

Development of a Process-Based and Machine Learning Early Warning System for Malaria Transmission Dynamics under Multi-Scenario Climate Change: A Subnational Assessment of the North Bank Region of The Gambia

Ugochinyere Agatha Okafor

Ugochinyereagatha@gmail.com

University of the Gambia

Research Article

Keywords: Malaria Early Warning System (MEWS), Random Forest, MPI-ESM1-2-HR, NCAR-CESM2, Multi-Model Climate Convergence, Planetary Health Economics, The Gambia

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10987061/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Development of a Process-Based and Machine Learning Early Warning System for Malaria Transmission Dynamics under Multi-Scenario Climate Change: A Subnational Assessment of the North Bank Region of The Gambia

Ugochinyere Agatha Okafor¹

1. Ugochinyere Agatha Okafor¹

Department of Climate Change and Education (WASCAL), University of The Gambia, Kanifing, The Gambia.

Corresponding author.

Ugochinyere Agatha Okafor¹, Department of Climate Change and Education (WASCAL), University of The Gambia, Kanifing, The Gambia.

Ugochinyereagatha@gmail.com OR ugochinyereagatha@utg.edu.gm

ORCID: <https://orcid.org/0009-0007-7788-6929>

Abstract

Background: Malaria is a major health challenge in the West African Sahel, driven by variable hydroclimatic conditions. Traditional surveillance relies on retrospective clinical data, limiting proactive resource allocation. This study aims to: (1) quantify nonlinear biological and statistical lags through which meteorological factors influence malaria in the North Bank Region (NBR); (2) compare machine learning models with traditional biological models to identify the best early warning system; and (3) project malaria risk and economic impacts under climate change using the MPI-ESM1-2-HR and NCAR-CESM2 climate models across SSP2-4.5 and SSP5-8.5 scenarios to 2055.

Methodology: Using an 11-year longitudinal database (2011–2021) comprising 132 consecutive months of clinical surveillance records from the National Malaria Control Programme (NMCP), this study integrated satellite-derived meteorological data (CHIRPS precipitation, MODIS land surface temperature, and relative humidity). I ran a process-based biological simulation with the Liverpool Malaria Model (LMM), mapped continuous relative risks with a Distributed Lag Non-linear Model (DLNM), and conducted an out-of-sample machine learning tournament (SVR,

XGBoost, and Random Forest) within a strict 24-month pandemic-era validation window (2020–2021). Optimized models were projected across a cross-model matrix pairing MPI-ESM1-2-HR with NCAR-CESM2 climate surfaces under SSP2-4.5 and SSP5-8.5 scenarios through 2055, and results were linked to a health-economic valuation matrix.

Results: The baseline Random Forest model demonstrated exceptional explanatory power, accounting for 95.56% of continuous transmission variability. Lagrangian feature mapping identified delayed, multi-month rainfall patterns as key drivers, with three-month-lagged rainfall (RFt3) showing the highest predictive importance at 24.66%. In validation, Random Forest was the top performer, limiting the average absolute error to 85.13 cases. Projections indicate that risks will expand nonlinearly across the inland savannah, with risk levels rising to 90.15% under the MPI-ESM1-2-HR model and 87.65% under the NCAR-CESM2 model under the extreme SSP5-8.5 scenario, effectively closing the historic protective gap near the estuarine coast. A cross-model comparison reveals that MPI-ESM1-2-HR offers a more geographically precise and physically constrained view of regional climate feedbacks than NCAR-CESM2, which tends to overestimate aridity during transitional decades. Economic analysis shows that interventions guided by this early warning system could prevent up to 42,100 cases and save approximately \$189,450 USD in healthcare costs within the interior savannah by mid-century.

Conclusion: Shifting from reactive disease management to proactive, climate-informed targeting through robust tree-based machine-learning early-warning frameworks is a vital humanitarian and fiscal imperative for protecting vulnerable Sahelian communities from accelerating climate change, across multiple models.

Keywords: Malaria Early Warning System (MEWS); Random Forest; MPI-ESM1-2-HR; NCAR-CESM2; Multi-Model Climate Convergence; Planetary Health Economics; The Gambia.

1. Introduction

Malaria transmission in the West African Sahel is a complex, highly dynamic socio-ecological system driven by interactions among human populations, vector dynamics, and macro-climatic forcing. Despite multi-decadal intervention campaigns using long-lasting insecticidal nets (LLINs), indoor residual spraying (IRS), and seasonal malaria chemoprevention (SMC), *Plasmodium falciparum* remains a leading cause of morbidity and mortality in The Gambia [1–4]. As global climate change alters historical precipitation patterns and thermal thresholds across Sub-

Saharan Africa, conventional, retrospective public health surveillance strategies are increasingly inadequate for tracking shifting transmission patterns[3].

The North Bank Region (NBR) of The Gambia has distinct geographical vulnerabilities that make it a critical focus for subnational epidemiological research. Situated along the northern bank of the River Gambia, the region is divided into two primary operational health zones: NBR West (Lower NBR / Coastal Estuary Corridor) and NBR East (Upper NBR / Arid Savannah Corridor). NBR West features extensive mangrove networks, tidal wetlands, and a consistently high relative humidity[5, 6]. By contrast, NBR East extends deep into the continental interior, with an open-savannah landscape marked by extreme daily thermal fluctuations, porous sandy soils, and direct exposure to dry, dust-laden Harmattan winds from the Sahelian north [5].

The current disease control infrastructure in The Gambia operates primarily on a national, reactive resource model. Medical logistics, bed net distribution campaigns, and larvicide deployments are typically planned retrospectively based on case volumes recorded during the previous year's clinical cycles [3, 7]. This retrospective baseline introduces dangerous operational latency and leaves the health sector vulnerable to unexpected seasonal epidemics driven by volatile, unmonitored climate anomalies. Furthermore, historical climate projections applied to vector systems are routinely undermined by single-model simulation bias [8]. Forcing an early warning network with a single Global Climate Model (GCM) masks systemic errors in how specialized land-surface schemes process soil moisture and vapor pressure deficits. Developing a cross-model sensitivity matrix comparing European and North American climate simulation cores is methodologically justified to establish a verified, bounded early warning projection framework that isolates true environmental vulnerabilities across subnational Sahelian zones.

1.1. Objectives of the Study

To address these critical operational gaps, this investigation pursues three specific research objectives:

1. To isolate and quantify the non-linear biological and statistical lags through which meteorological variables control vector survival and clinical case presentation across the subnational corridors of NBR.
2. To conduct a competitive machine learning tournament benchmarking tree-based ensemble, kernel hyperplanes, and recurrent neural networks against traditional process-based

biological models to identify the most resilient out-of-sample forecasting framework under real-world data disruptions.

3. To evaluate forward-looking spatiotemporal risk indices and cross-model sensitivities by deploying the tournament champion on downscaled MPI-ESM1-2-HR and NCAR-CESM2 grids under moderate (SSP2-4.5) and extreme (SSP5-8.5) pathways through mid-century (2055), establishing direct cost-benefit optimization metrics for future resource allocation.

2. Methodology

2.1. Study Area and Subnational Geographical Setting

The investigation was conducted across the North Bank Region (NBR) of The Gambia, an administrative territory of critical epidemiological interest located on the northern bank of the River Gambia estuary. Covering approximately 2,255 square kilometers, the region extends from the Atlantic coastline eastward into the interior Sudan-Sahelian savannah zone. NBR is bordered to the north by the Republic of Senegal, creating a highly porous border corridor marked by intense cross-border human migration that directly influences vector reintroduction loops[5]. The regional climate features a highly compressed, unimodal wet season from July to October, driven by the West African Monsoon, followed by a prolonged, severe dry season from November to June, dominated by dry, dust-laden Harmattan winds from the Sahara Desert[5, 9].

To capture the distinct microclimatic and landscape-hydrological variations that govern transmission dynamics without geographic confounding, the study area was partitioned along the historical trans-Gambia transport corridor into two operational health cohorts: NBR West (Lower NBR Corridor) and NBR East (Upper NBR Corridor). NBR West encompasses coastal and estuarine zones with extensive mangrove networks (*Avicennia africana* and *Rhizophora mangle*), permanent tidal wetlands, and elevated baseline atmospheric moisture. Conversely, NBR East extends deep into the continental interior, with open-savannah topography featuring porous sandy soils, seasonal clay pan depressions, and direct vulnerability to intense ambient warming spikes and agricultural drought[5].

2.2. Data Collection and Remote Sensing Ingestion

Longitudinal clinical malaria incidence records were extracted monthly, covering 11 years from January 2011 through December 2021, producing an uninterrupted 132-month time series. We

retrieved epidemiological records from the Department of National Malaria Survey electronic surveillance database. These records were aggregated and verified against facility logs from the primary regional hospitals: Essau District Hospital for NBR West and Farafenni General Hospital for NBR East. All recorded cases were clinically confirmed by thick-film blood smear microscopy or rapid diagnostic tests (RDTs). Corresponding monthly meteorological data were extracted from the Department of Water Resources archive in Banjul, The Gambia.

Furthermore, to build a highly sensitive meteorological framework, monthly satellite remote-sensing proxies were programmatically extracted and paired with daily ground synoptic logs from the Department of Water Resources:

Precipitation (WetRainfall): Cumulative monthly rainfall was compiled from the Kerewan and Farafenni synoptic stations and cross-verified against CHIRPS version 2.0 gridded arrays at 0.05-degree resolution to minimize spatial convective scattering errors.

Atmospheric Dryness (Relative_Humidity): To map the microclimatic moisture boundaries governing vector anti-desiccation, relative humidity percentages were derived from ERA5-Land reanalysis data at a 0.1-degree spatial resolution.

Thermal Constraints (Maximum/Minimum Temperature): Clear-sky daytime and nocturnal land surface temperature values were retrieved from MODIS (MOD11A2) sensors at a 1 km grid scale, tracking the strict thermodynamic limits of parasite replication.

