# Supplementary Materials for

# Predictive Modeling of Cholera Outbreaks in Bangladesh

Amanda A. Koepke, Ira M. Longini, Jr., M. Elizabeth Halloran, Jon Wakefield, and Vladimir N. Minin

## Appendix A: Simulation Details

### PMMH pseudocode

The following exposition of the algorithm follows closely the pseudocode of Andrieu et al. [2010] and Wilkinson [2011].

Step 1: initialization, for iteration $j = 0$,

(a) Set $\boldsymbol{\theta}(0)$ arbitrarily

(b) Run the following SMC algorithm to get $\hat{p}(\boldsymbol{y}|\boldsymbol{\theta}(0))$, an estimate of the marginal likelihood, and to produce a sample $\boldsymbol{X}_{\boldsymbol{t}_{0:n}}(0) \sim \hat{p}(\cdot|\boldsymbol{y}, \boldsymbol{\theta}(0))$.

Let the superscript $k \in \{1, \ldots, K\}$ denote the particle index, where $K$ is the total number of particles, and the subscript $t_i \in \{t_0, \ldots, t_n\}$ denote the time; thus, $\boldsymbol{X}^k_{t_i}$ denotes the $k$th particle at time $t_i$, and $\boldsymbol{X}^k_{\boldsymbol{t}_{0:i}} = \left(\boldsymbol{X}^k_{t_0}, \ldots, \boldsymbol{X}^k_{t_i}\right)$. At time $t_i = t_0$, sample $\boldsymbol{X}^k_{t_0} = (S^k_{t_0}, I^k_{t_0})$ for $k = 1, \ldots, K$ from the initial density of the hidden Markov state process. Specifically, sample $S^k_{t_0} \sim$ Poisson$(\phi_S)$ and $I^k_{t_0} \sim$ Poisson$(\phi_I)$. Compute the $k$ weights $w(\boldsymbol{X}^k_{t_0}) := \Pr(y_{t_0}|\boldsymbol{X}^k_{t_0}, \boldsymbol{\theta}(0)) = \binom{I^k_{t_0}}{y_{t_0}} \rho(0)^{y_{t_0}} (1 - \rho(0))^{I^k_{t_0} - y_{t_0}}$, and set $W(\boldsymbol{X}^k_{t_0}) = w(\boldsymbol{X}^k_{t_0}) / \sum_{k'=1}^{K} w(\boldsymbol{X}^{k'}_{t_0})$.

For $i = 1, \ldots, n$, resample $\bar{\boldsymbol{X}}^k_{t_{i-1}}$ from $\boldsymbol{X}^k_{t_{i-1}}$ with weights $W(\boldsymbol{X}^k_{t_{i-1}})$. Sample $K$ particles $\boldsymbol{X}^k_{t_i}$ from $p(\cdot|\bar{\boldsymbol{X}}^k_{t_{i-1}})$ (i.e. propagate resampled particles forward one time point). Assign weights

$w(\boldsymbol{X}_{t_i}^k) := \Pr(y_{t_i}|\boldsymbol{X}_{t_i}^k, \boldsymbol{\theta}(0))$ and compute normalized weights $W(\boldsymbol{X}_{t_i}^k) = w(\boldsymbol{X}_{t_i}^k)/\sum_{k'=1}^K w(\boldsymbol{X}_{t_i}^{k'})$. Set $\boldsymbol{X}_{\boldsymbol{t}_{0:i}}^k = (\bar{\boldsymbol{X}}_{\boldsymbol{t}_{0:i-1}}^k, \boldsymbol{X}_{t_i}^k)$.

It follows that

$$\hat{p}(y_{t_i}|\boldsymbol{y}_{\boldsymbol{t}_{0:i-1}}, \boldsymbol{\theta}(0)) = \frac{1}{K}\sum_{k=1}^K w(\boldsymbol{X}_{t_i}^k)$$

is an approximation to the likelihood $p(y_{t_i}|\boldsymbol{y}_{\boldsymbol{t}_{0:i-1}}, \boldsymbol{\theta}(0))$, and therefore an approximation to the total likelihood is

$$\hat{p}(\boldsymbol{y}|\boldsymbol{\theta}(0)) = \hat{p}(y_{t_0}|\boldsymbol{\theta}(0))\prod_{i=1}^n \hat{p}(y_{t_i}|\boldsymbol{y}_{\boldsymbol{t}_{0:i-1}}, \boldsymbol{\theta}(0)).$$

Thus we have a simple, sequential, likelihood-free algorithm which generates an unbiased estimate of the marginal likelihood, $p(\boldsymbol{y}|\boldsymbol{\theta}(0))$. A $\boldsymbol{X}_{\boldsymbol{t}_{0:n}}(0)$ trajectory is sampled from the $K$ trajectories ($\boldsymbol{X}_{\boldsymbol{t}_{0:n}}^k$, for $k = 1, \ldots, K$) based on the final set of particle weights, $W(\boldsymbol{X}_{t_n}^k)$.

Step 2: for iteration $j \geq 1$,

(a) Sample $\boldsymbol{\theta}^* \sim q\{\cdot|\boldsymbol{\theta}(j-1)\}$

(b) Run an SMC algorithm, as in step 1(b) with $\boldsymbol{\theta}^*$ instead of $\boldsymbol{\theta}(0)$, to get $\hat{p}(\boldsymbol{y}|\boldsymbol{\theta}^*)$ and $\boldsymbol{X}_{\boldsymbol{t}_{0:n}}^* \sim \hat{p}(\cdot|\boldsymbol{y}, \boldsymbol{\theta}^*)$

(c) With probability

$$\min\left\{1, \frac{\hat{p}(\boldsymbol{y}|\boldsymbol{\theta}^*)}{\hat{p}(\boldsymbol{y}|\boldsymbol{\theta}(j-1))}\frac{\Pr(\boldsymbol{\theta}^*)}{\Pr\{\boldsymbol{\theta}(j-1)\}}\frac{q\{\boldsymbol{\theta}(j-1)|\boldsymbol{\theta}^*\}}{q\{\boldsymbol{\theta}^*|\boldsymbol{\theta}(j-1)\}}\right\}$$

set $\boldsymbol{\theta}(j) = \boldsymbol{\theta}^*$, $\boldsymbol{X}_{\boldsymbol{t}_{0:n}}(j) = \boldsymbol{X}_{\boldsymbol{t}_{0:n}}^*$, and $\hat{p}(\boldsymbol{y}|\boldsymbol{\theta}(j)) = \hat{p}(\boldsymbol{y}|\boldsymbol{\theta}^*)$, otherwise set $\boldsymbol{\theta}(j) = \boldsymbol{\theta}(j-1)$, $\boldsymbol{X}_{\boldsymbol{t}_{0:n}}(j) = \boldsymbol{X}_{\boldsymbol{t}_{0:n}}(j-1)$, and $\hat{p}(\boldsymbol{y}|\boldsymbol{\theta}(j)) = \hat{p}(\boldsymbol{y}|\boldsymbol{\theta}(j-1))$.

### Simulating homogeneous SIRS

Gillespie's direct method [Gillespie, 1977] simulates the time to the next event and then determines which event happens at that time. The first reaction method [Gillespie, 1976] calculates the time to

the next reaction for each of the possible events, and the minimum time to next reaction determines the next step of the chain.

Using the direct method, we can think of our continuous-time Markov chain (CTMC) as a chemical system with three different reactions. These reactions and their rate functions are given by the infinitesimal rates

$$\lambda_{(S,I,R),(S',I',R')} = \begin{cases} (\beta I + \alpha)S & \text{if } S' = S - 1,\ I' = I + 1,\ R' = R, \\ \gamma I & \text{if } S' = S,\ I' = I - 1,\ R' = R + 1, \\ \mu R & \text{if } S' = S + 1,\ I' = I,\ R' = R - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Thus the three reactions have the rate functions $h_1(\boldsymbol{X}_t) = (\beta I_t + \alpha)S_t$, $h_2(\boldsymbol{X}_t) = \gamma I_t$, and $h_3(\boldsymbol{X}_t) = \mu R_t$, corresponding to the infinitesimal rates of the CTMC. Then the time to the next reaction, $\tau$, has an exponential distribution with rate $\lambda = h_1(\boldsymbol{X}_t) + h_2(\boldsymbol{X}_t) + h_3(\boldsymbol{X}_t)$, and the $k$th reaction occurs with probability $h_k(\boldsymbol{X}_t)/\lambda$, for $k = \{1, 2, 3\}$.

The first reaction method instead simulates the time $\tau_k$ that the $k$th reaction happens for $k = \{1, 2, 3\}$, given no other reactions happen in that time. Then the time to the next reaction $\tau = \min_k(\tau_k)$, and the reaction with the reaction time equal to $\tau$ is the event that happens.

Both the direct method and the first reaction method work only for homogeneous Markov chains. If we want to assume that the additional force of infection, $\alpha$, varies over time, the associated Markov chain is inhomogeneous and we must account for the fact that the transition rate could change before the next reaction occurs.

### Simulating inhomogeneous SIRS

Gibson and Bruck [2000] introduce the next reaction method, an efficient exact algorithm to simulate stochastic chemical systems. They extend this next reaction method to include time-dependent rates and non-Markov processes. Anderson [2007] deviates from these methods a bit, using Poisson

processes to represent the reaction times, with time to next reaction given by integrated rate functions. This leads to a more efficient modified next reaction method which they extend to systems with more complicated reaction dynamics.

Using the methods described by Gibson and Bruck [2000] and Anderson [2007], to incorporate a time-varying force of infection into the SIRS model we must integrate over the rate function $h_1(\boldsymbol{X}_t, s) = (\beta I_t + \alpha(s))\, S_t$. Thus, to find the time $\tau_1$ that the first reaction happens, given no other reactions happen in that time, we generate $u \sim \text{Uniform}(0, 1)$ and solve

$$\int_t^{\tau_1} h_1(\boldsymbol{X}_t, s) ds = \ln(1/u)$$

for $\tau_1$. Since the other two reactions have no time-varying parameters, we can solve for $\tau_2$ and $\tau_3$, the reaction times of the second and third reactions, using the methods of the previous section. Then we can continue, using the first reaction method to simulate the process.

We simplify this approach by assuming that the time-varying force of infection, $\alpha(t)$, remains constant each day. We define daily time intervals $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n-1\}$, and $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$. Then we can take advantage of the memoryless property of exponentials and propagate the chain forward in daily increments. Thus, we use the direct method, but when the time to next event exceeds the right end point of the current interval $A_i$, we restart CTMC simulation from the beginning of the interval $A_{i+1}$ using $\alpha_{A_{i+1}}$ in the waiting time distribution rate $\lambda(\alpha_A) = h_1(\boldsymbol{X}_t, \alpha_A) + h_2(\boldsymbol{X}_t) + h_3(\boldsymbol{X}_t)$, so $\tau \sim \text{Exp}(\lambda(\alpha_A))$. This modified Gillespie algorithm is depicted and detailed in Figure A-1.

### Selecting Tau

Unchecked, tau-leaping can lead to negative population sizes in a compartment if the compartment has a low number of individuals. To avoid this, we use a simplified version of the modified tau-leaping algorithm presented by Cao et al. [2005]. If the population of a compartment is lower than some prespecified critical size, a single step algorithm (like the Gillespie algorithm) is used until

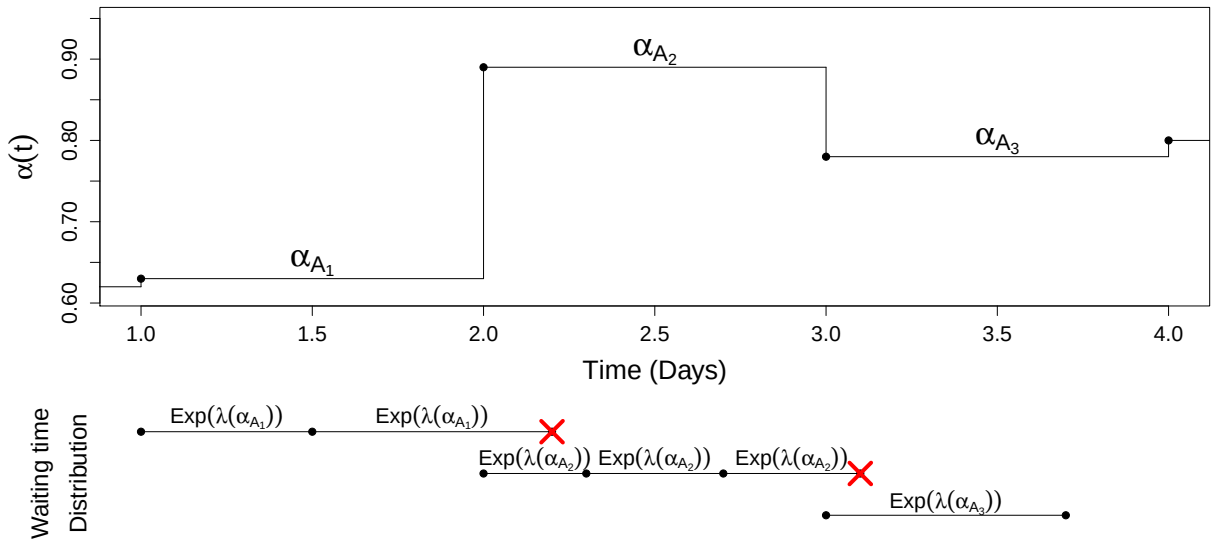


Figure A-1: Depiction of the modified Gillespie algorithm. We assume the environmental force of infection, $\alpha(t)$, is a step function which changes daily. Daily time intervals are denoted by $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n - 1\}$, so $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$. Starting at time $t = 1$, the waiting time to the next event, $\tau$, has an exponential distribution with rate $\lambda(\alpha_{A_1}) = h_1(\boldsymbol{X}_t, \alpha_{A_1}) + h_2(\boldsymbol{X}_t) + h_3(\boldsymbol{X}_t)$. In the depiction, $\tau = 0.5$. The simulated waiting time plus the current time, $t^* = t + \tau$, remains in the interval $A_1$, so we use $t^*$ as the next time in our CTMC and propagate $\boldsymbol{X}_t$ forward at that time using Gillespie's direct method. Since we are still in the interval $A_1$, we again simulate the time to the next event as an exponential random variable with rate $\lambda(\alpha_{A_1}) = h_1(\boldsymbol{X}_{t^*}, \alpha_{A_1}) + h_2(\boldsymbol{X}_{t^*}) + h_3(\boldsymbol{X}_{t^*})$. In this iteration, the waiting time plus the current time, $t^* + \tau$, exceeds the boundary of the interval $A_1$, so we discard this simulated waiting time $\tau$. Using the memoryless property of exponentials, we restart our simulation from the beginning of the interval $A_2$ using the new $\alpha(t)$ value, $\alpha_{A_2}$. We continue in this manner until we have simulated the Markov process $\boldsymbol{X}_t$ up to time $t_n$.

