Predicting Malaria Prevalence with Machine Learning Models Using Satellite-based Climate Information

Technical Report

December 2023
Colombo, Sri Lanka

Kaveesha Ileperuma, Mahesh Jampani, Uvindu Sellahewa, Shweta Panjwani and Giriraj Amarnath

CGIAR
INITIATIVE ON
Climate Resilience

Affilitation of authors

Kaveesha Ileperuma[1], Mahesh Jampani[1], Uvindu Sellahewa[1], Shweta Panjwani[2] and Giriraj Amarnath[1]

[1] International Water Management Institute (IWMI), Colombo, Sri Lanka

[2] International Water Management Institute (IWMI), New Delhi, India

Suggested Citation

Ileperuma, K.; Jampani, M.; Sellahewa, U.; Panjwani, S.; Amarnath, G. 2023. *Predicting malaria prevalence with machine learning models using satellite-based climate information: technical report.* Colombo, Sri Lanka: International Water Management Institute (IWMI). CGIAR Initiative on Climate Resilience. 32p.

Acknowledgments

This work was carried out with support from the CGIAR Initiative on Climate Resilience, ClimBeR. We would like to thank all funders who supported this research through their contributions to the CGIAR Trust Fund.

CGIAR Initiative on Climate Resilience

The CGIAR Initiative on Climate Resilience, also known as ClimBeR, aims to transform the climate adaptation capacity of food, land, and water systems and ultimately increase the resilience of smallholder production systems to better adapt to climate extremes. Its goal is to tackle vulnerability to climate change at its roots and support countries and local and indigenous communities in six low-and middle-income countries to better adapt and build equitable and sustainable futures.
Learn more about ClimBeR here: https://www.cgiar.org/initiative/climate-resilience/

Disclaimer

This publication has been prepared as an output of the CGIAR Initiative on Climate Resilience and has not been independently peer reviewed. Responsibility for editing, proofreading, and layout, opinions expressed and any possible errors lies with the authors and not the institutions involved.

Summary

The current report presents a machine learning model developed to predict malaria prevalence based on rainfall patterns, specifically tailored to different regions within Senegal. The developed model takes into account the varying climate conditions across regions to provide a more localized and accurate prediction. The primary input parameters used for prediction include rainfall, month, and year, allowing the model to capture each region's seasonal variations and trends. This research aims to enhance the precision of malaria predictions, contributing to more effective and targeted public health measures. The model is designed to provide future forecasts, offering valuable insights into early warning signals to help anticipate and mitigate the impact of malaria outbreaks. This proactive approach enables authorities and healthcare professionals to prepare and implement preventive measures in advance, potentially reducing the severity of malaria-related issues and aiding in the allocation of resources where they are most needed. By tailoring the prediction model to the unique characteristics of each region in Senegal, the current research addresses the localized nature of malaria outbreaks, recognizing that factors such as climate, geography, and environmental conditions can significantly influence the prevalence of malaria. The integration of predictive analytics and models in public health initiatives allows for a more strategic and responsive approach to malaria management, ultimately contributing to the overall well-being of the affected communities. This report includes an explanation of the methodology used for the development of the prediction model, along with the results obtained and their implications for public health in Senegal.

Table of Contents

SUMMARY **3**

1. INTRODUCTION **5**

2. DATA COLLECTION AND PROCESSING **7**

3. MACHINE LEARNING MODEL SELECTION **7**

4. RANDOM FOREST ALGORITHM **9**

5. MODEL DEVELOPMENT AND PERFORMANCE **10**

5.1. IMPORT LIBRARIES 10
5.2. IMPORT DATA FRAME 11
5.3. MODELS BUILDING 12
5.3.1. Year-wise Split model training 12
5.3.2. Random Split model training 13

6. EVALUATION MATRICES EXPLANATION **14**

7. MODEL RESULTS AND DATA INTERPRETATION **23**

8. LAG TIME CALCULATION **23**

9. SIGNIFICANCE AND FUTURE PERSPECTIVES **28**

10. CONCLUSIONS AND RECOMMENDATIONS **30**

REFERENCES **30**

1. Introduction

Malaria is a vector-borne infectious disease, typically transmitted through infected Anopheles mosquitoes, and has long been an embedded public health challenge, particularly in the Global South. Malaria prevalence and intensity are intricately tied to various environmental and climatic factors (Kulkarni et al., 2022; Samarasekera, 2023). The aftermath of water-related hazards often triggers the spread of vector-borne diseases like malaria (Coalson et al., 2021; Mouchet et al., 1996). Specifically for mosquito-borne diseases, the number of yearly reported cases at the global scale has risen by an estimated 247 million malaria cases, marking 0.6 million deaths (https://www.who.int/news-room/fact-sheets/detail/malaria). The nexus between malaria and climate has never been more critical as dwindling rainfall and rising global temperatures become increasingly erratic due to human-induced climate change. This evolving dynamic has potential consequences for regions where malaria is already endemic and areas outside the disease's traditional range. By diving into the climatic determinants of malaria transmission, such as temperature, precipitation, and relative humidity, and examining the complex interplay between them (Bationo et al., 2021; Santos-Vega et al., 2022; Wang et al., 2022), we can anticipate potential shifts in disease patterns and devise strategies to take preventive measures. This exploration aims to underscore the pressing need for collaborative, interdisciplinary action in adapting to and mitigating the impacts of climate on malaria and, by extension, global public health. This study adds a new dimension to a decade of CGIAR research on malaria led by IWMI (Mutero et al., 2005; Teklu et al., 2010; Kibret et al., 2015; Kibret et al., 2021).

Malaria has long presented a significant public health issue in many tropical and subtropical regions, and its incidence and distribution are especially relevant in Senegal, where the disease remains a primary health concern (Sallah et al., 2021). As the global community faces the multifaceted consequences of climate change, understanding its influence on malaria transmission within Senegal becomes essential. Senegal has a unique combination of Sahelian, Sudanian, and Guinean climate zones; even slight alterations in temperature and precipitation can produce marked effects on mosquito breeding habitats and malaria transmission cycles. These regional trends explore the intricate relationship between climatic variables and malaria

dynamics in Senegal (Diouf et al., 2013; Jampani et al., 2023). Factors such as changing rainfall patterns, increasing temperatures, and altered humidity levels can lead to shifts in disease prevalence, potentially causing outbreaks in previously low-transmission areas and altering peak transmission seasons (Dieng et al., 2020; Diouf et al., 2017; Ndiath et al., 2012). By studying these specific climatic impacts on malaria transmission in Senegal, it becomes evident that targeted, region-specific interventions are crucial. This understanding will guide policymakers, researchers, and public health officials in crafting adaptive, resilient strategies that safeguard communities from the augmented threats of a changing climate. In Senegal, the rainy season usually lasts from June to October, coinciding with a surge in malaria cases (Figure 1). Further, high humidity and increased rainfall provide ideal breeding conditions for mosquitoes, leading to an upsurge in malaria transmission.

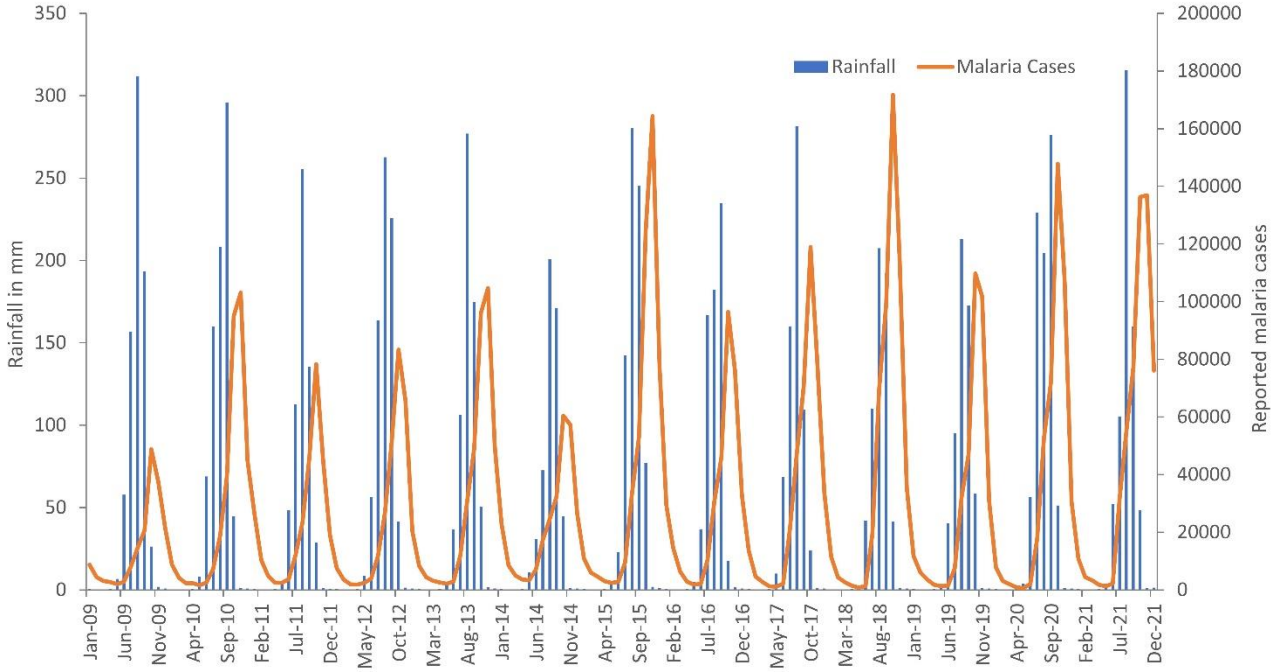


Figure 1: Distribution of rainfall variability and malaria prevalence for entire Senegal.

