An Integrated Approach for Immediate and Long-Term Air Quality Regulation and Monitoring in Mexico

Eliana Kai Juárez

Abstract

Air pollution is among the leading causes of premature deaths across the world and disproportionately impacts the Global South due to lesser air quality regulation and research on regional air pollution monitoring when compared to the United States, Europe, and East Asia. This project consists of three models to address air quality in different ways in Mexico. The first program includes data cleanup, visualization, training, and testing for a machine-learning model to provide hourly air pollution forecasts up to 48 hours in advance as an immediate means of minimizing the harm of hazardous air pollution events for numerous cities, with high seasonal accuracy scores up to R^2=0.77. Second, physical model simulations validated by a statistical analysis of historical relationships between ozone and its precursors present long-term policy suggestions to safely reduce ozone levels at each city, a challenge due to ozone's counter-intuitive behavior (highlighted during the pandemic lockdowns). Finally, to provide continuous estimates of air pollution in regions without monitoring stations, a machine-learning model that incorporates historical ground monitoring, meteorological parameters, and physical model simulations is developed and tested, with cross-validated R^2 scores up to 0.73 - the first of its kind covering all of Mexico.

1. Introduction

Responsible for one in five premature deaths each year, air pollution is a major concern for both environmental and human health. It is the single greatest environmental health risk according to the World Health Organization [1], and anthropogenic emissions are the primary drivers of climate change [2]. The United States upholds National Ambient Air Quality Standards (NAAQS) for six of the most common air pollutants (also known as criteria pollutants), including particle pollution, ground-level ozone, carbon monoxide, sulfur dioxide, nitrogen dioxide, and lead. Of the six pollutants, particle pollution and ground-level ozone pose the most immediate and severe health threats [3].

Air quality can be improved greatly through regulation. In the United States, successful implementation of the Environmental Protection Agency's Clean Air Act reduced air pollution by over 62% percent since the 1970s, avoiding an estimated 200,000 premature deaths since then [4]. This policy and similar regulations have been supported by hundreds of studies using numerous atmospheric chemistry models.[5].

Air pollution models may be used for various purposes, including forecasting future pollutant concentrations, predicting how air quality will change in response to different emission levels, and estimating pollutant concentrations where ground stations are not available. Each of these topics are crucial to informed policy making and public health, and may be achieved by using physical or statistical models [6].

Physics-based atmospheric chemistry models are capable of simulating air quality under new emission scenarios, allowing for a causal relationship (often indirect because of complex ozone chemistry) between emissions and air quality to be determined. However, these models face shortcomings due to their sensitivity to initial conditions (such as input emissions), characteristic of deterministic chaos and unpredictable behavior observed in nonlinear dynamical systems such as the weather [7]. This results in the issue of physical models having 1) extreme dependence on input initial conditions and 2) instability over time and high uncertainty as a standalone model of pollution modeling. These weaknesses are exacerbated in physical models of air pollution in less developed nations, as there are generally fewer ground stations available to measure initial conditions required for such models in the first place.

In addition to physical models, regression-based statistical analyses have been conducted to support policy research. More recently, increasingly advanced machine learning algorithms have been used to understand historical correlations and relationships between air pollutants and other variables [8]. Driven by large data and often hundreds of ground pollutant monitors, countries with fewer stations to begin with are at an additional disadvantage.

A combination of both physics-based and statistical models is a strong approach to overcome the disadvantages of each mode. However, the characteristics of each method present significant barriers to countries with fewer resources available for ground monitoring. This lack of research and regulation in the Global South is reflected by the disproportional mortality rates due to air pollution in these countries, as over 90% of air pollution-related deaths occur in low- and middle-income countries [9].

While US firms may appear to integrate "green" policies, recent studies describe a concerning trend of companies only outsourcing polluting activities to poorer countries with minimal environmental protections rather than meaningfully reducing emissions [10]. While the WHO presents guidelines for air quality, approximately 34% of countries entirely lack legally-mandated air quality standards [11]. There is no international alignment to address the pollution causing eight million annual premature deaths, a figure estimated to grow by over 50% by 2050 [12].

Mexico was chosen for its vast unmonitored regions and proximity to other countries in Latin America with even less monitoring. Additionally, the majority of studies on air quality in Mexico are focused on Monterrey or Greater Mexico City, which both continue to experience dangerous pollution levels posing a threat to their collective populations of over twenty million [13].

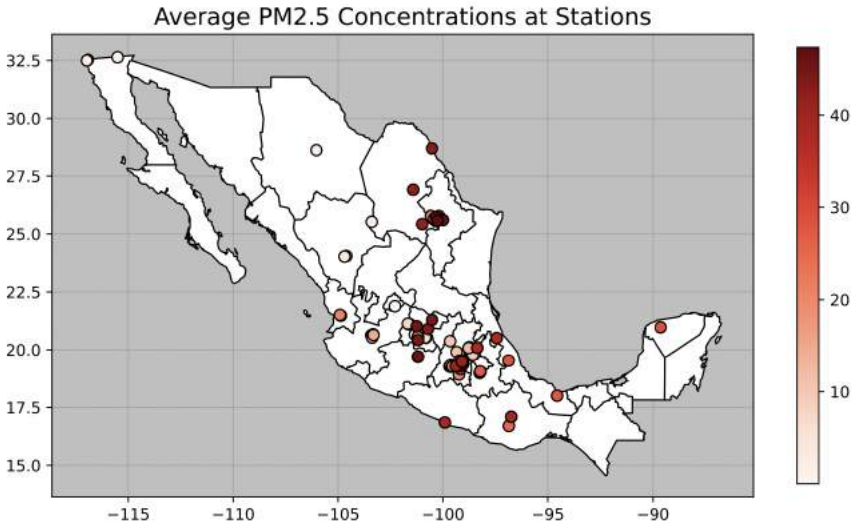

Figure 1. Average 24-hour particulate matter concentration at stations in Mexico. Most exceed the WHO limit of 15 $\mu g/m^3$ for no more than three to four days in an entire year [1].

To combat the disparity in geographic focus of research, this paper presents a three-part approach to both immediate and long-term means to reduce air pollution in Mexico, with data sources and methods that may be easily applied to other regions of the world as a baseline for more specialized regional studies.

1. First, a short-term ozone forecast predicts future concentrations for a public tool to mitigate exposure of vulnerable populations to upcoming hazardous ozone events.
2. Second, a long-term policy recommendation is made using various physical models to simulate ozone levels with different emissions scenarios, supported by historical analysis of pollutant relationships to determine the limiting factor of ozone in each city.
3. Finally, a nationwide model to estimate particulate matter in regions without ground stations is developed using satellite data, historical ground pollutant monitoring, and land use variables. This is compared to NASA's MERRA-2 hourly surface particulate matter estimates for the year of 2018.

2. Short-Term Ozone Forecasting

2.1. Background

In the troposphere, ozone (O_3) is a harmful pollutant formed by photochemical reactions between nitrogen oxides (NOx) and volatile organic compounds (VOCs), as well as sunlight [14]. Rather than being directly emitted, ozone is a secondary pollutant, making its concentrations more difficult to predict and control than a primary pollutant such as particulate matter. Ozone causes throat irritation, coughing, and increases the susceptibility to lung diseases such as asthma [4], but people may avoid dangerous ambient ozone levels with forecasting. Ozone is also known to harm vegetation [15], with major economic impacts as decreases in crop yield due to long-term ozone exposure equal approximately five hundred million Mexican pesos (about $25 million USD) per year [16].

Several studies to forecast ozone have taken place in Mexico. In Monterrey, time series forecasting involving meteorological parameters predicted short-term ozone concentrations within the day using a linear SARMA model [17]. The computational cost and accuracy of Support Vector Machine or SVM was evaluated in to predict real-time ozone, particulate matter, and nitrous oxide in Mexico City by Sotomayor et. al in 2013 [18]. This study differs in its longer pollutant forecasts up to 48 hours in advance, analysis of seasonal variance, and application to cities across the entire country of Mexico.

There are many machine learning algorithms capable of identifying non-linear relationships in complicated data, where various input (or independent variables) are trained to predict an output (or dependent variable). In this model, the current pollutant and meteorological variables will be an input, trained to predict a future ozone concentration as an output.

XGBoost, or extreme gradient boosted trees, is highly optimized for speed, designed to handle missing data, and supports regularization to reduce the potential of over-fitting [19]. XGBoost is widely used for regression problems for its high performance with short training times, and was used for the study.

