Supplementary Materials for "Semi-parametric modeling of SARS-CoV-2 transmission using tests, cases, deaths, and seroprevalence data"

S-1 Simulation study

We performed a simulation study on 200 datasets to validate our models. We use the same prior distributions for the parameters as in the main text. These distributions are presented in Table S-1-.2. We purposely chose parameter values that resulted in data similar to the Orange Country data used in the main text. Exact values for these parameters are presented in Table S-1.2. One of the 200 simulated datasets is presented in Figure S-1.1. Figures S-1.2–S-1.4 present the prior and posterior distribution for this single dataset. Figures S-1.5–S-1.8 show coverage and contraction properties for the whole simulation study. Contraction is calculated as one minus the ratio of standard deviation of the posterior and the prior. Commentary on these results is presented in Section 3 of the main text.

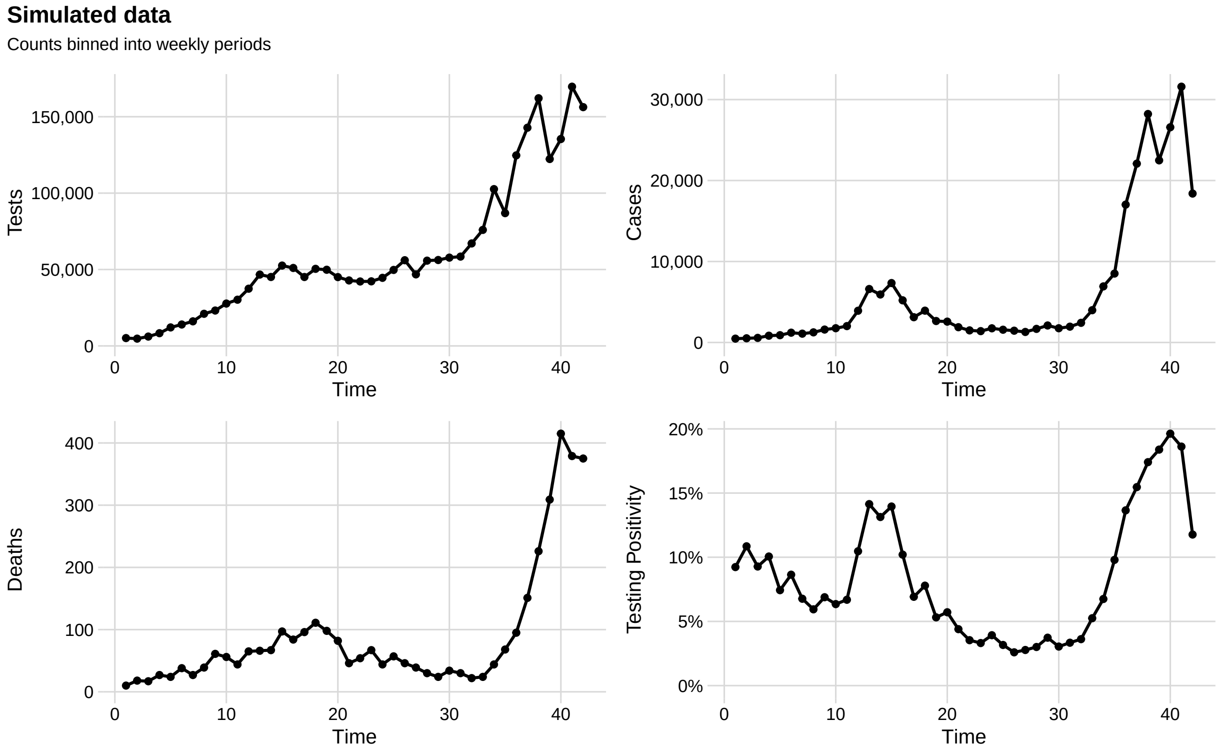


Figure S-1.1: Simulated data. The figure shows weekly counts of tests, cases (positive tests), reported deaths due to COVID-19, as well testing positivity.

Table S-1.1: Model parameters and their prior distributions.

Parameter	**Interpretation**	**Prior**	**Prior Median (95% Interval)**	**Source**
S_0	Initial susceptible proportion	Logit-Normal(6, 0.25)	0.998 (0.993, 0.999)	
$\tilde{I}_0$	Initial proportion of non-susceptibles who are infectious	Logit-Normal(0.6, 0.0009)	0.646 (0.632, 0.659)	
$\exp\left(\tilde{R}_{0,1}\right)$	Initial basic reproduction number	Log-Normal(0, 0.0625)	1.000 (0.613, 1.630)	
$1/\gamma$	Mean latent period (weeks)	Log-Normal(-0.25, 0.01)	0.779 (0.640, 0.947)	Xin et al. (2021)
$1/\nu$	Mean infectious period (weeks)	Log-Normal(0.15, 0.01)	1.160 (0.955, 1.410)	Byrne et al. (2020)
$\text{expit}\,(\tilde{\eta}_1)$	Initial infection fatality ratio	Logit-Normal(-5.3, 0.04)	0.00497 (0.00336, 0.00733)	Bruckner et al. (2021)
ρ^D	Mean death detection rate	Logit-Normal(2.3, 0.04)	0.909 (0.871, 0.937)	Bruckner et al. (2021)
ϕ_D	over-dispersion in observed deaths Negative-Binomial model	Log-Normal(4.16, 0.293)	63.9 (22.1, 185.0)	
$\exp\,(\tilde{\alpha}_1)$	Initial proportion in proportional odds test positivity model	Log-Normal(1.35, 0.0121)	3.86 (3.11, 4.79)	
ϕ_C	over-dispersion in observed cases beta-binomial model	Log-Normal(6.5, 0.0673)	665 (400, 1110)	
σ_{R_0}	Standard deviation of log-Guassian Markov random field for time-varying R_0	Log-Normal(-1.9, 0.09)	0.1500 (0.0831, 0.2690)	
σ_η	Standard deviation of logit-Guassian Markov random field for time-varying η	Log-Normal(-2.4, 0.0144)	0.0907 (0.0717, 0.1150)	
σ_α	Standard deviation of log-Guassian Markov random field for time-varying α	Log-Normal(-2.7, 0.0225)	0.0672 (0.0501, 0.0902)	
$\tilde{\rho}_1^Y$	Initial case detection rate	Logit-Normal(-2.5, 0.01)	0.0759 (0.0632, 0.0908)	Bruckner et al. (2021)
ϕ_Y	over-dispersion in observed cases Negative-Binomial model	Log-Normal(3.93, 0.0684)	51.1 (30.6, 85.3)	
σ_{ρ^Y}	Standard deviation of logit-Guassian Markov random field for time-varying ρ^Y	Log-Normal(-2.2, 0.04)	0.1110 (0.0749, 0.1640)	

Table S-1.2: Simulation parameters.

Parameter	**Interpretation**	**Value**
S_0	Initial susceptible proportion	0.9979
$\tilde{I}_0$	Initial proportion of non-susceptibles who are infectious	0.6455
$\exp\left(\tilde{R}_{0,1}\right)$	Initial basic reproduction number	1.2602
$1/\gamma$	Mean latent period (weeks)	0.7697
$1/\nu$	Mean infectious period (weeks)	1.1997
$\text{expit}\,(\tilde{\eta}_1)$	Initial infection fatality ratio	0.0005
ρ^D	Mean death detection rate	0.9061
ϕ_D	over-dispersion in observed deaths Negative-Binomial model	87.2776
$\exp\,(\tilde{\alpha}_1)$	Initial proportion in proportional odds test positivity model	4.3958
ϕ_C	over-dispersion in observed cases Beta-Binomial model	1,026.6765
σ_{R_0}	Standard deviation of log-Guassian Markov random field for time-varying R_0	0.1481
σ_η	Standard deviation of logit-Guassian Markov random field for time-varying η	0.0944
σ_α	Standard deviation of log-Guassian Markov random field for time-varying α	0.0696

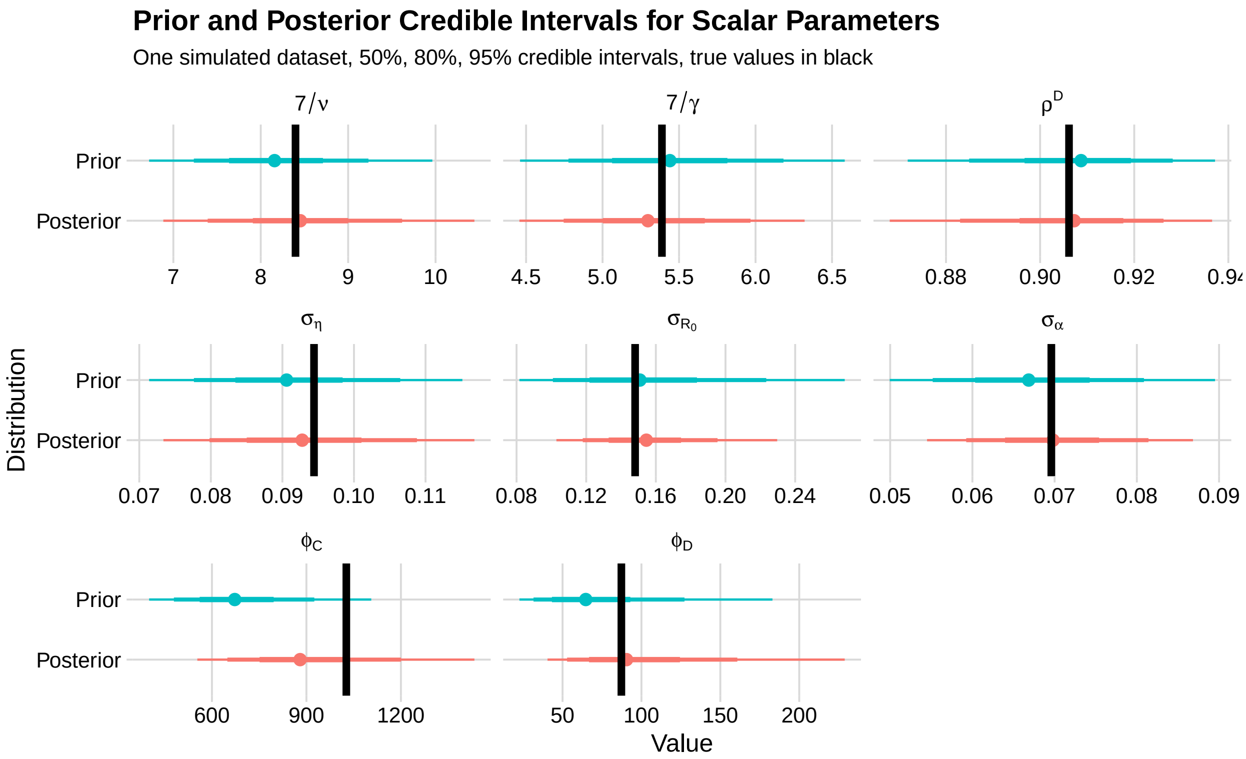


Figure S-1.2: Prior and posterior credible intervals for scalar parameters for a model fit to the dataset presented in Figure S-1.1. True values for the simulated parameters are indicated by solid black lines.

