Colorado State University

Honors Thesis

Zephyr: An Interactive Tool for Air Quality Research

Brenna Wolf

April 2025

Mentor:
Professor Shrideep Pallickara

Committee Members:

Nicole Vieira
Sarah Zwick-Tapley

1 Abstract

Air quality presents one of the biggest environmental risks to human health and has the potential to impact every organ. In fact, it is second only to communicable diseases in the risk that it poses to public health. Air pollution contributes to reduced lung function, oxidative stress, and immunosuppression among other adverse health risks. Those suffering from chronic and acute respiratory diseases are especially at risk. Poor air quality contributes to more than 100,000 premature deaths in the United States each year (NOAA, 2021), and 8.34 million excess deaths internationally each year (Lelieveld, 2023).

Accessible and continuous access to air quality is the first step to educating citizens and, eventually, legislative changes. The EPA has a large number of outdoor air quality monitoring sensors that harvest data regarding a range of airborne pollutants including particulate matter (PM2.5 and PM10), ozone, carbon monoxide, nitrogen dioxide, and sulfur dioxide. We have designed a browser-based tool, Zephyr, that allows users to interactively analyze longitudinal (starting in 1980) air quality data tracking over 1400 different pollutants. An added feature of Zephyr is integration with the 2020 US census data to explore how socioeconomic factors correlate with exposure to poor air quality. Zephyr also overlays infrastructure data relating to coal and gas fueled power plants, allowing users to assess how proximity to such power plants impacts air quality. We have also incorporated support for animating pollutant-specific variations so that users may assess air quality changes over time during transient, but often prolonged, events such as wildfires. Finally, the tool provides 7-day air quality forecasts using SARIMA models that provide insights into time-series trends. These models are fine-tuned for each metropolitan region to boost accuracy; this, in turn, would allow vulnerable populations to prepare for poor air quality days.

The tool is available for experimentation by anyone at the following link: Zephyr

2 Impact on Personal and Educational Career

Working in the Urban Sustain Research lab under Professor Shrideep Pallickara has impacted my educational career in multiple ways. First and foremost, this lab has provided me with the opportunity to work with very talented professors and students. Professor Shrideep Pallickara has significantly shaped both my college journey and perspective on the workforce. As a well recognized Professor at CSU, known for both his research and his undergraduate teaching, he has helped improve my critical thinking and problem solving skills, allowing me to approach challenges not as obstacles, but as opportunities for growth and learning. He has helped increase my confidence when starting new projects and always encouraged me to just get started and to get things moving. This in turn has led to a lot of trial and error which has helped further develop my programming skills as well as my software engineering skills. Additionally, this lab has introduced me to amazing students whom I was able to work with and discuss different approaches I was considering for a project as well as different problems I was running into. They have highlighted the importance of surrounding yourself with people who will support and encourage you to succeed, especially in a work setting. In research this made all the difference and kept me motivated when I wasn't getting the results I wanted.

Aside from meeting and getting to work with the amazing professors and people in the lab, this work has provided insight into how computer science and web development can be beneficial outside of the tech community. I love the critical thinking and problem solving aspect of computer science, but my main concern with working in such a progressive field was being stuck in a position that, while intellectually stimulating, didn't feel meaningful to me. I want my work to have a tangible impact, whether that's improving lives, driving innovation, or contributing to something bigger than myself. Working in this lab showed me that it's possible to pursue the technical side of computer science while also building tools that can make a meaningful difference in the world.

Throughout this project, I have expanded my knowledge on web development, machine learning, time complexity, and how to better work with spatial data. The process has challenged me in multiple ways and pushed me to become a better computer scientist. One of the most valuable skills I've gained from working in this lab is the ability to improve my self-learning while developing good judgment on when to seek help versus when to persist in solving a problem on my own. Both Professor Pallickara and my lab mates were always available to help and answer questions, this in turn taught me how to know when to ask questions and how to ask the right questions. In order to do this, it required a lot of self reflection on my own code and approach to different tasks, helping me to see all the possible options before going to someone for assistance. This has allowed me to build more character and confidence in an educational setting and will be a very useful skill in any job I have in the future.

