Global Climate Factors and Their Impact On Human Disease Burden

Contributors

Team 8: Carl Chan, Chiamaka Nnamani, Patrick Russell

Objective

Our objective for this project was to understand “How do interactions between climate change factors (e.g. air quality, temperature, precipitation, extreme weather) shape health risks across countries with differing environmental and socioeconomic conditions?”.

We wanted to understand the impacts of climate change factors for three reasons. The first was to provide credible information to help governments make informed decision-making for future healthcare planning activities. The second reason was to provide evidence for advocacy groups to aid their attempts in advocating for climate-responsible initiatives, policies, and technologies. The last reason was to add to the literature and collective understanding of the impact on humans across the globe.

In our research we made the following hypothesis: **The worsening of climate change factors would be associated with the worsening of disease burden rates, particularly deaths.** More specifically:

1. Respiratory illness rates would rise in relation to air pollution and heat stress factors.
2. Cardiovascular disease rates would rise in relation to heat waves and temperature factors.
3. Infectious disease rates would rise in relation to flooding and climate factors.

Data Preparation

Sources

We used two separate datasets together for our research. Both data sources were open on the web. The first dataset was the (“Global Climate-Health Impact Tracker (2015-2025)”) and the second was (“Global Burden of Disease (GBD)”), both linked in the works cited. This was done to provide a more novel and richer dataset to analyze. The intention was that by adding the GBD data, a more nuanced model that accounted for a variety of different diseases could be made.

During analysis, the GBD dataset was merged with the Global Climate-Health Impact Tracker (GCHIT) data. The purpose of this was to explore how climate and environmental variables

were associated with the burden of disease across countries and time. The analysis focused on three overall disease categories: Respiratory, Cardiovascular, and Infectious diseases.

Our notebook was designed to directly download the datasets from the web instead of assuming the files were in the local directory already.

Challenges and Data Quality

One challenge encountered with the datasets was the much lower time granularity in the GBD dataset. It was only reported on an annual basis, compared to the weekly breakdown for the Global Climate-Health Impact Tracker (GCHIT). This meant that when combining the two datasets together, the GCHIT data had to be aggregated on an annual basis as well. While this allowed the data to be merged, it lost the ability to represent seasonal factors within a year and reduced the total number of data points available for that portion of the analysis.

Additionally, it was found that there were errors in the air quality index (AQI) column in the GCHIT dataset. The value should only range from 0-500, but a few negative values were discovered. This was the only column found to have any data quality issues though.

Preprocessing

There were several basic preprocessing steps used to make the datasets more useful for analysis. Firstly, both datasets were immediately confirmed to have neither any missing values nor duplicate rows. Secondly, categorical columns such as country, region, and income level were converted to the category data type. That last categorical column, income level, was an ordinal category and so was manually given ordered values. Thirdly, the date field was converted into a proper date type, which was used to create a time-indexed copy of the data for temporal analysis. Fourthly, the previously mentioned errors in the air quality index data were fixed. That index should not have negative values, so the few values that did have those were clipped to zero.

Feature engineering was done to create several new columns in the GCHIT data. Both 4-week and 8-week rolling averages were taken for each country along several key climate factors related to temperature and air quality. This was to smooth out week-to-week volatility in an effort to demonstrate trends better. Interaction terms were also created by multiplying a climate factor with a social factor (e.g. temperature and income level) to help capture how the effect of one variable might depend on another.

Feature engineering was also done for the GBD data. Each cause of death in that dataset was grouped into one of the three main disease categories. Namely: respiratory, cardiovascular, and infectious disease. The dataset was filtered to have rows covering all ages and both sexes, then total deaths across all causes of death were aggregated by country, year, and disease type. This was normalized by dividing by each country’s population to reflect deaths per million.

Additionally, health outcomes do not always respond immediately to climate events. For example, we suspected floods might potentially lead to waterborne disease 1-3 weeks later. Thus, multiple copies of the GCHIT data were created where each of the target health outcomes were shifted 1-3 weeks into the future for their respective country. This aligned them with earlier feature values, to help determine if responses were simply lagged in time.

When visualizing the distributions of the data, it was found that a few of the columns did not seem to be normally distributed, such as the healthcare access and food security indices. It was not known if this would affect the performance of later modeling, so a copy of the GCHIT data was created where the columns with continuous values had a Box-Cox transformation applied to them. This would then be used to experiment during the machine learning process..

Finally, before passing the data to machine learning models, several other steps were applied. The nominal categorical fields (primarily related to country and region) were given one-hot encodings and expanded into many new columns to prevent the models from erroneously assuming the ordering of the original values held any significance. For both training/test set creation and cross-validation, a time series split was used instead of a regular K-fold to prevent data leakage from the future. This ensured the models always trained on past data and validated on future data. The features were scaled using normal distribution standardization before being fed to the models, as many of said models were distance-based. These scalers were fit on the training data only, to further prevent potential data leakage from the test set.

Outliers

The data was investigated for outliers. This was primarily done simply by looking for observations which had at least one column more than 3 standard deviations from their respective means. They were not removed from the dataset though as there was not enough statistical proof that they could be safely left out. They could not be proven to be input errors either (i.e. unlike the negative values in the AQI column). So these potential outliers were kept in the dataset to maintain the fullest representation of the data possible.

If there had been more time, these outliers could have been explored further by iteratively removing one at a time to determine the leverage they had on the models. Then if they very heavily impacted the performance of the resulting models they could be considered for removal.

Merging Datasets

The GCHIT climate data was aggregated from weekly to yearly averages to match the GBD data. Temperature, temperature anomalies, PM2.5 (hazardous particulate density), and air quality index were averaged over each year. Precipitation, heatwave days, drought/flood indicators, and extreme weather events were instead summed together on an annual basis.

The datasets were found to have slightly different names for some countries. After remapping those values to correct this, this resulted in a dataset covering 24 countries. The end result was a blended dataset where the annual climate factors for various countries were listed with their respective death burdens for each of our three designated categories of disease.

Analysis

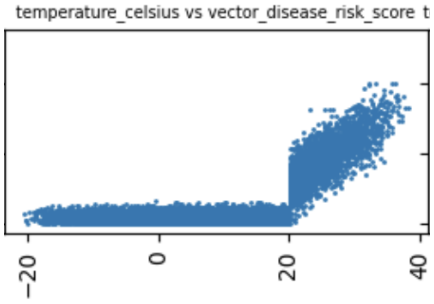
Global Climate-Health Impact Tracker (GCHIT) Data Only

Visualization

Distributions

The initial phase of our analysis focused on visualizing the GCHIT dataset to understand its structure and intuitively identify patterns. We examined the distributions of key variables, and discovered that several columns, such as healthcare access, exhibited non-normal distributions. This led us to consider Box-Cox transformations to improve model performance.

Exploratory Visualization

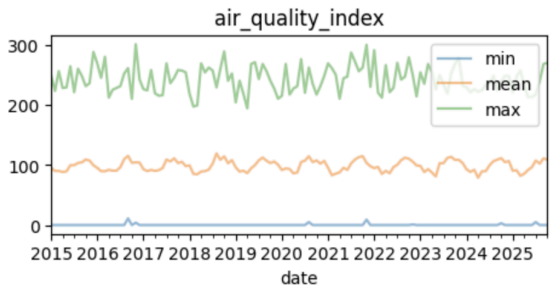


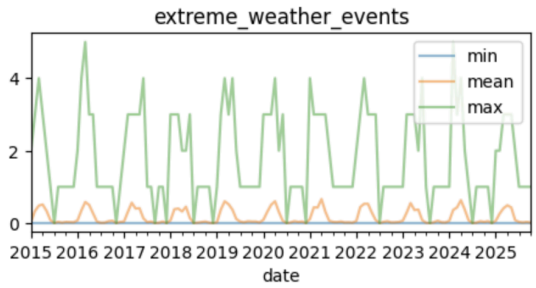
We began our analysis by creating a matrix of scatter plots which displayed each feature in the GCHIT dataset against each of our target health variables inside the same dataset (i.e. before adding in the fuller disease data of the GBD dataset). This gave some initial insight into the broad correlations between the variables, including some seemingly non-linear relationships such as the one between temperature and vector disease risk. This informed us early on that if these relationships were significant predictors of health outcomes, linear regressions might have difficulty properly capturing them and other models (e.g. tree-based) might be more appropriate.

Afterwards, several sets of pair plots were created. As part of our investigation into potential outliers, one pair plot was made to show the relationships between the climate factor columns we searched for outliers previously, and another similar pair plot was made for the health outcomes. While this ultimately did not inform on what might be outliers, after distinguishing the data points by income level we did see some clearer separation within some of the otherwise indistinct blobs in the plot. Refer to Appendix A.3, and note the row of plots for the healthcare access index which show well defined blocks occupied almost exclusively by each income level. This led to income level being a category we commonly split the data on in later analysis.

