A Gaussian-process approximation to a spatial SIR process using moment closures and emulators

Parker Trostle\textsuperscript{1}, Joseph Guinness\textsuperscript{2}, Brian J. Reich\textsuperscript{1,*}

\textsuperscript{1}Department of Statistics, North Carolina State University, Raleigh, NC, 27607, United States, \textsuperscript{2}Department of Statistics and Data Science, Cornell University, Ithaca, NY, 14853, United States

*Corresponding author: Brian J. Reich, Department of Statistics, North Carolina State University, 2311 Katharine Stinson Drive, Raleigh, NC, 27607, USA (bjreich@ncsu.edu).

ABSTRACT

The dynamics that govern disease spread are hard to model because infections are functions of both the underlying pathogen as well as human or animal behavior. This challenge is increased when modeling how diseases spread between different spatial locations. Many proposed spatial epidemiological models require trade-offs to fit, either by abstracting away theoretical spread dynamics, fitting a deterministic model, or by requiring large computational resources for many simulations. We propose an approach that approximates the complex spatial spread dynamics with a Gaussian process. We first propose a flexible spatial extension to the well-known SIR stochastic process, and then we derive a moment-closure approximation to this stochastic process. This moment-closure approximation yields ordinary differential equations for the evolution of the means and covariances of the susceptible and infectious through time. Because these ODEs are a bottleneck to fitting our model by MCMC, we approximate them using a low-rank emulator. This approximation serves as the basis for our hierarchical model for noisy, underreported counts of new infections by spatial location and time. We demonstrate using our model to conduct inference on simulated infections from the underlying, true spatial SIR jump process. We then apply our method to model counts of new Zika infections in Brazil from late 2015 through early 2016.

KEYWORDS: emulator models; moment-closure approximations; SIR models; spatiotemporal epidemiology.

1 INTRODUCTION

Understanding the dynamics of disease spread is vital for control efforts. Without a deep comprehension of what drives new infections, efforts and resources targeting prevention may be misdirected. Epidemiological modeling plays an important role in this process by allowing researchers to analyze historical infection data. Many of these models are proposed by subject-area experts, such as veterinarians using their understanding of the dynamics of within- and between-farm spread to develop models for porcine diseases. These theory-based models are often compartmental models (eg, Paeng and Lee, 2017; Jones et al., 2021; Galvis et al., 2022). The most realistic of these models are stochastic compartmental models, which are flexible enough to handle the variability in real infection data. Unfortunately, these models are computationally demanding to fit and often require simulation-based approaches such as approximate Bayesian computing (ABC; Beaumont, 2010). This challenge is even greater when the data are spatially indexed, leading to potentially long computational run times that could last weeks or more.

The goal of our work is to demonstrate a novel statistical method to fit spatial, stochastic compartmental models for disease spread without requiring ABC. Such an approach will permit researchers to fit these disease models more efficiently without long simulations, while still properly accounting for the variability inherent to their stochastic processes. The key to our methodology is to approximate the complex epidemiological process with a simpler Gaussian process. In this paper, we demonstrate our approach by proposing a stochastic, spatial susceptible-infectious-recovered (SIR) model based on work by Paeng and Lee (2017) and then using our approximation to conduct inference on the true process.

Modeling data from such a stochastic process is hard in part because characterizing the moments of a continuous-time Markov process is not simple (eg, see Allen, 2008, on the SIR process). We rely therefore on a moment-closure approximation, which allows us to approximate the moments of our spatial SIR process with those of a Gaussian process. The terminology refers to “closing” the dependency on higher-order moments because the higher-order moments in a Gaussian process are fully characterized by its mean and covariance function. Moment-closure approximations were originally proposed in Whittle (1957), and such an approximation was used in Isham (1991) for the non-spatial SIR process. Moment-closure approximations have been used extensively in the natural sciences (eg, Kuehn, 2016; Forgues et al., 2019). There have been some proposed spatial moment-closure models, particularly in ecology (eg, Murrell et al., 2004; Ernst et al., 2019) and for disease spread on
networks where nodes are binary-valued (eg, Sharkey et al., 2015; Chen et al., 2020). Our approach is more general and relies on fewer simplifying assumptions than some of the above literature. However, there is a significant drawback to the moment-closure approach. These approximations are characterized by coupled, ordinary differential equations (ODEs). These ODEs are used to calculate the means and spatial covariances for the Gaussian-process approximation. This is a computational bottleneck for even a modest number of spatial locations.

To solve this bottleneck, we implement an emulator-based approximation to the moment-closure forward equations. Emulators, also known as surrogate models, are used to approximate the output of computationally intensive models (Gramacy, 2020). Much of the framework for modern emulators can be traced to Kennedy and O’Hagan (2001), an influential paper that proposed a Bayesian approach to calibrating computer models. There are multiple methods building upon their framework (Higdon et al., 2004; Goldstein and Rougier, 2006; Qian et al., 2006; Bayarri et al., 2007). An influential follow-up methodologies was Higdon et al. (2008), which used an SVD-based approach to design emulators. This approach was extended in Hooten et al. (2011) and Leeds et al. (2014). Recently Pratola and Chkrebtii (2018) and Gopalan and Wikle (2022) extended SVD-based approaches to computational output stored in tensors, an approach we adapt in this paper. Many other approaches to constructing emulators are possible, however (eg, Reich et al., 2012; Massoud, 2019; Thakur and Chakraborty, 2022).

In this paper, we describe our methodology and show how it may be used in a spatiotemporal, Bayesian hierarchical model for retrospective statistical inference given noisy counts of new infections. We do so by modeling the latent number of susceptibles using a low-rank emulator model based on a moment-closure approximation to our underlying continuous-time spatial SIR process. Our work makes multiple novel contributions, including extending the moment-closure work of Isham (1991) to a spatial domain, demonstrating the use of emulators to approximate the resulting ODEs, and showing, to the best of our knowledge, the first demonstrated use of an emulator for a covariance function in a physical-statistical model.

2 SPATIAL SIR MODEL

We begin by describing our spatial extension to the SIR model. We base our model on that of Paeng and Lee (2017) but generalize the spatial infections to not necessarily depend on distance between spatial locations, to allow $\beta$ to vary spatially, and by making the process stochastic. We first propose our spatial SIR jump process in Section 2.1, and then we derive its moment-closure approximation and the resulting forward equations in Section 2.2. A background overview for the closed-population SIR model is provided in Web Appendix A.

2.1 Spatial jump process

Let $D$ be a spatial lattice with $n_i$ spatial coordinates $s$. Let $N(s)$ be the set of spatial locations adjacent to $s$. We denote the susceptibles at $s$ at time $t$ as $X(s, t)$, and we similarly denote the infectious at $s$ at time $t$ as $Y(s, t)$. Each location has population size $N(s)$ that does not vary in time. The number of recovered is therefore determined by $X(s, t), Y(s, t)$, and $N(s)$, so we do not model them directly.

Define the current state at time $t$ as $H(t) = \{X(s_1, t), ..., X(s_n, t), Y(s_1, t), ..., Y(s_n, t)\}$. Consider a sufficiently small $\Delta t$ such that only one new infection or recovery at most may occur at any spatial location. Then, there are $2n_i + 1$ possible events in the interval $(t, t + \Delta t)$ for this sufficiently small $\Delta t$: a new infection in one of the $n_i$ locations, a recovery in one of the $n_i$ locations, or no change. For $i \in \{1, ..., n_i\}$, let $I^+(s_i, t)$ denote the new infection event such that $H(t + \Delta t) = H(t)$ except $X(s_i, t + \Delta t) = X(s_i, t) - 1$ and $Y(s_i, t + \Delta t) = Y(s_i, t) + 1$, and let $R^+(s_i, t)$ denote the recovery event such that $H(t + \Delta t) = H(t)$ except $Y(s_i, t + \Delta t) = Y(s_i, t) - 1$. Then, our spatial SIR jump process is defined via the following conditional probabilities:

$$P \left\{ I^+(s_i, t) \mid H(t) \right\} \approx \frac{\beta(s_i) \Delta t}{N(s_i)} X(s_i, t) Y(s_i, t) + \frac{\phi \Delta t}{N(s_i)} \sum_{s_i \in N(s_i)} X(s_i, t) Y(s_i, t)$$

$$P \left\{ R^+(s_i, t) \mid H(t) \right\} \approx \eta Y(s_i, t) \Delta t$$

with the probability of no event, $H(t + \Delta) = H(t)$, equal to one minus the sum of the $2n_i$ probabilities in (1). These probabilities are controlled by local infection parameters $\beta(s_i)$, spatial infection parameters $\phi$, and a recovery parameter $\eta$.

Our model makes a few simplifying assumptions about the nature of the disease. First, we assume $\eta$ does not vary spatially because it reflects an intrinsic property of the pathogen, though this may be a simplification in cases where individuals have spatially varying access to healthcare. Second, we do not consider $\beta$, $\phi$, or $\eta$ to vary temporally, although in practice they may change as diseases evolve or as healthcare measures are put into place. In this paper, we focus on generally shorter time domains, but our methodology could be extended to allow for temporally varying parameters, e.g., by reassigning $\beta$ to spatiotemporal covariates. Furthermore, our model could be extended further to include common other compartments, such as an “exposed” compartment in an SEIR model (Abou-Ismail, 2020). We discuss this further in Web Appendix B.4.

2.2 Forward equations for the spatial SIR jump process

Working with the spatial SIR jump process directly to model new infection counts is not practical. Therefore, we now describe the moment-closure approximation to this jump process in which we use an approximating Gaussian process. As described in Isham (1991), there are two approaches to derive the governing ODEs for the means and spatial covariances through time, which we call the forward equations. We begin by demonstrating the simper of the two approaches, which we call the heuristic approach, by deriving the forward equation for the mean number of susceptibles.

As in Section 2.1, consider $H(t)$ and a sufficiently small $\Delta t$ that only the $2n_i + 1$ transition events described earlier may oc-
cur. For a particular $s_t$, there can be two events that occur:
\[
X(s_t, t + \Delta t) = \begin{cases} 
X(s_t, t) - 1 & \text{with probability } P[I^+(s_t, t)H(t)] \\
X(s_t, t) & \text{with probability } 1 - P[I^+(s_t, t)H(t)],
\end{cases}
\] (2)
where $P[I^+(s_t, t)H(t)]$ was defined above in (1). The expected change in $X(s_t, t)$ is therefore $-P[I^+(s_t, t)H(t)]$. Taking the expectation of this quantity with respect to the spatially indexed $H(t)$ following a Gaussian process and then taking the limit as $\Delta t \to 0$ yields:
\[
\frac{d\mu_X(s_t, t)}{dt} = -\frac{\beta(s_t)}{N(s_t)} (\mu_X(s_t, t) + \sigma_Y(s_t, s_t; t)) + \sigma_Y(s_t, s_t; t),
\] (3)
where $\mu_X(s_t, t)$ is the mean susceptibles at $s_t$ at time $t$, $\mu_Y(s_t, t)$ is the mean infectious at $s_t$ at time $t$, and $\sigma_Y(s_t, s_t; t)$ is the covariance between the susceptibles at $s_t$ and infectious at $s_t$ at time $t$. Though not present in (3), we define $\sigma_{XX}(s_t, s_t; t)$ and $\sigma_{XY}(s_t, s_t; t)$ similarly.

