Chronic Respiratory Disease Risk: Modeling Potential and Limitations

Abstract*—*Chronic Respiratory Diseases (CRDs), including Chronic Obstructive Pulmonary Disease (COPD) and asthma, are among the main causes of mortality worldwide, affecting 545 million people in 2017. Symptoms of noninfectious CRDs are mainly exacerbated by ambient air pollution, as well as changes in temperature and humidity. In this study, we explore a novel application of machine learning to forecast CRD risk and discuss its merits and limitations. We developed, trained, and tested a Random Forest regressor using datasets over the United States during 2000-2016, with mortality rate as the target variable. The final random forest regressor produced R-squared values of 0.7526 and 0.7528 for cross-validation and test dataset predictions, respectively, implying that our model generalizes well out-of-sample. The selected features comprise of location and temporal encoders, population density, and net primary production. The study reveals significant potential for modeling CRD risk, but highlights setbacks due to the primarily noninfectious nature of CRDs, phenomena only identifiable on finer spatiotemporal scales, and data limitations. We also identify methods that may refine our approach and describe future developments that may improve CRD risk modeling.

Keywords*—*chronic respiratory disease, regression, remote sensing, climate variables, carbon emissions, particulate matter, machine learning, random forest

I. Introduction

According to the Global Burden of Diseases Study (GBD) reports for 2017, there were an estimated 545 million prevalent cases and 62 million incident cases (new cases in 2017) of Chronic Respiratory Diseases (CRDs), the vast majority of which are due to Chronic Obstructive Pulmonary Disease (COPD) and asthma (Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017). In terms of mortality, there were 3.2 million deaths from COPD and 495,000 deaths from asthma (Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017).

Tobacco, air pollutants, airborne allergens, and occupational exposures are the main risk factors associated with lifetime increase of respiratory diseases and symptoms, according to reviews (Definition, epidemiology and natural history of COPD, Allergy and asthma: Effects of the exposure to particulate matter and biological allergens). Previous studies have also established that temperature and humidity play a role in respiratory disease exacerbation (The impact of cold on the respiratory tract and its consequences to respiratory health, Synergistic effects of temperature and humidity on the symptoms of COPD patients).

Furthermore, the World Health Organization emphasizes the potentially wide-ranging health impacts of climate change, with many being indirect (WHO | Preventing disease through healthy environments), such as increase of wildfire emissions and longer pollen seasons. However, due to limited datasets on a national, county-level scale, we are constrained to environmental factors, particularly air pollutants and climate variables, provided as datasets constructed with remote sensing.

The present study investigates the potential and limitations for a novel application of Machine Learning (ML) analyses with climate variables and indices, fire emissions, and particulate matter to predict mortality rates attributed to Chronic Lower Respiratory Diseases (CLRDs) in the contiguous United States during 2000-2016. In particular, we used random forest regression, performed feature selection and hyperparameter tuning with Cross-Validation (CV), and measured model performance with the coefficient of determination (R-squared, R^2) as the scoring metric. Finally, we draw insights from our approach, discuss its limitations, and describe future developments that may benefit the modeling of CRD risk.

II. Materials

The datasets included at the time of the present study overlap for the years 2000-2016, thus determining the period of interest. All datasets we describe are available in monthly temporal resolution unless otherwise specified.

A. Underlying Cause of Death – Chronic Lower Respiratory Diseases

The Underlying Cause of Death data available on CDC WONDER is county-level national mortality and population data, beginning from 1999. The data compilates death certificates for U.S. residents. Each death certificate identifies a single underlying cause of death and demographic data (Underlying Cause of Death 1999-2019). The data is subject to suppression constraints that omit all sub-national data representing less than 10 deaths.

We extracted death counts for monthly, county-level deaths caused by CLRDs from the request form for Underlying Cause of Death (Underlying Cause of Death, 1999-2019 Request), and calculated mortality rates per 100,000 people. To interpolate the suppressed data for counties in a given state and month, we subtract the sum of reported county deaths from the state total, then calculate mortality rate adjusted by county population. In doing so, we apply the assumption that mortality rates are equal between the counties with suppressed data.

CLRDs include emphysema, asthma, bronchiectasis, and other COPDs (ICD-10-CM Section J40-J47). We chose CLRDs over asthma since the asthma data is overly suppressed, with data available for very few counties. We chose them over respiratory diseases as a whole (ICD-10-CM

Chapter 10), which includes infectious respiratory diseases. CLRDs are more directly exacerbated by air pollution and climate variables, unlike influenza, pneumonia, and other respiratory infections. However, despite being not primarily infectious in etiology, CRDs have some aspects of their pathogenesis influenced by infectious organisms (Infections in "Noninfectious" Lung Diseases). This is an active area for investigation and outside the scope of the present study.

B. *Fine Particulate Matter*

Global estimates of fine particulate matter (PM2.5) concentrations, available at https://sites.wustl.edu/acag/datasets/surface-pm2-5/, are developed using advances in satellite observations, chemical transport modeling, and ground-based monitoring (Global Estimates and Long-Term Trends of Fine Particulate Matter Concentrations). We extracted the data files for monthly North American Regional Estimates (V4.NA.03).