the population gets above that critical size. If the size of the compartment is not critically low but the current value of $\tau$ still produces a negative population, we reject that simulation and try again with a smaller $\tau$ (reduced by a factor of 1/2). The subsequent value of $\tau$ is picked based on how long the current daily time-varying force of infection remains constant. We choose a value of $\tau$ that simulates what happens during the remainder of the day, until the value of the transition rate changes. This modified tau-leaping algorithm is depicted and detailed in Figure A-2.

For our simulations, we have chosen $\tau = 1$ day; we perform a simulation study to see if this value for $\tau$ is reasonable. Using the posterior estimates of the parameters, we simulate data forward in time

5000 times using both the modified Gillespie algorithm and the modified tau-leaping algorithm. We simulate data over the entire epidemic curve to see how the comparison changes for varying values of $\alpha(t)$. Figure A-3 shows estimates of the median and 95% intervals for the simulated values. The Monte Carlo standard error is very small for all estimates. For the numbers of susceptible individuals, the estimates under Gillespie and tau-leaping are almost identical over the entire epidemic. For the numbers of infected, the values are very close except at the epidemic peaks. However, the differences are very small. We conclude that for our application $\tau = 1$ day is a good compromise between computational efficiency and accuracy.

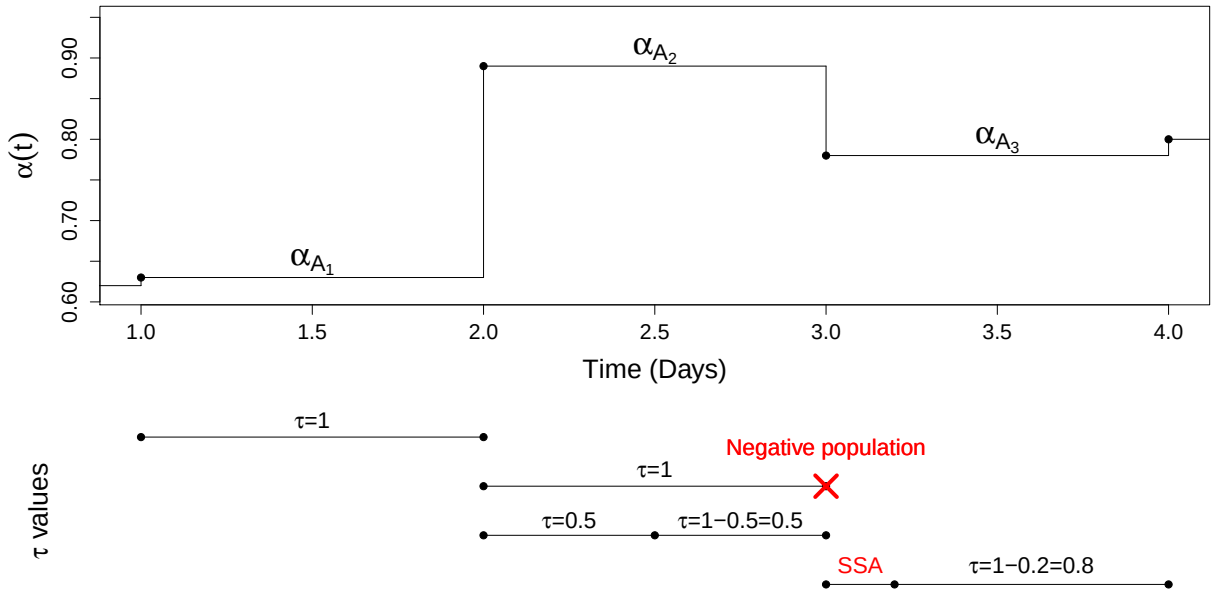


Figure A-2: Depiction of the modified tau-leaping algorithm. We assume the environmental force of infection, $\alpha(t)$, is a step function which changes daily. Daily time intervals are denoted by $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n - 1\}$, so $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$. As a default, we use $\tau = 1$ day. Starting at time $t = 1$, we simulate the changes in compartment populations over the interval $t \in [1, 2)$. At time $t = 2$, we again use $\tau = 1$ day to simulate the changes over the interval $t \in [2, 3)$. This value of $\tau$ produces a negative population so we reject that simulation and try again with a smaller $\tau$ (reduced by a factor of $1/2$). The next value of $\tau$ is then calculated based on how long the current daily time-varying force of infection remains constant, so $\tau = 0.5$. At time $t = 3$, the population of a compartment is lower than some prespecified critical size, so a single step algorithm (SSA), in our case the Gillespie algorithm, is used until the population gets above that critical size. Once the compartment populations are all above the critical size again, at time $t = 3.2$, the subsequent value of $\tau$ is again picked based on how long the current daily time-varying force of infection remains constant, so $\tau = 0.8$.

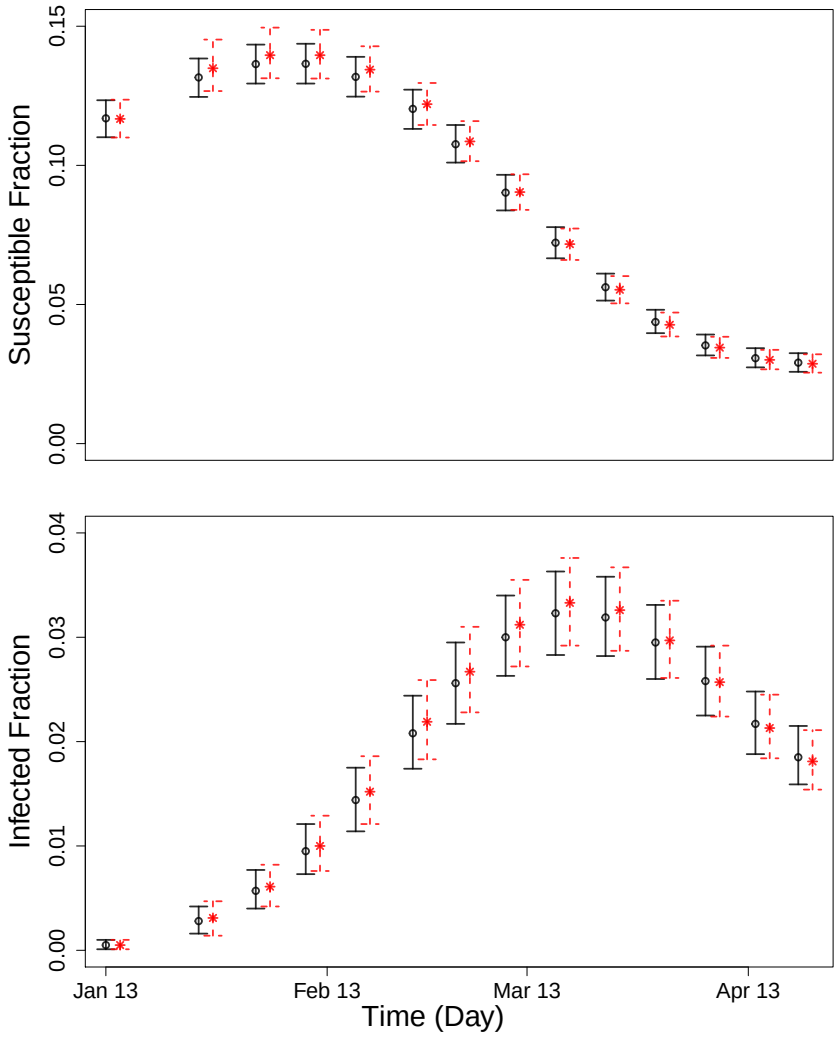


Figure A-3: Plots comparing the median and 95% intervals at different points during an epidemic, simulated using both the modified Gillespie algorithm and the modified tau-leaping algorithm with $\tau = 1$ day. The medians and 95% intervals for 5000 simulations using the Gillespie algorithm are given by the open circle and solid error bars. The medians and 95% intervals for 5000 simulations using the modified tau-leaping algorithm are given by the asterisk and dashed error bars.

## Appendix B: MCMC diagnostics

Using simulated data, we compare results from models with different assumptions on the values of $\phi_S$ and $\phi_I$; marginal posterior distributions for the parameters of the SIRS model from the final runs of PMMH algorithms are in Figure B-2. The posterior distributions are similar, regardless of assumptions about $\phi_S/N$ and $\phi_I/N$. Trace plots and autocorrelation plots for the parameters of the SIRS model assuming $\phi_S$ and $\phi_I$ are set at the true values are in Figure B-3, and Figure B-4 shows bivariate scatterplots of the parameters. Summary plots of the PMMH algorithm output for the parameters of the SIRS model with data from Mathbaria, Bangladesh are given in Figure B-6, and Figure B-7 shows bivariate scatterplots of the parameters. Effective sample sizes range from 928 to 7546 for the parameters of the SIRS model with a time-varying environmental force of infection and from 1590 to 3870 for the analysis of the data from Mathbaria. To test convergence, we varied the initial values for the parameters of the PMMH algorithm. Some of the initial values are shown in Table B-1 and the parameter estimates from the chains that started at these initial values are given in the top third of Table B-2.

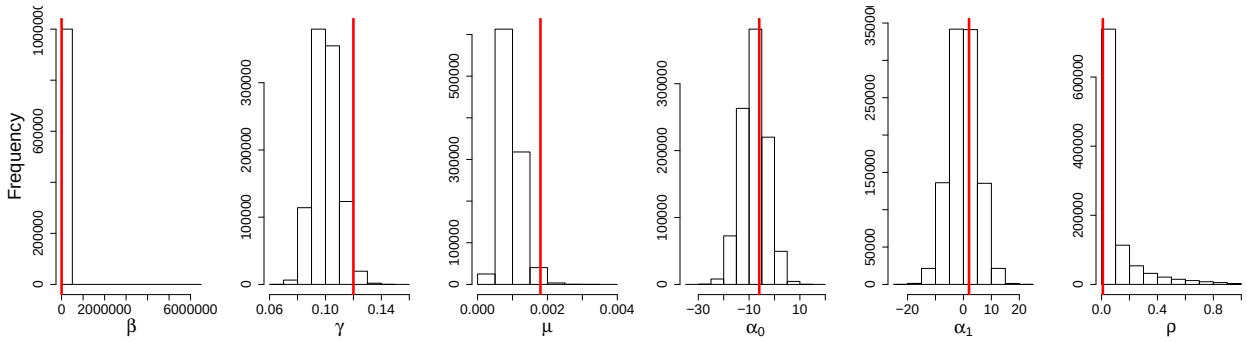


Figure B-1: Prior distributions for the parameters of the SIRS model used in simulated data example. The true values of the parameters are denoted by the red lines.