Deriving the remainder of the moment-closure forward equations to the spatial SIR jump process for $\mu_Y(s_t, t), \sigma_{XX}(s_t, s_t; t), \sigma_{YY}(s_t, s_t; t), \sigma_{XY}(s_t, s_t; t)$, $\sigma_{XX}(s_t, s_t; t)$, $\sigma_{XY}(s_t, s_t; t)$, $\sigma_{YY}(s_t, s_t; t)$, and $\sigma_{XY}(s_t, s_t; t)$ may be done in a similar fashion as shown in (3). In general, this is straightforward though tedious algebra. The full set of forward equations is provided in Web Appendix B along with derivation details. The alternative derivation approach is based on Whittle (1957) and involves moment and cumulant generating functions. Details on that method are also provided in Web Appendix B. The two derivation methods yield the same forward equations.

We assume the starting conditions of the outbreak do not correspond to a high probability of the infection dying out. This is a source of moment-closure approximations failing and is referred to as "epidemic fadeout" (Lloyd, 2004). Only a subspace of the parameter space yields realistic epidemic behavior. In practice checking the curves for the mean susceptibles to ensure they are monotonically nonincreasing serves as a check on these assumptions, so in practice our prior is truncated to the space of values that do not lead to epidemic fadeout.

3 EMULATOR FOR THE MEAN AND COVARIANCE FUNCTIONS

The moment-closure approximation for the number of susceptibles $X(s_t, t)$ is a Gaussian process with mean function $\mu_X(s_t, t; \theta)$ and spatial covariance function $\Sigma_{XX}(s_t, s_t; t, \theta)$. Both of these functions vary over space and time and depend on the model parameters $\theta = (\beta, \phi, S_0)^T$, where $\beta = (\beta(s_t), \ldots, \beta(s_n))^T$ and $S_0 \in \mathcal{D}$ is the source of the outbreak. We assume that the recovery rate $\eta$ and starting time of the outbreak $T_0$ are known. Typically $\eta$ may be estimated using patient reports, and $T_0$ may be imputed based on initial reported cases. In this section, we develop a statistical emulator to approximate the forward equations with an emulator that can be used for repeated function calls in an MCMC algorithm. To do so, we evaluate the forward equations of $K$ space-filling input parameters $\theta_1, \ldots, \theta_K$ and then use the realizations of the forward equations to build a statistical prediction of their values at other inputs $\theta^*$. We denote the output of running the forward equations for input $\theta_k$ as $\mu_X(\theta_k)$ and $\Sigma_{XX}(\theta_k) \in \mathbb{R}^{n \times n \times n}$, where $k \in \{1, \ldots, K\}$.

Much of the discussion in this section will use tensor terminology and notation. We provide a brief overview of the necessary background in Web Appendix C. We first describe our methodology in Sections 3.1 and 3.2 in terms of the higher-order singular value decompositions (HOSVD) of tensors, as in Gopalan and Wikle (2022). In Section 3.3, we discuss our imputation methodology for arbitrary $\theta^* \notin \{\theta_1, \ldots, \theta_K\}$. In Section 3.4, we describe our model for new infections.

3.1 Low-rank approximation to the mean function

We store our simulated output for the mean susceptibles in a third-order tensor $\mathcal{U} \in \mathbb{R}^{n \times n \times K}$, where the slice $\mathcal{U}_{r=1}$ is the matrix $\mu_X(\theta_k) \in \mathbb{R}^{n \times n}$. We construct a low-dimensional approximation to $\mathcal{U}$ using a HOSVD (De Lathauwer et al., 2000; Kolda and Bader, 2009), such that the low-dimensional approximation is $\tilde{\mathcal{U}} \in \mathbb{R}^{r \times r \times r}$ with $r_1 \leq n_1$ and $r_2 \leq n_2$. The HOSVD is frequently compared with the well-known SVD for matrices, and it is a method to calculate a "low-rank approximation [to a tensor] with small reconstruction error" (Zare et al., 2018).

We denote the spatial factor matrix of $\mathcal{U}$ as $\mathcal{Y} \in \mathbb{R}^{n \times K}$ and the temporal factor matrix of $\mathcal{U}$ as $\mathcal{D} \in \mathbb{R}^{R \times K}$ with columns $\mathcal{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_K]$ and $\mathcal{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K]$. This corresponds to a $J_1 \times J_2 \times K$ low-rank approximation to $\mathcal{U}$, with $J_1 \leq n_1$ and $J_2 \leq n_2$. The low-rank approximation to the mean susceptibles is therefore
\[
\nu_X(s_t, t; \theta) \approx \sum_{i=1}^{J_1} \sum_{j=1}^{J_2} \mathbf{y}_i(s_t) \mathbf{d}_j(t) m_{ij}(\theta).
\] (4)

The basis weights $m_{ij}(\theta)$ control the dependence of the mean function on the model parameters $\theta$ as well as the interactions between $\mathcal{Y}$ and $\mathcal{D}$. These weights are the focus of our interpolation in Section 3.3, and they result from calculating $\mathcal{U} \times_1 \mathcal{Y}^T \times_2 \mathcal{D}^T$, where $\times_1$ and $\times_2$ are the n-mode products. We note that if $J_1 = n_1$ and $J_2 = n_2$, then our approximation to $\mathcal{U}$ is exact (Kolda and Bader, 2009). Finally, we implement streaming algorithms for this low-rank approximation for large $n_1, n_2$, and $K$ to avoid storing the entire tensor $\mathcal{U}$. These streaming algorithms let us calculate $\mathcal{Y}$ and $\mathcal{D}$ by holding only one $\theta_k$ in memory at a time. See Web Appendix D.

3.2 Low-rank approximation to the covariance function

Our approach to emulating the spatial covariance function is similar to our approach for the mean function. We store our simulated output for the spatial covariance matrices in a tensor $\mathbf{S} \in \mathbb{R}^{n \times n \times n \times K}$. The marginal spatial covariance matrix for $\theta_k$ at time $t$ is therefore $S_{\ldots, k} = \Sigma_{XX}(t, \theta_k)$. We now wish to calculate a low-rank approximation $\tilde{S} \in \mathbb{R}^{L_1 \times L_2 \times L_3 \times K}$ to $S$ based on an HOSVD, where $L_1 \leq n_1$ and $L_2 \leq n_2$. 