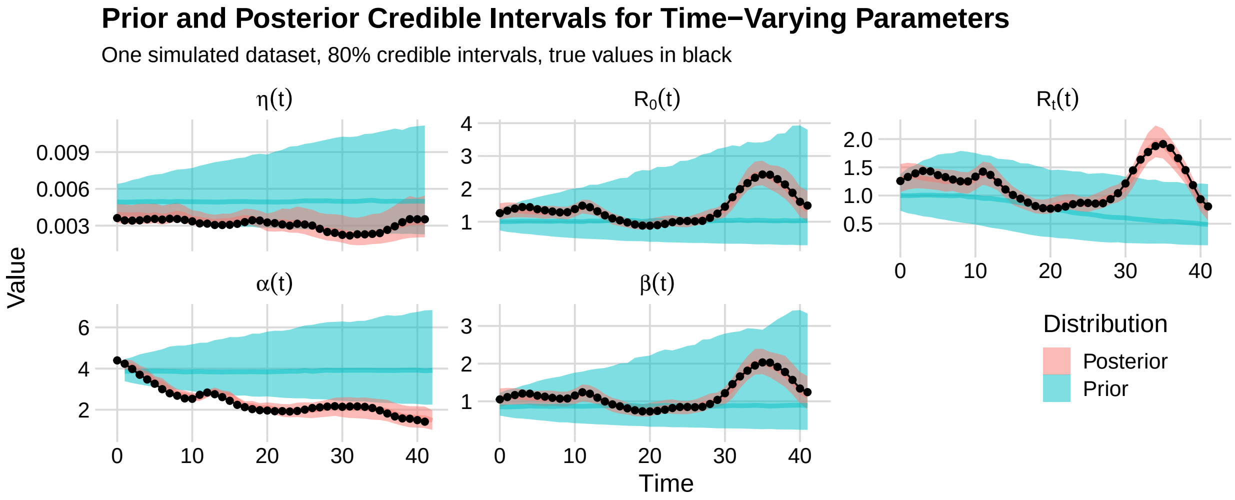


Figure S-1.3: Prior and posterior 80% credible intervals for time-varying parameters for a model fit to the dataset presented in Figure S-1.1. True values for the simulated parameters are indicated by solid black lines.

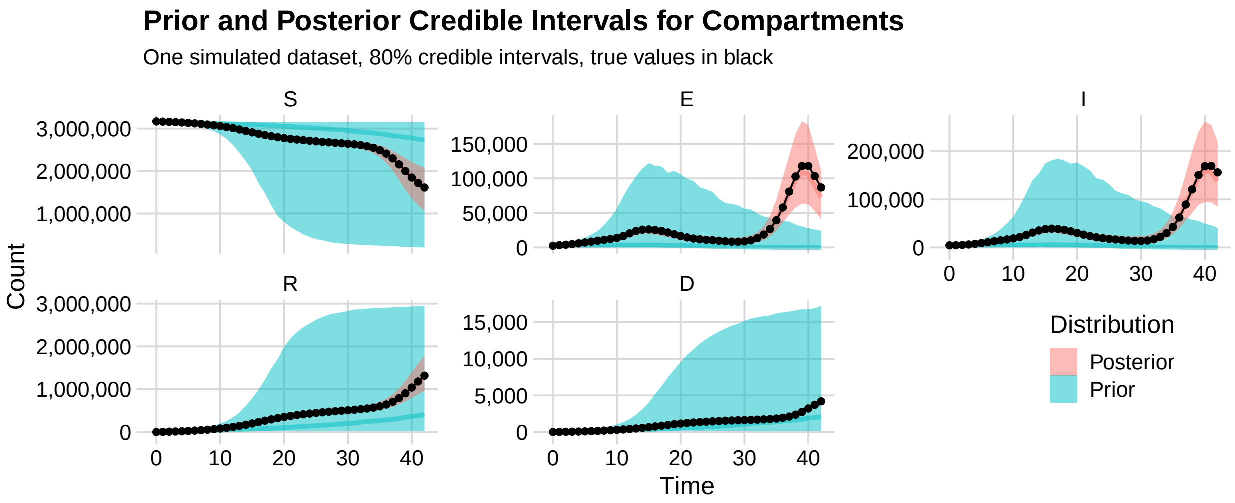


Figure S-1.4: Prior and posterior 80% credible intervals for latent compartments for a model fit to the dataset presented in Figure S-1.1. True values for the simulated compartment sizes are indicated by solid black lines.

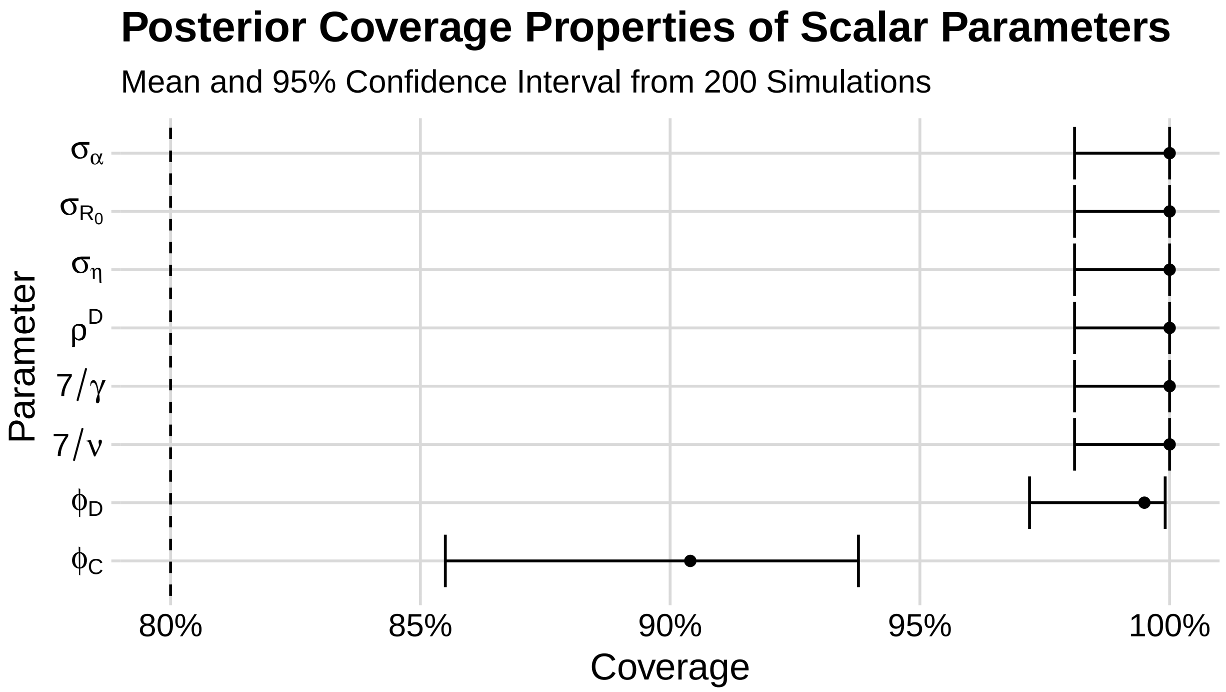


Figure S-1.5: Coverage properties of 80% posterior credible intervals for scalar parameters from models fit to 200 simulated datasets. Nominal coverage is indicated by the dashed line.

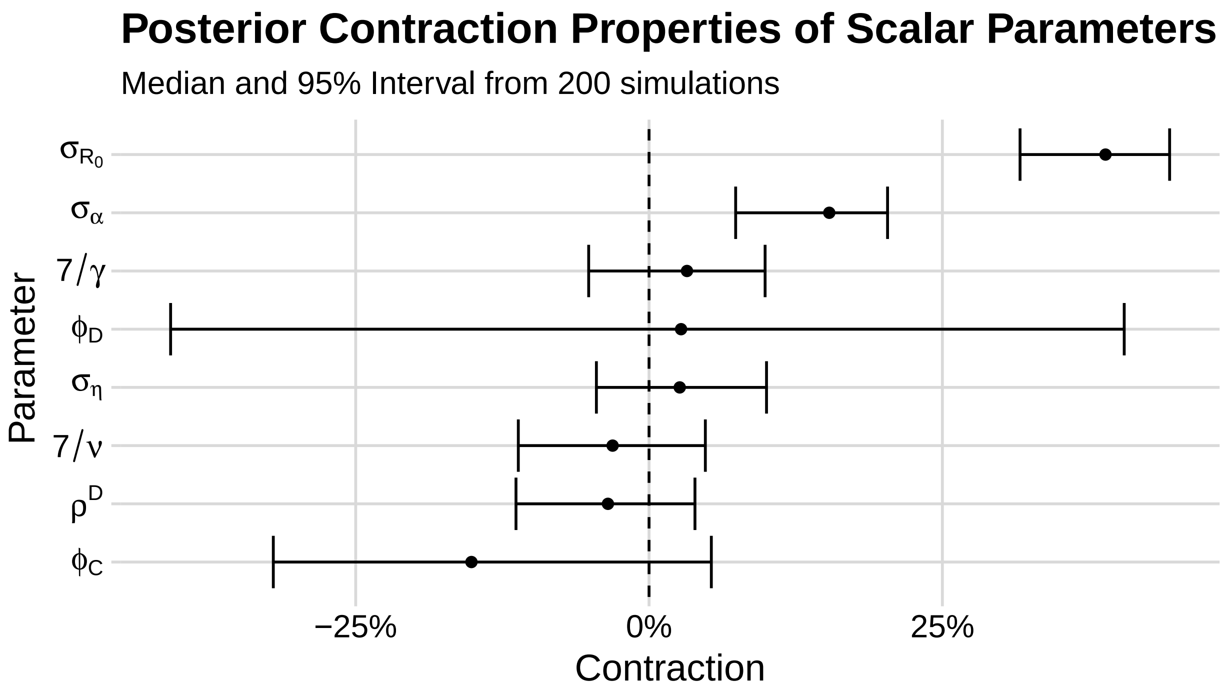


Figure S-1.6: Contraction properties of scalar parameters from models fit to 200 simulated datasets. Contraction is calculated as one minus the ratio of standard deviation of the posterior and the prior.

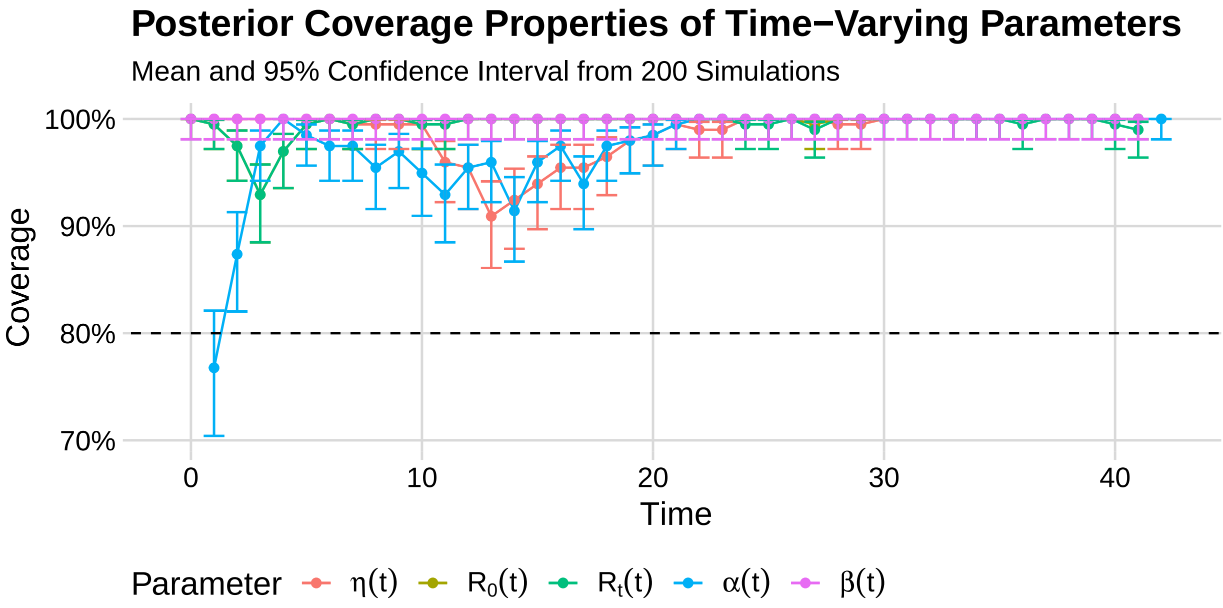


Figure S-1.7: Coverage properties of 80% posterior credible intervals for time-varying parameters from models fit to 200 simulated datasets. Nominal coverage is indicated by the dashed line.

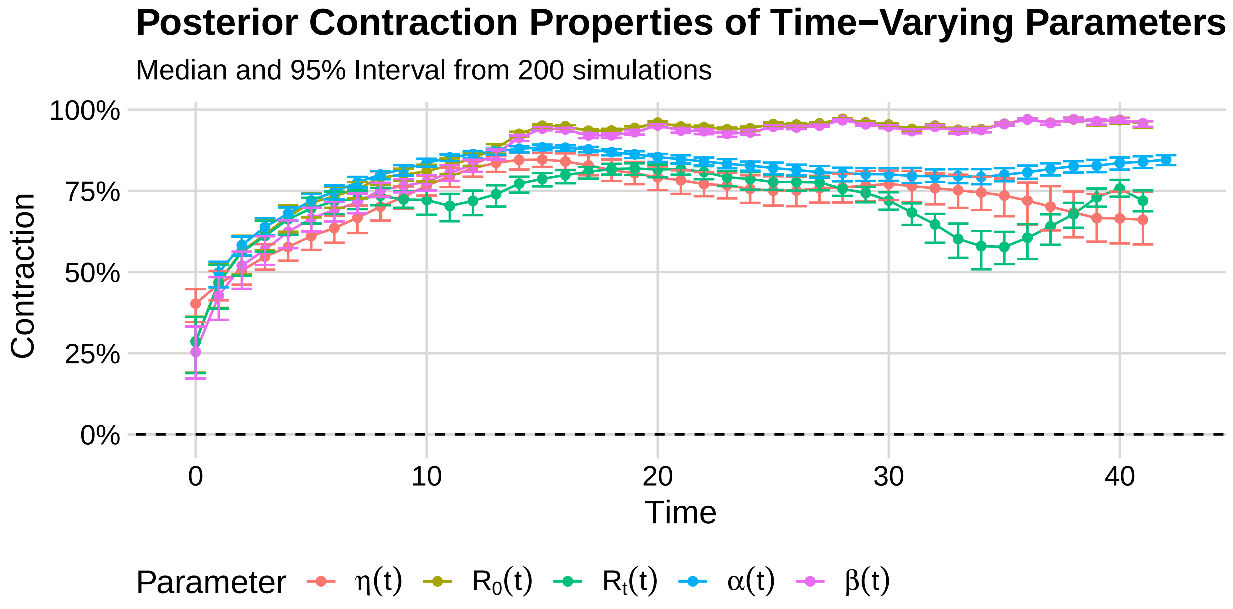


Figure S-1.8: Contraction properties of time-varying parameters from models fit to 200 simulated datasets. Contraction is calculated as one minus the ratio of standard deviation of the posterior and the prior.

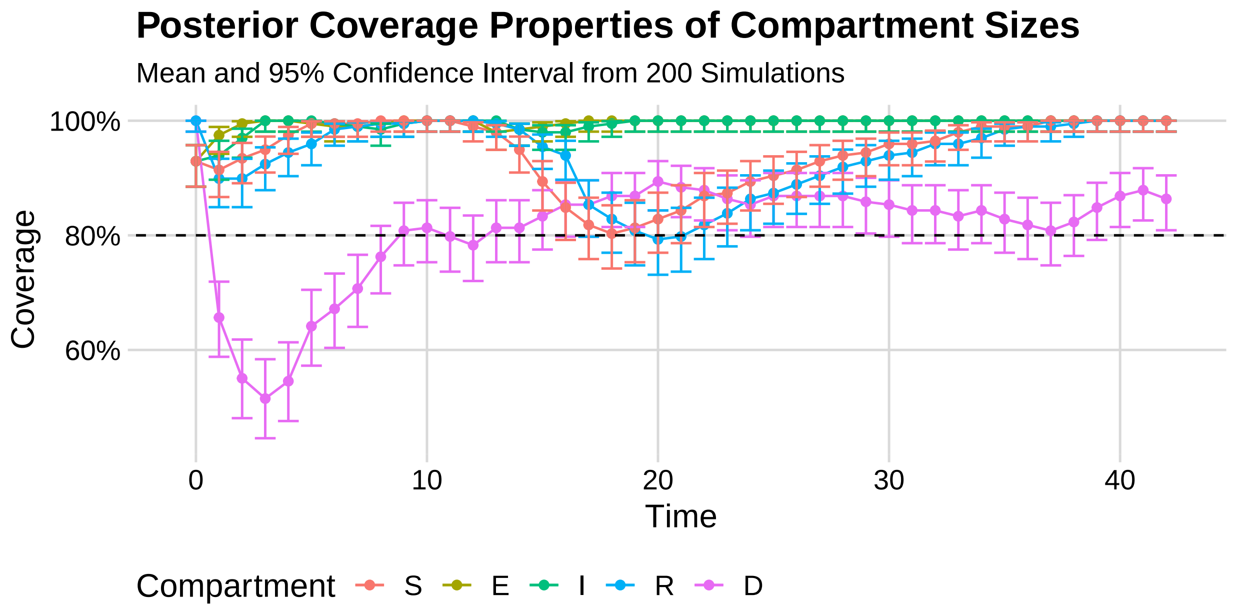


Figure S-1.9: Coverage properties of 80% posterior credible intervals for latent compartments from models fit to 200 simulated datasets. Nominal coverage is indicated by the dashed line.

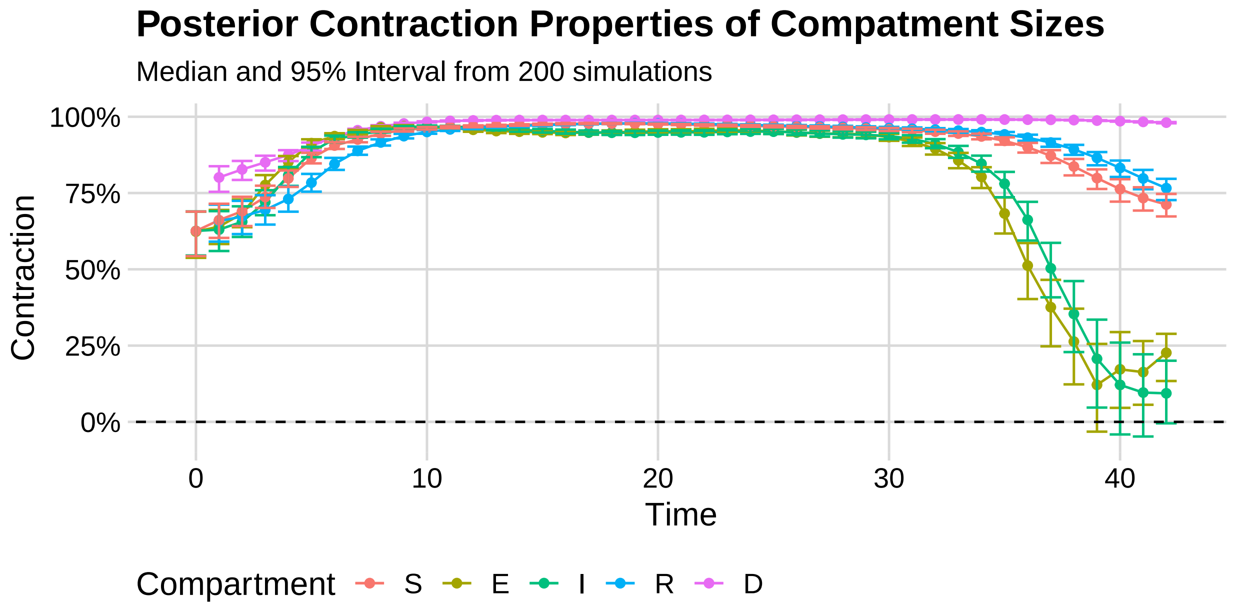


Figure S-1.10: Contraction properties of latent compartments from models fit to 200 simulated datasets. Contraction is calculated as one minus the ratio of standard deviation of the posterior and the prior.

S-2 Comparison with `epidemia`

We used the `epidemia` R package to infer R_t in the same 200 simulated datasets, as well as the Orange County data. Statistical details of these methods are presented below. Commentary on these results is presented in Section 3 of the main text.

The `epidemia` package can be used to create different branching process inspired models to estimate the effective reproduction number. In contrast to the compartmental model used in this paper, branching process inspired models have related the mean of current incidence to a weighted sum of previous incidence and the effective reproduction number R_t. Let I_t be the incidence at time t, R_t be the effective reproduction number at time t, and $g(t)$ be the probability density function of the generation time distribution (the time between an individual becoming infected and infecting another individual; under the compartmental model framework this is usually taken to be equivalent to the sum of the latent period and the infectious period). Then the mean relationship used is:

$$E[I_t|I_1, \ldots, I_{t-1}] = R_t \sum_{s=1}^{t-1} I_s g(t-s).$$

For the model we used in this study, we then added an observation model for new cases, modeled the effective reproduction number as a random walk, and modeled unobserved incidence as an autoregressive normal random variable with variance equal to the mean multiplied by an over-dispersion parameter.

$$
\begin{aligned}
\tau &\sim \exp(\lambda) \qquad \text{Hyperprior for unobserved incidence} \\
I_\nu &\sim \exp(\tau) \qquad \text{Prior on unobserved incidence } \nu \text{ days before observation} \\
I_{\nu+1}, \ldots, I_0 &= I_\nu \qquad \text{Unobserved incidence} \\
\sigma &\sim \text{Truncated-Normal}(0, 0.1^2) \\
\log R_0 &\sim \text{Normal}(\log 2, 0.2^2) \qquad \text{Prior on } R_0 \\
\log R_t | \log R_{t-1} &\sim \text{Normal}(\log R_{t-1}, \sigma) \qquad \text{Random Walk prior on } R_t \\
\psi &\sim \text{Normal}(10, 2) \qquad \text{Prior on variance parameter for incidence} \\
I_t | I_\nu, \ldots, I_{t-1} &\sim \text{Normal}(R_t \sum_{s<t} I_s g_{t-s}, \psi) \qquad \text{Model for incidence} \\
\alpha &\sim \text{Normal}(0.13, 0.7^2) \qquad \text{Prior on case detection rate} \\
y_t &= \alpha_t \sum_{s<t} I_s \pi_{t-s} \qquad \text{Mean of observed data model} \\
\phi &\sim P(\phi) \qquad \text{Prior on dispersion parameter for observed data} \\
Y_t &\sim \text{Neg-Binom}(y_t, \phi) \qquad \text{Observed data model}
\end{aligned}
$$

Here π_t are the values of the probability density function for the delay distribution, the time between an individual being infected and being observed. This distribution is assumed to be a gamma distribution with shape parameter one and mean equal to the true mean latent period. To sample from the posterior distribution, `epidemia` uses Hamiltonian Monte Carlo via the `Stan` simulation software (Stan Development Team, 2020). We draw 2000 posterior samples and discard the first 1000 for this analysis.

S-3 Comparison with modeling structured populations

Here, we demonstrate that semi-parametric modeling of key parameters can obviate the need for modeling heterogeneous populations with separate compartments. We construct a model wherein a disease spreads among two subpopulations: the "general" population and the "vulnerable" population, which interact with each other. Progression through compartments is governed by the following system of differential equations, with "g" subscripts denoting the general subpopulation and "v" subscripts denoting the vulnerable subpopulation and parameters having the same

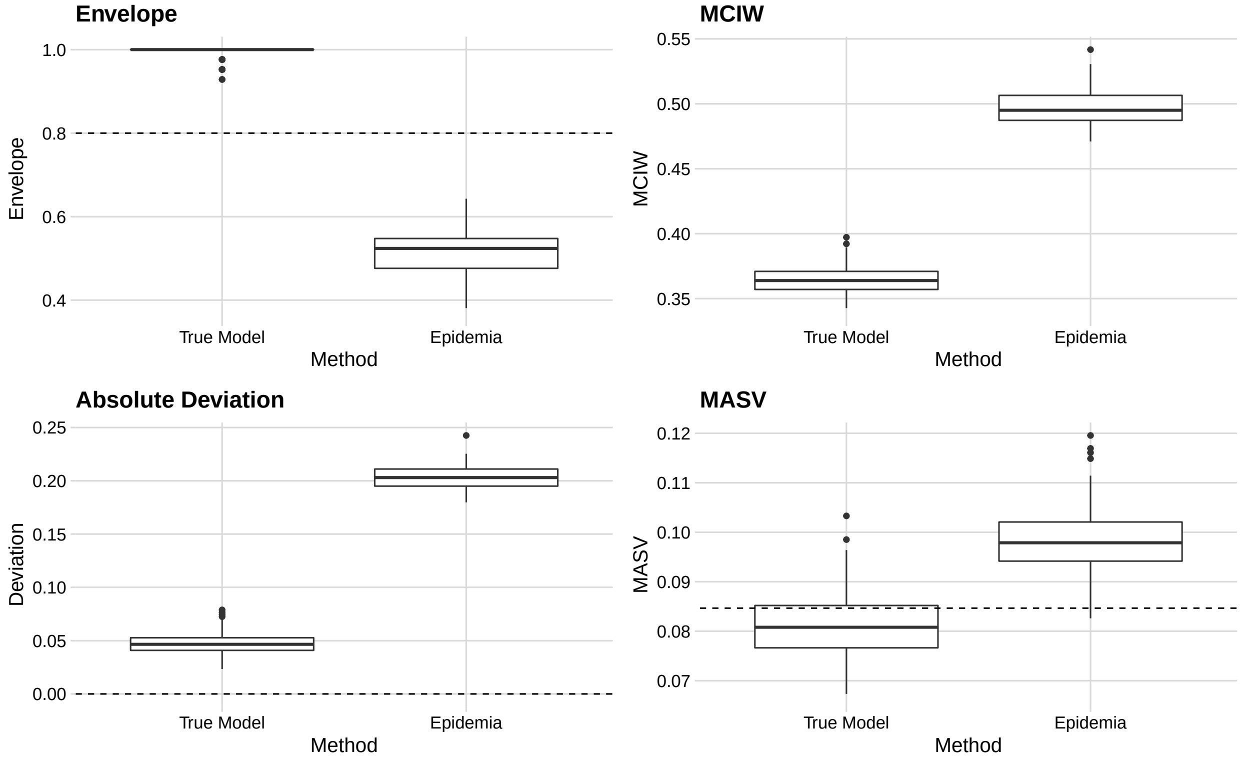


Figure S-2.11: Properties of R_t estimation from 200 simulated data sets. The envelope is the proportion of time points which the 80% posterior credible interval contains the true R_t value specified in the simulation. Mean credible interval width (MCIW) is the mean of credible interval widths across time points within a simulation replication. Absolute deviation is calculated as the mean of the absolute difference between the posterior median and the true R_t value at each time point. The mean absolute sequential variation (MASV) is the mean of the absolute difference between the posterior median at a time point and the posterior median at the previous time point.

interpretations as in the main text. The differential equations used for this model presented in (1).

$$
\begin{aligned}
\frac{\mathrm{d}S_v}{\mathrm{d}t} &= -\left(\beta_{vv} I_v + \beta_{vg} I_g\right)\frac{S_v}{N} & \frac{\mathrm{d}S_g}{\mathrm{d}t} &= -\left(\beta_{gg} I_g + \beta_{gv} I_v\right)\frac{S_g}{N} \\
\frac{\mathrm{d}E_v}{\mathrm{d}t} &= \left(\beta_{vv} I_v + \beta_{vg} I_g\right)\frac{S_v}{N} - \gamma E_v & \frac{\mathrm{d}E_g}{\mathrm{d}t} &= \left(\beta_{gg} I_g + \beta_{gv} I_v\right)\frac{S_g}{N} - \gamma E_g \\
\frac{\mathrm{d}I_v}{\mathrm{d}t} &= \gamma E_v - \nu I_v & \frac{\mathrm{d}I_g}{\mathrm{d}t} &= \gamma E_g - \nu I_g \\
\frac{\mathrm{d}R_v}{\mathrm{d}t} &= \nu(1-\eta_v) I_v & \frac{\mathrm{d}R_g}{\mathrm{d}t} &= \nu(1-\eta_g) I_g \\
\frac{\mathrm{d}D_v}{\mathrm{d}t} &= \nu\eta_v I_v & \frac{\mathrm{d}D_g}{\mathrm{d}t} &= \nu\eta_g I_g
\end{aligned}
\tag{1}
$$

subject to initial conditions $\mathbf{X}(t_0) = \mathbf{x}_0$ and $\mathbf{N}(t_0) = \mathbf{0}$, where $\mathbf{x}_0 = (S_{v0}, E_{v0}, I_{v0}, R_{v0}, D_{v0}, S_{g0}, E_{g0}, I_{g0}, R_{g0}, D_{g0})$ are initial compartment counts.

We only observed the unstratified case and death counts. Observed cases and deaths are Poisson distributed with the rate parameter equal to the number of latent cases and deaths, respectively.

$$
\begin{aligned}
Y_l &\sim \text{Poisson}(\Delta N_{E_v I_v}(t_l) + \Delta N_{E_g I_g}(t_l)) \\
M_l &\sim \text{Poisson}(\Delta N_{I_v D_v}(t_l) + \Delta N_{I_g D_g}(t_l))
\end{aligned}
$$

We construct a scenario where a disease outbreak occurs in a small vulnerable population with a true infection-fatality ratio of 10% before spreading to a larger general population with a true infection-fatality ratio of 1%. Because the outbreak spreads through the different populations at different times, the true population infection-fatality ratio varies in time. Figure S-3.12 shows the latent new cases and new deaths for each subpopulation, as well as the combined latent new cases and new deaths, and the observed new cases and new deaths for this constructed scenario.

Now, we fit a semi-parametric model, similar to the one in the main text, to this data. We model $\eta(t)$ with a logit-Guassian Markov random field, as in the main text. The differential equations used for this model presented in (2).

$$
\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta I \frac{S}{N} \\
\frac{\mathrm{d}E}{\mathrm{d}t} &= \beta I \frac{S}{N} - \gamma E \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= \gamma E - \nu I \\
\frac{\mathrm{d}R}{\mathrm{d}t} &= \nu(1-\eta(t)) I \\
\frac{\mathrm{d}D}{\mathrm{d}t} &= \nu\eta(t) I
\end{aligned}
\tag{2}
$$

subject to initial conditions $\mathbf{X}(t_0) = \mathbf{x}_0$ and $\mathbf{N}(t_0) = \mathbf{0}$, where $\mathbf{x}_0 = (S_0, E_0, I_0, R_0, D_0)$ are initial compartment counts.

Figures S-3.13–S-3.16, demonstrate, that when we fit our semi-parametric model, we can generally fit the data well and recover the true values of the parameters without modeling the two

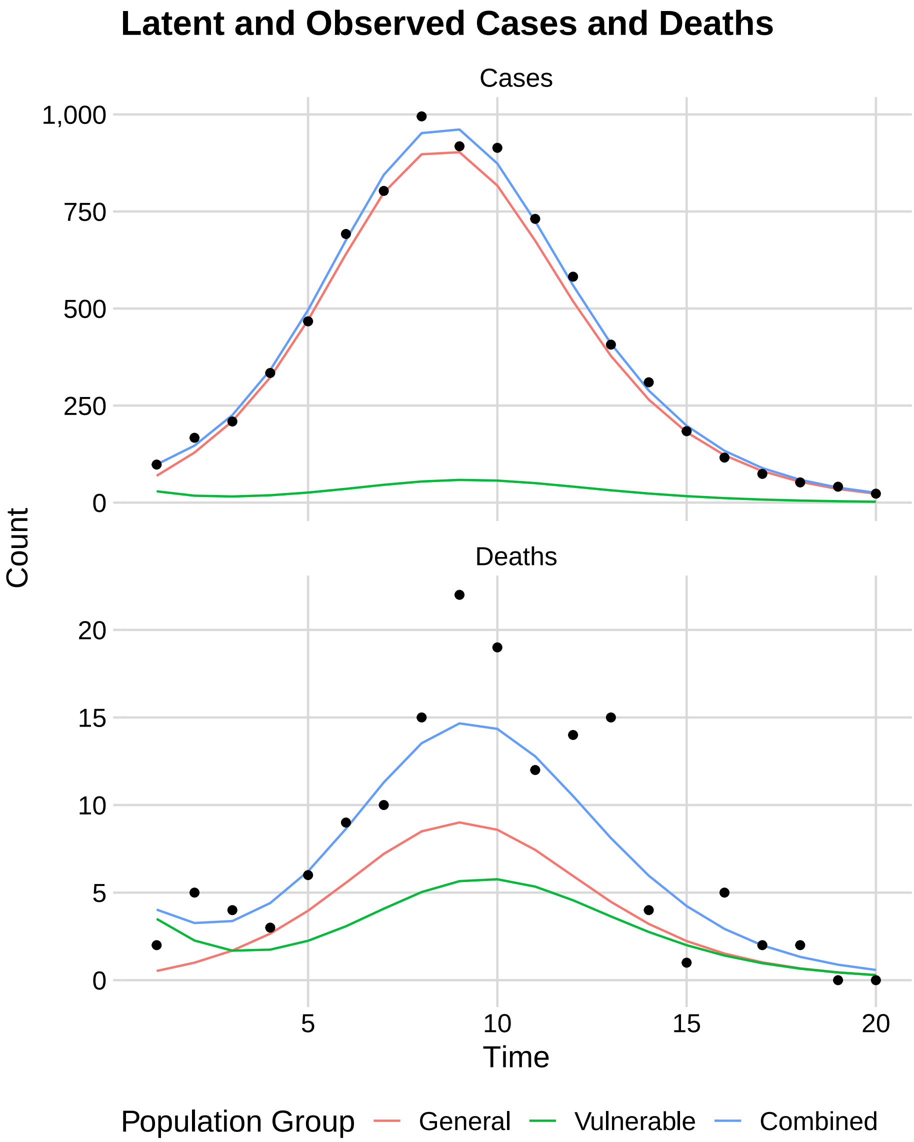


Figure S-3.12: Latent new cases and deaths for vulnerable and general subpopulations, along with combined latent new cases and new deaths and observed (combined) new cases and new deaths for a simulated scenario.

Table S-4.3: Model parameters and their prior distributions.

Analysis	**Parameter**	**Original Prior**	**Original Prior Median (95% Interval)**	**Sensitivity Prior**	**Sensitivity Prior Median (95% Interval)**
Half S_0	S_0	Logit-Normal(6, 0.25)	0.998 (0.993, 0.999)	Logit-Normal(5.31, 0.25)	0.995 (0.987, 0.998)
Half $\exp\left(\tilde{R}_{0,1}\right)$	$\exp\left(\tilde{R}_{0,1}\right)$	Log-Normal(0, 0.0625)	1.000 (0.613, 1.630)	Log-Normal(-0.693, 0.0625)	0.500 (0.306, 0.816)
Double expit $(\tilde{\eta}_1)$	expit $(\tilde{\eta}_1)$	Logit-Normal(-5.3, 0.04)	0.00497 (0.00336, 0.00733)	Logit-Normal(-4.61, 0.04)	0.00988 (0.00670, 0.01460)
Half $\exp(\tilde{\alpha}_1)$	Half $\exp(\tilde{\alpha}_1)$	Log-Normal(1.35, 0.0121)	3.86 (3.11, 4.79)	Log-Normal(0.657, 0.0121)	1.93 (1.55, 2.39)

heterogeneous populations.

S-4 Sensitivity analysis

We conducted four sensitivity analyses to see how our results change depending on the specified priors. In each additional analysis, we change only one aspect of the model priors. We perform one analysis where, *a priori*, twice the number of people are initially infected (denoted Half S_0), one with a lower initial basic reproduction number prior (denoted Half $\exp\left(\tilde{R}_{0,1}\right)$, one with a higher initial infection fatality ratio prior (denoted Double expit $(\tilde{\eta}_1)$, and one with a lower initial α prior (denoted Half $\exp(\tilde{\alpha}_1)$. Precise descriptions of the priors used in the sensitivity analyses are presented in Table S-4.3. Graphical results of the sensitivity analyses are presented in Figures S-4.17–S-4.19. We find that our model is typically robust to these alternative priors, and no alternative model leads to substantively different conclusions.

Additionally, we perform analyses where we modify the main model to, one at a time, fix each of the time-varying parameters, R_0, α, and η. As demonstrated in Figure S-4.20, fixing these parameters has no negative impact on the model's ability to properly fit the test positivity and death data, with each of the models exhibiting nearly identical posterior predictive distributions. However, these modified models do lead to substantially different inferences about the time-varying parameters themselves. This is shown in Figure S-4.21, where it appears that when one parameter is fixed, the others can become more flexible to still precisely match the observed data. The most dramatic effect is seen when fixing R_0, which leads to drastically different inferences about η and α. In contrast, there appears to be little impact from fixing the infection-fatality ratio, η as constant through time.

S-5 MCMC Diagnostics

Convergence diagnostics are presented in Tables S-5.4 and S-5.5, where $\hat{R}$ is the potential scale reduction factor (Vehtari et al., 2021), and ESS is the effective sample size, both as computed in the posterior R package (Bürkner et al., 2022). All parameters show potential scale reduction factors between 1 and 1.02, providing no evidence of lack of convergence. Additionally, all model parameters have effective sample sizes of multiple hundreds, which is sufficient for our inferences.

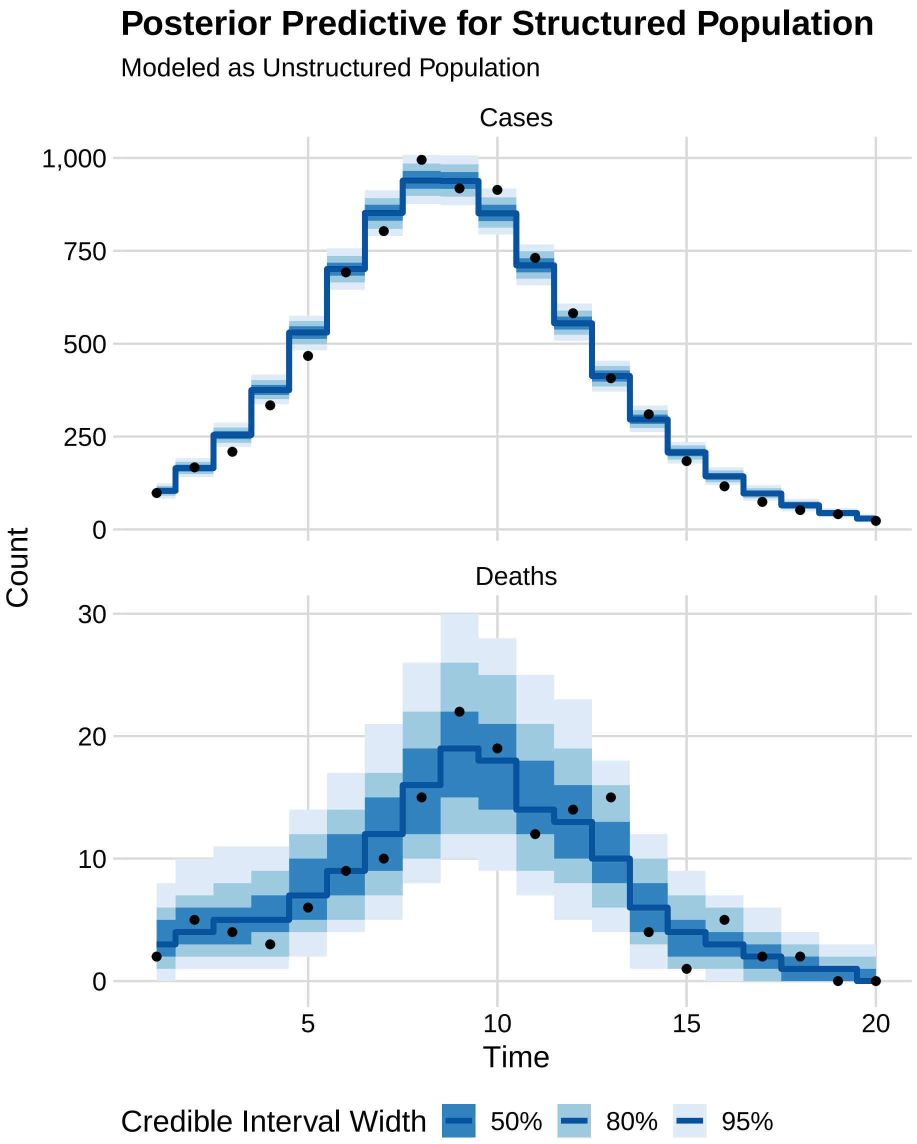


Figure S-3.13: Posterior predictive distributions for a model with non-parametric IFR fit to a simulated dataset with a heterogeneous population. The case and death data used are shown as black dots.

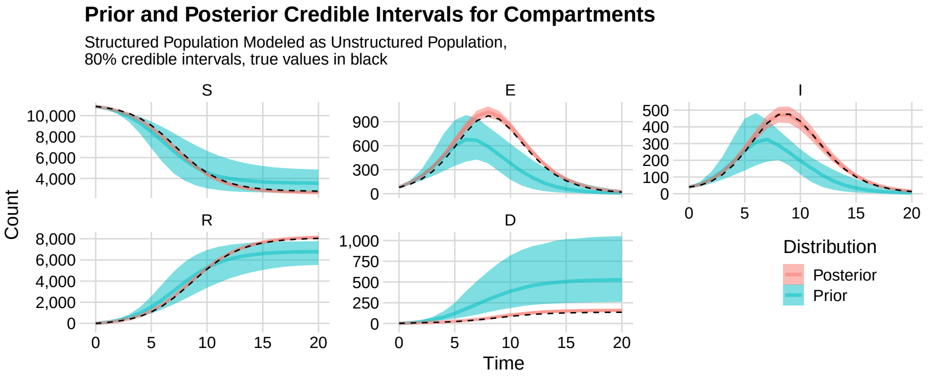


Figure S-3.14: Prior and Posterior distributions for latent compartments for a model with non-parametric IFR fit to a simulated dataset with a heterogeneous population. The true time-varying parameters are indicated by the dashed line.

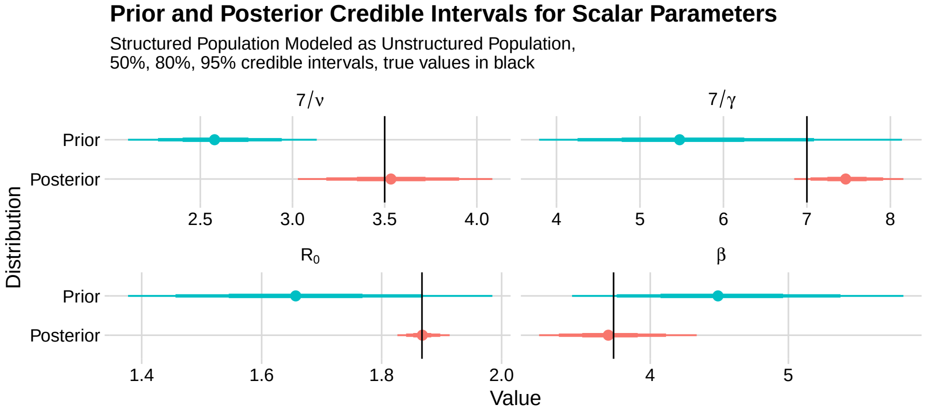


Figure S-3.15: Prior and posterior distributions for scalar parameters for a model with non-parametric IFR fit to a simulated dataset with a heterogeneous population. The true time-varying parameters are indicated by the vertical line.

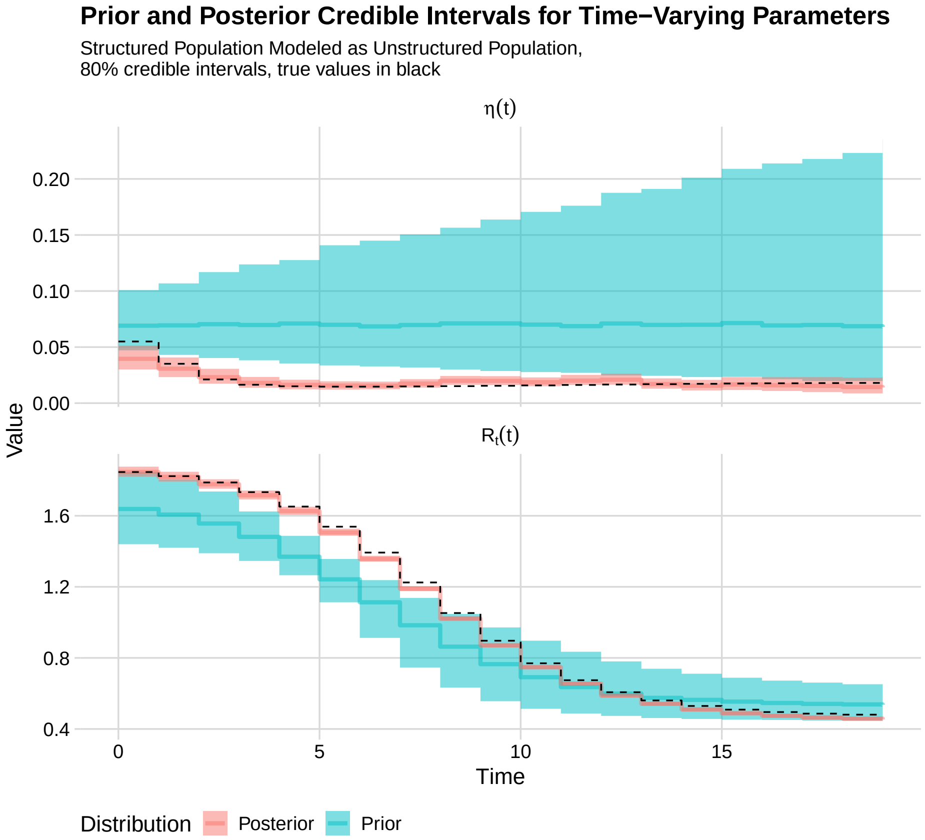


Figure S-3.16: Prior and Posterior distributions for time-varying parameters for a model with non-parametric IFR fit to a simulated dataset with a heterogeneous population. The true time-varying parameters are indicated by the dashed line.

