

RESEARCH

Open Access



Modeling sparse Rift Valley fever incidence data: a Bayesian perspective on zero-inflated self-exciting and autoregressive models

Alexandros Angelakis^{1,2}, Bryan O. Nyawanda^{1,2} and Penelope Vounatsou^{1,2*}

Abstract

Background Rift Valley fever (RVF) is a mosquito-borne zoonotic disease for which predictive modeling is often hindered by sparse data, particularly the high frequency of zero counts in both human and livestock surveillance systems. While zero-inflated models are commonly used for sparse data, several temporal count modelling frameworks exist, including less common self-exciting models that assume an initial case increases the likelihood of subsequent cases.

Methods This study compares three zero-inflated Bayesian models: the negative binomial (ZINB) with autoregressive temporal random effects, the self-exciting negative binomial (SE-NB) and the generalized autoregressive moving average negative binomial (GARMA-NB). The models were evaluated across simulated datasets with varying levels of sparsity.

Results We found that zero-inflation substantially improves predictive performance within specific sparsity thresholds: 29–94.5% (ZINB), 25–93% (SE-NB), and 30–95% (GARMA-NB). Applied to monthly RVF incidence data from northern Kenya (2018–2024), the ZINB model with a three-month rainfall lag provided the most accurate forecasts.

Conclusion These findings underscore the importance of zero-inflated negative binomial models and climate-based covariates in enhancing early warning systems for RVF-endemic regions.

Keywords Bayesian inference, Generalized autoregressive moving average (GARMA), Negative binomial, Self-exciting process, Surveillance, Time series, Zero-inflation

Introduction

Rift Valley fever virus (RVFV) is a zoonotic RNA virus from the *Phlebovirus* genus within the *Phenuiviridae* family [1]. It poses a significant threat to public health and livestock production systems, causing substantial economic losses due to high mortality in domesticated

ungulates and trade restrictions. Transmission to animals (e.g., cattle, sheep, goats, camels, buffaloes) typically occurs via bites from infected *Aedes* and *Culex* mosquitoes, leading to mortality rates of 5% to 20% in adult animals and 80% to 100% in newborns [2–4]. Human infection arises from contact with bodily fluids of infected animals or through mosquito bites [5, 6]. Most human cases are mild or asymptomatic resembling those of influenza or malaria [7], but approximately 10% can progress to severe symptoms (e.g. hepatitis, retinitis, encephalitis), and around 1% may develop haemorrhagic disease [2, 3] with a 50% fatality rate [8].

*Correspondence:

Penelope Vounatsou
penelope.vounatsou@swisstph.ch

¹Swiss Tropical and Public Health Institute, Kreuzstrasse 2,
Allschwil CH-4123, Basel-Land, Switzerland

²University of Basel, Peterspl. 1, Basel CH-4001, Basel-Stadt, Switzerland



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

First identified in Kenya in the early 20th century, the virus has predominantly affected the Horn of Africa (Kenya, Somalia, Tanzania), where outbreaks recur every 4–15 years, but it has also emerged in North Africa, the Arabian Peninsula, southern Africa, west Africa and Indian Ocean islands [9]. RVF outbreaks follow seasonal patterns linked to heavy rainfall, flooding, and El Niño-related water surface warming [10, 11]. However, this pattern has been increasingly challenged by recent studies reporting focal inter-epizootic transmission, even in the absence of major climatic anomalies. Localized virus circulation between large-scale outbreaks has been documented in regions such as Kenya [12] and South Africa [13], suggesting that RVFV may persist through cryptic enzootic cycles involving low-level transmission among livestock and wildlife reservoirs [14]. These findings imply that outbreak risk may be influenced not only by broad climatic triggers such as El Niño, but also by localized ecological and socio-economic factors, including livestock movement, vector habitat suitability, and gaps in surveillance. Recognizing and incorporating these dynamics is essential for improving the accuracy and relevance of predictive models and early warning systems.

In Kenya, Rift Valley fever (RVF) outbreaks have historically occurred every five to ten years. However, improvements in surveillance systems have led to more frequent detection of localized outbreaks in recent years [15]. A major national outbreak occurred in 2006–2007, resulting in approximately 340 confirmed human cases, 90 deaths, and economic losses exceeding US\$32 million [16–19]. According to the Kenyan Ministry of Health and Ministry of Agriculture, a ‘national outbreak’ typically refers to an event affecting multiple counties and requiring coordinated, cross-sectoral response at the national level. In contrast, smaller-scale outbreaks confined to specific districts or sub-counties may be managed locally without triggering national alerts. For example, the 2018 outbreak in Wajir County (northern Kenya) was geographically limited to a few sub-counties but was nonetheless declared a national outbreak due to its severity—marked by mass livestock abortions, high mortality among young animals, and four human deaths. This prompted the deployment of a multidisciplinary response team and the activation of national-level public health coordination mechanisms [19, 20].