There is symmetry in $S$ because the slice corresponding to $S_{\ldots,t,k}$ is a covariance matrix. Therefore, the factor matrices for the first and second mode are identical. To ensure symmetry while avoiding redundancies, we build an emulator for $\Phi(t, \theta)$ such that $\Phi(t, \theta)\Phi(t, \theta)^T = \tilde{\Sigma}XX(t, \theta)$. It is more efficient to emulate $\Phi(t, \theta)$ because no matrix decompositions are required while model fitting.

We first calculate the spatial factor matrix $\Gamma = [\Gamma_1, \ldots, \Gamma_{L_s}] \in \mathbb{R}^{n_s \times L_s}$, where $\Gamma_1, \ldots, \Gamma_{L_s}$ are the orthonormal spatial basis functions that capture spatial variability in the unfolded tensor along the first mode (equivalently, the second mode). We then calculate $Z = S \times_1 \Gamma^T \times_2 \Gamma^T$, which calculates time-and parameter-indexed weights for the interaction of the spatial basis functions. In the case of $L_s = n_s$, then we may write $\tilde{\Sigma}XX(t, \theta) = \Gamma Z \times_{t,k} \Gamma^T$. It follows there exists a Cholesky decomposition $C(t, \theta)$ such that $C(t, \theta)C(t, \theta)^T = Z \times_{t,k}$, which naturally implies that we may set $\Phi(t, \theta) = GC(t, \theta)$ (see Web Appendix E).

Approximating Cholesky decompositions in an analogous fashion serves as a further dimension reduction. This can be accomplished doing a low-rank approximation to $C(t, \theta)$. We perform this low-rank approximation by forming a tensor $C \in \mathbb{R}^{L_s \times n_s \times n_s \times K}$, where $C_{i\ldots,t,k} = C(t, \theta)$, and calculating a temporal factor matrix for $C$, i.e., we calculate the first $L_s$ left singular vectors of $C$ unfolded along its third, temporal mode. We denote this temporal factor matrix as $\Delta = [\Delta_1, \ldots, \Delta_{L_s}]$, where $\Delta_1, \ldots, \Delta_{L_s}$ are orthonormal temporal basis functions. We then calculate the weight tensor $M \in \mathbb{R}^{L_s \times n_s \times L_s \times K}$ by calculating $C \times_1 \Delta^T$.

Therefore, we approximate the covariance function as

$$\Phi_j(s, t; \theta) = \sum_{i=1}^{L_s} \sum_{j=1}^{L_s} \Gamma_i(s) \Delta_j(t) M_{ij}(\theta).$$

$$\tilde{\Sigma}XX(s, s'; t, \theta) \approx \sum_{j=1}^{L_s} \Phi_j(s, t; \theta) \Phi_j(s', t; \theta).$$

By modeling the covariance function as $\Phi(t, \theta)\Phi(t, \theta)^T$ with $\Phi(t, \theta) \in \mathbb{R}^{n_s \times L_s}$ comprised of elements $\Phi_i(s, t, \theta)$, we ensure that our approximation for $\tilde{\Sigma}XX$ is non-negative definite for all $t$ and $\theta$. We note as before that all our approximations are exact when $L_s = n_s$ and $L_t = n_t$ (Kolda and Bader, 2009), and we also note these calculations may be performed using streaming algorithms to avoid memory problems (see Web Appendix E).

### 3.3 Gaussian process regression for interpolation

To fit our data model using MCMC, we need to evaluate $\mu_X(\theta^*)$ and $\tilde{\Sigma}XX(\theta^*)$ for proposed parameters $\theta^*$. Given the low-rank approximations above, evaluating $\mu_X(\theta^*)$ and $\tilde{\Sigma}XX(\theta^*)$ reduces to evaluating $m_{ij}(\theta^*)$ and $M_{ij}(\theta^*)$. Therefore, we require a model for the weights as functions of the input parameters $\theta$. To do so, we build a Gaussian-process regression for $m_{ij}(\theta^*)$ and $M_{ij}(\theta^*)$ assuming independence between the indices $i$, $j$, $l$ and using the same covariance for all predictions. We then predict $m_{ij}(\theta^*)$ and $M_{ij}(\theta^*)$ using local kriging with a Matérn correlation function. We provide additional details in Web Appendix D.

### 3.4 Emulator model for new infections

Using the emulators described above, our model for the new infections $y(s, t)$ is

$$y(s, t) | X_s, p, v \sim NB(\rho [X(s, t - 1) - X(s, t)], \frac{p [X(s, t - 1) - X(s, t)]}{v - 1}).$$

$$X_t, \theta, \mu(t, \theta), \Phi(t, \theta), a_\epsilon = \mu_X(t, \theta) + \Phi(t, \theta)B a_\epsilon.$$
TABLE 1 Runtime estimates to perform all forward-equation and emulator computations for $n_t = 46$ and $K = 1, 000$ on a university cluster.

| $n_s$ | Mean-function runtime | Covariance-function runtime |
|-------|------------------------|-----------------------------|
| 100   | 8.6 sec                | 149.6 sec                   |
| 225   | 21.0 sec               | 15.3 min                    |
| 400   | 24.3 sec               | 59.1 min                    |

TABLE 2 Results for first set of simulation studies using the full emulator approximation to the moment-closure equations (“full”), using only the mean emulator (“mean only”), approximate Bayes computation (“ABC”), or least-squares optimization (“ODE”).

