Forecasting the Spread of Dengue Outbreaks with a Synthesis of Machine Learning Models Utilizing Exogenous Variables

Amulya Gottipati and Sreeja Iragavarapu

Academies of Loudoun

Corresponding author's Email: amulya.v.gottipati@gmail.com

Author Note: The authors are students enrolled at the Academies of Loudoun, a public institution and STEM magnet school in Leesburg, VA. A heartfelt acknowledgment is extended to Zachary Minchow-Proffitt, whose mentorship and guidance were pivotal in the development of this project.

Abstract: Dengue fever, a viral mosquito-borne disease, affects four billion people worldwide, posing economic and health burdens. Unfortunately, there are no antiviral drugs to treat dengue infections, requiring patients to rely solely on palliative treatment. Forecasting future epidemics will aid public officials in implementing mitigation efforts by predicting dengue cases. The purpose of this study was to develop a machine learning model that forecasts the incidence of dengue outbreaks temporally and geographically by utilizing eco-climatic and socioeconomic factors. Methods included preprocessing monthly dengue cases, precipitation, temperature, and socioeconomic datasets from seven countries (between 2014 and 2023) before performing a principal component analysis. A novel topographical feature applied to the model was stagnant water, a critical breeding ground for mosquitoes. A ridge regression technique was used to manage multicollinearity within the data before applying it to the seasonal autoregressive integrated moving average with exogenous variables (SARIMAX) model, which accounts for the seasonality aspect of the variables being examined. Overall, the forecasting algorithm was capable of accurately predicting dengue incidence up to at least six months in advance with a mean absolute error of 2.420e-6. When the novel feature of stagnant water was removed from the datasets, the prediction's accuracy significantly decreased when forecasting for the same time period of six months in advance, demonstrating its importance as a feature when forecasting dengue. Therefore, this algorithm can assist public health officials with planning proactive measures, significantly diminishing economic stress and dengue transmission, thus improving the quality of life in dengue-endemic countries.

Keywords: Dengue Fever, Forecasting, Exogenous Variables, Stagnant Water, Machine Learning

1. Introduction

Mosquitoes are one of the most prevalent insects and an important vector in the transmission of infectious diseases as they are able to easily transmit pathogens and parasites to various species (Dahmana & Mediannikov, 2017). Transmitted to humans through the bites of infected *Aedes aegypti* and *Aedes albopictus* mosquitoes, dengue fever is a tropical mosquito-borne disease that affects approximately four billion people worldwide ("How Dengue Spreads," 2024). In just the past two decades, there has been a significant rise in the worldwide occurrence of dengue fever, posing both economic and health burdens in developing countries. Recently, the historical record for the highest number of reported dengue cases has been surpassed with more than four million cases in 2023 alone ("Dengue – the Region of the Americas," 2023). South America, one of the most severely affected regions, spends three billion dollars annually on dengue treatment (Laserna, Barahona-Correa, Baquero, & Castañeda-Cardona, 2018). Unfortunately, there are no antiviral drugs to treat dengue. Thus, the need for forecasting dengue is especially important due to the delayed response time of public health officials and limited treatment options. Reducing lag time in the future can be done through the early forecasting of epidemics. Specifically, forecasting three months in advance has been shown to aid in preparation in Singapore (Hii et al., 2012).

Current dengue forecasting models account for eco-climatic drivers and socioeconomic factors that influence the transmission of dengue. Understanding these factors is vital for controlling and predicting the spread of dengue ("Dengue emergency in the Americas: time for a new continental eradication plan," 2023). Since mosquitoes thrive in hot and humid climates due to their cold-blooded nature, this makes temperature an essential eco-climatic driver to consider as it affects the growth of mosquitoes and virus reproduction (Farooq et al., 2022). Moreover, poverty influences the risk of dengue as there is an abundance of mosquito breeding grounds present in low-income communities due to poor waste management and sanitation systems (Morgan, Strode, & Salcedo-Sora, 2021). These variables exhibit a seasonal pattern, changing throughout the year, which is crucial for accurately forecasting dengue during specific periods since dengue is a seasonal disease.

It is important to note that precipitation is a common variable accounted for in dengue forecasting algorithms as humidity, which is affected by the relationship between temperature and rainfall, influences the lifespan of a mosquito and its transmission of dengue (Naish et al., 2014). However, to truly consider how precipitation affects the life cycle of a mosquito, a novel topographical feature was applied to the model developed: stagnant water. For the purposes of this study, stagnant water is defined as the seasonal change in water compared to the typical definition of a non-flowing body of water. All mosquitoes require stagnant water to reproduce as the female mosquito lays her eggs in areas that are prone to flooding from rain or irrigation ("Life cycle of Aedes mosquitoes," 2024). It only takes a small amount of water left in bowls, drains, or any open containers to create an inviting habitat for mosquitoes to hatch their eggs. Rainwater or any other human activity that adds water to a pool of stagnant water will trigger mosquito larvae to emerge. Thus, mosquitoes can develop quickly in suitable habitats as stagnant water acts as a breeding ground, making it an essential factor to research when predicting its effect on future dengue outbreaks.

The purpose of this study was to develop a machine learning algorithm capable of accurately forecasting the incidence of dengue outbreaks temporally and geographically in Central South America utilizing eco-climatic and socioeconomic factors acquired from countries in this region. To account for the multicollinearity and seasonality between the exogenous variables being examined, both a ridge regression model and a seasonal autoregressive integrated moving average model with exogenous variables (SARIMAX) were used to provide accurate predictions. Since forecasting three months in advance has been shown to aid public health officials in mitigating future dengue outbreaks in Singapore, it was assumed that doubling this timeframe could be applied to Central South America to account for the socioeconomic differences between the two regions. Central South America has a higher population density and fewer resources, such as healthcare facilities and funding, necessitating additional time for public health officials to allocate the essential resources to mitigate future dengue outbreaks compared to Singapore, a highly developed country, which can access and mobilize these resources more rapidly.

2. Methods

2.1 Data Preparation

Since Central South America was the scope of this research, monthly data from January 2014 to September 2023 was collected from Argentina, Bolivia, Brazil, Colombia, Paraguay, Peru, and Venezuela **(Figure 1)**.

Figure 1. Data Preprocessing Pipeline

(A) Data was collected from Argentina, Bolivia, Colombia, Paraguay, Peru, and Venezuela. The datasets collected include weekly dengue cases from PAHO, temperature data from NOAA, precipitation and humidity from NASA, stagnant water data from Freshwater Ecosystems Explorer, and socioeconomic data from The Global Economy. **(B)** Standard preprocessing was conducted, although a few datasets had additional steps to be formatted correctly. **(C)** Dengue data was decomposed into its trend, seasonality, and residual components to visualize the seasonality distribution of dengue. **(D)** Principal Component Analysis (PCA) reduced dimensionality and multicollinearity between features and created 15 new principal components while keeping a majority of the variance.

The collected data included various exogenous variables such as eco-climatic and socioeconomic factors that impact dengue transmission. In order to build a time series model that forecasts dengue cases across the entire Central South American region, the data points for each feature from each individual country were averaged across all the countries being assessed for the specific region of interest. For example, the average temperature data from Argentina, Bolivia, Brazil, Colombia, Paraguay, Peru, and Venezuela were combined and averaged to create a single average temperature value for the whole Central South American region. This process was repeated for all features. These data points were averaged to create a single, unified dataset representing the overall dengue trends in Central South America, thus allowing for a regional analysis rather than a country-specific one. These datasets were acquired from both national and international repositories and were appropriately formatted

using standard preprocessing techniques as further described. To begin, historical dengue data, which consisted of the number of dengue cases, was collected cumulatively from the Pan American Health Organization (Gutiérrez, L. A, n.d.). However, the cumulative dengue cases were converted to non-cumulative data so that the model was able to discern a relationship between the change in the number of dengue cases in comparison to the change in other exogenous variables. Precipitation and humidity data were acquired from NASA by outlining polygons of landmass corresponding to the shapes of the countries being explored in Central South America on a world map **(Figure 2)**.

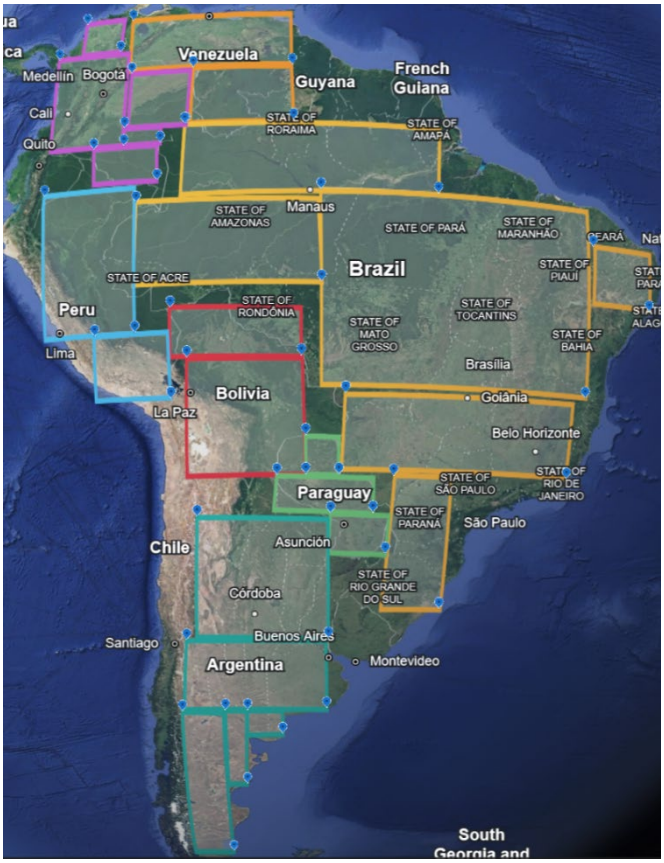

Figure 2. Acquisition of Precipitation & Humidity Data by Polygons of Landmass
Precipitation and humidity data was collected by creating polygons of landmass per country, although it was ultimately averaged to get an accurate representation of the weekly data in each country.

By creating these polygons, daily precipitation and humidity data were able to be downloaded specific to the outlined countries, thus providing an accurate representation of the overall conditions in each country (Data, M. N, n.d.). Daily temperature data measured in Fahrenheit was collected from the National Oceanic and Atmospheric Administration ("National Centers for Environmental Information," n.d.). Data on stagnant water, or the amount of seasonal change in water, data was collected from Freshwater Ecosystems Explorer, which measured the water bodies in square kilometers ("FreshExplorer," n.d.). For the purposes of this model, stagnant water is defined as the seasonal change in water compared to the typical definition of a non-flowing body of water. Finally, socioeconomic data, which included factors such as the average poverty ratio, was collected from The Global Economy ("Global Economy, World Economy," n.d.). Although eco-climatic drivers directly influence the transmission of dengue, it is also important to consider socioeconomic factors, particularly the average poverty ratio. Low-income communities often have inadequate housing and sanitation, which can lead to the accumulation of standing water and create breeding grounds for mosquitoes, facilitating the further spread of dengue. It is important to note that pivot tables in Excel were used to convert the original data format of variables without monthly data into a standardized monthly format, ensuring consistency across all variables.

After downloading the data for each exogenous variable and compiling them into one large dataset, there were several NaN values throughout the dataset. NaN is short for "Not a Number," and this could represent empty cells with no value or cells with unrepresentable values such as dividing by 0. Since machine learning models are unable to discern patterns from datasets with NaN values, the NaN values in each column for an exogenous variable were filled by taking the mean of all of the previous values in that column. However, if a year's worth of data was missing, then the NaN values were dropped as the seasonal patterns for that year were unknown and wouldn't make a significant contribution to making accurate dengue forecasts.

Once the historical dengue data was cleaned and pre-processed, it was later decomposed into its trend, seasonality, and residual components to visualize the seasonality distribution of dengue **(Figure 3)**.

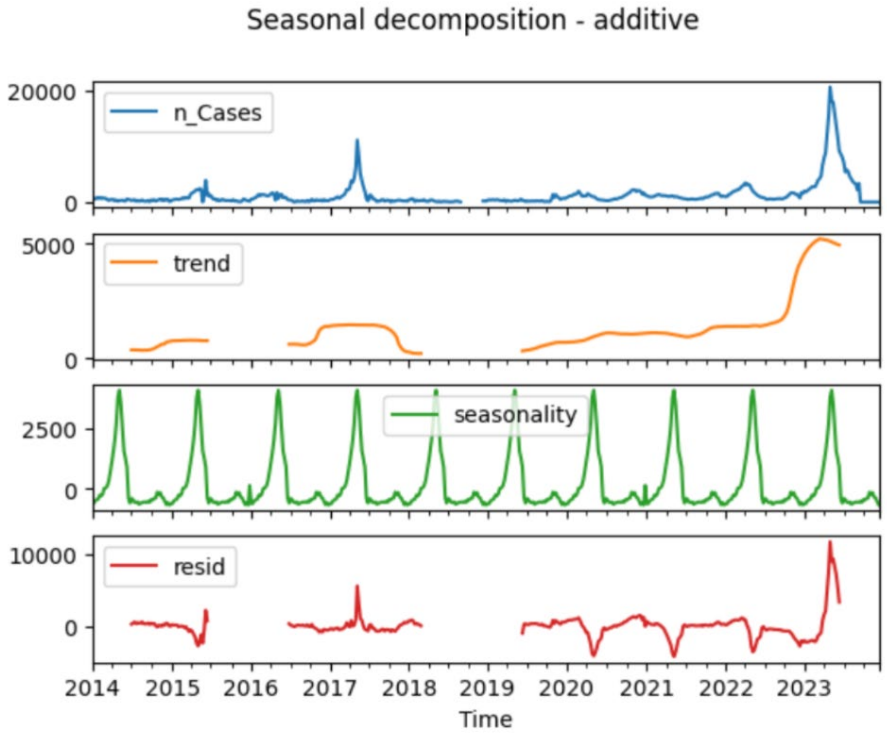

Figure 3. Seasonal Decomposition of Dengue Data in Peru

The seasonality element of dengue in Peru is shown through the seasonal decomposition of time. The historical dengue data was decomposed into its trend, seasonality, and residual components to visualize the seasonality distribution of dengue.

The trend component represents the overall direction and long-term movement of the dengue data. The seasonality component identifies the seasonal patterns in the dengue data by applying locally weighted scatterplot smoothing (LOESS) to the dataset. LOESS fits a smooth curve to the observed dengue data, allowing it to determine local patterns. By identifying the local patterns in the dataset, it is then able to identify the seasonal patterns. The residual component of the data captures the variability that is not accounted for by the seasonality and trend components. This is also known as the noise component as it captures the random fluctuations in the dataset ("Seasonal-Trend decomposition using LOESS (STL)," n.d.).

2.2 Principal Component Analysis

After the data was cleaned and pre-processed, there were a total of 44 features in the dataset **(see Appendix)**.

A principal component analysis was performed to simplify the large, compiled dataset of 44 features into a smaller set while still maintaining a majority of the variance and significant trends in the data. Another method of dropping features

from a dataset is by using a correlation matrix, which summarizes the correlations between each feature in the dataset through correlation coefficients **(Figure 4)**.

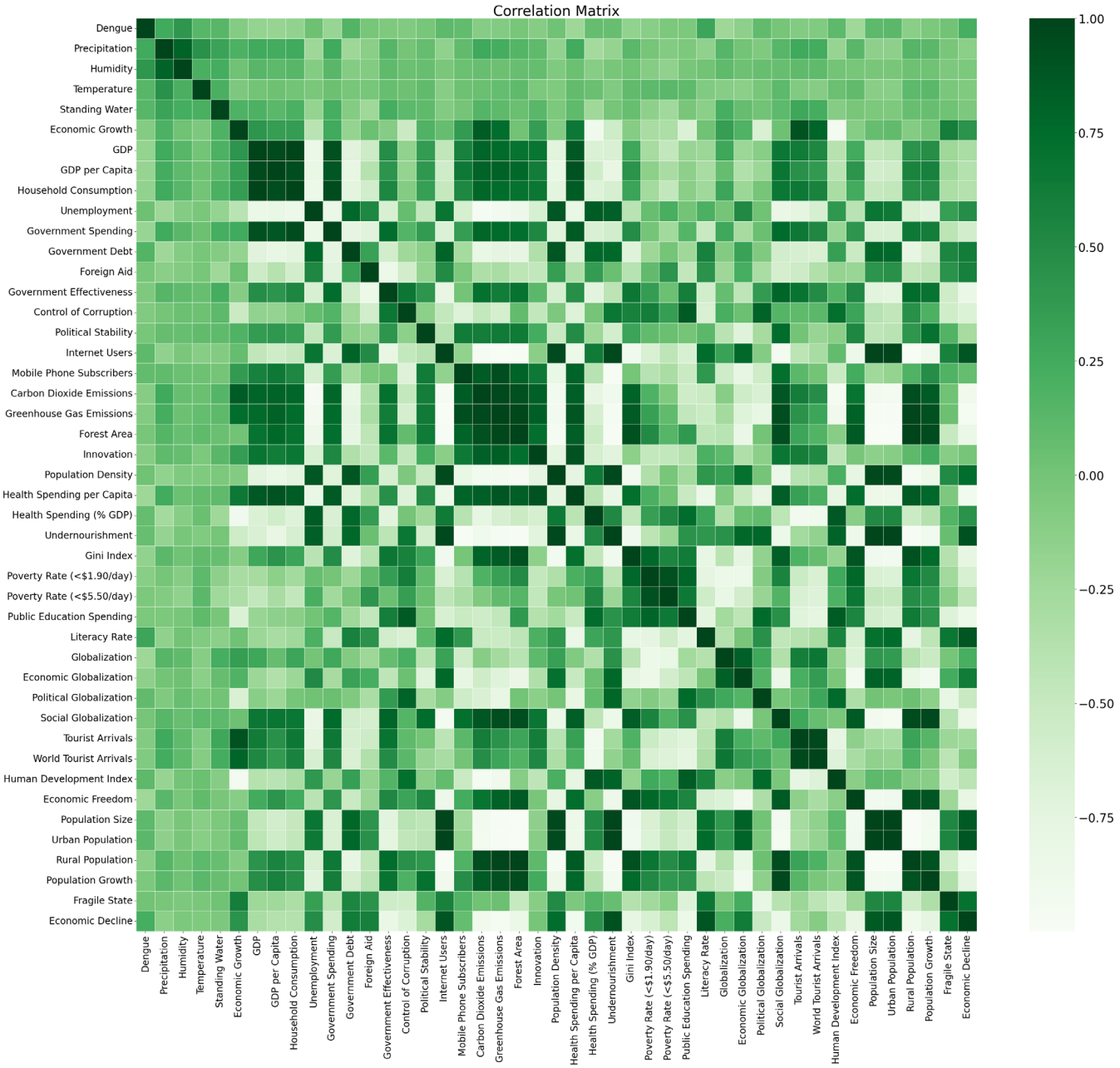

Figure 4. Correlation Matrix of Original Features in Compiled Dataset of Central South American Data

The correlation matrix shows the relationships and correlation coefficients between the original features in the dataset, where dark green indicates a high positive correlation and light green indicates a high negative correlation **(see Appendix for full feature names)**.

Correlation coefficients are used to measure the strength and direction of relationships between features and the target variable and between the independent variables (among features). A correlation coefficient greater than zero indicated a positive relationship, a value less than zero signified a negative relationship, and a value of zero indicated no relationship between the two variables. The correlation matrix indicates which features are highly correlated with each other, which is important during feature selection as one of the two highly correlated features can be dropped from the dataset if they have the same effect on the target variable. However, instead of manually determining what features to drop based on the correlation matrix, a principal component analysis was performed to reduce the dimensionality and multicollinearity between features, thus performing the same task. The principal component analysis began with standardizing the dataset to ensure that certain variables with different units are not over or under-exaggerated when checking for variance. Standardization was done such that each column's mean was 0 and standard deviation was 1. A covariance matrix was then created to measure each variable's relationship direction and strength. From the covariance matrix, the principal components and their variance were determined by eigenvectors and eigenvalues. Eigenvalues are the coefficients of each eigenvector and are determined through the covariance matrix, showing how much variance of the dataset is in each principal component (Jaadi, 2024). Then, the principal components were ordered through a scree plot so that 90% of the total variance of the original dataset was in the principal components that were being used. This resulted in the formation of a new dataset, consisting of the first 15 principal components from the scree plot.

2.3 Ridge Regression Analysis

The purpose of the ridge regression analysis is to reduce the multicollinearity in the dataset, ensuring that the forecasting model remains stable and reliable even when additional time series data is added, hence improving the accuracy of its predictions **(Figure 5)**.

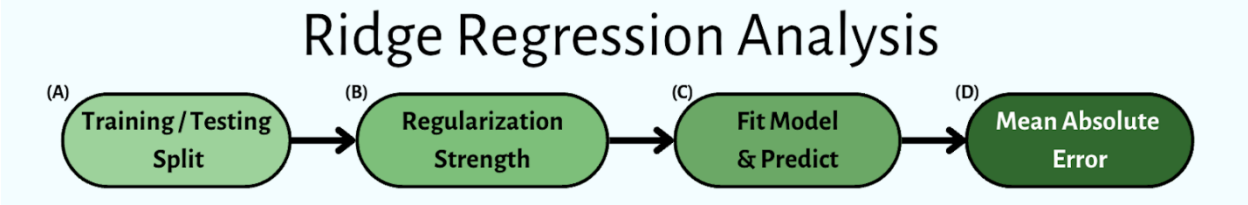

Figure 5. Ridge Regression Pipeline
(A) An 80/20 training-testing split was used. **(B)** The regularization strength determines penalties on the coefficients, which impacts the variance of the model. **(C)** The model was fit and predicted. **(D)** Mean absolute error was calculated after tuning regularization strength each time.

Multicollinearity is when different independent variables, the features that impact dengue transmission, within the dataset are largely correlated or interdependent, and can negatively impact forecasting through a lack of precision with the predictors and coefficients of the correlated variables.

The ridge regression analysis began with creating an 80-20 training-testing split within the data, which consisted of the 15 principal components created after performing the principal component analysis on the original dataset. The main parameter being changed to reduce multicollinearity was the alpha value or regularization strength parameter, lambda. A higher alpha value indicated greater multicollinearity in the dataset and applied a stronger penalty on the coefficients compared to a lower alpha value. A large alpha value leads to high bias, related to underfitting, and a small alpha value leads to high variance, related to overfitting (Navelski & Odongo, 2021). This means that an appropriate alpha value must be found that can properly trade off bias and variance. This penalizing hyperparameter was tuned by having the mean absolute error of the model be as low as possible. This was done because the mean absolute error can determine how well the model fits the data, and changing the alpha value affects the model's prediction. The mean absolute error is the mean size of mistakes in collected predictions and is ideal for assessing the model's accuracy rate as it is not sensitive to outliers. Thus, the specific alpha value that resulted in the lowest mean absolute error value was identified through tuning. For the purpose of this study, ridge regression was used to ensure that there was no more multicollinearity within the dataset. As the alpha value was almost 0, there was no need to further manipulate the data prior to forecasting using the SARIMAX model.

2.4 Seasonal Autoregressive Integrated Moving Average Model with Exogenous Variables

The SARIMAX model was used to account for the seasonality patterns of each exogenous variable and dengue itself in order to accurately forecast dengue incidence **(Figure 6)**.

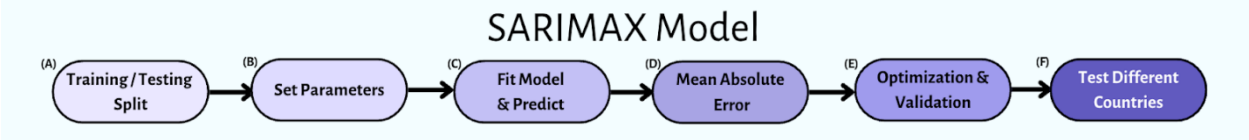

Figure 6. SARIMAX Model Pipeline

(A) A time 6 months before the end of the dataset was chosen as the end of the training data, but this time was changed many times to see the seasonal pattern of prediction, however, 6 months was the desired end to the training data. **(B)** The number of autoregressive, seasonal, and averaging terms are decided through (p, d, q) x (P, D, Q, s) values were set as parameters. **(C)** The model was fit and tested. **(D)** Mean absolute error (MAE) was calculated. **(E)** The model's parameters were optimized 100 times to get the lowest error possible and was then validated by changing the training set length to prevent overfitting. **(F)** The model was tested on data from different countries individually to see if it could adapt to specific data from each country and still obtain high accuracy.

The model is able to account for the seasonality components of dengue and the exogenous variables with its non-seasonal and seasonal parameters. The SARIMAX model is an extension of the SARIMA model, though both differ in their method of accounting for seasonality. The SARIMA model makes predictions by accounting for the seasonal component of the target variable, while the SARIMAX model makes predictions by not only considering the seasonality of the target variable but also the seasonality of exogenous variables. For this analysis, the SARIMAX model was selected for its ability to make accurate forecasts by accounting for the seasonality of dengue while also considering the seasonality of the eco-climatic drivers and socioeconomic factors and their impact on dengue transmission. It is important to note that the SARIMA model is an extension of the ARIMA model as both of the models include the (p, d, q) parameters, which capture the autocorrelation or non-seasonal components in time series data.

The "p" parameter is the order of the autoregressive part of the model, which is the number of lag observations included. This allows the model to capture the relationship between past and current observations by assuming the current value is a linear combination of the previous values. The "d" parameter is the degree of differencing or the number of times the data had its past values subtracted in order to make the time series stationary. If a time series is stationary, this means that its properties do not depend on the time at which the series is observed. So, the series should exhibit constant mean and variance over time. In order to make a time series stationary, the process of differencing is used to subtract the current observation from the previous observation. The "q" parameter is the order of the moving-average part of the model, which is the number of lagged forecast errors. This allows the model to model its error as a linear combination of error terms that occurred in past observations, thus allowing it to help smooth out the noise in the data (Wang, Yao, Hou, Zhao, & Zhao, 2021).

While the ARIMA model specializes in accounting for the non-seasonal components in data, the SARIMA model is designed to handle time series data with seasonal trends. If data has seasonal trends, this means that the data exhibits regular, predictable changes that occur every year (seasonal cycle), which includes daily, weekly, or monthly fluctuations. Thus, the SARIMA model is useful for forecasting as it is able to capture trends and seasonality in data to make accurate predictions. Along with the non-seasonal (p, d, q) parameters, the SARIMA model also has the seasonal (P, D, Q)m parameters.

The "P" parameter is the order of the seasonal autoregressive part of the model, which is the number of lag observations included. The "D" parameter is the seasonal degree of differencing or the number of times the data had its past values subtracted from the seasonal part of the model. The "Q" parameter is the seasonal order of the moving-average part of the model. Finally, the "m" parameter refers to the length of the seasonal period, which in this case is 12 as the SARIMA model is being trained with monthly data with an annual cycle. Essentially, the (P, D, Q) parameters are the same as the non-seasonal (p, d, q) parameters, but they are applied to the seasonal component of the time series data (Brownlee, 2019).

Before training the SARIMAX model, the data was split into training and testing sets in order to train the model on the training set and determine the model's accuracy on the testing set. Since the model is forecasting outwards into the future, a regular 80-20 training/testing split was not performed. Instead, the training set consisted of all of the data points until six months before the end of the dataset and the testing set consisted of all of the remaining data points after. Thus, the data was not split randomly into its training and testing sets and may have introduced bias in these datasets. After the training/testing

split was performed on the dataset, the (p, d, q) x (P, D, Q)m parameters were filled with common values. The model was then fit and tested, and the mean absolute error was later calculated to determine whether the model was able to make accurate forecasts compared to the actual values in the testing set. Since the parameters were filled with common values and other parameter values were not tested to determine if the mean absolute error of the model could be reduced, an optimization function was developed where the parameters were changed with 100 different values to get the lowest mean absolute error possible **(Figure 7)**.

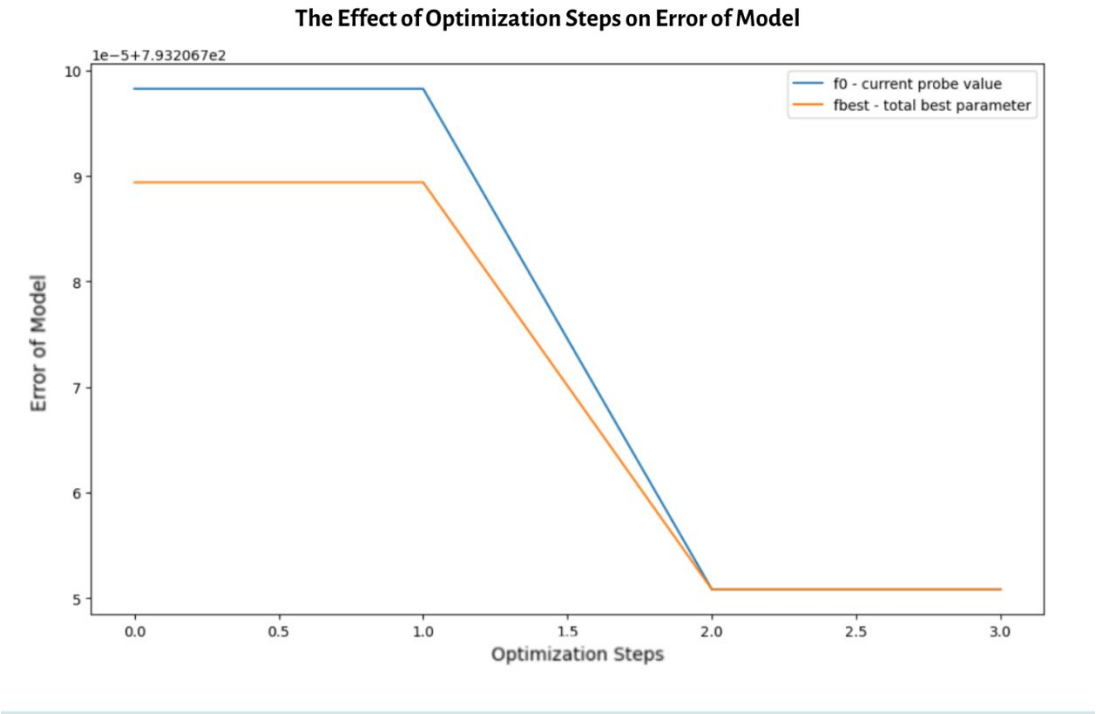

Figure 7. Optimization of Parameters in SARIMAX Model
The graph shows that the error of the model decreases as the parameters are further optimized.

Table 1. Effect of Forecasting Various Seasonal Periods on Error of Model
This displays a table showing different seasonal periods of forecasting and the mean absolute error associated with it, emphasizing the prediction at six months in advance.

Seasonal Period (months)	*Error*
3	1.583e-6
6	2.420e-6
9	2610.196
12	2075.835
18	1652.760
24	2020.065

These parameters were validated by changing the training set length to prevent overfitting. In addition to testing the model's accuracy in forecasting six months ahead in the future, the model was tested on forecasting ahead for various seasonal periods to determine how the seasonality of the model changes the mean absolute error of its predictions **(Table 1).**

This was done by changing the amount of data points in the training and testing datasets. For example, in order to test the model's accuracy in forecasting two years ahead in the future, the training set consisted of all of the data points until two years before the end of the dataset and the testing set consisted of all of the remaining data points after.

The model was also tested on whether it was able to be applied to individual countries in Central South America instead of forecasting dengue incidence in Central South America as a whole **(Table 2)**.

Table 2. Effect of Forecasting in Different Central South American Countries on Error of Model
This displays a table showing the model's mean absolute error when predicting individual countries 'dengue cases using the average of exogenous variables when forecasting six months in advance.

Countries	*Error*
Average of Countries	2314.320
Brazil	1390844.605
Bolivia	12984.806
Colombia	3165.738

This was done by changing the dengue data that the model was originally trained with to only include dengue data for that specific country while the average of the exogenous variables for the other countries in Central South America remained the same.

3. Results

After performing the principal component analysis on the compiled dataset of the Central South American data, a scree plot was used to visualize the amount of variability captured by the first few principal components. The scree plot helped to select the minimum number of principal components required for the ridge regression and SARIMAX models. This was done by determining the point at which the proportion of variance explained by each subsequent principal component dropped, which is also known as the elbow in a scree plot. In this study, 15 principal components were selected as approximately 90% of the total variance was captured with this number of principal components **(Figure 8)**.

Once the correct number of principal components was selected, this data was used to train and test the ridge regression model. As the purpose of the ridge regression model is to reduce the multicollinearity within the dataset by assigning a penalty term, various penalty terms were tested on the model to determine at which value the mean absolute error was at its lowest. It was found that the model had its least mean absolute error value of $1.766e^{-6}$, which is extremely close to 0, when it was set with a penalty term of 0.01. This is a relatively low penalty term, meaning that the dataset had very little multicollinearity. This is most likely due to the fact that the principal component analysis already reduced the majority of the multicollinearity within the dataset, thus resulting in a lower penalty term. Additionally, the graph of the ridge regression model indicates that very little error resulted in its predictions as the actual number of dengue cases in the dataset and the predicted number of dengue cases by the model were extremely close to each other **(Figure 9)**.

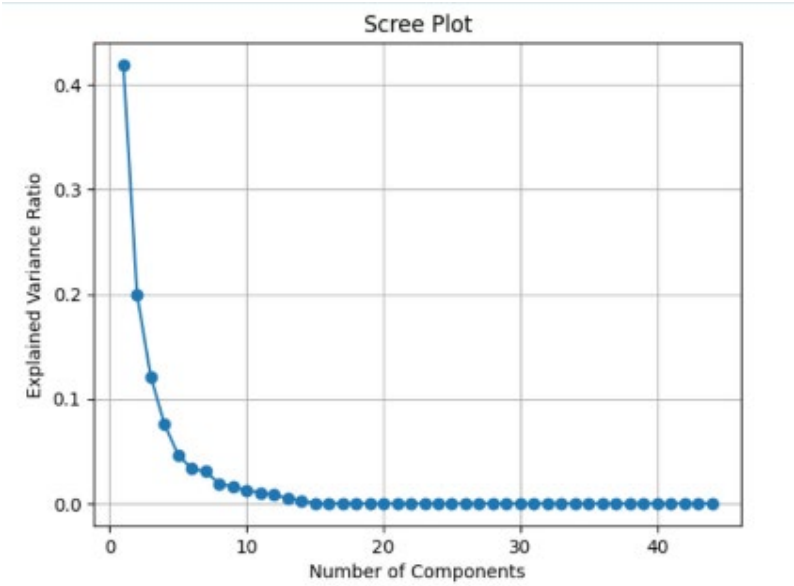

Figure 8. Principal Component Analysis Scree Plot
The scree plot shown was created while performing PCA. This plot shows the cumulative amount of variance captured with the addition of each principal component.

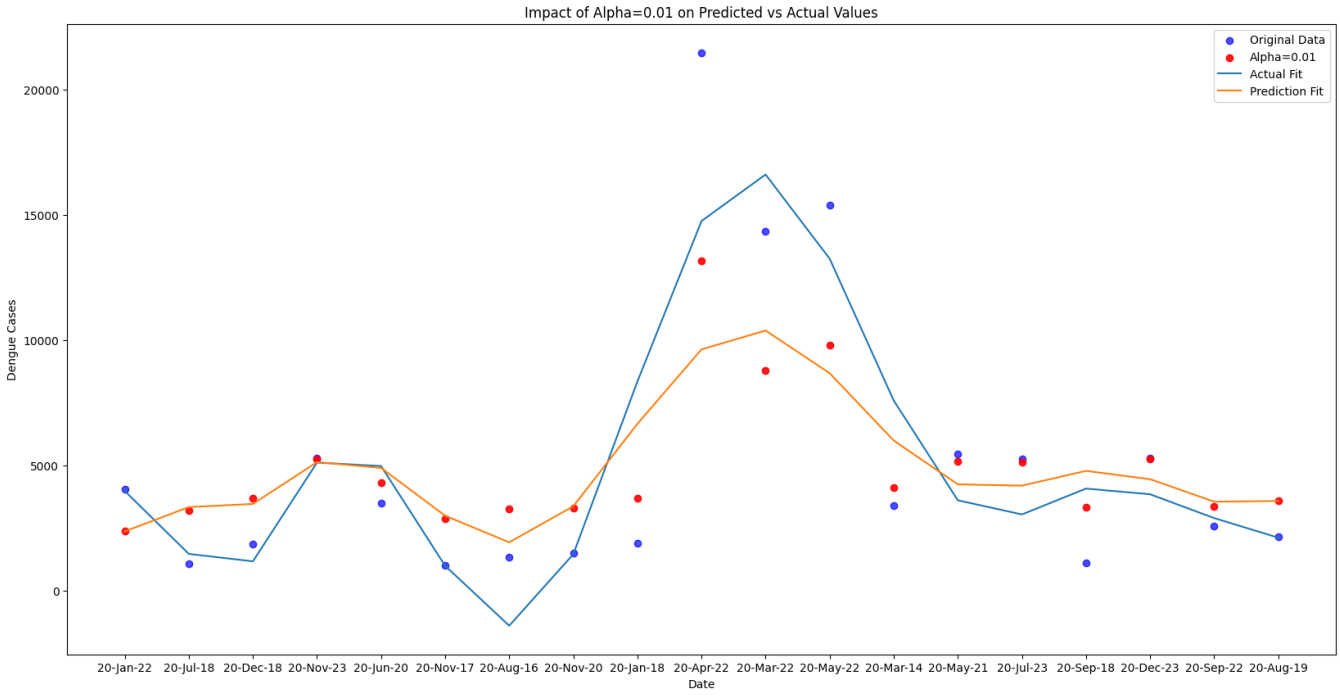

Figure 9. Prediction of Ridge Regression Model
The graph above shows the ridge regression model's prediction with an alpha value of 0.01 and the original forecast.

Since there was very little multicollinearity in the dataset due to the use of the principal component analysis, which can be supported by the graph of the ridge regression model, the 15 principal components were used as a dataset to train the SARIMAX model to forecast dengue. After the parameters were optimized in the SARIMAX model to have the least mean absolute error, the graph of the historical and predicted dengue case values was outputted to determine whether the model's forecasts were accurate. The SARIMAX model had a mean absolute error value of $2.420e^{-6}$, which is very close to 0 when it was used to forecast six months in advance. Additionally, the graph of the SARIMAX model indicates that very little error resulted in its predictions as the forecasted values from the model overlap the actual values in the dataset **(Figure 10)**.

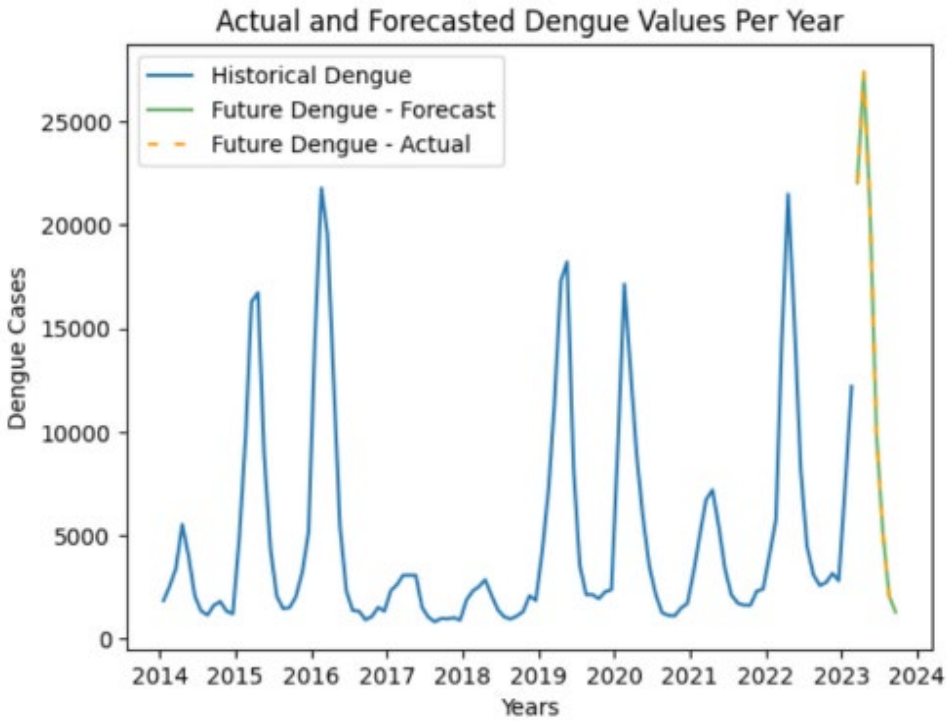

Figure 10. Prediction of SARIMAX Model Forecasting Six Months in Advance Using Average of Countries' Exogenous Variables - This shows the actual and forecasted dengue values when predicting six months in advance, with a mean absolute error of 2.420e-6. It is important to note that the figure above shows the model's prediction when trained with the average of each country's exogenous variables.

The SARIMAX model was also used to predict different seasonal periods such as predicting three months in the future vs. two years in the future in order to determine whether the length of seasonal periods affects the model's accuracy. When the model was used to predict three months in advance, it had a mean absolute error value of $1.583e^{-6}$, while when the model was to predict two years in advance, the mean absolute error value increased to 2020.065 **(Figure 11)**.

To determine whether the SARIMAX model could be applied to individual countries in Central South America instead of forecasting dengue incidence in the entire region of Central South America, the dengue data that the model was trained with was changed to only include dengue data for that specific country while the average of the exogenous variables remained the same. When the model was used to forecast dengue incidence in Brazil, it had a mean absolute error value of 1390844.605 while there was only a mean absolute error value of 3165.738 in Colombia and a mean absolute error value of 12984.806 in Bolivia **(Figure 12)**.

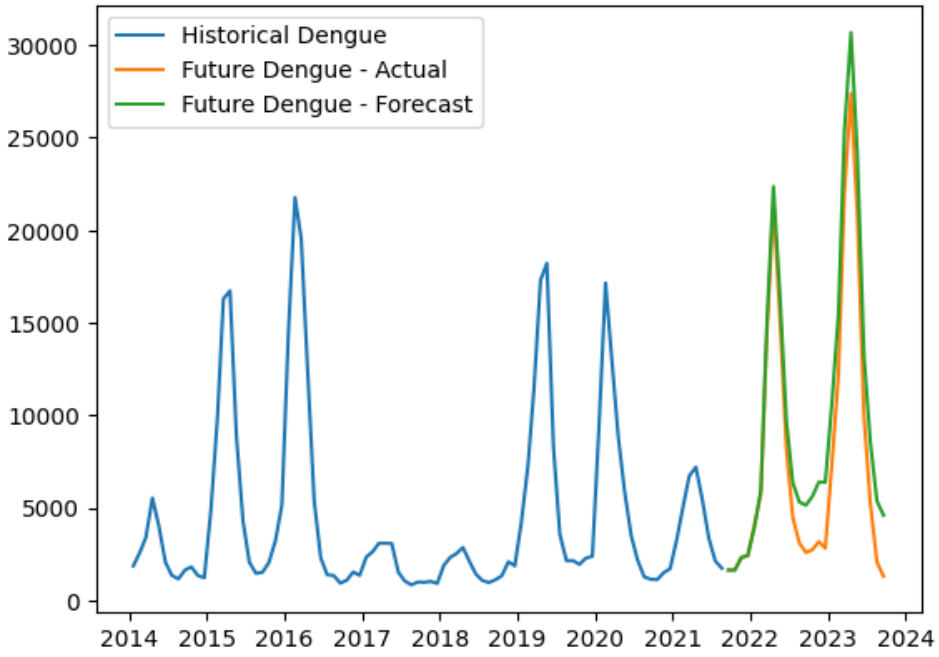

Figure 11. Prediction of SARIMAX Model Forecasting Two Years in Advance Using Average of Countries' Exogenous Variables – this plot shows the model forecasting two years in advance, with a mean absolute error of around 2020. It is important to note that the figure above shows the model's prediction when trained with the average of each country's exogenous variables.

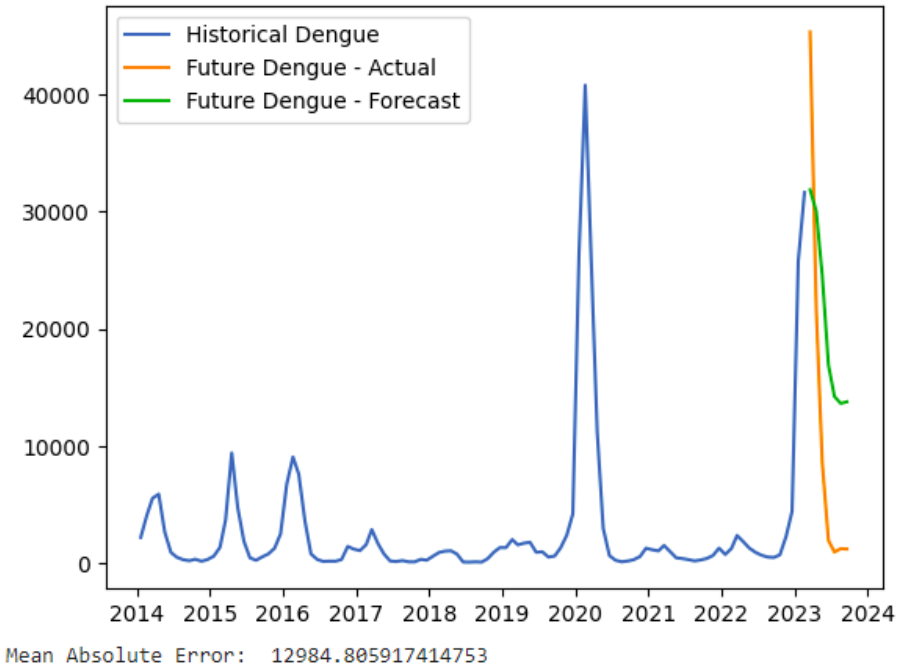

Figure 12. Prediction of SARIMAX Model Forecasting Two Years in Advance Using Dengue Data of Bolivia This shows the model forecasting six months in advance in Bolivia, with a mean absolute error of 12984.806. It is important to note that the figure above shows the model's prediction when trained with Bolivia's dengue data and the average of each country's other exogenous variables.

Finally, another important aspect of the model was stagnant water, which was the novel feature that was explored, and its contribution to making accurate dengue forecasts. It was observed that eliminating stagnant water from the original dataset prior to performing the principal component analysis resulted in a significant increase in the mean absolute error of 9884 after using the SARIMAX model. Additionally, the graph of the SARIMAX model indicates that there was a lot of error as the historical and forecasted dengue case values barely overlap each other **(Figure 13)**.

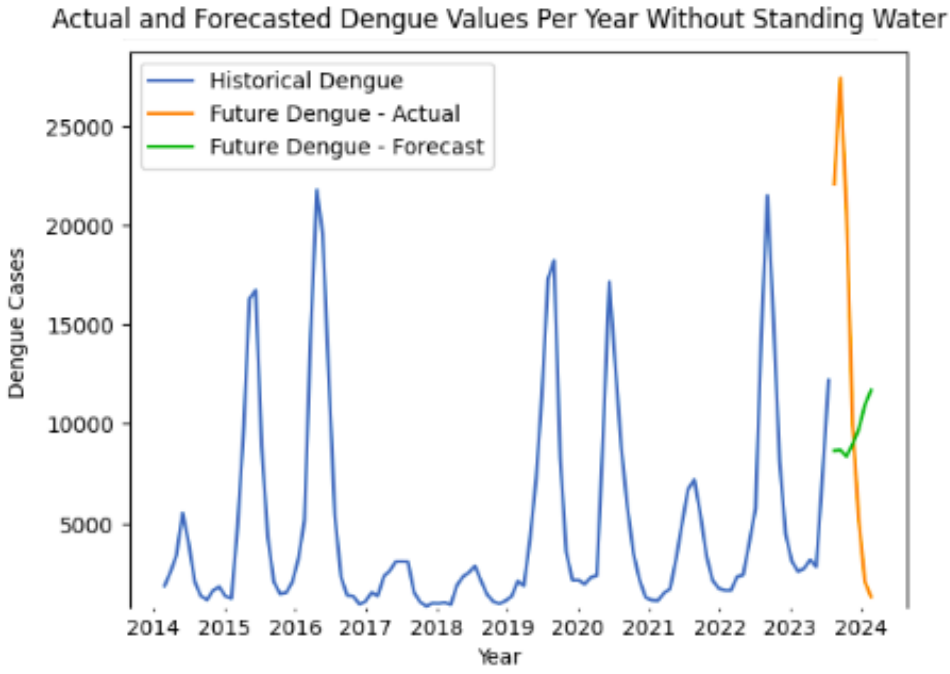

Figure 13. Prediction of SARIMAX Model Forecasting Six Months in Advance Without Standing Water
This displays the prediction when the original dataset does not include stagnant water when predicting six months in advance, with a mean absolute error of 9884.

4. Discussion

The ultimate goal of this study was to reduce lag time by forecasting dengue outbreaks in advance due to the delayed response time of public officials and limited treatment options. This was done by creating a machine learning model capable of accurately forecasting six months in advance in South America through the usage of eco-climatic and socioeconomic factors, particularly the novel factor of stagnant water. This purpose was achieved as the model had a mean absolute error value of 2.420e-6, which is extremely close to 0, when forecasting six months in advance, which is twice the lag time needed for public health officials to mitigate future dengue outbreaks, thus allowing our research to play a vital role in effectively curbing the spread of dengue. Therefore, this algorithm can assist public health officials with planning proactive measures, significantly diminishing economic stress and dengue transmission, and improving the quality of life in dengue-endemic countries.

Furthermore, the increase in the mean absolute error to 2610 at nine months in advance and beyond that time prediction is aligned with traditional forecasting models as the mean absolute error typically increases with a larger time frame to predict. The mean absolute error was also extremely high when the model was being applied to individual countries' dengue cases, which can conclude that exogenous variables need to be individual to the area of prediction for accurate forecasting. This is a significant limitation within this research as a widespread area of outbreak is not as valuable to know or to prepare with compared to a specific country. Moreover, as a principal component analysis was used to reduce multicollinearity and dimensionality between features, it is almost impossible to attribute the behavior of the model to the original features in the dataset. Although this does not impact the accuracy of the dengue forecasts, it limits the ability of public health officials to identify which factors significantly influence dengue transmission, hence hampering targeted mitigation efforts. Nevertheless, public health officials can gain insights into which variables are highly correlated with the number of dengue cases by

referencing Figure 4, the correlation matrix. Finally, when the novel feature of stagnant water was removed from the datasets, the prediction's accuracy significantly decreased when forecasting for the same time period of six months in advance. This demonstrates a significant need to use stagnant water for forecasting dengue and displays its importance as a feature to be considered when performing dengue and mosquito-borne disease research. This work can easily be expanded to different mosquito-borne diseases, such as Zika or Malaria, and can also be tested in different areas of the world that are dengue endemic such as Southeast Asia.

5. References

Brownlee, J. (2019, August 21). A Gentle Introduction to SARIMA for Time Series Forecasting in Python. Retrieved from https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/

Dahmana, H. & Mediannikov, O. (2017). Mosquito-borne diseases emergence/resurgence and how to effectively control it biologically. *Pathogens*, *9*(4). https://doi.org/10.3390/pathogens9040310

Data, M. N. (n.d.). Data collections: Earth System Data Explorer | My NASA Data. Retrieved from https://mynasadata.larc.nasa.gov/basic-page/data-collections-earth-system-data-explorer

Dengue emergency in the Americas: time for a new continental eradication plan. (2023). *The Lancet Regional Health - Americas, 22.* https://doi.org/10.1016/j.lana.2023.100539

Dengue – the Region of the Americas. (2023, July 19). Retrieved from https://www.who.int/emergencies/disease-outbreak-news/item/2023-DON475

Farooq, Z., Rocklöv, J., Wallin, J., Abiri, N., Sewe, M., Sjödin, H., & Semenza, J. (2022). Artificial intelligence to predict West Nile virus outbreaks with eco-climatic drivers. *The Lancet Regional Health - Europe*. https://doi.org/10.1016/j.lanepe.2022.100370

FreshExplorer. (n.d.). Retrieved from https://map.sdg661.app/#

Global economy, world economy. (n.d.). Retrieved from https://www.theglobaleconomy.com/

Gutiérrez, L. A. (n.d.). PAHO/WHO Data - National Dengue fever cases | PAHO/WHO. Retrieved from https://www3.paho.org/data/index.php/en/mnu-topics/indicadores-dengue-en/dengue-nacional-en/252-dengue-pais-ano-en.html

Hii, Y. L., Rocklöv, J., Wall, S., Ng, L. C., Tang, C. S., & Ng, N. (2012). Optimal lead time for dengue forecast. *PLOS Neglected Tropical Diseases, 6*(10). https://doi.org/10.1371/journal.pntd.0001848

How Dengue Spreads. (2024, May 14). Retrieved from https://www.cdc.gov/dengue/transmission/index.html

Jaadi, Z. (2024, February 23). Principal Component Analysis (PCA): A Step-by-Step Explanation. Retrieved from https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Laserna, A., Barahona-Correa, J., Baquero, L., & Castañeda-Cardona, C. (2018). Economic impact of dengue fever in Latin America and the Caribbean: a systematic review. *Revista Panamericana de Salud Pública, 42.* https://doi.org/10.26633/RPSP.2018.111

Life cycle of Aedes mosquitoes. (2024, April 16). Retrieved from https://www.cdc.gov/mosquitoes/about/life-cycle-of-aedes-mosquitoes.html

Morgan, J., Strode, C., & Salcedo-Sora, J. (2021). Climatic and socio-economic factors supporting the co-circulation of dengue, Zika and chikungunya in three different ecosystems in Colombia. *PLOS Neglected Tropical Diseases.* https://doi.org/10.1371/journal.pntd.0009259

Naish, S., Dale, P., Mackenzie, J. S., McBride, J., Mengersen, K., & Tong, S. (2014). Climate Change and Dengue: A Critical and Systematic Review of Quantitative Modelling Approaches. *BMC Infectious Diseases*, *14*(1). https://doi.org/10.1186/1471-2334-14-167

National Centers for Environmental Information (NCEI). (n.d.). Search | Climate Data Online (CDO) | National Climatic Data Center (NCDC). Retrieved from https://www.ncdc.noaa.gov/cdo-web/search

Navelski, J., & Odongo, K. (2021). Making Use of PCA in the Presence of Multicollinearity: An Application to Predicting Body Fat Percentage. *Washington State University.* https://s3.wp.wsu.edu/uploads/sites/2762/2022/10/PCA_and_Multicollinearity.pdf

Seasonal-Trend decomposition using LOESS (STL). (n.d.). Retrieved from https://www.statsmodels.org/dev/examples/notebooks/generated/stl_decomposition.html

Wang, H., Yao, R., Hou, L., Zhao, J., & Zhao, X. (2021). A Methodology for Calculating the Contribution of Exogenous Variables to ARIMAX Predictions. *Proceedings of the Canadian Conference on Artificial Intelligence.* https://doi.org/10.21428/594757db.2c2969c0

Appendix

Abbreviated Feature	*Full Feature Name*	*Description*
Dengue	Number of Reported Dengue Cases	Total number of dengue cases reported per month
Precipitation	Precipitation	Total monthly precipitation in millimeters
Humidity	Humidity	Average monthly humidity percentage
Temperature	Temperature	Average monthly temperature in degrees Fahrenheit
Standing Water	Standing Water	Areas with stagnant water, or seasonal change in water, measured in square kilometers
Economic Growth	Average of Economic Growth	Annual rate of change in real GDP
GDP	Average GDP	Total market value of goods/services, measured in billions of USD
GDP per Capita	Average GDP per Capita	Average economic output per person
Household Consumption	Average of Household Consumption	Average value of goods/services consumed by households
Unemployment	Average of Unemployment Rate	Average percentage of labor force unemployed
Government Spending	Average of Government Spending (% of GDP)	Average public expenditure, measured in billions of USD
Government Debt	Average of Government Debt	Average government debt as percent of GDP
Foreign Aid	Average of Foreign Aid and Official Development Assistance Received	Average international economic, military, and technical assistance received
Government Effectiveness	Average of Government Effectiveness Index	Average quality of public services and policy implementation by government
Control of Corruption	Average of Control of Corruption	Average measure of public power exercised for private gain by government
Political Stability	Average of Political Stability Index	Average likelihood of political instability and violence
Internet Users	Average of Internet Users	Percentage of population using the internet
Mobile Phone Subscribers	Average of Mobile Phone Subscribers	Average number of mobile phone subscriptions
Carbon Dioxide Emissions	Average of Carbon Dioxide Emissions per Capita	Average carbon dioxide emissions per person
Greenhouse Gas Emissions	Average of Greenhouse Gas Emissions	Average greenhouse gas emissions
Forest Area	Average of Forest Area	Percentage of land area covered by forests
Innovation	Average of Innovations Index	Average level of innovation in country
Population Density	Average of Population Density	Number of people per square kilometer
Health Spending per Capita	Average of Health Spending per Capita	Average health expenditure per person
Health Spending (% GDP)	Average of Health Spending (% of GDP)	Average health expenditure as percentage of GDP
Undernourishment	Average of Prevalence of Undernourishment	Percentage of population with insufficient food intake
Gini Index	Average of Gini Income Inequality Index	Measure of income inequality

Poverty Rate (<$1.90/day)	Average of Poverty Ratio (<$1.90/day)	Percentage of population living on less than $1.90 a day
Poverty Rate (<$5.50/day)	Average of Poverty Ratio (<$5.50/day)	Percentage of population living on less than $5.50 a day
Public Education Spending	Average of Public Spending on Education (% of GDP)	Public expenditure on education as percentage of GDP
Literacy Rate	Average of Literacy Rate	Percentage of population that can read and write
Globalization	Average of Globalization Index	Level of globalization (economic, social, political), measured on a scale from 0-100
Economic Globalization	Average of Economic Globalization Index	Trade and economic transactions index, measured on a scale from 0-100
Political Globalization	Average of Political Globalization Index	Political engagement index, measured on a scale from 0-100
Social Globalization	Average of Social Globalization Index	Social and cultural interactions index, measured on a scale from 0-100
Tourist Arrivals	Average of Tourist Arrivals	Average number of tourists
World Tourist Arrivals	Average of Percent of World Tourist Arrivals	Average percent share of global tourist arrivals
Human Development Index	Average of Human Development Index	Composite measure of human development, measured on a scale from 0-1
Economic Freedom	Average of Economic Freedom, Overall Index	Average measure of economic freedom, measured on a scale from 0-100
Population Size	Average of Population Size	Average population size, measured in millions
Urban Population	Average of Percent Urban Population	Percentage of total population living in urban areas
Rural Population	Average of Rural Population	Percentage of total population living in rural areas
Population Growth	Average of Population Growth	Annual percentage increase in population
Fragile State	Average of Fragile State Index	Average vulnerability of state for conflict or collapse, measured on a scale from 0 being low to 120 being high
Economic Decline	Average of Economic Decline Index	Average measure of economic decline severity, measured on a scale from 0 being low to 120 being high

Appendix. Provides abbreviated and full names, and brief descriptions of the original features used in the analysis of dengue cases in Central South America