Modelling the 2010-2011 Haiti Cholera Epidemic

A Study in Simulation-Based Inference for Infectious Disease Transmission

Anna Rosengart

Thesis Advisor
Professor Edward Ionides

Department of Statistics, University of Michigan

Abstract

At the onset of an epidemic, it is common for disease intervention methods to be evaluated for efficacy via mathematical modelling prior to deployment. Model specification and construction must be informed by both the case data of the disease under study and the scientific principles underlying its transmission. For any given epidemic, there are countless possible models that can be formulated and used for motivating public health action, thus underscoring the importance of model criticism and comparison. Although there may be many models that vary in implementation, complexity, and mechanistic and stochastic elements, it is imperative that models with the best available forecasting accuracy be used for informing policy in real-life health crises. To exemplify this point we use the 2010-2011 cholera epidemic in Haiti as a case study. Through the analysis of three different stochastic models, we show the wide variability in model quality and forecasting that can result from minor changes in model specification and calibration.

All materials and code can be downloaded at `https://github.com/aerosengart/haiti_thesis`.

1 Introduction

In late October of 2010, the first cases of cholera in over one hundred years were observed in Haiti. Within the next several weeks, case reports skyrocketed into the thousands and reached over 230,000 by mid-February of the next year [24, 3]. Exacerbated by the poor water infrastructure and the fact that the highly virulent strain introduced to the nation, O1 biotype El Tor serotype Ogawa, had associated antibiotic resistance, the threat the disease posed to public health was large [24]. Presently, there have been well over 800,000 cumulative cases and almost 10,000 deaths in Haiti since the onset of the epidemic [1]. Unfortunately, the opportunity for using vaccination to prevent cholera from becoming a chronic problem has passed, but cholera's decade-long presence in Haiti has provided much more data for the study of the progression of cholera in a population. Therefore, continued efforts in modelling this epidemiological event may prove instructive in responses to future outbreaks.

There was a multitude of opinions on the best and most effective course of action to take to mitigate cholera's spread at the start of the epidemic and throughout the disease's continued presence in Haiti. Vaccination was considered a promising response as two killed oral cholera vaccines (OCVs) were available at the time. However, the duration and level of protection acquired by these vaccines is dependent upon dosage, national coverage, and the age of the vaccine recipient as young children shed immunity faster than adults [3]. Questions continue to arise when considering the logistics of vaccine distribution and administration: When is the best time to begin vaccination? Should vaccine delivery be prioritized according to location-based risk? How many vaccines must be administered for the optimal amount of protection, and are there enough vaccines currently available for this? With all of these questions in mind, it was, and still is, challenging to determine whether investment in widespread vaccination would be an effective measure against cholera.

One means by which to overcome these uncertainties is epidemiological modelling, a powerful method of analysis of disease dynamics that can be used to inform vaccination policy and public health decision-making. In December of 2010, the CDC's modelling study predicted that the impact of vaccination would be relatively minor with the number of vaccines available at the time [6]. However, a later study by Chao et al. found that even relatively low vaccination coverage (around 30% of the population) was effective at controlling the spread of the disease if the administration was informed by the relative risk and exposure of a given community [3]. Fung et al. also showed that OCVs used in conjunction with improved sanitation practices and water infrastructure was most effective at reducing the

number of cases [8]. This paper provides an analysis of a state-space model proposed by Elizabeth C. Lee, Andrew S. Azman, and Justin Lessler of the Johns Hopkins Bloomberg School of Public Health. We begin with an analysis of the methods and results of Lee et al. and go on to propose improvements to its implementation. Through this case study, we aim to exhibit the dependence of the utility and quality of simulation-based inference in epidemiological contexts upon model examination and refinement.

2 Background

In many epidemiological settings, statistical modelling can be of immense use in motivating public health decision-making. For example, modelling reported cases of a given infectious disease can help inform the subsequent actions taken to mitigate the disease's spread. Due to the inherent randomness and complexity of population dynamics, there are numerous different methods for modelling disease, all of which have advantages and disadvantages. Yet the contributions of epidemiological models to our understanding of how infectious diseases evolve in a population are of high value, and it is worth tackling the challenge of developing a good and useful model.

It is well established that state-space models are appropriate and effective models when studying environmental and biological processes. At its core, a state-space model has two components: an unobserved state process and a dependent observation process [10]. The ability to use a state-space model to inform policy and public action depends upon the model's quality, which itself is dependent upon the ease of statistical inference with respect to the model's parameters. Fortunately several methods have been developed to facilitate estimation of unknown parameter values, one of which is maximum likelihood via iterated filtering (MIF) proposed by Ionides et al., a variant of which is addressed in section 2.3 [11].

Compartment models are another standard tool for modelling infectious diseases. By dividing a population into compartments, for example as (S)usceptible, (I)nfectious, (R)ecovered in the standard SIR compartment model, the spread of a disease can be described with much more specificity because attention is given to all stages of host infection. However, it is nearly impossible to know how many individuals populate a compartment or are transitioning between compartments at a given time. To overcome this uncertainty, compartment models can be coupled with state-space models to form a comprehensive representation of a disease's progression in which parameters can be more easily estimated via inference methods for state-space models. In the next few sections, we provide a brief overview of the foundational

concepts needed to understand these models.

2.1 Time Series and Markov Processes

Consider a sequence of N time points, $t_{1:N} = \{t_1, t_2, \ldots, t_N\}$, and a sequence of N observations made at each time point, $y_{1:N} = \{y_{t_1}, y_{t_2}, \ldots, y_{t_N}\}$. We call $Y_{1:N}$ a time series model with jointly defined random variables $Y_n, \quad \forall n \in 1:N$, and we can conceive of the data, $y_{1:N}$, as one realization of $Y_{1:N}$ [19].

We then describe a time series model, $X_{1:N}$, where $X_n = X(t_n)$ is a random process at time $n, \quad \forall n \in 1:N$. Should this time series model satisfy the condition that its state at time $n+1$ is conditional only on its state at time n, $X_{1:N}$ is called a Markov process model. Mathematically, this can be represented as the following equation stating that the conditional density of the process X_n given the processes $X_{1:n-1}$ is equivalent to the conditional density given only the process X_{n-1} [7]:

$$f_{X_n|X_{1:n-1}}(x_n|x_{1:n-1}) = f_{X|X_{n-1}}(x_n|x_{n-1}) \tag{1}$$

2.2 Partially Observed Markov Processes

Often the details of the mechanisms underlying the evolution of a natural system are unknown. In epidemiology, the exact number of individuals exposed to disease at a given time is usually unknown. We can work around the issue of missing information using partially observed Markov (POMP) models. We create a POMP model by joining two processes, one that is unobservable (latent) but of interest and one that is observable and dependent upon the first.

Let the random variables $X_{1:N}$ represent the latent state process where X_1 serves to initialize the process model, $f_{X|X_{n-1}}(x_n|x_{n-1})$. With the random variables $Y_{1:N}$ representing the observable measurement process, the measurement model is $f_{Y_n|X_n}(y_n|x_n)$, and the collected data $y_{1:N}$ are observations of this process. We assume that each Y_n depends only upon the latent process at time n, X_n, and is conditionally independent of the other variables representing the measurement and latent processes, Y_m and $X_m, \quad \forall m \in 1:N, \quad m \neq n$ [20]. Together, $X_{1:N}$ and $Y_{1:N}$ form our POMP model.

2.3 Likelihood and Iterated Filtering

In problems of statistical inference, it is common to use likelihood to inform parameter estimation and model selection. Given a model parameterized by vector θ in the m-dimensional parameter space Θ_m, the likelihood function is the joint probability density of the data, $y_{1:N}$, at θ:

$$\mathcal{L}(\theta) = f_{Y_{1:N}}(y_{1:N}; \theta) \tag{2}$$

We then aim to find an estimate of θ, $\hat{\theta}$, which maximizes this function, $\mathcal{L}(\hat{\theta})$, or its natural logarithm, $\ell(\hat{\theta})$ [16].

The utility of an epidemiological model of disease spread is dependent upon its ability to be used for forecasting cases or incidence. This ability is itself dependent upon our confidence in the model's prediction accuracy and our understanding of the ways in which the latent states change with time. Thus, we have two linked problems: identifying the distribution of X_n at time n given $y_{1:n}$ and finding parameter values, $\hat{\theta}$, which maximize the likelihood of our data. These problems are known as the filtering problem and the inference problem, respectively [5, 16].

Especially in the case of highly complex environments, both the likelihood function and the transition density of a POMP model can be difficult to write analytically, making these two problems quite hard. Many methods have been developed to surmount the inference and filtering problems, one of which is the particle filter. For the particle filter, we need only supply data, simulators for the initial density and the one time-step transition density of the latent process, and an evaluator for the density of the observation process conditional on the latent process to get maximum likelihood estimates for the model parameters.

We first initialize a swarm of M particles at time 1, $\{X_1^m; m \in 1:M\}$, each containing the necessary state information along with a vector of parameter values, θ. Then for each time $n \in 1:N$, we push the particles forward one time-step by drawing from the one time-step transition density, giving us an ensemble of particles representing the prediction distribution at time n, $f_{X_n|X_{n-1}}(\cdot|X_{n-1}^m; \theta)$. We weight the particles according to our data by evaluating the measurement density, so $w_{n,m} = f_{Y_n|X_n}(y_n|x_n^m)$. Finally we resample the particles according to these weights, which leads to an ensemble of particles representing the filtering distribution at time n, $f_{X_n|Y_{1:n}}(x_n|y_{1:n}; \theta)$.

