CHAPTER III

Methods

The recent scientific revolution in data science, acquisition and pre-processing for those outside the field of strict computer science doctrine has become readily available through unique online learning programs, literature, and university classrooms. Institutions, groups, and labs such as the EPA monitoring service, Tropospheric Ozone Assessment Report (TOAR), and Global Monitoring Laboratory at the National Oceanic and Atmospheric Administration (GML-NOAA) have all provided easy access to surface level data and monitoring networks (U. EPA 2013; Schultz et al. 2017). Many researchers have used these vast stores of data and ML/AI ensembles to assign exposures and deduce health trends (S. Abdullah et al. 2019; B. Chen et al. 2023; De Marco et al. 2022; Turner et al. 2016). Large databases generously produced by the wealth of disciplines using established GDAL libraries and ESRI suite programs for analysis of air pollutants, land use classification, and economic strategies offered unlimited development options for this project (Rouault et al. 2025; Dangermond and Goodchild 2020).

As technologies improve, advanced analytical methods become easier to understand, continually emerging into current and future disciplines (Miller and Goodchild 2015). ML and AI models are known to work better if pre-existing conditions and assumptions about the dataset are met, such as non-stochasticity and normalized distribution of training features, as stated in (Raschka and Mirjalili 2019). Ample preprocessing and separation of training features e.g. (Ilić, Popović, and Markić 2020; Napi et al. 2021; Staehle, Rieder, and Fiore 2023) was conducted due to the co-variance between metrological features and constituents driving surface O3 mentioned in Chapter II.. Each geo-atom was constructed of a property defined by a complex ensemble and assigned a value with its corresponding uncertainty predicted from an RK trend estimation of the same features, hence forth known as . Values for and additional data sources across available time periods can be viewed in Chapter IV. Table V.2 in Chapter V depicts the final selection of variables for each model and timeframe used in this project.

Depictions of this surface O3 are exhibited in PHOTUC by predicting the residual geospatial error at each location given values of , then applied to each geo-atom within a raster brick created in python. The number of variations in surface O3 predictive features from their true values are minimized via ’s like CTMs, emissions-based features, and geospatial uncertainty of at monitor locations. This thesis combines similarities in spatial-temporal cycles, like-minded sciences, and the three laws of geography to propose a novel geospatial method for incorporating geospatial uncertainty into ML/AI ensembles (Cao 2022; Goodchild 2004; Liu et al. 2018). Each of the proposed models in this thesis have been seen in a myriad of geographical studies utilizing similar methods with high success.

III.1. Pre-Processing

Linear regression, Ridge and LASSO optimization methods were omitted due to early results yielding significant errors greater than 15% due to their inability to capture cyclic trends over long periods of time. Features overtime can be seen in Figure VIII.2.4 having more seasonal tendencies and variation by county in the AOI. While regression methods with the incorporation of spatial uncertainties can appear more cyclic, their overall results were never better than today’s models e.g. (Hsu et al. 2024; S. Pan et al. 2017; Y. Wang et al. 2021) Temporal ranges of satellite technologies were utilized to create different datasets for testing. GRIDMet data is available from January 1st, 1979 and is continuously updated until a couple days before today by the ECMWF (European Centre For Medium-Range Weather Forecasts 2019), allowing for historical insights with the aforementioned feature transformations. This thesis covers the differences in data available from 2019-2024 and displays daily predictions for 5 randomly chosen months across the time frame, resulting in about 150 images. Full applications of the code will have more than ten thousand images depicting different AOIs across a 20-to-40-year timespan.

A Python script performed the pulling requests, JSON processing, and conversions to a common tabular format for further cleaning (Raschka and Mirjalili 2019). Monitoring data was narrowed down to the counties of Maricopa, Pinal, and Pima, used as a point feature in raster statistics, and saved as a CSV file providing the foundation for the datasets found at the end of section III.5. Variables used in this project can be found in Figure VIII.4.2. Further analyses and applications of these methods are expanded upon in Chapter VI with urban data. Future directions and the impacts of feature transformations proposed in this section are commented on in Chapter VII.

III.2. Vector Locations and Spatial Corrections

The AZ State shapefile (United States Census Bureau 2022) was downloaded at a census tract resolution for the corresponding years within the timeframe. Unique Identifiers (UID) were comprised of state, county, and census tract codes and spatially extracted in Python. ArcPRO was later utilized due to its extreme presence in geographic academia and technical settings. Python allowed for rapid cleaning, duplication, and distribution of project material with ease while ArcGIS pro allowed for high quality touch-up of the created data. Both programs performed spatial-joins for the TIGER-Line shapefile and raster representations of PHOTUC. The shapefile was also used for the creation of a mask for use in raster extraction, display, and overall uniformity of features during training. This mask was created at an initial resolution of 30m from a DEM provided by the U.S.G.S (U.S. Geological Survey 2024) and resampled to 250m for the final images.

The US Census Bureau also contains vast statistical collections of housing and income data that were joined to the study area. This was done to incorporate a visually appealing bivariate map of the AOI seen in Maps VIII.3.1 and VIII.3.2. 28 of the 35 monitors in the AOI were used, with another 65 monitors available for Arizona. Overall, the United States has thousands of monitors for use in future work. Maricopa, Pima, and Pinal correspond to county codes 013, 019, 021 respectively (United States Census Bureau 2025). Chapter II revealed that O3 mechanisms are heavily based on energy components, urban presence, and meteorological drivers (A. M. Abdullah et al. 2017; Ye, Wang, and Zhang 2022) which are common in these three counties and Arizona as a whole. Some features were used to create unique predictor variables for use in the ML/AI ensembles based on CTM and Physics concepts learned in the researcher’s early academic career.

III.3. Dependent and Independent Variable Extraction

Necessary corrections, transformations, and resampling methods were made to reduce the complexity of error and implement estimated trends into the final models. The assumed independent variable, daily maximum 8-Hour surface O3 detections, was selected based on extensive scientific evidence regarding the effects of surface ozone concentrations on public health and transport trends on the surface (Travis and Jacob 2019; M. Wang et al. 2016; W. Zhang et al. 2022); assumed potential dependent variables such and surface temperatures and precipitation were similarly gathered from known physical and chemical factors. Due to collaboration efforts among geospatial engineers, filtering is based on values commonly found in conventional GIS databases. The EPA (E. P. A. EPA 2025) Air Quality Systems (AQS) data for O3 were retrieved via Application Programming Interface (API) using the "dailyData”, “byState" endpoint via in Python. Monitor locations whose center spatially intersected with the PHOTUC region were gathered, filtered by site ID, and corrected based on their provided coordinate reference systems (CRS). Some monitors before 2010 were reported with the North American Datum 1983 (NAD83; EPSG:4269) and were transformed to the World Geodesic System 1984 (WGS84), then finally reprojected to a meter (m) based projection based on the European Petroleum Spatial Grid (EPSG) 32612 using data provided by epsg.io (Pridal et al. 2014). Filtering methods were done with spatial intersections between the latitudes and longitudes of monitor locations, satellite imagery, and shapefile of PHOTUC in EPSG 4326 and displayed in EPSG 32612.



Figure 1

Denotes the full description of the projection system used for AZ. This was done to minimize geospatial error and is a common tactic in many GIS based programs.

III.5. Transformations And Theory Based Features

Domain-inspired features were calculated to test if depictions based on physics and chemistry could improve the model without the full use of a CTM or GEOS-Chem model. Column ozone values in Dobson Units (DU) were converted from total column measurements to surface concentration estimates of mol/unit2 using area, volume, and unit normalization principles (e.g. 1 DU = 2.69x10¹⁶ mol/cm²). An estimation of the kinetic energy of O3 was derived from the ideal gas law (PV=nRT) using downward solar radiation, total ozone column values, surface temperature, and O3 column temperature to potentially estimate thermodynamic energy states for use in later works. Cloud base and top height differentials were combined with pressure gradients and overall cloud fraction to represent potential cloud energy radiating within the atmosphere. Weekly moving averages were applied to multiple variables to capture temporal smoothing and remove high-frequency noise. Each of feature implemented into the final model reduced spurious correlation between co-variates by combining their covariance into a representation of O3 based on transport and chemical mechanisms. The second law of classical mechanics by Issac Newton (F=mass\*acceleration; (Taylor 2005) can be applied to bonds between molecules; it is equal to the amount of heat an object generates over some distance, known as work (W=Fdx) (Borgnakke and Sonntag 2014). This section details the overall extraction process, feature transformations, and thermodynamics of chemical states used in the final algorithm.

III.5.1. Toms/Omi Dobson Unit (Du) Conversion to Surface Representation

One Dobson Unit is the number of molecules of ozone that would be required to create a layer of pure ozone 0.01 millimeters thick at a temperature of 0 degrees Celsius and a pressure of 1 atmosphere. Ground-based O3 measurements taken from the EPA monitors are represented in parts-per-million (ppm). As stated in the Data Sources section, the TOMS/OMI satellite total column estimates are in Dobson Units (DU). 1 Dobson Unit would contain about 2.69x1016 ozone molecules for every square centimeter of area at the base of the column. For an array O3 values in DU where 1 mole of Ozone is 47.9982 grams:

As parts-per-million (ppm) is a dimensionless quantity used to relate weight to volume, this rendition of DU requires some form of divisible meter metric to be an accurate representation of surface ozone at some location. It’s quite convenient as a geographer to have access to a representative surface that can be stated in terms of x, y, and z. If one liter (L) can be represented as 0.0001m3, the ppm representation of the Dobson Unit per m2 can be stated as:

Where the vector is the monitor value of TCO in Dobson Units. Using this transformation, the TCO measurement from TOMS/OMI was converted to a surface representation of ozone in . This was used to scale the DU down to a comparable measurement with Sentel 5-p Imagery.

III.5.2. Kinetic Energy Representations

Briefly switching into the realm of chemists and their amazing ability to provide concrete evidence of physical processes without the mathematical relationship via a tool called observation, chemical reactions always happen at a given rate. Calculated temperatures, concentrations, and unique combinations of constituents can be observed to do many different things, such as ice melting to water, then evaporating to vapor. Chemically, this is all H2O, simply in different thermodynamical states of a solid, liquid, and gas. As simplistically as possible (assuming no container (volume (V(L)(fluid + gas)) = 0), this can be mathematically represented as:

As mentioned in Chapter II.2, surface ozone reacts with certain VOC’s and emissions, oxidizing organic and inorganic material and causing what we know as surface O3. Chapter II.3.1 showed that the remaining parts of this pseudo-catalyst are typically preserved and reintegrated into the atmosphere if given enough energy overtime. This has many different chemical representations; however, this thesis mainly focuses on nitrous oxides and carbonous aerosols as these are two of the most chemically influential chemicals of surface O3 production (Cheng et al. 2018; Li et al. 2024; Richter 2009). Chapter II.3.3 revealed that carbon and nitrous based VoC emissions impact surface-level air quality, tropospheric photochemistry, and climate feedback mechanisms in addition to sustaining surface ozone concentrations:

CO + OH = O2 + Q(E) > CO2 + HO

NO2 + O2 + Q(E) = NO + O3 + Q(E) = NO2 + O2+ Q(E)

Chapter 2 revealed that carbon and nitrous based VoC emissions impact surface-level air quality, tropospheric photochemistry, and climate feedback mechanisms in addition to sustaining surface O3 reactions. These chemical relations are depictions of the sub-photosynthesis process, photolysis which has been found to drive O3 (C. Chen et al. 2021; He and Carmichael 1999; Shah et al. 2024). In more thermodynamical terms, it is an energy-based oxidation event. If a constituent is required to produce O3 and UV rays are necessary for ozone production, then it must exist in a constant reactionary state. The first law of thermodynamics states that matter can neither be created nor destroyed. This can be seen with the remaining O- that tend to titrate with VoCs. Atmospheric cycles attempt to rebalance the amount of newly seen negative oxygen molecules, in which energy is lost. This energy is measured by way of remote sensing and can theoretically be represented through a series of proper transformations.

III.5.3. A Basic Implementation of Thermodynamics

O3 is comprised of three oxygen (O-) molecules. Known as a triatomic molecule, if one mol of oxygen is roughly 15.999 grams, then 1 mole of Ozone is about 47.9982 grams. Chemicals can only be comprised of a certain number of elements; there is a limitation to its state before it’s something else entirely (i.e, H2O vs. H2O2 or water versus hydrogen peroxide). As air pollution exhibits a system of finite degrees of freedom, estimated kinetic energy must be representative of the chemical and physical properties of constituents. The basic take away from this is that theoretical kinetic energies of any reaction at a given state, in some space, at some time, can be represented as a combination of chemical-thermodynamics and derivations of the Ideal Gas Law.

For later use with a scaled field correlating to the predictor variable, the kinetic energy for a potential geo-atom could be represented as a combination of spatial pressure, temperature, and volume dynamics of a system. The proposed average translational kinetic energy for a geo-atom z(x,y) is started as:

Where at some time t, the estimated KE of a geo-atom can be represented as the and temperature () of the constituent and kb, Boltzmann’s Constant. The short of this theory is that on average, atoms exist with a thermal energy of , allowing the ideal gas law to be re-written as:

Where is pressure, is replace in lieu of the constant as a measurement of energy, R is the gas constant estimation used in the Boltzmann’s constant, and N(x,y) is number of O3 molecules estimated by remotely sensed imagery.

III.6. Interpolation and Imputation Strategies

Linear, modified 2-Dimensional Akima, 1-Dimensional Akima, and polynomials with degrees 0,1,2 and 3 were evaluated using the same RMSE, MAE, and MSE strategies as model evaluation methods:

* + Selecting all available monitor values
  + Randomly changing 10% of these values to Nan in python
  + Establishing trend with missing monitors
  + Calculating error statistics, results seen in Table 2
  + Placing known values back into the dataset
  + Predicting with the established trend for unknown values

The R2 correlation coefficient was added to interpolation-based methods, but not to regression-based imputation methods due it not being as valuable for low error in small predictions. A time-series based KNN imputation strategy was also developed for complex features missing in HD for further work on this project. Using the mean value between the best correlation methods, missing values were estimated with high precision and accuracy. Modified 2D Akima represents a moving piece-wise polynomial approximation over the given periods (Akima 1970). The trend estimated from training on replaced missing values was then used on the full data set, inclusive of the missing values taken out for KNN training. Resulting errors during training holds true for all tested values (Beretta and Santaniello 2016; Liao et al. 2014) when applied to the final dataset.

III.6.1. Missing Daily Raster Data

Some of the datasets only had either yearly averages/estimations before 2000 or required a set amount of time to pass before estimation of the variable became possible. For example, nighttime light imagery is only available with daily resolution from 2013 on, though it is available yearly via another source. NDVI imagery is subject to 16-day intervals to effectively “see” the change in wavelengths each source of vegetation is producing. To better account for daily changes, a function was created to calculate the difference between available temporal averages and daily differences to produce estimated daily changes. The function is simply a linear representation between the available data and number of days missing:

Where is a feature raster with a value, longitude, and latitude. is the difference in days from t1 to tmax:

Once the average daily change was calculated, a new, estimated disaster comprised of the summation of the first date and the average daily change can be:

While this dataset is bound to incur slightly repetitive values across monitor locations, this method allowed for the incorporation of a date-like index into the missing values.

For missing raster data, the time of each available raster was extracted. The difference between available raster data was divided by the number of missing days in between the extracted times. The resulting mean was added to the first raster, producing a new raster for each day missing in between the intervals. This had interesting effects on the outcome, working for features used in this these, but not so much for Nighttime light imagery. While the imputation of new rasters from a pre-existing set of models would have been ideal, this level of work would best be suited for a dissertation. This was only meant to introduce daily estimates into the algorithm to incorporate feature trends into each model.

III.7. Statistical Model and Residual Kriging Methodology

The Statistical Modeling and Residual Kriging (SMaRK) method is proposed as a methodology to exploit the flexibility of kriging with complex solutions from modern day ensembles to establish a geo-field via satellite and monitoring systems for use in geo-atoms comprised of satellite data. Chapter II found that nonparametric, decision tree-based, predictive models have been increasingly adopted in spatial statistical analysis due to their flexibility in handling heterogeneous covariates (Liu et al. 2018; Mu et al. 2023; Q. Pan, Harrou, and Sun 2023). Many of the selected features during each training process showed covariance (i.e. specific humidity, temperature created features among others), ML/AI methods to minimize the influence of similar trends were required to define the overall property for each pixel. The SMaRK methodology is comprised of five main steps:

1. Statistically model a trend at known points given *n,*
2. Measure error at known locations given ***f****():*
3. Apply RK to estimate an error trend at given : -
4. Reincorporate estimated error into the established trend.
5. Calculate the result as stationary pixels:

The resulting surface depicting O3 properties was produced at a resolution of 250m for each day within the selected time range. The underlying spatial concepts in SMaRK follow the establishment of initial geo-atoms at monitoring locations based on predicted trends from a chosen ensemble. The full definition of the geo-atom from (Goodchild, Yuan, and Cova 2007) in terms of statistical models and residually kriged uncertainty to represent a value for surface O3 geo-atoms in PHOTUC on a given day can be depicted as:

=

Where the definition of the property Z at some latitude and longitude and time is , established from the estimated complex trend and associated error relative to each monitor location. An overarching theory from some complex trend is corrected by the laws associated with in-situ measurements, making the value of the system estimated from some ***f***().

III.7.1. Model Parameters and Tuning

This thesis mainly overviews the sklearn package in Python; offering the most simplistic forms of modern complex analysis (Buitinck et al. 2013) at a CPU processing level. Implementations of PyTorch, Keras, and Tenserflow utilizing GPU components are saved for future work. The novelties within sklearn’s development schemes have allowed for basic integration of the latter packages as a rough start to advanced ML/AI modelling enhancements. XGBOOST is developed as a separate entity and was able to be utilized with a GPU in the middle stages of this project. While computations times are exponentially increased for this thesis, further incorporation of these methods would yield drastically reduced computation times.

In this study, models were trained using RandomSearchCV (RSCV) in Python. The search algorithm employed a parameter list derived from prior known distributions of the data. The initial dataset was reduced from approximately 2,760,000 to 656,000 rows in the 2018–2024 timeframe. Keeping reproducibility in mind, each model was trained with a unique parameter combination, ensuring comparability to the data size and preventing overfitting or underfitting. After training, residual values were kriged using geospatial monitor positions and associated residual values, following methodologies from similar works e.g (Anand et al. 2021; Liu et al. 2018; Oliver and Webster 2014). A summary of the five numerical models and their potential parameters randomly selected in RSCV is as follows:

1. Adaptive Boosted Sampling (ADABOOST): The learning rate was fine-tuned between 0.00001 to 1, with step values reflecting half the number of features. The number of estimators ranged from 50 to 500, with similar step adjustments. Loss calculations considered linear, square, or exponential trends. This model was chosen for its capacity to reduce variance and enhance accuracy via minimal tuning. While the computation time for ADABOOST is relatively quick, it is expected to yield high error as the most simplistic numerical model chosen in this project.
2. Gradient Boosting (GB): Parameters such as loss functions (squared error, huber), learning rate, maximum depth, and the number of estimators were adjusted. This model was included due to its effectiveness in minimizing loss and handling complex datasets.
3. Extreme Gradient Boosting (XGBoost): XGBoost was optimized using grow policies (depthwise, lossguide), learning rate adjustments, and regularization parameters. It was selected for its speed and high performance, especially with large datasets.
4. Random Forest (RF): This model utilized different criteria (squared error, absolute error, friedman\_mse) and was tuned for maximum depth and the number of estimators. Random Forest was included for its robustness and effectiveness in handling high-dimensional data.
5. Multi-Layered Perceptron (MLP): The MLP model's hidden layer sizes, learning rate, and other parameters were adjusted for optimal tuning. It was chosen for its flexibility in capturing non-linear relationships within the data.

After tuning, the in-situ value from the monitoring locations was used to calculate a residual which was predicted with a geospatial krige of the same parameters. The more complex the ensemble, the more simplistic the error trends are, especially within densely packed areas.

III.7.2. Residual Krige Tuning

In geo-statistics (Chiles and Delfiner, 2009), regression kriging (RK) (Hengl et al., 2007; Hengl, 2007) represents a hybrid approach for spatial prediction combining a statistical regression of some independent variable based on dependent variables for trend estimation, and a kriging term of the regression residuals representing spatial dependence. As a representation of geo-atoms, this can be amended to the overall geo-field associated with its property. For any location **s** with a vector of observations **x**(**s**), the RK estimator is written as:

Where is the regression term for the complex trend defined by the estimated parameters . The residual is spatially auto correlated and follows a selected variogram model. The RK method allows for various regression methods for use as the base trend estimation, making it suitable for leveraging complex trends estimated by ML/AI methods. The parameters for potential variograms are as follows:

* Variogram model: linear, spherical, gaussian
* Pseudo-inverse: False, True
* Pseudo-inverse type: None, pinvh
* Drift terms: function, specified, point
* Number of lags: 2, 4, 6
* Functional drift: Fourier Series, Elevation, Sinusoidal

Each statistical model was fit using the ground O3 concentration **z** as the dependent variable. The previously mentioned geographic datasets and feature transformations were used as predictors at each point. Given V1 as the first feature used in training, and **x** as the remaining features used in training:

For every O3 measurement at each monitoring station **z**(), there exists a corresponding residual which is compared to the associated estimation of the fitted ensemble. This can be used for interpolation due to inherit relationships between the points, properties, and values in each geo-atom. The residual term, can be obtained by kriging with residuals at N monitoring stations:

where the kriging weights can be represented as a matrix comprised of:

Where **C** is the covariance specified by among O3 monitor stations to account for the spatial dependence. The covariance vector, is the spatial dependence between some geo-atom in the geo-field, , and its associated complex trend with accounted drift. Tested drift functions included a sinusoidal estimation between latitude and the dependent variable, point specified drift between elevation and a Fourier series representation of the latitude, longitude, and maximum value to represent spatial drift at some point in time. This process allows for the incorporation of many trends into RK analysis estimations for a better representation of surface O3 based on spatially dependent in-situ values.

III.8. Model Evaluation Techniques

Models were selected based on numerous factors. The mean absolute percent error (MAPE) and root mean standard error (RMSE) were mainly utilized for comparison as studies prefer their use of units relative to the dataset (Hu et al. 2023; Mu et al. 2023; Q. Pan, Harrou, and Sun 2023; A. Zhang et al. 2023). Mean squared error (MSE), mean absolute error (MAE) and overall percent error were also calculated. Though not all metrics were required (e.g. RMSE is simply the square root of MSE so utilizing MSE seems redundant), utilizing all error metrics revealed stronger insights into the improvements made by including geospatial uncertainty. The best model operating with or without RK improvements was applied to the full dataset for prediction of the final rasters. RMSE and Mean Absolute Percent Error (MAPE) calculations along with statistical distributions of the outcome are used for the main determinants in model selection. Finally, the R2 correlation coefficient was used to represent the amount of variance in dependent variables that can be explained by the independent variable in a regression model. The calculation of R2 allowed for a better interpretation of SHAP and overall feature importance within each model. These error metrics were also used for the evaluation of interpolation and imputation of monitor values.

III.9. Conclusion

All calculations were done in ppm then later converted to ppb for a better representation of differences when writing. Models were tested with combinations of collected features to determine if data starting in 1980 can be utilized to make historical models as effective as those with modern data. Features were selected based on their correlation with respect to daily surface O3 values at monitoring locations, associated literature and learned knowledge through coursework. Initially, cutoffs for the number of features included in each model were determined by the ranking of variables across each of the three correlation equations, Pearson, Spearman, and Kendall. Upon final review, the ranking of feature variables before implementation into the model was used to select the best variables for inclusion in each dataset depicted in Chapter IV. Only one dataset used a cutoff of the top 24 features across Pearson correlations.

The first set of features contained the best correlating variables associated with imagery from GridMET and weekly moving averages. The second set of features depicts variables found in modern imagery and anthropogenic sources. The third set consists of energy, CTM and meteorological representations most seen in literature from Chapter II. The final set contains the top 24 correlating features out of all the datasets combined, essentially combining all variables in the first three sets with a few lower correlated variables which were also key representations of surface O3 mentioned in Chapter II. Table V.2 in Chapter V depicts the final selection of variables for each model and timeframe. Each set of variables contained metrological and/or domain inspired feature transformations and the top three correlating seasons, omitting the fourth season for use as a baseline category.

A raster dataset was created at a 500m spatial resolution from the imagery in GEE of the following geographic variables provided by the sources mentioned in Chapter IV. The stacked result was resampled to 250m, used as a surface for prediction given the best model, and added to a 250m resolution RK grid created by modeled uncertainty around the monitoring stations for display of the final daily images. Five ML/AI ensembles in total were used to find a trend yielding the least complex error for RK trend establishment. Overall predictive error was further reduced by adding the geospatial uncertainty of the selected features into the chosen ensemble. Even if the initial numerical model yielded no error, the RK method incorporated this and only enhanced areas which are further from the truth value. A case study of these methods is conducted over PHOTUC further separating the average results into county-based results.

Citations

Abdullah, Ahmad Makmom, Marzuki Ismail, Fong Si Yuen, Samsuri Abdullah, and Rasheida E. Elhadi. 2017. “The Relationship between Daily Maximum Temperature and Daily Maximum Ground Level Ozone Concentration.” *Polish Journal of Environmental Studies* 26: 517–23.

Abdullah, Samsuri, Najihah Husna Ahmad Nasir, Marzuki Ismail, Ali Najah Ahmed, and Mohamad Nor Khasbi Jarkoni. 2019. “Development of Ozone Prediction Model in Urban Area.” *International Journal of Innovative Technology and Exploring Engineering* 8 (10). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP: 2263–67. doi:10.35940/ijitee.j1127.0881019.

Akima, Hiroshi. 1970. “A New Method of Interpolation and Smooth Curve Fitting Based on Local Procedures.” *Journal of the ACM* 17 (4). Association for Computing Machinery (ACM): 589–602. doi:10.1145/321607.321609.

Anand, Akash, Prachi Singh, Prashant K. Srivastava, and Manika Gupta. 2021. “GIS-Based Analysis for Soil Moisture Estimation via Kriging with External Drift.” In *Agricultural Water Management*, 391–408. Elsevier. doi:10.1016/b978-0-12-812362-1.00019-9.

Beretta, Lorenzo, and Alessandro Santaniello. 2016. “Nearest Neighbor Imputation Algorithms: A Critical Evaluation.” *BMC Medical Informatics and Decision Making* 16 (S3): 74. doi:10.1186/s12911-016-0318-z.

Borgnakke, Claus, and Richard Edwin Sonntag. 2014. *Fundamentals of Thermodynamics*. 8. ed., SI version. Hoboken, NJ: Wiley.

Buitinck, Lars, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, et al. 2013. “API Design for Machine Learning Software: Experiences from the Scikit-Learn Project.” In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–22.

Cao, Guofeng. 2022. “Deep Learning of Big Geospatial Data: Challenges and Opportunities.” In *New Thinking in GIScience*, by Bin Li, A-Xing Zhu, Hui Lin, Xun Shi, and Cuizhen Wang, 159–69. Singapore: Springer Nature Singapore. doi:10.1007/978-981-19-3816-0\_18.

Chen, Bin, Yixuan Wang, Jianping Huang, Lin Zhao, Ruming Chen, Zhihao Song, and Jiashun Hu. 2023. “Estimation of Near-Surface Ozone Concentration and Analysis of Main Weather Situation in China Based on Machine Learning Model and Himawari-8 TOAR Data.” *Science of The Total Environment* 864 (March): 160928. doi:10.1016/j.scitotenv.2022.160928.

Chen, CH, TF Chen, SP Huang, and KH Chang. 2021. “Comparison of the RADM2 and RACM Chemical Mechanisms in O3 Simulations: Effect of the Photolysis Rate Constant.” *SCIENTIFIC REPORTS* 11 (1). doi:10.1038/s41598-021-84629-4.

Cheng, Y, YH Wang, YZ Zhang, JH Crawford, GS Diskin, AJ Weinheimer, and A Fried. 2018. “Estimator of Surface Ozone Using Formaldehyde and Carbon Monoxide Concentrations Over the Eastern United States in Summer.” *JOURNAL OF GEOPHYSICAL RESEARCH-ATMOSPHERES* 123 (14): 7642–55. doi:10.1029/2018JD028452.

Dangermond, Jack, and Michael F. Goodchild. 2020. “Building Geospatial Infrastructure.” *Geo-Spatial Information Science* 23 (1). Informa UK Limited: 1–9. doi:10.1080/10095020.2019.1698274.

De Marco, Alessandra, Hector Garcia-Gomez, Alessio Collalti, Yusef Omidi Khaniabadi, Zhaozhong Feng, Chiara Proietti, Pierre Sicard, Marcello Vitale, Alessandro Anav, and Elena Paoletti. 2022. “Ozone Modelling and Mapping for Risk Assessment: An Overview of Different Approaches for Human and Ecosystems Health.” *Environmental Research* 211 (August): 113048. doi:10.1016/j.envres.2022.113048.

EPA, Environmental Protection Agency. 2025. “Air Quality System (AQS) API.” Table. https://www.epa.gov/outdoor-air-quality-data/download-daily-data: Air Quality Data.

EPA, US. 2013. “Final Report: Integrated Science Assessment of Ozone and Related Photochemical Oxidants.” *US Environmental Protection Agency, Washington, DC*.

European Centre For Medium-Range Weather Forecasts. 2019. “ERA5 Reanalysis (0.25 Degree Latitude-Longitude Grid).” netCDF4. UCAR/NCAR - Research Data Archive. doi:10.5065/BH6N-5N20.

Goodchild, Michael F. 2004. “The Validity and Usefulness of Laws in Geographic Information Science and Geography.” *Annals of the Association of American Geographers* 94 (2). Informa UK Limited: 300–303. doi:10.1111/j.1467-8306.2004.09402008.x.

Goodchild, Michael F., May Yuan, and Thomas J. Cova. 2007. “Towards a General Theory of Geographic Representation in GIS.” *International Journal of Geographical Information Science* 21 (3): 239–60. doi:10.1080/13658810600965271.

He, Shan, and Gregory R. Carmichael. 1999. “Sensitivity of Photolysis Rates and Ozone Production in the Troposphere to Aerosol Properties.” *Journal of Geophysical Research: Atmospheres* 104 (D21): 26307–24. doi:10.1029/1999JD900789.

Hsu, CY, RQ Lee, PY Wong, SCC Lung, YC Chen, PC Chen, G Adamkiewicz, and CD Wu. 2024. “Estimating Morning and Evening Commute Period O3 Concentration in Taiwan Using a Fine Spatial-Temporal Resolution Ensemble Mixed Spatial Model with Geo-AI Technology.” *JOURNAL OF ENVIRONMENTAL MANAGEMENT* 351 (February). doi:10.1016/j.jenvman.2023.119725.

Hu, JT, YY Chen, W Wang, SC Zhang, C Cui, WK Ding, and Y Fang. 2023. “An Optimized Hybrid Deep Learning Model for PM2.5 and O3 Concentration Prediction.” *AIR QUALITY ATMOSPHERE AND HEALTH* 16 (4): 857–71. doi:10.1007/s11869-023-01317-0.

Ilić, Predrag, Zoran Popović, and Dragana Ne\vsković Markić. 2020. “Assessment of Meteorological Effects and Ozone Variation in Urban Area.” *Ecological Chemistry and Engineering S* 27: 373–85.

Li, YS, TJ Wang, QG Wang, MM Li, YW Qu, H Wu, and M Xie. 2024. “Exploring the Role of Aerosol-Ozone Interactions on O3 Surge and PM2.5 Decline during the Clean Air Action Period in Eastern China 2014-2020.” *ATMOSPHERIC RESEARCH* 302 (June). doi:10.1016/j.atmosres.2024.107294.

Liao, Serena G, Yan Lin, Dongwan D Kang, Divay Chandra, Jessica Bon, Naftali Kaminski, Frank C Sciurba, and George C Tseng. 2014. “Missing Value Imputation in High-Dimensional Phenomic Data: Imputable or Not, and How?” *BMC Bioinformatics* 15 (1): 346. doi:10.1186/s12859-014-0346-6.

Liu, Ying, Guofeng Cao, Naizhuo Zhao, Kevin Mulligan, and Xinyue Ye. 2018. “Improve Ground-Level PM2.5 Concentration Mapping Using a Random Forests-Based Geostatistical Approach.” *Environmental Pollution* 235 (April): 272–82. doi:10.1016/j.envpol.2017.12.070.

Miller, Harvey J., and Michael F. Goodchild. 2015. “Data-Driven Geography.” *GeoJournal* 80 (4): 449–61. doi:10.1007/s10708-014-9602-6.

Mu, X, SC Wang, P Jiang, and YL Wu. 2023. “Estimation of Surface Ozone Concentration over Jiangsu Province Using a High-Performance Deep Learning Model.” *JOURNAL OF ENVIRONMENTAL SCIENCES* 132 (October): 122–33. doi:10.1016/j.jes.2022.09.032.

Napi, NNLM, S Abdullah, AA Mansor, AN Ahmed, and M Ismail. 2021. “Development Of Models For Forecasting Of Seasonal Ground Level Ozone (O3).” *Journal Of Engineering Science And Technology* 16 (4): 3136–54.

Oliver, M.A., and R. Webster. 2014. “A Tutorial Guide to Geostatistics: Computing and Modelling Variograms and Kriging.” *CATENA* 113 (February). Elsevier BV: 56–69. doi:10.1016/j.catena.2013.09.006.

Pan, Qilong, Fouzi Harrou, and Ying Sun. 2023. “A Comparison of Machine Learning Methods for Ozone Pollution Prediction.” *Journal of Big Data* 10 (1). Springer Science and Business Media LLC. doi:10.1186/s40537-023-00748-x.

Pan, S, Y Choi, A Roy, and W Jeon. 2017. “Allocating Emissions to 4 Km and 1 Km Horizontal Spatial Resolutions and Its Impact on Simulated NOx and O3 in Houston, TX.” *ATMOSPHERIC ENVIRONMENT* 164 (September): 398–415. doi:10.1016/j.atmosenv.2017.06.026.

Pridal, Petr, Tomas Pohanka, Ali Ashraf, and Radim Kacer. 2014. “EPSG.Io: Coordinate Systems Worldwide.” *EPSG.Io From MapTiler Team*. https://epsg.io/32612.

Raschka, Sebastian, and Vahid Mirjalili. 2019. *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow 2*. Third edition. Expert Insight. Birmingham Mumbai: Packt.

Richter, A. 2009. “Nitrogen Oxides in the Troposphere – What Have We Learned from Satellite Measurements?” *The European Physical Journal Conferences* 1: 149–56. doi:10.1140/epjconf/e2009-00916-9.

Rouault, Even, Frank Warmerdam, Kurt Schwehr, Andrey Kiselev, Howard Butler, Mateusz Łoskot, Tamas Szekeres, et al. 2025. “GDAL.” Zenodo. doi:10.5281/ZENODO.15375292.

Schultz, Martin G, Sabine Schröder, Olga Lyapina, Owen R Cooper, Ian Galbally, Irina Petropavlovskikh, Erika von Schneidemesser, et al. 2017. “Tropospheric Ozone Assessment Report, Links to Global Surface Ozone Datasets.” Application/zip. PANGAEA. doi:10.1594/PANGAEA.876108.

Shah, Viral, Christoph A. Keller, K. Emma Knowland, Amy Christiansen, Lu Hu, Haolin Wang, Xiao Lu, Becky Alexander, and Daniel J. Jacob. 2024. “Particulate Nitrate Photolysis as a Possible Driver of Rising Tropospheric Ozone.” *Geophysical Research Letters* 51 (5): e2023GL107980. doi:10.1029/2023GL107980.

Staehle, Christoph, Harald E. Rieder, and Arlene M. Fiore. 2023. “Technical Note: An Assessment of the Performance of Statistical Bias Correction Techniques for Global Chemistry-Climate Model Surface Ozone Fields.” *EGUsphere*, November, 1–21. doi:10.5194/egusphere-2023-2743.

Taylor, John R. 2005. *Classical Mechanics*. Nachdr. Sausalito, Calif: University Science Books.

Travis, Katherine R., and Daniel J. Jacob. 2019. “Systematic Bias in Evaluating Chemical Transport Models with Maximum Daily 8 h Average (MDA8) Surface Ozone for Air Quality Applications: A Case Study with GEOS-Chem v9.02.” *Geoscientific Model Development* 12 (8): 3641–48. doi:10.5194/gmd-12-3641-2019.

Turner, Michelle C., Michael Jerrett, C. Arden Pope, Daniel Krewski, Susan M. Gapstur, W. Ryan Diver, Bernardo S. Beckerman, et al. 2016. “Long-Term Ozone Exposure and Mortality in a Large Prospective Study.” *American Journal of Respiratory and Critical Care Medicine* 193 (10). American Thoracic Society - AJRCCM: 1134–42. doi:10.1164/rccm.201508-1633OC.

United States Census Bureau. 2022. “TIGER/Line Shapefiles (Machine Readable Data Files).” Department of Commerce. https://www.census.gov/cgi-bin/geo/shapefiles/index.php.

———. 2025. “Arizona - Census Bureau Profile.” *Arizona*. https://data.census.gov/profile/Arizona?g=040XX00US04.

U.S. Geological Survey. 2024. “1 Meter Digital Elevation Models (DEMs) - USGS National Map 3DEP Downloadable Data Collection | USGS Science Data Catalog.” National Geospatial Technical Operations Center. https://data.usgs.gov/datacatalog/data/USGS:77ae0551-c61e-4979-aedd-d797abdcde0e.

Wang, Meng, Paul D. Sampson, Jianlin Hu, Michael Kleeman, Joshua P. Keller, Casey Olives, Adam A. Szpiro, Sverre Vedal, and Joel D. Kaufman. 2016. “Combining Land-Use Regression and Chemical Transport Modeling in a Spatiotemporal Geostatistical Model for Ozone and PM2.5.” *Environmental Science & Technology* 50 (10). American Chemical Society (ACS): 5111–18. doi:10.1021/acs.est.5b06001.

Wang, Y, QQ Yuan, TW Li, LY Zhu, and LP Zhang. 2021. “Estimating Daily Full-Coverage near Surface O3, CO, and NO2 Concentrations at a High Spatial Resolution over China Based on S5P-TROPOMI and GEOS-FP.” *ISPRS JOURNAL OF PHOTOGRAMMETRY AND REMOTE SENSING* 175 (May): 311–25. doi:10.1016/j.isprsjprs.2021.03.018.

Ye, XP, XL Wang, and L Zhang. 2022. “Diagnosing the Model Bias in Simulating Daily Surface Ozone Variability Using a Machine Learning Method: The Effects of Dry Deposition and Cloud Optical Depth.” *ENVIRONMENTAL SCIENCE & TECHNOLOGY* 56 (23): 16665–75. doi:10.1021/acs.est.2c05712.

Zhang, AX, TM Fu, X Feng, JF Guo, CF Liu, JK Chen, JJ Mo, et al. 2023. “Deep Learning-Based Ensemble Forecasts and Predictability Assessments for Surface Ozone Pollution.” *GEOPHYSICAL RESEARCH LETTERS* 50 (8). doi:10.1029/2022GL102611.

Zhang, Wenxiu, Di Liu, Hanqin Tian, Naiqin Pan, Ruqi Yang, Wenhan Tang, Jia Yang, et al. 2022. “Recurrent Mapping of Hourly Surface Ozone Data (HrSOD) across China during 2005-2020 for Ecosystem and Human Health Risk Assessment.” *Earth System Science Data Discussions*, December. Copernicus Gesellschaft mbH, 2–36. doi:10.5194/essd-2022-428.