NASA DEVELOP National Program
California- Jet Propulsion Laboratory (JPL)
Fall 2021

Oklahoma Health and Air Quality

Mapping Air Quality Using NASA Earth Observations to Investigate Recent Increases in Ozone Concentrations

DEVELOP Technical Report
Final – November 20th, 2021

Carolina Rosales (Project Lead)
Robert Alward
Kjirsten Coleman
Katherine Howell
Vanessa Machuca

Advisors:
Le Kuai, NASA Jet Propulsion Laboratory, California Institute of Technology (Science Advisor)
Heidar Thrastarson, NASA Jet Propulsion Laboratory, California Institute of Technology (Science Advisor)
Benjamin Holt, NASA Jet Propulsion Laboratory, California Institute of Technology (Science Advisor)

1. Abstract

Tropospheric ozone (O_3) is formed by anthropogenic pollutants interacting with sunlight and is considered harmful to human health in high concentrations. In the summer of 2018, the Oklahoma Department of Environmental Quality (DEQ) measured unexpected spikes in O_3 in Seiling, Oklahoma, with concentrations exceeding those measured in bustling Oklahoma City and Tulsa. The DEQ tracks air quality using ground monitors and does not utilize Earth observation data in its monitoring or analysis. This project used remotely sensed data to investigate these 2018 air quality anomalies, identifying possible causes. We analyzed atmospheric data from Terra and Aqua Moderate Resolution Imaging Spectroradiometer (MODIS), and Sentinel-5P Tropospheric Ozone Monitoring Instrument (TROPOMI) in conjunction with ground-based measurements of tropospheric ozone (O_3), nitrogen dioxide (NO_2), methane (CH_4), carbon monoxide (CO), formaldehyde (HCHO) and aerosol optical depth (AOD). We compared Hybrid Single-Particle Lagrangian Integrated Trajectory (HYSPLIT) model simulations and Earth observation visualizations to pinpoint ozone spike causes. We also generated models to identify contributing factors to variations in ground ozone concentrations in our study area. The results point to a variety of ozone spike causes, primarily from outside of the state, and support the placement of additional NO_2, O_3, and CO monitors to the southeast of Seiling. These analyses can help guide the placement of future monitors in the ground monitoring network and inform air quality regulations in Oklahoma.

Key Terms
HYSPLIT, MODIS, TROPOMI, back-trajectory analysis, emissions transport, atmospheric pollutants

2. Introduction

2.1 Background Information

Tropospheric ozone (O_3) is a greenhouse gas that is formed near Earth's surface when pollutants react with sunlight. These pollutants are emitted by cars, power plants, industrial boilers, chemical plants and other sources, and are referred to as ozone precursors (Butcher, 2020). Ozone in the troposphere is considered harmful to human health in concentrations above 0.07 parts per million (ppm). At these concentrations, it is associated with increased rates of lung cancer, asthma, and other lung-related diseases, as well as impaired cognitive function (American Lung Association, 2021). These adverse effects are a growing concern in states like Oklahoma, where uncharacteristic ozone spikes are being observed.

In 2011, Oklahoma suffered from some of the worst air quality in the country largely due to high ozone levels (American Lung Association, 2021), with multiple days surpassing the Environmental Protection Agency (EPA) limit for healthy air. The Oklahoma Department of Environmental Quality (DEQ) identified widespread drought as the likely main contributor to worsened air quality. By 2016 air quality improved, reaching a record 83% of days considered "good" according to EPA standards (Brown, 2019). However, in the summer of 2018 the DEQ noticed unexpected spikes in O_3 at a ground-monitoring station in Seiling, Oklahoma (Figure 1). While ozone spikes are expected in urban areas like Oklahoma City and

Tulsa, high levels in the rural town of Seiling were surprising, and prompted further investigation, especially given that the air quality monitor in Seiling was created to record background air quality measurements for the state. Unlike the event in 2011, this event was not preceded by a drought and the specific causes of the spikes were unclear. Since the spikes, Oklahoma's poor air quality has persisted, and in certain locations fails to meet the EPA's standard for particulate matter. According to reports by the Oklahoma DEQ and the American Lung Association (2021), in 2019 ozone levels peaked near the EPA's standard of 0.0705 ppm (DEQ, 2019).

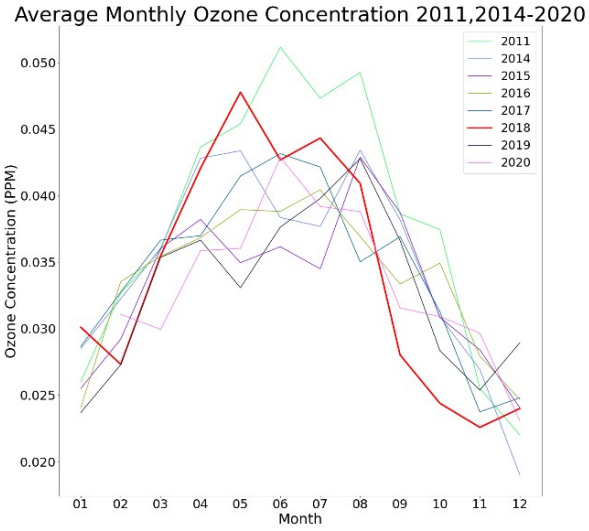


Figure 1. Average monthly ozone concentrations measured from ground stations in Seiling, Oklahoma. Ozone peaked in 2011 following a drought. Ozone levels declined over several years before spiking again in 2018.

To better understand Oklahoma's air quality anomalies, Earth observations (EO) of atmospheric data recorded by satellite-based instruments can be used to supplement ground-based measurement, providing insight into possible emission transportation scenarios. The satellite-based instruments Moderate Resolution Imaging Spectroradiometer (MODIS) and Tropospheric Monitoring Instrument (TROPOMI) have been tested against ground-based monitors globally (Ialongo et al., 2020). Additionally, algorithms for processing and validating data have been developed for linking ground and EO atmospheric data from these instruments (Garane et al., 2019). According to literature, the instruments perform well and are highly correlated with ground-based measurements, even at higher latitudes (Ialongo et al., 2020).

Although research suggests that some EO products are highly correlated with ground-based measurements, often more information is needed to predict surface O_3. Ozone precursors like nitrogen dioxide (NO_2), methane (CH_4), carbon monoxide (CO), formaldehyde (HCHO), and fine particulate matter (PM2.5) are often utilized in investigations of ozone (EPA, 2021). Furthermore, many research studies utilize meteorological, in-situ, and land-use variables. For example, Random Forest (RF) regression models have been used to predict ground-level ozone from satellite-derived and in-situ variables with a strong agreement between predictions and observations (Wang et al., 2022). These models shed light on the emissions and weather scenarios that lead to surface ozone formation.

Emission transport models are also important tools for air quality regulation and are necessary for identifying potential emission source locations for air pollutants. Since air pollutants can travel long distances, to track their source, researchers have developed publicly available models, including the National Oceanic and Atmospheric Administration (NOAA) Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT). This model provides forward and back trajectories, generating statistically likely paths of particles at specific times and locations (Crosman, 2021; Stein et al., 2015). HYSPLIT modeling, together with the qualitative EO data analysis and ground level ozone spike modeling from other data sources, can identify potential sources of particles. (Stein et al., 2015).

2.1.1. Study Area

This study encompassed Oklahoma and seven surrounding states: Texas, New Mexico, Colorado, Kansas, Missouri, Arkansas, and Louisiana (Figure 2). The study period was May 2018 through August 2018, covering the recorded ozone spike dates, as identified by the DEQ (May 26, May 30, June 5, July 17-19, July 22, July 28, and August 2-3, 2018), and allowing for the investigation of the days leading up to spike dates.

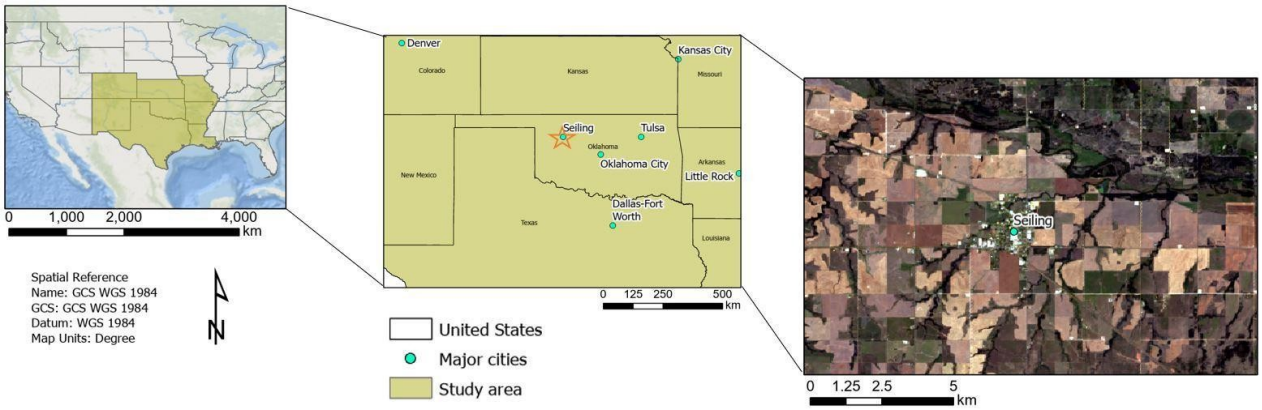


Figure 2. The study area encompassed Oklahoma and the bordering states, shaded in green. Seiling is a small, rural town located in Northwest Oklahoma.

2.2 Project Partners & Objectives

Our team partnered with the Oklahoma DEQ to investigate the sources of ozone and its precursors in Oklahoma and surrounding states. The DEQ tracks air quality measurements using ground-based stations and does not utilize NASA or other EO data in their monitoring or analysis of air quality. The objective of this analysis was to investigate the 2018 air quality anomalies and to better identify emission scenarios that might have contributed to these spikes. The work conducted during the term aided the DEQ in filling gaps in their ground monitoring network and can be used to guide future regulatory and monitoring policies.

3. Methodology

3.1 Data Acquisition

Our team at JPL downloaded all EO data products (Table 1) throughout the study range, May 10 to August 10, 2018. We acquired the products through the NASA

Earth data portal and Google Earth Engine. We then clipped all EO data products to a rectangle encompassing the study area shown in Figure 2.

3.1.1. Earth Observation Data

The TROPOMI and MODIS instruments utilize a passive imaging spectrometer to measure chemical species in the atmosphere with a spatial sampling resolution of 7km and 10km, respectively. Prior to downloading and analysis, we time-ordered, geolocated, and radiometrically corrected the data. We then included NO_2, aerosol optical depth (AOD), CO, HCHO, and O_3 column concentrations in this investigation, given that they are relevant in the ozone formation process, and serve as indicators of anthropogenic activity. Next, we team selected tropospheric column concentrations when available, as the troposphere exhibits similarities to ground pollutant concentrations. Where tropospheric column concentrations were not available, we selected total column concentrations.

Global Land Data Assimilation System (GLDAS) uses models to assimilate satellite and in-situ data to produce global datasets at 1km resolution. Our team derived air temperature, humidity, wind speed, and dew point temperature from GLDAS (Rodell et. al., 2004). Then we derived multispectral data from Sentinel-2 and the Landsat EOS from passive sensors which collect reflectance data across the light spectrum and store data in bands according to wavelength. Dominant land cover type from the National Land Cover Database (NLCD, 2016) and the normalized difference vegetation index (NDVI) from the Multispectral Instrument (MSI) onboard Sentinel-2 were also utilized. Additionally, we utilized the products of the Shuttle Radar Topography Mission (SRTM), an active sensor which utilizes microwave emissions to collect backscatter at a defined wavelength in order to map surface features for elevation.

Table 1
Earth observations, data products and source overview

Platform & Sensor	Products	Download Source
Sentinel-5P TROPOMI	Carbon Monoxide CO Column 1-Orbit L2 7km x 7km V1 Tropospheric NO2 1-Orbit L2 7km x 3.5km V1 Total Ozone Column 1-Orbit L2 3.5km x 3.5km V1 Tropospheric Formaldehyde Column 1-Orbit L2 7km x 3.5km V1 Methane CH_4 1-Orbit L2 7km x 7 km V1	NASA Earth Data Portal
Terra + Aqua MODIS	Terra Aerosol 5-Min L2 Swath 10km Aqua Aerosol 5-Min L2 Swath 10km Burned Area Monthly L3 Global 500m SIN Grid V0006	NASA Earth Data Portal

GLDAS	Noah Land Surface Model L4 3 Hourly 0.25 x 0.25 degree: Air Temperature, Humidity, Wind Speed, Dew Point Temperature	NASA Earth Data Portal
Landsat Earth Observing System (EOS)	National Land Cover Database (NLCD)	Google Earth Engine
Sentinel 2 Multispectral Instrument (MSI)	Normalized Difference Vegetation Index (NDVI) (Near Infrared Band - Red band) / (Near Infrared Band + Red Band)	Google Earth Engine
Shuttle Radar Topography Mission (SRTM)	Elevation	Google Earth Engine

3.1.2. Ancillary Data

Our team downloaded datasets of in-situ ozone, PM 2.5, CO, and NO_2 from the EPA Air Quality System (AQS) for May 10 through August 2018. These are measurements recorded by EPA ground monitors. We also downloaded datasets providing the locations of natural gas processing plants, power plants, and petroleum refineries from the U.S. Energy Information Administration (EIA) (United States, 2010). All data was subset by the noted study area.

3.2 Data Processing

The EO data were cloud masked prior to downloading, resulting in several missing values. Each TROPOMI dataset contains an array of quality flags for each pixel. We clipped all EO data to the study area and collocated with ground sensors. Then we averaged daily EO pollutant levels across pixels within an 11 km x 11 km box around each ground sensor (Figure 3). These average values were joined with ground values to create datasets for the ozone prediction models. We clipped land cover and elevation data sets by a 1km buffer around each ozone ground monitor prior to downloading and used a 20km buffer for average NDVI, which was only used in the Seiling model to capture seasonal vegetation changes. To match the study area, our team also filtered the EPA AQS and EIA point source data.

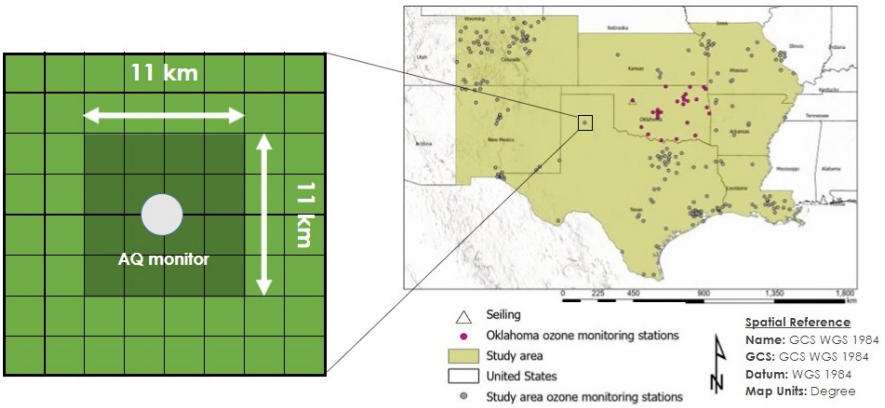


Figure 3. Sampling design to collocate EO and in-situ data for ground ozone formation modeling.

3.3 Data Analysis

3.3.1 Data Exploration

Our team began by exploring and visualizing in-situ data at various spatial and temporal scales within the study area on spike and baseline ozone dates. Next, we visualized EO data by pollutant across the time frame of interest. From static daily plots of EO data, we created animations to explore pollutant distributions. Subsequently, we used these plots in the qualitative analysis alongside HYSPLIT emissions transport model results.

3.3.2 Correlation and Time Lag Analysis

We investigated the correlation of EO data products with in-situ ozone readings to understand how EO data can be leveraged as a predictor of surface level ozone. The time period July – August 10, 2018, was selected to investigate the correlation and time lag between daily ground ozone distribution in Seiling versus nearby TROPOMI EO concentrations. Our team implemented this to exclude any seasonal variations within our study range and to focus on the time period with the most spike days. To obtain the correlation coefficient we collocated total ozone column (TOC) and tropospheric NO_2 EO time series data and regressed it onto ground time series data for June, July, and August. Then, using stationary correlation analysis, we performed time lagged cross correlation analysis to identify the time lag that maximized the correlation coefficient. While setting up the analysis we limited the maximum potential lag to 10 days due to the meteorological movements impacting the ground and EO data alignment.

3.3.3 HYSPLIT Back Trajectory Analysis

To further investigate the ozone spikes in Seiling, our team used the HYSPLIT (v5.1) model to analyze the backwards trajectory of particles in Seiling under the meteorological conditions in which the spikes occurred. We then computed simple particle trajectories, concentration plots, as well as dispersion and deposition simulations (Draxler et al., 2020). In this analysis, we initialized a fixed number of initial particles and subsequently calculated the backwards particle motion

through an autocorrelation function based on Lagrangian time scale and computer-generated random numbers.

The HYSPLIT model used gridded atmospheric data from NOAA's Air Resource Laboratory to calculate the probable location of the backwards dispersion of 1000 particles over 4 days prior to an ozone spike date. Our team set start dates to match a subset of the observed ozone spike dates in 2018 (May 26, June 5, July 19, July 28, August 3), and the location was set to the latitude and longitude of the Seiling sensor (36.14, -98.92). We then set the particles' start height at 100 meters above the ground, and set the top of the model to 1000m, assuming that the ground monitor measured the general ozone in the surrounding area. Our team ran both trajectory and concentration models to better understand the movement path of the ozone particles.

3.3.4 Qualitative cross-validation: HYSPLIT and EO data

Our team qualitatively assessed EO pollutant data with HYSPLIT trajectories by overlaying the two resulting visualizations using QGIS (version 3.16). HYSPLIT simulates particle movements twice per day, therefore, we selected both trajectory polygons for a given day of EO data. We then mapped the identified spike days with 4 preceding days of TOC data and one preceding day of NO2 data to capture the likelihood of pollutant transport based on relative species lifetimes following emission. Our team repeated this process for four randomly selected non-spike days to explore potential differences with baseline levels. For the baseline days, we used days with low to medium ground ozone levels reported in Seiling, where daily maximum levels did not exceed 0.055ppm. Dates fitting these criteria were chosen at random.

3.3.5 Ground-level ozone prediction model

In order to further investigate contributing factors to ground ozone formation within the study area, we utilized EO pollutant data, meteorological data, land cover data, burned area data, and metrics representing oil and gas activity as model predictors. It is known that temperature, pressure, humidity, solar radiation, and wind speed have significant effects on ground ozone levels (Jeong et al., 2020). Our team spatially and temporally collocated gridded meteorological weather data from GLDAS with ground monitors (Figure 3). Since ozone is a secondary pollutant, formed as a result of chemical reactions, we added in other pollutant species as model predictors. The full set of pollutant predictors EO data included the following column densities: Total O_3, Tropospheric NO_2, Total CO, and Tropospheric HCHO, along with AOD. According to the literature, ground ozone levels can vary by day of week due to automobile traffic. Therefore, we incorporated indicator variables into the model to capture weekdays versus weekends. We averaged in-situ time of day variations by using the daily maximum ozone values as the response variable. While variation in hourly air pollution is lost using daily aggregation, this aggregation reflects the daily temporal availability of the EO datasets in question.

The goals of ground level ozone modeling are to identify the most significant predictors for ground level ozone and quantify their impacts on influencing ozone variance, as well as to accurately predict ground ozone levels given weather and pollutant scenarios. With these goals in mind, our team created one set of

interpretable statistical models for the number of ozone particles and one set of models using machine learning techniques. For the statistical investigation, we utilized Python (ver. 3.9.7) and selected Multiple Linear Regression, Poisson Regression, Negative Binomial Regression, and Robust Linear Regression. For the machine learning investigation, we used Random Forest and XGBoost to accurately predict ground ozone. Our team selected forest and boosted machine learning techniques to capture the complex interactions between the explanatory variables used to predict ground ozone.

We created a separate model to predict ground level ozone in Seiling to further understand the contributing factors to Seiling specific ozone spikes. Here, we excluded features that change only with location, namely: land cover and proximity to oil and gas activity. Instead, we included NDVI to capture seasonal changes. We then applied the foregoing models to predict ozone and understand the interaction of various predictors.

To train and assess the performance of the models, our team split the dataset of response and explanatory variables into a training set with 70% of the data and a test set with the remaining 30% of the data. We compared these models using their corresponding mean square errors (MSE), r-squared, and mean absolute error (MAE) statistics to understand their ability to explain the variability in the dataset. Additionally, we assessed feature importance and the statistical significance of predictors to understand which explanatory variables most contributed to variance. Finally, our team compared the results of the study area model and the Seiling specific models to understand the effect of location on ground ozone predictability and the specific features which are the most important in predicting Seiling's ground ozone.

4. Results & Discussion

4.1 Data Exploration Results

The in-situ graphical visualizations showed spikes in summertime hourly ground ozone in Seiling (Dewey County) as compared to monitors in metropolitan counties (Oklahoma and Tulsa) in 2018 (Appendix D). When aggregated into monthly averages, 2018 was a spike year as compared to preceding and subsequent years (Figure 1). Visualizations of EO data showed hotspots of pollutants across the study area (Figure 4).

(a) **(b)**

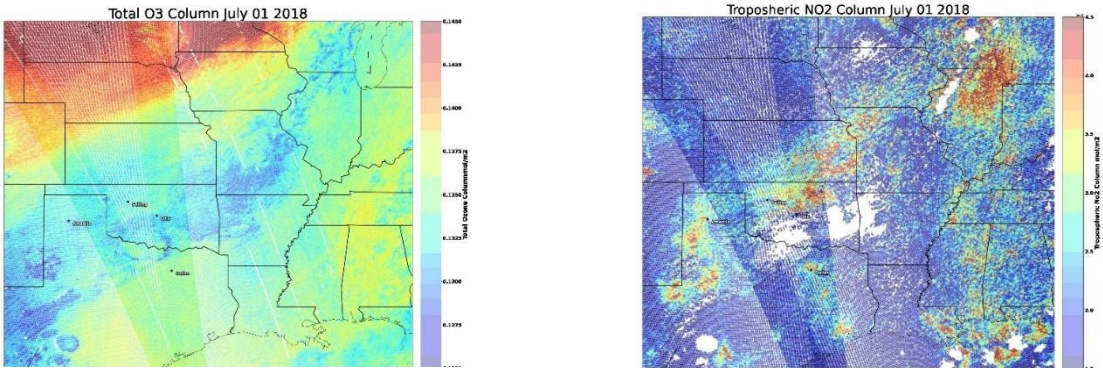


Figure 4. Visualizations of TOC data (a) and NO_2 data (b) showed pollutant hotspots across the study area. Red represents the highest concentrations while blue represents the lowest concentrations, in moles/m^2.

4.2 Correlation and time lag results

The outputs of the analysis for the 45-day period from July 1 to August 15 show for ground Ozone (Figure 5a) and NO_2 (Figure 5b), the ideal lag is 2 days with the EO NO_2 values leading. This lag resulted in a correlation coefficient of .16 between the ground Ozone and NO_2. For TOC Ozone and ground Ozone the maximum correlation coefficient was .39 and was maximized at a delay of 7 days with ground Ozone leading total column Ozone. These results show that time delays between EO data and ground data can influence the desired predictive levels. This exploration informed the HYSPLIT models, statistical models, and machine learning models. Ultimately, to utilize this time lag information more work is necessary.

(a) **(b)**

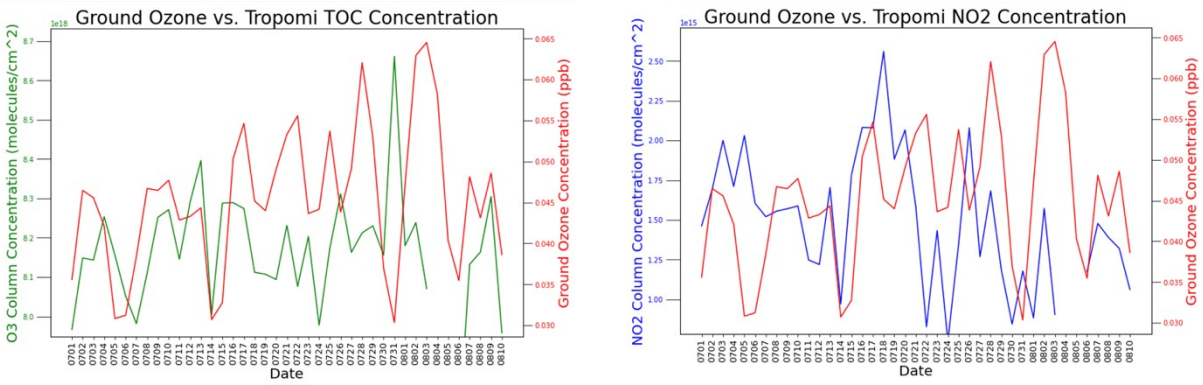


Figure 5. Comparison of ground ozone levels from the Seiling monitor and TROPOMI tropospheric O_3 data in areas surrounding this sensor (a), and TROPOMI NO_2 data (b).

4.3 HYSPLIT back trajectory results

4.3.1 Baseline dates

A random sample of four HYSPLIT back-trajectories preceding low-medium ozone days in Seiling showed consistent patterns among particle movement. However, slight variations exist as particles traveled from south or southeast of Seiling, avoiding Oklahoma City, Dallas, and the Texas panhandle. These are areas of consistent NO_2 hotspots regionally (Figure 6a-d). The trajectory results suggest that particles originated from southern Texas and the Gulf of Mexico (Appendix A).

(a) **(b)**

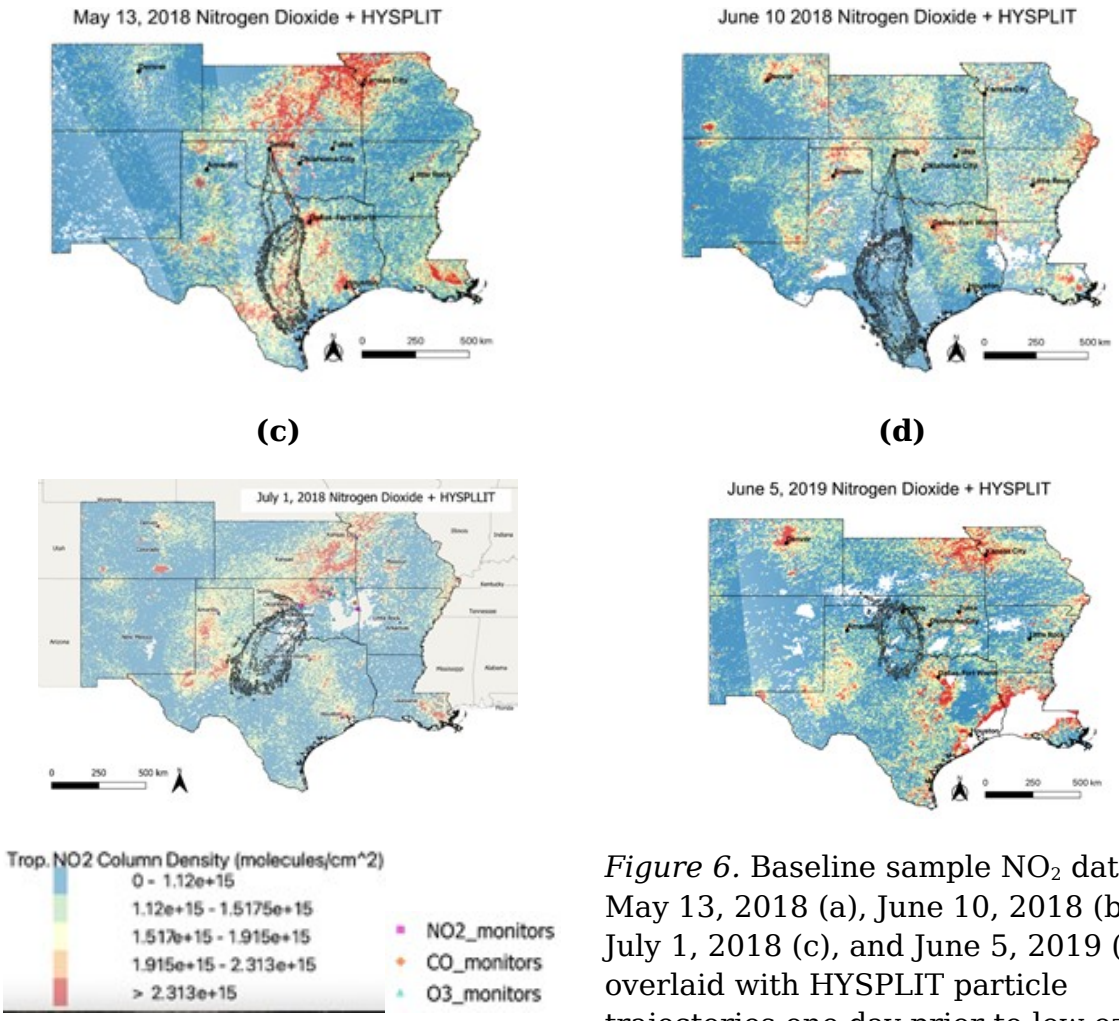


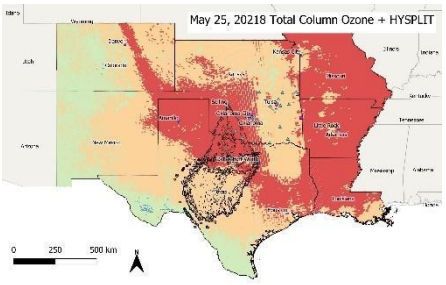
Figure 6. Baseline sample NO_2 data for May 13, 2018 (a), June 10, 2018 (b), July 1, 2018 (c), and June 5, 2019 (d), overlaid with HYSPLIT particle trajectories one day prior to low ozone measurements in Seiling.

4.3.2 Spike dates

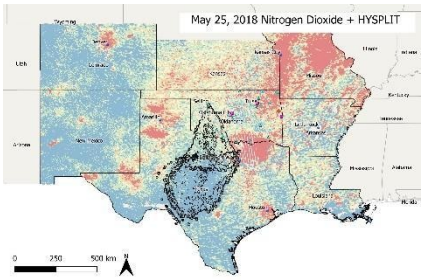
Below, the HYSPLIT trajectories (Appendix B) are presented overlaid with EO pollutant maps for NO_2 and TOC on days preceding spike days where potential interactions occurred. In general, we observed that HYSPLIT trajectories passed through regions of elevated NO_2 on the day prior to a spike day. There was no distinct pattern with TOC concentrations. On May 25, particles traveled through a region of high TOC values south and west of Dallas, while particles did not appear to travel directly through an area of high NO_2 concentration (Fig.7a, b). On June 4 particles moved from north to south across the midwestern states. The maps suggest interaction of high TOC and NO_2 values on June 4 in Oklahoma one day prior to spike day (Fig.7c, d). On July 17, the visualizations suggest an interaction of high TOC values in a region west of Dallas two days prior to spike day, while particles passed through a region of high NO_2 values on July 18, one day prior to a spike day (Fig.7e, f). Preceding the July 28 spike, visualizations suggest no clear interaction of particles with high TOC values along the trajectory, while on July 26, the particle trajectory passes an area of high NO_2 values northeast of Oklahoma (Fig.7g, h). Preceding the August 3 spike, visualizations suggest an interaction of particles and high TOC values along the trajectory with exception to the day prior

to the spike date, while particles pass through high levels of NO_2 in the Dallas area one day before a spike (Fig.7i, j). It is worth noting that spike date trajectories passing through Dallas (Fig 7b, d, j) had the highest daily maximum in-situ ozone levels in Seiling.

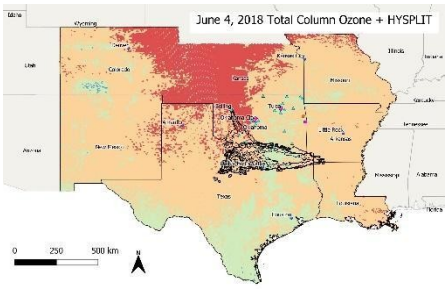
(a)


May 25, 20218 Total Column Ozone + HYSPLIT
0 250 500 km

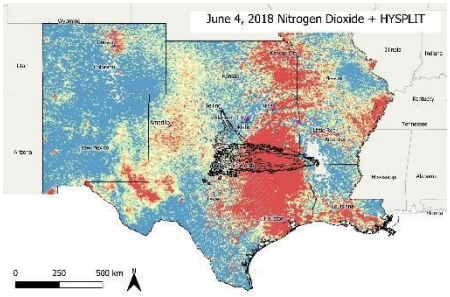
(b)


May 25, 2018 Nitrogen Dioxide + HYSPLIT
0 250 500 km

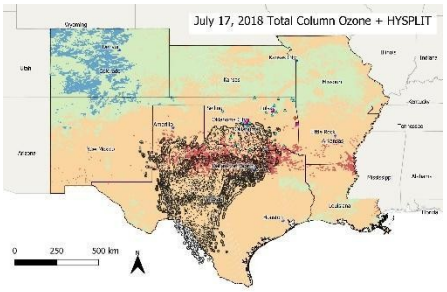
(c)


June 4, 2018 Total Column Ozone + HYSPLIT
0 250 500 km

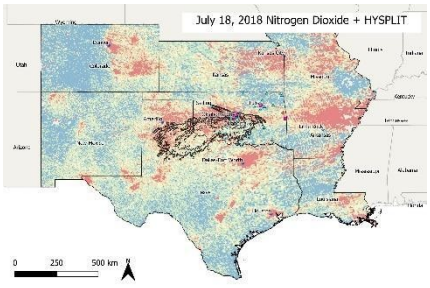
(d)


June 4, 2018 Nitrogen Dioxide + HYSPLIT
0 250 500 km

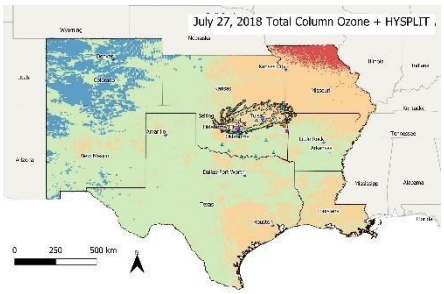
(e)


July 17, 2018 Total Column Ozone + HYSPLIT
0 250 500 km

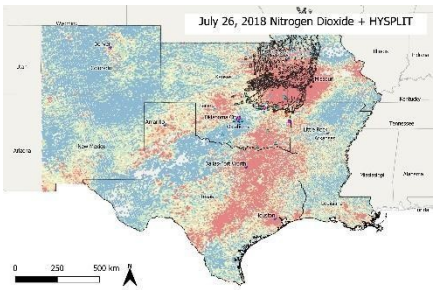
(f)


July 18, 2018 Nitrogen Dioxide + HYSPLIT
0 250 500 km

(g)


July 27, 2018 Total Column Ozone + HYSPLIT
0 250 500 km

(h)


July 26, 2018 Nitrogen Dioxide + HYSPLIT
0 250 500 km

(i)

(j)

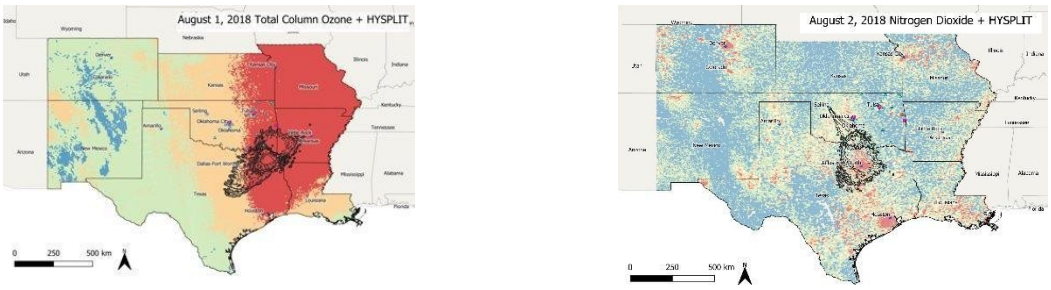


Figure 7. Potential interactions of TOC and NO_2 pollutants and concentrations of atmospheric particles depicted as polygons generated by the HYSPLIT software (a-j).

Given these comparisons and cross validation between daily NO_2 concentrations and hypothetical HYSPLIT trajectories, it is possible that some spikes in Seiling in 2018 could arise from weather patterns bringing NO_2 in from Kansas City, Amarillo, and Dallas and days where NO_2 is high in these areas. On days where wind patterns travel from the Gulf of Mexico, up through southern Texas and between Amarillo and Dallas, ground ozone levels in the summer of 2018 in Seiling were much lower. It is unclear where exactly ozone formation is occurring, and to what degree extremely hot and dry weather events could be affecting these scenarios.

4.3.3. Statistical and Machine Learning Model Results

Both the statistical and machine learning models for the study area had high accuracy with the lowest Mean Squared Error (MSE) coming from the random forest model at 0.36 (Table C1) and the lowest MSE from the statistical models coming from the robust linear regression model, 0.82 (Table C1). The ability of the random forest to learn complex non-linear relationships results in higher performance compared to the linear and general linear models.

The machine learning models identified humidity as the most important feature. This feature importance was determined by the amount of variation explained by the inclusion of the different predictor variables. Through this method of feature importance identification, humidity was the most important variable according to both the random forest and XGBoost models. Longitude and date were identified as the next two most important variables by both the random forest and XGBoost models. The earth observation values had mid-level feature importance falling from 0.0214 to 0.0625 (Figure C1). These features were also modeled qualitatively with the HYSPLIT model which displayed instances where particles aligned with the earth observations.

Regarding the statistical significance of variables, our team used the robust regression which had the lowest MSE to identify the key predictors of ground ozone. The robust regression model identified Latitude, Average TOC Ozone, Ground Temperature, Air Temperature, pressure, humidity, precipitation, and mean elevation as having been statistically significant at the 5% level (Table C4). Interestingly, the robust linear regression model did not identify date, windspeed,

or latitude which were in the top five most important features according to the machine learning models.

For the Seiling model, the linear statistical model was highly accurate with a R-squared of 74.8% (Table C2) which surpassed the r-squared values of the robust linear model for the study area. In the Seiling linear model, the most statistically significant terms were average carbon dioxide, precipitation, and wind speed (Table C5). The high comparative r-squared of the linear model points to the impact of location on ozone formation. Also, the r-squared of the random forest model was not above that of the linear model, yet the random forest had a lower mean square error of 0.5724 (Table C2) compared to the mean square error of .6811(Table C2) in the linear model. In conclusion, the high predictive power of the linear model indicates more straightforward interactions between pollutants and a high influence of location and regional properties on ozone formation

4.3.4. Feature Importance Evaluation
Of our six models, the most accurate was the random forest model, meanwhile the most accurate statistical model was the robust regression. It should be noted that across all the statistical models, predictors: air temperature, ground temperature, humidity, and the interaction between total column Ozone and humidity, explained the most variance in ground ozone, with p values of .005, .001, .020, .028, respectively. Predictors for fire intensity: fire event, weekend, and the number of power producers within 15 km of the monitor, showed limited to no contribution to changes in ground ozone.

4.4 Future Work
Our ozone model identified variables most likely to predict current ozone levels but did not capture ozone formation dynamics. While it can tell us how influential a given factor is on ozone concentrations in a certain area, it cannot tell us where and when that ozone was formed. In general, more statistical models could be developed to track how emission of primary pollutants, like methane, in one location could account for the variation of ground level ozone in another location.

One possible direction would be to develop a model that captures the spatiotemporal dynamics of ozone formation to further explore the influence of climate change. Our analyses indicate that ozone levels are highly correlated with humidity and energy sector activity. As summers become drier and warmer due to climate change, ozone formation dynamics may also shift, regardless of successful efforts to marginally curb emissions. Therefore, one line of inquiry would be to explore whether marginal decreases in precursor pollution levels in hotspot areas will be enough to reduce ozone levels, or does the influence of meteorological factors suggest that these efforts will be outweighed by increasing severity of weather? By developing a forecasting model to address this we could help predict where and when ozone spikes will occur.

Another element that could be further explored in future DEVELOP projects is developing a higher resolution temporal analysis. This could involve leveraging the temporal resolution of various reanalysis products or building downscaling models that can predict EO pollutant data at the hourly level, instead of the daily level. Since air quality fluctuates so rapidly, reinvestigating the questions set forward by

this paper would allow community and policy partners to gain a better understanding of emission transport and the duration of air quality spikes.

Finally, additional quantitative analysis could be applied to our qualitative analysis of HYSPLIT trajectories in conjunction with EO data measurements. This could be done by using the HYSPLIT polygons as input features to join EO measurements spatially. This design could aid in future modeling scenarios.

5. Conclusions

The work conducted throughout this term has showed novel insights into ground level ozone, its sources, and its causes. This study shed light on where Seiling's ground ozone could have originated from, meteorological and other EO product predictors of ground ozone, and the differences between modeling ground ozone in Seiling and the larger study area. This study also raised questions about the impact of climate change on future ozone spikes and methods to improve ground ozone models including temporal lag and spatial elements.

Through exploratory data analysis, our team determined general directions of precursors and total ozone column plumes corresponding to ozone spikes in Seiling. We were also able to identify emission sources and weather patterns that might be leading to anomalous ozone levels in Seiling, by using HYSPLIT back trajectories, and overlaying concentration plot outputs onto NO_2 and TOC column densities. This process was conducted for multiple baseline days where daily maximum ozone levels reported from the Seiling monitor did not exceed 0.055ppm. These baselines served as a comparison for patterns exhibited on spike dates. Through this comparison, our team was able to identify that wind patterns moving through Amarillo and the western panhandle, Dallas, as Kansas City 1-2 days prior to spikes, could have potentially brought high NO_2 concentrations towards Seiling leading to ground ozone formation. Our model indicated that extreme hot and dry weather are also strong pieces of the story.

Of the six models utilized for the study area pollutant regression analysis, the most accurate was the random forest model. Equated to the other models, the mean squared error (MSE) is 0.36, compared with the linear model, with an MSE of 0.82 respectively. It should be noted that across all three models, predictors of humidity, latitude, and distance to petroleum explain the most variance in ground ozone. These values are statistically significant with p values of < 0.00001, 0.002, and < 0.00001 respectively (Table C4). Such analysis assists in identifying which pollutants have the greatest impact on the creation and transport of ozone and should be traced accordingly. It is worth noting, that due to availability of TROPOMI data, pollutant concentrations used as model predictors varied in atmospheric column. While HCHO and NO_2 were tropospheric column densities, CO and O_3 were total column densities.

This work will assist the Oklahoma DEQ guide the placement of new air quality monitors throughout the state, regulate large emitters, strategize more effective emissions reduction programs in the future and, if needed, adopt guidelines that will reduce air pollutants. Given the results of integrated HYSPLIT and EO analysis, our team recommends that NO_2 and CO sensors be placed to the

Southeast of Seiling, in order to detect high precursor concentration levels coming from Dallas and western panhandle of Texas.

6. Acknowledgments

The NASA Develop Oklahoma Health and Air Quality team would like to thank our partners: Carrie Schroeder, Thomas Richardson, Camas Frey, Daniel Ross, Grant Loney, and Cecelia Kleman of the Oklahoma Department of Environmental Quality for their excitement and contributions to this project. We would also like to thank our Fellow Erica Carcelén, mentor Ben Holt, and advisors Le Kuai and Heidar Thrastarson for their guidance and support.

This material contains modified Copernicus Sentinel data 2018, processed by ESA.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Aeronautics and Space Administration. This material is based upon work supported by NASA through contract NNL16AA05C.

7. Glossary

AOD - aerosol optical depth
AQS – air quality sensor
Chemical reanalysis - technique that combines observational information from multiple satellite sensors and provides comprehensive information on tropospheric composition variations
Earth observations – satellites and sensors that collect information about the Earth's physical, chemical, and biological systems over space and time
In situ observations- observations made at the point where the measuring instrument is located
MERRA - Modern-Era Retrospective analysis for Research and Applications
MODIS – Moderate resolution Imaging Spectroradiometer
OMI – Ozone monitoring instrument
Remote sensing – science of obtaining information about objects or areas from a distance, typically from aircraft or satellites
Temporal resolution – denotes how frequently data of the same area is collected; Is typically referred to as the Revisit Time
TOC – total ozone column
Troposphere - the lowest region of the atmosphere, extending from the earth's surface to a height of about 3.7–6.2 miles, or 6–10 km, which is the lower boundary of the stratosphere
VOC - volatile organic compound

8. References

American Lung Association. (2021, April 20). *New Report: Oklahoma's Air Quality Still Failing.* Lung.org. https://www.lung.org/media/press-releases/ok-sota-2021

Beaudoing, H. and M. Rodell, NASA/GSFC/HSL (2020), *GLDAS Noah Land Surface Model L4 3 hourly 0.25 x 0.25-degree* (V2.1), Greenbelt, Maryland, USA,

Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: October 1st, 2021, https://doi.org/10.5067/E7TYRXPJKWOQ

Brown, T. (2019, June 29) *Oklahoma Air Quality Dips After Years of Steady Gains.* Oklahoma Watch. https://oklahomawatch.org/2019/06/29/oklahoma-air-quality-dips-after-years-of-progress/

Butcher, K. (2020, April 21) *2020 'State of the Air' Report: Oklahoma City's air quality worsened for ozone, particle pollution* Oklahoma's News 4. https://kfor.com/news/local/2020-state-of-the-air-report-oklahoma-citys-air-quality-worsened-for-ozone-particle-pollution/

Copernicus Sentinel data processed by ESA, Koninklijk Nederlands Meteorologisch Instituut (KNMI) (2019). *Sentinel-5P TROPOMI Tropospheric NO_2 1-Orbit L2 7km x 3.5km* (V1) [Data set]. https://doi.org/10.5270/S5P-s4ljg54

Copernicus Sentinel data (processed by ESA) (2019). *Sentinel-5P TROPOMI Methane CH_4 1-Orbit L2 7km x 7km* (V1)[Data set]. Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5270/S5P-3p6lnwd

Copernicus Sentinel data processed by ESA, Koninklijk Nederlands Meteorologisch Instituut (KNMI)/Netherlands Institute for Space Research (SRON) (2019). *Sentinel-5P TROPOMI Carbon Monoxide CO Column 1-Orbit L2 7km x 7km* (V1) [Data set]. Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Retrieved October 1st, 2021 from https://doi.org/10.5270/S5P-1hkp7rp

Copernicus Sentinel data processed by ESA, German Aerospace Center (DLR) (2019). *Sentinel-5P TROPOMI Tropospheric Formaldehyde HCHO 1-Orbit L2 7km x 3.5km* (V1) [Data set]. Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5270/S5P-tjlxfd2

Copernicus Sentinel data processed by ESA, German Aerospace Center (DLR) (2019), *Sentinel-5P TROPOMI Total Ozone Column 1-Orbit L2 7km x 3.5km* (V1) [Data set]. Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/110.5270/S5P-fqouvyz

Crosman, E. (2021). Meteorological drivers of Permian Basin methane anomalies derived from TROPOMI. *Remote Sensing, 13*(5), 896. https://doi.org/10.3390/rs13050896

Dewitz, J. (2019). *National Land Cover Database (NLCD) 2016 Products* (ver. 2.0, July 2020): U.S. Geological Survey. https://doi.org/10.5066/P96HHBIE.

Draxler, R., Rolph, G., Stein, A., Stunder, B., & Taylor, A. (2020). *HYSPLIT User's Guide 5.* https://www.arl.noaa.gov/documents/reports/hysplit_user_guide.pdf

Garane, K., Koukouli, M. E., Verhoelst, T., Lerot, C., Heue, K. P., Fioletov, V., & Zimmer, W. (2019). TROPOMI/S5P total ozone column data: Global ground-based validation and consistency with other satellite missions. *Atmospheric Measurement Techniques*, *12*(10), 5263–5287. https://doi.org/10.5194/amt-12-5263-2019

Giglio, L., Justice, C., Boschetti, L., Roy, D. (2015). MCD64A1 MODIS/Terra+Aqua Burned Area Monthly L3 Global 500m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. Retrieved October 1st, 2021 from https://doi.org/10.5067/MODIS/MCD64A1.006

Ialongo, I., Virta, H., Eskes, H., Hovila, J., & Douros, J. (2020). Comparison of TROPOMI/Sentinel-5 Precursor NO 2 observations with ground-based measurements in Helsinki. *Atmospheric Measurement Techniques*, *13*(1), 205-218. https://doi.org/10.5194/amt-13-205-2020

Jeong Y, Lee HW, & Jeon W. (2020) Regional Differences of Primary Meteorological Factors Impacting O3 Variability in South Korea. *Atmosphere,* 11(1):74. https://doi.org/10.3390/atmos11010074

Levy, R., Hsu, C., e., 2015. *MOD04_L2 - MODIS/Terra Aerosol 5-Min L2 Swath 10km* [Data set]. Adaptive NASA MODIS Processing System, Goddard Space Flight Center, USA. https://dx.doi.org/10.5067/MODIS/MOD04_L2.061

MODIS Science Team (2006), *MODIS/Aqua Aerosol 5-Min L2 Swath Subset 10km along MLS.* [Data set]. Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC. https://disc.gsfc.nasa.gov/datacollection/MAM04S0_002.html

Oklahoma Department of Environmental Quality (DEQ). (2019) *Air Data Report 2019.* https://www.deq.ok.gov/wp-content/uploads/air-division/Monitoring_Air_Data_Report_2019.pdf

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K.,Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J.K., Walker, J. P., Lohmann, D., and Toll, D. (2004). The Global Land Data Assimilation System, *Bulletin of the American Meteorological Society*, 85(3), 381-394, https://doi.org/10.1175/BAMS-85-3-381

Stein, A. F., Draxler, R. R., Rolph, G. D., Stunder, B. J., Cohen, M. D., & Ngan, F. (2015). NOAA's HYSPLIT atmospheric transport and dispersion modeling system. *Bulletin of the American Meteorological Society*, *96*(12), 2059-2077. https://doi.org/10.1175/BAMS-D-14-00110.1

United States. (2010) *U.S. Energy Information Administration EIA.* United States. [Web Archive]. Library of Congress. https://www.loc.gov/item/lcwaN0015422/

United State Environmental Protection Agency (EPA). (2021, October). *Ozone Concentrations.* Retrieved from Report on the Environment: https://cfpub.epa.gov/roe/indicator_pdf.cfm?i=8

United State Environmental Protection Agency (EPA). (2021, October). *Particulates, Daily CO, Daily NO2.* [Data set]. https://aqs.epa.gov/aqsweb/airdata/download_files.html

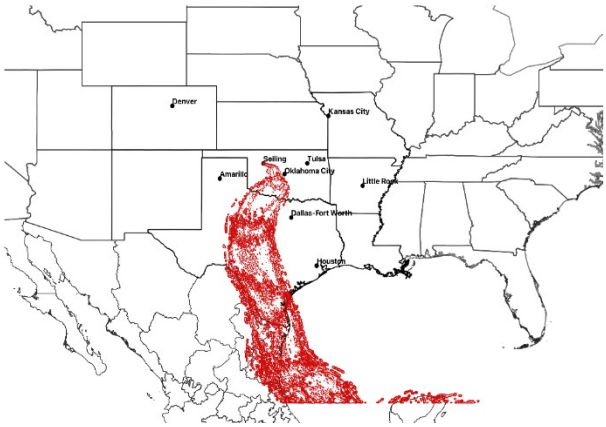
Wang, W., Liu, X., Bi, J., Liu, Y. (2022). A machine learning model to estimate ground-level ozone concentrations in California using TROPOMI data and high-resolution meteorology. *Environment International, 158*, 106917. https://doi.org/10.1016/j.envint.2021.106917

9. Appendices

Appendix A

Panels a – b depict baseline 4-Day HYSPLIT Trajectories

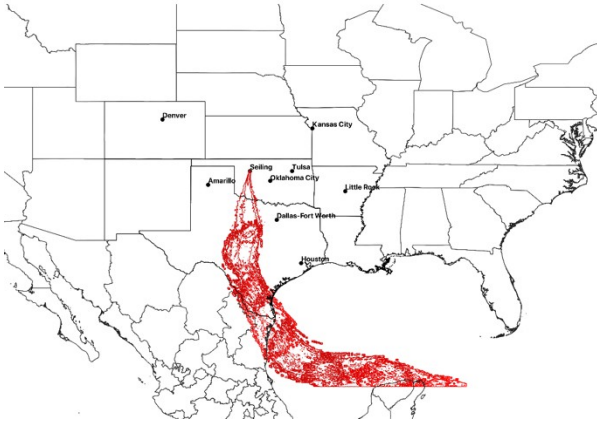
(a)



Simulation Start Date: July 2, 2018

Daily Max Ozone in Seiling: 0.054ppm

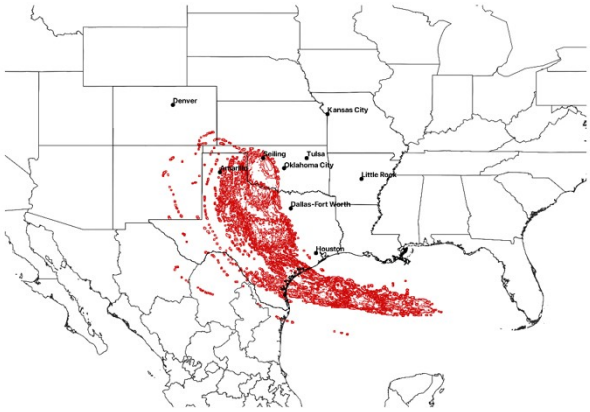
(b)



Simulation Start Date: June 11, 2018

Ground Ozone in Seiling: 0.046ppm

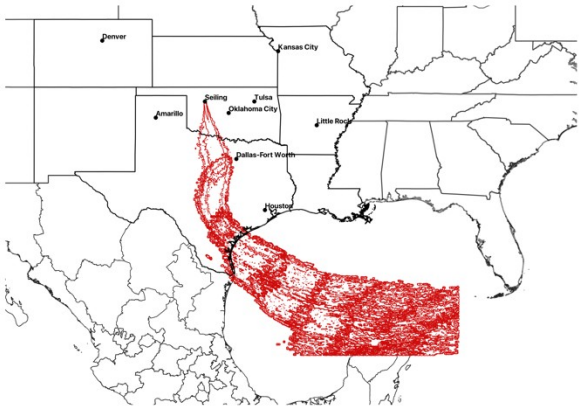
(c)



Simulation Start Date: June 6, 2019

Daily Max Ozone in Seiling: 0.034ppm

(d)



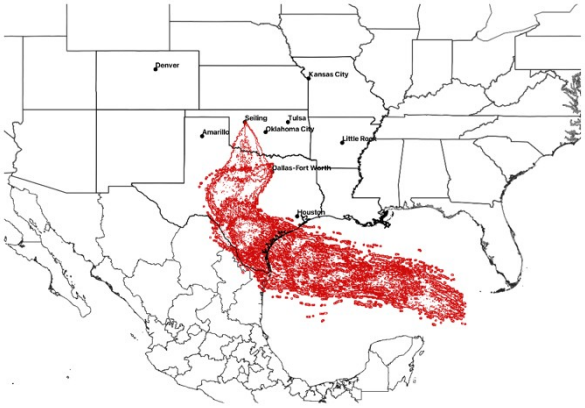
Simulation Start Date: May 14, 2018

Daily Max Ozone in Seiling: 0.037ppm

Appendix B

Panels a – e depict spike Date 4-Day HYSPLIT Trajectories

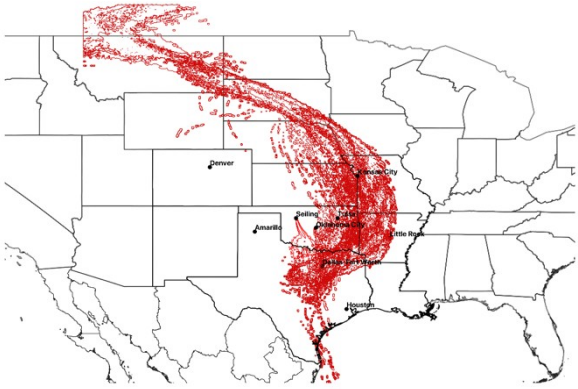
(a)



Simulation Start Date: May 26th, 2018

Daily Max Ozone in Seiling: 0.073ppm

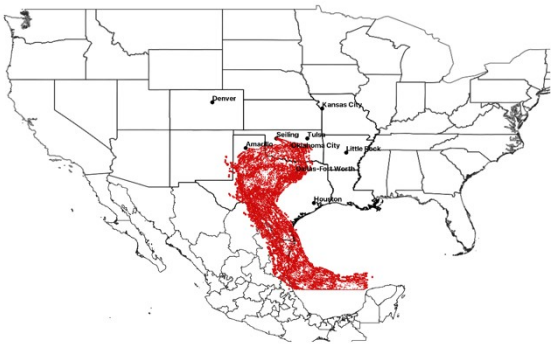
(b)



Simulation Start Date: June 5th, 2018

Ground Ozone in Seiling: 0.074ppm

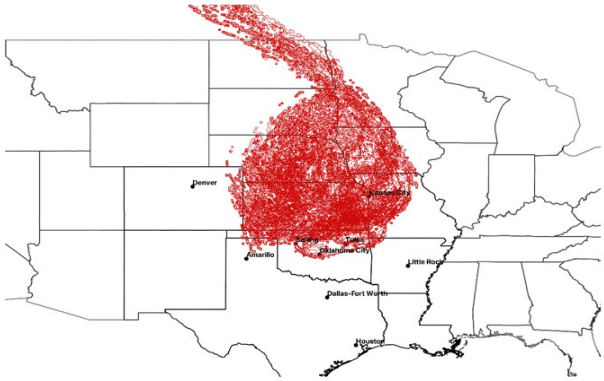
(c)



Simulation Start Date: July 18th, 2018

Ground Ozone in Seiling: 0.070ppm

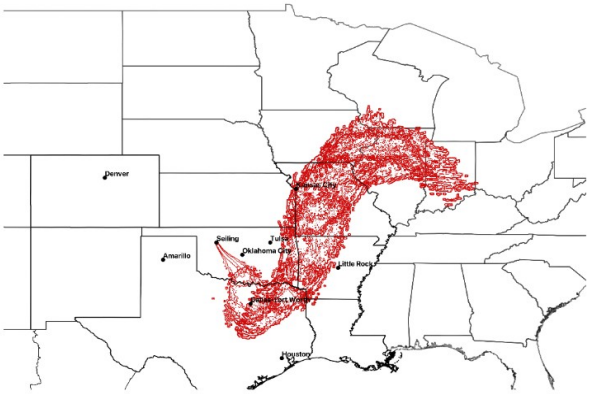
(d)



Simulation Start Date: July 28, 2018

Ground Ozone in Seiling: 0.070ppm

(e)



Simulation Start Date: August 3rd, 2018, Ground Ozone in Seiling: 0.075ppm

Appendix C

Table C1
Key evaluation metrics from the model outputs from the best performing models from the study area. (left) Results for random forest. (right) Results for robust regression.

Study area results for random Forest

Model Evaluation Metric	***Score***
Mean Square Error (MSE)	0.36
R-squared for Linear Model	71.72
Mean Absolute Error	0.44

Study area results for robust regression

Model Evaluation Metric	***Score***
Mean Square Error (MSE)	0.82
R-squared for Linear Model	36.87
Mean Absolute Error	0.64

Table C2
Key evaluation metrics from the model outputs from the best performing models from the Seiling. (left) Results for random forest. (right) Results for robust regression.

Seiling Model results for random forest

Model Evaluation Metric	***Score***
Mean Square Error	0.57

Seiling Model results for robust regression

Model Evaluation Metric	***Score***

(MSE)	
R-squared for Linear Model	65.84
Mean Absolute Error	0.59

Mean Square Error (MSE)	0.68
R-squared for Linear Model	74.8
Mean Absolute Error	0.62

(a) **(b)**

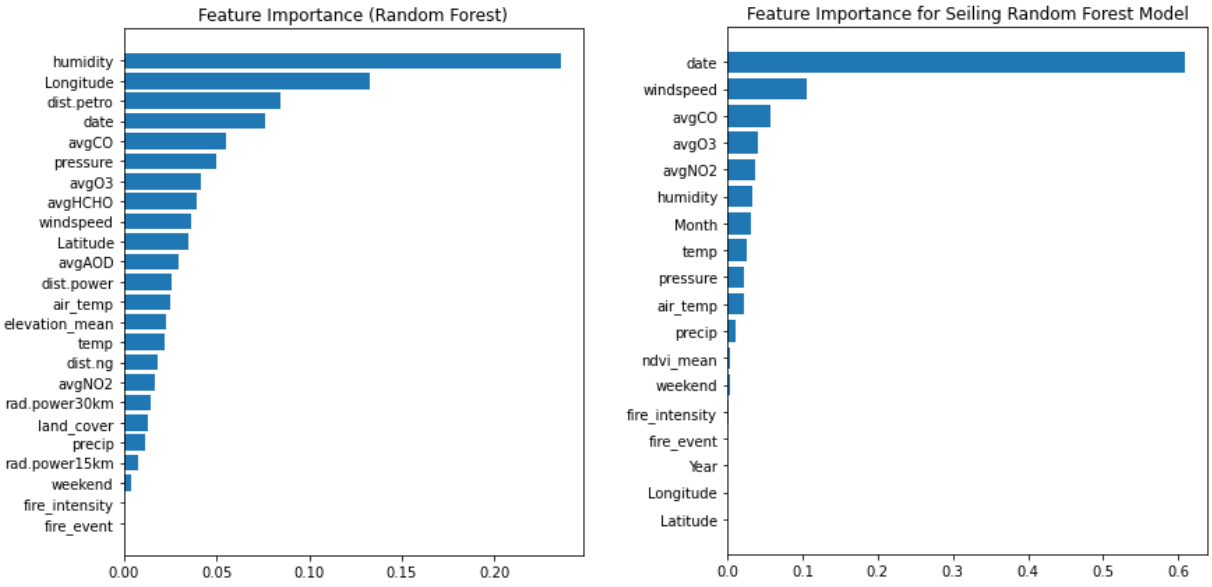


Figure C1. Feature Importance for Random Forest for Study Area (a) and Seiling (b)

The tables below are the key values from the statistical model outputs. The variables below are “coef” representing the coefficient value of the explanatory variable in the given regression, “std err” representing the standard error of the given explanatory variable, “Z” representing the Z-value of the coefficient of the given explanatory variable, and “P>|Z|” representing the P-value of the given Z-value for the explanatory variable.

Table C3
Study area linear regression output of all statistically significant variables at the .05 significance level.

variable	coef	std err	z	P>|z|

name				
Intercept	-44670	14600	-3.052	0.002
Latitude	0.0988	0.021	4.808	0
Longitude	-0.1099	0.028	-3.972	0
Date	0.0022	0.001	3.052	0.002
Avg. AOD	0.0028	0.001	2.979	0.003
Air temperature	-0.2374	0.041	-5.739	0
Temperature	0.1734	0.061	2.85	0.005
Pressure	1.00E-04	2.40E-05	4.841	0
Humidity	-1129.23	356.146	-3.171	0.002
Precipitation	5746.815	1190.166	4.829	0
Elevation mean	7.00E-04	0	4.156	0
Avg. O3: humidity	7947.549	2612.363	3.042	0.002
Avg. AOD: humidity	-0.2122	0.068	-3.112	0.002
Dist. to natural gas	1.54E-06	6.87E-07	2.248	0.025
Dist. to petroleum	-2.95E-06	6.17E-07	-4.774	0
Power in 15km Radius	0.0819	0.027	3.002	0.003

Table C4
Study area robust regression model full output

variable name	coef	std err	z	P>|z|

Intercept	-19680	13100	-1.504	0.133
Latitude	0.09	0.016	5.734	0
Longitude	-0.0185	0.018	-1.002	0.316
date	0.001	0.001	1.503	0.133
Avg AOD	2.00E-04	0	0.825	0.41
Avg CO	-4.8465	8.974	-0.54	0.589
Fire event	-3.8098	6.56	-0.581	0.561
Fire intensity	0.0745	0.085	0.878	0.38
Avg NO_2	-244700	185000	-1.326	0.185
Avg O_3	-79.5164	39.55	-2.011	0.044
Air temperature	-0.3059	0.036	-8.61	0
Ground temperature	0.3353	0.034	9.748	0
pressure	6.54E-05	2.03E-05	3.223	0.001
humidity	-1669.15	261.486	-6.383	0
precipitation	7285.493	1255.771	5.802	0
windspeed	0.0251	0.032	0.785	0.433
mean elevation	5.00E-	0	3.493	0

	04			
Land cover	7.00E-04	0.001	0.446	0.655
weekend	-0.3466	0.073	-4.741	0
Fire event: humidity	69.0523	106.36	0.649	0.516
Latitude: fire_event	0.0849	0.156	0.544	0.586
Avg NO_2: Avg O3	1853000	1350000	1.369	0.171
AvgO_3: humidity	11510	1902.22	6.049	0

Table C5
Seiling linear regression model full output

variable name	coef	std err	z	P>|z|
Intercept	19.8553	11.625	1.708	0.091
date	-0.0013	0.001	-1.711	0.09
avgCO	47.1696	23.66	1.994	0.049
fire_event	6.00E-04	0.001	1.127	0.262
fire_intensity	7.00E-04	0.001	1.209	0.23
avgNO2	-163300	134000	-1.22	0.225
avgO3	-31.948	27.696	-1.154	0.251

	8			
air_temp	-0.025	0.102	-0.246	0.807
temp	0.1215	0.102	1.186	0.238
pressure	8.15E-05	0	0.489	0.626
humidity	-443.007	258.732	-1.712	0.09
precip	3365.689	1652.147	2.037	0.044
windspeed	-0.1509	0.039	-3.833	0
weekend	-0.2637	0.149	-1.771	0.08
fire_event: humidity	9.29E-06	7.98E-06	1.164	0.247
Latitude	717.9255	420.34	1.708	0.091
Latitude:fire_event	0.0238	0.021	1.152	0.252
avgNO2: avgO3	1279000	1040000	1.235	0.22
avgO3: humidity	2395.277	2007.77	1.193	0.236

Appendix D

The graph below shows hourly ozone levels recorded by monitors at Seiling (Dewey County), Oklahoma City, Tulsa, and McAlester (Pittsburg County). The McAlester monitor was used as a baseline in this initial exploration, given that it is similarly unpopulated and rural to Seiling. The graph exhibits how ozone levels in Seiling were elevated compared to urban Oklahoma City and Tulsa, where such levels are to be expected.