectively. The means for establishing weights base on these methods are base in residual error similar to that of RK:

$$LASSO_{\lambda} = Error(Y - \widehat{Y}) + \lambda \sum_{1}^{n} |w_i| \tag{2.6}$$

$$Ridge_{\alpha} = Error(Y - \widehat{Y}) + \lambda \sum_{1}^{n} w_i^2 \tag{2.7}$$

In this thesis, three main types of sequential boost methods are tested with each regularization technique applied. Adaptive boosting (ADA) sequentially fits trees that focus on residual error by reweighing the data. Gradient boosting (GB) fits trees to linear residuals of prior trees via gradient descent of a chosen loss function. Extreme Gradient Boosting (XGRB) aims to optimize the use of computational resources for gradient-boosted ensembles. In addition to the standard L1 and L2 regularization techniques, specialized methods for error trend estimation have been proposed to enhance performance with datasets that contain outliers. These methods address the unique underlying spatial or temporal correlation structures that can distort global model assumptions if left unconsidered. In addition to the standard L1 and L2 regularization techniques, specialized methods for error trend estimation have been proposed to enhance performance with datasets containing outliers. These methods address the unique underlying spatial or temporal correlation structures that can distort global model assumptions if left unconsidered.

2.5.2 Parallel Boosting

This thesis utilizes the widely recognized parallel boosting technique known as Random Forest (RF). RF ensemble learning methods are primarily utilized for classification and regression tasks involving extensive datasets. Their application in geographic datasets has been thoroughly investigated and validated due to their robustness, accuracy, and capacity to manage complex data. Advancements in computer science, coupled with progress in geographic practices, support the use of RF models in contemporary geographic modeling and prediction systems. By design, they excel at managing high-dimensional datasets and uncovering complex trends by way of image analysis (Wright and Ziegler 2017).

Multiple base learners, typically decision trees, can be trained simultaneously on independent samples from the overall dataset, learning without being influenced by other models. Regression using these methods is achieved through majority averaging of predictions across each tree. This parallel approach benefits from faster training times when computational resources allow for parallel processing. Moreover, the presence of inherently independent features contributes to the reduction of overall variance, making parallel bagging particularly effective in mitigating overfitting, especially in scenarios characterized by high covariance.

Increasing the number of tuned metrics for these models exponentially increases the time for sorting and learning from each sample. Figure 2.4 shows how each tree must be analyzed after subsets and initialization parameters are considered.

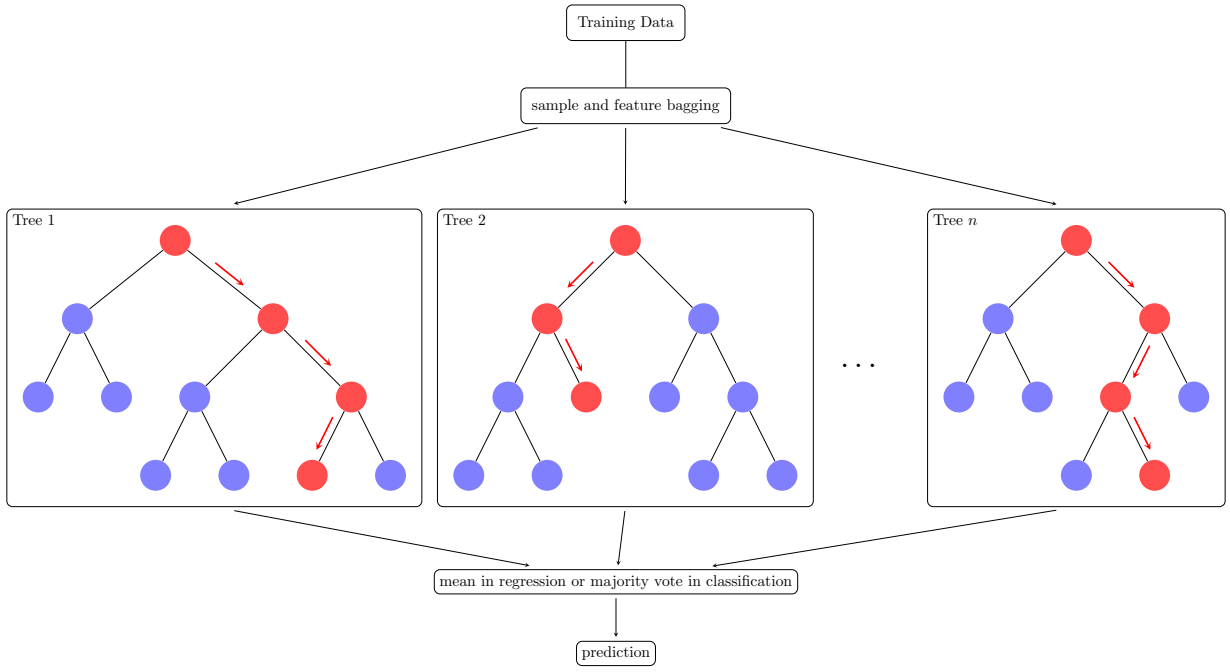


Figure 2.4: The illustration how complex an RF can be. Using many trees typically causes long computation times when training.

This complexity can be quickly determined through GPU integration for processing of the training data. This relatively new technology, and is not available for all machines. Although GPU integration in RF has steadily improved, AI has become nearly fully incorporated into conventional

machines. Stemmed in deep learning methods and activation nodes, AI integrations have natural GPU capabilities due to its prominence.

RF models offer significant advantages for analyzing intricate geographic trends, including high accuracy, resistance to noise, the ability to handle diverse data types, and the capacity to account for nonlinear relationships. Nonetheless, RF has limitations, such as computational complexity, lack of spatial explicitness, and potential challenges in interpretability when dealing with large datasets. While other parallel boosting strategies exist, basic RF strategies are sufficient for the purposes of this thesis. These alternatives are primarily minimal variations of the foundational RF ensemble. Despite the power of RF, computation times can become problematic unless a Graphics Processing Unit (GPU) is utilized (Schulz et al. 2015).

2.6 Artificial Intelligence

The full power of integrated GPU methods has elevated learning techniques, such as neural networks (NN), to the forefront of contemporary modeling systems. NN is a layered, learning-driven architecture that excels at capturing high-dimensional, nonlinear relationships in data through the implementation of activation functions. Before the integration of GPU technology, all processing was managed exclusively by the central processing unit (CPU), which also handled various other computational tasks. By incorporating a GPU into the model's training phase, the CPU is relieved of some processing burdens, allowing it to focus on directing the GPU, which handles data processing. Recent implementations have shown considerable success in modeling complex trends, including $PM_{2.5}$ emissions, disease transmission and hospitalizations, meteorological processes, and land use classification (S. Gao et al. 2021; Huang, C. Zhang, and Bi 2017; Kleinert, Leufen, and Schultz 2020). The introduction of non-linearity through activation functions enables NNs to approximate complex mappings that go beyond simple linear transformations:

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{1 - e^{-2x}}{1 + e^{-2x}} \tag{2.8}$$

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, 2, \ldots, K \tag{2.9}$$

$$Relu(z) = max(0, z) \tag{2.10}$$

Where equations 2.8, 2.9, and 2.10 denote the tangent, softmax, and ReLU activation functions, respectively (Maleki et al. 2019). The integration of these activation functions into GIScience has been pivotal for the simultaneous analysis of numerous layers. This advancement has enabled breakthroughs across various fields through the utilization of multi-spectral imagery (Janowicz et al. 2020).

This project utilized a common neural network architecture known as the Multilayer Perceptron (MLP), which is one of the most basic implementations available in the scikit-learn library (Buitinck et al. 2013). When appropriately trained with a variety of activation functions, layers, and ensemble learning methods, MLPs can produce predictions that are comparable to those made by the human brain (Z. Chen and Z. Zhang 2024; S. Gao et al. 2021; Q. Li et al. 2024). Despite the mathematical complexities involved in selecting activation functions and determining hyper-parameters, advancements in computer science have made the development, tuning, and application of neural networks in geography increasingly accessible.