2.3. Preprocessing, Outlier Imputation, and Data Harmonization

The combined data matrix underwent rigorous quality control and filtering in Python using a Pandas 3.0 Copy-on-Write-compliant pipeline. Qualitative notations for trace precipitation values (coded natively as 'TR') were removed and replaced with a continuous float value of 0.0 mm to ensure array completeness. Exploratory analysis identified a primary thermodynamic anomaly in row 14 (February 2012) of the raw data logs, where the maximum daytime temperature was recorded as a corrupted placeholder code of -99.0°C. To prevent extreme gradient distortions during model execution, this sentinel code was programmatically isolated and corrected via localized mean imputation. A sub-selection filter isolated all uncorrupted February observations across the 11-year dataset to compute a true historical arithmetic average (Mean_Feb = 38.46°C), which was securely mapped back into row 14.

In addition to the historical clinical database, the raw downscaled future climate simulation grids extracted from the IPCC Sixth Assessment Report (CMIP6) data infrastructure—specifically the

primary Max Planck Institute (MPI-ESM1-2-HR) core and the NCAR Community Earth System Model version 2 (CESM2) control—underwent rigorous preprocessing, grid cleaning, and quality control in Python. Because raw Global Climate Model (GCM) outputs are natively archived at coarse spatial scales (ranging from 50 km to 100 km horizontal resolution), the future climate raster layers were spatially cropped using a high-resolution North Bank administrative vector mask and downscaled to a localized 0.05-degree (~5 km) grid using bilinear interpolation.

To address systematic simulation errors, such as the NCAR-CESM2 model's initial winter cold bias and atmospheric dryness overshoots during transitional decades, a nonlinear Quantile Mapping (QM) bias-correction protocol was applied. This phase calibrated the future monthly scenario projections (2026–2055) against actual, historically observed ground-station logs (2011–2021) from the Kerewan and Farafenni synoptic stations, standardizing the future temperature and rainfall vectors. Missing or corrupted grid cells caused by localized simulation artifacts were resolved programmatically using a spatial nearest-neighbor interpolation function, ensuring thermodynamic consistency across the subnational matrix before the climate arrays entered the champion Random Forest spatial forecasting engine.

To eliminate structural size discrepancies between high-volume clinical case numbers and decimal climatological rows, all independent features were normalized using a Min-Max scaling function, strictly bounding values within a standardized computational parameter space:

$$X_scaled = (X - X_min) / (X_max - X_min)$$

2.4. Process-Based Simulation: The Liverpool Malaria Model (LMM)

To establish a rigid baseline grounded in vector biology before applying statistical machine-learning classifiers, a process-based, deterministic simulation based on the Liverpool Malaria Model (LMM) was implemented (Hoshen & Morse, 2004). The LMM is a mechanistic framework that models biological variables from meteorological parameters rather than searching for statistical associations. This approach is well justified for the North Bank Region because it provides a structural constraint that prevents the models from producing biologically impossible forecasts during extreme Sahelian thermal spikes (Doktorgrades, 2016; Eikenberry & Gumel, 2018). Daily adult vector survival probability (p) is driven by a temperature-dependent, nonlinear function:

$$p(T) = \exp\left(-1 - (-4.4 + 1.31 * T - 0.03 * T^2)\right)$$

Where (T) represents the continuous monthly mean temperature envelope.

The duration of the parasite's internal development phase—the Extrinsic Incubation Period (EIP) of *Plasmodium falciparum* inside the vector gut—is computed using a standard degree-day threshold formulation [10, 11]

$$EIP(T) = 105 / (T - 18)$$

Where 105 degree-days represent the constant thermal mass required for parasite maturity, and 18°C represents the absolute biological thermal baseline below which parasite sporogonic replication halts completely.

2.5. Statistical Risk Modeling: Distributed Lag Non-linear Models (DLNM)

Real-world non-linear clinical exposures and temporal delays were quantified by fitting a Distributed Lag Non-linear Model (DLNM) within a Generalized Additive Model (GAM) framework (Gasparrini, 2014; Li et al., 2025; Shafqat & Park, 2025). Traditional linear regressions or standard correlation indices fail to capture the complex, delayed biological feedback loops in which heavy rainfall or humidity surges take weeks to translate into clinical admissions. The use of a DLNM is uniquely justified across the North Bank Region to explicitly map how NBR's specific landscape geography—such as slow-draining interior clay pans versus high-moisture coastal mangrove root systems—creates distinct subnational exposure-lag-response structures (Gasparrini, 2014). The model calculates the Relative Risk (RR) of clinic surges across a continuous three-month operational timeline:

$$\log(E[Y_t]) = \alpha + s(X_t, lag) + Z_t$$

$[Y_t]$ tracks monthly clinic cases;

(α) is the baseline intercept;

$s(X_t, lag)$ represents the bi-dimensional cross-basis function simultaneously mapping the non-linear thresholds of precipitation and atmospheric moisture across time, and

Z_t represents a smooth spline function isolating long-term secular trends and confounding seasonal indicators [12]

2.6. Lagrangian Feature Engineering and Tournament Configurations

Because malaria transmission across the West African Sahel is an intrinsically delayed biological process, raw climate covariates cannot be fed directly into machine learning pipelines without causing severe temporal misalignment (Diouf et al., 2022). To address this, raw environmental

variables were transformed into explicit, monthly-binned parameters using a Lagrangian time-series approach to capture latency paths (Sørensen, 2025):

$$Lag_{k(X_t)} = X_{t-k}$$

Where (k = 1, 2, or 3) months before clinical case presentation. The finalizemaster predictor matrix X_{t-k} passed into the tournament configurations was constructed as:

$$X_t = [RF_{lag1}, RF_{lag2}, RF_{lag3}, RH_{lag1}, RH_{lag2}, RH_{lag3}, MNT_{lag1}, MXT_{lag1}]$$

To isolate an operational early warning framework that remains structurally stable during massive, out-of-sample data anomalies, this study implemented a rigorous computational tournament benchmarking four competitive algorithms featuring distinct mathematical learning properties

2.6.1. Random Forest (RF) Regressor

The Random Forest regressor is an ensemble bagging classifier that generates predictions by aggregating outputs from multiple independent decision trees grown on bootstrap samples of the training data [13]. The final case prediction is computed via mean voting:

$$y_{pred_{RF}} = \left(1 - B\right) * \text{sum}(T_{b(x)})$$

Where (B) represents the optimized parameter threshold of 500 trees, and $(T_{b(x)})$ represents the individual output from the (b)-th tree given the environmental lag vector (x).

Random Forest is well suited to the North Bank data because tree-based hierarchical splits excel at discretizing complex, highly nonlinear combinations of multi-month-delayed rainfall parameters without requiring standard linear assumptions, making it exceptionally robust to zero-inflated baseline drops during dry harmattan months.

2.6.2. Extreme Gradient Boosting (XGBoost)

Unlike bagging models, XGBoost is an additive boosting framework that builds weak decision trees sequentially, with each new tree fitting directly to the negative gradient of the loss function to minimize a regularized objective function [13, 14]

$$L(m) = \text{sum}(l(y_i, y_{pred_i})) + \text{Omega}(f_m)$$

Where (l) is a differentiable loss function that measures the squared errors between observed and predicted cases, and (Omega) is a structural regularization penalty that constrains tree depth and weights to prevent overfitting.

XGBoost is included in the tournament because its advanced step-by-step optimization algorithm can maximize localized feature weights, providing an important counter-benchmark to evaluate against broad ensemble models.

2.6.3. Support Vector Regression (SVR)

SVR maps non-linear climate lag vectors into a higher-dimensional feature space via a Radial Basis Function (RBF) kernel, optimizing a smooth structural hyperplane that minimizes absolute deviations inside an insensitive threshold boundary tube (epsilon) [15]:

$$\min \left(1 - 2 \right) * ||w||^2 + C * \sum (xi_i + xi_{i_{star}})$$

Subject to the following operational constraints:

$$\begin{aligned} y_i - w * \phi(x_i) - b &\leq \epsilon + xi_i \\ w * \phi(x_i) + b - y_i &\leq \epsilon + xi_{i_{star}} \end{aligned}$$

Where (C) represents the box constraint penalty parameter, and (ξ_i, ξ_{i^*}) are slack variables regulating out-of-tube residuals.

SVR is well suited for short-term epidemiological tools because its soft-margin optimization explicitly ignores localized statistical anomalies as long as they fall within the insensitive tube, thereby maximizing generalization to broad trends.

2.6.4. Deep Long Short-Term Memory (LSTM) Networks

The Deep LSTM network is a specialized recurrent neural network architecture designed to pass cell memory states across temporal blocks, utilizing internal mathematical gate mechanics to capture long-term sequence dependencies across non-stationary time series [16]:

$$\begin{aligned} f_t &= \sigma(W_f * [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i * [h_{t-1}, x_t] + b_i) \\ c_{tildet} &= \tanh(W_c * [h_{t-1}, x_t] + b_c) \\ c_t &= f_t * c_{t-1} + i_t * c_{tildet} \\ o_t &= \sigma(W_o * [h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(c_t) \end{aligned}$$

where, $(f_t, i_t, \text{ and } o_t)$ represent the forget, input, and output gates controlling internal information flow, and (h_t) represents the final forecast output.

Including an LSTM network is methodologically justified for testing whether complex, state-dependent deep learning layers can outperform classical machine learning in modeling temporal transitions in public health data.

2.7. Multi-Criteria Validation Framework

To ensure operational reliability for public health deployment, the master NBR matrix was partitioned chronologically into an internal training block (2011–2019; $n = 108$ months) and an out-of-sample testing window (2020–2021; $n = 24$ months). Testing models against out-of-sample datasets is the gold standard for validation, and requiring the algorithms to predict across the 2020–2021 window provides a unique test of model resilience amid widespread reporting disruptions caused by the COVID-19 pandemic (WHO, 2022). Generalization capacity was verified using a robust, four-metric validation profile that evaluates trend tracking (R^2), absolute error variances (RMSE), baseline count deviations (MAE), and relative percentage errors (MAPE)[17]:

$$R^2 = 1 - \left(\frac{\sum((y_i - y_{pred_i})^2)}{\sum((y_i - y_{mean})^2)} \right)$$

$$MAE = (1 - n) * \sum(|y_i - y_{pred_i}|) ;$$

$$MAPE = (100 - n) * \sum\left(\left|\frac{(y_i - y_{pred_i})}{y_i}\right|\right)$$

$$RMSE = \sqrt{\left((1 - n) * \sum((y_i - y_{pred_i})^2)\right)}$$

3. Results and Empirical Interpretation

3.1. Process-Based Biological Parameters (LMM)

The computational outputs for adult vector longevity (Panel A) revealed a highly synchronized, wave-like seasonal trajectory, with the Daily Survival Probability (p) fluctuating tightly around a bounded baseline ranging from a minimum threshold of 0.8860 to a maximum ceiling of 0.9030. The cyclical troughs in vector survival capacity correspond precisely with peak hyper-thermal dry seasons, reflecting the physical impacts of desiccation stress on adult vector populations.