2.2. Data Collection, Cleanup, & Analysis

Hourly data from 80 ground pollutant measuring stations for 2018 were downloaded from the Sistema Nacional de Información de Calidad del Aire (SINAICA) from Mexico's Instituto Nacional de Ecología y Cambio Climático (INECC) [20]. Due to the machine learning model's sensitivity to nan ("not a number" or missing) values, linear interpolation was applied to fill gaps for less than 6 consecutive hours between valid measurements, filling in no more than 5% of the original data for each variable. Columns with less than 80% completion were dropped, and measurements with irregular or invalid readings were removed.

While most stations also monitored weather parameters, some did not or were very inconsistent. To provide data where weather measurements were missing, ERA-5 ECMWF hourly data for land and single hourly levels over Mexico were downloaded for each station.

The ERA-5 climate reanalysis product has reliable hourly global estimations of weather variables [21] and was validated with station measurements with an average linear correlation between temperature (R=0.91), relative humidity (R=0.84), solar radiation (R=0.64), and surface pressure (R=0.78).

Preliminary correlations and statistical tests were conducted to ensure quality of data going into the model. Individual correlations between variables and ozone were made to understand what factors may be most predictive of future ozone levels, although the linear Pearson correlation coefficient does not capture non-linear relationships that may be useful in the final model. There was much temporal and spatial variation in linear relationships across locations.

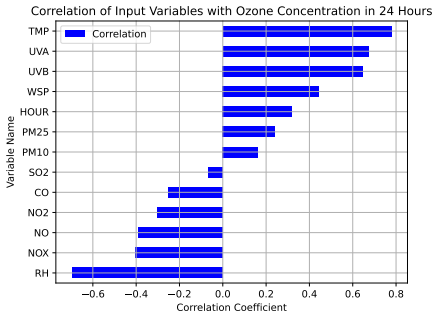

Figure 2. A histogram of the linear relationships between current pollutants and future ozone levels in Torreón, Mexico.

Due to seasonal variations in meteorological and anthropogenic patterns, ozone concentrations and relationships with precursor variables vary across the year. As a result, models may benefit by being trained specific to the seasons, in addition to a year-long model. Mexico's seasons consist of warm dry (March - May), rain (June - October), and cold dry (from November - February) [22].

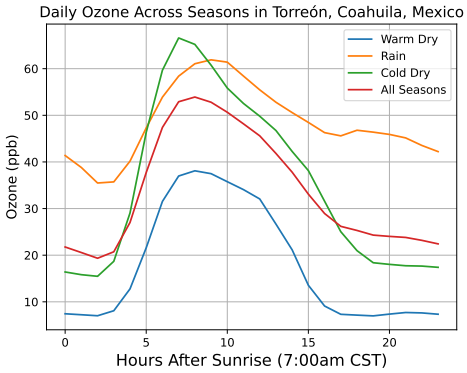

Figure 3. Average hourly ozone values over one day for different seasons in Torreón, Mexico.

Linear multiple regression is the simplest form of machine learning, and may be used as a baseline in ozone prediction studies [23]. To gauge the relative forecasting capabilities of the data, accuracy scores of linear regression models trained to predict ozone from 1 to 48 hours in advance were used to determine how far out in advance a more robust model could reliably predict.

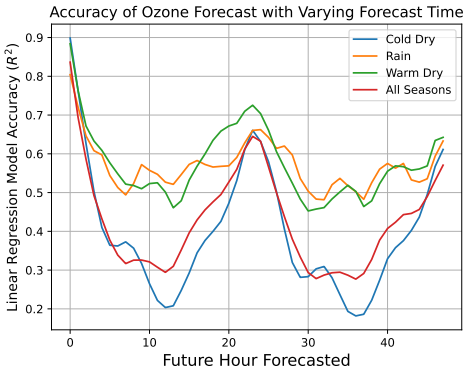

Figure 4. Accuracy of simple linear multiple regression ozone model with varying forecast time.

Figure 4 demonstrates a significant increase in predictive capability at the 24 hour and 48 hour marks across all seasons, suggesting diurnal patterns that improve prediction accuracy at daily intervals despite being further in advance.

2.3. Machine Learning Forecast Model

Because this is a time series problem, k-fold cross validation is not a valid method due to its use of training data ahead of time of the test sample [24]. Instead, a time series split was used to fairly evaluate the models, conducting cross-validation on a rolling basis to avoid bias from training on future values.

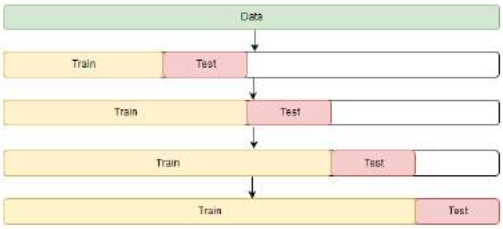

Figure 5. Time series cross-validation [24].

Models were evaluated using several metrics [25]. Mean absolute error, or MAE (1), represents the absolute value of the average difference between the expected and predicted values. Root Mean Square Error, or RMSE (2), is a measure of the standard deviation of residuals and is more affected by outliers than the MAE. For both these metrics, smaller values indicate greater accuracy. R-squared, or the coefficient of determination, is reported on a sample of the data unexposed to the model, allowing for an evaluation of how much of the variance in the dependence variable is explained in the model.

$$MAE = (\frac{1}{n}) \sum_{i=1}^{n} |y_e - y_p| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_e - y_p)^2} \tag{2}$$

To optimize the model for each training period, tuning for each model was conducted using an exhaustive search to find the optimal hyperparameters for learning rate (0.01-0.3), maximum tree depth (between 5-30), and the number of trees (50-250). Feature selection, or the decision of which variables to include, was conducted using a combination of automatic detection of unrelated variables and manual cutoff levels for percentage of missing data interpolated or dropped. These were both done on a seasonal and station-to-station basis.

2.4. Results

For each station, the models were trained and tested with different season and forecast abilities for 24 and 48 hours. The cross-validated metrics for the best performing model of each city are reported in Table 1.

Table 1. Ozone 24 hour forecast results.

Region	City	Season	MAE	RMSE	R^2	R
Northwest	Chihuahua	Cold Dry	0.0027	0.0035	0.5931	0.7701
	Mexicali	Warm Dry	0.0031	0.0039	0.5761	0.7590
	Tijuana	Warm Dry	0.0030	0.0039	0.5389	0.7341
Northeast	Durango	Cold Dry	0.0041	0.0055	0.7749	0.8803
	Torreón	Warm Dry	0.0029	0.0040	0.5747	0.7581
	Saltillo	Warm Dry	0.0030	0.0038	0.6220	0.7887
	Monterrey	Warm Dry	0.0030	0.0039	0.5406	0.7353
Central	Guerrero	Warm Dry	0.0037	0.0045	0.6756	0.8219
	Puebla	Cold Dry	0.0025	0.0033	0.6335	0.7959
	Valle de Mexico	Warm Dry	0.0030	0.0038	0.6255	0.7909
West	Guadalajara	Warm Dry	0.0025	0.0034	0.4654	0.6822
	Leon	Cold Dry	0.0039	0.0049	0.6637	0.8147
	Aguascalientes	Warm Dry	0.0034	0.0047	0.4574	0.6763
	Morelia	Cold Dry	0.0037	0.0037	0.5952	0.7715
Southeast	Minatitlan	Cold Dry	0.0035	0.0044	0.5931	0.8142
	Xalapa	Cold Dry	0.0024	0.0031	0.6359	0.7974
	Oaxaca	Cold Dry	0.0025	0.0036	0.7427	0.7427

R^2 accuracy values ranged temporally and spatially with highest overall scores in the City of Durango in the cold dry season and lowest in Aguascalientes. Score by season tended to be highest for cold dry (R^2=0.77, R=0.87) and lowest during the rain (R^2=0.46, R=0.67), possibly due to the increase in rainstorms which temporarily clear the air of pollution, decreasing the model's predictive ability.

2.5. Discussion

With quality control and simple implementation, this machine learning model maintains high accuracy across different cities with best performance when models are trained seasonally and by region. Cities within the same region consistently performed strongest in the same seasons, with warm dry seasonal models performing best in the northeast and central regions of Mexico, while cold dry seasonal models dominated in the southeast.

However, a data-driven model only presents estimations based on past events, and is not capable of capturing the changing chemistry that may result from differing emission amounts and sources. For a deeper understanding, physics-based models are a valid option.

3. Long-Term Ozone Policy Recommendation

3.1. Background

While data-driven forecast models have reliable results in predicting future ozone in current circumstances, they only present correlations rather than causation. The rate of ozone formation is a variable and complex function dependent on the underlying atmospheric chemistry, which may be understood using physical models to simulate possible emission scenarios for effective policy [26].