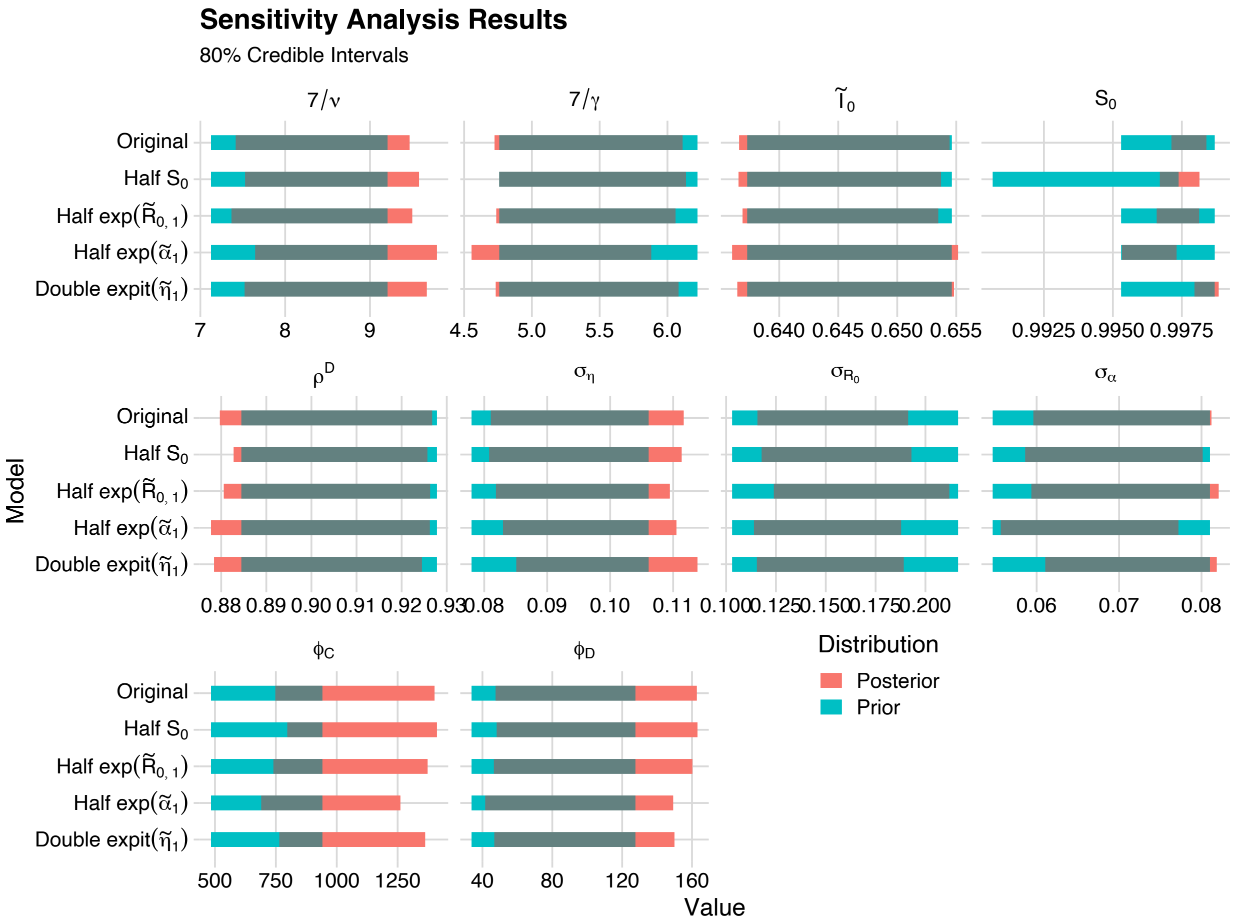


Figure S-4.17: Prior and posterior 80% credible intervals for scalar parameters from four sensitivity analyses and the original analysis.

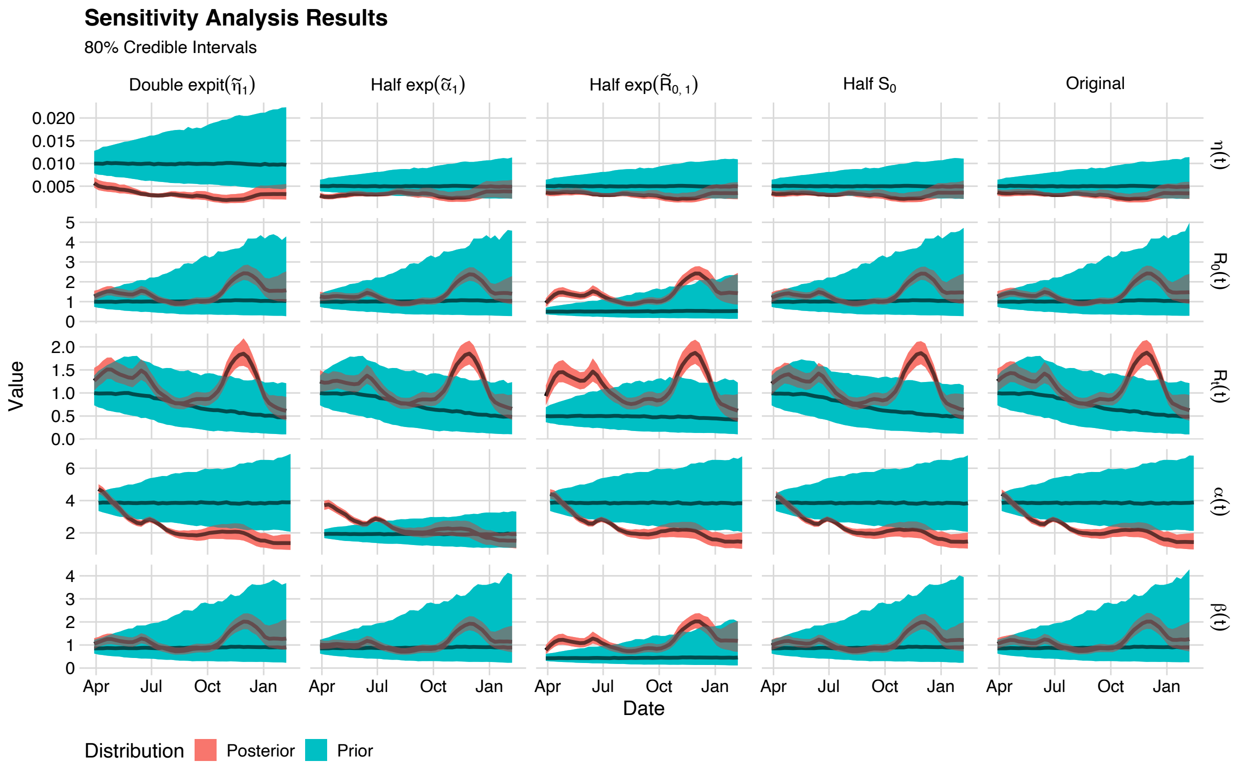


Figure S-4.18: Prior and posterior 80% credible intervals for time-varying parameters from four sensitivity analyses and the original analysis.

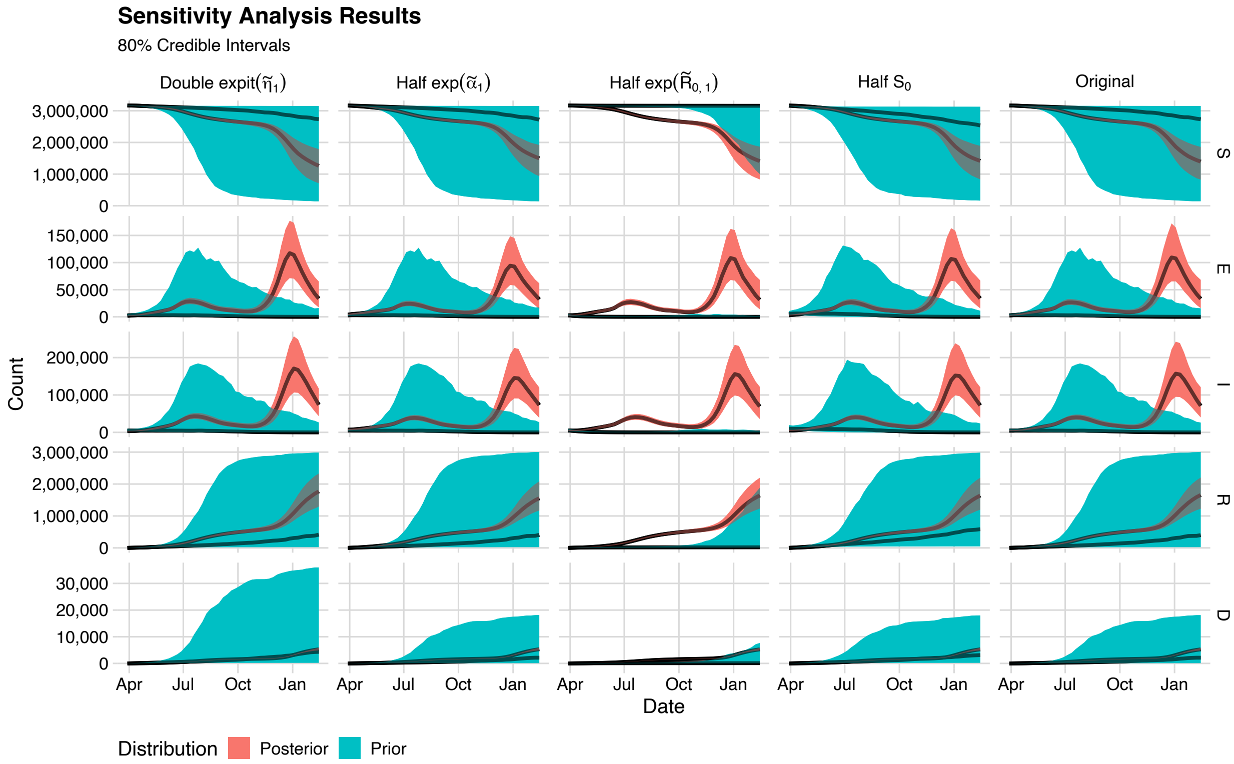
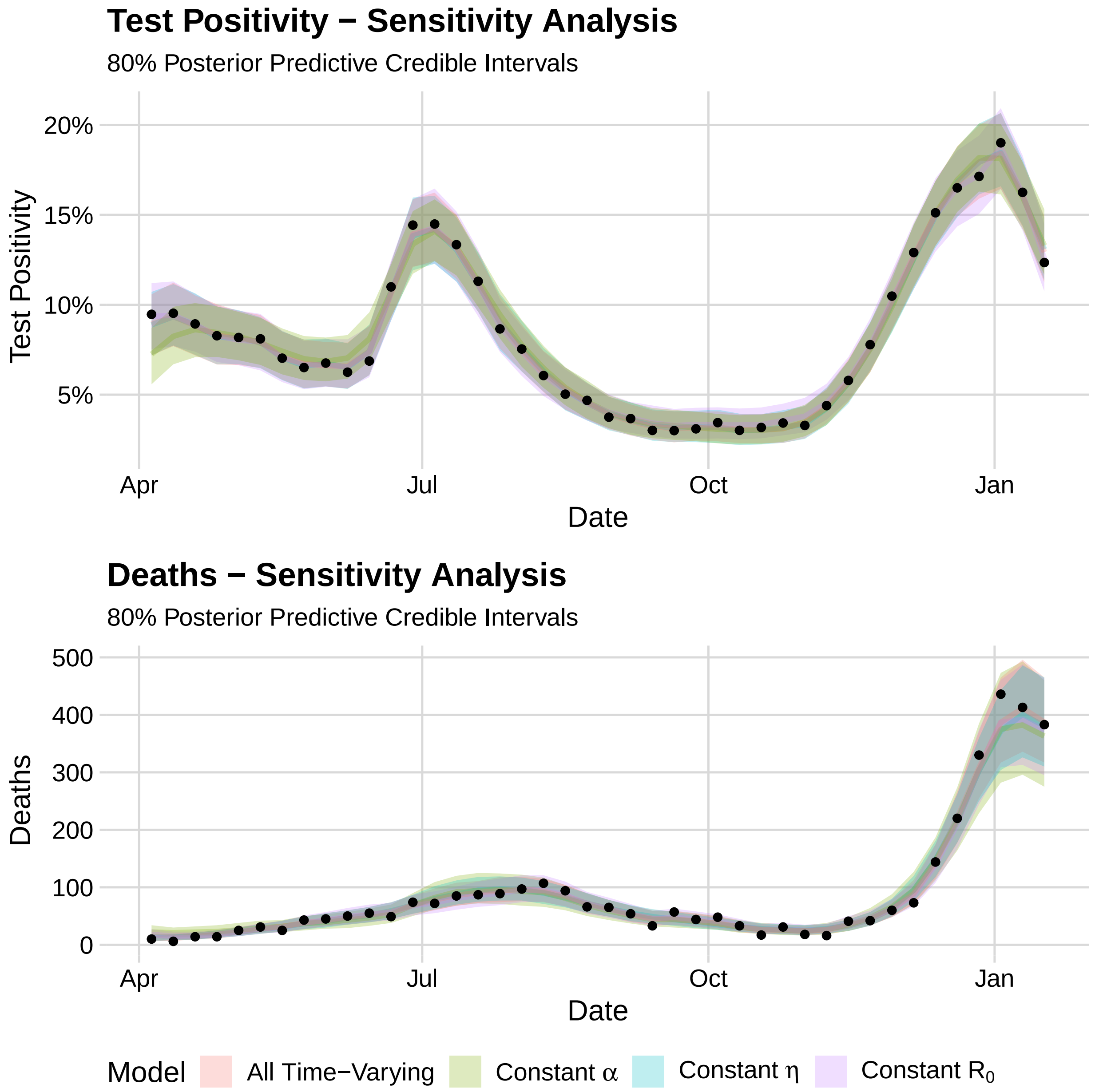


Figure S-4.19: Prior and posterior 80% credible intervals for time-varying parameters from four sensitivity analyses and the original analysis.



Test Positivity − Sensitivity Analysis

80% Posterior Predictive Credible Intervals

Deaths − Sensitivity Analysis

80% Posterior Predictive Credible Intervals

Figure S-4.20: Posterior predictive distributions when one of the typically time-varying parameters is made to be fixed through time.

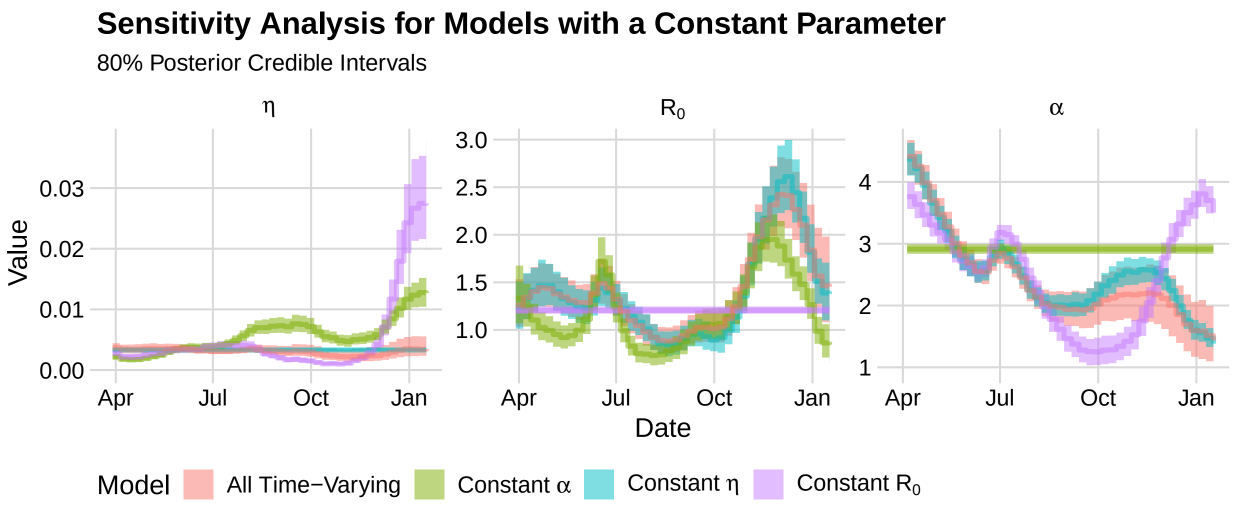


Figure S-4.21: Posterior inference for time-varying parameters when one of the typically time-varying parameters is made to be fixed through time.

We also produce a trace plot of the log-posterior probability for each chain in Figure S-5.22, which indicates that each chain explores a region of similar probability.

Table S-5.4: Convergence diagnostics for scalar parameters for the main model fit to the Orange County data set.

Parameter	$\hat{R}$	ESS
S_0	1.00	992.14
$\tilde{I}_0$	1.01	1128.48
$1/\gamma$	1.01	712.86
$1/\nu$	1.02	907.22
ϕ_D	1.00	876.28
ρ^D	1.00	812.16
ϕ_C	1.00	841.67
σ_{R_0}	1.00	544.21
σ_η	1.00	788.91
σ_α	1.01	567.30

Table S-5.5: Convergence diagnostics for scalar parameters for the main model fit to the Orange County data set.

Parameter	Min. $\hat{R}$	Avg. $\hat{R}$	Max. $\hat{R}$	Min. ESS	Avg. ESS	Max. ESS
$\exp(\tilde{\alpha}_t)$	1	1	1.01	630.78	827.20	1088.73
$\exp\left(\tilde{R}_{t,t}\right)$	1	1	1.02	693.98	977.45	1368.80
$\text{expit}(\tilde{\eta}_t)$	1	1	1.02	563.38	793.53	1259.37

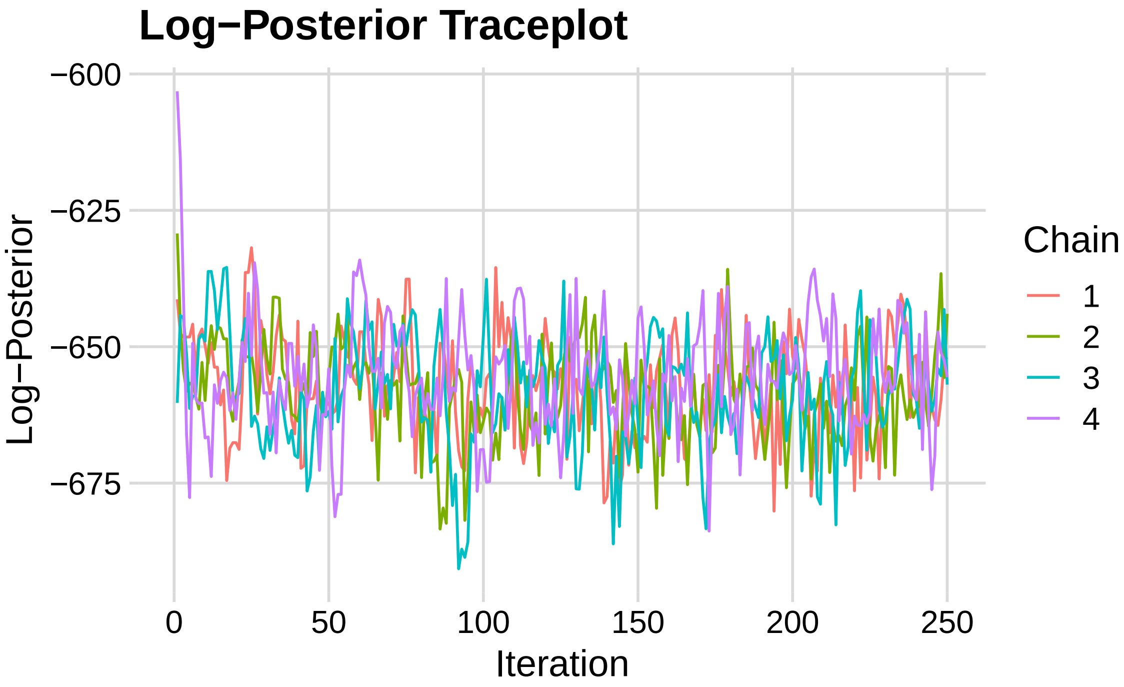


Figure S-5.22: Trace plot of log-posterior probability for the main model fit.