Over the past decade, various efforts have been made to develop effective interventions and preventive measures to mitigate the impact of malaria on public health in Senegal (Bicout et al., 2015; Diouf et al., 2017; Lucas et al., 2021). However, accurately predicting malaria outbreaks remains complex due to the influence of various environmental factors, particularly rainfall

patterns. To address this challenge, we leveraged advanced machine learning techniques to develop robust malaria prediction models that could provide early warning signals and prevent the spread of malaria. By grasping the links between rainfall patterns and malaria prevalence, we aim to forecast the number of malaria cases for the next three months across different regions in Senegal.

2. Data collection and processing

The monthly and province-wise dataset of recorded malaria cases was obtained from Senegal's National Malaria Control Program (NMCP - PNLP in French); the data collection information on malaria cases can be obtained from https://pnlp.sn/. When looking at the overall dataset, time series data for the period 2009 to 2021 are supplied in a tabular format Excel sheet in numerical format for 14 provinces of Senegal. When we visualize the dataset, it shows the non-linear relationship between the data variables. Each region has more than 150 data rows with rainfall and malaria cases between 2009 and 2021. It covers climatic and health data for 14 different regions and runs from 2009 to 2021. The two crucial elements assembled into this dataset are the historical rainfall data for each region and the associated number of malaria cases reported in each region. The combination of these variables provides a comprehensive picture of how environmental elements like rainfall interact with the prevalence of malaria cases.

The satellite-derived rainfall estimates are from CHIRPS (Climate Hazards Group InfraRed Precipitation with Station) data, processed to obtain monthly rainfall trends and validated with country-specific station rainfall datasets. The malaria data is also pre-processed to filter any noise or outliers. Both datasets are at a monthly temporal scale, and the spatial scale is provincial for Senegal. Further, lag time is calculated given the time between rainfall events and potential spikes in malaria (due to the mosquito life cycle and human incubation period), creating lag variables (e.g., rainfall data from the previous month).

3. Machine Learning Model Selection

Selecting a suitable machine learning algorithm is pivotal to the model's success in malaria prediction modeling. First, we evaluated whether the prediction task is a classification,

regression, or clustering problem. Malaria prediction likely involves binary classification (malaria outbreak or not), making classification algorithms like Random Forest, Support Vector Machines (SVM), and Gradient Boosting as potential choices. When working with datasets, it is crucial to analyze their size, dimensionality, and features. Gradient Boosting and Random Forest models have proven to perform well in large datasets. For high-dimensional data, dimensionality reduction techniques may be beneficial. In addition to that, it is crucial to understand the distribution of data classes. If the dataset is imbalanced, where one class is significantly more prevalent than the others, algorithms like Random Forest and SVM can handle it, or techniques like oversampling or under-sampling can be considered. It is also vital to consider feature importance, especially if there is prior knowledge about which features are most relevant to prediction. Algorithms like Random Forest offer built-in feature importance scores, aiding interpretability. However, one needs to decide the balance between model interpretability and predictive performance. While decision trees and Random Forest are interpretable, deep learning models may provide higher performance at the cost of interpretability. Lastly, it is crucial to identify the presence of noisy data or outliers in the dataset. Robust algorithms like Random Forest and SVM can handle noisy data better than other models.

Python is a widely regarded language for developing machine learning models for several compelling reasons. TensorFlow, PyTorch, sci-kit-learn, and Keras are some of the many libraries and frameworks that are part of Python's vast ecosystem, specifically designed for machine learning and data science. These libraries simplify the machine learning workflow by providing pre-built tools and job functions, including data pre-processing, model creation, and evaluation. Python is a simple and readable language, making it a brilliant choice for novice and experienced developers. Its syntax is like everyday English, making writing and understanding code easy. It is valuable when working with complex machine learning algorithms and models. Python's flexibility also makes integrating with other technologies and languages easy. The ability to build comprehensive machine learning pipelines, deploy models, and access data depends on this interoperability. Overall, Python is the leading language enabling developers to effectively leverage the potential of artificial intelligence and data science due to its extensive environment, readability, versatility, and robust community support.

Machine learning encompasses various paradigms, including supervised, unsupervised, and linear learning methods, each serving distinct purposes. The most popular method uses labeled data to guide the algorithm's learning. In this paradigm, the model is trained using a dataset containing input data and target labels, allowing it to make predictions or categorize data that has not yet been seen. Applications for supervised learning can be found in processes like sentiment analysis, spam detection, and image identification. While dealing with unlabeled data, unsupervised learning looks for patterns, structures, or groupings within the data. It is comparable to providing the program with a group of things without identifying them. Clustering and dimensionality reduction are frequent methods used in unsupervised learning. Applications like consumer segmentation and anomaly detection make use of it.

As the name implies, linear learning algorithms are a type of algorithm that represent relationships between variables as linear equations. For example, the straightforward yet effective linear technique of linear regression is used to forecast numerical results. It presumes a linear relationship exists between the input features and the desired outcome. For binary classification problems, however, linear classifiers like logistic regression are employed. These techniques are preferred when the underlying data relationships are straightforward and can be well represented by linear functions.

4. Random Forest Algorithm

Random forest is one of the best algorithms for regression problems. While selecting an appropriate algorithm, we need to understand the dataset well because sometimes it will be small or huge. According to this dataset amount, we cannot go to deep learning models to build this model because this dataset is too small. As well as the dataset having missing values, as a solution for that, we can use the random forest algorithm because it helps handle missing values. Random Forest provides excellent feature importance insights, high accuracy, robustness, and resistance against overfitting. It does, however, come with more complexity and less interpretability. After considerable deliberation and analysis, we strategically chose to use the Random Forest algorithm for predictive modeling. The aspects of Random Forest can handle both classification tasks and numerical features, its feature importance insights, and its power to

mitigate overfitting. This choice was determined after analyzing the advantages and disadvantages of several algorithms and considering variables like predicted accuracy, interpretability, and computing efficiency.

5. Model development and performance

Machine learning models such as Random Forest are used to perform predictive modeling of malaria cases based on rainfall data. Two model validation methods are employed: i) splitting the data to train and test the datasets and ii) random splitting, where random years were chosen for training and testing instead of sequential. Testing of the model that is trained on both these approaches is to assess the model's final performance. The model developed with parameter initialization by starting with a set of initial parameters for the Random Forest, including the number of trees, depth of trees, and criteria for splitting. The model's performance on the test dataset is evaluated using R-squared (to determine the proportion of variance explained by the model). We integrate the model into the AWARE (Early warning to Early action) platform only after the model's performance is satisfied. The model is updated based on continuous data updates as received to improve the model's predictions over time. As more recent data becomes available, retrain and update the model to ensure its continued accuracy and relevance. The respective machine learning approaches of Python code are also incorporated in this report below.

5.1. Import Libraries

We imported some libraries for data analysis and machine learning in Python. We are setting up Pandas, NumPy, Matplotlib, Scikit-Learn, and date time in our Python environment for data pre-processing, analysis, and potentially machine learning tasks.

```
import pandas as pd
import numpy as np
import math
import matplotlib.pyplot as plt
%matplotlib inline
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
import seaborn as sns
import plotly.graph_objects as go
from scipy.interpolate import splev, splrep
import datetime
from datetime import timedelta
```

5.2. Import Data Frame

Used Pandas to read two Excel files, 'RainfallSenegal.xlsx' and 'SenegalMalaria.xlsx' and stored their data in Data Frames named Rainfall data and Malaria data, respectively. The random forest regression algorithm was used from Scikit-Learn to build a machine-learning model and evaluate its performance across regions.

The dataset path should be changed according to the dataset:

```
RainFall_data = pd.read_excel("Filepath\\RainfallSenegal.xlsx")
Malaria_data = pd.read_excel("Filepath\\SenegalMalaria.xlsx")
```

The malaria data frame and adding new columns Year and Months:

```
Malaria_data['Years_Month'] = pd.to_datetime(Malaria_data['Years'],
format='%Y/%m/%d')
Malaria_data['Year'] =
pd.DatetimeIndex(Malaria_data['Years_Month']).year
Malaria_data['Month'] =
pd.DatetimeIndex(Malaria_data['Years_Month']).month
```

Creating a new Data Frame for each region:

```
Region_DataSet = pd.DataFrame()
Region_DataSet['Years'] = Malaria_data['Year']
Region_DataSet['Mon'] = Malaria_data['Month']
Region_DataSet['RainFall'] = RainFall_data['Diourbel']
Region_DataSet['Malaria'] = Malaria_data['DIOURBEL']
```

```
Region_DataSet['yearWithMonth'] = Region_DataSet['Years'].astype(str)
+"-"+ Region_DataSet['Mon'].astype(str)
Region_DataSet = Region_DataSet[Region_DataSet['Years']>2007]
```

5.3. Models Building

5.3.1. Year-wise Split model training

```
X_Year_train = Region_DataSet[Region_DataSet['Years']<2019]
Y_Year_train = Region_DataSet[Region_DataSet['Years']>=2019]
sample = Y_Year_train['yearWithMonth']
y_train = X_Year_train['Malaria']
x_train = X_Year_train.drop(['Malaria','yearWithMonth'], axis=1)
y_test = Y_Year_train['Malaria']
x_test = Y_Year_train.drop(['Malaria','yearWithMonth'], axis=1)

model_1 = RandomForestRegressor(n_estimators = 10, random_state = 3,
max_depth = 5, criterion ='poisson',min_samples_split = 1,
1.0).fit(x_train,y_train)
acc = model_1.score(x_test,y_test)
print('2008 to 2021 Accuracy :',acc)
```

```
y_pred = model_1.predict(x_test)
```

```
plt.figure(figsize=(15,5))
plt.plot(sample, y_test, label='Actual', color=line1_color)
plt.plot(sample, y_pred, label='Predicted', color=line2_color)
plt.title('Actual and Predicted')
plt.legend(['Actual', 'Predicted'])
plt.xticks(rotation=90)
plt.show()
```