2.6.1 Defining Complexity: Convolution and Recurrence

Although more complex models, such as Convolutional Neural Networks (CNNs) and Recursive Neural Networks (RNNs) were considered, the lack of proper GPU access necessitated deferring complex implementations of MLP to future work. MLPs offer tunable hidden layers that effectively recognize patterns across the dataset, making them suitable for tabular or flattened data. At their core, all NNs utilize fully connected layers. MLP can initialize pooling and memory retention methods to define various convolutions across designated areas in images. Recurrence can also be induced within indexed images, thereby mimicking complex spatial patterns through remote sensing (W. Zhang et al. 2022). Although MLP does not directly model spatial patterns, it utilizes corrected imagery as inputs to make deductions based on convolutions observed among related pixels. The lattice structure of input parameters depicts organization which MLP learns a series of activation functions to allocate predictions. This approach simulates spatial reasoning by leverag-

ing associated satellite-based lattice structures. Figure 2.5 provides an example of a convolutional layer, illustrating how one of the input features is assigned to nodes associated with each output layer. When sourced from multiple origins, large datasets can exponentially increase the number

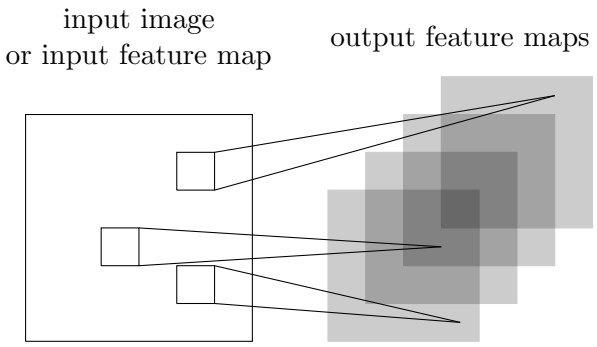


Figure 2.5: Any image (or list like object) can be sorted into a series of outputs for further analysis and sorting to deduce an outcome. This image represents how individual pixels selected at consistent intervals can reveal an initial trend from which activation nodes are created.

of nodes in convolution. Such large datasets are likely to benefit by re-incurring the learned trend from numerous images into a single outcome.

Recurrence in NN models (RNN) are designed for temporal or sequentially indexed data, effectively managing time dependencies. While representative of spatial data due to its basis in a grid structure, these models may make generalizations due to stagnation of dependent variables in imagery. They maintain hidden states that facilitate the integration of past information across an ordered sequence, contrasting with the moving window approach employed by CNNs. Combining CNN properties with an RNN leads to the development of Hierarchical Convolutional Recurrent Neural Networks (HCRNNs), which enhances classification accuracy in multi-spectral, time-varying geospatial datasets. The generalization of these models provides robust directions for this thesis, offering numerous strategies to address various gaps in surface O_3 literature, as Neural Networks (NNs) are grounded in lattice-based structures. MLP combined with spatiotemporal geo-statistical regression can be regarded as a fundamental implementation of an HCRNN, in which the geo-statistical kernels serve as a sequencing strategy. The number of nodes for each of the hidden layers follows the same general structure. The general MLP is depicted in Figure 2.6 where D input

units and C output units will be tested in this thesis. The full tuning strategies and hidden layers for this model are outlined in Chapter **??**. HCRNNs have demonstrated significant potential for

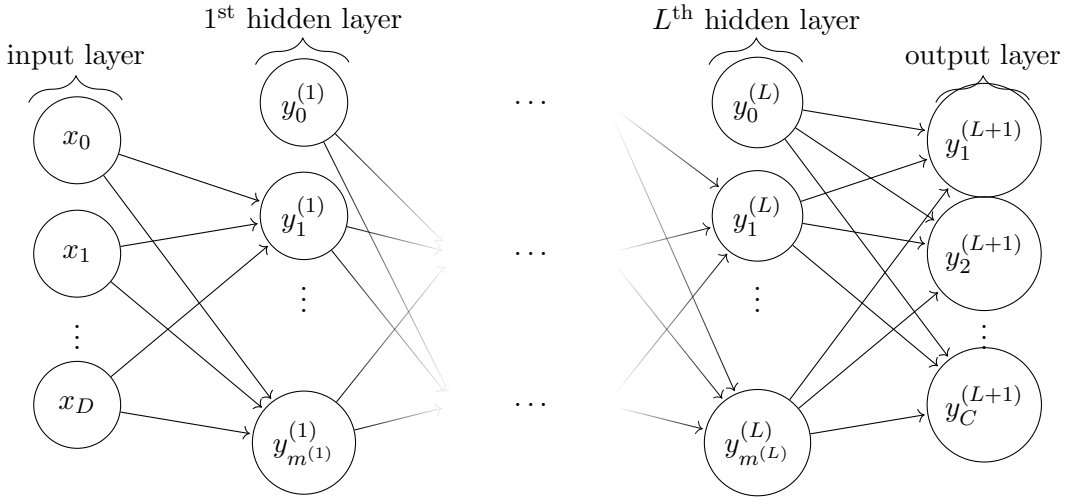


Figure 2.6: The generalized network graph shows a $(L+1)$-layer perceptron with D input units and C output units. The l^{th} hidden layer contains $m^{(l)}$ hidden units. When used with satellite imagery, this lattice notation does incorporate some spatial heterogeneity into the model, but can still reveal error if the learned trend is too general. Each hidden layer can be created form numerous NN combinations.

representing surface O_3 datasets and modeling air pollution trends. However, these models still exhibit substantial geospatial error in dense urban areas (Kleinert, Leufen, and Schultz 2020) due to heterogenetic geo-statistical associations.

2.7 Gaps in Surface Ozone Literature

As with all the ensembles referenced, the integration of geo-statistical uncertainty remains incomplete. Another major omission in current research is the insufficient comprehensive portrayal of raster data illustrating surface O_3 levels across the United States. Often, when dealing with high-resolution data, various specific forms of surface O_3 are present but have not been incorporated into high-resolution raster models that exclude heterogeneous features. Several high-resolution models have been created for the EU, China, India, Iran, and other urban areas that share concerns about the rising rates of surface O_3 reactions (Du et al. 2024; Z. Li et al. 2023; Tian et al. 2024).

There is potential for improvement when this spatial bias is treated as a separate phenomenon

in well-established models. Urban public health case studies can sidestep this issue by using point-based exposure measures derived from environmental aspects of O_3, correlating them with in situ observations. Typically, these studies assign O_3 exposure levels to participants by predicting personal exposure using geo-located addresses (Ekland et al. 2021; H. Ma et al. 2023; Turner et al. 2016). The results of these critical health-related exposure studies could be significantly accelerated by an automated system capable of determining exposures at specific points, thanks to the availability of emerging technology focused on cloud-based Big Data distribution. Studies on exposure often employ identical satellite data, such as the normalized difference vegetation index (NDVI) and temperature estimates, to determine surface O_3 exposure.

As noted retrospectively in Chapter 1.3.1, focusing on surface O_3 may imply that similar pollutant models necessitate additional adjustments depending on the selected study area. If the modeled complex trend approximates the real observed trend, these adjustments can be informed by RK-predicted error estimates when the regressed trend demonstrates spatial heterogeneity. In the case of surface O_3, such heterogeneity might arise from interactions with NO_x and $PM_2.5$, as discussed in Chapter 1.2. These interactions, along with VoCs, naturally occurring atmospheric chemicals, and the distribution of greenspace influenced by policy, suggest that desired features like temperature will often exhibit significant multi-collinearity with VoC emissions related to isoprene (Manzini et al. 2024; Nelson et al. 2023).

With careful consideration, models can accurately depict exposures even during wildfire events (Watson et al. 2019), underscoring the importance of a generalized approach for such representations. Furthermore, health studies have indicated that exposure to even low O_3 levels disrupts the regulation of oxidation-reduction systems, leading to chronic oxidative stress and causing irreversible damage to the cognitive and immune systems in susceptible populations (Barzeghar et al. 2020; Luan et al. 2024). Numerous studies have explored the diverse relationships between short-term and long-term surface O_3 exposure, which may have direct and indirect implications that are still under consideration. This trend could be attributed to the chemical instability of O_3, as oxidative stress is more easily triggered by the presence of O^- from unstable O_3 in O_2 systems (Krasensky

et al. 2017). In the context of health-related risk assessment and policy formation, coupled with rising concerns about the health impacts of climate change, there has been an increased need for high-resolution imagery around 300 m (L. Wang et al. 2023).

The most notable gap in O_3 literature relates to the challenges faced by all ML/AI representations of geospatial data. With the advent of Big Data and advancements in scientific technology, in situ surface measurements have become widely accessible worldwide (Gaudel et al. 2018). However, researchers specializing in hybrid ML/AI methods for geospatial representations often fail to incorporate geo-statistical relationships into their final predictions. Utilizing these measurements during the training process can establish accurate trends based on corrections made by remote sensing satellites. It is essential to recognize that these satellites operate independently of ground-based observations, which can lead to discrepancies in data interpretation. When applied to coarse resolutions, such as areas approximately 1 km and above, errors may be largely negligible due to the expansive regions these resolutions encompass.