3 Thesis Project

3.1 Introduction

Zephyr is a research tool that visualizes the air quality index (AQI) across the United States, everyday, back to 1980. Through aggregation pipelines and the use of multiple data collections, the user is able to compare the AQI between different geometric shapes and time periods, identify the number of people affected through a demographics analysis, and have the option to look at the affects of individual pollutants. Since Zephyr is a free, browser-based tool, it enhances accessibility and ensures that this information is readily available to a broad audience. Because it runs in a web browser, users don't need to install special software, making it more convenient for different types of users, regardless of their technical expertise. Additionally, since it works on both desktops and tablets, it offers flexibility in how and where people can access the information. Knowing this, working to develop Zephyr into a well presented research tool became a priority in order to grab the attention of the general public. To do so, my thesis included two parts; scaling up the website and making it into a predictive tool.

3.2 Part 1: Scaling the Website to Include Additional Pollutants

The first part of my thesis involved scaling this website up from including only the 6 criteria pollutants (PM2.5, PM10, Ozone, CO, NO2, and SO2) to now having access to over 1400 different pollutants. The first step to complete this task was to implement an auto-ingestion script that continuously downloaded and ingested data into the database. To do this, I utilized the EPA Air Quality System (AQS) API (Environmental Protection Agency, 2020) which records the raw measurements from air quality monitors, usually in units Parts per Million and Parts per Billion. The AQS records new data reported from monitors and can take up to 6 months to be published within the API. The delay in data is on the reporting and review periods that agencies have when submitting sample data to AQS. While the data is guaranteed to be public after 6 months, it may happen more often than that as reporting to the AQS is to be performed on a quarterly basis. To account for this delay, my auto-ingestor runs every 2 months. Another characteristic of this API is the monitor frequency. This tells the user how often each monitor is set to record values, which could be daily, every other day, or weekly. Since this information was not accessible in the query I used, I implemented a data filler in my ingestion script in order to handle days with no data. This data filler logs each day that returns an empty data list. Two months later, it rechecks those dates for new data. If no data is still found and six months have passed since the original date, the entry is removed. Otherwise, it remains on the list to be checked again during the next script run.

To optimize this script, I used threading to improve the time complexity. In order to use 5 threads, 5 different AQS API keys were required to ensure that one API key doesn't get overloaded with requests. This made it difficult for me to track when the ingestion script failed due to an inactive key since one thread could die and not affect the runtime of the other threads. To get around this, I have the ingestion script email me whenever one API key errors out in order to notify me that a new key must be generated and the script be restarted. By multithreading this ingestion script, it made the runtime 5 times faster. Benchmarks for this script can be found in Table 1.

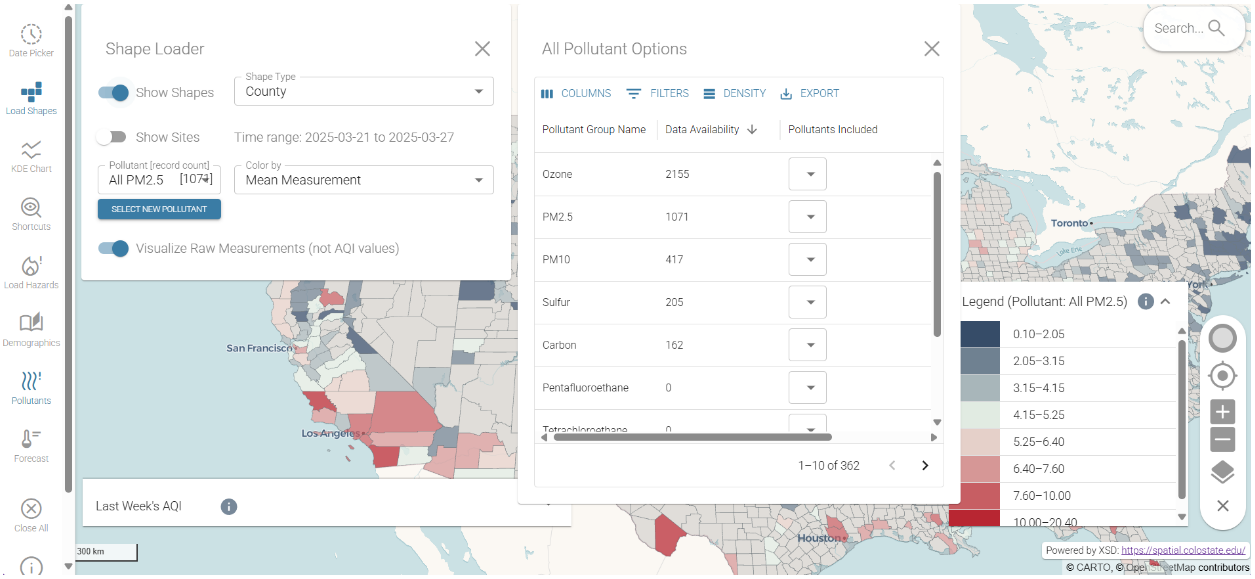
This auto-ingestion script was crucial to implement because it simplified the client side of the web application, making it faster and more interactive for the user. By continuously processing and integrating data on the server side, the client remains lightweight and efficient. More importantly,

Metric	Value
Total number of pollutants	1474
Total number of threads	5
Number of pollutants per thread	294
Time to finish 294 iterations	25 minutes
Time to finish 294 iterations 60 times (2 months)	25 hours

Table 1: AQS API Auto-Ingestion Script Benchmarks

the server can handle complexities such as rate limits, query restrictions, and other optimizations, ensuring smooth and reliable access to data without burdening the client. This concept is utilized in the second part of my thesis as well.

Finally, for the User Interface (UI) of this feature, I implemented a separate dashboard (Figure 1) that displays all pollutants with their raw measurements versus the home dashboard that displays the AQI values of the criteria pollutants. I utilized different Material UI tools, such as DataGrid, in order to have access to pagination which provides a clean and well organized, interactive table that is user friendly (MUI, 2024). In addition to the UI, I included the use of the ontology, ChEBI, in order to automate the process of grabbing a description of each of the 1474 pollutants. This provides the user with a short description that covers the chemical build and pollution effects of each pollutant (EMBL-EBI, 2018). Although this ontology doesn't return a description for every pollutant, it is a more efficient approach than having to manually write in a description for each pollutant. Furthermore, it allows the code to be kept easily scalable if more pollutants were to be added in the future.



Shape Loader

Show Shapes: on

Shape Type: County

Show Sites: off

Time range: 2025-03-21 to 2025-03-27

Pollutant [record count]: All PM2.5 [1071]

Color by: Mean Measurement

Visualize Raw Measurements (not AQI values): on

All Pollutant Options

Pollutant Group Name	Data Availability ↓	Pollutants Included
Ozone	2155	▾
PM2.5	1071	▾
PM10	417	▾
Sulfur	205	▾
Carbon	162	▾
Pentafluoroethane	0	▾
Tetrachloroethane	0	▾

1–10 of 362

Legend (Pollutant: All PM2.5)

■	0.10–2.05
■	2.05–3.15
■	3.15–4.15
■	4.15–5.25
■	5.25–6.40
■	6.40–7.60
■	7.60–10.00
■	10.00–20.40

Figure 1: Raw Measurements dashboard and Pagination table where the user can access all 1474 pollutants.

3.3 Part 2: Transforming the Website into a Predictive Tool

For the second part of my thesis, I implemented a forecasting feature to enhance Zephyr from a visualization tool to a predictive one. This addition was inspired by how people primarily use weather apps for future forecasts. By integrating this feature, Zephyr has become more useful and potentially more appealing to the general public. To do so, I utilized a SARIMA model, which stands for Seasonal Autoregressive Integrated Moving Average (statsmodels, 2025). This model is most generally used for making predictions given a time series dataset. By combining the concepts of an ARIMA model with a seasonal component, it efficiently captures both short and long term trends within the data.

This SARIMA model is well-suited for forecasting AQI values because autoregressive (AR) models account for dependencies on past values, such as this month's AQI being influenced by previous months. Additionally, the moving average (MA) component is appropriate since AQI levels are likely affected by past white noise error terms. Seasonal patterns (S) are evident as well, with higher pollution levels during the business week and increased emissions from power plants. Lastly, differencing (I) helps remove long-term trends, such as those caused by climate change or air pollution, ensuring a stationary time series for more reliable forecasting (Ozdogar, 2023).

The first step in using SARIMA was to determine the best set of parameters to use for all models. The SARIMA model utilizes 7 parameters: a seasonal and nonseasonal parameter for differencing (d, D), moving average (q, Q), and autoregression (p, P), and lastly a seasonality parameter (s). Due to the fact that data exists for 800 CBSA areas and 6 criteria pollutants, there could be a maximum of 4800 SARIMA models to train. Hence, determining a different parameter set for each model was unrealistic. To get around this, I chose three default parameter sets, trained all models using each set, then picked the parameters that resulted in the best metrics for each pollutant. Note that the seasonality parameter stayed the same in each parameter set since either way I would be predicting weekly values. The metrics for each parameter set can be seen in Figure 2, Figure 3, and Table 2. From these charts, I concluded that the parameter set containing all ones was by far the best option.

Pollutant	Parameter Set	MAE	MSE
PM2.5	(1,1,1,1,1,1,7)	0.069715815392215	0.010425834375712
PM10	(1,1,1,1,1,1,7)	0.071746136879228	0.012449941944686
Ozone	(1,1,1,1,1,1,7)	0.038246797270853	0.003450699580238
NO2	(1,1,1,1,1,1,7)	0.067956008860442	0.0111952081633003
SO2	(1,1,1,1,1,1,7)	0.033602551734987	0.0054593714181136
CO	(1,1,1,1,1,1,7)	0.0298470911709450	0.004015738832198

Table 2: Parameter Set Metrics Per Pollutant

Through the use of this model and the data available for each ¡pollutant, CBSA¿ pair, I was able to train and store 2521 models in order to make fine-tuned predictions for the next 7 days

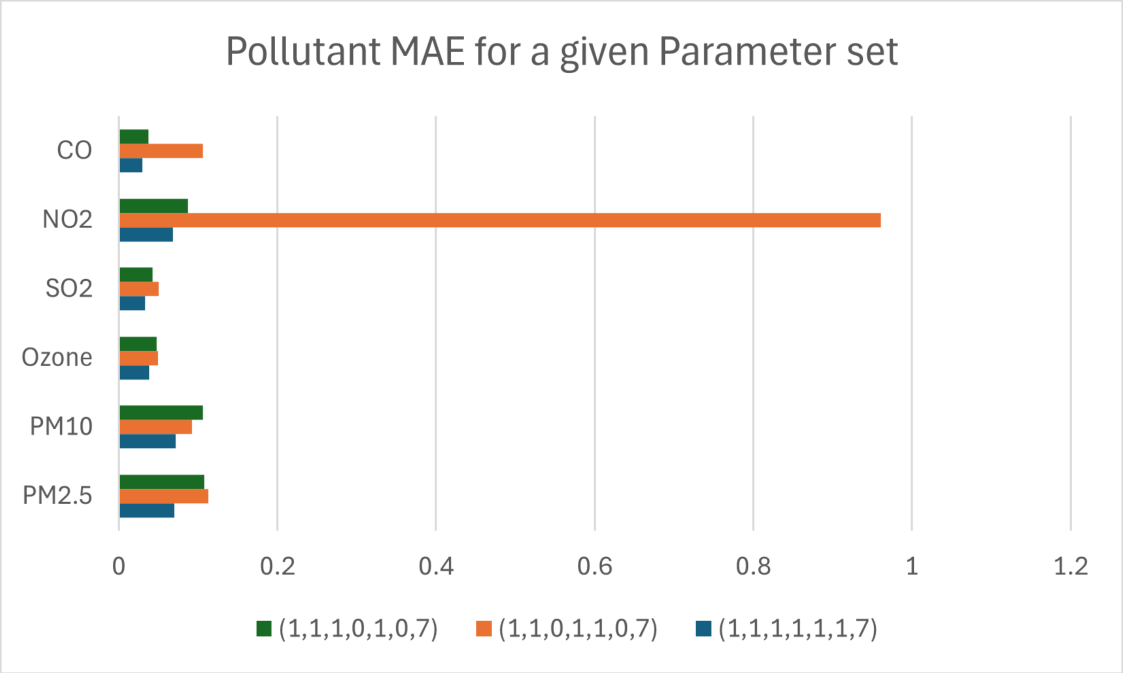


Figure 2: From this chart, the parameter set (1,1,1,1,1,1,7) results in the smallest MAE metric for all pollutants.

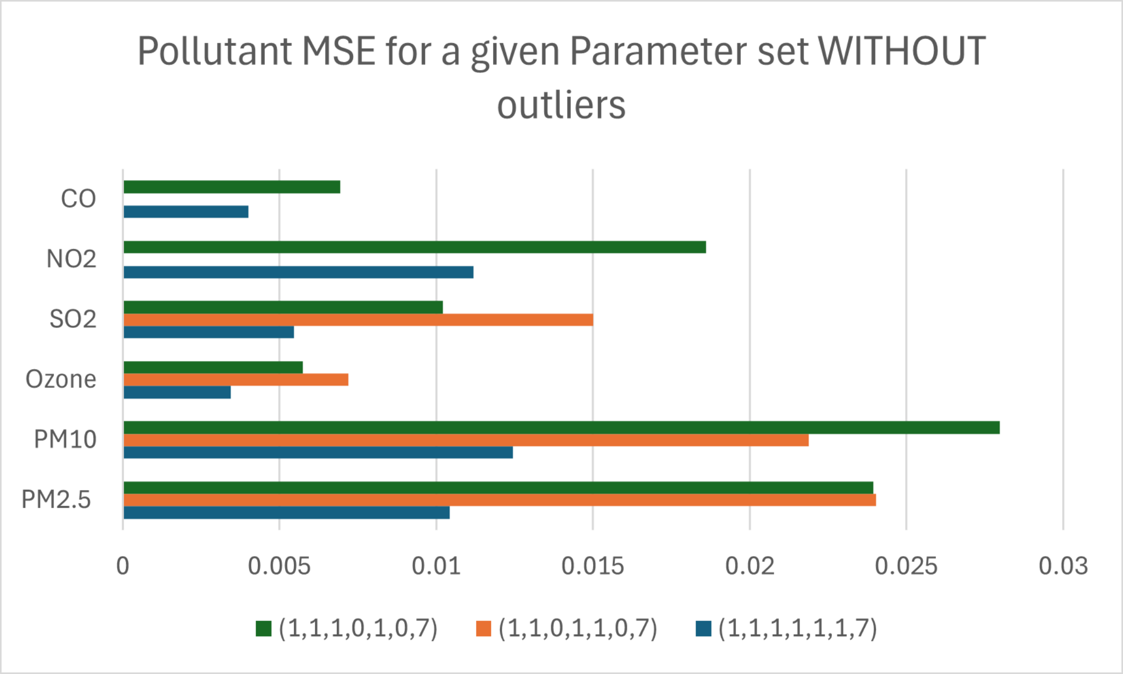


Figure 3: From this chart, the parameter set (1,1,1,1,1,1,7) results in the smallest MSE metric for all pollutants. Note that outliers for CO and NO2 for parameter set (1,1,0,1,1,0,7) are removed so that the chart is readable. Parameter set (1,1,0,1,1,0,7) will not be used for either pollutants.

for each criteria pollutant. I stored metadata for each model in MongoDB (MongoDB, 2024), and due to the size of each model object being too big to store in this database, I utilized MongoDB's built in file system, GridFS, to store the model objects (MongoDB GridFS, 2024). All models are consistently retrained and re-ingested into MongoDB and GridFS using an auto-ingestion script that runs every two weeks.

Once the models were saved, I then had to implement a script that would pull these models down, append new data to each model from when it was last trained, grab a forecast for each one, and send it back to the client each time the user requested a forecast. And most importantly, make it all happen in less than a second. Due to the size of each model being, on average, 600 MB on disk and even larger when loaded into memory, there were only so many threads that could be spawned before the machine would run out of local memory. Including all models for all pollutants, I had about a terabyte of data, therefore grabbing a forecast for each model was going to take longer than a second. To solve this problem, I implemented an auto-ingestion script that runs daily at 7am, grabs a forecast from every model, and stores it in a separate collection that can then be queried every time the user makes a request. This collection is indexed based on pollutant and CBSA area in order to make the query instantaneous. Additionally, when the script is run each morning, it grabs the new data that was ingested at 6am for the previous day and compares that value to the predicted value. This recorded error is stored in the same collection, allowing me to keep track of how well each model is performing. These errors can be found in Figure 4. Once scripts for both training and forecasting were completed, a front-end menu was implemented which is pictured in Figure 5. Benchmarks for these scripts can be found in Table 3.

Support for accurate forecasts is important because, similarly to predictive weather apps, it allows users to do advance planning. This is particularly important for vulnerable and sensitive groups who, depending on the AQI levels, could be advised to stay indoors for the next week. To help distinguish between what pollutants are worse for different sensitive groups, Zephyr provides tooltips that appear when you hover over the legend. These descriptions are pulled from the EPA and give an in-depth description of the effects each pollutant can have on human health and specifically who should be more cautious as the AQI increases.

After completing both parts of my thesis, Zephyr was redeployed and is currently up and running, free to the public.

Auto-Ingestion Script	Benchmark	Number of Threads
Forecasting Script	2 hours	17 threads
Training Script	11 hours	6 threads

Table 3: SARIMA ingestion script benchmarks

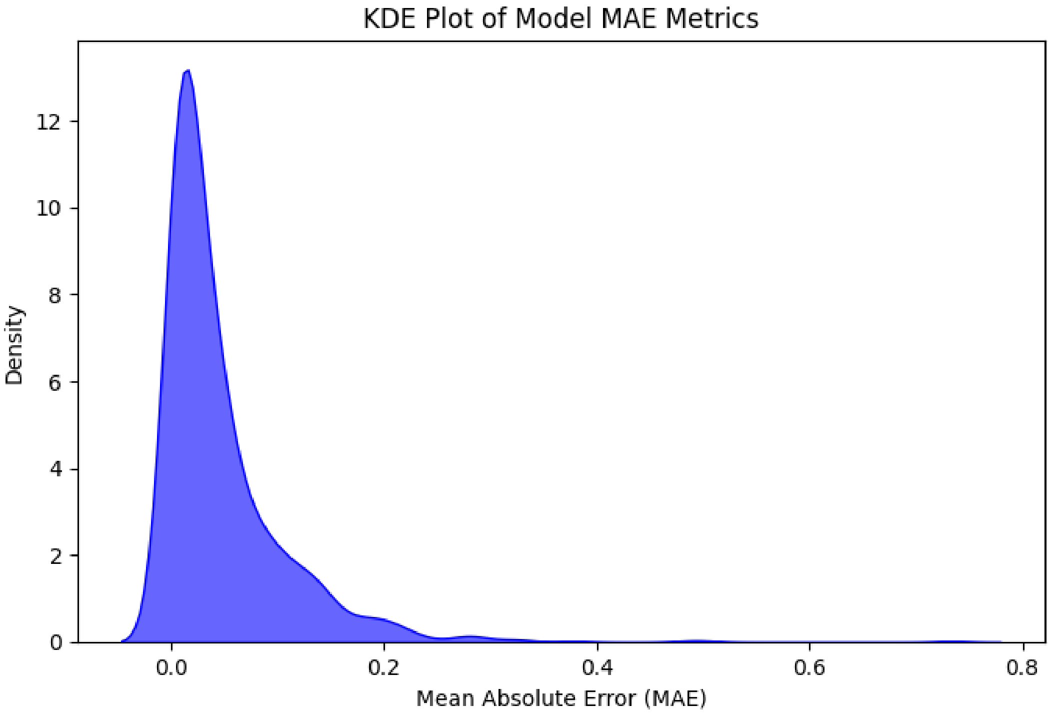
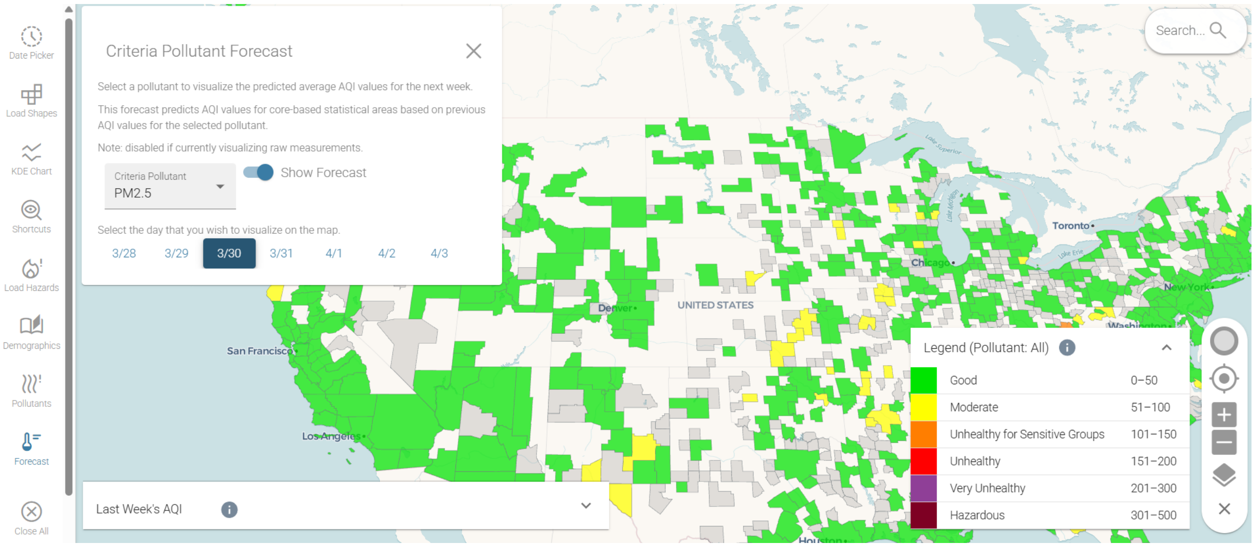


Figure 4: Displays the Mean Absolute Error distribution across the 1011 models I have error statistics for. Note that not all ¡pollutant, CBSA¿ pairs have data recorded each day so calculating an error for that model is impossible. The average MAE across these models is 0.048.



Criteria Pollutant Forecast

Select a pollutant to visualize the predicted average AQI values for the next week.

This forecast predicts AQI values for core-based statistical areas based on previous AQI values for the selected pollutant.

Note: disabled if currently visualizing raw measurements.

Criteria Pollutant: PM2.5

Show Forecast

Select the day that you wish to visualize on the map.

3/28 3/29 3/30 3/31 4/1 4/2 4/3

Last Week's AQI

Legend (Pollutant: All)

🟩	Good	0–50
🟨	Moderate	51–100
🟧	Unhealthy for Sensitive Groups	101–150
🟥	Unhealthy	151–200
🟪	Very Unhealthy	201–300
🟫	Hazardous	301–500

Figure 5: Forecasting Menu on Zephyr.

4 Broader Context

Regarding the general public, accessibility to this data is the first step towards change. Air quality and climate change are inseparable and both affect one another as all major pollutants share common sources with greenhouse gases. Not only do some pollutants have a direct effect on climate change, such as Ozone and Particulate Matter, but climate change can also affect the air quality (Schneidemesser, 2020). Specifically with gas-fired power plants, the most substantial impact of the operation and maintenance is emissions, including sulfur dioxide, nitrogen oxides, particulate matter, carbon dioxide, mercury, and other pollutants. These emissions don't only affect human health, but also wildlife health and the growth and survival of plants and other organisms, inherently reducing plant biodiversity (Global Energy Monitor, 2024). Zephyr has the potential to help bring attention to the location of these power plants and their proximity to citizen homes. For instance, between Boulder and Jefferson County in Colorado, four power plants exist; 3 Natural Gas power plants and one Bituminous Coal power plant.

The location of these plants is crucial for residents to be aware of, as it impacts both air and water quality. One of the 3 Natural Gas power plants mentioned above includes Boulder's Valmont Power Station that, just in 2017, had its last coal unit removed and is now solely burning natural gas. For over 90 years, this plant burned coal for electricity and has not only had an effect on the air quality but also the water quality, and is still a pressing issue. Something known as coal ash, which is the residue of burned coal, was removed over decades from inside smokestacks and boilers and moved into several nearby landfills and storage ponds on site. An article published in December of 2023 claims that this dump covers nearly 15 acres between Valmont Butte and Leggett Reservoir and holds about 1.6 million tons of coal ash waste (Hickman, 2024). To put this into perspective, this could fill almost 500 Olympic-sized swimming pools. This coal ash contains a mix of minerals such as quartz and clay, as well as toxic heavy metals, and resulted in contaminated groundwater. A cleanup plan is currently being put into place only after 6 years of Xcel reporting unsafe levels of groundwater contaminants. This information is not only important for Boulder residents to know but also for the 6 other communities in Colorado where Xcel has coal ash stored (Hickman, 2024).

Pollutant information also becomes more relevant when considering other health risks humans are exposed to. For instance, a study released in November 2023 from the Boston-based nonprofit research organization, Health Effects Institute (HEI), found that long-term exposure to air pollution, mainly particulate matter and nitrogen dioxide, can elevate your risk of hospitalization and death from COVID-19 (Global Energy Monitor, 2024). More specifically, this risk was strongest when exposure to these pollutants originated from combustion of fossil fuels, especially in motor vehicles. Whether we like to acknowledge it or not, air quality affects our daily lives and only enhances other diseases that pose a threat to human health.