Various models have been developed for the early detection and forecasting of RVF outbreaks, often in relation to climate anomalies (e.g., heavy rains and flooding) that create suitable habitats for vector proliferation. These models typically rely on a combination of climatological, environmental, and epidemiological data—such as precipitation, vegetation cover, and mosquito population surveillance—to identify periods when risk is elevated. Approaches have ranged from statistical

time-series models (e.g., ARIMA, Poisson, or negative binomial regression) and remote-sensing driven early warning systems to more complex frameworks that integrate mechanistic or agent-based models of vector ecology. Most studies, however, concentrate on monitoring vector dynamics in animals before spillover to humans [2, 9, 21]. Despite these advancements, challenges persist. One key issue in modeling zoonotic diseases like RVF is the high frequency of zero reporting, especially in areas where no cases are observed over extended periods. This zero inflation often renders standard time-series or regression methods less effective, as they assume data distributions that do not account for the high frequency of zero cases. One potential approach to addressing the data sparsity is zero-inflated or zero-modified models which involve assigning data samples to one of the mixture components, with the random probability of allocation being inferred. These models are combined with a Poisson or negative binomial count distribution for the cases and are fitted using Bayesian hierarchical frameworks to facilitate computation. Ideally, early detection and outbreak forecasting should rely on timely livestock surveillance. However, in many endemic regions, such data are often unavailable or incomplete due to limited resources and infrastructure within veterinary systems. As a result, much of the accessible and systematically reported data pertain to human cases. Although human infections occur downstream of transmission in livestock, modeling these data still provides value for public health preparedness, risk communication, and resource allocation. One of the key challenges in working with human case data, however, is the high frequency of zero reports across time and locations, particularly during inter-epidemic periods or in remote settings.

A promising alternative approach is the use of self-exciting models (often referred to as Hawkes processes) in forecasting RVF outbreaks. Self-exciting models are based on the assumption that the occurrence of an event (e.g., a reported case) increases the likelihood of subsequent events in later time periods [22, 23]. This feedback mechanism is especially relevant when disease clusters emerge following an initial case or when environmental conditions change abruptly (e.g., after flooding). By incorporating temporal dependencies, self-exciting models reveal patterns of localized disease amplification and more precisely model temporal clustering that follows case emergence [24]. In the context of sparse RVF data, self-exciting models can be adapted to integrate covariates such as precipitation and vegetation indices, allowing them to flexibly handle variable outbreak dynamics while still recognizing the increased risk following an initial case. Their capacity to capture “triggering” events can be particularly advantageous when outbreaks occur sporadically, as the model naturally responds to a single

detected case by increasing the probability of subsequent cases nearby in time. This makes them a valuable tool for anticipating rapid escalation in outbreak settings with high temporal variability.

In this study, we examined various modeling approaches for forecasting sparse RVF data, focusing on a self-exciting model, a negative binomial random effects model, and a generalized autoregressive moving average (GARMA) model. All three models include crucial climatic covariates (e.g., precipitation, vegetation) associated with RVF transmission and have been adapted into zero-modified variants to address the high number of zero observations. We evaluated each model's performance using actual RVF incidence data from Kenya and simulated datasets encompassing different levels of sparsity, enabling a thorough assessment of their predictive capabilities under diverse outbreak conditions.

Data sources

Rift valley fever incidence data

This study analyzed monthly RVF incidence data from the arid northern counties of Kenya. Specifically, RVF cases in humans from Mandera, Wajir, Marsabit, Samburu, Isiolo, and Baringo counties were aggregated [25]. The data were obtained from the Kenya Health Information System through the District Health Information Software 2 (DHIS2) and covered the period from January 2018 to April 2024, disaggregated by county. The dataset was highly sparse, with 81% of the observations being zero.

Climatic data

Daytime land surface temperature (LSTD) data were obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) on board NASA's Terra and Aqua satellites with a spatial resolution of 1 km² and a temporal resolution of 8 days [26]. Rainfall data were obtained from the Climate Hazards Group Infrared Precipitation with Station data (CHIRPS) at spatial resolution of 5.6 × 5.6 km² and a temporal resolution of 5 days [27]. Normalized Difference Vegetation Index (NDVI) data, extracted from MODIS satellite images were obtained at a spatial resolution of 1 km² and a temporal resolution of 15 days [28]. The monthly averages of temperature and NDVI and the cumulative rainfall were first calculated.

Population data

Population data by county were extracted from the 2019 Population and Housing Census results [29], and annual growth rates from the WorldPop database [30] were applied to obtain population data for subsequent years.

Model formulation

We assessed the forecasting performance of two types of models for sparse count data: (i) negative binomial models with temporal random effects, (ii) self-exciting negative binomial model and (iii) Generalised autoregressive moving average model. These models were compared with their analogues that account for zero-inflation. They were fitted to RVF human incidence data from Kenya as well as to simulated data with various degrees of sparsity. The description of the models is provided below.

This study adopts a Bayesian approach to statistical inference, which provides a coherent framework for incorporating prior knowledge and updating beliefs in light of observed data. Under the Bayesian paradigm, model parameters are treated as random variables, and inference is based on their posterior distributions, which combine prior distributions with the likelihood of the data. This approach enables a direct probabilistic interpretation of parameter uncertainty and is especially useful for complex hierarchical models, where classical methods may fall short.

Negative binomial model (NB)

We considered that the observed RVF cases $Y = \{y_1, \dots, y_n\}$, for time $t = 1, \dots, n$ follow a negative binomial distribution i.e. $y_t \sim NB(\mu_t, r)$ with probability mass function,

$$f(y_t; \mu_t, r) = \binom{y_t + r - 1}{y_t} \left(\frac{\mu_t}{\mu_t + r} \right)^{y_t} \left(\frac{r}{\mu_t + r} \right)^r$$

where μ_t is the mean at time t and r the dispersion parameter of the distribution. We modeled the mean on the log scale as:

$$\log(\mu_t) = \log(P_t) + \mathbf{X}_t \boldsymbol{\beta} + \epsilon_t \quad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)^T$ are the regression coefficients, $\mathbf{X}_t = (X_{t1}, X_{t2}, \dots, X_{tk})$ are the k predictors at time t . Here P_t denotes the population at time t .

Month-specific random effects, ϵ_t , were introduced on the log scale of the mean and modeled using a first order autoregressive process: $\epsilon_t \sim N(\rho \epsilon_{t-1}, \sigma^2)$ with $\epsilon_1 \sim N(0, \frac{\sigma^2}{1-\rho^2})$, where σ^2 and ρ represent the temporal variance and autocorrelation parameters, respectively. We assume a uniform prior distribution for ρ , $\rho \sim U(-1, 1)$, a standard non-informative choice that ensures stationarity of the autoregressive process and does not favor any particular correlation structure.

The temporal variance is assigned an inverse gamma prior, $\sigma \sim IG(2.01, 1.01)$, with parameters chosen intuitively to provide a weakly informative prior with finite mean and variance. Regression coefficients are

given weakly informative priors, $\beta_i \sim N(0, 100)$, centered at zero with a large variance, allowing the data to drive posterior estimates while maintaining stability. The dispersion parameter r is assigned a Gamma prior, $r \sim \text{Gamma}(1, 1)$, a flexible and commonly used choice in negative binomial models. This prior ensures positivity and provides broad support without being overly restrictive.

Self-exciting negative binomial model (SE-NB)

A self-exciting negative binomial model defines the mean of the process as:

$$\mu_t = \exp(\log(P_t) + \mathbf{X}_t\beta) + \eta y_{t-1}. \quad (2)$$

Similar to the random effects negative binomial model described above, the linear predictor includes climatic covariates through the term $\mathbf{X}_t\beta$. The self-exciting component is incorporated via the term ηy_{t-1} , which allows the expected number of cases at time t to depend on the observed case count at the previous time point. This structure captures temporal dependence not explained by covariates alone, such as residual local transmission or behavioral persistence effects.

The parameter η represents the self-excitation coefficient and quantifies the strength of temporal dependence in the incidence process. A value of $\eta > 0$ indicates that higher case counts in the previous time step increase the expected incidence at time t . When $\eta = 0$, the model simplifies to a standard negative binomial model without temporal dependence. This parameter is particularly relevant in epidemic settings, where recent case history often influences current transmission dynamics.

To ensure that η is constrained to the interval $[0, 1]$ and to reflect a non-informative prior belief, we assign it a Beta(1, 1) prior distribution. This choice which is equivalent to a uniform distribution, allows the model to flexibly estimate the strength of self-excitation from the data without biasing the estimate toward either weak or strong temporal dependence.

Generalized autoregressive moving average negative binomial model

In the generalized autoregressive moving average (GARMA) negative binomial model for count time series [31], the counts y_t , at time t , conditioned on the information available on responses up to $(t-1)$ and on the covariates up to time t , $H_t = \{X_t, \dots, X_1, y_{t-1}, \dots, y_1, \mu_{t-1}, \dots, \mu_1\}$, are modeled by a negative binomial distribution with mean:

$$\log(\mu_t) = \mathbf{X}_t\beta + Z_t.$$

The linear predictor depends on the covariates and the term Z_t , which represents the autoregressive and moving average parts of orders $p, q > 0$:

$$Z_t = \sum_{i=1}^p \phi_i [\log(y_{t-i}^*) - X_{t-i}^T\beta] + \sum_{j=1}^q \theta_j \left[\log\left(\frac{y_{t-j}^*}{\mu_{t-j}}\right) \right],$$

where $y_{t-i}^* = \max(y_{t-i}, c)$ for $0 < c < 1$. Hence, any zero count $y_{t-i} = 0$ is replaced by a small constant c . The values of ϕ_i and θ_j represent the coefficients of the autoregressive (AR) and moving average (MA) components, respectively. Specifically, the ϕ_i terms capture the influence of past values of the latent process (i.e., deviations between observed log-counts and the linear predictor) at previous time points $(t-i)$. In contrast, the θ_j parameters quantify the impact of past forecast errors, defined as the deviations between the log of the observed counts and their corresponding predicted means, at time $(t-j)$. Together, these components allow the model to account for both persistent temporal dependencies (via the AR terms) and short-term shocks or random fluctuations (via the MA terms). This structure enhances the model's ability to capture complex temporal dynamics beyond the explanatory power of covariates alone and contributes to improved forecasting performance by incorporating memory of recent behavior in the system.

