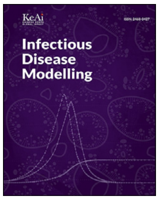
Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm



A parsimonious Bayesian predictive model for forecasting new reported cases of West Nile disease☆

Saman Hosseini [a], Lee W. Cohnstaedt [b, *], John M. Humphreys [c], Caterina Scoglio [a]

[a] Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS, USA
[b] Foreign Arthropod-Borne Animal Diseases Research Unit National Bio- and Agro-defense Facility, USDA ARS, Manhattan, KS, USA
[c] Foreign Animal Disease Research Unit, National Bio- and Agro-defense Facility, USDA ARS, Manhattan, KS, USA

ARTICLE INFO

Article history:
Received 3 November 2023
Received in revised form 18 March 2024
Accepted 15 June 2024
Available online 28 June 2024
Handling Editor: Dr Yijun Lou

Keywords:
West nile virus
ICC curve
Bayesian model
Logistic distribution
HPD credible interval

ABSTRACT

Upon researching predictive models related to West Nile virus disease, it is discovered that there are numerous parameters and extensive information in most models, thus contributing to unnecessary complexity. Another challenge frequently encountered is the lead time, which refers to the period for which predictions are made and often is too short. This paper addresses these issues by introducing a parsimonious method based on ICC curves, offering a logistic distribution model derived from the vector-borne SEIR model. Unlike existing models relying on diverse environmental data, our approach exclusively utilizes historical and present infected human cases (number of new cases). With a year-long lead time, the predictions extend throughout the 12 months, gaining precision as new data emerge. Theoretical conditions are derived to minimize Bayesian loss, enhancing predictive precision. We construct a Bayesian forecasting probability density function using carefully selected prior distributions. Applying these functions, we predict month-specific infections nationwide, rigorously evaluating accuracy with probabilistic metrics. Additionally, HPD credible intervals at 90%, 95%, and 99% levels is performed. Precision assessment is conducted for HPD intervals, measuring the proportion of intervals that does not include actual reported cases for 2020–2022.

1. Introduction

Disease outbreaks and epidemics have had a significant impact on human history and have shaped the trajectory of societies and civilizations (Aberth, 2011; Kenneth, 1993; Snowden, 2019). For instance, pandemics such as the bubonic plague, smallpox, and cholera devastated populations, altered social structures, and even contributed to the fall of empires (Piret & Boivin, 2021). Among infectious diseases, mosquito-borne diseases are of particular concern due to their high transmissibility, making it relatively easy for pathogens to spread from one local area to another within a region by the insect flight or host

☆ This document is the results of the research project funded by the USDA under agreement number 58-3022-1-010.

* Corresponding author.
E-mail address: lee.cohnstaedt@usda.gov (L.W. Cohnstaedt).

Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2024.06.004

movement (World Health Organization (2020)). The situation is exceedingly more complex for the West Nile virus, as it exhibits spreading behavior at two spatial scales: one, local transmission from mosquitoes to a human population, and two, pathogen spread from birds to mosquitoes over long distances (Fig. 1). West Nile disease is a zoonotic disease caused by a virus in the Flaviviridae family that is transmitted among mosquitoes (principally *Culex* spp.), birds, and mammals (Campbell, Marfin, Lanciotti, & Gubler, 2002). Humans and other mammals are not competent hosts and are, therefore, considered 'dead-end hosts' for the virus. Although birds are the primary virus reservoirs, vector mosquitoes can bridge birds to mammals, causing spillover and epidemics in humans (Ciota, 2017). In a few words, many mosquito species belonging to the Culex genus are recognized for their preference to primarily bite and feed on birds, but in the absence of avian hosts, they will opportunistically shift to other species.

The initial outbreak of WNV in the United States occurred in 1999, originating in New York (Kramer, Ciota, & Kilpatrick, 2019; Nash et al., 2001). Over time, the disease spread to affect all of the continental United States. Given the developments in the spread of the West Nile disease across the US and other parts of the world, there is an increased urgency to establish effective methods and policies aimed at preventing the propagation of the disease, reducing the toll on human lives, and mitigating financial consequences (Ronca, Ruff, & Murray, 2021).

Fortunately, in recent decades, various measures have been available to prevent or mitigate the impact of epidemics. If an outbreak occurs, there are methods to predict its dynamics and trajectory, enabling policymakers to implement effective strategies to reduce morbidity, mortality, and economic damage. Mathematical modeling (statistical and mechanistic) has proven to be one of the most effective ways to control and predict disease dynamics, including pandemics (Biggerstaff, Slayton, Johansson, & Butler, 2021). Both mechanistic and statistical models have been used to this purpose. Mechanistic models prove especially valuable in elucidating the interplay among diverse parameters, encompassing the disease and epidemic factors. They aid in forecasting epidemic dynamics and devising mitigation strategies. On the other hand, statistical models are effective for forecasting and prediction purposes. The most recent tangible situation is related to COVID-19. Scholars used many models to predict the virus's risk and spread, assess the effectiveness of interventions such as social distancing and vaccination, and estimate the potential impact on healthcare systems and the economy (Reich et al., 2022).

In the context of predicting the number of WNV infections, some excellent works have been published in recent years. These works are extremely diverse and can be divided based on methodological points of view. In the study by Kovach et al. (2018) (Kovach & Kilpatrick, 2018), a spatial analysis based on regression was conducted to explore potential correlations between land use and West Nile disease at the county level in California. Among the relevant studies in the field of early warning systems, Davis et al. (2018) is notable. This work employs a county-level distributed lag model in its methodology, incorporating human, weather, and mosquito infection data. Furthermore, Peper et al. (2018) utilize data from traps to predict the likelihood of encountering infected mosquitoes. This prediction is based on weather variables as influencing factors. In a similar vein, Poh et al. (2019) employ time series modeling and techniques to forecast mosquito WNV infections. Regarding early detection, we can refer to the works of Myer and Johnston (2019) and Humphreys, Young, Cohnstaedt, Hanley, and Peters (2021), who employed Bayesian models to enhance the accuracy of spatiotemporal predictions of West Nile disease. In the realm of mechanistic models for evaluating spatial risk assessment, notable contributions include the works of (Bergsman, Hyman, & Manore, 2016; Di Pol, Crotta, & Taylor, 2022). Also, Sifat has established a Bayesian network framework model for predicting the pattern of WNV spread in the United States in 2019 (Moon, Cohnstaedt, McVey, & Scoglio, 2019).

Models using different kinds of data vary in terms of their complexity and information they incorporate, such as temperature, humidity levels, weather data, mosquito infection data, and even economic indicators. While it is difficult to disregard the potential enhancement in accuracy that incorporating these data into the model could bring, it's important to note that, in many instances, these data types necessitate the estimation of varying numbers of parameters. This inherently introduces complexity and imposes additional computational overhead when executing the model. Furthermore, data

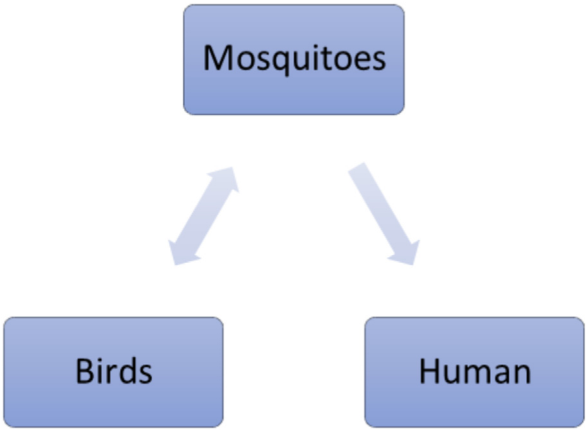


Fig. 1. The mechanism of West Nile virus transmission.

collection presents an additional hurdle. In numerous cases, securing precise data requires a significant expenditure of time and energy. Additionally, selecting the appropriate data introduces another challenge. For instance, if we intend to incorporate temperature data in a given location, we must ascertain whether to include directly the temperature or utilize a particular function derived from temperature. Another crucial issue that demands consideration is lead time. In many models, the lead time is typically within the range of days or weeks, which fails to provide a comprehensive perspective for longer periods, such as half a year or an entire year. Yet, these longer timescales are essential for developing and establishing significant preventive policies against epidemic outbreaks. Consequently, the pursuit of a parsimonious model, characterized by the least complexity while still delivering an all-encompassing predictive perspective for the entire year, takes on a role of paramount importance and significant value.

The objective of this study was to develop a parsimonious yet accurate Bayesian model that solely relied on human incidence data as input and demonstrated a capability to provide comprehensive year-long prediction. To accomplish this, we extended and implemented the ICC-Curve (Incidence vs. Cumulative Cases) technique to estimate the number of next-phase WNV infections (Lega, 2021). Incidence estimates were subsequently used to construct a probability density function (Bayesian perspective and a simplified probability model) for disease forecasting. Once these probability density functions were established, we derived Bayesian credible intervals. In a final phase, our method was applied to forecast and then assess results for all U.S. states through qualitative comparison to outcomes from the CDC (Centers for Disease Control and Prevention) WNV forecasting challenge (Holcomb et al., 2023) and numeric scoring using the logarithm of the probability of occurrence (Rosenfeld, Grefenstette, & Burke, 2012).

2. Estimation based on ICC curves for west nile virus

In this section, we will delve into the mathematical aspects of ICC curves and provide mathematical proof, showcasing their suitability for the SEIR vector-borne model of WNV. However, before presenting this method's mathematical verification, it is necessary to introduce the model that characterizes West Nile transmission. In this non-linear differential equation model, there are three populations of mammals (or humans), birds, and mosquitoes. The populations are denoted by H for humans, B for birds, and M for mosquitoes. The transmission rate from one population to another is represented by the symbol β, while the incubation rate is denoted by δ, and the recovery rate is represented by λ. The abbreviations S, E, I, and R represent the fraction of individuals in the susceptible, exposed, infectious, and recovered states, respectively. In the context of the human population, there isn't a distinct infectious phase. Instead, we have designated the "I_{Sy}" phase to represent the period during which an infected individual displays the symptoms of the disease. Alternatively, one can merge the exposed and infected (or symptomatic) phases into a single phase, denoted as I_H, where $I_H = E_H + I_{Sy}$.

The system of nonlinear equations provided below elucidates the mechanism behind this illness:

$$\begin{aligned}
\frac{dS_H}{dt} &= -\beta_{M\to H} S_H I_M;\\
\frac{dE_H}{dt} &= \beta_{M\to H} S_H I_M - \delta_H E_H;\\
\frac{dI_{Sy}}{dt} &= \delta_H E_H - \lambda_H I_{Sy};\\
\frac{dR_H}{dt} &= \lambda_H I_{Sy}.
\end{aligned}$$

Based on the previous explanation, it is possible to simplify these four compartmental equations into three, as follows:

$$\begin{aligned}
\frac{dS_H}{dt} &= -\beta_{M\to H} S_H I_M;\\
\frac{dI_H}{dt} &= \beta_{M\to H} S_H I_M - \lambda'_H I_H;\\
\frac{dR_H}{dt} &= \lambda'_H I_H.
\end{aligned}$$

For bird population:

$$\begin{aligned}
\frac{dS_B}{dt} &= -\beta_{M\to B} S_B I_M;\\
\frac{dE_B}{dt} &= \beta_{M\to B} S_B I_M - \delta_B E_B;\\
\frac{dI_B}{dt} &= \delta_B E_B - \lambda_B I_B;\\
\frac{dR_B}{dt} &= \lambda_B I_B,
\end{aligned}$$

and mosquito population:

$$\begin{aligned}
\frac{dS_M}{dt} &= -\beta_{B\to M} S_M I_B;\\
\frac{dE_M}{dt} &= \beta_{B\to M} S_M I_B - \delta_M E_M;\\
\frac{dI_M}{dt} &= \delta_M E_M - \lambda_M I_M;\\
\frac{dR_M}{dt} &= \lambda_M I_M.
\end{aligned}$$

It's important to highlight that within these equations, $\beta_{Population1\to Population2}$ signifies the infection rate from the infectious population 1 to the susceptible population 2. Previous research (Lega, 2021) has demonstrated the effectiveness of the ICC curve method in the SIR model. However, the mechanism of the West Nile disease is distinct as it involves at least three populations in the transmission process, and it follows the SEIR (Susceptible, Exposed, Infected, and Recovered) model. The subsequent theorem clarifies the functioning of the ICC curve method concerning the West Nile disease, taking into account the described transmission mechanism.