C. *Global Fire Emissions Database*

The Global Fire Emissions Database (GFED), available at https://globalfiredata.org/pages/data/, combines satellite information on fire activity and vegetation productivity to estimate gridded monthly burned area and fire emissions (Global fire emissions estimates during 1997–2016). The fourth version has several modifications from the previous version and uses higher quality input datasets. A notable upgrade is the inclusion of contributions from small fires. The GFED layers included in the present study are: fraction of area that burned (burned_frac), fraction of total emissions stemming from small fires (smallf_frac), total emissions measured in carbon (C) and dry matter (DM), net primary production (NPP), heterotrophic respiration (R_h), and fire emissions (BB). The GFED authors use 'biosphere fluxes' as the umbrella term referring to NPP, R_h, and BB.

D. *Air Quality Index*

Air Quality Index (AQI) data for the United States is available at https://aqs.epa.gov/aqsweb/airdata/download_files.html. We opted to use AQI instead of individual criteria pollutants, such as ground-level ozone; AQI was less spatially and temporally sparse and better for interpolation. For each site, we linearly interpolated dates with missing data using pandas.DataFrame.interpolate, then averaged by month. We then interpolated and rasterized AQI at 0.01° resolution using scipy.interpolate.griddata, before aggregating to county shapes. To choose the best interpolation parameters, we performed CV with R^2 to test the number of consecutive missing dates to interpolate between existing dates and the spatial interpolation method (linear, nearest, or cubic). Finally, we use zeros to replace months with incomplete interpolation, which was the best imputation strategy based on CV.

Due to the sparsity of the data, we considered the use of AQI experimental, as naive spatiotemporal interpolation is inherently erroneous. For this reason, we expected to eliminate AQI during feature selection.

E. *Population and Median Income*

Yearly estimates for population by county are available at https://www.census.gov/data/datasets.html under the "Population Estimates" filter. Population estimates of each year are for July 1st. We opted to apply the July 1st estimates for the entire month of July, then linearly interpolate the months in between.

Yearly estimates for median income by county are available at https://www.census.gov/programs-surveys/saipe/data/datasets.html. We applied the yearly values for all months in the calendar year, as no month or day is specified. Given this loss of temporal resolution, the use of median income was considered experimental. Thus, we expected to eliminate median income during feature elimination, though we may gain some insights during collinearity analysis.

F. *Climate Variables and Indices*

Monthly data for climate variables and indices is available from NOAA's Climate Divisional Database (nClimDiv) (NOAA's Climate Divisional Database (nCLIMDIV)). We extracted data for temperature, precipitation, and drought indices to best represent environmental factors described in (The impact of cold on the respiratory tract and its consequences to respiratory health, Synergistic effects of temperature and humidity on the symptoms of COPD patients, Adverse effects of increasing drought on air quality via natural processes, Drought-Induced Reduction in Global Terrestrial Net Primary Production from 2000 Through 2009). The included drought indices are Palmer Drought Severity Index [Palmer 1965] (PDSI) and monthly Standardized Precipitation Index [McKee et al., 1993, McKee et al., 1995] (monthly SPI; SP01). Unfortunately, the database does not have data for humidity, and we could not find humidity data that can be rasterized elsewhere, so the drought indices were considered a proxy for humidity.

The nClimDiv provides drought indices in climate division-level format. U.S. climate divisions do not follow county boundaries, so we rasterized the data at 0.01° resolution based on the climate divisions shapefile provided in the FTP repository, then aggregated by county.

G. *County Boundaries and Area*

The shapefile for U.S. county boundaries is available at https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html. The file also includes county land and water area data, which we used to calculate county-level population density in population per square kilometer. Since there are several changes to counties within the period of this study (2000-2016), we processed all county-level datasets to the most current boundaries based on changes described at https://www.census.gov/programs-surveys/geography/technical-documentation/county-changes.html.

We used the shapefile data to aggregate all gridded datasets at 0.01° resolution and adjust by the total area of 0.01° grid cells, which we calculated with spherical trigonometry. To simplify area calculation for complex

polygons, we approximated area by including an entire cell for a polygon if it contains the center of the cell.

III. Methods

A. Model Development

Similar to the use of monthly lagged values of climate variables in (Cholera Risk: A Machine Learning Approach Applied to Essential Climate Variables), we incorporated 1- and 2-month lagged values for the appropriate features in our model development. We used 10-fold CV with the training dataset and scoring with R^2 for feature selection, hyperparameter tuning, and overall evaluation of model performance. Scoring on an unseen test dataset excluded from training and feature selection was used to evaluate the generalization capability of the model. Specifically, we used a 70:30 split between the training and test datasets.

We conducted feature selection with SciKit Learn's recursive feature elimination and cross-validated selection (RFECV) function (sklearn.feature_selection.RFECV). A positive contribution to model performance equates to a higher mean R^2 from CV when a feature is included in the model versus when the feature is not included.

When using SciKit Learn's RFECV function, strange behavior occurs with more than 15 starting features, where the CV R^2 values are incorrect. Thus, we executed RFECV with random combinations of starting features. For each combination, an output array from the function call ranks features based on CV R^2 immediately before each feature is eliminated (when the feature is included but is the least important). We averaged rankings between each combination of starting features. Up to half of the features that ranked worst on average are removed. We then used the remaining features to rerun RFECV. These steps were performed iteratively until RFECV no longer eliminates any features. Furthermore, we did not include lagged values until after the second iteration; for the scope of the present study, we deemed this will increase efficiency while minimally affecting feature selection.

We performed random forest (RF) hyperparameter tuning with SciKit Learn's GridSearchCV function (sklearn.model_selection.GridSearchCV), which tests specified combinations of values for hyperparameters. Names of the tuned hyperparameters are listed in Table I and described in (sklearn.ensemble.RandomForestRegressor). For the present study, we performed RF hyperparameter tuning before and after feature selection.

B. Spearman Rank Correlation Coefficients

To assist in feature selection, we performed collinearity analysis using Spearman rank correlation with SciPy's 'scipy.stats.spearmanr' function (SciPy 1.0: fundamental algorithms for scientific computing in Python). The function also calculates statistical significance as two-sided p-values where the null hypothesis is that two datasets are uncorrelated. We used the returned correlation matrix (Fig. 9) to identify instances of collinearity. We considered absolute pairwise correlations to be high in collinearity if they are above a threshold ($| r | > 0.7$), as suggested by (Collinearity: a review of methods to deal with it and a simulation study evaluating their performance).

IV. Results

After feature selection and hyperparameter tuning, our model produced an R^2 of 0.7526 during CV and 0.7528 for test dataset predictions. When tuning hyperparameters solely to maximize R^2 on the test dataset, the model produced an R^2 of 0.7533 for test dataset predictions.

A. Feature Selection

RFECV was executed iteratively with random combinations of starting features, as described previously. For each combination, an output array from the RFECV function call ranks features based on CV R^2 immediately before each feature is eliminated. To characterize each iteration of RFECV, the CV R^2 for each rank is averaged between all combinations tested in the iteration (Fig. 3).

Among lagged/unlagged GFED layers, AQI, PM2.5, and climate variables/indices, the 1-month lagged NPP was the only feature not eliminated during RFECV. The other selected features are the location encoders for state and county (STATEFP and GEOID), temporal encoders for month and months from the start of the period (month and months_from_start), and land-area population density (popuDensity_ALAND_km2). Fig. 1 shows the Feature Importances for the model trained on these features.

By order of elimination, 1- and 2-month lagged temperature were the last and second last, respectively, 1-month lagged PM2.5 was the third last, followed by unlagged temperature, median income, and 2-month lagged PDSI (Fig. 2).

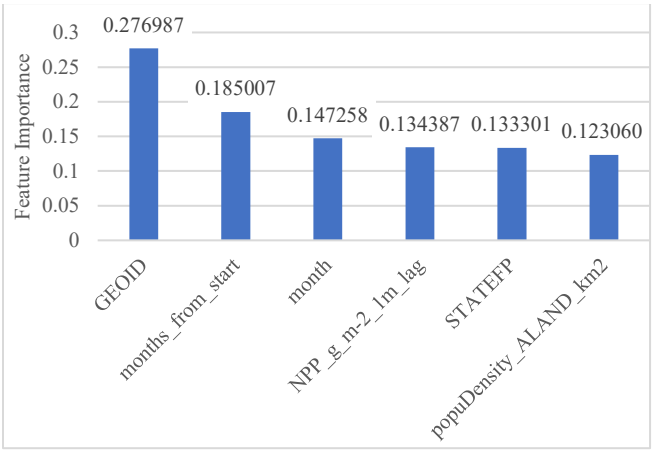


Fig. 1. Impurity-based feature importances for the model with the final selected features and optimal hyperparameters based on CV.

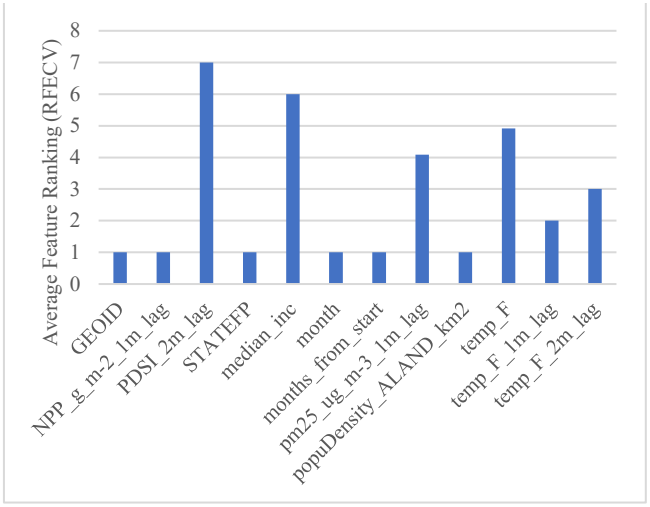


Fig. 2. Average feature rankings during RFECV (Iteration 4). Features are ranked by CV R^2 immediately before each feature is eliminated. For this iteration, selected features were the same across all RFECV function calls, thus having an average ranking of 1.

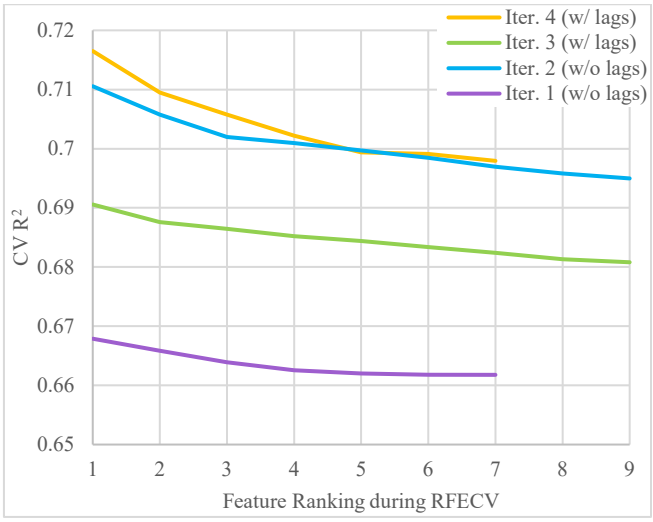


Fig. 3. CV R^2 by feature ranking during RFECV, regardless of feature. Features are ranked by CV R^2 immediately before each feature is eliminated. For any iteration, selected features are assigned rank 1.

B. *Collinearity Analysis*

We observed many cases of high collinearity ($|r| > 0.7$) between the lagged and unlagged values of the same variable, especially for PDSI, R_h, NPP, and temperature (Fig. 9). We also observed high collinearity between the GFED layers, with correlations equal or nearly equal to 1 between BB, C, and DM; and correlations of 0.867 to 0.988 between the other pairs of GFED layers.

Biosphere fluxes refer to NPP, R_h, and BB. Unlagged NPP and R_h are strongly correlated with temperature (0.869 and 0.753, respectively) and satisfy the threshold for high collinearity (Fig. 9). Out of these features, only 1-month lagged NPP was not eliminated during RFECV.

C. *Random Forest Hyperparameter Tuning*

We listed the tuned hyperparameters and their optimal tested values in Table I. We performed hyperparameter tuning before and after feature selection, as described previously. To create manageable runtime during feature selection, we set max_samples to 0.1 beforehand.

Increasing number of estimators only increased R^2 with diminishing returns (Fig. 4) and increased runtime; we determined 140 estimators to be sufficient.

Increasing min_impurity_decrease generally decreases R^2. However, there is significant noise when testing values below 5.0E-7 intervals (Fig. 5). This noise poses difficulty when selecting an optimal value, though for our model, simply using zero was near-optimal. Thus, we tuned the remaining hyperparameters with min_impurity_decrease set to zero, then tuned min_impurity_decrease afterward.

Tuning the remaining hyperparameters displayed the effects of "smoothing" the model and controlling overfitting, as reflected by the maximums observed for CV R^2 (Fig. 6). However, the improvement in R^2 is small for some hyperparameters and significant for others, with max_samples providing the largest improvement.

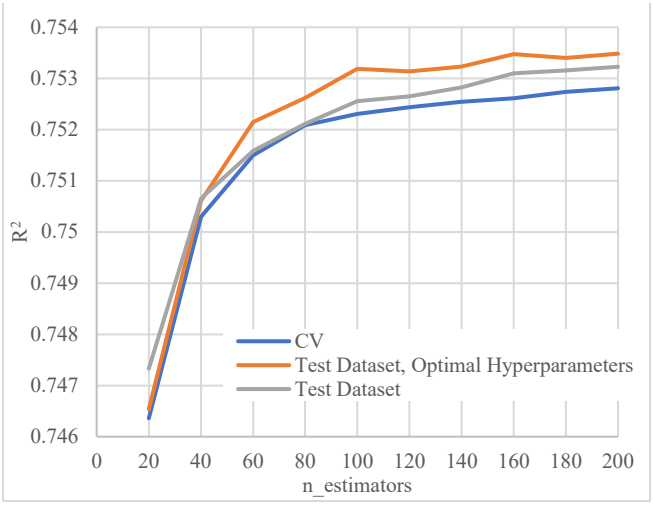


Fig. 4. R^2 by max_samples. The other hyperparameters are each at their optimal tested value (except for min_impurity_decrease = 0). "Optimal Hyperparameters" refers to the hyperparameters achieving the highest test dataset R^2.

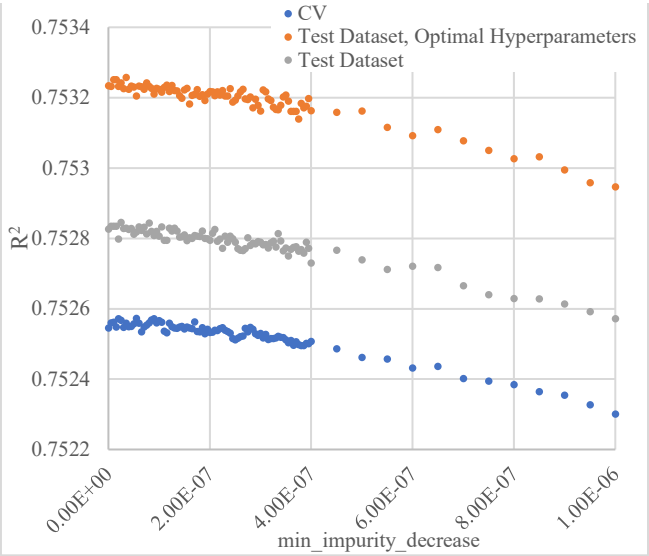
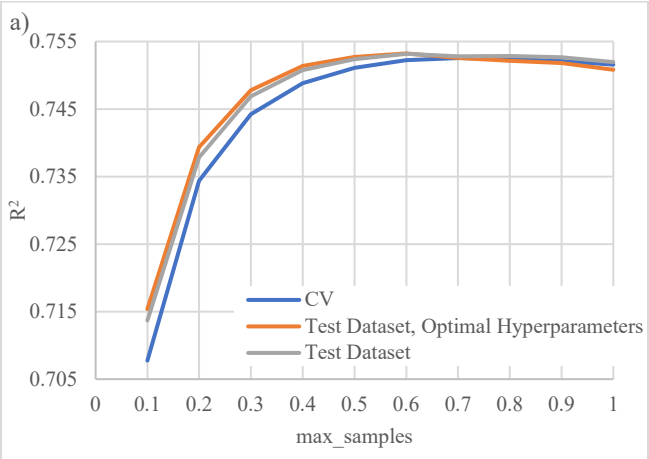


Fig. 5. R^2 by tested values for min_impurity_decrease. We tuned the other hyperparameters with min_impurity_decrease = 0, then tuned min_impurity_decrease afterwards. There is noise in R^2 when testing values for min_impurity_decrease below 5.0E-7 intervals. To narrow down the search, we tested values in 5.0E-9 intervals for values less than 4.0E-7. "Optimal Hyperparameters" refers to the hyperparameters achieving the highest test dataset R^2.

TABLE I. OPTIMAL TESTED HYPERPARAMETER VALUES

Name	Before RFECV (based on CV R^2)	After RFECV (based on CV R^2)	After RFECV (based on test dataset R^2)
n_estimators	140	140	140
max_samples	0.1	0.7	0.6
min_impurity_decrease [a]	0	5.5E-8	3.5E-8
min_samples_leaf	2	3	2
min_samples_split	4	8	9

a. There is noise in R^2 when testing values for min_impurity_decrease below 5.0E-7 intervals. Simply using 0 is very close to optimal. We tuned with n_estimators = 140 and min_impurity_decrease = 0, then tuned min_impurity_decrease at the final step.



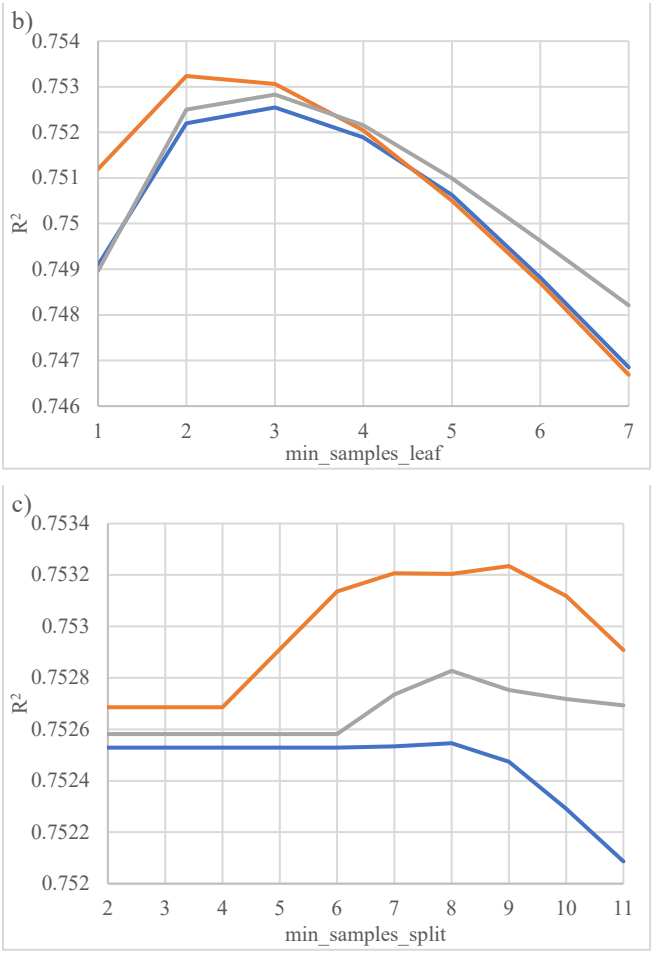


Fig. 6. R^2 by tested values for the hyperparameters a) max_samples, b) min_samples_leaf, and c) min_samples_split. We display each hyperparameter's train and test graph when other hyperparameters are each at their optimal tested value (except for min_impurity_decrease = 0). "Optimal Hyperparameters" refers to the hyperparameters achieving the highest test dataset R^2.

D. Test Dataset Scoring

Similar trends in R^2 were observed for CV and the test dataset during both feature selection (Fig. 7) and hyperparameter tuning (Fig. 8), suggesting that our model generalizes well for unseen data. Notably, R^2 tends to be slightly higher for the test dataset.

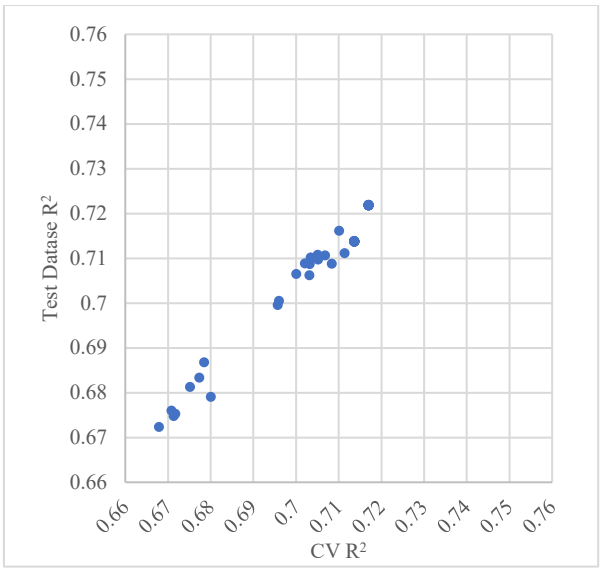


Fig. 7. Comparison between CV and test dataset R^2 for features with rank 1 across all RFECV executions (before hyperparameter tuning).

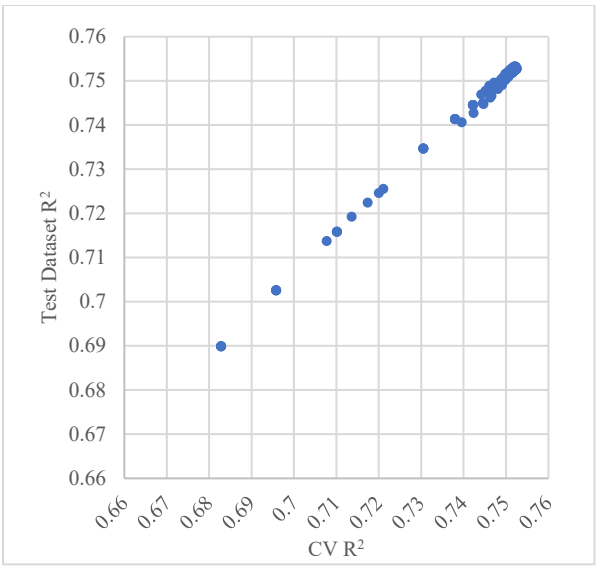


Fig. 8. Comparison between CV and test dataset R^2 for hyperparameter tuning after feature selection, with min_impurity_decrease = 0 and n_estimators = 140.

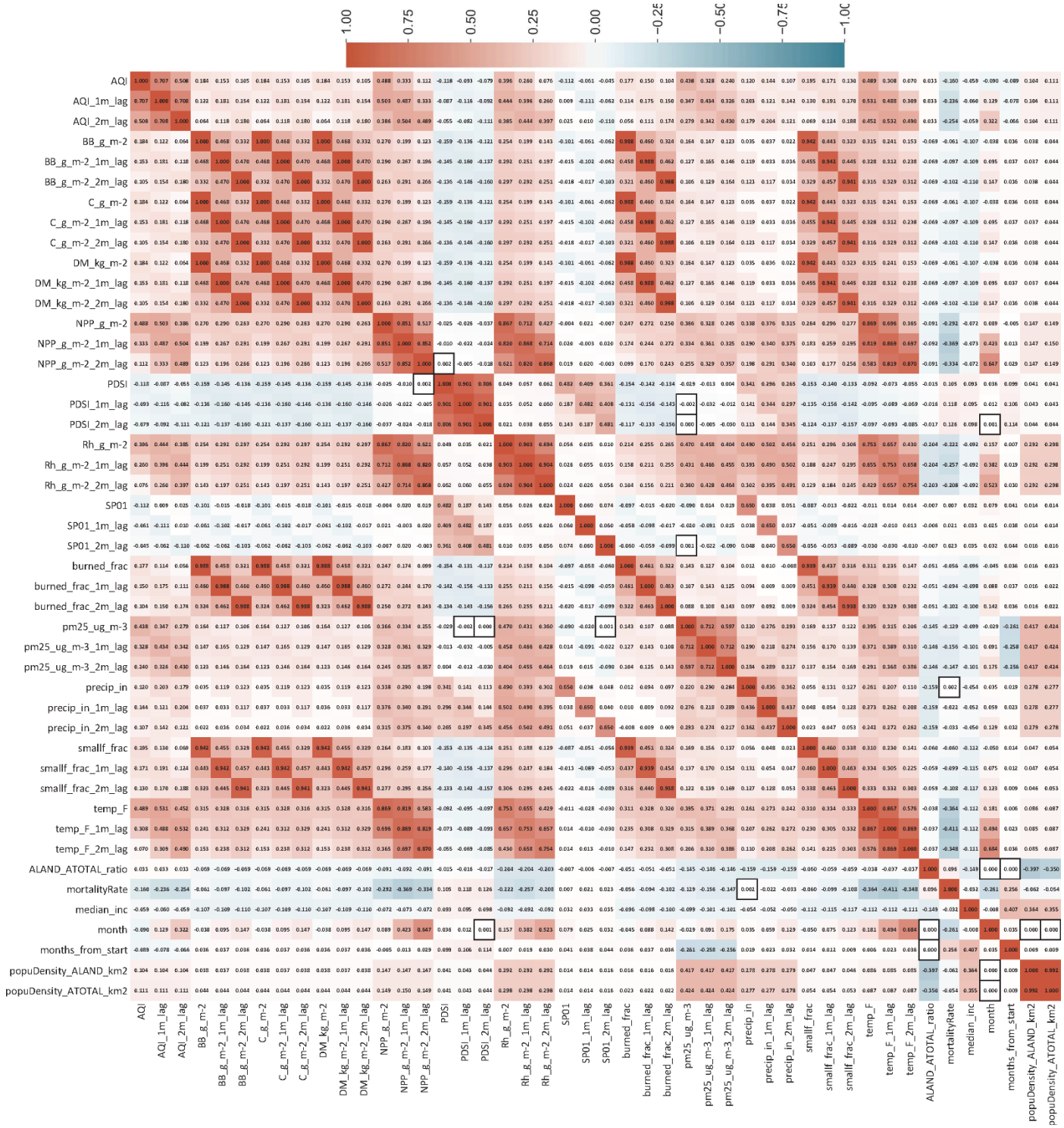


Fig. 9. Spearman rank correlation coefficient matrix for all essential climate variables and lagged values considered in the present study. Values with significant p-values (<0.05) are borderless. Statistical significance is calculated using a two-sided p-value where the null hypothesis is that two datasets are uncorrelated.

V. Discussion

A. RF Model Feature Performance

The high feature importance of the county encoder (GEOID) suggests there are relatively strong county-specific patterns in the form of complex relationships between features. The RF model likely splits the decision trees at intermediate stages so that these patterns are handled separately. Meanwhile, the state encoder (STATEFP) could be useful in handling patterns found on a lower spatial resolution. It may also improve model performance by creating a divide-and-conquer effect, as counties within the same state will have the most similar patterns.

Iteration 4 incorporates lagged values and produced the highest R^2 values, while Iteration 2 does not incorporate lagged values, yet produced the next highest R^2 values (Fig. 3). This suggests that incorporating lagged values for air pollutants, carbon emissions, and climate variables does not offer a substantial increase in our model's performance. We attributed this to the many cases of high collinearity between lagged and unlagged values of the same variable (Fig. 9).

B. Importance Measures

Impurity-based feature importances for RFs and other tree-based models suffer from two flaws. Firstly, they are biased towards high cardinality features (features with a large number of unique values, such as continuous variables). Secondly, they are computed on statistics derived from the training dataset and thus do not necessarily describe which features are the most important to make good predictions in unseen test data (Understanding Random Forests: From Theory to Practice).

Permutation importance is another feature importance measure commonly used for RFs that does not suffer from the flaws associated with impurity-based importance. However, SciKit Learn does not provide a recursive feature elimination function for permutation importance. Thus, we did not incorporate permutation importance in our model development. The close resemblance of R^2 between CV and test dataset scoring during feature selection (Fig. 7) suggests that the flaws of impurity-based importance were not pronounced in our model. This may compensate for our forgoing the use of permutation importance.

Among all the features we considered for selection, calendar month (month) and the state encoder have the lowest cardinality, yet they are among the final selected features. This suggests that the bias towards high cardinality features associated with impurity-based importance does not understate the importance of features that characterize location and seasonality.

C. Climate Variables and Drought Indices

During RFECV, 1- and 2-month lagged temperature were the last and second last, respectively, to be eliminated. This indicates that, among all eliminated features, the lagged temperature values have the lowest negative impact on model performance.

Humidity and temperature have synergistic effects on the symptoms of COPD patients, where high humidity enhances the risk of COPD due to low temperature (Synergistic effects of temperature and humidity on the symptoms of COPD patients). However, humidity data was not included, and we used the drought indices as a proxy for humidity.

PDSI's lagged/unlagged values have a weakly positive correlation with mortality rate (0.105 to 0.126), and temperature's lagged/unlagged values have a moderately negative correlation with mortality rate (-0.348 to -0.411) (Fig. 9). In other words, higher relative soil moisture (higher PDSI value) and lower temperatures are correlated with higher mortality rates. However, higher relative soil moisture does not always coincide with higher humidity (A Comparison of Weekly Monitoring Methods of the Palmer Drought Index, which cited (Heim 2002; Alley 1984)). This may explain the weak correlations between the lagged/unlagged PDSI values and mortality rate (Fig. 9).

D. Fire Emissions, Biosphere Fluxes, and PM2.5

Lagged/unlagged NPP was observed to have a significant negative correlation with mortality rate (-0.369 to -0.292) (Fig. 9). NPP is the net carbon gained by vegetation, calculated as the carbon gained by photosynthesis minus the carbon released by plant respiration (Biogeochemistry of Terrestrial Net Primary Production). In other words, higher NPP implies less carbon is released into the atmosphere, and this is correlated with a lower mortality rate.

Emission layers from GFED (C, DM, and BB), PM2.5 concentrations, and AQI have a weak negative correlation with mortality rate (Fig. 9). This seems to contradict commonly-known exacerbating effects of air pollution on respiratory disease (Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death, Characterization of PM 2.5, gaseous pollutants, and meteorological interactions in the context of time-series health effects models, Spatiotemporal relationship between particle air pollution and respiratory emergency hospital admissions in Brisbane, Australia). An explanation may be that pollutants in the form of fine particulate matter and carbon emissions do not trump climate variables as predictors of mortality.

E. Relationship Between Drought and Other Features

Elevated ozone and PM2.5 levels have been attributed to increasing drought (Adverse effects of increasing drought on air quality via natural processes), but the drought indices have a near-zero correlation with PM2.5 and AQI (Fig. 9). Estimates from (Drought-Induced Reduction in Global Terrestrial Net Primary Production from 2000 Through 2009) imply a reduction in global NPP due to droughts during 2000-2009, with increased NPP over the Northern Hemisphere offset by decreased NPP over the Southern Hemisphere. However, the present study is limited to the contiguous U.S., and a near-zero correlation between NPP and the drought indices (PDSI and SP01) was observed (Fig. 9).

Evidently, correlation coefficients cannot describe the relationship between drought and these features. Aside from the elimination of the drought indices during RFECV, the

impact of including drought indices on model performance are inconclusive.

F. Median Income

Based on our Spearman rank correlation analysis, median income has a weak negative correlation with all GFED layers, PM2.5, and AQI. This may be relevant for literature on environmental inequality, especially North American studies that indicate areas where low-socioeconomic-status communities dwell experience higher concentrations of pollutants (Socioeconomic Disparities and Air Pollution Exposure).

G. Data Limitations

Emergency Room Visits (ERVs) have been used as a proxy of disease exacerbation for a global study (Estimates of the Global Burden of Ambient PM2.5, Ozone, and NO2 on Asthma Incidence and Emergency Room Visits). However, data is limited for it on a by-county basis. Thus, we used mortality provided as death counts by CDC WONDER for every state and county for the present study. However, due to data suppression constraints (Underlying Cause of Death 1999-2019), we interpolated the unreported counts as described previously. Furthermore, deaths only capture the most acute cases of exacerbation and do not necessarily reflect non-mortality cases, not to mention self-treated cases that do not manifest in an ERV.

Not to confuse with total-column ozone, ground-level ozone data is available as annual gridded mean concentrations, which are estimates simulated by an ensemble of five chemical transport models (Coordination and harmonization of the multi-scale, multi-model activities HTAP2, AQMEII3, and MICS-Asia3). These annual mean concentrations were used to estimate global asthma ERVs attributable to ozone (Estimates of the Global Burden of Ambient PM2.5, Ozone, and NO2 on Asthma Incidence and Emergency Room Visits). Due to the lack of a monthly ground-level ozone dataset, we did not include ground-level ozone for the present study.

Data for daily AQI, ground-level ozone, and other pollutants from EPA Air Data is available in coordinate-specific format. This format is more beneficial for case studies on a localized scale, such as in Spatiotemporal relationship between particle air pollution and respiratory emergency hospital admissions in Brisbane, Australia, where monitoring stations distributed within a city can more closely capture the spatial variation of air pollution and subsequent health outcomes between different areas within the city. AQI was spatially interpolated between coordinates with available data and was eliminated in the RFECV as we expected. We can attribute this to the naive interpolation applied to AQI.

H. Other Limitations

Datasets on county-level and monthly resolutions cannot represent phenomena present in finer spatiotemporal resolutions, which is a significant source of limitations for the present study. Existing studies have analyzed various exacerbating effects on a more localized scale, such as within individual cities (Spatiotemporal relationship between particle air pollution and respiratory emergency hospital admissions in Brisbane, Australia). Thus, we may characterize these limitations.

A multi-city case-crossover study (Air pollution and emergency department visits for respiratory diseases) found that ERVs for respiratory diseases increase after days with higher concentrations of pollutants, even where pollutant concentrations are relatively low. However, the lagging effect on ERVs lasts only up to several days. For PM2.5 and ERVs for COPD, positive results were observed with lags of 1-8 days. Thus, the use of monthly predictive variables is likely a limitation of this study.

Excluding the predictors tested in this study, CLRDs are also attributable to other factors, particularly tobacco smoking and other airborne allergens (Global burden of COPD, Definition, epidemiology and natural history of COPD, Allergy and asthma: Effects of the exposure to particulate matter and biological allergens), but also infectious microorganisms (Infections in "Noninfectious" Lung Diseases).

Exacerbation of asthma and COPD also stems from sudden changes in temperature associated with the overuse of cold air conditioning during warmer months (The impact of cold on the respiratory tract and its consequences to respiratory health). Meanwhile, in cold and temperate climates, due to more time spent indoors, indoor pollutants are likely more important than outdoor ones (Influence of indoor factors in dwellings on the development of childhood asthma).

VI. Conclusions

This study has created some insights on the application of RFECV and RF as a ML technique for regression. It also demonstrates the difficulties in modeling CLRD risk on a national, county-level scale due to data availability limitations, factors not captured on the monthly timescale, and the nature of CLRD being not primarily infectious in etiology (Infections in "Noninfectious" Lung Diseases). It highlights the caveats of using ML methods to generating predictions from complex relationships between correlated variables. We can naively test for features that positively contribute to model performance, but the mechanisms behind their relationships remain unclear.

A similar approach has been effectively used in a recent study to classify cholera outbreaks in India using essential climate variables, also in monthly temporal resolution, with 89.5% of outbreaks correctly identified in the unseen test dataset (Cholera Risk: A Machine Learning Approach Applied to Essential Climate Variables). Pathogenic *Vibrio cholerae* bacteria are responsible for human cholera. Humans can be exposed to the pathogenic bacterium through consumption and drinking of contaminated seafood and water, and during recreational activities in contaminated waters. Strong relationships between essential climate variables and the coastal distribution and seasonal dynamics of *V. cholerae* allow for accurate predictions of cholera outbreaks.

Meanwhile, the role of microorganisms in "noninfectious" lung diseases, including COPD,

bronchiectasis, and asthma, is an active area for investigation. More sophisticated methods and analyses to investigate the interactions between the immune system, microbiota, and inflammatory pathways are needed to improve our understanding of pathogenesis for these diseases (Infections in "Noninfectious" Lung Diseases).

In Estimates of the Global Burden of Ambient PM2.5, Ozone, and NO2 on Asthma Incidence and Emergency Room Visits, health impact functions are used to estimate asthma ERVs and incidence attributable to each pollutant, stratified by country and age group. The health impact functions rely on relative risks (RRs) extracted from meta-analyses of epidemiological studies. However, they do not include climate variables such as temperature, which affect asthma and COPD rates, as existing studies indicate. The effects of climate are somewhat captured since climates tend to be country-specific and thus influence a country's baseline rate of ERVs/deaths, but this becomes an issue for larger countries. Thus, accessible data of climate variables may be used to improve estimates of the number of ERVs attributable to specific pollutants. Contributions of other factors, such as sudden changes in temperature and infectious determinants, remain open areas for future studies.

Immediate improvements to the present study may replace our use of grid-search hyperparameter optimization with randomized or gradient-based (Forward and Reverse Gradient-Based Hyperparameter Optimization) (Gradient-based Hyperparameter Optimization through Reversible Learning), compare RF with different ML techniques such as neural networks, and train with similar datasets but from alternative sources that may develop their datasets differently. By filtering the Underlying Cause of Death data, demographic-specific effects can also be analyzed.

Code Availability

The code developed for this study is available via GitHub at https://github.com/Unusuala1l2e3x4/Research-Spring2021. Datasets, and descriptions of sources for datasets too large for a GitHub repository, are included.

References

[1] D

[2] D

[3] D