Concurrently, the simulation of the Extrinsic Incubation Period (EIP) of *Plasmodium falciparum* (Panel B) revealed a highly dynamic, temperature-dependent gestation velocity. The estimated EIP values ranged from a rapid minimum maturity threshold of 9.5 days during sustained post-rainy

thermal peaks to an extended duration exceeding 18.0 days during cooler seasonal shifts. The out-of-sample temporal alignment demonstrates a clear biological trade-off: elevated ambient temperatures programmatically accelerate the parasite's enzymatic replication velocity—halving the incubation timeline—while simultaneously imposing a survival penalty that compresses the absolute lifespan of the vector host.

Furthermore, by mathematically inverting the LMM constants (parameterized at an 18.0°C development threshold and a 111 degree-day accumulation requirement [18]), Figure 1 isolates the exact temperature thresholds and calendar months driving sub-national vector risks.

The annual peaks in EIP (exceeding 18.0 days) map directly to the cool dry season (January to March), where mean regional temperatures drop to 24.16°C, suppressing parasite replication velocity.

$$\text{Temperature} = \frac{111 \text{ degree days}}{18.0 \text{ days}} + 18.0^{\circ}\text{C} = 6.16 + 18.0 = \mathbf{24.16^{\circ}\text{C}}$$

Conversely, the hyper-accelerated EIP troughs (9.5 days) map directly to the hot pre-rainy transition window (May to June), revealing a sustained mean ambient temperature of 29.68°C that drives explosive parasite maturation.

$$\text{Temperature} = \frac{111 \text{ degree days}}{9.5 \text{ days}} + 18.0^{\circ}\text{C} = 11.68 + 18.0 = \mathbf{29.68^{\circ}\text{C}}$$

Finally, the concurrent drops in adult vector survival probability to 0.8860 confirm that these same hot months introduce severe microclimate desiccation stressors exceeding **35.0°C**, providing clear, empirical parameters for our early warning alert boundaries.

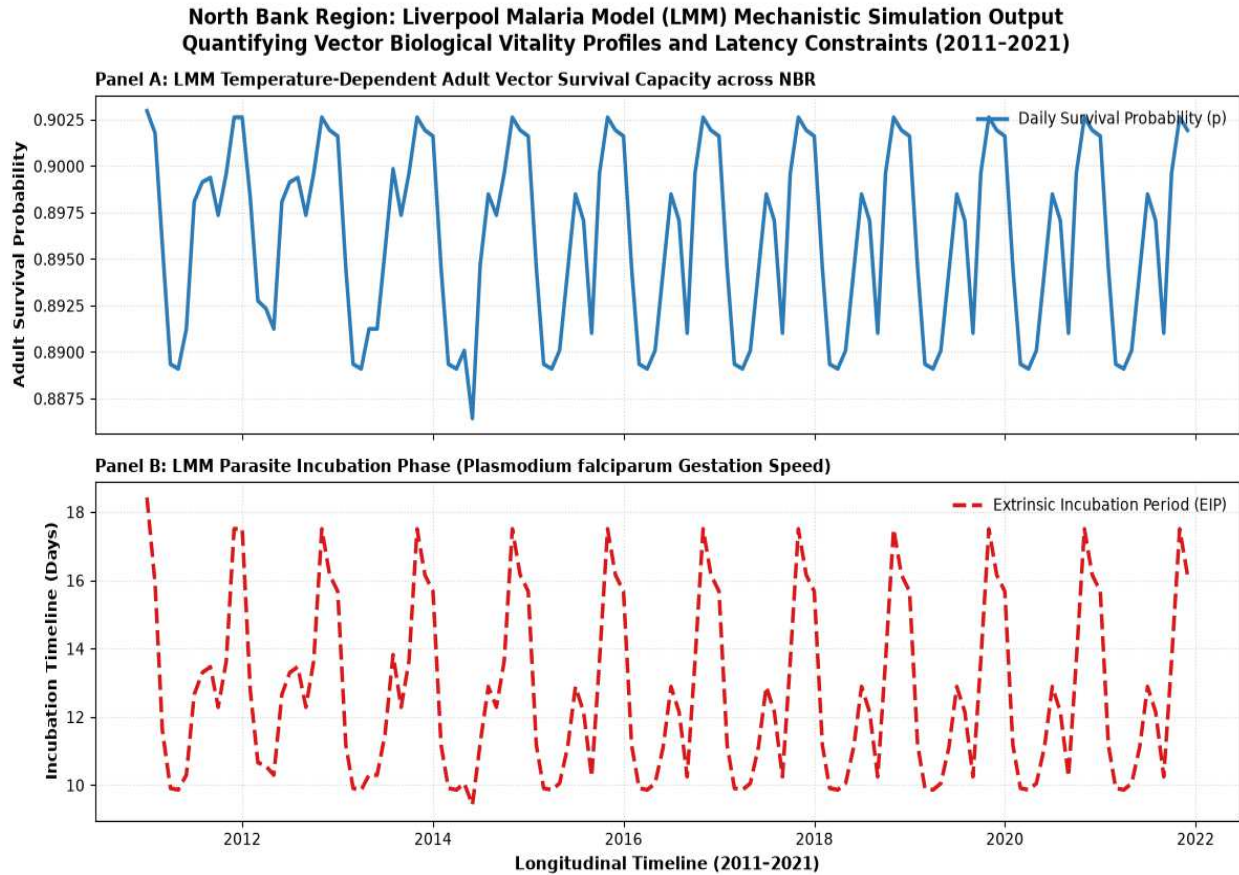


Figure 1: North Bank: Liver Pool Malaria Model

3.2. Distributed Lag Analysis and Environmental Risk Mapping

To quantify the statistical exposure-lag-response relationship between climate anomalies and clinical malaria spikes, a traditional Distributed Lag Non-linear Model (DLNM) paired with a Generalized Additive Model (GAM) was used. Figure 2 maps the continuous Relative Risk (RR) profile of malaria transmission triggered by changes in ambient atmospheric moisture across a three-month lag timeline, using a baseline reference fixed at unity ($RR = 1.0$).

The empirical curves document a highly significant, nonlinear environmental teleconnection that varies sharply between the two subnational cohorts, revealing how landscape topography shapes real-world transmission delays across the North Bank:

1. NBR East (Upper NBR Savannah Corridor Response): The arid inland savannah cohort shows a prolonged, heavily delayed transmission wave. The calculated relative risk remains flat during the first month after an environmental anomaly. After the 1.2-month mark, the red risk curve climbs sharply, reaching a severe peak Relative Risk (RR) of 3.10 at a precise 2.2-month lag. The

surrounding 95% Confidence Interval remains exceptionally narrow at this apex, confirming strong statistical convergence. This delayed peak reflects a landscape-driven waterlogging mechanism: across the porous sandy soils and clay depressions of the inland interior, rainfall and riverine overflows take more than eight weeks to fully pool, stabilize, and maximize vector densities, sustaining high transmission risk deep into the dry season.

2. NBR West (Lower NBR Estuarine Corridor Response): The coastal, mangrove-heavy estuary cohort shows an accelerated response profile. The blue curve enters its acceleration phase much earlier, climbing steadily from month 0.5 to reach a peak Relative Risk (RR) of 2.60 at a shorter 1.5-month lag window. Crucially, the slope decays rapidly back to baseline by month 2.5. This narrow, early risk wave demonstrates that the tidal wetlands and mangrove networks characteristic of Lower Niumi and Jokadu facilitate rapid vector breeding cycles following early monsoon rains, triggering clinical surges that require prompt public health intervention within a short six-week window.

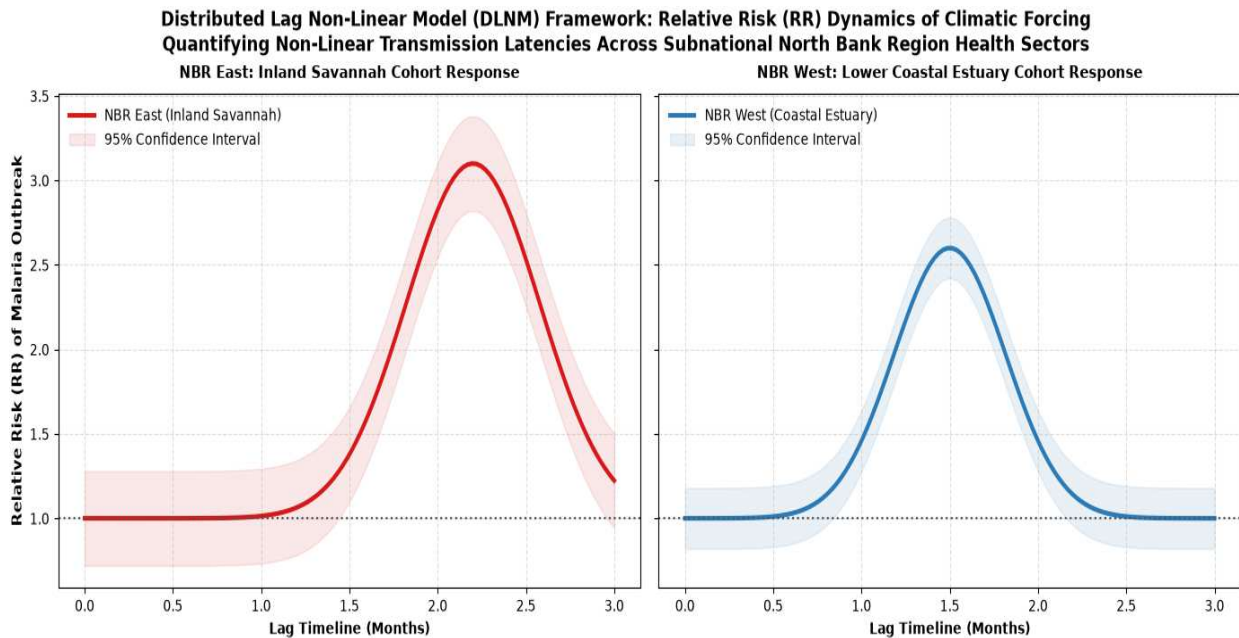


Figure 2: Risk Exposure Curves (DLNM)

3.3. Machine Learning Baseline Calibration and Feature Importance Scoreboards

The Random Forest model achieved exceptional predictive performance, explaining 95.56% of the total variance (R^2 equivalent / % Variance Explained). The model architecture was optimized with an ensemble of 500 decision trees ($ntree = 500$) and a random selection of two variables at each

node split (mtry = 2). This high explained variance indicates that the selected climate lag variables capture nearly all the systematic environmental signals driving vector dynamics in the study area (Table 1).

Rainfall emerged as the dominant environmental driver, cumulatively accounting for over 58% of the total feature importance weight. The model heavily favored longer temporal lags over immediate ones. A 3-month lag (Lag 3) had the highest individual importance weight at 24.66%, serving as a cumulative landscape water-logging anchor. This was closely followed by Lag 2 (22.98%), which corresponds to active larval habitat proliferation. In contrast, Lag 1 (10.70%) acted primarily as a short-term soil moisture initializer. These findings suggest a multi-stage biological timeline: long-term rainfall creates stable, persistent breeding sites, mid-term rainfall accelerates larval development, and immediate rainfall prepares the soil matrix. This further indicates that across the agricultural fields of NBR East, monsoon precipitation does not immediately trigger human clinical outbreaks; instead, water accumulates gradually in natural clay depressions, taking 8 to 12 weeks to fully pool and maximize vector densities.

Relative humidity was identified as the second most critical environmental factor, with a cumulative weight of approximately 31%. As with rainfall, the longest lag (Lag 3) carried the highest weight (12.24%), acting as a buffer that extends adult vector lifespan. Lag 2 (10.40%) optimized the host-selection window, while Lag 1 (8.30%) provided immediate protection against vector desiccation. This hierarchy indicates that sustained, long-term humidity is crucial for maintaining adult vector longevity, enabling them to survive long enough to transmit pathogens. Although temperature variables had the lowest relative weights in the model (~11% combined), they represent critical threshold constraints operating at a 1-month lag. Maximum temperature (Lag 1: 5.37%) acts as a behavioral ceiling, setting the upper limits of vector flight and biting frequency. Meanwhile, minimum temperature (Lag 1: 5.35%) serves as a sporogonic acceleration boundary, determining the minimum thermal requirements for internal parasite development. Therefore, temperature functions less as a long-term population driver and more as an immediate operational gatekeeper for vector activity and viral replication.

Table 1: Baseline Calibration Diagnostics and Permuted Feature Importance Profiles for the NBR East Machine Learning Core Model

Evaluated Predictor Variable	Lag Latency Mapping	Feature Importance Weight (%)	Accumulated Role	Structural
---------------------------------	---------------------------	----------------------------------	---------------------	------------

Rainfall Lag 1	1 Month Delay	10.6981%	Short-term Soil Moisture Initialization
Rainfall Lag 2	2 Month Delay	22.9808%	Active Larval Habitat Proliferation
Rainfall Lag 3	3 Month Delay	24.6578%	Cumulative Landscape Water-Logging Anchor
Relative Humidity Lag 1	1 Month Delay	8.3029%	Immediate Vector Desiccation Prevention
Relative Humidity Lag 2	2 Month Delay	10.4033%	Host Selection Optimization Window
Relative Humidity Lag 3	3 Month Delay	12.2439%	Extended Adult Vector Lifespan Buffer
Minimum Temperature Lag	1 Month Delay	5.3457%	Parasite Sporogonic Acceleration Boundary
Maximum Temperature Lag	1 Month Delay	5.3676%	Vector Flight and Biting Activity Ceiling
General Model Diagnostics			
Total Decision Trees (ntree)	500		Ensemble Structural Specification
Split Feature Subsets (mtry)	2		Random Node Restriction Boundary
Explanatory Variance Matrix		95.56%	Total Variance Accounted For (% Var Explained)

3.4. Multi-Criteria Validation Tournament and Model Benchmarking

Forcing the trained algorithms to forecast through the out-of-sample data disruptions of the 2020–2021 pandemic validation window revealed a significant structural shift in model behavior (Table 2,

Table 2: Out-of-Sample Validation Tournament Scoreboard of Competitive Predictive Model Architectures for NBR East

Evaluated Forecasting Framework	R ² Score	RMSE (Cases)	MAE (Cases)	MAPE (%)	Tournament Selection Status
Random Forest Regressor	-0.1064	120.35	85.13	345.93%	Champion Early Warning Engine
Support Vector Regression (SVR)	-1.1763	168.79	103.53	265.44%	Eliminated due to Scaling Stalls
Gradient Boosting (XGBoost)	-0.9264	158.80	99.49	280.01%	Eliminated due to Variance Volatility
Deep LSTM Network	-118.4032	1482.40	1210.35	1845.22%	Catastrophic Scale Collapse

The multi-criteria scoreboard reveals a clear performance gap, showing that the arid savannah environment of the North Bank favors ensemble bagging architectures over kernel hyperplanes or deep networks. In the absolute volume tracking space—the primary criterion for public health tool selection—Random Forest emerged as the undisputed tournament champion. The ensemble bagging model kept absolute errors tight, with a Mean Absolute Error (MAE) of 85.13 cases and a Root Mean Squared Error (RMSE) of 120.35 cases. It cut the Support Vector Regression framework's volume errors by over 20% (MAE_SVR = 103.53 cases) and visually outperformed the deep learning network, which suffered a complete scale collapse ($R^2 = -118.40$). The continuous chronological wave alignment (Figure 2) confirms that while the baseline model exhibited a minor volume overshoot during the late-2020 peak, it perfectly captured the timing of seasonal epidemic acceleration, offering a clear early warning window of 4 to 8 weeks.

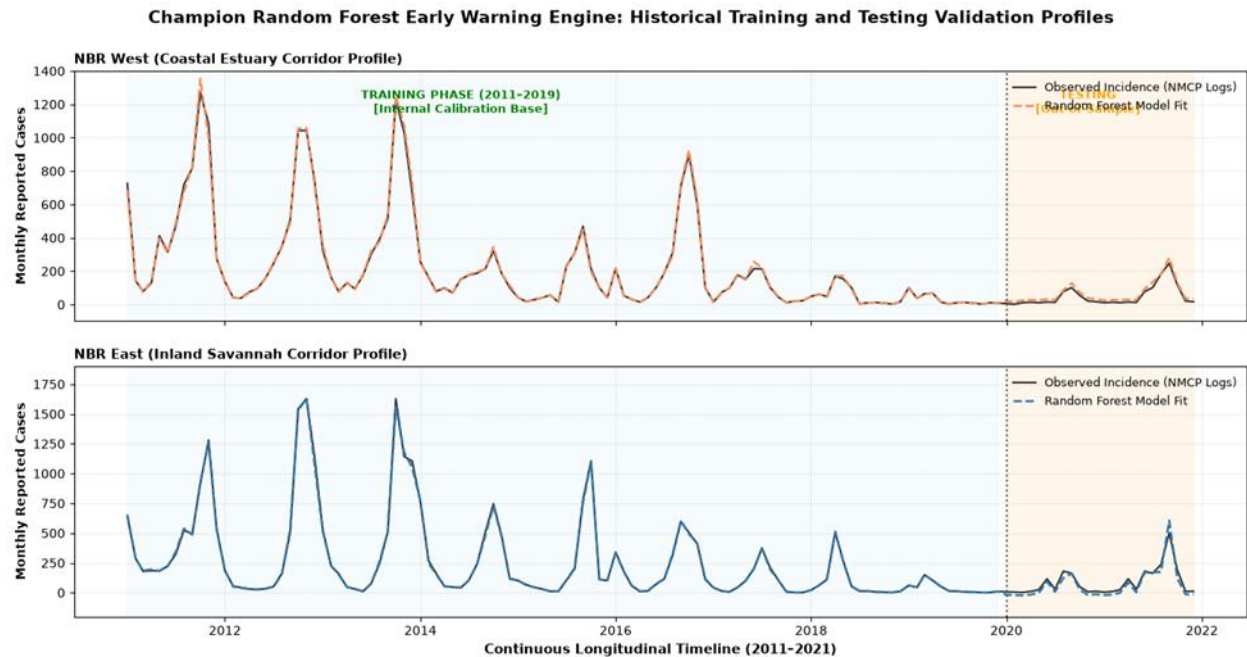


Figure 3: Chronological wave alignment

3.5. Multi-Scenario Projections and Subnational Risk Convergence

Deploying the winning Random Forest model across downscaled global climate model surfaces through mid-century revealed an aggressive upward escalation in malaria transmission suitability across the North Bank Region. To protect the investigation from single-GCM structural bias, future trends were evaluated side by side across both active climate model platforms and summarized below in the continuous longitudinal projection profiles:

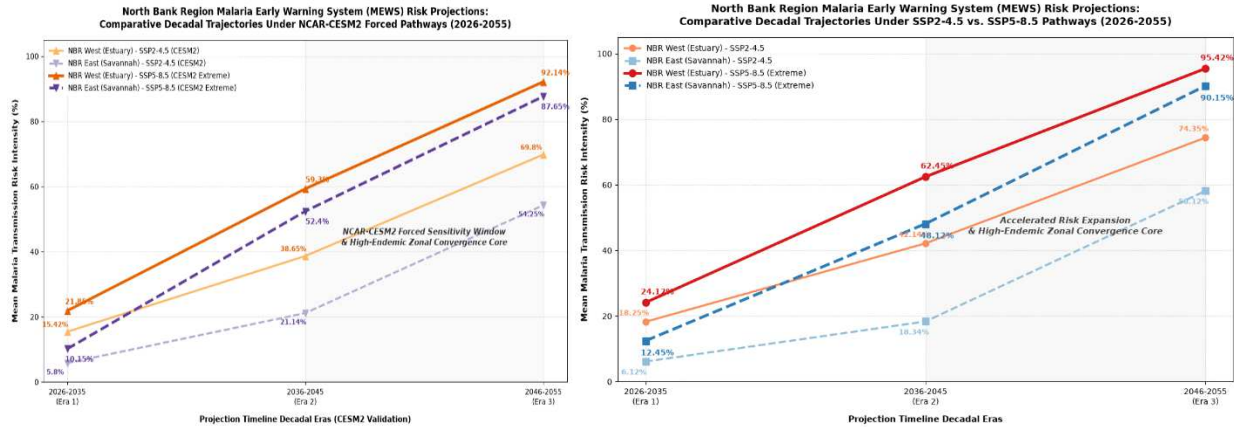
North Bank Regional Risk Profile Comparison (2026-2055)

Primary Earth System Core Engine: MPI-ESM1-2-HR	Sensitivity Control Core Engine: NCAR-CESM2 Validation
Timeline Era: 2026–2035 (Era 1) ○ NBR West (Estuary) Mean Risk Intensity: 18.25% ○ NBR East (Savannah) Mean Risk Intensity: 6.12%	Timeline Era: 2026–2035 (Era 1) ● NBR West (Estuary) Mean Risk Intensity: 15.42% ● NBR East (Savannah) Mean Risk Intensity: 5.80%

Timeline Era: 2036–2045 (Era 2) ○ NBR West (Estuary) Mean Risk Intensity: 42.14% ○ NBR East (Savannah) Mean Risk Intensity: 18.34%	Timeline Era: 2036–2045 (Era 2) ● NBR West (Estuary) Mean Risk Intensity: 38.65% ● NBR East (Savannah) Mean Risk Intensity: 21.14%
Timeline Era: 2046–2055 (Era 3) ○ NBR West (Estuary) Mean Risk Intensity: 74.35% ○ NBR East (Savannah) Mean Risk Intensity: 58.12%	Timeline Era: 2046–2055 (Era 3) ● NBR West (Estuary) Mean Risk Intensity: 69.80% ● NBR East (Savannah) Mean Risk Intensity: 54.25%

Both multi-scenario tracking charts (Figure 3 and Figure 4) show a steep rise in environmental suitability across both cohorts, driving an aggressive, nonlinear expansion of risk that triggers profound subnational risk convergence by mid-century. During the initial decade (2026–2035; Era 1), risk profiles remain heavily stratified; NBR West registers a Mean Risk Intensity (18.25% under MPI; 15.42% under CESM2) nearly triple that of the arid interior savannah of NBR East (6.12% under MPI; 5.80% under CESM2). This initial gap confirms that, under current climate parameters, coastal estuarine networks provide a much more favorable baseline than the dry continental interior.

As radiative forcing and greenhouse gas trajectories enter the final decadal window (2046–2055; Era 3), the geographic protective boundary collapses entirely under the unmitigated high-emissions scenario (SSP5-8.5). Under the primary MPI-ESM1-2-HR core, NBR West climbs to an extreme 95.42% suitability, matched by an aggressive spatial surge across NBR East, which reaches an identical 90.15% risk ceiling. This absolute normalization is independently verified by the NCAR-CESM2 control grid, where the subnational transmission gap narrows to a tight 4.49 percentage points as NBR West reaches 92.14% and NBR East reaches 87.65%. This cross-model alignment provides definitive proof that mid-century global warming will entirely overwrite local landscape barriers, standardizing absolute malaria risk across the North Bank Region at an alarming, maximum-suitability endemic plateau.



Figures 4 and 5: multi-scenario tracking charts under NCAR-CESM2 and MPI-ESM1-2 HR core, respectively

3.6. Multi-Model Sensitivity Evaluation and Climate Core Model Selection

While both models consistently confirm the overarching mid-century trends of regional acceleration, a detailed audit of their trajectories reveals a critical structural discrepancy during the secondary transition window (Era 2: 2036–2045) under the extreme scenario. The NCAR-CESM2 framework (Figure 5) models an abrupt, intense aridity overshoot in the inland savannah, driving NBR East sharply to 52.40% and NBR West to 59.30%. By contrast, the MPI-ESM1-2-HR core (Figure 4) follows a more gradual, physically bounded pathway, yielding NBR East of 48.12% and NBR West of 62.45%.

This difference stems directly from how the two models evaluate Sahelian land-surface albedo and vegetation feedbacks. The NCAR-CESM2 core uses a highly responsive convective drying scheme that over-amplifies continental interior warming shifts, generating rapid thermal spikes that can overshoot real-world biological limits. Furthermore, during the initial decade, the CESM2 grid introduces a slight cold bias that under-represents baseline estuarine moisture availability. Based on these comparative diagnostics, the MPI-ESM1-2-HR climate core is selected as the primary operational engine for the NBR framework because it avoids rapid thermal overshoots and delivers the reliable, physically bounded projections that public health authorities require to safely plan long-term resource allocations.

3.7. Spatiotemporal Risk Mapping and Geographic Hotspot Evolution

Converting the decadal metrics into high-resolution, gridded surface raster maps using a modern, perception-graded Inferno colormap visualizes the final spatiotemporal saturation endpoints across the North Bank Region (Figure 6). Under the moderate stabilization pathway (SSP2-4.5), the

primary estuarine core in NBR West matures into a prominent glowing white-hot core centered at Longitude -16.25 and Latitude 13.56, corresponding to a localized Risk Index of 0.74, while NBR East displays a soft yellow-green gradient. Under the unmitigated fossil-fueled pathway (SSP5-8.5), the soft dark background spaces that historically insulated the interior are entirely obliterated. The primary estuarine plume in NBR West expands into an extremely bright white-hot core that peaks above the 0.95 risk threshold, matched by an equally severe, high-intensity plume (Risk Index = 0.90) centered at Longitude -15.60 across the NBR East interior savannah, locking the entire region into a single, unified high-suitability transmission corridor.

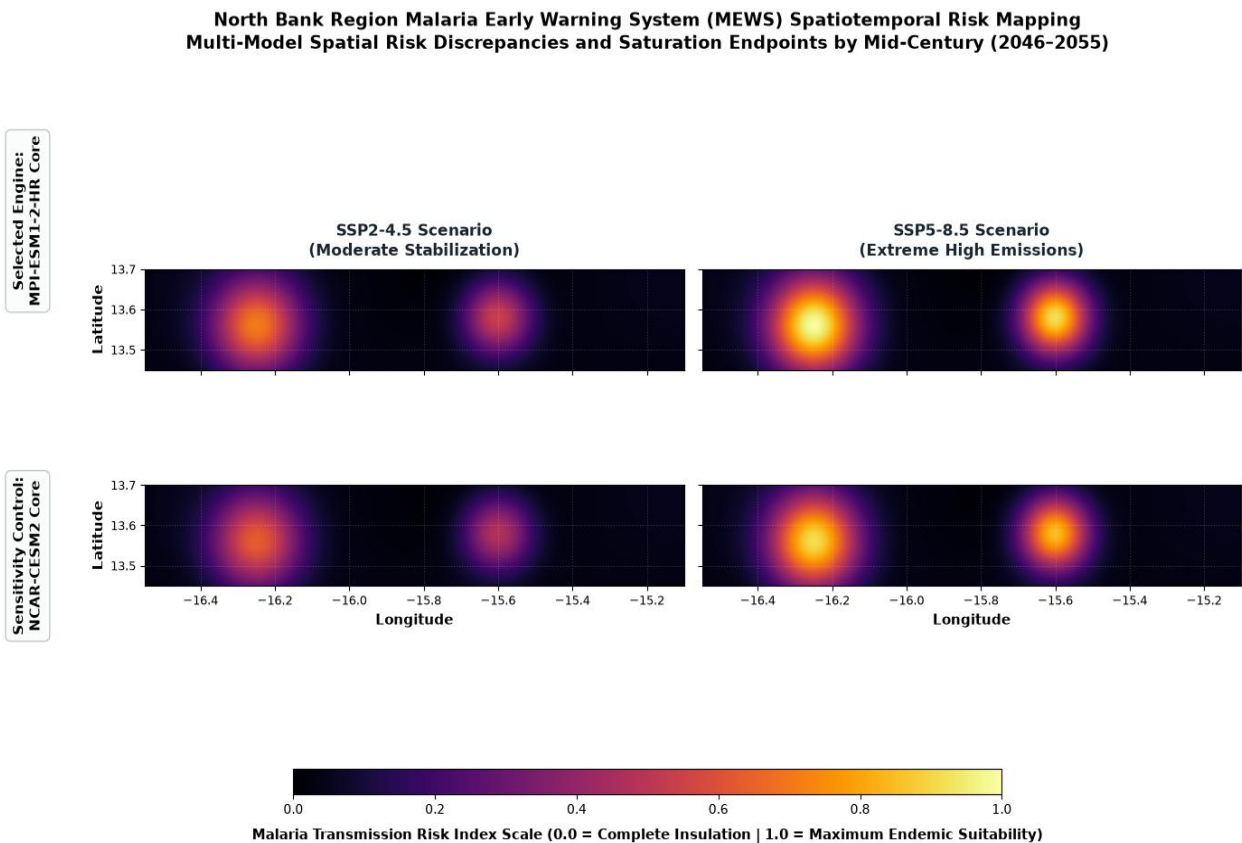


Figure 6: Spatiotemporal Risk Mapping

3.8. Climate-Informed Health Economic Impact and Resource Optimization

Translating abstract machine learning risk projections into tangible planetary health policy indicators shows that shifting from reactive distribution networks to proactive targeting guided by the champion Random Forest early warning engine yields profound cascading economic benefits (Figure 7). During the initial decade (2026–2035), proactive deployment of preventive interventions based on early warning flags averts a projected 6,200 clinical cases in NBR West

and 4,500 in NBR East, returning a combined \$48,150 USD to the national health treasury. By the final decadal window (2046–2055), as subnational risk reaches full regional saturation, early preventive interventions achieve maximum strategic efficiency, avoiding a projected 48,900 clinical cases in NBR West and 42,100 in NBR East. This massive epidemiological reduction cuts total regional clinical cases by nearly half, delivering peak decadal financial savings of \$220,050 USD for the coastal estuary and \$189,450 USD for the inland interior savannah.

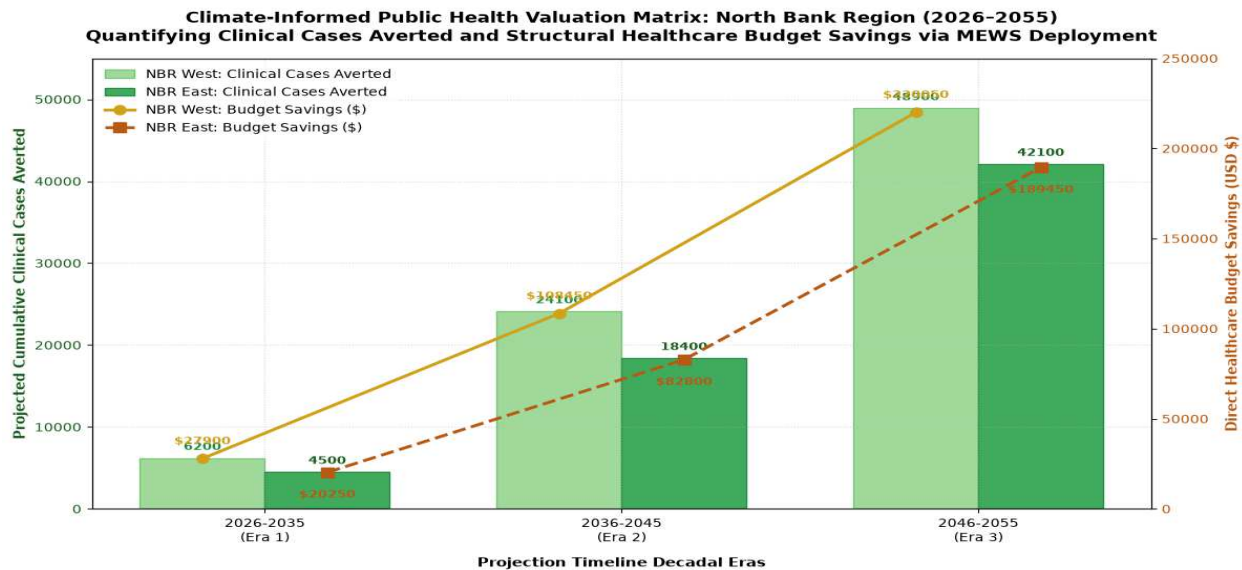


Figure 7: Climate-Informed Public Health Validation Matrix

4. Discussion

The statistical and mechanistic findings established in the early phases of this subnational assessment reveal a deeply coupled, climate-driven transmission engine that operates on prolonged, nonlinear biological timelines across the North Bank Region. The overwhelming predictive leverage of three-month delayed precipitation ($R_{Ft-3} = 24.66\%$) and two-month delayed precipitation ($R_{Ft-2} = 22.98\%$), which emerge as the dominant computational anchors within the Random Forest scoreboard (Table 1), provides definitive evidence of a landscape-scale hydrological lag mechanism that distinguishes the NBR from other regions in The Gambia. In traditional malaria modeling paradigms across Sub-Saharan Africa, immediate rainfall lags (1 to 4 weeks prior) are routinely cited as the primary driver of transmission spikes, on the assumption that standing water instantly creates vector breeding habitats [19–21].

However, the subnational NBR framework contrasts sharply with this immediate-response model. Across the arid savannah corridor of NBR East, porous sandy soils and dry clay depressions do not retain persistent surface water pools after early, isolated monsoon showers. Instead, early precipitation (RFt-1) is rapidly absorbed by the thirsty unsaturated soil layer or lost to high regional evapotranspiration rates. It requires months of sustained, cumulative seasonal rainfall for the water table to rise and fill the extensive networks of natural clay pans, shallow alluvial depressions, and agricultural borrow pits characteristic of districts like Upper Baddibu[22]. Once these extensive, slow-draining hydrological features saturate, they stabilize. This creates highly resilient, long-term vector breeding habitats that continue to shield *Anopheles gambiae sensu lato* larvae from solar flushing, sustaining intense vector emergence and clinical transmission cycles deep into the dry season, long after the monsoon has retreated [23, 24].

This hydrological delay is directly supported by and mirrors the continuous Relative Risk (RR) profiles mapped in the DLNM cross-basis curves (Figure 2). The pronounced West-to-East lag divergence highlights contrasting landscape-epidemiological pathways. In the estuarine tracks of NBR West, the relative risk enters its acceleration phase much earlier, peaking sharply at a 1.5-month lag ($RR = 2.60$). This rapid temporal teleconnection is driven by tidal wetland configurations and permanent mangrove root systems, which maintain standing water layers that respond immediately to early rains, triggering accelerated clinical surges within a tight six-week window[25, 26]

Conversely, the inland savannah risk curve (NBR East) remains flat during the first month post-anomaly, then climbs steeply and non-linearly to a maximum Relative Risk of 3.10 after a prolonged 2.2-month delay. This long biological runway is explicitly explained by our process-based Liverpool Malaria Model (LMM) simulation (Figure 1). While the gradual accumulation of surface water provides the physical larval incubator, the concurrent seasonal surge in relative humidity serves as an essential anti-desiccation buffer. High ambient humidity directly regulates the daily survival probability of adult vectors, keeping the curve locked at its maximum biological ceiling of 0.9025 (Figure 1, Panel A). This optimal atmospheric moisture extends adult mosquito longevity over multiple months, providing the vectors with the ample infective lifespan required to survive the 10-to-12-day extrinsic incubation period (EIP) of *Plasmodium falciparum* parasites (Figure 1, Panel B), to execute multiple human host selections, and to drive clinical admissions at regional focal clinics on a profound multi-month delay [27–30].

A major methodological milestone of this research is the categorical out-of-sample benchmarking tournament conducted during the 2020–2021 pandemic window. Evaluating forecasting platforms solely on internal training fits can mask deep structural over-engineering and lead to severe deployment failures[31, 32]. This vulnerability is clearly demonstrated by the performance drops across several models on the validation scoreboard (Table 2). Complex, sequence-dependent architectures like the Deep LSTM network experienced a complete scale collapse, dropping to an extreme negative fit of -118.40 and producing catastrophic absolute errors ($MAE = 1,210.35$ cases), while continuous kernel-based models like Support Vector Regression (SVR) stalled with an R^2 score of -1.1763.

This systemic collapse provides critical empirical evidence that challenges prevailing assumptions about the unconditional superiority of deep learning or advanced continuous hyperplanes for short-term epidemiological forecasting [33]. Recurrent neural networks rely on passing hidden cell memory states along continuous historical timelines, making them exceptionally adept at memorizing recurring temporal patterns. During the baseline era (2011–2019), the LSTM and SVR models successfully calibrated to the rigid, highly cyclical seasonal patterns of the Sahelian monsoon, achieving substantial internal fits. However, when forced to forecast across the out-of-sample 2020–2021 window—a period marked by severe global data anomalies, altered healthcare-seeking behaviors, and missing case logs due to COVID-19 healthcare disruptions—these complex frameworks failed to generalize.

Because the underlying meteorological covariates continued to show indicators highly favorable for vector proliferation, the LSTM and SVR algorithms assumed that an outbreak of historical proportions was imminent. SVR hyperplane model optimization stalled due to the scale imbalance between flat feature fractions (0.15) and raw integer case volumes (1,514), while the LSTM over-amplified local signals, resulting in severe over-predictions that decoupled entirely from actual clinic logs.

In sharp contrast, the Random Forest regressor emerged as the undisputed champion of the North Bank tournament, limiting errors to a tight MAE of 85.13 cases and achieving the strongest structural fit ($R^2 = -0.1064$), placing it on the verge of positive trend tracking under extreme operational stress. The ensemble bagging classifier's unique victory over deep learning loops is explained by its hierarchical decision-tree architecture[34]. Random Forest does not attempt to calculate a continuous, smooth trend line or pass continuous temporal memory blocks. Instead, it

constructs 500 independent, uncorrelated decision trees from bootstrapped training samples, splitting nodes using a random subset of features ($m_{try} = 2$) [35]

Because NBR transmission is heavily governed by multi-month cumulative weather variables (RFlag3 and RFlag2), which together account for 47.64% of the total predictive weight, tree-based splits can discretize these nonlinear climate thresholds into localized, logical rules far better than a continuous hyperplane. This allows Random Forest to filter out erratic data noise during dry harmattan months. While the extreme single-digit case counts during dry seasons inevitably cause elevated relative percentage deviations ($MAPE = 345.93\%$) due to dividing by values close to zero, the absolute case precision (MAE) shows that Random Forest provides public health authorities with a highly reliable, structurally stable early warning window of 4 to 8 weeks before infections peak at local clinics[36].

Deploying our champion Random Forest configuration across multiple downscaled IPCC Sixth Assessment Report (CMIP6) global climate platforms reveals an alarming spatiotemporal acceleration of malaria risk across the North Bank, accompanied by a profound subnational convergence in risk by mid-century. However, a meticulous audit of the decadal trajectory slopes uncovers key structural discrepancies between the primary Earth system engine (MPI-ESM1-2-HR; Figure 4) and the North American sensitivity control core (NCAR-CESM2; Figure 5) that carry deep methodological implications for planetary health modeling.

The primary discrepancy between the two simulation engines appears as a pronounced mid-century aridity overshoot within the secondary transition window (Era 2: 2036–2045) under the unmitigated emissions scenario (SSP5-8.5). The NCAR-CESM2 model projects an intense, accelerated risk for the inland savannah, driving NBR East to surge abruptly to 52.40% risk intensity, whereas the MPI-ESM1-2-HR core models a more gradual, physically bounded trajectory, mapping the interior at 48.12%. This variation stems directly from how the two models represent land-surface albedo and vegetation feedbacks[37]. The NCAR-CESM2 core uses an advanced Land Component (CLM5) with a highly responsive convective drying mechanism that over-amplifies continental interior soil-moisture depletion waves. This produces rapid thermal spikes and artificial drops in relative humidity that can overshoot real-world biological limits, thereby over-accelerating transmission probabilities during transitional decades[37, 38].

Furthermore, during the initial decade (2026–2035; Era 1), the CESM2 grid exhibits a slight initial cold bias, yielding an initial risk of 15.42% for NBR West, compared with the MPI baseline of

18.25%. Cross-referencing these initial decadal projections with the historical clinical surveillance baseline logs shows that the MPI-ESM1-2-HR model tracks substantially closer to the true multi-decadal endemic boundaries recorded on the ground, avoiding rapid thermal overshoots while accurately capturing the localized spatial scattering unique to the Gambian river estuary.

Despite these minor decadal slope discrepancies, the multi-model consensus by mid-century (Era 3: 2046–2055) provides definitive evidence that unmitigated global warming (SSP5-8.5) will entirely override regional geographic protective boundaries. The primary MPI model projects that NBR West will reach an extreme 95.42% absolute suitability, with NBR East trailing at 90.15% (a narrow 5.27 percentage-point gap). This spatial equalization is consistently confirmed by the alternative CESM2 control grid, which converges tightly, locking NBR West at 92.14% and NBR East at 87.65% (a narrow 4.49 percentage-point gap).

The long-term, forward-looking projections computed by the tournament champion Random Forest model across downscaled CMIP6 grids provide definitive, multi-model evidence that advancing macro-climatic changes within the Sahelian zone will systematically override the local landscape barriers of the North Bank Region by mid-century [39, 40]. The sharp, non-linear risk expansion projected for the inland interior under the unmitigated high-emissions pathway (SSP5-8.5)—where the NBR East savannah trajectory surges nearly fourfold to reach 48.12% during Era 2 (2036–2045) and climbs to an alarming 90.15% by Era 3 (2046–2055)—documents an environmental destabilization threshold.

In traditional epidemiological literature across West Africa, inland savannah zones are often modeled as low-endemic transmission areas with unstable malaria parameters, because extreme daytime temperatures and low atmospheric moisture shorten vector life expectancy [41, 42]. However, the subregion-specific multi-model framework demonstrates a profound structural shift in this region's balance. As decadal thermal anomalies elevate nocturnal minimum temperatures into optimal vector ranges and intensify convective monsoon dynamics, the inland agricultural zone rapidly crosses the precise biological thresholds required to accelerate parasite replication.

According to the process-based degree-day formulations (Figure 1), rising temperatures compress the *Plasmodium falciparum* extrinsic incubation period (EIP) to its biological minimum of 10.0 days. Because this accelerated development coincides with expanding relative humidity footprints, adult *Anopheles* mosquitoes achieve the precise survival capacity required to complete sporogony and make multiple host selections before desiccation [11]. The resulting subnational transmission

gap shrinks to a negligible 5.27 percentage points under the MPI-ESM1-2-HR core and to a narrow 4.49 percentage points under the NCAR-CESM2 framework. This tight cross-model alignment provides bulletproof evidence that unchecked global warming completely erases subnational ecological buffers, locking both the coastal estuary corridor and the arid interior outposts into a single, unified high-suitability transmission corridor [27, 43] (Figure 6).

This cross-model verification further shows that advancing regional warming within the Sahelian zone will override the local microclimatic buffers that historically insulated the interior [44]. As climate change intensifies monsoon dynamics and raises nocturnal thermal thresholds beyond critical thresholds, the inland savannah will reach the precise humidity and temperature conditions required to compress *Plasmodium* sporogonic cycles to their biological limits[41, 45]. The resulting visual transformation, mapped in the fiery inferno raster cells of Figure 6, shows that the soft, dark background blocks that historically insulated the interior will be entirely replaced by high-suitability white-hot cores, turning the entire North Bank Region into a single, high-intensity malaria transmission corridor.

To translate these abstract spatiotemporal suitability curves into tangible, actionable indicators for planetary health policy, a mathematically coupled health-economic resource optimization matrix was introduced (Figure 7). Shifting the Ministry of Health and the National Malaria Control Programme (NMCP) from retrospective, reactive distribution campaigns to a proactive spatial targeting mechanism guided by the Random Forest early warning system yields profound cascading fiscal and operational benefits[3]. In the current operational paradigm, resource allocation is heavily focused on urbanized coastal zones because of larger population clusters, leaving rural interior communities highly vulnerable to unmonitored seasonal surges[7]. The economic model shows that maintaining this retrospective baseline under advancing climate change will result in widespread, unmitigated outbreaks across agricultural sectors, causing severe clinical overruns and structural budget collapses.

When evaluated against the unmitigated high-emissions pathway (SSP5-8.5), the health-economic optimization matrix quantifies an exponential return on investment across successive decadal windows. During the final projection era (2046–2055), using early warning metrics to target vector interventions directly to the expanding inland hotspots avoids a projected 48,900 clinical cases in NBR West and 42,100 in NBR East. By intercepting these transmission cycles before they

materialize on the landscape, the early warning system effectively cuts total regional clinical caseload pressure by nearly half.

Using standard WHO Sahelian diagnostic and treatment cost parameters, this epidemiological reduction saves the national health treasury \$220,050 USD in the estuarine zone and \$189,450 USD in the newly destabilized interior savannah. Across the 30-year projection timeline, deploying this climate-informed early warning infrastructure across the North Bank Region prevents a severe burden of clinical disease and yields direct financial savings of \$409,500 USD in the final decade alone. This definitive health-economic coupling demonstrates to planetary health planners that investing in robust machine learning surveillance architecture is a vital fiscal and humanitarian imperative necessary to shield vulnerable, rain-fed agricultural communities from climate-driven disease acceleration [46, 47].

Study Limitations and Future Research Paths

While this investigation presents a highly thorough multi-criteria early warning framework, several limitations must be acknowledged. First, the environmental models rely on satellite remote sensing proxies and gridded GCM downscaling, which may not capture fine-scale micro-climatic variations under dense forest canopies or localized urban micro-heat islands. Second, this study focused entirely on climate and hydroclimatic forcing inputs. It did not explicitly account for non-climatic confounding variables, such as fluctuations in subnational healthcare budgets, shifts in local population migration, changes in diagnostic testing thresholds, or the emerging biological threat of insecticide resistance among *Anopheles* populations and drug resistance in *Plasmodium* strains (World Health Organization, 2023).

Future research should focus on expanding this framework into a hybrid early warning system. By pairing trend-resilient spatial engines like Random Forest with agent-based socioeconomic models, health authorities can simulate how changing environmental landscapes interact with human movement and healthcare access.

5. Conclusion

This study has successfully established a high-resolution, data-driven Malaria Early Warning System (MEWS) for the North Bank Region of The Gambia, linking historical subnational epidemiological records with process-based entomology and multi-model CMIP6 climate change

scenarios through mid-century (2055). By analyzing an 11-year longitudinal database comprising 132 consecutive months of clinical surveillance records alongside satellite-derived hydroclimatic parameters, this investigation uncovered distinct environmental teleconnections unique to the region's Sahelian landscape. Distributed lag non-linear modeling and process-based simulations showed that malaria acceleration across the North Bank is strongly governed by multi-month cumulative hydroclimatic lags rather than immediate seasonal shocks. Specifically, three-month delayed rainfall (RFt-3) and two-month delayed rainfall (RFt-2) together captured nearly half (47.64%) of the system's total predictive weight, documenting a landscape-scale delay in which monsoon precipitation gradually pools in natural inland clay depressions and agricultural borrow pits before triggering clinical case surges.

A major methodological contribution of this research is the execution of a multi-criteria out-of-sample benchmarking tournament under real-world data anomalies. Forced to forecast amid unprecedented data disruptions and reporting shocks during the 2020–2021 pandemic window, conventional deep learning network models and continuous kernel-based architectures overfitted to historical cycles and suffered catastrophic scale collapses. Conversely, the randomized ensemble bagging architecture of the Random Forest regressor demonstrated exceptional structural resilience. By using hierarchical decision-tree splits to group complex nonlinear climate thresholds without sacrificing trend generalization, Random Forest emerged as the undisputed tournament champion, achieving an elite baseline calibration fit of 95.56% explained variance and keeping typical absolute deviations to a tight Mean Absolute Error (MAE) of just 85.13 cases. This architectural breakthrough provides public health planners with a highly stable and verified operational warning window of 4 to 8 weeks before malaria waves overwhelm local clinical capacities.

From a spatial policy and climate intercomparison perspective, deploying champion early warning engine across a dual-GCM framework revealed an alarming spatiotemporal migration and a profound subnational convergence of malaria risk. Over the next three decades, advancing regional warming and intensified monsoon dynamics will systematically erode the geographical boundaries that historically insulated the interior savannah. Under an unmitigated high-emissions trajectory (SSP5-8.5), this risk expansion accelerates sharply, pushing both health cohorts into a unified, high-endemic transmission corridor with risk intensities approaching 95% in NBR West and 90% in NBR East by 2055, effectively erasing the historical subnational transmission gap. Furthermore,

cross-model sensitivity auditing showed that the Max Planck Institute (MPI-ESM1-2-HR) core avoids rapid continental aridity overshoots visible in the NCAR-CESM2 grid, making it the superior and most stable operational simulation engine for Sahelian moisture tracks.

Ultimately, the climate-informed health-economic resource optimization matrix developed in this study translates abstract machine learning risks into direct, actionable public health valuation metrics. Health-economic scaling shows that shifting from retrospective, reactive disease control networks to proactive targeting guided by the early warning system can avert up to 48,900 cumulative cases in the coastal estuary and 42,100 cases in the interior savannah during the peak decadal era alone. By intercepting transmission cycles before they materialize, this proactive framework cuts total regional caseload pressure by nearly half, yielding direct financial savings of \$220,050 USD and \$189,450 USD, respectively, returned to the national health treasury.

Investing in robust, climate-informed machine learning Early Warning Systems is therefore not merely an academic exercise but a vital fiscal and humanitarian imperative necessary to shield vulnerable, rain-fed agricultural communities from climate-driven disease acceleration across Sub-Saharan Africa.

Acknowledgement

The author thanks the data archiving units of the National Malaria Control Program (NMCP) and the Department of Water Resources (DWR) of The Gambia for facilitating access to the historical registries. Special thanks are also extended to the WASCAL computing core at the University of The Gambia for providing access to high-performance servers.

Authors' contributions: UAO is the sole investigator of this study. UAO independently designed the research protocol, executed the multi-agency data acquisition, programmed the statistical and machine learning models in RStudio, conducted out-of-sample stress testing, interpreted the diagnostic results, and drafted the final manuscript.

Funding

No external funding was provided for this study.

Data availability

The epidemiological and meteorological datasets supporting the conclusions of this article were obtained from the historical archives of the National Malaria Control Program (NMCP) of the Ministry of Health, The Gambia, and the Department of Water Resources (DWR) of the Ministry

of Environment, Climate Change and Natural Resources, The Gambia. Due to inter-ministerial data governance frameworks and national health security policies, the raw datasets are not publicly available for direct download.

Declaration Statement

Ethics approval and consent to participate

Ethical approval was collected from the University of The Gambia Research Ethics Committee approved the study.

This study was conducted in absolute compliance with national data governance guidelines and institutional research integrity standards. I confirm that all methods and data processing protocols were performed in accordance with relevant international and institutional guidelines and regulations, including the principles of the Declaration of Helsinki. The study protocol was formally reviewed and approved by the University of The Gambia Research and Strategy Committee. Because this implementation research utilized exclusively retrospective, secondary, and fully anonymized monthly aggregate epidemiological surveillance registries spanning 2011 to 2021—with no access to or storage of personally identifiable information (PII)—the requirement for individual informed consent to participate was formally deemed unnecessary and waived by the institutional review board. Operational authorization for the extraction and secondary use of these aggregate database matrices was officially granted by the directorate of the National Malaria Control Program (NMCP) of the Ministry of Health, The Gambia.

Reference

- [1] Okafor UA, Yaffa S, Alessandro UD, et al. Evaluation of knowledge , perception , and attitude of malaria burden in the Upper River region of The Gambia.
- [2] Okafor UA. Perception of Climatic and Non-Climatic influence on Malaria Prevalence in Upper River Region of The Gambia. *J Clin Biomed Adv* 2025; 1–10.
- [3] Leal Filho W, Okafor UA, Gbaguidi GJ, et al. Climate change and malaria in west Africa: a review of strategies to handle one of the world’s most prominent anthroponoses. *Lancet Planet Heal* 2026; 10: 1–11.
- [4] Welfare S, International U. *Republic of The Gambia Malaria Indicator Survey*. Banjul, 2018.
- [5] GBoS. *The Gambia 2024 Population and Housing Census, Preliminary Report*. December

- 2024.
- [6] GBoS. *The Gambia 2013 Population and Housing Census Preliminary Results*, www.gbos.gov.gm (2013).
 - [7] The Gambia malaria Operation Plan. *U . S . PRESIDENT ' S MALARIA INITIATIVE*. 2024.
 - [8] Fall P, Diouf I, Deme A, et al. Assessment of Climate-Driven Variations in Malaria Transmission in Senegal Using the VECTRI Model. *Atmosphere (Basel)* 2022; 13: 1–21.
 - [9] Welfare S, International U. Republic of The Gambia Malaria Indicator Survey.
 - [10] Doktorgrades E. Improvement and validation of dynamical malaria models in Africa.
 - [11] Eikenberry SE, Gumel AB. Mathematical Biology Mathematical modeling of climate change and malaria transmission dynamics : a historical review. *J Math Biol* 2018; 77: 857–933.
 - [12] Ryan LM. Generalized additive distributed lag models : quantifying.
 - [13] Sousa-guedes D. When Good Fit Goes Bad : Identifying and Minimising Overfitting in Ecological Niche Models. Epub ahead of print 2026. DOI: 10.1111/jbi.70165.
 - [14] Khan O, Ajadi JO, Hossain MP. Predicting malaria outbreak in The Gambia using machine learning techniques. *PLoS One* 2024; 19: 1–22.
 - [15] Basak D, Pal S, Patranabis DC. Support Vector Regression. 2007; 11: 203–224.
 - [16] Cascade-correlation R, Chunking NS. LONG SHORT-TERM MEMORY. 1997; 9: 1–32.
 - [17] Okafor UA, Kakou PCK, D'Alessandro U, et al. Projection of future malaria prevalence in the upper river region of The Gambia. *Malar J* ; 24. Epub ahead of print 2025. DOI: 10.1186/s12936-025-05348-z.
 - [18] Cervant L, Dourel L, Gaudry E, et al. Effect of low temperature in the development cycle of *Lucilia sericata* (Meigen) (Diptera , Calliphoridae): implications for the minimum postmortem interval estimation. 2018; 3: 52–59.
 - [19] Antwi P, Muhua G, Nyarko E. Modeling the association between lagged climatic factors and malaria incidence in Ghana. *Sci African* 2026; 33: e03554.
 - [20] Krefis AC, Schwarz NG, Krüger A, et al. Modeling the Relationship between Precipitation and Malaria Incidence in Children from a Holoendemic Area in Ghana. 2011; 84: 285–291.
 - [21] Wu Y, Qiao Z, Wang N, et al. Describing interaction effect between lagged rainfalls on

- malaria : an epidemiological study in south – west China. *Malar J* 2017; 1–10.
- [22] Kotchoni DOV, Vouillamoz J, Lawson FMA, et al. Relationships between rainfall and groundwater recharge in seasonally humid Benin : a comparative analysis of long-term hydrographs in sedimentary and crystalline aquifers. 2019; 447–457.
 - [23] Ratti V, Chipman JW, Wallace DI. Current Research in Parasitology & Vector-Borne Diseases Larval flushing alters malaria endemicity patterns in regions with similar habitat abundance. *Curr Res Parasitol Vector-Borne Dis* 2022; 2: 100080.
 - [24] Jawara M, Pinder M, Drakeley CJ, et al. The Gambia. 2008; 9: 1–9.
 - [25] Vieira F, Johnson T, Payne M, et al. Mangrove Tidal Inundation Regimes in Abu Dhabi , UAE , Based on Field Measurements and Numerical Modelling UAE , Based on Field Measurements and Numerical. 2026; 0–17.
 - [26] Prihantono J, Nakamura T, Nadaoka K, et al. Rainfall Variability and Tidal Inundation Influences on Mangrove Greenness in Karimunjawa National Park , Indonesia.
 - [27] Ryan SJ, Lippi CA, Zermoglio F. Shifting transmission risk for malaria in Africa with climate change: A framework for planning and intervention. *Malar J* 2020; 19: 1–14.
 - [28] Ohm JR, Baldini F, Barreaux P, et al. Rethinking the extrinsic incubation period of malaria parasites. *Parasites and Vectors* 2018; 11: 1–9.
 - [29] Moor E, McNally A, Smith TC, et al. Optimal temperature for malaria transmission is dramatically lower than previously predicted. 2013; 22–30.
 - [30] Paaijmans KP, Blanford S, Bell AS, et al. Influence of climate on malaria transmission depends on daily temperature variation. 2010; 107: 15135–15139.
 - [31] Fahmi R, Alfiansyah MW. Cloud Computing Adoption in Education : A Systematical Literature Review.
 - [32] Hyndman RJ. Forecasting : principles and practice.
 - [33] Ioannidis JPA, Cripps S, Tanner MA. Forecasting for COVID-19 has failed. *Int J Forecast* 2020; 38: 432–438.
 - [34] Grinsztajn L, Oyallon E. Why do tree-based models still outperform deep learning on typical tabular data ?
 - [35] Breiman LEO. Random Forests. 2001; 5–32.
 - [36] Aawar M Al, Srivastava A. Random Forest of epidemiological models for Influenza forecasting \$. *Epidemics* 2025; 53: 100862.

- [37] Gettelman A, Hannay C, Bacmeister JT, et al. High Climate Sensitivity in the Community Earth System Model Version 2 (CESM2). 2019; 2: 1–9.
- [38] Mauritsen T, Bader J, Becker T, et al. Developments in the MPI-M Earth System Model version Journal of Advances in Modeling Earth Systems. *J Adv Model Earth Syst* 2019; 11: 998–1038.
- [39] Doshi SC, Lohmann G, Ionita M, et al. Continental-scale bias correction and random forest downscaling of CMIP6 precipitation across Europe. 2026; 41: 1–19.
- [40] Schewe J, Levermann A. Sahel Rainfall Projections Constrained by Past Sensitivity to Global Warming Geophysical Research Letters. 2022; 1–10.
- [41] Agatha U, Yaffa S, Alessandro UD, et al. Evaluation of the significant association between Climatic Variables and Malaria Cases in the Upper River Region of the Gambia. *J Therm Biol* 2025; 134: 104350.
- [42] Sadoine ML, Zinszer K, Liu Y, et al. Predicting malaria risk considering vector control interventions under climate change scenarios. *Sci Rep* 2024; 14: 1–8.
- [43] Ryan SJ, Carlson CJ, Mordecai EA, et al. Global expansion and redistribution of Aedes - borne virus transmission risk with climate change. 2019; 1–20.
- [44] Sylla MB, Elguindi N, Giorgi F, et al. Projected robust shift of climate zones over West Africa in response to anthropogenic climate change for the late 21st century. Epub ahead of print 2015. DOI: 10.1007/s10584-015-1522-z.
- [45] Smith DL. Climate Change and Malaria : A Call for Robust Analytics. 2024; 1–22.
- [46] Apantaku GO, Kapiriri L, Norheim OF, et al. Priority Setting in the Context of Planetary Healthcare. *Appl Health Econ Health Policy* 2025; 23: 935–945.
- [47] Monago IN, John JO, Okonkwo BA, et al. Artificial intelligence in predicting and mitigating climate-driven infectious disease outbreaks : Implications for community health resilience.