Theorem I. *If $C_T(t)$ represents the cumulative number of infected cases up to time t in the SEIR model of West Nile disease, where $C_T(t) = E_H + I_{Sy} + R_H$, then:*

i) The cumulative number of cases at time t, $C_T(t)$, and the rate density function $r_T(t)$ can be expressed as follows:

$$C_T(t) = \frac{L}{1+\exp\{-\delta(t-\mu)\}}; \quad t>\mu, \delta>0 \tag{1}$$

and

$$r_T(t) = \frac{L\delta \exp\{-\delta(t-\mu)\}}{(1+\exp\{-\delta(t-\mu)\})^2}, \tag{2}$$

where L represents the total number of cases at the end of the epidemic.

ii) $r_T(t)$ is a parabolic function of $C_T(t)$.

Proof: Taking into account the subsequent non-linear model:

$$\frac{dS_H}{dt} = -\beta_{M\to H} S_H I_M; \tag{3}$$

$$\frac{dI_H}{dt} = \beta_{M\to H} S_H I_M - \lambda'_H I_H; \tag{4}$$

$$\frac{dR_H}{dt} = \lambda'_H I'_H. \tag{5}$$

The Cumulative Number of cases is defined as $C_H = I'_H + R_H$ (or $C_H = E_H + Sy_H + R_H$). By summing the non-linear model equations (4) and (5), we can derive the following relationship for C_H:

$$\frac{dC_H}{dt} = \frac{dI'_H}{dt} + \frac{dR_H}{dt} = \beta_{M\to H} S_H I_M. \tag{6}$$

Furthermore, by dividing (3) by (4),

$$\left(\frac{dS_H}{dt}\right) \Big/ \left(\frac{dR_H}{dt}\right) = \frac{dS_H}{dR_H} = \frac{-\beta_{M\to H} S_H I_M}{\lambda_H I'_H}. \tag{7}$$

Additionally, a clear connection exists between the infectious compartment of mosquitoes and the infected compartment of humans. In essence, as the number of infectious mosquitoes increases, the number of infected individuals among humans also tends to rise. Hence, it is reasonable to posit that there exists a functional relationship between these two variables, denoted as $I_M = g(I'_H)$. Here, g represents a real and continuous function. To further understand this relationship over time, we can employ a linear approximation, such as a Taylor expansion, by expressing I_M as approximately equal to a constant multiplier 'm' times I'_H, i.e., $I_M \simeq mI'_H$. This linear approximation helps us to capture the proportional change in the number of infectious mosquitoes concerning the infected humans. Considering this note, we can reformulate (6) and (7) as follows:

$$\frac{dC_H}{dt} = m\beta_{M\to H} S_H I'_H, \tag{8}$$

and consequently, we have:

$$\frac{dC_T(t)}{dt} = \frac{dN_C}{dt} = \frac{m}{N}\beta_{M\to H} N_S N_{I'}. \tag{9}$$

Additionally:

$$\frac{dS_H}{dR_H} = \frac{-m\beta_{M\to H} S_H}{\lambda_H}. \tag{10}$$

We should highlight that C_H represents the proportion of cumulative cases, whereas $C_T(t) = N_C = N \cdot C_H$ (where N is the population size) represents the number of cumulative cases.From (10), it is concluded that:

$$S_H = \exp\left(-\frac{m\beta_{M\to H}}{\lambda_H} R_H\right), \tag{11}$$

and we know that the number of susceptible equals $N_S = NS_H$. Also, it is clear that $C_T(t) = N - NS_H = N - N_S$, so by expanding exponential function (11):

$$\begin{aligned} C_T(t) &= N(1 - S_H) \\ &= \frac{m\beta_{M\to H}}{\lambda_H} NR_H \\ &= \frac{m\beta_{M\to H}}{\lambda_H} N_R \to N_R = \frac{\lambda_H}{m\beta_{M\to H}} C_T(t). \end{aligned}$$

Also, we know that $I'_H = C_H - R_H$ and consequently:

$$N_{I'} = C_T(t) - N_R = \left(1 - \frac{\lambda_H}{m\beta_{M\to H}}\right) C_T(t). \tag{12}$$

Taking into account (9), at the beginning of the epidemic ($S = N$) and (12):

$$\frac{dC_T(t)}{dt} = \lambda_H \left(\frac{m\beta_{M\to H}}{\lambda_H} - 1\right) C_T(t). \tag{13}$$

Also, we know that if L is the total number of infected at the end of the epidemic, the differential equation at that point must be zero. It means:

$$\frac{dC_T(t)}{dt} = W(N_C) \times (L - C_T(t))$$

or equivalently:

$$\frac{dC_T(t)}{dt} = L \times W(C_T(t)) \times \left(1 - \frac{C_T(t)}{L}\right). \tag{14}$$

So, the differential equation must satisfy both the linear relations (13) and non-linear equation (14) at the beginning and end of the epidemic. It's evident that by multiplying $\left(1 - \frac{N_C}{L}\right)$ by (13), we can derive the desired differential equation:

$$\frac{dC_T(t)}{dt} = \lambda_H \left(\frac{m\beta_{M\to H}}{\lambda_H} - 1\right) C_T(t) \times \left(1 - \frac{C_T(t)}{L}\right). \tag{15}$$

As a result, we can conclude that this is indeed a logistic differential equation and consequently, part i of Theorem I is proven. Furthermore, it becomes apparent from the right hand of equation (15) that the rate $\frac{dN_C}{dt}$ (denoted as $r_T(t)$) exhibits an inverse parabolic relationship with respect to the cumulative number of cases $C_T(t)$.

In practice, we can use this theorem to fit a parabola to the rate-cumulative coordinates at the beginning of the epidemic. This parabola will help us estimate the total number of cases at the end of the epidemic, denoted by L. We can then fit the logistic function (1) to the available data and use L as the maximum value of the logistic function to estimate the cumulative number of cases at time t, and consequently, the number of new cases. Fig. 2 illustrates this process.

3. Bayesian approach

There is a distinction between prediction and forecasting in this paper. The term prediction refers to a specific value estimation (point estimation), whereas forecasting entails the use of a probability density function to describe the likelihood associated with each value of the prediction. When we forecast, we simultaneously provide uncertainty along with the predictions. In other words, forecasting represents the probability density function associated with the point estimation. Forecasting is often considered more interesting because relying solely on a single estimated value for the future of a system, like the estimated number of newly reported cases, may not be dependable. So, using a probability density function is preferable as it allows us to focus on specific domains of the parameter range (predicted values) with higher density. However, it's important to recognize that prediction and forecasting are valuable for gaining insights into the system's future. In practice, the prediction value is used to refine the forecasting distribution, underscoring the significance of accuracy in both the prediction and the proposed distribution. After obtaining a point prediction as described in section 2, we have proposed both a Bayesian approach and a straightforward method to construct a probability density function. This was achieved by updating our beliefs toward the distribution of numbers on new cases based on the value of the point prediction. To do so, we initially acknowledged that the total number of new cases L during each outbreak can be represented by a Poisson random variable. Each outbreak can result in either a natural number of infected cases or zero infected cases. In addition to the Poisson distribution, another viable option for the prior probability density function is the normal distribution. It is conceivable to assume that the maximum number of infected individuals during an outbreak follows a normal distribution (in fact, a censored normal distribution because number of the new cases should be non-negative), as is the case with various natural phenomena (Frank, 2009).

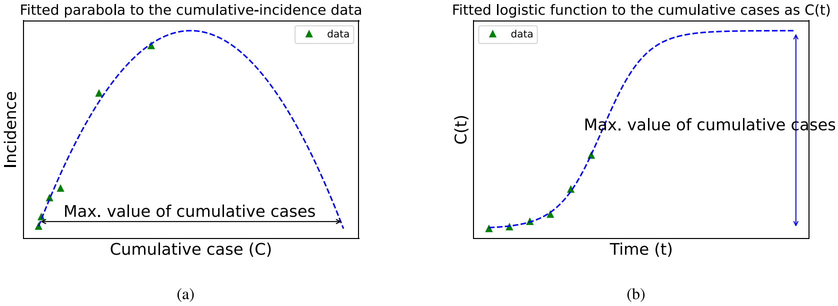


Fig. 2. Panel (a) displays the parabola fitted to the cumulative versus new cases, and Panel (b) exhibits the fitted logistic function to the data.

3.1. Forecasting and probability density function

To commence, it is essential to mention a noteworthy aspect related to fitting curves to the data and their impact within the Bayesian approach. The point estimate derived by the ICC-Curve method possesses certain distinctive characteristics that make it suitable for reliable use.

Based on the details presented in the preceding section, the estimation of cumulative cases at a given time ‘t’ using the ICC-curve approach can be expressed as follows:

$$C_T(t) = \frac{L_0}{1+e^{-\delta_0(t-\mu_0)}}, \tag{16}$$

The estimated parameters L_0, δ_0, and μ_0 are obtained through the ICC-curve method. However, it is worth mentioning that there are alternative approaches to fitting the parabola and logistic function. For instance, one can employ regression-based methods or built-in functions in Python libraries like NumPy, or opt for fitting using the MCMC (Markov Chain Monte Carlo) method. It is important to note that using different methods to fit the parabola and logistic function will yield different distributions for L_0, δ_0, and μ_0, consequently leading to different distributions for the estimations on the number of cumulative and new cases. This variation in estimation methods can result in different outcomes, such as bias or unbiasedness, which will be discussed in the subsequent sections. The key point to grasp from this explanation is that the distributions of L_0, δ_0, and μ_0 are influenced by the chosen method of curve fitting. In general, if we use delta method to fit the curves, it implies that the estimations will have distributions based on this particular method. It means if we have a sample $\underline{X} = (X_1, \ldots, X_n)$, the probability density function of $L_0(\underline{X}, \Delta)$ (estimation based on this sample and delta method) will vary depending on the chosen method delta or;

$$L_0(\underline{X}, \Delta) \sim g_\Delta, \tag{17}$$

in which g_Δ is the probability density function of L_0 regarding delta method. It has to be mentioned that we are not delving into this subject in depth, and we won't establish Bayesian estimations and posteriors through fitting techniques. Instead, we'll simply employ the estimated value of $L_0(\underline{X})$ as a parameter for other prior density functions, which will be elaborated upon in the next theorem.