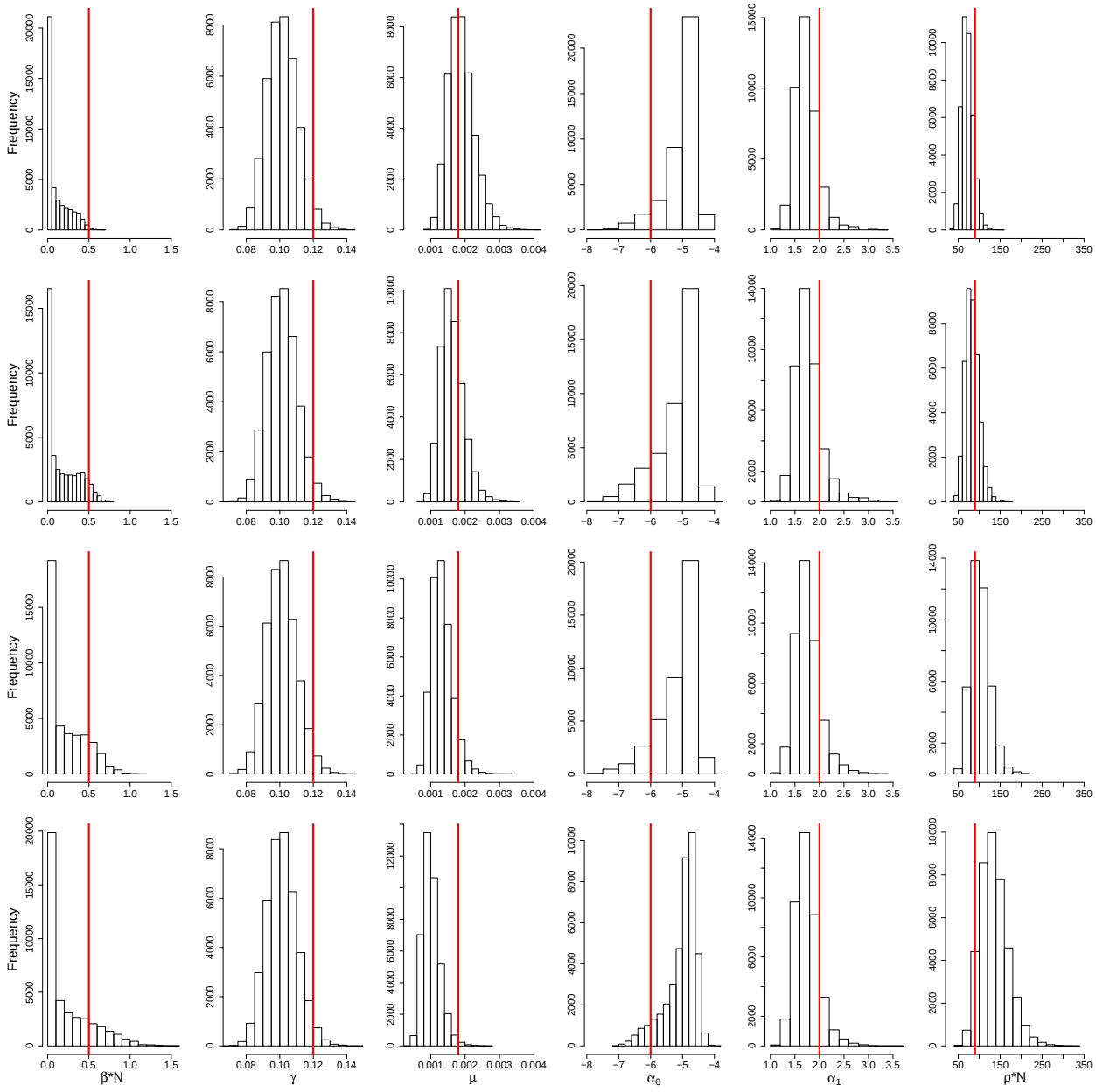


Figure B-2: Posterior distributions for the parameters of the SIRS model, based on simulated data. From top to bottom, the rows have $\phi_S/N$ and $\phi_I/N$ above the true values (0.39 and 0.0168), at the true values (0.29 and 0.0084), below the true values (0.19 and 0.0042), and further below (0.095 and 0.0021). The true values of the parameters are denoted by the red lines.

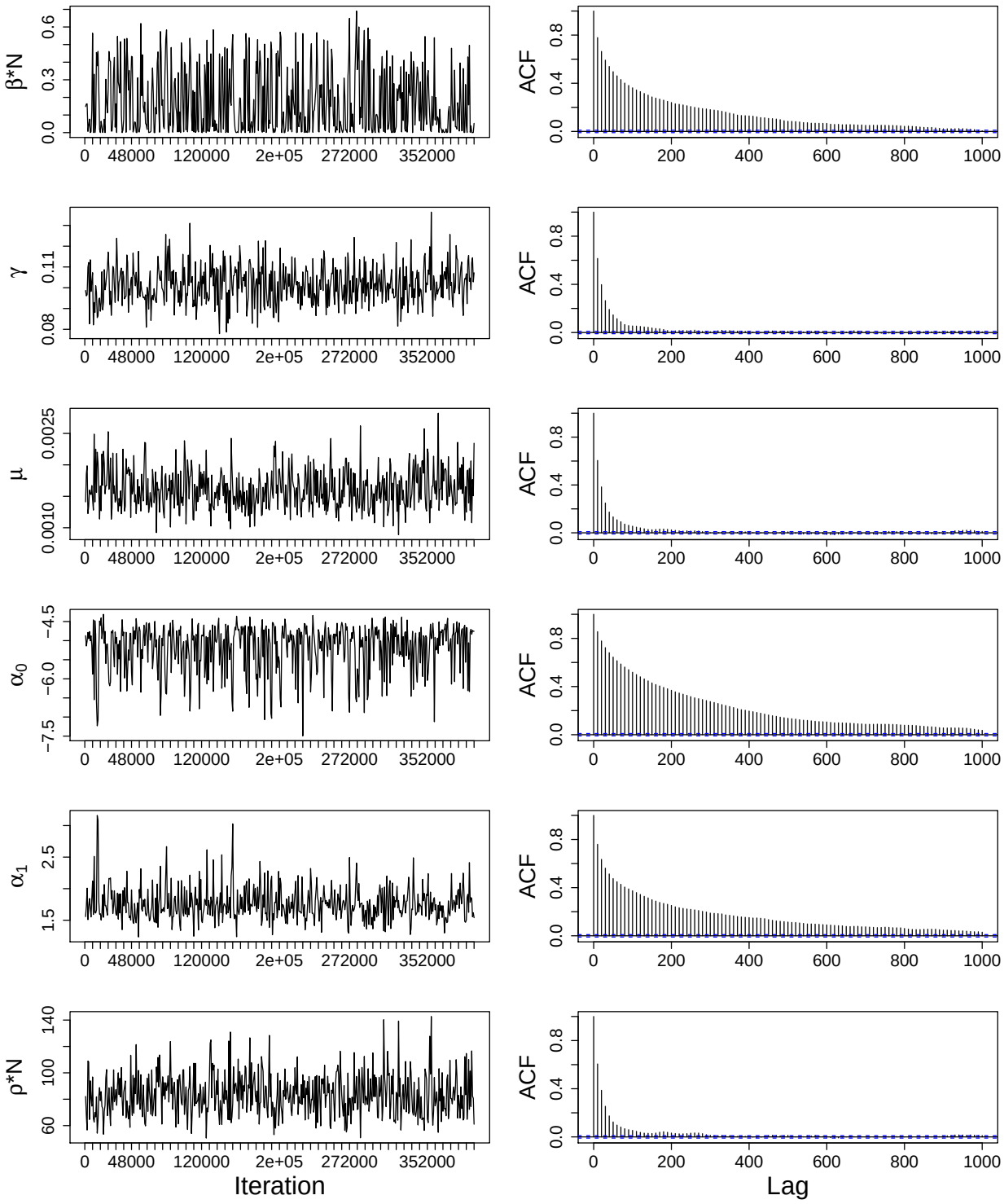


Figure B-3: Summary plots of the PMMH algorithm output (final run of 400000 iterations) for the parameters of the SIRS model, based on simulated data. ACF plots are thinned to 40000 iterations and trace plots are thinned to display only 500 iterations.

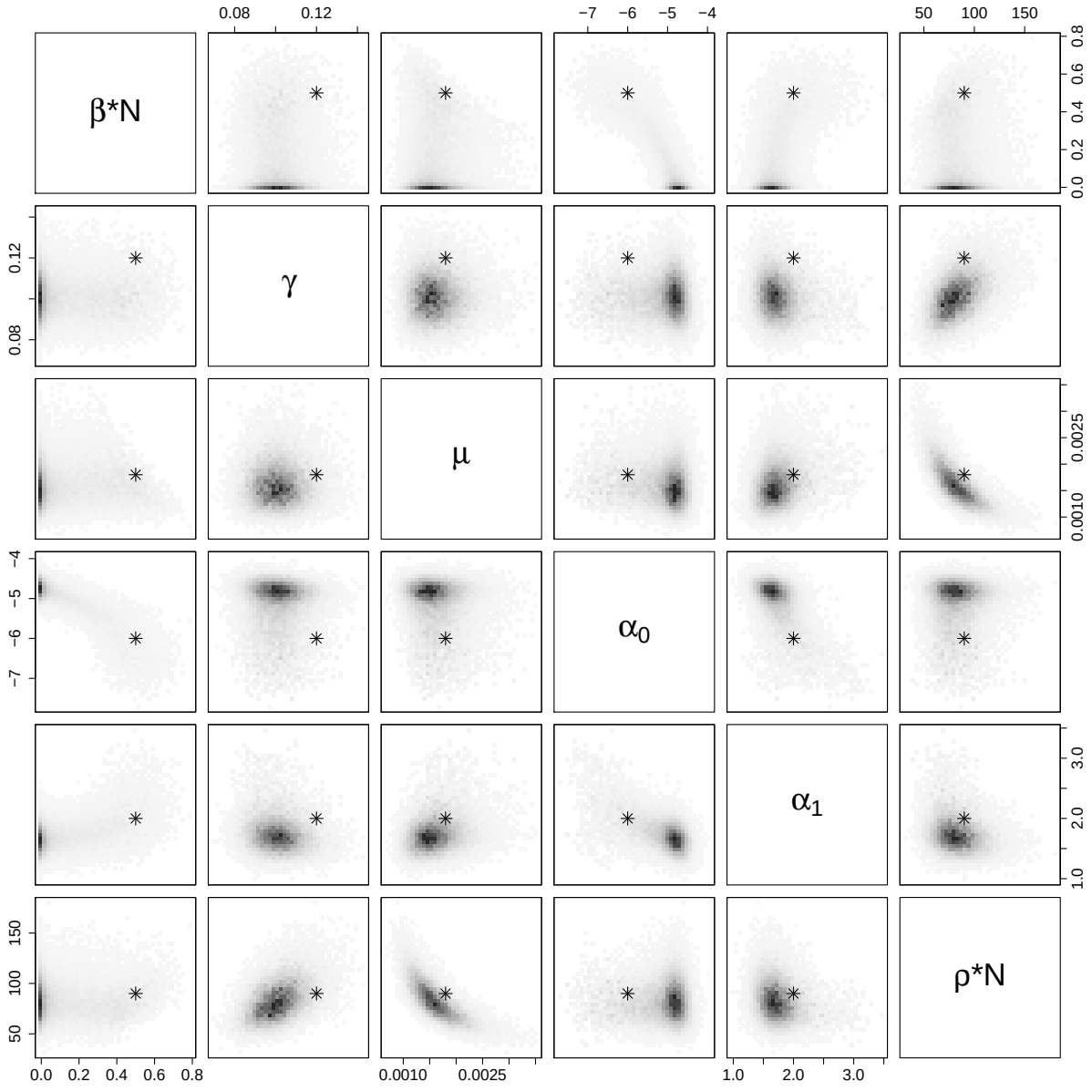


Figure B-4: Smoothed bivariate scatterplots of parameters of the SIRS model, based on simulated data. True values of the parameters are denoted by the asterisks.

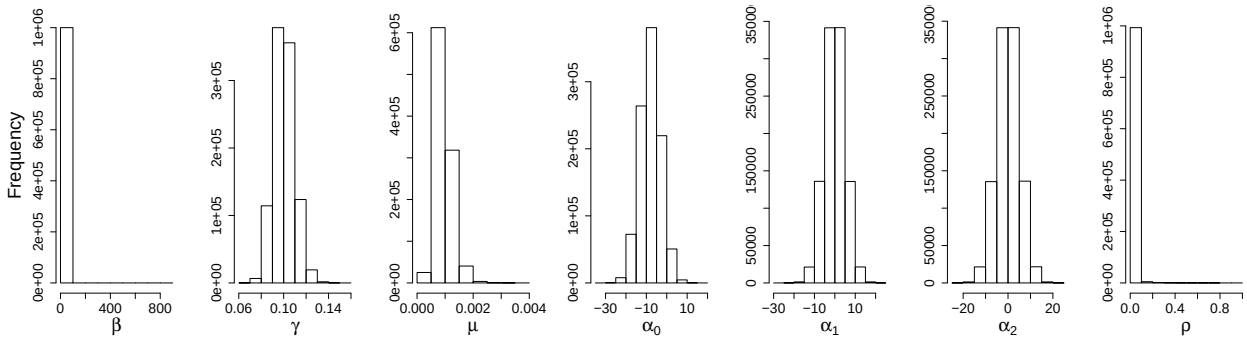


Figure B-5: Prior distributions for the parameters of the SIRS model used in analysis of data from Mathbaria.

| Coefficient | Starting value set 1 | Starting value set 2 | Starting value set 3 |
|---|---|---|---|
| $\beta \times N$ | 0.6 | 0.8 | 0.4 |
| $\gamma$ | 0.11 | 0.1 | 0.12 |
| $\mu$ | 0.0009 | 0.0012 | 0.0006 |
| $\alpha_0$ | $-7.11$ | $-8$ | $-3$ |
| $\alpha_1$ | 0 | 0 | 1 |
| $\alpha_2$ | 0 | 0 | $-1$ |
| $\rho \times N$ | 60 | 6 | 100 |

Table B-1: Initial values used for separate runs of the PMMH algorithm on the data from Mathbaria. We assume $N = 10000$.

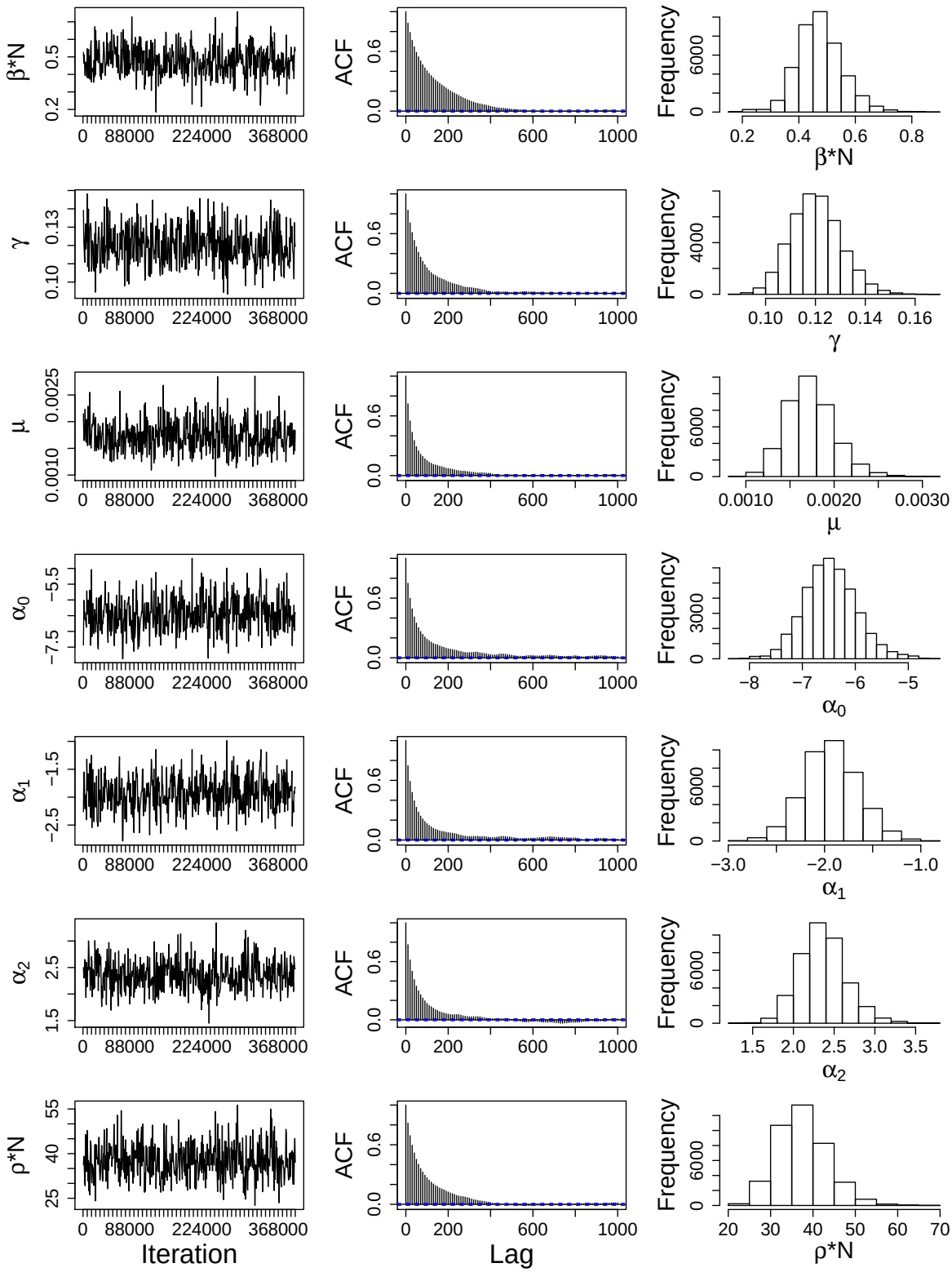


Figure B-6: Summary plots of the PMMH algorithm output (final run of 400000 iterations) for the parameters of the SIRS model, based on data from Mathbaria, Bangladesh. ACF plots and histograms are thinned to 40000 iterations and trace plots are thinned to display only 500 iterations.

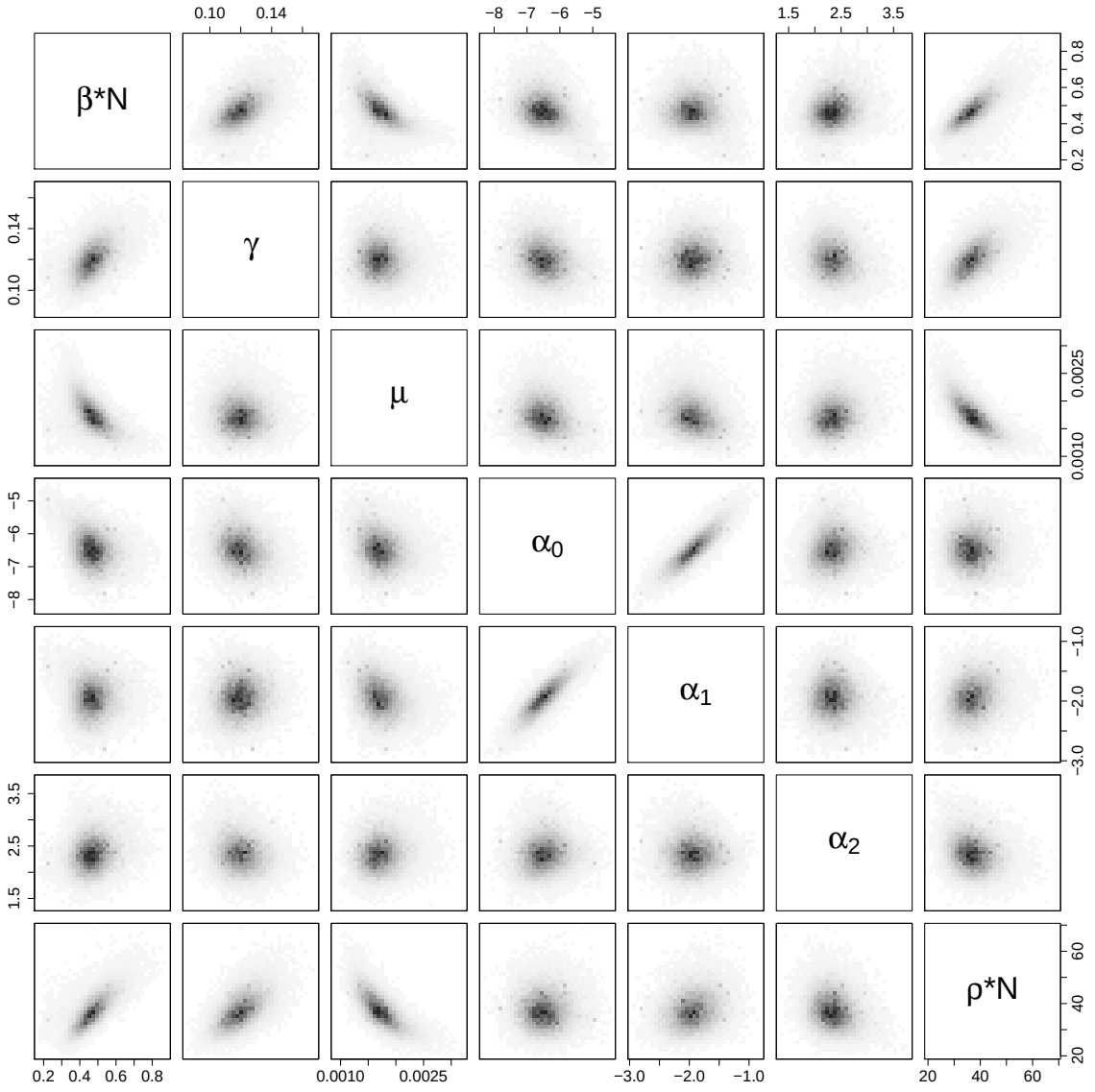


Figure B-7: Smoothed bivariate scatterplots of parameters of the SIRS model estimated using data from Mathbaria.

| | Starting value set 1 | | Starting value set 2 | | Starting value set 3 | |
|---|---|---|---|---|---|---|
| Coefficient | Estimate | 95% CIs | Estimate | 95% CIs | Estimate | 95% CIs |
| $\beta \times N$ | 0.47 | (0.33 , 0.65) | 0.47 | (0.33 , 0.65) | 0.47 | (0.34 , 0.65) |
| $\gamma$ | 0.12 | (0.1 , 0.14) | 0.12 | (0.1 , 0.14) | 0.12 | (0.1 , 0.14) |
| $\mu$ | 0.002 | (0.001, 0.002) | 0.002 | (0.001, 0.002) | 0.002 | (0.001, 0.002) |
| $(\beta \times N)/\gamma$ | 3.92 | (2.84 , 5.19) | 3.96 | (2.82 , 5.13) | 3.95 | (2.93 , 5.11) |
| $\alpha_0$ | -6.49 | (-7.39 , -5.41) | -6.5 | (-7.37 , -5.34) | -6.5 | (-7.37 , -5.46) |
| $\alpha_1$ | -1.94 | (-2.49 , -1.37) | -1.94 | (-2.44 , -1.34) | -1.94 | (-2.46 , -1.39) |
| $\alpha_2$ | 2.35 | (1.85 , 2.98) | 2.35 | (1.84 , 2.95) | 2.36 | (1.84 , 2.96) |
| $\rho \times N$ | 37.1 | (27.2 , 50.2) | 37.3 | (27.5 , 50.8) | 37 | (27.2 , 50.4) |
| | N=10000, K=1000 | | N=5000, K=100 | | N=50000, K=100 | |
| | $\phi_S/N = 0.2$, $\phi_I/N = 0.02$ | | $\phi_S/N = 0.2$, $\phi_I/N = 0.02$ | | $\phi_S/N = 0.2$, $\phi_I/N = 0.02$ | |
| Coefficient | Estimate | 95% CIs | Estimate | 95% CIs | Estimate | 95% CIs |
| $\beta \times N$ | 0.47 | (0.34 , 0.64) | 0.51 | (0.35 , 0.69) | 0.47 | (0.36 , 0.63) |
| $\gamma$ | 0.12 | (0.1 , 0.14) | 0.12 | (0.1 , 0.14) | 0.12 | (0.1 , 0.14) |
| $\mu$ | 0.002 | (0.001, 0.002) | 0.002 | (0.001, 0.002) | 0.002 | (0.001, 0.002) |
| $(\beta \times N)/\gamma$ | 3.95 | (2.92 , 5.09) | 4.17 | (3.03 , 5.41) | 3.93 | (3.08 , 4.99) |
| $\alpha_0$ | -6.5 | (-7.38 , -5.46) | -6.67 | (-7.63 , -5.56) | -6.86 | (-7.59 , -5.98) |
| $\alpha_1$ | -1.94 | (-2.47 , -1.37) | -1.96 | (-2.49 , -1.43) | -2.19 | (-2.67 , -1.67) |
| $\alpha_2$ | 2.35 | (1.83 , 2.94) | 2.26 | (1.71 , 2.86) | 2.43 | (1.9 , 3.09) |
| $\rho \times N$ | 37.1 | (27.4 , 50.2) | 37.7 | (27.6 , 51.8) | 37 | (27.3 , 49.9) |
| | N=10000, K=100 | | N=10000, K=100 | | N=10000, K=100 | |
| | $\phi_S/N = 0.4$, $\phi_I/N = 0.04$ | | $\phi_S/N = 0.1$, $\phi_I/N = 0.01$ | | $\phi_S/N = 0.4$, $\phi_I/N = 0.01$ | |
| Coefficient | Estimate | 95% CIs | Estimate | 95% CIs | Estimate | 95% CIs |
| $\beta \times N$ | 0.29 | (0.01 , 0.43) | 0.78 | (0.51 , 1.11) | 0.3 | (0.09 , 0.39) |
| $\gamma$ | 0.12 | (0.1 , 0.14) | 0.12 | (0.1 , 0.14) | 0.12 | (0.1 , 0.14) |
| $\mu$ | 0.002 | (0.002, 0.004) | 0.001 | (0.001, 0.002) | 0.003 | (0.002, 0.003) |
| $(\beta \times N)/\gamma$ | 2.37 | (0.1 , 3.29) | 6.54 | (4.43 , 8.91) | 2.5 | (0.79 , 3.14) |
| $\alpha_0$ | -5.9 | (-7.06 , -4.65) | -6.66 | (-7.52 , -5.64) | -5.98 | (-7.11 , -4.84) |
| $\alpha_1$ | -1.68 | (-2.36 , -1.13) | -1.96 | (-2.45 , -1.43) | -1.72 | (-2.39 , -1.15) |
| $\alpha_2$ | 2.33 | (1.87 , 2.9) | 2.33 | (1.82 , 2.94) | 2.42 | (1.93 , 3.04) |
| $\rho \times N$ | 26.5 | (19.6 , 35.8) | 57.2 | (39.8 , 80.4) | 25.6 | (19.3 , 34) |

Table B-2: Convergence diagnostics and sensitivity analysis: Posterior medians and 95% equitailed credible intervals (CIs) under different initial values and assumptions for the parameters of the SIRS model estimated using clinical and environmental data sampled from Mathbaria, Bangladesh. The PMMH algorithm is run from different initial values using $N = 10000$, $K = 100$, $\phi_S/N = 0.2$, and $\phi_I/N = 0.02$, and also run using different values for the population size, $N$, the total number of particles, $K$, and the means of the Poisson initial distributions, $\phi_S$ and $\phi_I$. Agreement of parameter estimates and credible intervals across columns within each row indicates convergence of MCMC runs started from different initial conditions to the same distribution (the first block of parameter rows), robustness of parameter estimation to population size misspecification (the second block of parameter rows), and robustness to prior assumptions about the numbers of susceptible and infected individuals at the first observation time point (the third block of parameter rows).

