NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Predicting Dengue Fever Incidence and Disease Dynamics under Climate Change in Southeast Asia

Josephine Lutter

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Predicting Dengue Fever Incidence and Disease Dynamics under Climate Change in Southeast Asia

by

Josephine Lutter

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Roberto Henriques, PhD, Nova Information Management School

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Berlin, July 1st 2024]

ABSTRACT

Dengue fever is a climate-sensitive vector-borne disease primarily transmitted by Aedes mosquitoes, A. aegypti and A. albopictus. Previous research has analyzed the relationship between climate and disease, with varying outcomes. Temperature and precipitation have been demonstrated as relevant predictors in most studies. The effects of climate change on dengue fever were found to be uncertain, highlighting the need for further study.

This study analyzed how environmental variables interact with disease transmission, enabling predictive modeling to forecast dengue incidence. For deployment, climate change simulations were used as a framework to assess the disease's response to changing environmental factors. The incidence and environmental data for 17 Southeast Asian locations were collected from the national Ministries of Health and the National Oceanic and Atmospheric Administration (NOAA) from 2016-2023. Traditional machine learning and deep learning models were used to forecast dengue incidence based on ten input features of temperature, precipitation, and lagged observations.

The predictive ability was evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Deep and machine learning models showed similar results for predicting dengue incidence. The Convolutional Neural Network (CNN) achieved the lowest error with an average MAE of 10.10 and RMSE of 13.61 on the validation set. Models showed varying predictive abilities across locations. Despite extensive data preparation, some locations performed worse on all models, indicating potential issues with initial data quality. Errors were reduced for all models on the test set, with CNN demonstrating superior with an average MAE of 5.06 and RMSE of 7.09. Although errors decreased with additional data, model performance could benefit from additional variables that were not included.

Lastly, CNN was deployed to assess the disease's response to climate change. The predicted model was found to be sensitive to simulated changes in total precipitation and mean temperature. Results show positive and negative changes in the annual incidence rates for both emission scenarios, with a positive linear trend observed for mean temperature.

KEYWORDS

Dengue Fever; Incidence Forecast; Deep Learning, Machine Learning; Climate Change

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

A.	Aedes
AR5	Fifth Assessment Report
CNN	Convolutional Neural Network
CRISP-DM	Cross-Industry Standard Process for Data Mining
DENV	Dengue-Virus
DT	Decision Tree
EIP	Extrinsic incubation period
EMA	Exponential Moving Average
FNN	Feedforward Neural Network
GRU	Gated Recurrent Unit
IPCC	Intergovernmental Panel on Climate Change
IQR	Interquartile Range
KNN	K-Nearest Neighbor
LOCF	Last Observation Carried Forward
LSTM	Long Short-Term Memory
LSTM-ATT	Long Short-Term Memory with Attention
MAE	Mean Absolute Error
MCAR	Missing Completely at Random
MICE	Multivariate Imputation by Chained Equations
MLP	Multilayer Perceptron
NOAA	National Oceanic and Atmospheric Administration
NOCB	Next Observation Carried Backward
RCP	Representative Concentration Pathways
RF	Random Forest
RFE	Recursive Feature Elimination

RNN	Recurrent Neural Network
RMSE	Root Mean Square Error
SDG	Sustainable Development Goals
SMA	Simple Moving Average
STL	Seasonal and Trend Decomposition using Loess
SVR	Support Vector Regression
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

1. INTRODUCTION

Vector-borne diseases present a significant public health threat, with dengue fever being the fastest-spreading disease with a known sensitivity to climate (Nuraini et al., 2021). As of early December 2023, global reports indicated over 5 million dengue cases and more than 5,000 related deaths for the year (European Centre for Disease Prevention and Control, 2023). The prevalence of dengue transmission in tropical regions is significant, with Asia carrying the highest risk (Bhatt et al., 2013), representing around 70% of the global dengue burden (World Health Organization, 2024b). Recent research has increasingly highlighted the influence of climate change on dengue transmission in Southeast Asia (Kulkarni et al., 2022). This focus arose due to climate change driving the geographical expansion of the two species of Aedes mosquitoes, Aedes (A.) aegypti and A. albopictus, the primary dengue vector (Ebi & Nealon, 2016).

This research builds upon prior work that explored the climate-disease relationship in South and Southeast Asia. A unified framework has been established by combining research from multiple regions, environmental factors, and climate change with predictive modeling. A distinguishing factor is the self-obtained data, which provides a real-life setting for examining disease associations and developing the dengue fever incidence prediction. This research is relevant as developing a reliable forecast and analyzing disease dynamics enables the healthcare system to adapt to future health needs. Furthermore, the research provides valuable insights for research on other vector-borne diseases that exhibit correlations with changing weather conditions.

This research addresses the question: "How can predictive modeling using environmental variables assess dengue fever incidence and its response to climate change in Southeast Asia?". The primary research objective is to identify significant environmental factors influencing disease spread. This involves analyzing previous research and the historical climate-disease relationship. Subsequently, the research aims to forecast dengue incidence for multiple Southeast Asian locations using environmental data. The third objective is to assess the impact of changing climate on dengue fever in Southeast Asia. The changes in annual incidence rates are predicted based on simulated changes in mean temperature and total precipitation, considering two projected emission scenarios.

In Chapter 2, an extensive literature review assesses the existing research in the field and derives relevant implications shaping this study. Chapter 3 introduces the research methodology that aligns with the CRISP-DM framework, enhancing the project's structure and robustness. Historical secondary data on regional dengue incidence and environmental variables, precipitation and temperature, are used for data analysis and preparation. A predictive model for forecasting dengue fever incidence is developed using Python and supporting libraries. After evaluation, the validated model is deployed to the climate change simulations framework to assess the disease's response to climate change. Chapter 4 presents

and discusses the research findings, followed by conclusions derived in Chapter 5. Lastly, Chapter 6 outlines the study's limitations and provides recommendations for future work. The entire code and data used are available in Appendix B to ensure transparency and enable project replication.

2. LITERATURE REVIEW

The following chapter starts by explaining how dengue fever is transmitted and outlines key vector characteristics of the two Aedes species that directly influence disease transmission. It then discusses the impact of environmental variables, extreme weather events, and climate change on vector ecology and dengue fever incidence in Southeast Asia.

The literature search was conducted using two primary databases, Google Scholar and PubMed. Keyword and author-based searches were used, along with snowball sampling prioritizing recent research published since 2019. The search terms used included a combination of the keywords *Forecasting*, *Prediction*, *Dengue Fever, Climate Change, Southeast Asia, Data Mining*, and *Time Series Analysis*. The keywords were chosen for their direct relevance to the research question. Studies without full-text availability in English were not considered. Thirty-four papers were identified as relevant during the selection process, with 20 being included in the literature review, given their focus on dengue fever vectors and the disease's relation to environmental factors and climate change.

2.1. DENGUE TRANSMISSION AND VECTOR

Dengue fever is transmitted through the bite of an infected Aedes mosquito, primarily A. aegypti and A. albopictus (Wongkoon et al., 2013). These mosquitoes are vectors, transferring the virus from one infected individual to another through subsequent bites. Female mosquitoes play a crucial role in the transmission process, requiring a blood meal to develop their eggs (Ebi & Nealon, 2016). This mode of transmission, known as vector-borne, emphasizes the importance of managing mosquito vectors to prevent the spread of the disease.

Female mosquitoes lay their eggs on water-holding containers; after rain or flooding, eggs hatch into larvae, transform into pupae, and emerge as adult mosquitoes (Ebi & Nealon, 2016). Understanding mosquito development is crucial for vector control and prioritizing locations, given the observed relationship between larval indices and the incidence of dengue (Wongkoon et al., 2013). Consequently, the infection rates among female mosquitoes and the correlation of larval infection rates with subsequent rainy seasons emerge as relevant factors predicting dengue outbreaks (Kesorn et al., 2015).

Water-holding containers in and around houses support the mosquito lifecycle as they become breeding grounds (Ebi & Nealon, 2016). A study during the Philippines' rainy season revealed that both Aedes species use similar breeding sites, with artificial breeding sites such as plastic drums dominating natural ones (Edillo et al., 2012), indicating a preference for metropolitan areas. The importance of non-natural breeding sites is supported by research that examined larval habitats and distribution during Thailand's dry season. Although revealing some distributional distinctions between the species, drought generally does not

limit dense populations, with larvae residing in central areas utilizing breeding habitats created by humans (Chareonviriyaphap et al., 2003).

Ebi and Nealon (2016) reported that Aedes mosquitoes resist changing conditions. At higher temperatures, Aedes mosquitoes adapt by seeking shelter and exhibiting increased feeding activity. The resilience of mosquito eggs is a crucial factor, as they can endure drying and remain viable for months without water. Specifically, A. albopictus eggs can survive over winter and during long transportation, allowing expansion into new areas. In contrast, A. aegypti is characterized by a broader temperature tolerance, suitable for urban areas (Ebi & Nealon, 2016). Given their distinct characteristics, it is recommended to predict the distribution of A. aegypti and A. albopictus separately (Messina et al., 2015). To conclude, assessing dengue incidence by distinguishing between both Aedes species would enhance understanding disease dynamics.

Diverse approaches have been used to develop national dengue fever forecasts, with varying outcomes due to differences in modeling and regional distinctions. In Vietnam, for instance, dengue incidence varies across provinces, each characterized by unique geographical and climate conditions (Tran et al., 2022). Consequently, applying a single threshold for an epidemic outbreak at a national scale is insufficient, reinforcing the need for localized forecasting to enhance accuracy. Disregarding regional differences in statistical modeling may introduce bias, resulting in inaccurate relationships between disease occurrence and environmental predictors (Messina et al., 2015). These variations underline the necessity of identifying significant location-specific discriminators before establishing a global dengue model (Cheng et al., 2021).

Moreover, varying outcomes are evident within locations. This variability might be attributed to differences in model specifications, study period, and variables considered. For instance, some research considers nationally reported dengue incidence (Tran et al., 2022), while others leverage mosquito characteristics, predicting vector density to assess disease risk (Bonnin et al., 2022), making generalizations difficult. In addition, reporting bias could stem from broad disease definitions, misdiagnosis, and failure of disease identification (Bhatt et al., 2013). Lastly, distinguishing locally acquired and travel-imported cases could enhance accuracy in disease capture (Cheng et al., 2021).

The nature of the disease spread has been extensively analyzed. Accordingly, elevated dengue risk is associated with urbanization and climate change-induced rise in temperatures and precipitation (Ebi & Nealon, 2016). With a growing human population at risk, the future distribution of Aedes mosquitoes will continuously expand in climatic suitable urban areas (Kraemer et al., 2019). Changing climate affects environmental systems, encompassing water storage, land use, and irrigation, as well as human factors, driving migration and unplanned urbanization, increasing human density and suitability for larval habitats (Ebi & Nealon, 2016). An increased dengue risk has been identified in connected low-income urban areas where population movement favors dengue spread (Bhatt et al., 2013). Populations in economically

disadvantaged areas are exposed to significant and disproportionate dengue risk (Messina et al., 2019). In conclusion, climate change and the growing global interconnectivity intensify the challenge of containing disease vectors.

2.2. ENVIRONMENTAL VARIABLES

Temperature significantly impacts the lifecycle of Aedes mosquitoes, their density, biting activity, distribution, and flying distance, with increasing temperatures accelerating reproduction and reducing the virus's extrinsic incubation period (EIP), supporting disease transmission (Li et al., 2018). At ambient temperatures between 25-28°C, the EIP lasts approximately 8-12 days after an infected person has been bitten (World Health Organization, 2024b). The total dengue infection cycle lasts 4-7 weeks (Seah et al., 2021), considering an intrinsic incubation period of 4-10 days after infection until the symptoms appear, which last up to 7 days (World Health Organization, 2024b). Research exhibited a mid-year peak of disease occurrence, with a similar trend observed for temperature (Seah et al., 2021). The World Health Organization (WHO) underlines the importance of seasonality in the transmission process, with excessive occurrence during and after the rainy season (World Health Organization, 2024a).

The rainy season in Southeast Asia occurs from May to October, with seasonality impacting temperatures ranging from 24-28°C between April and June and from 16°C-24°C between December and February (Bonnin et al., 2022), indicating associations between environmental factors. For instance, humidity is related to precipitation and temperature, which influence incidence (Li et al., 2018), with increasing temperatures inducing a rise in humidity (Ebi & Nealon, 2016). Research demonstrated an insignificant negative association between dengue risk and relative humidity (Wang et al., 2022). A seasonal analysis in Thailand reinforces this connection, emphasizing the inverse relationship between relative humidity and dengue risk with an increase in rainy days (Wongkoon et al., 2013). In contrast, additional evidence indicates a statistically significant, positive association between absolute humidity and dengue infections (Seah et al., 2021). Specific conditions in the respective studies might differ, leading to contradictions and non-comparability. Previous research analyzed multiple environmental factors, including temperature, precipitation, humidity, wind velocity, sunshine, and air pressure, with contrary results. Given insufficiencies in reliable data (Davis et al., 2021; Wang et al., 2023), variables received varying attention.

The relationship between dengue and precipitation is ambiguous. Precipitation and evaporation influence water sources, affecting mosquito populations and biting behavior (Ebi & Nealon, 2016). In the absence of precipitation, the dry environment limits breeding sites for mosquito eggs, reducing their population (Wang et al., 2022). Increasing precipitation positively correlates with dengue cases, creating suitable conditions through the availability of more breeding sites (Li et al., 2018). The effect of extreme precipitation, so-called flushing, has been analyzed in Singapore. As a result, flushing significantly decreases the risk of dengue outbreaks up to 6 weeks after the excessive rainfall as breeding sites are flooded, disrupting

the breeding process (Benedum et al., 2018). Reviewed literature showcased the disease's sensitivity to climate variability and the complexity of concluding disease associations.

Small environmental changes affect disease transmission, while rising temperatures expand the vector habitat but constrain virus vitality, mosquito reproduction, and development in already warm regions (Li et al., 2018). Seah et al. (2021) examined the impact of maximum ambient temperature and heatwaves in Singapore. Accordingly, the relative risk of dengue infections linearly increases with rising maximum temperatures but exhibits a non-linear decrease beyond a threshold of 31 °C. Higher temperatures may enhance transmission due to increased biting rates and a shorter virus EIP. However, prolonged high temperatures reduce mosquito lifespan and egg-to-adult survival, leading to a decline in the Aedes population and lower dengue transmission risk over time. Similar results were reported for heatwaves. Heatwaves commonly last 2-3 days, with 1-2 heat days per week having no effect, while 3-6 heatwave days lead to a decrease in reported dengue infections. During extreme heat, people may spend more time indoors, reducing vector-host contact and lowering transmission risk. Contrary to other research, their study identified that the impact of cumulative precipitation on dengue incidence is constrained due to the significance of non-natural breeding sites (Seah et al., 2021). To summarize, temperature has a direct and indirect influence on disease transmission.

Findings emphasize the need for a time series analysis with lagged observations to understand delayed disease response. Wang et al. (2022) explored short-term associations between extreme weather events and dengue fever infection risk in South and Southeast Asia. Extremely low temperatures reduced transmission risk with a 1-3 week lag but not for extended lag observations. Extremely high temperatures were associated with an increased infection risk with a lag of 2-3 weeks, depending on the number of extremely hot days in a week. Extreme rainfall was associated with a significant decrease in infection risk with a lag of 0-4 weeks, with the lowest risk at a lag of 2 weeks and an increasing risk denoted at a lag of 7 weeks. Their research concluded the importance of understanding the short-term effects of climate variability on dengue transmission (Wang et al., 2022). In contrast, long-term forecasts were associated with inaccurate predictions due to decreased accuracy (Nuraini et al., 2021), denoting the difference in the significance and reliability of short- and long-term analyses.