5.3.2. Random Split model training

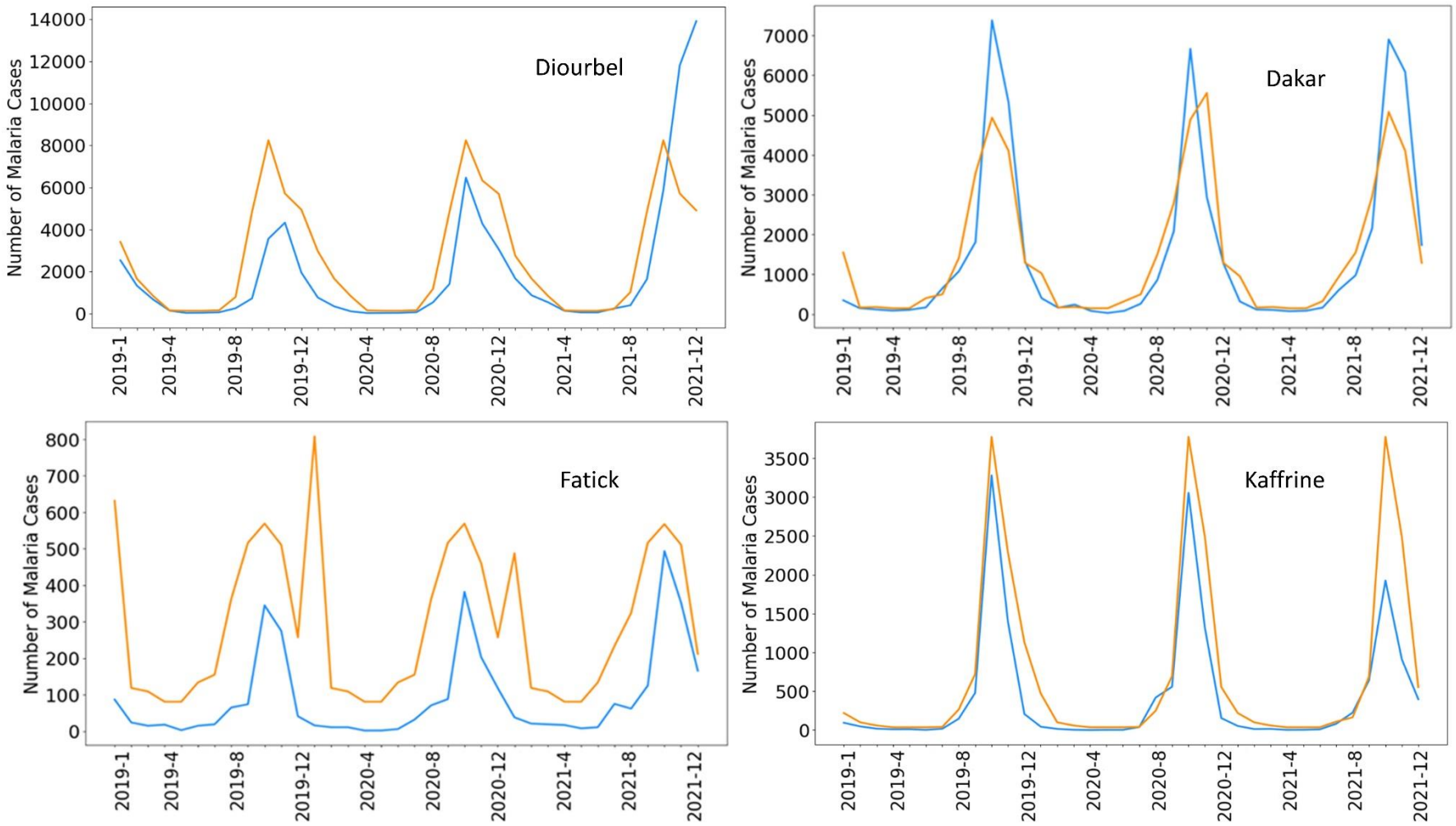
```
X_region = Region_DataSet.drop(['Malaria'], axis=1)
Y_region = Region_DataSet['Malaria']

X_region_train, X_region_test, Y_region_train, Y_region_test =
train_test_split(X_region, Y_region, test_size=0.1, random_state=42)

Sample_date = X_region_test['yearWithMonth']
X_region_train = X_region_train.drop(['yearWithMonth'], axis=1)
X_region_test = X_region_test.drop(['yearWithMonth'], axis=1)

model_2 = RandomForestRegressor(n_estimators = 10, random_state = 3,
max_depth = 5, criterion ='poisson',min_samples_split = 1,
                                min_samples_leaf = 1, max_features =
1.0).fit(X_region_train,Y_region_train)
acc = model_2.score(X_region_test,Y_region_test)
print('2008 to 2018 Accuracy :',acc)

y_region_pred = model_2.predict(X_region_test)

import matplotlib.pyplot as plt
# Sample data - predicted and actual values

# Set the width of the bars
#bar_width = 0.45
plt.figure(figsize=(20, 8))
# Create an array of indices for the x-axis ticks
#x = range(len(predicted))
line1_color = 'lightcoral'
line2_color = 'chocolate'
# Plotting the bars
plt.bar(X_region_test['Years'], Y_region_test, align='center',
label='Predicted',color=line1_color)
plt.bar(X_region_test['Years'], y_region_pred, align='edge',
label='Actual',color=line2_color)

# Set labels and title
plt.xlabel('Years')
plt.ylabel('Cases')
plt.title('Predicted vs Actual Values')

plt.xticks(x_test_data['years&Month'])  # Set x-axis tick labels

# Add legend
plt.legend()

# Display the plot
```