## Appendix C: Model fit

To select a lag for the environmental covariates in the Mathbaria analysis, we compare prediction results from models assuming three different lags: $\kappa = 14$, $\kappa = 18$, and $\kappa = 21$. These are shown in Figures C-1 and C-2. The predictive distributions of the hidden states look similar across lags, so we use the 21 day lag model in order to predict an upcoming epidemic furthest in advance. With a three week lag, we would be able to make predictions three weeks in advance.

Figure C-3 shows plots of standardized residuals versus time for each of the two phases of data collection in Mathbaria, Bangladesh. Standardized residuals are calculated as $\epsilon_{t_i} = (y_{t_i} - \mathrm{E}(y_{t_i}))\,/\mathrm{sd}(y_{t_i})$, where $y_{t_i}$ is the number of observed infections at time $t_i$ for observation $i \in \{0, 1, ..., n\}$. $\mathrm{E}(y_{t_i})$ and $\mathrm{sd}(y_{t_i})$ are approximated via simulation by fixing the model parameters to the posterior medians, running the SIRS model forward 5000 times, and computing the average and sample standard deviation of the 5000 realizations of the case counts at each time point. Residuals are furthest from zero during the epidemic peaks; the inflation of residuals during times of high case counts is probably due to the model being off in terms of the timing of the epidemic peak or the latent states not being predicted correctly.

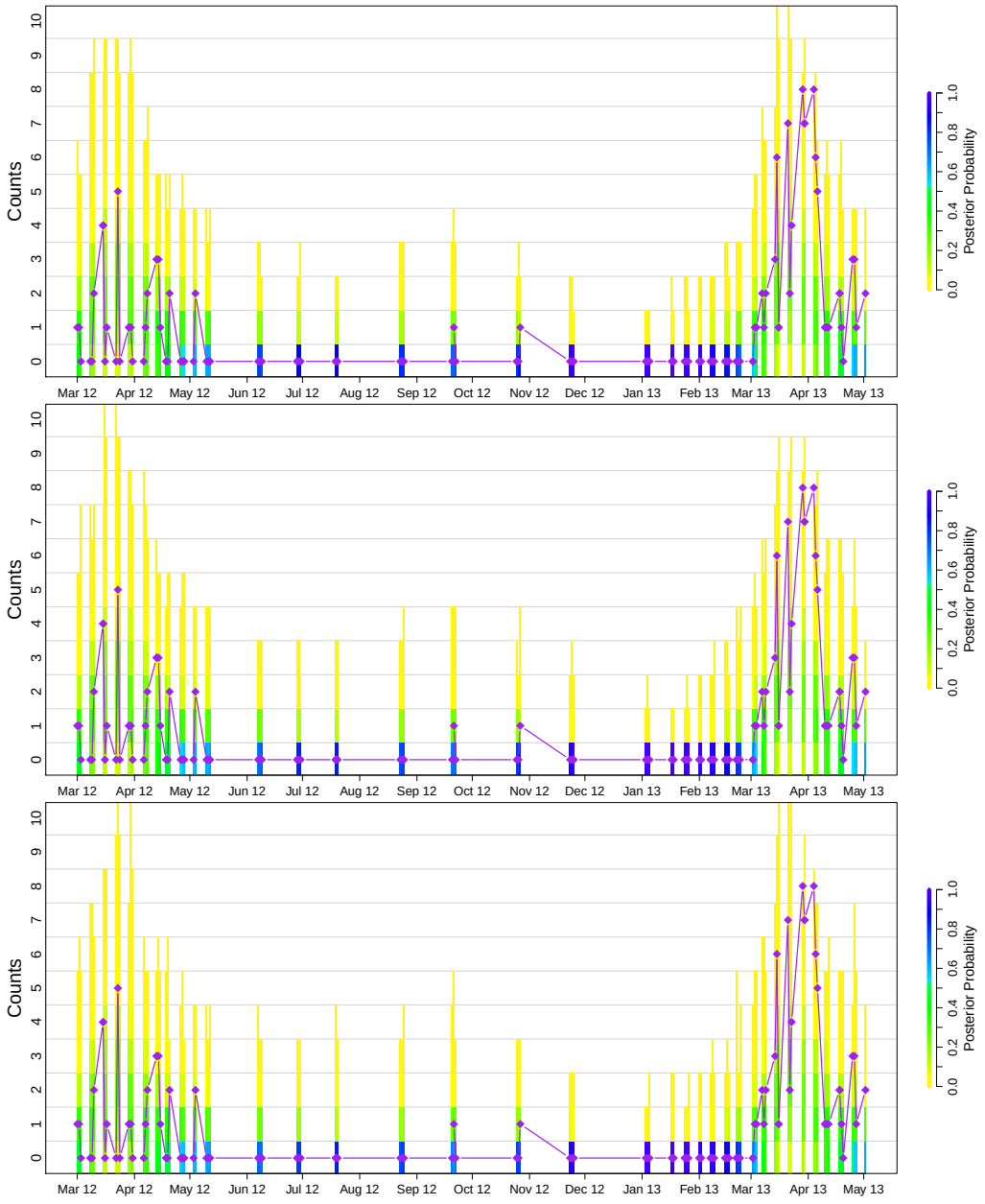


Figure C-1: Distributions of predicted reported cases under models assuming a covariate lag of 14 days (top), 18 days (middle), and 21 days (bottom). The posterior probability of the predicted counts is compared to the test data (diamonds connected by straight line). The coloring of the bars is determined by the frequency of each set of counts in the predicted data for each time point. The distributions are similar, regardless of lag choice.

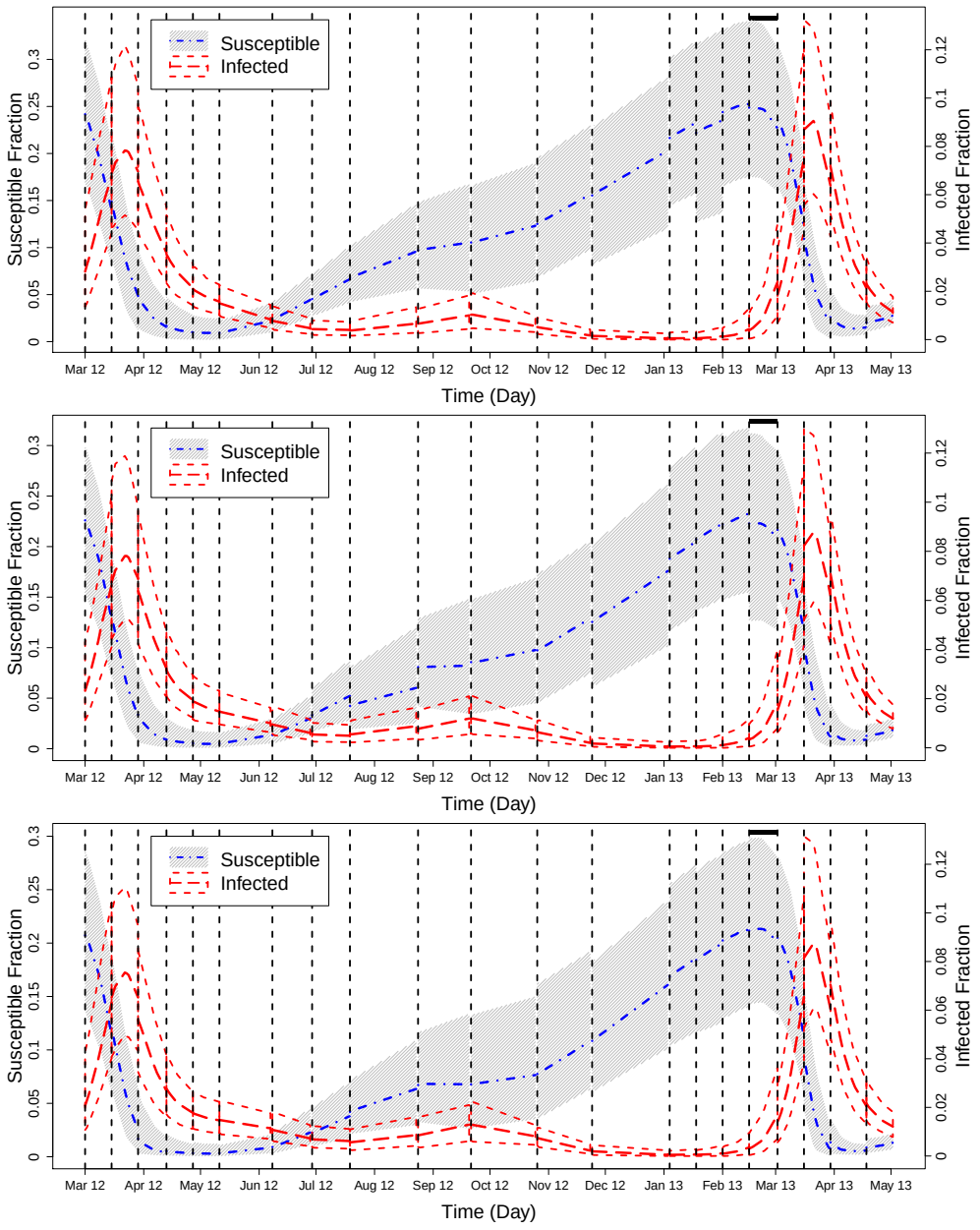


Figure C-2: Predictive distributions of the hidden states, under models assuming a covariate lag of 14 days (top), 18 days (middle), and 21 days (bottom). The gray area and the dot-and-dash line denote the 95% quantiles and median of the predictive distributions for the fraction of susceptibles. The short dashed lines and the long dashed line denote the 95% quantiles and median of the predictive distributions for the fraction of infected individuals. Differences between the distributions under the different lags are difficult to distinguish.

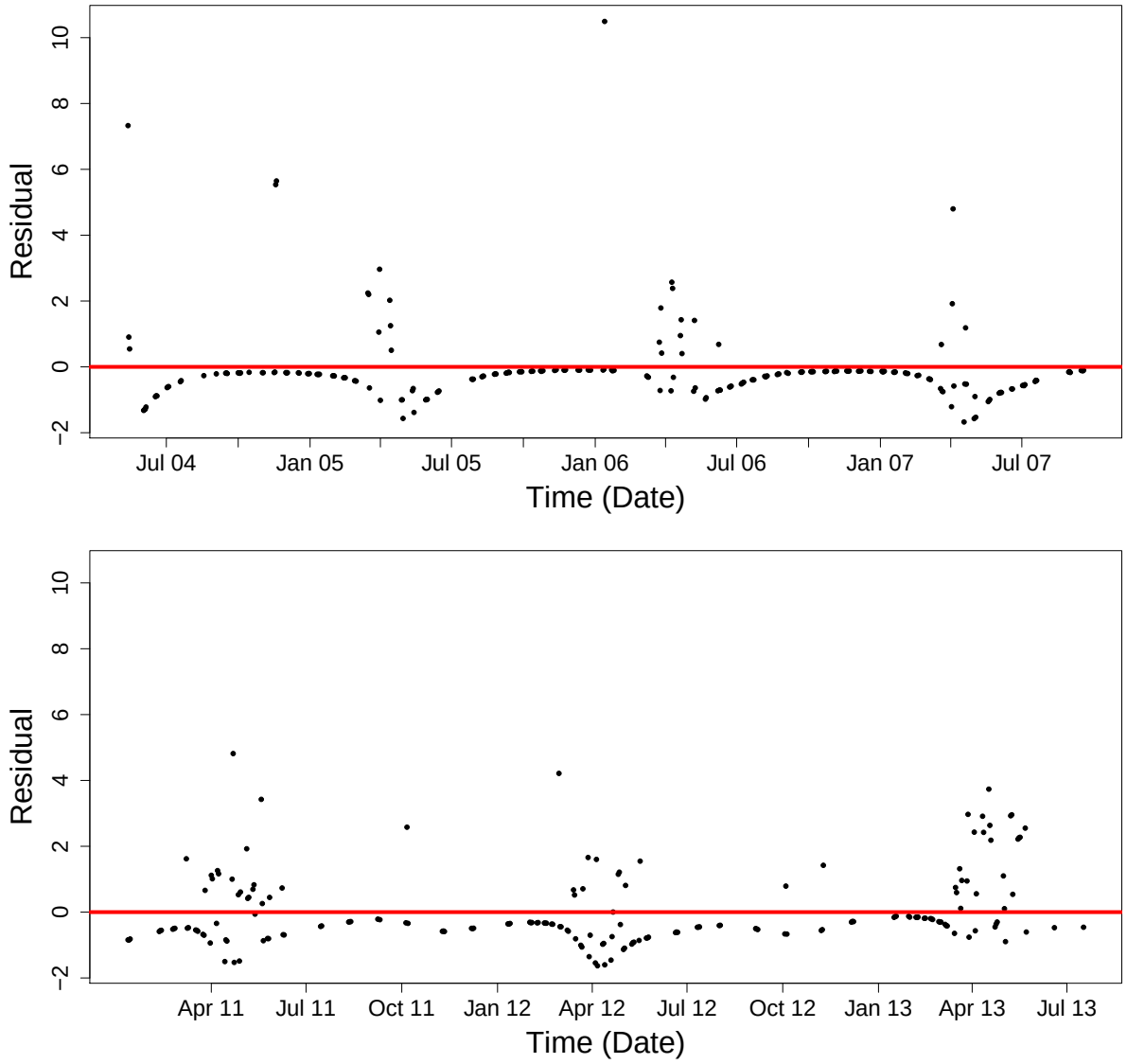


Figure C-3: Plot of standardized residuals versus time. The top figure shows the residuals for the first three years of data collected from Bangladesh, and the bottom figure shows the residuals for the second three years of data collection. The red line is drawn through zero for reference.

## Appendix D: Sensitivity analysis

In our analysis of the data from Mathbaria, we assumed the size of the population, $N = 10000$ and the means of the Poisson initial distributions, $\phi_S = .2 \times N$ and $\phi_I = 0.02 \times N$, are known. We studied sensitivity to these assumptions by setting all of these parameters to different values, and the results are shown in the bottom two-thirds of Table B-2. We report $\beta \times N$ and $\rho \times N$ since we found these parameter estimates to be robust to changes in the population size $N$. As seen in Table B-2, estimates are similar over different values of $N$, $\phi_S$, and $\phi_I$. There does appear to be some sensitivity to assumptions about the means of the Poisson initial distributions $\phi_S$ and $\phi_I$, seen in estimates of $\beta$ and $\rho$.

We also checked sensitivity to the number of particles in the SMC algorithm, using 1000 particles instead of 100. Since it is very computationally expensive to use 1000 particles, we used a final PMMH run of only 200000 iterations. However, effective sample sizes for the parameters using this shorter run look similar to effective sample sizes for the longer run with only 100 particles. As seen in Table B-2, posterior distributions look very similar to analysis with 100 particles.

We tested the effect of incorrect values for $\phi_S$ and $\phi_I$ on prediction using simulated data, as seen in Figure D-1. Values for $\phi_S/N$ and $\phi_I/N$ are set above the true values, (0.39 and 0.0168), at the truth (0.29 and 0.0084), below the true values (0.19 and 0.0042), or further below the true values (0.095 and 0.0021). Predicted distributions look similar for all values of $\phi_S$ and $\phi_I$. Uncertainty is greatest when $\phi_S$ and $\phi_I$ are set at higher values than the truth. For the lowest values of $\phi_S$ and $\phi_I$, the fraction of susceptible individuals is lower and the fraction of infected is higher than those predicted fractions under other settings. However, important information, like the timing of the epidemic, remains intact.

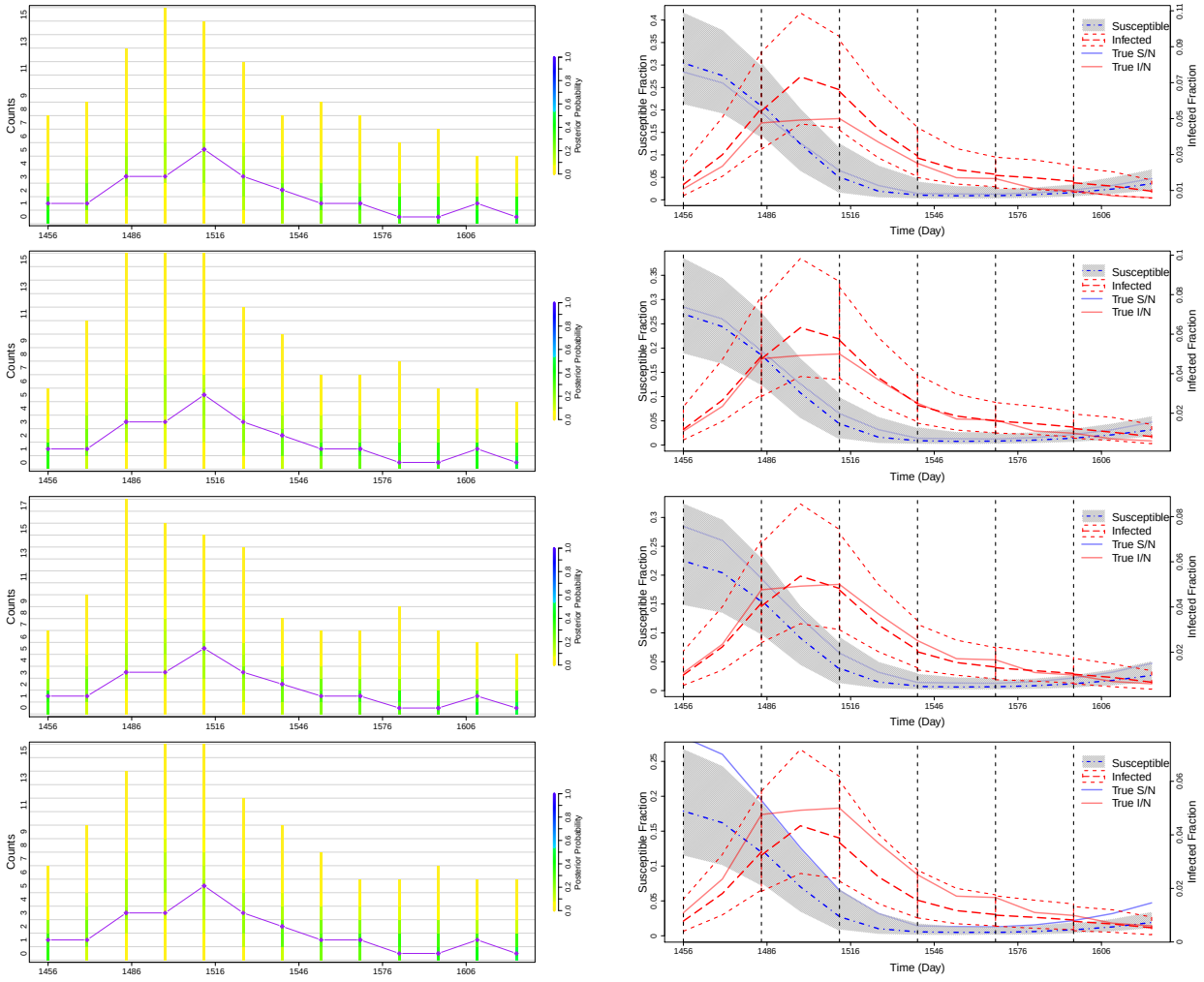


Figure D-1: Summary of prediction results for simulated data; each row shows prediction results under different assumptions about the values of $\phi_s/N$ and $\phi_I/N$. From top to bottom, values for $\phi_S/N$ and $\phi_I/N$ are set above the true values (0.39 and 0.0168), at the true values (0.29 and 0.0084), below the true values (0.19 and 0.0042), or further below the true values (0.095 and 0.0021). Plots on the left compare the posterior probability of the predicted counts to the test data (diamonds connected by straight line). The coloring of the bars is determined by the frequency of each set of counts in the predicted data for each time point. The plots on the right show how the trajectory of the predicted hidden states change over the course of the epidemic. The gray area and the dot-and-dash line denote the 95% quantiles and median of the predictive distribution for the fraction of susceptibles. The short dashed lines and the long dashed line denote the 95% quantiles and median of the predictive distribution for the fraction of infected individuals. The solid blue and red lines denote the true simulated fraction of susceptible and infected individuals.

## Appendix E: Prediction

We compare SIRS predictive distributions to predictions made from a lagged quasi-Poisson regression model, similar to the one used by Huq et al. [2005]. For the two predictors, water temperature (WT) and water depth (WD), we have

$$\ln \mathrm{E}(Y_t|C_{WD}(t-\kappa), C_{WT}(t-\kappa)) = \beta_0 + \beta_1 C_{WD}(t-\kappa) + \beta_2 C_{WT}(t-\kappa),$$

where $\kappa = 21$ days. The quasi-Poisson model accounts for overdispersion in the data [McCullagh and Nelder, 1989]. Figure E-1 shows the predicted means and 95% intervals under the quasi-Poisson model. Test data are again cut off at different points during the 2012 and 2013 epidemic peaks and predictions are run until the next cut off point, with cut off points chosen approximately every two weeks. Predicted mean number of reported cases and 95% intervals from the hidden SIRS model are also shown for comparison. To calculate these, we sample 1000 sets of parameter values from the posterior. For each set of parameters, we simulate data forward until the next cut off point 100 times and then calculate the mean of the predicted counts at each observation time. Using these 1000 means from the 1000 parameter sets, we calculate the overall predicted means and 95% intervals. Both models predict well the timing of epidemic peaks. However, the quasi-Poisson regression framework does not provide any information about the underlying fraction of infected individuals in the population, which may be important for resource allocation.

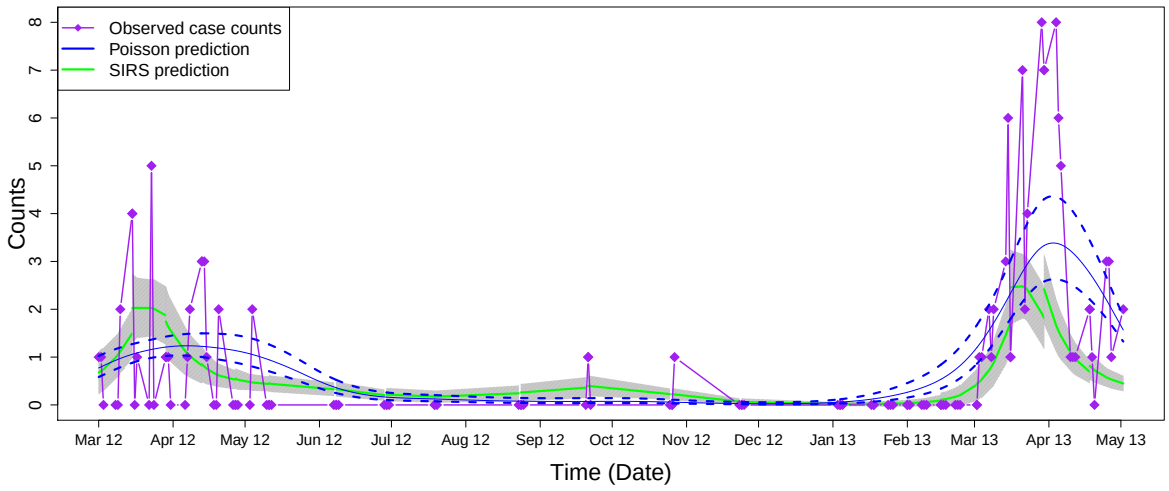


Figure E-1: Comparison of predicted means for number of reported cases. The solid blue lines and the dashed blue lines denote the predicted means and 95% intervals under the quasi-Poisson model. The green line and gray area denote the predicted means and 95% intervals under the SIRS model. Predictions are started and stopped using identical cut-off points for the training and test data to those in Figure 5. Test data are denoted by the purple diamonds connected by straight lines.

## Appendix F: Routes of transmission

An important question in cholera modeling is: what is the relative contribution of different routes of transmission at different points of the epidemic? We hypothesized that environmental forces trigger the seasonal cholera epidemics and that infectious contact between susceptible and infected individuals drives the epidemics. To examine this possible dynamic, we compare the forces of infection from the environment, $\alpha(t)$, to that from infected individuals, $\beta \times I_t$, over time. Values are computed by sampling 5000 sets of parameter values from the posterior. For each set of parameters, we generate data using our hidden SIRS model. Figure F-1 shows median and 95% quantiles for $\alpha(t)$ vs $\beta \times I_t$ plotted over time. The median values of $\alpha(t)$ are almost always higher than values of $\beta \times I_t$, except at the very beginning of the seasonal outbreaks when both forces of infection are small. However, when $I_t$ is largest, the posterior median for $\alpha(t)$ is larger than the posterior median for $\beta \times I_t$. This supports the hypothesis that the epidemics are driven by the environmental force of infection.

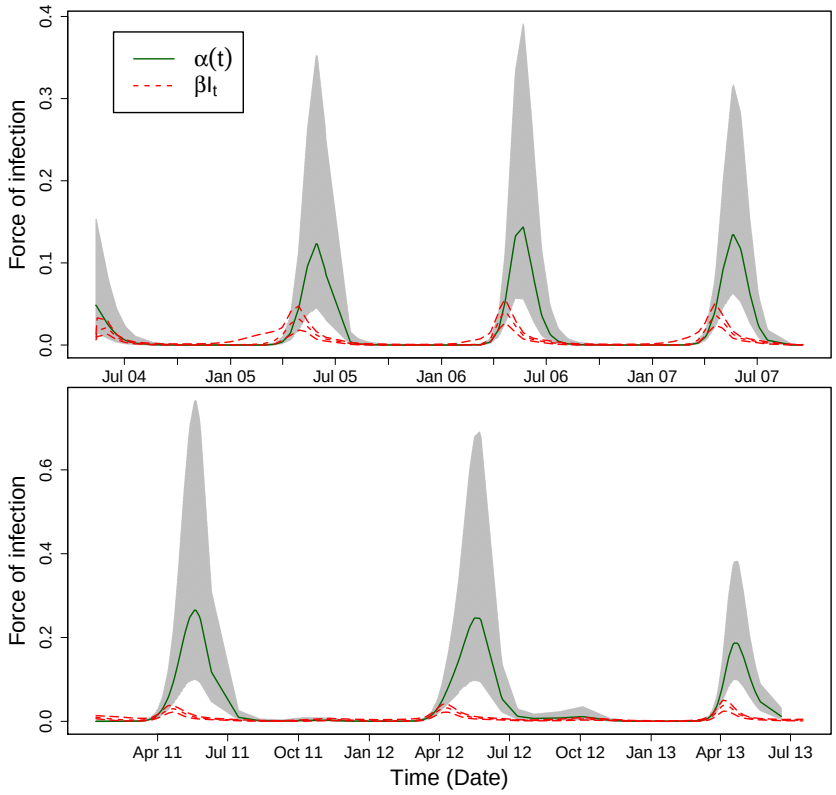


Figure F-1: The relative contribution of different routes of transmission at different points of the epidemic curves. The gray area and the solid line denote the 95% quantiles and median of the force of infection from the environment, $\alpha(t)$. The long dashed lines and the short dashed line denote the 95% quantiles and median of the force of infection from infected individuals, $\beta \times I_t$.

## Appendix G: Bayesian analysis using SIRS model on data simulated from an SIWR model

In the hidden SIRS model, we assume that the hazard rate of infection is $\beta I_t + \alpha(t)$ for each time $t$, where $\beta$ is the infectious contact rate between infected individuals and susceptible individuals and $\alpha(t)$ is the time-varying environmental force of infection. In that model, infectious contact incorporates both direct person-to-person transmission of cholera and consumption of contaminated water. We now separate these contributions to transmission from infected individuals and explore models which incorporate a feedback loop from the infected individuals back into the environment to capture the effect of infected individuals excreting *V. cholerae* into the environment.

To accomplish this, we add a water compartment, $W$, that quantifies the concentration of *V. cholerae* in the environment. Instead of using an environmental force of infection, we incorporate the environmental covariates using the same function, $\alpha(t)$, as the rate of seasonal increase in water *V. cholerae* concentration. This SIWR model is similar to the SIWR model of Tien and Earn [2010] and Eisenberg et al. [2013], but unique in the way it incorporates the environmental covariates. The hazard rate of infection is $\beta_I I_t + \beta_W W_t$ for each time $t$, where $\beta_I$ represents the infectious contact rate between infected individuals and susceptible individuals and $\beta_W$ represents force of infection from contact with or consumption of contaminated water. Infected individuals excrete *V. cholerae* into the environment/water compartment at rate $\kappa$. The time-varying function $\alpha(t)$ also contributes to the increase of the *V. cholerae* concentration in the water compartment. This concentration decays at rate $\eta$. Again, infected individuals recover from infection at rate $\gamma$, and recovered individuals lose immunity to infection and become susceptible at rate $\mu$.

We model $\boldsymbol{X}_t = (S_t, I_t, R_t, W_t)$ as an inhomogeneous Markov process [Taylor and Karlin, 1998]

with infinitesimal rates

$$\lambda_{(S,I,R,W),(S',I',R',W')}(t) = \begin{cases} (\beta_I I + \beta_W W)\, S & \text{if } S' = S-1,\ I' = I+1,\ R' = R,\ W' = W, \\ \gamma I & \text{if } S' = S,\ I' = I-1,\ R' = R+1,\ W' = W, \\ \mu R & \text{if } S' = S+1,\ I' = I,\ R' = R-1,\ W' = W, \\ \kappa I + \alpha(t) & \text{if } S' = S,\ I' = I,\ R' = R,\ W' = W+1, \\ \eta W & \text{if } S' = S,\ I' = I,\ R' = R,\ W' = W-1, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\boldsymbol{X} = (S, I, R, W)$ is the current state and $\boldsymbol{X}' = (S', I', R', W')$ is a new state. We do not keep track of the number of recovered individuals because $R_t = N - S_t - I_t$.

The water compartment has no scale; it is used to quantify water contamination but may not necessarily be the exact amount of *V. cholerae* in the water. In fact, for a constant $c$ the dynamics of the process are invariant if one makes the change of variables $c\beta_W$ and $(\kappa I + \alpha(t))/c$. This also controls the range of $W$, which can speed up simulation.

Again, we assume that $\boldsymbol{X}_t = (S_t, I_t, R_t, W_t)$ is not directly observable. Instead, we only observe $y_t$, the number of observed infections at time $t$, and assume $y_t$ has a binomial distribution with size $I_t$, the number of infected individuals at time $t$, and success probability $\rho$, the probability of infected individuals seeking treatment; thus, $y_{t_i}|\boldsymbol{X}_{t_i} = (S_{t_i}, I_{t_i}, R_{t_i}, W_{t_i}), \rho \sim \text{Binomial}(I_{t_i}, \rho)$.

We simulate from the hidden SIWR model using a population size of $N = 10000$ and assume independent Poisson initial distributions for $S_{t_0}$, $I_{t_0}$, and $W_{t_0}$, with means $\phi_S = 3450$, $\phi_I = 6$, and $\phi_W = 12$. The other parameters are set at $\beta_I = 1.072 \times 10^{-5}$, $\beta_W = 7 \times 10^{-6}$, $\alpha_0 \approx 0.39$, $\alpha_1 = 2$, $\gamma = 0.12$, $\kappa = 0.02$, $\eta = 1/30 \approx 0.03$, and $\mu = 0.0018$. Rates are measured in the number of events per day. We use the daily time intervals $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n - 1\}$ and define

$\alpha(t) = \alpha_{A_i}$ for $t \in A_i$ where $\alpha_{A_i} = \exp[\alpha_0 + \alpha_1 C(i)]$ and

$$C(i) = \begin{cases} 2.1 \sin(2\pi i/365) & \text{if } 0 \leq i \leq 365, \\ 1.8 \sin(2\pi i/365) & \text{if } 365 < i \leq 730, \\ 2 \sin(2\pi i/365) & \text{if } 730 < i \leq 1095, \\ 2.2 \sin(2\pi i/365) & \text{if } 1095 < i \leq 1460, \\ 2 \sin(2\pi i/365) & \text{if } i > 1460. \end{cases}$$

Using the modified Gillespie algorithm to simulate from the SIWR model, as described in Appendix A, the resulting $(S_t, I_t, W_t)$ chain is given in Figure G-1. From the hidden data, we simulate the observed number of infections, $y_t$, as $y_t \sim \text{Binomial}(I_t, \rho)$, where $\rho = 0.016$. Here case observations occur once every two weeks.

We use this simulated data to study the effects of misspecification of the data generating process on estimation and prediction. Using the data simulated under the SIWR model, we implement the PMMH algorithm under the assumption of a hidden SIRS model. We test the effects on parameter estimation and prediction when the model that we use to fit the data does not match the data generating process. We assume the same settings for the PMMH algorithm and prior distributions for the parameters that we used in the simulated data example for the SIRS analysis in Section 5.

Posterior medians and 95% equitailed credible intervals for the parameters of the SIRS model estimated using data simulated from an SIWR model are shown in Table G-1, along with true values for parameters that are comparable between the two models. Figure G-2 shows summary plots of the PMMH algorithm output. The credible intervals for $\gamma$, $\mu$, and $\rho \times N$ do not include the true values of these parameters. Surprisingly, the true value of $\alpha_1 = 2$ for the SIWR model is in the credible interval for the parameter $\alpha_1$ from the SIRS model. This means that the covariate effect is still being captured in this simulation, regardless of how we model the relationship of environmental covariates to cholera outbreaks. The posterior estimate of $\alpha_0$ seems to compensate for the model misspecification, as the true value of $\alpha_0 = 0.39$ for the SIWR simulated data is outside the credible interval for the SIRS $\alpha_0$ parameter. The chains are mixing well, as seen in the trace plots in Figure

G-2.

Table G-1: Posterior medians and 95% equitailed credible intervals for the parameters of the SIRS model estimated using data simulated from an SIWR model. True values for parameters that are comparable between the two models are shown.

| SIRS Parameter | SIWR True value | Estimate | 95% CIs |
|---|---|---|---|
| $\beta \times N$ | — | 0.26 | (0.18 , 0.36) |
| $\gamma$ | 0.12 | 0.09 | (0.08 , 0.11) |
| $\mu$ | 0.0018 | 0.0008 | (0.0005, 0.0012) |
| $\alpha_0$ | — | -10.02 | (-11.66 , -8.99) |
| $\alpha_1$ | — | 1.95 | (1.41 , 2.82) |
| $\rho \times N$ | 160 | 266.9 | (169.8 , 431.7) |

We test the predictive ability of the SIRS model on data generated from an SIWR model using staggered training and test sets of data, as described in Section 5.1. Despite the model misspecification, the model predictions look good, as seen in Figure G-4. The general trend and important features of the epidemic curve are captured. We find that we can still predict outbreaks well when we fit the parameters of the SIRS model to data generated by the SIWR model. This suggests that the predictions made with the SIRS model using the data from Mathbaria in Section 6.1 may not be too far off, even if the model is not biologically realistic.

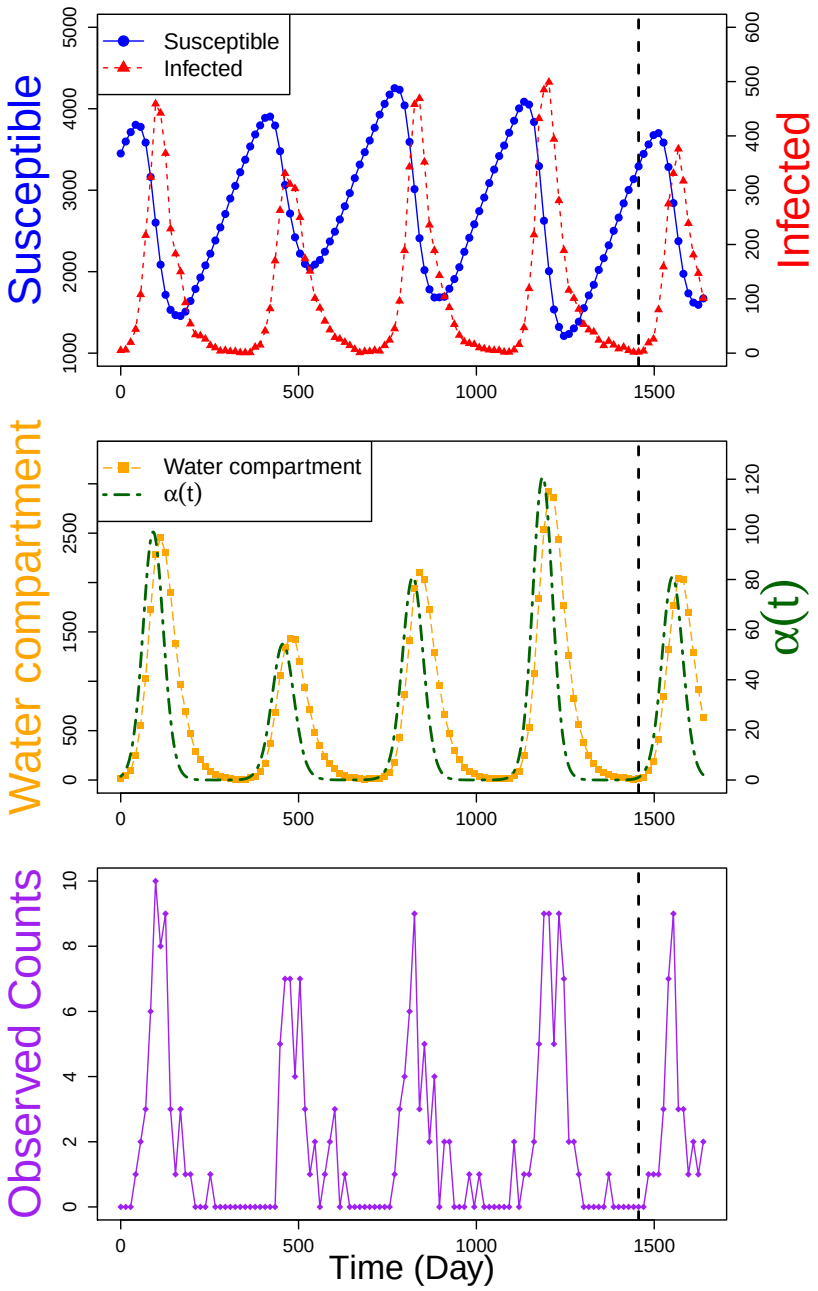


Figure G-1: Plots of simulated hidden states (counts of susceptible individuals, $S_t$, infected individuals, $I_t$, and the water compartment, $W_t$) and the observed data ($\alpha(t)$ and $y_t \sim \text{Binomial}(I_t, \rho) =$ number of observed infections) plotted over time ($t$). The top plot shows the dynamics of the number of susceptible and infected individuals over time. The middle plot shows the dynamics of the water compartment and rate of seasonal increase in water *V. cholerae* concentration, $\alpha(t)$. There is a slight delay in the increase in the water compartment after $\alpha(t)$ increases, and the decay of the water compartment is more gradual than the decrease in $\alpha(t)$. The dashed vertical black lines represent cut offs between the training sets and test data.

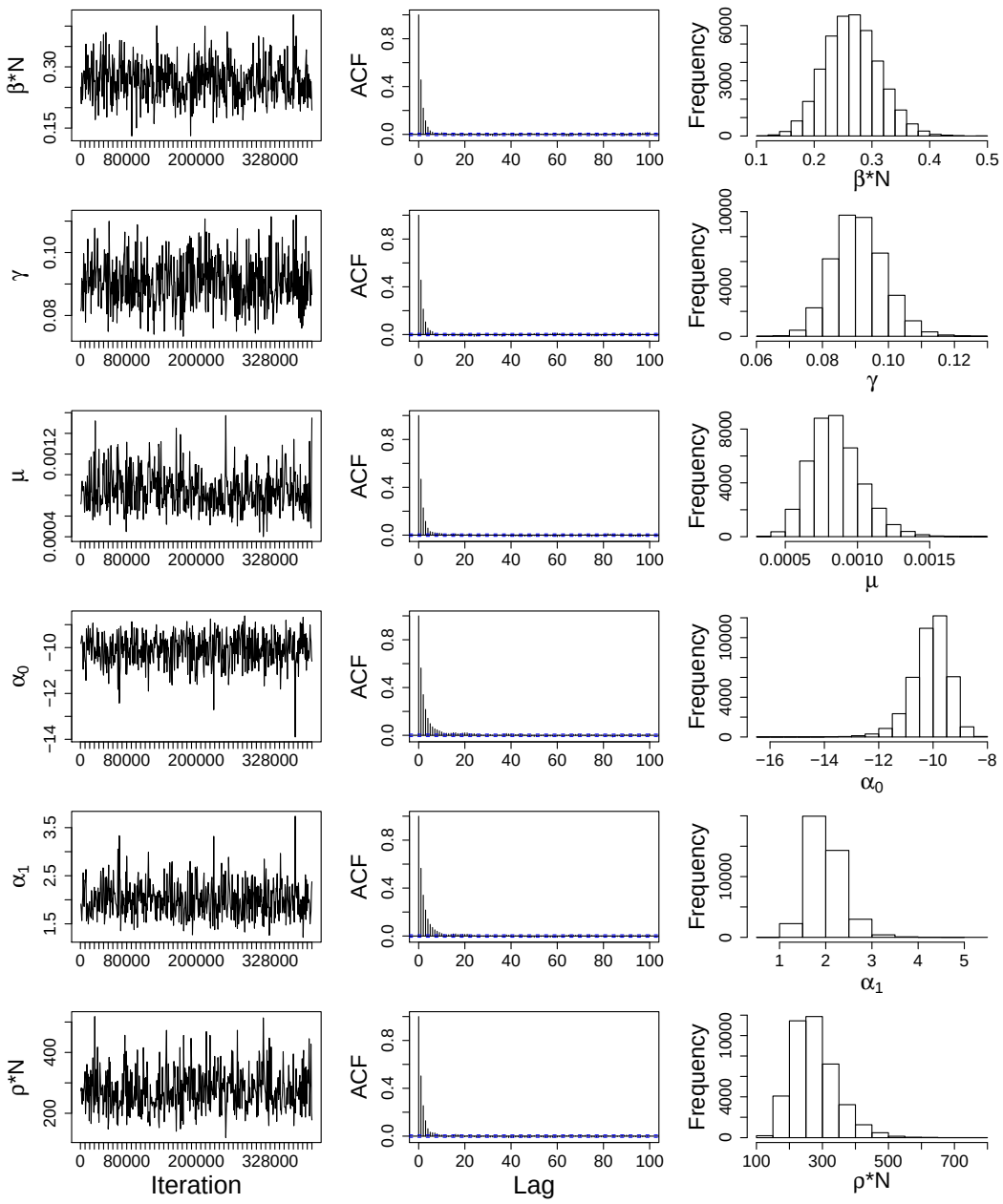


Figure G-2: Summary plots of the PMMH algorithm output (final run of 400000 iterations) for the parameters of the SIRS model estimated using data simulated from an SIWR model. Trace plots are thinned to display only 500 iterations; autocorrelation plots and posterior histograms are thinned to display 40000 iterations.

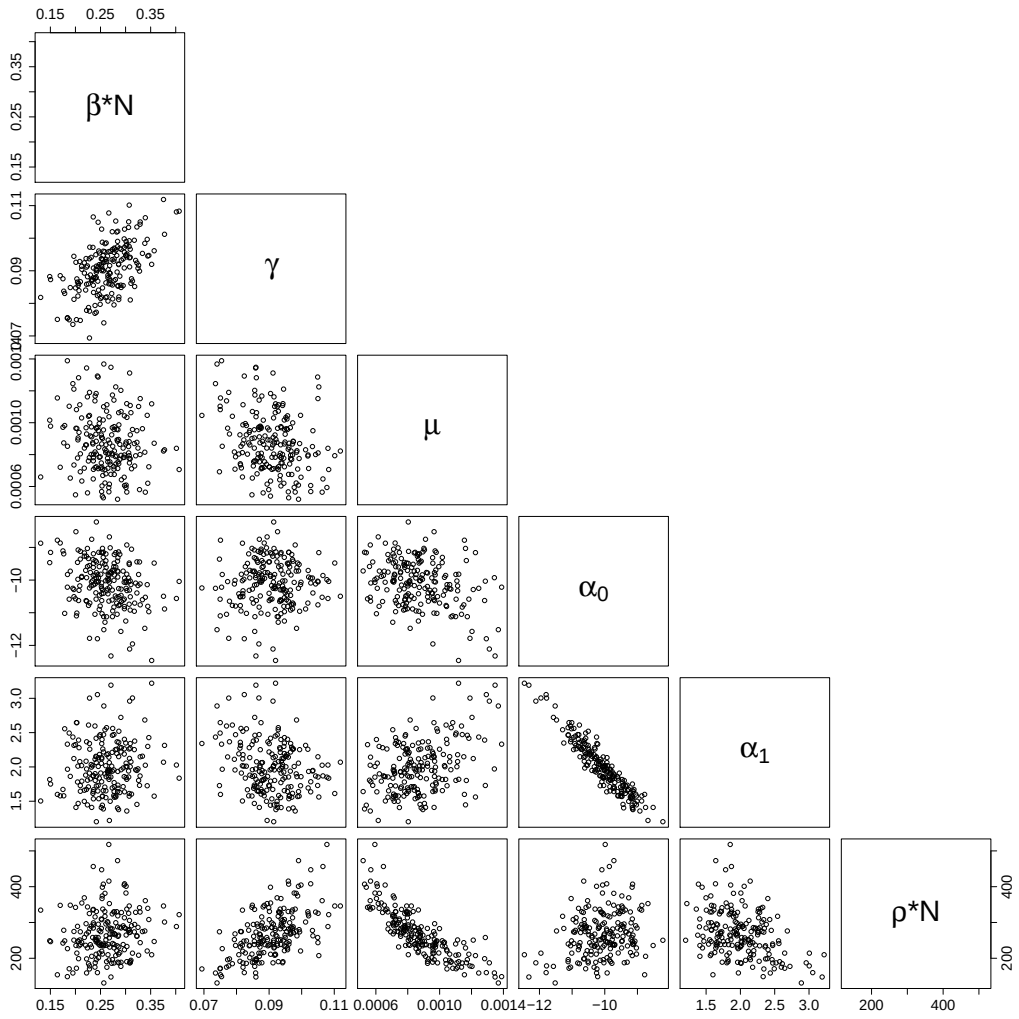


Figure G-3: Bivariate scatterplots of parameters of the SIRS model estimated using data simulated from an SIWR model. Scatterplots are thinned to display only 200 samples, so only every 2000th sample from the posterior distribution is plotted.

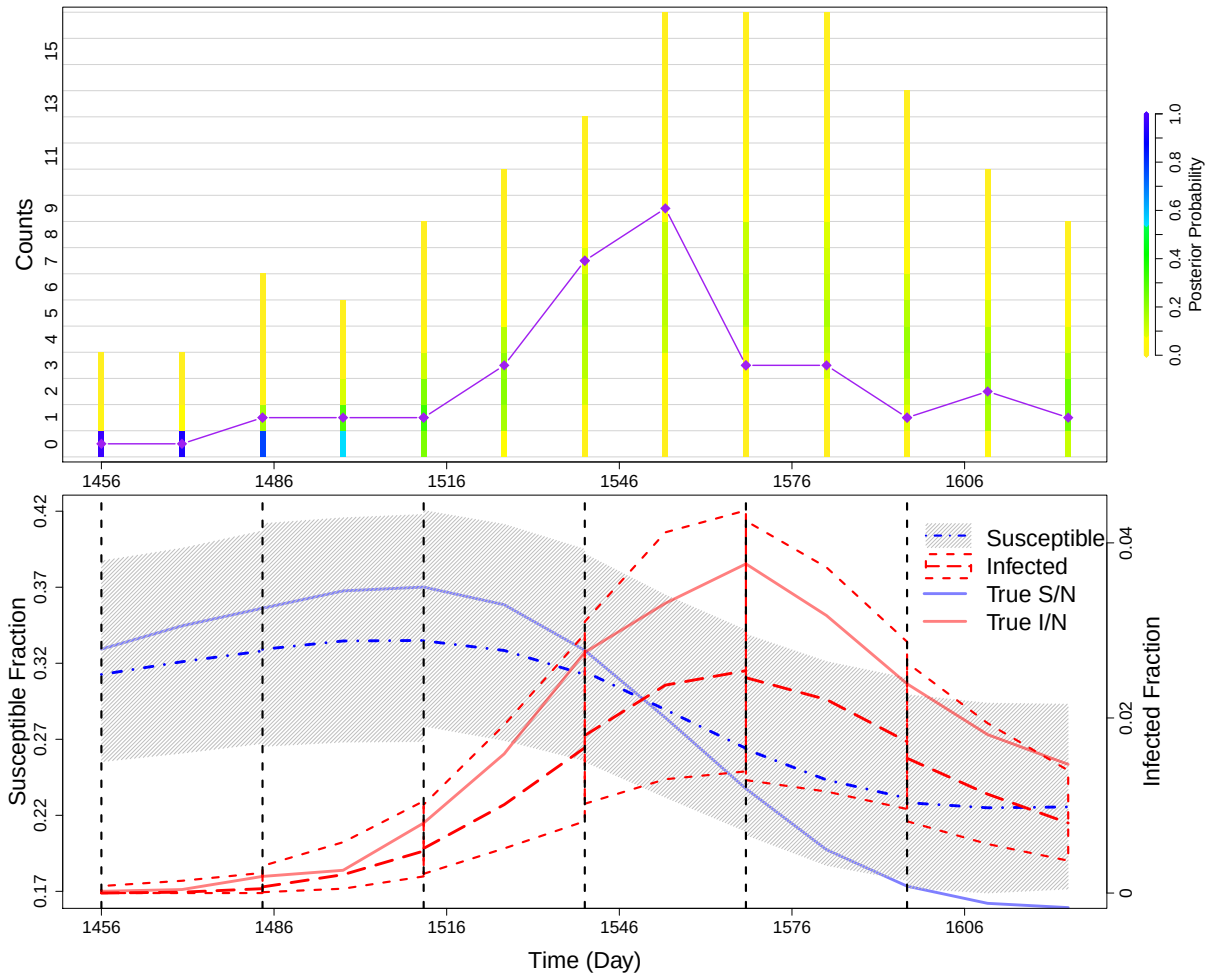


Figure G-4: Summary of prediction results for data simulated from an SIWR model. We approximate the posterior distribution of the parameters of the hidden SIRS model using training sets of the data, which are cut off at each of the dashed black lines in the bottom plot, and future cases are predicted until the next cut off. The top plot compares the posterior probability of the predicted counts to the test data (purple diamonds and line), and the bottom plot shows how the trajectory of the predicted hidden states changes over the course of the epidemic.

## References


D. F. Anderson. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *Journal of Chemical Physics*, 127(21):214107, 2007.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

Y. Cao, D. T. Gillespie, and L. R. Petzold. Avoiding negative populations in explicit Poisson tau-leaping. *The Journal of Chemical Physics*, 123(5):054104, 2005.

M. C. Eisenberg, S. L. Robertson, and J. H. Tien. Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. *Journal of Theoretical Biology*, 324 (0):84–102, 2013.

M. A. Gibson and J. Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The Journal of Physical Chemistry A*, 104(9):1876–1889, 2000.

D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

A. Huq, R. B. Sack, A. Nizam, I. M. Longini, G. B. Nair, A. Ali, J. G. Morris Jr, M. N. Khan, A. K. Siddique, M. Yunus, M. J. Albert, D. A. Sack, and R. R. Colwell. Critical factors influencing the occurrence of *Vibrio cholerae* in the environment of Bangladesh. *Applied and Environmental Microbiology*, 71(8):4645–4654, 2005.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, 1989.

H. M. Taylor and S. Karlin. *An Introduction to Stochastic Modeling*. Academic Press, 3rd edition, 1998.

J. H. Tien and D. J. D. Earn. Multiple transmission pathways and disease dynamics in a waterborne pathogen model. *Bulletin of Mathematical Biology*, 72(6):1506–1533, 2010.

D. J. Wilkinson. *Stochastic Modelling for Systems Biology*. CRC press, 2nd edition, 2011.