Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/140210

Climate-Informed Malaria Prediction Models: A Bayesian Approach For South African Endemic Provinces

Makwelantle Asnath Sehlabana[1,*], *Daniel Maposa*[1], *Alexander Boateng*[2] *and Sonali Das*[3]

[1]Department of Statistics and Operations Research, University of Limpopo, South Africa
[2]Department of Mathematics and Computer Science, Modern College of Business and Science, Muscat, Sultanate of Oman
[3]Department of Business Management, University of Pretoria, South Africa

Received: 30 Jul. 2024, Revised: 2 Oct. 2024, Accepted: 19 Jan. 2025
Published online: 1 May 2025

Abstract: In this study, we predict malaria cases using climate factors and Bayesian methods. Climate change plays a pivotal role in determining both the geographic spread and severity of malaria outbreaks. Recent research underscores that climate-related factors outweigh other contributors, such as epidemiological, socio-economic, and environmental factors, in the resurgence of malaria cases. The coronavirus disease of 2019 (COVID-19) pandemic have caused a setback in the global strides made towards malaria control and elimination. South Africa has not met its malaria elimination targets, despite strategic plans like the National Malaria Elimination Strategic Plan (NMESP), which emphasises strengthening surveillance systems. Researchers are developing malaria forecasting and prediction models incorporating climate factors, primarily using time series and machine learning techniques. While time series models exhibit shortcomings in long-term forecasting, machine learning models have shown promise in prediction but did not prove granularity in delineating critical malaria seasons or providing climate-specific predictions. This study seeks to edify these models using a Bayesian framework to predict malaria in South Africa's endemic provinces based on climate and environmental factors. The study found that malaria transmission is high in regions with temperatures of 20-30°C, rainfall of 0-200 mm, and normalised difference vegetation index (NDVI) levels of 0.5-0.8, predicting 200 to 1000 malaria cases in these conditions. The Ehlanzeni district in Mpumalanga and the uMkhanyakude district in KwaZulu-Natal are identified as high-risk areas with elevated malaria counts. Targeted prevention and control measures are recommended for these districts. Future research should explore malaria prediction using subjective informative prior distributions for deeper insights.

Keywords: Bayesian framework, climate change, climate factors, malaria control, malaria elimination

1 Introduction

The effectiveness of the global response to malaria faces a critical threat from climate change. Numerous studies indicate that various climate and environmental factors play a significant role in influencing malaria transmission, leading to its resurgence worldwide. In South Africa, malaria transmissions primarily occur during the warm and rainy summer months from September to May, with considerably fewer cases reported in June, July, and August [1]. Research has shown that the transmission rate of malaria is moderate within a temperature range of approximately 3°C to 24°C, but significantly increases at higher temperatures, around 27°C and above [2]. Moreover, rainfall patterns have emerged as robust predictors of seasonal malaria incidence, particularly in regions like Limpopo province [3]. Climate-related factors are identified as major contributors to the resurgence of malaria, surpassing the influence of other factors such as epidemiological, socio-economic, and environmental factors [4]. The persistent outbreaks of malaria in South Africa are attributed to natural forces, including climate change, and the movement of migrant human populations [5].

Similar to other malaria-endemic countries, South Africa is actively pursuing the goal of malaria elimination. While the World Health Organization (WHO)'s E-2020 initiative aimed to eliminate malaria in 21 countries identified as having the potential to achieve this by 2020, South Africa set its sights on eliminating malaria by 2018 [6]. However, adverse effects of climate change, such as increased rainfall, temperatures, and milder winter seasons, led to a rise in malaria cases

* Corresponding author e-mail: asnath.sehlabana@ul.ac.za

in 2018 [7]. Consequently, since the target of malaria elimination by 2018 was not met, South Africa revised its strategic plan, which aimed for malaria elimination by 2023. Under the revised plan, Limpopo province was in the control phase, Mpumalanga province in the early stages of malaria elimination, and KwaZulu-Natal province was the furthest along in implementing elimination strategies[8]. However, the target was still not met.

The progress in malaria control and elimination globally was significantly impeded by the coronavirus disease of 2019 (COVID-19) pandemic. According to a report by the WHO in 2021, there was a steady reduction in malaria incidence and mortality between 2000 and 2019. However, the disruptions to malaria services caused by the COVID-19 pandemic led to a reversal of this trend. In the WHO report for African region, malaria incidence increased from 222 per 1000 population at risk to 232 per 1000 population at risk in 2020, while malaria mortality rose from 534,000 in 2019 to 602,000 in 2020. Furthermore, plans were in place to distribute a total of 171 million Insecticide-Treated Mosquito Nets (ITNs) in 2021. However, due to the disruptions in malaria services caused by COVID-19, only 128 million (75%) ITNs were actually distributed. Consequently, there was an estimated 247 million malaria cases globally in 2021, marking an increase of 13 million cases compared to 2020. The WHO African region accounted for a staggering 95% of global malaria cases and 96% of deaths related to malaria [9].

In 2022, approximately 249 million malaria cases were reported globally, spanning 85 countries and regions endemic to malaria. This marked a significant uptick from 2021, with prominent contributors to this surge including Pakistan (reporting around 2.1 million cases), Ethiopia (with about 1.3 million cases), Nigeria (also tallying approximately 1.3 million cases), Uganda (reporting about 597,000 cases), and Papua New Guinea (with approximately 423,000 cases) [10]. The global malaria case count in 2022 even surpassed pre-COVID-19 pandemic levels seen in 2019 [11]. In South Africa, the Notifiable Medical Conditions System (NMC-SS) recorded a total of 5,813 case notifications between September 2022 and August 2023. Of these notifications, 4,137 originated from malaria-endemic provinces, with Limpopo province accounting for 49%, Mpumalanga province for 15%, and KwaZulu-Natal for 7%. Intriguingly, about 1,095 case notifications came from Gauteng province, which is not considered a malaria-endemic region [12].

Objective two of the South African National Malaria Elimination Strategic Plan (NMESP) prioritises strengthening the surveillance system to bolster the country's elimination efforts [12]. Integrating climate data into malaria modelling emerges as a pivotal strategy, moving beyond mere estimations and predictions to become a potent system that not only anticipates the disease but also provides the foresight necessary for informed surveillance and effective defence against this deadly threat. Many malaria prediction models rely on forecasted climate data, while others serve as forecasting models themselves. Traditional time series or forecasting models typically provide an overarching malaria trend expected in the upcoming years. For instance, Kim et al. [2] developed a malaria prediction model utilising both observed and seasonal climate forecasts, demonstrating effective short-term prediction capabilities of one to two weeks, with reasonable performance extending up to 16 weeks. However, studies such as Wang et al. [13] have highlighted the limitations of forecasting models compared to other models like ensemble architecture models. In the case of South Africa, Landman et al. [3] utilised forecasted seasonal rainfall totals and seasonal mean maximum temperatures, along with malaria data, to devise a malaria prediction model specifically for Limpopo province. However, this study was constrained by its reliance on forecast models and underscored the importance of accessing climate data from robust monitoring systems rather than relying solely on rainfall predictions.

Recently, researchers have turned to machine learning algorithms to develop malaria prediction models, as evidenced by studies conducted by Adamu [14], Lee et al. [15], and Ntikura et al. [16]. While Lee et al.[15] focused on clinical information for their model, both Adamu [14] and Ntikura et al. [16] incorporated climate factors into their malaria prediction models. However, these studies primarily demonstrated the predictive capabilities of their models without delving into the stratification of important malaria seasons or providing specific predictions related to the predictor variables or climate factors. This study aims to fill this gap by developing a malaria prediction model using a Bayesian framework. Within this framework, the robustness of objective prior distributions will be assessed, and the malaria prediction model will be utilised to predict future malaria cases relative to climate change factors such as rainfall, daytime, and nighttime temperatures. Additionally, an environmental factor, normalised difference vegetation index (NDVI), will be included as one of the predictors in the malaria prediction model. The outcomes of this study will support malaria control and elimination strategic plans by serving as an early malaria warning system. Furthermore, the predictions generated by this study may offer valuable insights for the distribution of ITNs and other malaria preventive resources or measures in endemic provinces.

2 Material and Methods

2.1 Study frame and data sources

The study investigates malaria cases in three South African provinces: KwaZulu-Natal, Limpopo, and Mpumalanga. KwaZulu-Natal, the second most populous province, features three malaria-endemic districts (King Cetshwayo,

uMkhanyakude, and Zululand) with moderate climate. Limpopo, known for its warmth, has two malaria-endemic districts (Mopani and Vhembe) and diverse landscapes. Mpumalanga, despite being the second smallest province, experiences notable population growth and hosts one malaria-endemic district (Ehlanzeni), characterised by warm temperatures and plateau grasslands.

The study relies on secondary data sources, with malaria data sourced from the South African National Department of Health, and climate data obtained from the South African Weather Service (SAWS). The malaria dataset includes parasitological confirmed cases, accounting for both locally transmitted and imported instances. Climate data comprises daily rainfall measurements and hourly temperature readings, which are aggregated to derive monthly averages. Furthermore, the study incorporates the NDVI, accessed through EarthExplorer (EE), an online platform supported by the United States Geological Survey (USGS). The dataset spans the period from 2018 to 2022.

2.2 Analytical techniques

Bayesian analysis relies on two core elements: the sampling model, which captures vital information from the data under examination, and the prior distribution, integrating existing knowledge or information about the unknown parameters into the sampling model. These components are combined using Bayes theorem to produce the posterior distribution. From this posterior distribution, a predictive distribution is derived, enabling the prediction of future observations of the response variable, which in this study pertains to count data. Given the nature of the response variable as count data, the selection of the sampling model involves exploring various count models. Additionally, for the prior distribution, this study examines two approaches: Jeffreys prior and the conjugate prior distribution of the sampling model. These considerations are pivotal in constructing a robust Bayesian framework for the analysis.

2.2.1 Count models

In this study, we explore several models for analysing count data. We begin with the Poisson regression model, a common choice that assumes the mean count is a linear function of covariates and requires equi-dispersion (where the mean equals the variance). The generalised Poisson (GP) model extends this by allowing the variance to vary, thus providing more flexibility to capture data dispersion. Similarly, the negative binomial model extends the Poisson model by including a dispersion parameter to handle additional variation and unobserved heterogeneity. The hurdle model, developed by Mullay [17], diverges from these GLMs by addressing excess zeros in the data through a two-component approach, one component follows a zero-truncated distribution (such as Poisson, negative binomial or geometric), and the other follows a logit model for the zero mass component. In contrast, zero-inflated models also employ a two-component approach but differ by having a point mass component consisting solely of zeros and a count component containing both zeros and non-zero counts. This model combines a logit model to determine whether an observation belongs to the point mass or count component with a model for count observations. Lastly, the mixture model framework represents different subpopulations within a larger population without explicitly identifying each subpopulation. These models enhance the flexibility and accuracy of statistical analysis for count data by accommodating varying dispersion and excess zeros.

2.2.2 Prior distributions and posterior distributions

i) Jeffreys Prior Distribution

Non-informative priors, also known as conventional default choices, are often employed when prior information is either insufficient or ambiguous, making it challenging to elicit an adequate subjective prior distribution. This idea of selecting a prior based on convention is credited to Jeffreys [18]. What sets Jeffreys' prior apart from other prior selection methods is its connection to information theory, and most importantly, its parameterization invariance. This means that the relative probability assigned to a probability space using Jeffreys' prior remains the same regardless of the parameterization used to define it. Jeffreys prior is defined as follows:

$$\pi(\theta) \propto |I(\theta)|^{1/2}, \tag{1}$$

where $|\cdot|$ represents the determinant and $I(\theta)$ denotes the expected Fisher information matrix. The Fisher information matrix, based on the likelihood function $f(y_i|\theta)$, is given by:

$$I(\theta) = -E_{y_i|\theta}\left[\frac{d^2 \ln f(y_i|\theta)}{d\theta^2}\right]. \tag{2}$$

Now, considering the negative binomial density function of y_i defined as:

$$f(n|\theta) = \begin{cases} \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}, & n = r, r+1, r+2, \ldots, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where the number of successes r is constant, the sample size n is stochastic, the probability of success is denoted by θ, and the mean is given by r/θ. This implies that $E_{n|\theta}(r) = r$, and $E_{n|\theta}(n-r) = E_{n|\theta}(n) - r = \frac{r}{\theta} - r$. Therefore, the Fisher information is computed as:

$$I(\theta) = --E_{n|\theta}\left[\frac{d^2 \ln \pi(n|\theta)}{d\theta^2}\right] = \frac{E_{n|\theta}(r)}{\theta^2} + \frac{E_{n|\theta}(n-r)}{(1-\theta)^2} = \frac{r}{\theta^2} + \frac{r/\theta - r}{(1-\theta)^2} = \frac{r}{\theta^2(1-\theta)}.$$

Thus, the Jeffreys prior for the negative binomial distribution is given by:

$$\pi(\theta) \propto |I(\theta)|^{1/2} \propto \left[\theta^2(1-\theta)^{-1}\right]^{1/2} \propto \theta^{-1}(1-\theta)^{-1/2}, \tag{4}$$

recognising that this is a Beta(a,b) distribution, where $a = \varepsilon$ and $b = 0.5$ as $\varepsilon \to 0$. Multiplying the likelihood function in equation (3) by the Jeffreys prior in equation (4) gives the posterior:

$$\pi(\theta|y_i) = \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}\theta^{-1}(1-\theta)^{-1/2} \propto \theta^{r-1}(1-\theta)^{s-1/2}, \tag{5}$$

where $s = n - r$. This corresponds to a Beta(a,b) distribution with $a = r$ and $b = s + 0.5$. This density is proper when $r \geq 1$.

ii) Conjugate Prior Distribution

Conjugate priors play a crucial role in the Bayesian framework just like any other type of prior distribution. One notable advantage of employing a conjugate prior distribution is that it yields a posterior with the same functional form and identical properties as the prior distribution. This characteristic simplifies data analysis, computations, and interpretation of the model. Conjugate priors are applicable to exponential family models, including the normal, Poisson, gamma, binomial, and negative binomial distributions. The distribution $Beta(\alpha,\beta)$ serves as a conjugate prior for the negative binomial model. However, the parameter values for θ are unknown. According to Bolstad [19], in cases where there is no information available about certain parameters in a model, a prior that assigns equal weight to all possible values of the parameter of interest can be employed. This type of prior is represented by a uniform distribution, $U[0,1]$. Additionally, for a $Beta(\alpha,\beta)$ prior where $\alpha = 1 = \beta$, the uniform prior is equivalent to $Beta(1,1)$. The uniform prior is expressed as:

$$\pi(\theta) = \frac{1}{Beta(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{6}$$

Therefore, the posterior is given by:

$$\pi(\theta|y_i) = \frac{1}{Beta(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}. \tag{7}$$

Now, considering that $\alpha = 1 = \beta$ and recognising that the part not dependent on θ is constant for all values of θ, the posterior in equation (7) simplifies to:

$$\pi(\theta|y_i) = \frac{1}{Beta(1,1)}\binom{n-1}{r-1}\theta^r(1-\theta)^{n-r} \propto \theta^r(1-\theta)^{n-r}. \tag{8}$$

We recognise this as a $Beta(a,b)$ distribution with $a = r+1$ and $b = n-r+1$. This implies that $U[0,1] = Beta(1,1) \Rightarrow \pi(\theta) = 1$ for $0 \leq \theta \leq 1$. This kind of prior is also known as a flat prior.

2.2.3 Posterior predictive distribution

The probability distributions for parameters sometimes enable the calculation of probabilities for future observations of the response variable. Like other unknowns in the Bayesian framework, these probabilities are represented by a probability distribution known as a predictive probability distribution. Predictive distributions are valuable for communicating the

results and aid in making informed decisions. Consider an independent future observation y. The predictive distribution of y is defined as:

$$g(y \mid x) = \frac{f(y,x)}{f(x)} = \frac{\int f(y,x|\theta)\pi(\theta)d\theta}{\int f(x|\theta)\pi(\theta)d\theta} = \frac{\int f(y|\theta)f(x|\theta)\pi(\theta)d\theta}{\int f(x|\theta)\pi(\theta)d\theta} = \int f(y|\theta)\pi(\theta|x)d\theta. \tag{9}$$

Given that the observations on x follow the negative binomial model, future independent observations will similarly adhere to this distribution. Employing Jeffreys' prior in equation (9), the predictive probability distribution is expressed as:

$$g(y \mid n) = \int f(y|\theta)\pi(\theta|x)d\theta = \int \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}\theta^{r-1}(1-\theta)^{s-1/2}d\theta \propto \theta^{2r-2}(1-\theta)^{y-2r+n+1/2}. \tag{10}$$

We recognise equation (10) as a beta distribution, $Beta(a,b)$, where $a = 2r$ and $b = y-2r+n+1/2$. Suppose a uniform prior, $U[0,1]$ (approximating Beta$(1,1)$), is used instead of a Jeffreys prior. In this case, the predictive probability distribution is expressed as:

$$g(y \mid n) = \int f(y|\theta)\pi(\theta|x)d\theta = \int \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}\theta^{\alpha-1+r}(1-\theta)^{\beta-1+s}d\theta \propto \theta^{2r+\alpha-1}(1-\theta)^{\beta+y+n-2r}. \tag{11}$$

We recognize equation (11) to be proportional to a beta distribution, $Beta(a,b)$, where $a = \alpha+2r$ and $b = \beta+y+n-2r+1$.

Graphical representations are a powerful tool for presenting predicted values derived from a posterior distribution. In this study, we employ effect plots generated using SAS software to illustrate predicted responses as a function of specific covariates while keeping other covariates constant. Examples of these effect plots include contour plots, fit panels, interaction plots, and slice fits.

2.2.4 Model selection and comparison

In model evaluation, the concept of model predictive precision, as highlighted by Turkman et al. [20], is the prominent idea. In this section, we discuss predictive performance measures known as information criteria (IC). These criteria come highly recommended both theoretically and practically. According to the specific IC definition, lower values signify better model suitability. Let's consider y to represent future observations and x to represent the observed data.

i) Akaike Information Criterion (AIC)
Akaike [21] introduced a criterion where predictive accuracy is defined as:

$$\tilde{A}_{AIC} = \ln f(x|\hat{\theta}) - \rho, \tag{12}$$

where $\hat{\theta}$ represents the maximum likelihood estimate of θ. The AIC measure of information is obtained through a linear transformation of $\tilde{A}_{AIC}$ in equation (12), such that:

$$AIC = -2\ln f(x|\hat{\theta}) + 2\rho. \tag{13}$$

ii) Bayesian Information Criterion (BIC)
The BIC, credited to Schwarz [22], selects models by approximating a marginal distribution of the data, particularly with large samples. This distribution is represented as:

$$P(x) = E_{h(\theta)}[f(x|\theta)]. \tag{14}$$

In contrast to the AIC, the BIC integrates the sampling distribution with the prior density. It is defined as:

$$BIC = -2\ln f(x|\hat{\theta}) + \rho\ln n, \tag{15}$$

where $\hat{\theta}$ represents the maximum likelihood estimate of θ.

iii) Deviance Information Criterion (DIC)
The DIC criterion, introduced by Spiegelhalter et al. [23], modifies the measures of predictive accuracy $\tilde{A}_{AIC}$ in equation (12) by replacing the maximum likelihood estimate $\hat{\theta}$ with a Bayesian estimate $\bar{\theta}$ and substituting the number of parameters with an effective model dimension P_D. This criterion is expressed as:

$$DIC = D(\bar{\theta}) + 2P_D = \overline{D(\theta)} + P_D = 2\overline{D(\theta)} - D(\bar{\theta}), \tag{16}$$

where $D(\bar{\theta})$ is Bayesian deviance, P_D is the effective number of parameters, and $\overline{D(\theta)}$ is the posterior expectation.

3 Results and Discussion

3.1 Exploratory data analysis

Within this section, we offer a detailed description of the data employed in our study. All variables, accompanied by the corresponding codes used for representation within the software employed, are outlined in Table 1. Additionally, a comprehensive summary of the data through both concise summary tables and visually engaging presentations is provided.

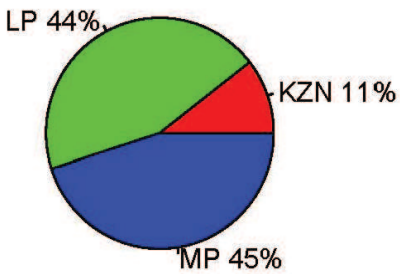
Table 1: Variables and Code Names

Variable	Dataset Code and Description	Data Type
District	District: Districts in the three Provinces of interest. These districts encompass all regions within KwaZulu-Natal, Limpopo, and Mpumalanga: Amaj: Amajuba district, Capr: Capricorn district, Ehlan: Ehlanzeni district, Gert: Gert Semenya district, Harry G: Harry Gwala district, iLembe: iLembe district, KingC: King Cetshwayo district, Mop: Mopani district, Nkang: Nkangala district, Sekh: Sekhukhune district, Ugu: Ugu district, uMgun: uMgungundlovu, Umkha: uMkhanyakude, Umzin: uMzinyathi, Uthu: uThukela, Vhem: Vhembe district, Water: Waterberg district, Zulu: Zululand district	Character
Malaria counts	Mal: Malaria counts	Number
Maximum Temperatures	MaxTemp: Temperature during the day (in $^\circ$ C)	Number
Minimum Temperatures	MinTemp: Temperature during the night (in $^\circ$ C)	Number
Month	Month: Months of the year, where: Jan: January, Feb: February, Mar: March, Apr: April, May: May, Jun: June, Jul: July, Aug: August, Sep: September, Oct: October, Nov: November, Dec: December	Character
Normalised Difference Vegetation Index	NDVI: Normalised Difference Vegetation Index, which is a widely used metric for quantifying the health and density of vegetation using sensor data.	Number
Province	Province: The three malaria endemic provinces, namely: KZN: KwaZulu-Natal, LP: Limpopo, MP: Mpumalanga	Character
Rainfall	Rain: Rainfall (in mm)	Number
Year	Year: This study focuses on events that occurred between 2018 and 2022.	Character

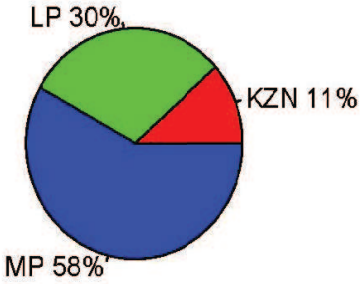
Table 2: Summary Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
Mal	1044	58.35	185.99	0	3005.00
MaxTemp	855	26.18	3.40	0	37.60
MinTemp	854	12.66	5.02	-0.20	22.30
Rain	924	59.95	67.22	0	570.50
NDVI	1044	0.53	0.14	0.24	0.79

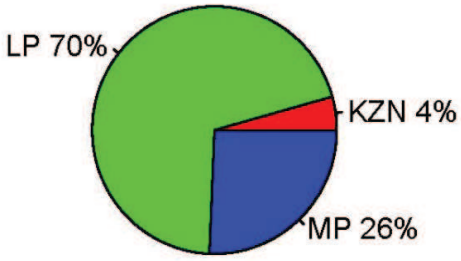
Table 2 presents summary statistics for key numeric variables in the dataset. The table includes data on monthly malaria counts, monthly average daily minimum and maximum temperatures, monthly rainfall, and NDVI values across three provinces. For malaria counts, there were 1044 observations with an average of 58 cases per month and a large standard deviation of 186, indicating significant dispersion. Average daytime temperature was 26° C with a small standard deviation, suggesting data clustered around the mean. Night-time temperature averaged 13° C with a slightly higher standard deviation. Monthly rainfall averaged 60 mm with a high standard deviation and range, indicating widely dispersed observations. NDVI values averaged 0.5, indicating substantial vegetation coverage, with a smaller standard deviation, suggesting tighter clustering around the mean.



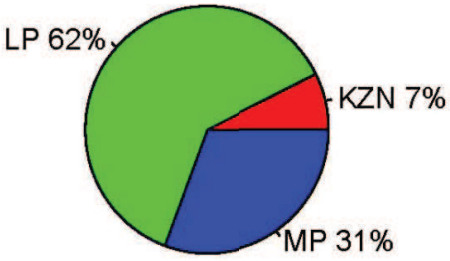
1a: Malaria cases across provinces in 2018



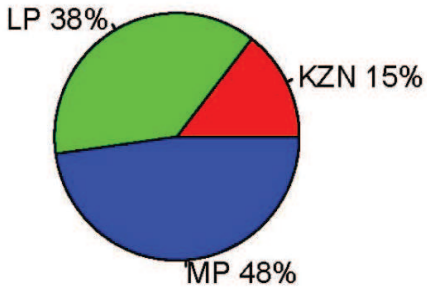
1b: Malaria cases across provinces in 2019



1c: Malaria cases across provinces in 2020



1d: Malaria cases across provinces in 2021



1e: Malaria cases across provinces in 2022

Fig. 1: Malaria cases across provinces

Figures 1 illustrate the malaria profile in provinces endemic to malaria from 2018 to 2022. Throughout the study period, a total of 60,913 malaria cases were recorded across the three provinces. The highest number of malaria cases, totalling 20,528, was reported in 2018. Each year, KwaZulu-Natal consistently contributed less than 16% of the total malaria cases. However, the dynamics shifted between Limpopo and Mpumalanga. In 2018, 2019, and 2022, Mpumalanga took the lead, accounting for 45% (9245), 58% (7611), and 48% (3198) of cases across the three provinces. Conversely, Limpopo surged ahead in 2020 and 2021 representing 70% (8514) and 62% (5213) of cases, respectively. During these years, Limpopo emerged as the most malaria-endemic province, surpassing both KwaZulu-Natal and Mpumalanga.

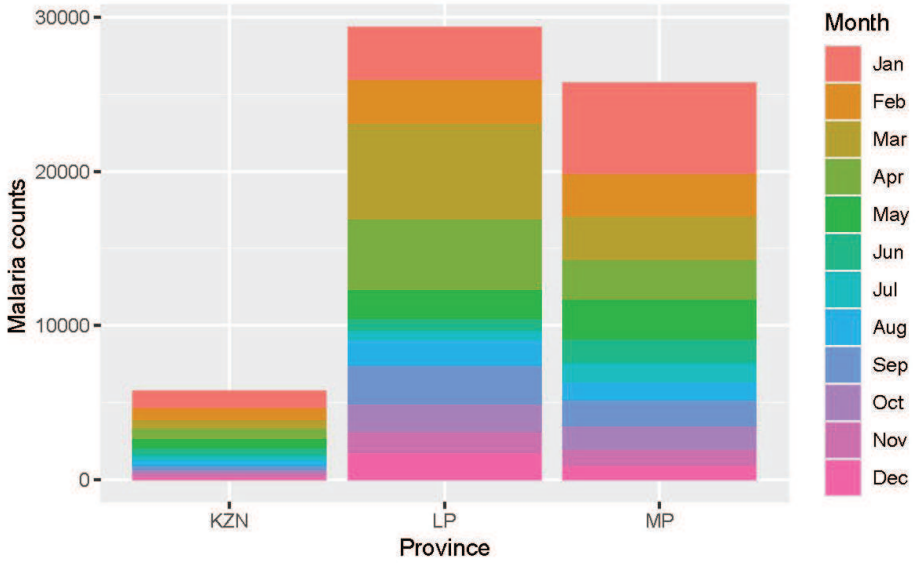


Fig. 2: Monthly distribution of malaria cases across provinces

In Figure 2, the stacked bar graph depicts the monthly distribution of malaria cases across the three provinces. It's notable that KwaZulu-Natal and Mpumalanga report the highest number of cases in January, followed by relatively consistent numbers from February to May. In contrast, Limpopo shows a peak in cases during March, followed by April. Generally, malaria transmission peaks between January and April across the provinces, gradually declining from May onwards.

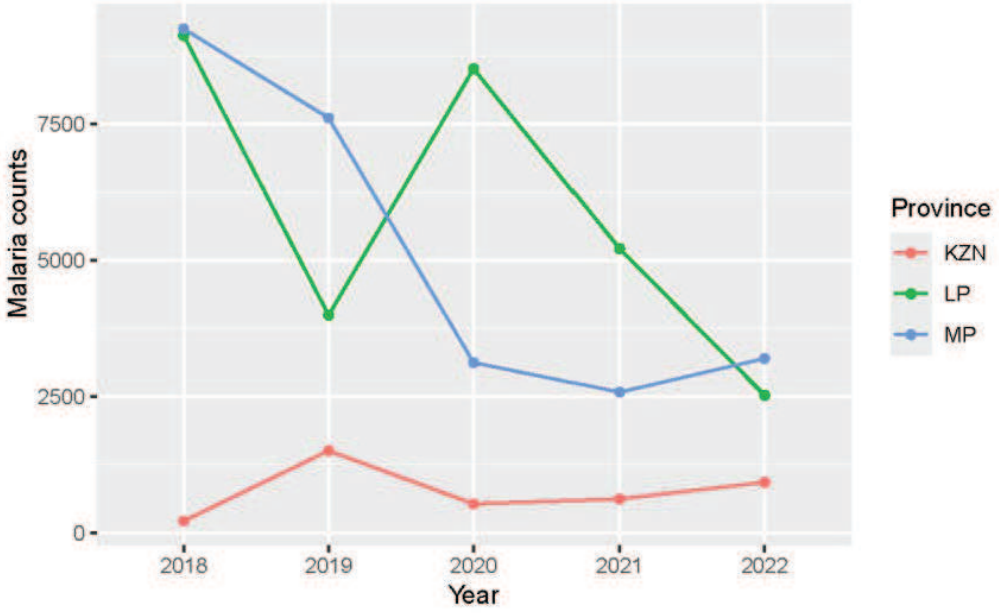


Fig. 3: Malaria patterns across provinces over time

In Figure 3, the malaria trends over the years are presented. KwaZulu-Natal reported its highest number of malaria cases (approximately 1250) in 2019 and its lowest in 2020 (about 625), with cases gradually increasing between 2020 and 2022. Limpopo and Mpumalanga both recorded their highest number of malaria cases in 2018 (more than 8750). However, there was a gradual decrease in malaria cases in Mpumalanga between 2018 and 2021. Conversely, in Limpopo, malaria cases drastically dropped in 2019 to about 3750, spiked in 2020 to about 8740, and then decreased sharply between 2021 and 2022. Mpumalanga and Limpopo reported their lowest number of malaria cases, approximately 2500 in 2021 and 2022, respectively. Overall, Mpumalanga has shown a considerable decrease in malaria transmission over the years.

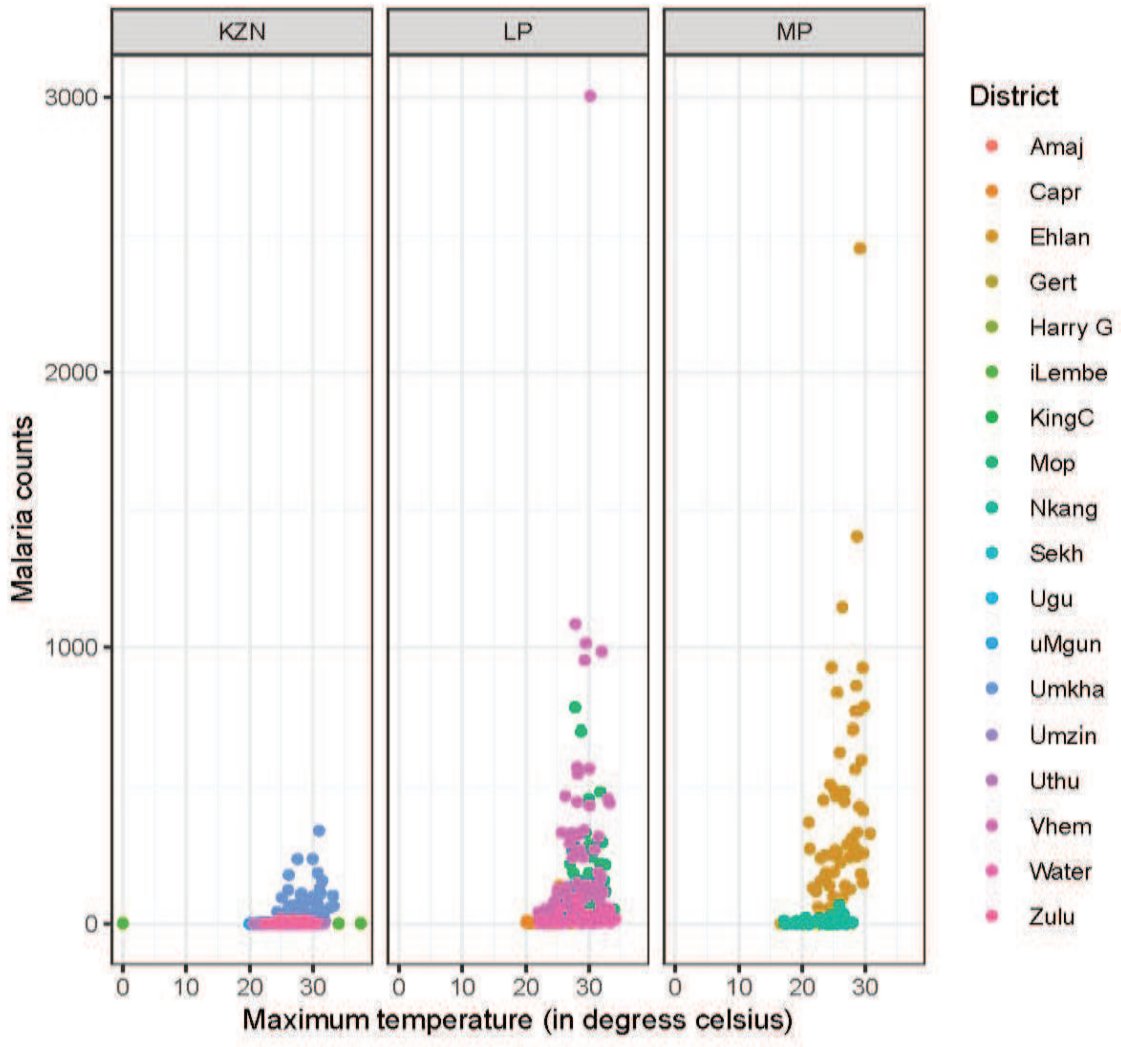


Fig. 4: Malaria counts by maximum temperature in three Provinces

The scatter plot in Figure 4 illustrates the relationship between the average daily maximum temperature (in degrees Celsius) and malaria counts across three provinces: KwaZulu-Natal (KZN), Limpopo (LP), and Mpumalanga (MP). The data points are color-coded according to districts within these provinces, allowing for a clear visual distinction of regional variations. In KwaZulu-Natal, malaria counts remain relatively low across all temperature values, with no extreme outbreaks observed. However, most cases are concentrated in the uMkhanyakude district. In contrast, Limpopo exhibits a notable concentration of higher malaria counts, particularly when maximum temperatures range between 25°C

and 30°C, suggesting a potential relationship between warmer temperatures and increased malaria incidence. The majority of cases in Limpopo are recorded in the Vhembe and Mopani districts. Similarly, Mpumalanga shows instances of high malaria counts, especially as temperatures approach 30°C, indicating that the region may be more susceptible to malaria outbreaks under higher temperatures. The highest number of cases in Mpumalanga occurs in the Ehlanzeni district. The distribution of malaria cases varies across districts, as highlighted by the color-coded legend, with some districts experiencing significantly higher case numbers than others. This variation suggests that additional environmental or socio-economic factors may influence malaria transmission beyond temperature alone.

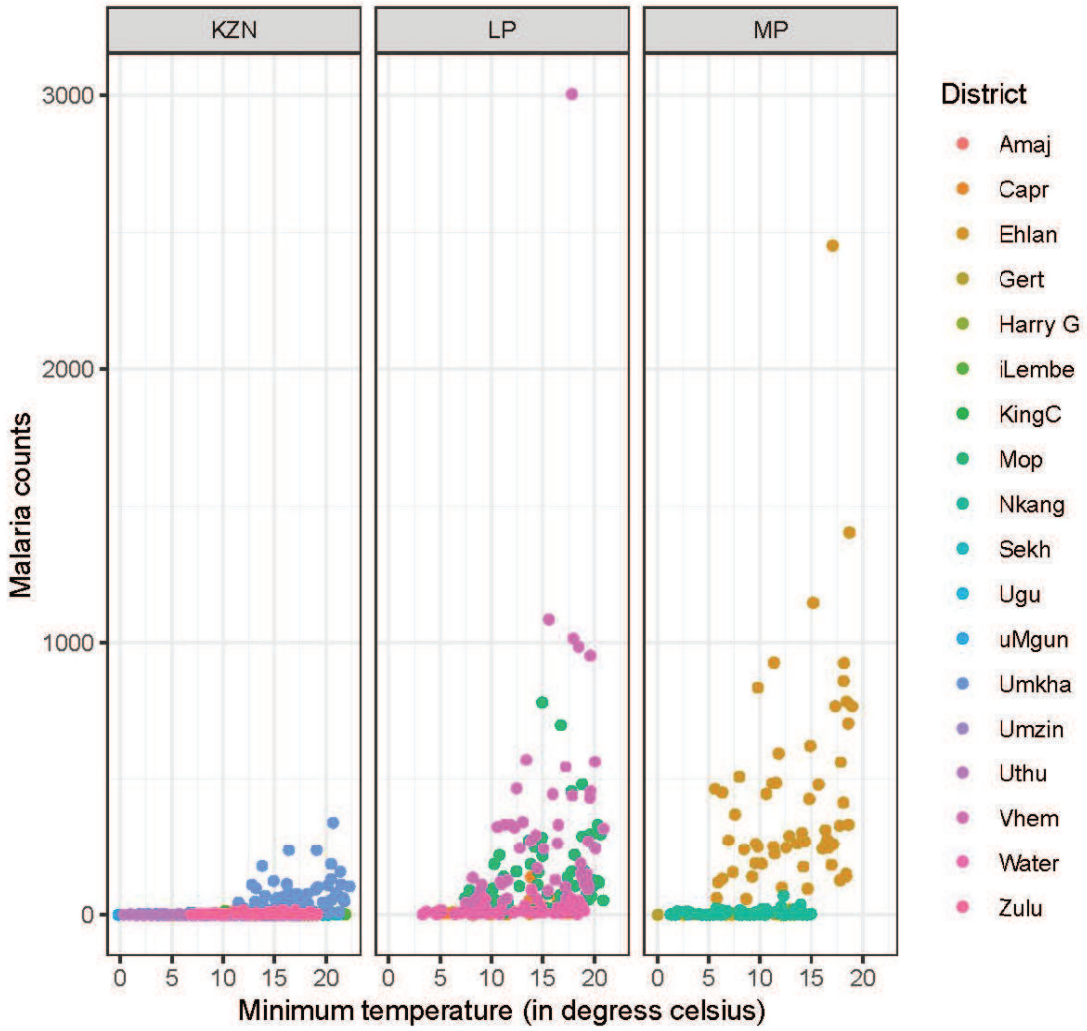


Fig. 5: Malaria counts by manimum temperature in three Provinces

The scatter plot in Figure 5 shows the relationship between average minimum temperature (in degrees Celsius) and malaria counts across three provinces: KwaZulu-Natal (KZN), Limpopo (LP), and Mpumalanga (MP). Each data point is color-coded to represent different districts within these provinces. In KwaZulu-Natal, malaria counts remain relatively low across all minimum temperature values, with most cases concentrated in the uMkhanyakude district. In Limpopo, malaria counts increase as the minimum temperature rises, particularly between 10°C and 20°C, with the highest case numbers recorded in the Vhembe and Mopani districts. Similarly, Mpumalanga shows a notable increase in malaria cases

at higher minimum temperatures, with the highest number of cases observed in the Ehlanzeni district, especially when temperatures exceed 15°C. Overall, malaria counts tend to rise with increasing minimum temperatures, particularly in Limpopo and Mpumalanga, suggesting that warmer temperatures at night may contribute to higher malaria transmission.

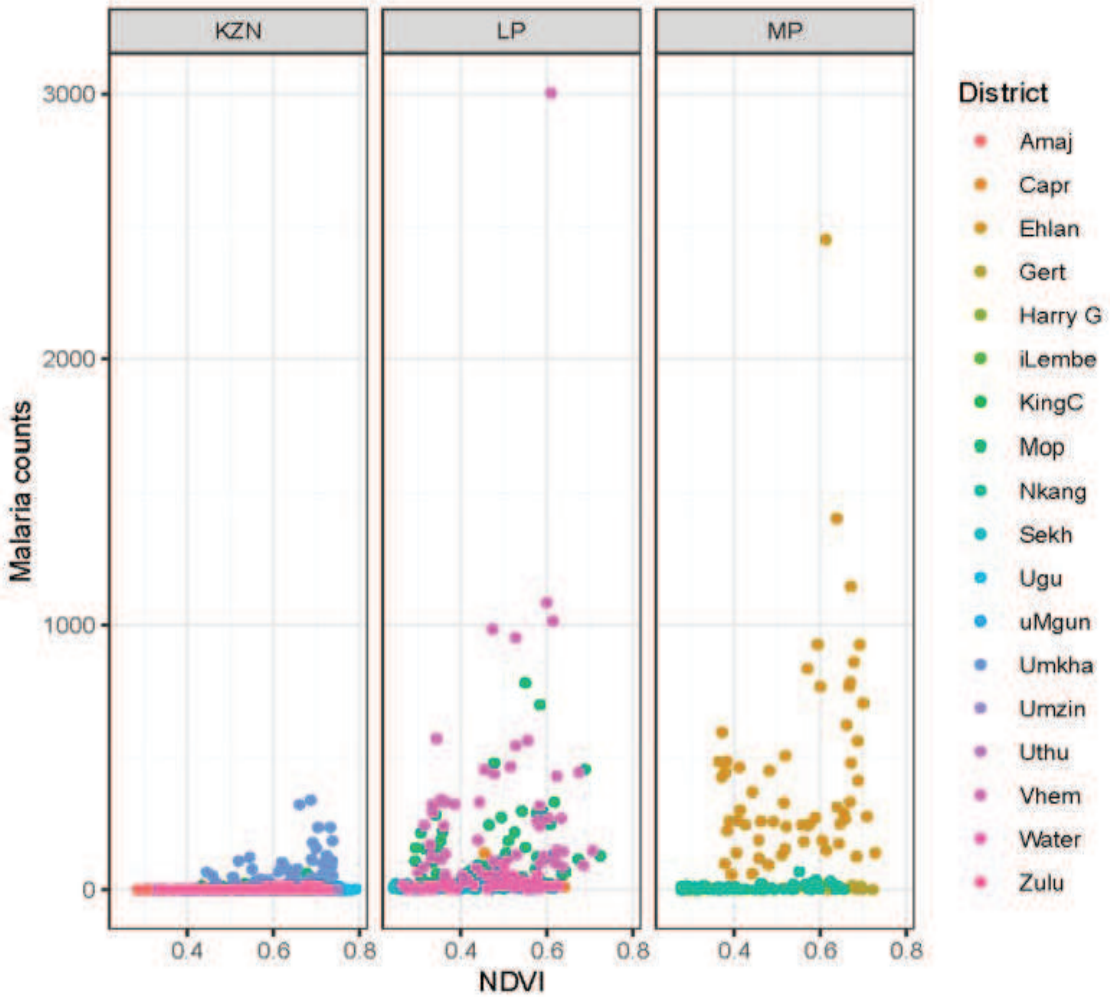


Fig. 6: Malaria counts by NDVI in three Provinces

Figure 6 depicts the relationship between the NDVI and malaria counts across three provinces: KwaZulu-Natal (KZN), Limpopo (LP), and Mpumalanga (MP). Data points are color-coded by district, providing a clear visual representation of regional variations. In KwaZulu-Natal, malaria counts remain consistently low across all NDVI values, with cases concentrated in the uMkhanyakude district. In Limpopo, malaria counts are more variable, with the highest case numbers occurring when NDVI values range between 0.4 and 0.8, suggesting a possible association between vegetation density and malaria transmission. The Vhembe and Mopani districts contribute significantly to these high counts. Similarly, in Mpumalanga, malaria cases increase with NDVI, particularly in the Ehlanzeni district, where higher vegetation density (NDVI$>$ 0.5) corresponds to elevated malaria counts. Overall, the plot suggests that malaria incidence tends to rise in areas with moderate to high vegetation density, particularly in Limpopo and Mpumalanga. This trend may indicate that malaria transmission is influenced by environmental conditions that favor mosquito breeding, such as the presence of vegetation and water bodies.

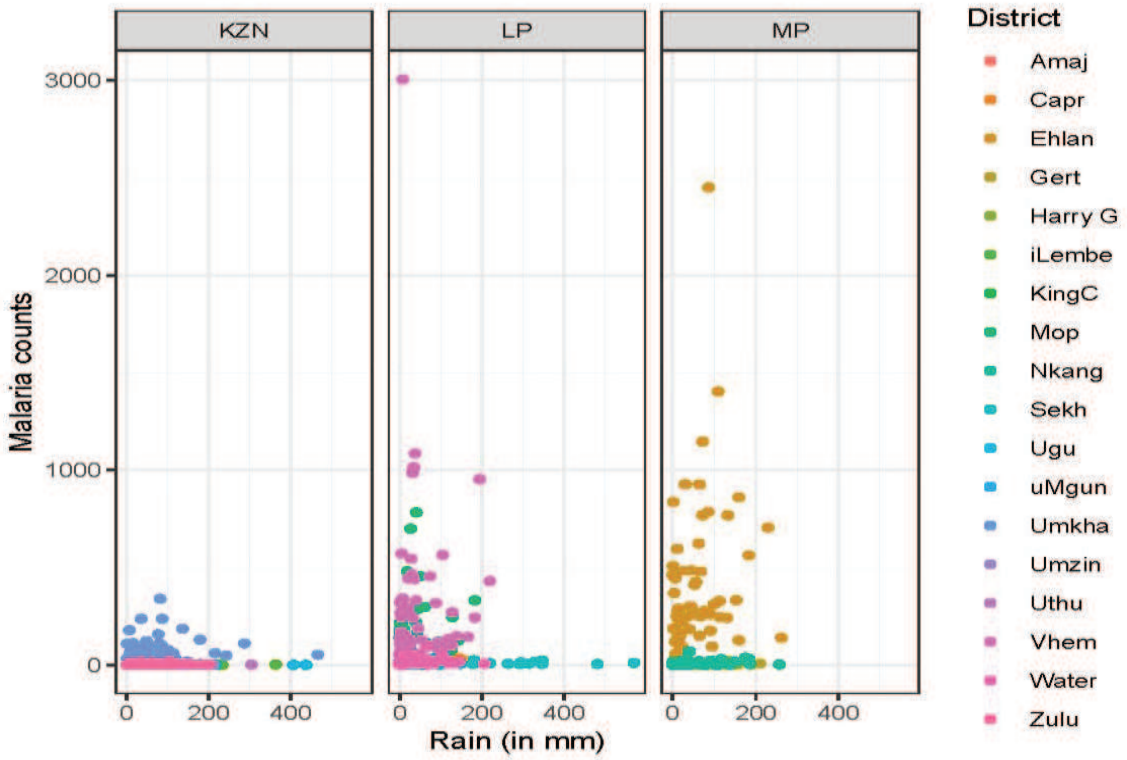


Fig. 7: Malaria counts by rain in three Provinces

Figure 7 presents a faceted scatter plot illustrating the relationship between rainfall (in mm) and malaria counts across different districts within three provinces: KwaZulu-Natal (KZN), Limpopo (LP), and Mpumalanga (MP). Each district is represented by a distinct color, as indicated in the legend. The plot reveals that malaria counts tend to be higher at lower rainfall levels, though variability exists. In KZN, most data points are clustered at low malaria counts, with rainfall values spread across a broader range. LP exhibits a wider spread in malaria counts, with some notably high values, while MP shows a concentration of lower malaria counts, particularly at higher rainfall levels. District-level variations are evident, with certain districts (Ehlanzeni and Vhembe) experiencing higher malaria counts. These patterns suggest that while rainfall may influence malaria transmission, additional environmental or socio-ecological factors likely contribute to the observed trends.

3.2 Model Fitting, Evaluation, and Prediction Results

Table 3: Count models fit statistics

Model type	**GLMs**			**Zero-augmented**				**Mixture models**	
Distribution	Poisson	GP	NB	ZIP	ZINB	HP	HNB	MixP	MixNB
AIC	59151.21	6001.66	5244.20	56581.12	5933.00	56551.79	5889.02	19929.35	6044.92
BIC	59255.58	6118.77	5410.22	56789.85	6146.49	56760.53	6102.50	20142.83	6248.91

In our comparison of various count models (Table 3), including GLMs, zero-augmented models, and mixture models, we evaluated their performance using both the AIC and the BIC. Our findings indicated that the negative binomial model within the GLM framework obtained the lowest AIC value at 5244.20, and the lowest BIC value at 5410.22. This suggests that the negative binomial model of the GLMs outperforms all other fitted count models in terms of

performance. Therefore, we infer that the GLM's negative binomial model stands as the superior count model within the scope of this study.

Table 4: Bayesian NB models fit statistics

Prior distribution	Jeffreys	Beta/Uniform prior
DIC (smaller is better)	5244.326	5244.841
P_D (effective number of parameters)	34.999	34.832

By utilising the fit statistics of beta/uniform negative binomial model and the Jeffreys negative binomial model presented in Table 4, we can effectively compare their performance. It becomes evident that the Jeffreys negative binomial model exhibits a smaller BIC value of 5244.325 and a higher number of effective parameters compared to the beta/uniform negative binomial model, which has a BIC value of 5244.841 and 34.832 effective parameters. Consequently, within the scope of this study, it can be concluded that the Jeffreys negative binomial model outperforms the beta/uniform negative binomial model in terms of performance and predictive accuracy. For future predictions of malaria cases, the negative binomial model developed using Jeffreys prior, as shown in Table 5 is employed.

Table 5: Posterior Summaries

Parameter	**Mean**	**Std. Dev.**	**HPD Interval**	**MCSE**
Intercept	-0.6316	0.8987	(-2.4394 , 1.1196)	0.0215
LP	2.3312	0.2531	(1.8694 , 2.8610)*	0.00570
MP	1.8941	0.2161	(1.4806 , 2.3216)*	0.00466
Ehlan	3.5990	0.2030	(3.2294 , 4.0042)*	0.00481
Gert	-0.7339	0.1592	(-1.0451 , -0.4305)*	0.00347
Harry_G	-2.4225	0.3883	(-3.1718 , -1.6842)*	0.0113
Mop	2.1552	0.1728	(1.8430 , 2.5085)*	0.00424
Ugu	-0.4845	0.2942	(-1.0913 , 0.0426)	0.00718
Umkha	3.3936	0.2830	(2.8718 , 3.9409)*	0.00689
Umzin	-1.5252	0.2937	(-2.0706 , -0.9457)*	0.00758
Uthu	-1.0767	0.2707	(-1.6510 , -0.5868)*	0.00645
Vhem	2.9315	0.1713	(2.5985 , 3.2699)*	0.00439
Water	-0.0103	0.1654	(-0.3439 , 0.2879)	0.00398
Zulu	0.8704	0.2303	(0.4111 , 1.3049)*	0.00516
iLembe	-0.5930	0.2954	(-1.2097 , -0.0429)*	0.00680
Feb	-0.4219	0.1454	(-0.7099 , -0.1400)*	0.00321
Mar	-0.3585	0.1488	(-0.6395 , -0.0633)*	0.00369
Jun	-0.6593	0.3538	(-1.3621 , -0.0194)*	0.00847
Jul	-0.7081	0.3712	(-1.4318 , -0.0119)*	0.00845
Aug	-0.6371	0.3209	(-1.2891 , -0.0527)*	0.00779
Oct	-0.6760	0.2389	(-1.1322 , -0.2042)*	0.00568
Nov	-0.8455	0.1884	(-1.1985 , -0.4625)*	0.00448
Dec	-1.1863	0.1611	(-1.5227 , -0.8885)*	0.00378
2019	-0.3110	0.0952	(-0.5023 , -0.1348)*	0.00237
2020	-0.7528	0.0955	(-0.9465 , -0.5761)*	0.00231
2021	-0.9052	0.0983	(-1.0938 , -0.7145)*	0.00209
2022	-0.7957	0.1121	(-1.0260 , -0.5884)*	0.00280
MaxTemp	0.0296	0.0254	(-0.0193 , 0.0796)	0.000610
MinTemp	0.0363	0.0320	(-0.0302 , 0.0962)	0.000738
Rain	-0.00162	0.000783	(-0.00310 , -0.00006)*	0.000019
NDVI	1.6078	0.7899	(0.1200 , 3.1895)*	0.0184
Dispersion	0.4923	0.0354	(0.4252 , 0.5630)*	0.000397

The Bayesian Negative Binomial model , developed using Jeffreys prior, provides insights into malaria occurrence across spatial, environmental and climatic factors (Table 5). Spatially, malaria risk is significantly higher in Ehlanzeni (3.5990), Umkhanyakude (3.3936), and Vhembe (2.9315), confirming these districts as high-risk areas. In contrast, Harry Gwala (-2.4225), Umzinyathi (-1.5252), and Uthukela (-1.0767) exhibit significantly lower malaria risk, suggesting protective factors in these regions. The significance of these effects is determined by their HPD intervals, which do not include zero. Seasonally, malaria cases decline during the cooler months, with significant reductions observed in February (-0.4219), June (-0.6593), and December (-1.1863), indicating lower transmission risk compared to

the reference period. A declining trend over time is also evident, with 2020 (-0.7528), 2021 (-0.9052), and 2022 (-0.7957) all showing significantly lower malaria incidence, suggesting the effectiveness of control interventions or environmental changes. Environmental factors further influence malaria transmission. NDVI (1.6078) is positively associated with malaria, implying that dense vegetation fosters mosquito breeding. Meanwhile, rainfall (-0.00162) has a small but significant negative effect, suggesting that excessive rain may disrupt breeding sites and reduce mosquito populations. The dispersion parameter (0.4923) is significantly greater than zero, confirming overdispersion in malaria cases and justifying the use of the Negative Binomial model. These findings highlight the need for targeted malaria interventions, particularly in high-risk districts, during peak transmission months, and in response to environmental conditions. In this study, it has been established that the model developed using Jeffreys prior distribution outperforms the one constructed with the conjugate prior distribution, namely a beta/uniform prior distribution. Therefore, the malaria predictions rely on the posterior distribution within the Jeffreys prior distribution model. These predictions of future malaria counts are based on one or more explanatory variables within the dataset. The anticipated malaria counts are depicted through effects plots, where contour plots showcase malaria predictions based on numeric variables, and interaction plot illustrate malaria predictions based on categorical explanatory variables.

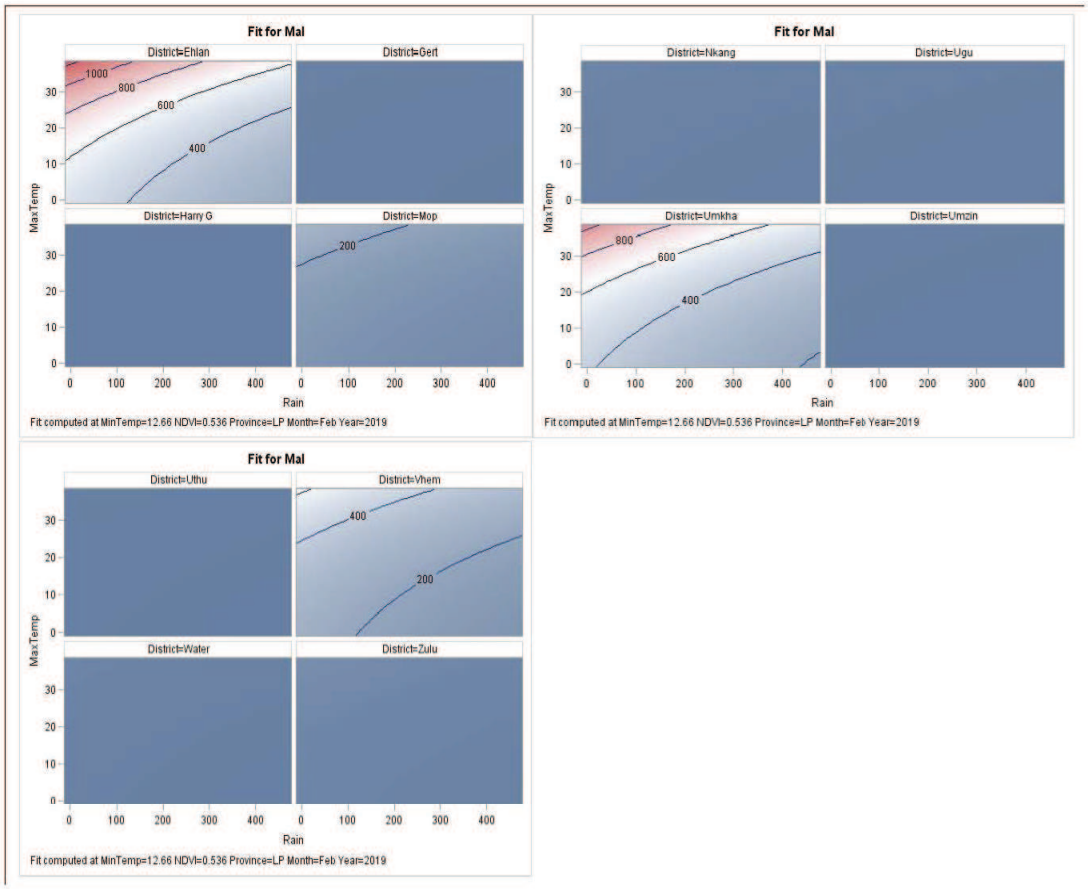


Fig. 8: Malaria predictions relative to rain and temperature across provinces

Figure 8 provide an overview of malaria case predictions concerning rainfall and temperature across all districts within three provinces of interest. Notably, the Ehlanzeni district of Mpumalanga province stands out with predicted counts peeking at 1000 cases, attributed to temperatures between 20 °Cand 30 °Cand rainfall ranging from 0 mm to 200 mm. Similarly, the uMkhanyakude district, in KwaZulu-Natal province, exhibits high predicted malaria counts under identical climate conditions. In contrast, Mopani and Vhembe districts of Limpopo province, experiencing the same climate conditions, are predicted to have a maximum of approximately 200 and 400 malaria cases, respectively.

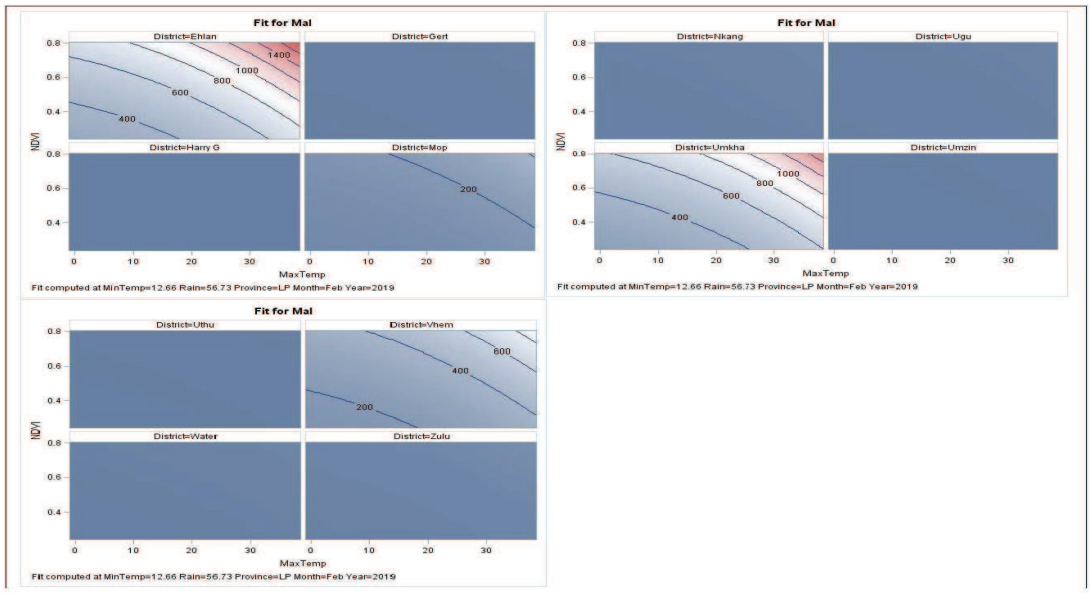


Fig. 9: Malaria predictions relative to NDVI and temperature across provinces

Figures 9 illustrate malaria count predictions relative to temperature and NDVI across various districts. Notably, Ehlanzeni district exhibits high malaria counts, peaking at approximately 1400 cases, followed closely by uMkhanyakude district in KwaZulu-Natal with counts reaching 1000. These counts are correlated with temperatures ranging from 25,$^\circ$Cto 30,$^\circ$C and NDVI levels hovering around 0.7 to 0.8. In contrast, under identical climate and environmental conditions, Mopani and Vhembe districts in Limpopo province are predicted to have lower maximum counts, approximately 200 and 600 malaria cases, respectively.

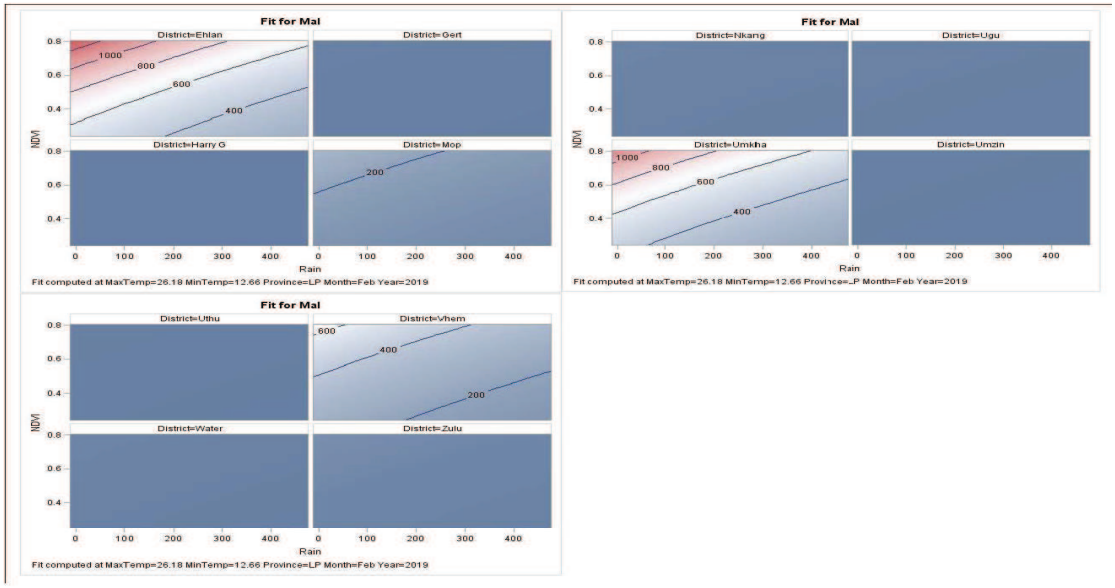


Fig. 10: Malaria predictions relative to rain and NDVI across provinces

Figures 10 illustrate predictions of malaria transmissions relative to rainfall and NDVI conditions. The analysis reveals considerable malaria transmission predicted in the Ehlanzeni district of Mpumalanga and uMkhanyakude district of KwaZulu-Natal province. These regions are expected to experience a high volume of transmissions, with estimates ranging from 800 to 1000, corresponding to rainfall levels between 0 mm and approximately 200 mm, and NDVI values ranging from 0.5 to 0.8. Contrastingly, under similar conditions, the Mopani district of Limpopo province is expected to see a maximum of 200 malaria transmissions, while projections for the Vhembe district of the same province range from 400 to 600. Notably, no malaria transmissions are projected for the remaining districts included in the study.

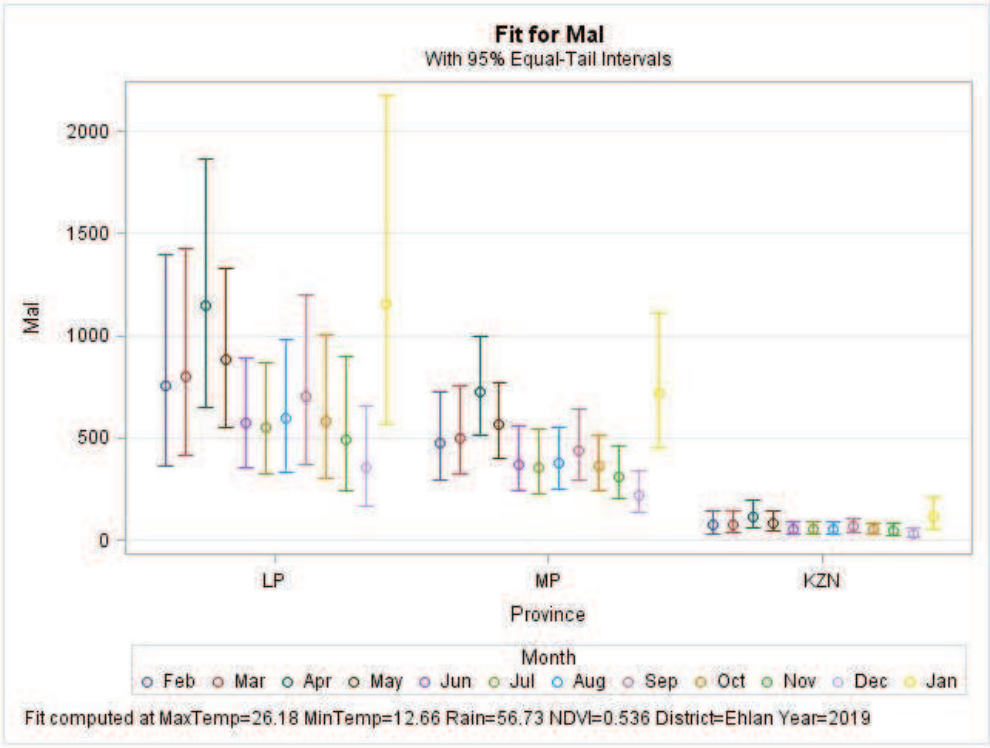


Fig. 11: Malaria predictions across provinces and months

The interaction plot showcased in Figure 11 provides insights into malaria count predictions associated with specific climate and environmental factors: a daily average temperature of 26.18,° C, night time average temperature of 12.66,° C, average rainfall of 56.73 mm, and an NDVI of 0.536. In terms of regional predictions, Limpopo province emerges with the highest anticipated malaria counts, particularly in January, where projections exceed 2000 cases. April follows closely behind, with estimates surpassing 1500 cases, while February and March anticipate fewer cases, still remaining below the 1500 mark. Following Limpopo, Mpumalanga province exhibits elevated malaria counts, particularly in January, projecting just above 1000 cases. However, throughout the other months, anticipated counts fall below the 1000 threshold. In contrast, KwaZulu-Natal province is expected to experience overall lower malaria counts. Notably, projections indicate a seasonal pattern: from January to May, malaria counts are anticipated to rise, followed by a decrease from June to August, a subsequent increase in September, and finally a decline from October to December.

4 Conclusion and Recommendations

The analysis of malaria cases based on rainfall, temperature, and NDVI across various districts within the three provinces provides valuable insights. The negative binomial model represented the data well compared to other GLMs, zero-augmented, and mixture count models. Jeffreys' prior distribution better represented the prior information

compared to the Beta distribution. Therefore, the posterior distribution used for both inference and malaria predictions was developed using the negative binomial model and Jeffreys' prior distribution.

Malaria transmission is predicted to be high in areas with temperatures ranging from 20,°C to 30,°C, rainfall between 0mm and 200mm, and NDVI levels around 0.5 to 0.8, with case projections ranging from 200 to 1000. These results differ somewhat from those of Kim et al. [2], who predicted high malaria transmission in regions with temperatures ranging from 23,°C to 24,°C, with the highest transmission at approximately 27,°C and at around 12,°C. The Ehlanzeni district of Mpumalanga province exhibits notably high predicted malaria counts, followed by the uMkhanyakude district of KwaZulu-Natal. In contrast, the Mopani and Vhembe districts of Limpopo province anticipate moderate malaria counts. The results also reveal a seasonal pattern: malaria counts are predicted to be high from January to May and in September, moderate from October to December, and low from June to August. These seasonal trends are consistent with those outlined by previous studies, such as Abiodun et al. [1]. However, they contrast with the findings of Landman et al. [3], which identified December as a month with high malaria incidence.

Targeted malaria prevention and control measures are recommended for high-risk districts, such as Ehlanzeni and uMkhanyakude in KwaZulu-Natal province, where elevated malaria counts are consistently predicted. These measures include intensified mosquito control, distribution of insecticide-treated bed nets, and community awareness campaigns. Timely interventions aligned with seasonal patterns observed in malaria counts are essential, involving enhanced surveillance, resource allocation for healthcare facilities, and proactive measures to mitigate potential outbreaks. Continued monitoring and research are advised to assess the effectiveness of existing strategies and adapt interventions based on evolving climate and environmental factors, emphasising collaboration between health authorities, research institutions, and local communities.

The study encountered limitations due to the lack of climate data for additional districts, namely Amajuba, King Cetshwayo, and Sekhukhune. Consequently, the research was unable to predict malaria cases for these areas. Moreover, the study revealed constraints associated with objective prior distributions, which are typically chosen according to formal rules. These distributions tend to be either non-informative or weakly informative, failing to capture the relationship between malaria transmission and climate change, as established by numerous researchers. This void hinders the integration of relevant findings and knowledge into malaria modelling processes when relying solely on objective priors. To address this challenge, future research endeavours may explore the development of predictive malaria models using subjective informative elicited prior distributions. Such an approach would incorporate insights from existing research, thereby enriching the predictive accuracy and robustness of malaria modelling efforts.

Acknowledgement

The first author acknowledges the financial support received from the University Capacity Development Program (UCDP), a division of the Department of Higher Education and Training (DHET).

The authors also extend their sincere thanks to the anonymous referees, the Associate Editor, and the Editor for their valuable and constructive feedback, which significantly enhanced the quality of this paper.

References

[1] G.J. Abiodun, P. Witbooi, and K.O. Okosun, Hacettepe Journal of Mathematics and Statistics **47**, 219-235 (2018).
[2] Y. Kim, J.V. Ratnam, T. Doi, Y. Morioka, S. Behera, A. Tsuzuki, N. Minakawa, N. Sweijd, P. Kruger, R. Maharaj, and C.C. Imai, Scientific Reports **9**, 17882 (2019).
[3] W.A. Landman, N. Sweijd, N. Masedi, and N. Minakawa, Environmental Development **35**, 100522 (2020).
[4] G.J. Abiodun, B. Adebiyi, R.O. Abiodun, O. Oladimeji, K.E. Oladimeji, A.M. Adeola, O.S. Makinde, K.O. Okosun, R. Djidjou-Demasse, Y.J. Semegni, and K.Y. Njabo, The Open Public Health Journal **13**, 1 (2020).
[5] B.D. Brooke, J. Raman, J. Frean, K. Rundle, F. Maartens, E. Misiani, A. Mabuza, K.I. Barnes, D.P. Moonasar, Q. Dlamini, and S. Charles, SAMJ: South African Medical Journal **110**, 1072-1076 (2020).
[6] World Health Organization, Update on the E-2020 initiative of 21 malaria-eliminating countries: report and country briefs (No. WHO/CDS/GMP/2018.13), World Health Organization (2018).
[7] J.F. Baker, Bayesian spatiotemporal modelling of chronic disease outcomes, Doctoral dissertation, Queensland University of Technology (2017).
[8] R. Maharaj, A. Ward, B. Didier, I. Seocharan, F. N. Firas, R. Balawanth, D. Lucero, N. Morris, M. Shandukani, E. Raswiswi, and G. Malatjie, Malaria Journal **22**, 107 (2023).
[9] World Health Organization, World malaria report 2022, World Health Organization (2022).
[10] World Health Organization, WHO Guidelines for malaria, 14 March 2023 (No. WHO/UCN/GMP/2023.01), World Health Organization (2023).

[11] P. Venkatesan, The 2023 WHO World malaria report, The Lancet Microbe (2024).
[12] M. Mabona, T. Zwane, J. Raman, L. Kuonza, B. Mhlongo, and P. Phafane, Malaria Journal **23**, 47 (2024).
[13] M. Wang, H. Wang, J. Wang, H. Liu, R. Lu, T. Duan, X. Gong, S. Feng, Y. Liu, Z. Cui, and C. Li, PLOS ONE **14**, e0226910 (2019).
[14] Y.A. Adamu, Malaria prediction model using machine learning algorithms, Turkish Journal of Computer and Mathematics Education (TURCOMAT) **12**, 7488-7496 (2021).
[15] Y.W. Lee, J.W. Choi, and E.H. Shin, Machine learning model for predicting malaria using clinical information, Computers in Biology and Medicine **129**, 104151 (2021).
[16] O. Nkiruka, R. Prasad, and O. Clement, Prediction of malaria incidence using climate variability and machine learning, Informatics in Medicine Unlocked **22**, 100508 (2021).
[17] J. Mullahy, Specification and testing of some modified count data models, J. Econ. **33**, 341–365 (1986).
[18] R.E. Kass and L. Wasserman, The selection of prior distributions by formal rules, Journal of the American Statistical Association **91**, 1343-1370 (1996).
[19] W.M. Bolstad, *Understanding computational Bayesian statistics* (Vol. 644), John Wiley and Sons (2009).
[20] M.A.A. Turkman, C.D. Paulino, and P. Müller, Computational Bayesian statistics: an introduction (Vol. 11), Cambridge University Press (2019).
[21] H. Akaike, Information theory and an extension of the maximum likelihood principle, in B. N. Petrov and F. Caski (Eds.), Proceedings of the Second International Symposium on Information Theory, pp. 267-281, Budapest: Akademiai Kiado (1973).
[22] G. Schwartz, Estimating the dimension of a model, The Annals of Statistics **6**, 461-464 (1978).
[23] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. Van Der Linde, Bayesian measures of model complexity and fit, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**, 583-639 (2002).

Appendix: Model diagnostics