Zero-inflation (ZI)

To address the excess zero counts issue that the models above are not able to accurately estimate, all models were extended to zero inflated analogues. In particular, We assumed that the observations y_t follows a hierarchical mixture distribution

$$y_t|H_t \sim \begin{cases} 0, & \text{with probability } \pi_t, \\ NB(\mu_t, r), & \text{with probability } (1 - \pi_t), \end{cases}$$

where there are two means, the negative binomial mean defined in (1) and (2), and the π_t is defined as:

$$\text{logit}(\pi_t) = \log\left(\frac{\pi_t}{1 - \pi_t}\right) = \delta.$$

A normal prior was assigned to δ , such that $\delta \sim N(0, 100)$, with mean 0 and large variance chosen intuitively to reflect a weakly informative belief centered at zero. This specification allows the data to influence the estimate while avoiding overly strong prior assumptions.

Software

All models were fitted using Markov chain Monte Carlo (MCMC) simulation implemented in JAGS, accessed via

Table 1 Mean absolute error (MAE) and root mean square error (RMSE) of the negative binomial model with temporal effects and its zero-modified counterpart, on simulated data with different sparsity levels

Sparsity level	Model	MAE	RMSE
26.8%	NB	3.99	4.9
	ZINB	5.5	6.14
28.7%	NB	4.03	4.68
	ZINB	4.29	4.95
29.6%	NB	4.38	5.23
	ZINB	3.43	4.58
84.2%	NB	8.5	17.7
	ZINB	3.05	5.32
90.7%	NB	9.4	20.1
	ZINB	1.86	2.45
94.4%	NB	2.26	3.4
	ZINB	1.9	3.19
95.3%	NB	2.1	2.5
	ZINB	2.21	3.61
96.2%	NB	2.09	2.57
	ZINB	2.7	4.38

Table 2 Mean absolute error (MAE) and root mean square error (RMSE) of the self-exciting negative binomial model and its zero-modified counterpart, on simulated data with different sparsity levels

Sparsity level	Model	MAE	RMSE
20.3%	SE-NB	20.5	49.7
	SE-ZINB	29	54.9
24%	SE-NB	19.2	38.6
	SE-ZINB	21.9	41.9
25%	SE-NB	17.1	30.7
	SE-ZINB	16.7	30.5
84.2%	SE-NB	15	35.9
	SE-ZINB	5.4	9.7
88.8%	SE-NB	16.6	29.7
	SE-ZINB	6.6	10.6
92.5%	SE-NB	7.1	11.2
	SE-ZINB	5.4	6.1
93.5%	SE-NB	2.1	2.9
	SE-ZINB	2.7	3.4

the R2jags package in R [32]. We ran three independent chains, each with 100,000 iterations, discarding the first 10,000 as burn-in.

Results

Simulated data

Monthly count data were simulated for a period of nine years, as illustrated in supplement information, with the aim of evaluating the threshold at which the zero-inflation models outperform the simple non-inflated model. To assess the forecasting ability of each model, they were fitted on a training set consisting of eight years of data and evaluated on the ninth year (test data) by comparing

Table 3 Mean absolute error (MAE) and root mean square error (RMSE) of the generalized autoregressive moving average (GARMA(2,2)) model and its zero-modified counterpart, on simulated data with different sparsity levels

Sparsity level	Model	MAE	RMSE
27.7%	GARMA-NB	6.0	8.3
	GARMA-ZINB	9.55	11.75
29.6%	GARMA-NB	6.31	8.68
	GARMA-ZINB	9.2	11.62
30.5%	GARMA-NB	8.46	10.17
	GARMA-ZINB	7.05	9.26
84.2%	GARMA-NB	4.54	6.84
	GARMA-ZINB	1.46	4.02
91.6%	GARMA-NB	3.01	5.83
	GARMA-ZINB	0.725	2.61
94.4%	GARMA-NB	2.59	5.66
	GARMA-ZINB	0.43	2.07
95.3%	GARMA-NB	1.7	3.65
	GARMA-ZINB	0.45	2.31
96.2%	GARMA-NB	0.89	1.92
	GARMA-ZINB	1.04	2.95

model-based forecasts with the observed data. For the comparison, we used the mean absolute error (MAE) and root mean square error (RMSE) measures. Tables 1, 2 and 3 demonstrate that the forecasting performance of the standard model and of the zero-modified model shows consistent trends across the same sparsity thresholds, regardless of whether a random effects, self-exciting model or GARMA model is used.

More specifically, Table 1 shows that the forecasting performance of the zero-inflated negative binomial model is better than that of the negative binomial model in sparse data scenarios ranging from 29% to 94.4%. Similarly, Table 2 shows that the zero-inflated self-exciting model outperforms the non-inflated model at sparsity levels ranging from 25% to 93%. Finally, Table 3 demonstrates that the zero-inflated GARMA model exhibits improved forecasting performance at sparsity levels ranging from 30% to 96%.