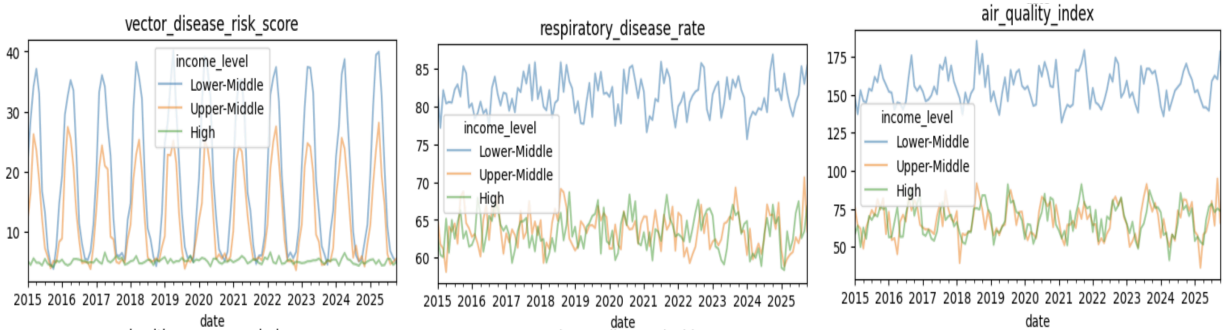
Temporal Analysis

Temporal trends were analyzed through time-series plots, resampled to monthly intervals to smooth out week-to-week volatility and help illustrate overall trends. These plots illustrated the global progression of variables over time by averaging data across the countries in our dataset and compared monthly averages to minimum and maximum values to assess variability. We noted several variables (e.g. air quality index and extreme weather events) which seemed to have some underlying cyclic nature.

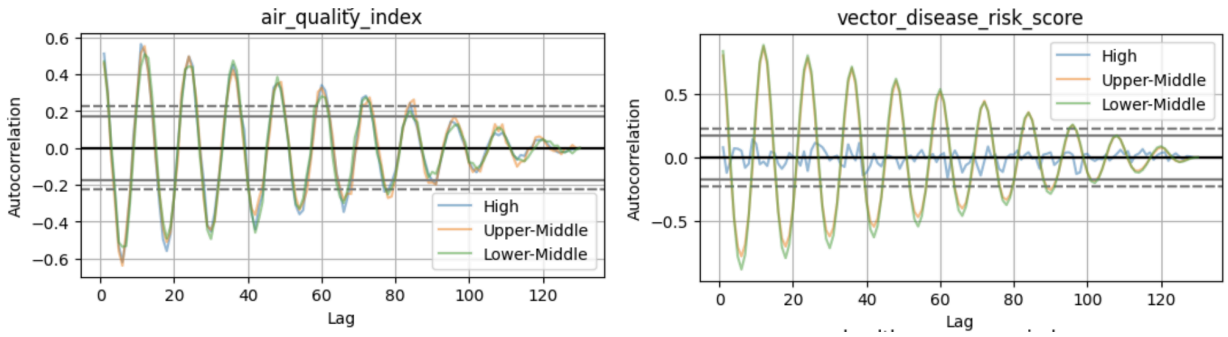




To provide further insight, we produced similar monthly plots segmented by country income level. In several variables, for all other things being equal, there was clear separation between the income levels. For instance, for the vector disease risk score, countries in the highest tier of income level had very little variance over time, while other countries tended to follow the same annual cycles where the risk spikes to very high levels. Conversely, rates of respiratory disease were very similar between high and upper-middle income countries, with the lower-middle income countries being the unusual case. This was quite similar to the plot for air quality index (in which higher values correspond to worse air quality), suggesting that a trend of the poorest nations having very unhealthy air was the real driver rather than the wealth of the nation itself.

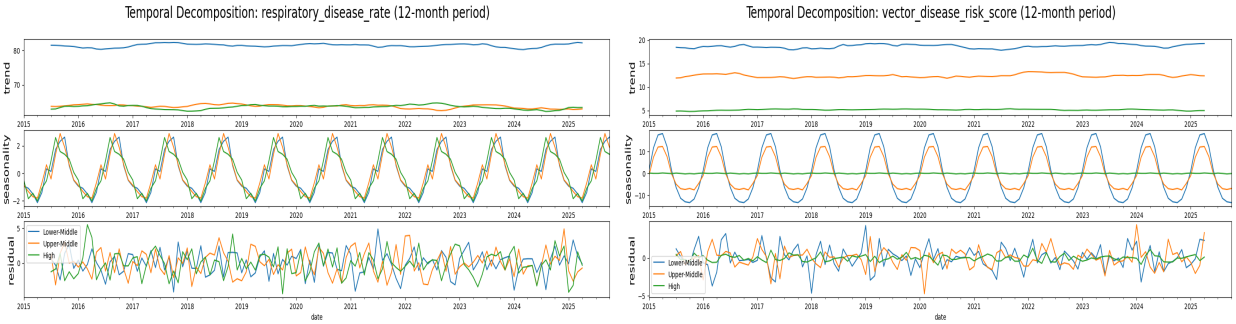


To explore the cyclic patterns observed in several temporal plots, an autocorrelation analysis was conducted on key climate and health indicators, again resampled to a monthly basis. Many of the variables exhibited statistically significant (i.e. beyond the 99% confidence bands) autocorrelations on lags spaced roughly 12 months apart. This was expected for many climate factors, though a somewhat surprising one of those was the air quality index. Unlike variables such as temperature, it was much less apparent why air quality would vary so predictably over the course of a year. This was observed across all income levels, so while the least wealthy countries had by far the worst air, all countries experienced roughly the same annual cyclic patterns for it.



When interpreting the autocorrelation plots for the health indicators, the highest income countries never had any statistically significant autocorrelation for vector disease risk. This was despite all the less wealthy countries having clear cycles at roughly 12-month periods. Thus not only did the richest countries have the lowest average risk of vector-borne disease, that category of nation also maintained an extremely stable baseline for that health indicator while other countries were subject to much more volatile seasonal patterns.

With the seasonality of many variables appearing to be on a 12-month period, a temporal decomposition was done for several key variables on that period. Due to the wealth-based patterns found previously, these decompositions were also broken down by income level. Two variables that demonstrated both statistically significant autocorrelations at regular intervals and informative temporal decompositions were respiratory disease rate and vector-borne disease risk. Their trends were fairly stable and they had well formed seasonality components, suggesting that time (e.g. week of the year) should be used as a feature in later machine learning steps.



Regression Models and Statistical Analysis

For the GCHIT dataset, we implemented a series of regression models attempting to predict the health indicators using features present in the data. This began with basic linear regressions (with cross-validation), followed by regularized approaches including Ridge (L2), Lasso (L1), and ElasticNet (L1 + L2) to discourage the models from overfitting on superfluous features and control for collinearity between the large number of new one-hot encoded categorical features. As this was time-series data, we used a time-based split to create the training and test sets instead of a random split to prevent data leakage from future data points.

The best model for each health-related target exhibited substantially varying performances when put against the 2024-2025 test set. Predictions for heat-related admissions were strong and well-calibrated ($R^2 \approx 0.79$), while respiratory disease rates and vector-borne risk scores show

moderate accuracy (R^2≈0.58 and 0.56, respectively), though with a tendency to underpredict high values and overpredict lows. Modeling for waterborne disease incidents performed more poorly (R^2≈0.41), and cardio mortality was very poorly explained (R^2≈0.17). Modeling for mental health index held essentially no predictive power (R^2≈0.01), suggesting high noise and/or feature mismatch. Across most targets, residuals increased with the actual values, indicating heteroscedasticity. That is to say the variances of actual values were higher for countries with very high average disease burden, which increased the range of errors for them. This also suggested that the models tended to underestimate extreme values and perform best near average conditions.

Overall, the models performed well for outcomes that were strongly influenced by climate and seasonal patterns, but performance declined when trying to model certain other health indicators. For these outcomes, predicted values increasingly diverged from observed values, indicating that climate variables alone were insufficient to fully explain disease burden. These consistent error patterns suggest the influence of additional, unmeasured factors not captured in the models. Model performance was primarily assessed via both R-squared value and by visually comparing observed and predicted values for each health indicator on plots, with the size of prediction errors colourized based on RMSE.

Using a 5-fold TimeSeriesSplit, the cross-validated R^2 showed moderate and stable performance for respiratory disease rates (mean $R^2 \approx 0.48$), weak but consistent signal for waterborne incidents ($R^2 \approx 0.21$), and low with notable instability for vector-borne disease ($R^2 \approx 0.20$, std ≈ 0.24) and heat-related admissions ($R^2 \approx 0.06$, std ≈ 0.44). Modeling for cardiovascular mortality remained weak ($R^2 \approx 0.10$), and mental health was even negative ($R^2 \approx -0.06$), which meant it underperformed even a mean baseline. The high time-based variability, especially for heat-related admissions, indicated the models performed well in some periods but poorly in others even when time (i.e. week number of the year) was added as a feature.

Comparisons were made between models trained with and without our engineered features, as well as those trained on the Box-Cox transformed copy of the dataset created previously. We also experimented with filtering out superfluous features in each model using univariate feature selection, under the suspicion that the models might be capturing noise from them. Feature engineering produced meaningful improvement for the vector-borne disease risk score (R^2 increased from 0.565 to 0.660, +0.095), while changes for respiratory, waterborne, heat-related, cardio mortality, and mental health were negligible to slightly negative (≤ +0.003 in R^2). The overall R^2 increased modestly (0.422 to 0.437, +0.015), driven primarily by the vector outcome.

Additional models were developed using the original dataset segmented by income level. Three separate models were made, one each trained on a subset of the data representing a single income level. However, this experiment did not seem to improve performance.