Remote sensing data employed in ML ensembles utilize total column estimates to predict intricate interactions at the surface level (W. Wang et al. 2022). Accuracy is maintained within resolutions ranging from 1 km to 10 km; however, this resolution is primarily intended to facilitate global transport modeling. Discrepancies persist in numerical models that utilize tropospheric ozone (O_3 monitoring instruments in both the upper and lower hemispheres of Earth when compared to surface ozone (O_3) representations below 500 m. This thesis argues that both simple and complex numerical models—such as linear models, boosted sampling techniques, and image analytics—can benefit from the incorporation of geo-statistical relationships. The proposed solution aims to enhance the final predicted outcomes of tested numerical models, making them more representative of geographic tendencies.

2.8 Conclusion

This thesis argues that both simple and complex numerical models (i.e., linear, boosted sampling, or image analytics) can benefit from the incorporation of geo-statistical relationships into

their final predictions. This study also examines the integration of these geo-statistical relationships into contemporary models utilizing a regression kriging (RK) model. The models discussed in this chapter enhance the general understanding of the complex processes governing O_3 formation and transport, clarifying the methodologies used for feature establishment and production. As similar models are employed by decision-makers at various scales, the literature provides a framework for identifying notable trends in the training sets, which reflect actions taken to mitigate the impacts of constituents contributing to surface O_3 prominence (Du et al. 2024; F. Wang et al. 2021).

The recent revolution in Big Data, along with advanced modeling techniques, has enabled approaches that extend beyond traditional linear frameworks (G. Cao 2022). The methods developed for this thesis incorporate transport, anthropogenic, and thermodynamical mechanisms into feature bases for urban locations. Modeling techniques for training, validation, prediction, and implementation using the residual kriging method are discussed further in the methods section. Studies such as (Yu et al. 2018; Zhou et al. 2018) utilize CTM-based methods to combine atmospheric dynamics and satellite imagery. Many tropospheric O_3 transport models assume a constant distribution of the gas across the system, resulting in predictions with a resolution of approximately 4 km. CTM models are widely used in various studies due to their methodological robustness; however, significant computational expenses arise from the considerable effort required to optimize these models, with the complexity of the gas in question influencing the extent of these expenses.

For O_3, these models are often essential for ensuring an accurate representation of its transport and distribution. Despite this, improvements in high-resolution surface O_3 modeling are still necessary for both current and upcoming generations of studies. The health impacts and concerns associated with surface O_3 exposure cannot be accurately assessed using CTM-based models. Their coarse results tend to underestimate or overestimate concentrations at the surface, particularly in higher and lower latitudes, respectively (Brown-Steiner and Hess 2011; Kan Yi et al. 2016; Xing et al. 2016). Researchers have developed interpolated exposure rates through temporal analysis of extrapolated statistical learning models, creating unique activity space exposures (Turner et al. 2016). These models possess an extremely fine resolution; however, they represent exposures at the

individual level and not an overall transport model.

Based on the analyzed literature, it was anticipated that the dataset would need to capture the complexity of the ecological factors involved. CAS modeling necessitates multiple transformations to accurately represent the independent variable of choice. Studies that provide a brief overview or focus primarily on surface O_3 suggest that known interactions with both natural and built environments cause non-linear trends. The most suitable models for this project are assumed to be tree-based models and/or neural network models due to this complexity. Trends and transportation mechanisms should be representative of mechanics noted in this chapter.

The models in this thesis are trained on data gathered from similar exposure analyses and satellite observations. Developing a large dataset that integrates these techniques may expedite exposure assignment times by minimizing the coding work necessary to incorporate O_3 as a variable of interest. The culmination of this literature review contributes to the features, ensembles, and residual kriging methodology. This approach utilizes various feature transformations and spatial-temporal analytics to test unique datasets across the AOI. The selection of these features, their respective data sources, and the rationale for their combination into the final model are informed by similar ideas found in existing literature and suggestions from colleagues.

Bibliography

Abdullah, Ahmad Makmom et al. (2017). "The Relationship between Daily Maximum Temperature and Daily Maximum Ground Level Ozone Concentration". In: **Polish Journal of Environmental Studies** 26, pp. 517–523. DOI: 10.15244/pjoes/65366.

Adebayo-Ojo, TC et al. (July 2022). "Short-Term Effects of PM10, NO2, SO2 and O3 on Cardio-Respiratory Mortality in Cape Town, South Africa, 2006-2015". In: **International Journal of Environmental Research and Public Health** 19.13. ISSN: 1660-4601. DOI: 10.3390/ijerph19138078.

Alvarez-Mendoza, CI, A Teodoro, and L Ramirez-Cando (Mar. 2019). "Spatial Estimation of Surface Ozone Concentrations in Quito Ecuador with Remote Sensing Data, Air Pollution Measurements and Meteorological Variables". In: **Environmental Monitoring and Assessment** 191.3. ISSN: 0167-6369. DOI: 10.1007/s10661-019-7286-6.

Ambec, Stefan and Philippe Barla (May 2002). "A Theoretical Foundation of the Porter Hypothesis". en. In: **Economics Letters** 75.3, pp. 355–360. ISSN: 01651765. DOI: 10.1016/S0165-1765(02)00005-8.

American Lung Association (Nov. 2023). **Disparities in the Impact of Air Pollution**. English.

Balk, Deborah et al. (2018). "Understanding Urbanization: A Study of Census and Satellite-Derived Urban Classes in the United States, 1990-2010". eng. In: **PloS One** 13.12, e0208487. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0208487.

Barzeghar, V et al. (Mar. 2020). "Long-Term Trend of Ambient Air PM10, PM2.5, and O3 and Their Health Effects in Tabriz City, Iran, during 2006-2017". In: **Sustainable Cities and Society** 54. ISSN: 2210-6707. DOI: 10.1016/j.scs.2019.101988.

Brauer, Michael et al. (May 2024). "Global Burden and Strength of Evidence for 88 Risk Factors in 204 Countries and 811 Subnational Locations, 1990–2021: A Systematic Analysis for the Global Burden of Disease Study 2021". en. In: **The Lancet** 403.10440, pp. 2162–2203. ISSN: 0140-6736. DOI: 10.1016/s0140-6736(24)00933-4.

Brown-Steiner, B and P Hess (Sept. 2011). "Asian Influence on Surface Ozone in the United States: A Comparison of Chemistry, Seasonality, and Transport Mechanisms". In: **Journal of Geophysical Research-Atmospheres** 116. ISSN: 2169-897X. DOI: 10.1029/2011JD015846.

Bughin, Jacques (Dec. 2016). "Big Data, Big Bang?" en. In: **Journal of Big Data** 3.1. ISSN: 2196-1115. DOI: 10.1186/s40537-015-0014-3.

Buitinck, Lars et al. (2013). "API Design for Machine Learning Software: Experiences from the Scikit-Learn Project". In: **ECML PKDD Workshop: Languages for Data Mining and Machine Learning**, pp. 108–122.

Cai, CX et al. (Mar. 2019). "Simulating the Weekly Cycle of NOx-VOC-HOx-O3 Photochemical System in the South Coast of California During CalNex-2010 Campaign". In: **Journal of Geophysical Research-Atmospheres** 124.6, pp. 3532–3555. ISSN: 2169-897X. DOI: 10.1029/2018JD029859.

Cao, Guofeng (2022). "Deep Learning of Big Geospatial Data: Challenges and Opportunities". en. In: **New Thinking in GIScience**. Singapore: Springer Nature Singapore, pp. 159–169. ISBN: 978-981-19-3815-3 978-981-19-3816-0. DOI: 10.1007/978-981-19-3816-0_18.

Cao, RH et al. (Oct. 2024). "Using Complex Systems Theory to Comprehend the Coordinated Control Effects of PM2.5 and O3 in Yangtze River Delta Industrial Base in China". In: **Stochastic Environmental Research and Risk Assessment** 38.10, pp. 4027–4041. ISSN: 1436-3240. DOI: 10.1007/s00477-024-02791-3.

Carvalho, RB et al. (Aug. 2022). "O3 Concentration and Duration of Exposure Are Factors Influencing the Environmental Health Risk of Exercising in Rio Grande, Brazil". In: **Environmental Geochemistry and Health** 44.8, pp. 2733–2742. ISSN: 0269-4042. DOI: 10.1007/s10653-021-01060-4.

Cavaliere, A et al. (Oct. 2023). "Development of Low-Cost Air Quality Stations for next-Generation Monitoring Networks: Calibration and Validation of NO2 and O3 Sensors". In: **Atmospheric Measurement Techniques** 16.20, pp. 4723–4740. ISSN: 1867-1381. DOI: 10.5194/amt-16-4723-2023.

CDC, U.S Centers for Disease Control (Feb. 2024). **Air Pollutants**. en-us. Information. (Visited on 07/17/2025).

Chapleski, Robert C. et al. (July 2016). "Heterogeneous Chemistry and Reaction Dynamics of the Atmospheric Oxidants, O3, NO3, and OH, on Organic Surfaces." eng. In: **Chemical Society Reviews** 45.13, pp. 3731–3746. ISSN: 0306-0012. DOI: 10.1039/c5cs00375j.