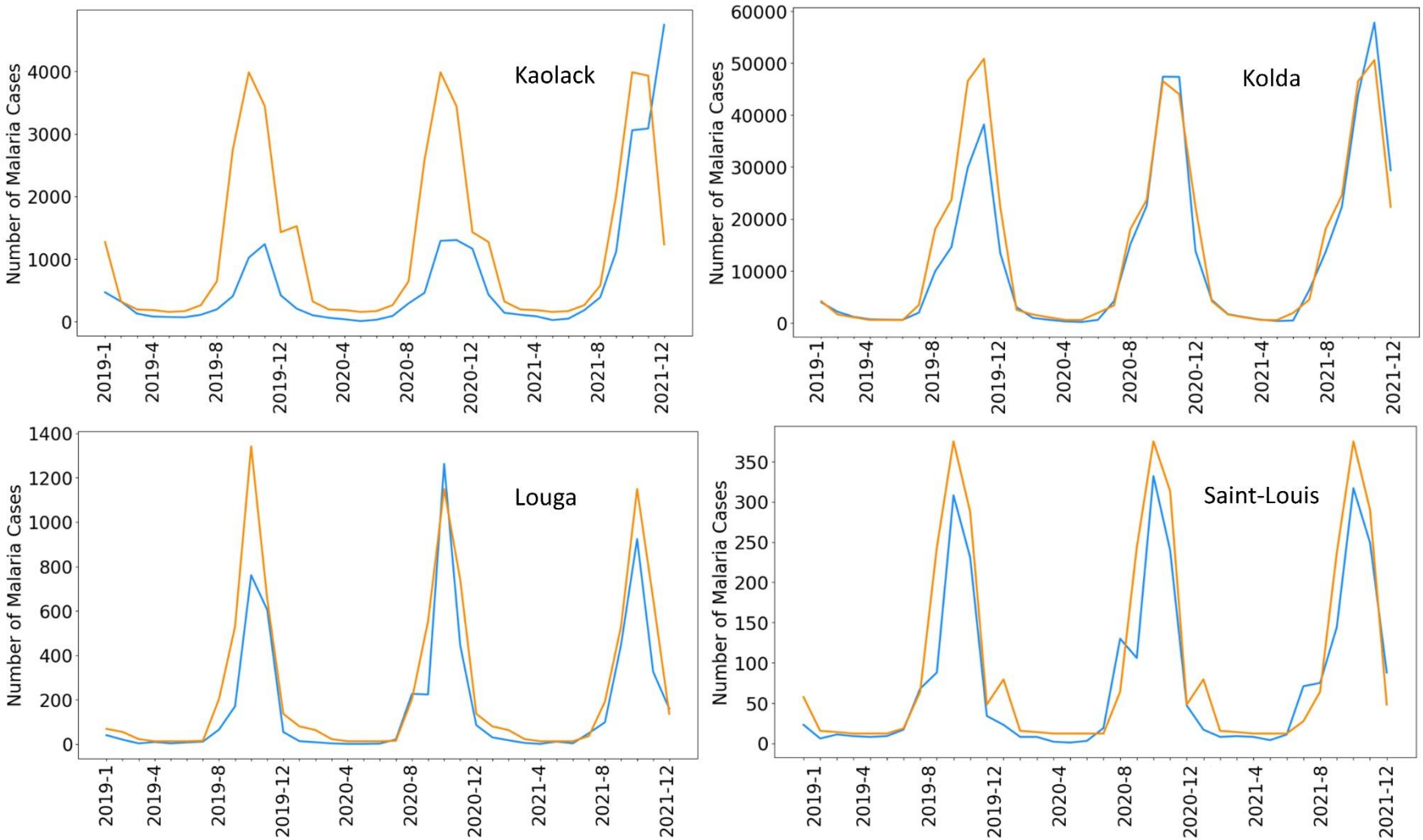
6. Evaluation Matrices Explanation

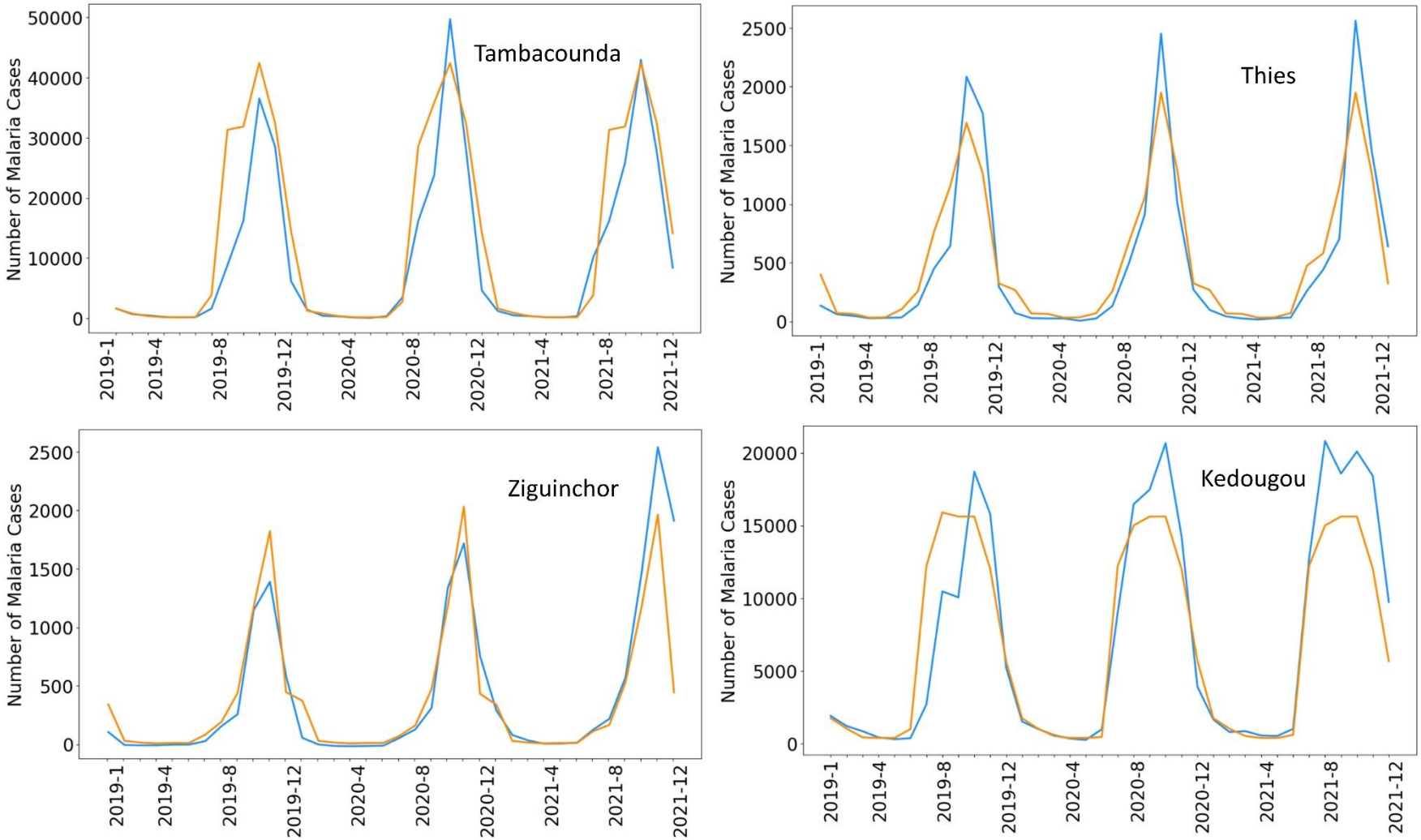
When evaluating the performance of a model, it is crucial to select appropriate metrics based on the nature of the problem. One standard metric is R-squared, which measures the proportion of the variance in the target variable and can be explained by the model's predictors. A higher value of R-squared indicates a better fit of the developed model. It is also crucial to evaluate the model's sensitivity to changes in specific variables, particularly when analyzing the impact of climate factors on the transmission of diseases such as malaria. Additionally, considering the model's behavior on different spatial and temporal scales, such as local, regional, national, monthly, or annually, can provide valuable insights into its effectiveness. To present the results of model training, two types of approaches were used: year-wise split and random split models. The corresponding accuracy of these models at the province level is presented in Table 1, while Figures 2 and 3 illustrate their performance.

Table 1: The province-wise model accuracies for Senegal of the Year-wise split and Random split models

Region	Year-wise Split Model Accuracy	Random Split Model Accuracy
Dakar	0.87	0.95
Diourbel	0.62	0.82
Fatick	0.62	0.97
Kaffrine	0.83	0.95
Kaolack	0.81	0.88
Kedougou	0.91	0.96
Kolda	0.92	0.95
Louga	0.93	0.69
Matam	0.89	0.71
Saint-Louis	0.95	0.87
Sedhiou	0.94	0.53
Tambacounda	0.92	0.92
Thies	0.80	0.74
Ziguinchor	0.90	0.82


Diourbel
Dakar
Fatick
Kaffrine
Number of Malaria Cases
2019-1
2019-4
2019-8
2019-12
2020-4
2020-8
2020-12
2021-4
2021-8
2021-12


Kaolack
Kolda
Louga
Saint-Louis
Number of Malaria Cases
0
1000
2000
3000
4000
10000
20000
30000
40000
50000
60000
200
400
600
800
1000
1200
1400
50
100
150
200
250
300
350
2019-1
2019-4
2019-8
2019-12
2020-4
2020-8
2020-12
2021-4
2021-8
2021-12


Tambacounda
Thies
Ziguinchor
Kedougou
Number of Malaria Cases
2019-1
2019-4
2019-8
2019-12
2020-4
2020-8
2020-12
2021-4
2021-8
2021-12