Because of the assumed independence of the measurement process variables and their

dependence upon the latent process variables in a POMP model, we have that:

$$\begin{aligned}
\mathcal{L}(\theta) &= f_{Y_{1:N}}(y_{1:N};\theta) \\
&= \prod_{n=1}^{N} f_{Y_n|Y_{1:n-1}}(y_n|y_{n-1};\theta) \\
&= \prod_{n=1}^{N} \int f_{Y_n|Y_{1:n-1},X_n}(y_n|y_{1:n-1},x_n;\theta) f_{X_n|Y_{1:n-1}}(x_n|y_{1:n-1};\theta) dx_n \\
&= \prod_{n=1}^{N} f_{Y_n|X_n}(y_n|x_n)
\end{aligned} \tag{3}$$

Notice that the weights used in the particle resampling are $w_{n,m} = f_{Y_n|X_n}(y_n|x_n^m)$ for each particle m at time n. If we take the average of $f_{Y_n|X_n}(y_n|x_n^m)$ over all M particles, we can approximate $f_{Y_n|X_n}(y_n|x_n;\theta)$. Therefore:

$$\mathcal{L}(\theta) = \prod_{n=1}^{N} f_{Y_n|X_n}(y_n|x_n;\theta) \approx \prod_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} f_{Y_n|X_n}(y_n|x_n^m) \tag{4}$$

In other words, by the Monte Carlo principle we can approximate the conditional likelihood at time n with $w_{n,m}$. Thus the particle filter provides a much easier way to estimate the likelihood of the data given our model and to approximate the distribution of X_n at time n given $y_{1:n}$ [14, 12, 11].

An extension of the particle filter is the improved iterated filtering algorithm (IF2) developed by Ionides et al. [10]. As a plug-and-play method, IF2 is a computationally efficient, simulation-based means for maximum likelihood estimation and inference. IF2 takes an initialized swarm of particles and, using a combination of particle filtering, small changes to the parameter values, and particle resampling, estimates the parameter values which achieve the maximum likelihood [12]. With the particle filter and IF2, we are able to approximate solutions to the inference and filtering problems.

3 Methodology

3.1 Model Structure

Lee et al. used an SEIAR compartmental model (S: Susceptible, E: Exposed, I: Infectious, A: Asymptomatic Infectious, R: Recovered) for the Haiti cholera epidemic. In their formulation, at a given time point t each compartment contains some unobserved number

of individuals from the total population of Haiti. Between two time points t and $t+1$, individuals can transition into the system by birth, out of the system by death, or between compartments at rates that are either specified or estimated. We define these transition rates with the following series of equations:

$$q_{S_k E_k} = \lambda(t) \tag{5}$$

$$q_{E_k I_k} = \sigma(1 - \theta_0(t)) \tag{6}$$

$$q_{E_k A_k} = \sigma\theta_0(t) \tag{7}$$

$$q_{I_k R_k} = q_{A_k R_k} = \lambda \tag{8}$$

$$q_{R_k S_k} = \alpha q_{S_0 S_k} = q_{E_0 E_k} = q_{I_0 I_k} = q_{A_0 A_k} = q_{R_0 R_k} = \eta_k(t) \tag{9}$$

$$q_{\cdot S_0} = \mu q_{S_k \cdot} = q_{E_k \cdot} = q_{I_k \cdot} = q_{A_k \cdot} = q_{R_k \cdot} = \delta \tag{10}$$

where $q_{X_k Y_k}$ indicates the one time-step transition rate from compartment X to compartment Y, and $k \in [0, 10]$ denotes vaccination cohort with $k = 0$ indicating the cohort that did not receive vaccinations. At time t, $\eta_k(t)$ is the vaccination rate of cohort k, and $\lambda(t)$ is the force of infection, calculated as $\lambda(t) = \frac{\beta(I(t)+(1-\kappa)A(t))^\nu}{N(t)}$. The seasonal transmission term is $\beta = \sum_{i=1}^{6} \beta_i s_i$, which consists of six degree six periodic B-spline terms, $s_{1:6}$, multiplied by estimated seasonality parameters, $\beta_{1:6}$. $I(t)$ is the proportion of the population that is infectious at time t, $A(t)$ is the proportion of the population that is asymptomatic at time t, $N(t)$ is the population of Haiti at time t with $N_0 = 10911819$, $\kappa = 0.95$ is the assumed reduction in infectiousness of asymptomatic individuals, ν is an estimated population mixing coefficient, and $\theta_0(t) = 0$ is the proportion of non-vaccinated, exposed individuals who become infected but are asymptomatic. Not dependent on time are $\frac{1}{\alpha} = 8$, the mean duration of natural immunity in years; $\frac{1}{\sigma} = 1.4$, the latent period of cholera in days; $\frac{1}{\gamma} = 2$, the infectious period of cholera in days; $\mu = 0.43$, the birth rate per 1000 individuals per week; and $\delta = 0.14$, the natural death rate per 1000 individuals per week. $q_{\cdot S_0}$ and $q_{X_k \cdot}$ denote the transition rates into and out of the system's compartments via birth and death, respectively. Below is a figure based upon the model diagram from Lee et al. [15]. It illustrates the compartmental model with one vaccination cohort. Transitioning out of the system due to death is omitted for legibility.

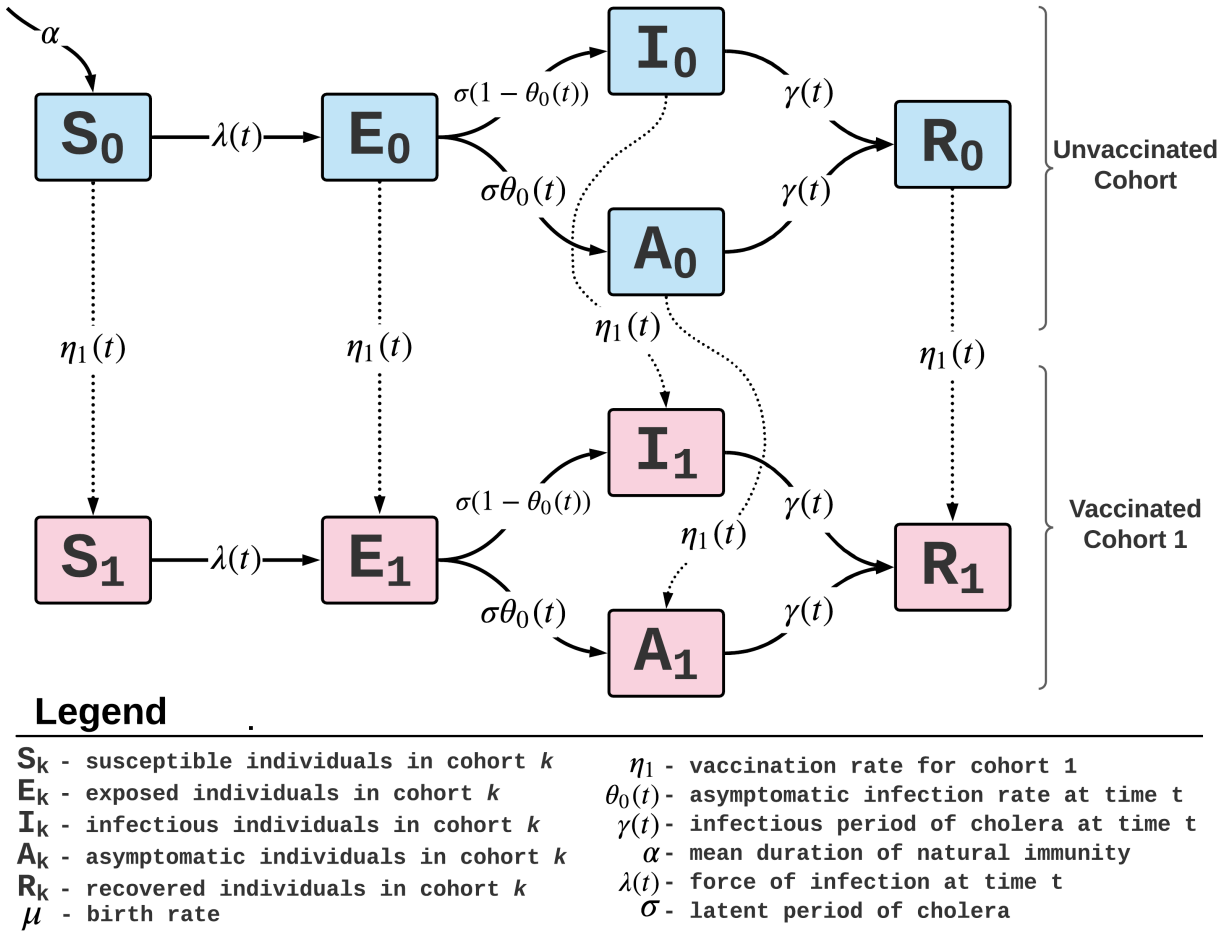

Figure 1: SEIAR compartment model diagram with one vaccination cohort adapted from Lee et al. [15].