Chauhan, A, SK Gupta, and YA Liou (Apr. 2023). "Rising Surface Ozone Due to Anthropogenic Activities and Its Impact on COVID-19 Related Deaths in Delhi, India". In: **Heliyon** 9.4. ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2023.e14975.

Chen, CH et al. (Mar. 2021). "Comparison of the RADM2 and RACM Chemical Mechanisms in O3 Simulations: Effect of the Photolysis Rate Constant". In: **Scientific Reports** 11.1. ISSN: 2045-2322. DOI: 10.1038/s41598-021-84629-4.

Chen, ZW and Z Zhang (Oct. 2024). "Analysis of Spatiotemporal Variation and Relationship to Land Use - Landscape Pattern of PM2.5 and O3 in Typical Arid Zone". In: **Sustainable Cities and Society** 113. ISSN: 2210-6707. DOI: 10.1016/j.scs.2024.105689.

Cheng, PY et al. (Jan. 2022). "Improvement of Summertime Surface Ozone Prediction by Assimilating Geostationary Operational Environmental Satellite Cloud Observations". In: **Atmospheric Environment** 268. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2021.118751.

Cheng, Y et al. (July 2018). "Estimator of Surface Ozone Using Formaldehyde and Carbon Monoxide Concentrations Over the Eastern United States in Summer". In: **Journal of Geophysical Research-Atmospheres** 123.14, pp. 7642–7655. ISSN: 2169-897X. DOI: 10.1029/2018JD028452.

Choi, Y et al. (2012). "Summertime Weekly Cycles of Observed and Modeled NOx and O3 Concentrations as a Function of Satellite-Derived Ozone Production Sensitivity and Land Use Types over the Continental United States". In: **Atmospheric Chemistry and Physics** 12.14, pp. 6291–6307. ISSN: 1680-7316. DOI: 10.5194/acp-12-6291-2012.

Claeyman, M et al. (2011). "A Thermal Infrared Instrument Onboard a Geostationary Platform for CO and O3 Measurements in the Lowermost Troposphere: Observing System Simulation Experiments (OSSE)". In: **Atmospheric Measurement Techniques** 4.8, pp. 1637–1661. ISSN: 1867-1381. DOI: 10.5194/amt-4-1637-2011.

Djalalova, I et al. (Feb. 2010). "Ensemble and Bias-Correction Techniques for Air Quality Model Forecasts of Surface O3 and PM2.5 during the TEXAQS-II Experiment of 2006". In: **Atmospheric Environment** 44.4, pp. 455–467. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2009.11.007.

Dou, XD et al. (Feb. 2024). "The WRF-CMAQ Simulation of a Complex Pollution Episode with High-Level O3 and PM2.5 over the North China Plain: Pollution Characteristics and Causes". In: **Atmosphere** 15.2. ISSN: 2073-4433. DOI: 10.3390/atmos15020198.

Du, SW et al. (Mar. 2024). "Policy Implications for Synergistic Management of PM2.5 and O3 Pollution from a Pattern-Process-Sustainability Perspective in China". In: **Science of the Total Environment** 916. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2024.170210.

Ekland, J et al. (May 2021). "The Effect of Current and Future Maternal Exposure to Near-Surface Ozone on Preterm Birth in 30 European Countries-an EU-wide Health Impact Assessment". In: **Environmental Research Letters** 16.5. ISSN: 1748-9326. DOI: 10.1088/1748-9326/abe6c4.

Emmons, L. K. et al. (Jan. 2010). "Description and Evaluation of the Model for Ozone and Related Chemical Tracers, Version 4 (MOZART-4)". en. In: **Geoscientific Model Development** 3.1, pp. 43–67. ISSN: 1991-9603. DOI: 10.5194/gmd-3-43-2010.

EPA, Environmental Protection Agency (Apr. 2020). **Integrated Science Assessment for Ozone and Related Photochemical Oxidants**. Assessment. Office of Research and Development.

— (Feb. 2023). **Air Quality Guide for Ozone**.

— (June 2025). **Air Quality System (AQS) API**. https://www.epa.gov/outdoor-air-quality-data/download-daily-data.

Farkas, E (1979). "Surface Ozone Variations in the Auckland Region". In: **New Zealand Journal of Science** 22.1, pp. 63–76. ISSN: 0028-8365.

Flandorfer, Claudia (Jan. 2019). "Evaluation and Comparison of O3 and PM10 Forecasts of ALARO-CAMx and WRF-Chem." eng. In: **Geophysical Research Abstracts** 21, pp. 1–1. ISSN: 1029-7006.

Flynn, CM et al. (Aug. 2014). "Relationship between Column-Density and Surface Mixing Ratio: Statistical Analysis of O3 and NO2 Data from the July 2011 Maryland DISCOVER-AQ Mission". In: **Atmospheric Environment** 92, pp. 429–441. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2014.04.041.

Gao, JH et al. (Feb. 2016). "A Case Study of Surface Ozone Source Apportionment during a High Concentration Episode, under Frequent Shifting Wind Conditions over the Yangtze River Delta, China". In: **Science of the Total Environment** 544, pp. 853–863. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2015.12.039.

Gao, S et al. (Sept. 2021). "Simulation of Surface Ozone over Hebei Province, China Using Kolmogorov-Zurbenko and Artificial Neural Network (KZ-ANN) Combined Model". In: **Atmospheric Environment** 261. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2021.118599.

Gaudel, A. et al. (Jan. 2018). "Tropospheric Ozone Assessment Report: Present-day Distribution and Trends of Tropospheric Ozone Relevant to Climate and Global Atmospheric Chemistry Model Evaluation". en. In: **Elementa: Science of the Anthropocene** 6. Ed. by Detlev Helmig and Alastair Lewis, p. 39. ISSN: 2325-1026. DOI: 10.1525/elementa.291.

Gilliland, AB et al. (June 2008). "Dynamic Evaluation of Regional Air Quality Models:: Assessing Changes in O3 Stemming from Changes in Emissions and Meteorology". In: **Atmospheric**

Environment 42.20, pp. 5110–5123. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2008.02.018.

Girach, Imran A. et al. (Aug. 2023). "Influences of Downward Transport and Photochemistry on Surface Ozone over East Antarctica during Austral Summer: In Situ Observations and Model Simulations." English. In: **EGUsphere**, pp. 1–36. DOI: 10.5194/egusphere-2023-1524.

Goodchild, Michael F. (Jan. 1992). "Geographical Information Science". en. In: **International journal of geographical information systems** 6.1, pp. 31–45. ISSN: 0269-3798. DOI: 10.1080/02693799208901893.

He, L et al. (Jan. 2024). "Effects of VOC Emissions from Chemical Industrial Parks on Regional O3-PM2.5 Compound Pollution in the Yangtze River Delta". In: **Science of the Total Environment** 906. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2023.167503.

He, Shan and Gregory R. Carmichael (Nov. 1999). "Sensitivity of Photolysis Rates and Ozone Production in the Troposphere to Aerosol Properties". en. In: **Journal of Geophysical Research: Atmospheres** 104.D21, pp. 26307–26324. ISSN: 0148-0227. DOI: 10.1029/1999JD900789.

Hodzic, A. and S. Madronich (Dec. 2018). "Response of Surface Ozone over the Continental United States to UV Radiation Declines from the Expected Recovery of Stratospheric Ozone." eng. In: **NPJ Climate & Atmospheric Science** 1.1, N.PAG–N.PAG. ISSN: 2397-3722. DOI: 10.1038/s41612-018-0045-5.

Huang, L, C Zhang, and J Bi (Oct. 2017). "Development of Land Use Regression Models for PM2.5, SO2, NO2 and O3 in Nanjing, China". In: **Environmental Research** 158, pp. 542–552. ISSN: 0013-9351. DOI: 10.1016/j.envres.2017.07.010.

Iglesias, Virginia et al. (July 2021). "Risky Development: Increasing Exposure to Natural Hazards in the United States". en. In: **Earth's Future** 9.7, e2020EF001795. ISSN: 2328-4277, 2328-4277. DOI: 10.1029/2020EF001795.

Janowicz, Krzysztof et al. (Apr. 2020). "GeoAI: Spatially Explicit Artificial Intelligence Techniques for Geographic Knowledge Discovery and Beyond". en. In: **International Journal**

of Geographical Information Science 34.4, pp. 625–636. ISSN: 1365-8816, 1362-3087. DOI: 10.1080/13658816.2019.1684500.