| Scenario | Model | $\beta$, 95% cov. | $\phi$, 95% cov. | $\beta$, 99% cov. | $\phi$, 99% cov. | $\beta$, MSE (SE) | $\phi$, MSE (SE) |
|----------|-------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Baseline | Full  | 96.0%             | 86.0%             | 99.0%             | 96.5%             | 3.507 (0.366)     | 0.547 (0.058)     |
| Baseline | Mean only | 41.0% | 31.0% | 51.5%             | 38.5%             | 26.153 (5.375)     | 3.363 (0.590)     |
| Baseline | ABC   | 100.0%            | 100.0%            | 100.0%            | 100.0%            | 6.304 (0.585)     | 0.737 (0.068)     |
| Baseline | ODE   | 38.0%             | 39.0%             | 59.0%             | 64.0%             | 19.773 (2.224)     | 2.698 (0.301)     |
| $p = 0.75$ | Full  | 96.0%             | 89.0%             | 99.0%             | 95.5%             | 4.523 (0.434)     | 0.595 (0.058)     |
| $p = 0.75$ | Mean only | 46.5% | 38.0% | 60.0%             | 48.5%             | 15.005 (1.644)     | 2.018 (0.217)     |
| $p = 0.75$ | ABC   | 100.0%            | 100.0%            | 100.0%            | 100.0%            | 7.276 (0.638)     | 0.771 (0.070)     |
| $p = 0.75$ | ODE   | 41.5%             | 44.0%             | 59.0%             | 65.0%             | 15.634 (1.681)     | 2.054 (0.218)     |
| $\nu = 10$ | Full  | 95.5%             | 92.0%             | 97.0%             | 96.5%             | 7.873 (0.819)     | 0.832 (0.081)     |
| $\nu = 10$ | Mean only | 63.0% | 49.5% | 74.0%             | 61.0%             | 18.574 (2.934)     | 2.354 (0.321)     |
| $\nu = 10$ | ABC   | 99.5%             | 99.5%             | 100.0%            | 100.0%            | 12.493 (1.189)     | 1.210 (0.117)     |
| $\nu = 10$ | ODE   | 40.5%             | 43.5%             | 60.0%             | 63.0%             | 17.203 (1.739)     | 2.181 (0.213)     |

Empirical interval coverage is shown along with MSE and its standard error, multiplied by 10^4.

(iii) Fitting the model using MCMC requires storing the weights for the tensor-based approximations in memory (on the order of $J_f J_r K$ for $m$ and $L^2 f L_r K$ for $M$). The slowest step is the repeated $O(K \log(K))$ search for the nearest neighbors in the parameter space for proposed values of $\theta$.

We provide additional suggestions on implementation details as well as our MCMC algorithm in Web Appendix F. Though the emulator setup requires nontrivial computational resources, it yields about a 40-fold speed-up in computing $\hat{\mu}_x(\theta^*)$ and $\Phi$ compared with running the forward equations directly when fitting the model. Table 1 provides some example computational times for preparing the emulator. These times are based on a university cluster with nodes consisting mostly of Intel Xeon Gold 6130, 6226R, and Platinum 8358 processors.

4 SIMULATION STUDY

We analyze the performance of our method on data simulated from the spatial SIR jump process described in Section 2.1. We simulate counts of daily susceptibles from this jump process using the Gillespie stochastic simulation algorithm (Gillespie, 1976, 1977). The simulated response data $y(s, t)$ are generated by drawing from an overdispersed negative-binomial distribution with mean $p(X(s, t) - 1) - X(s, t)$ and with overdispersion $\nu > 1$.

4.1 Simulation study with constant $\beta$

We define a regular lattice of 25 spatial locations with rook adjacencies. The population size is uniformly 100,000, and the center of the square grid is $S_0$. We assume 100 individuals were infected at time 0 at $S_0$ and all others were susceptible. We use $\beta = 0.043$, $\gamma = 0.025$, $\eta = 0.019$, $n_t = 80$, and $\nu = 3.2$. We begin our analysis at day 61, so $t \in [61, \ldots, 140]$, reflecting data not being collected in the early days of an outbreak. We simulate 200 data sets for each simulation study.

We tune our emulators to use $J_f = 20$, $J_r = 10$, $L_r = 10$, and $L_f = 10$ (see Web Appendix G for details). For the first simulation, we use $p = 1$ so there is no underreporting, and we use $\nu = 3.2$. For the second simulation, we use $p = 0.75$ so not all new infections get reported. For the third simulation, we use $p = 1$ but set $\nu = 10$. For all simulations, we also fit a simpler model that only uses the mean function, i.e., sets all $\alpha$ equal to 0. Table 2 shows our results. It takes 2.0 minutes to draw 1000 posterior samples for these simulations.

We compare our results in Table 2 with two standard approaches for spatial epidemiological data. First, we fit the model using ABC (Beaumont et al., 2009). We use the R package SimInf to simulate from the stochastic spatial SIR process (Bauer et al., 2016; Widgren et al., 2019), and we use the same priors as in our proposed methodology. We use the means and standard deviations of the spatial infection counts to accept or reject proposed parameter draws, and we follow Beaumont et al. (2009) by repeatedly decreasing the acceptance threshold by 10% over ten iterations of the algorithm. Second, we fit the model analogously to Paeng and Lee (2017) via least-squares optimization using ODE curves based on (1). We also estimate confidence intervals for these parameters using parametric bootstrapping (Chowell, 2017).

As Table 2 shows, our good coverage and parameter estimates come from the covariance emulator. Fitting only mean curves results in worse coverage, a consequence of failing to allow sufficient model flexibility. As expected, our mean squared errors (MSEs) increase as overdispersion increases or reporting rates decrease. Our results compare favorably to the competing meth-
TABLE 3 Empirical credible-interval coverage results for second set of simulation studies using the full emulator approximation (“full model”) to the moment-closure equations or using only the mean emulator (“mean only”) as well as an approximate Bayesian computation (“ABC”) fit and a least-squares fit using the ODEs (“ODE”).

| Study   | $\beta_0$, 95% cov. | $\beta_0$, 99% cov. | $\beta_1$, 95% cov. | $\beta_1$, 99% cov. | $\phi$, 95% cov. | $\phi$, 99% cov. |
|---------|----------------------|----------------------|----------------------|----------------------|-------------------|-------------------|
| Full model | 92.0%                | 96.0%                | 89.5%                | 96.0%                | 88.5%             | 95.5%             |
| Only means | 24.5%                | 27.5%                | 14.0%                | 17.0%                | 21.5%             | 28.0%             |
| ABC     | 100.0%               | 100.0%               | 100.0%               | 100.0%               | 100.0%            | 100.0%            |
| ODE     | 100.0%               | 100.0%               | 59.5%                | 65.5%                | 99.0%             | 100.0%            |

In this study, $\beta(s) = \exp[\beta_0 + \beta_1 x(s)]$ where $x(s)$ is a spatially varying covariate.