If we have a close look at the ICC-curve estimator for the cumulative number of infected $\left(C(T) = \frac{L}{1+e^{-\delta_0(T-\mu_0)}}\right)$, it is evident that this estimator is distributed uniformly, that is, $C(T) \sim U(0, L)$. Considering $\pi_L(l)$ as a prior for L it is possible to obtain the general form of the posterior probability density function:

$$P(L=l|C(T)=c) = \frac{f_{C(T)}(c|L)\pi_L(l)}{\sum\limits_{l\in\Theta} f_{C(T)}(c|L)\pi_L(l)} = \frac{\frac{1}{l}\pi_L(l)}{\sum\limits_{l\in\Theta}\frac{1}{l}\pi_L(l)} \tag{18}$$

It should be mentioned that the maximum number of cumulative cases at the end of the epidemic (L) can be any natural number theoretically, but we expect to have a number around the value that is estimated by the ICC-curve method ($L_0(X_1, X_2, \ldots, X_n)$), so, for example, it makes sense to choose a censored Poisson at zero (because based on (18) the value of L can not take zero) with the parameter of $L_0(X_1, \ldots, X_n)$ as the prior density function. Taking all of them into consideration:

$$P(L=l|C(T)=c) = \frac{f_{C(T)}(c|L)\pi_L(l)}{\sum\limits_{l\in\Theta} f_{C(T)}(c|L)\pi_L(l)} = \frac{\frac{\left(L_0\left(\underset{\sim}{x}\right)\right)^l}{l\times l!}}{\sum\limits_{l=1}^{+\infty}\frac{\left(L_0\left(\underset{\sim}{x}\right)\right)^l}{l\times l!}} \tag{19}$$

We have to notice that (19) is the probability density function of $L|C(T) = c$, but what we need is to obtain the probability density function of $\left(\frac{L}{1+e^{-\delta(t-\mu)}}\middle|C(T) = c\right)$.

Before proceeding, it is important to note that in many cases, the prediction value $\left(\frac{L}{1+e^{-\delta(t-\mu)}} = y\right)$ is a real number, as it corresponds to the value of a quantity (1), which functions as the cumulative distribution function (CDF) for a continuous variable. However, since the number of cumulative cases is a natural number, we need to consider only the natural values of this function. Taking this note into consideration, the probability density function for forecasting can be derived in the following manner. The subsequent relationship aids in providing a clear understanding of this distinction.

If $A = 1 + e^{-\delta(t-\mu)}$ and $B = \sum\limits_{l=1}^{+\infty}\frac{\left(L_0\left(\underset{\sim}{x}\right)\right)^l}{l\times l!}$

$$P\left(\frac{L}{1+e^{-\delta(T-\mu)}}=y\middle|C(T)=c\right)=$$

$$P\left(\frac{L}{1+e^{-\delta(t-\mu)}}=y\middle|C(T)=c\right)=$$

$$\sum_{\left\{z\middle|\lfloor\frac{z}{A}\rfloor=y\in\mathbb{N}\right\}}\frac{\left(L_0\left(\underset{\sim}{x}\right)\right)^z}{B\times z\times z!}$$

For example if $A = 1.56$ then:

$$P\left(\frac{L}{1+e^{-\delta(T-\mu)}}=1\middle|C(T)=c\right)=$$

$$\sum_{\left\{z\middle|\left\lfloor\frac{z}{A}\right\rfloor=1\in\mathbb{N}\right\}}\frac{\left(L_0\left(\underset{\sim}{x}\right)\right)^l}{B\times l\times l!}=$$

$$\sum_{\{2,3\}}\frac{\left(L_0\left(\underset{\sim}{x}\right)\right)^z}{B\times z\times z!}\qquad=\left\{\frac{\left(L_0\left(\underset{\sim}{x}\right)\right)^2}{4B}+\frac{\left(L_0\left(\underset{\sim}{x}\right)\right)^3}{18B}\right\}.$$

Because

$$\left\{z\middle|\lfloor\frac{2}{A}\rfloor=\lfloor\frac{2}{1.56}\rfloor=1\in\mathbb{N}\right\}=\{2\}$$

$$\left\{z\middle|\lfloor\frac{z}{A}\rfloor=\lfloor\frac{3}{1.56}\rfloor=1\in\mathbb{N}\right\}=\{3\}.$$

Note:

When we possess a probability density function for $\hat{C}(t)$, we can calculate the density of new cases $(\hat{N}_T(t))$ as follows:

$$\pi_{\hat{N}(t)(y|C(T)=c)}=P(\hat{C}(t)-c(t-1)=y|C(T)=c). \tag{20}$$

where $c(t-1)$ is the real value for cumulative cases up to time $t-1$.

3.2. Bayesian estimation and posterior density function based on a Poisson prior

This subsection shall be initiated with the introduction of a theorem. This theorem will lay out the fundamental concept behind the process of selecting priors.

Theorem II. *The ICC-Curve estimator, with respect to the quadratic loss function and prior $\pi_L(l)$, represents a Bayesian estimation with the least value of risk function for a cumulative number of cases at time t if $E\left(\frac{1}{L}\right)=\frac{1}{L_0(x_1,x_2,\ldots,x_n)}$.*

Proof: The theorem states that $\hat{C}(t)=\frac{L_0(X_1,\ldots,X_n)}{1+e^{-\delta(t-\mu)}}$ minimizes the Bayesian risk function of the form

$$r(\pi,C(t))=\sum_{l\in D_L}\left\{\int_{t\in D_T}Loss(\gamma(l),\hat{C}(t))f_T(t;l)d\nu(t)\right\}\pi(l), \tag{21}$$

for estimation of $\gamma(L)=\frac{L}{1+e^{-\delta(t-\mu)}}$ if the prior density function satisfies equality of $E\left(\frac{1}{L}\right)=\frac{1}{L_0(x_1,x_2,\ldots,x_n)}$.
The value that minimizes the Bayesian risk function (21) is $E(\gamma(L)|C(T)=c)$ (Berger (2013)), so:

$$\begin{aligned} E(\gamma(L)|C(T)=c) &= E\left(\frac{L}{1+e^{-\delta(T-\mu)}}\middle|C(T)=c\right)\\ &= E\left(\frac{L}{1+e^{-\delta(t-\mu)}}\middle|C(T)=c\right)\\ &= \frac{1}{1+e^{-\delta(t-\mu)}}E(L|C(T)=c). \end{aligned} \tag{22}$$

Using (18) it is possible to find out the value of (22):

$$E(L|C(T)=c) =$$

$$\sum_{l\in\Theta} l\cdot f_{(L|C(T)=c)}(l) \quad = \sum_{l\in\Theta} l\frac{\frac{1}{l}\pi_L(l)}{\sum_{l\in\Theta}\frac{1}{l}\pi_L(l)} = \frac{1}{E\left(\frac{1}{L}\right)}$$

This signifies that if $E\left(\frac{1}{L}\right)\simeq\frac{1}{L_0(X_1,\ldots,X_n)}$, then we have the following relations:

$$\begin{aligned} E(\gamma(L)|T=t) &= \frac{E(L|T=t)}{\left(1+e^{-\delta(t-\mu)}\right)}\\ &= \frac{1}{E\left(\frac{1}{L}\right)\left(1+e^{-\delta(t-\mu)}\right)} \qquad = \frac{1}{\frac{1}{L_0(X_1,\ldots,X_n)}\left(1+e^{-\delta(t-\mu)}\right)}\\ &= \frac{L_0(X_1,\ldots,X_n)}{\left(1+e^{-\delta(t-\mu)}\right)}, \end{aligned}$$

which equals the estimator derived from the ICC-curve technique.

For instance, consider a censored Poisson at zero. $\left(\pi_L(l) = \frac{e^{-\lambda}\lambda^l}{l!(1-e^{-\lambda})}I_{\{1,2,\ldots\}}(l)\right)$ as a prior distribution. If the estimated value, based on the ICC-curve method, for.$L_0(X_1,\ldots,X_n)$ is 50, a simple calculation reveals that $E\left(\frac{1}{L}\right) = 0.02042$ so in this case:

$$E\left(\frac{1}{L}\right)\simeq\frac{1}{L_0(X_1,\ldots,X_n)}.$$

Consequently, in this scenario, by selecting this prior and employing a quadratic loss function, estimation based on the ICC-curve method is a Bayesian estimation with the least value of Bayesian risk. Generally, it is possible to prove the above results about censored Poisson. Theorem III Given a sufficiently large value of estimation for L (practically greater than 5), estimation based on the ICC curve method with respect to zero-censored Poisson prior and quadratic loss function is a Bayesian estimation with the least value of risk function.

Proof: Referring to Theorem II, our task now simplifies to demonstrating: $E\left(\frac{1}{L}\right)\simeq\frac{1}{L_0\left(\underset{\sim}{x}\right)}$, so:

$$E\left(\frac{1}{L}\right) = \sum_{l=1}^{+\infty}\frac{1}{l}\,\frac{e^{-L_0\left(\underset{\sim}{x}\right)}\left(L_0\left(\underset{\sim}{x}\right)\right)^l}{l!\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}.$$

We know that if the value of the parameter from Poisson distribution is large its density function is almost symmetrically distributed around its parameter (Fig. 3). From this fact, we can find that the probability of the small natural values of the Poisson variable (beginning of the domain) is negligible, and it is possible to consider them as zeros. So, just large values have a density (it is called effective domain W).

Additionally, from the assumption of largeness of $L_0(X_1, \ldots, X_n)$, we can assume $L \simeq L+1$, so:

$$
\begin{aligned}
E\left(\frac{1}{L}\right) &= \sum_{l=1}^{+\infty}\frac{1}{l}\,\frac{e^{-L_0\left(\underset{\sim}{x}\right)}\left(L_0\left(\underset{\sim}{x}\right)\right)^l}{l!\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}\\
&= \sum_{l\in W}\frac{1}{l}\,\frac{e^{-L_0\left(\underset{\sim}{x}\right)}\left(L_0\left(\underset{\sim}{x}\right)\right)^l}{l!\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}\\
&= \sum_{l\in W}\frac{e^{-L_0\left(\underset{\sim}{x}\right)}\left(L_0\left(\underset{\sim}{x}\right)\right)^{l+1}}{(l+1)!\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}\,\frac{1}{L_0\left(\underset{\sim}{x}\right)}\\
&= \frac{1}{L_0\left(\underset{\sim}{x}\right)\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}\sum_{l\in W}\frac{e^{-L_0\left(\underset{\sim}{x}\right)}\left(L_0\left(\underset{\sim}{x}\right)\right)^{l+1}}{(l+1)!}\\
&= \frac{1}{L_0\left(\underset{\sim}{x}\right)\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}.
\end{aligned}
$$

But for large values of L_0, it is clear that $e^{-L_0\left(\underset{\sim}{x}\right)} \simeq 0$, and consequently, $E\left(\frac{1}{L}\right) \simeq \frac{1}{L_0\left(\underset{\sim}{x}\right)}$.

Table (1) shows that this approximation works very well, even for small values. **Note**: It can be shown that when the estimated value of $L_0(X_1, \ldots, X_n)$ is a large value and the prior probability density function follows a zero-censored Poisson distribution, then $E\left(\frac{1}{L}\right) \simeq \frac{1}{E(L)}$.