3.2 Reproduction

The preference for complex over simple models has been growing for several decades despite the fact that it has been shown that complexity is associated with decreases in forecasting accuracy [9]. Because epidemiological modelling is motivated by the need to accurately forecast disease prevalence to inform policy, we first establish a point of comparison for the evaluation of our model fit and quality. We elect to use a linear, Gaussian autoregressive moving average (ARMA) model of order (2,1) as it is a fairly simple model in which the current state depends only on previous states and white noise [21]. We can then compare the likelihood of the data under this model to the likelihoods achieved under our proposed

models to evaluate whether the additional complexity is truly beneficial.

Lee et al. divided the case data into two periods: epidemic (October 23rd, 2010 through March 31st, 2015) and endemic (April 1st, 2015 through January 12th, 2019) [15]. We adopted this breakpoint in our analyses. The ARMA(2,1) benchmark model achieved log-likelihoods of -1616.678, -1139.238, and -2800.808 for the epidemic, endemic, and the combined time period, respectively.

After establishing benchmark log-likelihoods, we attempted to reproduce the results of Lee et al. as closely as possible in order to facilitate the evaluation of their model and parameter estimates. Lee et al. implemented their model in the R package `pomp` v1.19 and started the model calibration by generating 300 different sets of starting parameter values. They then used trajectory matching followed by iterated filtering to find a maximum likelihood estimate for the parameter values using each of the 300 sets. From the epidemic calibration, they pruned away sets resulting in filtering failures or extreme outlying values. The remaining sets were used as starting values for the endemic calibration in which all parameters were reestimated, excluding the initial state values (E_0 and I_0) [15].

We repeated most of this process with some minor changes. We did not perform trajectory matching as it assumes a deterministic latent process, which is not assumed in the forecasting model. Additionally, Lee et al. did not publish their initial starting sets, so we created our own using the schema provided in their supplemental code. We left weeks with missing data as `NA` rather than `0` as the `pomp` package is capable of working with missing data. We also filtered out epidemic parameter sets with $\nu \leq 0.9$ and $\beta_1 \geq 100$ and endemic parameter sets with log-likelihoods of -3000 units or less to avoid outlying parameter values similar to Lee et al.'s pruning process. Our reproduction (fig. A1) does seem to visually match the results of Lee et al. in figure S7 of their supplement [15].

4 Model and Method Adjustments

Lee et al. did not report parameter point estimates as part of their findings. Rather, they used a cloud of parameter values to repeatedly simulate reported cases and then summarized over these simulations when plotting. This provided a relatively good visual match to the observed reported cases, but a cloud of parameter sets is not entirely helpful for inference and prediction as there is no clear method by which to evaluate the model's performance. For this reason, we propose slight adjustments to the model and methods in order to more rigorously predict cholera elimination and vaccination campaign efficacy in Haiti.

We first suggest alternative algorithmic parameters when estimating model parameter values with the end goal of identifying maximum likelihood estimates. We used the improved iterated filtering algorithm as implemented in the `pomp` function `mif2()`. We elected to use 5000 particles and iterated 100 times in order to reduce variability in the log-likelihood estimates and to avoid potential particle depletion. Moreover, we increased the random walk standard deviations and created a larger set of starting values for the parameters for our global search for the MLE. In the following sections, we describe the structural alterations we made to the model.

4.1 Overdispersion

Creating an equidispersed model can have inappropriate implications on the biological processes assumed to be driving the model. Especially in the case of epidemiological model development for the purpose of forecasting, the importance of incorporating enough stochasticity to explain the collected data has previously been discussed [2, 13]. The model proposed by Lee et al. did take stochasticity into consideration by estimating τ, the inverse dispersion parameter for the negative binomial distribution simulator used in the measurement process [14, 15].

$$\mathcal{C}_t = \text{NegBinom}(\rho\xi, \tau); \qquad \mathcal{C}_t = \text{cases at time } t,\ \rho = \text{force of infection},\ \xi = \text{incidence} \tag{11}$$

However, their latent process remained equidispersed. Calibrations to the epidemic period achieved log-likelihoods of no greater than -1823.403 units, and calibrations to the endemic period only -1143.416. Though their model's log-likelihood was greater than that of the ARMA benchmark model in the endemic period, their model was not competitive in the epidemic period as it achieved a log-likelihood over 200 units below the ARMA model's. In an attempt to achieve higher likelihoods in the epidemic period, we included the addition of another parameter, σ^2, the variance of a gamma white noise process that provided a multiplicative effect upon the force of infection:

$$\omega \sim \text{Gamma}(\sigma^2, dt); \quad \lambda(t) = \frac{(I(t) + (1-\kappa)A(t))^\nu \beta}{N(t)} \times \frac{\omega}{dt} \tag{12}$$

where dt is the time interval, $\lambda(t)$ is the force of infection at time t, κ is the assumed reduction in infectiousness of asymptomatic individuals, ν is a population mixing coefficient, $I(t)$ is the proportion of the population that is infectious at time t, $A(t)$ is the proportion of the

population that is asymptomatic at time t, and $N(t)$ is the total population of Haiti at time t.

4.2 Model Fitting

Using the maximum likelihood estimates of the parameters from this adjusted model's calibration, we plotted the simulated case reports against the case data (fig. 2). Using point estimates has the added advantage of facilitating model evaluation and interpretability. For example, having a single value for the reporting rate, ρ, rather than a range enables us to calculate a single likelihood value for model comparison and provides more specific information about the disease's progression through the population.

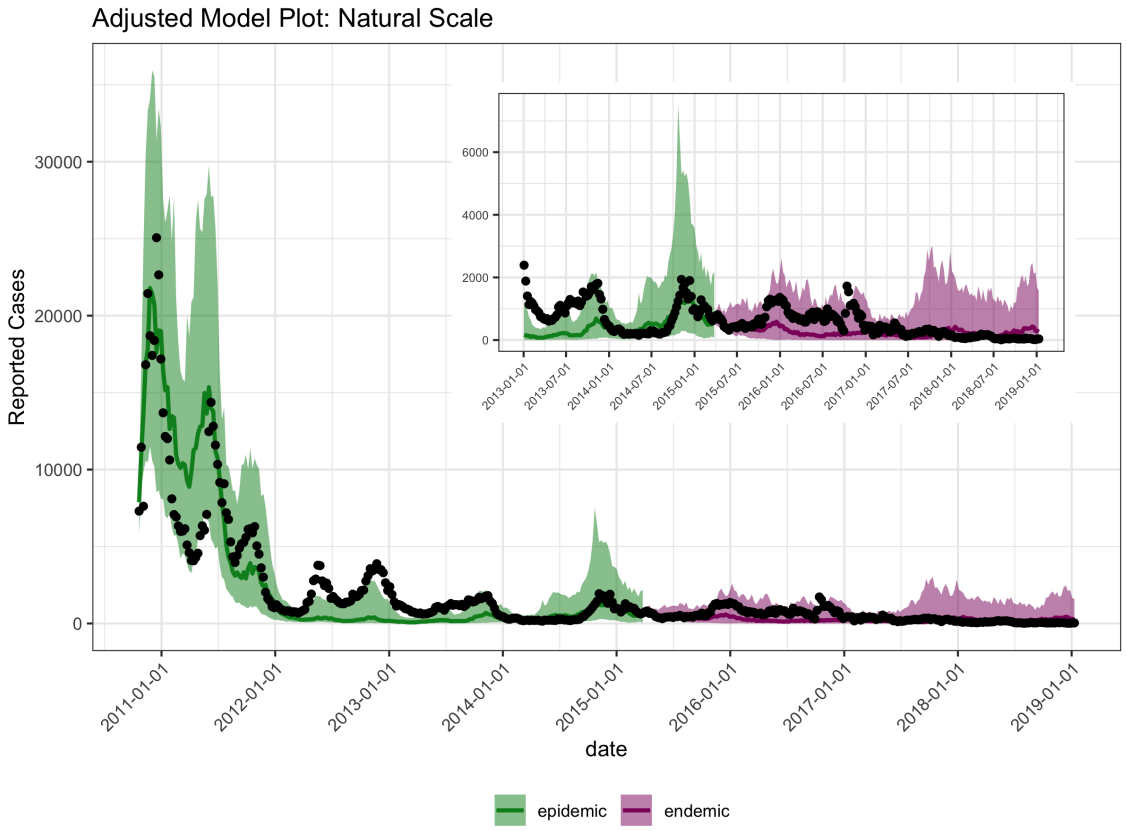

Figure 2: Plot of model simulations with overdispersion in the latent process using the parameter MLEs. Solid line indicates the median simulated reported cases across parameter sets. Ribbon indicates the 2.5th and 97.5th percentiles for the simulated reported cases. Inset provides a closer view of the model fit from 2013 through 2018.

4.3 Profile Likelihood

It is less than ideal to evaluate model fit or perform model selection using solely visual convergence of simulations. Because of this, we reestimated all parameters with the exception of σ^2 to create a profile log-likelihood plot over σ^2 and get a better idea of the MLE for this parameter.

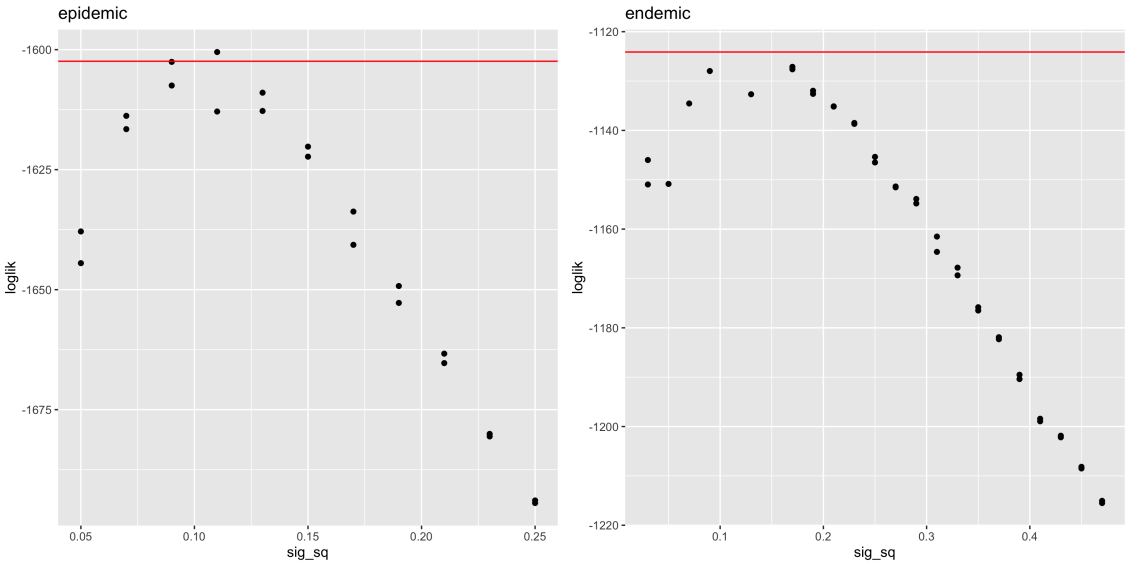

Figure 3: Profile log-likelihood plot over σ^2 for the epidemic and endemic periods. The locations at which the red line intersects the curve connecting the points (not pictured) indicate the 95% confidence interval for σ^2.

From this figure, we see that σ^2 values of around 0.09 to 0.13 in the epidemic period and 0.09 to 0.16 in the endemic period are most consistent with the case data and have associated log-likelihoods in the epidemic and endemic periods of upwards of -1610 and -1130, respectively. This profile log-likelihood plot provides a better understanding of the amount of uncertainty accompanying our estimate of σ^2 [4]. The narrow confidence interval, lack of identifiability issues, and improvement in log-likelihood support the inclusion of this parameter in the model.

4.4 Joint Estimation

Lee et al. proceeded by fitting the model to the epidemic period, simulating to the end of the epidemic period with each set of parameter estimates, and then they used the final states to reestimate all of the parameters for the endemic period. Though this method did

achieve relatively good looking simulations with reasonably high log-likelihoods after adding σ^2, there is justification for linking the two estimation procedures more closely. For one, it is mechanistically logical to use all of the available data to inform the estimation of parameters that are present in the model during both periods. As the designation of the break-point between periods is somewhat arbitrary and the epidemiological system exists without respect for this breakpoint, it also may be better to estimate demographic and seasonal parameters without respect for this break-point.

We fit the model to the epidemic and endemic periods simultaneously, estimating only ρ, τ, and σ^2 separately. This jointly estimated model performed quite well with a log-likelihood of about -2734.761, beating the ARMA benchmark by over 50 log-likelihood units.

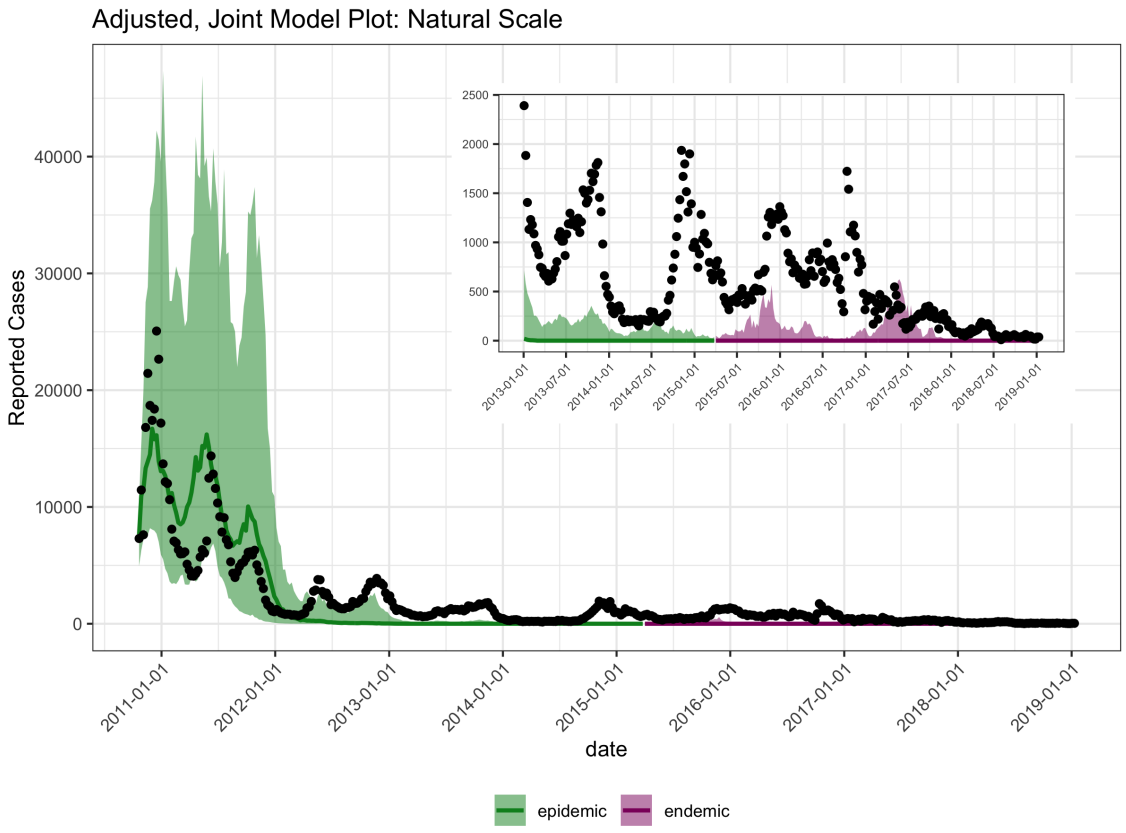

Figure 4: Plot of simulations from the jointly estimated model with additional overdispersion in the latent process using the maximum likelihood estimates of the model parameters. Solid line indicates the median number of reported cases from simulations. Ribbon indicates the 2.5th and 97.5th percentiles for the simulated reported cases. Inset provides a closer view of the model fit from 2013 through 2018.

4.5 Model Comparison

Though these models are all closely related, they are not nested and so we cannot select a model based upon a likelihood ratio test. However, a comparison of likelihoods is still informative so long as we keep in mind the nature of the models' relationship. The Akaike Information Criterion (AIC) is a likelihood-based measure that can be used for model selection that applies a penalty to a model's score according to its complexity. Below we supply an AIC table (table 1) for comparison of the four models discussed above: the ARMA(2, 1) benchmark model, the original model proposed by Lee et al., the altered model with latent process noise, and the jointly estimated altered model.

We estimated the log-likelihoods for the two altered models using repeated particle filters and the parameter maximum likelihood estimates. For Lee et al.'s original model, we report the best possible likelihood achieved across all calibration parameter sets. The full log-likelihoods were calculated as the sum of the likelihoods from the epidemic and endemic periods for the models without joint estimation. The joint model log-likelihoods for the epidemic and endemic periods were calculated as the sum of the conditional log-likelihoods at each time point in the corresponding time periods. We took the average of 20 replications of the particle filter.

Table 1: Table of log-likelihoods by period, number of estimated parameters, and AIC are reported for the ARMA(2,1) benchmark model, original Lee et al. model, the altered model, and the jointly estimated altered model.

Measure	ARMA Model	Lee et al. Model	Alt. Model	Joint Model
Epi. Log-Lik.	-1616.678	-1823.403	-1600.133	1612.962
End. Log-Lik.	-1139.238	-1143.416	-1121.02	-1127.888
Full Log-Lik.	-2800.808	-2966.819	-2721.153	-2735.623
Num. of Params.	5	11+9=20	12+10=22	15
AIC	5611.616	5973.638	5486.306	5499.522

According to AIC, the altered model with the lowest score is the ideal model; however, bearing in mind the limitations of likelihood-based model selection discussed above, we opt to use the jointly estimated altered model in the following forecasting section. This decision was made based upon the model's greater log-likelihood compared to the original model as

well as its lower complexity compared to the altered model without joint fitting.

5 Forecasting

The impetus behind many epidemiological modelling studies is the need to predict future outbreaks of a disease in order to avert widespread infection and death. For this reason, our final section concerns model forecasting with respect to the variety of vaccination campaigns studied by Lee et al. We used the jointly estimated altered model for our forecasting and did so for the following 6 scenarios:

- No Vaccinations
- Two Department: vaccination carried out in the Artibonite and Centre departments over two years
- Three Department: vaccination carried out in the Artibonite, Centre, and Ouest departments over 2 years
- Slow National: vaccination carried out in all 10 departments over 5 years
- Fast National: vaccination carried out in all 10 departments over 2 years
- Fast National, High-Coverage: vaccination carried out in all 10 departments over 2 years

The first four vaccination campaigns assumed target population coverage of 70% with two doses, 10% with one dose, and 20% with zero doses, while the Fast National, High-Coverage scenario assumed 95%, 1.67%, and 3.33% for two-dose, one-dose, and zero-dose coverage, respectively. All vaccination efficacy and roll-out specifications were kept the same as in Lee et al. [15].

A simulation was said to achieve cholera elimination if its true incidence fell below one case for at least fifty-two consecutive weeks after the beginning of the vaccination campaign and remained below one case for the rest of the ten-year forecasting period. A simulation was said to achieve cholera elimination at x years if the fifty-two consecutive week period began before the end of year x. Below we provide a table of the predicted probabilities of elimination of cholera after five years in all six scenarios for our jointly estimated altered model along with the probabilities found using the original model as reported by Lee et al.

[15]. The probability of elimination was calculated as the proportion of simulations achieving five-year elimination. One thousand simulations were carried out for the joint model, but the number of simulations was not reported for the original model as it used a collection of parameter sets rather than repeated simulations with the parameter MLEs.

Table 2: Table of estimated probabilities of elimination of cholera by vaccination scenario for the original Lee et al. model and the jointly estimated, altered model [15].

Model	No Vac.	2-Dept.	3-Dept.	Slow Nat.	Fast Nat.	Fast Nat., High Cov.
Original	5.8%	32.7%	64.5%	71.6%	79.6%	88.2%
Altered	99.0%	100%	100%	100%	100%	100%

The jointly estimated altered model predicted the elimination of cholera in all scenarios with very high probability. Plots of the simulated reported cases and true incidence by vaccination scenario (fig. A2, fig. A3) show that the elimination time was reduced as vaccination administration grew more rigorous. Interestingly, even in the case of no intervention, the model predicted fade-out of the disease by the start of 2021. These results contrast with the forecasting conducted by Lee et al. which predicted seasonal variation in the number of cases (zero to upwards of two thousand cases) across all scenarios as well as very low probability of elimination in the No Vaccination scenario with the possibility of reemergence [15].

6 Discussion

Using a combination of Monte Carlo, maximum likelihood, and simulation-based methods, we evaluated a stochastic state-space model for the 2010-2011 Haiti cholera epidemic proposed by Lee et al. at the Bloomberg School of Public Health along with two of our own variations on the model [1]. Though only consisting of minor adjustments in the calibration methodology and latent process specification, these altered models achieved improvements in likelihood over the original model.

These varied likelihood estimates coupled with the differences in forecasting results emphasize the importance of rigorous model fitting and interrogation. Epidemiological models can be of great importance when faced with an infectious disease outbreak as they can provide insights into the potential effectiveness or failure of intervention plans to mitigate spread.

However, the utility of these models hinges upon the assumption that they are appropriate for the situation; that is, that they are able to come close to describing the data-generating process(es). Whether due to improper parameter estimation or model misspecification, a model that is unable to explain the data is not likely to produce accurate predictions and is ultimately of little use to researchers and public health officials alike.

6.1 Limitations

After the onset of the epidemic, the Haiti Ministry of Health and Population began a national plan for combating cholera. The plan included an assortment of interventions including improvements to sanitation, water accessibility, and strict case monitoring [22]. Because of the time needed to implement large-scale changes to the water infrastructure of the country, many small vaccination campaigns were carried out from 2012 to 2018 as a first step to mitigating the spread. The number of individuals receiving at least one dose totaled over 1.5 million, or approximately 9% of the population of Haiti. Case-area targeted interventions (CATIs) consisting of cholera education, disinfecting spray, soap, and chlorine tablets were also supplied to over 48,000 locations across the country [17].

In this case study, the data used to fit all the models included 430 weeks of case data spanning October 23rd, 2010 through January 12th, 2019. This interval overlaps significantly with the periods in which the series of vaccination campaigns and CATIs were deployed. However, none of these models included mechanisms for the campaigns or the intervention programs. As a consequence of this, we remain critical of the final model parameter estimates reported in table A1 despite the fact that our adjusted models are competitive with the simple ARMA model. Maximum likelihood estimation attempts to find the parameter values for a model that maximize the probability of observing the data. This can come with unintended consequences related to the mechanisms underlying the processes of interest. There is the possibility that the MLE we found for any given parameter is not a good estimate of the true parameter value as the MLE may be biologically or ecologically improbable. In addition, the model construction can greatly influence the parameter estimates. In the case of our study, the exclusion of the vaccination campaigns and intervention programs may have led to other parameters compensating for the omitted model components in order to explain the data. Although our altered models outperformed that of the original authors according to likelihood-based measures, we must be mindful of the relationships between model construction, model fitting, and model quality.

Looking more deeply at figure 2 and figure 3, it is clear that the simulations do not

match the case reports in the period after the start of 2013 as closely as in the years prior to this point. The discrepancy between the data and the model simulations is more clearly visible when plotting on the natural logarithmic scale as in fig. A4, fig. A5, fig. A6). To understand why the change of scale highlights this disparity, suppose we have a dataset of two measurements: 1 case in week one and 100 cases in week two. We simulate cases and get 6 cases in week one and 105 cases in week two. Plotting this on the natural scale would show the same relative difference of 5 cases between the simulations and measurements. However, in week one we overestimated the number of cases by a factor of 6, while in week two we only overestimated by a factor of 1.05. Transforming the cases with the natural logarithm allows us to better diagnose model problems by illustrating these differences in magnitude.

It is also important to note that there are many other methods of model criticism that can be carried out to even more thoroughly evaluate model quality and fit [16, 4]. Additional likelihood profiles and model diagnostics such as plots of the autocorrelation between cases at different time points as well as analysis of the spatial distribution of cases and the variability of case forecasts through time can be instrumental in assessing a model's appropriateness in context [13]. Due to the nature of maximum likelihood estimation and the iterated filtering algorithm, there is also the option to carry out a more exhaustive search throughout the parameter space when fitting the model. The algorithmic parameters of IF2 include starting values for all estimated parameters, the number of particles used when filtering, and the number of iterations [12]. In an attempt to balance the price of computation with rigor of results, we used 450 (jointly estimated adjusted model) and 500 (adjusted model) sets of starting values, 5000 particles, and 200 iterations when fitting the adjusted models. But by using a larger number of starting points, particles, and iterations, one might be able to identify parameter values that achieve a higher likelihood.

Despite the fact that our models failed to produce simulations that visually matched the case data, we maintain that our study was successful. We aimed to exhibit crucial steps in epidemiological model construction and application by building upon the framework established by Lee et al. [1]. We therefore restricted our adjustments of the original model to minor alterations in order to underscore this important relationship between model development and model quality. Though the poorly matching simulations indicate that additional improvements can be made, the greater likelihoods and vastly contrasting forecasts still illustrate the significance of thorough model fitting procedures and proper model specification.

7 Conclusion

There have been no laboratory confirmed cases of or deaths caused by cholera in Haiti since early 2019, most likely due to the fast and widespread deployment of CATIs because of the limited distribution of vaccinations [17]. Similar to the reality of the last few years and Haiti's response to the epidemic, our jointly estimated altered model predicted the natural elimination of cholera by early 2021 without any vaccination interventions. Though conducted post hoc, these results have promising implications for the model's quality and, more broadly, for the application of mathematical modelling for disease transmission to informing public health decision-making.

Many of the techniques and methods used in this demonstration of the power of simulation-based inference can be extended to current and future epidemiological contexts such as the recent magnitude 7.2 earthquake that hit Haiti in August of 2021. Over 1,800 water supply systems, 53 healthcare sites, and 130,000 homes have been damaged by the earthquake and Tropical Depression Grace [23]. Though cholera has not been observed in Haiti for upwards of three years, the threat of its reemergence is present, especially considering the fact that the 2010-2011 epidemic arose in the aftermath of a magnitude 7.0 earthquake that ravaged the country and its infrastructure [18]. Hopefully, proper modelling can be helpful in informing policy by providing insights into the efficacy of potential response plans in an effort to avert a second cholera epidemic.

8 Appendix

Table A1: Table of parameter values achieving the maximum likelihood for the epidemic (epi) and endemic (end) calibrations separately with the exception of the jointly estimated altered model in which the two periods were combined. Parameters reported include reporting rate (ρ), measurement process overdispersion (τ), latent process overdispersion (σ^2), population mixing coefficient (ν), and seasonality terms ($\beta_{1:6}$). All parameters are reported to 4 decimal places.

