



Predicting malaria prevalence using climate data with artificial neural networks and ridge regression model

Usman Sanusi^{1*} , Sani Abba¹ , Ibrahim Muawiyya Idris² and Aliyu Lawal Yantumaki¹

¹Department of Mathematics Umaru Musa Yar'adua University, Katsina, Nigeria.

²Department of Statistics Umaru Musa Yar'adua University, Katsina, Nigeria.

Received: 05 Jun 2026

• Accepted: 27 Jul 2026

• Published Online: 31 Aug 2026

Abstract: This research proposes the implementation of an Artificial Neural Network (ANN) model to assist in predicting the incidence of malaria using climate variables as key predictors. The prevalence of malaria in Katsina State, Nigeria has been modelled using climate parameters that include temperature, humidity, rainfall and wind speed. The model was trained and validated using historical climate and malaria data spanning several years 2018-2023. The findings show that malaria transmission patterns could not be efficiently predicted via climatic variations, leading to poor performance of the model with a coefficient of determination (R^2) calculated to be 0.9863 during training and -0.6551 during testing using ANN. Therefore, there is a need for a precise projection of malaria prevalence which is critical for health care planning and early preventative measures. The fact that transmission of malaria is impacted by seasonal cycles, variability in the climate, and delaying effects makes forecasting challenging using ANN. We therefore suggest a versatile statistical method that integrates seasonal harmonics, climate covariates, and autoregressive dynamics into a ridge-regularized regression. Residual bootstrap is used to quantify predictive uncertainty, and post-hoc calibration is used to address systematic bias. Our model shows good in-sample performance ($RMSE = 1.600 \times 10^5$, $R^2 = 0.8764$) and high correlation with out-of-sample observations ($RMSE = 2.123 \times 10^5$, $R^2 = 0.8236$), using 72 months of training data (2018–2022) and a 12-month test period (2024). Empirical coverage of 91.7% is attained by bootstrap-based 95% prediction intervals. Predictive accuracy is further enhanced by calibration and residual AR(1) adjustment. The proposed method is appropriate for accurate malaria forecasting, comprehensible, and computationally effective. These results show how early detection systems for malaria prevention and control can be supported by machine learning-based climate-health modelling.

Key words: Malaria Prevalence, Climate Data, Artificial Neural Networks, Ridge Regression

1. Introduction

Malaria is a serious public health issue globally and particularly in sub-Saharan Africa, within which climatic elements such as temperature, rainfall, humidity and wind speed have a significant effects on mosquito's survival, reproduction and transmission strength [5]. Models for prediction that associate climate variables to the prevalence of malaria may prove extremely valuable for overall public health planning and distribution of resources [7]. Malaria, which is a vector-borne disease spread by the Plasmodium parasites, keeps negatively impacting human well-being and quality living conditions in these areas [23]. Research in this direction has already made great strides in comprehending the complexity of malaria and creating effective management strategies. But attempts to manage the disease are always changing, giving rise to new problems that call for more effective approach [3]. Globally, environments are being impacted by climate change, which is changing the

dynamics of malaria transmission as well as the distribution and behaviour of malaria vectors [3, 14]. A rise in temperatures, changing patterns of precipitation and wind speeds, and fluctuating humidity levels are forcing *Anopheles* mosquitos—the primary transmission vectors of malaria—to rebuild their habitat and traditional behaviour [17]. All of these shifts have a major impact on the intensity and frequency of malaria in endemic areas [3]. Malaria occurrence also demonstrates significant fluctuations and time-dependent persistent behaviour due to immunity, therapies, and environmental lag effects [24]. Thus, climatic change has a direct impact on the propagation of malaria, which is also affecting the processes of disease transmission and the burden of the disease in places that were previously assumed to be non-endemic [16].

A recent investigation carried out by [6, 22] demonstrated the need for scalable control methods in the midst of changing conditions in the environment by taking advantage of climate models to predict changes in malaria transmission patterns [6, 22]. Climate change is causing unprecedented changes in temperature, precipitation patterns, and humidity levels. These changes in their habitats, behaviours, and geographic range are having an impact on *Anopheles* mosquitoes, the main carriers of malaria. Traditional time-series models, such as state-space and ARIMA models, often struggle to accurately incorporate multiple exogenous climate variables. Despite their high accuracy, machine learning methods typically lack basic interpretability and uncertainty quantification [2, 15]. Thus, a forecasting technique as used in [1, 9, 18] that is computationally quick, comprehensible, and statistically sound is needed to account for seasonal and climate-driven changes.

