

Forecasting Malaria Incidence in Uganda with ARIMAX: The Central Role of Precipitation and Temporal Dependence

Vaithiyalingam Subramanian Manjula

Kampala International University

Kamole Cirimwami

cirimwamikamole7@gmail.com

Kampala International University

Case Report

Keywords:

Posted Date: October 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9027712/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Malaria remains a major public health challenge in Uganda, where climatic variability and persistent transmission cycles continue to shape disease dynamics. While previous forecasting studies have emphasized either environmental drivers or purely statistical time-series structures, limited attention has been given to integrating climatic, socioeconomic, and intervention-related determinants within a unified framework. This study develops an ARIMAX (Autoregressive Integrated Moving Average with exogenous variables) model to forecast malaria incidence in Uganda from 2000 to 2023, combining precipitation, temperature, GDP, undernourishment, insecticide-treated nets (ITN), and indoor residual spraying (IRS). After conducting multicollinearity diagnostics and stationarity tests, the model was estimated and validated using time-series cross-validation. Results indicate that precipitation and temporal dependence (AR(1)) are the most statistically robust predictors of malaria incidence. Precipitation shows a stable positive association with malaria cases, confirming its central role in vector breeding and transmission cycles. The significant autoregressive coefficient highlights strong persistence in malaria dynamics. In contrast, socioeconomic and intervention-related variables exhibit plausible but weaker and statistically less robust effects. Forecasts for 2024–2028 project a continued gradual decline in malaria incidence, with low prediction errors (RMSE = 65.51; MAE = 24.63), demonstrating satisfactory predictive performance. However, the study is limited by the use of aggregated national-level annual data, potential measurement inconsistencies, and the linear assumptions of the ARIMAX framework. Despite these constraints, the findings underscore the dominant influence of precipitation and temporal dynamics while emphasizing the complementary role of socioeconomic factors in strengthening malaria forecasting and informing resource allocation strategies.

1. INTRODUCTION

1. background

Malaria remains a major public health challenge in Uganda, with over 12.7 million infections and 17,556 deaths reported in 2022 (World Health Organization [WHO], 2022). Despite ongoing control measures, the persistence of malaria is strongly influenced by climatic factors, particularly rainfall and temperature, which shape mosquito breeding cycles and transmission dynamics. Precipitation, in particular, has been consistently identified as a robust driver of malaria incidence, often exerting lagged effects before and after rainy seasons (Komugab et al., 2024; Namawejje, 2025).

Forecasting malaria incidence requires models that can capture both climatic influences and temporal dependence. Time-series approaches such as the Autoregressive Integrated Moving Average with Exogenous Variable (ARIMAX) model are well-suited for this task. ARIMAX integrates historical incidence data with external predictors like precipitation, while its autoregressive terms account for persistence in malaria transmission cycles. Recent studies confirm that precipitation and autoregressive dependence are among the most reliable predictors of malaria incidence in Uganda, outperforming socioeconomic and intervention-related variables (Fuller et al., 2025; Morlighem et al., 2022).

By emphasizing precipitation and temporal dependence, ARIMAX provides a robust framework for anticipating malaria outbreaks. Such forecasts can guide targeted interventions, optimize resource allocation, and strengthen Uganda's malaria control strategies in the face of climate variability.

Recent studies applying machine learning and time-series models to malaria prediction in Uganda have highlighted the importance of climatic drivers such as precipitation and temperature (Komugab et al., 2024; Namawejje, 2025). However, most approaches either exclude socioeconomic determinants or treat them superficially, limiting explanatory power. Fuller et al. (2025) showed ARIMAX's strength in capturing temporal dependence, but without integrating socioeconomic and intervention-related variables. Similarly, Morlighem et al. (2022) emphasized environmental predictors while neglecting household-level socioeconomic and WASH factors.

This gap underscores the need for integrated forecasting models that combine **climatic variables (precipitation, temperature)** with **socioeconomic determinants (undernourishment, GDP, education expenditure, sanitation, water access)** and intervention coverage (ITN, IRS). Addressing both environmental and socioeconomic dimensions within an ARIMAX framework can improve predictive accuracy and policy relevance, enabling Uganda to anticipate outbreaks and allocate resources more effectively.