Model	ρ	τ	σ^2	ν	$\beta_{1:6}$
Original (epi)	0.8220	15.2111	NA	0.9811	3.0048, 3.8524, 2.4215, 3.7633, 3.2372, 3.4116
Original (end)	0.9968	22.2507	NA	0.9925	3.1299, 3.3738, 2.1317, 3.3409, 2.8502, 2.7652
Altered (epi)	0.3148	376.7802	0.1016	0.9841	5.3324, 2.6566, 3.8325, 2.7666, 5.0974, 1.8034
Altered (end)	0.9517	85.4460	0.01122	0.9869	2.4296, 4.1216, 2.0811, 3.7738, 2.4402, 3.6037
Joint, Altered	0.4765 (epi) 0.4497 (end)	688.7796 (epi) 105.3583(end)	0.1106 (epi) 0.1677 (end)	0.9976	4.0148, 2.7089, 2.7423, 3.0589, 3.5747, 2.2309

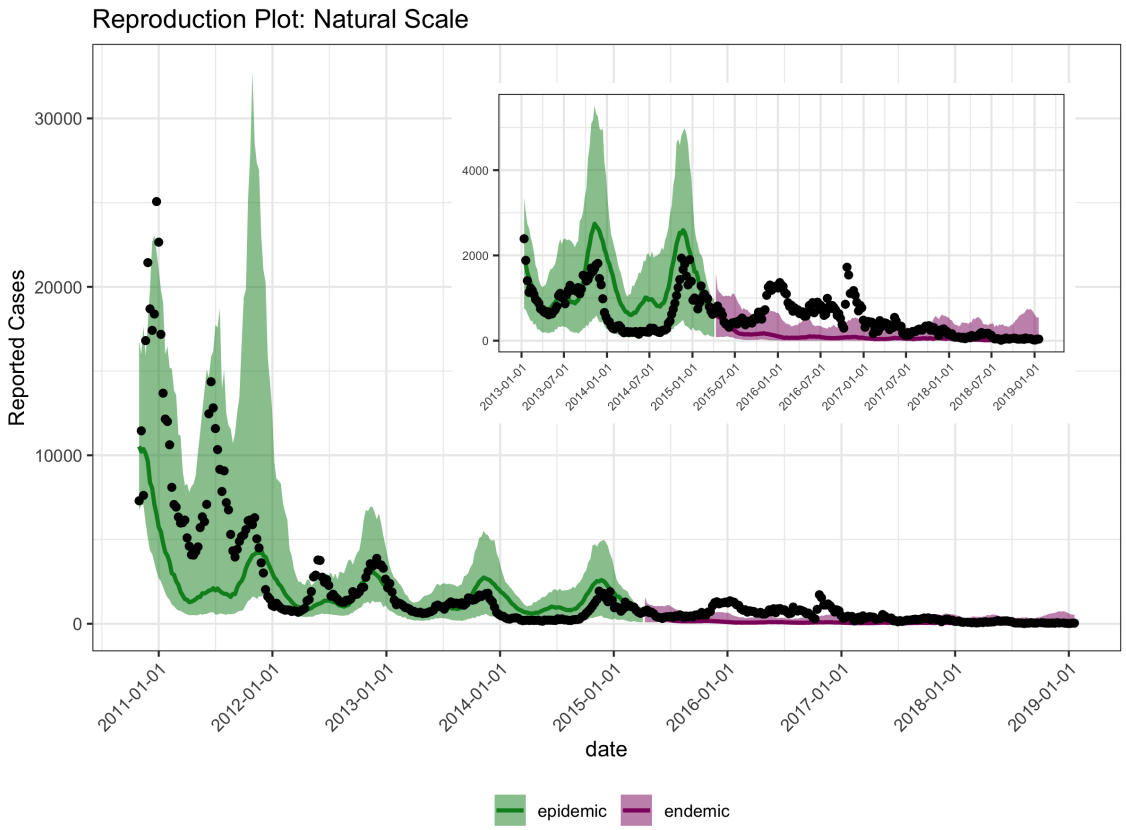

Figure A1: A reproduction of figure S7 of Lee et al. [15]. The solid line indicates the median number of reported cases from the simulations across sets of parameters, and the ribbon indicates the 2.5th and 97.5th percentiles for the simulated reported cases. Inset provides a closer view of the model fit from 2013 through 2018.

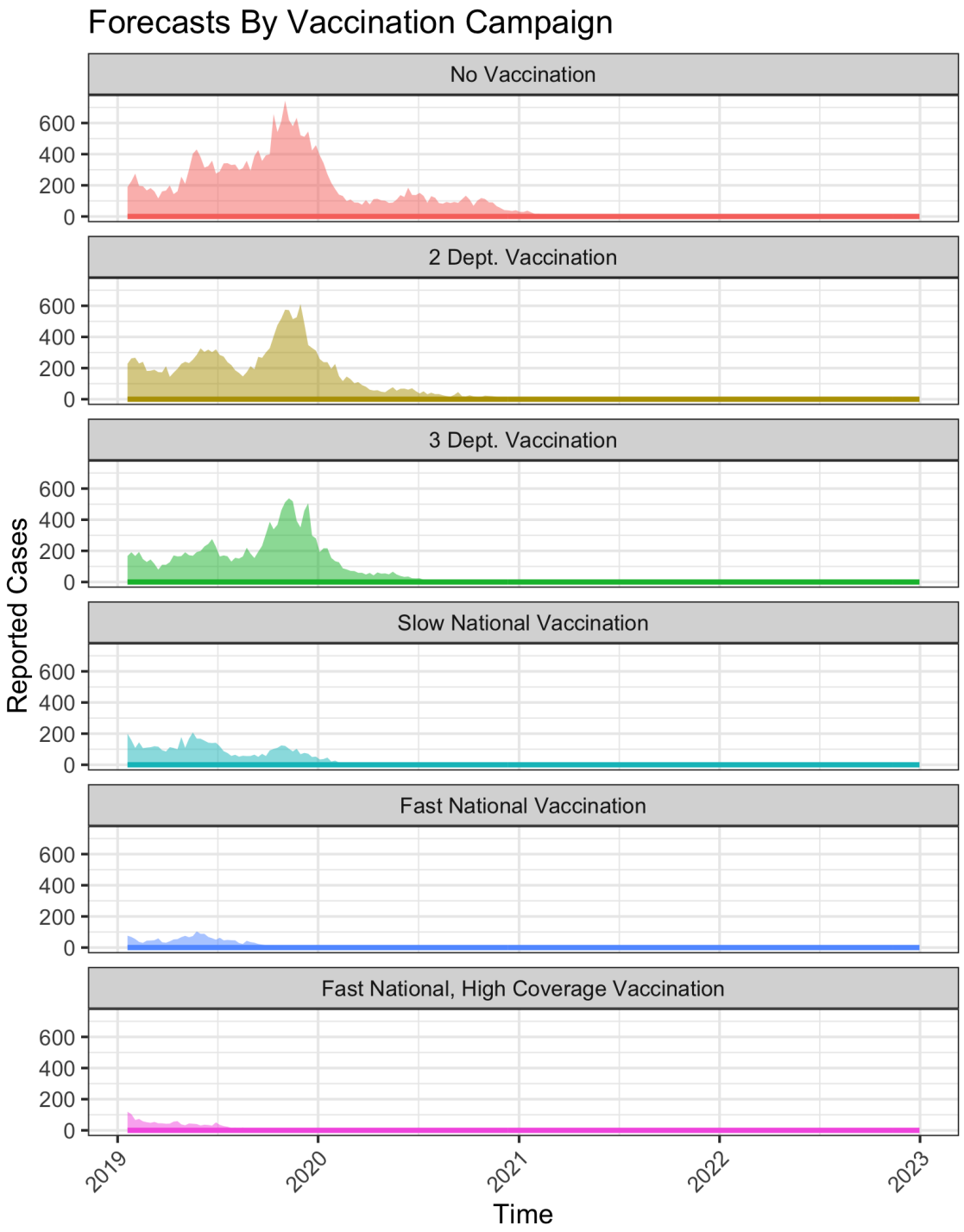

Figure A2: Forecast reported cases by vaccination campaign. Solid line indicates the median number of reported cases from the simulations across sets of parameters. Ribbon indicates the 2.5th and 97.5th percentiles for the simulated reported cases. Note that the plot shows forecasts until 2023 while the forecasting was conducted for a ten-year period extending until 2030. Reemergence did not occur in any campaign.

