Predictive Modeling for Ozone in the Colorado Springs Area:

Application of Machine Learning Techniques

Paul Trygstad

University of Wisconsin Green-Bay, MSDS Program

DS785: Capstone

Prof. Alex Korogodsky

December 10, 2023

Abstract

Colorado Springs, CO, has experienced high levels of the pollutant tropospheric ozone and faces noncompliance with the United States Environmental Agency's 2015 Nation Ambient Air Quality standards. Surface ozone can be deleterious to human and environmental health and noncompliance designation may strain local economy. Colorado Department of Public Health and Environment forecasting teams publish ozone determinations to protect local health and avoid regulatory noncompliance. Sufficiently accurate predictive modeling presents an opportunity to strengthen ozone forecasting. Data mining techniques were used to model ozone with local atmospheric data gathered from the United States Environmental Agency and The National Oceanic & Atmospheric Administration. Mixing layer height data was incorporated from the Integrated Global Radiosonde Archive. Two final datasets were created: one with original variable values, one with variables transformed for linear modeling assumptions. Penalty hyperparameters were tuned in preliminary modeling. Ozone levels at two local monitors AFA and MAN were modeled across a suite of 5-fold inner and outer cross-validated machine learning methods including linear models, generalized additive models, random forest, gradient boosted trees, and artificial neural nets. Gradient boosted tree models were found to the most accurate. The dataset with original values resulted in more accurate predictions. Best fits produced accuracies of 0.66 and 0.71 for AFA and MAN, respectively. Temperature, relative humidity, and calendar date were found to be the most influential variables across best fit models. Mixing layer height displayed moderate relative influence in the final models. Results were sufficiently accurate to present a feasible advantage to ozone forecasting efforts and local stakeholders.

Table of Contents

Tables

Figures

Chapter 1: Introduction

Colorado Springs and Ground-level Ozone

Located in south central Colorado along the Front Range of the Rocky Mountains, Colorado Springs is an expanding city and home to nearly 500,000 residents (United States Census Bureau, 2022). The Colorado Springs area hosts installations from multiple branches of the armed services and boasts such natural attractions as Pikes Peak and Garden of the Gods. Colorado Springs is becoming increasingly developed, having gained almost 109,000 people over the last 20 years (Laden, 2023).

Like much of the Front Range Urban Corridor, Colorado Springs has experienced air quality issues. One of the main air contaminants of concern in Colorado Springs has been ground-level ozone. Separate from the stratospheric ozone layer, which filters harmful solar radiation in the upper atmosphere, tropospheric or ground-level ozone occurs at or near the surface and is considered deleterious to human health and the environment (United States Environmental Protection Agency, 2023d). Ground-level ozone has been associated with a variety of negative health effects and is especially hazardous for individuals with pre-existing medical conditions (United States Environmental Protection Agency, 2023c). Prolonged exposure to high ozone levels may result in decreased respiratory function, asthma, cardiovascular issues, and overall higher mortality (Gold & Samet, 2013). Symptoms of ozone exposure include coughing, irritation, inflammation, and asthma attacks for individuals with that condition (United States Environmental Protection Agency, 2023b).

Ozone is a secondary pollutant, resulting from the chemical reactions of directly emitted primary pollutants catalyzed by sunlight and heat (United States Environmental Protection

Agency, 2023d). The main precursor chemicals that break down to form ozone are nitrates of oxide (NOx) and volatile organic compounds (VOCs) (United States Environmental Protection Agency, 2023d). These chemicals are emitted through anthropogenic processes such as industrial operations, energy production, vehicle discharge, and natural processes like forest fires.

Unhealthy ozone levels have been most common in urban environments, such as Colorado Springs, particularly on hot, sunny days. Through wind transport, ozone pollution can affect outlying communities as well as urban production zones (United States Environmental Protection Agency, 2023d). In the Colorado Springs region, ozone has been shown to increase with elevation (Flynn et al., 2021). This makes ozone a concern in the urban center and surrounding communities, particularly those at higher elevations.

The Colorado Springs area currently faces increased regulatory scrutiny and potentially tightened restrictions (Pikes Peak Area Council of Governments, 2021). The United States Environmental Protection Agency (EPA) uses a three-year rolling average of the fourth highest ozone measurements in an area to determine compliance with regulatory thresholds (Boyce, 2023). The current National Ambient Air Quality (NAAQ) standard for tropospheric ozone established by EPA is 70 parts per billion (ppb) (United States Environmental Protection Agency, 2023f). The three-year rolling average for Colorado Springs was 73 ppb in 2021 and 74 ppb in 2022, higher than the NAAQ standard for two consecutive years (Boyce, 2023). Increased restrictions and associated expenses may be imposed on emission producing businesses if the area is classified as being in a state of nonattainment, or ongoing noncompliance with regulatory standards.

Accurate predictions of high-ozone days enable warnings to be disseminated to affected populations and help business and government decision-makers act conscientiously regarding air quality. To protect public health and assist emissions-producing stakeholders to make responsible decisions, a team at the Colorado Department of Public Health and Environment (CDPHE) produces publicly available daily ground-level ozone forecasts. According to supervisor Scott Landes, the team reviews meteorological measurements and models with knowledge of local ozone trends to make each ozone determination (S. Landes, personal communication, September 5, 2023). CDPHE publishes the forecasts online and distributes information through Colorado Air Quality Alerts, an opt-in daily emailing list (Colorado Department of Public Health and Environment, 2023a).

The Challenge of an Accurate Forecast

Due to the variety of sources, wind transport, and complexity of ozone-producing reactions, accurately predicting daily ground-level ozone levels has been a challenging task (Flynn et al., 2021; S. Landes, personal communication, September 15, 2023). Atmospheric concentrations are dependent on complex relationships between the different chemical and meteorological variables. Existing forecasting methods have examined a wide range of meteorological data and published models but have also relied significantly on human expertise to interpret this information and make predictions.

While the CDPHE team considers a variety of available meteorological models for forecast determination, an in-house data-mining initiative has not been established. Machine- learning techniques could supplement and improve current forecasting methods, as data mining application has produced accurate ozone models for a wide variety of study environments around

the world (Camalier et al., 2007; Gong et al., 2017; Mohan & Saranya, 2018; Cheng & Huang, 2021). As an assistive tool, predictive models for ozone could be valuable to forecasting teams by increasing the precision of forecasts, supplying quantification for decision-making, and automating predictions. The downstream effects of model-enhanced forecasting could save Colorado Springs residents and businesses thousands to millions of dollars annually in healthcare costs and regulation mandated control measure expenses. To provide utility, the model must be sufficiently accurate.

Purpose and Research Questions

The purpose of this study was to determine if machine learning predictive models for tropospheric ozone in the Colorado Springs area could be produced with sufficient accuracy. Quantitative models were constructed to yield precise predictions that could be easily compared with measured values, especially during high ozone periods. The models examined years' worth of data to facilitate short- and long-term forecasting for ozone, so results may inform public health notifications and strategies to achieve regulatory compliance.

This project used historical atmospheric data to train and test a suite of candidate models for ozone concentrations at each of the two local ozone monitors and select the most accurate fit for results. Supplementarily, two datasets were tested to determine if original or mathematically normalized data yielded the best predictions. The most influential predictor variables in the final model were identified. The highest achieved accuracies of final fits for each monitor were used to evaluate the feasibility of data mining as a forecasting tool.

The results of this study addressed the following questions:

1. What models produced the best fit for overall ozone in the Colorado Springs area?

2. Does original or mathematically transformed data produce more accurate results?

3. How accurate was the final selected models from all tests?

4. What were the most influential variables in the final models?

5. Did atmospheric mixing height data have any predictive value across results?

6. What does the accuracy of the final model suggest about the viability of predictive modeling in tropospheric ozone forecasting?

Project objectives were designed to deliver results to answer each research question. The first objective was to collect all necessary data from identified sources. The next objective was to prepare final datasets for modeling, first by combining collected data, then by subsetting and transforming variables. The following objectives were to conduct preliminary modeling to tune hyperparameters, and then run final modeling tests with cross-validation, selecting the best fit model. The final targets were to fit the selected model to the entire modeling dataset, analyze results, and create recommendations for future research.

Terms Relating to the Study

The models in this project were designed to predict average 8-hour daily maximum ozone concentrations pursuant to the 2015 NAAQs. As predictors of ozone concentration, this project examined precursor chemicals as predictors. Precursor chemicals are the compounds that break down in heat and sunlight to produce ground-level ozone, the two most prominent being NOx and VOCs (United States Environmental Protection Agency, 2023d). Additional important contaminant data were concentrations of particulate matter of 10 microns or less (PM10) and 2.5 microns or less (PM2.5), the latter of which has been used to identify wildfire smoke days in historical datasets (Flynn et al., 2021).

This study considered atmospheric mixing layer height (MLH) as an explanatory variable. MLH is the altitude at which lower levels of the atmosphere undergo turbulent mixing. MLH has been an important measure in determining ozone forecast for the CDPHE Team (S. Landes, personal communication, September 15, 2023). Investigating the influence of MLH as a predictor variable was a query of this project.

Modeling Techniques and Data

Data was assembled and cleaned in the Python programming language, making noted use of the 'pandas' software package. 8 distinct machine learning were run with cross validation in the language R, facilitated by the 'caret' data-mining package. Best fit was determined by the lowest root mean square error (RMSE) metric. Models were evaluated for daily averages at both ozone monitors located within the study region. Additionally, two datasets were prepared for modeling: one with original values and one with select variables transformed to meet linear modeling assumptions. Models from both sensors were evaluated using both datasets, and the overall best model was selected. Model accuracy was additionally assessed by examining the R^2 coefficient of determination.

Considered models included standard and modified linear regression algorithms. Generalized additive models were included to follow previous modeling efforts in the study region (Flynn et al., 2021). This project also considered random forest, gradient boosted trees, and neural net methods, as similar techniques have been deployed elsewhere with accurate results (Siwek & Osowski, 2016; Bhuiyan et al., 2020; Du et al., 2022).

The modeling datasets consisted of combined air quality, meteorological, and radiosonde measurements. The source for air quality data was the EPA's Air Quality System (AQS) (United

States Environmental Protection Agency, 2023e). The source for meteorological and radiosonde data was The National Oceanic & Atmospheric Administration's (NOAA) National Centers for Environmental Information (NCEI) Climate Data Online (CDO) and Integrated Global Radiosonde Archive (IGRA) databases, respectively. Mean daily values were selected from the broader datasets for modeling.

Significance of Modeling Efforts

Previous attempts to model ozone with data mining techniques in the Colorado Springs area have not produced sufficiently accurate results to provide useful predictions or benefit to forecasting teams. Flynn et al. (2021) generated predictive models for the two ozone monitors in the Colorado Springs area resulting in a best R^2 accuracy score of 0.45, indicating only 45% of variance in ozone concentration explained by the model. An R^2 of 0.70 or above has been considered a strong result in scientific application (Frost, 2023).

By considering a wide variety of modeling and tuning techniques, this project hoped to establish a new best fit model for ozone in the study area. If sufficiently accurate, predictive modeling could present an advantageous supplement to current ozone forecasting methods. An accurate model could be a valuable tool for ozone forecasting teams, regulatory specialists, business leaders, and government officials, as well as any other agencies seeking to promote citizen wellbeing.

Beyond the intrinsic value of protecting human health, sufficiently accurate ozone forecasting models could provide the economic value of saving residents healthcare costs. Accurate modeling could also help quantify the effect of uncontrollable events, like wildfire activity, to high ozone levels (Gong et al., 2017; Flynn et al., 2021). Additionally, this project

may serve as guidance for future modeling efforts in the Colorado Springs area or application in new study locations.

Project Limitations

This project faced several limitations without feasible resolution, many of which related to the source data. The historical air quality and meteorological measurements were aggregated from disparate data sources separated by distance and terrain. Data was only available for two permanent ozone sensors in the Colorado Springs area. The nearest station with current MLH data was in Denver, which may not accurately represent the study area's conditions. Additionally, MLH data did not exist at the daily summary level, and daily mean values had to be calculated. The inconsistent number of records for each calendar date may compromise the overall accuracy of the resultant averages.

After the datasets were combined, many parameters contained many missing values. Missing values are likely attributable to instrumentation absence, maintenance, or failure. Variables with significant quantities of null values were eliminated from consideration, except the average MLH variable. Remaining gaps in data required imputation techniques. In particular, the average MLH variable exhibited significant missingness prior to imputation. Despite missing values, the feature was included with the goal assessing the variable’s prognostic value.

As noted by previous efforts, accurately modeling ozone levels in the Colorado Springs area has proven difficult. Research has suggested that regional sources, like international transport, controlled local ozone concentrations more than photochemical production (Flynn et al., 2021; Ochse, 2022). The atmospheric variables used in this study may not include sufficient data to predict regional events.

This project's scope was designed not to exceed the computational resources of personally owned devices. Considerable computational resources could result in model accuracy beyond what was examined here.

Project Organization

The next section summarizes a literature search into the problems posed by ozone and ozone modeling with data mining techniques. A detailed explanation of the methodology follows. Results summarizes the final model selection and relative variable influence in the final model fits. Finally, the Discussion and Conclusions section examines the potential value of the findings and makes suggestions for future research.

Chapter 2: Literature Review

Introduction

Across the globe, there have been efforts to protect health and assist compliance efforts through accurate ozone projections (Bhuiyan et al., 2020; Cheng & Huang, 2021; Colorado Department of Public Health and Environment, 2023a). The literature review for this study searched for the health effects of ozone, costs of ozone nonattainment, and previous ozone modeling projects. This review focused primarily on ozone predictive modeling cases, informing the methodology in this study.

As a hazardous air pollutant, ozone has been documented to do harm to human health and the environment. Noncompliance designations due to high ozone have resulted in significant control measure expenses and economic strain (Gilman, 2017). Industries contributing to ozone levels in designated noncompliant areas have been found to decline significantly, a finding which has stark implications for regional oil and gas production (Cheadle et al., 2017; Jaffe, 2022).

There have been many published attempts to forecast ground-level ozone with data mining techniques, especially in the past decade (Yafouz et al., 2021). The use of common data types, datasets, and machine learning techniques throughout the literature has established a foundation for predictive ozone modeling. Investigation of unique variables and modeling methodology has laid groundwork for the pursuit of accurate results across a variety of study environments. This project took much guidance from published literature, particularly in data selection and modeling suite design.

The Threat Posed by Ozone

High ozone has been shown to cause serious harm to human health and the environment, especially plant life (United States Environmental Protection Agency, 2023d). Additionally, expensive control measures and restrictions on emissions-associated business have produced economic stress in areas designated noncompliant (Blankenheim, 2013; Gilman, 2017; Ochse, 2022)

Danger to Human and Plant Life

Like all air pollution, ground-level ozone is considered dangerous to human health (Gold & Samet, 2013). Ozone has been shown to inflame conditions like asthma; long-term ozone exposure may lead to the development of respiratory illness (United States Environmental Protection Agency, 2023b). Ozone exposure has also been positively correlated with increased risk of heart attack, in addition to overall mortality (Gold & Samet, 2013; Cox, 2017; United States Environmental Protection Agency, 2023b).

In addition to human health, ozone is harmful to plant life (United States Environmental Protection Agency, 2023d). High ozone levels have a deleterious effect on agriculture including crop yield and grade (Avnery et al., 2011). By diminishing agricultural yields, tropospheric ozone poses a threat to global food security (Mills & Harmens, 2011).

Economic Consequences of Noncompliance Designation

Control Measures. Noncompliance designations have induced strain on local economies (Gilman, 2017; Ochse, 2022). In the Minnesota's Minneapolis–Saint Paul area, compliance nonattainment required control measures were found to incur total costs between $140 million - $237 million in 2013, equivalent to $182 million - $308 million in 2023 (Blankenheim, 2013; Federal Reserve Bank of Minneapolis, 2023).

70 miles north of Colorado Springs along the Front range, the Denver metropolitan area was downgraded to severe nonattainment in November 2023, triggering reformulated gasoline (RFG) requirements in the area (Reformulated Gasoline Covered Areas, 2023). EPA has estimated RFG costs in Denver to equal $13.3 million annually, but other estimates suggest costs of $800 million to $1 billion for Colorado residents and businesses (Cummings & Hill, 2022; Determinations of Attainment by the Attainment Date, Extensions of the Attainment Date, and Reclassification of Areas Classified as Serious for the 2008 Ozone National Ambient Air Quality Standards, 2022). Implementation of these RFG production requirements in the Denver area have been calculated to cost individual oil refineries between $250 million and $710 million, and result in fuel prices 11 cents to 1 dollar more per gallon (Gilman, 2017; Ochse, 2022). Exacerbating the issue, recent research show that RFG implementation may not provide significant advantage to air quality (Bishop et al., 2018).

Oil and Gas Industries. Along the Front Range, petroleum production has been the major emissions-producing industry contributing to ozone levels (Cheadle et al., 2017). Analysis suggests that emission producing businesses, such as oil and gas production, could experience a reduction up to 24% under nonattainment designation (Stanley, 2017). Colorado oil and gas production accounted for nearly 720,000 jobs statewide in 2022 (Bureau of Labor Statistics, 2023a). Of the wells approved in 2022, two thirds were along the Front Range (Jaffe, 2022). As such, petroleum industry shrinkage could have harsh economic consequences.

Colorado Springs may be especially vulnerable to industry decline as the area has already experienced an increase in unemployment from 2.9% to 3.3% between 2022 and 2023 (Bureau

of Labor Statistics, 2023b). Both unemployment values were higher than state averages for their respective years.

With the problems posed by ozone defined, there is a clear need for accurate projections to help protect health and avoid regulatory penalties. The rest of the literature search focused on previous modeling efforts to inform this project's design.

Modeling Trends

Data mining has proven to be a powerful tool for modeling ozone. Early data mining applications used more basic linear techniques, while complex nonlinear and hybrid models have been used increasingly in recent studies (Camalier, et al., 2007; Yafouz et al., 2021). Across the literature, a wide variety of model types have produced the best results, indicating that testing a suite of models is the best practice, as was constructed in this study (Siwek & Osowski, 2016). Common air quality and meteorological variables, including those featured in this project, were ubiquitously included across modeling efforts (Yafouz et al., 2021). Gauging the predictive quality of a unique variable or variables was commonly a secondary objective of research (Gong et al., 2017; Du et al., 2022, Wang et al., 2021). Following this convention, assessment of an MLH feature was an objective of this study.

Common and Divergent Theory

The same basic statistical theory underlies any data mining approach for modeling ground-level ozone (Yafouz et al., 2021). Historical data were used to train and test models. Models were evaluated on error metrics, and the model with the least error was selected as the best fit. In most studies, models were cross validated to avoid overfitting and generate honest predictions.

Differences in theory presented as details in model construction and data consideration. While air quality and meteorological data have been consistently included, supplemental features such as traffic information, mixing height, and wind transport features have been additionally considered as factors potentially influential to ozone levels (Gong et al., 2017; Du et al., 2022; Wang et al., 2021). Additionally, the use of different models has implied different assumptions about the relationships between variables. The relationship between atmospheric variables and ozone has widely thought to be nonlinear, a conclusion supported by the performance of advanced techniques that, unlike linear models, do not require underlying conditions of normality (Di. Et al, 2016; Du et al., 2022).

Major variations across literature consisted of study area, model design, data included for modeling, research scope, and machine learning methods employed. The next sections give an overview of these decisions across literature as pertaining to the build of this project in the pursuit of achieving a sufficiently accurate model to benefit ozone forecasting.

Ozone Modeling Methodology and Design

Previous studies have frequently incorporated common air quality and meteorological attributes in their models; the advantage of supplemental data sources or unique variables has often been an investigative inquiry (Wang et al., 2021; Du et al., 2022). Various feature selection techniques have been utilized to filter only the most prognostic data, although this has not proven unilaterally superior to using unfiltered variable sets (Mohan & Saranya, 2018). While different studies have examined ozone concentrations in regional and local scope with varying temporal resolution, the best predictions are generated by algorithms optimized for each monitor or city, emphasizing the importance of local optimization for accuracy in regional or case studies

(Camalier et al., 2007; Cheng & Huang, 2021). A broad range of machine learning models have been tested and compared with varying performance based on study area and prediction resolution. Overall, advanced techniques that can model nonlinear interactions have produced the most accurate results (Di et al., 2016; Gong et al., 2017; Yafouz et al., 2021). A comprehensive review of these factors informed the design of the models in this project.

Explanatory Variables

As photochemical ozone production involves the interaction of precursor chemicals with heat and sunlight, air pollution and meteorological information have been ubiquitously included in modeling (Yafouz et al., 2021). A supplemental objective of many ozone modeling efforts, including this study, was to identify which features were the most influential for generating predictions (Ishak et al., 2017; Camalier et al., 2007; Wang et al., 2021). In select instances, model results have been improved by the addition of derived or supplemental features (Di et al., 2016; Gong et al., 2017; Flynn et al., 2021; Wang et al., 2021; Du et al., 2022).

Air Pollution Data

Historical Ozone Data. An essential part of any ozone predictive modeling project was historical ozone data. Ground-level ozone data was gathered through permanent or temporary ozone monitoring equipment stations deployed as single point or air quality measurement networks (Camalier et al., 2007; Mohan and Saranya, 2018; Flynn et al., 2021; Yafouz et al., 2021). Ozone data was frequently analyzed at the hourly or daily resolution scale (Camalier et al., 2007; Lee et al., 2018; Flynn et al., 2021; Du et al., 2022). As in this study, ozone data was frequently sourced from the EPA AQS for study areas within the continental United States

(Camalier et al., 2007; Di et al., 2016; Flynn et al., 2021; Wang et al., 2021; United States Environmental Protection Agency, 2023e).

Nitrates of Oxides (NOx). Studies have shown that precursor chemicals, especially nitrates of oxides (NOx), are important variables in forecasting ground level ozone (Geiß et al., 2017; Bhuiyan, 2020; Wang et al., 2021). Complicated chemical interactions can render NOx difficult to model; inclusion as a predictor may introduce uncertainty. Mohan & Saranya (2018) improved their results by removing NOx data from their final modeling dataset. Neither traffic information nor NOx measurements were available for this project, so no complications were encountered.

Transport Data

Ground-level ozone transport by wind is a potential local contributor. When considered as a model feature, wind data has delivered mixed results (Camalier et al., 2007; Gong et al., 2017). Gong et al. (2017) found prevailing wind direction necessary to accurately forecast ozone distributions affected by regional wildfires. The relative influence of wind data, however, has not been consistent between study areas (Camalier et al., 2007). In Colorado Springs, transport from the nearby metropolis of Denver was found not to be a significant contributor to local ozone, although transport of precursors produced from wildfires was shown to be significantly influential (Flynn et al., 2021). Wind speed and directional data were included in both meteorological and air quality datasets collected for this study.

Meteorological Variables

Temperature and Relative Humidity. Camalier et al. (2007) examined large scale ozone distributions in the Eastern US and showed that temperature and relative humidity were among

the most influential variables in regional ozone distributions, additionally confirming that temperature and relative humidity were positively and negatively correlated, respectively, with surface ozone. Relative humidity and temperature have been moderately to highly influential variables in advanced predictive machine learning techniques across a variety of climates including Seoul in Korea, the Juarez, Mexico / El Paso area in Texas, Amman in Jordan, and Colorado Springs (Lee et al., 2018; Aljanabi, 2020; Bhuiyan, 2020; Flynn et al. 2021). This study incorporated both variables.

Mixing Layer Height (MLH). MLH is a derived feature calculated from measurements collected with ceilometer or radiosonde equipment (Wang & Wang, 2014; Geiß et al., 2017). MLH is a measure of vertical turbulent mixing within the boundary layer of the atmosphere and has been shown to affect surface ozone levels (Geiß et al., 2017). Derivation methods from direct or remote sensor measurements introduce inherent uncertainty that can undermine accuracy, however, and there has been debate over the best calculation methodology (Wang & Wang, 2014; Wang et al., 2021). Additionally, relatively sparse sensor coverage complicates representative characterization (Camalier, 2007; Di et al., 2016; Geiß et al., 2017). This study faces the limitation of sparse MLH measurements. Nevertheless, investigating MLH as a predictor in the Colorado Springs area was one of this study's queries. No prior literature existed assessing the influence of MLH to local ozone.

Feature Selection

Feature selection algorithms have been frequently used to subset predictors before model fitting. Stepwise linear fit is a popular feature selection technique that selects local optima by iteratively adding variables from a model with no features (forward selection) or removing from

a model with all available features (backwards selection) (CRAN, n.d.). Variations for stepwise fitting have been utilized to obtain best results across machine learning techniques including artificial neural nets and generalized additive models (Siwek & Osowski, 2016; Gong et al., 2017). Stepwise techniques have been shown to outperform more complicated feature selection methods such as the genetic algorithms on a case basis (Siwek & Osowski, 2016; Aljanabi et al., 2020). Mohan and Saranya (2018), however, found results produced more accurate results with unfiltered data than with those created by stepwise selection and more advanced techniques. This project used intrinsic feature selection methods in the R software package 'caret' in the following algorithms: linear regression with forward selection, linear regression with least absolute shrinkage and selection operator (LASSO) penalization, general additive modeling, random forest, and gradient boosted trees.

Resolution and Scale

Ground-level ozone modeling has been applied at both large-scale and case study scope. While predictor algorithms are fitted to local conditions in most cases regardless of overall study scope, study area scale and measurement resolution varied depending on the objective (Cheng & Huang, 2021; Du. et al., 2022). This project aims to produce an overall predictive model for ozone. Assessed on historical data, the models constructed in this study were a long-term prediction models.

Ozone Resolution

Ozone is often predicted in units of hourly average part per billion (ppb) concentration or daily maximum 8-hour average ppb; the latter unit of measure has supplied EPA's ozone standards since 1997 (Ishak et al., 2017; Camalier et al., 2007; Wang et al., 2021; Gong et al.,

2017; United State Environmental Protection Agency 2023e). Ozone predictions can be sorted into three discrete categories: short-term forecasting, peak seasonal, and long-term forecasts.

Short-term Predictions. Short-term predictions consist of modeling next-day or next-week average ozone concentrations. Short-term predictions can achieve notable accuracy, and hence are especially important to short interval forecasting (Siwek & Osowski, 2016; Aljanabi et al., 2020; Cheng & Huang, 2021). While short-term predictions could provide advantages to forecasting teams, this project constructed overall predictive models that could be used as the foundation for targeted short-term tools or long-term compliance planning.

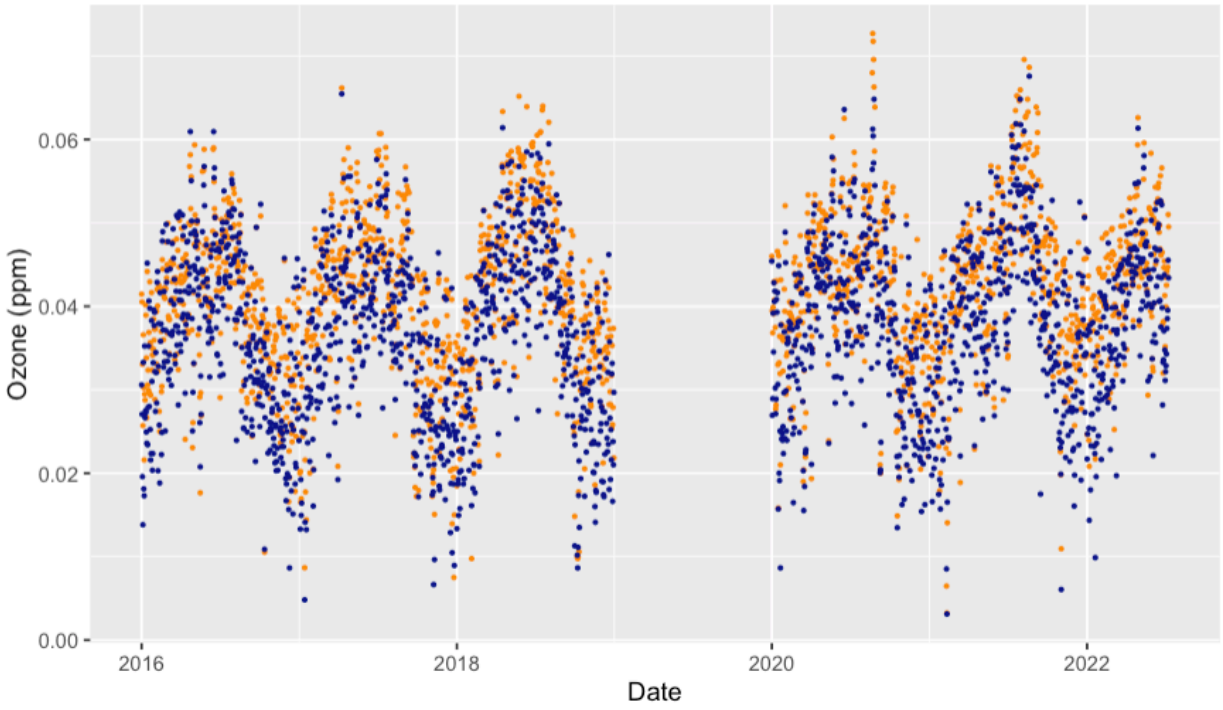
Peak Ozone Season. Many ozone modeling efforts have focused solely on the most prominent seasonal interval, typically the warm spring and summer months (Ishak et al., 2017; Camalier et al., 2007; Mohan & Saranya, 2018; Flynn et al., 2021; Wang et al., 2021). Figure 1 shows ozone seasonal variability for the monitors in Colorado Springs, which notable increases in the summer months. Modeling peak ozone season removed seasonal variability and allowed more accurate predictions when ozone is most harmful (Mohan & Saranya, 2018). Pursuant to the objective of investigating overall ozone levels, entire calendar years were considered in this study instead of only peak ozone data.

Long-term Predictions. Finally, ozone concentrations have been predicted in long-term intervals, spanning years. Predictive modeling at this scale has aimed to surmount short-term limitations and yield a strong model for short- and long-term projections (Di. et al., 2016; Wang et al., 2021; Du et al., 2022). While this study evaluated models on historical data, the final model in this project could most accurately be classified as a long-term prediction model.

Spatial Resolution

Figure 1

Ozone Concentrations at AFA and MAN Monitors by Year



Note. Orange data points represent measurements at Manitou Springs (MAN) ozone monitor. Blue points represent readings from the Air Force Academy (AFA) sensor. The annual curve for both sensors represents seasonal patterns. 2019 data was omitted from this modeling dataset due to elevated missingness.

Ozone has been modeled at regional and local scale (Di et al., 2016; Mohan & Saranya, 2018; Wang et al., 2021). The scale of the ozone distribution characterized affected the complexity of the study; both case study and large-scale spatial distributions frequently have relied on individualized models for each monitor or city (Camalier et al., 2007; Siwek & Osowski, 2016; Gong et al., 2017; Aljanabi et al., 2020; Bhuiyan et al., 2020; Flynn et al., 2021).

Local characterization and feature selection has been paramount to accurate predictions on any scale and this project was no exception.

Machine Learning Approaches

To model ground-level ozone concentrations, a breadth of machine learning methods has been employed. The interactions of chemistry, meteorology, and atmospheric mechanics are often reflected by complex, nonlinear relationships between variables (Di et al., 2016; Du et al., 2022). Advanced data mining techniques have attempted to characterize these relationships, frequently with success (Mohan & Saranya, 2018; Aljanabi et al., 2020; Cheng & Huang, 2021). The following sections provide an overview of modeling efforts by type as they relate to this project.

Linear Modeling

As noted, nonlinear relationships between meteorological and air pollution variables has been well documented. However, advanced linear regression techniques, such as LASSO penalization or generalized linear modeling, can accommodate nonlinear effects while preserving interpretability. Camalier et al. 2007 used generalized linear modeling and achieved coefficient of determination (R^2) scores as high as 0.80. Linear modeling was included in this project following previous work.

Generalized Additive Modeling

Generalized additive models (GAMs) have been selected due to their flexibility and ability to incorporate linear, nonlinear, and categorical elements (Flynn et al., 2021). Gong et al. (2017) used GAMs to investigate wildfire smoke contribution to ozone concentration and produced a maximum R^2 result of 0.81 for one modeled location. Flynn et al. (2021) applied GAM

techniques to predict ozone at the MAN and AFA Colorado Springs monitoring sites but reported the low R^2 scores of 0.45 and 0.42, respectively. GAM were included in this study following the modeling efforts by Flynn et al. (2021).

Random Forest

Random forest algorithm is a powerful algorithm that can manage nonlinear associations without requiring feature independence or distribution assumptions (Wang et al., 2021). In some instances, random forest has outperformed competing algorithms for classification, quantitative modeling, and base-learner application in ensemble methods relating to ozone predictions (Siwek & Osowski, 2016; Ishak et al., 2017; Mohan & Saranya, 2018; Bhuiyan et al., 2020). Random Forest was included in this project as a favorable candidate for best fit.

Support Vector Regression

Support vector machines (SVMs) are proven algorithms for classification; when applied to quantitative problem solving, support vector regression (SVR) is employed (Siwek & Osowski, 2016; Ishak et al., 2017; Aljanabi 2020). While SVM and SVR have been useful as elements in hybrid techniques, individual SVM models have been matched or outperformed by more flexible techniques for ozone modeling, and hence were not included in this study (Aljanabi et al., 2020; Bhuiyan et al., 2020; Cheng 2021).

Artificial Neural Nets

Artificial neural nets are powerful, flexible machine learning models exhibit the weakness of overfitting (Mohan & Saranya, 2018). Neural nets have produced impressive results; Aljanabi et al. (2020) used a multilayer perceptron (MLP) neural net to produce R^2 results as high as 0.98 for next day ozone in Amman, Jordan. However, neural nets have been outperformed by

competing methods for ozone forecasting, demonstrating they are not universally superior to other machine learning techniques (Siwek & Osowski, 2016; Mohan & Saranya, 2018). Beyond single-learner methods, neural nets have been frequently incorporated as a component of hybrid models (Cheng & Huang, 2021; Yafouz et al., 2021).

Ensemble Methods and Hybrid Models

Ensemble methods and hybrid models combine the results of two or more algorithms to produce final predictions (Mohan & Saranya, 2018; Yafouz et al., 2021; Du et al., 2022). Flexible, accurate, and robust against overfitting, ensemble and hybrid models are becoming popular methods for predicting ozone concentrations (Yafouz et al., 2021). Mohan & Saranya (2018) used ensemble methods to model summertime ozone and generated a bagged random forest model with an R^2 of 0.86. Cheng & Huang (2021) employed a hybrid model scheme that incorporated a wavelet decomposition algorithm, neural networks, and SVR for five-day ozone forecasting, resulting in an impressive range of R^2 scores from 0.90 to 0.97. Du et al. (2022) used extreme gradient boosting to produce a long-term model with an R^2 of 0.776. Similar to the technique employed by Du et al. (2022), this project examined the gradient boosted trees ensemble method.

Design of this Study

There has been no demonstrated one-size-fits-all design approach to predicting surface ozone levels. Research has proven the need for fitting and comparing a variety of machine learning models. Additionally, recent studies underscore the use of ensemble or hybrid techniques to increase forecast accuracy (Mohan & Saranya 2018; Cheng & Huang, 2021;

Yafouz et al., 2021). The methods used in this study were principally informed by literature containing analogous variables and scope to this case study.

Chapter 3: Methodology

Introduction

This project aimed to create a suite of predictive models for tropospheric ozone concentration in the Colorado Springs area with mean summary data and select the most accurate model from the results. Modeling suite performance was used to assess the feasibility of data mining application as a forecasting tool. Data processing and modeling techniques were used to pursue the most accurate model within project constraints.

This project employed methods that can be broadly separated into two phases: data cleaning and modeling. Data cleaning was accomplished in Python and R; the former language was used to assemble and filter a combined dataset. Remaining predictors were examined and further subsetted in R. Mathematical transformations replaced select variables in a copy of the final dataset to meet underlying linear modeling assumptions. Preliminary modeling tests were run to establish tuning values for models containing adjustable hyperparameters. A suite of 8 predictive models were trained and tested on both original and transformed datasets. The model with the least error was selected as the final model and fit to the entire data set for final assessment, and subsequent evaluation of viability of predictive modeling in local ozone forecasting. Models examined include standard linear regression, generalized linear regression, linear regression with forward selection, penalized linear regression with LASSO optimization, generalized additive modeling, random forest, gradient boosted trees, and artificial neural nets.

Data Sources and Specifications

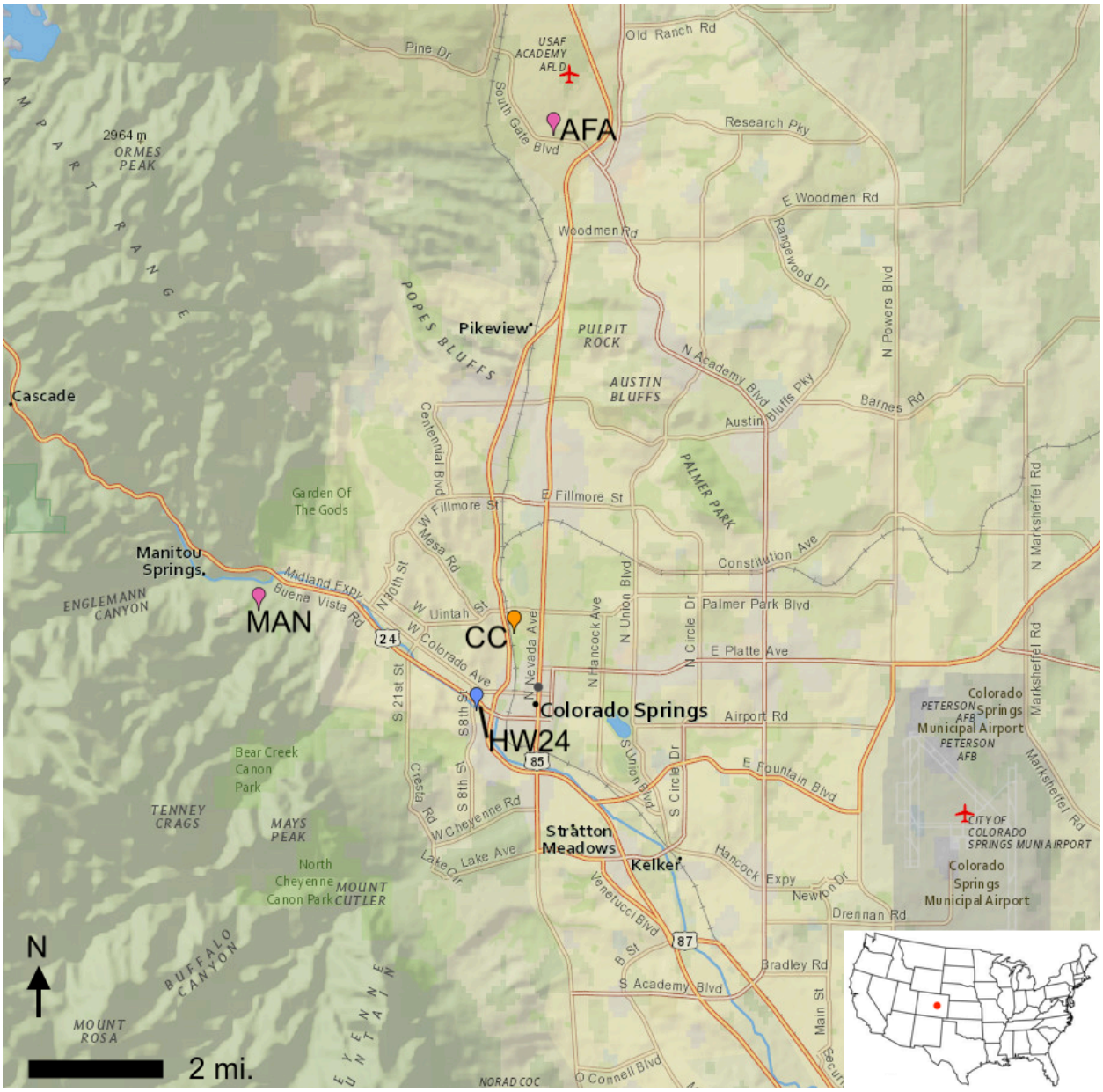
The wide variety of physical and chemical factors affecting tropospheric ozone concentration required aggregation of air quality and meteorological features from various

sources. Air Quality data was retrieved from United States Environmental Protection Agency (USEPA) Air Quality System (AQS). The AQS Interactive Map of Air Monitors tool facilitated air quality monitor selection, and AQS data mart tools helped to filter observations from January 1, 2011, to the most recent available data (United States Environmental Protection Agency, 2023e). Similarly, NOAA National Centers for Environmental Information (NCEI) Climate Data Online (CDO) was used to retrieve meteorological records from January 1, 2011-October 5, 2023, for the singular station in the CDO database located in Colorado Springs (National Oceanic and Atmospheric Administration, n.d.-a). MLH data was acquired from NOAA Integrated Global Radiosonde Archive (IGRA) database (National Oceanic and Atmospheric Administration, n.d.-b). As IGRA has no tool to specify an interval, Radiosonde data was retrieved in bulk from October 16, 1948-July 9, 2022.

The collected AQS pollutant data reflected four permanent air quality sensors installed and managed by CDPHE in the Colorado Springs area (Flynn et al., 2021). Figure 2 shows the relative location of the sensors. Located southwest of the study area, a monitor along State Highway 24 (HW24) captures carbon monoxide and sulfur dioxide, temperature, relative humidity, wind speed, and direction. Particulate matter, temperature, and pressure data were recorded in downtown Colorado Springs by the monitor at Colorado College campus. Air quality data also contained readings from two ozone monitors. One ozone monitor (identified as MAN) is located 5 miles northwest of downtown in the community of Manitou Springs, and the other (AFA) in the northern extent of the Colorado Springs area at the United States Air Force Academy.

Figure 2

Atmospheric Data Collection Points in Colorado Springs



Note. Ozone monitors are shown in purple, labeled Manitou Springs (MAN) and Air Force Academy (AFA). The orange point shows the Colorado College air quality sensor (CC). The blue point is the atmospheric monitoring array located along Highway 24 (HW24). The Colorado Springs Municipal Airport is located in the lower left. Image from AQS Interactive Map of Air Quality Monitors.

CDO daily meteorological summaries were recorded by equipment at the Colorado Springs Municipal Airport (COS), southeast of downtown Colorado Springs. Derived feature MLH radiosonde data recorded in the Colorado Springs region was not available past 1998, so readings taken at the former site of the Stapleton International Airport in Denver were used as the nearest available option for current data.

Combining and Cleaning Data in Python

Preparation for modeling started with combining all records across files and formats into a single data structure. This process was accomplished with the Python programming language, run in Spyder Interactive Development Environment from Anaconda open-source software package. For this phase of the project, Python was the ideal choice as the Pandas module "Data.Frame" object has built-in methods that enable efficient construction and modification of tabular data structures. Particularly useful for combining datasets, the Pandas package also contains functions for joining tables analogous to Standard Query Language (SQL) commands, with arguably more flexibility and convenience (Parameswaran, 2022).

Air Quality Records

Tall Table. Arrangement of the final data structure began with air quality data. AQS records consist of common fields, including measured parameter, measurement interval, average and maximum values, and monitor information. For each parameter and interval available, data were combined into a single data-frame containing a unique row for each measurement. At this stage, the data existed in "tall table" form, containing a relatively large number of rows compared to columns, with multiple rows for each calendar date.

Before further data transformation, the tall table was filtered to include only records after January 1st, 2015, and before July 9, 2022. These cutoff dates were established in subsequent phases of data exploration and based on missingness caused by availability of records. The data was retroactively filtered by date while in the tall table structure to increase computational efficiency and reduce the number of missing values generated.

Wide Table. For final modeling, data were transformed into "wide table" format, storing each measurement type for each measured parameter at each monitor site as a separate column with rows indexed by date. The result is that each calendar day is represented by only one row, with the columns containing all the available air quality measurements for that day.

Missingness Distribution. Missingness was examined for each field in the wide table. Missingness increased in the dataset for most variables in periods before 2015, so a cutoff point of January 1st, 2015, was established. As reported previously, this cutoff was retroactively used to filter the tall-table data structure before creation of the wide table. Wide table columns were removed if more than 10% of the records were missing.

MLH Records

Combining Records. The air quality wide table and IGRA data were combined into a single dataset. IGRA data was not available in pre-formatted daily summaries. Instead, radiosonde measurements included single, multiple, or missing measurements for each calendar date. After filtering to include only records from January 1, 2015, to most recent, a daily average MLH value was calculated for each day in the IGRA data set. The result was a new dataset with one row per date. Daily maximum and minimum MLH values were also calculated, but only the mean values used. Daily average MLH was added with the AQS data wide table as a new

column, rows matched by date. The most recent date in the IGRA dataset was July 9, 2022. To reduce the number of missing values created by joining datasets, the table structure was retroactively filtered to include dates only up to the IGRA cutoff.

Missing Values. Although a missingness threshold of 10% was used to remove columns earlier, the daily average MLH variable was retained despite consisting of approximately 28% missing values. Retaining daily average MLH was considered warranted by investigating the variable's predictive potential. The high number of missing values, however, complicated interpretations along this feature.

Meteorological Records

Combined Structure. CDO meteorological data were joined to the combined dataset. As retrieved, CDO data were already organized with one row per date. To match the interval of the combined AQS and IGRA wide table, CDO data was filtered to only include records in the interval from January 1, 2015, through July 9, 2022. Linking rows by date, meteorological variables from the filtered CDO dataset were joined as new columns to the combined data frame.

Missing Values. After integration of the meteorological data, missing values were concentrated in the combined structure during calendar years 2015 and 2019. Records from these calendar years were removed from the combined dataset.

Mean Values Subset

The combined dataset consisted of 2017 rows and 50 columns. Many of the parameter columns consisted of maximum, minimum, or mean daily readings. Since average maximum daily ozone is the parameter being calculated, mean values were selected to generate ground-level ozone predictions. Columns beginning or ending with the words "maximum," or

"minimum" were removed from the dataset. The filtered dataset was reduced to 30 columns, ordered by custom format for consistent output, then exported as a comma separated value file for preparation and modeling in R.

Data Preparation in R

R programming language offers convenient data modeling and visualization tools (Nunes, 2022). This project especially benefitted from the R software package 'caret,' which streamlines machine learning model training and testing (Hsu, 2020). Designed for statistical analysis, R enables in-depth analysis of variable significance, predictor relationships, and model performance. With modular code chunks, RStudio provided a convenient interface for final dataset preparation and model evaluation.

Imputation

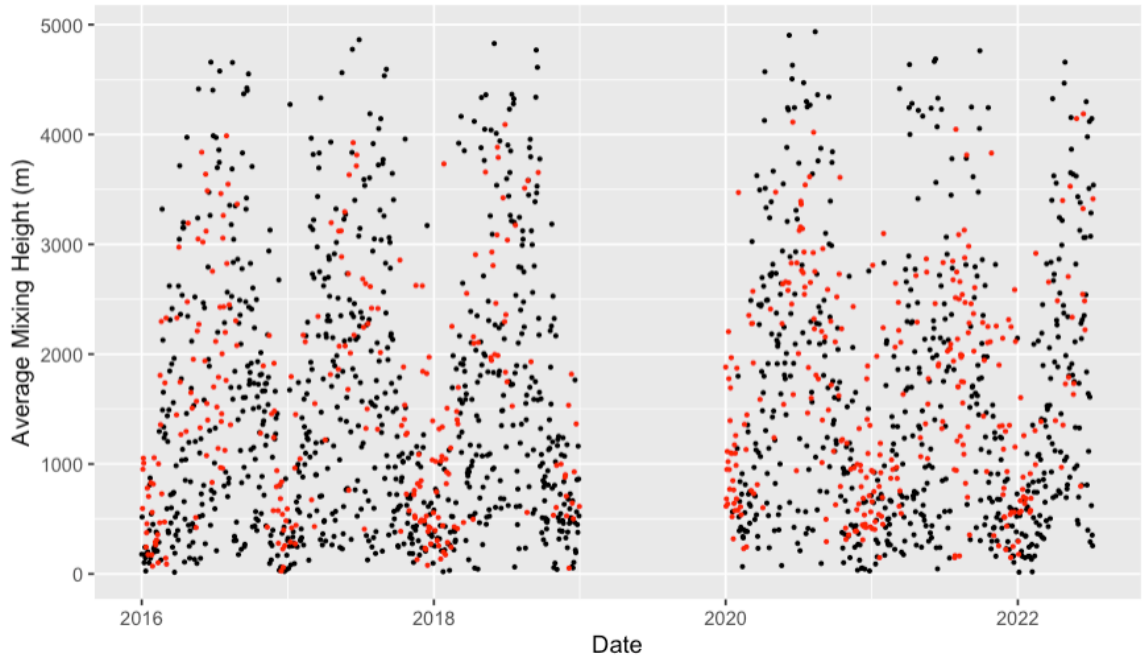
Prior to modeling, remaining missing values were imputed using a *k*-Nearest Neighbors (kNN) algorithm available in the 'DMwR2' package. This method uses *k* nearest data points to compute values for missing data. For average daily mixing height, which displayed a high degree of variance, higher values of *k* result in imputed values that approached the column mean. A value of 2 was selected as the value for *k* to preserve seasonal patterns. Figure 3 shows the imputed data points with seasonal data. After imputation, no missing values remained in the dataset.

Collinear Features

A final set of filters was applied to the dataset before modeling. The static parameters monitor latitude, longitude, and elevation were removed. A correlation plots identified collinear features in the dataset. Wind, temperature, and separate measurement types for the same

Figure 3

Imputed Values of Daily Average Mixing Height using 2-Nearest Neighbors



Note. Average MLH values calculated from the IGRA dataset are shown in black. Red points are the imputed values. 2-nearest neighbors was selected to preserve the seasonal variability.

pollutant (e.g., carbon monoxide daily average 8-hour maximum versus daily average 1-hour maximum) exhibited high degrees of collinearity.

Variance Inflation Factor (VIF) was used as criterion to select between collinear predictors. VIF was calculated by running preliminary linear models and passing the results to the 'vif' function in the 'car' package in R. Two preliminary models were created: one modeling daily 8-hour average ozone concentration at AFA as a function of all the other predictors except MAN ozone readings, and the other similarly modeling MAN ozone excluding AFA values. Results for both models showed the same relative VIF scores for collinear predictors; the collinear term with

higher VIF was eliminated. Correlation was re-examined after variable removal; additional variables were removed until no collinear factors remained. Figure 4 shows the correlation plot of the final dataset.

Final Dataset

The final data set consisted of 2017 rows and 15 variables. The final variables included date, AFA and MAN ozone measurements, HW24 measurements including daily maximum 8 hour average carbon monoxide, maximum 1-hour resultant wind speed, sulfur dioxide daily maximum 5-minute average, average hourly relative humidity, average hourly temperature, and mean hourly standard deviation of horizontal wind direction, average atmospheric mixing height from IGRA, and CDO daily measurements including precipitation, snowfall, snow depth, and fasted 2-minute direction and speed.

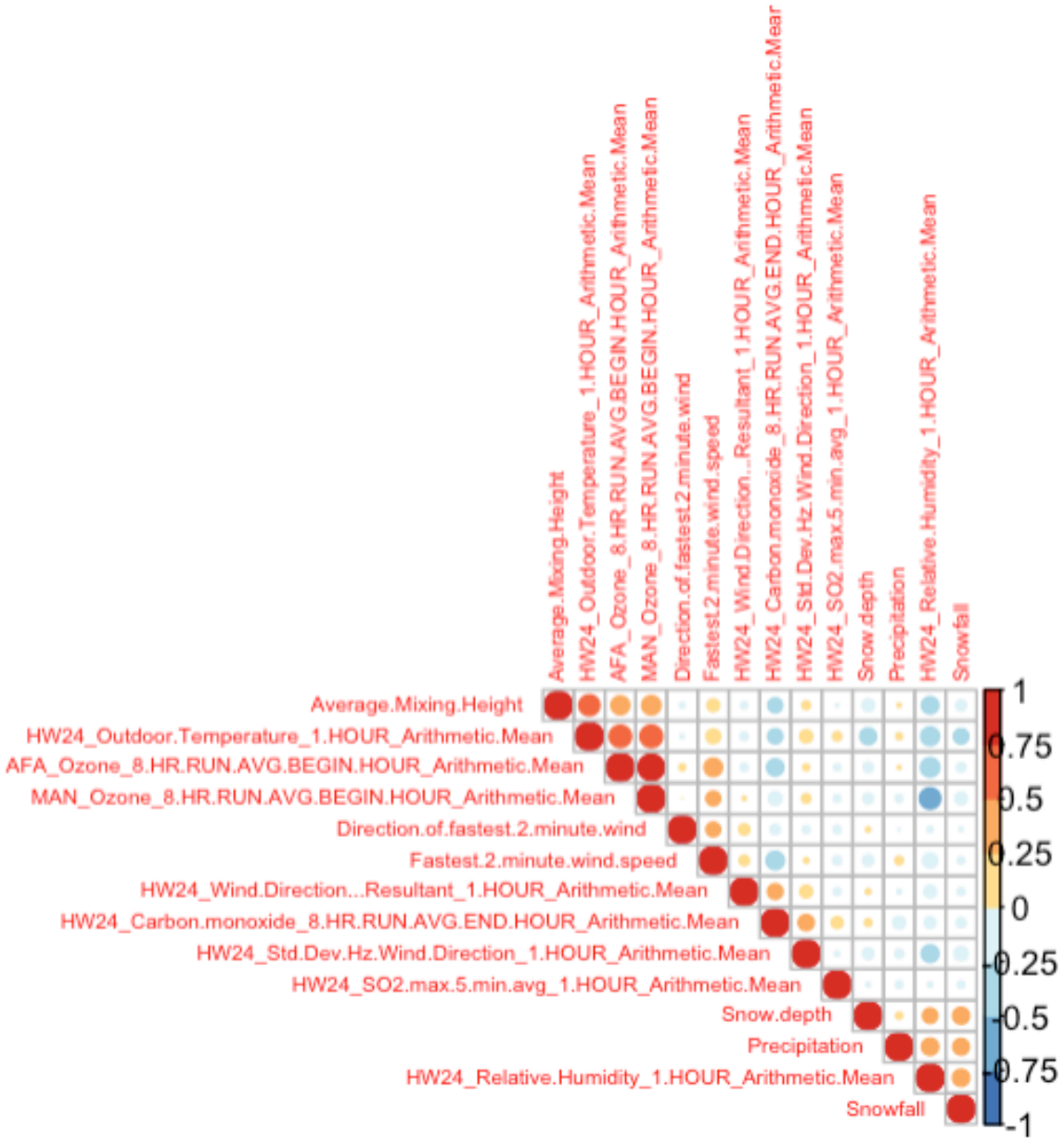
Transformed Variable Dataset

The relationship between ozone confrontation and air quality/meteorological variables is notedly nonlinear (Di et al., 2016; Du et al., 2022). Linear models were included in this project, however, as linear modeling techniques have yielded accurate predictions in select cases (Camalier et al, 2007).

Using histograms and additional diagnostic plots, distributions for each variable in the final dataset were examined for the underlying assumptions of linear modeling: linearity, normality, homoscedasticity, and independence of error terms (Kassambara, 2018). Carbon monoxide, sulfur dioxide, average mixing height, precipitation, snowfall, and snow depth exhibited distributions that violated one or several linear assumptions. Figure 5 shows an example of normal and non-normal variable distributions.

Figure 4

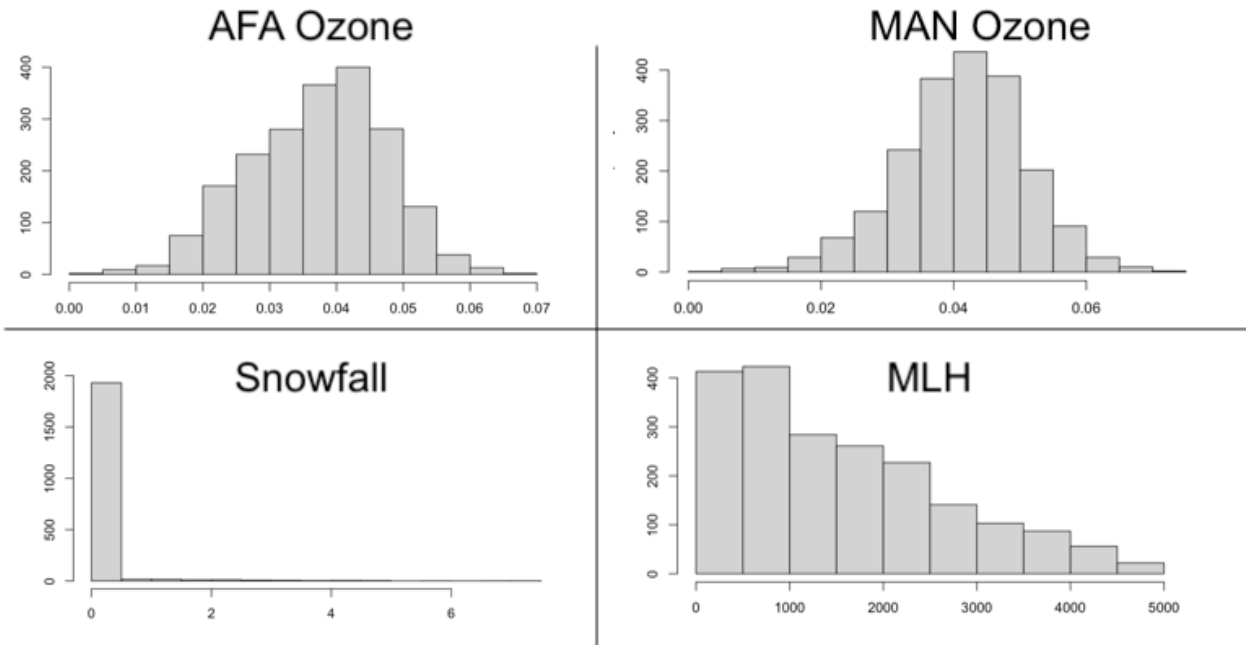
Correlation Plot of the Final Modeling Dataset



Note. AFA and MAN ozone levels display strong collinearity. Relatively strong correlations between ozone and temperature and humidity agree with relative influence of variables in the final model fits.

Figure 5

Distributions of Normal and Non-normal Variables

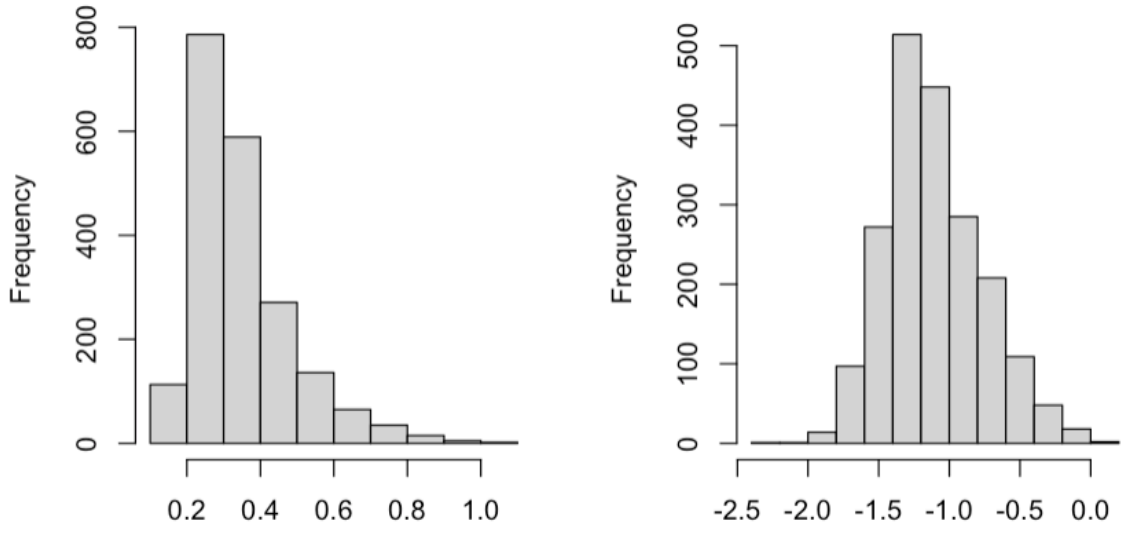


Note. Histograms showing select variable distributions. Ozone concentrations at AFA and MAN followed normal distributions. Other predictors violated assumptions of normality, such as the snowfall and average MLH distributions shown here.

To better fit the assumptions of linear models, select non-linear terms were replaced by mathematical transformations in a copy of the final dataset. Average daily mixing height was replaced by a power transformation of the original variable to remove nonlinearity (Jim, 2022). The values for daily maximum 8-hour average carbon monoxide concentrations and daily maximum average 5-minute sulfur dioxide concentration were replaced by the logarithm of those features. As shown in Figure 6, these transformed variables exhibited normal distributed data points.

Figure 6

Average 8-hr Carbon Monoxide Before and After Logarithmic Transformation



Note. The histogram of the untransformed average 8-hr carbon monoxide concentration is shown on the left, exhibiting pronounced right skew. The histogram on the right shows the variable normally distributed after values were logarithmically transformed.

Transformations for the precipitation, snowfall, and snow depth variables did not show significant improvements towards linear modeling assumptions. Neither the original nor transformed values for these variables were included in the final transformed dataset.

Preliminary Modeling and Hyperparameter Tuning

Before the final model suite was run, preliminary testing was conducted to determine the best hyperparameter values. Three models that have an adjustable penalty or shrinkage hyperparameter in caret were selected: linear regression with LASSO penalty, gradient boosted trees, and artificial neural net. These three models were run in 5-fold inner and 5-fold outer cross

validation to assess the best lambda at each outer fold. Iterative testing was performed until the best values of the respective hyperparameter were located firmly within the distribution and not near the fringes. Hyperparameter vectors were tested for AFA and MAN ozone models using both the raw and transformed datasets, resulting in four unique sets of tuning values for each run of the model suite.

Final Modeling Suite

8 models were selected to consider a wide variety of techniques and compare results to previous modeling efforts. All models were trained with 5-fold cross-validation using the 'caret' package in R. An outer layer of 5-fold cross validation was applied to assess the error attributable to the model-fitting process. For both layers of cross validation, 5-folds was selected as a reasonable tradeoff between accuracy and computational feasibility.

Linear Models

Linear models in the final modeling suite were linear regression, regression with forward selection, LASSO penalized regression, and generalized linear regression. Linear models are easily interpretable and can provide an advantageous option in instances of similar accuracy between models. Also, the transformed dataset was prepared for linear modeling to achieve accuracy despite the nonlinearity of original data.

GAM

Generalized additive models (GAM) were included to follow modeling efforts by Flynn et al. (2021). The GAM in this project used restrictive maximum likelihood method to determine penalized p-spline smoothing parameters (Larsen, 2015).

Random Forest, Gradient Boosted trees, and Artificial Neural Net

Random forest, gradient boosted trees, and artificial neural net models require no assumptions about variable distributions. Each of these models was fitted with model parameters deemed reasonable tradeoffs for model complexity and computation feasibility, shown in Table 1.

Final Model Fitting

For thoroughness, the entire suite of machine learning models was used to generate predictions for MAN and AFA ozone levels using both the original and transformed datasets. The outer layer of cross validation was used to evaluate the amount of variation in the dependent variable explained by the entire model fitting process. For each execution of the molding suite, the best model was selected and fit to the entire dataset using 5-fold cross validation for assessment.

Table 1

Model Hyperparameter Values For Nonlinear Models

Model	Hyperparameter Name	Hyperparameter Description	Setting or Range
RF	mtry	Input features considered	1 through 5
GBT	n.trees	Number of trees	300
GBT	n.minobsinnode	Observations in terminal nodes	10
ANN	size	Units in hidden layer	1 through 3

Note. Unique hyperparameters are shown for random forest (RF), gradient boosted trees (GBT), and artificial neural net (ANN) models. Additional lambda shrinkage hyperparameters were separately tuned for accuracy.

The machine learning models considered were include as comparison with previous efforts or display the potential to produce an accurate ozone model. Random forest, gradient boosted trees, and neural net models were considered promising candidates based on published literature and represent novel technique applications in the study area (Siwek & Osowski, 2016; Mohan & Saranya, 2018; Yafouz et al., 2021; Aljanabi et al., 2020). Additional methods considered across literature but not found to produce best fit, such as support vector regression, were not included here (Aljanabi et al., 2020; Bhuiyan et al., 2020; Cheng& Huang, 2021). With the inclusion of both techniques for comparison and advanced, locally untested approaches, this study was designed with the potential for the most accurate ozone model yet produced for the Colorado Springs area.

Chapter 4: Results

Overview

Four modeling options were investigated: two each that modeled MAN and AFA ozone as the dependent variables, and two each that used datasets with original or transformed variables. Before model testing, hyperparameter values were established. The modeling suite was executed four times, resulting in a best model selection for each monitor-dataset combination. Each best model was fit to the entire dataset using 10-fold cross validation. Answering the six research questions stated in Chapter 1, the results identified the most accurate overall models, examined which dataset produced the most accurate results, scored the accuracy of the final models, identified which variables were most influential in the final models, reported information relating the predictive quality of atmospheric mixing height, and offered insight to the feasibility of modeling as a forecasting tool.

Hyperparameter Tuning Values

Prior to final predictive modeling, unique vectors of hyperparameter values were established for AFA and MAN models using transformed and untransformed datasets, resulting in four unique sets of tuned hyperparameter values. Four distinct sets were necessary to achieve the best results possible in each execution of the modeling suite. The hyperparameters included in each set consisted of lambda shrinkage factors for LASSO regression, gradient boosted trees, and artificial neural nets. Table 2 summarizes the statistical measures that resulted in the lowest RMSE from the overall hyperparameter tuning process.

Modeling Suite

Table 2

Final Lambda Values After Hyperparameter Tuning

Model	AFA Original	AFA Transformed	MAN Original	MAN Transformed
LASSO	0.000012	0.000017	0.000028	0.000015
Boosted Tree	0.1	0.15	0.13	0.14
ANN	0.0008	0.0008	0.0006	0.006

Note. LASSO lambda rounded to the nearest 0.000001, gradient boosted tree rounded to nearest 0.1, and neural net rounded to nearest 0.0001.

A common suite of 8 machine learning techniques was used to model ozone at AFA and MAN sensors. As noted, 3 of those methods used tuned hyperparameter values to iteratively re-calculate the model and select the fit of least error. Models were constructed using the original and transformed datasets for each predictor. Results for the overall run of each suite are summarized in Table 3.

Best Model

For each test, gradient boosted tree models produced the best results. Use of this technique constitutes a novel application for predicting ozone in this area, and the method's superior performance supports nonlinear variable relationships. Predictions for MAN were more accurate than for AFA. The R^2 scores ranged from 0.64 to 0.69 for the gradient boosted trees best models selected from each run of the modeling suite. While not exceptionally robust, these scores approach the threshold for strong results ($R^2 = 0.7$). The research question of final accuracy was assessed final fits of each selected best model to the entire dataset.

Table 3

Results of 5-fold Cross-validated Modeling Suite with Each Monitor and Dataset

Measure	AFA Original	AFA Transformed	MAN Original	MAN Transformed
RMSE	0.000031	0.000047	0.000020	0.000039
R^2	0.65	0.64	0.69	0.68

Note. RSME and R2 are rounded to the nearest 0.000001 and .01, respectively.

Dataset Selection

For both sensors, the original dataset yielded more accurate predictions than models built with the transformed datasets. These results further support the consensus that ozone exhibits nonlinear associations to atmospheric parameters. Achieving the lowest RMSE of 0.00002 and highest R^2 of 0.69, models predicting MAN with the original dataset performed the best of all monitor-dataset combinations.

Final Fit

The best model from each modeling suite was fitted to the entire data set with 10-fold cross validation using the methods from the 'caret' software package. For each fit, the hyperparameters of the single best model produced in the modeling suite were used over the whole dataset. The measures of each fit are summarized in Table 4, along with the lambda shrinkage hyperparameter. The boosted tree model hyperparameters of interaction depth and number of trees were 3 and 300, respectively, for each best fit model.

Final Accuracy

Table 4

Comparison of the Final Fit for Each Predictor and Dataset

Measure	AFA Original	AFA Transformed	MAN Original	MAN Transformed
RMSE	0.005964741	0.006051406	0.005230584	0.005338716
R^2	0.66	0.65	0.71	0.69
MAE	0.004751047	0.004801719	0.004119809	0.004176643
Lambda (shrinkage)	0.1	0.13	0.13	0.1

Note. RSME and R2 are rounded to the nearest 0.000001 and .01, respectively.

Like the 5-fold cross validated tests, the model for MAN using the original dataset yielded the highest R^2 with the least RMSE. Also, like the 5-fold model suite results, unmodified datasets outperformed their transformed counterparts. With 10-fold cross validation, the final model fits produced better results than the 5-fold model selection process.

The best R^2 produced for AFA was 0.66, and for MAN 0.71, the latter crossing the threshold for robust results. With this accuracy, machine learning applications appeared promising as a forecasting tool. Further discussion on final model forecasting value is included in the next chapter.

Variable Relative Influence

For all final model fits, by a substantial margin the three most influential variables were Highway 24 (HW24) daily average temperature, HW24 relative humidity, and date. Table 5

Table 5

Relative Influence of Variables in Final Model Set for Each Dataset

Variable	AFA	AFA Transformed	MAN	MAN Transformed
Carbon monoxide (Log)	4.7647504	5.771690	3.733167	4.119937
Resultant Wind Direction	3.0480237	3.909948	3.105744	2.614680
Sulfur Dioxide (Log)	4.7069591	5.564657	**5.020135**	**5.316116**
Relative Humidity	**12.5769072**	**12.544015**	**22.328834**	**23.376478**
Temperature	**32.5891029**	**33.562502**	**34.221854**	**33.421554**
Standard Deviation of Horizontal Wind Direction	2.8631704	2.995859	3.247660	3.122461
Average Mixing Height (Power)	**8.2405598**	**6.639725**	**5.818712**	**8.500795**
Direction of Fastest 2-minute Wind	2.2546282	2.502889	1.294473	1.463324
Speed of Fastest 2-minute Wind	**10.3256592**	**9.633455**	1.566406	1.304768
Precipitation	1.1694869	—	1.111527	—
Snowfall	0.9622248	—	1.220196	—
Snow Depth	1.3205780	—	1.584503	—
Date	**15.1779493**	**16.875260**	**15.746790**	**16.759888**

Note. The top 5 values for each column are shown in bold. For each of the transformed variables the transformation type is shown in parentheses. The relative influence of precipitation, snowfall, and snow depth were not available in the transformed datasets as those variables were omitted.

illustrates the relative influence value of each variable across each final model fit. The top three variables are in bold.

For AFA and MAN, the final fits of original and transformed datasets exhibited the same ranking of variables by relative influence. Temperature was the most influential variable in each of the final modeling fits, positively correlated with ozone levels (Figure 3). Relative humidity, negatively correlated with ozone levels and temperature, was the second most influential variable for the MAN models, and third for AFA. Date ranked as the second for AFA and third for MAN.

For the final fits of all datasets, the daily average MLH variable or mathematical transformation thereof was the fourth ranked variable on influence, displaying moderate predictive value. For the AFA models, the fastest 2-minute wind speed also ranked in the top five influential features. The fifth most predictive variable in the MAN models was sulfur dioxide.

Summary

The best predictions for each monitor were generated with a gradient boosted tree model using the dataset with original variables. With R^2 values of 0.66 and 0.71 for AFA and MAN, respectively, the final fitted models in this study explain 26-29% more of the variance in the ozone concentrations than previous modeling attempts for the Colorado Springs area (Flynn et al., 2021). These results indicated promising implications for forecasting application. Supplementally, MLH showed moderate predictive value, ranking as the fourth most influential variable in each final model.

In the next chapter, the implications of these results are explored and the value of predictive modeling in ozone forecasting quantified.

Chapter 5: Discussion and Conclusion

Introduction

Gradient boosted trees model was selected as the final model type, pointing to nonlinear relationships between ozone and the predictor variables. The best fit model performed above the R^2 threshold of 0.70, indicating strong results. As forecasting tools, these models could reduce residential healthcare costs and save the area hundreds of thousands to millions of dollars in control measure costs by avoiding regulatory noncompliance.

Relative variable influences is mostly consistent with established theory (Camalier et al., 2007). As an investigated feature, MLH showed moderate prognostic value. More representative values for MLH would be needed to fully evaluate the relationship to ozone. Additional project limitations include the scope, available and selected data, missing values, reasonability considerations. A variety of options exist for future research, including new modeling targets, or refinement of Colorado Springs models through expanded model parameters, increased cross validation levels, additional predictive features, or a more representative MLH parameter.

Summary of Findings

These findings answered the research objectives set forth in Chapter 1. Out of the models considered, gradient boosted trees produced the most accurate results. Models produced with the original dataset performed better than those trained on the transformed dataset. The 10-fold cross-validated fit for the best model for AFA reported an R^2 score of 0.65. The 10-fold cross-validated fit for the best model for MAN and best model overall resulted in an R^2 of 0.71. Temperature, relative humidity, calendar date and MLH features were the most influential variables in the final models, along with 2-minute fastest wind speed for AFA and sulfur dioxide

measurements for MAN. As the variable ranked fourth for relative influence in each final model fit, MLH demonstrated moderate but notable predictive value. Project results indicate that data mining techniques are sufficiently accurate for forecasting application.

Best Model

For every modeling suite tested, gradient boosted trees generated the best results. This method is known for reducing bias errors in predictions. Simple models are sequentially constructed and assembled into a compound model that can produce robust predictions even when aggregating individually weak learners (History of Data Science, 2021). Reduced bias may potentially explain the relative performance of this method compared to the rest of the other models tested.

Original vs. Transformed Dataset

For both AFA and MAN ozone monitors, models generated more accurate predictions when built on datasets featuring original predictor values as opposed to transformed values. The transformed datasets were primarily included to meet the assumptions of linear modeling. With the notably nonlinear relationship between ozone and atmospheric predictors, however, it is unsurprising that nonlinear techniques outperformed the regression models(Di et al., 2016). The best results were produced from these nonlinear modeling techniques using the dataset with untransformed variables.

Accuracy of Best Fit Models

The R^2 scores of the final fit of best models for AFA and MAN were 0.65 and 0.71, respectively. While not exceptionally accurate, these models performed considerably better than previously produced predictive models for the area (Flynn et al., 2021). Results from the 5-fold

cross-validated modeling suite tests were not as accurate as the 10-fold cross-validated final fit. Further increased cross-validation levels may have produced more accurate results.

Most Influential Variables

The most influential variable in each of the final models was temperature. Elevated temperature is known to catalyze ozone producing reactions, possibly explaining the strong relative influence and positive correlation in the dataset (Figure 4) (Camalier et al, 2007; United States Environmental Protection Agency, 2023d). Furthermore, high temperature days in the dataset likely reflected sunny days. Solar radiation additionally catalyzes ozone production.

Ranked second most influential variable of MAN and third for AFA, relative humidity is negatively correlated with ozone (Figure 4). Conditions associated with increased humidity, including cooler temperatures and cloudy weather, are associated with decreased ozone production (Camalier, et al. 2007; United States Environmental Protection Agency, 2023g). The calendar date field was found to the second most influential variable for the AFA models and third for MAN. This influence likely stemmed from seasonable ozone patterns, as the highest ozone levels in the dataset occurred during the warm and sunny summer months (Figure 1). The ‘date’ datatype in R allowed the model to interpret and use this data.

Additionally included in the top 5 most influential features for AFA and MAN, respectively, were fastest 2-minute wind speed and sulfur dioxide concentration, respectively. Positive correlation between fastest 2-minute wind speed and ozone levels may indicate that transport plays a role in ozone levels at the AFA monitor. Although thought of as a precursor chemical, sulfur dioxide exhibited a negative correlation with ozone at both monitors.

MLH Predictive Value

A supplementary objective of this study was to ascertain the prognostic value unique MLH. In each final model fit, MLH or transformations thereof ranked the fourth most influential variable, demonstrating moderate predictive quality. Untransformed MLH was positively correlated with ozone at both sensors, but a more detailed interpretation of this relationship would require a thorough analysis of regional atmospheric patterns.

Forecasting Feasibility

The results of this study indicate that data mining techniques show strong potential as forecasting tools. An R^2 score of 0.71 was a promising increase from previous modeling attempts and may be accurate enough to assist stakeholders in safeguarding public health and managing regulatory compliance. Further testing may result in increased accuracy for gradient boosted trees or additional methods to the benefit of ozone forecasting.

Implications for Stakeholders

The results of this study indicated that data mining applications present feasible forecasting advantages. Model-enhanced ozone forecasting could benefit government, business, and residential stakeholders. Model-assisted forecasting could advantage Colorado Springs residents and businesses by mitigating the incurrence of respiratory healthcare costs and supporting regulatory compliance to avoid stringent restrictions.

Benefitted Parties

Air Quality Professionals. The ozone models produced in this study could benefit air quality professionals and researchers. Models would be most useful to frontline ozone forecasters, such as the team at CDPHE. State, federal or academic environmental researchers

may also use predictive ozone modeling for long-term projections and examination of pollutant trends.

Colorado Springs Residents. As a local benefit, accurate forecasting could provide value to the entire Colorado Springs community. Model-enhanced forecasting could provide private citizens precise air quality information that may result in reduced healthcare costs, especially for residents with respiratory conditions. Predictive models may also represent value to government officials and business leaders as they strive for compliance with air quality regulations to avoid the economic consequences of noncompliance designation.

Local Valuation of an Accurate Forecast

The monetary value of model-enhanced forecasting may be estimated through mitigation of both citizen healthcare expenses and associated costs for nonattainment compliance actions. Value-added to affected citizens and businesses by model-enhanced ozone forecasting is quantified in the next paragraphs.

Healthcare Costs. Ground-level ozone can precipitate many health defects; here the focus was narrowed to respiratory effects, specifically aggravated asthma. Ozone has been shown to trigger asthma attacks which may require extended use of an individual's medical inhaler. Inhalers recently underwent a price increase due to tighter environmental regulations (Plain, 2015). Increased usage due to ozone-induced asthma attacks may require the acquisition of additional inhalers, further straining affordability. In 2022, El Paso County, CO, where Colorado Springs is located, recorded an asthma rate of 23.3 in every 10,000 persons, or 0.233% (Colorado Department of Public Health and Environment, 2023b). Using this rate for the population of Colorado Springs with the prices of Advair and Flovent, the two most popular inhaler brands, the

total overall cost of high ozone resulting in each asthmatic individual requiring a single additional inhaler per year was estimated between $329,264 and $559,298 (Marsh & van Meijgaard, 2020; United States Census Bureau, 2022).

Accurate ozone forecasting can reduce the economic burden on asthmatic Colorado Springs residents and their families by correctly identifying and warning of ozone levels that might trigger an attack. Indoor ozone concentrations may be as low as 20% of outdoor levels, so mitigation of outdoor air exposure on high ozone days may mitigate the need for increased inhaler usage (United States Environmental Protection Agency, 2023b).

Noncompliance Costs. Noncompliance designations have proven quite expensive and disruptive to emissions producing businesses in designated areas (Blankenheim, 2013; Stanley, 2017). Control measures required for listed areas may cost millions to businesses, municipalities, and residents (Cummings & Hill, 2022; Ochse, 2022). RFG, a control measure currently being implemented in the Denver metropolitan area, may cause fuel prices to increase as much as $1 per gallon and incur an overall cost of $800 million to $1 billion. The population of Colorado Springs is roughly 70% that of Denver, but even at a proportionate cost, fuel restrictions in the Colorado Springs could cost as much as $560 million overall. Additionally, increased restrictions on oil and gas production could jeopardize 720,000 jobs throughout Colorado, many of which are along the front range (Jaffe, 2022; Bureau of Labor Statistics, 2023a). Analysis of the economic makeup of areas in nonattainment revealed that emission-producing industries declined by almost 25%. As the major ozone contributing emitter, the petroleum industry would face the greatest regulatory restrictions (Cheadle et al., 2017). Forecasting may provide a tool to avoid noncompliance classification and subsequent economic consequences.

Model-enhanced forecasting could help government and business leaders coordinate to reduce emissions and avoid triggering regulatory violations for ozone. If ozone is predicted to be high, emissions-producing businesses like petroleum production could temporarily pause some portion of their operations until conditions no longer threaten an NAAQ ozone violation. Predictive modeling could help preserve up to a quarter of oil and gas jobs in the area, as well as save the local economy tens to hundreds of millions of dollars (Stanley, 2017; Determinations of Attainment by the Attainment Date, Extensions of the Attainment Date, and Reclassification of Areas Classified as Serious for the 2008 Ozone National Ambient Air Quality Standards, 2022).

Modeling results may also help prove uncontrollable air quality conditions in the event of elevated ozone levels. CDPHE Air Pollution Control Division estimated that 71% of the ozone experienced statewide was beyond control, with wildfires contributing as much as 10% (Ochse, 2022). In 2020, the five days with the highest ozone concentrations in Colorado Springs all occurred during significant wildfire pollution events (Flynn et al., 2021). As the predictive models in this study did not incorporate smoke data, differences between the predictions and actual observations during observed high ozone events could help quantify the ozone contribution of smoke pollution or other uncontrollable sources (Gong et al., 2017). Exemptions exist if pollution levels were attributable to an uncontrollable, regional air quality event.

Implications for Further Application

The final selected models in this study predicted ozone levels only for the Colorado Springs area. In broader implementation, a series of locally optimized models could assist ozone forecasting across the state. Predictive modeling could be especially useful to protect public

health by supporting predictions for additional urban areas that experience high ozone levels, similar to conditions in Colorado Springs.

Result Limitations

The results faced many limitations in the study design and available resources. As noted, scope was restricted to the Colorado Springs area. Ozone was modeled at only two monitors. Only air quality and meteorological data were used to construct the modeling datasets. Some variables were not complete enough to be included in the study, such as particulate matter and NOx measurements. This reduced the overall number of predictors available for modeling and omitted potentially prognostic data.

Mean Values

This study used only mean daily summary parameters in the modeling datasets. Features summarizing additional statistics, such as daily maximum and minimum readings, were omitted. Exclusively mean values were selected to manage collinearity that may have resulted from including multiple statistics for the same parameter. Additionally, mean values matched the dependent variable of maximum daily 8-hour average ozone concentration. The predictive qualities of additional statistical values were unexplored.

MLH Measurement Source

MLH readings taken in the study area for the examined 2015-2022 interval were not available, so the nearest published data was sourced from the former site of the Stapleton Airport in Denver. The geographical distance, however, may compromise the representativeness of the parameter for Colorado Springs. Furthermore, daily summary values were not available for MLH, so averages were computed based on available data. Before averaging, each calendar date

included an inconsistent number of readings - including dates with no measurements. This introduced uncertainty, increased variance, and generated quantities of missing data in the resultant daily average parameter. A significant quantity of missing MLH data was imputed using 2 nearest neighbors, introducing a final layer of uncertainty. While MLH was found to be the fourth most influential parameter in all best fit gradient boosted trees models, the feature may not be representative of actual study area conditions.

Reasonability Considerations

Modeling suite settings, including individual model parameters and cross validation levels, were selected with consideration of computational resources and execution time. The number of trees used in gradient boosted tree models was held constant at 300 throughout evaluation, and interaction depth only tested from integer values 1 to 3. Further adjustment of these parameters may have resulted in more accurate models. Only a 5-fold cross validation level was applied to inner and outer folds of the modeling suite. In the final fits of the selected models, a portion of increased accuracy was likely attributable to the increased 10-fold cross validation level. Increased cross validation levels would have potentially increased accuracy across all models.

Suggestions for Future Research

Future research may apply a similar approach to this project in a new study area to determine the best fit models and achievable accuracy with available data. Alternatively, further efforts could drill down on accuracy for Colorado Springs ozone models. In the interest of producing more accurate predictions, future work on gradient boosted modeling could investigate higher interaction depths, different number of trees available to the model, and increased cross validation levels. Furthermore, additional data parameters could be incorporated

into modeling sets, such as emissions (e.g., traffic) or wildfire data. Minimum and maximum air quality and meteorological variables could be examined for advantage to accuracy. Either through a new data source or calculation method, new MLH values could be established that are more representative of real-world conditions. Finally, future research could explore novel machine learning techniques that may generate more accurate results.

Conclusion

Predictive modeling shows advantageous potential for ozone forecasting in the Colorado Springs area. The selected gradient boosted tree models showed significant improvement from previous molding attempts and proved that a robust local ozone model ($R^2 = 0.71$) was feasible (Flynn et al., 2021). MLH has shown moderate influence in the final models, but values may not be sufficiently representative of study area conditions for a precise assessment.

Forecasting application of the resultant models could save asthmatic Colorado Springs residents an overall total of $300,000-$500,000 annually. Robust ozone predictions could help local leaders avoid noncompliance classification, sparing the local economy as much as $560 million in mandated control measures while safeguarding regional oil and gas production against detrimental restrictions.

Result limitations stem from the decision to include mean values exclusively in the modeling datasets, omitting additional statistics. This design of this project incorporated reasonable time and computational resource constraints.

Future research may consist of data mining application in novel study areas, or continued work towards refining the accuracy of Colorado Springs predictions. Many avenues are possible

for future research that may overcome the limitations of this study and produce a more accurate model for local ozone forecasting.

References

Aljanabi, M., Shkoukani, M., & Hijjawi, M. (2020). Ground-level ozone prediction using machine learning techniques: A case study in Amman, Jordan. *International Journal of Automation and Computing, 17*, 667-677. https://doi.org/10.1007/s11633-020-1233-4

Avnery, S., Mauzerall, D. L., Liu, J., & Horowitz, L. W. (2011, April). Global crop yield reductions due to surface ozone exposure: 1. Year 2000 crop production losses and economic damage. *Atmospheric Environment*, *45*(13), 2284-2296. https://doi.org10.1016/j.atmosenv.2010.11.045

Bhuiyan, M., Mahmud, S., Sarmin, N., & Elahee, S. (2020, October 7). A study on statistical data mining algorithms for the prediction of ground-level ozone concentration in the El Paso–Juarez area. *Aerosol Science and Engineering, 4*, 293-305. https://doi.org/10.1007/s41810-020-00074-2

Bishop, G., Hoekman, S., & Broch, A. (2018). *Evaluation of emissions benefits of federal reformulated gasoline versus conventional gasoline*. University of Denver. https://digitalcommons.du.edu/cgi/viewcontent.cgi?article=1085&context=feat_publications

Blankenheim, C. (2013, April 10). *Estimating the economic impact of ozone and fine particulate nonattainment in the twin cities*. [Unpublished master's thesis]. The University of Minnesota.

Boyce, D. (2023, July 7). Colorado Springs is breaking the EPA's ozone limit. Officials say wildfire smoke is partly to blame. *CPR News.* https://www.cpr.org/2023/07/07/colorado-springs-is-breaking-the-epas-ozone-limit-officials-say-wildfire-smoke-is-partly-to-blame/

Bureau of Labor Statistics. (2023a, April 25). *OEWS research estimates by state and industry.*

https://www.bls.gov/regions/mountain-plains/summary/blssummary_coloradosprings.pdf

Bureau of Labor Statistics. (2023b, November 30). *Colorado Springs Area Economic Summary.* https://www.bls.gov/oes/current/oes_research_estimates.htm

Camalier, L., Cox, W., & Dolwick, P. (2007, October). The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. *Atmospheric Environment*, *41*(33), 7127-7137. https://doi.org/10.1016/j.atmosenv.2007.04.061

Capital Area Council of Governments. (n.d.). What is ground-level ozone? *Air Central Texas*. https://aircentraltexas.org/en/regional-air-quality/what-is-ground-level-ozone

Cheadle, L., Oltmans, S., Pétron, G., Schnell, R., Mattson, E., Herndon, S., Thompson, A., Blake, D., & McClure-Begley, A. (2017, November 3). Surface ozone in the Colorado northern Front Range and the influence of oil and gas development during FRAPPE/ DISCOVER-AQ in summer 2014. *Elementa: Science of the Anthropocene* 5(61). https://doi.org/10.1525/elementa.254

Cheng, Y., He, L. Y., & Huang, X. F. (2021, August 31). Development of a high-performance machine learning model to predict ground ozone pollution in typical cities of China. *Journal of Environmental Management*, *299*. https://doi.org/10.1016/j.jenvman.2021.113670

Colorado Department of Public Health and Environment. (2023a). *Air quality advisories*. https://cdphe.colorado.gov/public-information/air-quality-advisories

Colorado Department of Public Health and Environment. (2023b). Asthma data. *Environmental Public Health Tracking*. https://coepht.colorado.gov/asthma-data

Cox Jr, L. A. (2017, May). Socioeconomic and air pollution correlates of adult asthma, heart

attack, and stroke risks in the United States, 2010–2013. *Environmental Research*, *155*, 92-107. https://doi.org/10.1016/j.envres.2017.01.003

CRAN. (n.d.). *Variable selection methods*. https://cran.r-project.org/web/packages/olsrr/vignettes/variable_selection.html

Cummings, J. & Hill, T. (2022, May 10). Without EPA waiver, price at pump will skyrocket. *Colorado Politics*. https://www.coloradopolitics.com/opinion/without-epa-waiver-price-at-pump-will-skyrocket/article_29f884f6-cfef-11ec-8189-afaaef6c1584.html

Determinations of Attainment by the Attainment Date, Extensions of the Attainment Date, and Reclassification of Areas Classified as Serious for the 2008 Ozone National Ambient Air Quality Standards, 87 FR 60926 (proposed 2022, October 7) (to be codified at 40 C.F.R. § 1090.285). https://www.govinfo.gov/content/pkg/FR-2022-10-07/pdf/2022-20458.pdf

Di, Q., Rowland, S., Koutrakis, P., & Schwartz, J. (2017). A hybrid model for spatially and temporally resolved ozone exposures in the continental United States. *Journal of the Air & Waste Management Association, 67*(1), 39-52. https://doi.org/10.1080/10962247.2016.1200159

Du, J., Qiao, F., Lu, P., & Yu, L. (2022, September). Forecasting ground-level ozone concentration levels using machine learning. *Resources, Conservation and Recycling, 184.* https://doi.org/10.1016/j.resconrec.2022.106380

Federal Reserve Bank of Minneapolis. (2023). *Inflation calculator.* https://www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator

Flynn, M. T., Mattson, E. J., Jaffe, D. A., & Gratz, L. E. (2021, May 19). Spatial patterns in summertime surface ozone in the Southern Front Range of the U.S. Rocky Mountains.

Elementa: Science of the Anthropocene 9(1). https://doi.org/10.1525/elementa.2020.00104

Frost, J. (2023). How to interpret adjusted R-squared and predicted R-squared in regression analysis. *Statistics by Jim*. https://statisticsbyjim.com/regression/interpret-adjusted-r-squared-predicted-r-squared-regression/

Geiß, A., Wiegner, M., Bonn, B., Schäfer, K., Forkel, R., von Schneidemesser, E., Münkel, C., Chan, K., & Nothard, R. (2017, August 18). Mixing layer height as an indicator for urban air quality? *Atmospheric Measurement Techniques, 10*(8), 2969-2988. https://doi.org/10.5194/amt-10-2969-2017

Gilman, S. (2017). Colorado's Ground-Level Ozone Burden. *Colo. Nat. Resources Energy & Envtl. L. Rev.*, *28*(1), 283-310. https://www.colorado.edu/law/sites/default/files/attached-files/gilman_final.pdf

Gold, D. R., & Samet, J. M. (2013). Air pollution, climate, and heart disease. *Circulation*, *128*(21), e411-e414. https://doi.org/10.1161/CIRCULATIONAHA.113.003988

Gong, X., Kaulfus, A., Nair, U., & Jaffe, D. A. (2017, October 25). Quantifying O3 impacts in urban areas due to wildfires using a generalized additive model. *Environmental science & technology, 51*(22), 13216-13223. https://doi.org/10.1021/acs.est.7b03130

History of Data Science (2021, August 27). *Gradient boosting algorithms: busting bias error*. https://www.historyofdatascience.com/gradient-boosting-algorithms-busting-bias-error/

Hsu, Y. (2020, December 28). Create Predictive Classification Models in R with Caret. *Medium*. https://yuenhsu.medium.com/create-predictive-classification-models-in-r-with-caret-19a83c1b742

Ishak, A., Daoud, M., & Trabelsi, a. (2017). Ozone concentration forecasting using statistical learning approaches. *Journal of Materials and Environmental Sciences*, *8*(12), 74532-4543. https://doi.org/10.26872/jmes.2017.8.12.478

Jaffe, M. (2022, April 11). Air emissions, water demands skyrocket as 72% of Colorado's new oil and gas activity centers on the Front Range. *The Colorado Sun*. https://coloradosun.com/2023/04/11/oil-gas-industry-air-pollution-water-2022/

Jim. (2022, October 23). Box Cox transformation in R. *R-Bloggers*. https://www.r-bloggers.com/2022/10/box-cox-transformation-in-r/

Kassambara, A. (2018, November 3). Linear regression assumptions and diagnostics in R: essentials. *Statistical tools for high-throughput data analysis*. http://www.sthda.com/English/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

Laden, R. (2023, September 5). Got growth? Colorado Springs residents, developers debate as decades-long trends continue. *The Gazette*. https://gazette.com/business/got-growth-colorado-springs-residents-developers-debate-as-decades-long-trends-continue/article_a078b182-4136-11ee-bd04-0b80cdd2449c.html

Larsen, K. (2015, July 30). GAM: the predictive modeling silver bullet. *Multithreaded*. https://multithreaded.stitchfix.com/blog/2015/07/30/gam/

Lee, K. J., Kahng, H., Kim, S. B., & Park, S. K. (2018, December 2). Improving environmental sustainability by characterizing spatial and temporal concentrations of ozone. *Sustainability, 10*(12). https://doi.org/10.3390/su10124551

Marsh, T., & van Meijgaard, J. (2022, June 8). How much asthma inhalers cost and how to

save. *GoodRx*. https://www.goodrx.com/conditions/asthma/heres-why-asthma-inhalers-are-so-expensive

Mills, G., & Harmens, H. (2011, September). *Ozone pollution: A hidden threat to food security*. National Environmental Research Council Centre for Ecology & Hydrology. https://nora.nerc.ac.uk/id/eprint/15071/1/N015071CR.pdf

Mohan, S., & Saranya, P. (2019). A novel bagging ensemble approach for predicting summertime ground-level ozone concentration. *Journal of the Air & Waste Management Association, 69(*2), 220-233. https://doi.org/10.1080/10962247.2018.1534701

National Oceanic and Atmospheric Administration. (n.d.-a). Climate data online. *National Centers for Environmental Information*. https://www.ncei.noaa.gov/cdo-web/

National Oceanic and Atmospheric Administration. (n.d.-b). Integrated Global Radiosonde Archive. *National Centers for Environmental Information*. https://www.ncei.noaa.gov/cdo-web/

Nunes, L. (2022, June 30). Methods: how to do data visualization using R - even if you don't use R. *Observer, 35*(4). https://www.psychologicalscience.org/observer/methods-data-visualization-using-r

Ochse, C. (2022, August 3). *EPA decision impacts business permits and gas prices for Colorado*. Denver Metro Chamber of Commerce. https://denverchamber.org/2022/08/03/epa-decision-impacts-business-permits-and-gas-prices-for-colorado/

Parameswaran, A. (2022, Nov 24). Pandas vs. SQL — part 4: pandas is more convenient. *Towards Data Science*. https://towardsdatascience.com/pandas-vs-sql-part-4-pandas-is-more-convenient-8e9744e2cd10

Pikes Peak Area Council of Governments. (2021, January). Pikes Peak Region ozone advance program 2021. https://www.epa.gov/sites/default/files/2021-01/documents/co_pikes_peak_2020_path_forward.pdf

Plain, C. (2015, May 13). Inhaler ban increases costs for asthma patients. *University of Minnesota School of Public Health*. https://www.sph.umn.edu/news/inhaler-ban-increases-costs-for-asthma-patients/

Reformulated Gasoline Covered Areas, 40 C.F.R. § 1090.285 (2023). https://www.ecfr.gov/current/title-40/chapter-I/subchapter-U/part-1090

Siwek, K., & Osowski, S. (2016, July 1). Data mining methods for prediction of air pollution. *International Journal of Applied Mathematics and Computer Science*, *26*(2), 467-478. https://doi.org/10.1515/amcs-2016-0033

Stanley, J. (2017, November 23). Labor market impacts from ozone nonattainment status: a regression discontinuity analysis. *Environmental Economics and Policy Studies, 20*(3), 527-546. https://doi.org/10.1007/s10018-017-0204-7

United States Census Bureau. (2022, July 1). *QuickFacts Colorado Springs city Colorado*. https://www.census.gov/quickfacts/fact/table/coloradospringscitycolorado

United States Environmental Protection Agency. (2023a, March 15). *NAAQs Table*. https://www.epa.gov/criteria-air-pollutants/naaqs-table

United States Environmental Protection Agency.(2023b, April 20). *Health Effects of Ozone in the General Population*. https://www.epa.gov/ozone-pollution-and-your-patients-health/health-effects-ozone-general-population

United States Environmental Protection Agency. (2023c, May 24). *Health effects of ozone pollution*. https://www.epa.gov/ground-level-ozone-pollution/health-effects-ozone-pollution

United States Environmental Protection Agency. (2023d, June 2). *Ground-level ozone basics.* https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics#effects

United States Environmental Protection Agency. (2023e, August 22). Interactive Map of Air Quality Monitors. *Outdoor Air Quality Data*. https://www.epa.gov/outdoor-air-quality-data/interactive-map-air-quality-monitors

United States Environmental Protection Agency. (2023f, October 4). *Eight hour ozone concentrations.* https://www3.epa.gov/region1/airquality/avg8hr.html

United States Environmental Protection Agency. (2023g, October 4). *Trends in ozone adjusted for weather conditions.* https://www.epa.gov/air-trends/trends-ozone-adjusted-weather-conditions

Wang, W., Liu, X., Bi, J., & Liu, Y. (2022). A machine learning model to estimate ground-level ozone concentrations in California using TROPOMI data and high-resolution meteorology. *Environment International, 158*. https://doi. org/10.1016/j.envint.2021 .106917

Wang, X. Y., & Wang, K. C. (2014, June 12). Estimation of atmospheric mixing layer height from radiosonde data. *Atmospheric Measurement Techniques, 7*(6), 1701-1709. https://doi.org/10.5194/amt-7-1701-2014

Yafouz, A., Ahmed, A. N., Zaini, N. A., & El-Shafie, A. (2021). Ozone concentration forecasting based on artificial intelligence techniques: A systematic review. *Water, Air, & Soil Pollution, 232*. https://doi.org/10.1007/s11270-021-04989-5