TABLE 4 MSEs and their standard errors for second set of simulation studies using the full emulator approximation (“full model”) to the moment-closure equations or using only the mean emulator (“mean only”) as well as an approximate Bayesian computation (“ABC”) fit and a least-squares fit using the ODEs (“ODE”).

| Study   | $\beta_0$, MSE (SE) | $\beta_1$, MSE (SE) | $\phi$, MSE (SE) |
|---------|---------------------|---------------------|------------------|
| Full model | 25.202 (2.843)      | 1.298 (0.130)       | 0.010 (0.001)    |
| Only means | 171.197 (7.682)     | 7.682 (1.301)       | 0.071 (0.005)    |
| ABC     | 5,946.657 (187.692) | 435.759 (12.136)    | 0.554 (0.014)    |
| ODE     | 110.416 (14.678)    | 17.660 (1.266)      | 0.033 (0.005)    |

In this study, $\beta(s) = \exp[\beta_0 + \beta_1 x(s)]$ where $x(s)$ is a spatially varying covariate. Both MSEs and SEs are multiplied by $10^4$.

4.2 Simulation study with spatially varying $\beta(s)$

For our second set of simulations, we allow $\beta$ to vary spatially such that $\beta(s) = \exp[\beta_0 + \beta_1 x(s)]$. We draw $\exp[x(s)] \sim U(0, 3000)$ for each $s \in D$ and then calculate $x(s) = x(s) - x_t(s)$ to center the spatially-varying covariate. We use $\beta_0 = -2.83$, $\beta_1 = 0.1$, $\phi = 0.045$, $\eta = 0.04$, $\nu = 2.5$, and $p = 1$ while keeping $n_t = 25$ and $n_i = 80$. We begin our analysis at day 21, i.e., $t \in [21, \ldots, 100]$. We again use $J_t = 20$, $J_i = 10$, $L_t = 10$, and $L_i = 10$ (see Web Appendix G).

Table 3 shows coverage compared with a mean-emulator-only model, and Table 4 similarly shows the MSEs for $\beta_0$, $\beta_1$, and $\phi$. The difference between including or excluding the covariance emulator is striking. The problem with the mean-only model is insufficient flexibility to estimate the parameters well. The inclusion of the covariance emulator yields good coverage and MSEs. It takes 5.5 minutes to draw 1000 posterior samples for these simulations.

We again compare our results with ABC and a least-squares fit using ODE curves. As before, ABC was slow, and we were decreased the acceptance tolerance threshold only five times to get runtime under four days. This led to wide credible intervals and poor MSE, suggesting yet more time was needed to fit this model. Unlike the first simulation study, the ODE approach sometimes yielded 100% coverage. We found the estimated sampling distributions exhibited strong skew with long tails, caused by unlikely simulated infections leading to wildly different parameter estimates.

5 DATA ANALYSIS

We demonstrate the application of our model to Zika virus outbreak data in Brazil. The Zika virus is typically a mild virus, causing fevers, rashes, and various pains (CDC, 2019b). Unfortunately, there are serious complications for pregnant women, whose children may be born with multiple severe birth defects including microcephaly (CDC, 2019a). Several papers have studied the outbreak in Brazil as a whole but left spatial and geographic heterogeneities for future work (Wang et al., 2017; Sadeghieh et al., 2021). One paper fit separate models for the outbreaks in eight different states independently (Zhao et al., 2019).

The Pan American Health Organization (PAHO) reports government-supplied data on weekly new cases of the Zika virus (PAHO, 2021). The data are reported for all 27 states of Brazil, including the Federal District. The first significant outbreak began in Maranhao in early 2015 with infections peaking in the second half of 2015 and into 2016 (PAHO, 2016). We fit our model to the 40 weeks of outbreak data beginning in the 41st week of 2015 through the 28th week of 2016, corresponding roughly to October 2015 through July 2016 (PAHO, 2021).

We relied on the estimate of weekly recovery rate $\eta = 1.2$ from Sadeghieh et al. (2021), which studied Zika spread by mosquitoes. We assumed the outbreak began in the first week of 2015 in Maranhao with 100 infections. We expected the reported weekly new infections to be underreported; one paper suggests the reporting rate for Zika infections was between 7% and 17% (Shutt et al., 2017) for Central and South America as a whole. These rates likely varied by country. To pick starting values and to evaluate reporting rates, we fit a simple model to the Zika data with a Poisson likelihood and mean function based on a deterministic SIR model. We found that even the lower bound of 7% yielded poor fits to the data, and we estimated that the reporting rates were often closer to 1% and varied by state. We fixed those reporting rates for all subsequent analysis. Additional suggestions on using available literature and estimating initial pa-
paramaters is in Web Appendix H, which may be useful in cases when the disease is not as well documented as Zika.

Using the above starting values, we prepared our emulator simulations. We model $\beta(s) = \exp(\beta_0 + \beta_1 x(s))$, where $x(s)$ is the scaled and centered log population density of state $s$ (IBGE, 2021). We used $J_1 = 20$, $J_2 = 10$, $L_1 = 10$, and $L_2 = 10$, and we used the 20 nearest neighbors when kriging for both emulators. This was to help with the large parameter space relative to $K$.

We estimate $\hat{\beta}_0 = -0.077 (-0.109, -0.052)$, $\hat{\beta}_1 = 0.202 (0.193, 0.213)$, and $\hat{\phi} = 0.087 (0.082, 0.092)$, where the estimates are posterior means and the numbers in parentheses are 95% credible intervals. Fitting the emulator took 232 minutes using 20 cores, and it took 2.5 minutes to draw 1000 posterior samples during model fitting (totaling 14.5 hours for 250,000 samples after discarding 100,000 as burn-in). By comparison, drawing fifty posterior samples using ten iterations of ABC took 4.5 days and yielded $\hat{\beta}_0 = -0.647 (-1.356, 0.009)$, $\hat{\beta}_1 = 0.457 (0.150, 0.835)$, and $\hat{\phi} = 0.161 (0.076, 0.225)$, leading to much wider credible intervals. As a further comparison, we fit a non-spatial model using the original Isham (1991) Gaussian Process approximation to the non-spatial, stochastic SIR model (see Web Appendix A). Because this model only takes 38 seconds to fit, we do not rely on an emulator, though otherwise the methodology is analogous to our approach described in Section 3.4. We estimate $\hat{\beta} = 1.373 (1.368, 1.378)$, where $\beta$ is the sole transmission parameter. As expected, $\hat{\beta}$ is much larger here because all transmission are local to Brazil as a whole. We again fix $\eta = 1.2$.

There are two implications from our model fit. First is that, as expected, higher population densities are associated with higher rates of transmission. The second is that there is evidence of strong spatial spread of Zika as evidenced by the credible interval for $\phi$. We note that the parameter estimates and credible intervals were not in the emulator design space corresponding to problematically low $\phi$ and high $\beta_1$ as discussed previously.

Our model fit best in states with the highest counts of outbreaks. In Figure 1, we plot the new counts of Zika infections in Rio de Janeiro and Bahia, the two states with the highest counts of reported new infections, along with overall state-level Zika infections during this time. Additional discussion of these results are available in Web Appendix I.

6 DISCUSSION

We have proposed a novel method to analyze spatial infection data based on combining moment-closure approximations with emulators. Both components of our method are vital. The emulator approach is needed because of the computational run time of the forward equations. However, without the forward equations the emulator-based approximation would need to be based on the stochastic, spatial SIR jump process, a much harder modeling problem.

Though this paper has been about a spatial SIR jump process, our approach can be extended to other continuous-time Markov processes. There is a need for efficient approximations in domains like the veterinary literature, which often relies on repeated simulations of complex compartmental models for epidemiological studies (eg, Jones et al., 2021; Galvis et al., 2022). In our experience, even simple spatial models there may take weeks to fit. Our method could allow a significant reduction in the computational burden in this work.

There are several areas for improvement. First, we assume there is no epidemic fadeout. A more sophisticated moment closure may not need this assumption. The second is that as the dimension of $\Theta$ grows, $K$ may have to grow as well, suggesting a need for a more sophisticated experimental design. The third is we assume $\eta$, $\beta$, and the starting time of the infection are known or could be estimated in preliminary analysis. The final is practical: our approach requires a significant amount of hands-on work, requiring the user to derive forward equations, tune emulator designs, and then tune an MCMC algorithm. Finding ways to simplify the approach would allow for its wider adoption.

ACKNOWLEDGMENTS

The authors thank the associate editor and the referee for their constructive and insightful comments and suggestions that greatly improved the paper. The authors also acknowledge the
Supplementary Material

Supplementary material is available at Biometrics online.

Web Appendices referenced in Sections 2–5 along with all code for Sections 4–5 are available with this paper at the Biometrics website on Oxford Academic. The code is also available at: https://github.com/jptostle/SpatialSIRGPMC.

Funding

This work was partially supported by the National Science Foundation (DMS2152887) and the National Institutes of Health (R01ES031651-01).

Conflict of Interest

None declared.

Data Availability

The data that support the findings in this paper are available at the Pan American Health Organization (PAHO) Health Information Platform for the Americas (PLISA) website at https://www3.paho.org/data/index.php/en/mnu-topics/zika.html.

References

Abou-Ismail, A. (2020). Compartmental models of the Covid-19 pandemic for physicians and physicians-scientists. SN Comprehensive Clinical Medicine, 2, 852–858.

Allen, L. J. S. (2008). An introduction to stochastic epidemic models. In: Mathematical Epidemiology, (eds. P., F. Brauer), 81–130. Berlin: Springer.

Bauer, P., Engblom, S. and Widgren, S. (2016). Fast event-based epidemiological simulations on national scales. The International Journal of High Performance Computing Applications, 30, 438–453.

Bayarri, M. J., Walsh, D., Berger, J. O., Cafoe, J., Garcia-Donato, G., Liu, F. et al. (2007). Computer model validation with functional output. The Annals of Statistics, 35, 1874–1906.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics, 41, 379–406.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M. and Robert, C. P. (2009). Adaptive approximate Bayesian computation. Biometrika, 96, 983–990.

CDC (2019a). Birth Defects | Zika virus | CDC https://www.cdc.gov/zika/healtheffects/birth_defects.html. [Accessed 20 July 2022].

CDC (2019b). Symptoms | Zika Virus | CDC https://www.cdc.gov/zika/symptoms.html. [Accessed 20 July 2022].

Chen, X., Ogura, M. and Preciado, V. M. (2020). SDP-based moment closure for epidemic processes on networks. IEEE Transactions on Network Science and Engineering, 7, 2850–2865.

Chowell, G. (2017). Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. Infectious Disease Modelling, 2, 379–398.

Davis, T. A. (2006). Direct Methods for Sparse Linear Systems. Philadelphia: Society for Industrial and Applied Mathematics.

De Lathauwer, L., De Moor, B. and Vandewalle, J. (2000). A multilinear singular value decomposition. SIAM Journal on Matrix Analysis and Applications, 21, 1253–1278.

Ernst, O. K., Bartol, T. M., Sejnowski, T. J. and Mjolsness, E. (2019). Learning moment closure in reaction-diffusion systems with spatial dynamic Boltzmann distributions. Physical Review E, 99, 063315.

Forgues, F., Ivan, L., Trottier, A. and McDonald, J. G. (2019). A Gaussian moment method for polydisperse multiphase flow modelling. Journal of Computational Physics, 398, 108839.

Galvis, J. A., Corzo, C. A. and Machado, G. (2022). Modelling and assessing additional transmission routes for porcine reproductive and respiratory syndrome virus: Vehicle movements and feed ingredients. Transboundary and Emerging Diseases, 69, e1549–e1560.

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

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

Goldstein, M. and Rougier, J. (2006). Bayes linear calibrated prediction for complex systems. Journal of the American Statistical Association, 101, 1132–1143.

Gopalan, G. and Wikle, C. K. (2022). A higher-order singular value decomposition tensor emulator for spatiotemporal simulators. Journal of Agricultural, Biological and Environmental Statistics, 27, 22–45.

Gracy, R. B. (2020). Surrogates : Gaussian Process Modeling, Design, and Optimization for the Applied Sciences. Boca Raton: Chapman and Hall/CRC.

Higdon, D., Gattiker, J., Williams, B. and Rightley, M. (2008). Computer model calibration using high-dimensional output. Journal of the American Statistical Association, 103, 570–583.

Higdon, D., Kennedy, M., Cavendish, J. C., Caofo, J. A. and Ryne, R. D. (2004). Combining field data and computer simulations for calibration and prediction. SIAM Journal on Scientific Computing, 26, 448–466.

Hooten, M. B., Leeds, W. B., Fiechter, J. and Wikle, C. K. (2011). Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models. Journal of Agricultural, Biological, and Environmental Statistics, 16, 475–494.

IBGE (2021). Portal do IBGE. https://www.ibge.gov.br/en/home-eng.html. [Accessed 21 July 2022].

Isham, V. (1991). Assessing the variability of stochastic epidemics. Mathematical Biosciences, 107, 209–224.

Jones, C. M., Jones, S., Petrasova, A., Petras, V., Gaydos, D., Skrip, M. M. et al. (2021). Iteratively forecasting biological invasions with PoPS and a little help from our friends. Frontiers in Ecology and the Environment, 19, 411–418.

Kennedy, M. C. and O’Hagan, A. (2001). Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 425–464.

Kolod, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. SIAM Review, 51, 455–500.

Kuehn, C. (2016). Moment closure—a brief review. In: Control of Self-Organizing Nonlinear Systems, Understanding Complex Systems, Cham, Switzerland: Springer.

Leeds, W. B., Wikle, C. K. and Fiechter, J. (2014). Emulator-assisted reduced-rank ecological data assimilation for nonlinear multivariate dynamical spatio-temporal processes. Statistical Methodology, 17, 126–138.

Lloyd, A. L. (2004). Estimating variability in models for recurrent epidemics: Assessing the use of moment closure techniques. Theoretical Population Biology, 65, 49–65.

Massoud, E. C. (2019). Emulation of environmental models using polynomial chaos expansion. Environmental Modelling & Software, 111, 421–431.

Murrell, D. J., Dieckmann, U. and Law, R. (2004). On moment closures for population dynamics in continuous space. Journal of Theoretical Biology, 229, 421–432.
Paeng, S.-H. and Lee, J. (2017). Continuous and discrete SIR-models with spatial distributions. *Journal of Mathematical Biology*, 74, 1709–1727.

PAHO (2016). Timeline of the emergence of Zika virus in the Americas. www.paho.org. [Accessed 1 August 2022].

PAHO (2021). Data - BRA Zika Report. www.paho.org. [Accessed 1 August 2022].

Pratola, M. and Chkrebtii, O. (2018). Bayesian calibration of multistate stochastic simulators. Statistica Sinica, 28, 693–719.

Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K. and Jeff Wu, C. F. (2006). Building surrogate models based on detailed and approximate simulations. *Journal of Mechanical Design*, 128, 668–677.

Reich, B. J., Kalendra, E., Storlie, C. B., Bondell, H. D. and Fuentes, M. (2012). Variable selection for high dimensional Bayesian density estimation: Application to human exposure simulation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61, 47–66.

Sadeghieh, T., Sargeant, J. M., Greer, A. L., Berke, O., Dueymes, G., Gachon, P. et al. (2021). Zika virus outbreak in Brazil under current and future climate. *Epidemics*, 37, 100491.

Sharkey, K., Kiss, I., Wilkinson, R. and Simon, P. (2015). Exact equations for SIR epidemics on tree graphs. *Bulletin of Mathematical Biology*, 77, 614–645.

Shutt, D. P., Manore, C. A., Pankavich, S., Porter, A. T. and Del Valle, S. Y. (2017). Estimating the reproductive number, total outbreak size, and reporting rates for Zika epidemics in South and Central America. *Epidemics*, 21, 63–79.

Thakur, A. and Chakraborty, S. (2022). A deep learning based surrogate model for stochastic simulators. *Probabilistic Engineering Mechanics*, 68, 103248.

Wang, L., Zhao, H., Oliva, S. M. and Zhu, H. (2017). Modeling the transmission and control of zika in brazil. *Scientific Reports*, 7, 1–14.

Whittle, P. (1957). On the use of the normal approximation in the treatment of stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 19, 268–281.

Widgren, S., Bauer, P., Eriksson, R. and Engblom, S. (2019). Siminf: An R package for data-driven stochastic disease spread simulations. *Journal of Statistical Software*, 91, 1–42.

Zare, A., Ozdemir, A., Iwen, M. A. and Aviyente, S. (2018). Extension of PCA to higher order data structures: An introduction to tensors, tensor decompositions, and tensor PCA. *Proceedings of the IEEE*, 106, 1341–1358.

Zhao, S., Musa, S. S., Fu, H., He, D. and Qin, J. (2019). Simple framework for real-time forecast in a data-limited situation: The zika virus (zikv) outbreaks in brazil from 2015 to 2016 as an example. *Parasites & Vectors*, 12, 1–13.