This web application is important because it brings attention to our air quality and specifically what sensitive groups are put in danger by it. Air pollution is one of the greatest environmental threats to public health and contributes globally to an estimated 7 million premature deaths each year (UNEP, 2023). Not only is air quality important for healthy lungs, but it also plays a significant role in the current state of our home and every living organism on planet Earth. This is not a problem to ignore, and Zephyr has the potential to help spread awareness as well as help scientists look into past air quality patterns. Unlike other sources that map daily air quality indexes, Zephyr allows the user to look at past affects of power plants or natural disasters, for example, how past fires affected surrounding areas and for how long. This information can be extremely crucial, especially when looking into patterns that could take form in the future or just simply what to expect from the next natural disaster.

5 Conclusion

Research has played a pivotal role in shaping my college career, providing me with invaluable problem-solving skills, critical thinking abilities, and hands-on experience. Through working in the Urban Sustain research lab, I have gained a deeper understanding of my field, specifically software engineering and the importance of breaking up big tasks in order to make consistent progress. It is sometimes hard to see the end when working on a project that spans multiple months, but by setting weekly checkpoints and meeting with peers and mentors, I have learned how to better approach this process and hence become a better computer scientist. This experience has not only reinforced my passion for innovation but has also prepared me for future challenges in both academia and industry.

6 References

[1] "Air Pollution Note – Data You Need to Know." UN Environment Programme, UNEP, 6 Sep. 2023, www.unep.org/interactives/air-pollution-note/

[2] Environmental Protection Agency. (2024, August 16). U.S. Environmental Protection Agency. EPA. https://www.epa.gov/

[3] Environmental Protection Agency. (2020). Air Quality System (AQI) API. EPA. https://aqs.epa.gov/aqsweb/documents/data_api.html#signup

[4] EMBL-EBI. (2018). ChEBI. Chemical Entities of Biological Interest (ChEBI). https://www.ebi.ac.uk/chebi/libchebi.do

[5] Global Energy Monitor. (2024, May 17). Health effects of gas plants.https://www.gem.wiki/Health_Effects_of_Gas_Plants#:%7E:text=The%20most%20substantial%20impact%20of,are%20nitrogen%20oxides%20or%20NOx.

[6] Lelieveld J, Haines A, Burnett R, Tonne C, Klingmüller K, Münzel T, Pozzer A. Air pollution deaths attributable to fossil fuels: observational and modelling study. BMJ. 2023 Nov 29;383:e077784. doi: 10.1136/bmj-2023-077784. PMID: 38030155; PMCID: PMC10686100.

[7] MUI. (2024). Overview - Material UI. Material UI. https://mui.com/material-ui/getting-started/

[8] MongoDB. (2024). GridFS for Self-Managed Deployments. GridFS for Self-Managed Deployments - Database Manual v8.0 - MongoDB Docs. https://www.mongodb.com/docs/manual/core/gridfs/

[9] MongoDB. (2024). The Developer Data Platform. MongoDB.

[10] Ozdogar, C. (2023, August 10). Time Series Forecasting using Sarima (python). Medium. https://medium.com/@ozdogar/time-series-forecasting-using-sarima-python-8db28f1d8cfc

[11] NOAA. (2021, November). State of the Science FACT SHEET. NOAA. https://sciencecouncil.noaa.gov/wp-content/uploads/2022/07/Air-Quality-SoS-Fact-Sheet-FINAL-2022.01.18-1.pdf

[12] Schneidemesser, E., Driscoll, C., Rieder, H., & Schiferl, L. (2020, September 28). How will air quality effects on human health, crops and ecosystems change in the future?. The Royal Society. https://royalsocietypublishing.org/doi/10.1098/rsta.2019.0330

[13] statsmodels. (2025). Statsmodels.tsa.statespace.sarimax.SARIMAX¶. statsmodels.tsa.statespace.sarimax.SARIMAX - statsmodels 0.15.0 (+638). https://www.statsmodels.org/dev/generated/statsmodels.tsa.statespace.sarimax.SARIMAX.html

[14] Tyler Hickman, G. A. (2024, April 22). Hidden hazard: Boulder's million-ton coal ash problem. The Boulder Reporting Lab. https://boulderreportinglab.org/2023/12/11/hidden-hazard-boulders-million-ton-coal-ash-problem-has-no-local-watchdog/