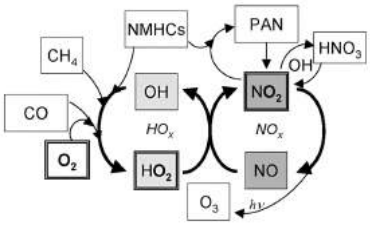

Figure 6. A diagram of several equations involved in ozone formation [27].

Tropospheric ozone is a dangerous and unpredictable pollutant formed by reactions between sunlight and primary pollutants. Volatile organic compounds, or VOCs, are a variety of chemicals released by natural biogenic processes in trees, plants, and animals, with background concentrations throughout the world [28] in addition to man-made sources such as in paints, cleaning agents, and petroleum fuels. When anthropogenic emissions of nitrous oxides (NOx) from fossil fuel combustion are added to the atmosphere, NOx begin to interact with VOCs and sunlight to form ozone. At lesser emitted quantities, nitrous oxides are the limiting factor of ozone formation, which would make lowering NOx the policy goal. However, in many urban locations the amount of emitted NOx is so high that it consumes all available VOCs, in turn making VOCs the limiting factor.

While real-world experimentation is often not possible, a unique opportunity to observe ozone behavior arose in 2020. When COVID lockdowns across the world significantly reduced economic activity and transportation, primary pollutant emission levels fell to record lows [29]. While ozone precursor pollutants such as nitrogen oxides (NOx) drastically decreased, ozone remained high and even increased in Mexico City, as can be noted in Figure 7.

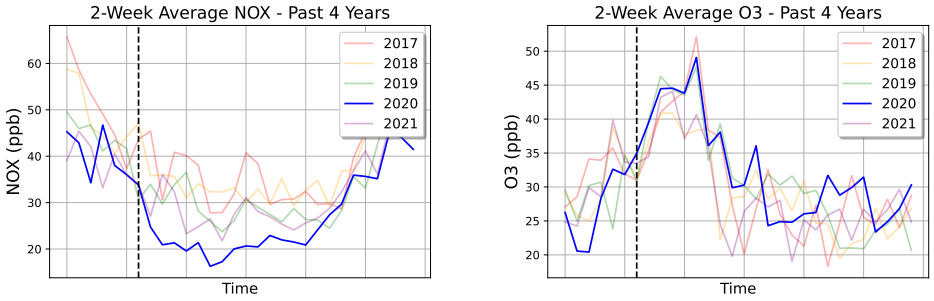

Figure 7. Levels of pollutants over one year- 2020 (dark blue) had lower NOx (left) but higher O_3 (right) compared to previous years. Beginning of lockdowns in March are marked with the dotted line.

This was unexpected, since ozone is a secondary pollutant formed by chemical reactions involving nitrogen oxides which decreased. Yet this trend has been noted in other urban locations that have enacted policies in attempt to reduce ozone, such as China when in 2013 the Clean Air Action Plan led to rapid decreases in particulate matter (PM) and NOx concentrations. There, surface ozone skyrocketed in major cities, which offset some of the health benefits from the PM and NOx emission decreases [30].

The goal of this section is to determine the limiting factor of urban ozone formation with photochemical box models in Mexico City and a country-wide statistical analysis to suggest long-term policy recommendations for numerous cities.

3.2. Eulerian Model

The first model is based on a steady-state solution of a uniformly mixed "box" (Figure 8), assuming negligible deposition from the bottom and transport out the top. This model had a fixed location, or a Eulerian frame [31], which included terms to account for emission rates and wind removal.

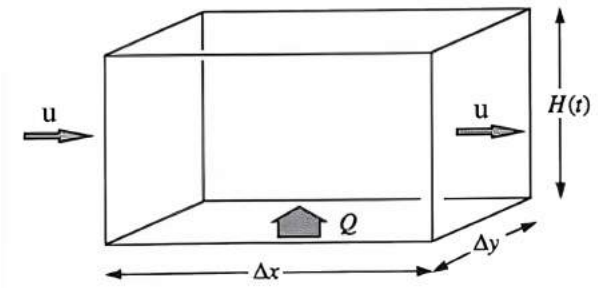

Figure 8. Simplified Eulerian "box" model, where Q represents input emissions [32].

Initially, the ozone formation equations used by Laird et al. [33] were used in the model, originating from the basic advection-diffusion chemistry conservation of mass equation:

$$\frac{d[O_3]}{dt} = k_1[NO_2] + k_3[NO][O_3] - \frac{U}{D}[O_3] \tag{3}$$

Where NO, NO_2, and O_3 are volume averages and U is horizontal wind normal to the box. Ozone volume average concentration for any daytime hour was assumed to be in steady state, and the equations for other chemicals in the model were derived in the same manner.

Using Mexico City's 2018 emissions inventory [34], tons per year of NO, NO_2, CO, propene, and ethylene were converted to molecules per cm^3 per second as inputs to the model. However, the initial emissions model (left, Figure 9) returned negative ozone values (in red), so the equations from Laird involving RO_2, a placeholder for organic peroxyl radicals where R is an organic group, were re-derived. Terms originally dropped and assumed to be negligible were re-incorporated in order to make the equation for RO_2 a quadratic. This resulted in all positive values. The amount of ozone produced after one hour was plotted against varying amounts of initial NOx (NO + NO_2) in Figure 9.

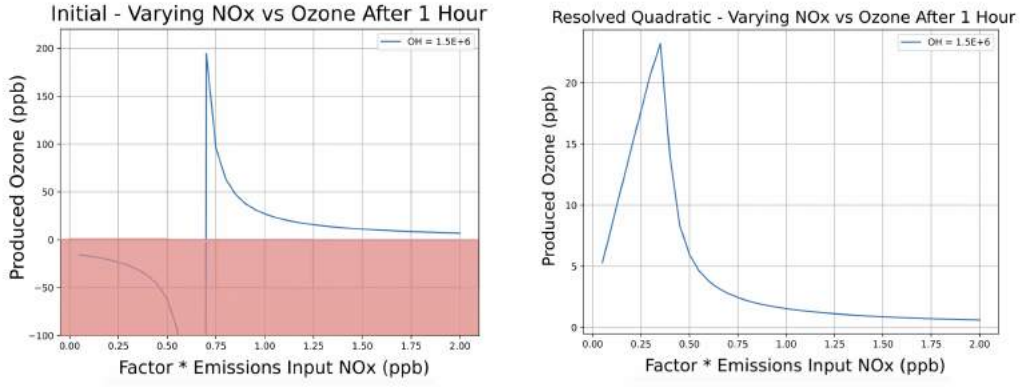

Figure 9. Quadratic RO_2 Eulerian emissions model (right) returns only positive values.

Interestingly, an inverse relationship between ozone and initial NOx concentration is noted in the upper half of initial NOx levels, indicating that the location is VOC-limited.

This modified Eulerian box model provides plausible ozone values and incorporates real emissions data from Mexico City. However, it assumes steady state conditions for all pollutants which does not hold true in reality as various chemicals critical to ozone behavior have varying rates of formation across the day. To address this, a complex Lagrangian model was created for a more realistic simulation.

3.3. Lagrangian Model

The complex model follows an air parcel through space, known as a Lagrangian frame where input concentrations undergo photochemical reactions for a set period of time. Eight essential ozone formation reactions from Seinfeld and Pandis [35] are listed in Figure 10. With a constant source of HOx = HO_2 + OH + RO_2, the other seven rate equations are derived from these chemical equations in Figure 11; three species (OH, HO_2, and RO_2) are steady state [36], leaving four differential equations (for RH, NO, NO_2, and O_3) to be solved for by stepping forward in time.