To demonstrate this, consider a zero-censored Poisson distribution with the parameter $L_0(X_1, \ldots, X_n)$:

$$
\begin{aligned}
E(L) &= \sum_{l=1}^{+\infty} l\,\frac{e^{-L_0\left(\underset{\sim}{x}\right)}\left(L_0\left(\underset{\sim}{x}\right)\right)^l}{l!\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}\\
&= \frac{L_0\left(\underset{\sim}{x}\right)}{\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}\sum_{l=1}^{+\infty}\frac{e^{-L_0\left(\underset{\sim}{x}\right)}\left(L_0\left(\underset{\sim}{x}\right)\right)^{l-1}}{(l-1)!} \simeq\\
&= \frac{L_0\left(\underset{\sim}{x}\right)}{\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}\sum_{l=0}^{+\infty}\frac{e^{-L_0\left(\underset{\sim}{x}\right)}\left(L_0\left(\underset{\sim}{x}\right)\right)^{l}}{(l)!} = \frac{L_0\left(\underset{\sim}{x}\right)}{\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}.
\end{aligned}
$$

But the term $\frac{1}{\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}$ approximately equals to one, as:

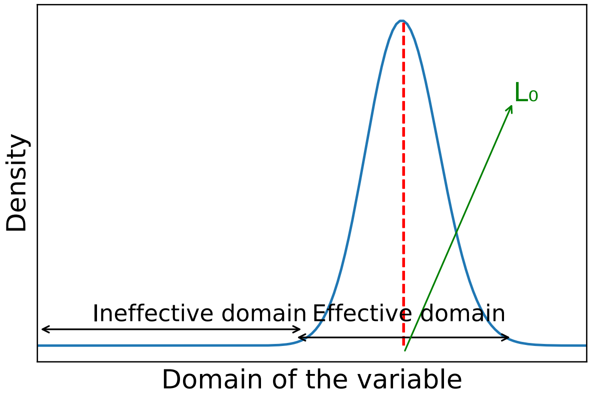


Fig. 3. A Poisson probability density function with a large parameter value.

$$L_0\left(\underset{\sim}{x}\right)\to+\infty\Rightarrow e^{-L_0\left(\underset{\sim}{x}\right)}\to 0\Rightarrow\frac{1}{\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}\to 1,$$

and means $E(L)\ =\frac{L_0\left(\underset{\sim}{x}\right)}{\left(1-e^{-L_0\left(\underset{\sim}{x}\right)}\right)}\simeq L_0\left(\underset{\sim}{x}\right)$. Additionally, during the proof of Theorem III, we observed that $E\left(\frac{1}{L}\right)\simeq\frac{1}{L_0\left(\underset{\sim}{x}\right)}$ and it means $E\left(\frac{1}{L}\right)\simeq\frac{1}{E(L)}$.

3.3. A simple forecasting model using Poisson probability density function

In this part, we'll establish a predictive model utilizing the Poisson probability density function, bypassing the calculation of the Bayesian posterior. Indeed, the objective of this subsection is to demonstrate that even simple models can yield effective results, and it will be demonstrated that the accuracy of these models is exceptionally high, as highlighted in the concluding section of the paper. This showcases the potential of achieving favorable outcomes by employing straightforward models.

To establish the probability density function, we make the assumption that the total number of infected people at the epidemic's conclusion denoted as L, follows a Poisson distribution with a parameter L_0. Here, L_0 is determined using the ICC-Curve method for prediction. Indirectly, this implies that if we were to repeatedly experience an epidemic and record the total number of infected cases at its conclusion, these recorded values would closely align with L_0. Building upon this introductory explanation, we can derive the probability density function for the cumulative number of cases at the time 't', denoted as $\hat{C}_T(t)$. To demonstrate this mathematically, we proceed with the following steps:

$$P\left(\overset{\wedge}{C_T}(t)\ =y\right)=P\left(\frac{L}{1+e^{-\delta(t-\mu)}}=y\right).$$

Given that y is a whole number, we need to select the corresponding values of $L=0,1,2,3,\ldots$for each y in such a way that $C(T)$ approximates y. Recognizing that $\frac{L}{1+e^{-\delta(t-\mu)}}$ is typically a real value and needs to approximate integers, we proceed as follows:

$$P(\hat{C}(T)=y)=\sum_{\left\{z\left|\left\lfloor\frac{z}{1+e^{-\delta(t-\mu)}}\right\rfloor=y\right.\right\}}\frac{e^{-L_0}{L_0}^z}{z!},$$

and consequently

$$P(\hat{N}(T)=y)=P(\hat{C}_T(t)-c(t-1)=y). \tag{23}$$

In the final section, we will observe the remarkable outcomes achieved through the utilization of a simple forecasting probability density function.

3.4. Highest posterior density (HPD) credible intervals for cumulative and new reported cases

The highest posterior density credible intervals are obtained by solving the following equation and inequality for the cumulative number of cases and number of new cases, respectively:

$$
\begin{aligned}
I_k &= \left\{ y \middle| \pi_{\left(C_T\hat{(t)}|T=t\right)}(y) \geq k, \right. \\
&\left. \sum_{\left\{ y \middle| \pi_{\left(C_T\hat{(t)}|T=t\right)}(y) \geq k \right\}} \pi_{\left(C_T\hat{(t)}|T=t\right)}(y) \leq \alpha \right\} \\
I'_k &= \left\{ y \middle| \pi_{\left(N_T\hat{(t)}|\, T=t\right)}(y) \geq k, \right. \\
&\left. \sum_{\left\{ y \middle| \pi_{\left(N_T\hat{(t)}|\, T=t\right)}(y) \geq k \right\}} \pi_{\left(N_T\hat{(t)}|\, T=t\right)}(y) \leq \alpha \right\}
\end{aligned}
\tag{24}
$$

By utilizing the posterior obtained in equation (20) and applying the definition of HPD (Highest Posterior Density) credible intervals in equation (24), it becomes feasible to determine the intervals numerically (Fig. 4).

4. Modified probability density function for west nile virus

Having established a probability density function for the predicted values, assessing these predictive PDFs using appropriate measures is only logical. As mentioned in the introduction of this study, our evaluation has been conducted using probabilistic measures, specifically the logarithm of the probability for accurately predicted values.

To gauge the accuracy of our measurements, we adhered to established standards, aligning our approach with the guidelines set forth by the Centers for Disease Control and Prevention (CDC). Following the guidelines, the domain of the predictive PDF has been divided into multiple bins. It is important to note that these bins have practical significance and were derived from the expertise of national health agents, representing average values. The bins are defined as individual values of zero, ranges [1,5], [6,10], …, [46,50], [50,100], [100,150], …, [200,250]. The logarithmic accuracy score is calculated as the logarithm of the probability corresponding to the bin within which the reported value falls, and if the value falls out of these bins, the score is set to −10. A favorable prediction outcome corresponds to negative values that approach zero, as the logarithm of the probability inherently yields negative values.

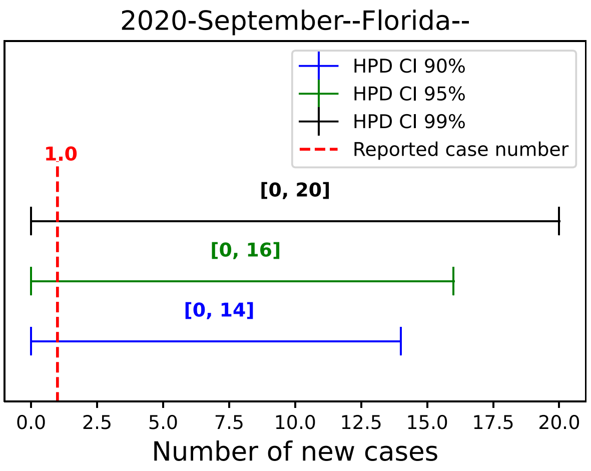


Fig. 4. The highest posterior density credible intervals for the number of new cases in Florida during September and August 2020, with confidence levels of 99%, 95%, and 90%.

Based on information from the CDC, a significant portion of counties (88%) report zero cases of the disease, approximately 11.5% report 1–10 cases, and 0.4% report 11–50 cases. The yearly maximum case count per county ranges from 18 to 239 cases during the period from 2005 to 2020 (Holcomb et al., 2023). This observation suggests that the density of zero value in the predictive PDF should be adjusted to account for excess zeros. We have adopted a strategy of modifying the PDF(s) by adjusting the density at zero in two distinct scenarios.

4.1. Zero-inflated probability density function

Based on the explanation in section 4, we increased the density of the zero value in the forecasting probability density functions to improve forecast accuracy.

In probability density functions (20) and (23), we multiplied the density of zero by the factor α, resulting in modified PDFs as:

$$\tilde{\pi}_{\mathrm{N}(\hat{\mathrm{t}})}(y|T=t)=\begin{cases}\alpha\pi \\ \quad {}_{\hat{N}(t)(y|T=t)} \qquad \text{if } y=01-\alpha\dfrac{\pi_{\hat{N}(t)(0|T=t)}}{1-\pi_{\hat{N}(t)(0|T=t)}}\pi_{\hat{N}(t)(y|T=t)} \qquad \text{if } y>0, \\ \end{cases} \tag{25}$$

and

$$\tilde{\boldsymbol{P}}(\hat{N}(t)=y)=\begin{cases}\alpha P(\hat{N}(t)=0) \qquad \text{if } y=01-\alpha P(\dfrac{\hat{N}(t)=0)}{1-P(\hat{N}(t)=0)}P(\hat{N}(t)=y) \qquad \text{if } y>0.\end{cases} \tag{26}$$

Determining the appropriate value of $\alpha > 0$ can be achieved through various approaches and methods. In our initial attempt, increasing the emphasis on zero from 1 to 3 yielded a significant improvement in the result, enhancing it by 40%. The strategy for estimating the optimal value of α (for the upcoming year) involves estimating the optimal α values for previous years to project the α value for the upcoming year. We have the number of infected cases ($x_{t,A}$) for each month from 2003 to 2022: $x_{2003,A}, x_{2004,A}, x_{2005,A}, x_{2006,A}, \ldots, x_{2021,A}, x_{2022,A}$.

This data set is divided into different vectors as follows;

1st:

$$x_{2003,A}, x_{2004,A}, x_{2005,A}, x_{2006,A}, \cdots, x_{2010,A}$$

2nd:

$$x_{2003,A}, x_{2004,A}, x_{2005,A}, x_{2006,A}, \cdots, x_{2010,A}, x_{2011,A}$$

3rd:

$$x_{2003,A}, x_{2005,A}, x_{2006,A}, \cdots, x_{2010,A}, x_{2011,A}, x_{2012,A}$$

23rd:

$$x_{2003,A}, \cdots, x_{2012,A}, x_{2013,A}, \cdots, x_{2020,A}, x_{2021,A}, x_{2022,A}$$

The first vector is used to find out predictive zero concentrated, which has been shown by $\tilde{\pi}$ and $\tilde{P}$ (26 and 25), and consequently prediction. We have access to the actual number of infected cases for which we conducted forecasting using the first vector, which is $x_{2011,A}$. It means the logarithmic score can be calculated as a function of α (because PDFs are functions of α). By varying the values of α, the optimal value that maximizes the score can be determined ($\alpha_{2010\rightarrow2011}$). Using other vectors and doing the same process,

$$\alpha_{2011\rightarrow2012}, \alpha_{2012\rightarrow2013}, \alpha_{2013\rightarrow2014}, \cdots, \alpha_{2021\rightarrow2022}$$

will yield. The general process we are about to undertake has been elucidated in Fig. 7. The values in this sample exhibit two distinct scenarios (Fig. 5). The score function demonstrates an ascending trend in May and June when the number of new cases is zero, followed by a descending trend in July when the number of infected cases is nonzero. When $\alpha_i = 1$, means that the confirmed number of cases for month A of that year is not zero, and we do not have to increase the concentration on the probability of zero, and when $\alpha_i > 1$ which is related to when the confirmed number of cases is zero and by increasing

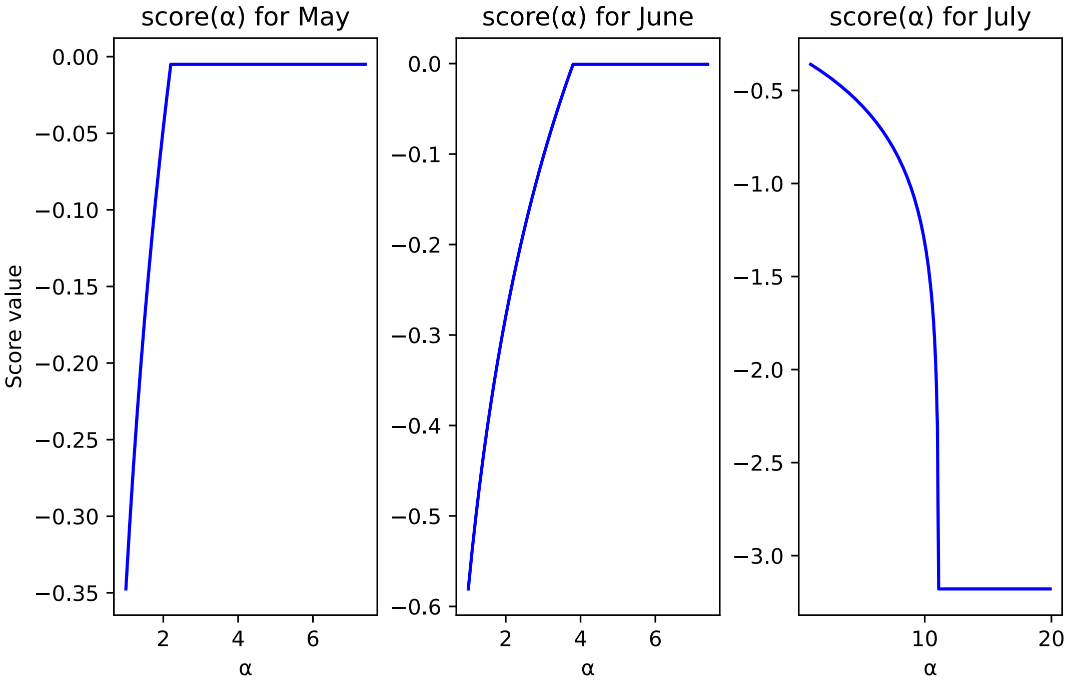


Fig. 5. The score function of Colorado, depicted as a function of α, for May, June, and July.

