

Article

Not peer-reviewed version

---

# Leveraging Climate Data for Dengue Forecasting in Ba Ria Vung Tau Province, Vietnam: An Advanced Machine Learning Approach

---

[Dang Anh Tuan](#) \* and [Tran Ngoc Dang](#)

Posted Date: 23 September 2024

doi: 10.20944/preprints202409.1535.v3

Keywords: Dengue fever; Machine learning; Climate forecasting; Negative Binomial Regression; SARIMAX; XGBoost; LSTM; Ba Ria Vung Tau; Vietnam



Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Article

# Leveraging Climate Data for Dengue Forecasting in Ba Ria Vung Tau Province, Vietnam: An Advanced Machine Learning Approach

Dang Anh Tuan \* and Tran Ngoc Dang

Faculty of Public Health, University of Medicine and Pharmacy at Ho Chi Minh City,  
Ho Chi Minh City, 70000, Vietnam

\* Correspondence: tuanda222@gmail.com; Tel.: +84-987-066-222

**Abstract:** Dengue fever is a persistent public health issue in tropical regions, including Vietnam, where climate variability plays a crucial role in disease transmission dynamics. This study focuses on developing climate-based machine learning models to forecast dengue outbreaks in Ba Ria Vung Tau (BRVT) province, Vietnam, using meteorological data from 2003 to 2022. We utilized four predictive models—Negative Binomial Regression (NBR), Seasonal AutoRegressive Integrated Moving Average with Exogenous Regressors (SARIMAX), Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM)—to predict weekly dengue incidence. Key climate variables, including temperature, humidity, precipitation, and wind speed, were integrated into these models, with lagged variables included to capture delayed climatic effects on dengue transmission. The NBR model demonstrated the best performance in terms of predictive accuracy, achieving the lowest Mean Absolute Error (MAE) compared to other models. The inclusion of lagged climate variables significantly enhanced the model's ability to predict dengue cases. SARIMAX and LSTM models, although effective in capturing seasonal trends, struggled with overfitting and failed to accurately predict short-term outbreaks. XGBoost exhibited moderate predictive power but was sensitive to overfitting, particularly without fine-tuning. Our findings confirm that climate-based machine learning models, particularly the NBR model, offer valuable tools for forecasting dengue outbreaks in BRVT. However, improving the models' ability to predict short-term peaks remains a challenge. The integration of meteorological data into early warning systems is crucial for public health authorities to plan timely and effective interventions. This research contributes to the growing body of literature on climate-based disease forecasting and underscores the need for further model refinement to address the complexities of dengue transmission in highly endemic regions.

**Keywords:** Dengue fever; Machine learning; Climate forecasting; Negative Binomial Regression; SARIMAX; XGBoost; LSTM; Ba Ria Vung Tau; Vietnam

## 1. Introduction

Dengue fever (DF), a mosquito-borne viral infection, continues to pose a significant public health threat in tropical and subtropical regions worldwide, including Vietnam. The disease is transmitted predominantly by *Aedes aegypti* mosquitoes, whose breeding, survival, and transmission dynamics are strongly influenced by environmental factors such as temperature, precipitation, and humidity [1]. Ba Ria Vung Tau (BRVT), a coastal province in Southern Vietnam, regularly experiences outbreaks of dengue fever, making the development of accurate predictive models essential for the timely implementation of disease control measures [2]. Climate-based forecasting models, which incorporate meteorological data, have shown great promise in improving the accuracy of dengue outbreak predictions, thereby enabling more effective public health responses [3].

Several studies have confirmed the critical role of climatic factors in influencing the transmission of dengue. A study conducted in Southern China demonstrated that temperature, humidity, and rainfall are the most significant variables affecting the mosquito population and, consequently, the incidence of dengue fever [1]. Similarly, research in Queensland, Australia, using Bayesian spatial analysis, found a strong correlation between climatic variability and dengue transmission, with

increased rainfall and higher temperatures significantly boosting the risk of outbreaks [3]. These findings highlight the importance of integrating climate data into predictive models to forecast dengue outbreaks effectively.

Machine learning (ML) and deep learning (DL) approaches have recently revolutionized the field of disease forecasting by providing more sophisticated tools for handling complex, non-linear relationships in time series data. Long Short-Term Memory (LSTM) models, for instance, have proven highly effective in forecasting dengue outbreaks based on climate factors. A study conducted in China demonstrated that LSTM models outperformed traditional time-series models by reducing prediction error by as much as 24.91% when forecasting dengue cases one month in advance [4]. Furthermore, a dynamic ensemble approach combining weather and population susceptibility cycles was able to accurately predict dengue outbreaks in Brazil months ahead of time, demonstrating the significant potential of integrating meteorological and population data [5].

In Vietnam, probabilistic seasonal forecasting models have also proven valuable in predicting dengue outbreaks up to six months in advance. A superensemble model that integrates Earth observation data and seasonal climate forecasts has been shown to outperform baseline models, particularly in regions with semi-regular dengue transmission such as Southern Vietnam. This model successfully predicted outbreak periods with an accuracy significantly higher than models that rely solely on historical incidence rates [2].

The integration of advanced machine learning techniques, including neural networks and hybrid models, has further refined dengue forecasting models. A study in Mexico employed a multi-stage machine learning approach to predict dengue outbreaks by analyzing the temporal dynamics of temperature and dengue data. This model outperformed several state-of-the-art forecasting methods and provided accurate predictions of dengue cases on an annual scale [6]. Another study in Malaysia demonstrated that support vector machine (SVM) models, when applied to meteorological data, achieved an accuracy rate of 70% in predicting dengue outbreaks, highlighting the potential of machine learning techniques in dengue forecasting [7].

Given the success of climate-based predictive models in forecasting dengue outbreaks, this study aims to develop a robust model for predicting dengue outbreaks in BRVT using meteorological data. By leveraging advanced machine learning techniques and incorporating key climatic variables such as temperature, humidity, wind speed, rainfall, and some additional variables such as surface pressure, wind direction, sea surface temperature (SST) this model seeks to provide accurate, early warnings of potential dengue outbreaks. The findings of this study will contribute to the growing body of research focused on climate-based early warning systems for vector-borne diseases and provide a practical tool for public health officials in dengue-endemic regions.

## 2. Material and Methods

### 2.1. Study Area

This study was conducted in BRVT, a coastal province located in Southern Vietnam (approximately between latitudes 10°20'N and 10°40'N, and longitudes 107°00'E and 107°30'E), which is highly vulnerable to dengue outbreaks due to its tropical monsoon climate. The province spans an area of about 1,982 square kilometers, with a population exceeding 1.1 million as of the 2020 census. The region experiences distinct wet and dry seasons, with the rainy season typically occurring from May to October, characterized by high humidity and heavy rainfall, providing optimal breeding conditions for *Aedes aegypti* mosquitoes, the primary vector for dengue transmission [8].

BRVT's geographic diversity—featuring urban centers, rural farmlands, and coastal zones—creates a range of microclimates that influence local weather patterns, including temperature variability and humidity levels, both of which are known to impact mosquito breeding and dengue transmission dynamics [9]. The province's varying elevation and proximity to the coast further modify its climate, making it particularly susceptible to dengue outbreaks. The frequent occurrence of dengue cases in recent years, coupled with the province's exposure to climate variability, underscores its suitability for studies focused on climate-based dengue forecasting.

2.2. Data Collection

*Epidemiological Data:* Weekly data on confirmed dengue fever cases reported during the study period from January 1, 2003, to December 31, 2022, were obtained from the Center for Disease Control of Ba Ria Vung Tau Province, Vietnam.

*Climate Data:* The climate parameters presented in Table 1 were used in our model. These variables were selected based on previous studies and their assumed relationship with dengue fever. The data were collected from daily datasets with original resolution from NASA/POWER CERES/MERRA2, spanning from January 1, 2003, to December 31, 2022. Location coordinates: Latitude 10.5417, Longitude 107.243. Altitude from MERRA-2: Average for a region of 0.5 x 0.625 degrees latitude/longitude = 35.52 meters. To assess the accuracy of the models in determining the relationship between these variables and dengue cases, we used the number of dengue cases as a benchmark for both regression and classification tasks.

**Table 1.** Climate parameters used in the models of this study.

| No. | Parameter                       | Symbol      | Unit    |
|-----|---------------------------------|-------------|---------|
| 1   | Temperature at 2 Meters Range   | T2M_RANGE   | °C      |
| 2   | Temperature at 2 Meters Maximum | T2M_MAX     | °C      |
| 3   | Temperature at 2 Meters Minimum | T2M_MIN     | °C      |
| 4   | Temperature at 2 Meters         | T2M         | °C      |
| 5   | Relative Humidity at 2 Meters   | RH2M        | %       |
| 6   | Precipitation Corrected         | PRECTOTCORR | mm/day  |
| 7   | Surface Pressure                | PS          | kPa     |
| 8   | Wind Speed at 10 Meters         | WS10M       | m/s     |
| 9   | Wind Speed at 10 Meters Maximum | WS10M_MAX   | m/s     |
| 10  | Wind Speed at 10 Meters Minimum | WS10M_MIN   | m/s     |
| 11  | Wind Speed at 10 Meters Range   | WS10M_RANGE | m/s     |
| 12  | Wind Direction at 10 Meters     | WD10M       | Degrees |
| 13  | Sea Surface Temperature         | SST         | °C      |

2.3. Data Preprocessing

2.3.1. Missing Data Imputation

Due to occasional gaps in the meteorological datasets, two imputation techniques were employed:

*Linear Interpolation:* Missing values for continuous variables such as temperature and humidity were filled using linear interpolation to maintain the temporal continuity of the data [1].

*K-Nearest Neighbors (KNN) Imputation:* For more substantial gaps, the KNN imputation technique was used. This method estimates missing values based on the similarity to neighboring data points in the feature space, ensuring robust predictions [10].

2.3.2. Feature Engineering

Several new features were engineered to enhance the predictive power of the model:

*Lagged Variables:* Climate variables (temperature, precipitation, humidity) were shifted by 2 to 20 weeks to capture the delayed effects of weather conditions on dengue transmission. This accounts for the lifecycle of mosquitoes and the incubation period of the virus [11].

*Rolling Averages:* To capture the cumulative effects of climate conditions, rolling averages were computed for key variables across different time windows (ranging from 1 to 20 weeks), helping to model sustained climate effects on dengue outbreaks.

*Seasonal and Monthly Indicators:* Categorical variables representing the season and month were included to account for the strong seasonality in dengue incidence, with dummy variables for each season (spring, summer, autumn, and winter).

#### 2.4. Feature Scaling

To ensure that all features contributed equally to the model, standardization was applied to the dataset. Using `StandardScaler` from `scikit-learn`, all continuous features were transformed to have a mean of 0 and a standard deviation of 1. This process prevents features with larger numerical ranges from dominating the model training process [12].

#### 2.5. Feature Selection

A correlation matrix was computed to examine the relationships between the climate variables and the target variable, total dengue cases. The features showing the highest correlation with dengue cases were selected for further analysis. These lagged features were selected as they demonstrated the strongest correlations with the target variable (total dengue cases).