TABLE 6.3 Generalized VOC/NO_x Mechanism

Reaction	Rate Constant (298 K)
1. $RH + OH \xrightarrow{O_2} RO_2 + H_2O$	26.3×10^{-12} [a]
2. $RO_2 + NO \xrightarrow{O_2} NO_2 + R'CHO + HO_2$	7.7×10^{-12} [b]
3. $HO_2 + NO \longrightarrow NO_2 + OH$	8.1×10^{-12}
4. $OH + NO_2 \xrightarrow{M} HNO_3$	1.1×10^{-11} (at 1 atm)
5. $HO_2 + HO_2 \longrightarrow H_2O_2 + O_2$	2.9×10^{-12}
6. $RO_2 + HO_2 \longrightarrow ROOH + O_2$	5.2×10^{-12} [c]
7. $NO_2 + h\nu \xrightarrow{O_2} NO + O_3$	Depends on light intensity[d]
8. $O_3 + NO \longrightarrow NO_2 + O_2$	1.9×10^{-14}

Figure 10. Essential ozone formation reactions from Seinfeld and Pandis [35]used in Lagrangian model.

$$\frac{d[O_3]}{dt} = k_7[NO_2] - k_8[O_3][NO] \quad (1)$$

$$\frac{d[NO]}{dt} = k_7[NO_2] - k_2[RO_2][NO] - k_3[HO_2][NO] - k_8[O_3][NO] \quad (2)$$

$$\frac{d[NO_2]}{dt} = k_2[RO_2][NO] + k_3[HO_2][NO] + k_8[O_3][NO] - k_4[OH][NO_2] - k_7[NO_2] \quad (3)$$

$$\frac{d[RH]}{dt} = -k_1[RH][OH] \quad (4)$$

Steady State

$$\frac{d[RO_2]}{dt} = k_1[RH][OH] - k_2[RO_2][NO] - k_6[RO_2][HO_2] = 0 \quad (5)$$

$$\frac{d[HO_2]}{dt} = k_2[RO_2][NO] - k_3[HO_2][NO] - k_5[HO_2][HO_2] - k_6[RO_2][HO_2] = 0 \quad (6)$$

$$\frac{d[OH]}{dt} = k_3[HO_2][NO] - k_1[RH][OH] - k_4[OH][NO_2] = 0 \quad (7)$$

Figure 11. Differential ozone equations derived from equations in Figure 10.

The model was ran with varying initial NOx and RH (VOC) concentrations for a 10 hour period. On the resulting isopleth in Figure 14, ozone in parts per billion (ppb) is represented by contours, with initial NOx and VOC concentrations on the Y and X axes, respectively. The dotted red line depicts the shift from negative correlation between ozone and NOx (above) and positive correlation (below).

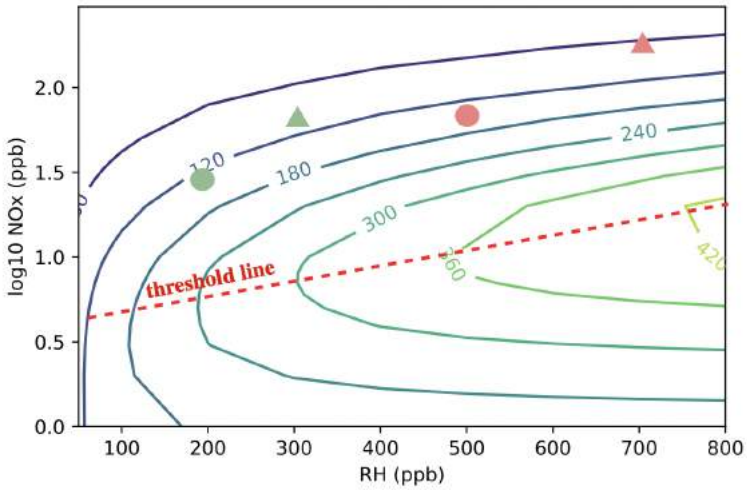

Figure 12. Isopleth Lagrangian model results, including datapoints from Mexico City in 1997 (triangles) and 2017 (circles) for their maximum (red) and minumum (green) ranges at each time.

Year	Level	Reactive VOC, RH	NOx Quartile	NOx Max / Min	Shape
1997	High	700 ppb	84 ppb (75%)	Maximum 370 ppb	▲
1997	Low	300 ppb	40 ppb (25%)	Minimum 6 ppb	▲
2017	High	500 ppb	40 ppb (75%)	Maximum 84 ppb	●
2017	Low	200 ppb	20 ppb (25%)	Minimum 3 ppb	●

Figure 13. Key for historical placements on the isopleth.

Below the threshold line (with lower initial NOx concentrations), decreasing NOx may help decrease ozone formation as well. However if the initial NOx concentration is above that threshold line, ozone will increase as NOx decreases.

Requiring lab equipment and manual testing, the high volatility of VOCs makes their measurement difficult [37], so only data for the observed VOC and NOx levels in Mexico City in 1997 [38] and 2017 [39] were plotted on the isopleth chart. Despite the overall improvements in air quality, all measured points on the isopleth are above the threshold line, indicating that Mexico City has been VOC-limited for the past several decades. While NOx concentrations have decreased slightly, ozone has remained high and possibly increased due to insufficient decreases in VOC concentration.

Taking only a cross-section of the isopleth with a set VOC concentration of 300 ppb (within the range of recent VOC measurements in Mexico City), the change in ozone produced after one hour in relation to NOx change is very similar to that of the Quadratic Emissions Model, where ozone increases with NOx up until a maxima where afterwards the relationships is inverse. This indicates that the simpler model was able to capture aspects of the more complex behavior.

3.4. Statistical Analysis

While Mexico City's location in a well-mixed valley allowed the physical models to simulate plausible ozone behavior, they are very idealized. A statistical analysis on observed relationships between ozone and its precursors is necessary to 1) validate the placement of Mexico City on the isopleth and its status as a VOC-limited location with an inverse relationship between ozone and NOx and 2) suggest the limiting factor in cities without measurements of VOCs, as not all locations have the resources or equipment necessary to do so.

Due to the lack of VOC measurements in most Mexican cities, the linear relationships between NOx and ozone were focused on for the statistical analysis because hourly data of NOx and ozone concentrations are more readily available and may be used to identify the limiting factor of ozone [40].

Mexico City's placement on the isopleth above the threshold indicating a VOC-limited condition is supported in Table 2, where the Ciudad de México (Mexico City) has a strong negative correlation (R=-0.74) between ozone and NOx.

3.5. Results

Table 2. Statistical analysis results, where a strong negative correlation between O_3 and NOx suggests a VOC-limited condition (often the case in urban areas).

Region	City	O_3 to NOx
Northwest	Chihuahua	-0.5734
	Mexicali	-0.2519
	Tijuana	-0.3734
Northeast	Durango	-0.502
	Torreón	0.3750
	Saltillo	-0.3827
	Monterrey	0.018
Central	Guerrero	-0.5499
	Puebla	-0.2189
	Ciudad de México	-0.7435
West	Guadalajara	-0.4516
	Leon	-0.3799
	Aguascalientes	-0.4050
	Morelia	-0.037
Southeast	Minatitlan	-0.0707
	Xalapa	-0.3540
	Oaxaca	-0.2120

Additionally, the relationships between NOx and ozone vary greatly by location, indicating different limiting factors of ozone formation with significant implications for policy in each location. For instance, a location with a strongly negative correlation between NOx and O_3 is likely to have initially high NOx concentrations (such as a city with many cars), and is at risk of increasing ozone when NOx is lowered without also decreasing VOC appropriately. This is similar to what was seen during the COVID lockdown in Mexico City [41].

3.6. Discussion

Although the physical models were very idealized, even the simpler steady-state emissions model was successful in capturing the relationships between ozone and its precursors, indicating a limiting factor and appropriate strategy for lowering ozone.

This program presents multiple physical models that may be run using regional emissions data where available, and finds the historical correlation to support the physical model's estimations of conditions for each city.

4. Surface Particulate Matter Estimation

4.1. Background

Of all criteria pollutants, particulate matter with a diameter of less than 2.5 micrometers, or PM2.5, is the most dangerous to human health due to its microscopic size and ability to infiltrate the bloodstream and respiratory system [42]. Both short and long term exposure to PM2.5 has been proven to increase risk of heart and respiratory disease, and asthma, strongly linked to premature deaths [43]. Particulate matter consists of suspended solid and liquid particles, and is a primary pollutant emitted by fossil fuel combustion, wood burning, and a variety of other physical and chemical sources [44].

Research studies on mortality, health, and economic costs of particulate matter often require data on the continuous distribution of pollution, as point-sources from ground monitors alone are insufficient in supporting proper policy change [45]. Since the 1990s, aerosol maps have been created for more developed nations to calculate the distribution of air pollution at high temporal (often daily) and spatial (from 1 to 10km^2) resolutions [46].