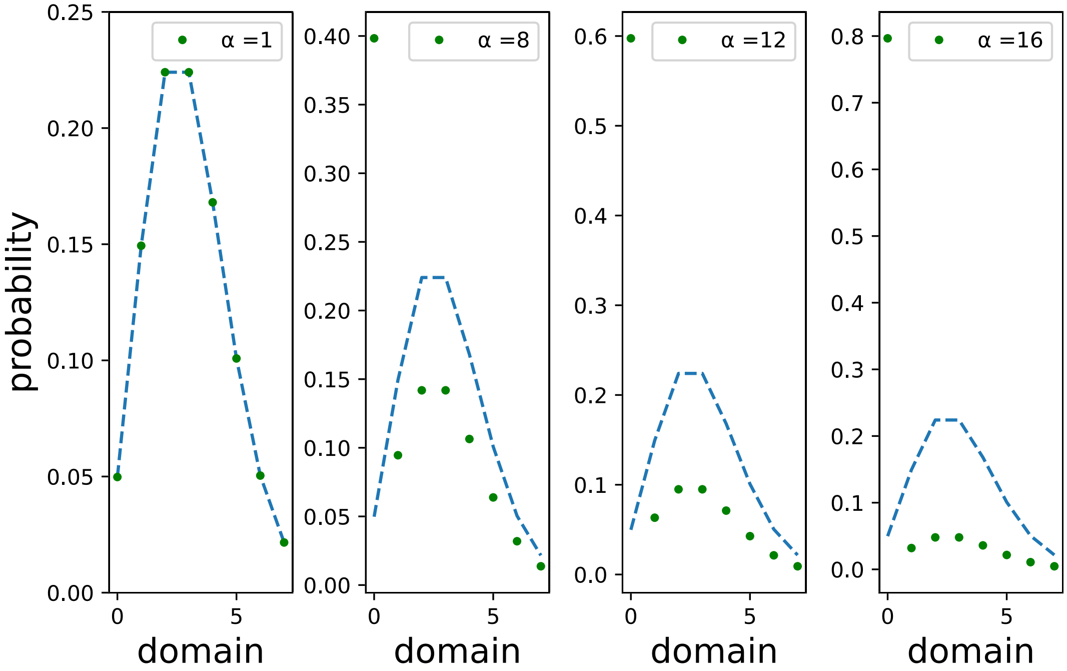


Fig. 6. The probability density function of predictions for various levels of zero inflation.

concentration on the probability of zero the accuracy of the forecast model is increased (Fig. 6). To estimate α for the target forecasting year, we opted for the simplest estimator: the average of these α values.

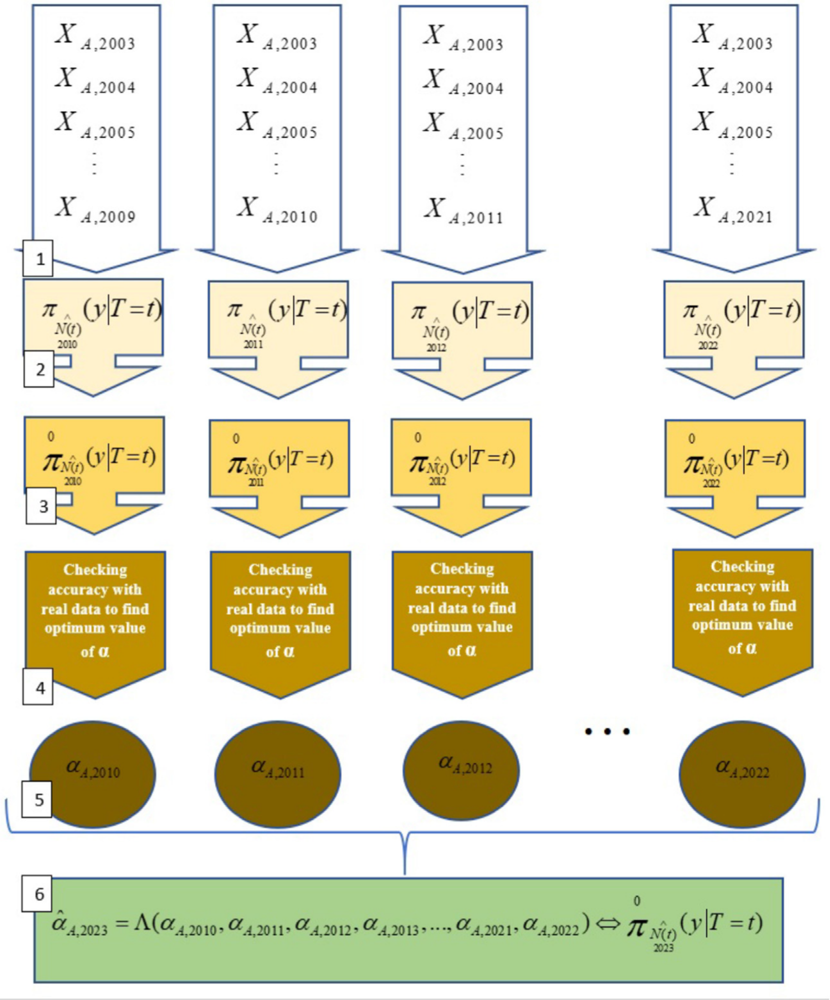

$X_{A,2003}$
$X_{A,2004}$
$X_{A,2005}$
$X_{A,2009}$
$X_{A,2010}$
$X_{A,2011}$
$X_{A,2021}$
1
$\pi_{\hat{N}(t)_{2010}}(y|T=t)$
$\pi_{\hat{N}(t)_{2011}}(y|T=t)$
$\pi_{\hat{N}(t)_{2012}}(y|T=t)$
$\pi_{\hat{N}(t)_{2022}}(y|T=t)$
2
$\overset{0}{\pi}_{\hat{N}(t)_{2010}}(y|T=t)$
$\overset{0}{\pi}_{\hat{N}(t)_{2011}}(y|T=t)$
$\overset{0}{\pi}_{\hat{N}(t)_{2012}}(y|T=t)$
$\overset{0}{\pi}_{\hat{N}(t)_{2022}}(y|T=t)$
3
Checking accuracy with real data to find optimum value of α
4
$\alpha_{A,2010}$
$\alpha_{A,2011}$
$\alpha_{A,2012}$
$\alpha_{A,2022}$
5
6
$\hat{\alpha}_{A,2023} = \Lambda(\alpha_{A,2010}, \alpha_{A,2011}, \alpha_{A,2012}, \alpha_{A,2013}, \ldots, \alpha_{A,2021}, \alpha_{A,2022}) \Leftrightarrow \overset{0}{\pi}_{\hat{N}(t)_{2023}}(y|T=t)$

Fig. 7. The procedure for deriving a sample of $\alpha_{A,t}$ from historical data (steps 1 to 5) and its subsequent application in estimating $\alpha_{targetyear}$, with the target year being 2023 (step 6).

4.2. Zero-deflated posterior

Here, in this subsection, it's important to reiterate that in certain states, such as Arizona, there are occurrences of infected individuals for the majority of the year. Similarly, in various other states, instances of infected individuals are observed during specific months, such as August. This implies that by increasing the density of zeros in the forecasting probability density functions, we inadvertently decrease the value of logarithmic metrics. In simpler terms, this leads to a loss of forecast accuracy. In such scenarios, the solution is straightforward: we need to eliminate or decrease the density of zero values to achieve more favorable results. To do this, we derived the optimal value of α for previous years in month A using historical data. Building upon the set of α_t values, we formulated a hypothesis-testing approach as follows:

$$H_0 : \alpha = 1 \quad \text{versus} \quad H_1 : \alpha > 1.$$

Similarly, this process can be likened to conducting the following test for a binomial distribution:

$$H_0 : p = \frac{1}{2} \quad \text{versus} \quad H_1 : p > \frac{1}{2}.$$

This entails our interest in determining whether the occurrence of these one and non-one values is purely random with equal probabilities (H_0) or if there exists a discernible pattern where the probability of encountering a one is higher than that of a non-one (H_1). The P-value for this test is computed as follows:

$$P_{H_0}(B \geq b) = \sum_{i=b}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n,$$

where the variable B represents the count of ones and follows a binomial distribution with parameters n and 0.5.It has to be munitioned that, n is number of the year we have used to do the hypothesis testing and b is number of the ones (number of the year we non-zero cases for that month). To elucidate this matter, let us consider the following data set, which consists of the values of α_t for 14 years from 2007 to 2022 in month A. These values were used to calculate α_{2023}:

$A = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$, so the value of the p-value is calculated as:

$$P_{H_0}(B \geq 14) = \sum_{i=b}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n = \left(\frac{1}{2}\right)^{14} \approx 0,$$

Hence, the computed p-value of 0 is less than the significance level of 0.05. Consequently, we reject the null hypothesis (H_0). This suggests that we anticipate that the value of α_{2023} to be greater than 1 for the upcoming year, indicating the potential to reduce or diminish the density of zero values. The essence of this subsection can be summarized as follows:

When a consecutive sequence of ones is observed for $\alpha_{A,t}$, it signifies a consistent presence of confirmed cases based on historical data for that particular month in a specific state. In such cases, increasing the density of zero in the posterior PDF or the simple PDF is not logical. Instead, the focus should be on reducing the density of zero value and redistributing it across other values within the probability density function's domain. This situation introduces the concept of a "zero-diminished posterior," which is defined as follows.

$$\pi^{+}{}_{\mathrm{N}(\hat{\mathrm{t}})}(y|T=t) = \begin{cases} 0 & \text{if } y = 0 \\ \dfrac{\pi_{\hat{N}(t)(y|T=t)}}{1 - \pi_{\hat{N}(t)(0|T=t)}} & \text{if } y > 0. \end{cases}$$

5. West nile virus incidence forecast

While the ICC-curve method proved to be highly effective for SEIR and SEIR vector-borne diseases, as demonstrated by the information presented thus far, our approach encounters limitations attributed to data constraints or the small sample size. In practical applications, diseases such as West Nile often exhibit a limited number of infected cases, with some months at the beginning and end of the year registering zero cases due to factors like low temperature and other parameters influencing the mosquito population dynamics and disease transmission. Consequently, at the onset of the year, data are absent (although technically present, all values are zero), making it challenging to predict the number of new infections for the upcoming months because sufficient data are essential to accurately fit a suitable parabola and subsequently estimate the total number of cases following a specific period or outbreak. To address the initial step, such as the beginning of the year when there are either no infections or limited cases, we fitted a logistic function to the number of the new infected cases from the year 2003 to the year 2022 for a certain month (month A). Applying this logistic function allowed us to predict the cumulative number of infected cases (and consequently number of the new infected cases) in the year 2023, specifically for month A.

More accurately, we assumed that the data for past years during the same month follows a logistic distribution. Relying on this assumption, we fit a parabola and logistic function to the data set, allowing predictions based on historical data. As time advances and new cases of infection are recorded, the simulated data can be replaced with real data, enabling adjustments to the trajectory of the epidemic and the predictive curve. It means that at the beginning of the epidemic, we express our prediction regarding the future of the epidemic by drawing upon our past reports and analyzing the historical data. As time progresses, we continuously update this belief with the most current data available. To generate the virtual data (initial prediction for the year 2023), we employed monthly historical data encompassing the number of newly reported cases over 20 years (from 2003 to 2022), which was sourced from the CDC. Fig. 8 will provide further clarification. Panel A represents the initial step of synthesizing data, referred to as virtual data, using historical data when no current data are available for the current year. In panel B, corresponding to the moment when the first new cases are reported, adjustments are made to both the parabola and logistic functions to accurately fit the data. Similarly, in each subsequent step (C, D, E, and F), as new cases are reported, the parabolas and logistic functions continue to be adjusted and fitted, ensuring increasingly realistic and accurate predictions.

Having discussed all the generalities and theories, we next assessed the overall goodness of our model by estimating the number of infected cases for all states across the US. To achieve this goal, we compared our model outcomes to forecasts made by past CDC WNV Forecasting Challenge contributors. We opted for the 2020 CDC challenge results because it represented the most recent year for which the CDC challenge for forecasting has been officially announced and published up until the present

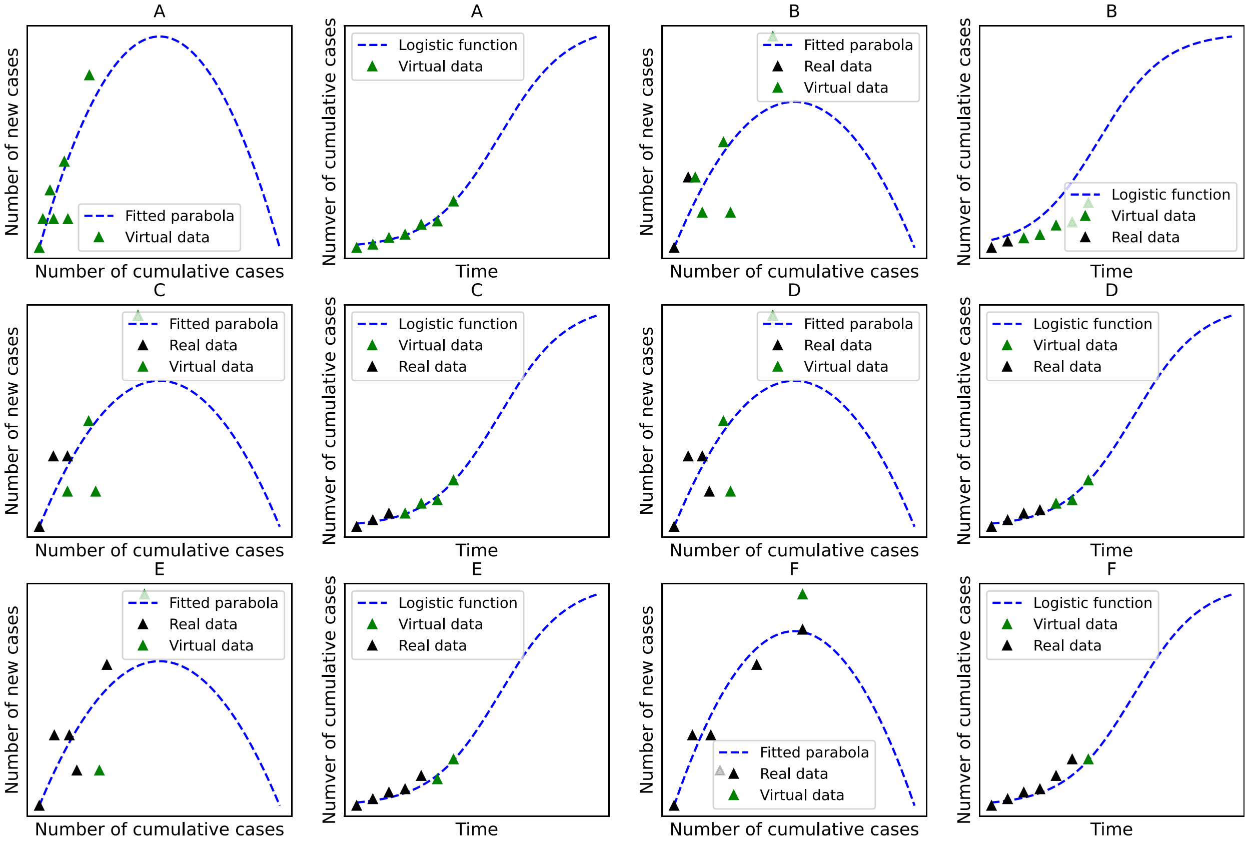


Fig. 8. The process of fitting the parabola and logistic function to the data and predicting in different stages.

time. We utilized monthly data spanning from 2003 to 2019. With this historical data, we performed forecasts based on the Bayesian posterior and the simple probability density function for all states. The logarithmic scoring system (Rosenfeld et al., 2012) was used to evaluate the precision of our approach and then compared to other models that contributed to the 2020 forecasting challenge (Holcomb et al., 2023). In this section, our exclusive emphasis is on the initial stage, where we relied on predictions derived from historical data to demonstrate the potential accuracy of this approach. For the initial step, we employed three distinct approaches to make predictions and subsequently establish a probability density function derived from those predictions.

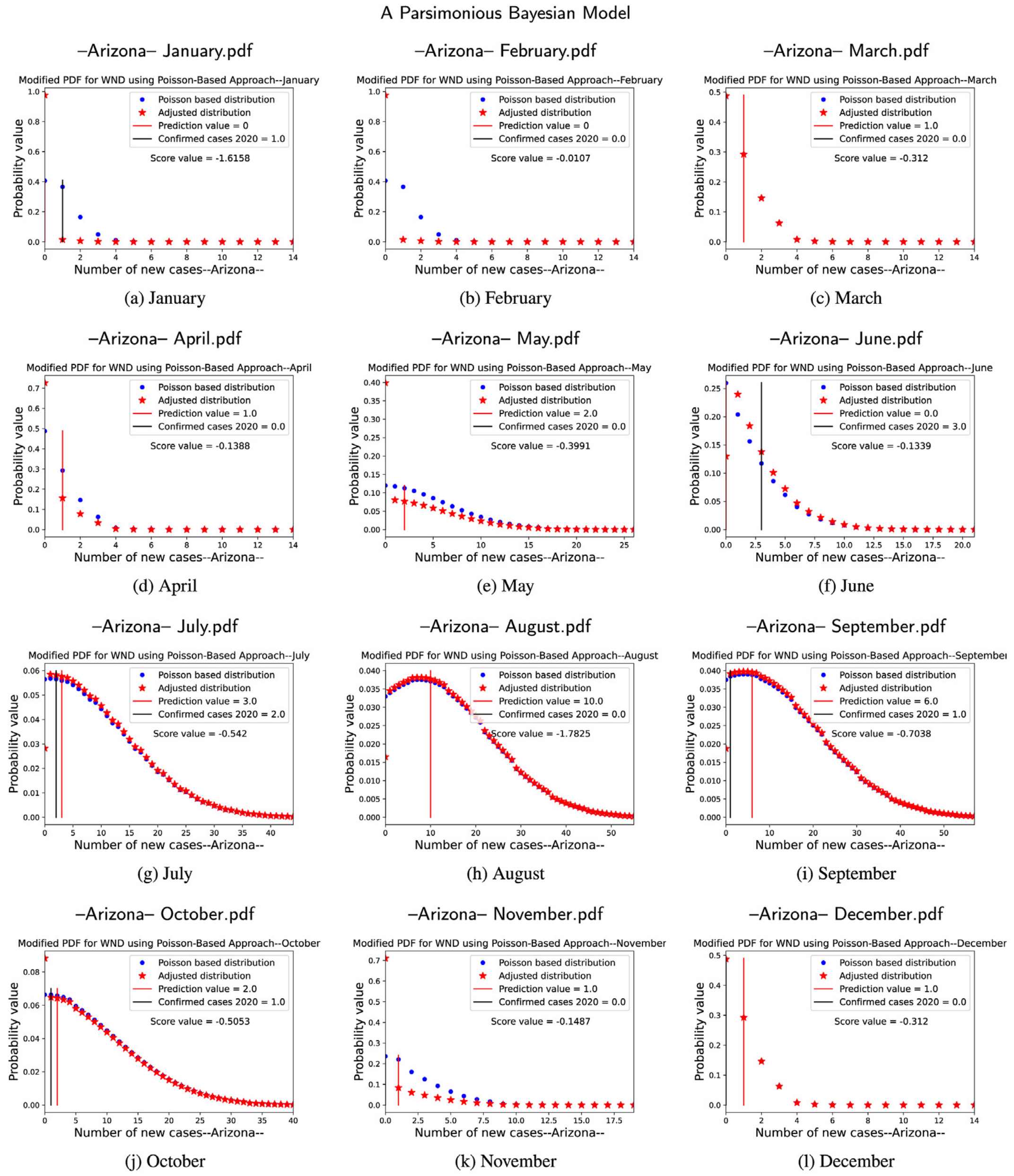
The first kind of prediction relied on historical data, encompassing the entirety of data available from the onset of the epidemic in the US in 2003 up until the present time (in this instance, 2019). This approach captured the epidemic's overall average behavior.

The second prediction method utilized the tail-end of the historical data, which we refer to as the “tail of the epidemic.” This prediction captures the recent behavior of the epidemic, influenced by the impact of, for example, climate change, which has been on the rise in recent years. For this type of prediction, we only utilize the tail end of the historical data (the last two or three years) and determine the logistic function based on the concepts explained in the previous sections.

The third approach involved calculating the average number of predictions derived from both historical data and the tail of the epidemic. This method provided us with a more conservative prediction, capturing the behavior that lies between the patterns observed in the first and second approaches. In the cases where there is a significant leap between values of two consecutive years for a given month A, historical data has been employed for prediction and forecasting, while in instances of no notable leap, predictions are made based on the tail-end of the epidemic. We saved all the probability density functions that were derived from the modified Poisson density and utilized them for forecasting. The prediction and forecasting process was carried out not only for 2020 but also for 2021 and 2022. The saved results can be found in the appendix-1 file.

To exemplify the general forecasting approach, Figures (9) and (10) display the results for the state of Arizona in 2020. We note that for each month, the value of the score has been calculated for the alpha-adjusted density. By observing the plots for the 12 months, it becomes evident that the predictions predominantly fall within the region exhibiting the highest density.