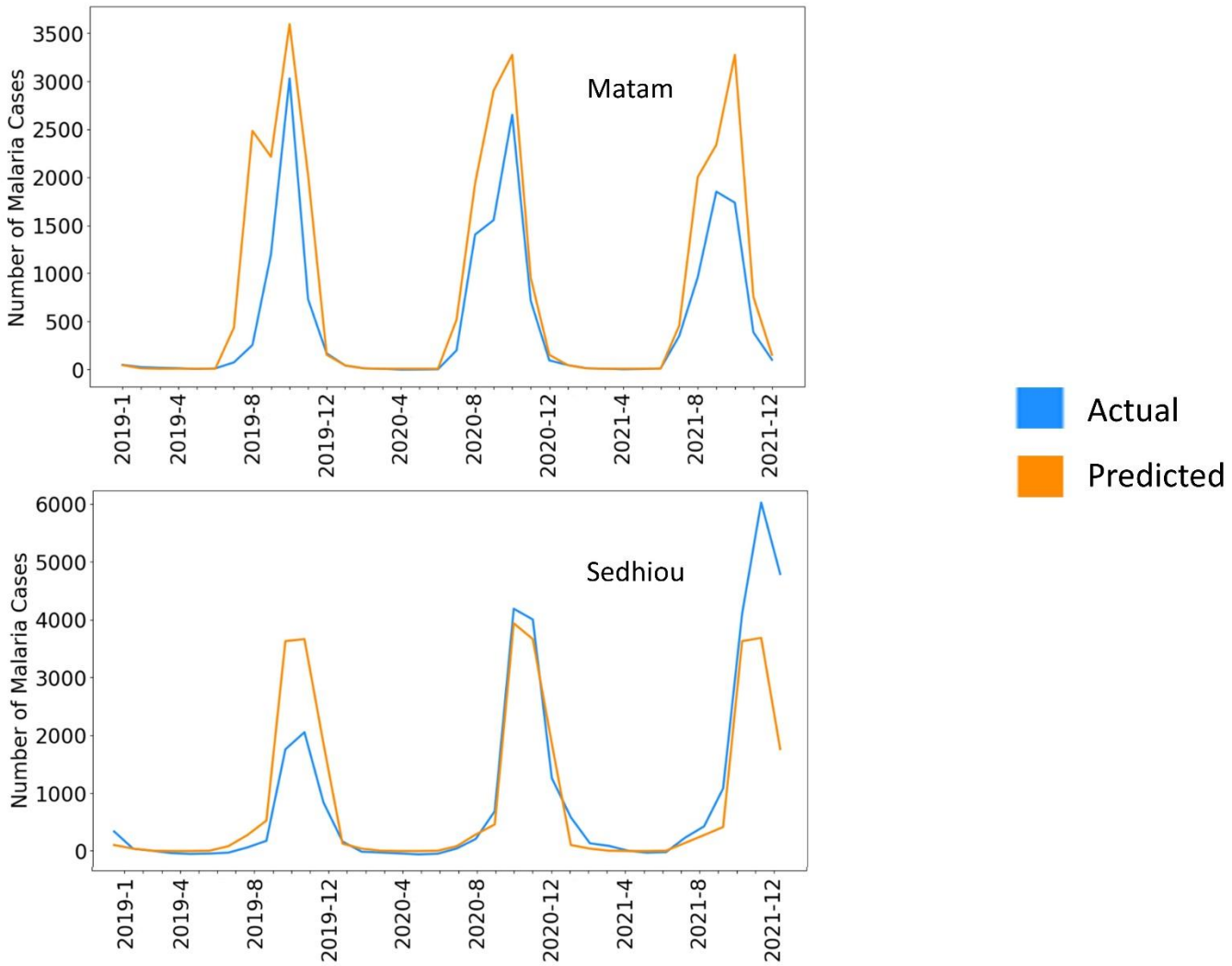
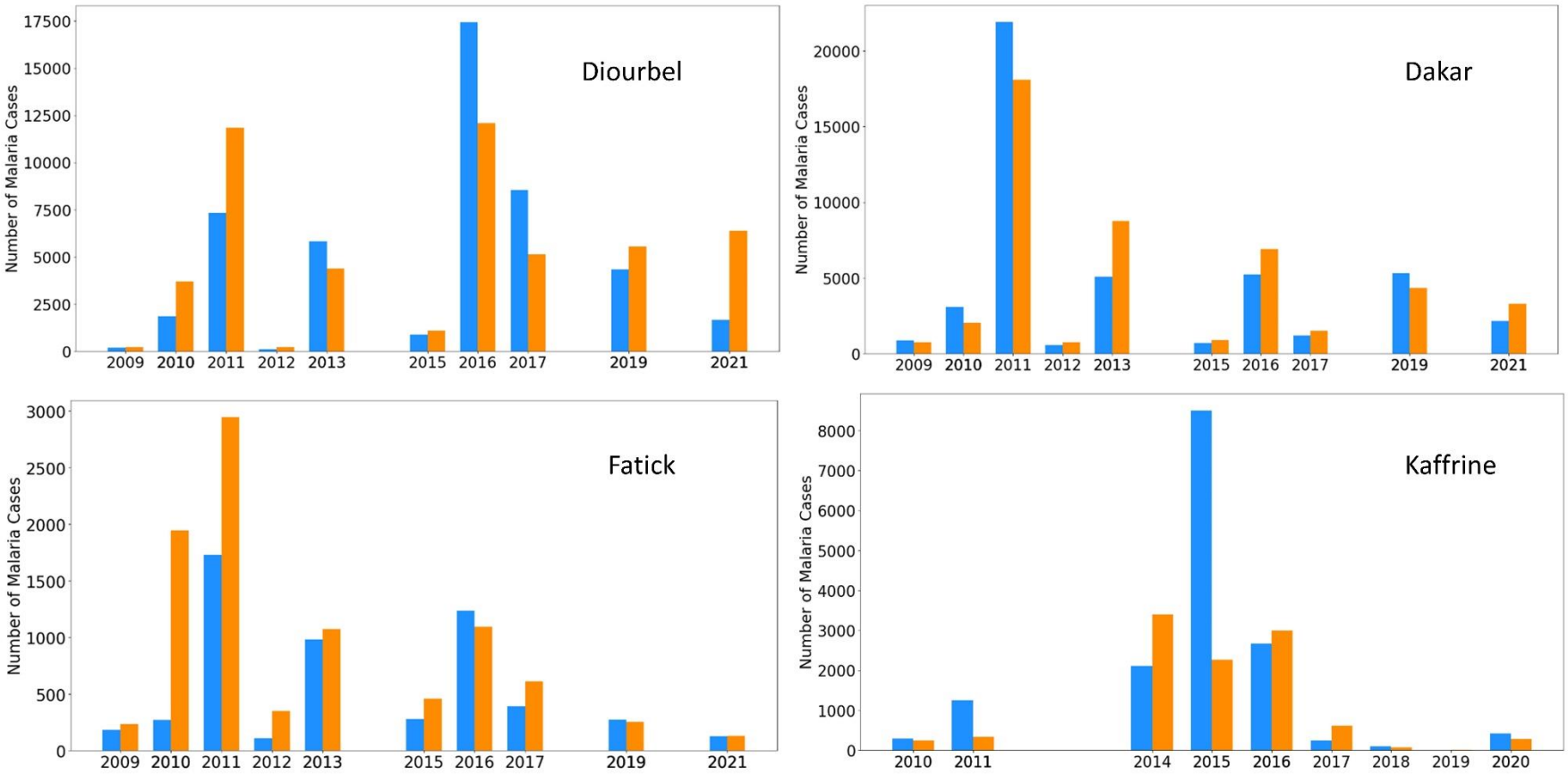
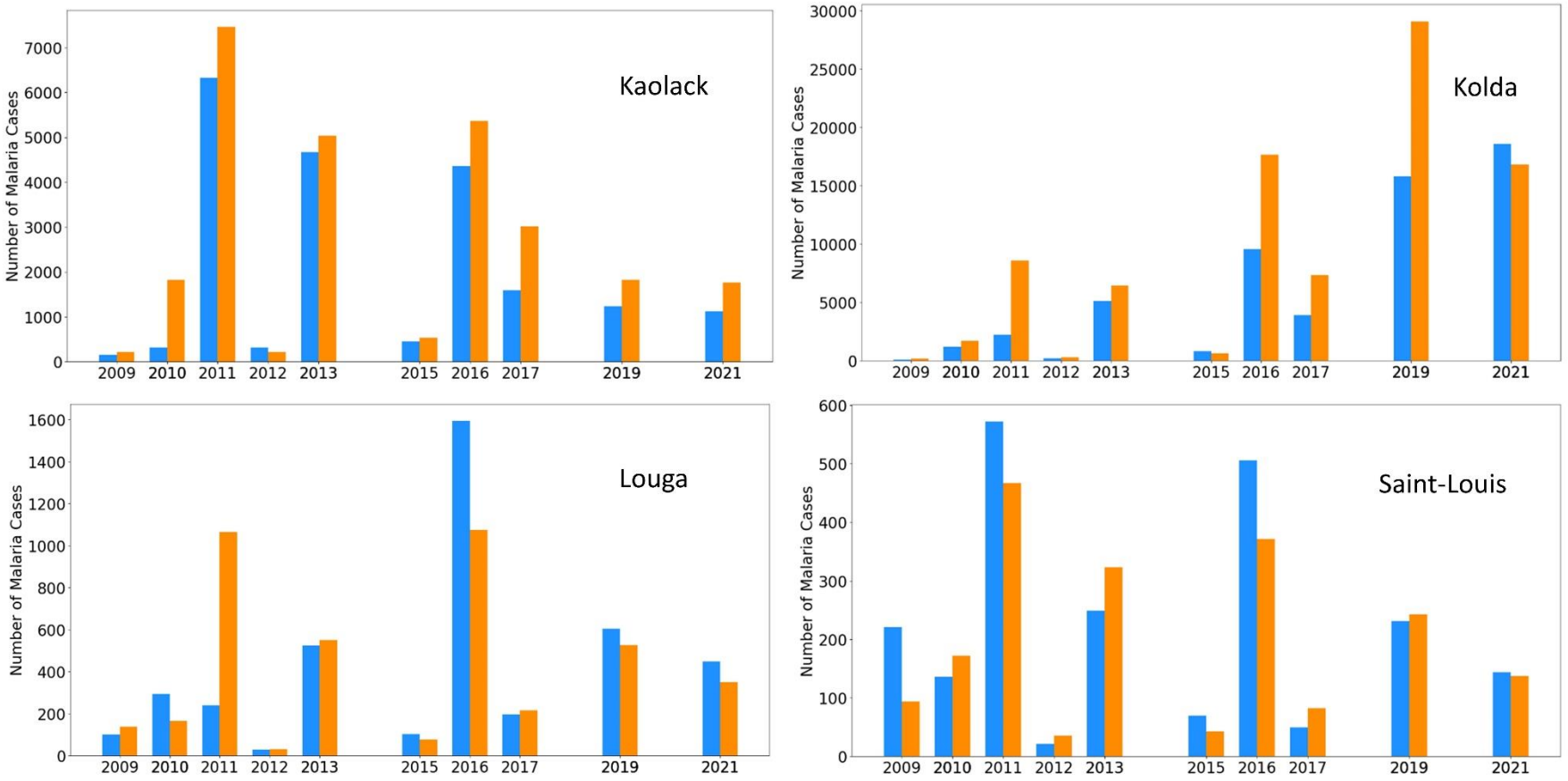
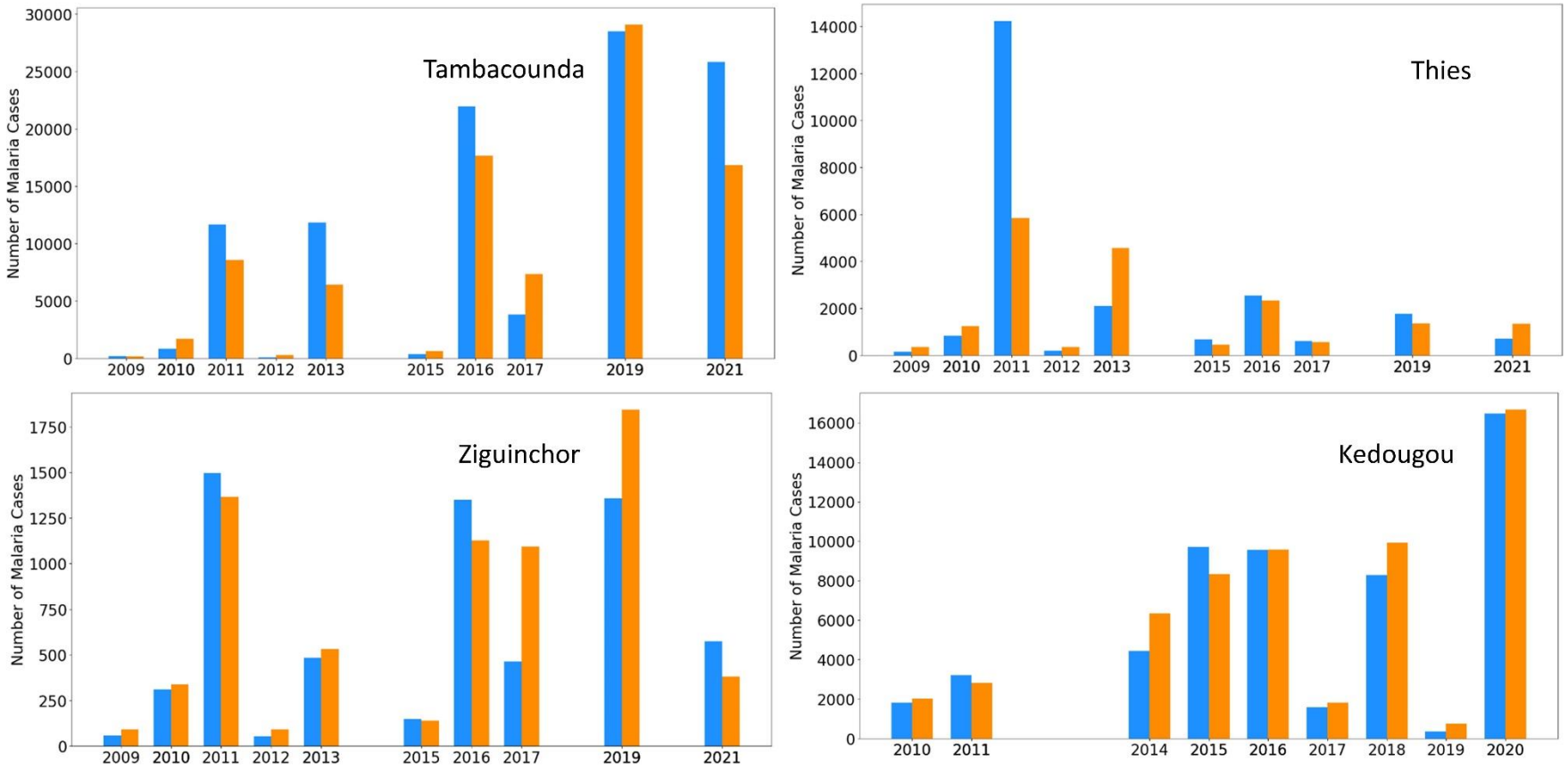