Kan Yi et al. (Dec. 2016). "Response of Global Surface Ozone Distribution to Northern Hemispheric Sea Surface Temperature Changes: Implication for Long-Range Transport." eng. In: **Atmospheric Chemistry & Physics Discussions**, pp. 1–29. ISSN: 1680-7367. DOI: 10.5194/acp-2016-1001.

Kang, Y et al. (Nov. 2021). "Estimation of Surface-Level NO2 and O3 Concentrations Using TROPOMI Data and Machine Learning over East Asia". In: **Environmental Pollution** 288. ISSN: 0269-7491. DOI: 10.1016/j.envpol.2021.117711.

Kleinert, Felix, Lukas H. Leufen, and Martin G. Schultz (Aug. 2020). "IntelliO3-ts v1.0: A Neural Network Approach to Predict near-Surface Ozone Concentrations in Germany." eng. In: **Geoscientific Model Development Discussions**, pp. 1–69. ISSN: 1991-9611. DOI: 10.5194/gmd-2020-169.

Ko, K, S Cho, and RR Rao (Oct. 2022). "Machine-Learning-Based Near-Surface Ozone Forecasting Model with Planetary Boundary Layer Information". In: **Sensors** 22.20. ISSN: 1424-8220. DOI: 10.3390/s22207864.

Krasensky, Julia et al. (Mar. 2017). "Ozone and Reactive Oxygen Species". en. In: **Encyclopedia of Life Sciences**. 1st ed. Wiley, pp. 1–9. ISBN: 978-0-470-01617-6 978-0-470-01590-2. DOI: 10.1002/9780470015902.a0001299.pub3.

Li, Bin et al. (Mar. 2024). "Spatiotemporal Patterns of Surface Ozone Exposure Inequality in China". In: **Environmental Monitoring and Assessment** 196.3. ISSN: 0167-6369. DOI: 10.1007/s10661-024-12426-3.

Li, JC et al. (Dec. 2023). "Double Trouble: The Interaction of PM2.5 and O3 on Respiratory Hospital Admissions". In: **Environmental Pollution** 338. ISSN: 0269-7491. DOI: 10.1016/j.envpol.2023.122665.

Li, QY et al. (Dec. 2024). "Development of a City-Level Surface Ozone Forecasting System Using Deep Learning Techniques and Air Quality Model: Application in Eastern China". In: **Atmospheric Environment** 339. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2024.120865.

Li, Z et al. (2023). "Estimation of Near-Ground Ozone With High Spatio-Temporal Resolution in the Yangtze River Delta Region of China Based on a Temporally Ensemble Model". In: **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing** 16, pp. 7051–7061. ISSN: 1939-1404. DOI: 10.1109/JSTARS.2023.3298996.

Lin, J.-T. et al. (June 2012). "Model Uncertainties Affecting Satellite-Based Inverse Modeling of Nitrogen Oxides Emissions and Implications for Surface Ozone Simulation." eng. In: **Atmospheric Chemistry & Physics Discussions** 12.6, pp. 14269–14327. ISSN: 1680-7367. DOI: 10.5194/acpd-12-14269-2012.

Lin, MY et al. (Oct. 2012). "Springtime High Surface Ozone Events over the Western United States: Quantifying the Role of Stratospheric Intrusions". In: **Journal of Geophysical Research-Atmospheres** 117. ISSN: 2169-897X. DOI: 10.1029/2012JD018151.

Liu, Qian et al. (Aug. 2022). "Carbonyl Compounds in the Atmosphere: A Review of Abundance, Source and Their Contributions to O3 and SOA Formation." eng. In: **Atmospheric Research** 274, N.PAG–N.PAG. ISSN: 0169-8095. DOI: 10.1016/j.atmosres.2022.106184.

Liu, S et al. (July 2022). "Impact of Climate-Driven Land-Use Change on O3 and PM Pollution by Driving BVOC Emissions in China in 2050". In: **Atmosphere** 13.7. ISSN: 2073-4433. DOI: 10.3390/atmos13071086.

Liu, Ying et al. (Apr. 2018). "Improve Ground-Level PM2.5 Concentration Mapping Using a Random Forests-Based Geostatistical Approach". en. In: **Environmental Pollution** 235, pp. 272–282. ISSN: 02697491. DOI: 10.1016/j.envpol.2017.12.070.

Luan, Y et al. (Sept. 2024). "Assessment and Prediction of Health and Agricultural Impact from Combined PM2.5 and O3 Pollution in China". In: **Sustainability** 16.17. ISSN: 2071-1050. DOI: 10.3390/su16177391.

Ma, HF et al. (Sept. 2023). "Short-Term Exposure to PM2.5 and O3 Impairs Liver Function in HIV/AIDS Patients: Evidence from a Repeated Measurements Study". In: **Toxics** 11.9. ISSN: 2305-6304. DOI: 10.3390/toxics11090729.

Ma, SM et al. (Oct. 2021). "Sensitivity of PM2.5 and O3 Pollution Episodes to Meteorological Factors over the North China Plain". In: **Science of the Total Environment** 792. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2021.148474.

Maleki, Heidar et al. (Aug. 2019). "Air Pollution Prediction by Using an Artificial Neural Network Model". en. In: **Clean Technologies and Environmental Policy** 21.6, pp. 1341–1352. ISSN: 1618-954X, 1618-9558. DOI: 10.1007/s10098-019-01709-w. (Visited on 09/02/2025).

Manzini, J et al. (July 2024). "Detection of Morphological and Eco-Physiological Traits of Ornamental Woody Species to Assess Their Potential Net O3 Uptake". In: **Environmental Research** 252. ISSN: 0013-9351. DOI: 10.1016/j.envres.2024.118844.

Meng, Kai et al. (Oct. 2022). "Influence of Stratosphere-to-Troposphere Transport on Summertime Surface O3 Changes in North China Plain in 2019." eng. In: **Atmospheric Research** 276, N.PAG–N.PAG. ISSN: 0169-8095. DOI: 10.1016/j.atmosres.2022.106271.

Meo, SA et al. (Dec. 2021). "Effect of Green Space Environment on Air Pollutants PM2.5, PM10, CO, O3, and Incidence and Mortality of SARS-CoV-2 in Highly Green and Less-Green Countries". In: **International Journal of Environmental Research and Public Health** 18.24. ISSN: 1660-4601. DOI: 10.3390/ijerph182413151.

Minghu Ding et al. (Aug. 2020). "Year-Round Record of near-Surface Ozone and "O3 Enhancement Events" (OEEs) at Dome A,East Antarctica." eng. In: **Earth System Science Data Discussions**, pp. 1–31. ISSN: 1866-3591. DOI: 10.5194/essd-2020-130.

Nadzir, MSM et al. (Jan. 2018). "Spatial-Temporal Variations in Surface Ozone over Ushuaia and the Antarctic Region: Observations from in Situ Measurements, Satellite Data, and Global Models". In: **Environmental Science and Pollution Research** 25.3, pp. 2194–2210. ISSN: 0944-1344. DOI: 10.1007/s11356-017-0521-1.

Napi, NNLM et al. (Aug. 2021). "Development Of Models For Forecasting Of Seasonal Ground Level Ozone (O3)". In: **Journal Of Engineering Science And Technology** 16.4, pp. 3136–3154. ISSN: 1823-4690.

Nawaz, MO et al. (Jan. 2023). "A Source Apportionment and Emission Scenario Assessment of PM2.5- and O3-Related Health Impacts in G20 Countries". In: **GeoHealth** 7.1. ISSN: 2471-1403. DOI: 10.1029/2022GH000713.

Nelson, D et al. (Oct. 2023). "A Comprehensive Approach Combining Positive Matrix Factorization Modeling, Meteorology, and Machine Learning for Source Apportionment of Surface Ozone Precursors: Underlying Factors Contributing to Ozone Formation in Houston, Texas". In: **Environmental Pollution** 334. ISSN: 0269-7491. DOI: 10.1016/j.envpol.2023.122223.

Nixon, Richard (July 1970). **Reorganization Plan No. 3 of 1970**. en. (Visited on 07/03/2025).

Pan, Qilong, Fouzi Harrou, and Ying Sun (May 2023). "A Comparison of Machine Learning Methods for Ozone Pollution Prediction". en. In: **Journal of Big Data** 10.1. ISSN: 2196-1115. DOI: 10.1186/s40537-023-00748-x.

Pan, Y et al. (Nov. 2024). "Vertical Structure and Transport Characteristic of Aerosol and O3 during the Emergency Control Period in Wuhan, China, Using Vehicle-Lidar Observations". In: **Atmospheric Environment** 337. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2024.120762.

Place, BK et al. (Aug. 2023). "Sensitivity of Northeastern US Surface Ozone Predictions to the Representation of Atmospheric Chemistry in the Community Regional Atmospheric Chemistry Multiphase Mechanism (CRACMMv1.0)". In: **Atmospheric Chemistry and Physics** 23.16, pp. 9173–9190. ISSN: 1680-7316. DOI: 10.5194/acp-23-9173-2023.