While ground stations have limited geographic coverage as point sources of data, satellite coverage is global and frequent. Aerosol optical depth, or AOD, is a unitless optical measurement of light extinction, such as by aerosols or haze. Remote sensing is often used in models estimating continuous distribution of PM2.5, as it can make up a large component of haze and smog measured by AOD. This has been done in the Northeastern United States, where an ensemble-based machine learning model trained on satellite data, weather parameters, and land-use variables predicted particulate matter at a 1x1km^2 daily resolution with an R^2 ranging from 0.52 to 0.81 [47]. The GOCI geostationary satellite provides high-resolution measurements of AOD specifically over the Korean peninsula, and was used used by Pendergrass et. al to estimate surface particulate matter concentrations across Korea with an R^2 of 0.89 for daily PM2.5 estimations at a 6x6km^2 resolution using a Random Forest model [48]. Another Random Forest model in Japan used MAIAC AOD to map continuous particulate matter across the country, had an R^2 of 0.86 for daily estimations at a 1x1km^2 resolution [49].

Previous studies in Latin America are limited to a citywide scope. In 2019, Carmona et al. used MODIS data, WRF-Chem simulations and ERA-5 ECMWF data in an ensemble neural network to estimate PM 2.5 in Lima, Peru at the 1x1km^2 daily resolution, with an R^2 of 0.70 [18]. Gutiérrez et. al used an XGBoost model to estimate daily mean and maximum PM 2.5 concentrations in Mexico City, with R^2 ranging from 0.64 to 0.86 for daily mean values [50]. In Monterrey, an ensemble neural network was trained to use Merra-2 meteorological fields and aerosol components to predict monthly mean particulate matter concentrations at three stations, with R^2 scores ranging from 0.29 to 0.85 [51].

This study differs in its approach to map the entire country at a high temporal (hourly) and spatial (1x1km^2) resolution, as continuous surface particulate matter prediction has not yet been attempted across all of Mexico. Performance is compared to that of NASA's Merra-2 estimates of particulate matter as a baseline, and an in-depth analysis of model building on prediction time is conducted.

4.2. Data Collection

Ground-monitored station data for hourly particulate matter was collected and cleaned in addition to climate variables from ERA5 using methods detailed in Section 2.2.

MODIS (or Moderate Resolution Imaging Spectroradiometer) is an instrument aboard the Terra and Aqua satellites. MAIAC, or Multi-Angle Implementation of Atmospheric Correction, utilizes data from MODIS instruments on both satellites to provide atmospheric measurements multiple times per day (depending on the number of orbits over each location). Variables include AOD at 470 and 550 nanometers, AOD uncertainty, and column water vapor at a 1x1km^2 resolution [52], each of which may be related to surface particulate matter concentration [53]. In South America, MAIAC's AOD was validated in 2017 by Martins et. al [54], suggesting similar accuracy over Central America as well. Data for the four MODIS sinusinoidal tiles covering Mexico were downloaded for the year of 2018, with overpass times from roughly 16:00-20:00 UTC.

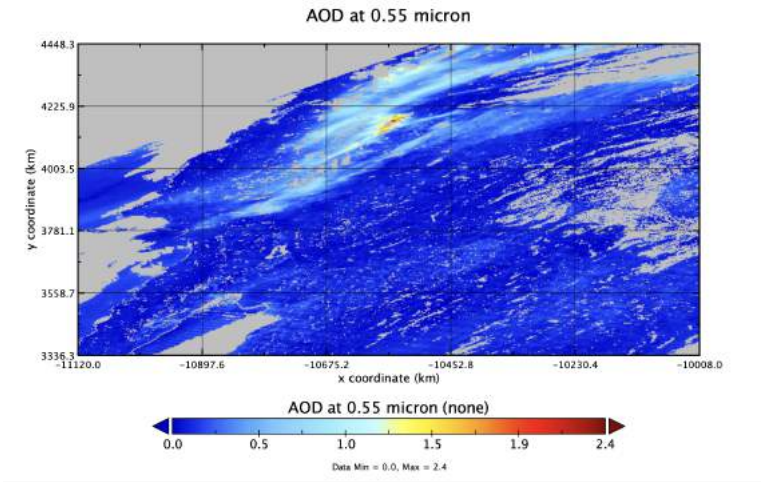

Figure 14. Plot of MAIAC AOD over northwest Mexico on August 4, 2018.

4.3. Data Exploration

The goal of creating a high-resolution model is to estimate particulate matter in locations without monitoring. Due to the complex non-linear relationships between particulate matter and its related variables, machine learning may be used to train models using data from locations that do monitor particulate matter, as long as all the input variables are available for the entire area of Mexico. The variables and data sources used in the model are listed in Figure 15.

Use	Data Source	Variables and Units/Abbreviations	Resolution
Input (Given data)	MAIAC MCD19A2 (MODIS instrument aboard Aqua & Terra Satellites)	• AOD at 550 nm • AOD at 470 nm • AOD Uncertainty • Column Water (cm)	• Daily • Global coverage, 1kmx1km
	ERA-5 Single Hourly Levels from 1959 to Present	• boundary layer height (BLH) • mean sea level pressure (mslp) • total column water (TCW) • total column ozone (TCO)	• Hourly • Global coverage, 25x25km
	ERA5-Land Hourly Data from 1950 to Present	• 2m dewpoint temp (dew) • 2m temperature • surface solar radiation downwards • near-surface pressure (sp) • 10 m meridional wind (v) • 10 m zonal wind (u)	• Hourly • Global coverage, 9kmx9km
Output (Predicted value)	SINAICA	• Particulate Matter (ug/m3)	• Hourly • Point Sources in Mexico

Figure 15. Parameters used, where input data are available for all of Mexico and used by models to predict the output hourly PM2.5 where stations are not available.

Over Mexico, the AOD value is a single daily average of three orbits during a short timeframe (16:00-20:00 UTC), which presents many possibilities for optimizing model accuracy by determining 1) the ideal range of hours when the model can be used for and 2) at which hour can particulate matter be predicted most accurately.

An in-depth exploration of the model's predictive ability with various time frames was conducted primarily in Mexico City due to the completeness of the data. While all stations had relatively similar patterns in training window ability, these methods are capable of being reproduced for stations at an individual basis for increased optimization. They also may be applied to similar high-resolution PM2.5 remote sensing studies in other locations.

First, while models could be trained to predict surface particulate matter at all 24 hours of the day, these may have lower accuracy than models trained to specifically predict hours during the time AOD is measured. Since particulate matter pollution typically peaks in the daytime within the period of the AOD measurement (10:00 am - 2:00 pm local time), it is a viable option to create models focused on estimating the most relevant and accurate hours.

To test this, models were trained on different subsets of a day. It was found that across all seasons, models trained to predict hourly concentrations between 4:00-23:00 UTC performed best (R^2=0.78), compared to those trained to predict all 24 hours (R^2=0.71) or only the hours of satellite overpass (R^2=0.61).

While particulate matter concentration is measured at ground stations on the surface of the Earth, aerosol optical depth is a measurement of the optical properties of the entire atmospheric column, meaning it may be more accurate at estimating the particulate matter concentrations from earlier times of the day because of vertical aerosol transportation. While in practice this may result in slightly delayed estimations, a higher accuracy would be valued for most studies involving long-term analysis of particulate matter distribution.

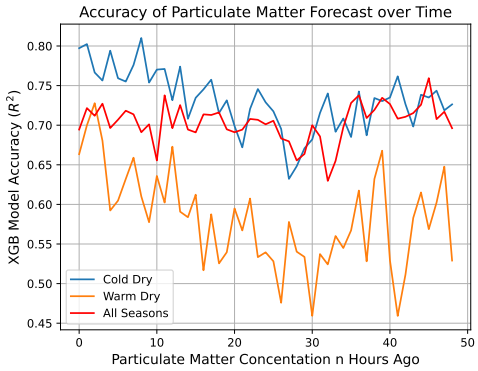

Figure 16. Accuracy of XGBoost model across seasons for predicting particulate matter n hours ago in Mexico City.

Figure 16 demonstrates the ability of seasonal models to estimate the particulate matter concentrations up to 48 hours in the past. For all models, accuracy is highest when predicting particulate matter concentrations 0-2 hours ago, but maintain strong predictive ability up to 48 hours in the past.

As of now, there are no high-resolution continuous distribution maps of particulate matter over Mexico. However, NASA's Merra-2 global historical reanalysis model provides global hourly estimates on the aerosol subspecies [55] that may be used to calculate particulate matter concentration using:

$$[PM2.5] = 1.375 * [SO_4] + [OC] + [BC] + [DU2.5] + [SS2.5] \tag{4}$$

Where sulfate, organic carbon, black carbon, dust, and sea-salt concentrations are combined [56]. The linear correlations between actual PM2.5 and Merra-2 estimates are listed in Table 3, in addition to the cross-validated results of year-long model performance and time taken (in seconds) at each station.

4.4. Results

Table 3. Particulate matter estimation for year-long model results across all cities.