To capture temporal dynamics, we created a series of lagged models by shifting health targets from one week to three weeks into the future. This was as the health impact of certain climate conditions might not be immediately apparent, such as flooding potentially contributing to waterborne disease. This did not seem to meaningfully improve performance either though.

Tree-based models were explored, primarily Random Forest, trained on both the original data and its lagged variants. The original data performed best, and for some health targets resulting in substantially better results than the linear models (e.g. vector disease R^2≈0.912, +0.252). Refer to Appendix B.1 for a list of the most important features for predicting each health target.

K-Nearest Neighbours regressions were also tried, but with negligible performance improvements, if any.

While several of our models demonstrated reasonable performance and consistency though, we still felt that the GCHIT dataset alone was not sufficient to fully meet our research objectives. This was especially in terms of meaningfully adding novel insight to climate change literature. One limitation was that several GCHIT health indicators were broadly defined and lacked clear transparency regarding their underlying disease composition. Additionally, these indicators were not consistently focused on mortality outcomes, which were central to our analysis. Due to these limitations, we incorporated the Global Burden of Disease (GBD) dataset from the Institute for Health Metrics and Evaluation (IHME), which provided standardized, disease-specific burden and mortality data, which enabled a more direct and richer connection between climate factors and health outcomes.

Global Burden of Disease Data: Climate Disease Analysis

Correlation Analysis: Climate Factors vs Disease Burden

We analyzed the connections between different climate factors and various categories of disease burden, specifically respiratory, cardiovascular, and infectious diseases. A number of consistent trends came to light.

Higher temperatures, more rainfall, and an uptick in extreme weather events were all consistently linked to an increase in vector-borne and infectious diseases. These trends suggested that such diseases were highly sensitive to climate conditions. For example, temperature was moderately to strongly associated with vector disease risk ($r \approx 0.52 – 0.60$). Similarly, rainfall ($r \approx 0.47 – 0.55$) and extreme weather events ($r \approx 0.40 – 0.50$) also had positive correlations with disease rates.

Respiratory diseases, on the other hand, were most strongly tied to air quality. Indicators like PM2.5 and the Air Quality Index (AQI) showed strong correlations with respiratory disease rates (PM2.5: $r \approx 0.68$; AQI: $r \approx 0.63$). Furthermore, although air pollution showed some connection to vector-borne diseases, it played a smaller role compared to other environmental factors.

When it came to infectious diseases, variability in climate seemed to be a more critical factor than static conditions. Temperature anomalies had a meaningful correlation with infectious disease rates ($r \approx 0.51$), and precipitation also showed a moderate link ($r \approx 0.49$). This underscores the impact of shifting and unpredictable climate patterns on disease transmission.

Conversely, cardiovascular disease outcomes showed very weak correlations with climate variables ($r < 0.15$). This suggested that these health issues were more strongly influenced by demographic, lifestyle, and healthcare factors than by environmental ones.

In summary, our analysis highlighted that vector-borne and infectious diseases were notably affected by climate factors, respiratory diseases were mainly driven by air pollution, and cardiovascular diseases were relatively insensitive to climate influences. These distinctions informed our later modeling and interpretations of how climate impacts health.

National Income, Air Quality, and Disease Burden

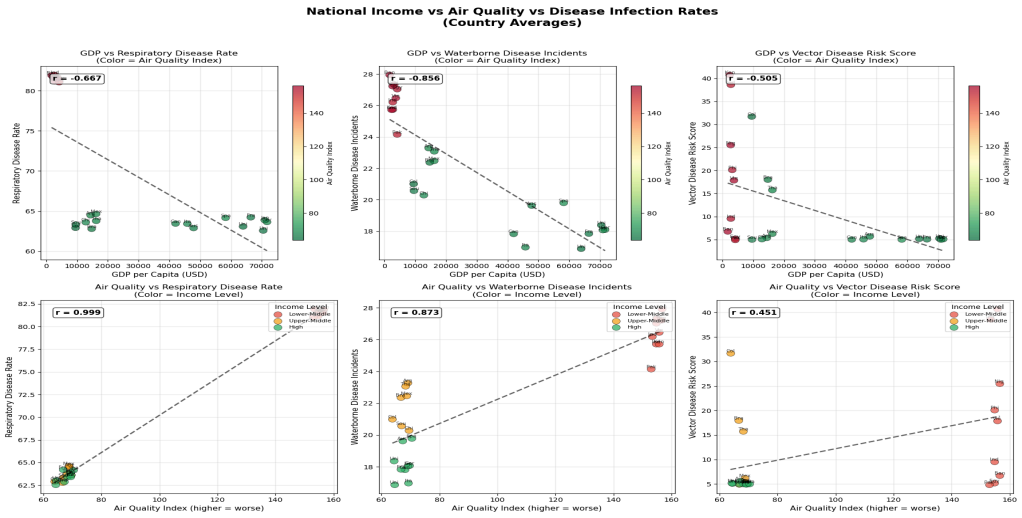
Our analysis also examined how national income (GDP per capita) related to environmental quality and disease burden across countries. Results from the merged dataset indicated that income played a substantial role in shaping both air pollution exposure and overall health outcomes, particularly for respiratory diseases. This possibly meant income level could be a confounding factor influencing both air quality and health, rather than the latter two being strongly linked directly.

Across income groups, higher GDP per capita was consistently associated with lower levels of air pollution and lower disease burden. Countries with higher income tended to have reduced PM2.5 concentrations and AQI values, as well as lower overall disease rates. This pattern was supported by correlation analysis, which showed moderately negative relationships between GDP per capita and PM2.5 ($r \approx -0.62$), as well as between GDP per capita and overall disease burden ($r \approx -0.55$). These findings indicated that as national income increased, pollution levels and disease burden generally both also decreased.

Lower-middle income countries experienced the most severe combined environmental and health burdens. In this income group, average PM2.5 concentrations were in the range of approximately 48-55 μg/m³, with AQI values between 110 and 135. Correspondingly, respiratory disease rates in these countries were approximately two to three times higher than those observed in high income countries. These results highlight the disproportionate impact of poor air quality on respiratory health in economically constrained settings.

While income was strongly related to pollution and respiratory disease outcomes, it did not fully explain all observed disease patterns. This suggested that infectious disease burden was influenced by factors beyond income and pollution alone. Climate-related conditions such as temperature, precipitation, and extreme weather which were analyzed separately in the climate-disease sections of this study appeared to play a more direct role in shaping infectious and vector-borne disease risk.

Overall, these findings indicate that national income modifies health risks primarily through its influence on environmental quality, especially air pollution, but does not eliminate disease risks associated with climate-sensitive pathways. This distinction is important for interpreting climate-health relationships and for understanding the limits of economic development in reducing certain categories of disease burden.



Feature Assignment by Disease Category

A slightly different subset of the environmental features was assigned as inputs to each of the machine learning models meant to predict the burden of one of the three disease categories previously outlined. Causes of death were then grouped into these disease categories and modeled based on established epidemiological pathways linking environmental exposure to health outcomes. In this analysis, air pollution was considered a primary influence of respiratory disease through direct inhalation, while vector borne diseases were driven by climate mediated pathways involving temperature and precipitation rather than air quality. This feature assignment is informed by both the correlation analysis and prior biological and environmental evidence.

Among the infectious, vector-borne, and waterborne diseases the focus was on the climate variables of temperature, precipitation, and extreme weather events. These diseases were highly sensitive to the surrounding environment factors that influenced pathogen survival, transmission, and exposure. This assignment was well-justified by very strong model performance, including moderate predictive power in the GCHIT models and a near-complete explanation of variance in the best GBD infectious disease model ($R^2 \approx 0.96$). Experimentally, these feature assignments resulted in good performance, where in our model we found climate variables accounted for approximately 97% of feature importance in our tree-based models.

Respiratory diseases were modeled using both climate and air quality variables. While climate factors showed a certain level of association, given the direct physiological impacts on lung health due to pollution it was expected that pollution indicators, including PM2.5 and AQI, would be the dominant drivers. This was confirmed by the model performances, with climate-only models showing only moderate explanatory power ($R^2 \approx 0.46$), indicating that respiratory outcomes might be driven mainly by pollution and behavioural risk factors rather than by climate alone.

Cardiovascular diseases were analyzed separately because they could be more strongly influenced by demographic, lifestyle, and healthcare-related factors than by climate variables. Consistent with this understanding, climate-based models showed only middling predictive power ($R^2 \approx 0.5613$), indicating that climate alone might not be a primary driver of cardiovascular disease burden at the national level.

Together, these findings justified our feature assignment approach and indicated that climate influenced disease categories in different and biologically relevant ways.

Predictive Modeling: Climate Impact on Disease

Based on the objective of this study, we applied predictive models to assess health outcomes across the disease categories using climate-related variables: temperature, precipitation, extreme weather events, and air quality. Model performance was assessed using R^2 as a measure of the proportion of variability in disease burden explained by our models using climate factors. Disease-specific indicators were used where appropriate, such as drought and flood indicators for infectious diseases and heat wave days for cardiovascular diseases, to better capture relevant climate-health relationships.