Reid, Colleen E. et al. (Dec. 2012). "The Role of Ambient Ozone in Epidemiologic Studies of Heat-Related Mortality". en. In: **Environmental Health Perspectives** 120.12, pp. 1627–1630. ISSN: 0091-6765, 1552-9924. DOI: 10.1289/ehp.1205251.

Rovira, J, JL Domingo, and M Schuhmacher (Feb. 2020). "Air Quality, Health Impacts and Burden of Disease Due to Air Pollution (PM10, PM2.5, NO2 and O3): Application of AirQ plus Model to

the Camp de Tarragona County (Catalonia, Spain)". In: **Science of the Total Environment** 703. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2019.135538.

Sadiq, Mehliyar et al. (Sept. 2016). "Effects of Ozone-Vegetation Coupling on Surface Ozone Air Quality via Biogeochemical and Meteorological Feedbacks." eng. In: **Atmospheric Chemistry & Physics Discussions**, pp. 1–21. ISSN: 1680-7367. DOI: 10.5194/acp-2016-642.

Schultz, Martin G. et al. (Oct. 2017). "Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations". In: **Elementa: Science of the Anthropocene** 5. Ed. by Michael E. Chang and Alastair Lewis, p. 58. ISSN: 2325-1026. DOI: 10.1525/elementa.244.

Schulz, Hannes et al. (2015). "CURFIL: Random Forests for Image Labeling on GPU:" in: **Proceedings of the 10th International Conference on Computer Vision Theory and Applications**. International Conference on Computer Vision Theory and Applications. Berlin, Germany: SCITEPRESS - Science and and Technology Publications, pp. 156–164. ISBN: 978-989-758-089-5 978-989-758-090-1 978-989-758-091-8. DOI: 10.5220/0005316201560164.

Sofen, E. D. et al. (July 2015). "Gridded Global Surface Ozone Metrics for Atmospheric Chemistry Model Evaluation." eng. In: **Earth System Science Data Discussions** 8.2, pp. 603–647. ISSN: 1866-3591. DOI: 10.5194/essdd-8-603-2015.

Staehle, C et al. (Nov. 2022). "Quantifying Changes in Ambient NOx, O3 and PM10 Concentrations in Austria during the COVID-19 Related Lockdown in Spring 2020". In: **Air Quality Atmosphere and Health** 15.11, pp. 1993–2007. ISSN: 1873-9318. DOI: 10.1007/s11869-022-01232-w.

Stuhr, Robin, Patrick Bayer, and Axel Jacobi Von Wangelin (Dec. 2022). "The Diverse Modes of Oxygen Reactivity in Life & Chemistry". en. In: **ChemSusChem** 15.24, e202201323. ISSN: 1864-5631, 1864-564X. DOI: 10.1002/cssc.202201323. (Visited on 08/15/2025).

Tang, ZQ et al. (Jan. 2024). "The Impact of Short-Term Exposures to Ambient NO2, O3, and Their Combined Oxidative Potential on Daily Mortality". In: **Environmental Research** 241. ISSN: 0013-9351. DOI: 10.1016/j.envres.2023.117634.

Tian, XY et al. (Jan. 2024). "Assessing the Short-Term Effects of PM2.5 and O3 on Cardiovascular Mortality Using High-Resolution Exposure: A Time-Stratified Case Cross-over Study in Southwestern China". In: **Environmental Science and Pollution Research** 31.3, pp. 3775–3785. ISSN: 0944-1344. DOI: 10.1007/s11356-023-31276-z.

Travis, Katherine R. and Daniel J. Jacob (Aug. 2019). "Systematic Bias in Evaluating Chemical Transport Models with Maximum Daily 8 h Average (MDA8) Surface Ozone for Air Quality Applications: A Case Study with GEOS-Chem v9.02". en. In: **Geoscientific Model Development** 12.8, pp. 3641–3648. ISSN: 1991-9603. DOI: 10.5194/gmd-12-3641-2019.

Turner, Michelle C. et al. (May 2016). "Long-Term Ozone Exposure and Mortality in a Large Prospective Study". en. In: **American Journal of Respiratory and Critical Care Medicine** 193.10, pp. 1134–1142. ISSN: 1073-449X, 1535-4970. DOI: 10.1164/rccm.201508-1633OC.

Venkanna, R et al. (May 2015). "Environmental Monitoring of Surface Ozone and Other Trace Gases over Different Time Scales: Chemistry, Transport and Modeling". In: **International Journal of Environmental Science and Technology** 12.5, pp. 1749–1758. ISSN: 1735-1472. DOI: 10.1007/s13762-014-0537-8.

Wang, FY et al. (Feb. 2021). "Policy-Driven Changes in the Health Risk of PM2.5 and O3 Exposure in China during 2013-2018". In: **Science of the Total Environment** 757. ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2020.143775.

Wang, LL et al. (Dec. 2023). "Evolution of Surface Ozone Pollution Pattern in Eastern China and Its Relationship with Different Intensity Heatwaves". In: **Environmental Pollution** 338. ISSN: 0269-7491. DOI: 10.1016/j.envpol.2023.122725.

Wang, Wenhao et al. (Jan. 2022). "A Machine Learning Model to Estimate Ground-Level Ozone Concentrations in California Using TROPOMI Data and High-Resolution Meteorology". In: **Environment International** 158, p. 106917. ISSN: 0160-4120. DOI: 10.1016/j.envint.2021.106917.

Wang, Yuting et al. (Dec. 2023). "Does Downscaling Improve the Performance of Urban Ozone Modeling?" en. In: **Geophysical Research Letters** 50.23, e2023GL104761. ISSN: 0094-8276, 1944-8007. DOI: 10.1029/2023GL104761.

Watson, Gregory L. et al. (Nov. 2019). "Machine Learning Models Accurately Predict Ozone Exposure during Wildfire Events". en. In: **Environmental Pollution** 254, p. 112792. ISSN: 0269-7491. DOI: 10.1016/j.envpol.2019.06.088.

Wen, W et al. (Sept. 2021). "Comparative Analysis of PM2.5 and O3 Source in Beijing Using a Chemical Transport Model". In: **Remote Sensing** 13.17. ISSN: 2072-4292. DOI: 10.3390/rs13173457.

Wright, Marvin N. and Andreas Ziegler (2017). "**Ranger** : A Fast Implementation of Random Forests for High Dimensional Data in **C++** and **R**". en. In: **Journal of Statistical Software** 77.1. ISSN: 1548-7660. DOI: 10.18637/jss.v077.i01.

Wu, CL et al. (Mar. 2023). "A Hybrid Deep Learning Model for Regional O3 and NO2 Concentrations Prediction Based on Spatiotemporal Dependencies in Air Quality Monitoring Network*". In: **Environmental Pollution** 320. ISSN: 0269-7491. DOI: 10.1016/j.envpol.2023.121075.

Xiao Lu et al. (Mar. 2019). "Exploring 2016–2017 Surface Ozone Pollution over China: Source Contributions and Meteorological Influences." eng. In: **Atmospheric Chemistry & Physics Discussions**, pp. 1–45. ISSN: 1680-7367. DOI: 10.5194/acp-2019-98.

Xing, J et al. (Sept. 2016). "Representing the Effects of Stratosphere-Troposphere Exchange on 3-D O3 Distributions in Chemistry Transport Models Using a Potential Vorticity-Based Parameterization". In: **Atmospheric Chemistry and Physics** 16.17, pp. 10865–10877. ISSN: 1680-7316. DOI: 10.5194/acp-16-10865-2016.

Xiong, KL, XD Xie, L Huang, et al. (Feb. 2024). "Improved O3 Predictions in China by Combining Chemical Transport Model and Multi-Source Data with Machining Learning Techniques". In: **Atmospheric Environment** 318. ISSN: 1352-2310. DOI: 10.1016/j.atmosenv.2023.120269.

Xiong, KL, XD Xie, JJ Mao, et al. (Feb. 2023). "Improving the Accuracy of O3 Prediction from a Chemical Transport Model with a Random Forest Model in the River Delta China". In: **Environmental Pollution** 319. ISSN: 0269-7491. DOI: 10.1016/j.envpol.2022.120926.

Yu, Karen et al. (Jan. 2018). "Errors and Improvements in the Use of Archived Meteorological Data for Chemical Transport Modeling: An Analysis Using GEOS-Chem V11-01 Driven by GEOS-5 Meteorology". en. In: **Geoscientific Model Development** 11.1, pp. 305–319. ISSN: 1991-9603. DOI: 10.5194/gmd-11-305-2018.