Region	City	Merra-2 R	R	R^2	RMSE	MAE	Test R^2	Time
Northwest	Mexicali	0.0984	0.857	0.734	17.103	9.367	0.816	1312.6
Northeast	Durango	0.066	0.799	0.639	0.7749	5.606	0.689	1157.73
	Torreón	-0.035	0.784	0.616	0.5747	6.752	0.625	1065.56
Central	Ciudad de México	-0.104	0.881	0.776	7.032	5.160	0.855	211.64
	Tlaxcala	-0.009	0.786	0.618	2.456	1.744	0.550	226.93
	Cuernavaca	0.052	0.751	0.564	7.957	4.597	0.134	854.92
	Puebla	0.080	0.699	0.489	8.707	6.089	0.296	1673.51
West	Guadalajara	-0.063	0.820	0.673	11.154	6.695	0.753	1180.94
	Salamanca	-0.033	0.819	0.671	9.039	5.496	0.759	100.17
	Morelia	0.0296	0.792	0.628	7.811	5.044	0.758	476.45
	Tepic	0.010	0.741	0.549	4.769	3.350	0.636	530.18
	Aguascalientes	-0.053	0.610	0.372	10.460	6.714	0.419	1176.58
Southeast	Mérida	-0.023	0.834	0.695	7.657	5.161	0.834	708.36
	Xalapa	0.196	0.811	0.657	7.583	3.981	0.712	958.3
	Oaxaca	-0.043	0.624	0.390	9.092	4.181	0.598	520.1

The model maintained high performance across many regions of Mexico, and Figure 17 compares the performance of the predicted vs actual concentrations for Merra-2 and the new model, respectively.

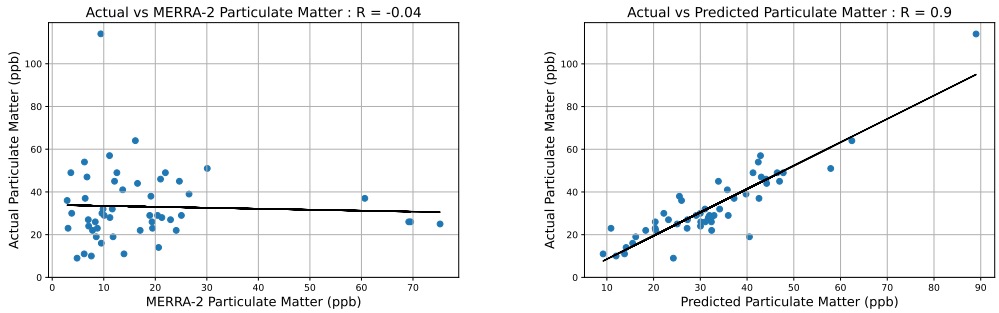

Figure 17. Tested on the same sample, the new model (right) has a high level of accuracy in predicting hourly particulate matter in Mexico City.

As noted in Table 3, the correlation between the actual vs new model's predicted PM2.5 (average R=0.77) is much stronger than Merra-2's PM2.5 yearlong estimate (R=0.011, statistically insignificant with $p < .05$), though it should be noted that Merra-2's estimates are slightly improved during the cold-dry season (averaging an R=0.18). However, physical simulations used by Merra-2 with regional bias-correction may be useful to increase accuracy in future studies.

This model improves upon current methods of particulate matter estimation and has many future applications for long-term policy making, epidemiological, health, and environmental impact studies, and for optimal future sensor placement recommendations, such as how Kelp et al. used continuous particulate matter maps in the US to quantify uncertainty and variability [57]. With reliable results at each city tested, the country-wide model's future incorporation of road types and densities, population distribution, and other land use variables has great potential to further improve results.

5. Discussion

These three approaches serve as an accessible baseline for further research at a regional scale for eventually global applications, which may be achieved through citizen science. The European Environment Agency reported that citizen science has been successful in widespread measurement and awareness of pollutants using low-cost sensors [58], and competitions such as NASA's Airathon: Predict Air Quality challenged teams from across the world to create the most accurate models for predicting particulate matter and NO_2 using AOD measurements over certain cities. Also from NASA, the Applied Remote Sensing Training (ARSET) programs offer many online resources and lectures on how beginners can access and use remote sensing data, and has already reached thousands of participants in over 117 countries with multilingual trainings on air quality monitoring and other environmental applications of remote sensing [59].

These initiatives may inspire citizens to raise awareness of air pollution in areas with little to no previous region-specific research. The detailed methodology and explanation of various approaches to modelling air pollution in this paper present many opportunities for future development, as each component may be improved by:

1. Short-Term Ozone Forecasting
    - Optimizing models for regional seasons
    - Incorporating other variables with hourly data such as local emissions or traffic levels
    - Testing of other machine learning algorithms such as Random Forest or neural networks
2. Long-Term Ozone Policy Recommendation
    - Developing simplified Eulerian and Lagrangian model simulations using regional emissions data
    - Analyzing historical correlations between O_3 and NOx
3. Surface Particulate Matter Estimation
    - Incorporating ground pollution data from low-cost sensors, available from platforms such as OpenAQ [60]
    - Using the best available satellite data depending on region, such as GOCI over East Asia and the future TEMPO satellite over North America [61]
    - Using a similar framework for mapping other pollutants including ozone, CO_2, or NOx with the goal to identify key sources and decrease global warming from greenhouse gases

6. Conclusion

Air pollution remains one of the deadliest threats to humanity due to its high toll on human health and the environment, with increasing relevance in policy making as the main driver of climate change. Because of the great spatial variation in infrastructure, climate, and population characteristics across the Earth, air quality monitoring is most accurate when studied at a regional scale. Increased research focus on the Global South is crucial as wealthy countries continue to push environmentally harmful activities to poorer countries with few current regulations and monitoring.

This project presents a three-part approach to addressing air pollution: first, the process of creating a data-driven machine learning model to forecast ozone concentrations is described as an immediate tool for public health. With an average R correlation of 0.79 across all tested stations, this forecast model has reliable estimates of ozone up to 48 hours in advance.

Next, several physical photochemical box models incorporating local emissions data were developed to determine the limiting factor of ozone formation, validated with observed statistical relationships between ozone and its precursors at numerous cities.

Last, the development of a machine learning model to map continuous particulate matter concentration serves as a means to estimate the distribution of the most harmful pollutant across the entire country of Mexico, crucial to understanding the sources and where policy must be enforced. Compared to the Merra-2's highest accuracy during the cold dry season (R=0.18), the model displayed a high accuracy with average R=0.77. The explanation on extensive testing in training hours may also yield improved results for existing continuous particulate matter maps in other countries.

As discussed by Molina et al. [62], in addition to research, political and social efforts are necessary to drive lasting and effective improvements in air quality. However, these both are built upon interdisciplinary studies conducted on the region of focus. Designed for countries with fewer monitoring stations and previous research, these models have widespread application using publicly available data and accessible programs that require minimal computing power. The codes for each component are available in a public repository online, and aim to provide researchers focusing on geographically understudied areas with the resources to overcome historical barriers and inspire worldwide movement towards improving air quality.