After the passage of time and the documentation of initial cases for the upcoming year (the year targeted for prediction), we can leverage the newly acquired data to establish both a parabolic and logistic function for predictive modeling. We have attempted to clarify this process by utilizing the new case numbers for the state of Colorado in 2023, as illustrated in



(a) January (b) February (c) March

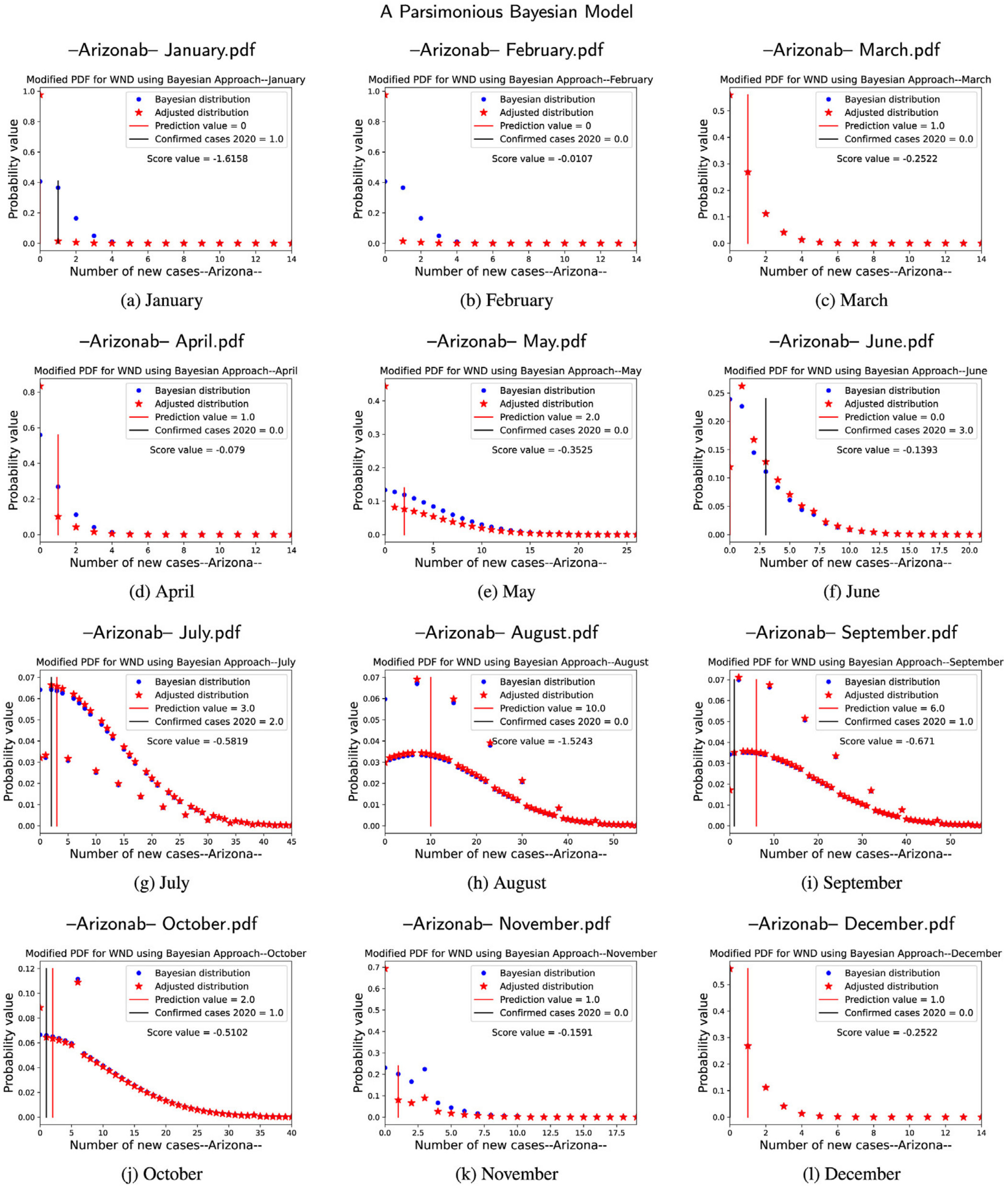
(d) April (e) May (f) June

(g) July (h) August (i) September

(j) October (k) November (l) December

Fig. 9. Forecast for the state of Arizona in 2020 for different months using the simple Poisson forecasting probability density function and its inflated-deflated version.

Figure (11). It has to be mentioned that, practically, up to week number 26, there is no data. This means that the Colorado Department of Public Health and Environment has not recorded any number of infected cases for this period (It's important to acknowledge that this data can be changed or updated over time by the department). After week 25, the first cases were recorded, indicating the onset of the epidemic. The strategy to forecast the number of new cases involves utilizing data from the initial weeks of the epidemic, establishing the parabola, and logistic function as previously explained. Subsequently, the predictive logistic curve is compared with the actual curve of real cases for validation. Figure (11) elucidates the prediction process in a few steps; panel (a) displays the fitted parabola and logistic function to the data from week 26 to week 32. It is evident that the limited amount of data during this period hinders the accurate establishment of a logistic function (depicted by the dashed blue curve) to predict the trajectory of the epidemic, as illustrated by the black curve. In such instances, the fitted predictive logistic curve is employed primarily for short-term predictions, such as forecasting for the next week or the subsequent two weeks. In panel (b), utilizing data from week 26 to week 33, we encounter a similar challenge with the logistic function. Once more, due to the limited data, the prediction is confined to a short period, typically one or two weeks. However,



(a) January (b) February (c) March

(d) April (e) May (f) June

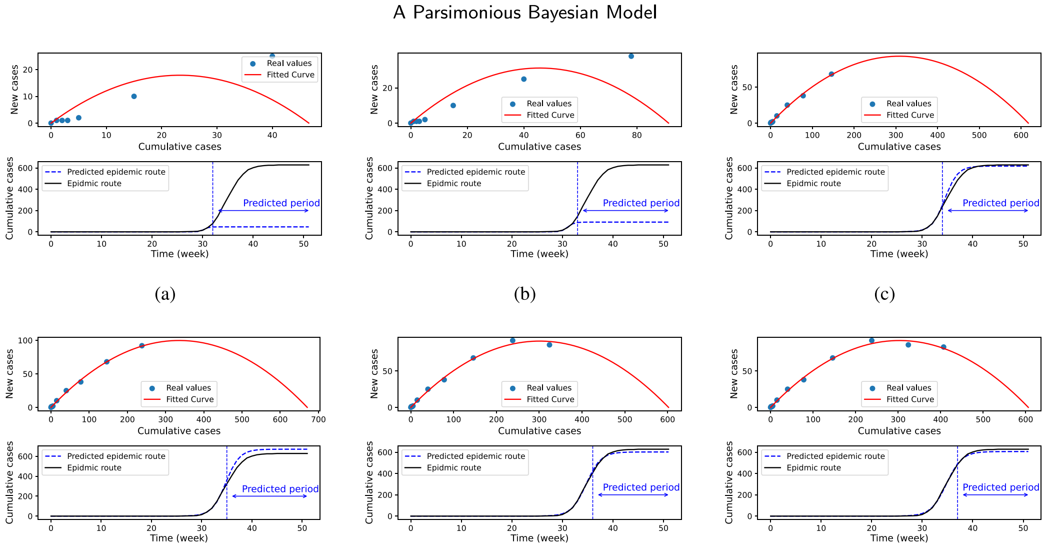
(g) July (h) August (i) September

(j) October (k) November (l) December

Fig. 10. Forecast for the state of Arizona in 2020 for different months using the Bayesian posterior probability density function based on censored Poisson and its modified version.

the accuracy of the prediction improves significantly when we include one more week in the raw data (from week 26 to week 34) to fit the parabola and logistic function. In panel (c), we observe that the predictive logistic curve can closely approximate the trajectory of the epidemic quite effectively. Panels (d), (e), and (f) demonstrate that with the inclusion of more data (practically up to week 37), the accuracy of the prediction is significantly improved, reaching a very satisfactory level. To understand the effectiveness of the methodology, it's crucial to focus on the predictive period marked by a blue double-point arrow. The vertical line indicates the utilization of data up to that specific moment. Based on this period, predictions are generated and presented as the predicted period. This underscores a crucial point: even with a relatively small dataset (essentially consisting of eight numbers) used to train the model (fitting parabola and logistic function), it demonstrates proficiency in delivering accurate predictions for an extended period of the epidemic.

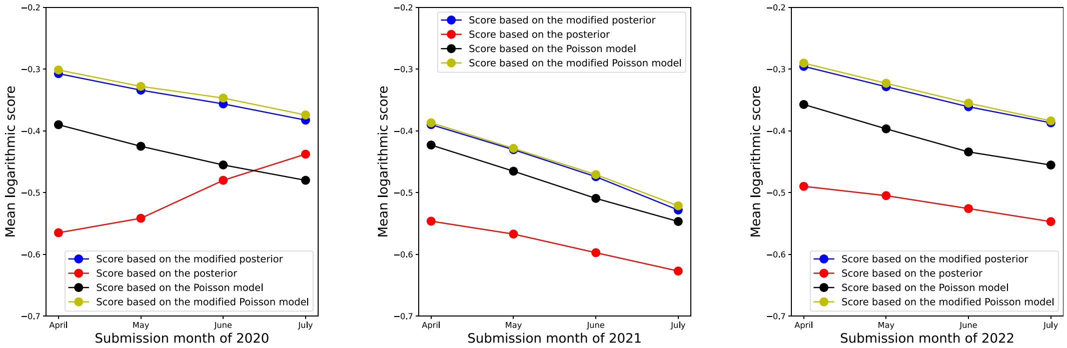
Continuing the evaluation of our model's performance with other approaches, Fig. 12 illustrates the outcomes for four different submissions that have been started from April, May, June, and July to the end of the year. A simple comparison with the CDC results of the 2020 challenge (Holcomb et al. (2023), Fig. 2) reveals its good performance for both models, particularly



(a) (b) (c)

(d) (e) (f)

Fig. 11. Prediction for the cumulative number of cases in the state of Colorado throughout 2023, utilizing data from the same year.



(a) (b) (c)

Fig. 12. (a) Results for the year 2020. (b) Results for the year 2021. (c) Results for the year 2022.

considering that this is an initial step that has not been updated by the 2020 data and the model does not incorporate any temperature parameters. For example, in the April submission, the results of both models were better than the results of the ten participant teams in the challenge. In panel (b) of Figure (12), we made forecasts for the year 2021 using both the Poisson and Bayesian models (modified and unmodified) and calculated the corresponding scores based on the aforementioned four submissions. While it may not be directly comparable to panel (a) and (c), as it pertains to the results of different years, panels (a) and (c) provide insight into the model's performance across various years and offers an understanding of its effectiveness in different scenarios. All three panels demonstrate the effectiveness of the modification approach outlined in Section 4, as they reveal a noticeable improvement in accuracy. Notably, the modified versions of the models (inflated-deflated PDF) exhibit the highest performance in terms of probabilistic measurement, and this serves as practical evidence affirming the efficacy of the technique.