Climate-Health Predictive Models

Initial models including climate and environmental variables were of varying predictive strength across disease outcomes:

- Heat-related hospital admissions showed a strong climate signal, $R^2 = 0.79$, thus temperature and extreme heat were highly predictive for heat-related health outcomes.
- Respiratory diseases resulted in moderate predictive performance, $R^2 = 0.58$, which was indicative of the joint influence of climate conditions and air pollution.
- Vector-borne diseases showed a modest climate sensitivity, with $R^2 = 0.56$, consistent with their dependence on temperature and precipitation patterns.
- Waterborne diseases showed weaker albeit detectable climate influence, with a value of $R^2 = 0.41$, which indicated that climate factors contributed but were not the sole drivers.

GBD Climate-Only Disease Burden Models (Country-Year Level)

Below, we retrained models with features limited to only climate predictors and using death data from the GBD data to examine how much of the national variation in disease burden can be explained by climate variables in and of themselves.

Infectious Disease Model

The best infectious disease model used a random forest and performed extremely well ($R^2 \approx 0.96$), which indicated nearly all observed variation in averaged infectious disease burden could be explained by climate variables. Feature importance analysis showed that this relationship was overwhelmingly driven by temperature, which accounted for roughly 97% of the model's feature importance, while precipitation, drought, and flooding contributed only marginally.

- Temperature contributing ~97.18%
- Precipitation contributing ~2.42%
- Extreme weather events (including flood and drought) contributing ~0.41%

These findings confirmed that the burden of infectious diseases was directly and strongly related to climate conditions.

Respiratory Disease Model

In contrast, the model for respiratory diseases, when restricted to only climate predictors, had a much lower explanatory power, with $R^2 \approx 0.46$. This indicated that climate alone did not adequately explain respiratory disease outcomes. Instead, non-climatic factors, most notably air pollution and smoking, were suspected to be the dominant drivers.

Cardiovascular Disease Model

The model of cardiovascular disease demonstrated middling performance using climate variables, $R^2 \approx 0.56$. This finding was expected because cardiovascular diseases had many other possible contributing influences based on chronic lifestyle factors, aging, and access to healthcare outside of direct climate exposure. Nonetheless, this was still a notable improvement in performance when compared the equivalent models trained on only the GCHIT dataset.

Summary

Collectively, these models indicated a strong link between climate and infectious disease, as well as moderate climate contribution to respiratory disease outcomes and cardiovascular disease burden. These results emphasized that the health impacts from climate change can be

highly disease specific and should be appropriately considered in predictive health assessments.

Visualizing Climate Disease Links

Climate Patterns and Vector-Borne Disease Clustering

Our visualizations identified strong geographic and climate patterns in vector-borne disease distributions. Regions with average temperatures above 24°C and total annual precipitation over 1500 mm consistently showed a disease risk greater than 0.70. This clustering remained regardless of air quality, suggesting that climate factors rather than pollution were the dominant driver of vector disease distribution. Please refer to the plots in Appendix A.6 for more information.

Income and Pollution Disparities

The overlays of income, air quality, and disease rates showed clear environmental inequalities. Low-income countries had PM2.5 ≈ 50-65 μg/m³, while high-income countries had PM2.5 ≈ 12–18 μg/m³. The disparities in pollution correspond with higher respiratory disease burdens in low-income countries For more information, please refer to the plots in Appendix A.9.

Drivers of Respiratory Disease

Our analysis showed that respiratory disease burden was primarily motivated by air pollution and income level, not climate factors. Our plots demonstrated that PM2.5 levels were strongly associated with respiratory deaths ($r \approx 0.68$), while temperature had virtually no link ($r \approx 0.01$). This distinction helped separate pollution-driven diseases from climate-driven ones in a visually intuitive way. For more information, please refer to the plots in Appendices A.6, A.9, and A.10.

Key Messages: Climate and Health Interactions

Climate-Disease Relationships

Infectious and vector-borne diseases exhibited strong climate sensitivity, with correlations of $r \approx 0.45$ - 0.60 linked to higher temperatures, rainfall, and extreme weather. These conditions foster unhealthy environments where disease-carrying vectors such as mosquitos and ticks thrive. Some tropical regions maintain vector-risk levels above 0.70 even when PM2.5 was low (below 20 μg/m³), which demonstrated that climate alone can sustain high disease risk even in regions with good air quality.

Pollution-Health Relationships

Respiratory diseases were encouraged primarily by air pollution, not climate. Strong correlations with PM2.5 ($r \approx 0.68$) and AQI ($r \approx 0.60$) highlighted pollution as a major determinant of respiratory disease burden, overshadowing the influence of factors like temperature and other climate factors.

Climate-Insensitive Diseases

Cardiovascular diseases appear to be largely climate-insensitive, showing a very weak correlation with climate variables ($r < 0.20$). Climate features contributed less than. Cardiovascular outcomes were predominantly shaped by factors such as healthcare access, and socioeconomic status.

Income and Exposure Patterns

While higher income typically reduced overall disease burden, extreme climatic conditions could override socioeconomic protection. For instance, in high-income countries with cleaner air, bouts of intense heat and heavy rainfall elevated infectious and vector-borne disease risks.

Limitations

There were several limitations to this analysis that are important to keep in mind. Although the GBD and GCHIT datasets were successfully combined and all 648 country-year combinations matched correctly between them, some important constraints remained due to how the data were structured and what information was available.

Firstly, climate and air quality data had to be averaged at the national level and by year. This meant the analysis on the merged dataset could not capture seasonal or short-term changes, such as heat waves, extreme rainfall events, or sudden pollution spikes, which are often closely linked to health outcomes. In addition, national averages could have hidden large differences within countries. For example, PM2.5 air pollution levels can vary by roughly 20-40 μg/m³ within a single country, yet the models had to be trained with only one national value per year.

Secondly, several major factors that strongly influenced health outcomes were not included in the dataset. These include smoking prevalence, as well as healthcare access, health-system capacity, vaccination coverage, and vector-control efforts. As these factors were missing, it was difficult to fully separate the effects of climate from other non-climate drivers of disease.

Lastly, broader structural and social factors such as healthcare infrastructure, economic conditions, and a country's ability to adapt to climate risks were not directly modeled. As a result, while the analysis identified clear relationships between climate variables and health outcomes, it should be interpreted as showing associations rather than direct cause-and-effect relationships.

Conclusion

This integrated analysis of the GCHIT and GBD climate-health data underlined the strong and complex association between ambient environmental conditions and disease burden. Overall, these results demonstrated that infectious and vector-borne diseases are highly sensitive to climate variables, especially temperature, precipitation, and extreme weather, whereas respiratory diseases tend to be more associated with air pollution. Cardiovascular outcomes seem to be relatively independent of direct climate factors, but rather influenced by socioeconomic and healthcare-related factors.

The income level represented a crucial mediating factor, since generally, richer countries face lower pollution and have lower rates of many diseases. Yet, the presence of vector-borne diseases in warm, humid climates indicated that climate-related health risks still exist even in some of the wealthier countries with lower pollution.

Overall for hypotheses 1 and 3, proposing that specific environmental factors were key drivers of respiratory and infectious disease deaths, respectively, there was sufficient evidence to accept them. For hypothesis 2, however, proposing a link between temperature and cardiovascular disease, the evidence was in favour of rejecting it.

While predictive models hold promise, especially for infectious diseases, limitations in data granularity and geographic scope preclude overreliance on forecasting. Recommendations for future work include a need for more granular and timely data at a subnational level, especially with respect to health system capacity and policy interventions. These additions could improve model performance and support more reliable extensions of this framework toward future disease burden projections under climate change. Furthermore, the use of more advanced modeling approaches and targeted outlier analyses may help capture complex relationships and explain deviations from observed global trends.

This report calls for the urgent inclusion of climate factors into global strategies on health and policy decisions. Integrating environmental data into health planning will more fully prepare countries to meet the challenges of a changing climate in terms of health.