Rift valley fever data

The incidence of RVF was modelled from 2018 to 2023, and the forecasting ability was tested using data from the first four months of 2024. Random effects, self-exciting and GARMA models, each with and without zero-inflation were fitted. We evaluated all possible combinations of rainfall, LSTD, and NDVI at various time lags for each model, selecting the combination with the best forecasting performance (based on MAE and RMSE) for presentation in Table 4.

Table 4 shows that the ZINB model with only rainfall yielded the best forecasting performance.

It successfully forecasts increased RVF incidence in January (36 cases), followed by a larger increase in

Table 4 The four best forecasting performances, in terms of mean absolute error (MAE) and root mean square error (RMSE), for the zero-inflated negative binomial (ZINB), the GARMA(2,2) negative binomial, and the zero-inflated self-exciting negative binomial (SE-ZINB) model using different combinations of climatic variables

Model	Climatic variables	MAE	RMSE
ZINB	Rainfall lag 3	26	46
ZINB	Rainfall lag 3, Temperature lag 1	35	53
GARMA-NB	Rainfall lag 3, Temperature lag 1	39.5	67
SE-ZINB	Rainfall lag 3	82	106

February (224 cases), associated with high rainfall in November and December of 2023, and then projects a significant decrease in March and April (Fig. 1a), aligning well with the actual reported cases: 26 in January; 132 in February; 0 in March; and 0 in April. Including temperature as an additional covariate (Fig. 1b) leads to similar performance, except for an overestimation in January (77 cases).

The zero-inflated self-exciting model with rainfall as a covariate also forecasts the initial two-month outbreak but overestimates cases in March and April (Fig. 1c). The GARMA(*p*,*q*) model with different orders, *p* ∈ [2, 5] and *q* ∈ [1, 5], always underestimates cases during the whole forecasting period for every combination of covariates, predicting always zero cases for the first 4 months of 2024.

Furthermore, Table B2 summarizes the posterior estimates and 95% Bayesian credible intervals for key parameters across the four best-performing models. Notably, rainfall lagged by three weeks emerged as a consistently significant predictor in all models, while temperature (lagged by one week) showed weak or uncertain effects. The self-excitation parameter *η* in the SE-ZINB model was positive, with a credible interval excluding zero, providing evidence for temporal clustering in RVF cases. Dispersion and zero-inflation parameters varied across models, reflecting differences in how they capture count variability and excess zeros.

Discussion

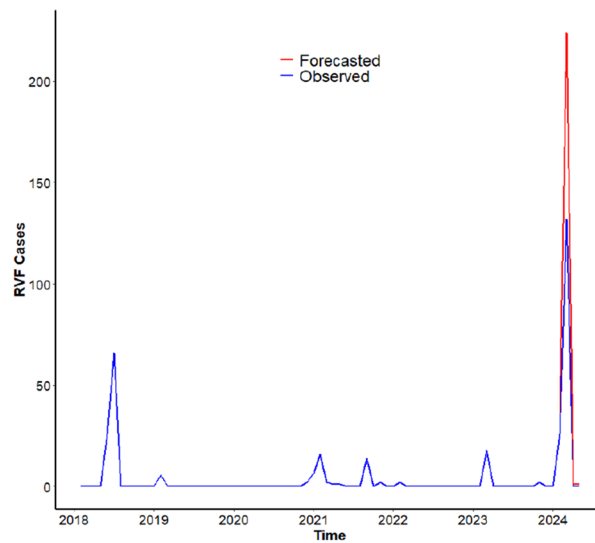
Transmission of many infectious diseases is influenced by climatic conditions. Diseases like Rift Valley Fever, which spend part of their life cycle in mosquitoes, outside of humans or other warm-blooded hosts are especially sensitive to climate. In many regions, these diseases manifest as epidemics, sometimes triggered by sharp variability of climatic factors that increase transmission rates. While early detection of epidemics is critical for the timely implementation of control interventions, reliance on human case data — as is often the case in resource-limited settings — presents a major challenge. Human cases typically represent late-stage indicators of transmission,

by which point opportunities for preventive measures, such as livestock vaccination, may have already been missed, as highlighted in the FAO Action Planning Guidelines. Nevertheless, forecasting based on human surveillance data can still offer value in several ways: it can support anticipation of healthcare needs, inform public risk communication, and enhance preparedness planning in the absence of timely livestock or entomological surveillance. In endemic regions where RVFV transmission is under-detected and persistent, even delayed signals from human case data can serve to trigger reactive interventions and guide future investments in more proactive surveillance systems.

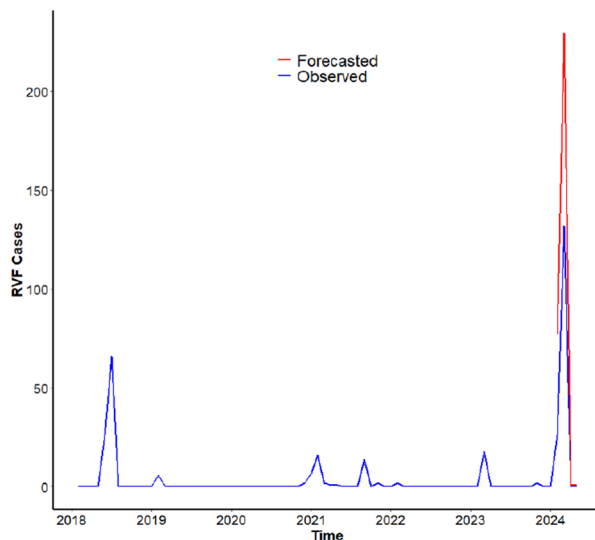
For this reason, this study identifies models that can successfully forecast sparse data. More specifically, negative binomial models with temporal random effects and self-exciting models were assessed for forecasting RVF incidence. The negative binomial distribution was deemed the most appropriate for modelling RVF count data, given the observed over-dispersion, where the conditional variance exceeds the conditional mean. Self-exciting temporal point process models are used to predict the rate of events as a function of time and the previous history of events. Generalized autoregressive moving average process models, in which the mean of the conditional exponential family distribution is depended on the past history of the process, can be applied to any type of quantitative time series. These models are well-suited to capturing triggering and clustering behaviour and have been widely employed in fields where temporal clustering of events is observed.

The models are zero-modified to facilitate the fitting of sparse vector-borne disease data, such as RVF incidence data. A zero-inflated model is appropriate when the observed zero in the data exceed the probability of zeros that the data distribution can estimate [33]. However, there are scenarios in which zero-inflated models may not offer improved performance over conventional models, particularly when the excess zeros are not substantial or when model complexity outweighs the benefits. Our simulation results across varying levels of sparsity further support this observation, showing that zero-inflated models tend to perform relatively better than standard count models when excess zeros are moderate, but not when they are either too few or substantial many.

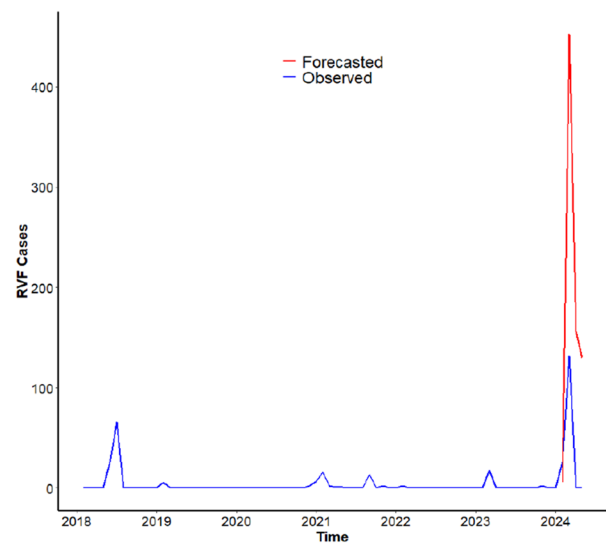
The analysis of the negative binomial model revealed that for a dataset exhibiting sparsity between 29% and 94.5%, the zero-modified version demonstrated a superior forecasting ability. In the case of the self-exciting model, the zero-modified analogues performed better when the data sparsity ranged between 25% and 93%, and for the GARMA model between 30% and roughly 95%. These broader ranges for the self-exciting and GARMA models are likely due to the additional parameters,



(a) Forecasting generated by the zero-inflated negative binomial (ZINB) model using only rainfall as a predictor.



(b) Forecasting generated by the zero-inflated negative binomial (ZINB) model using rainfall and daytime land surface temperature as predictors.



(c) Forecasting generated by the zero-inflated self-exciting model using only rainfall as a predictor.

Fig. 1 Monthly RVF cases observed in blue, and the forecasted cases for the first four months of 2024 in red

denoted as η and Z_t which is specific to the self-excitation part and the autoregressive and moving average parts of the models, respectively, and allows it to address the overdispersion of zeros in the data more effectively.

The data for Rift Valley Fever in the arid north counties of Kenya between 2018 and 2023 was used to fit the models, and the first four months of 2024 were used to forecast the incidence of the disease in light of future climatic

conditions. Additionally, a variety of combinations of daytime land surface temperature, rainfall, and vegetation indices were employed to determine which was the most effective predictor. The analysis demonstrated that the zero-inflated negative binomial model with temporal random effects exhibited the most optimal performance. Furthermore, the analysis indicated that rainfall occurring three months prior was the most effective predictor

of RVF cases. In particular, the ZINB model with rainfall variables demonstrated ability to capture an increase in cases during the months of January and February, which was attributed to the exceptional rainfall that occurred from October to November in the region, resulting in flooding. The ZINB model, which combined rainfall with a three-month lag and temperature with a one-month lag, demonstrated comparable forecasting performance; however, it exhibited a tendency to overforecast in January. The slight overprediction in the 2024 forecasts may reflect underreporting of human RVF cases during the outbreak, which would make the observed incidence appear lower than the model predictions, even though the model accurately captures the timing and overall magnitude of the outbreak. Both models indicated a positive correlation between rainfall and RVF incidence. In the second best model, a negative association was observed between temperature and RVF incidence, though the results were not statistically important.

This study has few limitations that are offered for consideration. While socio-economic and potential intervention factors were not considered in the analysis, the model allows for the integration and evaluation of additional predictors. Notably, this could facilitate a more comprehensive understanding of RVFV, which in turn could inform the development of more effective forecasting and early warning systems.

Future work will aim to extend the current temporal forecasting framework to a spatiotemporal model, enabling the joint analysis of temporal trends and geographic variation in RVF incidence. This would support the identification of high-risk areas across both space and time. Additionally, incorporating more detailed environmental and land use variables, such as proximity to water bodies, vegetation type, or livestock density may improve the ecological realism of the model. Integrating socio-economic and behavioral factors, including livestock movement patterns, human mobility, and vector control practices, could further enhance understanding of localized transmission dynamics.

Abbreviations

ARIMA	Autoregressive Integrated Moving Average
GARMA	Generalised autoregressive moving average
LSTD	Daytime land surface temperature
MAE	Mean absolute error
NB	Negative Binomial
NDVI	Normalized Difference Vegetation Index
PBIDS	Population Based Infectious Disease Surveillance
RMSE	Root mean square error
RVF	Rift Valley Fever
SE	Self-exciting
ZI	Zero inflation
ZINB	Zero Inflated Negative Binomial

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12879-025-11506-0>.

Supplementary Material 1.

Acknowledgements

Not applicable.

Authors' contributions

P.V. Conceptualization. A.A. and P.V. designed the study and model experiments. A.A. processed and analysed the data, developed the codes and prepared figures. B.O.N. supported data management. A.A. drafted the manuscript then all authors critically reviewed the manuscript for its intellectual content and approved the final version.

Funding

Open access funding provided by University of Basel. This work is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 101000365, project PREPARE4VBD (A Cross-Disciplinary Alliance to Identify, PREDict and prePARE for Emerging Vector-Borne Diseases).

Data availability

Data supporting the conclusions of this article are included within the article.

Declarations

Ethics approval and consent to participate

This study used anonymized, aggregated county-level data obtained from the Health Information System platform. The dataset contained no individual-level identifiers or personal health information. As the study involved secondary analysis of de-identified, routinely collected health data without interaction with human subjects, ethics approval and informed consent were not required. This approach is consistent with institutional and national data protection guidelines and adheres to the ethical principles outlined in the Declaration of Helsinki for research involving minimal risk and non-identifiable data.

Consent for publication

Not applicable.

Competing interest

The authors declare no competing interests.

Received: 4 April 2025 / Accepted: 8 August 2025

Published online: 29 September 2025

References

1. Pepin M, Bouloy M, Bird BH, Kemp A, Paweska J. Rift valley fever virus (Bunyaviridae: Phlebovirus): an update on pathogenesis, molecular epidemiology, vectors, diagnostics and prevention. *Vet Res*. 2010;41(6):61. <https://doi.org/10.1051/vetres/2010033>.
2. Telford C, Nyakarahuka L, Waller L, Kitron U, Shoemaker T. Geostatistical modeling and prediction of Rift Valley fever seroprevalence among livestock in Uganda. *Am J Trop Med Hyg*. 2023;108(4):712–21. <https://doi.org/10.4269/ajtmh.22-0555>.
3. Daubney R, Hudson J. Enzootic hepatitis or Rift Valley fever; an un-described virus disease of sheep, cattle and man from East Africa. *J Pathol Bacteriol*. 1931;34:545–79.
4. Centers for Disease Control and Prevention. About Rift Valley fever. 2024. Accessed February 2025. <https://www.cdc.gov/rift-valley-fever/about/index.html>
5. Nyakarahuka L, et al. Prevalence and risk factors of Rift Valley fever in humans and animals from Kabale district in Southwestern Uganda. *PLoS Negl Trop Dis*. 2016;12:e0006412.
6. Mohamed M, et al. Epidemiologic and clinical aspects of a Rift Valley fever outbreak in humans in Tanzania. *Am J Trop Med Hyg*. 2010;83:22–7.

7. European Centre for Disease Prevention and Control. Facts about Rift Valley fever. <https://www.ecdc.europa.eu/en/rift-valley-fever/facts>. Accessed 14 Mar 2025.
8. de St Maurice A, et al. Rift Valley fever response: Kabale District, Uganda, March 2016. *MMWR Morb Mortal Wkly Rep*. 2016;65:1200–1.
9. Joint FAO- WHO Experts Consultation. Rift Valley fever outbreaks forecasting models. Food and Agriculture Organisation of the United Nations (FAO) and World Health Organisation (WHO), Rome; 2009. WHO/HSE/GAR/BDP/20092.
10. Linthicum K, Davies F, Kairo A, Bailey C. Rift Valley fever virus (family Bunyaviridae, genus Phlebovirus): isolations from Diptera collected during an inter-epizootic period in Kenya. *Epidemiol Infect*. 1985;95:197–209.
11. Anyamba A, Chretien J, Small J, Tucker C, Formenty P, Richardson J, et al. Prediction of a rift valley fever outbreak. *Proc Natl Acad Sci U S A*. 2009;106:955–9.
12. LaBeaud AD, Muchiri EM, Ndzovu M, Mwanje MT, Muiruri S, Peters CJ, et al. Interepidemic rift valley fever virus seropositivity, northeastern Kenya. *Emerg Infect Dis*. 2008;14(8):1240–6. <https://doi.org/10.3201/eid1408.080082>.
13. LaBeaud AD, Cross PC, Getz WM, Glinka A, King CH. Rift valley fever virus infection in African buffalo (*Syncerus caffer*) herds in rural South Africa: evidence of interepidemic transmission. *Am J Trop Med Hyg*. 2011;84:641–6. <https://doi.org/10.4269/ajtmh.2011.10-0187>.
14. Olive MM, Goodman SM, Reynes JM. The role of wild mammals in the maintenance of Rift Valley fever virus. *J Wildl Dis*. 2012;48(2):241–66. <https://doi.org/10.7589/0090-3558-48.2.241>.
15. Rissmann M, Stoek F, Pickin MJ, Groschup MH. Mechanisms of inter-epidemic maintenance of Rift Valley fever phlebovirus. *Antivir Res*. 2020;174: 104692. <https://doi.org/10.1016/j.antiviral.2019.104692>.
16. Nguku P, et al. An investigation of a major outbreak of Rift Valley fever in Kenya: 2006–2007. *Am J Trop Med Hyg*. 2010;83(Suppl 2):5–13.
17. Rich K, Wanyoike F. An assessment of the regional and national socio-economic impacts of the 2007 Rift Valley fever outbreak in Kenya. *Am J Trop Med Hyg*. 2010;83(Suppl 2):52–7.
18. KMD. Kenya meteorology department-Review of March-April-May (MAM) 2018 seasonal rainfall. Ministry of Environment, Government of Kenya; 2018. Accessed February 2025 <http://www.meteo.go.ke/pdf/seasonal.pdf>
19. Hassan A, Muturi M, Mwatondo A, Omolo J, Bett B, Gikundi S, et al. Epidemiological investigation of a Rift Valley fever outbreak in humans and livestock in Kenya, 2018. *Am J Trop Med Hyg*. 2020;103(4):1649–55. <https://doi.org/10.4269/ajtmh.20-0387>.
20. Oyas H, et al. Enhanced surveillance for Rift Valley fever in livestock during El Niño rains and threat of RVF outbreak, Kenya, 2015–2016. *PLoS Negl Trop Dis*. 2018;12: e0006353.
21. Gibson S, Noronha LE, Tubbs H, Cohnstaedt LW, Wilson WC, Mire C, et al. The increasing threat of Rift Valley fever virus globalization: strategic guidance for protection and preparation. *J Med Entomol*. 2023;60(6):1197–213. <https://doi.org/10.1093/jme/tjad113>.
22. Hawkes AG. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*. 1971;58(1):83–90. <https://doi.org/10.1093/biomet/58.1.83>.
23. Reinhardt A. A review of self-exciting spatio-temporal point processes and their applications. *Stat Sci*. 2018;33(3):299–318. <https://doi.org/10.1214/17-ST-5629>.
24. Meyer S, Held L. Incorporating social contact data in spatio-temporal models for infectious disease spread. *Biostatistics*. 2017;18(2):338–51. <https://doi.org/10.1093/biostatistics/kxw051>.
25. Vigani M, Dudu H, Ferrari E, Mainar A. Estimation of food demand parameters in Kenya A Quadratic Almost Ideal Demand System (QUAIDS) approach. 2019. <https://doi.org/10.2760/479781>.
26. Wan, Z., Hook, S., & Hulley, G. MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V061 [Data set]. NASA Land Processes Distributed Active Archive Center. 2021. Accessed June 2024. <https://doi.org/10.5067/MODIS/MOD11A2.061>.
27. Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J, Shukla S, et al. The climate hazards infrared precipitation with stations - a new environmental record for monitoring extremes. *Sci Data*. 2015;2: 150066. <https://doi.org/10.1038/sdata.2015.66>.
28. Didan K. MODIS/Terra Vegetation Indices Monthly L3 Global 1km SIN Grid V061. 2021. <https://doi.org/10.5067/MODIS/MOD13A3.061>.
29. Kenya National Bureau of Statistics. 2019 Kenya population and housing census. Kenya National Bureau of Statistics (KNBS), vol I. Nairobi; 2019. ISBN: 978-9966-102-09-6.
30. World population annual growth rate. 2020. Accessed February 2025. <https://www.worldpop.org>
31. Benjamin MA, Rigby RA, Stasinopoulos MD. Generalized autoregressive moving average models. *J Am Stat Assoc*. 2003;98:214–23.
32. Yu-Sung S, Masanao Y. R2jags: Using R to Run 'JAGS'. 2021. R package version 0.7-1. Accessed February 2025. <https://cran.r-project.org/web/packages/R2jags/index.html>
33. Allison PD. Logistic regression using SAS: theory and application. 2nd ed. Cary: SAS Institute Inc; 2012.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.