Figure 2: Province-wise results of malaria cases of the Year-wise Split model predictions for Senegal.


Diourbel
Number of Malaria Cases
17500
15000
12500
10000
7500
5000
2500
0
2009 2010 2011 2012 2013 2015 2016 2017 2019 2021
Dakar
Number of Malaria Cases
20000
15000
10000
5000
0
2009 2010 2011 2012 2013 2015 2016 2017 2019 2021
Fatick
Number of Malaria Cases
3000
2500
2000
1500
1000
500
0
2009 2010 2011 2012 2013 2015 2016 2017 2019 2021
Kaffrine
Number of Malaria Cases
8000
7000
6000
5000
4000
3000
2000
1000
0
2010 2011 2014 2015 2016 2017 2018 2019 2020


Kaolack
Kolda
Louga
Saint-Louis
Number of Malaria Cases
2009
2010
2011
2012
2013
2015
2016
2017
2019
2021


Tambacounda
Number of Malaria Cases
0
5000
10000
15000
20000
25000
30000
2009
2010
2011
2012
2013
2015
2016
2017
2019
2021
Thies
Number of Malaria Cases
0
2000
4000
6000
8000
10000
12000
14000
2009
2010
2011
2012
2013
2015
2016
2017
2019
2021
Ziguinchor
Number of Malaria Cases
0
250
500
750
1000
1250
1500
1750
2009
2010
2011
2012
2013
2015
2016
2017
2019
2021
Kedougou
Number of Malaria Cases
0
2000
4000
6000
8000
10000
12000
14000
16000
2010
2011
2014
2015
2016
2017
2018
2019
2020

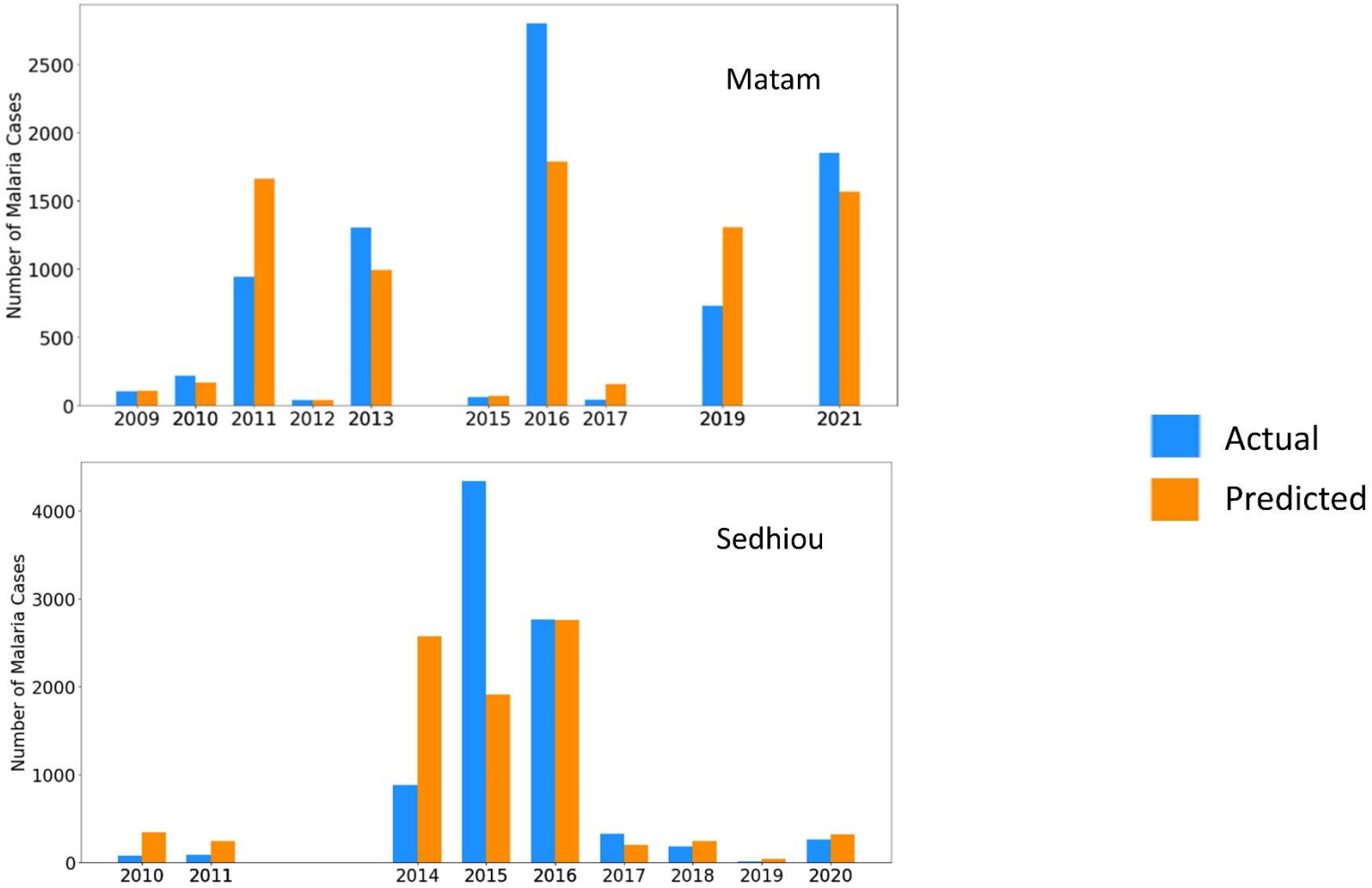


Figure 3: Province-wise results of malaria cases of the Random Split model predictions for Senegal.

7. Model Results And Data Interpretation

Interpreting data on rainfall-induced malaria prevalence requires a nuanced understanding of the relationship between precipitation patterns and mosquito breeding habitats (Figure 2 and Figure 3). Rainfall creates suitable breeding grounds for the Anopheles mosquitoes – the primary malaria vectors – in the form of puddles, waterlogged areas, and freshwater collections. As such, an increase in rainfall can often lead to a surge in mosquito populations and, subsequently, a rise in malaria transmission. However, it is crucial to differentiate between moderate rains, which provide ideal breeding grounds, and heavy downpours, which can wash away larval habitats. Moreover, the latency between increased rainfall and a subsequent rise in malaria cases – typically around one to two months – must be factored into data interpretations. It is also essential to consider the local ecosystem, infrastructure, and interventions in place. For instance, effective water management or rapid response mechanisms can mitigate the impact of increased rainfall on malaria transmission. When examining graphs or datasets, sharp spikes in malaria cases following periods of consistent or increased rainfall might be observed. However, the interpretation must be contextual, considering regional specificities, existing health infrastructure, and any other external interventions or events. The predictive modeling results are displayed for the next three months, considering historical rainfall and malaria prevalence trends (Figure 4).

8. Lag Time calculation

To calculate the lag time for the first month, the initial rows were removed from the malaria dataset in the region, and the last row was excluded from the rainfall dataset. After this, the dataset was split into training and testing sets, and a specific Random Forest regression model was developed and saved.

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score
import pickle

malariaSet_1 = malaria_data['DIOURBEL']
otherSet_1 = Rainfall_data[['YEAR','MON','DIOURBEL']]

print(malariaSet_1)
print(otherSet_1)

malariaSet_1 = malariaSet_1.drop(labels=0, axis=0)
otherSet_1 = otherSet_1.drop(labels=155, axis=0)

print(malariaSet_1)
print(otherSet_1)


X_region_test = otherSet_1[otherSet_1['YEAR']>=2018]
X_region_train = otherSet_1[otherSet_1['YEAR']<2018]

print(X_region_test)
print(X_region_train)

Y_region_train = malariaSet_1.iloc[:len(X_region_train.MON)]
Y_region_test = malariaSet_1.iloc[len(X_region_train.MON):]

print(Y_region_train)
print(Y_region_test)

model_1 = RandomForestRegressor(n_estimators = 19, random_state = 16,
max_depth = 9, criterion ='absolute_error',min_samples_split = 5,
                                min_samples_leaf = 20, max_features =
'log2').fit(X_region_train,Y_region_train).fit(X_region_train,Y_region
_train)
acc = model_1.score(X_region_test,Y_region_test)
print('First LagTime Accuracy :',acc)

pickle.dump(model_1, open('DIOURBEL/1_month_model.pkl','wb'))


print('First Malaria Future Prediction',model_1.predict(df))
```

To account for a two-month delay, first exclude the first and second rows from the malaria data set in the region, and next, exclude the last two rows from the rainfall data set. Finally, the data set is split for training and testing, and a specific Random Forest regression model is developed.

```
malariaSet_2 = malaria_data['DIOURBEL']
otherSet_2 = Rainfall_data[['YEAR','MON','DIOURBEL']]

print(malariaSet_2)
print(otherSet_2)
```

To account for a third-month lag time, remove the first, second, and third rows of the malaria dataset and the last three rows of the rainfall dataset. Then, the dataset was split for training and testing purposes, and a specific Random Forest regression model was further developed.

```
malariaSet_3 = malaria_data['DIOURBEL']
otherSet_3 = Rainfall_data[['YEAR','MON','DIOURBEL']]

print(malariaSet_3)
print(otherSet_3)


malariaSet_3 = malariaSet_3.drop([0,1,2], axis=0)
otherSet_3 = otherSet_3.drop([153,154,155], axis=0)

print(malariaSet_3)
print(otherSet_3)

X_region_test = otherSet_3[otherSet_3['YEAR']>=2018]
X_region_train = otherSet_3[otherSet_3['YEAR']<2018]

print(X_region_test)
print(X_region_train)

Y_region_train = malariaSet_3.iloc[:len(X_region_train.MON)]
Y_region_test = malariaSet_3.iloc[len(X_region_train.MON):]

print(Y_region_train)
print(Y_region_test)

model_3 = RandomForestRegressor(n_estimators = 19, random_state = 13,
max_depth = 1, criterion ='absolute_error',min_samples_split = 7,
                                min_samples_leaf = 12, max_features =
None).fit(X_region_train,Y_region_train)
b = model_3.score(X_region_test,Y_region_test)
print('Third LagTime Accuracy :',b)

pickle.dump(model_3, open('DIOURBEL/3_month_model.pkl','wb'))
print('Third Malaria Future Prediction',model_3.predict(df))
```

Saving model PKL files of monthly lag times:

```
def LoadPklFile(Year, Month,Rainfall):
    input_data = [[Year, Month, Rainfall]]
    with open('DIOURBEL/1_month_model.pkl', 'rb') as file:
        loaded_1_Month_Model = pickle.load(file)
        predictions_1 = loaded_1_Month_Model.predict(input_data)
        print('1 Month Malaria Cases :'+str(predictions_1[0]))
    with open('DIOURBEL/2_month_model.pkl', 'rb') as file:
        loaded_2_Month_Model = pickle.load(file)


        predictions_2 = loaded_2_Month_Model.predict(input_data)
        print('2 Month Malaria Cases :'+str(predictions_2[0]))
    with open('DIOURBEL/3_month_model.pkl', 'rb') as file:
        loaded_1_Month_Model = pickle.load(file)
        predictions_3 = loaded_1_Month_Model.predict(input_data)
        print('3 Month Malaria Cases :'+str(predictions_3[0]))

LoadPklFile(2022,1,2.35)
```

Future Prediction Model Results using Lag time for Diourbel region:

```
#results
1 Month Malaria Cases :4330.95
2 Month Malaria Cases :1477.61
3 Month Malaria Cases :1264.54
```

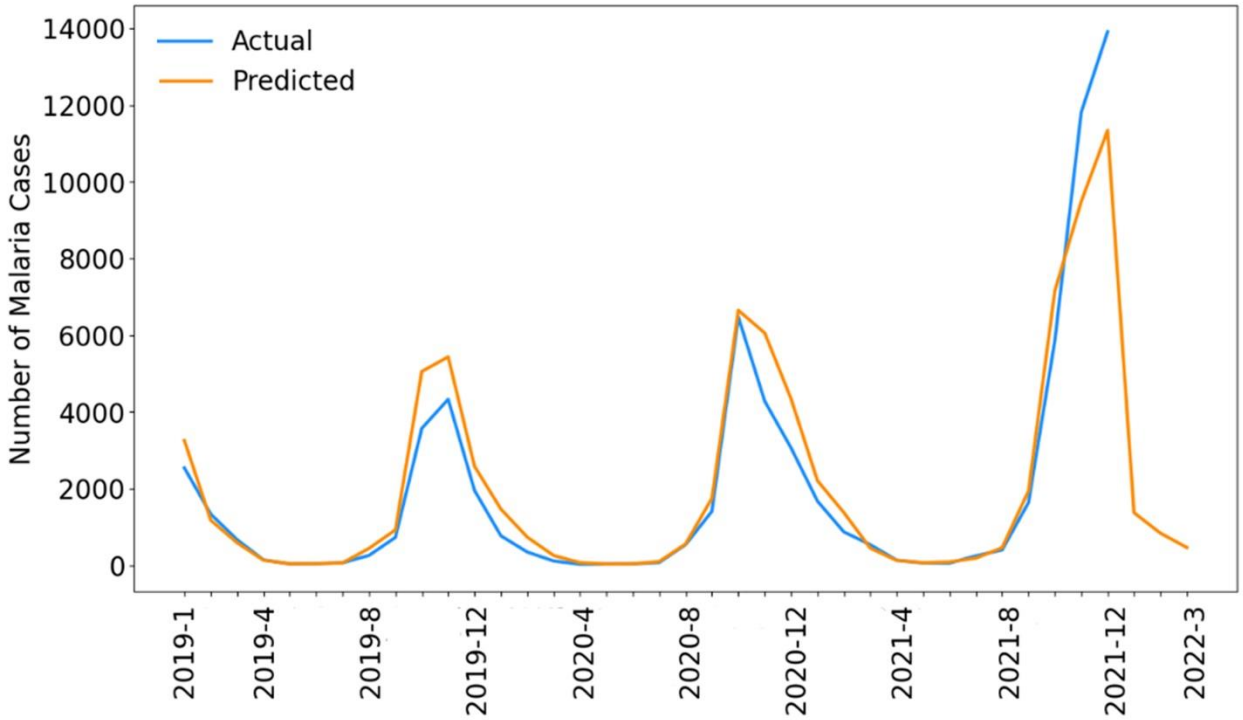


Figure 4: Three-month forecast of malaria prevalence for the Diourbel region.

9. Significance and Future Perspectives

Most predictive models employ limited variables, often overlooking socioeconomic indicators that can significantly enhance prediction accuracy. Our research seeks to transform the fight against malaria by establishing a platform that integrates clinical, meteorological, and ecological variables. This merger will create a robust data ecosystem for model development, strengthening malaria prevention, diagnosis, and treatment (Fletcher et al., 2022; Samarasekera, 2023). Future research should aim to develop predictive models for every region using multiparametric datasets to evaluate malaria prevalence and develop accurate prevention and control measures (Figure 5).

The application of climate data in predicting and managing malaria prevalence has significant implications for public health and disease control. Climatic variables such as rainfall increasingly influence malaria transmission, and the information presented can be used to evaluate early warning signals. This data displays potential trends and predicted results of potential outbreaks based on rainfall, allowing for timely interventions and resource allocations. Understanding the potential impacts of climate on malaria can guide the allocation of resources like bed nets, antimalarial drugs, and diagnostics to regions most likely to be affected during certain climatic conditions and inform public awareness campaigns. Climate data can also guide environmental interventions, such as creating better water drainage systems in vulnerable areas and habitat management, such as introducing larvivorous fish in stagnant waters or applying larvicides to control mosquito breeding. Insights derived from climate-malaria linkages can inform national and regional policies for health, environment, and urban planning, ensuring a holistic approach to health and development considering the future challenges of climate change. Collaborative research between climate and health researchers can lead to the development of innovative solutions and strategies tailored to the unique climatic conditions of different regions. Incorporating climate data into malaria management strategies represents a convergence of environmental science and public health, underscoring the need for a multidisciplinary approach to address global health challenges in an era of rapid climate change.

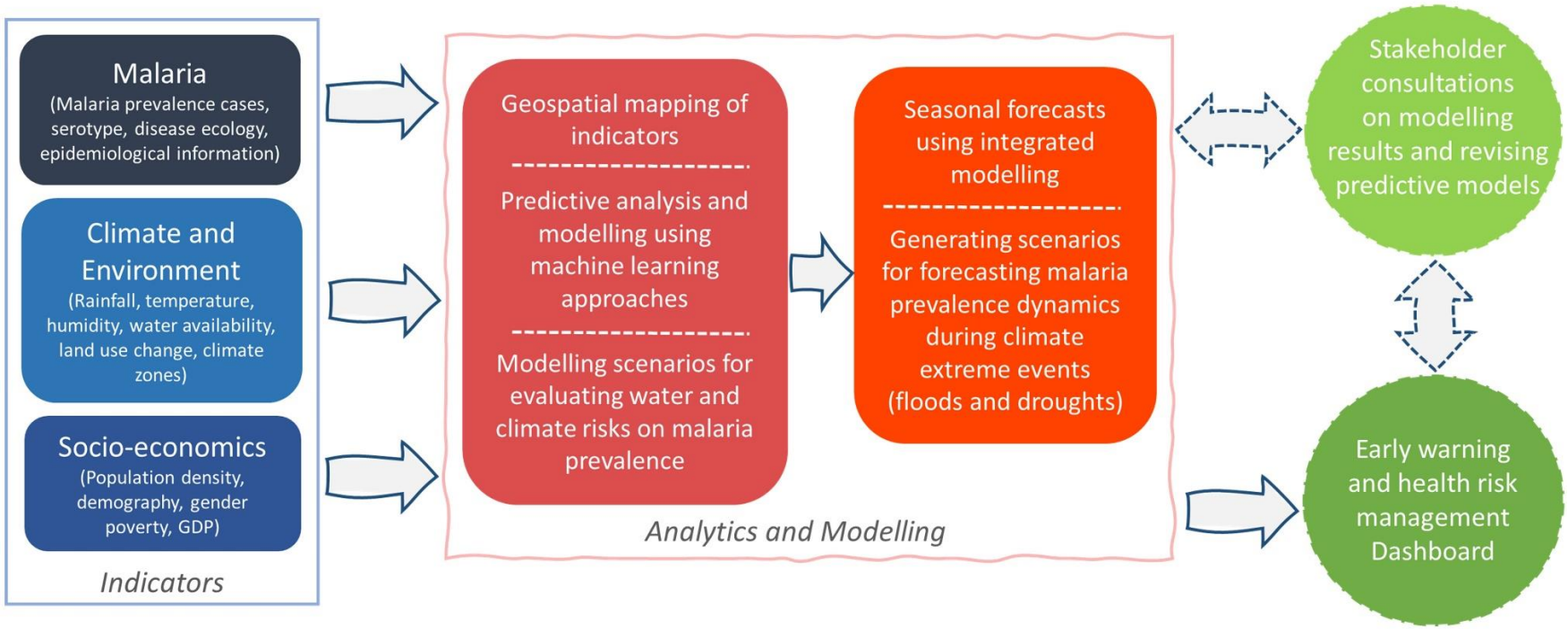


Figure 5: Comprehensive methodological framework for developing multiparametric climate-induced predictive models for malaria prevalence dynamics. Source: Authors.

10. Conclusions and Recommendations

The objective of the research was to create a prediction model for malaria cases in different regions of Senegal using the machine learning Random Forest Algorithm based on data related to malaria and rainfall. The study evaluated the model's performance for various splits across the regions and drew necessary conclusions.

- The study discovered that the model's performance was better when trained using random splits compared to year-wise splits for most regions, suggesting that a random split trained model can produce more accurate predictions of future malaria cases.
- Additionally, the prediction model's accuracy was good (> 0.9) for regions in Senegal with a higher malaria prevalence, indicating that the model can help predict future malaria cases in these regions.
- Lastly, the study predicted malaria cases for the next three months with a one-month lag time. This prediction can help stakeholders generate early warning signals and take timely and effective preventive measures to control the spread of malaria.
- In summary, this research shows the potential of machine learning algorithms in predicting malaria cases and providing early warning signals to policymakers.

References

Bationo, C.S., Gaudart, J., Dieng, S., Cissoko, M., Taconet, P., Ouedraogo, B., Somé, A., Zongo, I., Soma, D.D., Tougri, G., Dabiré, R.K., Koffi, A., Pennetier, C., Moiroux, N., 2021. Spatio-temporal analysis and prediction of malaria cases using remote sensing meteorological data in Diébougou health district, Burkina Faso, 2016–2017. *Sci Rep* 11, 20027. https://doi.org/10.1038/s41598-021-99457-9

Bicout, D.J., Vautrin, M., Vignolles, C., Sabatier, P., 2015. Modeling the dynamics of mosquito breeding sites vs rainfall in Barkedji area, Senegal. *Ecological Modelling* 317, 41–49. https://doi.org/10.1016/j.ecolmodel.2015.08.027

Coalson, J.E., Anderson, E.J., Santos, E.M., Madera, G.V., Romine, J.K., Luzingu, J.K., Dominguez, B., Richard, D.M., Little, A.C., Hayden, M.H., Ernst, K.C., 2021. The Complex Epidemiological

Relationship between Flooding Events and Human Outbreaks of Mosquito-Borne Diseases: A Scoping Review. *Environmental Health Perspectives* 129, 096002. https://doi.org/10.1289/EHP8887

Dieng, S., Ba, E.H., Cissé, B., Sallah, K., Guindo, A., Ouedraogo, B., Piarroux, M., Rebaudet, S., Piarroux, R., Landier, J., Sokhna, C., Gaudart, J., 2020. Spatio-temporal variation of malaria hotspots in Central Senegal, 2008–2012. *BMC Infectious Diseases* 20, 424. https://doi.org/10.1186/s12879-020-05145-w

Diouf, I., Deme, A., Ndione, J.-A., Gaye, A.T., Rodríguez-Fonseca, B., Cissé, M., 2013. Climate and health: Observation and modeling of malaria in the Ferlo (Senegal). *Comptes Rendus Biologies*, Sciences, enseignement et technologie pour le développement de l'Afrique 336, 253–260. https://doi.org/10.1016/j.crvi.2013.04.001

Diouf, I., Rodriguez-Fonseca, B., Deme, A., Caminade, C., Morse, A.P., Cisse, M., Sy, I., Dia, I., Ermert, V., Ndione, J.-A., Gaye, A.T., 2017. Comparison of Malaria Simulations Driven by Meteorological Observations and Reanalysis Products in Senegal. *International Journal of Environmental Research and Public Health* 14, 1119. https://doi.org/10.3390/ijerph14101119

Fletcher, I.K., Grillet, M.E., Moreno, J.E., Drakeley, C., Hernández-Villena, J., Jones, K.E., Lowe, R., 2022. Synergies between environmental degradation and climate variation on malaria re-emergence in southern Venezuela: a spatiotemporal modelling study. *The Lancet Planetary Health* 6, e739–e748. https://doi.org/10.1016/S2542-5196(22)00192-9

Jampani, M., Panjwani, S., Ghosh, S., Sambou, M.H.A., Amarnath, G., 2023. Climate variability and extremes impact on seasonal occurrence patterns of malaria cases in Senegal. American Geophysical Union (AGU), Chapman Conference, Washington, D. C., USA, 12-15 June 2023.

Kibret, S., Lautze, J., McCartney, M., Wilson, G.G., Nhamo, L., 2015. Malaria impact of large dams in sub-Saharan Africa: maps, estimates and predictions. Malaria Journal 14, 339. https://doi.org/10.1186/s12936-015-0873-2\

Kibret, S., McCartney, M., Lautze, J., Nhamo, L., Yan, G., 2021. The impact of large and small dams on malaria transmission in four basins in Africa. Sci Rep 11, 13355. https://doi.org/10.1038/s41598-021-92924-3\

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. https://doi.org/10.1186/s12992-021-00793-2

Lucas, T.C.D., Nandi, A.K., Chestnutt, E.G., Twohig, K.A., Keddie, S.H., Collins, E.L., Howes, R.E., Nguyen, M., Rumisha, S.F., Python, A., Arambepola, R., Bertozzi-Villa, A., Hancock, P., Amratia,

P., Battle, K.E., Cameron, E., Gething, P.W., Weiss, D.J., 2021. Mapping malaria by sharing spatial information between incidence and prevalence data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 70, 733–749. https://doi.org/10.1111/rssc.12484

Mouchet, J., Faye, O., Julvez, J., Manguin, S., 1996. Drought and malaria retreat in the Sahel, West Africa. *The Lancet* 348, 1735–1736. https://doi.org/10.1016/S0140-6736(05)65860-6

Mutero, C., Amerasinghe, F., Boelee, E., Konradsen, F., Van der Hoek, W.; Nevondo, T., and F. Rijsberman. 2005. Systemwide Initiative on Malaria and Agriculture: An Innovative Framework for Research and Capacity Building. EcoHealth. 2. 11-16. 10.1007/s10393-004-0088-4.

Ndiath, M.O., Sarr, J.-B., Gaayeb, L., Mazenot, C., Sougoufara, S., Konate, L., Remoue, F., Hermann, E., Trape, J., Riveau, G., Sokhna, C., 2012. Low and seasonal malaria transmission in the middle Senegal River basin: identification and characteristics of Anopheles vectors. *Parasites & Vectors* 5, 21. https://doi.org/10.1186/1756-3305-5-21

Sallah, K., Giorgi, R., Ba, E.-H., Piarroux, M., Piarroux, R., Cisse, B., Gaudart, J., 2021. Targeting Malaria Hotspots to Reduce Transmission Incidence in Senegal. *International Journal of Environmental Research and Public Health* 18, 76. https://doi.org/10.3390/ijerph18010076

Samarasekera, U., 2023. Climate change and malaria: predictions becoming reality. *The Lancet* 402, 361–362. https://doi.org/10.1016/S0140-6736(23)01569-6

Santos-Vega, M., Martinez, P.P., Vaishnav, K.G., Kohli, V., Desai, V., Bouma, M.J., Pascual, M., 2022. The neglected role of relative humidity in the interannual variability of urban malaria in Indian cities. *Nat Commun* 13, 533. https://doi.org/10.1038/s41467-022-28145-7

Teklu, B.M., Tekie, H., McCartney, M., Kibret, S., 2010. The effect of physical water quality and water level changes on the occurrence and density of Anopheles mosquito larvae around the shoreline of the Koka reservoir, central Ethiopia. Hydrology and Earth System Sciences 14, 2595-2603. https://doi.org/10.5194/hess-14-2595-2010\

Wang, Z., Liu, Y., Li, Yapin, Wang, G., Lourenço, J., Kraemer, M., He, Q., Cazelles, B., Li, Yidan, Wang, R., Gao, D., Li, Yuchun, Song, W., Sun, D., Dong, L., Pybus, O.G., Stenseth, N.C., Tian, H., 2022. The relationship between rising temperatures and malaria incidence in Hainan, China, from 1984 to 2010: a longitudinal cohort study. *The Lancet Planetary Health* 6, e350–e358. https://doi.org/10.1016/S2542-5196(22)00039-0