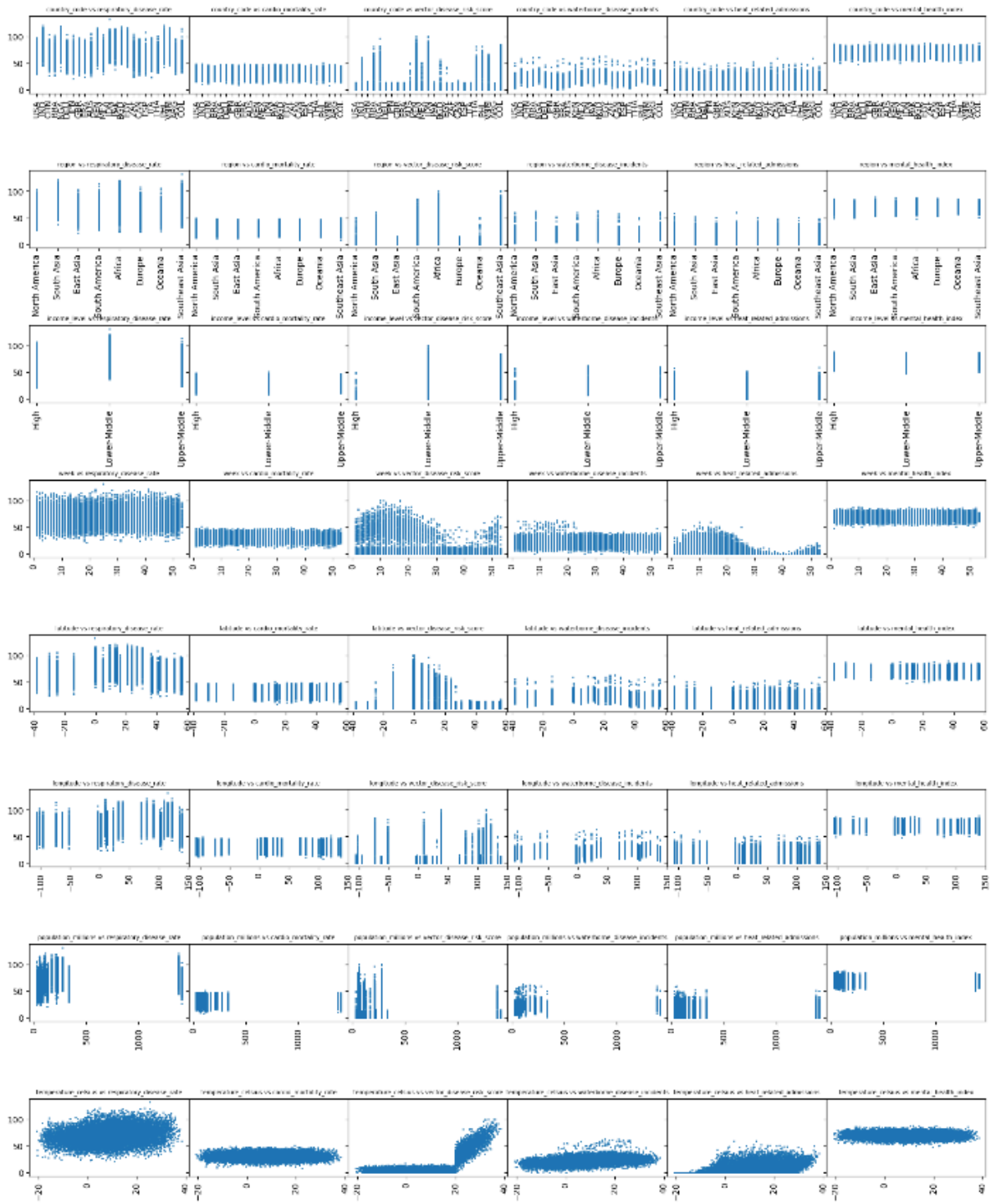
Works Cited

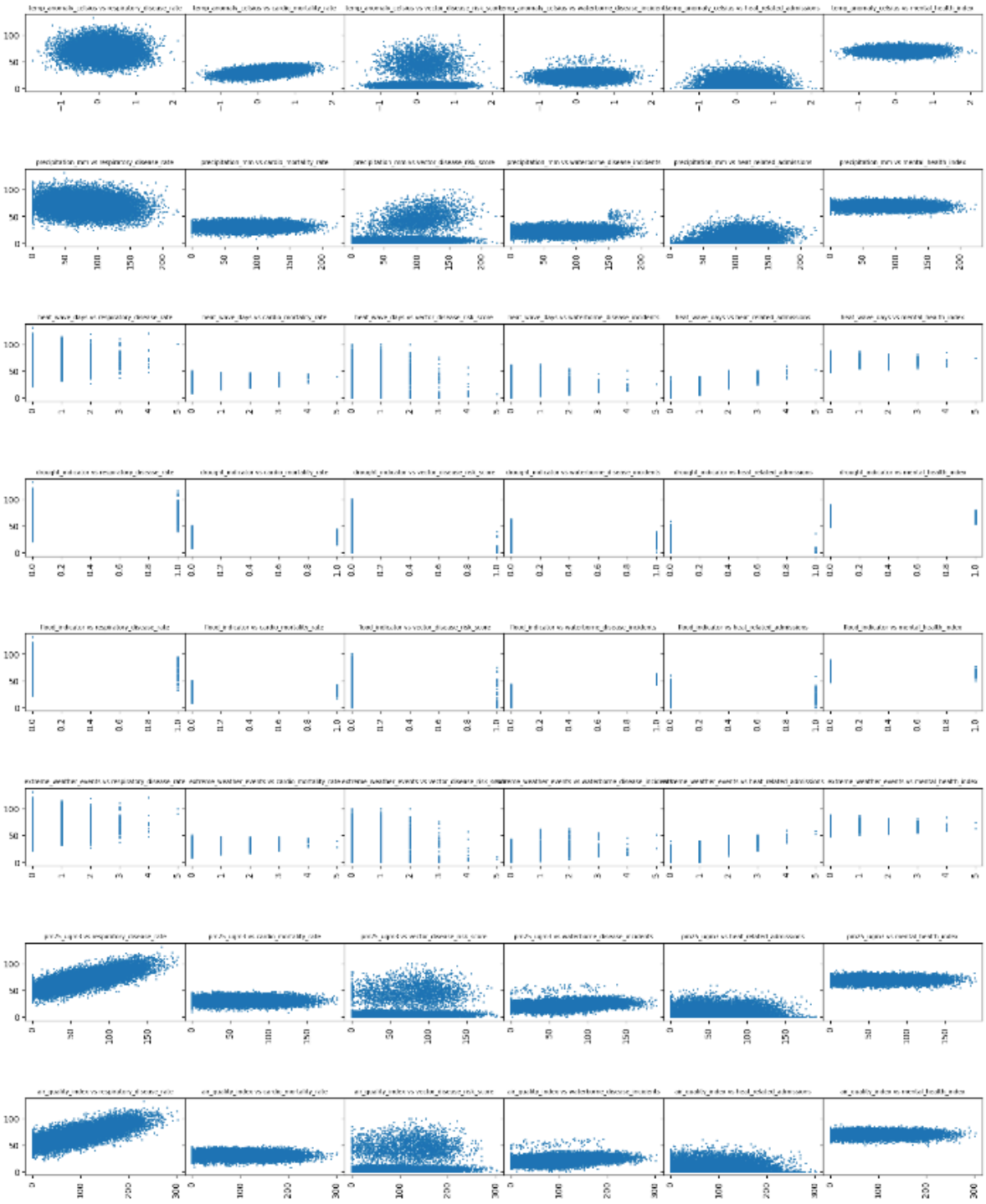
"Global Burden of Disease (GBD)." *Interactive data visuals*, Institute for Health Metrics and Evaluation, https://vizhub.healthdata.org/gbd-results/. Accessed 9 December 2025.

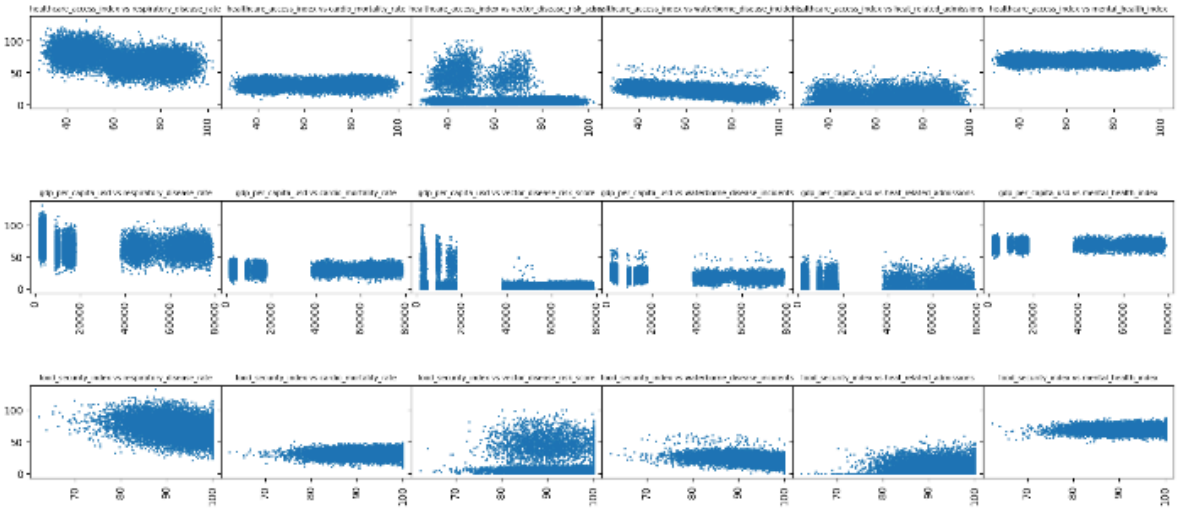
"Global Climate-Health Impact Tracker (2015-2025)." *Kaggle*, https://www.kaggle.com/datasets/sohumgokhale/global-climate-health-impact-tracker-2015-2025. Accessed 9 December 2025.

Appendix A: Plots

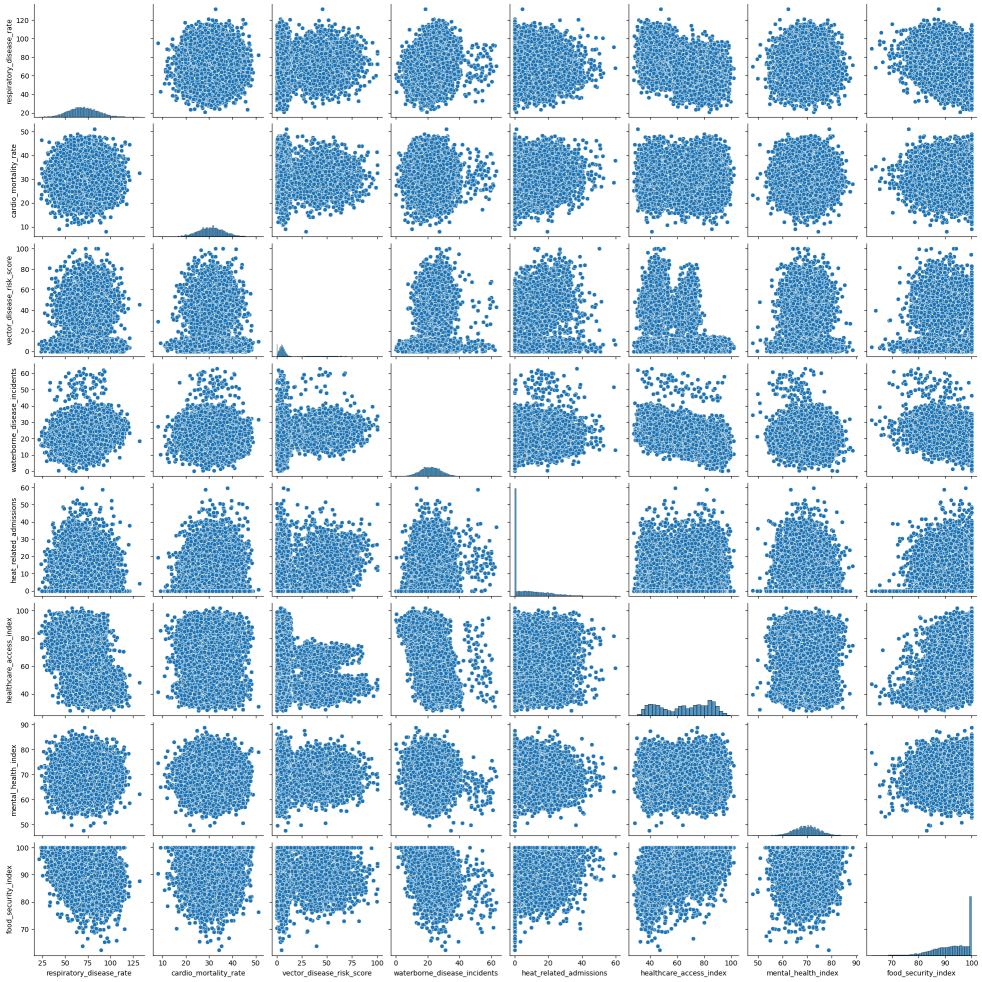
A.1: GCHIT - All Features vs. All Health Targets



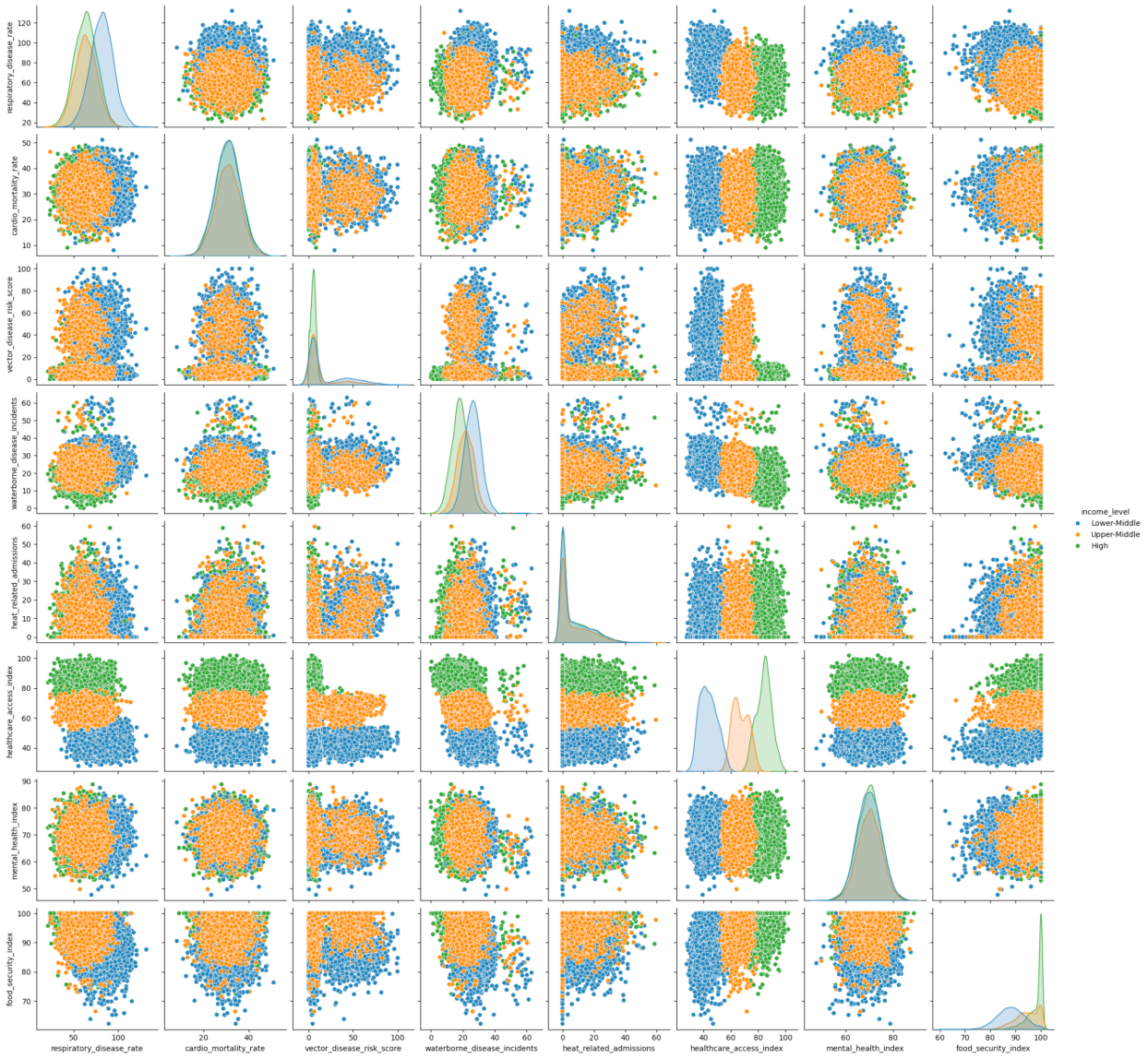




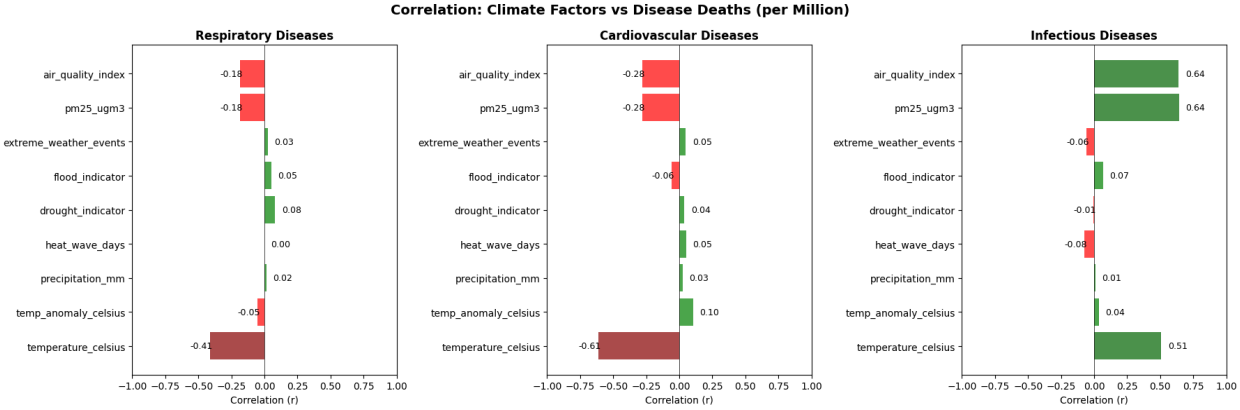
A.2: GCHIT Pair Plots - Health Indicators