7. References

1. Organization, W.H. *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*; World Health Organization, 2021; pp. xxi, 273 p.
2. Kinney, P.L. Interactions of climate change, air pollution, and human health. *Current Environmental Health Reports* **2018**, *5*, 179–186. doi:10.1007/s40572-018-0188-x.
3. National Ambient Air Quality Standards. https://www.epa.gov/criteria-air-pollutants/naaqs-table. Accessed: 2021-07-01.
4. Chen, T.M.; Kuschner, W.G.; Gokhale, J.; Shofer, S. Outdoor Air Pollution: Ozone Health Effects. *The American Journal of the Medical Sciences* **2007**, *333*, 244–248. doi:10.1097/MAJ.0b013e31803b8e8c.
5. Health Organization, W. Ambient air pollution: A global assessment of exposure and burden of disease. *Clean Air Journal* **2016**, *26*, 6. doi:10.17159/2410-972x/2016/v26n2a4.
6. Air Pollution Modeling. *Air Pollution* **2010**, p. 69–130. doi:10.1201/ebk1439809624-9.
7. Zeng, X. *Chaos theory and its application in the atmosphere*; UMI Dissertation Services, 1998.
8. Chiwewe, T.M.; Ditsela, J. Machine learning based estimation of Ozone using spatio-temporal data from air quality monitoring stations. 2016 IEEE 14th International Conference on Industrial Informatics (INDIN), 2016, pp. 58–63. ISSN: 2378-363X, doi:10.1109/INDIN.2016.7819134.
9. 9 out of 10 people worldwide breathe polluted air, but more countries are taking action.
10. Dai, R.; Duan, R.; Liang, H.; Ng, L. Outsourcing climate change. *SSRN Electronic Journal* **2021**. doi:10.2139/ssrn.3765485.
11. Programme, U.N.E. Regulating Air Quality: The First Global Assessment of Air Pollution Legislation, 2021.
12. Landrigan, P.J.; Fuller, R.; Acosta, N.J.; Adeyi, O.; Arnold, R.; Basu, N.N.; Baldé, A.B.; Bertollini, R.; Bose-O'Reilly, S.; Boufford, J.I.; et al.. The Lancet Commission on pollution and health. *The Lancet* **2018**, *391*, 462–512. doi:10.1016/s0140-6736(17)32345-0.
13. **2019**doi:10.18356/6255ead2-en.
14. Haagen-Smit, A.J.; Bradley, C.; Fox, M. Ozone formation in photochemical oxidation of organic substances. *Industrial & Engineering Chemistry* **1953**, *45*, 2086–2089.
15. Käffer, M.I.; Domingos, M.; Lieske, I.; Vargas, V.M. Predicting ozone levels from climatic parameters and leaf traits of Bel-W3 tobacco variety. *Environmental Pollution* **2019**, *248*, 471–477. doi:10.1016/j.envpol.2019.01.130.
16. Ruiz-Suárez, L.; Mar-Morales, B.; García-Reynoso, J.; Andraca-Ayala, G.; Torres-Jardón, R.; García-Yee, J.; Barrera-Huertas, H.; Gavilán-García, A.; Basaldud Cruz, R. Estimation of the impact of ozone on four economically important crops in the city belt of central Mexico. *Atmosphere* **2018**, *9*, 223. doi:10.3390/atmos9060223.
17. Iglesias-González, S.; Huertas, M.; Hernández-Paniagua, I.; Mendoza, A. Time series forecasting of ozone levels in the metropolitan area of Monterrey, Mexico. *IOP Conference Series: Earth and Environmental Science* **2020**, *489*, 012020. doi:10.1088/1755-1315/489/1/012020.
18. Sotomayor-Olmedo, A.; Aceves-Fernández, M.A.; Gorrostieta-Hurtado, E.; Pedraza-Ortega, C.; Ramos-Arreguín, J.M.; Vargas-Soto, J.E. Forecast Urban Air Pollution in Mexico City by using support vector machines: A kernel performance approach. *International Journal of Intelligence Science* **2013**, *03*, 126–135. doi:10.4236/ijis.2013.33014.
19. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery: New York, NY, USA, 2016; KDD '16, p. 785–794. doi:10.1145/2939672.2939785.
20. Instituto Nacional de Ecología y Cambio Climático.
21. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; Simmons, A.; Soci, C.; Abdalla, S.; Abellan, X.; Balsamo, G.; Bechtold, P.; Biavati, G.; Bidlot, J.; Bonavita, M.; De Chiara, G.; Dahlgren, P.; Dee, D.; Diamantakis, M.; Dragani, R.; Flemming, J.; Forbes, R.; Fuentes, M.; Geer, A.; Haimberger, L.; Healy, S.; Hogan, R.J.; Hólm, E.; Janisková, M.; Keeley, S.;

Laloyaux, P.; Lopez, P.; Lupu, C.; Radnoti, G.; de Rosnay, P.; Rozum, I.; Vamborg, F.; Villaume, S.; Thépaut, J.N. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* **2020**, *146*, 1999–2049, [https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803]. doi:https://doi.org/10.1002/qj.3803.

22. Escoto, J.A.V., 6. Weather and Climate of Mexico and Central America. In *Handbook of Middle American Indians, Volume 1*; Wauchope, R.; West, R.C., Eds.; University of Texas Press: New York, USA, 2021; pp. 187–215. doi:doi:10.7560/732599-006.
23. Jumin, E.; Zaini, N.; Ahmed, A.N.; Abdullah, S.; Ismail, M.; Sherif, M.; Sefelnasr, A.; El-Shafie, A. Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction. *Engineering Applications of Computational Fluid Mechanics* **2020**, *14*, 713–725. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19942060.2020.1758792, doi:10.1080/19942060.2020.1758792.
24. Bergmeir, C.; Hyndman, R.J.; Koo, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics Data Analysis* **2018**, *120*, 70–83. doi:https://doi.org/10.1016/j.csda.2017.11.003.
25. Petris, G.; Petrone, S.; Campagnoli, P. Dynamic Linear Models, 1970.
26. White, J. Prediction of long-term exposure to indoor air pollutants using short-term measurements. *Design and Protocol for Monitoring Indoor Air Quality*. doi:10.1520/stp10154s.
27. Koch, A. Climate Impact Mitigation Potential given by Flight Profile and Aircraft Optimization. PhD thesis, 2013. doi:10.13140/RG.2.1.4896.9047.
28. Jin, X.; Fiore, A.M.; Murray, L.T.; Valin, L.C.; Lamsal, L.N.; Duncan, B.; Folkert Boersma, K.; De Smedt, I.; Abad, G.G.; Chance, K.; Tonnesen, G.S. Evaluating a Space-Based Indicator of Surface Ozone-NOx-VOC Sensitivity Over Midlatitude Source Regions and Application to Decadal Trends. *Journal of Geophysical Research: Atmospheres* **2017**, *122*, 10,439–10,461. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2017JD026720, doi:10.1002/2017JD026720.
29. Lovrić, M.; Pavlović, K.; Vuković, M.; Grange, S.K.; Haberl, M.; Kern, R. Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning. *Environmental Pollution* **2021**, *274*, 115900. doi:10.1016/j.envpol.2020.115900.
30. Li, K.; Jacob, D.J.; Shen, L.; Lu, X.; De Smedt, I.; Liao, H. Increases in surface ozone pollution in China from 2013 to 2019: Anthropogenic and meteorological influences. *Atmospheric Chemistry and Physics* **2020**, *20*, 11423–11433. doi:10.5194/acp-20-11423-2020.
31. Zannetti, P. Eulerian dispersion models. *Air Pollution Modeling* **1990**, p. 107–139. doi:10.1007/978-1-4757-4465-1_6.
32. McHenry, J.N.; Ryan, W.F.; Seaman, N.L.; Coats, C.J.; Pudykiewicz, J.; Arunachalam, S.; Vukovich, J.M. A REAL-TIME EULERIAN PHOTOCHEMICAL MODEL FORECAST SYSTEM: Overview and Initial Ozone Forecast Performance in the Northeast U.S. Corridor. *Bulletin of the American Meteorological Society* **2004**, *85*, 525–548.
33. Laird, A.R.; Miksad, R.W.; Middleton, P. A simple model for urban ozone impact predictions. *Journal of the Air Pollution Control Association* **1982**, *32*, 1221–1225. doi:10.1080/00022470.1982.10465535.
34. Pardo, C.S.; Rodríguez, P.C.; Hernández Villaseñor, S.Z.; García, M.R. Inventario de Emisiones de la Zona Metropolitana del Valle de México 2018. *Dirección General de Calidad del Aire, Dirección de Proyectos de Calidad del Aire* **2021**.
35. Seinfeld, J.H.; Pandis, S.N. *Atmospheric Chemistry and Physics of air pollution: From air pollution to climate change*; John Wiley, 2006.
36. Westenberg, A.A.; DeHaas, N. Steady-state intermediate concentrations and rate constants. ho2 results. *The Journal of Physical Chemistry* **1972**, *76*, 1586–1593. doi:10.1021/j100655a018.
37. Molina, M.J.; Molina, L.T.; West, J.; Sosa, G.; Pardo, C.S.; San Martini, F.; Zavala, M.A.; McRae, G. Air Pollution Science in the MCMA: Understanding source-receptor relationships through emissions inventories, measurements, and modeling. *Air Quality in the Mexico Megacity* **2002**, p. 137–212. doi:10.1007/978-94-010-0454-1_5.
38. Arriaga-Colina, J.; West, J.; Sosa, G.; Escalona, S.; Ordúñez, R.; Cervantes, A. Measurements of vocs in Mexico City (1992–2001) and evaluation of vocs and co in the emissions inventory. *Atmospheric Environment* **2004**, *38*, 2523–2533. doi:10.1016/j.atmosenv.2004.01.033.

39. Zavala, M.; Brune, W.H.; Velasco, E.; Retama, A.; Cruz-Alavez, L.A.; Molina, L.T. Changes in ozone production and VOC reactivity in the atmosphere of the Mexico City Metropolitan Area. *Atmospheric Environment* **2020**, *238*, 117747. doi:10.1016/j.atmosenv.2020.117747.
40. Cohan, D.S.; Hu, Y.; Russell, A.G. Alternative approaches to diagnosing ozone production regime. *Air Pollution Modeling and Its Application XVII*, p. 140–148. doi:10.1007/978-0-387-68854-1_16.
41. Peralta, O.; Ortínez-Alvarez, A.; Torres-Jardón, R.; Suárez-Lastra, M.; Castro, T.; Ruíz-Suárez, L.G. Ozone over Mexico City during the COVID-19 pandemic. *Science of The Total Environment* **2021**, *761*, 143183. doi:10.1016/j.scitotenv.2020.143183.
42. Kim, K.H.; Kabir, E.; Kabir, S. A review on the human health impact of airborne particulate matter. *Environment International* **2015**, *74*, 136–143. doi:https://doi.org/10.1016/j.envint.2014.10.005.
43. Chen, Y.; Wild, O.; Ryan, E.; Sahu, S.K.; Lowe, D.; Archer-Nicholls, S.; Wang, Y.; McFiggans, G.; Ansari, T.; Singh, V.; Sokhi, R.S.; Archibald, A.; Beig, G. Mitigation of $PM_{2.5}$ and ozone pollution in Delhi: a sensitivity study during the pre-monsoon period. *Atmospheric Chemistry and Physics* **2020**, *20*, 499–514. Publisher: Copernicus GmbH, doi:10.5194/acp-20-499-2020.
44. Humbert, S.; Fantke, P.; Jolliet, O. Particulate matter formation. In *Life cycle impact assessment*; Springer, 2015; pp. 97–113.
45. Narita, D.; Oanh, N.T.K.; Sato, K.; Huo, M.; Permadi, D.A.; Chi, N.N.H.; Ratanajaratroj, T.; Pawarmart, I. Pollution Characteristics and Policy Actions on Fine Particulate Matter in a Growing Asian Economy: The Case of Bangkok Metropolitan Region. *Atmosphere* **2019**, *10*. doi:10.3390/atmos10050227.
46. Sethi, J.K.; Mittal, M. A new feature selection method based on machine learning technique for air quality dataset. *Journal of Statistics and Management Systems* **2019**, *22*, 697–705. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/09720510.2019.1609726, doi:10.1080/09720510.2019.1609726.
47. Hu, X.; Waller, L.A.; Lyapustin, A.; Wang, Y.; Al-Hamdan, M.Z.; Crosson, W.L.; Estes, M.G.; Estes, S.M.; Quattrochi, D.A.; Puttaswamy, S.J.; et al.. Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. *Remote Sensing of Environment* **2014**, *140*, 220–232. doi:10.1016/j.rse.2013.08.032.
48. Pendergrass, D.C.; Zhai, S.; Kim, J.; Koo, J.H.; Lee, S.; Bae, M.; Kim, S.; Liao, H.; Jacob, D.J. Continuous mapping of Fine Particulate matter (pm2.5) Air Quality in East Asia at daily 6×6km2 resolution by application of a random forest algorithm to 2011–2019 GOCI Geostationary Satellite Data. *Atmospheric Measurement Techniques* **2022**, *15*, 1075–1091. doi:10.5194/amt-15-1075-2022.
49. Jung, C.R.; Chen, W.T.; Nakayama, S.F. A national-scale 1-KM Resolution PM2.5 estimation model over Japan using MAIAC AOD and a two-stage random forest model. *Remote Sensing* **2021**, *13*, 3657. doi:10.3390/rs13183657.
50. Gutiérrez-Avila, I.; Arfer, K.B.; Carrión, D.; Rush, J.; Kloog, I.; Naeger, A.R.; Grutter, M.; Páramo-Figueroa, V.H.; Riojas-Rodríguez, H.; Just, A.C.; et al.. Prediction of daily mean and one-hour maximum PM2.5 concentrations and applications in central Mexico using satellite-based machine-learning models, 2022.
51. Carmona, J.M.; Gupta, P.; Lozano-García, D.F.; Vanoye, A.Y.; Yépez, F.D.; Mendoza, A. Spatial and Temporal Distribution of PM2.5 Pollution over Northeastern Mexico: Application of MERRA-2 Reanalysis Datasets. *Remote Sensing* **2020**, *12*. doi:10.3390/rs12142286.
52. Lyapustin, A.; Wang, Y. MCD19A2 MODIS/Terra+ aqua land aerosol optical depth daily L2G global 1km SIN grid V006 [data set]. *NASA EOSDIS land processes DAAC* **2018**.
53. Liu, Y. Relationships of wind speed and precipitable water vapor with regional PM 2.5 based on WRF-Chem Model. *Natural Resource Modeling* **2021**, *34*. doi:10.1111/nrm.12306.
54. Martins, V.S.; Lyapustin, A.; Carvalho, L.A.; Barbosa, C.C.; Novo, E.M. Validation of high-resolution Maiac Aerosol product over South America. *Journal of Geophysical Research: Atmospheres* **2017**, *122*, 7537–7559. doi:10.1002/2016jd026301.
55. Randles, C.A.; da Silva, A.M.; Buchard, V.; Colarco, P.R.; Darmenov, A.; Govindaraju, R.; Smirnov, A.; Holben, B.; Ferrare, R.; Hair, J.; et al.. The merra-2 aerosol reanalysis, 1980 onward. part I: System description and data assimilation evaluation. *Journal of Climate* **2017**, *30*, 6823–6850. doi:10.1175/jcli-d-16-0609.1.
56. MERRA-2 FAQ. https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/FAQ/#Q4.

57. Kelp, M.M.; Lin, S.; Kutz, J.N.; Mickley, L.J. A new approach for determining optimal placement of PM2.5 air quality sensors: Case study for the contiguous United States. *Environmental Research Letters* **2022**, *17*, 034034. doi:10.1088/1748-9326/ac548f.
58. Rodrigues, V.; Gama, C.; Ascenso, A.; Oliveira, K.; Coelho, S.; Monteiro, A.; Hayes, E.; Lopes, M. Assessing air pollution in European cities to support a citizen centered approach to air quality management. *Science of The Total Environment* **2021**, *799*, 149311. doi:10.1016/j.scitotenv.2021.149311.
59. Prados, A.I.; Carleton-Hug, A.; Gupta, P.; Mehta, A.; Blevins, B.; Schmidt, C.; Barbato, D.G.; McCullum, A.J.; Hook, E.; Podest, E.; et al.. Impact of the ARSET program on use of remote-sensing data. *ISPRS International Journal of Geo-Information* **2019**, *8*, 261. doi:10.3390/ijgi8060261.
60. Hasenkopf, C.A.; Flasher, J.C.; Veerman, O.; DeWitt, H.L. OpenAQ: A Platform to Aggregate and Freely Share Global Air Quality Data. AGU Fall Meeting Abstracts, 2015, Vol. 2015, pp. A31D–0097.
61. Zoogman, P.; Liu, X.; Suleiman, R.; Pennington, W.; Flittner, D.; Al-Saadi, J.; Hilton, B.; Nicks, D.; Newchurch, M.; Carr, J.; Janz, S.; Andraschko, M.; Arola, A.; Baker, B.; Canova, B.; Chan Miller, C.; Cohen, R.; Davis, J.; Dussault, M.; Edwards, D.; Fishman, J.; Ghulam, A.; González Abad, G.; Grutter, M.; Herman, J.; Houck, J.; Jacob, D.; Joiner, J.; Kerridge, B.; Kim, J.; Krotkov, N.; Lamsal, L.; Li, C.; Lindfors, A.; Martin, R.; McElroy, C.; McLinden, C.; Natraj, V.; Neil, D.; Nowlan, C.; OSullivan, E.; Palmer, P.; Pierce, R.; Pippin, M.; Saiz-Lopez, A.; Spurr, R.; Szykman, J.; Torres, O.; Veefkind, J.; Veihelmann, B.; Wang, H.; Wang, J.; Chance, K. Tropospheric emissions: Monitoring of pollution (TEMPO). *Journal of Quantitative Spectroscopy and Radiative Transfer* **2017**, *186*, 17–39. Satellite Remote Sensing and Spectroscopy: Joint ACE-Odin Meeting, October 2015, doi:https://doi.org/10.1016/j.jqsrt.2016.05.008.
62. Molina.; Velasco.; Retama.; Zavala. Experience from integrated air quality management in the mexico city metropolitan area and Singapore. *Atmosphere* **2019**, *10*, 512. doi:10.3390/atmos10090512.