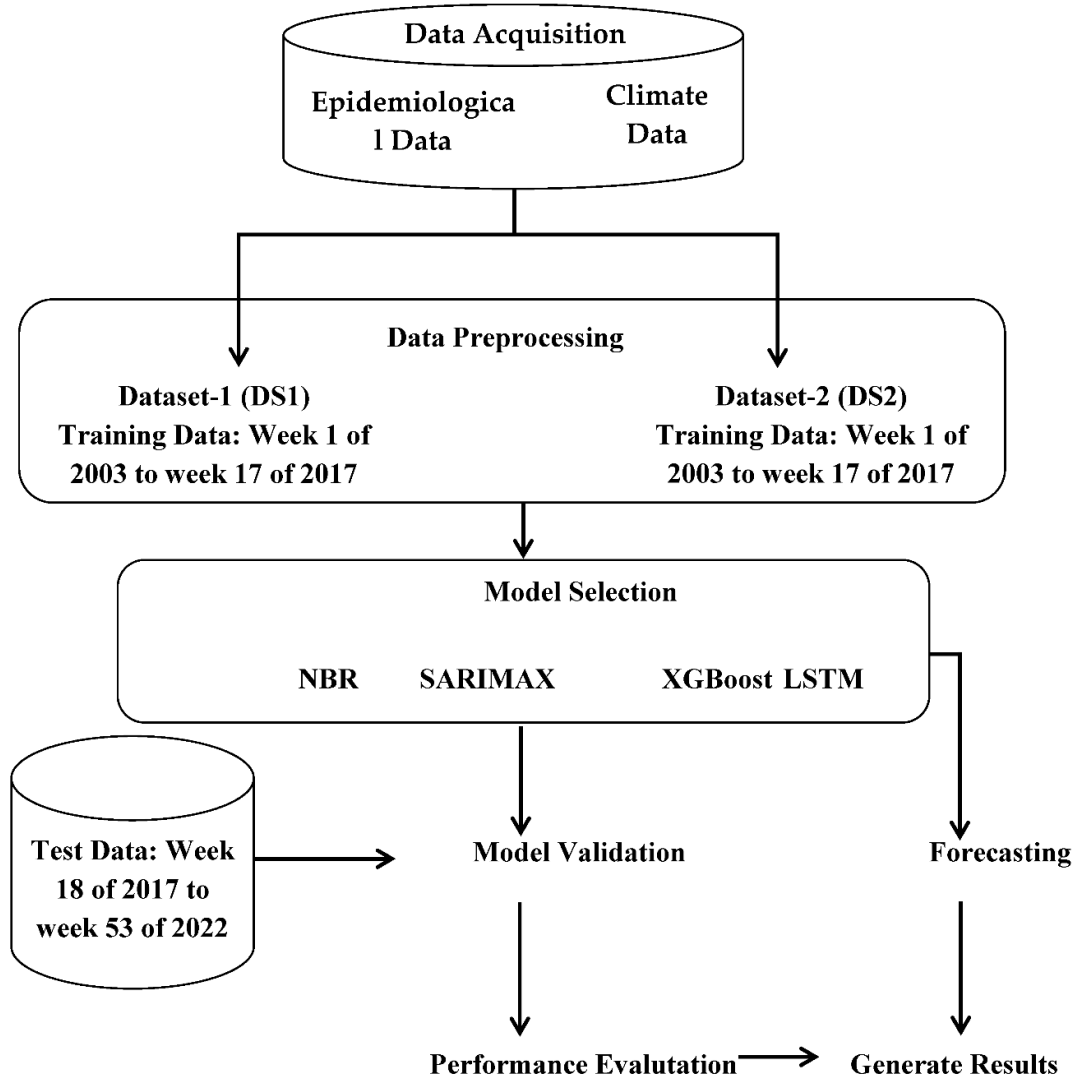
#### 2.6. Data Splitting

The dataset, which includes dengue fever cases and weather variables from week 01 of 2003 to week 53 of 2022, was used for this study. Before analysis, feature engineering was performed to create new features from the variables or to transform the original data into more informative formats. Two datasets were considered: the first dataset consisted of raw data including weather parameters (Table 1) and time variables, referred to as DS1. The second dataset contained time variables and the number of dengue cases, referred to as DS2. To capture the delayed effects of climatic factors on dengue transmission, cross-correlation analysis was conducted to determine the appropriate lag time for each meteorological variable, ranging from 0 to 12 months. Similarly, autocorrelation analysis (ACF) was conducted to evaluate the impact of past dengue cases on current cases.

Before conducting the analysis, the datasets were split into training and test sets. The training data consisted of observations from week 01 of 2003 to week 17 of 2017 for both DS1 and DS2. Likewise, the test data covered the period from week 18 of 2017 to week 53 of 2022 for both DS1 and DS2. The training dataset was used to build the models, while the test dataset was used to validate the models.

#### 2.7. Data Modeling

In this study, statistical, machine learning, and deep learning models such as Negative Binomial Regression (NBR), Seasonal AutoRegressive Integrated Moving Average with Exogenous Regressors (SARIMAX), Extreme Gradient Boosting (XGBoost) Regression, and Long Short-Term Memory (LSTM) were deployed to predict dengue cases in BRVT. The overall framework of the study and step-by-step modeling workflow is illustrated in **Figure 1**. Dengue cases were treated as the dependent or response variable, while the climatic variables and their respective lags were used as independent variables.



**Figure 1.** Overall process of DF case prediction study.

### 2.7.1. Negative Binomial Regression (NBR)

Negative Binomial Regression (NBR) was selected to predict dengue case counts due to the overdispersed nature of the data, where the variance substantially exceeded the mean. NBR is preferable to Poisson regression in such cases, as it introduces a dispersion parameter to account for overdispersion [13].

The NBR model relates the expected number of dengue cases ( $\mu_i$ ) to the predictor variables ( $X_i$ ) through a log link function:

$$\log(\mu_i) = \beta_0 + \sum_{k=1}^p \beta_k X_{ik}$$

where:  $\mu_i$  is the expected count of dengue cases,  $X_{ik}$  are the predictor variables (e.g., temperature, humidity, lagged climatic variables),  $\beta_k$  are the regression coefficients,  $\beta_0$  is the intercept.

NBR introduces a dispersion parameter  $\alpha$  to account for overdispersion, which modifies the variance function:

$$Var(Y_i) = \mu_i + \alpha \mu_i^2$$

This variance function allows the model to adjust for the extra variability observed in dengue case counts [13].

Model fitting was performed using maximum likelihood estimation (MLE). The dispersion parameter  $\alpha$  was tuned via cross-validation. A grid search was conducted over a range of  $\alpha$  values (0.001 to 2.0) to optimize model performance based on the Mean Absolute Error (MAE) on the test set.

The steps for model tuning included: (i) Train-Test Split—Data was used for training (80%), and data was used for testing (20%) to evaluate the model's predictive performance; (ii) Cross-Validation—A rolling window time-series cross-validation approach was employed to prevent data leakage and ensure temporal consistency [14]; (iii) Hyperparameter Optimization— GridSearchCV was used to identify the best-fitting model by minimizing MAE [15].

### 2.7.2. Seasonal AutoRegressive Integrated Moving Average with eXogenous Regressors (SARIMAX)

The SARIMAX model was selected for dengue case prediction due to its ability to model both temporal dynamics and the influence of exogenous variables, such as climate factors. SARIMAX extends the classical ARIMA (Autoregressive Integrated Moving Average) model by incorporating seasonal components and exogenous variables, which makes it particularly well-suited for modeling seasonal data with external influences.

The SARIMAX model is defined by the following components:

*AR (Autoregressive) component:* This component captures the relationship between an observation and its previous values.

*I (Integrated) component:* This accounts for differencing the series to make it stationary, reducing the impact of trends.

*MA (Moving Average) component:* This models the error terms as a linear combination of past error terms.

*Seasonality:* SARIMAX accounts for seasonality through seasonal AR, I, and MA terms.

The general SARIMAX model is written as:

$$Y_t = \sum_{i=1}^p \phi_i Y_{t-i} + \sum_{j=1}^q \theta_j \epsilon_{t-j} + X_t \beta + \epsilon_t$$

where:  $Y_t$  is the dengue case count at time  $t$ ,  $\phi_i$  are the autoregressive coefficients,  $\theta_j$  are the moving average coefficients,  $\epsilon_t$  is the error term (white noise),  $X_t$  represents the exogenous variables (e.g., climate variables such as temperature, humidity, precipitation),  $\beta$  is the vector of coefficients associated with the exogenous variables.

For the seasonal component, the model is extended by introducing seasonal autoregressive (SAR), seasonal moving average (SMA), and seasonal differencing terms, denoted as  $P$ ,  $D$ , and  $Q$ , respectively. The seasonal component is added with periodicity  $s$ , representing the number of time steps in a season.

The full SARIMAX model is typically written as:

$$SARIMA(p, d, q)(P, D, Q)_s$$

where:  $p$ ,  $d$ ,  $q$  represent the non-seasonal parameters (autoregressive, differencing, and moving average),  $P$ ,  $D$ ,  $Q$  are the seasonal counterparts,  $s$  is the length of the seasonal cycle.

The exogenous variables ( $X_t$ ) included in the SARIMAX model were climate factors such as temperature, humidity, and precipitation, along with their lagged versions. These variables were selected based on their known impact on dengue transmission and mosquito population dynamics [16]. Lagged versions of these variables (1 to 4 weeks) were included to capture delayed climatic effects on dengue incidence.

The SARIMAX model was tuned using a grid search approach to identify the optimal values for the parameters  $p$ ,  $d$ ,  $q$ ,  $P$ ,  $D$ ,  $Q$ , and  $s$ . The following steps were undertaken: (i) Train-Test Split—The dataset was split into training and testing sets. The training set was used to fit the model, while the

test set was used for out-of-sample evaluation; (ii) Grid Search—A grid search was conducted over a range of possible values for the hyperparameters  $p$ ,  $d$ ,  $q$ ,  $P$ ,  $D$ , and  $Q$ , along with seasonal periodicity  $s = 52$ , to find the optimal combination that minimized the Akaike Information Criterion (AIC) [17]; (iii) Cross-Validation—Time series cross-validation using a rolling window approach was implemented to ensure that temporal dependencies were preserved and to prevent data leakage during model evaluation [14].

### 2.7.3. Extreme Gradient Boosting (XGBoost) Regression

XGBoost (Extreme Gradient Boosting) was selected for this study due to its efficiency and performance in handling large, complex datasets with non-linear relationships. XGBoost is a supervised machine learning algorithm that builds an ensemble of decision trees using a boosting technique to improve prediction accuracy. The algorithm works by iteratively adding weak learners (trees) to minimize the residual errors from previous iterations.

XGBoost operates by minimizing a loss function, typically the Mean Squared Error (MSE) for regression tasks, along with a regularization term to control model complexity and prevent overfitting. The objective function is given by:

$$Obj(\theta) = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

where:  $L(y_i, \hat{y}_i)$  represents the loss function (e.g., squared error) between the actual values  $y_i$  and predicted values  $\hat{y}_i$ ,  $\Omega(f_k)$  is the regularization term applied to each decision tree  $f_k$ , which penalizes the complexity of the model to avoid overfitting [18].

The regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2$$

where  $T$  is the number of leaves in the tree,  $w_j$  are the leaf weights,  $\gamma$  and  $\lambda$  are regularization hyperparameters that control the trade-off between model complexity and training performance.

The XGBoost model includes several key hyperparameters that were tuned using grid search to optimize the model's performance: (i) Learning Rate ( $\eta$ )—Controls the step size at each iteration of boosting. A lower learning rate requires more trees to achieve convergence but typically leads to better generalization; (ii) Max Depth—Controls the maximum depth of each decision tree. A deeper tree captures more complex patterns but increases the risk of overfitting; (iii) Subsample—Fraction of the training data used to build each tree. Subsampling helps prevent overfitting by introducing randomness into the training process; (iv) Gamma ( $\gamma$ )—Minimum loss reduction required to make a further partition on a leaf node of the tree; (v) Min Child Weight—Minimum sum of instance weights needed in a child node.

The hyperparameters were optimized using GridSearchCV with 5-fold cross-validation to ensure robustness. The combination of hyperparameters that minimized the Root Mean Square Error (RMSE) was selected for the final model [18].

#### 2.7.4. Long Short-Term Memory (LSTM)

Long short-term memory (LSTM) is the most effective approach for time series forecasting [19], where classical linear methods are difficult to use multivariate forecasting issues. Among various machine learning methods, the recurrent neural network (RNN) uses feedback loops; subsequently, the network can learn the sequence of information. However, standard RNN fails to remember the long-term information; hence, the present study proposes LSTM network for dengue prediction.

LSTM is the recurrent neural network (RNN) architecture and it was designed by the Hochreiter and Schmidhuber to address the vanishing and exploding gradient problems of traditional RNNs [20]. These vanishing and gradient issues hinder the model accuracy. The LSTM consists of different blocks called memory cells. Memory block contains memory cells and gates. Memory cells are able to remember and manipulating the temporal state of the memory or network by self-connections and the process of memory or information controlled through three gates. The first gate called as “Forget” gate ( $f_j$ ), which is responsible for removing information. The second gate is “Input” gate ( $i_j$ ), which is responsible for addition of information to the cell state. The final gate is “Output” gate ( $o_j$ ), which selects the useful information from current cell and shows it as an output [21]. These forget gate, input gate, and output gate of LSTM are described by [20] using the following notations mentioned below:

$$f_j = \sigma(w_f \cdot [h_{j-1}, x_j] + b_f)$$

$$i_j = \sigma(w_i \cdot [h_{j-1}, x_j] + b_i)$$

$$o_j = \sigma(w_o \cdot [h_{j-1}, x_j] + b_o)$$

$$\hat{C} = \tanh(w_c \cdot [h_{j-1}, x_j] + b_c)$$

$$C_j = f_j * C_{j-1} + i_j * \hat{C}$$

$$h_j = o_j * \tanh(C_j)$$

where  $f, i, o$  are forget gate, input gate and output gate respectively. The input gate is used to decide the information to be stored in the cell, forget gate is used to decide what kind of information is to be dropped from cell and output gate is used to decide output from the cell.  $\sigma$  represents the logistic sigmoid function.  $w, b$ , and  $h$  are weights, biases, and value of hidden layers respectively.  $\hat{C}$  and  $C$  are vector of new candidate values and cell state respectively.

#### 2.8. Predictive Performance and Model Validation

This study employed a comprehensive approach to evaluate the predictive performance of models developed for dengue case forecasting. Rigorous validation processes were implemented to ensure robustness, accuracy, and generalizability. Several error metrics, cross-validation techniques, and hyperparameter tuning strategies were used to optimize model performance.

##### 2.8.1. Evaluation Metrics

The predictive models were evaluated using key metrics designed to measure accuracy and error distribution:

*Mean Absolute Error (MAE):* MAE is calculated as the mean of the absolute differences between predicted and actual values, providing a direct measure of prediction accuracy. The formula for MAE is:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

where  $y_i$  represents the actual values,  $\hat{y}_i$  the predicted values, and  $n$  the number of observations [22]

*Root Mean Square Error (RMSE)*: RMSE was used to give more weight to larger errors, and is expressed as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

RMSE is particularly sensitive to outliers and thus provides insight into the error distribution [23].

For time-series models such as SARIMAX, the Akaike Information Criterion (AIC) was employed to evaluate model complexity relative to goodness of fit. A lower AIC indicates a better fit when comparing models of different complexity [17]. The AIC formula is:

$$AIC = 2k - 2 \ln(\hat{L})$$

where  $k$  is the number of parameters in the model, and  $\hat{L}$  is the maximized likelihood.

### 2.8.2. Model Validation Approach

The dengue dataset, spanning from week 1 of 2003 to week 17 of 2017, was split into two subsets:

*Training Set (80%)*: The initial 80% of the data was reserved for training the models.

*Testing Set (20%)*: The final 20% of the data was used to evaluate the models' out-of-sample predictive performance.

For time-series forecasting models, including SARIMAX and LSTM, temporal integrity was preserved by ensuring that future observations were not included in the training phase, thereby replicating real-world forecasting conditions [24].

### 2.8.3. Cross-Validation Strategy

To prevent data leakage and ensure model robustness, we employed TimeSeriesSplit cross-validation for machine learning models like XGBoost. This method expands the training set sequentially while moving the validation set forward in time, preserving the temporal structure of the data [14]

Hyperparameter optimization was performed using GridSearchCV, which systematically explored the hyperparameter space to find the optimal combination of parameters. This search was combined with 5-fold cross-validation to ensure robustness and generalizability [18].

### 2.8.4. Overfitting and Generalization

To monitor overfitting, the models' performance on both training and test datasets was compared:

*Negative Binomial Regression*: The model demonstrated consistent performance across training and testing data, with similar MAE scores indicating strong generalization capability. Negative Binomial Regression is particularly effective for count data with overdispersion, where the variance exceeds the mean [13]. The formula for Negative Binomial Regression is:

$$\log(\mu_i) = X_i \beta$$

where  $\mu_i$  represents the expected number of cases,  $X_i$  the predictor variables, and  $\beta$  the coefficients.

**SARIMAX:** The baseline SARIMAX model exhibited overfitting, with significantly lower MAE and RMSE on the training set compared to the test set. The inclusion of exogenous variables (e.g., climate factors) in the multivariate SARIMAX model improved generalization and reduced overfitting [23].

**XGBoost:** Initially, the XGBoost model captured seasonal patterns but struggled with the prediction of outbreak peaks. The introduction of lagged climate variables significantly improved model performance, reducing generalization error and enhancing predictive accuracy [18].

#### 2.8.5. Model Refinement and Iteration

Several iterations were conducted to optimize the models:

**Negative Binomial Regression:** Incorporating time-shifted climate variables improved the model's ability to predict weekly dengue case counts more accurately.

**SARIMAX:** The addition of lagged exogenous variables helped improve the model's accuracy in capturing both seasonal trends and outbreak peaks.

**XGBoost:** The introduction of lagged variables in XGBoost further enhanced its ability to generalize across both training and testing datasets, resulting in improved MAE and RMSE scores.

#### 2.8.6. Final Model Selection

After iterative optimization, the final models were selected based on their performance in both training and testing datasets. These models were then retrained on the full dataset to maximize their predictive capabilities for future dengue cases. Feature importance analysis was also performed to identify the most significant predictors of dengue outbreaks [18].

### 2.9. Statistical Analysis

Once the final dataset was prepared, various machine learning models were trained to predict weekly dengue cases based on the selected climate features. The models' performance was evaluated using metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The analysis was performed using Python 3.12.4, leveraging libraries such as pandas, numpy, scikit-learn, and seaborn for data processing, modeling, and visualization.

## 3. Results

### 3.1. Trends and Fluctuations in Annual Dengue Incidence (2003-2022)

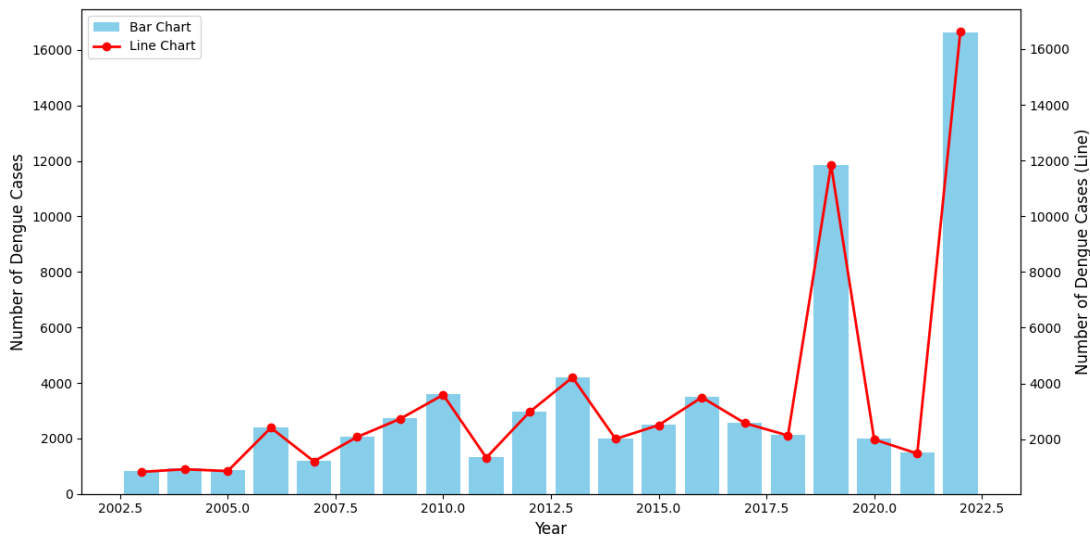
The presented Figure 2 chart highlights the yearly distribution of dengue cases from 2003 to 2022, offering significant insights into the trends and fluctuations of the disease. Over the years, the overall trend indicates a progressive increase in the number of dengue cases, with notable peaks occurring in specific years. In the early period (2003-2010), the number of cases exhibited a gradual increase, reflecting the growing prevalence of the disease. However, between 2010 and 2017, the trend stabilized, with minor fluctuations observed across the years.

The most prominent spike occurred in 2019, where the number of dengue cases surged dramatically, reaching the highest level recorded up to that point. This was followed by an unexpected sharp decline in 2020 and 2021. The significant reduction in cases during these years may be attributed to external factors, such as global public health interventions or behavioral changes linked to the COVID-19 pandemic, which potentially mitigated the spread of vector-borne diseases. However, the data for 2022 shows a striking resurgence, with the number of cases surpassing even the 2019 peak, suggesting a renewed outbreak with potentially more severe implications.

The dual-axis chart employs both bar and line graphs, allowing for a clear comparison between annual trends and individual fluctuations in case numbers. The consistency between the two visual representations underscores the accuracy of the observed peaks and troughs over time.

In summary, this chart reveals an alarming long-term upward trend in dengue incidence, punctuated by periodic outbreaks, particularly in 2019 and 2022. The data suggest that without effective, sustained public health interventions, the frequency and intensity of dengue outbreaks may

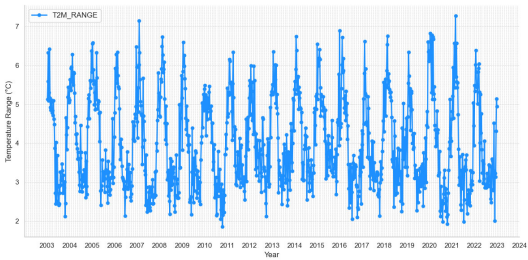
continue to rise, posing a significant challenge to global health systems. These findings emphasize the urgent need for reinforced surveillance, prevention, and control measures to curb the further spread of dengue fever.



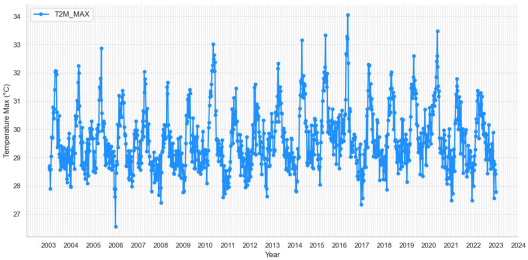
**Figure 2.** Annual distribution of dengue cases in Ba Ria Vung Tau, Vietnam (2003-2022).

3.2. Descriptive Statistics

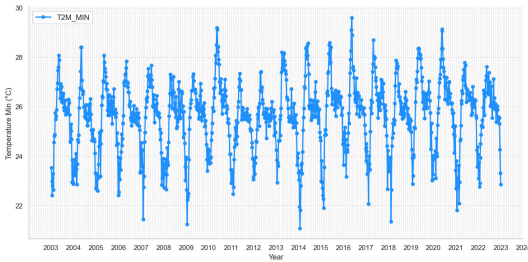
The time series plots of weekly 13 weather variables from 2003 to 2022 of BRVT are shown in Figure 3. Table 2 provides a descriptive summary of the key variables observed in the dataset. The weekly dengue cases (DF\_case) show a high level of variability, with an average of 64.94 cases per week and a large standard deviation of 111.99, indicating significant fluctuations in case numbers across different weeks. The minimum number of weekly cases is 0, while the maximum recorded is 913, with 75% of weeks having fewer than 68 cases. This suggests that dengue outbreaks can be sporadic but severe in certain weeks. Temperature-related variables exhibit more stability. The mean maximum temperature (T2M\_MAX) is 29.63°C, with a standard deviation of 1.08°C, while the minimum temperature (T2M\_MIN) averages 25.63°C, and the average temperature (T2M) stands at 27.38°C. The temperature range (T2M\_RANGE) fluctuates from 1.86°C to 7.26°C, indicating relatively mild variability in daily temperature swings across weeks. Humidity levels (RH2M) are generally high, with a mean of 79.28% and a range from 58.29% to 89.13%, reflecting the tropical climate conducive to mosquito breeding. Precipitation (PRECTOTCORR) is highly variable, with a mean of 4.15 mm but a large standard deviation of 4.78 mm, suggesting that rainfall can be inconsistent, with some weeks experiencing no rainfall and others reaching up to 42.53 mm. Wind speed (WS10M) averages 5.76 m/s, with the maximum wind speed (WS10M\_MAX) reaching 12.51 m/s and the minimum wind speed (WS10M\_MIN) dropping as low as 0.84 m/s. Wind direction (WD10M) has considerable variability, with an average of 154.46 degrees and a range extending from 49.03 to 269.12 degrees. Finally, sea surface temperature (SST) is relatively stable, with a mean of 27°C and a range from 24.7°C to 29.8°C. These environmental factors—temperature, humidity, wind, and precipitation—are critical in understanding the transmission dynamics of dengue, as they directly influence mosquito populations and behavior.



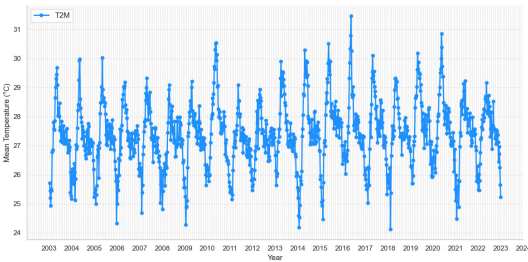
(A)



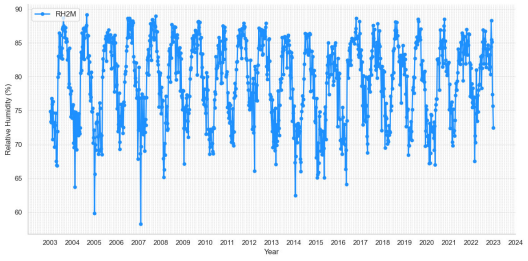
(B)



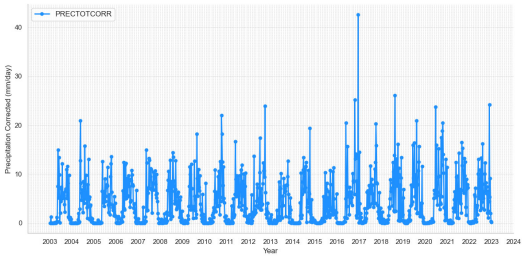
(C)



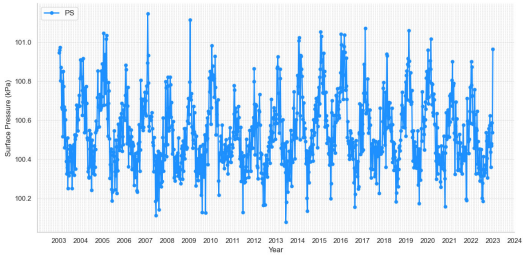
(D)



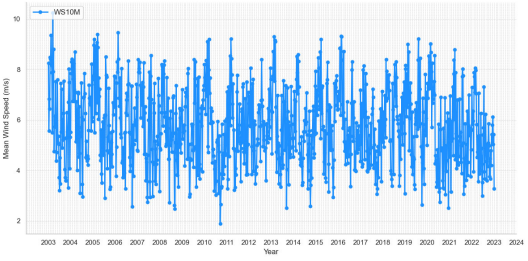
(E)



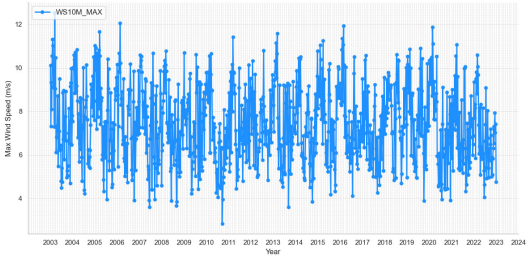
(F)



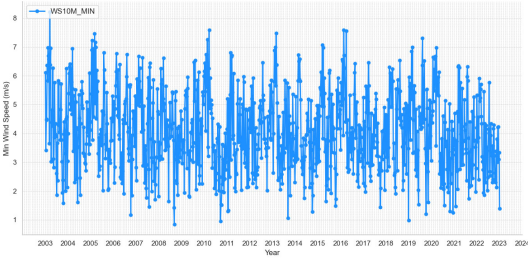
(G)



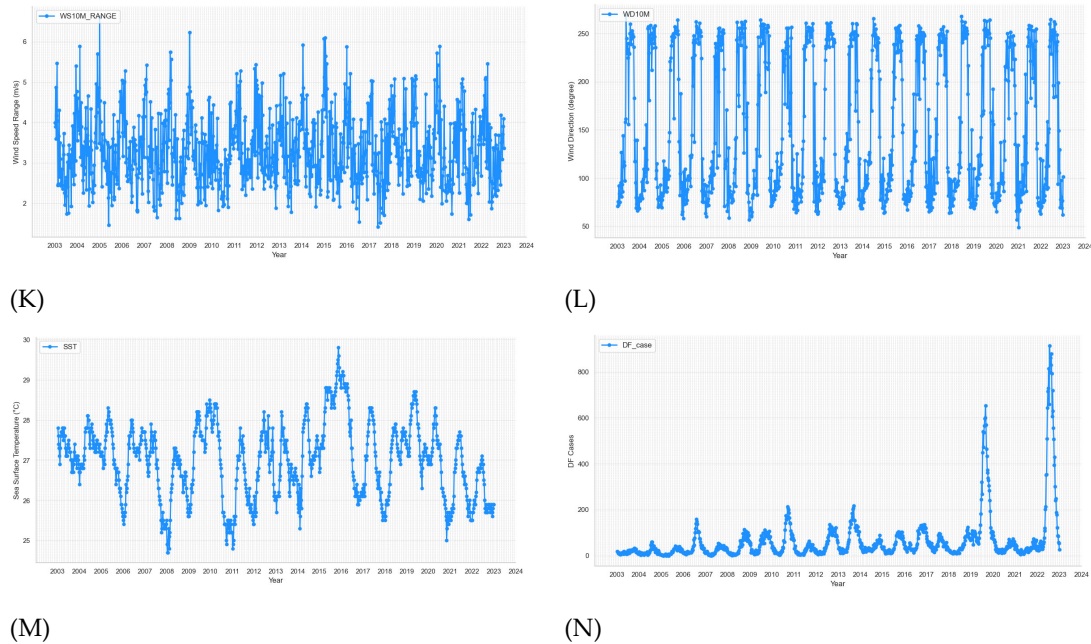
(H)



(I)



(J)



**Figure 3.** Time series plots of weekly climate variables and dengue cases from week 1 - 2003 to week 53 - 2022. (A) temperature range, (B) temperature max, (C) temperature min, (D) mean temperature, (E) relative humidity, (F) precipitation corrected, (G) surface pressure, (H) mean wind speed, (I) max wind speed, (J) min wind speed, (K) wind speed range, (L) wind direction, (M) sea surface temperature, (N) dengue cases.

**Table 2.** Descriptive statistics of weekly weather variables and dengue cases from 2003 to 2022 in Ba Ria Vung Tau, Vietnam.

| Variables   | count | mean    | std     | min    | 25%     | 50%     | 75%     | max     |
|-------------|-------|---------|---------|--------|---------|---------|---------|---------|
| DF_case     | 1044  | 64.943  | 111.994 | 0      | 15      | 32      | 68      | 913     |
| T2M_RANGE   | 1044  | 3.998   | 1.166   | 1.86   | 3.057   | 3.769   | 4.926   | 7.263   |
| T2M_MAX     | 1044  | 29.63   | 1.082   | 26.57  | 28.852  | 29.427  | 30.242  | 34.044  |
| T2M_MIN     | 1044  | 25.633  | 1.404   | 21.071 | 24.863  | 25.826  | 26.527  | 29.596  |
| T2M         | 1044  | 27.376  | 1.131   | 24.126 | 26.768  | 27.34   | 28.012  | 31.453  |
| RH2M        | 1044  | 79.275  | 5.937   | 58.286 | 74.646  | 80.393  | 84.312  | 89.134  |
| PRECTOTCORR | 1044  | 4.145   | 4.783   | 0      | 0.197   | 2.457   | 6.729   | 42.527  |
| PS          | 1044  | 100.538 | 0.184   | 100.08 | 100.407 | 100.509 | 100.661 | 101.144 |
| WS10M       | 1044  | 5.764   | 1.538   | 1.897  | 4.631   | 5.656   | 6.936   | 10.234  |
| WS10M_MAX   | 1044  | 7.406   | 1.772   | 2.851  | 6.064   | 7.304   | 8.776   | 12.514  |
| WS10M_MIN   | 1044  | 4.093   | 1.367   | 0.844  | 3.106   | 4.004   | 5.099   | 8.21    |
| WS10M_RANGE | 1044  | 3.313   | 0.88    | 1.42   | 2.657   | 3.2     | 3.896   | 6.61    |
| WD10M       | 1044  | 154.463 | 72.843  | 49.03  | 86.743  | 123.339 | 239.295 | 269.124 |
| SST         | 1044  | 26.999  | 0.934   | 24.7   | 26.3    | 27.1    | 27.6    | 29.8    |

3.3. Correlation Analysis

The Pearson correlation analysis in Figure 4 provides insights into the relationships between various climatic variables relevant to the study. A strong positive correlation is observed between

WS10M and WS10M\_MAX ( $r = 0.97, p < 0.01$ ), indicating that wind speed and its maximum values tend to move together. Similarly, T2M and T2M\_MIN exhibit a strong positive correlation ( $r = 0.95, p < 0.01$ ), suggesting that daily mean and minimum temperatures are highly related. Conversely, a moderate negative correlation is seen between T2M\_RANGE and T2M\_MAX ( $r = -0.48, p < 0.01$ ), implying that greater fluctuations in temperature range are associated with lower maximum temperatures. Additionally, PS shows weaker, but notable, negative correlations with several temperature-related variables, such as T2M\_RANGE ( $r = -0.32, p < 0.01$ ), indicating that higher pressure might be linked with smaller temperature variations. PRECTOTCORR is moderately correlated with T2M\_MIN ( $r = 0.66, p < 0.01$ ) and RH2M ( $r = 0.64, p < 0.01$ ), highlighting that temperature, particularly the minimum temperature, and humidity are key drivers of precipitation patterns. This relationship is further supported by a positive correlation between RH2M and T2M\_MIN ( $r = 0.52, p < 0.01$ ), confirming the role of temperature in influencing humidity levels.

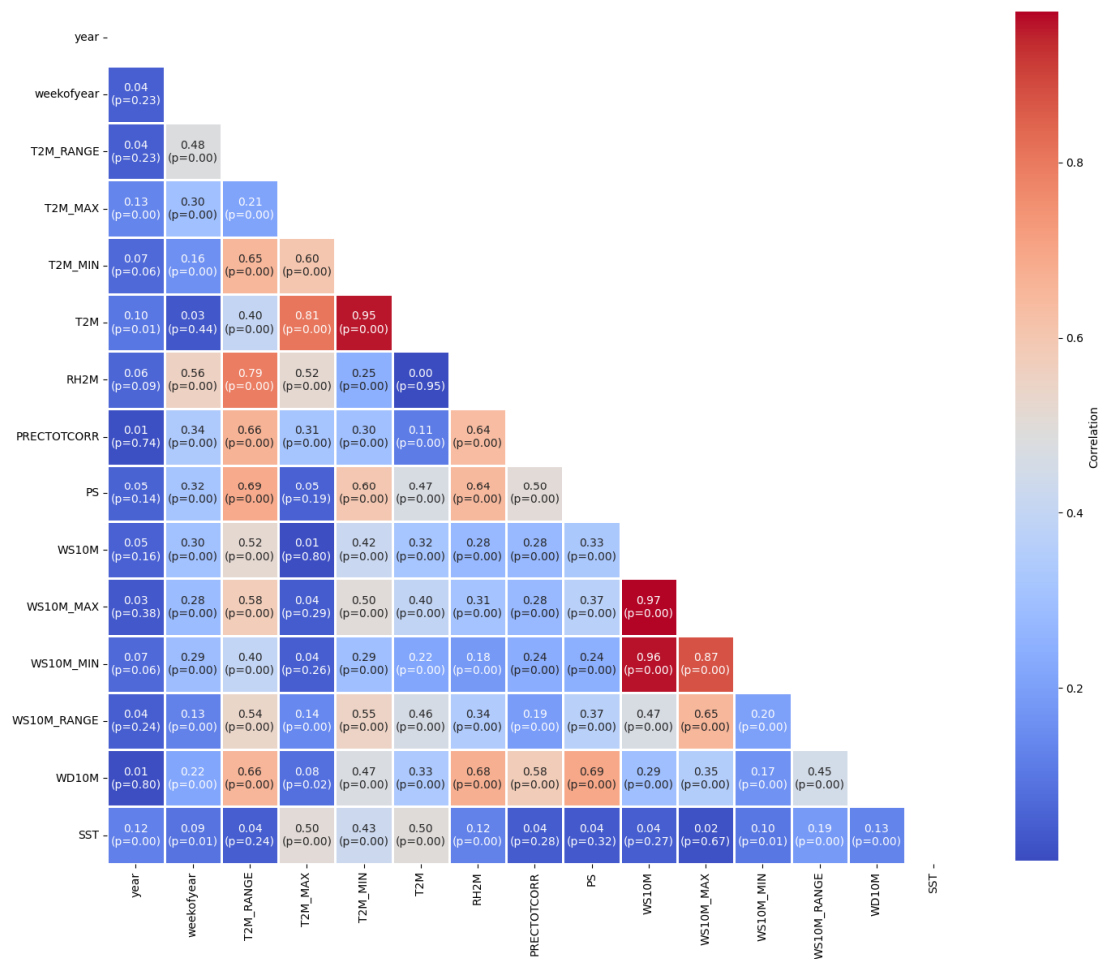
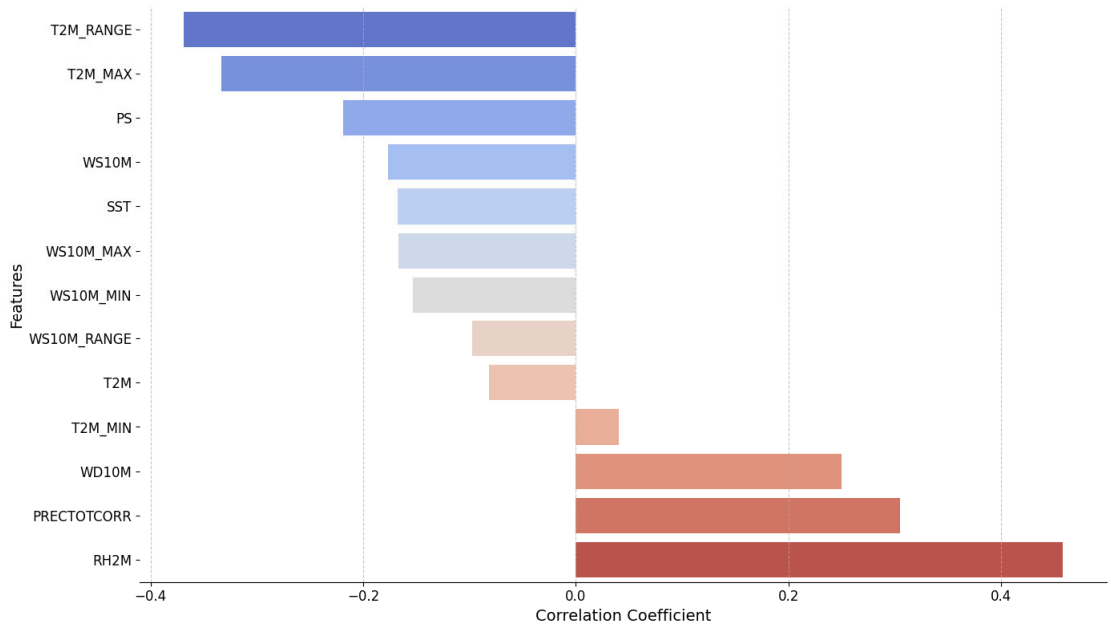


Figure 4. Correlation Matrix of Climatic Variables.

The correlation analysis between climatic variables and total cases is presented in Figure 5. The most positively correlated factor with total cases is the RH2M, with a correlation coefficient of approximately 0.4. This suggests that an increase in relative humidity is closely associated with a rise in total cases. Additionally, PRECTOTCORR, WD10M, and T2M\_MIN also show positive correlations, albeit weaker. On the other hand, the daily T2M\_RANGE has the strongest negative correlation ( $\sim -0.35$ ), indicating that an increase in this variable tends to reduce the total cases. Other factors, such as T2M\_MAX, PS, and WS10M, also exhibit negative correlations, but with less pronounced effects. Some factors, like SST, WS10M\_RANGE, and T2M, have very weak or almost negligible correlations with total cases. Therefore, we will focus on the following variables as they

provide the highest correlations including RH2M, PRECTOTCORR, WD10M, T2M\_RANGE, T2M\_MAX and PS, and further examine the correlation of their lagged versions with total cases.



**Figure 5.** Correlation of climatic variables (features) with total cases.

The analysis of the climatic factors reveals distinct relationships between specific variables and the total cases (Figure S1). Temperature-related variables demonstrate a notable negative correlation with case counts. Specifically, as temperature range increases, the total number of cases decreases, indicating that larger variations in daily temperature may contribute to a reduction in cases. Similarly, an increase in maximum temperature is associated with a slight decline in case numbers, although the effect is less pronounced. Relative humidity, on the other hand, shows a clear positive correlation with the total number of cases. Higher humidity levels are consistently linked with an increase in cases, suggesting that humidity could play a significant role in disease transmission or exacerbation. Precipitation does not display a consistent trend; however, there is a marked increase in cases at certain points, followed by fluctuations, which implies a possible non-linear or threshold-dependent relationship. Surface pressure appears to have a negligible effect, as total cases remain relatively constant across varying pressure levels. Finally, wind direction exhibits considerable variability, with no discernible trend, indicating that its influence on case counts may be highly context-specific or modulated by other factors.

Figure S2 shows that the variables with the highest correlation for total cases are T2M\_RANGE\_shift\_14, T2M\_MAX\_shift\_20, RH2M\_shift\_6, PRECTOTCORR\_shift\_10, PS\_shift\_16, WD10M\_shift\_14.

3.4. Negative Binomial Regression (NBR)

Negative Binomial Regression (NBR) was the chosen method in this study to predict the number of dengue cases (total\_cases). The target variable is a count variable, and the dataset exhibited overdispersion (variance much larger than the mean), which makes NBR a more appropriate model than Poisson regression.

Distribution analysis of the target variable showed that the mean number of dengue cases was 43.73, while the variance was 1692.84, confirming that the data had high variability (Figure S3). The histogram of the target variable showed a right-skewed distribution, with most values concentrated on the left and tapering off towards the right. The KDE curve (Figure S3) also reflected this dispersion,

which aligns with the negative binomial distribution, where the probability of occurrence decreases as the number of cases increases.

The first negative binomial regression model was built with six initial variables. The results indicated that the model had an optimal alpha of approximately 0.034, with a Mean Absolute Error (MAE) of 26.035 on the test set and 26.179 on the training set (Table 3 and Figure 11A). The small difference between these MAE values suggests that the model does not overfit and generalizes well.

Table 3. Generalized Linear Model Regression Results for Model 1.

|             | coef    | std err | z       | P> z  | [0.025 | 0.975]   |
|-------------|---------|---------|---------|-------|--------|----------|
| Intercept   | 20.2737 | 10.862  | 1.866   | 0.062 | -1.016 | 41.563   |
| T2M_RANGE   | -0.1264 | 0.017   | -7.242  | 0     | -0.161 | -0.092   |
| T2M_MAX     | -0.2771 | 0.016   | -16.913 | 0     | -0.309 | -0.245   |
| RH2M        | 0.0563  | 0.005   | 12.386  | 0     | 0.047  | 0.065    |
| PRECTOTCORR | -0.0141 | 0.003   | -4.037  | 0     | -0.021 | -0.007   |
| PS          | -0.1235 | 0.104   | -1.185  | 0.236 | -0.328 | 0.081    |
| WD10M       | -0.0004 | 0       | -1.615  | 0.106 | -0.001 | 8.00E-05 |

In the second model, additional variables were included, such as the time variable (month), which significantly improved the model’s performance. The optimal alpha decreased to 0.022, and the MAE values for both the training and test sets dropped to 25.916 and 25.556, respectively. This indicates an improvement in predictive accuracy (Figure 11B).

The third model introduced time-shifted variables, which further improved prediction accuracy, reducing the test set MAE to 24.845. Time-shifted variables helped the model better capture temporal trends and fluctuations in the data, improving its overall performance (Figure 11C).

Finally, the fourth model, which incorporated time-shifted variables with lag, produced the best prediction results, with an alpha = 1.95. The MAE for the test set decreased to 21.409, significantly lower than the previous models. This improvement indicates that using lagged variables allowed the model to capture more complex fluctuations in the data, resulting in more accurate predictions (Figure 11D).

In summary, the negative binomial regression model with time-shifted and lagged variables demonstrated superior performance, offering more accurate predictions than the previous models. Climate variables such as T2M\_RANGE\_shift\_14 and PS\_shift\_16 were identified as crucial factors in predicting dengue cases, along with seasonal factors (Figure 6).

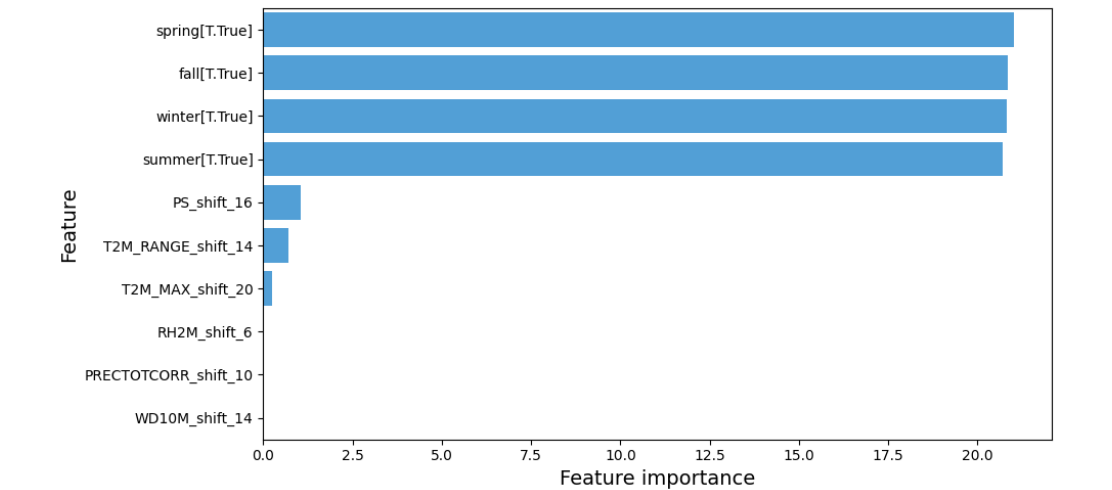


Figure 6. Relative importance of features in predicting dengue cases.

3.5. SARIMAX Model

To address seasonality and autocorrelation in the dataset, the SARIMAX model was selected to predict dengue cases. The data was resampled from weekly to monthly, as hyperparameter tuning using `pm.auto_arma` was not effective with weekly data (Figure 7).

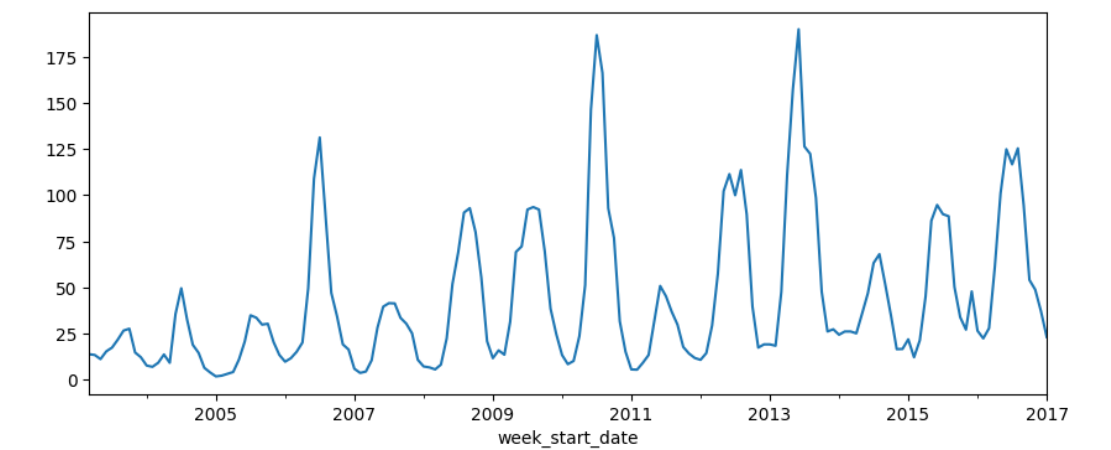


Figure 7. Time series of monthly dengue cases from the training set.

The analysis began with Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to assess the degree of autocorrelation within the time series. These plots helped determine the order of the AutoRegressive (AR) and Moving Average (MA) terms. The PACF suggested that the AR order ( $p$ ) could be 1 or 2, and the ACF indicated the MA order ( $q$ ) could be 1, 2, or 3 (Figure 8).

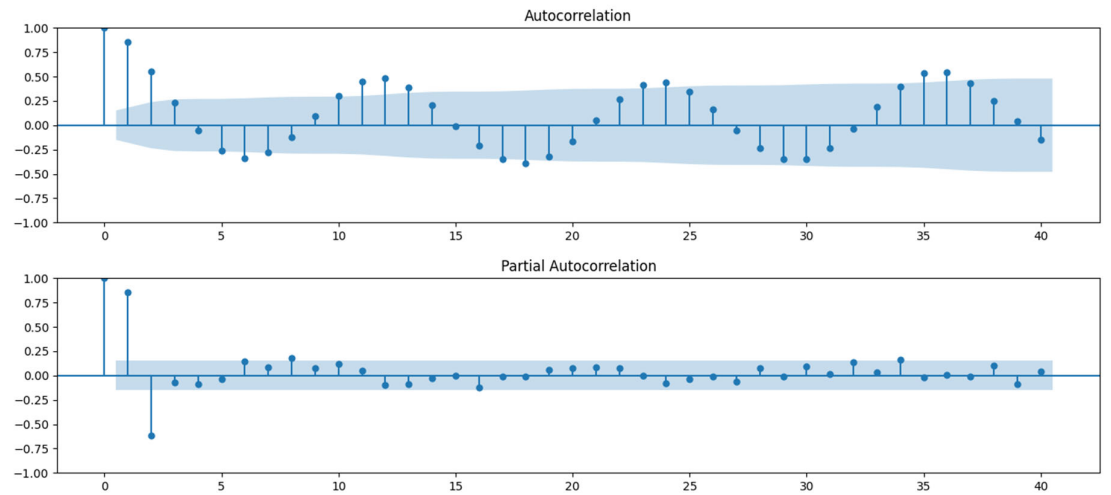


Figure 8. ACF (top) and PACF (bottom) plots.

Next, seasonal decomposition was applied to examine the trend and seasonality of the data, which clearly showed a seasonal pattern with no strong trend (Figure S4). A stationarity test confirmed that the time series was stationary, allowing  $d$  and  $D$  to be set to 0 in the SARIMAX model.

3.5.1. SARIMAX #1 Baseline Model

With 80% of the data allocated for training and 20% for testing (Figure 9) is applied to all models in this study. The first SARIMAX model used only `total_cases` as the predictor variable. The

auto\_arima() function identified the best-fitting model as ARIMA(2 0 0)(1 0 1)[12], which captures both autoregressive and seasonal moving average components (Table S1).

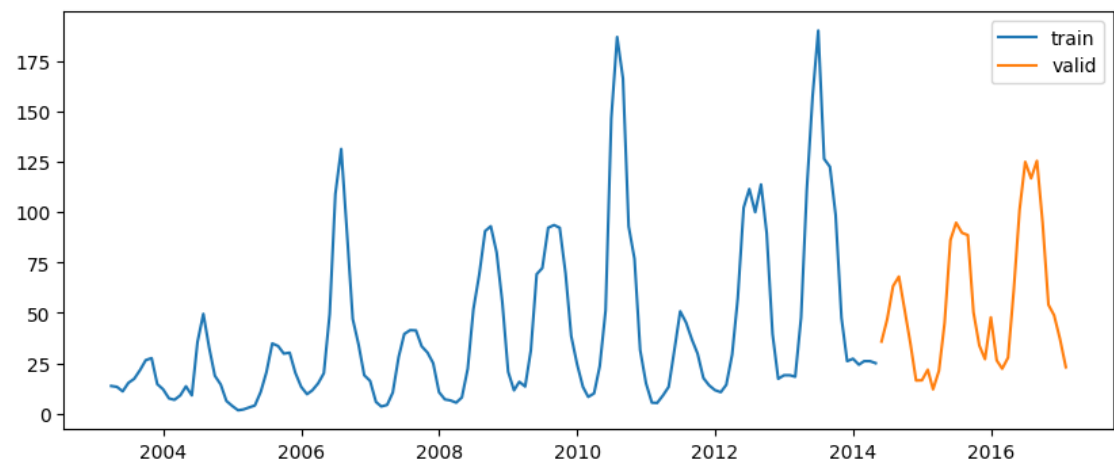


Figure 9. Train-test data split for all models in this study.

The SARIMAX model fit the data well, with statistically significant AR and MA parameters, capturing the autocorrelation and seasonality of the dataset. However, diagnostic tests revealed that while the model fit was good (no significant autocorrelation left in the residuals), the residuals showed positive skewness and high kurtosis (Table S1, Figure S5). This suggests that some outliers or non-normal distribution of residuals may be present.

The MAE and RMSE scores for the training and test sets highlight a significant overfitting issue: MAE\_train = 10.54 vs. MAE\_test = 20.31, and RMSE\_train = 16.10 vs. RMSE\_test = 27.19 (Figure 11E). The model performed well on the training data but struggled with the test data, indicating that the baseline SARIMAX model is overfitted.

3.5.2. SARIMAX #2 Full Multivariate Model

A second SARIMAX model was developed using lagged exogenous variables such as T2M\_RANGE\_shift\_14, T2M\_MAX\_shift\_20, RH2M\_shift\_6, and others as predictors (Figure 9). The best model identified was ARIMA(2 0 0)(1 0 0)[12].

However, the model did not converge, suggesting that either the parameter orders or the initial parameter values were not optimal. The statistical results (Table S2) showed that most independent variables were not statistically significant, with the exception of a few AR parameters. Despite the improved model, it still exhibited issues such as heteroskedasticity and residuals not following a normal distribution (Figure S6).

The MAE and RMSE scores for SARIMAX #2, MAE\_train = 11.37 and MAE\_test = 17.02, showed slight improvement compared to the baseline model. However, the model still failed to predict the test data accurately (Figure 11F).

3.6. XGBoost Regression

3.6.1. Model #1

For the first XGBoost regression model, we used the original variables: total\_cases, month, T2M\_RANGE, T2M\_MAX, RH2M, PRECTOTCORR, PS, and WD10M. The dataset was split into 80% training and 20% testing sets, with a 5-fold cross-validation performed (Figure 11G). Hyperparameter tuning was conducted using GridSearchCV, which optimized for neg\_mean\_squared\_error during cross-validation. The best parameters obtained were as follows: colsample\_bytree: 0.5, gamma: 1, learning\_rate: 0.2, max\_depth: 8, min\_child\_weight: 3, objective: reg, subsample: 1

The Mean Absolute Error (MAE) for the best model was -1596.43, and the best score showed that the model had overfitting issues, with a much lower error on the training set than the test set. Specifically, the training MAE was 1.07, while the test MAE was 21.77. Similarly, the Root Mean Squared Error (RMSE) was 1.46 for training and 29.73 for testing, further highlighting overfitting (Figure 11G).

These results suggest that the model fits the training data well but fails to generalize to new data, likely due to overfitting caused by a high max\_depth and min\_child\_weight, which made the model overly complex.

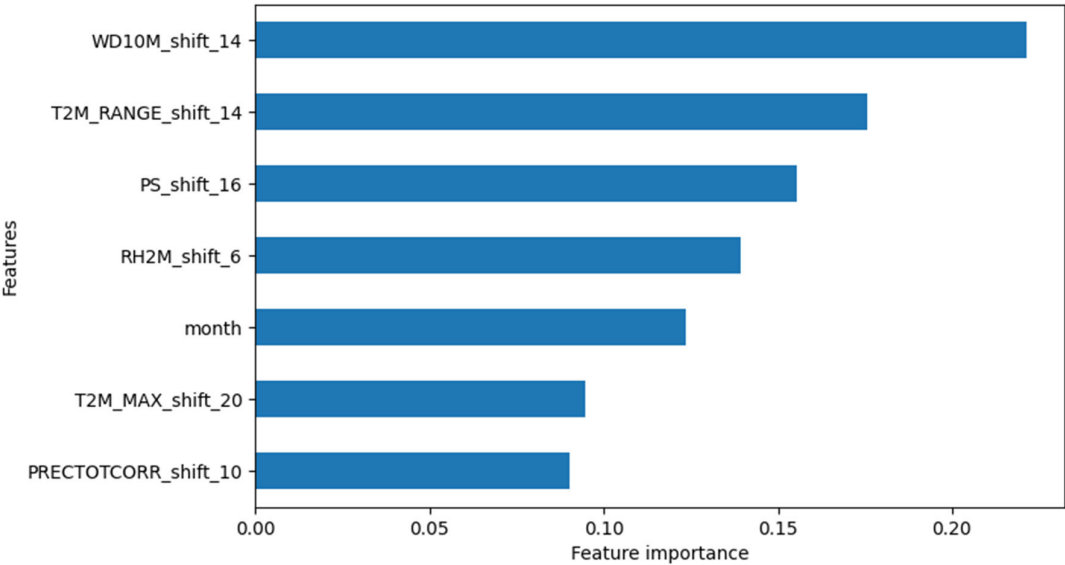
3.6.2. Model #2

In the second model, we introduced lagged variables such as T2M\_RANGE\_shift\_14, T2M\_MAX\_shift\_20, RH2M\_shift\_6, PRECTOTCORR\_shift\_10, and others (Figure 11H). The best parameters for this model were: colsample\_bytree: 0.75, gamma: 1, learning\_rate: 0.2, max\_depth: 4, min\_child\_weight: 1, objective: reg, subsample: 0.75.

The MAE for the best model was -1355.61, and while the performance improved slightly, the model still exhibited overfitting. The training MAE was 6.63, while the test MAE increased to 24.45. The RMSE was 13.04 for training and 30.97 for testing (Figure 11H).

After refitting the model using the entire dataset, predictions showed two additional dengue peaks in early 2018 and 2022 (Figure 11H' and Figure S7).

In terms of feature importance, WD10M\_shift\_14 (14-week lagged wind direction) and T2M\_RANGE\_shift\_14 (14-week lagged temperature range) were identified as the most important variables for predicting dengue cases (Figure 10).



**Figure 10.** Relative feature importance in predicting dengue cases for XGB Regression Model #2.

3.7. LSTM Neural Network

For the LSTM model, the input data was structured as a 3D tensor with shape (batch\_size, timesteps, input\_dim). The data preprocessing involved scaling and converting both the training and testing sets using a scaler before applying predictions (Figure 9).

3.7.1. LSTM Model #1

A simple LSTM model with a single hidden layer of LSTM units and one output layer was implemented. After 35 epochs, the model showed convergence in both the training and testing sets, as visualized in **Figure 10A**. However, despite the convergence, the model’s performance metrics

indicate room for improvement. The MAE was 23.73 for the training set and 28.86 for the test set. The RMSE was 38.03 for the training set and 38.65 for the test set (Figure 11I).

While the model captured the overall seasonality of the data, it failed to detect individual peaks and outbreaks in both the training and test sets. This suggests that a more complex architecture is required to capture these finer details.

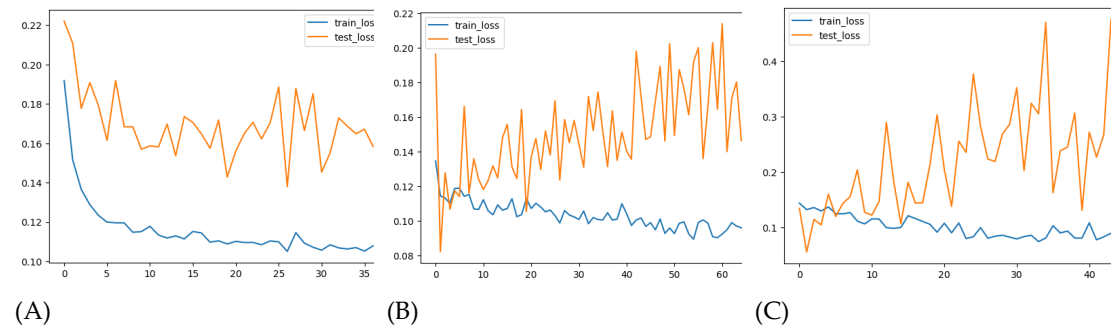
3.7.2. LSTM Model #2

A deeper LSTM model with more layers and neurons was trained to improve performance. While the model converged well on the training data (Figure 10B), it struggled to generalize to the test set, as evidenced by fluctuations in the test loss. The MAE and RMSE slightly improved but remained unsatisfactory, with MAE\_train at 18.70 and MAE\_test at 18.14. The RMSE was 29.92 for training and 24.37 for testing (Figure 11J).

This model showed some improvement, but it still could not capture individual peaks or outbreaks, suggesting the need for additional tuning or alternative input features.

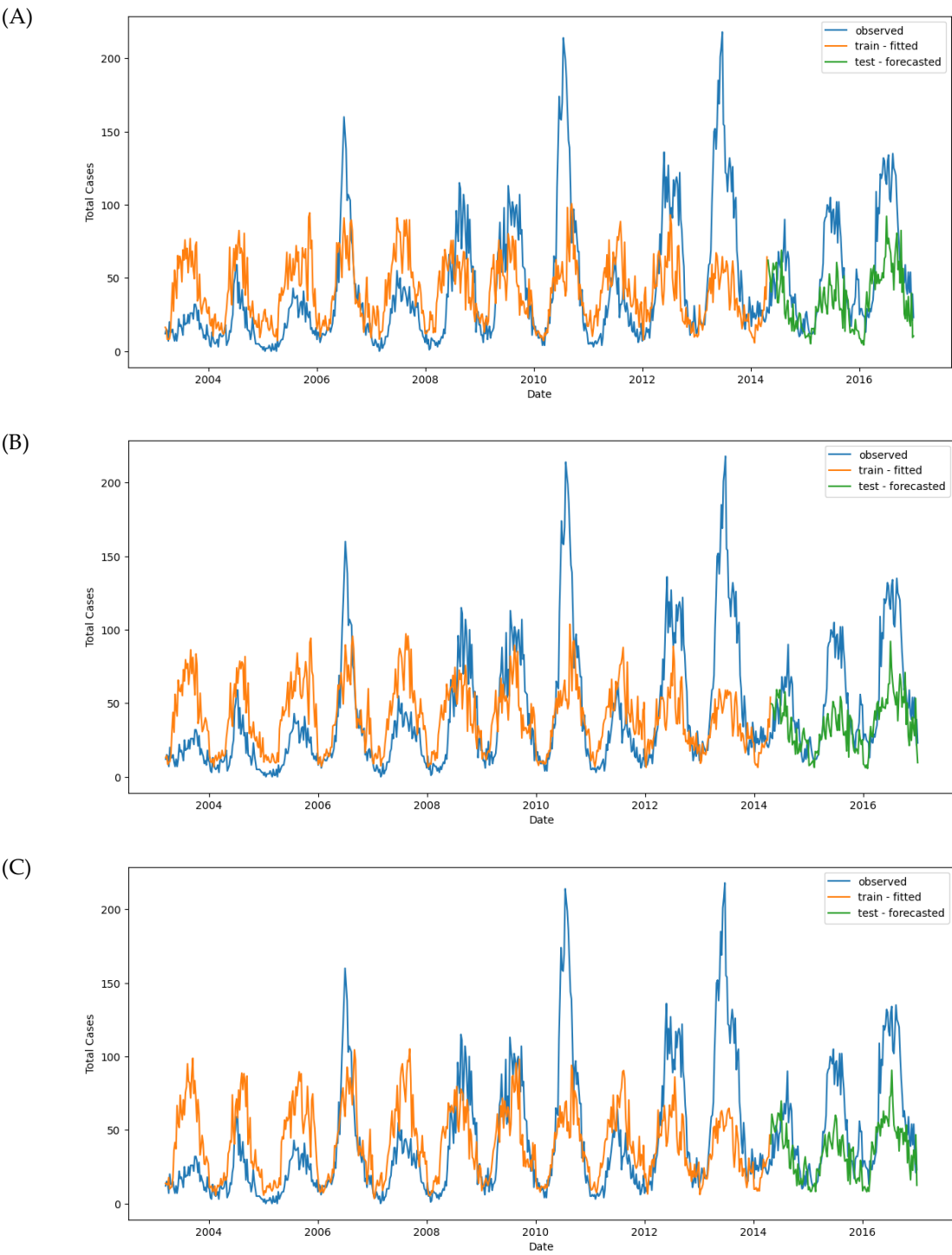
3.7.3. LSTM Model #3

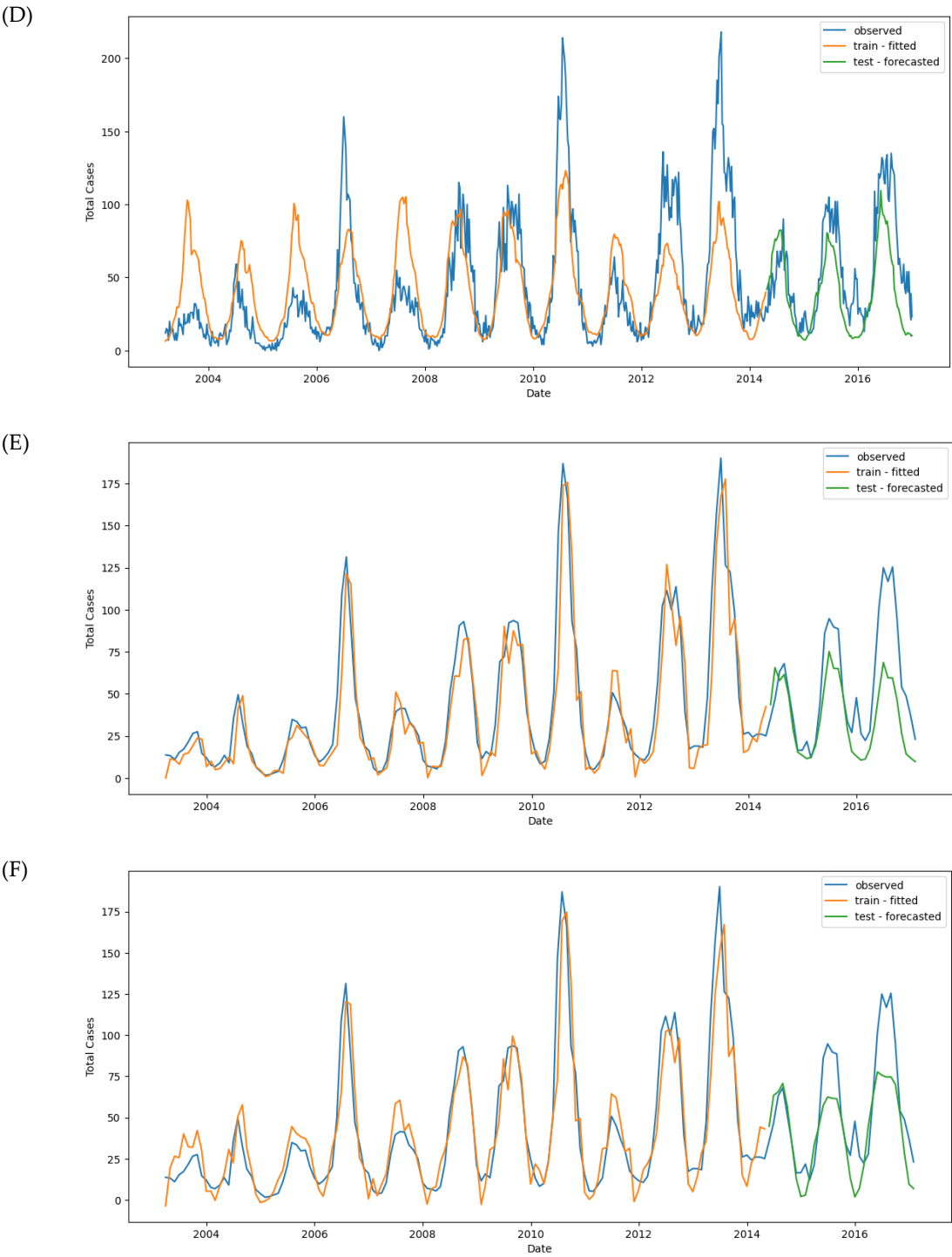
In this iteration, lagged variables were introduced into the model, which included total\_cases, month, T2M\_RANGE\_shift\_14, T2M\_MAX\_shift\_20, RH2M\_shift\_6, PRECTOTCORR\_shift\_10, PS\_shift\_16, and WD10M\_shift\_14. Despite adding these features, the model encountered difficulty generalizing on the test set, as shown in Figure 10C. The MAE on the training set was 13.89, while the test MAE increased to 24.86. Similarly, the RMSE was 20.54 for training and 34.92 for testing (Figure 11K).

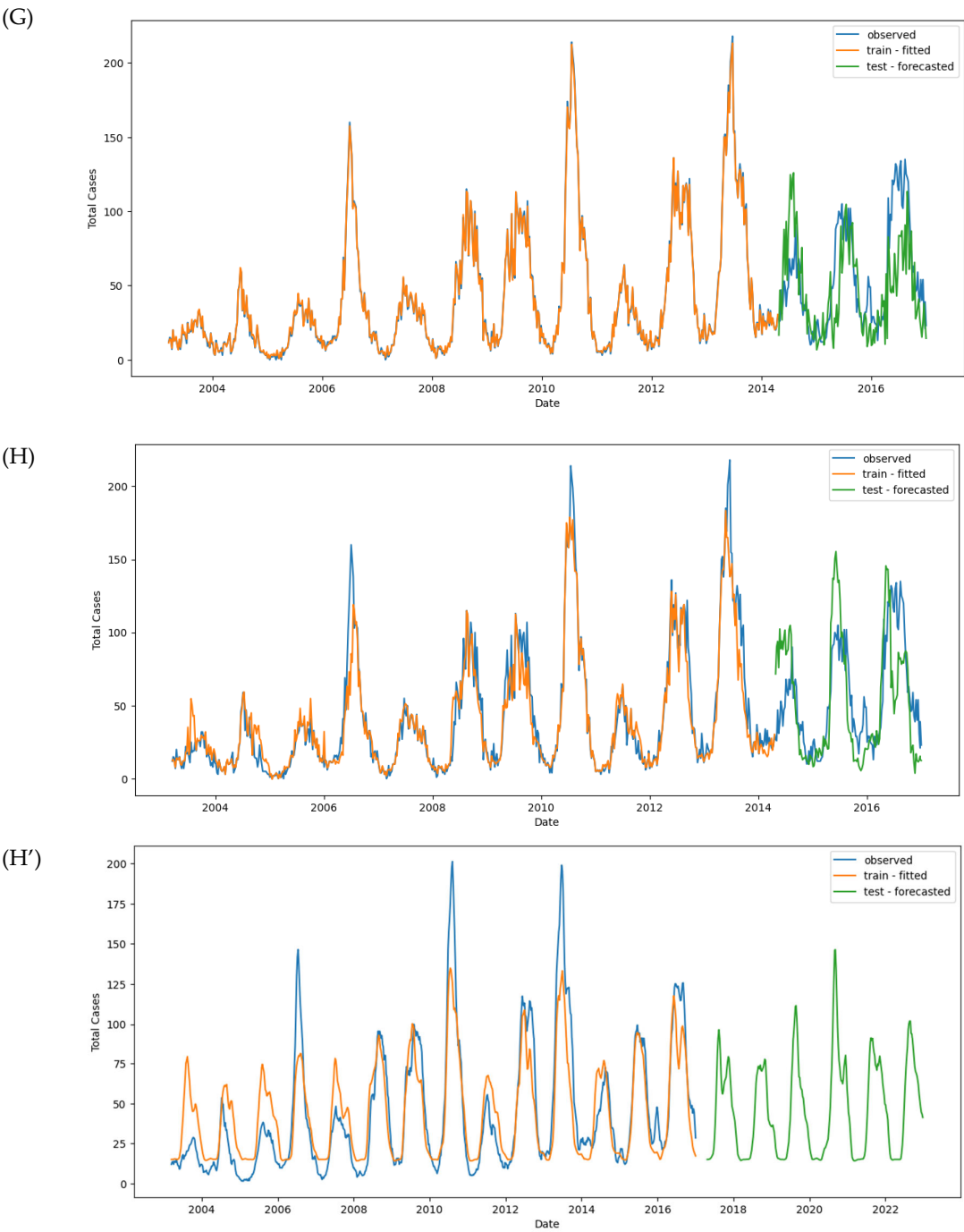


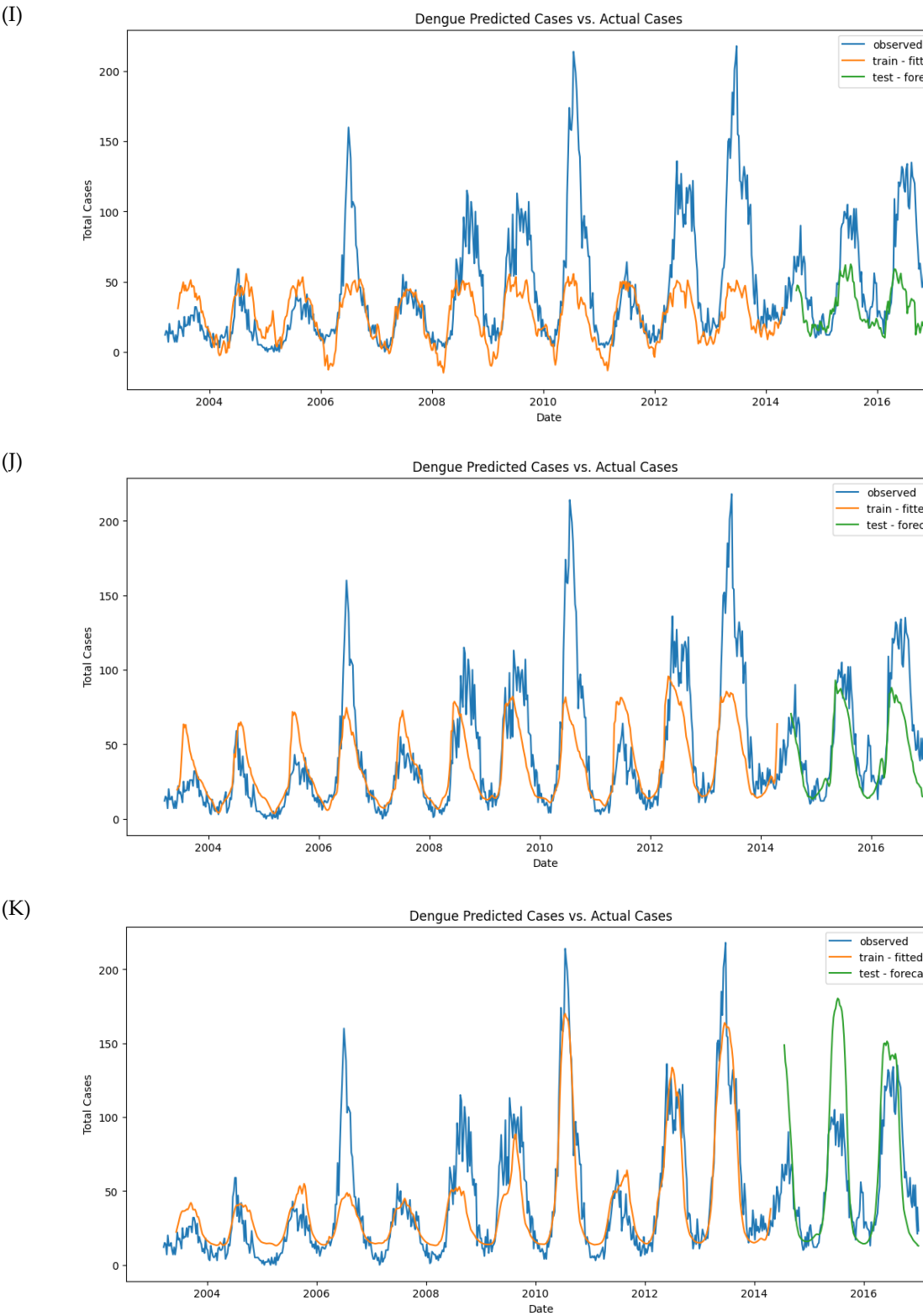
**Figure 10.** Loss convergence on training and test sets for LSTM. (A) Model #1, (B) Model #2, (C) Model #3.

Despite using a more complex LSTM model and incorporating lagged variables, the model still struggled to predict the large values in the test set accurately. While the model improved over the previous versions, it still failed to detect individual peaks or outbreaks in the data.









**Figure 11.** Observed and predicted dengue cases in training (week 1 of 2003 to week 17 of 2017) and test data (week 18 of 2017 to week 53 of 2022) of different models. (A) NBR #1, (B) NBR #2, (C) NBR #3, (D) NBR #4, (E) SARIMAX #1, (F) SARIMAX #2, (G) XGBoost #1, (H) XGBoost #2, (H') XGBoost #2 (refit), (I) LSTM #1, (J) LSTM #2, (K) LSTM #3.

3.8. Assessment of Predictive Performance

This section evaluates the predictive performance of four models: NBR, SARIMAX, XGBoost, and LSTM Neural Network, in predicting dengue cases. Key performance metrics include Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which were computed for both the training and testing datasets. The Table 6 below summarizes the performance of each model.

**Table 6.** Summary of Predictive Performance for NBR, SARIMAX, XGBoost, and LSTM Models.

| Model      | MAE<br>(Train) | MAE<br>(Test) | RMSE<br>(Train) | RMSE<br>(Test) | Key Observations                                                                |
|------------|----------------|---------------|-----------------|----------------|---------------------------------------------------------------------------------|
| NBR #1     | 26.321         | 25.822        | 37.380          | 33.413         | Reasonable generalization but struggles with larger fluctuations.               |
| NBR #2     | 25.916         | 25.556        | 37.397          | 33.502         | Slight improvement; still lacks peak detection.                                 |
| NBR #3     | 25.426         | 24.846        | 37.101          | 32.364         | Better generalization, but key outbreak peaks remain undetected.                |
| NBR #4     | 20.748         | 21.409        | 29.755          | 26.427         | Significant improvement in predictive accuracy and generalization.              |
| SARIMAX #1 | 10.539         | 20.307        | 16.101          | 27.190         | Strong overfitting; poor generalization to test data.                           |
| SARIMAX #2 | 11.367         | 17.017        | 15.786          | 22.635         | Reduced overfitting; still requires fine-tuning for better test performance.    |
| XGBoost #1 | 1.074          | 21.767        | 1.461           | 29.732         | Severe overfitting; excellent train performance but poor test generalization.   |
| XGBoost #2 | 6.631          | 24.450        | 13.035          | 30.973         | Minor improvement with lagged variables, overfitting persists.                  |
| LSTM #1    | 23.731         | 28.856        | 38.031          | 38.650         | Captures seasonality but fails to detect individual peaks or outbreaks.         |
| LSTM #2    | 18.704         | 18.143        | 29.922          | 24.368         | Deeper model improves fit, but test performance remains inconsistent.           |
| LSTM #3    | 13.890         | 24.859        | 20.544          | 34.920         | Improved training results, but test performance is weak, especially with peaks. |

Each of the four models showed distinct advantages and limitations in predicting dengue cases. **NBR** demonstrated steady improvement across all iterations, with Model #4 providing the most balanced performance, showing improvements in both predictive accuracy and generalization. **SARIMAX** performed well in capturing seasonality but suffered from overfitting, particularly in Model #1, while Model #2 reduced this issue but still required further refinement to enhance test performance. **XGBoost** demonstrated excellent performance on training data but failed to generalize well to unseen data due to overfitting, despite slight improvements with lagged variables in Model

#2. **LSTM** showed potential in capturing temporal dependencies and seasonality, especially in Model #2, but consistently struggled with generalizing to test data and accurately predicting outbreaks or peaks. In summary, **NBR** Model #4 emerged as the most reliable and well-rounded model for dengue case prediction, whereas **SARIMAX** and **XGBoost** models could benefit from further optimization. **LSTM**, despite its promise for long-term temporal predictions, requires additional refinement to effectively capture short-term peaks and outbreak patterns.

#### 4. Discussion

The results of this study align with previous research, reinforcing the critical role of climate variables such as temperature, humidity, and precipitation in driving dengue transmission. Notably, RH2M, PRECTOTCORR, and T2M\_RANGE, T2M\_MAX were identified as key factors affecting dengue incidence in BRVT. These findings are consistent with studies conducted in other tropical regions, which have highlighted the influence of climate conditions on mosquito breeding and virus transmission.

Several studies have confirmed that climatic factors such as temperature and humidity play a pivotal role in shaping the dynamics of dengue outbreaks. For instance, research in Southern China demonstrated that higher temperatures and humidity are directly correlated with increases in mosquito populations, which in turn escalates dengue transmission [1]. Similarly, a Bayesian analysis in Queensland, Australia, found that increases in rainfall and higher temperatures significantly raised the likelihood of dengue outbreaks [3]. Our study corroborates these findings, with RH2M and PRECTOTCORR both showing strong positive correlations with dengue cases in BRVT, suggesting that humid and wet conditions provide favorable environments for *Aedes aegypti* mosquitoes, the primary vector for dengue.

In comparison to models used in other regions, such as Brazil and Mexico, where dynamic ensemble learning and multi-stage machine learning models have been effectively utilized to predict dengue outbreaks [5], our study introduces a novel approach by integrating both meteorological data and time-lagged variables. This approach allows for the modeling of delayed effects of climatic conditions on dengue incidence, which is particularly crucial given the mosquito lifecycle and the virus incubation period.

Among the models evaluated, the Negative Binomial Regression (NBR) model performed the best, particularly when incorporating lagged climate variables, achieving the lowest mean absolute error (MAE) on both the training and test sets. The effectiveness of NBR in handling overdispersed count data aligns with its successful application in similar studies focused on vector-borne diseases. For instance, a study in Mexico employed a multi-stage machine learning approach to predict dengue, with lagged climate data significantly improving model performance [6].

Our study's use of lagged temperature and humidity variables (e.g., T2M\_MAX\_shift\_20, RH2M\_shift\_6) reflects a common strategy in dengue forecasting, as demonstrated in research conducted in central Vietnam and Southern China, where lagged climatic variables proved crucial in capturing the delayed impacts of weather on dengue transmission [8]. The inclusion of lagged variables further enhanced the generalization capabilities of our models, reducing overfitting issues observed in other models such as SARIMAX and LSTM.

This is consistent with findings from other studies that applied SARIMA models to time-series data, where inclusion of external variables often improves the model's predictive accuracy, as observed in the study of climate-based dengue forecasting in Vietnam [2]. However, the introduction of exogenous variables in our SARIMAX model only provided marginal improvements, indicating the need for further tuning of seasonal components and better feature selection.

On the other hand, the LSTM model, while promising for long-term forecasting, struggled to capture outbreak peaks, which are crucial for timely public health interventions. Similar challenges were observed in other studies applying LSTM models for forecasting dengue, where capturing outbreak peaks remains challenging [10].

The successful application of machine learning techniques such as NBR and XGBoost in this study demonstrates the potential of these models to provide accurate, climate-based early warning

systems for dengue outbreaks. Accurate prediction models enable health authorities to implement timely control measures, such as mosquito population reduction and public awareness campaigns, before outbreaks occur. The inclusion of climate variables such as temperature, humidity, and precipitation provides a more nuanced understanding of dengue transmission, aligning with global efforts to integrate meteorological data into vector-borne disease forecasting systems [2].

However, there are limitations to our study. First, the reliance on climate data from NASA/POWER CERES/MERRA2 may introduce uncertainties, particularly with imputation techniques used for missing data. This issue is common in studies using remote sensing data, as gaps in data often require sophisticated imputation methods, which can affect model accuracy [10]. Moreover, while the models were able to capture broad trends in dengue incidence, they were less successful in predicting exact outbreak peaks, particularly in years like 2019 and 2022, where dengue cases spiked unexpectedly. This indicates that additional variables, such as socio-environmental factors and mosquito population dynamics, should be integrated into future models to improve prediction accuracy.

Future research should focus on integrating additional environmental and socio-demographic factors, such as population density, land use, and mosquito abundance, to further enhance the predictive power of dengue forecasting models. Ensemble approaches, which combine the strengths of multiple models, could also be explored to improve accuracy. Moreover, the application of more sophisticated deep learning architectures, such as hybrid LSTM models, could help overcome challenges in detecting short-term dengue outbreaks, as seen in recent research on vector-borne diseases [10].

## 5. Conclusion

This study has demonstrated the effectiveness of climate-based machine learning models in forecasting dengue outbreaks in BRVT, Vietnam. By leveraging advanced models such as NBR, SARIMAX, XGBoost, and LSTM, we were able to capture key climatic factors—such as temperature, humidity, wind speed, and precipitation—that significantly impact dengue transmission. Among the models tested, the NBR model incorporating lagged climate variables proved to be the most accurate and generalizable in predicting weekly dengue cases, highlighting the value of considering delayed climate effects in disease forecasting.

While this research adds to the growing body of literature supporting the integration of meteorological data into public health models, there remain challenges, particularly in capturing short-term dengue outbreaks. Models like SARIMAX and LSTM exhibited overfitting and struggled with predicting peak outbreaks. This suggests the need for further refinement of these models, possibly through the inclusion of additional socio-environmental factors and ensemble modeling approaches.

**Supplementary Materials:** The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

**Funding:** This research received no external funding.

**Compliance with Ethical Standards:** This article does not involve any studies conducted by the authors that included human participants.

**Acknowledgments:** The completion of this research work was made possible through the collaborative efforts and dedication of a multidisciplinary team. We extend our sincere appreciation to each member for their invaluable contributions.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Chen, J., et al., *Collaboration between meteorology and public health: Predicting the dengue epidemic in Guangzhou, China, by meteorological parameters*. *Frontiers in Cellular and Infection Microbiology*, 2022. 12: p. 881745.

2. Colón-González, F.J., et al., *Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles*. PLoS medicine, 2021. **18**(3): p. e1003542.
3. Akter, R., et al., *Climate variability, socio-ecological factors and dengue transmission in tropical Queensland, Australia: A Bayesian spatial analysis*. Environmental Research, 2021. **195**: p. 110285.
4. Xu, J., et al., *Forecast of dengue cases in 20 Chinese cities based on the deep learning method*. International journal of environmental research and public health, 2020. **17**(2): p. 453.
5. McGough, S.F., et al., *A dynamic, ensemble learning approach to forecast dengue fever epidemic years in Brazil using weather and population susceptibility cycles*. Journal of The Royal Society Interface, 2021. **18**(179): p. 20201006.
6. Appice, A., et al., *A multi-stage machine learning approach to predict dengue incidence: a case study in Mexico*. Ieee Access, 2020. **8**: p. 52713-52725.
7. Salim, N.A.M., et al., *Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques*. Scientific reports, 2021. **11**(1): p. 939.
8. Nguyen, L.T., et al., *Impact of climate variability and abundance of mosquitoes on dengue transmission in central Vietnam*. International journal of environmental research and public health, 2020. **17**(7): p. 2453.
9. Pham, N.T., C.T. Nguyen, and H.H. Vu, *Assessing and modelling vulnerability to dengue in the Mekong Delta of Vietnam by geospatial and time-series approaches*. Environmental Research, 2020. **186**: p. 109545.
10. Mudele, O., et al., *Dengue vector population forecasting using multisource earth observation products and recurrent neural networks*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021. **14**: p. 4390-4404.
11. Patil, S. and S. Pandya, *Forecasting dengue hotspots associated with variation in meteorological parameters using regression and time series models*. Frontiers in public health, 2021. **9**: p. 798034.
12. Thiruchelvam, L., et al. *Inclusion of Climate Variables for Dengue Prediction Model: Preliminary Analysis*. in 2021 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). 2021. IEEE.
13. Hilbe, J., *Negative binomial regression*. 2011, Cambridge University Press.
14. Bergmeir, C., R.J. Hyndman, and J.M. Benítez, *Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation*. International journal of forecasting, 2016. **32**(2): p. 303-312.
15. Hastie, T., R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. 2017, Springer.
16. Lowe, R., et al., *Climate services for health: predicting the evolution of the 2016 dengue season in Machala, Ecuador*. The lancet Planetary health, 2017. **1**(4): p. e142-e151.
17. Sutherland, C., et al., *Practical advice on variable selection and reporting using Akaike information criterion*. Proceedings of the Royal Society B, 2023. **290**(2007): p. 20231261.
18. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system*. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.
19. Choi, J.Y. and B. Lee, *Combining LSTM network ensemble via adaptive weighting for improved time series forecasting*. Mathematical problems in engineering, 2018. **2018**(1): p. 2470171.
20. Hochreiter, S., *Long Short-term Memory*. Neural Computation MIT-Press, 1997.
21. Wang, Y., et al. *Water quality prediction method based on LSTM neural network*. in *2017 12th international conference on intelligent systems and knowledge engineering (ISKE)*. 2017. IEEE.
22. Kuhn, M., *Applied predictive modeling*. 2013, Springer.
23. Hyndman, R., *Forecasting: principles and practice*. 2018: OTexts.
24. Shmueli, G. and J. Polak, *Practical time series forecasting with r: A hands-on guide*. 2024: Axelrod schnall publishers.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.