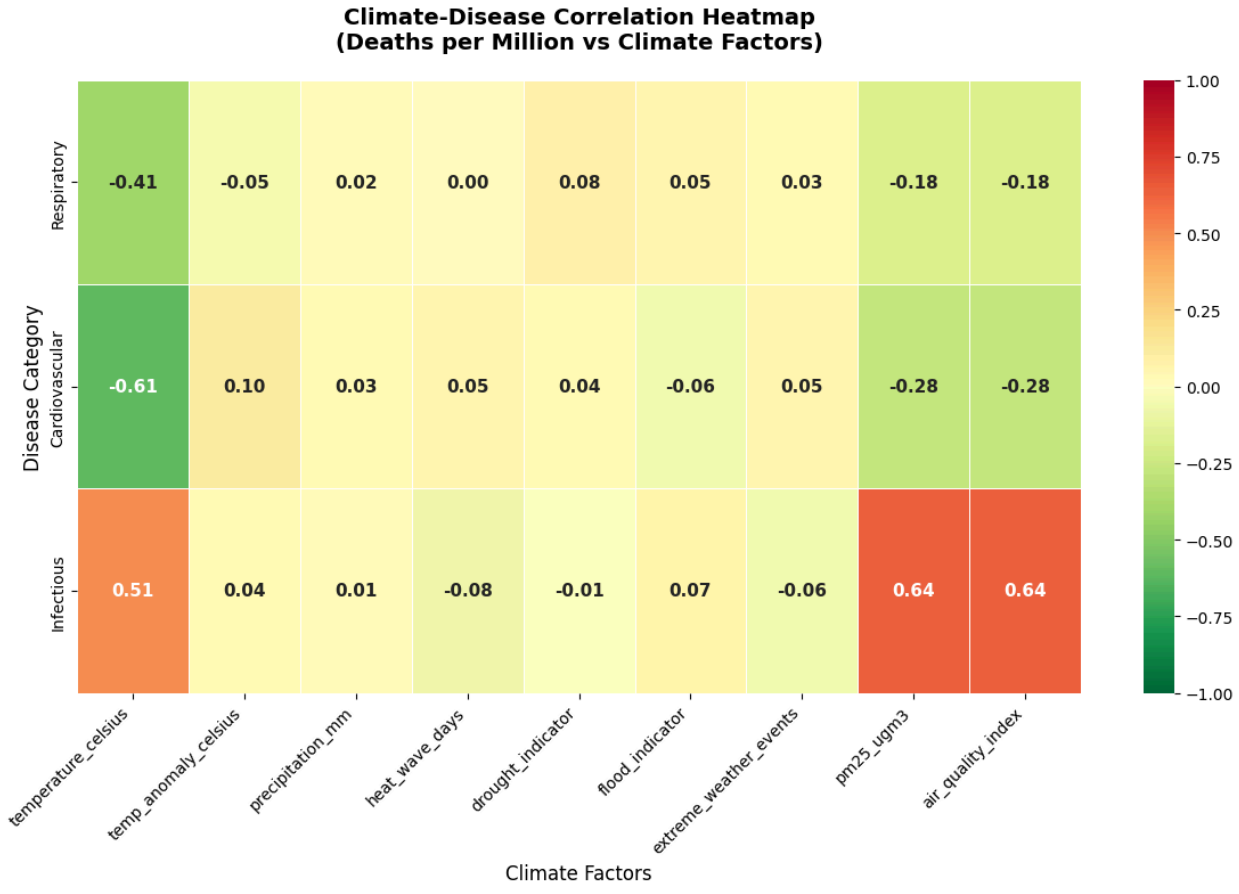
A.3: GCHIT Pair Plots - Health Indicators by Income Level



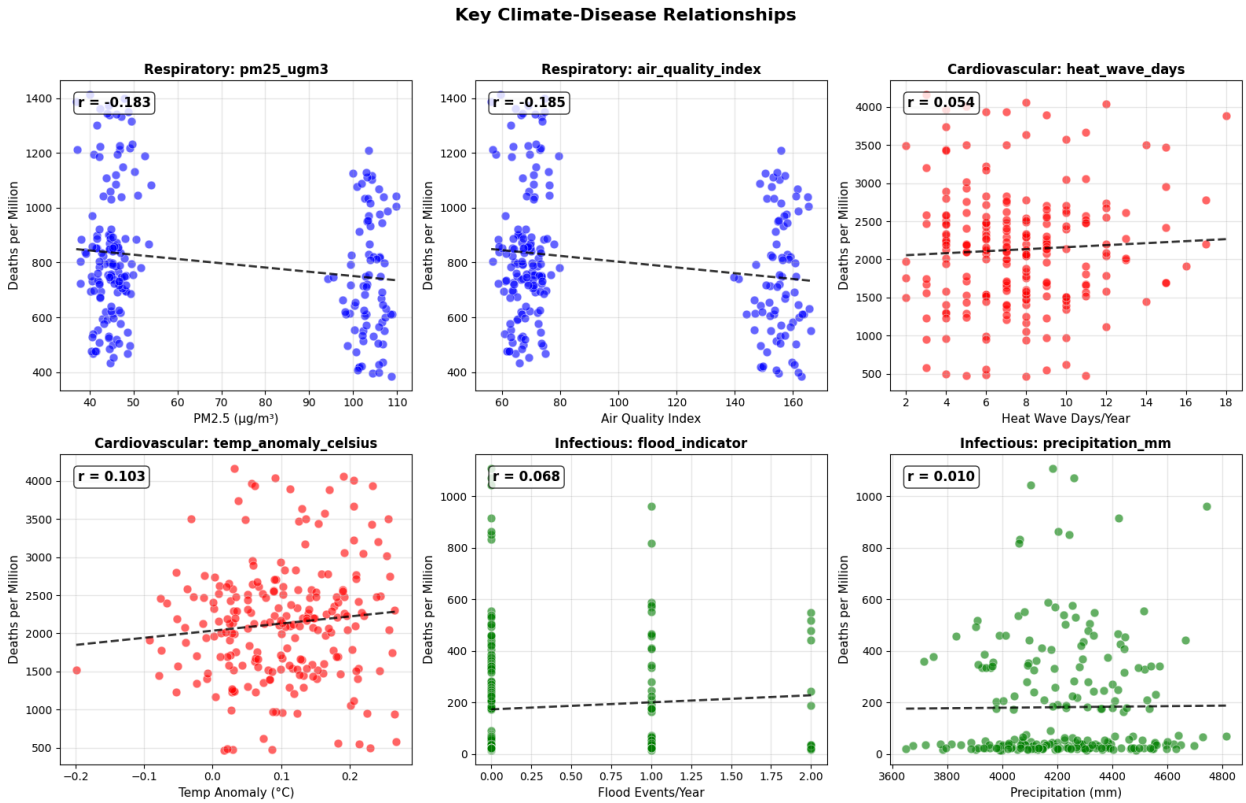
A.4: Correlation of Climate Factors vs. Disease Deaths by Type



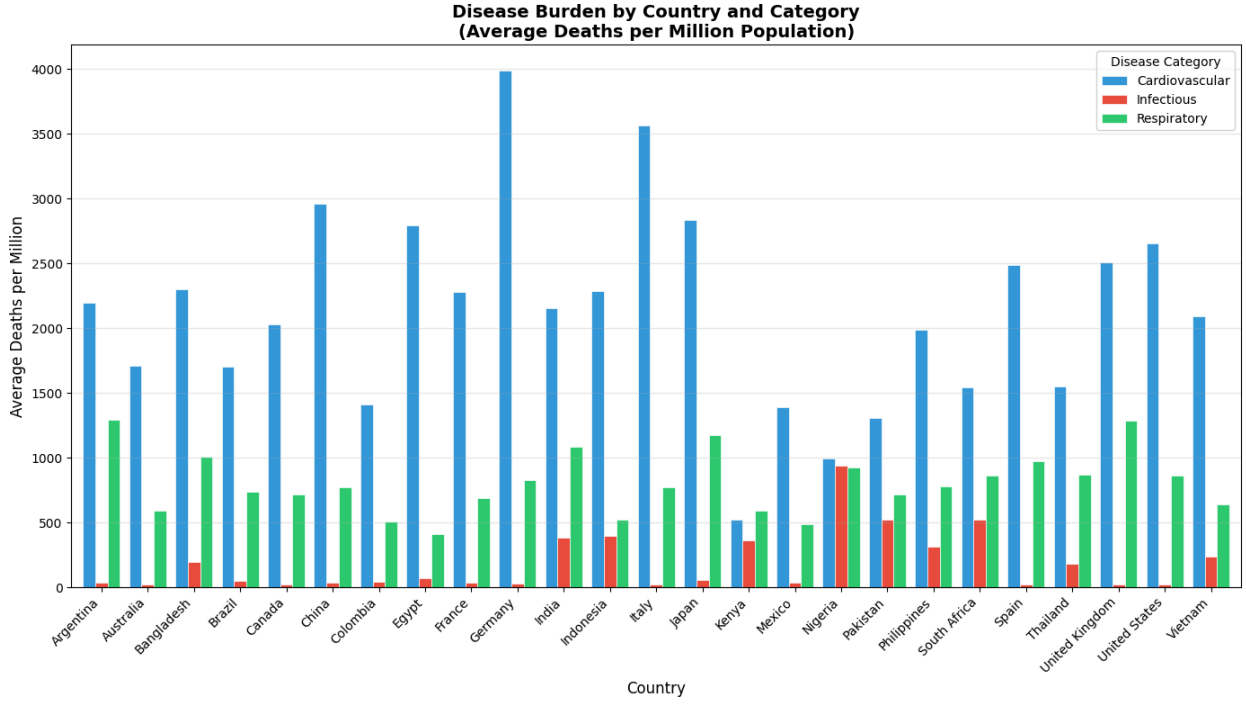
A.5: Climate-Disease Correlation Heatmap



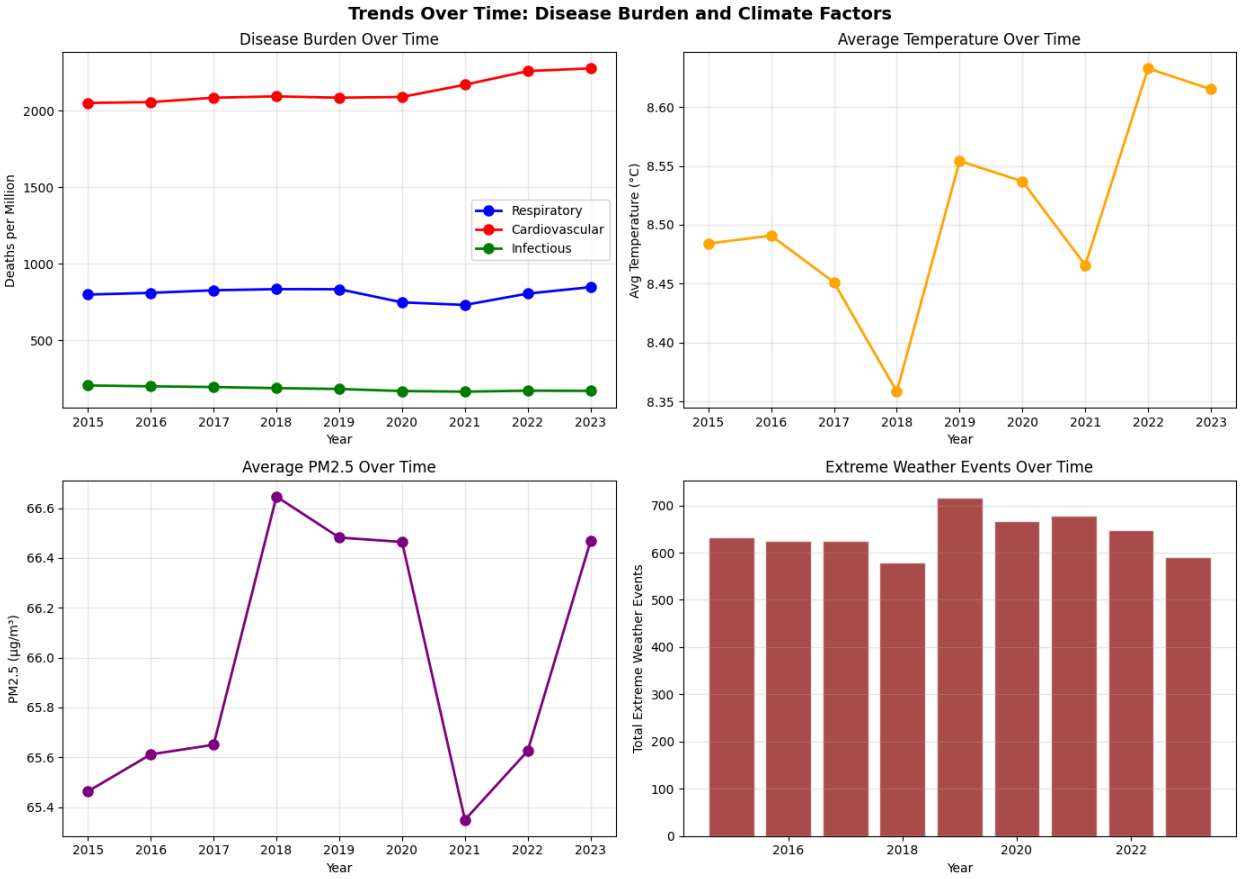
A.6: Key Climate-Disease Relationship Plots



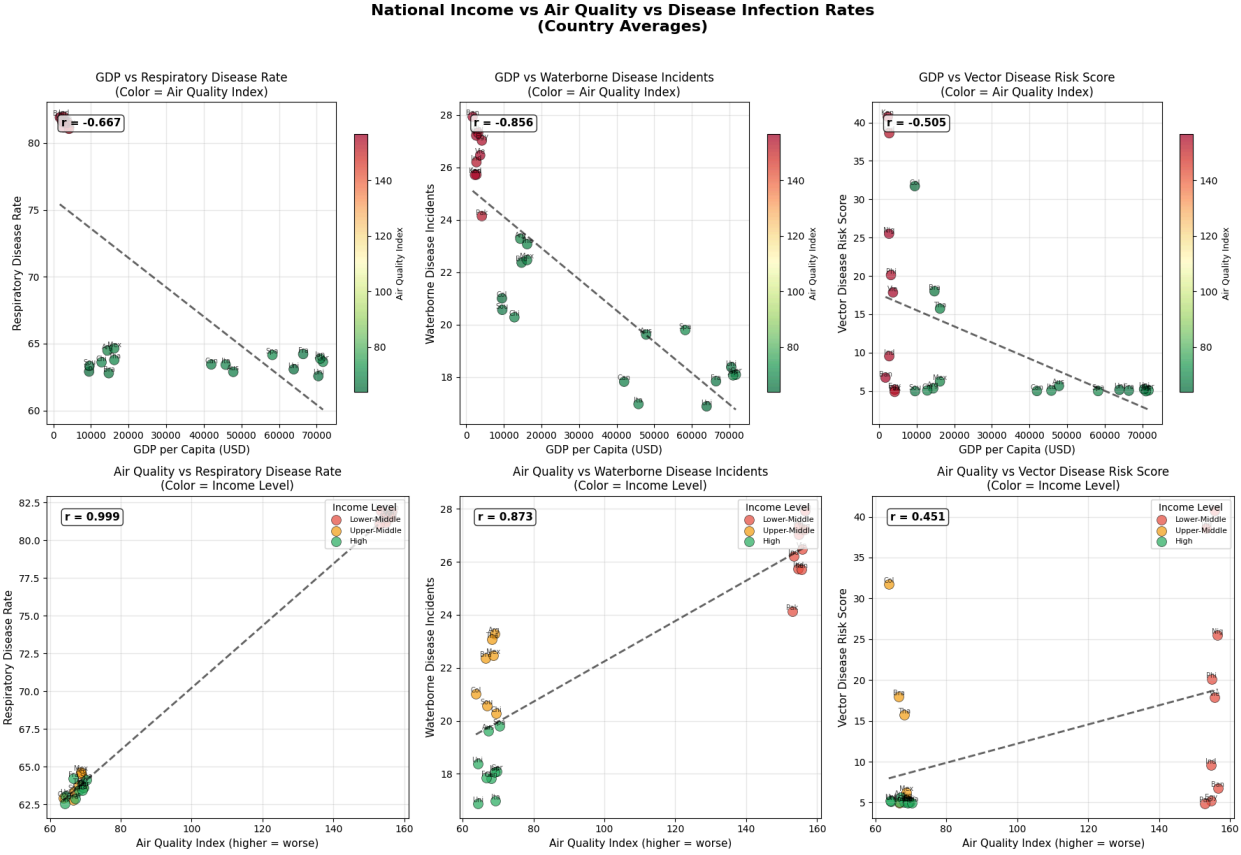
A.7: Disease Burden by Country and Disease Category



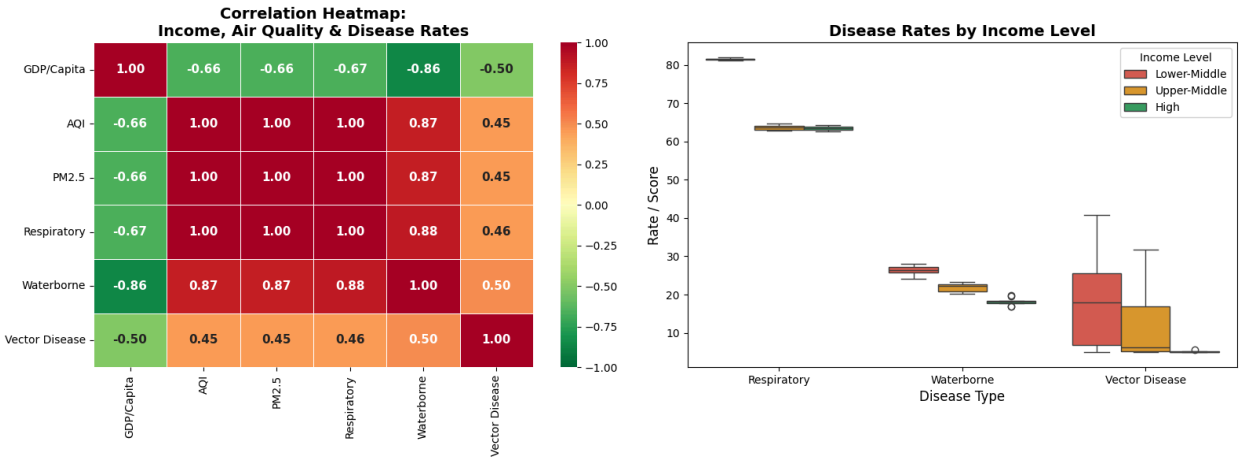
A.8: Disease Burden and Climate Trends Over Time



A.9: National Income vs Air Quality vs Disease Infection Rates



A.10: Correlation Heatmap and Disease Rates by Income Level



Appendix B: Tables

B.1: Most Important Features in Best Random Forest Models

respiratory_disease_rate	cardio_mortality_rate	vector_disease_risk_score	waterborne_disease_incidents	heat_related_admissions	mental_health_index
pm25_ugm3	temp_anomaly_celsius	temperature_celsius	healthcare_access_index	week	precipitation_mm
precipitation_mm	precipitation_mm	precipitation_mm	flood_indicator	heat_wave_days	temperature_celsius
temp_anomaly_celsius	temperature_celsius	temp_anomaly_celsius	precipitation_mm	temperature_celsius	temp_anomaly_celsius
temperature_celsius	pm25_ugm3	healthcare_access_index	temp_anomaly_celsius	precipitation_mm	healthcare_access_index
air_quality_index	gdp_per_capita_usd	food_security_index	temperature_celsius	temp_anomaly_celsius	pm25_ugm3