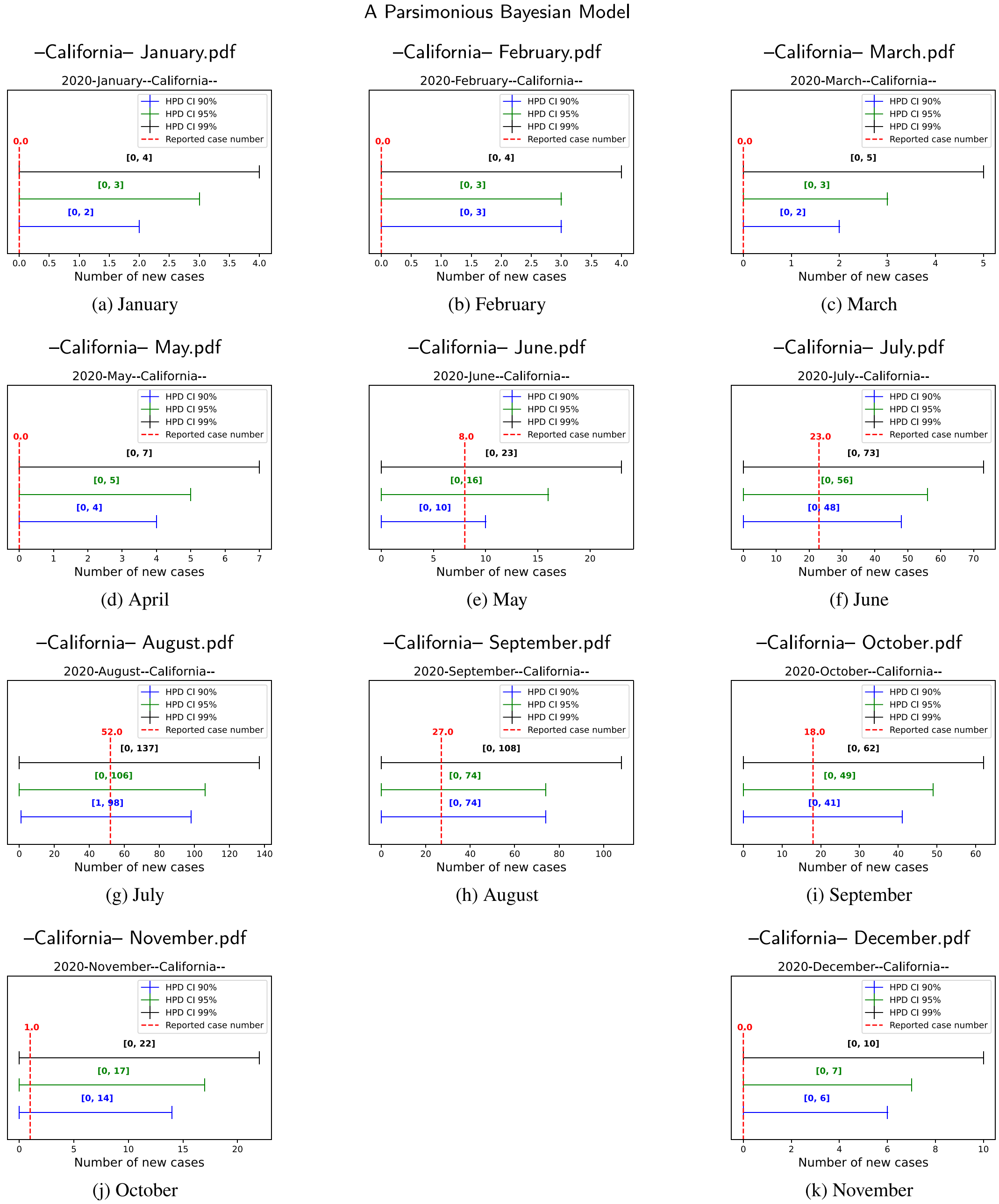
Table 1
Comparing the inverse Expectation and Expectation of the inverse for a censored Poisson variable.

$L_0(\underline{X})$	5	10	15	20	30	40	100	500	2000
$E[\frac{1}{L}]$	0.26	0.11	0.71	0.52	0.33	0.025	0.0101	0.002004	0.0005002
$\frac{1}{E[L]}$	0.2	0.1	0.67	0.5	0.35	0.026	0.01	0.002	0.0005

Table 2
Accuracy values for all states of the U.S. for 2020, 2021, and 2022.

Year/Err.	$= 0$	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5	≤ 6	≤ 7	≤ 8	≤ 9	≥ 10
2020	59.34	83.85	90.1	94.79	96.53	97.92	98.26	98.44	99.13	99.48	2.56
2021	62.28	85.42	89.85	93.23	94.97	96.18	96.86	97.40	98.1	98.61	3.40
2022	61.98	84.72	90.1	93.75	95.14	96.7	97.4	98.09	98.44	98.96	3.06

Table 2 displays the accuracy of the results (percentage of error, defined by differences between the predicted value and real reported cases) for the three-year predictions of 2020, 2021, and 2022 for all states of the US. As evident from these tables, the values of predictions, and consequently forecasting probability density function are reasonably accurate. In this context, the results of (Table 2) indicate that 59.34%, 61.28%, and 61.98% of the predictions aligned with the confirmed reported cases



(a) January (b) February (c) March

(d) April (e) May (f) June

(g) July (h) August (i) September

(j) October (k) November

Fig. 13. Highest Posterior Density (HPD) intervals extracted from the Poisson posterior distribution for the state of California in the year 2022.

for the years 2020, 2021, and 2022, respectively. Moreover, in the case of 2020, 2021, and 2022, 83.85%, 85.42%, and 84.72% of the predictions, respectively, were only one person less than the reported cases. Furthermore, for the subsequent years of 2020, 2021 and 2022, 90.1%, 89.85%, and 90.1% of the predictions deviated by two persons from the confirmed reported cases.

For error values ranging from 3 to 9, the percentage of accurate predictions is consistently above 95%–99%. In conclusion, by adopting a meticulous approach to prediction and forecasting and considering an outlier prediction as the difference between real reported cases and predicted values exceeding 10 cases (even though they may not be true outliers), the percentage of outliers for the years 2020, 2021, and 2022 are 2. 56%, 3. 40%, and 3. 06%, respectively (the last column of the table).

Furthermore, in addition to the forecasting, using historical data, the Highest Posterior Density (HPD) intervals based on Poisson posterior were calculated for all states of the US for the years 2020, 2021, and 2022 at three distinct credible levels: 90 %, 95%, and 99% (see Appendix II). It is worth noting that the percentage of interval estimations that do not encompass the actual reported case numbers are represented by 0.7% (2 out of 260), 4.1% (11 out of 266), and 1.85% (5 out 269) for the three respective years 2020, 2021, and 2022. Fig. 13 illustrates the outcomes pertaining to the state of California in the year 2022. The visualization showcases all three intervals alongside a box plot for the Poisson posterior. Upon observation, it becomes apparent that, for December, all values across various years from 2000 to 2021 are consistently zero. As a result, establishing an interval estimation for this specific month is not practically viable.

6. Conclusions

Our work demonstrated that an accurate and parsimonious model can be constructed using only two parameters and relying solely on a sample data set. We showed theoretically, that this prediction method is not only parsimonious but also exhibits favorable properties within the Bayesian framework, taking the minimum value of the risk function. Finally, using this prediction method, we established forecasting probability density functions (both Bayesian and non-Bayesian). After creating the probability density functions (PDFs) and drawing insights from statistics published by healthcare institutions like the CDC, we refined these PDFs. Additionally, we proposed a method to address the challenge of lacking data at the beginning of the year when forecasting for all months of the year. Based on this suggested method, we obtained Bayesian credible intervals and forecasting probability density functions for all states of the US. Finally, using logarithmic-probabilistic metrics, we offered empirical evidence of the accurate results obtained with the proposed method. Although we have initially configured our forecasting model based on the Poisson probability density function, future research may involve exploring the optimal prior distribution to maximize accuracy.

CRediT authorship contribution statement

Saman Hosseini: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lee W. Cohnstaedt:** Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. **John M. Humphreys:** Writing – review & editing, Methodology, Investigation. **Caterina Scoglio:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

I/We declare we have no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.idm.2024.06.004.

References

Aberth, J. (2011). *Plagues in world history*. Rowman & Littlefield Publishers.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Bergsman, L. D., Hyman, J. M., & Manore, C. A. (2016). A mathematical model for the spread of west nile virus in migratory and resident birds. *Mathematical Biosciences and Engineering, 13*, 401–424.

Biggerstaff, M., Slayton, R. B., Johansson, M. A., & Butler, J. C. (2021). Improving pandemic response: Employing mathematical modeling to confront coronavirus disease 2019. *Clinical Infectious Diseases, 74*, 913–917.

Campbell, G. L., Marfin, A. A., Lanciotti, R. S., & Gubler, D. J. (2002). West nile virus. *The Lancet Infectious Diseases, 2*, 519–529.

Ciota, A. T. (2017). West nile virus and its vectors. *Current opinion in insect science, 22*, 28–36.

Davis, J. K., Vincent, G. P., Hildreth, M. B., Kightlinger, L., Carlson, C., & Wimberly, M. C. (2018). Improving the prediction of arbovirus outbreaks: A comparison of climate-driven models for West Nile virus in an endemic region of the United States. *Acta Tropica, 185*, 242–250.

Di Pol, G., Crotta, M., & Taylor, R. A. (2022). *Modelling the temperature suitability for the risk of west nile virus establishment in european culex pipiens populations, 69* (pp. e1787–e1799). Transboundary and Emerging Diseases.

Frank, S. A. (2009). The common patterns of nature. *Journal of Evolutionary Biology, 22*, 1563–1585.

Holcomb, K. M., Mathis, S., Staples, J. E., Fischer, M., Barker, C. M., Beard, C. B., et al. (2023). Evaluation of an open forecasting challenge to assess skill of west nile virus neuroinvasive disease prediction. *Parasites & Vectors, 16*, 1–13.

Humphreys, J. M., Young, K. I., Cohnstaedt, L. W., Hanley, K. A., & Peters, D. P. C. (2021). Vector surveillance, host species richness, and demographic factors as west nile disease risk indicators. *Viruses*, 13.

Kenneth, F. K. (1993). *The cambridge world history of human disease, 197* (pp. 9–23). Cambridge: Cambridge University Press.

Kovach, T. J., & Kilpatrick, A. M. (2018). Increased human incidence of west nile virus disease near rice fields in California but not in southern United States. *The American Journal of Tropical Medicine and Hygiene, 99*, 222.

Kramer, L. D., Ciota, A. T., & Kilpatrick, A. M. (2019). Introduction, spread, and establishment of west nile virus in the americas. *Journal of Medical Entomology, 56*, 1448–1455.

Lega, J. (2021). Parameter estimation from icc curves. *Journal of Biological Dynamics, 15*, 195–212.

Moon, S. A., Cohnstaedt, L. W., McVey, D. S., & Scoglio, C. M. (2019). A spatio-temporal individual-based network framework for west nile virus in the USA: Spreading pattern of west nile virus. *PLoS Computational Biology, 15*, Article e1006875.

Myer, M. H., & Johnston, J. M. (2019). Spatiotemporal bayesian modeling of west nile virus: Identifying risk of infection in mosquitoes with local-scale predictors. *Science of the Total Environment, 650*, 2818–2829.

Nash, D., Mostashari, F., Fine, A., Miller, J., O'leary, D., Murray, K., et al. (2001). The outbreak of west nile virus infection in the new york city area in 1999. *New England Journal of Medicine, 344*, 1807–1814.

Peper, S. T., Dawson, D. E., Dacko, N., Athanasiou, K., Hunter, J., Loko, F., et al. (2018). Predictive modeling for west nile virus and mosquito surveillance in lubbock, Texas. *Journal of the American Mosquito Control Association, 34*, 18–24.

Piret, J., & Boivin, G. (2021). Pandemics throughout history. *Frontiers in Microbiology, 11*, Article 631736.

Poh, K. C., Chaves, L. F., Reyna-Nava, M., Roberts, C. M., Fredregill, C., Bueno, R., Jr., et al. (2019). The influence of weather and weather variability on mosquito abundance and infection with west nile virus in harris county, Texas, USA. *Science of the Total Environment, 675*, 260–272.

Reich, N. G., Lessler, J., Funk, S., Viboud, C., Vespignani, A., Tibshirani, R. J., et al. (2022). Collaborative hubs: Making the most of predictive epidemic modeling. *American Journal of Public Health, 112*, 839–842. PMID: 35420897.

Ronca, S. E., Ruff, J. C., & Murray, K. O. (2021). A 20-year historical review of west nile virus since its initial emergence in north America: Has west nile virus become a neglected tropical disease? *PLoS Neglected Tropical Diseases, 15*, Article e0009190.

Rosenfeld, R., Grefenstette, J., & Burke, D. (2012). *A proposal for standardized evaluation of epidemiological models*.

Snowden, F. M. (2019). *Epidemics and society: From the black death to the present*. Yale University Press.

World Health Organization. (2020). Vector-borne diseases. URL: https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases.