A Machine Learning Approach to Forecasting and Estimating Ground Level Ozone

William Keenan

Department of Geography, University of Denver

Master's Thesis Proposal

1. Introduction

Ground-level ozone (GLO) is a major air pollutant that negatively impacts both human health and ecological health across the globe. It is one of six “criteria air pollutants” defined by the EPA and is generally considered to have one of the worst impacts on human health alongside particulate matter. Exposure to high levels of ozone can cause acute shortness of breath, asthma attacks, increased risk for pulmonary infection, and greatly increase risk of hospitalization for those with pre-existing conditions. Long term exposure has been linked to premature death in a plethora of studies (Zhang 2019, Cohen 2015).

Efforts to forecast and estimate ozone pollution levels have historically relied on computationally intensive numerical models that produce low accuracy and low-resolution results (Jiang 2020). The major detrimental health effects of GLO make it imperative that both policy makers and ordinary citizens have access to high accuracy, high-resolution forecasts of their local area. Better forecasts could be leveraged by environmental protection agencies to issue timely warnings to the citizens they serve. These forecasts would also allow health conscious and pollution-sensitive citizens to make better decisions about the activities they do and the locations they do them. Additionally, public health and medical researchers studying the negative impacts of air pollution could use high resolution maps to create accurate exposure estimates. These exposure estimates could be used alongside the mountain of health and spatial data many people collect on their smartwatches and smartphones to correlate asthmatic episodes with certain levels of ozone at the individual or population level.

Published studies using machine learning (ML) to forecast air pollution levels have increased dramatically in recent years and shown promising results (Masood 2021). This research will use a multi-source, ML approach to improve upon the current ozone forecast and estimation models. A long-short-term-memory (LSTM) machine learning model will predict ozone levels at each EPA monitoring site in the Denver Metro Area several hours in advance. Advanced interpolation techniques will then be applied to produce continuous surfaces at high spatial resolution and 1-hour temporal resolution. Model inputs will primarily be meteorological variables derived from NOAA's high-resolution rapid refresh (HRRR) weather model. Pollutant data will be acquired from the EPA's Air Quality System API and will include ozone (O_3), nitrogen dioxide (NO_2), and Particulate Matter 2.5 ($PM_{2.5}$). Static variables such as elevation, land use, and population density will be used to train models using geographically similar data. The resulting models and maps will test the feasibility of high-resolution ozone forecasting and demonstrate novel data source integrations.

2 Literature Review

In this section, I summarize the sources and patterns of GLO, the current state of GLO AI (artificial intelligence) forecasting and estimation research, and the research gaps my thesis intends to tackle.

2.1 Sources and Patterns of Ground Level Ozone

Ozone, O_3, is formed in the atmosphere when oxygen in its most common molecular structure, O_2, reacts with a single oxygen atom, O (Zhang 2019). In the stratosphere, this reaction occurs at a much higher rate than at ground level, forming the "good" ozone layer we rely on to protect us from UV radiation. In the troposphere, ozone exists naturally at about 20 – 45 parts per billion (ppb), a much lower concentration than is in the stratosphere or in highly polluted air (Chen 2017). GLO concentrations above this "background level ozone" occur because the anthropogenic burning of fossil fuels releases volatile organic compounds (VOCs) and nitrous oxides (NOx) into the atmosphere. These VOCs and NO_xs undergo a complex photochemical reaction in which one oxygen molecule is traded to O_2, forming O_3. This reaction is greatly hastened by sunlight and higher temperatures making GLO levels highly seasonal and diurnal. GLO concentrations peak in the summer and during the hottest, sunniest parts of the day. To illustrate this, figure 1 shows GLO levels for one EPA testing site in downtown Denver on June 19, 2021.

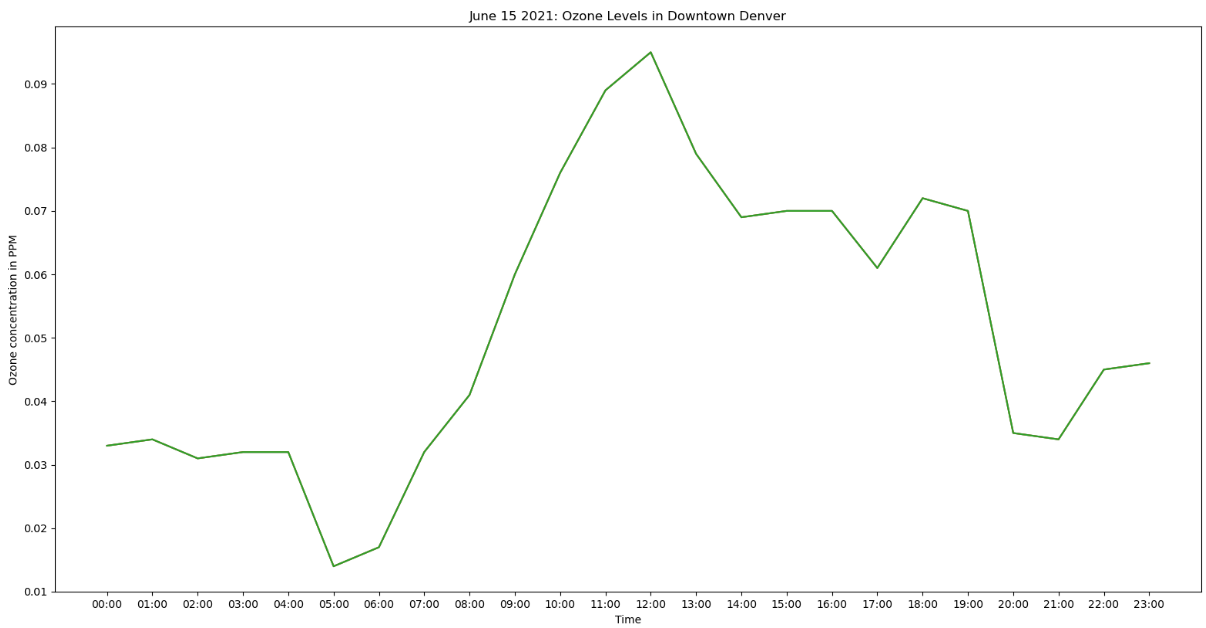


Figure 1. Data from the EPA's Air Quality System API (AQS)

This diurnal pattern is seen at all Ozone monitoring stations, regardless of pollution level. Areas such as this one with significant pollution, however, consistently peak at levels over the EPA's Ozone threshold of 70 ppb in the summer months.

The seasonal pattern of Ozone is shown in Figure 2.

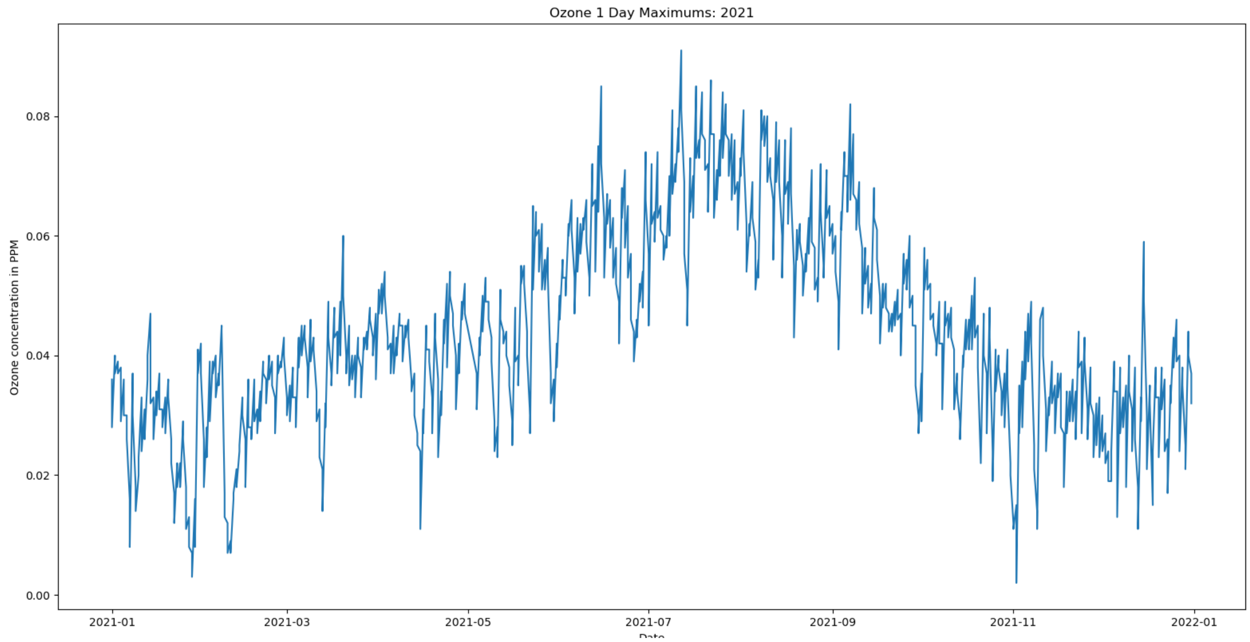


Figure 2. Data from the EPA's Air Quality System API (AQS)

While the current state of global ozone pollution is a direct result of anthropogenic pollution, high ozone levels are not necessarily spatially autocorrelated with the emission sources, such as traffic or oil and gas activity. In other words, an EPA monitoring station near a major highway does not necessarily record higher levels of ozone pollution than one farther from it. This is due to the "ozone paradox": there is a higher availability of NO (produced by traffic) close to the road, which reacts readily with O_3 to produce NO_2 and O_2 (Amato 2022). Another layer of complexity is added by the fact that the ozone a local population is exposed to may have been generated from precursors emitted hundreds of kilometers away due to a process called long-range transport (Glavas and Sazakli 2011). Additionally, because ozone is a naturally occurring particle in the atmosphere, it can be hard to differentiate between human-caused ozone spikes and natural ones. In summary, ozone's complex relationship with primary pollutants (VOCs and NOx), meteorological factors, and naturally occurring ozone make it especially hard to forecast.

2.2 Forecasting vs Estimation

Ground level ozone studies can be classified by whether their primary motivation is *forecasting* or *estimation*. Forecasting papers[1] primary goal is to predict ozone concentrations that will occur sometime in the future at known locations, i.e. over monitoring stations. These papers typically pick pollutant monitoring stations that have been operational for a long time and therefore have very continuous data. Additionally, a station is more likely to be used if meteorological data is easily acquired for that location. Freeman 2018 conducted one of the first studies applying deep learning techniques to ozone forecasting. The study used data from one air quality station and forecasted 8-hour averages using different forecast horizons (1 hour, 12 hour, and 24 hour). This study along with Krishan 2019, Du et al 2022, and Marvin 2022 characterize the forecasting category, which focuses on using relatively high temporal resolution, but extremely low or no spatial resolution. Papers in this category focus on optimizing the input variables and the ML model to achieve extremely high accuracy for a single location or set of locations but ignore the problem of estimating ozone at locations between monitoring stations. While these papers do provide some valuable insights, forecasting values at a single location is not very useful for decision makers who care about how ozone is distributed across their area of influence, and neither is it useful for citizens concerned about air quality in their neighborhood.

Studies in the estimation category[2] focus instead on estimating ozone values at unknown locations (locations in between ozone ground monitors), but do not try to forecast future values. These papers create many continuous surfaces at a certain time interval and spatial resolution. For example, Wang 2022 created daily maximum 8-hour average rasters over California at

[1] List of forecasting references: Li 2017, Freeman 2018, Krishan 2019, Du et al 2022, Ko 2022, and Marvin 2022

[2] List of estimation references: Wang 2022, Jiajia 2022, Crooks 2022, and Wang, Liu 2022

10km^2 resolution. A greater emphasis is placed on achieving high accuracy for unknown locations at high spatial resolutions, but low temporal resolutions. This area of research often combines remotely sensed data to fill in the gaps between monitoring stations and leverages deep learning to interpolate the surface from a variety of data sources. Estimation studies also face the challenge that different locations respond to independent variables in different ways. Unlike forecasting studies which often focus on using a single monitoring station, estimation studies use a network of stations to create a more general model which is not overfit to a single location. Similar to forecasting studies, the products of estimation studies are not useful for end users who care about air quality in real time. These studies don't attempt to forecast future values of ozone and often rely on data sources that aren't available soon enough to do so (i.e. remotely sensed data that is released days after it's collected).

Combining the fields of air quality estimation and forecasting with an ML approach is a challenging task that has not been fully explored. It requires fusing extremely heterogeneous data sources with different spatiotemporal resolutions, formats, and levels of precision. Once the data is compiled, the researcher must employ a variety of AI and statistical techniques to both forecast future values for known locations (monitoring sites) and interpolate the surface.

2.3 Air Pollution Modelling Summary

In addition to categorizing studies by forecasting or estimation, models for air quality research typically fit into three categories: statistical models, numerical models, and AI / ML models. In meteorology, numerical models are simulations of the atmosphere based on physical and chemicals laws. Two major numerical models which typically forecast other meteorological variables have been applied to forecast ozone: the Community Multiscale Air Quality (CMAQ) model and the Weather Research and Forecasting/Chemistry (WRF-chem) model (Zhou 2017, Jiang & Yoo 2018). These models must be run at the global scale and require massive computational resources (Wang, Liu 2022). Because of this, they typically have much lower spatial resolution than AI models. Jiang & Yoo 2018 attempted to forecast PM 2.5 at $12km^2$ and $4km^2$ resolutions using the CMAQ model and highlighted the high computational burden and error that comes with resolutions as low as $4km^2$. In contrast, AI models such as geo-intelligent light gradient boosting can be done without super-computers and achieve spatial resolutions of $1km^2$ (Jiajia 2022).

Statistical models for air pollution forecasting generally fall under the category of spatial linear Land Use Regression (LUR) models (Ren 2020). Ren et al conducted a comparison study between several ML and LUR ozone forecast models and concluded that ML models performed better in every category other than simplicity and processing time. LUR models fail to capture non-linear relationships between the variables which is critical in accurately forecasting GLO. For these reasons, the state-run air pollution forecast models of the future will likely depart from traditional statistical and numerical models and adopt artificial intelligence.

2.4 GLO Forecasting with Deep Neural Networks

Advancements in computing power, artificial intelligence, and data availability have led to an explosion of research interest in forecasting air pollutants such as ozone. Masood et al conducted a systematic literature review to capture this explosion and found that between 2014 and March 2021, fifty-eight publications were released using AI for air pollution forecasting, a large increase compared to the previous decade (Masood 2021). Most of this research took place in Asia and Europe, with very few papers utilizing study areas in North America or data sources with coverage over the US, which reveals a gap in the current literature.

Masood breaks down the studies into four categories of AI models: artificial neural networks (ANN), support vector machines, fuzzy logic systems, and deep neural networks (DNN). These categories are then compared across five major pollutants, including ozone, and ranked based on their R^2 and RMSE values. Four ozone studies were reviewed that utilize DNN forecasting models, and on average they greatly outperformed all other types of models. DNN models are very similar to ANN models but consist of more than 2 hidden layers. Because of their increased complexity, they require more computational time than ANNs, but can achieve higher levels of abstraction. One of the biggest benefits of DNNs is their ability to achieve feature extraction from raw inputs (Masood 2021). Among DNNs, the long short-term memory (LSTM) and convolutional neural network (CNN) models have been most successful and widely used (Kim 2018, Krishan 2019, Freeman 2018). Furthermore, compared to ANNs, applying DNNs to the problem of ozone concentration forecasting is a relatively new field.

2.5 Research Gaps of Interest

To date, no studies have combined the research strides in deep-learning driven forecasting and estimation to produce temporally and spatially continuous, high resolution ozone forecast maps. My research intends to bridge the gap between both areas of study to test the limits of data driven air quality forecasting and estimation. From the forecasting category, I will draw heavily from research in Krishan 2019 where an LSTM model was used to achieve highly accurate results (RMSE < 3 $\mu g/m^3$). From the estimation category, I will draw heavily from Jiajia 2022 and Wang 2022 where highly accurate estimations were created. Using this literature as general guidelines, I will attempt to achieve similar accuracy levels for both the forecasting and estimation steps.

Secondly, I will make methodological contributions to the field by merging heterogenous data sources and pushing the limits of spatiotemporal resolution. My usage of spatially continuous data from the High-Resolution Rapid Refresh (HRRR) system will be the first application of it to forecasting ozone. The HRRR model is a numerical weather model produced by NOAA over the continental United States at $3km^2$ spatial resolution. In most of the forecasting literature (Krishan 2019, Marvin 2022, Freeman 2018), data from individual weather stations that are collocated with the air quality monitors are used. While this makes the forecasts more accurate and the data pre-processing more convenient for studies that only forecast for one location, this method is not viable for forecasting the air quality of an entire area. None of the weather stations and air quality monitors in Denver are collocated, so using station data requires making assumptions that the conditions at a particular weather station apply to a very large geographic area. This assumption does not hold true, especially in areas with extremely variable topography like Colorado. Conversely, data from HRRR is spatially continuous across the entire continental US and is based on decades of atmospheric science research at NOAA.

Using HRRR data will improve my model by allowing me to forecast ozone at locations without weather stations. By doing so, the spatial resolution of my interpolations will be greatly improved compared to research which uses numerical models to make spatially continuous forecasts (Zhou 2017, Jiang & Yoo 2018, Wang, Liu 2022). The only other usage of HRRR I have found is from Wang 2022 who achieved highly accurate estimations using this data, but did not use the data for forecasting. I will use the insights from this paper to guide my integration of HRRR data.

Lastly, although I will not use real-time forecasted data, the HRRR model does produce forecasts in real time for all the variables I will use. The pollutant data from the AQS system is also available in real-time, therefore (unlike most research which uses data sources that are not released in real time) my model could theoretically be operationalized for real-time forecasting. This is one of the proposal's strongest contributions because the model created will represent a realistic approach to AI driven air quality forecasting that does not depend on the generation of new infrastructure such as real time satellite data feeds or increased monitoring sites.

3. Study Area and Data Sources

3.1 Study Area

The Denver Metro Area has exceeded safe levels of ozone pollution since at least the 1980s when the EPA began systematic monitoring. A range of factors, both anthropogenic and not, contribute to Colorado's susceptibility to dangerous Ozone levels. To start, ozone concentrations generally increase alongside elevation because of the sunlight intensity (Lu 2017). These two photochemical facts make it no surprise that the "mile-high city" in the state famously known for 300 days of sunshine per year has an ozone problem. Semi-anthropogenic sources of ozone include biomass burning from agriculture and wildfires, which are more common in the high desert of CO and produce ozone precursor emissions (Crooks 2022). Additionally, a substantial body of research shows that oil and natural gas drilling release Ozone precursors such as methane and NO_x (Cheadle 2017). Colorado ranks 5th as a state for number of oil and natural gas wells. A high concentration of these are on the front range in the vicinity of Greeley (Cheadle 2017). Lastly, the Front Range is the mountain west region's largest and densest population center (alongside Salt Lake City, which also has a substantial GLO problem) and consequently has more anthropogenic sources of ozone such as car exhaust and fossil-fuel-burning energy production than most of the region.

Currently, the Colorado Department of Public Health issues air quality forecasts and nowcasts for four major pollutants, including ozone (CDPHE). This data is then integrated into the EPA's AirNow system and made available for third party weather companies through the EPA's Air Quality System (AQS) API. While these forecasts are relatively accurate, their spatial and temporal resolution is too coarse to allow accurate assessments at the local scale. For example, the forecasts on the EPA's AirNow application are contours maps that utilize a simple inverse distance weighted interpolation from highly dispersed monitoring stations. Data such as topography and weather conditions for locations without a monitoring site are not considered in these forecasts. This can leave areas as disparate as the peaks in Rocky Mountain National Park, the I-25 corridor, and everything in between with the same rough estimate of ozone pollution. Temporally, these forecasts are also generalized for the 8-hour maximum value, but don't reveal what hour the peak is projected to occur (AirNow).

My study area encapsulates all the EPA air pollutant monitoring stations in the Denver Metro Area. Although there are a few stations to the north in Fort Collins and to the south in Colorado Springs, these areas are typically considered separate metropolitan areas and are not used in my study. Unfortunately, the EPA does not monitor all pollutants at all locations. NO_2 and $PM_{2.5}$ monitors are concentrated in downtown Denver, while ozone monitors have a more consistent and wider distribution. and A map of the study area and EPA monitors is shown in Figure 3:

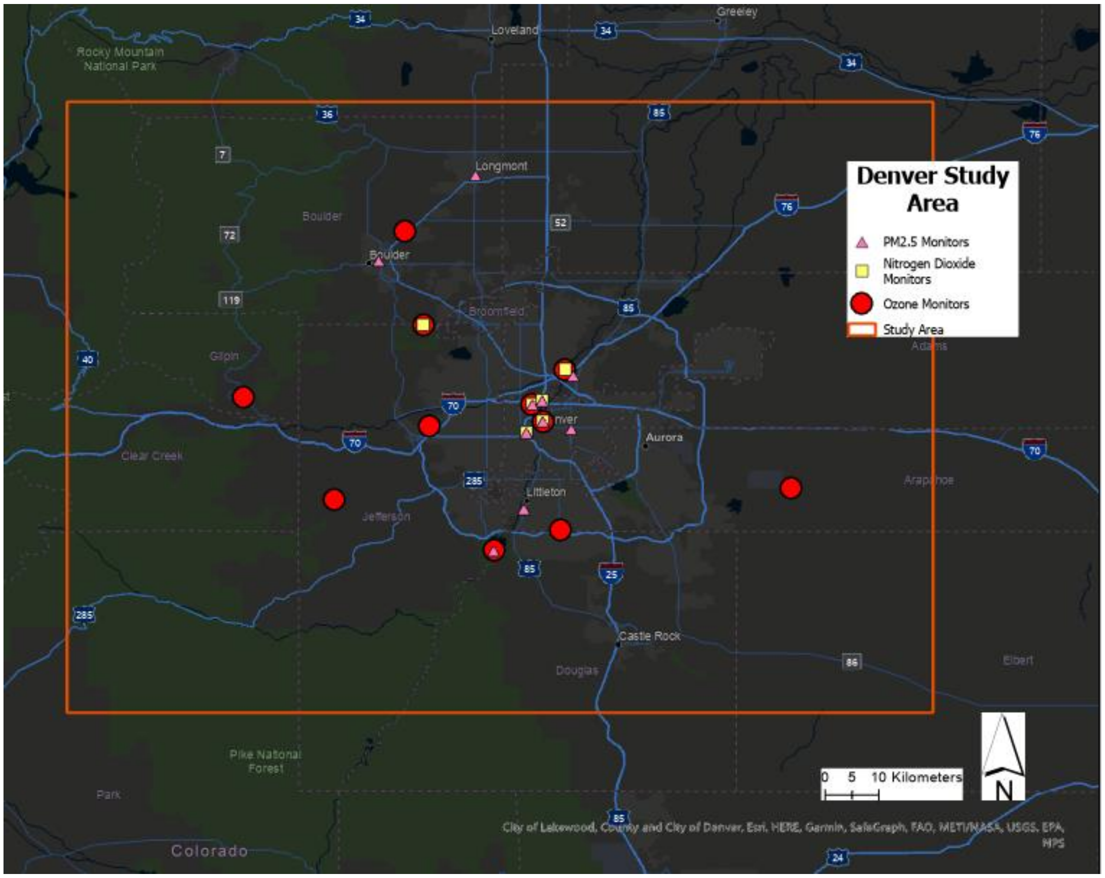


Figure 3: Map of the Study Area

Lastly, using Denver, CO as my study area addresses the literature gap for research using US data sources. As I stated earlier, the vast majority of AI air quality forecasting research takes place in Asia and Europe (Masood 2021). While studies using Asian and European data are theoretically guiding, the future of ozone forecasting and estimation in the US will ultimately rely on US data.

The study area bounding box shown in figure 3 will be used for both the input and output data rasters. Its coordinates are: minimum latitude: 39.2598°, minimum longitude: -105.6330°, maximum latitude: 40.1730°, maximum longitude: -104.2377°. The time period of the analysis will be the summer months of 2021.

3.2 Data Sources

3.2.1 Pollutant data

Pollutant data will be acquired from the EPA AQS API using the requests module in Python 3.9. Data will be downloaded at the hourly level, which is the finest temporal resolution the EPA releases. There are 11 ozone monitors, 9 particulate matter 2.5 monitors, and 6 nitrogen dioxide monitors in the study area.

3.2.2 Meteorological Data

The High Resolution Rapid Refresh model is a real-time, 3-km resolution, hourly updated, cloud-resolving, convection-allowing atmospheric model, initialized by 3km grids with 3km radar assimilation (NOAA). To start, I plan on using the most important predictors from the results of Wang 2022 including maximum downward velocity, maximum upward velocity, surface temperature, relative humidity, planetary boundary layer height (elevation), wind speed, and incoming short-wave radiation. There are over 200 variables produced by the HRRR, so the ones used in my final model will likely change depending on the results of my analysis.

To download this data, I will use the Herbie module in Python 3.9. The Herbie module allows users to filter historic HRRR data and manipulate it in xarray, numpy, or pandas data structures before downloading the extremely large GRIB2 datasets. I will use this functionality to spatially and temporally clip my datasets before downloading them locally.

3.2.3 Static Data

Static data such as population density, elevation, and land cover will also be used to assess geographic similarity between forecast points. Table 1 describes all of my data sources.

Data	Source	Spatial/ Temporal Resolution, Data Availability	Variables
Pollutant Concentrations	EPA AQS API	By station, 1 hour, Real Time	• O_3 • NO_2 • $PM_{2.5}$
Meteorological Data	NOAA HRRR	3 km^2, 1 hour, Real Time	• max downward velocity • max upward velocity • surface temperature • relative humidity • planetary boundary layer height • wind speed • incoming short-wave radiation
Land Cover	NLCD	30 meters, resampled to 1 km^2, Static	• Land Cover
Elevation	USGS DEMs	30 meters, resampled to 1 km^2, Static	• Elevation
Population	US Census	Census Tracts, resampled to 1 km^2, Static	• Population Density

Table 1: Data Sources

4 Methods

In the following section I present the proposed methods of this study. Unless otherwise stated, the primary software I will use to accomplish these methods will be open-source packages in Python and QGIS. The workflow is graphically described in figure 4 below.

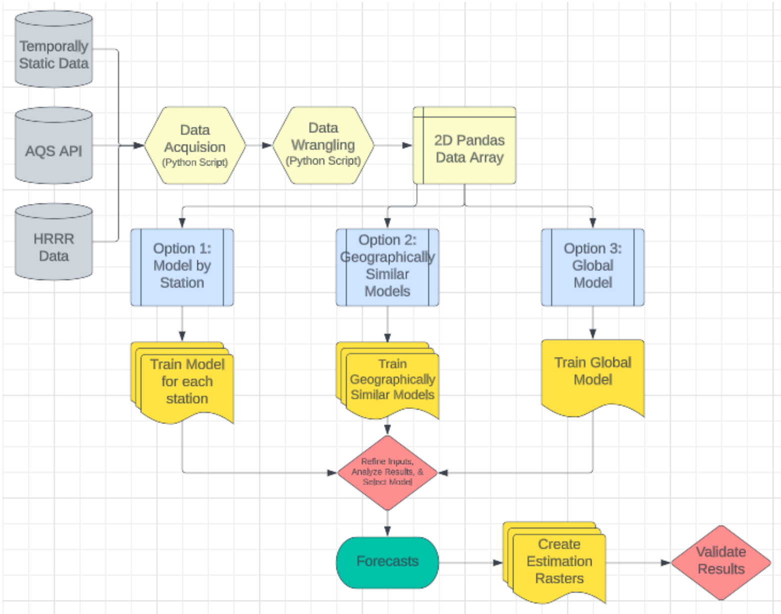


Figure 4: Flowchart

4.1 Data Preprocessing

In order to train an LSTM model with multiple predictors, all the data must first be aligned in 2-dimensional arrays. HRRR meteorological data in GRIB2 format must be extracted and aligned with pollutant data at all 11 ozone monitoring sites. This means that from each GRIB2 file representing one hour of weather conditions in Denver, data from 11 pixels should be extracted and assigned to the correct time and location in a large 2D array. This will be done for each weather variable. Next, data for NO_2 and $PM_{2.5}$ must be aligned with the ozone data. A nearest neighbor approach will be used to assign $PM_{2.5}$ and NO_2 values for ozone locations without a collocated $PM_{2.5}$ or NO_2 monitor. This result of this preprocessing is a large 2D array. A subsect of this result is presented below in Figure 5. Note that this is an example which is missing many columns.

Datetime	Latitude	Longitude	Station ID	Ozone	Temperature	NO_2
01-01-2021-0100	39.7756	-104.98	1	0.45	28	0.21
01-01-2021-0200	39.7756	-104.98	1	0.46	29	0.24
01-01-2021-0300	39.7756	-104.98	1	0.48	32	0.26
01-01-2021-0400	39.7756	-104.98	1	0.56	32	0.40

01-01-2021-0500	39.8456	-104.50	2	0.60	33	0.45
01-01-2021-0600	39.8456	-104.50	2	0.61	36	0.45

Table 2

The data in this array will be fed to an LSTM model using Keras, a deep learning API written in Python. Keras is used extensively in research that leverages AI and thus has a large user base and extensive documentation.

4.2 Forecasting and Estimation Methods

In this stage of the research, many options for how to train the LSTM model, interpolate the values, and validate the results will be explored. Decisions need to be made such as:

- Should a model be made for each site or should one model be made for the region?
- What forecast horizon should be used to balance usability and accuracy?
- Should I use point forecasting or multi-step forecasting?
- What interpolation technique should be used?
- How should I validate the results?

It is impossible to know exactly which methods will be most effective at this stage, but I explore these questions in detail and outline a general plan of action below.

4.2.1 Geographically Similar Models vs Regional Model

Forecasting studies that use a single forecasting location typically achieve higher R^2 and low RMSE values (Du et al 2022, Marvin 2022). Training the model on one location allows it to learn the characteristics of a single point in space. Although the model is not generalized for the entire study area, it is extremely accurate for that location. Because all things or phenomena are connected in geography (Goodchild 2009), including atmospheric components such as pollutant concentration, we can assume some level of spatial autocorrelation between model accuracy and distance from the location it was trained on. Additionally, no matter how robust a model is, there will always be unexplained variance due to the unique geographic characteristics of each point in space. While a perfect model that has high accuracy for all points in space is ideal, it is unrealistic.

For these reasons, I propose that a separate LSTM model be trained for each ozone monitor in the study area. Each forecasting site will be trained only on the pollutant and meteorological data for that site in order to maximize the amount of variance it can explain. The models created will be unique to each of the eleven monitoring sites in the study area and may weight inputs differently. For example, sites in downtown Denver may use NO_2 as the strongest predictor, while sites in the foothills may regard incoming shortwave radiation as the strongest predictor.

While forecasting by site should theoretically produce more accurate results, it also reduces the amount of data points that can be trained on for each model, and therefore may have the opposite effect. To mitigate this, another option is to develop a weighted criteria model to rank pixels on their geographic similarity using data such as land cover class, population density, and elevation. Each $3km^2$ pixel will be assigned a value between 0 and 1 based on these criteria. Using these values, pixels will be assigned into of 3 groups: mountainous, plains, and urban. Then, 3 separate models will be trained using data from monitoring sites in the respective regions.

These options increase the complexity and computation resources of the data pre-processing by requiring more than one LSTM models be trained. For this reason, it may actually be better to train one LSTM model for all of the data in the study area and time period. This option will be explored in addition to creating a model for each station or each region.

4.2.2 Forecast Horizons and Multi-step vs Point Forecasting

In the field of time series forecasting, a forecast horizon is the number of time steps into the future we want to forecast at any point in time (Joseph 2022). A 6 hour forecast horizon predicts 6 hours in advance, while a 12 hour forecast predicts 12 hours in advance. In between this horizon there are likely many other timesteps which we may or may not forecast for and use as inputs. Whether or not the model calculates these unknown values and uses them as inputs depends on if a recursive multi-step or point strategy is used.

A point strategy model is trained to predict for the desired forecast horizon without predicting the timesteps in between. In the example below in table 2, the point strategy model takes inputs t_1, t_2, t_3… t_6 to predict the ozone value at forecast horizon t_9 without predicting t_7 or t_8.

Timestep	t_1	t_2	t_3	t_4	t_5	t_6	t_9
Hour	0000	0100	0200	0300	0400	0500	0800
Ozone	0.62	0.65	0.69	0.75	0.75	0.77	0.80*
Temp	45	46	47	50	52	55	60

Table 3: Point forecasting * = target forecast, forecast horizon = 3 hours, window size = 6 hours[3]

In contrast, a multi-step recursive model predicts each timestep in between the inputs and the forecast horizon. It starts by predicting one timestep ahead and then uses this predicted value as input to forecast the next hour, as in the example below (table 4). In iteration I_1, the model

[3] The window size of the model refers to how many timesteps are used as inputs. In tables 3 and 4 below, a window size of 6 indicates that 6 hours of known data points are used as inputs to calculate the value for the forecast horizon. The window size of the model will affect both the accuracy and computation time. My model will use a window size that balances feasibility and accuracy.

uses 6 hours of ozone values ($t_1 \ldots t_6$) to predict the very next hour, t_7. Then in iteration I_2 it uses the predicted value t_7 as an input to predict t_8. In this example, it does this for five iterations until the target forecast value is calculated.

Iteration	I_1	I_2	I_3	I_4	I_5
Input Timesteps	$t_1 \ldots t_6$	$t_2 \ldots t_7$	$t_3 \ldots t_8$	$t_4 \ldots t_9$	$t_5 \ldots t_{10}$
Predicted Timestep	t_7	t_8	t_9	t_{10}	t_{11}*

Table 4: Multi-step forecasting * =target forecast, forecast horizon = 5 hours, window size = 6 hours

The major benefit to using a point strategy instead of a multi-step strategy is the simplicity of the model and decreased computation (Joseph 2022). In the example above, the recursive multi-step strategy uses approximately five times more processing power because it takes five iterations to reach the forecast horizon. The major benefit to multi-step forecasting is increased accuracy, especially as we increase the forecast horizon. For example, using a point strategy with a forecast horizon of 4 hours means that the model can only use data points from over 4 hours ago as input, which could lead to extremely inaccurate results. A multi-step model on the other hand will use the all the values in the table below as inputs to predict the forecast horizon.

Timestep	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
Hour	0000	0100	0200	0300	0400	0500	0600	0700	0800	0900
Ozone	0.62	0.65	0.69	0.75	0.75	0.77	0.80^+	0.81^+	0.83^+	unk*
Temp	45	46	47	50	52	55	$60^\dagger$	$60^\dagger$	$61^\dagger$	$65^\dagger$

Table 6: Multi-step forecasting inputs † = forecasted values from HRRR, * = target value, + = forecasted values from multi-step iterations, window size = 6 hours

In addition to forecasting strategy, the size of the forecast horizon is an important considerations. Longer forecast horizons allow for more time to react to upcoming conditions and thus have obvious benefits for stakeholders and decision makers, but have obvious drawbacks for accuracy. In a recursive multi-step model, long forecast horizons also require increased computation. My research will attempt to use a forecast horizons between 2 and 12 hours, as forecast horizons less than 2 hours are not very useful because it's not enough time for decision makers to react to upcoming conditions. Forecast horizons greater than 12 hours are likely to suffer large accuracy losses and may have unacceptable computation time.

Ultimately, optimizing the model will require extensively testing and evaluating the effects of different forecasting strategies, forecast horizons, and window sizes. Fortunately, by using the built in LSTM framework in the keras python library, changing these variables will not be particulary time intensive. Testing and evaluating them may be computationally expensive if done on the entire dataset, so I will subset my data to perform preliminary testing.

4.2.3 Estimation

Once I have optimized my models, the next step is to use the models to forecast ozone values at multiple locations across the study area and interpolate them to create continuous surfaces.

In section 4.2.1 I described my plan to train forecasting models for each ozone monitoring station as opposed to a regional model. If the accuracy of this method proves to be higher, then I propose a similar approach be applied to the estimation step of this research. One of the main benefits to using HRRR data as opposed to discrete weather stations is that I will have continuous grids of all my meteorological variables. Because of this, I can apply the LSTM forecasting models to areas between pollutant monitoring sites provided I make some assumptions about the pollutant concentrations in those areas.

I first calculate how many discrete forecast points I would need to interpolate a surface at approximately 1km^2 spatial resolution. Equation 1 from Hengl 2006 describes a commonly used method to determine an appropriate pixel size given a number of point observations and study area.

$$p = 0.0791 \cdot \sqrt{\frac{A}{N}}$$

Equation 1: where… A is the area of the study area, N is the total number of observations, and p is the pixel size

My study area is approximately 14200 km^2 and my target pixel size is $1km^2$, meaning I need at least 89 observations. To acquire this many observations, I will use a fishnet point grid with points spaced 10 km apart to create 142 observations. Then, I will forecast ozone at each point using the nearest LSTM model (as shown by the Thiessen polygons in Figure 5), the HRRR data from the underlying pixel, and pollutant data from the nearest monitoring site. Figure 5 shows a map of my study area with Thiessen polygons created from ozone monitoring sites, the 10km fishnet grid of sample points, and the $3km^2$ grid representing HRRR weather data.

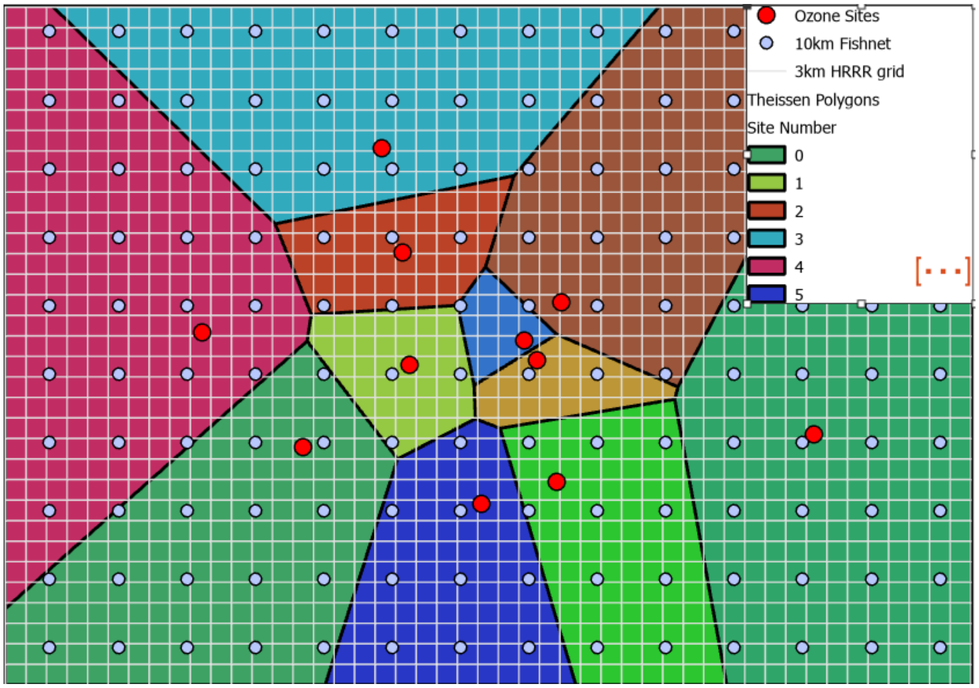


Figure 5: Estimation Strategy

Each blue dot on figure 5 will be a point at which an ozone forecast is calculated. The resulting grid will be interpolated using a simple Kriging method at 1km^2 resolution.

In the case that my forecasts are more accurate using the geographic similarity approach I described in section 4.2.1, a second option for estimation will be explored. Instead of using Thiessen polygons to determine which LSTM models to use at unknown locations, I will use the same weighted criteria model approach I used to train the models to choose which model to use at each fishnet point.

These methods have several advantages over a simple interpolation from my ozone monitoring sites: the model preserves as much spatial autocorrelation as possible, the evenly spaced points will create a smooth interpolation with higher accuracy at the edges, and I take full advantage of the HRRR continuous weather data rather than only using data that intersect the ozone monitoring stations.

4.2.4 Validation

The results of both the forecasting and estimation methods will need to be validated for accuracy. The most common metrics to assess accuracy in forecasting research are R^2 and root-mean-square-error (RMSE) (Masood 2021). I will use the same metrics so that my research can easily be compared to past and future work done in the field.

To validate my forecast models, I will use the most common strategy in machine learning: k-fold cross-validation (Joseph 2022). This method randomly splits the data into training and testing samples k number of times and then tests the resulting models for accuracy. This leads to k number of RMSE and R^2 measurements which can be averaged to reach an error estimate that is more accurate than performing only iteration of training and testing.

To validate my estimation method, I will use leave one out cross validation (ESRI). This method is similar to k-fold cross validation in that it iteratively separates the data into training and testing sets. In each iteration it will remove one of the eleven monitoring station and test the accuracy of the interpolation using that known point.

Traditional interpolation methods such as Inverse Distance Weighting, Kriging, and Global Polynomial Interpolation are available in ArcGIS Pro through the Geostatistical Wizard toolset. These tools are easy to implement, require less computation, and automatically produce accuracy assessments. I will compare the results of my estimation techniques with these tools to see if I achieved higher accuracy.

4.3 Deliverables

Upon completion of this research, I will have the following deliverables. The data, code, and documentation will be stored in a publicly available GitHub repository to facilitate the reproduction of my methods.

1. All input datasets.
2. Source code for downloading and preprocessing the data, creating the forecast models, interpolating the results.
3. Documentation on using the data and source code.
4. A dataset of ozone pollution forecasts in the Denver study area for at least the summer of 2021.
5. A completed master's thesis writeup.
6. A presentation for the Association of American Geographer's conference and other conferences.

5 Contribution to Current Research

The unfortunate reality of climate change means that ozone pollution on the Colorado front range will likely get worse before it gets better (Crooks 2022). While efforts by Colorado Public Health and the EPA's Air Now program have improved the public's ability to make informed health decisions regarding air pollution, more can and should be done.

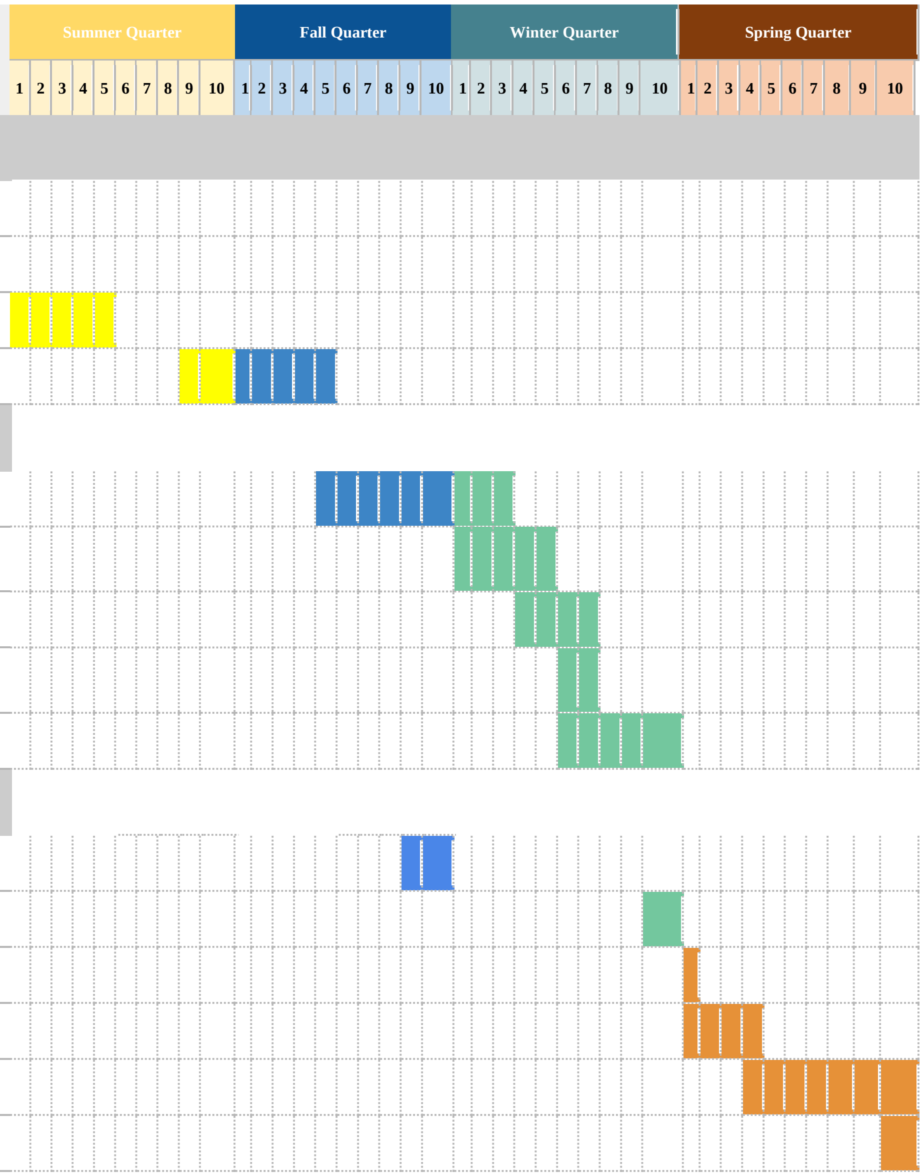
This research will contribute to a project currently being worked on by the National Jewish Health Center, Colorado Public Health, and the Department of Geography at the University of Denver. This project attempts create high resolution air quality assessments for use in a personalized air pollution exposure tool. The proposed tool will allow users to input a spatiotemporal vector path and calculate their exposure to the three criteria air pollutants: particulate matter, nitrogen dioxide, and ozone. Figure 6 below demonstrates how a high resolution $PM_{2.5}$ map can be used to quantify exposure from a short walk through downtown Denver. My research contributes to this body of work by creating maps and techniques that could be used to assess ozone exposure in real time.

Figure 6: Air Pollution Exposure Tool (Li, Jing 2022)

6 Project Timeline

The Gantt chart below shows the timeline for the completion of the project.

WBS	TASK TITLE	START DATE	DUE DATE	DAYS
1	**Data Collection and Preprocessing**			
1.1	Proposal Defense	5/25/23	5/25/23	0
1.1.1	Proposal Revisions	5/25/23	6/6/23	11
1.2	Data Collection	6/6/23	7/15/23	39
1.3	Data Preprocessing	8/8/23	10/1/23	53
2	**Modelling and Analysis**			
2.1	Forecast Modelling	10/1/23	1/29/24	118
2.1.1	Forecast Model Refinement	1/8/24	2/8/24	30
2.2	Estimation	1/15/24	3/1/24	46
2.2.1	Estimation Methods Refinement	2/15/24	3/1/24	16
2.3	Analysis of Results	3/1/24	4/1/24	30
3	**Writing and Presenting**			
3.1	Write AAG Abstract	11/1/23	11/21/23	20
3.1.1	Present at AAG	3/23/24	3/31/24	8
3.2	Write Intro	4/1/24	4/7/24	6
3.3	Writing Methods	4/4/24	4/30/24	26
3.4	Writing Results	4/25/24	5/31/24	36
3.5	Thesis Defense	5/31/24	5/31/24	0



References

AirNow.gov, U.S. EPA. (n.d.). About airnow. About AirNow | AirNow.gov. Retrieved May 3, 2023, from https://www.airnow.gov/about-airnow/

Amato, Laib, M., Guignard, F., & Kanevski, M. (2020). Analysis of air pollution time series using complexity-invariant distance and information measures. Physica A, 547, 124391–. https://doi.org/10.1016/j.physa.2020.124391

Brasch, S. (2022, April 13). *The EPA moves to declare the front range a 'severe' air quality violator. here's why that matters.* Colorado Public Radio. Retrieved March 3, 2023, from https://www.cpr.org/2022/04/12/front-range-air-quality-ozone-violations-epa/#:~:text=A%202021%20study%20by%20National,worse%20as%20global%20warming%20intensifies

Cheadle, Oltmans, S. J., Pétron, G., Schnell, R. C., Mattson, E. J., Herndon, S. C., Thompson, A. M., Blake, D. R., & McClure-Begley, A. (2017). Surface Ozone in the Colorado Northern Front Range and the Influence of Oil and Gas Development During FRAPPE/DISCOVER-AQ in Summer 2014. *Elementa (Washington, D.C.)*, *5*. https://doi.org/10.1525/elementa.254

Chen, Zhou, L., Chen, X., Bi, J., & Kinney, P. L. (2017). Acute effect of ozone exposure on daily mortality in seven cities of Jiangsu Province, China: No clear evidence for threshold. *Environmental Research*, *155*, 235–241. https://doi.org/10.1016/j.envres.2017.02.009

Cohen, Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., … Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet (British Edition)*, *389*(10082), 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6

Colorado Air Quality Summary. CDPHE. (n.d.). Retrieved May 3, 2023, from https://www.colorado.gov/airquality/colorado_summary.aspx

Crooks, J.L., Licker, R., Hollis, A.L. *et al.* (2022). The ozone climate penalty, NAAQS attainment, and health equity along the Colorado Front Range. *J Expo Sci Environ Epidemiol* **32**, 545–553 https://doi.org/10.1038/s41370-021-00375-9

Du, Qiao, F., Lu, P., & Yu, L. (2022). Forecasting ground-level ozone concentration levels using machine learning. Resources, Conservation and Recycling, 184, 106380–. https://doi.org/10.1016/j.resconrec.2022.106380

Environmental Protection Agency. (n.d.). Ozone Trends. EPA. Retrieved March 3, 2023, from https://www.epa.gov/air-trends/ozone-trends

Freeman, Taylor, G., Gharabaghi, B., & Thé, J. (2018). Forecasting air quality time series using deep learning. Journal of the Air & Waste Management Association (1995), 68(8), 866–886. https://doi.org/10.1080/10962247.2018.1459956

Glavas, & Sazakli, E. (2011). Ozone long-range transport in the Balkans. Atmospheric Environment (1994), 45(8), 1615–1626. https://doi.org/10.1016/j.atmosenv.2010.11.030

Goodchild, M. F. (2009). International Encyclopedia of Human Geography: First Law o f Geography, 179–182.

Hengl. (2006). Finding the right pixel size. Computers & Geosciences, 32(9), 1283–1298. https://doi.org/10.1016/j.cageo.2005.11.008

High-Resolution Rapid Refresh (HRRR). NOAA Global Systems Laboratory. (n.d.). Retrieved May 3, 2023, from https://rapidrefresh.noaa.gov/hrrr/

History of ozone in Colorado. Department of Public Health & Environment. (n.d.). Retrieved March 3, 2023, from https://cdphe.colorado.gov/history-of-ozone-in-colorado#:~:text=Under%20the%20standard%2C%20the%20Denver,designated%20as%20nonattainment%20in%201978

Jiajia Chen, Huanfeng Shen, Xinghua Li, Tongwen Li, & Ying Wei. (2022). Ground-level ozone estimation based on geo-intelligent machine learning by fusing in-situ observations, remote sensing data, and model simulation data. International Journal of Applied Earth Observation and Geoinformation, 112, 102955–. https://doi.org/10.1016/j.jag.2022.102955

Jiang, & Yoo, E. (2018). The importance of spatial resolutions of Community Multiscale Air Quality (CMAQ) models on health impact assessment. The Science of the Total Environment, 627, 1528–1543. https://doi.org/10.1016/j.scitotenv.2018.01.228

Joseph. (2022). Modern time series forecasting with Python : explore industry-ready time series forecasting using modern machine learning and deep learning / (1st ed.). Packt Publishing, Ltd.

Kim, Kim, D.-K., Noh, J., & Kim, M. (2018). Stable Forecasting of Environmental Time Series via Long Short Term Memory Recurrent Neural Network. IEEE Access, 6, 75216–75228. https://doi.org/10.1109/ACCESS.2018.2884827

Ko, Cho, S., & Rao, R. R. (2022). Machine-Learning-Based Near-Surface Ozone Forecasting Model with Planetary Boundary Layer Information. Sensors (Basel, Switzerland), 22(20), 7864–. https://doi.org/10.3390/s22207864

Krishan, Jha, S., Das, J., Singh, A., Goyal, M. K., & Sekar, C. (2019). Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India. Air Quality, Atmosphere and Health, 12(8), 899–908. https://doi.org/10.1007/s11869-019-00696-7

Li, Jing. (2022, October 24). *Using data-driven, AI techniques to model exposure risks to ambient air pollution at localized scales in the Denver-metro area* [Conference presentation]. DU/NJH Research Luncheon. Denver, CO.

Li, Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental Pollution (1987), 231(Pt 1), 997–1004. https://doi.org/10.1016/j.envpol.2017.08.114

Lu, Ai, T., Zhang, X., & He, Y. (2017). An Interactive Web Mapping Visualization of Urban Air Quality Monitoring Data of China. Atmosphere, 8(8), 148–. https://doi.org/10.3390/atmos8080148

Marvin, Nespoli, L., Strepparava, D., & Medici, V. (2022). A data-driven approach to forecasting ground-level ozone concentration. International Journal of Forecasting, 38(3), 970–987. https://doi.org/10.1016/j.ijforecast.2021.07.008

Masood, & Ahmad, K. (2021). A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: Fundamentals, application and performance. Journal of Cleaner Production, 322, 129072–. https://doi.org/10.1016/j.jclepro.2021.129072

Pepin, Arnone, E., Gobiet, A., Haslinger, K., Kotlarski, S., Notarnicola, C., Palazzi, E., Seibert, P., Serafin, S., Schöner, W., Terzago, S., Thornton, J. M., Vuille, M., & Adler, C. (2022). Climate Changes and Their Elevational Patterns in the Mountains of the World. *Reviews of Geophysics (1985)*, *60*(1). https://doi.org/10.1029/2020RG000730

Ren, Mi, Z., & Georgopoulos, P. G. (2020). Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. Environment International, 142, 105827–. https://doi.org/10.1016/j.envint.2020.105827

Using cross validation to assess interpolation results. (n.d.). ESRI ArcGIS Pro Help. Retrieved May 1, 2023, from https://pro.arcgis.com/en/pro-app/3.0/help/analysis/geostatistical-analyst/performing-cross-validation-and-validation.htm

Tan, Zhang, Y., Ma, W., Yu, Q., Wang, Q., Fu, Q., Zhou, B., Chen, J., & Chen, L. (2017). Evaluation and potential improvements of WRF/CMAQ in simulating multi-levels air pollution in megacity Shanghai, China. Stochastic Environmental Research and Risk Assessment, 31(10), 2513–2526. https://doi.org/10.1007/s00477-016-1342-3

Wang, 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, 106917–. https://doi.org/10.1016/j.envint.2021.106917

Wang, Huo, Y., Mu, X., Jiang, P., Xun, S., He, B., Wu, W., Liu, L., & Wang, Y. (2022). A High-Performance Convolutional Neural Network for Ground-Level Ozone Estimation in Eastern China. Remote Sensing (Basel, Switzerland), 14(7), 1640–. https://doi.org/10.3390/rs14071640

Zhang, Wei, Y., & Fang, Z. (2019). Ozone Pollution: A Major Health Hazard Worldwide. Frontiers in Immunology, 10, 2518–2518. https://doi.org/10.3389/fimmu.2019.02518

Zhou, Xu, J., Xie, Y., Chang, L., Gao, W., Gu, Y., & Zhou, J. (2017). Numerical air quality forecasting over eastern China: An operational application of WRF-Chem. Atmospheric Environment (1994), 153, 94–108. https://doi.org/10.1016/j.atmosenv.2017.01.020