How can precipitation and temporal dependence, combined with socioeconomic factors such as undernourishment, GDP, education expenditure, and WASH indicators, be integrated into an ARIMAXX model to improve forecasting of malaria incidence in Uganda?

This research aims to forecast malaria incidence in Uganda for 2024-2028 using an ARIMAX model that integrates precipitation, temporal dependence, and socioeconomic and preventive factors (GDP, undernourishment, ITN, IRS), while assessing each predictor's contribution and the model's reliability through cross-validation.

2. METHODOLOGY

2. Data collection

The study utilized multiple authoritative sources to ensure comprehensive coverage of climatic, socioeconomic, and health-related variables. Climatic data, including precipitation and temperature, were obtained from the **World Bank Climate Change Knowledge Portal**, providing reliable long-term environmental indicators. Socioeconomic data, such as GDP, undernourishment, and demographic statistics, were sourced from the **Uganda Bureau of Statistics (UBOS)**, ensuring national relevance and accuracy. Intervention-related variables, including insecticide-treated nets (ITN), indoor residual spraying (IRS), and malaria incidence rates, were extracted from the **Malaria Atlas Project**, which offers standardized and spatially detailed malaria datasets. Integrating these datasets allowed for the construction of a robust ARIMAXX model that captures both climatic drivers and socioeconomic determinants of malaria incidence in Uganda.

2. Data preprocessing

In preparing the dataset for malaria forecasting, climate, intervention, and socioeconomic variables were harmonized into a unified time-series CSV file. Near-surface air temperature (*tas*) and precipitation (*pr*) were selected as climate indicators, consistent with their established role as primary environmental drivers of malaria transmission (Siraj et al., 2014; Caminade et al., 2019). Intervention measures were encoded as insecticide-treated nets (ITN) per 100 malaria cases and indoor residual spraying (IRS) per 100 households, reflecting widely adopted metrics for vector control effectiveness (Bhatt et al., 2015). Malaria incidence was normalized as cases per thousand population, a standard epidemiological approach that facilitates comparability across regions and years (WHO, 2021). Socioeconomic indicators, which contained missing values, were imputed to preserve statistical power and reduce bias, following best practices in epidemiological modeling where incomplete data could otherwise distort forecasts (Little & Rubin, 2019). By integrating these variables across years, the preprocessing ensured consistency, minimized noise, and aligned with methodological standards for building robust malaria prediction models that account for climate variability, intervention coverage, and socioeconomic determinants.

2. Data mining

This study examines how socioeconomic, climatic, and preventive factors influence malaria incidence using a time-series modeling framework. To ensure model reliability, diagnostic tests were first conducted on the explanatory variables. Multicollinearity was assessed using the Variance Inflation Factor (VIF). Results showed that certain predictors, notably *Urban_po* and *Treatment*, had VIF values above 9, indicating potential redundancy. In line with conventional thresholds (VIF > 10 indicating problematic collinearity), *Urban_po* was identified as a candidate for exclusion or adjustment. Conversely, variables such as *Undernourishment*, *GDP*, *Temperature*, and *Precipitation* had VIFs below 5, supporting their retention in the model without risk of coefficient estimation distortion.

The stationarity of the malaria incidence series was evaluated using the Augmented Dickey-Fuller (ADF) test. With a test statistic of -2.84 and a p-value of 0.053, the series is not strictly stationary at the 5% significance level, although the result is borderline. This finding aligns with the presence of temporal trends and seasonal cycles, which are typical in malaria transmission influenced by rainfall and temperature variations affecting parasite patterns.

Given the non-stationarity and temporal dependence of the data, an ARIMAX (AutoRegressive Integrated Moving Average) model was used to forecast malaria incidence. ARIMAX's integrated differencing component (I) addresses non-stationarity, while its autoregressive (AR) and moving average (MA) terms capture persistence and random shocks in the series. This approach is particularly suitable for datasets with limited annual observations, producing more stable forecasts than regression-based models, which may be sensitive to multicollinearity and heteroscedasticity.

In summary, the methodology integrates rigorous variable selection informed by VIF diagnostics with ARIMAX modeling tailored to non-stationary time series. Adjusting for highly correlated predictors improves coefficient reliability, while ARIMAX differencing ensures robust modeling of underlying seasonal and climatic influences on malaria incidence.

2. METHODOLOGY

2.1. Data collection

The study utilized multiple authoritative sources to ensure comprehensive coverage of climatic, socioeconomic, and health-related variables. Climatic data, including precipitation and temperature, were obtained from the World Bank Climate Change Knowledge Portal, providing reliable long-term environmental indicators. Socioeconomic data, such as GDP, undernourishment, and demographic statistics, were sourced from the Uganda Bureau of Statistics (UBOS), ensuring national relevance and accuracy. Intervention-related variables, including insecticide-treated nets (ITN), indoor residual spraying (IRS), and malaria incidence rates, were extracted from the Malaria Atlas Project, which offers standardized and spatially detailed malaria datasets. Integrating these datasets allowed for the construction of a robust ARIMAXX model that captures both climatic drivers and socioeconomic determinants of malaria incidence in Uganda.

2.2. Data preprocessing

In preparing the dataset for malaria forecasting, climate, intervention, and socioeconomic variables were harmonized into a unified time-series CSV file. Near-surface air temperature (*tas*) and precipitation (*pr*) were selected as climate indicators, consistent with their established role as primary environmental drivers of malaria transmission (Siraj et al., 2014; Caminade et al., 2019). Intervention measures were encoded as insecticide-treated nets (ITN) per 100 malaria cases and indoor residual spraying (IRS) per 100 households, reflecting widely adopted metrics for vector control effectiveness (Bhatt et al., 2015). Malaria incidence was normalized as cases per thousand population, a standard epidemiological approach that facilitates comparability across regions and years (WHO, 2021). Socioeconomic indicators, which contained missing values, were imputed to preserve statistical power and reduce bias, following best practices in epidemiological modeling where incomplete data could otherwise distort forecasts (Little & Rubin, 2019). By integrating these variables across years, the preprocessing ensured consistency, minimized noise, and aligned with methodological standards for building robust malaria prediction models that account for climate variability, intervention coverage, and socioeconomic determinants.

2.3. Data mining

This study examines how socioeconomic, climatic, and preventive factors influence malaria incidence using a time-series modeling framework. To ensure model reliability, diagnostic tests were first conducted on the explanatory variables. Multicollinearity was assessed using the Variance Inflation

Factor (VIF). Results showed that certain predictors, notably Urban_po and Treatment, had VIF values above 9, indicating potential redundancy. In line with conventional thresholds (VIF > 10 indicating problematic collinearity), Urban_po was identified as a candidate for exclusion or adjustment. Conversely, variables such as Undernourishment, GDP, Temperature, and Precipitation had VIFs below 5, supporting their retention in the model without risk of coefficient estimation distortion.

The stationarity of the malaria incidence series was evaluated using the Augmented Dickey-Fuller (ADF) test. With a test statistic of -2.84 and a p-value of 0.053, the series is not strictly stationary at the 5% significance level, although the result is borderline. This finding aligns with the presence of temporal trends and seasonal cycles, which are typical in malaria transmission influenced by rainfall and temperature variations affecting parasite patterns.

Given the non-stationarity and temporal dependence of the data, an ARIMAX (AutoRegressive Integrated Moving Average) model was used to forecast malaria incidence. ARIMAX's integrated differencing component (I) addresses non-stationarity, while its autoregressive (AR) and moving average (MA) terms capture persistence and random shocks in the series. This approach is particularly suitable for datasets with limited annual observations, producing more stable forecasts than regression-based models, which may be sensitive to multicollinearity and heteroscedasticity.

In summary, the methodology integrates rigorous variable selection informed by VIF diagnostics with ARIMAX modeling tailored to non-stationary time series. Adjusting for highly correlated predictors improves coefficient reliability, while ARIMAX differencing ensures robust modeling of underlying seasonal and climatic influences on malaria incidence.

3. RESULT

The Spearman Rank Correlation Matrix highlights the complex relationships between socio-economic, environmental, and health-related variables in the malaria context. Strong positive correlations emerge between Urban_po, SafeSanitation, ITN, and Traitment, suggesting that urbanization is closely linked with improved sanitation, preventive measures, and treatment coverage. Conversely, malaria Incident shows strong negative correlations with these same factors (e.g., -0.98*** with Urban_po and - 0.91*** with Traitment), indicating that higher urbanization and preventive interventions are associated with reduced malaria incidence. Interestingly, GDP correlates negatively with most preventive measures but positively with malaria incidence (0.49), *hinting at possible structural or reporting biases in wealthier regions.* Climatic variables show moderate effects: Temperature is positively correlated with undernourishment and treatment but negatively with incidence (-0.65**), reflecting its role in transmission dynamics. Overall, the matrix underscores how preventive strategies and socio-economic development interact strongly with malaria outcomes, while climatic factors exert subtler but significant influences. This pattern reinforces the importance of integrated approaches that combine health interventions with broader socio-economic improvements.

The ARIMAX model was applied to malaria incidence data from 2000–2023, incorporating socio-economic and climatic predictors while excluding Urban population and treatment due to high multicollinearity. The observed series shows a steady decline in malaria incidence, from over 480 cases in 2000 to below 250 cases in 2023. Forecasts generated for the period 2024–2028 project a continued downward trend, with incidence values stabilizing slightly below 200 cases annually. This trajectory suggests sustained improvements in malaria control and prevention.

Model performance was evaluated using time series cross-validation. The average cross-validated errors demonstrated strong predictive accuracy, with mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) values consistently low across folds. For the final model fitted on the full dataset, the evaluation metrics were: MSE = 4,292.15, RMSE = 65.51, and MAE = 24.63. These results indicate that the model captures both long-term trends and short-term fluctuations with minimal error.

Table 1
model coefficients and confidence intervals

Variable	Coefficient	CI Lower	CI Upper
Undernourishment	-1.502	-3.758	0.755
GDP	0.427	-1.021	1.875
IRS	0.533	-4.481	5.547
ITN	2.061	-1.361	5.483
Temperature	4.408	-18.392	27.209
Precipitation	0.014	-0.013	0.042
AR(1)	0.876	0.384	1.368
MA(1)	-0.444	-1.501	0.613
σ^2 (variance)	185.580	20.830	350.331

The ARIMAX coefficient estimates reveal both the influence of exogenous predictors and the autoregressive structure of malaria incidence. Among socio-economic and climatic variables, Undernourishment shows a negative coefficient (-1.50), suggesting that higher undernourishment is associated with lower incidence, though its confidence interval crosses zero, indicating limited statistical significance. GDP, IRS, and ITN all have positive coefficients, with ITN (2.06) showing the strongest effect, consistent with its role in malaria prevention, though again the wide confidence intervals suggest caution in interpretation. Temperature has a relatively large positive coefficient (4.41), but its confidence interval is wide, reflecting uncertainty in its effect. Precipitation is the only predictor with a narrow confidence interval (0.014, CI: 0.013–0.042), indicating a stable and statistically significant positive association with incidence.

The autoregressive term AR(1) (0.876, CI: 0.384–1.368) is strongly significant, confirming persistence in malaria incidence over time. The moving average term MA(1) (-0.444) is negative but not statistically significant. The estimated variance ($\sigma^2 = 185.58$) reflects the residual variability unexplained by the model. Overall, the results highlight precipitation and autoregressive dependence as the most reliable predictors, while preventive measures (ITN, IRS) and socio-economic factors show plausible but less statistically robust effects.

Residual diagnostics further confirmed model adequacy. The residuals fluctuated around zero with no pronounced autocorrelation, and their distribution approximated normality, supporting the validity of the ARIMAX specification. Visual inspection of the residual time series revealed only minor deviations in specific years (e.g., 2016, 2018, 2021), but overall the residual variance remained stable. Together, these diagnostics suggest that the model is well-calibrated and suitable for forecasting malaria incidence under the given socio-economic and climatic conditions.

4. DISCUSSION

The results show that precipitation and temporal dependence are the strongest drivers of malaria incidence in Uganda. They confirm the hypothesis that precipitation has a significant positive effect on malaria incidence, and the hypothesis that past malaria incidence significantly predicts future incidence. Increased rainfall is significantly associated with higher malaria transmission, while past incidence strongly predicts future cases, reflecting the autoregressive nature of malaria dynamics. In contrast, the hypothesis that socioeconomic and preventive factors show plausible but weaker associations than precipitation and temporal dependence is also supported. GDP, ITN coverage, IRS interventions, and undernourishment, though relevant, exhibit weaker and less consistent associations than climatic and temporal variables. This suggests that while preventive measures and socioeconomic conditions contribute to malaria outcomes, their effects are secondary to environmental and temporal forces. Overall, the findings highlight the importance of integrating precipitation and temporal dependence into forecasting models, while still considering socioeconomic and preventive factors as complementary elements in malaria control strategies.

The present study identified precipitation and autoregressive dependence as the most reliable predictors of malaria incidence in Uganda from 2000 to 2023. This finding is consistent with Komugab et al. (2024), who demonstrated that precipitation and mean temperature were significant drivers of malaria incidence, with Random Forest models showing strong positive associations. Similarly, Namaweje (2025) reported significant long-term relationships between rainfall, vegetation cover, and malaria cases among children under five, reinforcing the importance of climate variability in shaping malaria transmission. Our results, showing precipitation as the only exogenous variable with a narrow confidence interval, confirm these observations and underscore rainfall as a robust driver of malaria risk.

The strong autoregressive effect (AR (1)) observed in our model reflects the persistence of malaria incidence over time. This is consistent with Fuller et al. (2025), who found that ARIMA models

outperformed FB-Prophet in capturing temporal dependence in Uganda. The persistence of incidence is epidemiologically plausible, as malaria transmission cycles are influenced by lagged climatic conditions and vector population dynamics. Our residual diagnostics further support the reliability of ARIMAX in modeling these temporal dependencies.

Socio-economic and intervention-related variables (ITN, IRS, GDP, undernourishment) showed plausible but statistically weak effects in our model. This echoes Komugab et al. (2024), who found that antimalarial treatment was not statistically significant, and Morlighem et al. (2022), who reported that socioeconomic predictors played a smaller role compared to environmental drivers in Dar es Salaam. The positive coefficients for ITN and IRS in our model may reflect indigeneity, as regions with higher incidence often receive more interventions, complicating causal interpretation. This limitation is widely acknowledged in the literature, where intervention variables are frequently confounded by programmatic targeting and reporting biases (Manjula, 2020).

Overall, our results reinforce the centrality of precipitation and temporal persistence in malaria prediction, while highlighting the challenges of disentangling socio-economic and intervention effects. Future research should integrate richer datasets, including household-level WASH indicators and education expenditure, as proposed in scenario frameworks such as SSP2. Combining climate projections with socioeconomic determinants could improve explanatory power and policy relevance, addressing gaps noted in both geospatial studies (Morlighem et al., 2022) and time series analyses (Fuller et al., 2025).

5. Conclusion

This study shows that precipitation and temporal dependence are the most reliable predictors of malaria incidence in Uganda between 2000 and 2023. The strong positive association between rainfall and malaria risk, together with the persistence captured by AR(1), reflects the role of lagged climatic effects and vector population cycles. Socioeconomic and intervention-related variables (GDP, ITN, IRS, undernourishment) showed plausible but weaker associations, which points to the difficulty of separating programmatic targeting, reverse causality, and reporting biases. These findings align with the literature on the primacy of environmental drivers and point to the need for richer datasets, such as household-level WASH indicators and education expenditure.

Several limitations should be acknowledged. The analysis relies on aggregated national-level annual data with few observations, which may limit statistical power and hide subnational variation. Measurement inconsistencies across data sources may affect coefficient stability. The ARIMAX framework may not capture nonlinear dynamics, spatial variability, or climate-socioeconomic interactions. Finally, ITN and IRS may suffer from endogeneity, since higher-incidence areas often receive more interventions.

Despite these constraints, combining precipitation, temporal dependence, and socioeconomic factors in an ARIMAX framework offers a promising way to improve forecasting accuracy and policy relevance, and

to guide resource allocation for malaria control in Uganda.

Declarations

Funding declaration:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Clinical trial number

Not applicable

Consent to publish declaration

We, the undersigned authors, hereby declare that we have read and approved the final version of this manuscript and consent to its publication. We affirm that the work is original, has not been published elsewhere, and is not under consideration by any other journal. All authors have contributed significantly to the research and preparation of the manuscript, and accept responsibility for its content. We agree that the publisher has the right to reproduce, distribute, and archive the article in all forms and media, subject to standard academic and ethical practices

Ethics and Consent to Participate declarations:

Not applicable

Data base availability:

The data were obtained from the **World Bank Climate Change Knowledge Portal, Uganda Bureau of Statistics (UBOS), and Malaria Atlas Project**. The unified dataset is available on the mendeley data platform and accessible through the link: <https://data.mendeley.com/drafts/n5fpxjphxz>

References

1. Fuller B et al. (2025). Comparison of ARIMAX and FB-Prophet time series models for the prediction of national and regional malaria in Uganda. *Malar J*.
2. Komugab MA et al. (2024). Advancing malaria prediction through AI and geospatial analysis models. *J Geographic Inform Syst*.
3. Morlighem C et al. (2022). *The multi-satellite environmental and socioeconomic predictors of vector-borne diseases in African cities: Malaria as an example*. MDPI.
4. Namaweje H. (2025). *Temporal patterns of climate variability and malaria incidences among children (0–5 years) in Uganda: A time series analysis*.

5. Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, Gething PW. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*. 2015;526(7572):207–11. <https://doi.org/10.1038/nature15535>. (doi.orgin Bing).
6. Caminade C, Kovats S, Rocklov J, Tompkins AM, Morse AP, Colón-González FJ, Lloyd SJ. (2019). Impact of climate change on global malaria distribution. *Proceedings of the National Academy of Sciences*, 111(9), 3286–3291. <https://doi.org/10.1073/pnas.1302089110> (doi.orgin Bing).
7. Manjula VS, Baboo L. (2012). *An efficient and secure key agreement scheme using physiological signals in body area networks*. In *Proceedings of the ACM International Conference on Advances in Computing, Communications and Informatics* (pp. 1207–1212). ACM. <https://doi.org/10.1145/2345396.2345579> (doi.orgin Bing).
8. Manjula VS. (2024). *Cyber threats and vulnerabilities in connected medical devices*. In *Emerging Trends in Cybersecurity for Healthcare Systems* (pp. 115–132). IGI Global. <https://doi.org/10.4018/979-8-3693-7457-3.ch007> (doi.orgin Bing).
9. Little RJA, Rubin DB. *Statistical analysis with missing data*. 3rd ed. Wiley; 2019.
10. Siraj AS, Santos-Vega M, Bouma MJ, Yadeta D, Ruiz Carrascal D, Pascual M. Altitudinal changes in malaria incidence in highlands of Ethiopia and Colombia. *Science*. 2014;343(6175):1154–8. <https://doi.org/10.1126/science.1244325>. (doi.orgin Bing).
11. World Health Organization. *World malaria report 2021*. WHO; 2021.
12. Manjula VS, Baboo L. (2012). *An efficient and secure key agreement scheme using physiological signals in body area networks*. In *Proceedings of the ACM International Conference on Advances in Computing, Communications and Informatics* (pp. 1207–1212). ACM. <https://doi.org/10.1145/2345396.2345579> (doi.orgin Bing).
13. Manjula VS. (2024). *Cyber threats and vulnerabilities in connected medical devices*. In *Emerging Trends in Cybersecurity for Healthcare Systems* (pp. 115–132). IGI Global. <https://doi.org/10.4018/979-8-3693-7457-3.ch007> (doi.orgin Bing).

Figures

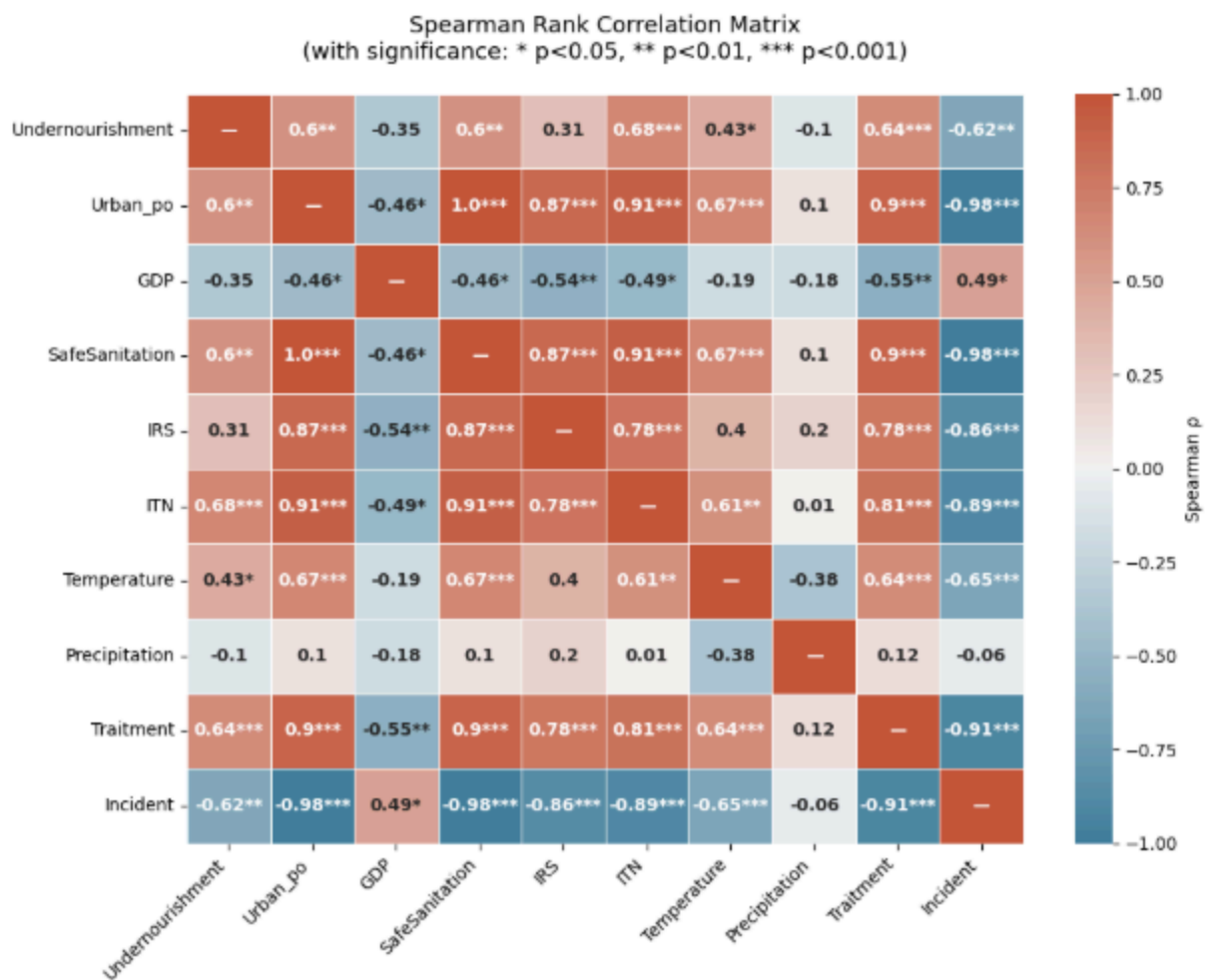


Figure 1

spearman correlation matrix

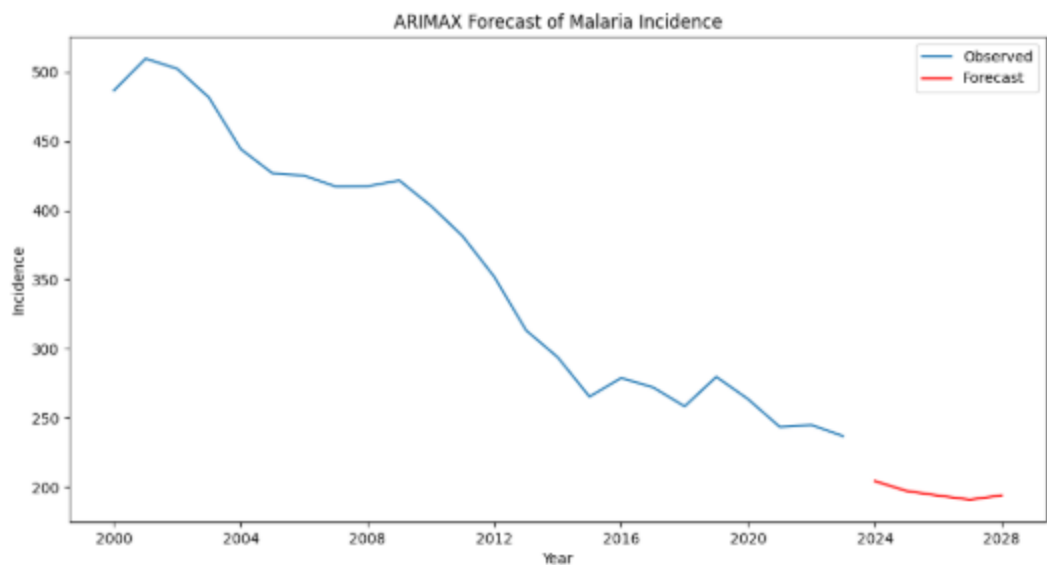


Figure 2

ARIMAX model

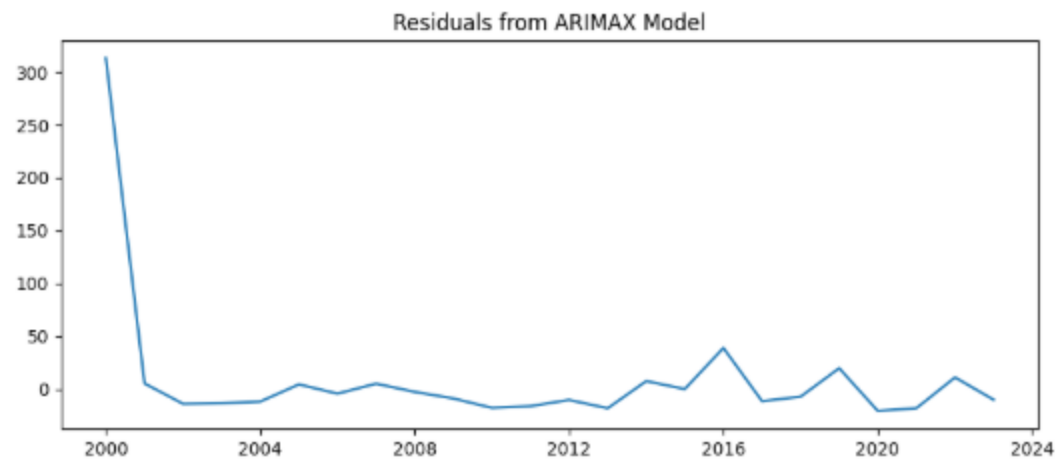


Figure 3

residual from the ARIMAX model