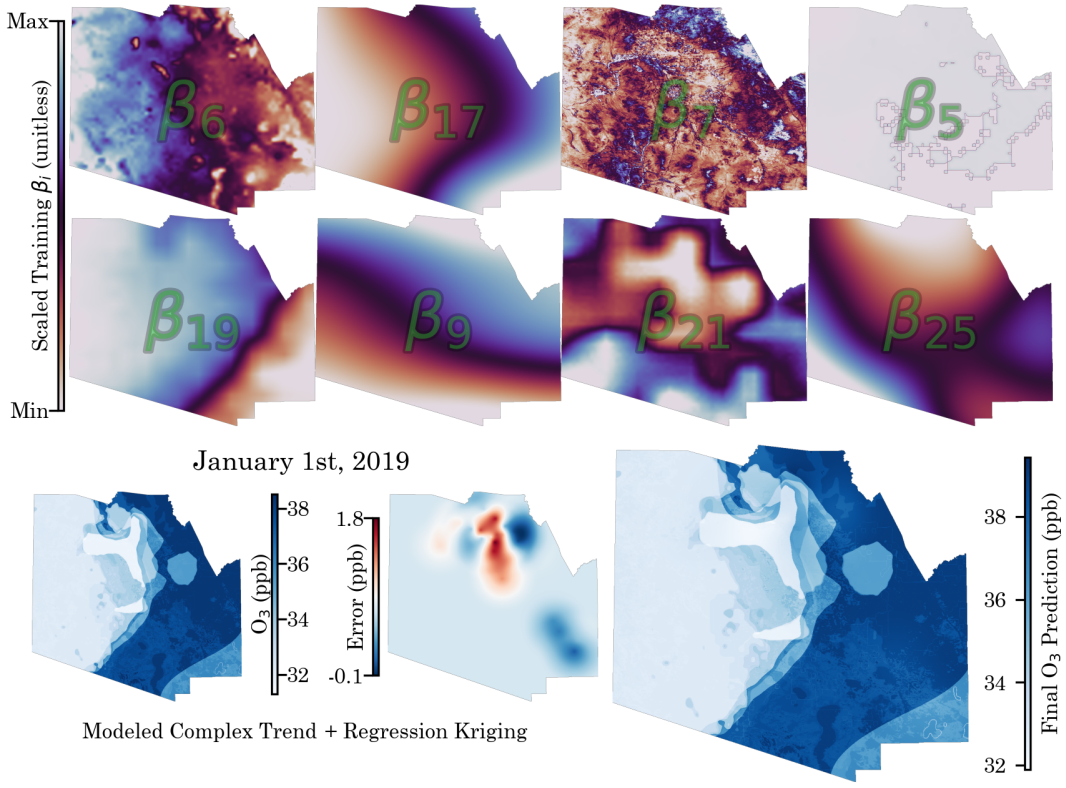
Zhang, Wenxiu et al. (Dec. 2022). "Recurrent Mapping of Hourly Surface Ozone Data (HrSOD) across China during 2005-2020 for Ecosystem and Human Health Risk Assessment." eng. In: **Earth System Science Data Discussions**, pp. 2–36. ISSN: 1866-3591. DOI: 10.5194/essd-2022-428.

Zhao, Naizhuo et al. (Dec. 2021). "Long-Term Ozone Exposure and Mortality from Neurological Diseases in Canada". en. In: **Environment International** 157, p. 106817. ISSN: 01604120. DOI: 10.1016/j.envint.2021.106817. (Visited on 09/12/2025).

Zhou, SS et al. (Oct. 2018). "Coupling between Surface Ozone and Leaf Area Index in a Chemical Transport Model: Strength of Feedback and Implications for Ozone Air Quality and Vegetation Health". In: **Atmospheric Chemistry and Physics** 18.19, pp. 14133–14148. ISSN: 1680-7316. DOI: 10.5194/acp-18-14133-2018.

Appendix A

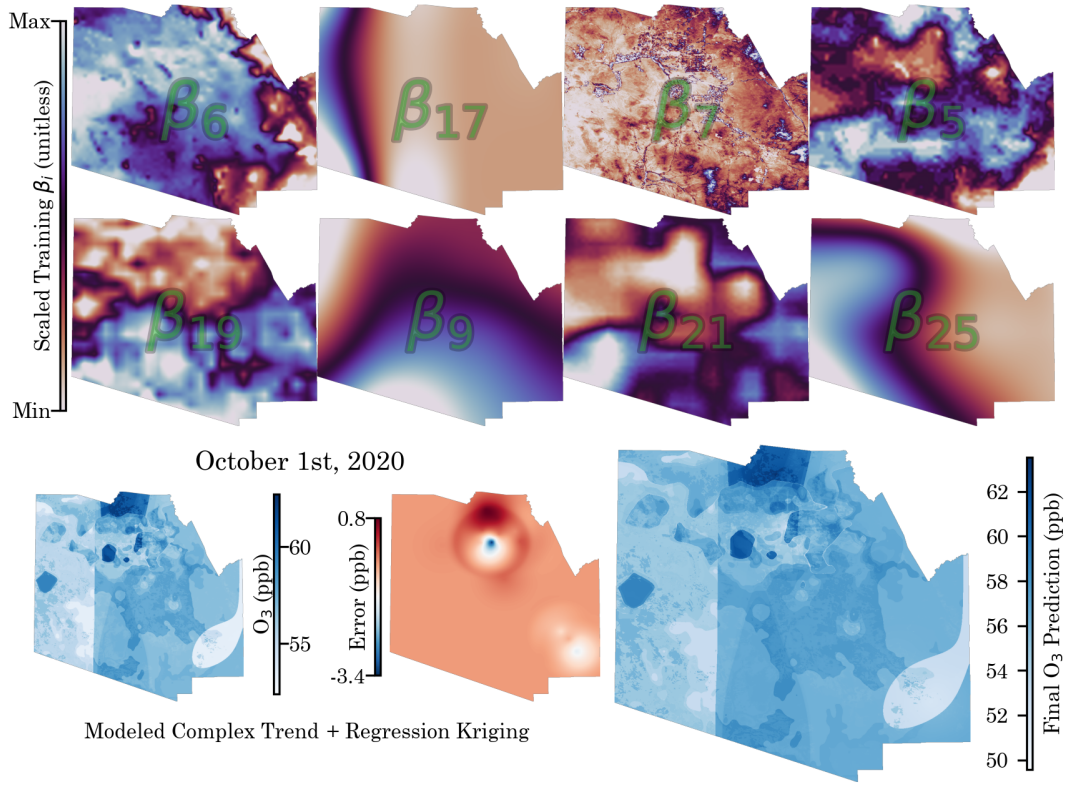
Surface Ozone Predictions: January, 2019



The results for the predicted average maximum 8-Hr surface ozone concentration features for PHOTUC on January 1st, 2019. Statistics for β_n codes are found at the beginning of Chapter 4. Each figure contains the names for these and depicts the learned complex trend, predicted regression krige, and final outcome for the represented day. The date is noted for each image in the bottom left above the predicted trends.

Appendix B

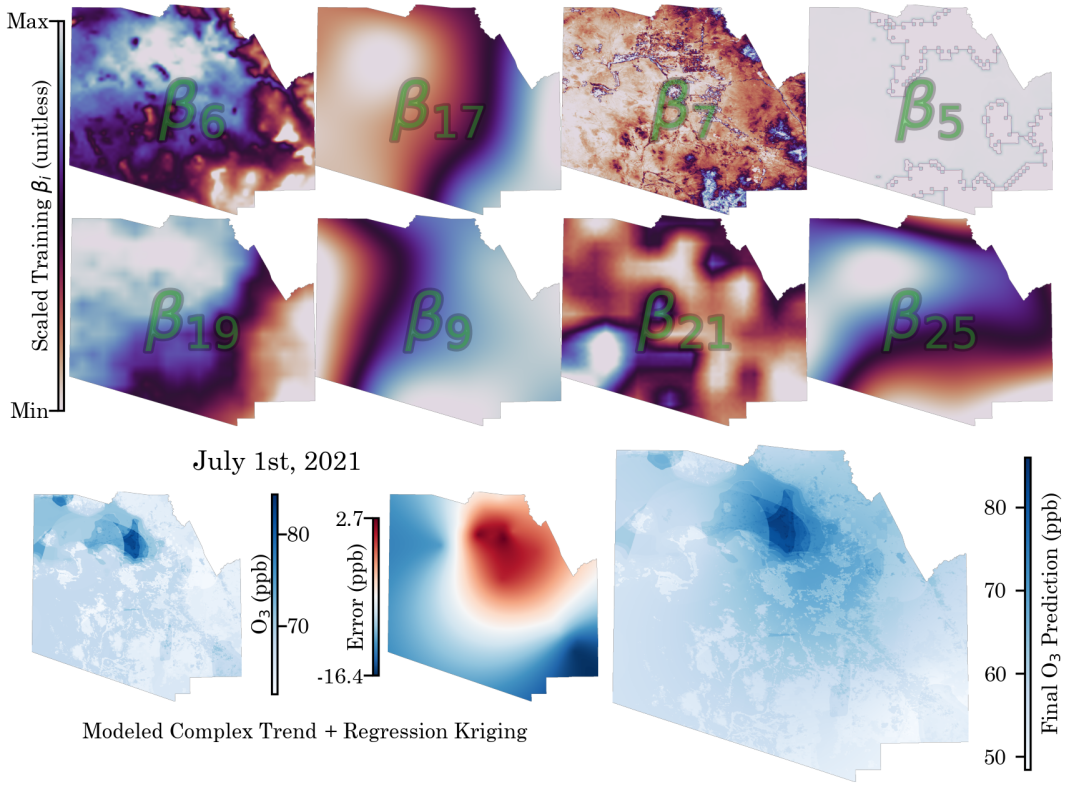
Surface Ozone Maps: October, 2020



The results for the predicted average maximum 8-Hr surface ozone concentration features for PHOTUC on October 1st, 2020. Statistics for β_n codes are found at the beginning of Chapter 4. Each figure contains the names for these and depicts the learned complex trend, predicted regression krige, and final outcome for the represented day. The date is noted for each image in the bottom left above the predicted trends.

Appendix C

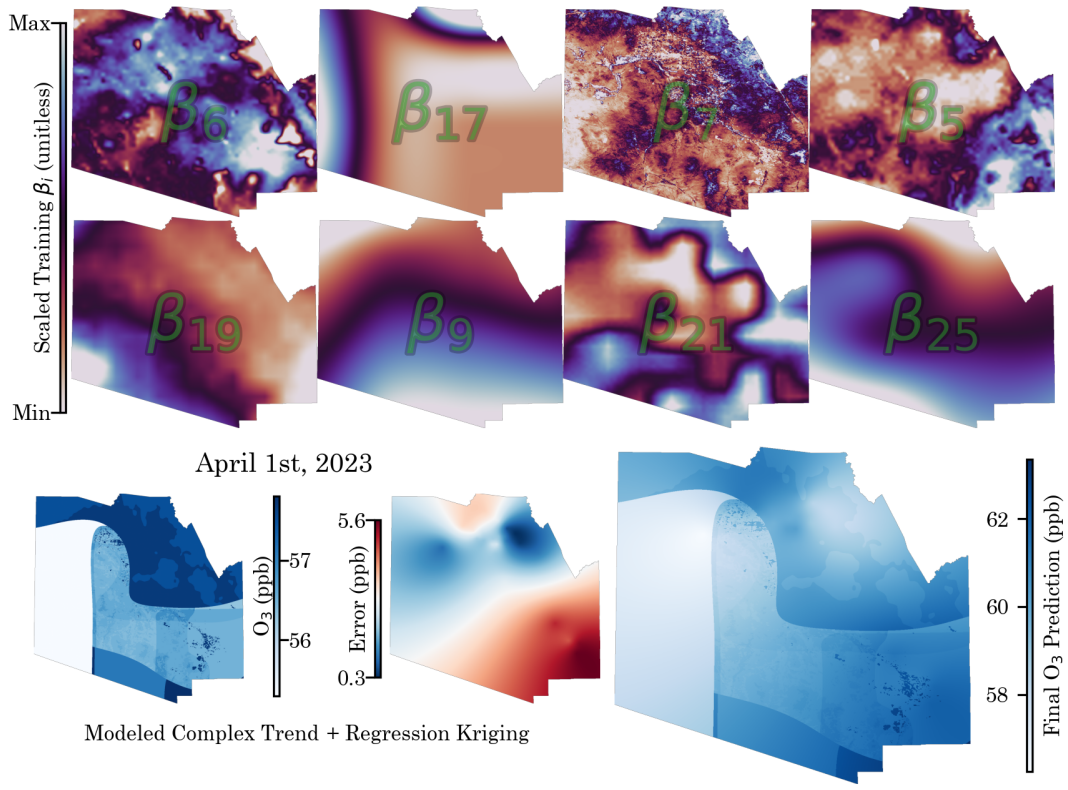
Surface Ozone Maps: July, 2021



Average maximum 8-Hr surface ozone concentration features for PHOTUC on October 1st, 2020. Statistics for β_n codes are found at the beginning of Chapter 4. Each figure contains the names for these and depicts the learned complex trend, predicted regression krige, and final outcome for the represented day. The date is noted for each image in the bottom left above the predicted trends.

Appendix D

Surface Ozone Prediction: Spring - April, 2023



Average maximum 8-Hr surface ozone concentration features for PHOTUC on October 1st, 2020 Statistics for β_n codes are found at the beginning of Chapter 4. Each figure contains the names for these and depicts the learned complex trend, predicted regression krige, and final outcome for the represented day. The date is noted for each image in the bottom left above the predicted trends.

Appendix E

List of Acronyms, Units, and Molecules

ADA Adaptive boost

AQI Air Quality Index

API Application Programming Interface

ArcPRO ArcGIS PRO

AOI Area of interest

AI Artificial Intelligence

ANN Artificial Neural Network

BAS Boundary and Annexation Survey

CDC Center for Disease Control and Prevention

CO_2 Carbon Dioxide

CO Carbon Monoxide

CAS Complex Adaptive System

CPU Computer Processing Unit

CSV Comma-separated values

CESM Community Earth System Model

CMAQ Community Multiscale Air Quality modeling system

CNN Convolutional Neural Network

CRS Coordinate reference system

CAMS Copernicus Atmosphere Monitoring Service

C3S Copernicus Climate Change Service

$\mathbf{R}^2$ correlation coefficient

CTM Chemical Transport Model

DAG Directed Acyclic Graph

DALY Disability-Adjusted Life Year

DAMSO Daily Averaged 8-hour Maximum Surface Ozone

DU Dobson Unit

EBSCO EBSCOhost Database

EVI Enhanced Vegetation Index

EPA Environmental Protection Agency

ECMRWF European Centre For Medium-Range Weather Forecast

EPSG European Petroleum Spatial Grid

ERA5 European Re-Analysis 5th Generation

KED External drift kriging

XGRB Extreme Gradient boost

CH_2O Formaldehyde

GIScience Geographic Information Science

GBD Global Burden of Disease

GMAO Global Modeling and Assimilation Office

GML Global Monitoring Laboratory

GEE Google Earth Engine

GB Gradient boost

GUI Graphical User Interface

GPU Graphics Processing Unit

g Grams

GOAT24 Greatest correlations over all time periods (Top 24)

GCP Ground Control Point

HCRNN Hierarchical Convolutional Recurrent Neural Network

HD Historical Data

K Kelvin

km Kilometer

KE Kinetic Energy

KNN K-Nearest Neighbors

LBD Literature Based Data

L Liters

ML Machine Learning

MAE Mean absolute error

MSE Mean square error

CH_4 Methane

m Meter

MD Modern Data

MASO Monthly Averaged Surface Ozone

MTDB Movement and Transportation data base

MLP Multi-Layered Perceptron

NOAA National Oceanic and Atmospheric Administration

NASA National Aeronautics and Space Administration

NCEP National Center for Environmental Prediction

NIR Near-infrared

NRT Near-real-time

NN Neural Network

NO_2 Nitrogen Dioxide

NO Nitrogen Monoxide

NO_x Nitrogen oxide

NDVI Normalized Difference Vegetation Index

NAD83 North American Datum 1983

OHE One-Hot Encoding

O_2 Oxygen

O_3 Ozone

OMI Ozone Monitoring Instrument

PM Particulate Matter

ppb Parts-per-billion

ppm Parts-per-million

PHO Phoenix

PHOTUC Phoenix and Tucson

PBL Planetary boundary layer

PCA Principal Component Analysis

RF Random Forest

RK Regression Krige

RKED Regression kriging with external drift

RNN Recursive Neural Network

RMSE Root mean square error

S5P Sentinel-5P

SM Statistical Model

SMaRK Statistical Model and Regression Krige

TIGER Topologically Integrated Geographic Encoding and Referencing

TOMS Total Ozone Monitoring Spectrometer

TCO_3 Total Column Ozone

TOAR Tropospheric Ozone Assessment Report

TROPOMI Tropospheric Monitoring Instrument

TUC Tucson

UV Ultra Violet

UID Unique Identifier

US United States

UK Universal kriging

UCB University of Colorado, Boulder

VoC Volatile organic compound

H_2O Water

WkAvg Weekly Moving Average

WGS84 World Geodesic System 1984

WHO World Health Organization

WoS Web of Science Database