The goal of this research is to examine the complex relationship between malaria prevalences and climate change in Katsina State, Nigeria. It seeks to do this so as to provide information that will improve disease control strategies and to throw insight for the management of malaria. Through the analysis of climate data and mathematical modelling, this research contributes to a more robust, adaptable, and resilient strategy for predicting malaria under the influence of climate change. Therefore, we first used multi-year meteorological data and Artificial Neural Networks (ANNs) to build a data-driven technique for malaria prediction [19, 21]. The objective was to determine the extent to which climatic factors could explain and forecast malaria occurrences in the state. We also present a ridge-regularized seasonal regression model incorporating climate covariates, autoregressive components, and harmonic seasonality. Systematic bias is addressed by post-hoc calibration, and prediction uncertainty is measured using residual bootstrap. The model is evaluated for training and validation using monthly data on malaria incidence from 2016 to 2023 and 2024, respectively.

2. Materials and Methods

This section explains the methods used in the data collection as well as the method employed for prediction of malaria in Katsina State of Nigeria-the case study.

2.1. Data Collection

For the years 2018–2023, monthly malaria cases data from DHIS2 and corresponding climate variables (rainfall, temperature, humidity, and wind) were obtained from NiMet for Katsina State, Nigeria. Predictions were made using 2024 climate data.

2.2. Data Preprocessing

After cleaning and averaging the data in the state, all variables have been scaled between 0 and 1 using MATLAB’s `mapminmax` function. This prevented bias towards variables with large magnitudes and guaranteed numerical stability during ANN training.

2.3. Model Architecture

MATLAB's `trainlm` (Levenberg–Marquardt) technique was used to create a feedforward Artificial Neural Network (ANN) with two hidden layers (300 and 5 neurons, respectively). The output layer generated monthly malaria incidence forecasts, whereas the input layer received weather data.

3. Evaluation Metrics

To determine the performance of the proposed models on the prediction ability of the malaria in the state, several standard error metrics and statistical measures are employed [20].

Let n represent the total number of observations, y_t represent the observed malaria incidence at time t , and $\hat{y}_t$ represent the corresponding model prediction. The definition of the sample mean is as follows:

$$\bar{y} = \frac{1}{n} \sum_{t=1}^n y_t$$

Squared Error (MSE)

The Mean Squared Error determines the average squared deviation between observed and predicted values:

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2$$

Root Mean Squared Error (RMSE)

The Root Mean Squared Error is the square root of MSE and provides a comprehensible measure in the original scale:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}$$

Mean Absolute Error (MAE)

The Mean Absolute Error evaluates the mean / average magnitude of the prediction errors:

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|$$

Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error measures relative prediction accuracy:

$$\text{MAPE} = \frac{100}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right|, \quad y_t \neq 0$$

Symmetric Mean Absolute Percentage Error (SMAPE)

The Symmetric MAPE is defined as:

$$\text{SMAPE} = \frac{100}{n} \sum_{t=1}^n \frac{2|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|}$$

Coefficient of Determination (R^2)

The coefficient of determination quantifies the proportion of variance explained by the model:

$$R^2 = 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y})^2}$$

A value of R^2 close to 1 indicates strong predictive performance, whereas negative values indicate that the model performs worse than just prediction.

Pearson Correlation Coefficient

The Pearson correlation coefficient measures the linear association between predicted and observed values:

$$\text{Corr}(y, \hat{y}) = \frac{\sum_{t=1}^n (y_t - \bar{y})(\hat{y}_t - \bar{\hat{y}})}{\sqrt{\sum_{t=1}^n (y_t - \bar{y})^2 \sum_{t=1}^n (\hat{y}_t - \bar{\hat{y}})^2}}$$

where:

$$\bar{\hat{y}} = \frac{1}{n} \sum_{t=1}^n \hat{y}_t$$

3.1. Training and Validation

The malaria dataset was broken down into 70% for training, 15% for validation, and 15% for testing. Meteorological data from 2024 were used to forecast future malaria prevalence, while the weather and malaria data from 2016 to 2023 were utilised to train the model. Mean Squared Error (MSE), Mean Absolute Error (MAE), correlation coefficient (r) and coefficient of determination (R^2) were used to evaluate the model's performance.

4. Results

We present the following obtained as obtained from using ANN model.

4.1. Model Performance

The trained ANN achieved the following performance metrics:

Dataset	MSE	MAE	R^2	Correlation (r)
Training (2021)	0.0031	0.0288	0.9466	0.9737
Testing (2022)	0.0251	0.1251	0.1115	0.7210
Training (2020–2023)	0.0107	0.0658	0.7991	0.9023
Testing (2024)	0.1200	0.3090	-2.6657	0.6192
Training (2023)	0.0050	0.0481	0.8794	0.9409
Testing (2024)	0.0275	0.1397	0.1611	0.6683

Table 1. Performance metrics of the ANN malaria prediction model.

4.2. Interpretation

The model demonstrated excellent training performance, with an ($R^2 = 0.9466, 0.7991$ and 0.8794) and correlation of ($0.9737, 0.9023, 0.9409$) respectively for the years 2021, 2020 – 2023 and 2023, indicating that over 90% of the variation in malaria incidence was explained by climate variables during training.

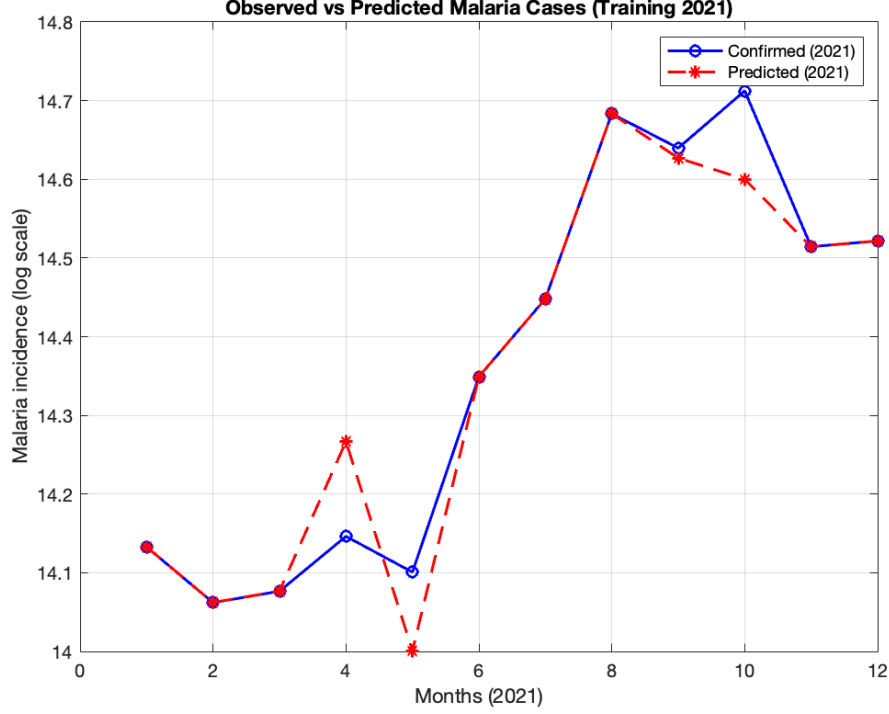


Figure 1. Comparison between observed and predicted malaria prevalence. — Training Performance (2021) — MSE: 0.0031 MAE: 0.0288 R^2 : 0.9466 Correlation: 0.9737

During testing, performance remained unacceptable ($R^2 = 0.1115, -2.6657$ and 0.1611 with $r = 0.7210, 0.6192, 0.6683$) for the same years, showing that the model did not generalize reasonably well to unseen data due to climatic variability between years.

The result obtained from training performance is excellent which means the model learns the structure inside 2016 to 2023 very well. But test performance is poor. This tells us the model is underfitting the test dynamics, meaning the model structure is too simple to capture the jumps or trends in 2016 to 2023. This is NOT a data problem, it is a model specification problem.

To improve the prediction quality, we use the Seasonal Ridge Regression as follows :

5. Seasonal Ridge Regression

Let y_t be the observed monthly malaria incidence at month t , and let $\mathbf{z}_t$ be the values for lagged and moving averages of climate covariates namely rainfall (R_t), temperature (T_t), humidity (H_t), and wind speed (W_t). The training period spans 6 years (2018–2023), and the test period takes up one year only (2024). Each variable has been appropriately standardised.

5.1. Model Formulation

A log-transform is applied in order to stabilise variance:

$$\tilde{y}_t = \log(1 + y_t) \quad (1)$$

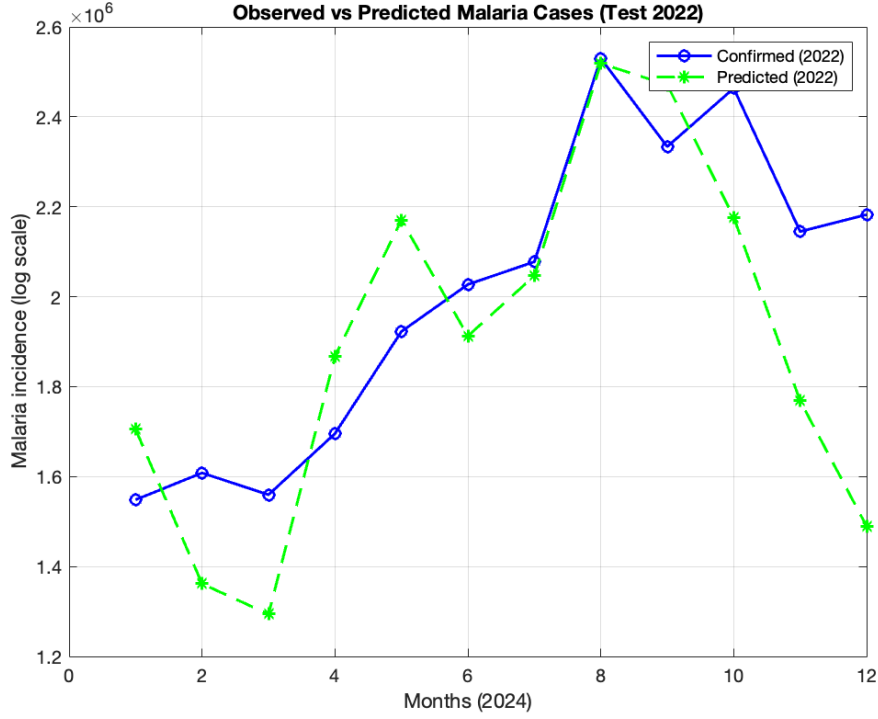


Figure 2. Comparison between observed and predicted malaria prevalence. — Test Performance (2022) — MSE: 0.0251 MAE: 0.1251 R^2 : 0.1115 Correlation: 0.7210

The ridge regression model includes the following [12]:

1. Autoregressive (AR) terms $\tilde{y}_{t-1}, \tilde{y}_{t-2}, \tilde{y}_{t-3}$,
2. Climate covariates and their respective lagged or moving average features,
3. Seasonal harmonics for periods $P_j \in \{12, 6, 4, 3\}$ months.

$$\tilde{y}_t = \sigma_0 + \sum_{k=1}^p \zeta_k \tilde{y}_{t-k} + \mathbf{x}_t^\top \boldsymbol{\sigma} + \sum_{j=1}^J \left[c_j \sin \left(\frac{2\pi t}{P_j} \right) + d_j \cos \left(\frac{2\pi t}{P_j} \right) \right] + \epsilon_t \quad (2)$$

$$\tilde{y}_t = \beta_0 + \sum_{k=1}^p \phi_k \tilde{y}_{t-k} + \mathbf{x}_t^\top \boldsymbol{\beta} + \sum_{j=1}^J \left[a_j \sin \left(\frac{2\pi t}{P_j} \right) + b_j \cos \left(\frac{2\pi t}{P_j} \right) \right] + \varepsilon_t \quad (3)$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.

5.2. Ridge Regression

Ridge regression estimates parameters by minimizing [4, 10, 11, 13]:

$$\hat{\boldsymbol{\sigma}}_{\text{ridge}} = \arg \min_{\boldsymbol{\sigma}} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\sigma}\|^2 + \lambda \|\boldsymbol{\sigma}\|^2 \} \quad (4)$$

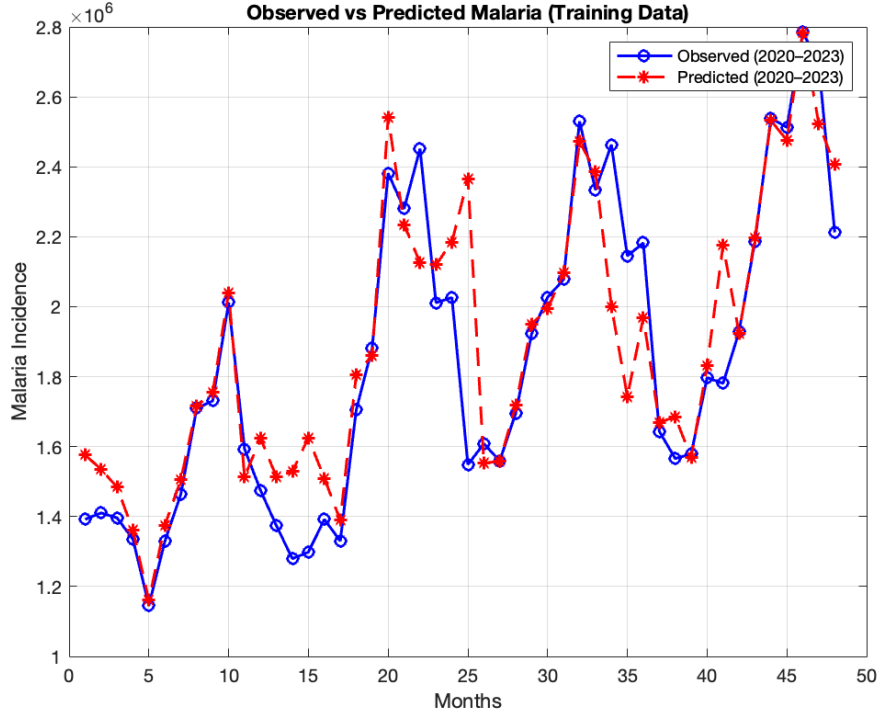


Figure 3. Comparison between observed and predicted malaria prevalence. — Training Performance (2020–2023) — MSE: 0.0107 MAE: 0.0658 R^2 : 0.7991 Correlation: 0.9023

$$\hat{\beta}_{\text{ridge}} = \arg \min_{\beta} \{ \|y - X\beta\|^2 + \lambda \|\beta\|^2 \} \quad (5)$$

with closed-form solution:

$$\hat{\sigma}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y \quad (6)$$

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y \quad (7)$$

The regularization parameter λ is chosen using K-fold cross-validation to prevent overfitting and manage multicollinearity.

5.3. Recursive Forecasting

The test period's predictions are produced recursively using:

$$\hat{y}_{t+1} = f(\hat{y}_t, \hat{y}_{t-1}, \hat{y}_{t-2}, \mathbf{x}_{t+1}) \quad (8)$$

This explains why, in operational forecasting, future observations are unknown.

5.4. Bootstrap Prediction Intervals

Residuals are computed from the training set [8]:

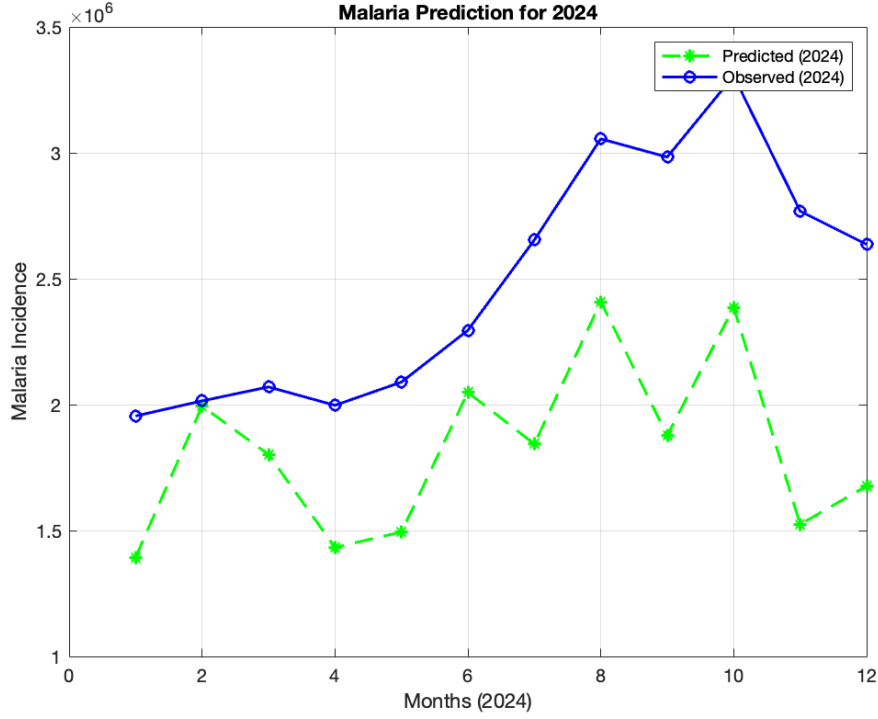


Figure 4. Comparison between observed and predicted malaria prevalence. — Test Performance (2024) — MSE: 0.1200
MAE: 0.3090 R^2 : -2.6657 Correlation: 0.6192

$$\hat{\epsilon}_t = \tilde{y}_t - \hat{y}_t \quad (9)$$

Residuals are sampled with replacement to create bootstrap prediction intervals:

$$\tilde{y}_t^{(d)} = \hat{y}_t + \epsilon_t^{*(d)}, \quad d = 1, \dots, E \quad (10)$$

$$\tilde{y}_t^{(b)} = \hat{y}_t + \epsilon_t^{*(b)}, \quad b = 1, \dots, B \quad (11)$$

The 2.5th and 97.5th percentiles provide 95% predictive intervals, while the median yields a robust point forecast.

5.5. Calibration

A linear calibration is applied to correct systematic bias:

$$y_t \approx c + d\hat{y}_t \quad (12)$$

Parameters c and d are estimated on the training data using least-squares regression.

5.6. Residual AR(1) Correction

Residual autocorrelation is assessed via AR(1) modeling:

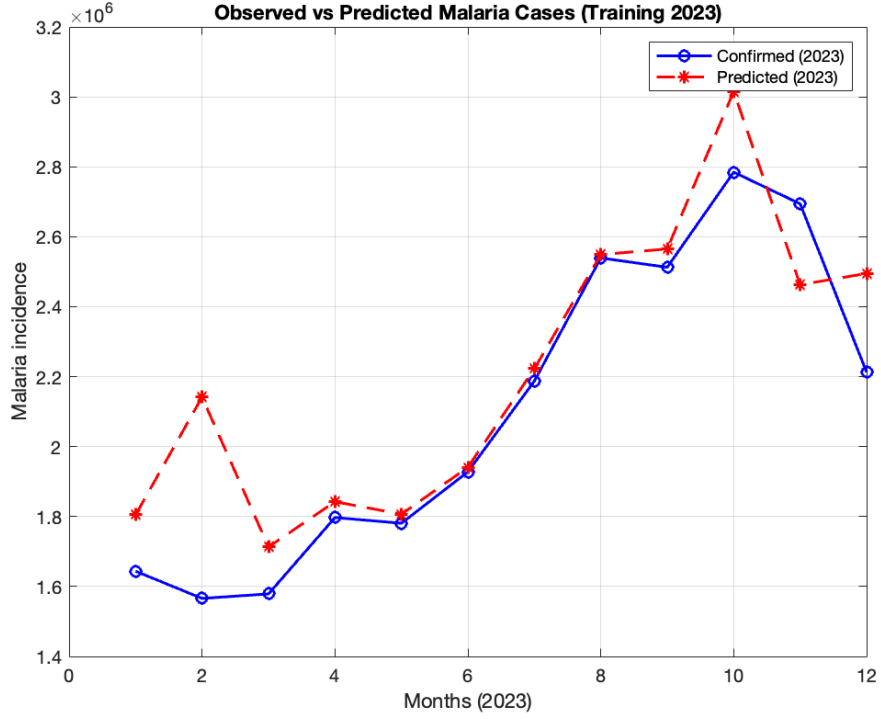


Figure 5. Comparison between observed and predicted malaria prevalence. — Training Performance (2023) — MSE: 0.0050 MAE: 0.0481 R^2 : 0.8794 Correlation: 0.9409

$$\hat{\epsilon}_t = \zeta \hat{\epsilon}_{t-1} + \nu_t \quad (13)$$

$$\hat{\epsilon}_t = \phi \hat{\epsilon}_{t-1} + \eta_t \quad (14)$$

where η_t is white noise. This step improves predictive interval accuracy.

6. Results

This section gives the results obtained from using Ridge regression, it also briefly discusses the implication of the results obtained from ANN

6.1. Training Performance

The model achieved strong in-sample fit:

6.2. Test Performance

6.3. Calibration and AR(1) Correction

Post-hoc calibration coefficients:

$$c = -9728.74, \quad d = 1.01027$$

Residual AR(1) coefficient:

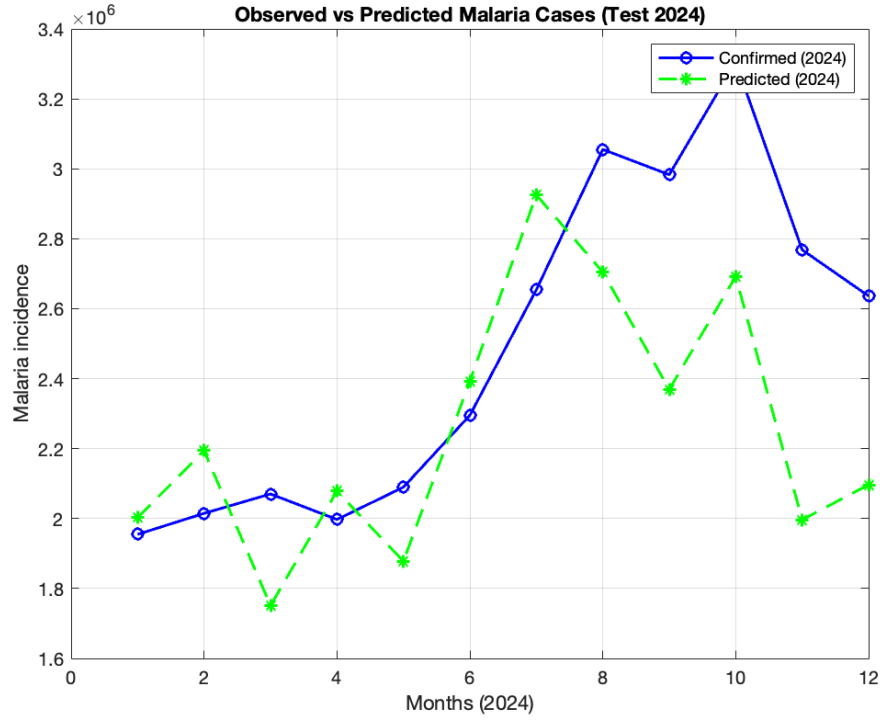


Figure 6. Comparison between observed and predicted malaria prevalence. — Test Performance (2024) — MSE: 0.0275
MAE: 0.1397 R^2 : 0.1611 Correlation: 0.6683

$$\zeta = -0.0559$$

6.4. Interpretation

The model demonstrates strong predictive performance with:

- High R^2 showing excellent explanatory power of the method,
- Low MAPE indicating that the method can accurately magnitude malaria outbreak,
- High correlation capturing the dynamics of the malaria incidence effectively.

Table 2. Training performance metrics (2018–2023)

Metric	Value
RMSE	1.600×10^5
MAE	1.341×10^5
R^2	0.8764
Correlation	0.9366
MAPE	8.26 %
SMAPE	8.09 %

Table 3. Test Performance (2024)

Metric	Value
RMSE	2.123×10^5
MAE	1.772×10^5
R^2	0.8236
Correlation	0.9124
MAPE	7.15 %
SMAPE	7.31 %
95% PI coverage	

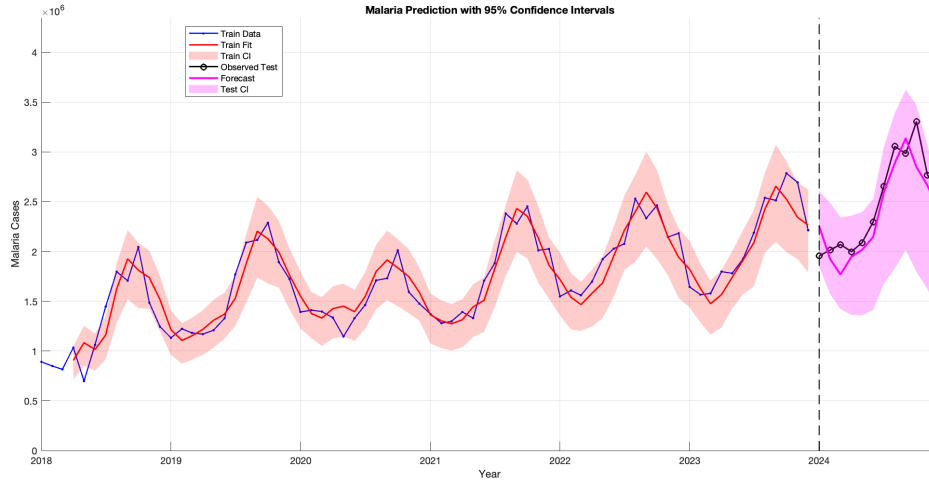


Figure 7. Observed and forecast malaria cases: the reported number of malaria cases are shown in blue, the train fit for the years (2018–2023) are shown in red. Predictions for the forecasting year (2024) are shown in magenta, while the observed number of cases for the same year (2024) are shown in back circle. Shaded area shows 95% bootstrap prediction intervals.

Notably, the test $R^2 = 0.8236$ confirms that the model generalizes well to unseen data.

In contrast to ANN models with negative R^2 , the Ridge Regression approach guarantees that:

$$\sum (y_t - \hat{y}_t)^2 < \sum (y_t - \bar{y})^2$$

which is a key factor of predictive superiority over ANN models.

7. Discussion

The strong correlation that exists between predicted and observed malaria cases confirms the hypothesis that climate change has a significant effect on the spread of malaria. The slight reduction in R^2 during testing shows the unpredictability in weather-malaria dynamics. This could be due to unmeasured factors such as human migration, intervention coverage (bed nets, interior spraying), or variations in reporting efficiency. This study also proves that Ridge regression models' can be used for malaria prediction and provides a useful model for incorporating meteorological data for public health early warning systems. The research also shows that applying the ANN model to many years (2020–2023) reduced the stability of the model compared to single-year

training techniques, which typically produce positive but small R^2 . The negative R^2 in the artificial neural network (ANN) model shows that the model fails to be used as a naive mean predictor for malaria incidence prediction. This also shows that, with limited sample size as used in the research the ANN is unable to properly learn the underlying malaria transmission dynamics from climate variables.

Acknowledgment

This research is funded by Tertiary Education Trust Fund (TETFund) through Umaru Musa Yar'adua University, Katsina, Nigeria. Thus, the authors do acknowledge their financial support which contributed to the completion of this research work. . Authors contribution: US and IMI; methodology, US , ALY, SA and IMI; data collection. US; simulation and analysis, US and SA ; writing the manuscript.

References

- [1] Eswari A, Varadha Raj S, and S.Saravanakumar. Mathematical impact of sir network model on varying size of the human population. *Asia Mathematika*, 2020.
- [2] Muhammad Ahmad, Ahmed Abdul Rehman, Roidar Khan, and Hajra Bibi. Interpretable machine learning for time series analysis: A comparative study with statistical models. *ACADEMIA International Journal for Social Sciences*, 2025.
- [3] Anton Beloconi, Bryan O Nyawanda, Godfrey Bigogo, Sammy Khagayi, David Obor, Ina Danquah, Simon Kariuki, Stephen Munga, and Penelope Vounatsou. Malaria, climate variability, and interventions: modelling transmission dynamics. *Scientific Reports*, 2023.
- [4] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. Time series analysis: forecasting and control. *John Wiley & Sons*, 2015.
- [5] Biobele J Brown, Petru Manescu, Alexander A Przybylski, Fabio Caccioli, Gbeminiyi Oyinloye, Muna Elmi, Michael J Shaw, Vijay Pawar, Remy Claveau, John Shawe-Taylor, et al. Data-driven malaria prevalence prediction in large densely populated urban holoendemic sub-saharan west africa. *Scientific Reports*, 10, 2020.
- [6] Cyril Caminade, Diego Ayala, Thibaud de Chevigny, Olivia Ngou, André Tchouatieu, Florian Girond, Gildas A Yahouedo, Corinne S Merle, Emilie Pothin, Ibrahima Diouf, et al. Climate change and malaria control: a call to urgent action from africa’s frontlines. *Malaria Journal*, 2025.
- [7] Ibrahima Diouf, Abiodun M Adeola, Gbenga J Abiodun, Christopher Lennard, Joyce M Shirinde, Pascal Yaka, Jacques-André Ndione, and Emiola O Gbobaniyi. Impact of future climate change on malaria in west africa: Impact of future climate change on malaria in west africa. *Theoretical and Applied Climatology*, 147, 3, 853–865, 2022.
- [8] Bradley Efron and Robert J Tibshirani. An introduction to the bootstrap new york. *NY: Chapman and Hall*, 1993.
- [9] Devipriya G. Analytical solution of dengue model with maturation delay by homotopy perturbation method. *Asia Mathematika*, 2017.
- [10] Andrew C Harvey and Neil Shephard. 10 structural time series models. *Elsevier*, 1993.
- [11] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 1970.
- [12] Roger W Hoerl. Ridge regression: a historical context. *Technometrics*, 2020.
- [13] Rob J Hyndman and George Athanasopoulos. Forecasting: principles and practice. *OTexts*, 2018.
- [14] Monica C Jackson, Laura Johansen, Cathy Furlong, Abigail Colson, and Kimberly F Sellers. Modelling the effect of climate change on prevalence of malaria in western africa. *Statistica Neerlandica*, 2010.
- [15] Siyan Liu, Dan Lu, Scott L Painter, Natalie A Griffiths, and Eric M Pierce. Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions. *Frontiers in Water*, 2023.
- [16] WJ Martens, Louis W Niessen, Jan Rotmans, Theo H Jetten, and Anthony J McMichael. Potential impact of global climate change on malaria risk. *Environmental health perspectives*, 1995.
- [17] Erin A Mordecai, Jamie M Caldwell, Marissa K Grossman, Catherine A Lippi, Leah R Johnson, Marco Neira, Jason R Rohr, Sadie J Ryan, Van Savage, Marta S Shocket, et al. Thermal biology of mosquito-borne disease. *Ecology letters*, 2019.
- [18] Manigandan P, Alagirisamy K, Lokesh K, Pachiyappan D, and Hajira Banu G. Forecasting the prevalence of covid-19 outbreak in world using the arima and exponential smoothing model. *Asia Mathematika*, 2020.
- [19] Rahila Parveen, Akhtar Hussain Jalbani, Mohsin Shaikh, Kashif Hussain Memon, Saima Siraj, Mairaj Nabi, and Shamshad Lakho. Prediction of malaria using artificial neural network. *Int J Comput Sci Netw Secur*, 2017.

- [20] Manliura Datilo Philemon, Zuhaimy Ismail, and Jayeola Dare. A review of epidemic forecasting using artificial neural networks. *Epidemiology and Health System Journal*, 2019.
- [21] Muhammad A Razi and Kuriakose Athappilly. A comparative predictive analysis of neural networks (nns), nonlinear regression and classification and regression tree (cart) models. *Expert systems with applications*, 2005, 29, 65–74, Elsevier.
- [22] Sadie J Ryan, Catherine A Lippi, Oswaldo C Villena, Aspen Singh, Courtney C Murdock, and Leah R Johnson. Mapping current and future thermal limits to suitability for malaria transmission by the invasive mosquito *Anopheles stephensi*. *Malaria Journal*, 2023.
- [23] Eloise B Skinner, Marissa L Childs, Matthew B Thomas, Jackie Cook, Eleanore D Sternberg, Alphonsine A Koffi, Raphael N’Guessan, Rosine Z Wolie, Welbeck A Oumbouke, Ludovic P Ahoua Alou, et al. Global malaria predictors at a localized scale. *Frontiers in Malaria*, 2024.
- [24] Israel Ukawuba and Jeffrey Shaman. Inference and dynamic simulation of malaria using a simple climate-driven entomological model of malaria transmission. *PLoS computational biology*, 2022.