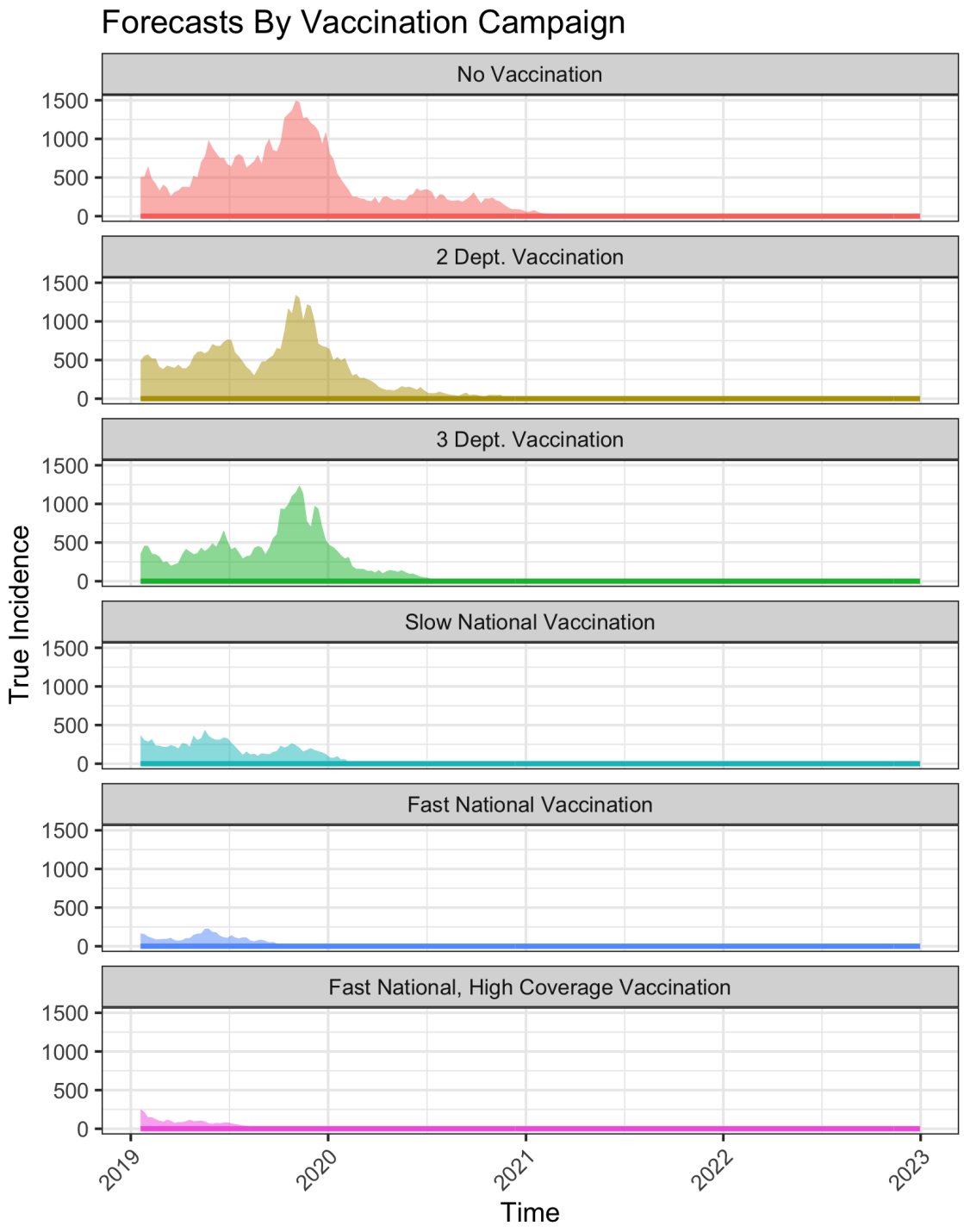

Figure A3: Forecast true incidence by vaccination campaign. Solid line indicates the median number of reported cases from the simulations across sets of parameters. Ribbon indicates the 2.5th and 97.5th percentiles for the simulated incidence. Note that the plot shows forecasts until 2023 while the forecasting was conducted for a ten-year period extending until 2030. Reemergence did not occur in any campaign.

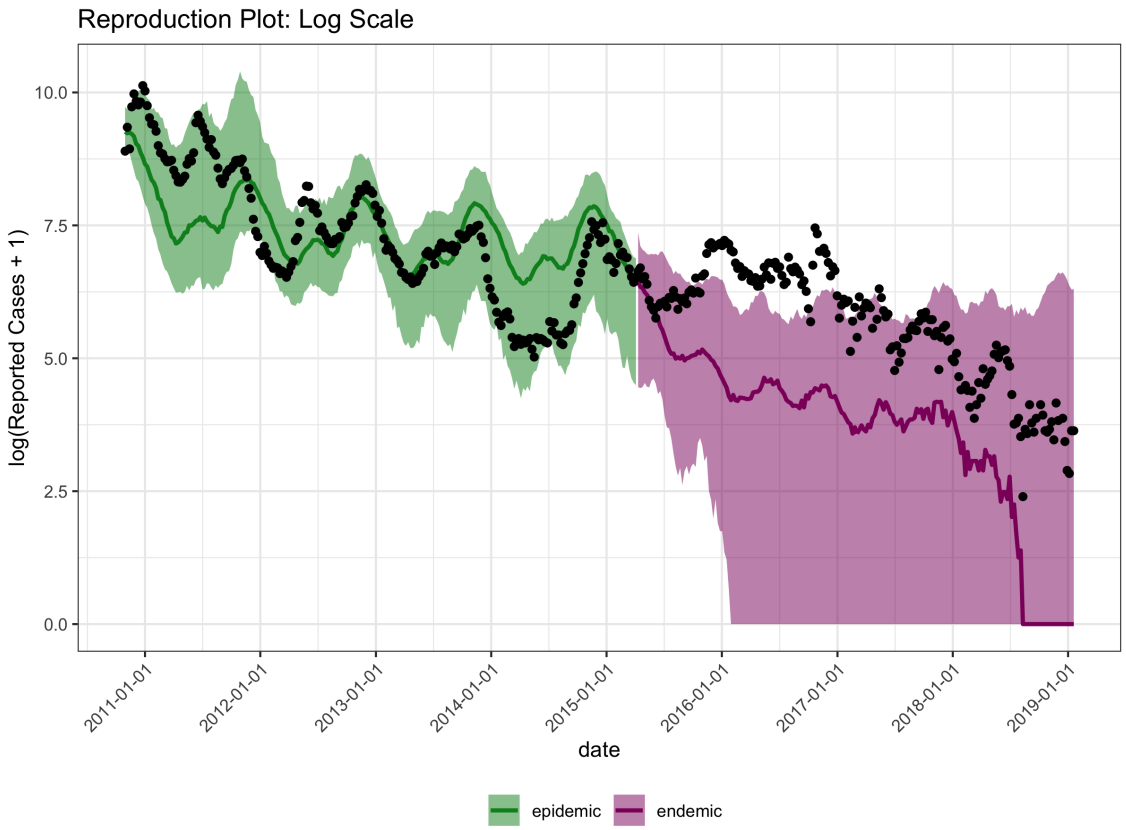

Figure A4: Plot of original model simulations on the log scale. Solid line indicates the median simulated reported cases across parameter sets. Ribbon indicates the 2.5th and 97.5th percentiles for the simulated reported cases.

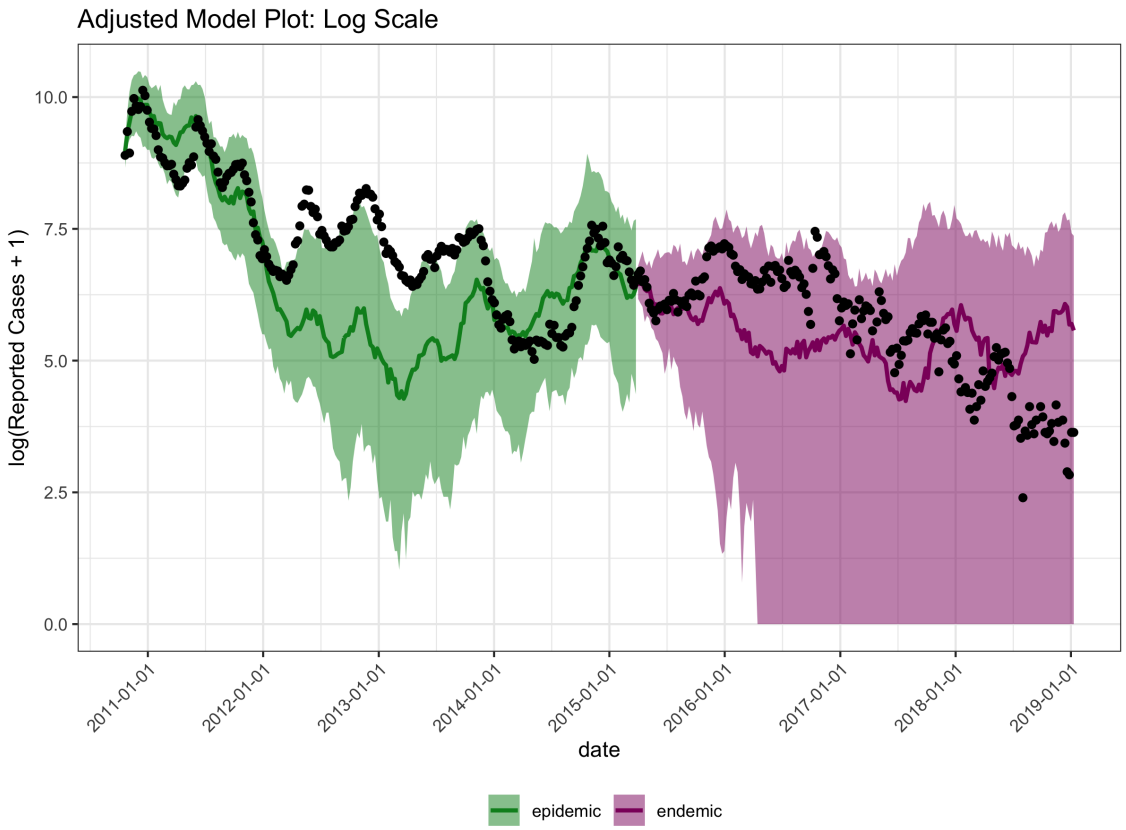

Figure A5: Plot of model simulations on the log scale with overdispersion in the latent process using the parameter MLEs. Solid line indicates the median simulated reported cases across parameter sets. Ribbon indicates the 2.5th and 97.5th percentiles for the simulated reported cases.

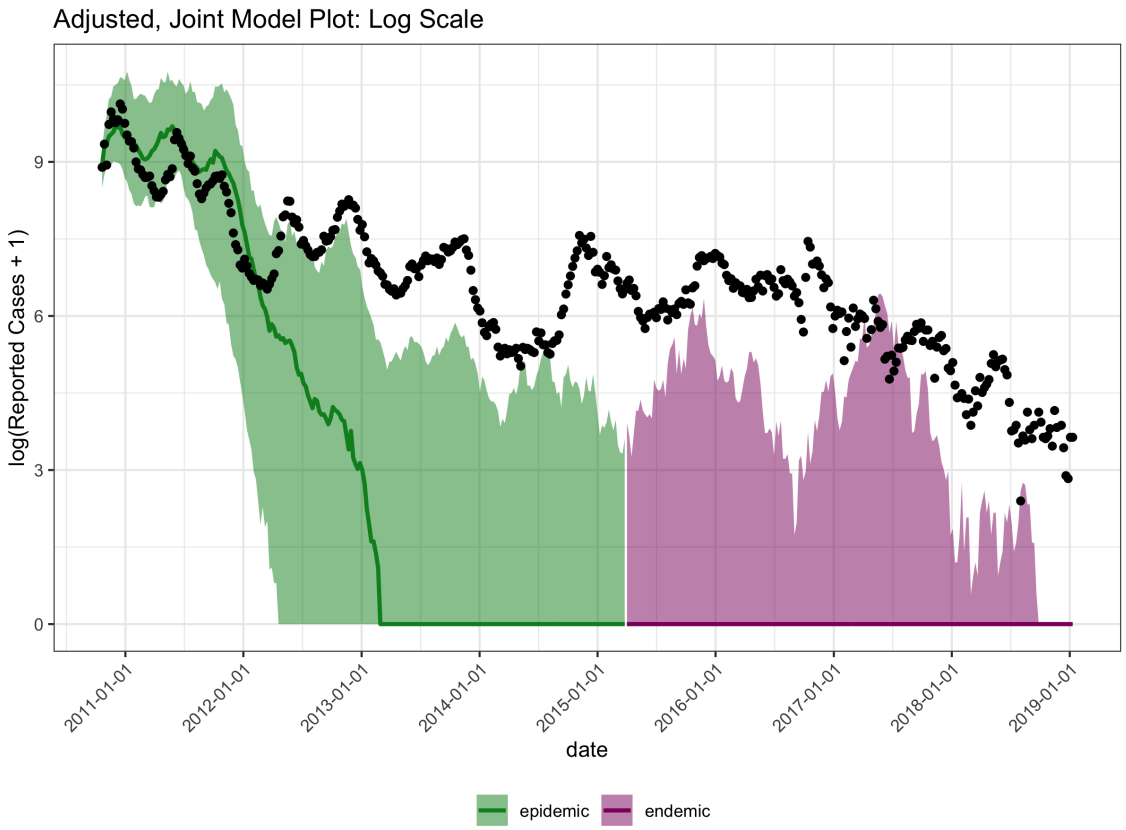

Figure A6: Plot of simulations from the jointly estimated model on the log scale with additional overdispersion in the latent process using the maximum likelihood estimates of the model parameters. Solid line indicates the median number of reported cases from simulations. Ribbon indicates the 2.5th and 97.5th percentiles for the simulated reported cases.

References

[1] Achieving coordinated national immunity and cholera elimination in Haiti through vaccination: a modelling study. *The Lancet Global Health*, 8(8):e1081–e1089, 2020.

[2] Carles Bretó, Daihai He, Edward L. Ionides, and Aaron A. King. Time series analysis via mechanistic models. *Annals of Applied Statistics*, 3(1):319–348, 2009.

[3] Dennis L. Chao, M. Elizabeth Halloran, and Ira M. Longini. Vaccination strategies for epidemic cholera in Haiti with implications for the developing world. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):7081–7085, 2011.

[4] Stephen R. Cole, Haitao Chu, and Sander Greenland. Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *American Journal of Epidemiology*, 179(2):252–260, 2014.

[5] Dan Crisan. The stochastic filtering problem: a brief historical account. *Journal of Applied Probability*, 51(A):13–22, 2014.

[6] Kashmira A. Date, Andrea Vicari, Terri B. Hyde, Eric Mintz, M. Carolina Danovaro-Holliday, Ariel Henry, Jordan W. Tappero, Thierry H. Roels, Joseph Abrams, Brenton T. Burkholder, Cuauhtémoc Ruiz-Matus, Jon Andrus, and Vance Dietz. Considerations for oral cholera vaccine use during outbreak after earthquake in Haiti, 2010-2011. *Emerging Infectious Diseases*, 17(11):2105–2112, 2011.

[7] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Cham, 2018.

[8] Isaac Chun Hai Fung, David L. Fitter, Rebekah H. Borse, Martin I. Meltzer, and Jordan W. Tappero. Modeling the effect of water, sanitation, and hygiene and oral cholera vaccine implementation in Haiti. *American Journal of Tropical Medicine and Hygiene*, 89(4):633–640, 2013.

[9] Kesten C. Green and J. Scott Armstrong. Simple versus complex forecasting: The evidence. *Journal of Business Research*, 68(8):1678–1685, 2015.

[10] E. L. Ionides, C. Bretó, and A. A. King. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 103(49):18438–18443, 2006.

[11] Edward L. Ionides, Anindya Bhadra, Yves Atchadé, and Aaron King. Iterated filtering. *Annals of Statistics*, 39(3):1776–1802, 2011.

[12] Edward L Ionides, Dao Nguyen, Yves Atchadé, Stilian Stoev, and Aaron A King. Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719 LP – 724, jan 2015.

[13] Aaron A. King, Matthieu Domenech De Cellés, Felicia M.G. Magpantay, and Pejman Rohani. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society B: Biological Sciences*, 282(1806):0–6, 2015.

[14] Aaron A. King, Dao Nguyen, and Edward L. Ionides. Statistical inference for partially observed markov processes via the R package pomp. *Journal of Statistical Software*, 69:1–43, 2016.

[15] Elizabeth C. Lee, Dennis L. Chao, Joseph C. Lemaitre, Laura Matrajt, Damiano Pasetto, Javier Perez-Saez, Flavio Finger, Andrea Rinaldo, Jonathan D. Sugimoto, M. Elizabeth Halloran, Ira M. Longini, Ralph Ternier, Kenia Vissieres, Andrew S. Azman, Justin Lessler, and Louise C. Ivers. Supplementary appendix 3: Achieving coordinated national immunity and cholera elimination in Haiti through vaccination: a modelling study. *The Lancet Global Health*, 8(8):e1081–e1089, 2020.

[16] Russell B. Millar. *Maximum Likelihood Estimation and Inference: with exmaples in R, SAS, and ADMB*. John Wiley & Sons, Ltd., 2011.

[17] Stanislas Rebaudet, Patrick Dély, Jacques Boncy, Jean Hugues Henrys, and Renaud Piarroux. Toward Cholera Elimination, Haiti. *Emerging Infectious Diseases*, 27(11):2932–2936, 2021.

[18] Valentine Sanon, Frantz Sainvil, R Awan, George P Einstein, and L Orien. Haiti's cholera epidemic: will it return in 2021? *Gastroenterology & Hepatology*, 12(4):124–126, 2021.

[19] Robert H. Shumway and David S. Stoffer. Time Series Analysis and Its Applications. In George Casella, Stephen Fienberg, and Ingram Olkin, editors, *Time Series Analysis and Its Applications*, chapter Characteristics of Time Series. Springer Science+Business Media Inc., New York, second edition, 2006.

[20] Robert H. Shumway and David S. Stoffer. Time Series Analysis and Its Applications. In George Casella, Stephen Fienberg, and Ingram Olkin, editors, *Time Series Analysis and Its Applications*, chapter State-Space Models. Springer Science+Business Media Inc., New York, second edition, 2006.

[21] Robert H. Shumway and David S. Stoffer. Time Series Analysis and Its Applications. In George Casella, Stephen Fienberg, and Ingram Olkin, editors, *Time Series Analysis and Its Applications*, chapter ARIMA Models. Springer Science+Business Media Inc., New York, second edition, 2006.

[22] Rania A. Tohme, Jeannot François, Kathleen Wannemuehler, Preetha Iyengar, Amber Dismer, Paul Adrien, Terri B. Hyde, Barbara J. Marston, Kashmira Date, Eric Mintz, and Mark A. Katz. Oral cholera vaccine coverage, barriers to vaccination, and adverse events following vaccination, Haiti, 2013. *Emerging Infectious Diseases*, 21(6):984–991, 2015.

[23] UNICEF. Emergency Response Haiti Earthquake, 2021.

[24] John Zarocostas. Cholera outbreak in Haiti-from 2010 to today. *Lancet (London, England)*, 389(10086):2274–2275, 2017.