2.3. Climate Change and Extreme Weather Events

The climate-disease relationship has been thoroughly researched, particularly in Southeast Asia, with recommendations to further assess the impacts of climate change (Kulkarni et al., 2022). Having examined the geographical expansion of dengue associated with climate change, daily mean temperature and temperature variations stand out as the main factors influencing dengue distribution, followed by precipitation (Ebi & Nealon, 2016). Understanding the significance of the changing climate on dengue fever is relevant for assessing the disease's risk and adapting mitigation strategies (Davis et al., 2021). A relevant impulse for this research was established by projecting dengue epidemics in tropical areas in

South and Southeast Asia under different emission scenarios. Results demonstrated that transmission risk is expected to increase for all scenarios, showing a peak in dengue fever epidemic size and outbreak duration (Wang et al., 2023). Most research incorporating climate change refers to the Intergovernmental Panel on Climate Change (IPCC). Its Fifth Assessment Report (AR5) projects climate change, impact, and risk until 2100 based on four Representative Concentration Pathways (RCP) scenarios. Each scenario describes different level of greenhouse gas emissions and radiative forcing, with a low-emission scenario (RCP2.6), two intermediate scenarios (RCP4.5 and RCP6.0), and a high emission scenario (RCP8.5) (IPCC, 2014).

Climate change refers to an alteration of extreme weather events, such as El Niño, typhoons, floods, droughts, and heatwaves, in frequency and magnitude and a long-term change in environmental variables influencing the survival, replication, development, and distribution of the vector and virus (Li et al., 2018). El Niño refers to the periodic warming of sea surface temperatures, affecting global weather patterns (Sundari & Krishnamoorthy, 2019). An interaction of environmental variations and long-term climate change might foster transmission (Ebi & Nealon, 2016). Extreme weather events are predicted to become more frequent (Wang et al., 2022), affecting Aedes mosquitoes and ultimately promoting disease transmission (Li et al., 2018). However, extreme weather events do not exhibit a uniform definition. For instance, Seah et al. (2021) tested different heatwave definitions by adjusting the percentiles and number of consecutive heatwave days.

Cheng et al. (2021) explored the relationship between dengue outbreaks and climate change-induced extreme weather events in China, observing a positive association and delayed effects. During or immediately after extreme weather events, the risk of dengue outbreaks did not homogenously increase, with decreases in some instances. A peak in dengue outbreak risk was observed approximately 1.5 months post-heatwaves and 1.5-3 months after extreme precipitation and humidity, with effects lasting 2-3 months. The temporal delay showcases the duration of negative and positive effects of environmental conditions on dengue transmission. In conclusion, the importance of further exploring extreme weather events in dengue-prone Asia-Pacific countries and their outbreak scale was emphasized (Cheng et al., 2021).

Bonnin et al. (2022) integrated future climate scenarios into developing a process-based dengue model for Southeast Asia. Their study examined seasonal correlations, revealing that a future increase in temperature is the main driver for increasing seasonal densities of Aedes mosquitoes. The optimal temperature for A. aegypti was identified as 33°C and for A. albopictus as 29°C. Beyond these temperatures, rising temperatures led to lower densities of adult female mosquitoes. Precipitation exhibits lower significance, yet its increase correlates with higher adult female densities for both mosquito species. Despite some area-specific differences, both densities are projected to increase in all future climate scenarios. The study concludes that even with a rigorous implementation of greenhouse gas emissions reduction

strategies, mosquito densities will not decline (Bonnin et al., 2022). However, other research argues that reduced greenhouse gas emissions would decrease the suitability of vector habitats (Kraemer et al., 2019), with distribution suitability remaining consistent or slightly decreasing (Davis et al., 2021). The validity of these contradictory statements is addressed in the subsequent analysis.

Messina et al. (2019) leveraged IPCC projections for 2020, 2050, and 2080 to assess dengue risk based on 2015 using a boosted regression tree. Accordingly, temperature and annual cumulative precipitation were determined as the primary factors influencing dengue suitability. Projections until 2080 indicate variable shifts in dengue suitability's geographical and temporal distribution, with minimal global changes but significant subnational variations. Asian cities in coastal eastern China and Japan are expected to become more suitable by 2050. The model outcome for the risk assessment depended on the integrated climate scenario. The scenario RCP6.0 suggested a global increase in dengue risk, potentially affecting an additional 2.25 billion people between 2015 and 2080, encompassing over 60% of the world's population. This increase would result from population growth in already endemic areas instead of the suggested spread of dengue in new populations. Contrarily, scenario RCP4.5 indicated a potential reduction in dengue risk between 2050 and 2080 (Messina et al., 2019). Climate change scenarios are leveraged in the subsequent analysis to evaluate disease dynamics in Southeast Asia further.

2.4. SUMMARY

The literature review provides a structured exploration, highlighting fundamental studies, disease characteristics, and contradictions. By addressing the complexity of dengue transmission, vector ecology, and the impact of environmental variables, seasonality, and climate change, the aim was to establish a profound understanding and address knowledge gaps related to the research question.

Due to the elaborated relevance of disease characteristics and associations, qualitative reported differences, and non-comparable study settings of different locations, study periods, and variables, the literature review did not focus on comparing performance metrics. The reviewed literature built the foundation for the geographical focus, data collection, and variable selection, which is crucial for model development.

Defining the research scope proved challenging given the variety of spatial coverage ranging from urban settings to cross-border regions. The geographical focus of this study has been determined through a scoping review that emphasized the impact of climate change on dengue incidence in Southeast Asia (Kulkarni et al., 2022). The framework of Li et al. (2018) was an inspiration to follow a multidimensional approach by integrating climate change to drive the meaningfulness of the underlying research. Given the repeated emphasis on the significance of temperature and precipitation on disease transmission (Bonnin et al., 2022; Ebi

& Nealon, 2016; Messina et al., 2019), these factors are used as predictors for the time series analysis.

3. METHODOLOGY

3.1. INTRODUCTION

The research methodology aligns with the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework established by Chapman (2000). The schema follows a structured sequence that forms a cycle of six phases: (1) *Business Understanding*, (2) *Data Understanding*, (3) *Data Preparation*, (4) *Modeling*, (5) *Evaluation*, and (6) *Deployment* (Chapman et al., 2000). The cyclical process is illustrated in Figure 1.

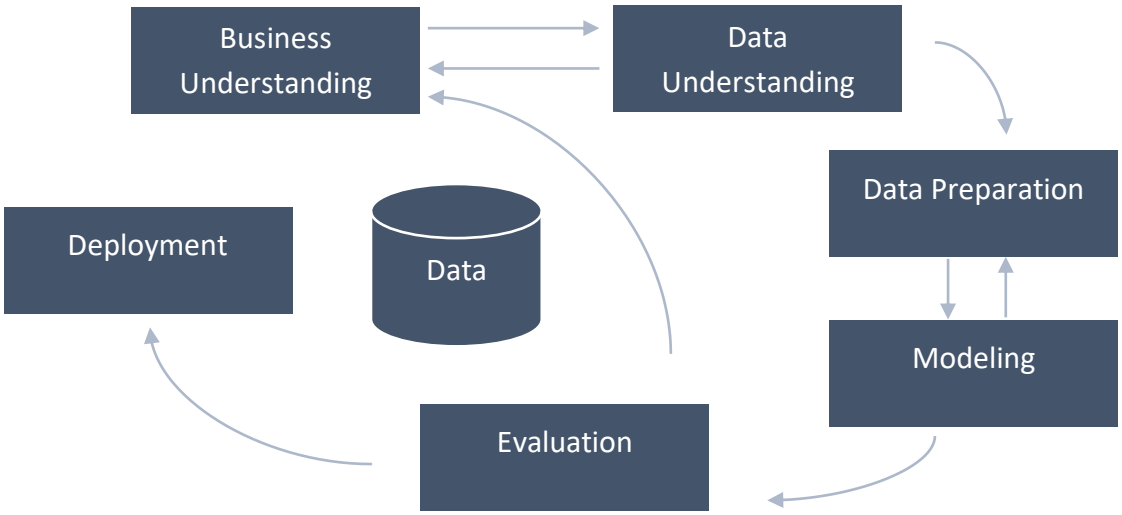


Figure 1 - CRISP-DM methodology

In the following, *Business Understanding* will be referred to as *Domain Understanding* to better align with the research framework. In this phase, the current state of research was analyzed to determine the research need, direction, and scope of this study. Subsequently, the project was initialized by defining the research question and objectives.

Within the *Data Understanding*, raw data was analyzed to verify suitable format and quality. Following the data extraction, a profound exploratory analysis was performed to identify data limitations and inconsistencies shaping the *Data Preparation*.

The *Data Preparation* phase includes data preprocessing, feature engineering, and selection. The aim was to clean the data and address data quality issues observed during *Data Understanding*. Data from the different locations was standardized, aggregated, and combined in a format suitable for *Modeling*.

The *Modeling* encompasses repetitive model training and evaluation. Models were developed, hyperparameter tuning was conducted, and various modeling and validation approaches were tested. In addition, the performance of applied data preparation techniques was assessed to determine the best performance.

Within the *Evaluation,* the research process was reviewed to prevent irregularities, and the success of the domain objectives and data mining goals were confirmed before *Deployment*.

The *Deployment* phase began with additional research that provided suitable climate change simulations for Southeast Asia. Finally, the forecast was deployed into the selected climate change framework. The project has been examined in detail to ensure seamless monitoring and maintenance in the future.

Various methods have been applied within these stages, presented in Figure 2 and further explained in the subsequent chapters.

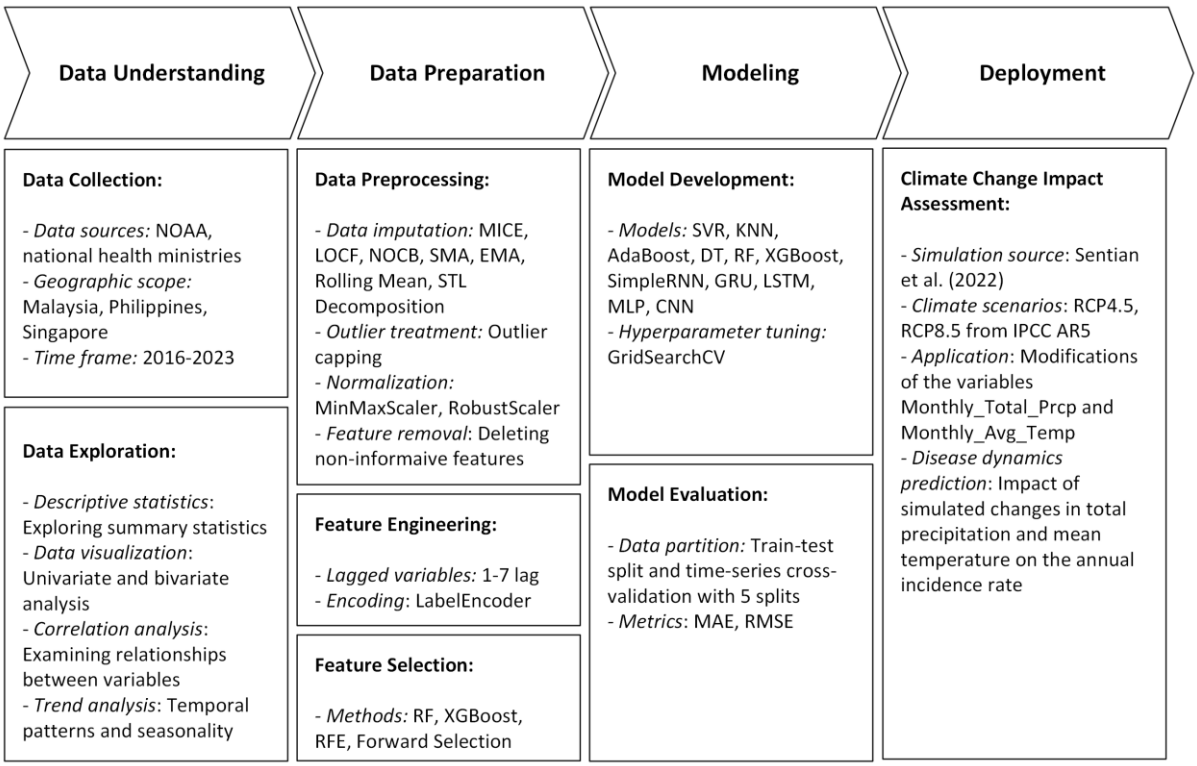


Figure 2 - Detailed overview of applied methods

3.2. DOMAIN UNDERSTANDING

Dengue fever represents a significant health burden in Southeast Asia, with environmental factors such as temperature and precipitation favoring its spread. Understanding disease associations and predicting transmission is crucial for developing prevention strategies. Building upon the research background and need discussed in the literature review, this project aimed to combine open-source data from locations within Southeast Asia with data science tools to enhance the current state of research. The project used Python programming language and supporting libraries, whose respective versions are linked in Appendix A.

This research addressed the question: "How can predictive modeling using environmental variables assess dengue fever incidence and its response to climate change in Southeast Asia?". Three domain objectives were formulated (Figure 3). The first objective was to identify

significant environmental drivers of disease transmission. The second was to develop the predictive model for dengue incidence using temperature and precipitation variables. Finally, the third objective was to assess how the disease responds to climate change.

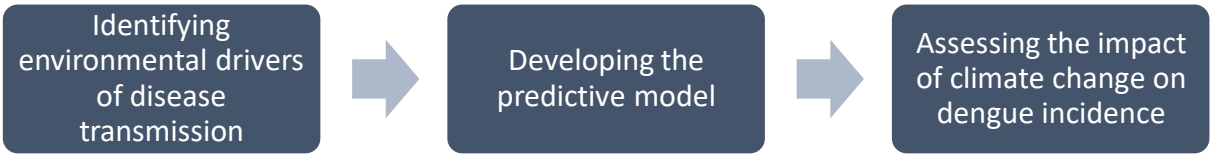


Figure 3 - Domain objectives

From a technical perspective, the data mining goals were defined (Figure 4). First, collect data from multiple Southeast Asian locations. Second, analyze data as part of a profound exploratory analysis. Third, prepare and transform raw environmental and incidence data to enhance data quality and interpretability. Fourth, develop and optimize selected machine and deep learning models in the multivariate time series analysis to create a highly accurate dengue incidence forecast that generalizes well on unseen data. Lastly, evaluate model performance to ensure deployment readiness.

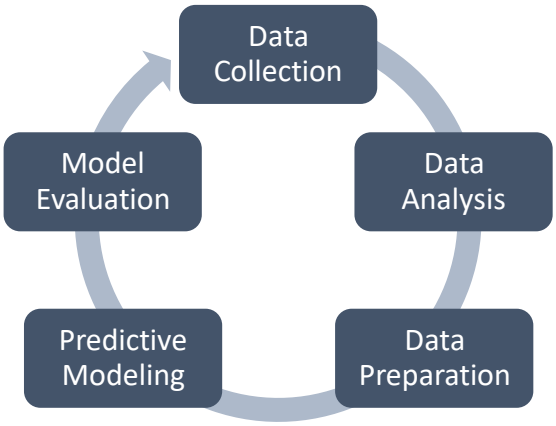


Figure 4 - Data mining goals

3.3. DATA UNDERSTANDING

The geographical scope within Southeast Asia was determined by data availability and sufficiency. Key insights for data search were gathered through the literature review. Environmental data was explored on the websites of national meteorological, environmental, and climatological departments. Dengue fever incidence data was collected from national health ministries, disease control divisions, and WHO country offices. If data was not publicly available, it was requested individually. Unfortunately, digital requests for nine other Southeast Asian locations did not receive a positive response.

To establish a foundation for the model development, cohesive data points on a regional level in daily, weekly, or monthly intervals were sought for environmental variables and dengue fever incidence. Yearly data was not considered, as the goal was to capture patterns and

seasonal trends over time to understand disease associations and effectively address the research question.

Chapter 2 discusses key environmental factors influencing disease spread, revealing variations in the choice of variables. Some studies examined dengue correlations with environmental variables, while others incorporated extreme weather events, considering their excessive occurrence as a potential indicator of climate change. Furthermore, as highlighted in the previous chapter, analyses yielded incohesive results. While certain environmental variables emerged influential in some instances, others failed to confirm such assertions, adding complexity to the variable selection. Given their demonstrated relevance in reviewed literature, temperature and precipitation have been chosen for the model establishment.

Weather and climate extremes are not directly integrated through rules of definitions. Extremes can be identified through patterns and characteristics such as the duration and intensity learned during the training phase. Dengue fever was not differentiated from more severe forms of the virus infection, including dengue hemorrhagic fever and dengue shock syndrome. Data points for these severe forms were considered dengue fever and summed up. Additionally, serotypes of dengue virus (DENV) DENV 1, DENV 2, DENV 3, and DENV 4 were not distinguished. The secondary data was acquired from multiple sources at quality and accuracy that cannot be guaranteed or verified. The data exploration revealed some inconsistencies, partially visualized in subsequent phases, facilitating the Python libraries Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021). The entire exploratory analysis, along with the corresponding Jupyter notebooks, is provided in Appendix B.

In conclusion, comprehensive data points were established for 17 locations within Malaysia, the Philippines, and Singapore. Sequential weekly dengue incidence was acquired from national departments. Daily environmental data, including precipitation, maximum temperature, minimum temperature, and average temperature, was obtained from the National Oceanic and Atmospheric Administration (NOAA) for each country, enhancing data consistency (Table 1).

Table 1 - Data collection sources

Country	Environmental variables		Dengue incidence		Time frame
	Data source	Missing values	Data source	Missing values	
Malaysia	NOAA	Yes	Ministry of Health, Malaysia	Yes	2017-2022
Philippines	NOAA	Yes	Department of Health-Epidemiology Bureau, Philippines	No	2016-2020
Singapore	NOAA	Yes	Ministry of Health, Singapore	No	2016-2023

3.4. DATA PREPARATION

3.4.1. Data Preprocessing

At the beginning of the *Data Preparation* phase, the data was transformed into a format suitable for the modeling. For each location, the Excel file was standardized as follows:

- The first sheet, df_1, contained daily environmental data.
- The second sheet, df_2, contained weekly incidence data.
- The third sheet, df_3, combined and aggregated the weekly incidence and daily environmental data to a standard monthly scale to enhance interpretability.

Consequently, the data quality and preparation of the first two datasets directly affected the third dataset. Lastly, the monthly environmental data was transformed into ten meaningful variables. Table 2 displays the final format of the third dataset, which variables were used for the modeling.

Table 2 - Transformed environmental variables

Variable	Unit	Description
Min_Daily_Prcp	Millimeters (mm)	Minimum daily precipitation
Max_Daily_Prcp	Millimeters (mm)	Maximum daily precipitation
Monthly_Avg_Prcp	Millimeters (mm)	Monthly average precipitation based on daily temperature reporting
Monthly_Total_Prcp	Millimeters (mm)	Monthly total precipitation based on cumulated daily precipitation
Monthly_Avg_Temp	Degrees Celsius (°C)	Monthly average temperature based on daily temperature reporting
Min_Daily_Temp	Degrees Celsius (°C)	Minimum daily precipitation of the respective month
Max_Daily_Temp	Degrees Celsius (°C)	Maximum daily precipitation of the respective month
Min_Average_Temp	Degrees Celsius (°C)	Minimum value of the average daily temperature
Max_Average_Temp	Degrees Celsius (°C)	Maximum value of the average daily temperature
N_Raining_Days	Degrees Celsius (°C)	Cumulative number of rainy days in the respective month

The initial preprocessing focused on the first two datasets holding raw data using Pandas (The pandas development team, 2024) and NumPy (Harris et al., 2020) libraries. The main goal of dealing with time series data was to maintain cohesive data points considering the temporal order of observations while capturing seasonality and trends to explain disease interactions.

At first, missing values were imputed to ensure consistent data. The type of missing values was classified as Missing Completely at Random (MCAR) based on the absence of distinct patterns at specific locations, times, or variables. The probability of missingness was assumed to be the same for all samples. Each variable exhibited distinct patterns of missing values, with some showing continuous gaps. Consequently, multiple methods were explored and tested individually for all variables. These methods included Last Observation Carried Forward (LOCF) and Next Observation Carried Backward (NOCB), Simple Moving Average (SMA), Exponential Moving Average (EMA), Rolling Mean Statistics, with Seasonal and Trend Decomposition using Loess (STL) Decomposition, and Multivariate Imputation by Chained Equations (MICE) Imputation using IterativeImputer from Scikit-learn (Pedregosa et al., 2011).

MICE yielded the best results for imputing continuous gaps in the temperature variables, incorporating all three variables to predict missing values simultaneously (Figure 5). Retaining the relationships among variables reduced bias in processing large amounts of missing values. However, integrating the variable *PRCP_mm*, daily precipitation in millimeters, into this selection introduced noise with negative values in some instances, suggesting MICE's suitability for homogenous input variables (Figure 6). The imputation process is illustrated using Singapore as an example.

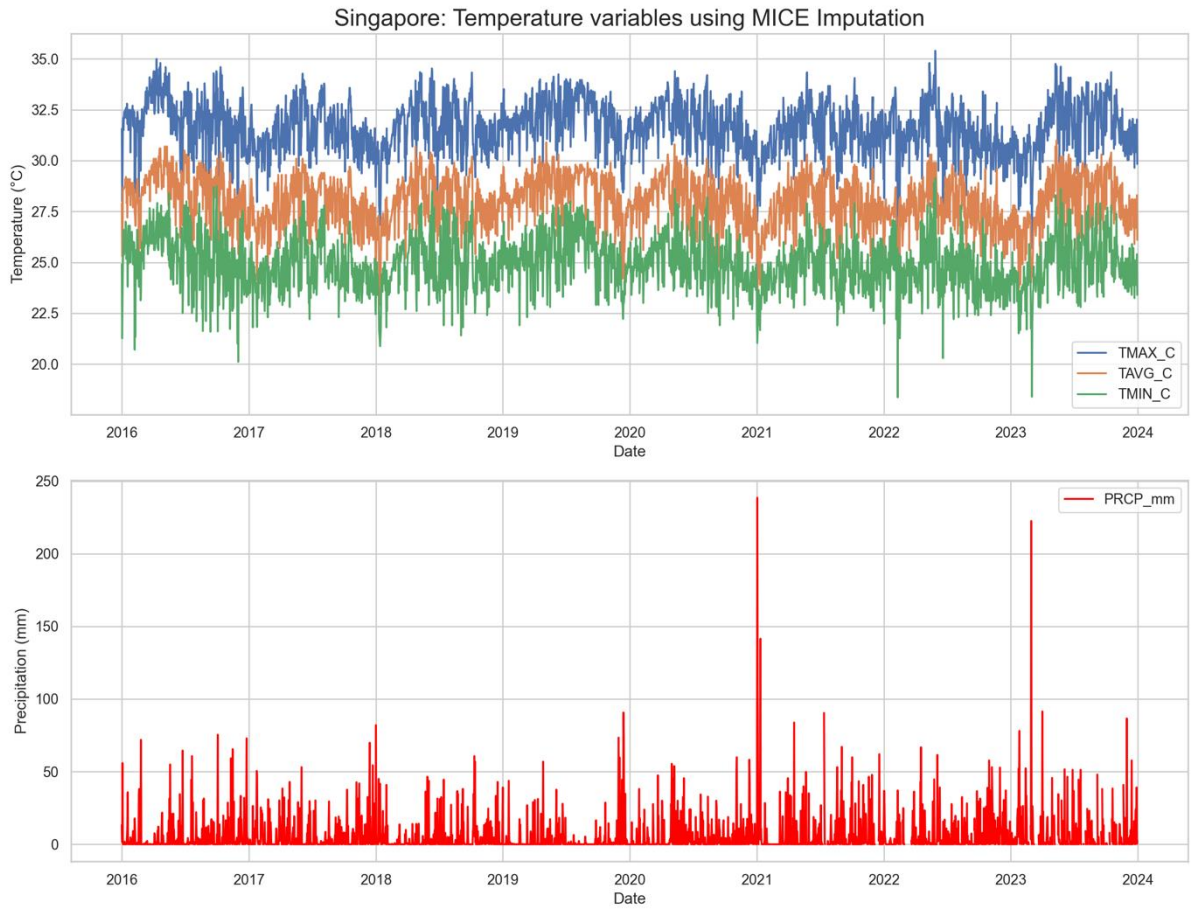


Figure 5 - Singapore: MICE imputation for the temperature variables

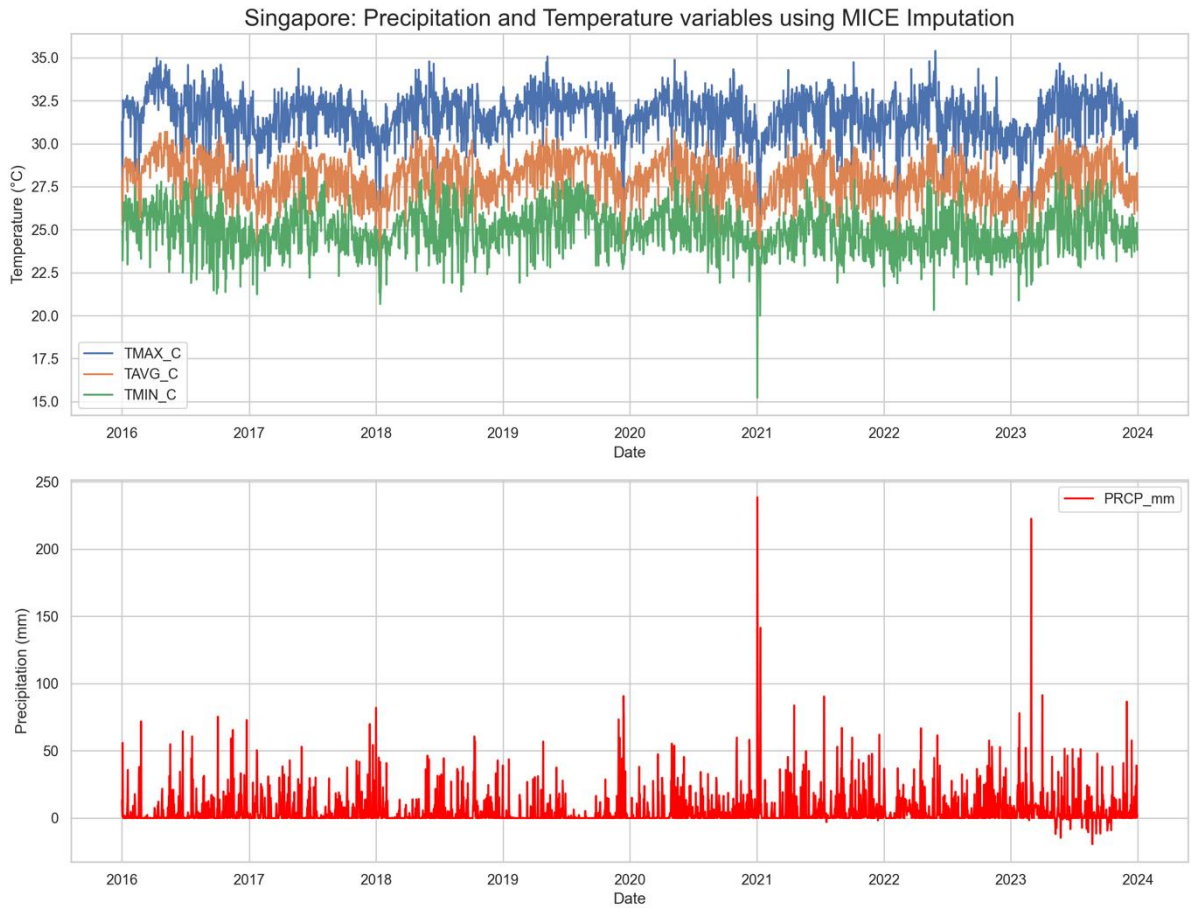


Figure 6 - Singapore: MICE imputation for the precipitation and temperature variables

LOCF showed stable results for *PRCP_mm,* preserving the data structure and statistical properties after imputation. Missing values of the variable *Incidence* were addressed with an iterative Rolling Mean Imputation, which computed the mean of a fixed window of daily observations and iteratively filled missing values until convergence, considering local trends. Experimentation with different parameter values revealed that a window size of 7 and a minimum of 4 non-missing values within the window performed best.

Missing values in the variable *Incidence* were solely found within Malaysian datasets. However, the data exploration revealed some data validity concerns. Figure 7 depicts the raw weekly incidence data in Manila, Philippines, suggesting that potential missing entries in 2016 could have been imputed with a placeholder value of zero. This finding underlines the necessity of extensive data preparation, with monthly aggregation and outlier treatment addressing the observed inconsistency.

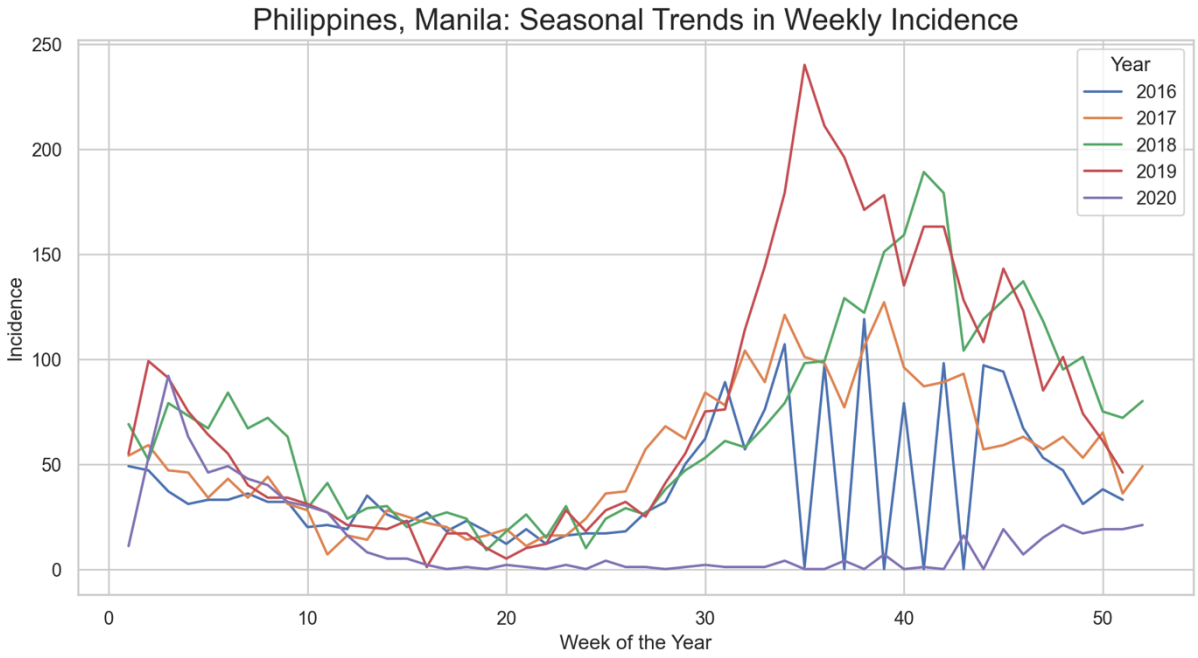


Figure 7 - Manila, Philippines: Potential missing values

Having imputed missing values for each location, outlier capping was implemented within location-specific datasets and on the third dataset comprising all locations. The primary goal was to reduce noise caused by outliers while preserving environmental dynamics, as extremes in weather affect mosquito habitat and disease transmission. Applied approaches aimed to mitigate bias and information loss, enabling data integrity and interpretation. Consequently, outlier removal techniques were disregarded to maintain sequential data and data sufficiency.

Outliers influence statistical distributions, directly affecting the modeling. Capping outliers across all locations prioritized broader fluctuations but might not have addressed relative extremes. Conversely, capping outliers within each dataset addressed location-specific inconsistencies but could have reduced informational significance due to different distributional ranges across the locations. The significance of these contrasting methods is examined in Chapter 4. Additionally, a combination of both approaches was performed.

Different thresholds for the standard deviation and z-score methods were tested as part of the outlier detection, with a threshold of 3 to 4 proved optimal. Results were aligned with the choice of quantiles of the Interquartile Range (IQR) method that was utilized for the outlier removal. Experimentation revealed that a 25/75 split resulted in severe outlier capping, while a 20/80 and 15/85 split provided narrower bounds, still encompassing significant outliers. Balancing sensitivity to extreme values with robust outlier treatment was crucial. In conclusion, the 15/85 split was chosen to maintain analytical integrity while tolerating the complexities of environmental data.

The variable *Min_Daily_Prcp* was excluded from the analysis due to its limited significance, with 0 mm in most cases. The variables *Monthly_Average_Prcp, Min_Average_Temp,* and

Max_Average_Temp were removed to avoid multicollinearity due to high correlation, which added redundancy (Appendix C).

Feature scaling was performed using Scikit-learn (Pedregosa et al., 2011) to ensure a proportional impact of all variables in distance-based models and neural networks. MinMaxScaler was implemented to normalize the features to a range between 0 and 1. RobustScaler, which uses the interquartile range (IQR) for normalization, was employed to enhance robustness against outliers. The performance of both methods is evaluated in Chapter 4. In addition, statsmodels library (Seabold & Perktold, 2010) was used to conduct an Augmented Dickey-Fuller test to assess stationarity. The test results indicated that the features in df_3 were stationary, eliminating the need to further transform the time series for modeling.

3.4.2. Feature Engineering

In feature engineering, lagged variables were created for the independent variables spanning 1-7 months. 6-month lag intervals were chosen to capture the total duration of the dengue infection cycle of up to 7 weeks (Seah et al., 2021) and delayed direct and indirect environmental influences on the mosquito lifecycle and habitat. The incidence rate was calculated per 100,000 population to standardize locations and capture changes over time by adjusting population sizes (Equation 1). This decision was based on the explored variability of incidence across years and locations. Using the example of Dagupan, Philippines, and Singapore, the varying seasonal trends of weekly incidence are demonstrated in Appendix D. Consequently, the variable *Incidence Rate* was defined as the target variable.

$$(1)\ Incidence\ Rate = \frac{Incidence}{Population} \times 100{,}000$$

Lastly, the categorical variable *Name* was transformed into *Location Code* in a numerical format using LabelEncoder from Scikit-learn (Pedregosa et al., 2011) to ensure compatibility within the modeling.

3.4.3. Feature Selection

Four feature selection techniques were explored to define significant environmental features that interact with disease transmission utilizing Scikit-learn (Pedregosa et al., 2011). Feature importance analysis was conducted using Random Forest (RF) and Extreme Gradient Boosting (XGBoost) algorithms for selecting variables based on their importance weights, yielding varying results in terms of the number and impact of selected features (Appendix E).

In addition, two wrapper methods, Recursive Feature Elimination (RFE) and Forward Feature Selection, were tested with RF Regressor. These wrapper methods provide flexibility by allowing manual specification of the number of features. The methods were tested with 5, 10, and 15 input features, with 10 features yielding the best score in the modeling. The outcomes

of the four applied feature selection methods are presented in Appendix F, which shows the total number of times each feature was selected across all methods.

3.5. MODELING

3.5.1. Model Development

This project used predictive modeling to forecast dengue incidence based on environmental variables using supervised learning for regression. The complexity of this multivariate time series analysis lies in processing data from multiple Southeast Asia locations, each with different time ranges.

The initial approach was to train a holistic model covering all locations that predicts dengue incidence based on date, environmental variables, and the differentiator *Location Code*. The assumption was based on dependency, as the dengue incidence of each location was predicted using the same input features. Therefore, it was assumed that the model may benefit from past values of the other locations to learn inherent disease characteristics and seasonal patterns. The initially implemented time-based split for each location raised the concern of data leakage due to non-aligned time ranges within the training process (Figure 8).

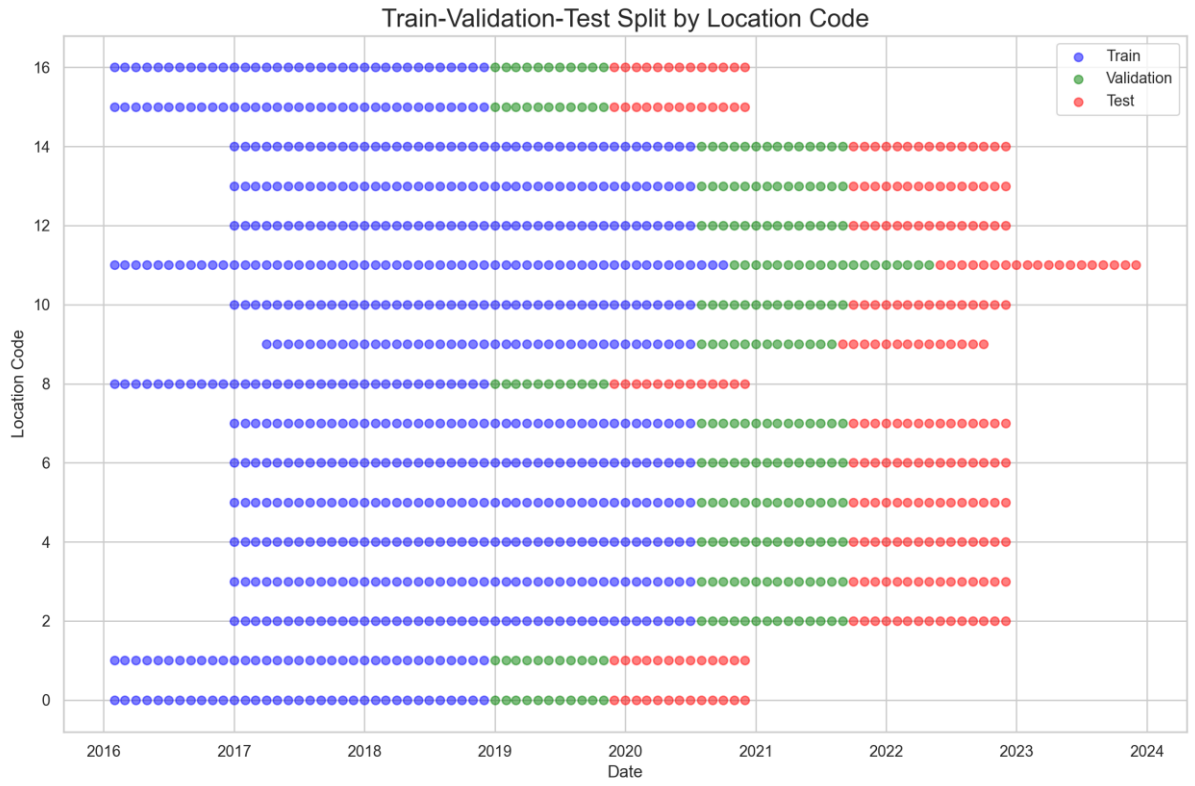


Figure 8 - Holistic model: Non-aligned time-based split

Consequently, a split was performed at the timestamps 2019-01-01 and 2020-01-01 to differentiate training, validation, and test sets (Figure 9). The test set was saved for assessing the disease's response to climate change in the final phase, requiring 12 monthly instances.

Having addressed potential data leakage decreased the training set substantially. Another concern was that the model's ability to learn location-specific characteristics was restricted due to the multitude of locations. Moreover, locations had a different number of observations in the training set, which led to an unbalanced representation. As the model predicted on labeled data, this may induce issues in supervised learning.

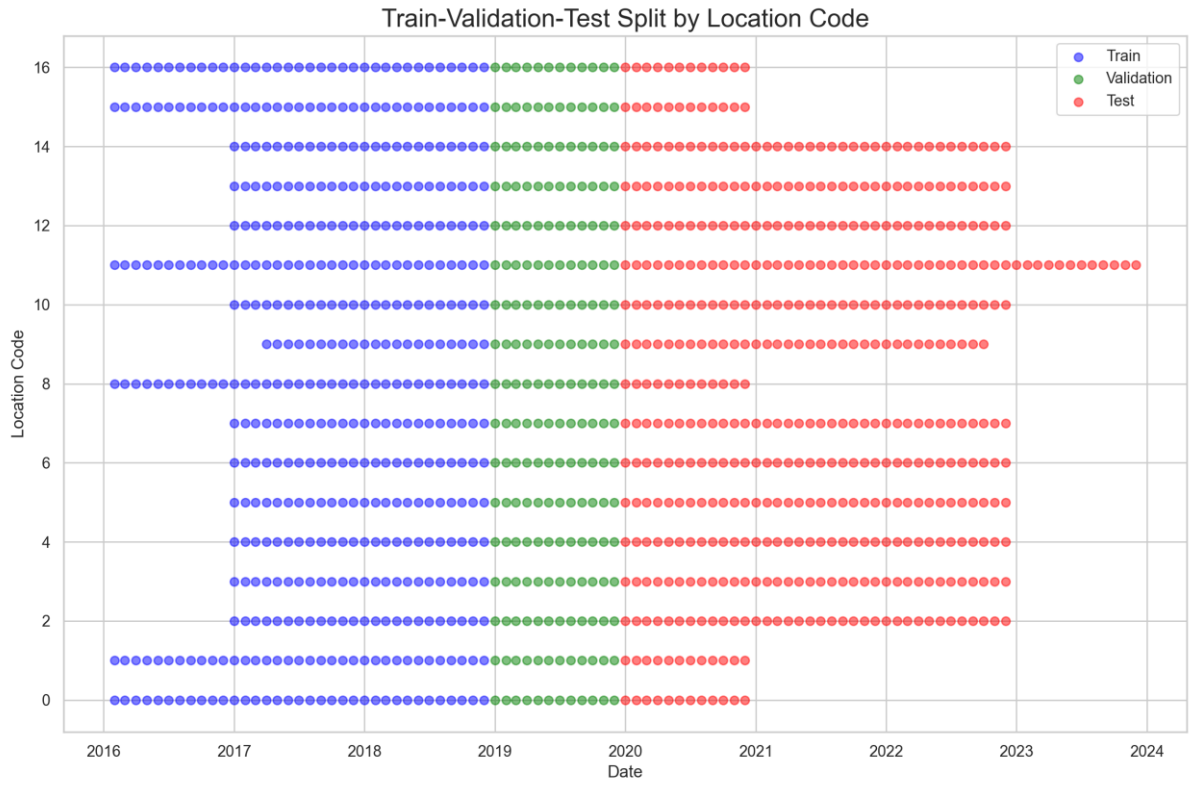


Figure 9 – Holistic model: Aligned time-based split

Consequently, an independent approach was added with multiple location-specific time series. Here, the number of observations in the training set was comparably less, yet the model was trained with unique location-specific characteristics. A train-test split was performed with the test set comprising 12 data points for each location relevant to the deployment (Figure 10).

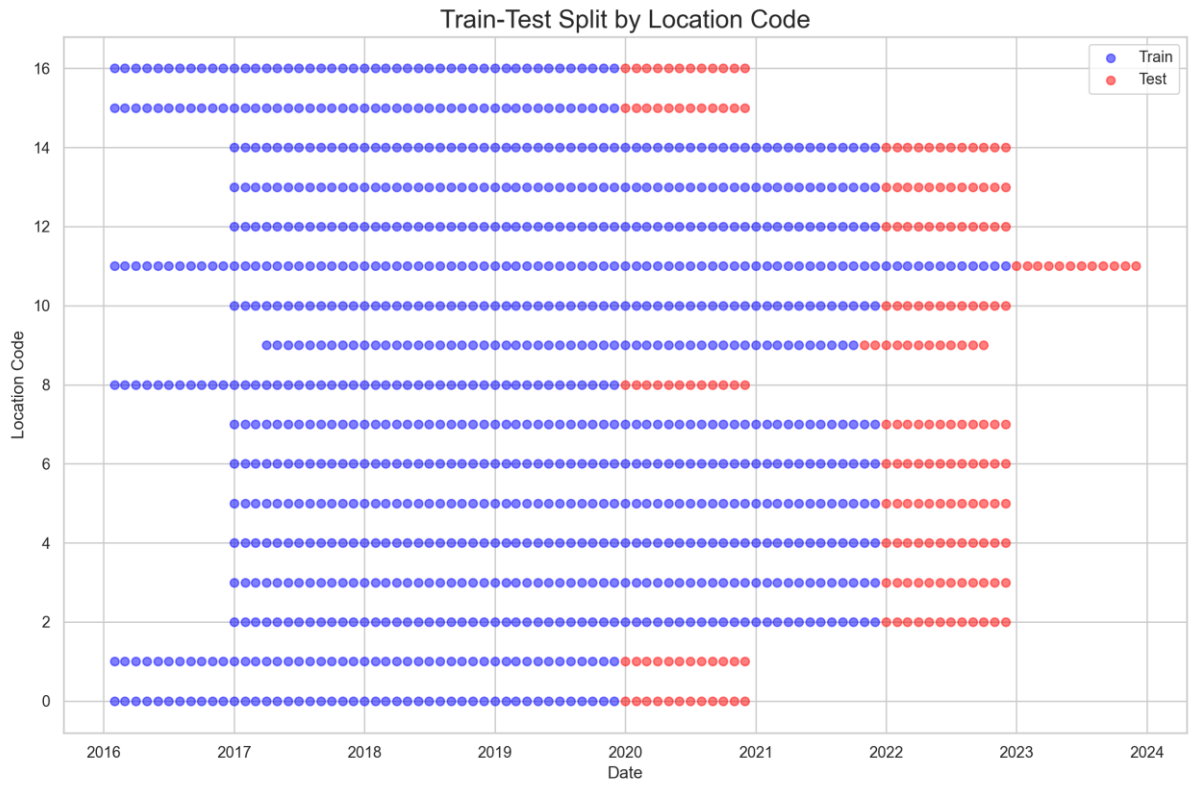


Figure 10 - Location-specific time-based split

The training phase included a time series cross-validation using TimeSeriesSplit from the Scikit-learn library (Pedregosa et al., 2011). Here, the model was fitted to the training data and evaluated on the validation data in 5 splits, using an expanding training window to ensure robust generalization on unseen data. The split is displayed in Appendix G using the example of Singapore, *Location Code* 11. The encoded variable *Location Code* was dropped to mitigate the bias of ordinal ranking.

Forecasting models were developed with the aim of processing time series regression data with a non-linear relationship between target and independent variables. Time series data was transformed into a supervised learning problem through appropriate data preparation and splitting, making it suitable for traditional machine learning models. Support Vector Regression (SVR), K-Nearest Neighbor (KNN), AdaBoost Regressor, Decision Tree (DT), and its ensemble, RF, were implemented using Scikit-learn (Pedregosa et al., 2011). XGBoost was integrated using the XGBoost library (Chen & Guestrin, 2016). These models were initially employed with default parameters to assess different capping approaches, feature selection techniques, and scaling methods based on the validation set. Subsequently, hyperparameter tuning was conducted using GridSearchCV from the Scikit-learn library (Pedregosa et al., 2011) to optimize model performance by adjusting parameters while preventing overfitting. Tuned parameters are presented in Appendix H.

The initial research employing deep learning models using climate data to forecast long- and short-term dengue incidence and outbreaks was conducted by Tran et al. (2022) in Vietnam. Their time series analysis revealed that Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Long Short-Term Memory with Attention (LSTM-ATT) were the best-performing models (Tran et al., 2022), informing the decision to incorporate deep learning models into the underlying analysis to ensure alignment with the latest research on dengue fever prediction. Deep Learning models based on neural networks were implemented using Keras API with TensorFlow (Abadi et al., 2015). Applied deep learning models were divided into Recurrent Neural Network (RNN), Feedforward Neural Network (FNN), and CNN. SimpleRNN, Gated Recurrent Unit (GRU), and LSTM were used for the RNN. A Multi-Layer Perceptron (MLP) with two hidden layers was implemented for the FNN, with 1-dimensional CNN being used for the CNN, passing information forward. The models were trained and evaluated using time series cross-validation with 5 splits. Early stopping criteria were implemented to avoid overfitting, which would stop the training process if no improvement was observed after 5 consecutive epochs. The configuration of each model is provided in Appendix I.

3.5.2. Model Evaluation

The examined validation methods have been discussed in the model development. In conclusion, the train-test split and time series cross-validator with 5 splits were applied to assess model performance, ensuring temporal order to avoid data leakage while splitting location-specific data. Evaluation metrics for regression mean absolute error (MAE) (Equation 2) and root mean squared error (RMSE) (Equation 3) from Scikit-learn (Pedregosa et al., 2011) were employed to quantify the models' predictive ability on unseen data. The RMSE penalizes larger errors through the square root, making it particularly relevant for regression tasks. MAE was chosen due to its easy interpretability, directly representing the average error.

$$(2)\ MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (3)\ RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- n is the number of observations.
- y_i is the actual value.
- $\hat{y}_i$ is the predicted value.
- $|y_i - \hat{y}_i|$ is the absolute difference between the actual and predicted values.
- $(y_i - \hat{y}_i)^2$ is the squared difference between the actual and the predicted values.

3.6. EVALUATION

The CRISP-DM framework established a systematic and robust foundation to address the domain objectives. Environmental factors driving disease spread were analyzed. The project's aim to develop a forecast using dengue incidence and environmental variables in Southeast Asia was achieved. Subsequently, the project proceeds to the deployment phase to address the last objective.

The final review revealed that all data mining goals were successfully realized. The data search yielded sufficient and reliable data points for 17 Southeast Asian locations. Guided by the initial exploratory analysis, various data preparation techniques were tested. Multiple machine and deep learning models were developed using different approaches, followed by model optimization through hyperparameter tuning. Lastly, predictive capability was evaluated through repetitive training and validation using the evaluation metrics MAE and RMSE.

3.7. DEPLOYMENT

In the last phase, the forecast was deployed to assess the disease's response to changing climate. The integrated climate change simulations were obtained from Sentian et al. (2022), who analyzed climate change interactions with monsoon seasons to assess the impact of temperature and rainfall variability in Southeast Asia. Their research leveraged the projected emission scenarios RCP4.5 and RCP8.5 from the IPCC AR5. Based on these two scenarios, they simulated the percentage change of Southeast Asia's monthly mean total precipitation and mean surface temperature in January and July for 2030, 2050, 2070, and 2100 relative to 2013 (Sentian et al., 2022).

The underlying research employed these simulated changes as a framework to predict respective changes in annual incidence rates. The variables *Monthly_Total_Prcp* and *Monthly_Avg_Temp* were modified with the simulated changes in the test set, ceteris paribus. If the simulated changes in total precipitation resulted in negative values, they were set to 0. After the training phase, the disease response assessment was performed on the entire test set comprising 12 data points, representing a year. Given the relevance of lagged observations, their simulated changes were captured by predicting subsequent months.

Finally, the predicted changes in annual incidence rates were compared to the actual predictions of the target variable to assess disease dynamics under climate change given two scenarios. In conclusion, the final objective was addressed.

4. RESULTS AND DISCUSSION

The research question led to three domain objectives.

4.1. ENVIRONMENTAL DRIVERS

The first objective was to explore the environmental drivers of disease spread. Although the literature discusses the importance of environmental variables, emphasizing temperature and precipitation, statistical techniques did not yield significant results. In the exploratory analysis, independent variables demonstrated unsatisfactory outcomes in the correlation analysis, with *Max_Daily_Temp_lag_4* showing the highest correlation of 0.11 with the target variable (Appendix J). Contrary to other studies conducted in Singapore (Seah et al., 2021), maximum daily temperature positively correlated with dengue incidence across all lags (Appendix J). However, research findings are consistent with Wang et al. (2022), showing a decreased correlation of minimum daily temperature at lag 2, followed by an increase in correlation at subsequent lags. Similar to their observation of extremely high temperatures being associated with increased risk at a 2-3 weeks lag (Wang et al., 2022), this research yielded comparable results, showing an increasing correlation between maximum temperature and incidence until lag 4, followed by a decrease in correlation (Appendix J). The four feature selection methods repeatedly highlighted some variables, indicating their significance (Appendix F). Lastly, varying incidence was observed across different years and locations (Appendix D), leading to no evidence of a causal relationship with the rainy season as outlined in previous literature (Li et al., 2018; World Health Organization, 2024a).

4.2. PREDICTIVE MODELING

For the second objective, the initial approach was establishing one holistic model across all locations to create a dengue incidence forecast for Southeast Asia. The primary evaluation was based on the validation set without parameter optimization to maintain a comparable setting for analyzing the performance of different data preparation techniques and models. In the preprocessing, two outlier capping logics were applied to mitigate bias and enhance data quality. The combination of both approaches revealed the best performance. Consequently, aside from data aggregation and the capping of relative outliers, the model performance advanced from addressing extremes on a multi-location level for the holistic model.

Training all variables, the broader research framework resulted in underfitting, with low predictive power for all models. Without hyperparameter tuning, KNN and RF emerged as the best-performing models, with an MAE of 17.62 and 18.90 and RMSE of 41.58 and 42.13 across locations, representing large errors. Although feature importance analysis showed high relevance for the variable *Location Code* in both methods (Appendix E), the large error suggests that the model could not differentiate between locations effectively. Consequently, the difficulty in learning location-specific patterns arose, supported by decreased and varying

training lengths (Figure 9), which might have led to the underrepresentation of some locations.

The independent approach encompassing multiple location-specific predictions showed better results. Evaluation metrics were averaged across 5 splits and locations to determine superior model performance. For each time series model, location-specific outlier capping was performed. After hyperparameter tuning, AdaBoost achieved the best result on the validation set for scale-invariant models with an average MAE of 13.19 and an RMSE of 16.75 using RF feature selection. For scale-variant models, SVR emerged as the best model with an average MAE of 13.03 and an RMSE of 16.35 using RobustScaler. The feature selection methods exhibited increased performance for certain models and methods. However, all four methods demonstrated inferior results compared to training all variables with SVR, which yielded the best performance across all machine learning models.

Deep learning models were trained, with SimpleRNN showing an average MAE of 15.88 and an RMSE of 19.31, outperforming GRU and LSTM. For the FNN, MLP yielded an average MAE of 13.25 and an RMSE of 16.83. CNN emerged as the best deep learning model for predicting dengue incidence, with an average MAE of 10.10 and an RMSE of 13.61. The final assessment revealed that deep learning models outperformed machine learning models. Appendix K presents the location-specific MAE and RMSE scores on the validation sets of the four superior models.

All project phases have been thoughtfully performed to mitigate bias, enhance data quality, and yield an effective prediction. However, the performance of the predictive models showed consistently inferior results for certain locations on the validation set (Appendix K). Reviewing the data exploration did not explain varying predictive abilities, as neither a common pattern, such as in the case of missing values, nor a country-specific correlation was observed. Underreporting, as argued by Bhatt et al. (2013), may have compromised the integrity of the raw data. Consequently, predictions might have been affected by inaccuracies in the secondary open-source data.

Models showed an overall decrease in errors on the test set, which is a good indicator for avoiding overfitting. AdaBoost and SVR showed an MAE of 10.22 and 9.31 and RMSE of 12.49 and 11.34 on the test set. MLP yielded an MAE of 6.82 and an RMSE of 8.87. CNN was chosen for the deployment, with an MAE of 5.06 and an RMSE of 7.09 on the test set. Large errors were reduced in almost all locations, which could be explained by increased data sufficiency. However, a location-specific shift in errors was observed in some instances. This suggests that reliable prediction was not tied to certain locations. Instead, the varying errors could have been related to other factors that were not included. The location-specific evaluation metrics on the test set are detailed in Appendix L.

4.3. CLIMATE CHANGE IMPACT ASSESSMENT

For the third objective, CNN was leveraged to predict the changes in annual incidence rates based on the simulated changes in mean temperature and total precipitation of Sentian et al. (2022). Kuching, Malaysia, was chosen for the deployment with the lowest predictive error across all locations. The percentage changes in the annual incidence rates in Kuching, Malaysia, for both emission scenarios are presented in Appendix M.

For both scenarios, a positive linear trend between the change in the annual dengue incidence rate in Kuching, Malaysia, and the mean temperature was observed (Figure 11). However, no clear trend was found between the changes in total precipitation and the target variable, showing scattered data points in Figure 12. These observations are consistent with Ebi and Nealon (2016), who reported that daily mean temperature and temperature variations are the main drivers of dengue incidence.

In conclusion, assessing disease dynamics based on changes in total precipitation, mean temperature, and lagged observations represents a substantial achievement. This success underscores the model's sensitivity in predicting dengue incidence and demonstrates its ability to assess the impact of climate change on disease transmission.

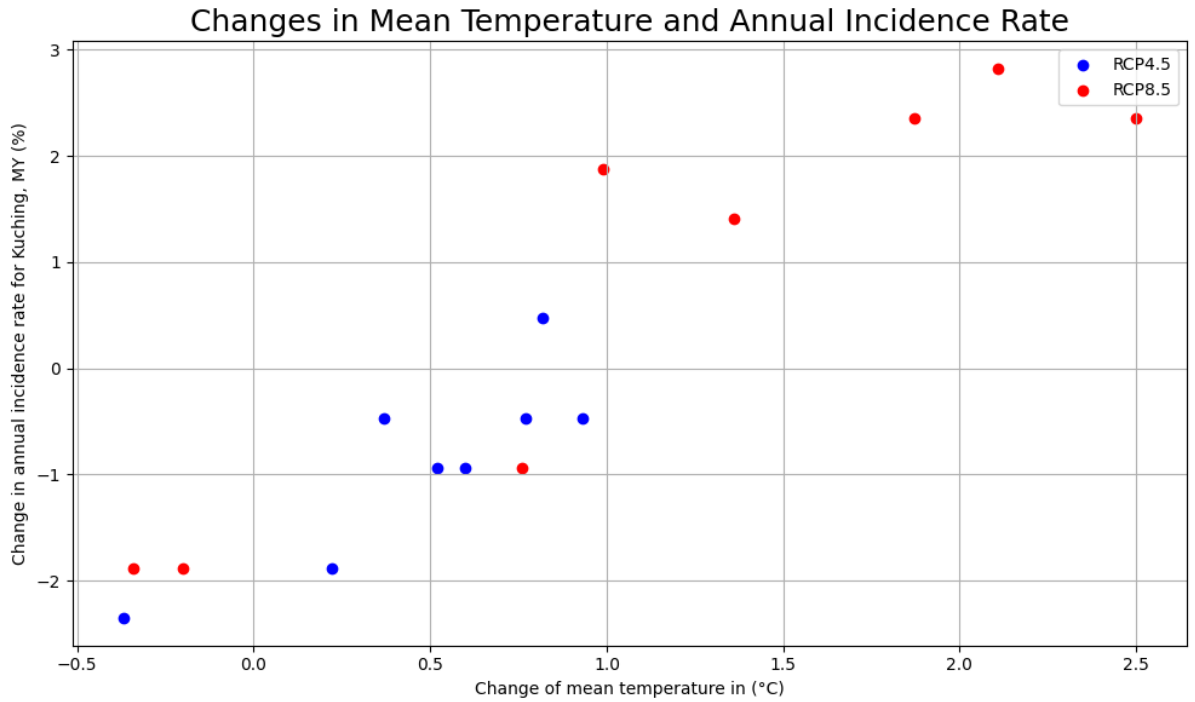


Figure 11 - Simulated changes in mean temperature and annual incidence rate

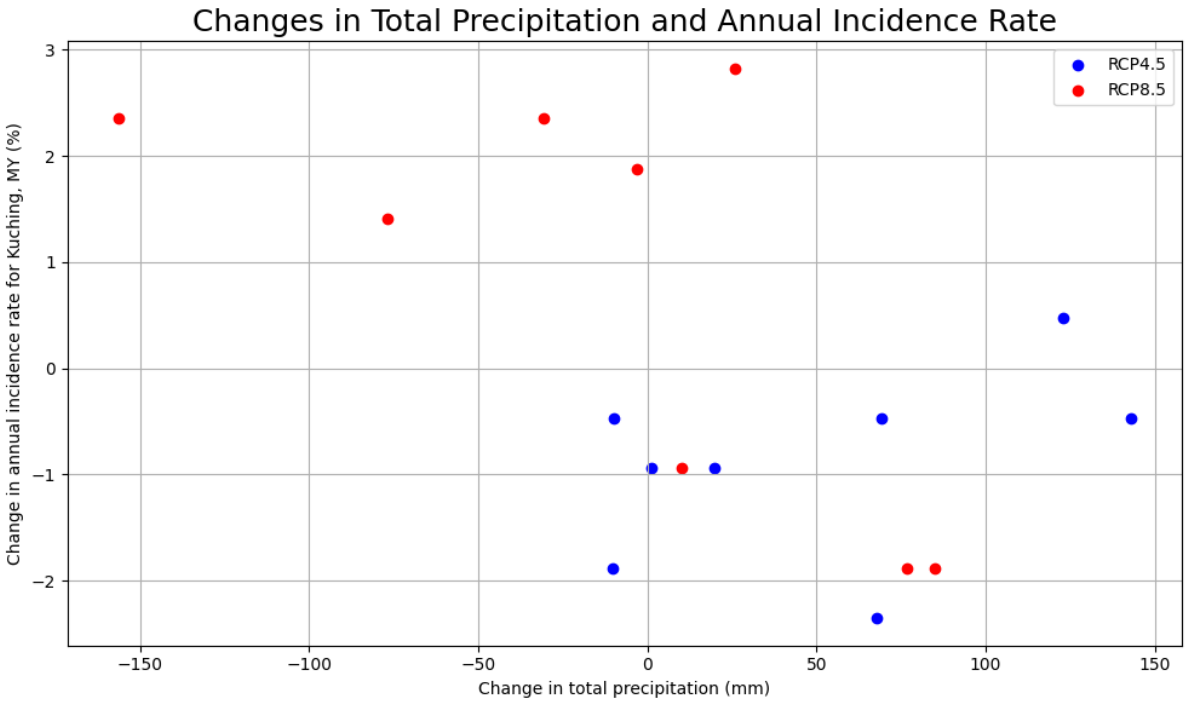


Figure 12 - Simulated changes in total precipitation and annual incidence rate

5. CONCLUSIONS

This study aimed to align with the latest research on dengue fever prediction, combining open-source data with data science tools. The CRISP-DM schema employed a systematic approach and standardized structure, enhancing the project's robustness and future maintenance. The variety of applied methods contributes to successfully answering the research question.

CNN emerged as the best-performing model with an MAE of 5.06 and an RMSE of 7.09 on the test set, demonstrating that predictive modeling using environmental variables can effectively assess dengue fever incidence. Deep learning models yielded a slight decrease in errors compared to traditional machine learning models. The reduction of errors achieved throughout the project highlights the necessity of profound data preparation and the effectiveness of the chosen approach.

Applied models consistently showed inferior results for certain locations on the validation set. The review of the exploratory analysis did not explain varying predictive abilities. Extensive data preparation and the multitude of applied models led to the assumption that high predictive errors for certain locations were due to secondary data obtained. For instance, Bhatt et al. (2013) argued that there is underreporting of dengue capture. In contrast, a location-specific shift in errors was observed on the test set with overall decreased errors, showing variations with increased errors in some instances. This result indicated that the models generalized well with increasing data sufficiency, suggesting that varying errors were related to other unconsidered factors.

The climate change impact assessment results showed that total precipitation and mean temperature changes enabled a disease response, indicating a sensitive model. The simulated changes showed varying effects on annual incidence rates for both scenarios, with a positive linear trend observed for mean temperature. In conclusion, the model effectively assessed disease dynamics under climate change.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORK

This work encountered some notable limitations. Firstly, previously applied control measures that influenced regional conditions and dengue distribution were not considered, which might have affected the predictive ability. Additionally, limited data availability could have reduced predictive power. Decreased training sets due to multiple independent time series may have limited the model's ability to learn sequential dynamics. As decreased errors were observed on the test sets, increased data volume through continuous maintenance could improve future location-specific model performance.

The independent approach of location-specific time series makes the forecast regionally flexible. However, this project was based on environmental variables that have demonstrated relevance for predicting dengue fever in Southeast Asia. Therefore, it is not recommended to expand the forecast globally. Instead, future work could enhance the Southeast Asian analysis and build on Tran et al. (2022), who demonstrated superior model performance by adding an attention layer after the LSTM network. In addition, further work may expand the analysis with additional observations. The literature outlines a positive association between dengue fever and socioeconomic factors, which might enhance modeling efforts. Varying incidences across locations may result from unique geographical conditions directly affecting mosquito habitats. Adding demographic and topographic variables could improve the prediction of dengue incidence. Moreover, improved data documentation might allow the differentiation of dengue-incidence-induced Aedes species to enhance the interpretability of disease dynamics.

The climate change simulations and their application encompass limitations. While total precipitation and mean temperature changes enabled a disease response, their effects might be interrelated. Future work should analyze associations between independent variables and their interrelated effects to enhance dengue risk assessment. In this project, the impact of climate change on the annual incidence rates was assessed on the test set and not temporally aligned with the projections until 2100 due to the assumption of continuously increasing error of the underlying prediction and required weather forecasts. A long-term goal for the third objective would be to integrate a regional climate model rather than relying on secondary climate change simulations. This way, time periods could be homogenized to increase representativeness. Furthermore, the additional emission scenarios RCP2.6 and RCP6.0 could be incorporated. In addition, climate change simulations were limited to two variables at specific timestamps. Since independent variables in the underlying analysis are related, simultaneous change is expected. Simulating changes for the remaining variables would enhance the model's predictive ability, allowing a comprehensive assessment of disease dynamics. Lastly, climate change was defined as a long-term change in environmental variables (Li et al., 2018). Therefore, simulated changes at the specified timestamps do not represent a real-world scenario.

Lastly, understanding the disease cycle, initiating factors, and identifying risk areas is crucial to mitigate future dengue risk. Data documentation and storage need to be enhanced globally and made publicly available to facilitate knowledge exchange and enable research. Data collection revealed that national incidence reporting demands more standardization. Furthermore, cross-national collaboration could improve disease capture and analysis. In this regard, data science could unlock great potential to address the future disease burden.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems* [White paper]. Google Research. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45166.pdf

Benedum, C. M., Seidahmed, O. M. E., Eltahir, E. A. B., & Markuzon, N. (2018). Statistical modeling of the effect of rainfall flushing on dengue transmission in Singapore. *PLoS Neglected Tropical Diseases*, *12*(12), e0006935. https://doi.org/10.1371/journal.pntd.0006935

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., Drake, J. M., Brownstein, J. S., Hoen, A. G., Sankoh, O., Myers, M. F., George, D. B., Jaenisch, T., Wint, G. R. W., Simmons, C. P., Scott, T. W., Farrar, J. J., & Hay, S. I. (2013). The global distribution and burden of dengue. *Nature*, *496*(7446), 504–507. https://doi.org/10.1038/nature12060

Bonnin, L., Tran, A., Herbreteau, V., Marcombe, S., Boyer, S., Mangeas, M., & Menkes, C. (2022). Predicting the Effects of Climate Change on Dengue Vector Densities in Southeast Asia through Process-Based Modeling. *Environmental Health Perspectives*, *130*(12), 127002. https://doi.org/10.1289/EHP11068

Chapman, P., Clinton, J., Khabaza, T., & Reinartz, T. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS. http://www.crisp-dm.org/CRISPWP-0800.pdf

Chareonviriyaphap, T., Akratanakul, P., Nettanomsak, S., & Huntamai, S. (2003). Larval habitats and distribution patterns of Aedes aegypti (Linnaeus) and Aedes albopictus (Skuse), in Thailand. *The Southeast Asian Journal of Tropical Medicine and Public Health*, *34*(3), 529–535.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, J., Bambrick, H., Frentiu, F., Devine, G., Yakob, L., Xu, Z., Li, Z., Yang, W., & Hu, W. (2021). Extreme weather events and dengue outbreaks in Guangzhou, China: A time-series quasi-binomial distributed lag non-linear model. *International Journal of Biometeorology*, *65*. https://doi.org/10.1007/s00484-021-02085-1

Davis, C., Murphy, A. K., Bambrick, H., Devine, G. J., Frentiu, F. D., Yakob, L., Huang, X., Li, Z., Yang, W., Williams, G., & Hu, W. (2021). A regional suitable conditions index to

forecast the impact of climate change on dengue vectorial capacity. *Environmental Research, 195*, 110849. https://doi.org/10.1016/j.envres.2021.110849

Ebi, K. L., & Nealon, J. (2016). Dengue in a changing climate. *Environmental Research*, *151*, 115–123. https://doi.org/10.1016/j.envres.2016.07.026

Edillo, F. E., Roble, N. D., & Otero, N. D. (2012). The key breeding sites by pupal survey for dengue mosquito vectors, Aedes aegypti (Linnaeus) and Aedes albopictus (Skuse), in Guba, Cebu City, Philippines. *The Southeast Asian Journal of Tropical Medicine and Public Health*, *43*(6), 1365–1374.

European Centre for Disease Prevention and Control. (2023, December). *Dengue Worldwide Overview—Situation update*. https://www.ecdc.europa.eu/en/dengue-monthly

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. Computing in Science & Engineering. https://doi.org/10.1109/MCSE.2007.55

IPCC. (2014). *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (R. K. Pachauri & L. A. Meyer, Eds.). IPCC.

Kesorn, K., Ongruk, P., Chompoosri, J., Phumee, A., Thavara, U., Tawatsin, A., & Siriyasatien, P. (2015). Morbidity Rate Prediction of Dengue Hemorrhagic Fever (DHF) Using the Support Vector Machine and the Aedes aegypti Infection Rate in Similar Climates and Geographical Areas. *PLoS ONE*, *10*(5), e0125049. https://doi.org/10.1371/journal.pone.0125049

Kraemer, M. U. G., Reiner, R. C., Brady, O. J., Messina, J. P., Gilbert, M., Pigott, D. M., Yi, D., Johnson, K., Earl, L., Marczak, L. B., Shirude, S., Davis Weaver, N., Bisanzio, D., Perkins, T. A., Lai, S., Lu, X., Jones, P., Coelho, G. E., Carvalho, R. G., … Golding, N. (2019). Past and future spread of the arbovirus vectors Aedes aegypti and Aedes albopictus. *Nature Microbiology*, *4*(5), 854–863. https://doi.org/10.1038/s41564-019-0376-y

Kulkarni, M. A., Duguay, C., & Ost, K. (2022). Charting the evidence for climate change impacts on the global spread of malaria and dengue and adaptive responses: A scoping review of reviews. *Globalization and Health*, *18*(1), 1. https://doi.org/10.1186/s12992-021-00793-2

Li, C., Lu, Y., Liu, J., & Wu, X. (2018). Climate change and dengue fever transmission in China: Evidences and challenges. *Science of The Total Environment*, *622–623*, 493–501. https://doi.org/10.1016/j.scitotenv.2017.11.326

Messina, J. P., Brady, O. J., Golding, N., Kraemer, M. U. G., Wint, G. R. W., Ray, S. E., Pigott, D. M., Shearer, F. M., Johnson, K., Earl, L., Marczak, L. B., Shirude, S., Davis Weaver, N., Gilbert, M., Velayudhan, R., Jones, P., Jaenisch, T., Scott, T. W., Reiner, R. C., & Hay, S. I. (2019). The current and future global distribution and population at risk of dengue. *Nature Microbiology*, *4*(9), 1508–1515. https://doi.org/10.1038/s41564-019-0476-8

Messina, J. P., Brady, O. J., Pigott, D. M., Golding, N., Kraemer, M. U. G., Scott, T. W., Wint, G. R. W., Smith, D. L., & Hay, S. I. (2015). The many projected futures of dengue. *Nature Reviews. Microbiology*, *13*(4), 230–239. https://doi.org/10.1038/nrmicro3430

Nuraini, N., Fauzi, I. S., Fakhruddin, M., Sopaheluwakan, A., & Soewono, E. (2021). Climate-based dengue model in Semarang, Indonesia: Predictions and descriptive analysis. *Infectious Disease Modelling*, *6*, 598–611. https://doi.org/10.1016/j.idm.2021.03.005

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference*, 92–96. https://doi.org/10.25080/Majora-92bf1922-011

Seah, A., Aik, J., Ng, L.-C., & Tam, C. C. (2021). The effects of maximum ambient temperature and heatwaves on dengue infections in the tropical city-state of Singapore – A time series analysis. *Science of The Total Environment*, *775*, 145117. https://doi.org/10.1016/j.scitotenv.2021.145117

Sentian, J., Payus, C., Herman, F., & Kong, V. (2022). Climate change scenarios over Southeast Asia. *APN Science Bulletin*, *2022*, 103–110. https://doi.org/10.30852/sb.2022.1927

Sundari, B., & Krishnamoorthy, M. (2019). Factors to Predict Dengue Fever using Data Mining Techniques: A Review. *International Journal of Scientific Research and Engineering Development*, *2*(4), 154–160.

The pandas development team. (2024). *pandas-dev/pandas: Pandas* (v2.2.2) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.10957263

Tran, T.-H., Minh, H., Thi Trang Nhung, N., Ngoc-Bich, N., & Tran, N. Q. L. (2022). Deep learning models for forecasting dengue fever based on climate data in Vietnam. *PLoS Neglected Tropical Diseases*, *16*. https://doi.org/10.1371/journal.pntd.0010509

Wang, Y., Wei, Y., Li, K., Jiang, X., Li, C., Yue, Q., Zee, B. C., & Chong, K. C. (2022). Impact of extreme weather on dengue fever infection in four Asian countries: A modelling analysis. *Environment International, 169*, 107518. https://doi.org/10.1016/j.envint.2022.107518

Wang, Y., Zhao, S., Wei, Y., Li, K., Jiang, X., Li, C., Ren, C., Yin, S., Ho, J., Ran, J., Han, L., Zee, B. C., & Chong, K. C. (2023). Impact of climate change on dengue fever epidemics in South and Southeast Asian settings: A modelling study. *Infectious Disease Modelling, 8*(3), 645–655. https://doi.org/10.1016/j.idm.2023.05.008

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wongkoon, S., Jaroensutasinee, M., & Jaroensutasinee, K. (2013). Distribution, seasonal variation & dengue transmission prediction in Sisaket, Thailand. *The Indian Journal of Medical Research, 138*(3), 347–353.

World Health Organization. (2024a). *Health topics, Dengue and severe dengue*. https://www.who.int/health-topics/dengue-and-severe-dengue

World Health Organization. (2024b, April 23). *Fact sheets, Dengue and severe dengue*. https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

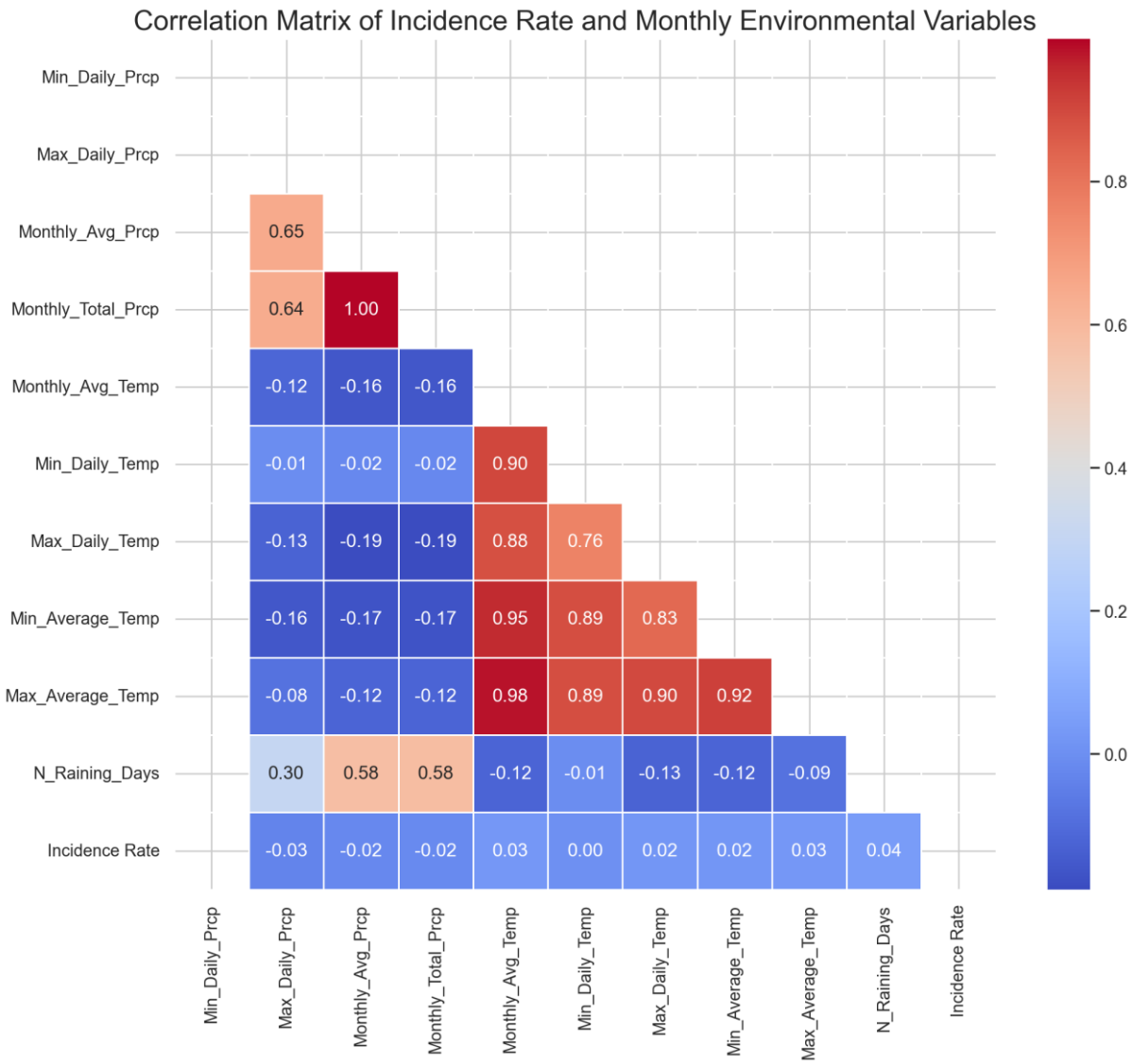
APPENDIX A

Libraries	Version
Python	3.11.5
Matplotlib	3.8.2
NumPy	1.26.2
Pandas	2.1.4
Scikit-learn	1.4.0
Seaborn	0.13.1
Statsmodels	0.14.1
TensorFlow	2.15.0
XGBoost	2.0.3

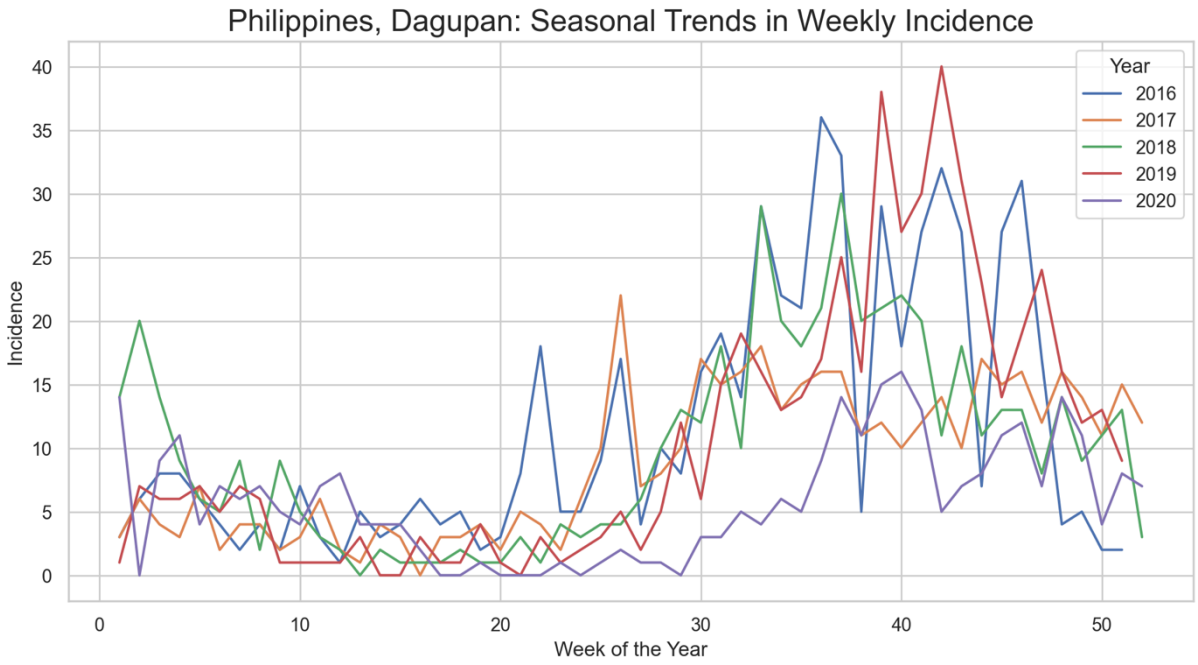
APPENDIX B

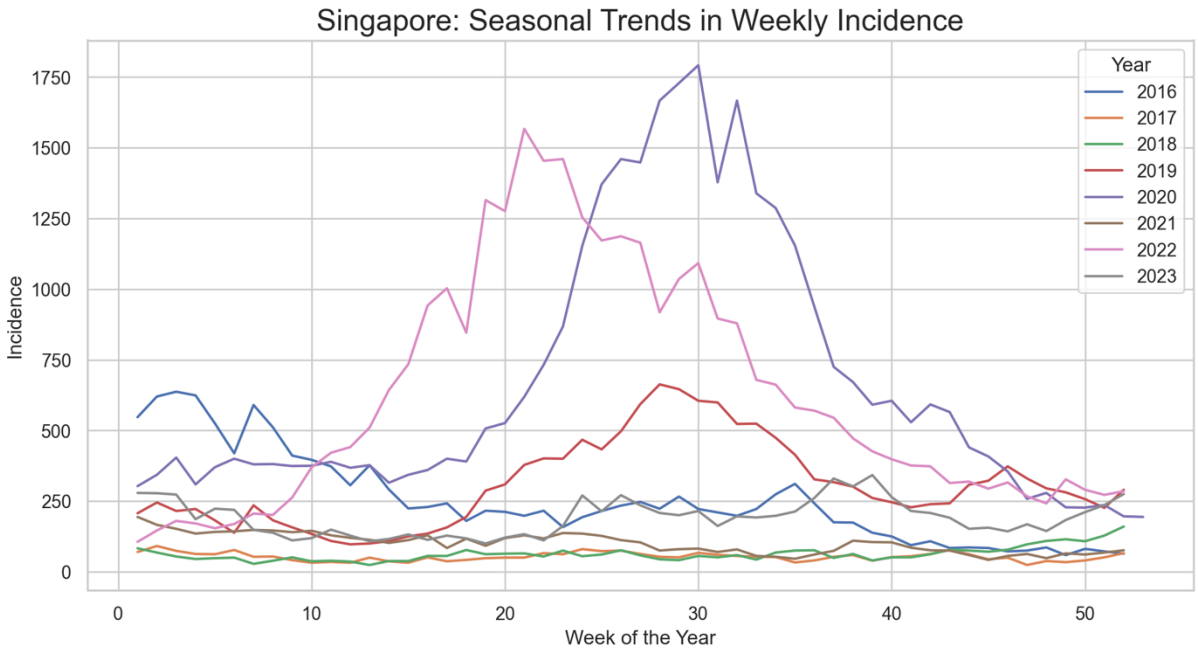
Code and data availability statement: The entire code and data used are available at https://github.com/josephinelutter/master-thesis.

APPENDIX C

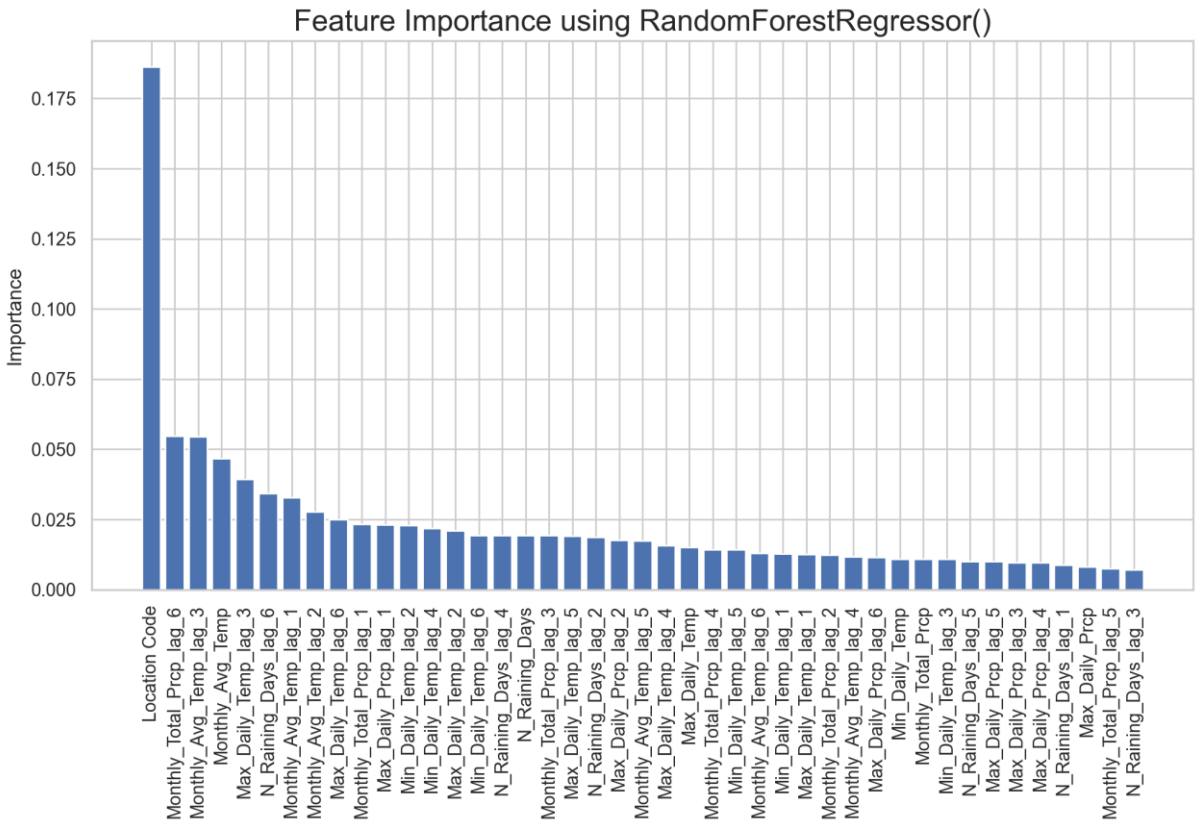

Correlation Matrix of Incidence Rate and Monthly Environmental Variables
Min_Daily_Prcp
Max_Daily_Prcp
Monthly_Avg_Prcp
Monthly_Total_Prcp
Monthly_Avg_Temp
Min_Daily_Temp
Max_Daily_Temp
Min_Average_Temp
Max_Average_Temp
N_Raining_Days
Incidence Rate
0.65
0.64 1.00
-0.12 -0.16 -0.16
-0.01 -0.02 -0.02 0.90
-0.13 -0.19 -0.19 0.88 0.76
-0.16 -0.17 -0.17 0.95 0.89 0.83
-0.08 -0.12 -0.12 0.98 0.89 0.90 0.92
0.30 0.58 0.58 -0.12 -0.01 -0.13 -0.12 -0.09
-0.03 -0.02 -0.02 0.03 0.00 0.02 0.02 0.03 0.04
0.8
0.6
0.4
0.2
0.0

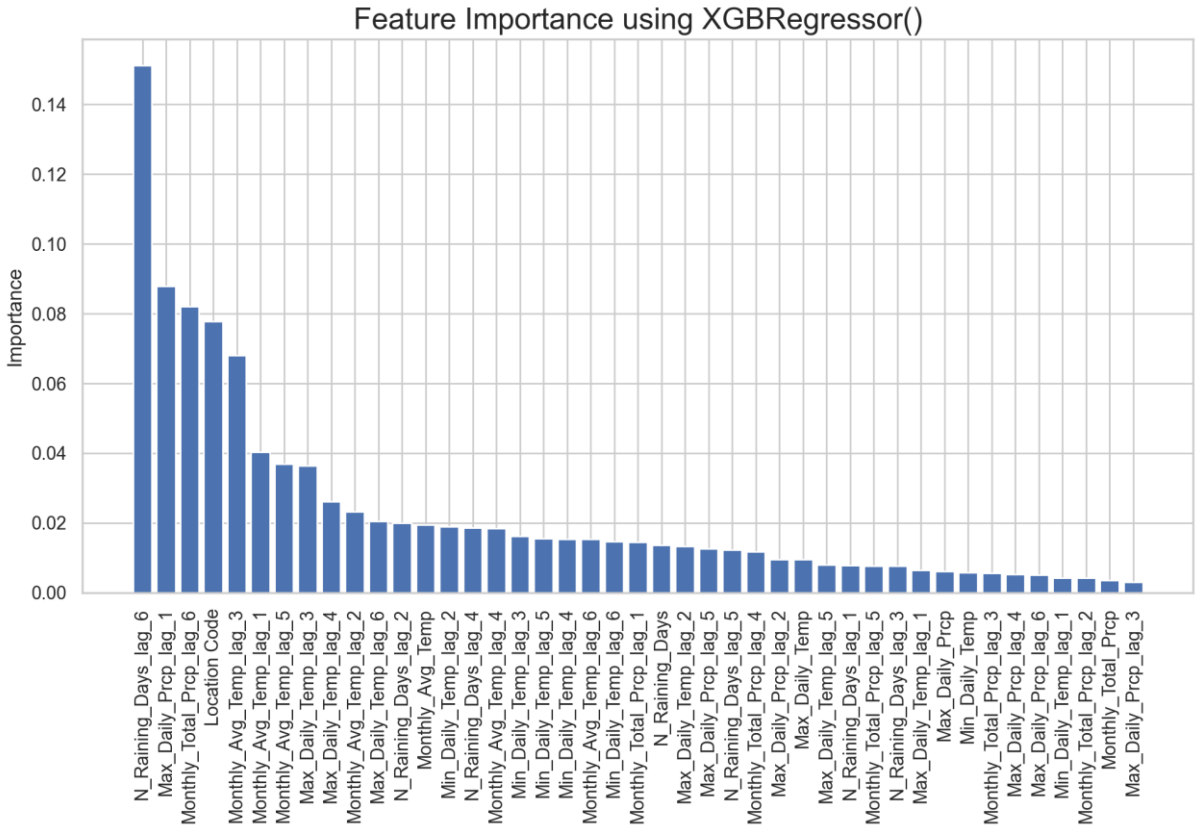
APPENDIX D


Philippines, Dagupan: Seasonal Trends in Weekly Incidence
Incidence
Week of the Year
Year
2016
2017
2018
2019
2020


Singapore: Seasonal Trends in Weekly Incidence
Incidence
Week of the Year
Year
2016
2017
2018
2019
2020
2021
2022
2023

APPENDIX E


Feature Importance using RandomForestRegressor()
Importance
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Location Code
Monthly_Total_Prcp_lag_6
Monthly_Avg_Temp_lag_3
Monthly_Avg_Temp
Max_Daily_Temp_lag_3
N_Raining_Days_lag_6
Monthly_Avg_Temp_lag_1
Monthly_Avg_Temp_lag_2
Max_Daily_Temp_lag_6
Monthly_Total_Prcp_lag_1
Max_Daily_Prcp_lag_1
Min_Daily_Temp_lag_2
Min_Daily_Temp_lag_4
Max_Daily_Temp_lag_2
Min_Daily_Temp_lag_6
N_Raining_Days_lag_4
N_Raining_Days
Monthly_Total_Prcp_lag_3
Max_Daily_Temp_lag_5
N_Raining_Days_lag_2
Max_Daily_Prcp_lag_2
Monthly_Avg_Temp_lag_5
Max_Daily_Temp_lag_4
Max_Daily_Temp
Monthly_Total_Prcp_lag_4
Min_Daily_Temp_lag_5
Monthly_Avg_Temp_lag_6
Min_Daily_Temp_lag_1
Max_Daily_Temp_lag_1
Monthly_Total_Prcp_lag_2
Monthly_Avg_Temp_lag_4
Max_Daily_Prcp_lag_6
Min_Daily_Temp
Monthly_Total_Prcp
Min_Daily_Temp_lag_3
N_Raining_Days_lag_5
Max_Daily_Prcp_lag_5
Max_Daily_Prcp_lag_3
Max_Daily_Prcp_lag_4
N_Raining_Days_lag_1
Max_Daily_Prcp
Monthly_Total_Prcp_lag_5
N_Raining_Days_lag_3

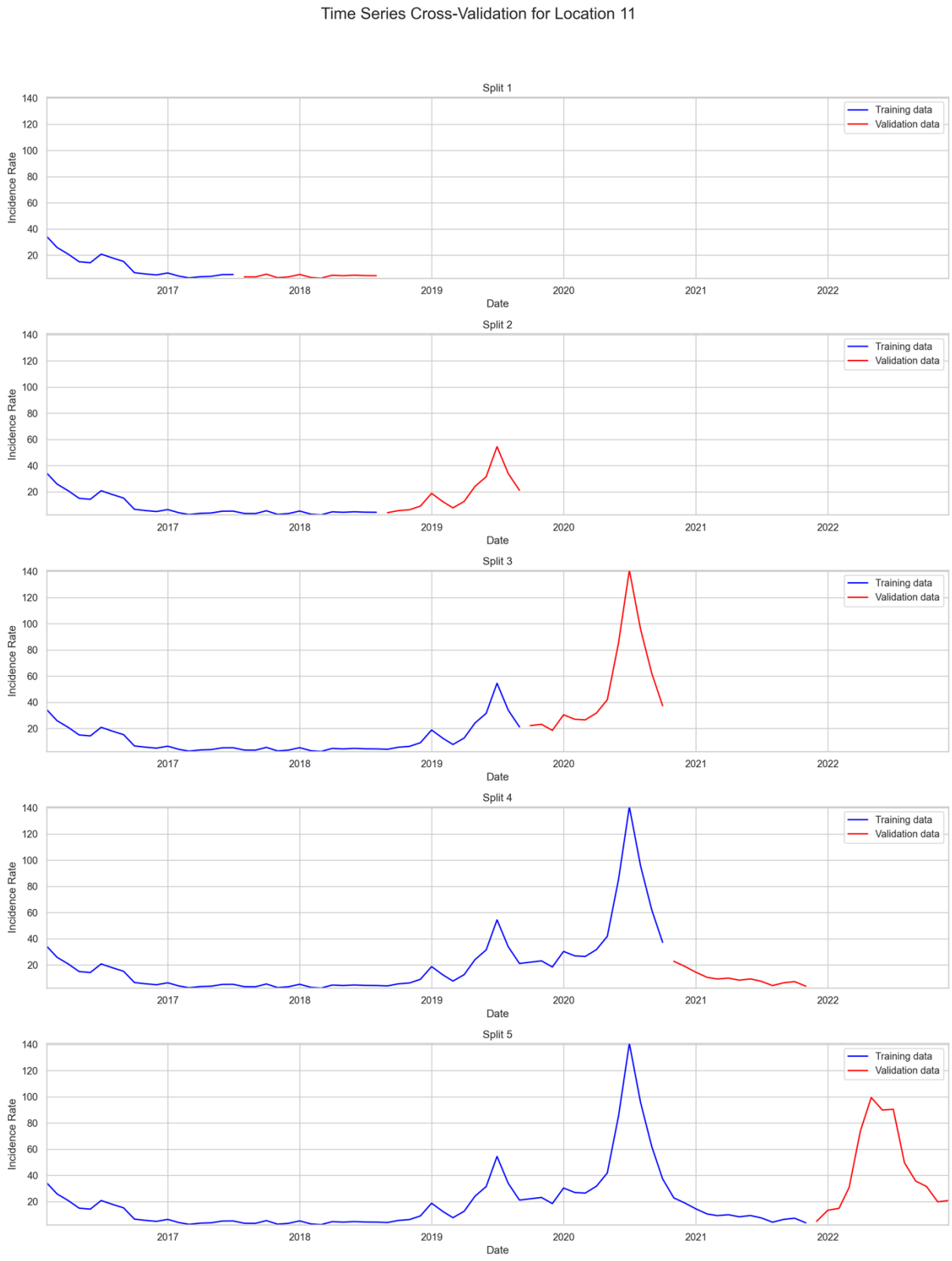

Feature Importance using XGBRegressor()
Importance
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
N_Raining_Days_lag_6
Max_Daily_Prcp_lag_1
Monthly_Total_Prcp_lag_6
Location Code
Monthly_Avg_Temp_lag_3
Monthly_Avg_Temp_lag_1
Monthly_Avg_Temp_lag_5
Max_Daily_Temp_lag_3
Max_Daily_Temp_lag_4
Monthly_Avg_Temp_lag_2
Max_Daily_Temp_lag_6
N_Raining_Days_lag_2
Monthly_Avg_Temp
Min_Daily_Temp_lag_2
N_Raining_Days_lag_4
Monthly_Avg_Temp_lag_4
Min_Daily_Temp_lag_3
Min_Daily_Temp_lag_5
Min_Daily_Temp_lag_4
Monthly_Avg_Temp_lag_6
Min_Daily_Temp_lag_6
Monthly_Total_Prcp_lag_1
N_Raining_Days
Max_Daily_Temp_lag_2
Max_Daily_Prcp_lag_5
N_Raining_Days_lag_5
Monthly_Total_Prcp_lag_4
Max_Daily_Prcp_lag_2
Max_Daily_Temp
Max_Daily_Temp_lag_5
N_Raining_Days_lag_1
Monthly_Total_Prcp_lag_5
N_Raining_Days_lag_3
Max_Daily_Temp_lag_1
Max_Daily_Prcp
Min_Daily_Temp
Monthly_Total_Prcp_lag_3
Max_Daily_Prcp_lag_4
Max_Daily_Prcp_lag_6
Min_Daily_Temp_lag_1
Monthly_Total_Prcp_lag_2
Monthly_Total_Prcp
Max_Daily_Prcp_lag_3

APPENDIX F

Feature	Method				Total Count
	Select From Model using RF	Select From Model using XG Boost	RFE using RF	Forward Feature Selection using RF	
Location Code	1	1	1	1	4
Max_Daily_Temp_lag_3	1	1	1	0	3
Monthly_Avg_Temp_lag_5	0	1	1	1	3
Monthly_Avg_Temp_lag_2	1	0	1	1	3
Monthly_Avg_Temp	1	0	1	1	3
Monthly_Total_Prcp_lag_6	1	1	1	0	3
Monthly_Avg_Temp_lag_3	1	1	1	0	3
N_Raining_Days_lag_6	1	1	1	0	3
Max_Daily_Temp_lag_4	0	1	0	1	2
Monthly_Avg_Temp_lag_1	1	1	0	0	2
Monthly_Total_Prcp	0	0	0	1	1
Monthly_Total_Prcp_lag_1	0	0	1	0	1
Monthly_Avg_Temp_lag_6	0	0	0	1	1
Max_Daily_Prcp_lag_1	0	1	0	0	1
Max_Daily_Temp_lag_6	1	0	0	0	1
N_Raining_Days_lag_1	0	0	0	1	1
Min_Daily_Temp_lag_2	0	0	0	1	1
N_Raining_Days	0	0	0	1	1
Max_Daily_Prcp_lag_2	0	0	1	0	1

Note. 1: Feature was selected by the corresponding method, 0: Feature was not selected.

APPENDIX G


Time Series Cross-Validation for Location 11
Split 1
Split 2
Split 3
Split 4
Split 5
Training data
Validation data
Incidence Rate
Date
140
120
100
80
60
40
20
2017
2018
2019
2020
2021
2022

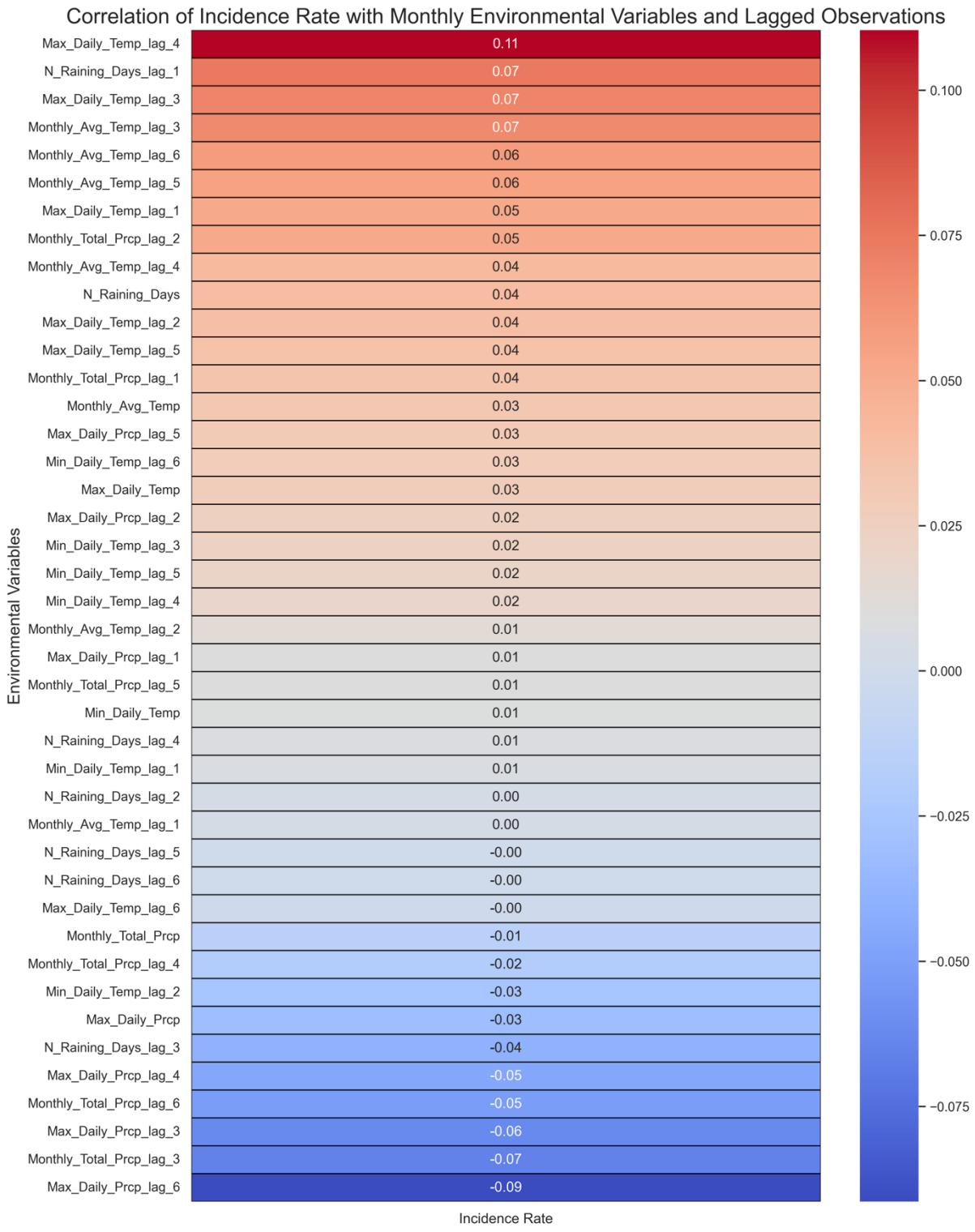
APPENDIX H

Model	Hyperparameters
DT	{}
RF	{max_depth: 20, max_features: 'sqrt', min_samples_leaf: 4, min_samples_split: 10}
XGBoost	{colsample_bytree: 0.7, learning_rate: 0.014, max_depth: 9, min_child_weight: 2, n_estimators: 141, reg_lambda: 3, subsample: 0.956}
AdaBoost	{learning_rate: 0.01, loss: 'linear'}
KNN	{metric: 'euclidean', n_neighbors: 10, weights: 'distance'}
SVR	{C: 100, kernel: 'rbf'}

APPENDIX I

Model	Model Configurations
SimpleRNN	{Model: Sequential, Layer 1: SimpleRNN, 50 units, Output Layer: Dense, 1 unit, Loss Function: MAE, Optimizer: Adam, Epochs: 50, Batch Size: 72}
GRU	{Model: Sequential, Layer 1: GRU, 50 units, Output Layer: Dense, 1 unit, Loss Function: MAE, Optimizer: Adam, Epochs: 50, Batch Size: 72}
LSTM	{Model: Sequential, Layer 1: LSTM, 50 units, Output Layer: Dense, 1 unit, Loss Function: MAE, Optimizer: Adam, Epochs: 50, Batch Size: 72}
MLP	{Model: Sequential, Layer 1: Dense, 64 units, ReLU activation, Layer 2: Dense, 32 units, ReLU activation, Output Layer: Dense, 1 unit, Loss Function: MAE, Optimizer: Adam, Epochs: 50, Batch Size: 72}
CNN	{Model: Sequential, Layer 1: Dense, 64 units, ReLU activation, Layer 2: Dense, 31 units, ReLU activation, Output Layer: Dense, 1 unit, Loss Function: MAE, Optimizer: Adam, Epochs: 50, Batch Size: 72}

APPENDIX J


Correlation of Incidence Rate with Monthly Environmental Variables and Lagged Observations
Environmental Variables
Incidence Rate
Max_Daily_Temp_lag_4 0.11
N_Raining_Days_lag_1 0.07
Max_Daily_Temp_lag_3 0.07
Monthly_Avg_Temp_lag_3 0.07
Monthly_Avg_Temp_lag_6 0.06
Monthly_Avg_Temp_lag_5 0.06
Max_Daily_Temp_lag_1 0.05
Monthly_Total_Prcp_lag_2 0.05
Monthly_Avg_Temp_lag_4 0.04
N_Raining_Days 0.04
Max_Daily_Temp_lag_2 0.04
Max_Daily_Temp_lag_5 0.04
Monthly_Total_Prcp_lag_1 0.04
Monthly_Avg_Temp 0.03
Max_Daily_Prcp_lag_5 0.03
Min_Daily_Temp_lag_6 0.03
Max_Daily_Temp 0.03
Max_Daily_Prcp_lag_2 0.02
Min_Daily_Temp_lag_3 0.02
Min_Daily_Temp_lag_5 0.02
Min_Daily_Temp_lag_4 0.02
Monthly_Avg_Temp_lag_2 0.01
Max_Daily_Prcp_lag_1 0.01
Monthly_Total_Prcp_lag_5 0.01
Min_Daily_Temp 0.01
N_Raining_Days_lag_4 0.01
Min_Daily_Temp_lag_1 0.01
N_Raining_Days_lag_2 0.00
Monthly_Avg_Temp_lag_1 0.00
N_Raining_Days_lag_5 -0.00
N_Raining_Days_lag_6 -0.00
Max_Daily_Temp_lag_6 -0.00
Monthly_Total_Prcp -0.01
Monthly_Total_Prcp_lag_4 -0.02
Min_Daily_Temp_lag_2 -0.03
Max_Daily_Prcp -0.03
N_Raining_Days_lag_3 -0.04
Max_Daily_Prcp_lag_4 -0.05
Monthly_Total_Prcp_lag_6 -0.05
Max_Daily_Prcp_lag_3 -0.06
Monthly_Total_Prcp_lag_3 -0.07
Max_Daily_Prcp_lag_6 -0.09
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075

APPENDIX K

		MAE on the validation set			
Location Code	**Name**	**AdaBoost**	**SVR**	**CNN**	**MLP**
0	Baguio, RP	10.99	11.06	8.56	16.93
1	Dagupan, RP	13.07	10.39	14.51	20.45
2	Kota Kinabalu International, MY	3.75	3.53	2.46	12.47
3	Kuala Lumpur International, MY	27.11	29.14	22.09	13.60
4	Kuantan, MY	6.89	7.45	4.14	13.44
5	Kuching, MY	3.18	2.86	2.03	11.83
6	Labuan, MY	5.51	4.10	1.83	14.44
7	Malacca, MY	9.25	8.07	5.05	16.01
8	Manila, RP	8.10	8.36	9.04	18.91
9	Mersing, MY	10.17	9.86	4.66	13.25
10	Penang International, MY	12.80	11.37	5.74	12.52
11	Singapore Changi International, SN	19.17	19.99	18.63	18.92
12	Sitiawan, MY	6.77	5.57	1.96	11.68
13	Sultan Abdul Aziz Shah International, MY	20.71	22.84	15.65	13.00
14	Sultan Ismail Petra, MY	11.63	10.80	7.77	17.44
15	Surigao, RP	21.71	24.01	18.68	23.48
16	Zamboanga, RP	33.48	32.10	28.84	21.36
Average MAE		13.19	13.03	10.10	13.25

Location Code	Name	RMSE on the validation set			
		AdaBoost	**SVR**	**CNN**	**MLP**
0	Baguio, RP	14.56	13.13	11.40	21.62
1	Dagupan, RP	15.73	13.27	17.65	25.91
2	Kota Kinabalu International, MY	4.56	4.26	3.22	16.87
3	Kuala Lumpur International, MY	31.21	33.07	26.07	17.32
4	Kuantan, MY	7.73	8.30	4.79	18.20
5	Kuching, MY	3.71	3.26	2.26	15.56
6	Labuan, MY	6.66	4.92	2.49	19.65
7	Malacca, MY	10.80	9.31	6.11	21.79
8	Manila, RP	9.79	10.27	11.43	26.00
9	Mersing, MY	11.62	11.34	5.92	16.83
10	Penang International, MY	14.50	13.9	7.60	16.50
11	Singapore Changi International, SN	25.63	25.81	25.05	25.89
12	Sitiawan, MY	7.50	6.59	2.69	15.78
13	Sultan Abdul Aziz Shah International, MY	25.34	26.93	21.27	17.14
14	Sultan Ismail Petra, MY	13.81	12.84	10.37	23.86
15	Surigao, RP	28.68	30.51	24.66	29.73
16	Zamboanga, RP	52.85	50.16	48.39	29.52
Average MAE		16.75	16.35	13.61	16.83

APPENDIX L

		MAE on the test set			
Location Code	**Name**	**AdaBoost**	**SVR**	**CNN**	**MLP**
0	Baguio, RP	14.78	10.18	2.36	2.30
1	Dagupan, RP	16.43	9.82	8.23	9.48
2	Kota Kinabalu International, MY	12.4	11.68	11.43	11.55
3	Kuala Lumpur International, MY	14.39	14.36	11.13	34.74
4	Kuantan, MY	2.04	4.28	2.02	1.99
5	Kuching, MY	0.79	1.87	0.62	0.62
6	Labuan, MY	2.1	2.28	1.37	1.42
7	Malacca, MY	4.67	2.69	1.66	1.58
8	Manila, RP	9.33	12.01	2.78	2.84
9	Mersing, MY	4.22	3.59	3.83	3.78
10	Penang International, MY	5.48	6.91	4.86	4.79
11	Singapore Changi International, SN	11.26	10.51	3.74	3.84
12	Sitiawan, MY	3.37	3.82	1.40	1.36
13	Sultan Abdul Aziz Shah International, MY	9.31	15.03	8.64	13.48
14	Sultan Ismail Petra, MY	3.34	5.88	2.84	2.77
15	Surigao, RP	26.41	24.79	15.99	16.19
16	Zamboanga, RP	33.37	18.53	3.10	3.19
MAE		10.22	9.31	5.06	6.82

Location Code	Name	RMSE on the test set			
		AdaBoost	**SVR**	**CNN**	**MLP**
0	Baguio, RP	23.16	12.19	3.37	3.21
1	Dagupan, RP	21.08	13.88	10.45	11.72
2	Kota Kinabalu International, MY	14.4	13.94	13.52	13.75
3	Kuala Lumpur International, MY	16.8	17.53	12.75	36.42
4	Kuantan, MY	2.56	5.11	2.67	2.69
5	Kuching, MY	0.93	2.14	0.71	0.72
6	Labuan, MY	2.41	2.58	1.83	1.87
7	Malacca, MY	5.61	3.14	2.11	2.03
8	Manila, RP	10.31	14.68	4.76	4.90
9	Mersing, MY	5.69	4.97	4.56	4.70
10	Penang International, MY	6.55	10.74	7.69	7.75
11	Singapore Changi International, SN	13.33	11.46	4.25	4.37
12	Sitiawan, MY	4.12	4.58	1.86	1.76
13	Sultan Abdul Aziz Shah International, MY	11.4	18.59	11.46	15.14
14	Sultan Ismail Petra, MY	4.59	6.81	3.37	3.65
15	Surigao, RP	33.25	29.64	29.87	30.56
16	Zamboanga, RP	36.15	20.86	5.21	5.54
Average RMSE		12.49	11.34	7.09	8.87

APPENDIX M

RCP 4.5				
Year	**Month**	**Change of mean temperature in (°C)**	**Change in total precipitation (mm)**	**Change in annual incidence rate for Kuching, MY (%)**
2030	January	-0.37	67.72	-2,35
2030	July	0.22	-10.12	-1,88
2050	January	0.60	1.29	-0,94
2050	July	0.77	69.40	-0,47
2070	January	0.37	143.05	-0,47
2070	July	0.52	19.69	-0,94
2100	January	0.82	123.08	0,47
2100	July	0.93	-9.76	-0,47

RCP 8.5				
Year	**Month**	**Change of mean temperature in (°C)**	**Change of total precipitation (mm)**	**Change in annual incidence rate for Kuching, MY (%)**
2030	January	-0.34	76.69	-1,88
2030	July	0.76	10.09	-0,94
2050	January	-0.20	84.82	-1,88
2050	July	0.99	-3.09	1,88
2070	January	1.36	-76.87	1,41
2070	July	1.87	-30.61	2,35
2100	January	2.11	25.80	2,82
2100	July	2.50	-156.44	2,35

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa