SIAM J. SCI. COMPUT.
Vol. 41, No. 4, pp. A2212–A2238




# PARAMETER AND UNCERTAINTY ESTIMATION FOR DYNAMICAL SYSTEMS USING SURROGATE STOCHASTIC PROCESSES*

MATTHIAS CHUNG†, MICKAËL BINOIS‡, ROBERT B. GRAMACY§, JOHNATHAN M. BARDSLEY¶, DAVID J. MOQUIN‖, AMANDA P. SMITH#, AND AMBER M. SMITH#

**Abstract.** Inference on unknown quantities in dynamical systems via observational data is essential for providing meaningful insight, furnishing accurate predictions, enabling robust control, and establishing appropriate designs for future experiments. Merging mathematical theory with empirical measurements in a statistically coherent way is critical and challenges abound, e.g., ill-posedness of the parameter estimation problem, proper regularization and incorporation of prior knowledge, and computational limitations. To address these issues, we propose a new method for learning parameterized dynamical systems from data. We first customize and fit a surrogate stochastic process directly to observational data, front-loading with statistical learning to respect prior knowledge (e.g., smoothness), cope with challenging data features like heteroskedasticity, heavy tails, and censoring. Then, samples of the stochastic process are used as "surrogate data" and point estimates are computed via ordinary point estimation methods in a modular fashion. Attractive features of this two-step approach include modularity and trivial parallelizability. We demonstrate its advantages on a predator-prey simulation study and on a real-world application involving within-host influenza virus infection data paired with a viral kinetic model, with comparisons to a more conventional Markov chain Monte Carlo (MCMC) based Bayesian approach.

**Key words.** inverse problems, dynamical systems, Gaussian process, parameter estimation, uncertainty estimation, viral kinetic model

**AMS subject classifications.** 60G15, 62F10, 62F15, 65L09, 65L05, 92-08

**DOI.** 10.1137/18M1213403

## 1. Introduction and background.

Standard data fitting for dynamical systems uses a least squares framework in which the Euclidean distance between data and model, via a computer-implemented solver, is minimized. Parameter estimation schemes often begin with an initial guess on the values of each coordinate, followed by repeated solving of the dynamical system via numerical integration techniques as directed by a search algorithm, until a termination criterion is satisfied. With many data sets, including the influenza virus infection data discussed here, the optimization

*Submitted to the journal's Methods and Algorithms for Scientific Computing section September 11, 2018; accepted for publication (in revised form) May 7, 2019; published electronically July 16, 2019.

https://doi.org/10.1137/18M1213403

**Funding:** This work was supported by USDA National Institute of Food and Agriculture (grant 2016-08687), by NIH National Institute of Allergy and Infectious Diseases (grants AI125324 and AI100946), and by ALSAC. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of these entities.

†Department of Mathematics, Computational Modeling and Data Analytics Division, Academy of Integrated Science, Virginia Tech, Blacksburg, VA 24061 (mcchung@vt.edu).

‡Booth School of Business, University of Chicago, Chicago, IL 60637 (mickael.binois@chicagobooth.edu).

§Department of Statistics, Virginia Tech, Blacksburg, VA 24061 (rbgramacy@vt.edu).

¶Department of Mathematical Sciences, University of Montana, Missoula, MT 59812 (bardsleyj@mso.umt.edu).

‖Department of Internal Medicine, University of Tennessee Health Science Center, Memphis, TN 38103 (dmoquin@uthsc.edu).

#Department of Pediatrics, University of Tennessee Health Science Center, Memphis, TN 38103 (amanda.smith@uthsc.edu, amber.smith@uthsc.edu).

problem is ill-posed. As a result, meaningful parameter and uncertainty estimates can remain elusive [25].

Our proposed method incorporates data and prior knowledge to generate surrogate data, which includes the statistics of the data as well as prior assumptions on the dynamics of the system. Here, we propose a new apparatus, which front-loads with statistical modeling as a means of enhancing transparency in the flow of data-derived information and prior assumptions through to inference and uncertainty quantification, while simultaneously preserving computational tractability. We begin with a continuous, fitted stochastic process to generate feasible data approximations, e.g., smoothness-preserving dynamics. Then, rather than data fitting, we perform sample process fitting. That is, we use a least squares framework to fit samples of the stochastic process (i.e., solution trajectory) rather than fitting the data itself. A major advantage of this framework is that regularization may be introduced in a more intuitive and "up-front" manner regarding the propagation of uncertainty via mapping posterior predictive surfaces to distributions on model parameters. Finally, relative to similar alternatives, such as straightforward Bayesian modeling [64] or more elaborate Bayesian computer model calibration schemes [39, 37], both of which require serial computation via Markov chain Monte Carlo (MCMC), our methodology is a simple Monte Carlo scheme. Therefore, it offers the potential for vast (and embarrassingly parallel) distributed computation.

Our methodological developments are motivated by experimental data and a parameterized dynamical model describing in vivo viral load kinetics during influenza virus infection. As we illustrate, the data present some parameter estimation challenges that can thwart straightforward fitting methods typically used to infer unknown model parameters. One challenge is that the samples from murine infection are destructive and serial time sampling of individuals is unavailable to determine viral loads [58, 63]. To address this challenge, data are collected for several animals and at many times after infection. The small heterogeneity in sampling indicates high reproducibility and a robust curve that can be used to estimate the average time course of virus dynamics [63]. Even so, we often rely on statistically formulated prior beliefs about the dynamics in order to match the data with rigid assumptions posed by the mathematical models. Additional complications arise when the data exhibit input-dependent noise (i.e., they are heteroskedastic), have heavy tails (i.e., they are leptokurtic), and/or there is a lower limit of detection (LOD) within the measurement assay (i.e., censoring). Our aim is to develop an inferential scheme that can cope with such nuances and data deficiencies yet remain computationally tractable and offer a transparent filtering of domain knowledge and information from data.

We highlight three pillars of novelty in this paper. The first is a new framework based on stochastic processes for parameter and uncertainty estimation. Although the spirit of our approach here bears some resemblance to some others in the recent literature [23, 13, 31], there are some important differences which are discussed further in our review. The second is how we construct, and fit, the stochastic process. We illustrate how the canonical surrogate modeling framework, via stationary Gaussian processes (GPs), is overly simplistic except for the most basic examples one could entertain in this framework. To accommodate our motivating influenza setting, we developed extensions to a recently developed library called `hetGP` [7] to support heavy-tailed observations, and we designed a custom scheme for coping with missing data under monotonicity assumptions which are crucial for obtaining reasonable surrogate fits from a biological perspective. The final pillar is our application of the framework to real data on an influenza virus, offering stark contrast to results obtained from a

more conventional MCMC-based Bayesian framework.

The remainder of the paper is organized as follows. We complete our introductory section below by providing background on parameter estimation methods for dynamical systems by way of motivating our new approach, which is briefly outlined before a full treatment in section 2. Here we detail the influenza virus infection data, the within-host viral kinetic model, and the challenges associated with heteroskedasticity, outliers, and censored observations to motivate the development of a tailored modeling scheme, based on GPs, in section 3. Section 4 provides detailed empirical work, first via a simulation study as a means of illustration, and then on our motivating influenza example. We conclude with a brief discussion in section 5.

**1.1. Solvers and inference.** Solving ordinary differential equation (ODE) model-constrained parameter estimation problems may be computationally challenging for various reasons, including having a limited number of observations, high levels of noise in the data, chaotic system dynamics, nonlinear system models, and large numbers of unknown parameters. Many of these challenges appear in biological systems [71, 3]; hence, parameter estimation for dynamical systems for biological applications is of high interest and an active area of research [72, 9, 50, 51, 19].

A classic parameter estimation problem may be stated as follows:

$$\min_{\mathbf{p}\in\mathcal{P}} \mathcal{J}(\mathbf{p}) = \|\mathbf{s}(\mathbf{y}(\mathbf{p})) - \mathbf{d}\|_2^2 + \mathcal{R}(\mathbf{p}) \qquad \text{subject to } \mathbf{y}' = \mathbf{f}(t, \mathbf{y}, \mathbf{p}). \tag{1}$$

A diagram of this process is shown in Figure 1. Here, $\mathbf{y} : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^n$ is a solution of a parameter-dependent ODE $\mathbf{y}' = \mathbf{f}(t, \mathbf{y}, \mathbf{p})$. We assume that the state solution $\mathbf{y}$ is uniquely determined for any $\mathbf{p} \in \mathcal{P} \subset \mathbb{R}^m$, where $\mathcal{P}$ is the feasible parameter space, e.g., $\mathbf{p}$ may contain nonnegative growth rates $p_j \geq 0$. In biological systems, we often consider initial value problems and may include the initial condition $\mathbf{y}(t_0) \in \mathbb{R}^n$ in addition to the unknown model parameters, $\mathbf{p}$. Throughout this work, we assume that $\mathbf{p}$ contains model parameters and initial conditions $\mathbf{y}(t_0)$ such that the initial value problem is uniquely determined. The functional $\mathbf{s}$ is an operator that maps the state solution $\mathbf{y}$ onto an observation space $\mathcal{D}$, here $\mathcal{D} \subset \mathbb{R}^N$, where $\mathbf{d} = [d_1, \ldots, d_N] \in \mathcal{D}$ are the available experimental measurements at (time) points $\mathbf{t} = [t_1, \ldots, t_N]^\top$. For example, in predator-prey systems, one may be able to monitor the predator population at certain times while observation of the prey is not available. In other applications, such as acoustics, observations $\mathbf{d}$ are often made in the frequency domain and not on the state space. Here, $\mathbf{s}$ represents a Fourier transform by mapping the state dynamics into the frequency domain.

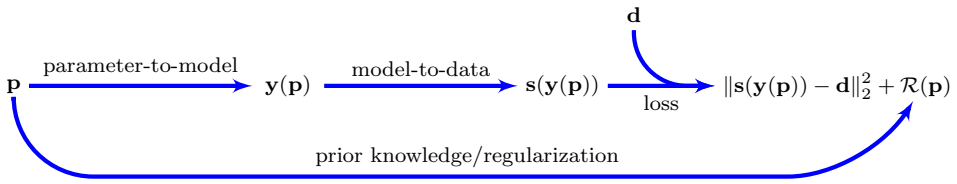


FIG. 1. *Illustration of the ingredients of a classical parameter estimation problem. Parameter* $\mathbf{p}$ *defines the specific model* $\mathbf{y}$*, which then gets mapped by* $\mathbf{s}$ *onto the data* $\mathbf{d}$*, and finally measured with a quality measure* $\mathcal{J}$*, which may include prior knowledge* $\mathcal{R}$ *about* $\mathbf{p}$*. For the inverse problem we want to find a* $\widehat{\mathbf{p}}$ *with the best quality measure* $\mathcal{J}$*, given* $\mathbf{y}$*,* $\mathbf{s}$*, regularization* $\mathcal{R}$*, and data* $\mathbf{d}$*.*

The overall quality $\mathcal{J}$ as stated in (1) is a measure which combines the data fit and prior knowledge. In (1), we assume that the discrepancy between the model

prediction $\mathbf{s}(\mathbf{y}(\mathbf{p}))$ and the data $\mathbf{d}$ is given by the Euclidean norm $\|\cdot\|_2$. The methodology we present is not tailored to this choice. Other loss functions, or norms, may be utilized. However, we prefer the $\ell_2$-norm for its computational advantages and familiarity. Prior knowledge on the parameters $\mathbf{p}$ may also be included in the form of an additive regularization term $\mathcal{R}(\mathbf{p})$, where regularization prevents ill-posed problems from overfitting the data $\mathbf{d}$ (see, for example, [35, 3]). In a Bayesian framework, the minimization of $\mathcal{J}$ can be seen as an evaluation of a maximum a posteriori (MAP) estimate, where the data fitting term originates from a negative log-likelihood function $\|\mathbf{s}(\mathbf{y}(\mathbf{p})) - \mathbf{d}\|_2^2$ and the regularization term $\mathcal{R}$ arises from the negative log of a prior distribution; see [15] for details. Here, the 2-norm in the data fitting term then corresponds to Gaussian assumptions on the noise $\mathbf{s}(\mathbf{y}(\mathbf{p})) - \mathbf{d}$.

Parameter estimation for the setup described in Figure 1 and (1) amounts to solving an *inverse problem*. Our inferential scheme is tailored to the setting where one has *repeated* observations $\mathbf{d} = [d_1, \ldots, d_N]^\top$ of the states given at discrete time points $\mathbf{t} = [t_1, \ldots, t_N]^\top$, as such a setup appears frequently in a diverse set of biological applications (e.g., as in [19, 27, 58, 60, 63]). However, even with such regularity in the data, establishing a coherent parameter and uncertainty estimation scheme with underlying dynamical systems for such observations is challenging. Thus, new computationally tractable methodologies could be of great value.

In the literature, various numerical methods have been proposed to solve model constrained optimization problems such as (1). Focusing in particular on biological systems, "single shooting" approaches are often utilized [65, 5]. For this strategy, first an initial guess for $\mathbf{p}^{(0)}$ is used to numerically solve the initial value problem using numerical ODE solvers like Runge–Kutta and Adams–Bashforth [33, 34]. Next, the misfit $\mathcal{J}(\mathbf{p}^{(0)})$ is computed, and depending on the optimization strategy (e.g., gradient based or direct search, say), a new parameter vector $\mathbf{p}^{(1)}$ is chosen such that $\mathcal{J}(\mathbf{p}^{(1)}) < \mathcal{J}(\mathbf{p}^{(0)})$. Then, in an iterative fashion, $\mathbf{p}^{(k)}$, $k = 1, \ldots,$ is updated until a $\widehat{\mathbf{p}}$ is found subject to predetermined optimality criteria [48]. Notice that each optimization step requires at least one ODE solve. However, this may significantly increase depending on the optimization scheme employed. One faces the trade-off between sophisticated optimization schemes with improved convergence properties versus simplistic optimization schemes with potential modest improvement. If local optimization methods are utilized, globalization is often attempted by Monte Carlo sampling of the search space, i.e., repeated local optimization with random initial guesses [20]. Empirically, the global minimizer $\widehat{\mathbf{p}}$ is chosen from the set of local minimizers obtaining the minimal objective function value. Certainly, other global optimization strategies may be employed, e.g., simulated annealing [41, 63], evolutionary algorithms [57], or particle swarm optimization methods [38], to name a few.

Besides single shooting methods to solve the optimization problem (1), which benefit from a straightforward implementation, more sophisticated multiple shooting methods and principal differential analysis offer the potential for improved robustness of the point estimates [28, 4, 10, 72, 50, 19, 18]. However, the implementation and computation present unique challenges, and thus, we restrict our attention to single shooting methods here. Nevertheless, each of the methods mentioned above only provides point estimates $\widehat{\mathbf{p}}$, and therefore extra machinery is required in order to quantify uncertainty. One approach in the literature is to deploy local sensitivity analysis [53, 46, 16]; another utilizes bootstrapping methods [66]. However, Bayesian methods are also gaining traction by using accelerated MCMC [14, 64, 68]. Despite their advantages of naturally providing uncertainty estimates, MCMC methods in this context can be particularly computationally burdensome because numerical ODE

solvers are embedded in accept-reject calculations and the Markov property limits the scope for potential relief via parallelization.

Most data assimilation problems can be stated as a parameter estimation problem of the form in (1), where data assimilation methods such as variational methods, Kalman filters, ensemble Kalman filters, etc., integrate data with mathematical models to perform prediction, typically on future states; see [2, 42]. Although our methods may be applied and bear similarities to variational data assimilation methods, a full discussion of these methods is beyond the scope of this work. The interested reader is referred to various recent developments [21, 22, 12].

**1.2. A new approach.** A statistically sound strategy avoiding ill-posedness[1] of the problem and, therefore, multiple local minima—mitigating the extent to which Monte Carlo search must be utilized—is to introduce prior knowledge in terms of regularization $\mathcal{R}$. Note that $\mathcal{R}$ may be derived from the Bayesian framework and correspond to the negative logarithm of the prior probability density function of $\mathbf{p}$; see [14]. Without proper tuning, however, regularization may bias heavily towards prior knowledge, and, thus, parameter estimates are of little value. Moreover, regularization terms, such as standard Tikhonov $\mathcal{R}(\mathbf{p}) = \lambda \|\mathbf{p}\|_2^2$ with $\lambda \geq 0$, may be inappropriate. This is particularly true for dynamical systems (e.g., biological systems) because parameter values are largely unknown. On the other hand, prior knowledge of state variable dynamics is often readily available, where the dynamics of $\mathbf{y}$ may be expected to have certain behavior. For instance, state dynamics might be monotonic, some states may have limited scales, or smoothness or other known dynamics are encountered, such as fast transient behavior and a slow steady state convergence. As an example, insulin dynamics may have a limited length scale upon intravenous glucose administration due to limited storage capacities and production rates of $\beta$ cells [27].

Gong et al. [30] proposed regularization terms for dynamical systems which introduce smoothness on the state variables $\mathbf{y}$. However, as mentioned above, tuning associated regularization parameters may be computationally challenging, and determining how to propagate uncertainty arising from such procedures is not straightforward.

By way of a potential remedy, we propose to implicitly introduce regularization on the state dynamics $\mathbf{y}$ by imposing a surrogate stochastic process on observations from the "data-generating process." Then, using sample realizations from the fitted surrogate, we perform inference and estimate uncertainties of $\widehat{\mathbf{p}}$ using standard parameter estimation methods, e.g., via "single shooting" with numerical ODE and least squares solvers. This blends prior assumptions on the dynamics with the observed dynamics. Although such a scheme is simple to describe at a high level, the devil is in the details when it comes to diligent application. With the ultimate goal of providing meaningful uncertainty estimates around $\widehat{\mathbf{p}}$, our new method demands three key ingredients that require substantial methodological development: (1) determining scientifically appropriate yet implementationally pragmatic mechanisms for including prior knowledge about the biological systems of interest in the surrogate stochastic process; (2) coping with input-dependent noise under potentially heavy-tailed error distributions efficiently, in both statistical and computational senses; and (3) handling censored observations in the data.

Toward that end, we developed extensions to a so-called heteroskedastic Gaussian process (`hetGP`) [8] to encourage trajectories of a certain shape, handled Student-$t$

[1] A problem is ill-posed if a solution does not exist, is not unique, or does not depend continuously on the data [32].

noise distributions, and developed an imputation (or data augmentation) scheme to cope with censored observations. With an appropriate surrogate stochastic process in hand, we illustrate how inference over unknown parameters to the dynamical system is a straightforward application of "single shooting" Monte Carlo, mapping posterior predictive samples into samples of parameters via the ergodic theorem.

**2. Problem setup and review.** The basic setup of our proposed methodology is as follows. We fit a surrogate to data $\mathbf{d}$, comprised of measured observations from a physical (in our case, biological) process at a discrete/limited number of indices (in our case time). The fitting mechanism incorporates prior information about the physical process as a means of regularization. Then, we generate a set of continuous surrogate data $\{\mathbf{g}_j(t)\}_{j=1}^J$, drawn as samples from the fitted stochastic process. Each realization of this sample, indexed by $j$, possesses the essence of the data without qualities that challenge its direct use in ODE solvers. That is, surrogate data derived from the nonparametric posterior enjoy an added degree of regularity owing to the prior, e.g., smoothness or monotonicity, can be furnished without intrinsic noise, and yet exhibit all other sources of variability extrinsic to the true data generating mechanism. For each sample $\mathbf{g}_j(t)$ we obtain an estimate $\widehat{\mathbf{p}}_j$ by solving the optimization problem

$$\widehat{\mathbf{p}}_j = \arg\min_{\mathbf{p}} \|\mathbf{s}(\mathbf{y}(\mathbf{p})) - \mathbf{g}_j\|_{\mathcal{L}_2}^2 \quad \text{subject to } \mathbf{y}' = \mathbf{f}(t, \mathbf{y}, \mathbf{p}) \quad \text{for } j = 1, \ldots, J. \tag{2}$$

Here, $\|\cdot\|_{\mathcal{L}_2}$ denotes the continuous $\mathcal{L}_2$-norm on a closed interval $[a, b]$. Note that in contrast to (1), where $\mathbf{s}$ projects the state solution onto a data vector, here $\mathbf{s}$ projects $\mathbf{y}$ from the state space onto a function space. In the special case where all states are observed, $\mathbf{s}$ may simplify to the identity map. We assume that optimization problem (2) has a unique solution $\mathbf{p}_j$. Hence, we assume that the stochastic process adds regularity to the optimization problem (2), as empirically observed by our numerical investigations in section 4. Standard regularization terms $\mathcal{R}(\mathbf{p})$ may be additionally included in optimization problem (2) to obtain a unique solution $\widehat{\mathbf{p}}_j$ as introduced in (1); see [19]. Problems with nonunique solutions to (2) are beyond the scope of this work; here, the interested reader is referred to [43]. The set $\{\widehat{\mathbf{p}}_j\}_{j=1}^J$ defines a distribution of estimates of the underlying $\mathbf{p}_{\text{true}}$.

Our approach bears similarities to utilizing stochastic processes for parameter estimation to existing methods such as the ones discussed in [23, 13]. One key difference, however, is that while we use stochastic processes on the data and aim to parallelize parameter and uncertainty estimation, the methods mentioned above utilize stochastic processes on the model and aim to accelerate model evaluations. Our method has much in common with *randomized maximum likelihood* and bootstrapping [17, 24], which both randomize the data and use Monte Carlo samples to gain uncertainty estimates. A key additional component in our method is the introduction of a suitable family of stochastic processes, encoding domain-specific (i.e., biologically defensible) prior knowledge of the system, to extend the discretely observed data to "continuous" measurements.

Through the surrogate and subsequent optimization(s), our method filters data from prior and likelihood to posterior distribution on the unknown parameters $\mathbf{p}_{\text{true}}$. The model (combining likelihood and prior) is comprised of a statistical component via the fitted surrogate and a mathematical component via the choice of $\mathbf{s}(\mathbf{y}(\cdot))$. Note that priors are not being placed directly on $\mathbf{p}_{\text{true}}$. Prior knowledge about the entire system is encapsulated in choices made for the underlying stochastic process and, thereby, transfers into the samples $\mathbf{g}_j$ and carries over to the estimated state variable $\mathbf{y}(\cdot, \widehat{\mathbf{p}}_j)$.

For instance, if the stochastic process exhibits smooth dynamics in certain areas, $\widehat{\mathbf{p}}_j$ will be selected, for which this feature is prominent in the state variable $\mathbf{y}(\cdot, \widehat{\mathbf{p}}_j)$. If samples from the surrogate come from a Bayesian posterior, then the set $\{\widehat{\mathbf{p}}_j\}_{j=1}^{J}$ is comprised of samples from the posterior distribution of $\mathbf{p}_{\text{true}}$ via the ergodic theorem, as the latter optimization step(s) can be interpreted as a deterministic function, $\mathbf{h}$, of the surrogate draws, $\widehat{\mathbf{p}}_j = \mathbf{h}(\mathbf{g}_j)$.

Characterizing prior-to-posterior updating is a little less direct in this setup, compared to the canonical Bayesian framework, because the prior measure is not placed directly on $\mathbf{p}_{\text{true}}$ but rather on the function space of $\mathbf{g}$; however, that does not make posterior inferences any less Bayesian. In fact, this setup offers distinct advantages. Rather than asking a practitioner to choose a family of densities for $\mathbf{p}_{\text{true}}$ encoding their prior beliefs, which can be a daunting and inexact exercise in elucidation [47], especially when the relationship between parameter and the underlying physical/biological process is weak or theoretical, our framework instead asks the practitioner to loosely characterize otherwise nonparametric random functions (via smoothness, noise, monotonicity, etc.) that respect qualities of the data generating mechanism. Our experience is that this latter form of prior knowledge elucidation is much easier to carry out. Smoothness and other covariance properties may be prescribed by selecting appropriate GP kernel functions like the Matèrn. Other modifications like monotonicity in the dynamics require additional effort and are detailed in section 3.3 in response to needs arising in our motivating influenza application. If so inclined, one could remove a prior measure on $\mathbf{p}_{\text{true}}$ by applying the same posterior "surrogate data" sampling procedure we describe to sample paths obtained instead from prior processes, i.e., without first conditioning on the data. An example is provided in section 4.

In this setup, all of the learning transpires in the first step (i.e., fitting the surrogate data) because this is the only place data observations are involved. Thus, one can argue that the optimization steps (equation (2)) amount to post (learning) processing. From an implementation or algorithmic perspective that characterization is prescriptive. This is what provides the Bayesian posterior interpretation described above. However, that description unfairly diminishes the role of those latter steps in the overall inferential procedure. Moreover, there is a feedback loop between the prior elements from the surrogate and those from the mathematical modeling components. The former is tasked with producing surrogate data of the form assumed as prerequisites for the latter. Therefore, at least conceptually, the two prior elements are intimately linked. Care is required when choosing appropriate components for each stage in the process, in particular depending on the nature of the data generating mechanism. Therefore, in what follows, we describe the data $\mathbf{d}$, appropriate mathematical models for that process $\mathbf{y}$, and choices for appropriate surrogates in our setting—essentially inverting the order of operations described above.

**2.1. Influenza data.** Influenza viruses are a frequent cause of lower respiratory tract infections and cause over 15 million infections that result in 200,000 hospitalizations each year [69]. Infections vary in severity from mild to lethal, where the H1N1 strain in the 1918 "Spanish Flu" pandemic claimed over 40 million lives. Despite the prevalence of influenza viruses, the interactions between the virus and host remain poorly understood.

Influenza virus infections are typically acute and self-limiting with short incubation (about 2 days) and infectious periods (typically 4–7 days) [67]. Although the majority of infections are confined to the upper respiratory tract, migration to the

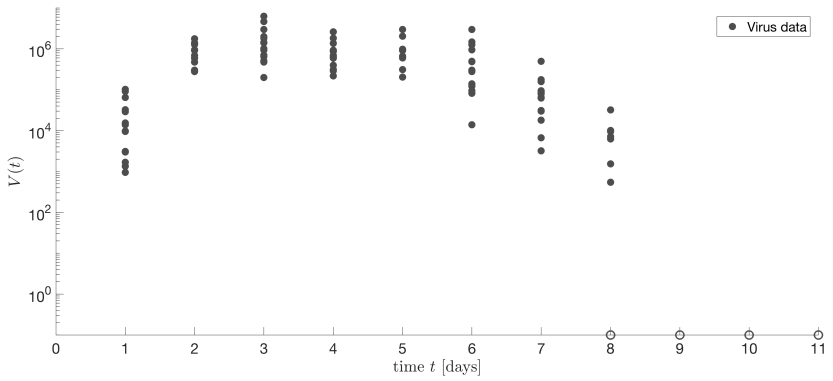


FIG. 2. *Viral titers from the lungs of individual mice (black dots) infected with* 75 $\text{TCID}_{50}$ *influenza A/Puerto Rico/8/34 (H1N1) (PR8)* [63]. *Virus is undetectable in some mice at* 8 *d post-infection (pi) and for all mice at* $t = 9, 10,$ and 11 *d pi (black circles). These are plotted at* $10^{-1}$ $\text{TCID}_{50}$ *for visualization.*

lower respiratory tract can result in severe pneumonia. To gain deeper insight into infection mechanisms and the rates of viral growth and decay, mathematical models have been developed and paired with experimental data from humans and animals (e.g., reviewed in [62]). Parameter estimation remains an important aspect of these studies in order to extract meaningful insight about the infection kinetics. Here, we use one data set and a simple model that accurately describes influenza virus kinetics to illustrate the method here [63]. The viral load data that we use was obtained from infecting groups of BALB/cJ mice with 75 $\text{TCID}_{50}$ (50% tissue culture infectious dose, the dose at which 50% of cell cultures are infected with virus) influenza A/Puerto Rico/8/34 (PR8) at time $t = 0$ (time of infection) and measuring viral loads via $\text{TCID}_{50}$ at daily time points $t = 1, \ldots, 11$ *days postinfection,* typically abbreviated as "d pi" [63]. At each time point (1–11 d pi), samples were collected from 15 individual mice. Thus, the data vector $\mathbf{d} \in \mathbb{R}^{165}$ comprises the viral load data. The data is shown in Figure 2. The open circle at 8 d pi may be censored values or true zero values, while the data at 9–11 d pi are thought to be exclusively and reproducibly zeros. For simplicity and consistency, we will treat all zero data points as censored values because the $\text{TCID}_{50}$ assay is typically not sensitive enough to capture low viral loads. We arbitrarily set the LOD at 200 $\text{TCID}_{50}$.

Anticipating the modeling developments (discussed later), observe the following from the figure: The trajectory would appear to be unimodal, which is supported by the biology, although there is no clear way to "trace" this from the raw data. This is because the samples are destructive, meaning it is not possible to collect a trajectory of measurements for a single mouse over time. Hence, simple interpolation strategies of the data averages are not suitable; they will introduce unwanted bias. In addition, observe that the data exhibit a degree of heteroskedasticity and/or are heavy-tailed. This is exemplified by the narrow spread of points nearby time point $t = 4$ d, versus a wider spread at time $t = 1$ d and $t \geq 6$ d. The heterogeneity in these time points is true biological heterogeneity and corresponds to the times when the virus is rapidly increasing/decreasing [63]. Although this is expected, the pattern of spread is at odds with a typical mean-variance relationship (i.e., spread goes up when mean goes up). Thus, simple transformations are unlikely to help. Finally, it is known that the

viral load starts at zero and, therefore, so does the variance [63]. This is because the virus rapidly infects cells or is cleared in the early stages of infection (0–4 hours postinfection) [63]. Similarly, the viral load declines to zero as time increases past 8 d pi. Although these features may easily be accommodated via latent *mean* quantities (described later), this exacerbates the heteroskedastic nature of the data and requires more nuanced treatment.

**2.2. Mathematical modeling.** Because viral dynamics are rapid and complex, studying influenza virus infections with experimental models alone is challenging. Thus, mathematical models have been employed to help identify and detail the mechanisms responsible for controlling viral growth and resolving the infection (reviewed in [62]). These studies have shown that viral load dynamics can be accurately described using 3–4 equations without inclusion of specific innate and adaptive immune responses. Indeed, we recently developed a model that accurately recapitulates the viral load data, including the rapid clearance of virus between 7–9 d pi [63]. The model tracks susceptible epithelial ("target") cells $T$, two classes of infected cells $I_1$ and $I_2$, and virus $V$. In this model, target cells become infected with the virus at rate $\beta V$ per cell. Once infected, these cells enter an eclipse phase $I_1$ at rate $\kappa$ per cell before transitioning to produce the virus at rate $\rho$ per cell $I_2$. Viral loads are cleared at rate $c$ and virus-producing infected cells $I_2$ are cleared in a density-dependent manner with maximal rate $\delta$, where $K_d$ is the half-saturation constant. The following system of differential equations describes these dynamics [63]:

$$\begin{aligned} T' &= -\beta TV, \\ I_1' &= \beta TV - \kappa I_1, \\ I_2' &= \kappa I_1 - \frac{\delta I_2}{K_d + I_2}, \\ V' &= \rho I_2 - cV. \end{aligned} \tag{3}$$

The system is of the form $\mathbf{y}' = \mathbf{f}(t, \mathbf{y}, \mathbf{p})$, as considered in (1). The state variables are given by $\mathbf{y}(t) = [T(t), I_1(t), I_2(t), V(t)]^\top$. The parameters $\beta, \kappa, \delta, K_d, \rho, c$ and initial conditions $T(0), I_1(0), I_2(0), V(0)$ uniquely determine the initial value problem. Given an initial condition, this system can be solved numerically with standard ODE solvers; see [33, 34].

**2.3. Gaussian process surrogate modeling.** Gaussian process (GP) regression, or surrogate modeling, offers a nonparametric framework for estimating functions. GPs are typically trained on a vector of $N$ observations or outputs, $\mathbf{d} = [d_1, \ldots, d_N]^\top$, observed at design of input locations $\mathbf{t} = [t_1, \ldots, t_N]^\top \in \mathbb{R}^N$. The input and output spaces may be multidimensional; however, they are both scalar in the application here. The training data need not be ordered, but as in our discussion here where the inputs are times, we will presume that $t_j \leq t_{j+1}$. A GP is completely defined by its mean and covariance structure, which defines a multivariate normal (MVN) distribution on a finite collection of realizations of the outputs $\mathbf{d}$ as a function of inputs $\mathbf{t}$. Note that we generate GP posteriors $\mathbf{g}$ separately for each observed state $y_j$. For clarity of notation, however, we only present derivations for GP posteriors for a single state.

We assume a zero-mean GP prior for functions $\mathbf{g}$ generating $\mathbf{d}$, which is a common simplifying assumption in the computer simulation modeling literature, e.g., [54]. This has the effect of moving the modeling effort exclusively into the covariance structure, which is defined by a positive definite kernel $k(\cdot, \cdot)$ function and yields the

$N \times N$ covariance matrix of the MVN. The kernel is usually determined by spatial (e.g., Euclidean) distances in the input $t$-space, and within that realm there are many common forms. The power exponential and Matérn families are the most popular. Both of these contain a small number of hyperparameters that are usually learned from the data (for details see [52, 54]). The specification is completed by choosing a noise process on the output $\mathbf{d}$-variables, comprised of the likelihood component in this Bayesian model specification. Typically, this is independent and identically distributed (i.i.d.) Gaussian with variance $v(t)$. In some (mainly historical) computer modeling contexts, the simulations are deterministic. In this case, $v(t_j) = 0$. Stochastic simulations are increasingly common, and GP-based surrogate models usually treat the variance function $v$ as constant, which leads to a homoskedastic fit. However, our influenza virus infection example will benefit from a *heteroskedastic* modeling capability in addition to heavier-tailed noise distribution—a detail we will return to below. For now, we treat $v(t)$ generically.

The above description can be summarized by the following data-generating specification:

$$\mathbf{d} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_N), \quad \text{where} \quad \mathbf{K}_N \in \mathbb{R}^{N \times N} \text{ with } \quad (\mathbf{K}_N)_{ij} = k(t_i, t_j) + \delta_{ij} v(t_i), \tag{4}$$

where $\delta_{ij}$ is the Kronecker delta such that the diagonal of $\mathbf{K}_N$ is augmented with the "noise variance." This specification bundles GP prior $\mathbf{g}(\mathbf{t}) \sim \mathcal{N}(0, \mathbf{C}_N)$, where $\mathbf{C}_N$ is the same as $\mathbf{K}_N$ without the $\delta_{ij} v(t_i)$ term, and i.i.d. Gaussian likelihood term for the noise with variance $v(t_i)$. Any hyperparameters to $k(\cdot, \cdot)$ can be inferred through the likelihood implied by the MVN above, say, via maximum likelihood estimation (MLE). Perhaps the most distinctive feature of this setup is that, due to simple MVN conditioning rules, the posterior predictive distribution at a new input location site $t$ $(d(t)|\mathbf{d})$ is (univariate) Gaussian with parameters

$$\begin{aligned} \mu(t) &= \mathbb{E}\left(d(t) \mid \mathbf{t}, \mathbf{d}\right) = \mathbf{k}_N(t)^\top \mathbf{K}_N^{-1} \mathbf{d} \text{ and} \\ \sigma^2(t) &= \mathbb{V}\mathrm{ar}(d(t) \mid \mathbf{t}, \mathbf{d}) = k(t,t) + v(t) - \mathbf{k}_N(t)^\top \mathbf{K}_N^{-1} \mathbf{k}_N(t), \text{ with} \\ k_N(t, t') &= \mathbb{C}\mathrm{ov}(d(t), d(t') \mid \mathbf{t}, \mathbf{d}) = k(t,t') - \mathbf{k}_N(t)^\top \mathbf{K}_N^{-1} \mathbf{k}_N(t') + \delta_{t=t'} v(t), \end{aligned} \tag{5}$$

where $\mathbf{k}_N(t) = [k(t, t_1), \ldots, k(t, t_N)]^\top$. Denoised versions, providing the posterior distribution on states $\mathbf{g}|\mathbf{d}$, follow the same moments given in (2.3) with $v(t) = 0$. Vectorized versions, essentially tabulating the covariance structure described above into a full MVN structure, generalize these equations to a joint predictive distribution over a set $\mathcal{T}$ of new locations. One disadvantage, which is apparent from inspecting the equations above, is that the method can be computationally (and storage) intensive in the presence of moderately large data sizes (large $N$), owing to the cubic cost of matrix decomposition and the quadratic cost of storage for $\mathbf{K}_N$. Similar bottlenecks are in play when working with the likelihood for hyperparameter inference.

Excluding the deterministic case (i.e., using $v(t_j) = 0$), having replicate $d_i$ and $d_j$ observations at the same input locations, $t_i = t_j$, can be useful for separating signal from noise [8]. Our influenza virus infection data naturally has this feature. It turns out that having replication in the design also yields computational advantages [8]. Let $\bar{t}_i$, $i = 1, \ldots, n$, be the $n \leq N$ *unique* input locations, and let $d_i^{(j)}$ be the $j$th out of $a_i \geq 1$ replicates, i.e., $j = 1, \ldots, a_i$, observed at $\bar{t}_i$, where $\sum_{i=1}^n a_i = N$. Also, let $\bar{\mathbf{d}} = [\bar{d}_1, \ldots, \bar{d}_n]^\top$ stand for averages over replicates, $\bar{d}_i = \frac{1}{a_i} \sum_{j=1}^{a_i} d_i^{(j)}$. Then, it can be shown [8] that predictive equations (5) may be applied with $\bar{\mathbf{d}}$ in place of $\mathbf{d}$ and unique-$n$ matrices and vectors in place of full-$N$ ones using $\mathbf{k}_n(t) = [k(t, \bar{t}_1), \ldots, k(t, \bar{t}_n)]^\top$

and $(\mathbf{K}_n)_{ij} = k(\bar{t}_i, \bar{t}_j) + \delta_{ij} v(\bar{t}_i)/a_i$. In effect, this unique-$n$ scheme is utilizing an $\mathcal{O}(n)$ number of sufficient statistics for $\mathcal{O}(N)$ degrees of freedom.

**3. Influenza surrogate.** The standard GP setup falls short in the context of our motivating influenza virus infection data provided in Figure 2. Although replications are well handled by sufficient statistics, the variance function is unknown and may be changing throughout the input space. Furthermore, the tails may be heavier than Gaussian, and a portion of the data falls below the LOD and is censored. In the following, we provide extensions to remedy these shortcomings: first, we leverage recent advances in heteroskedastic regression (section 3.1); second, we present a novel approach to Student-$t$ errors in the heteroskedastic GP context (section 3.2). Finally, we develop a data-augmentation scheme for handling censored observations in section 3.3.

**3.1. Heteroskedastic GP modeling.** In practice, $v$ is seldom known, and the noise variance must be estimated from data, as with any other unknown quantity with the exception of [49]. A simple, yet effective, way of estimating the variance from data in a GP setting under replication is to use empirical, or moment-based, estimators. That is, select the diagonal matrix $\{\delta_{ij} v(t_i)\}$ in (4) as

$$\widehat{\mathbf{\Sigma}}_n = \operatorname{diag}\left(\hat{\sigma}_1^2/a_1, \ldots, \hat{\sigma}_n^2/a_n\right), \quad \text{where} \quad \hat{\sigma}_i^2 = \frac{1}{a_i - 1} \sum_{j=1}^{a_i} (d_i^{(j)} - \bar{d}_i)^2. \tag{6}$$

The resulting predictor is known in the literature as stochastic kriging (SK) [1]. SK has the benefit of accommodating heteroskedastic data, as each input's variance is estimated independently of the rest. It is also computationally advantageous due to working with $n$ rather than $N$ quantities. However, two disadvantages are (i) the method requires a minimum amount of replication ($a_i \gg 10$ is recommended) both for stability and for its asymptotic properties; and, perhaps more importantly for our needs, (ii) it yields no direct estimate of $v(t)$ out-of-sample, i.e., for a $t$ not in the training design $\mathbf{t}$. For the latter, it is recommended to fit a second independent GP to the logarithm of the empirical variances $(\hat{\sigma}_1^2/a_1, \ldots, \hat{\sigma}_n^2/a_n)$.

If two processes, one for the mean and another for the variance, are to be fit from the same data, then ideally the inference for those two unknowns would be performed jointly. Such approaches predate SK in the machine learning literature [29], where the (log) variances are treated as latent variables under a GP prior. Here, inference requires cumbersome MCMC calculations over all $N$ unknowns and implies a complexity of $\mathcal{O}(N^4)$, which is prohibitive even for modest $N$. Subsequently, researchers replaced the MCMC with point-based alternatives, e.g., using expectation maximization (EM), and related methods [40, 11]. However, none of these methods leveraged the computational savings that comes from having replicates in the design as in the case of SK.

A new method called `hetGP`, detailed in [8], offers a hybrid between SK and a joint modeling approach. For a stationary kernel, such as the ones mentioned above, we may equivalently write $k(t, t') = \nu c(t - t')$ such that $\mathbf{K} = \nu(\mathbf{C} + \mathbf{\Lambda}_n)$ with $(\mathbf{C})_{ij} = c(\bar{t}_i - \bar{t}_j)$ and $\log \mathbf{\Lambda}_n = \mathbf{C}_{(g)}(\mathbf{C}_{(g)} + g\mathbf{A}_n^{-1})^{-1}\mathbf{\Delta}_n$, where $\mathbf{C}_{(g)}$ is the analogue of $\mathbf{C}$ for the second GP with kernel $k_{(g)}$ and $\mathbf{A}_n = \operatorname{diag}(a_1, \ldots, a_n)$. That is, $\log \mathbf{\Lambda}_n$ is the prediction given by a GP based on latent variables $\mathbf{\Delta}_n = (\delta_1, \ldots, \delta_n)$ learned as additional hyperparameters alongside the second GP hyperparameters of $k_{(g)}$ and

its noise $g$. Using this notation, the predictive equations (5) can be represented as

$$\begin{aligned} \mu_n(t) &= \mathbf{c}_n(t)^\top \left(\mathbf{C}_n + \mathbf{\Lambda}_n \mathbf{A}_n^{-1}\right)^{-1} \bar{\mathbf{d}}, \\ \sigma_n^2(t) &= \nu \left(1 - \mathbf{c}_n(t)^\top \left(\mathbf{C}_n + \mathbf{\Lambda}_n \mathbf{A}_n^{-1}\right)^{-1} \mathbf{c}_n(t)\right). \end{aligned} \tag{7}$$

Unknown quantities may be inferred via the likelihood as follows. The MLE of $\nu$ is

$$\hat{\nu}_N = N^{-1} \left( N^{-1} \sum_{i=1}^{n} \frac{a_i}{\lambda_i} s_i^2 + \bar{\mathbf{d}}^\top (\mathbf{C} + \mathbf{A}_n^{-1} \mathbf{\Lambda}_n)^{-1} \bar{\mathbf{d}} \right), \tag{8}$$

with $s_i^2 = \frac{1}{a_i} \sum_{j=1}^{a_i} (d_i^{(j)} - \bar{d}_i)^2$. The remaining hyperparameters can be optimized using the concentrated joint log-likelihood

$$\begin{aligned} \log \tilde{L} = & -\frac{N}{2} \log \hat{\nu}_N - \frac{1}{2} \sum_{i=1}^{n} \left[ (a_i - 1) \log \lambda_i + \log a_i \right] - \frac{1}{2} \log |\mathbf{C} + \mathbf{A}_n^{-1} \mathbf{\Delta}_n| \\ & - \frac{n}{2} \log \hat{\nu}_{(g)} - \frac{1}{2} \log |\mathbf{C}_{(g)} + g \mathbf{A}_n^{-1}| + \text{const}, \end{aligned}$$

with $\hat{\nu}_{(g)} = n^{-1} \mathbf{\Delta}_n^\top (\mathbf{C} + g \mathbf{A}_n^{-1})^{-1} \mathbf{\Delta}_n$. Closed-form expressions of the derivatives are given in [8]. Besides being able to cope with input-dependent noise, the coupling of the processes means that no minimum amount of replication is required to perform inference and rely on the resulting predictions.

To provide an illustration of this method in a more controlled setting—we shall return to the motivating influenza example shortly—consider the motorcycle accident data [56], a stochastic simulation modeling the acceleration of the helmet of a motorcycle rider as a function of time just before and after an impact. It possesses both of the features targeted by the method above: light replication and input-dependent noise.

The left panel of Figure 3 shows the predictive surface from an ordinary GP fit (5), while the middle panel shows equivalently the predictive surface for `hetGP`, accommodating heteroskedasticity (7). We observe how the former is unable to capture the noise dynamics, whereas the latter captures the data better. The mean fits are similar but not identical. The right panel in Figure 3 shows the estimate of the noise level more directly, via estimates of the latent $\mathbf{\Delta}_n$ values. The dots in this panel show the empirical variances for the input locations which have two or more replicates. Basing an estimate of spatial variance on these values (i.e., following SK) would clearly leave something to be desired. For starters, these values are "choppy" over time—a feature not evident in a cursory inspection of the data. The `hetGP` method furnishes a smooth alternative within a unified mean-variance stochastic modeling framework without the SK requirement of a high degree of replication: all $a_i \gg 10$.

**3.2. Heavy-tailed heteroskedastic GP.** Sometimes the Gaussianity assumption can be overly rigid, particularly in the presence of heavy-tailed perturbations or outliers. Recall that the influenza virus infection data in Figure 2 exhibit this feature. In this setting, Shah, Wilson, and Ghahramani [55] showed that Student-$t$ processes generalize GPs and share most of their practical appeal. The trick to overcoming limitations from previous attempts [52] is to augment an existing covariance kernel $k(\cdot, \cdot)$ with white noise. Likelihood and predictive equations remain tractable with simple adjustments to the usual closed-form expressions. After specifying the

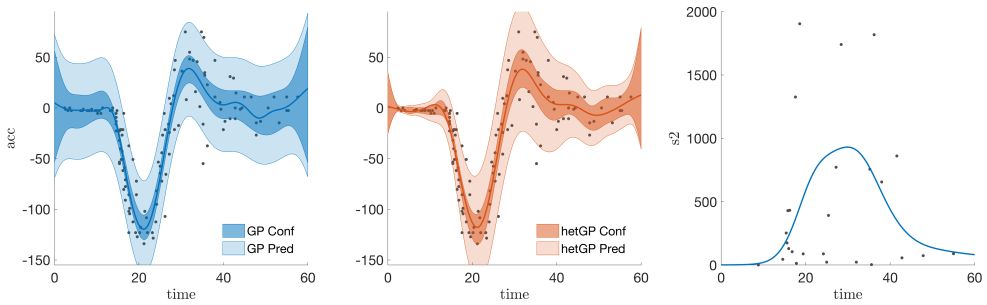


FIG. 3. *The left panel shows an ordinary GP prediction, whereas the middle panel shows the* `hetGP` *alternative. Mean uncertainty (dark), i.e., confidence intervals, and total uncertainty (light + dark), i.e., predictive intervals, are shown. The right panel illustrates the estimated variances* ($s2 \equiv \sigma_N^2(x)$) *from the heteroskedastic fit (black line), versus time, against the empirical variance estimates for those inputs with two or more replicates.*

degrees-of-freedom parameter $\alpha \in \mathbb{R}_+ \setminus [0,2]$ the predictive equations of a Student-$t$ process have the following moments:

$$
\begin{aligned}
\alpha_N &= \alpha + N, \\
\mu(t) &= \mathbb{E}\left(d(t) \mid \mathbf{t}, \mathbf{d}\right) = \mathbf{k}_N(t)^\top \mathbf{K}_N^{-1}\mathbf{d}, \\
\sigma^2(t) &= \mathbb{V}\mathrm{ar}(d(t) \mid \mathbf{t}, \mathbf{d}) = \frac{\alpha+\beta-2}{\alpha+N-2}\left(k(t,t) - \mathbf{k}_N(t)^\top \mathbf{K}_N^{-1}\mathbf{k}_N(t)\right) + v(t), \text{ and} \\
\mathbb{C}\mathrm{ov}(d(t), d(t') \mid \mathbf{t}, \mathbf{d}) &= \frac{\alpha+\beta-2}{\alpha+N-2}\left(k(t,t') - \mathbf{k}_N(t)^\top \mathbf{K}_N^{-1}\mathbf{k}_N(t')\right) + \delta_{t=t'}v(t),
\end{aligned}
\tag{9}
$$

with $\beta = \mathbf{d}^\top \mathbf{K}_N^{-1}\mathbf{d}$. Note that the predictive covariance depends on the observed $\mathbf{d}$ values. This is in contrast to the Gaussian case of (5).

The corresponding log-likelihood is then given by

$$
\begin{aligned}
\log(L) = &-\frac{N}{2}\log((\alpha-2)\pi) - \frac{1}{2}\log(|\mathbf{K}_N|) \\
&+ \log\left(\frac{\Gamma\left(\frac{\alpha+N}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\right) - \frac{(\alpha+N)}{2}\log\left(1 + \frac{\beta}{\alpha-2}\right),
\end{aligned}
$$

where $\Gamma$ denotes the Gamma function. In the presence of replicates, it is possible to show that the full-$N$ equations can be expressed by unique-$n$ analogues via the following equations:

$$
\begin{aligned}
\beta &= \mathbf{d}^\top(\tau^2\mathbf{C}_N + \mathbf{\Sigma}_N)^{-1}\mathbf{d} = \mathbf{d}^\top\mathbf{\Sigma}_N^{-1}\mathbf{d} - \bar{\mathbf{d}}^\top\mathbf{A}_n\mathbf{\Sigma}_n^{-1}\bar{\mathbf{d}} + \bar{\mathbf{d}}^\top(\tau^2\mathbf{C}_n + \mathbf{A}_n^{-1}\mathbf{\Sigma}_n)^{-1}\bar{\mathbf{d}}, \\
\log|\mathbf{K}_N| &= \log|\tau^2\mathbf{C}_N + \mathbf{\Sigma}_N| = \log|\tau^2\mathbf{C}_n + \mathbf{A}_n^{-1}\mathbf{\Sigma}_n| + \log|\mathbf{\Sigma}_N| - \log|\mathbf{A}_n^{-1}\mathbf{\Sigma}_n|.
\end{aligned}
$$

An SK-style moment-based estimator of the variance calculated from replicates can be used in these equations to cope with heteroskedasticity. However, this suffers the same drawbacks as in the GP case. For instance, if there are not sufficient replicates, then the result is highly unstable. Additionally, in the Student-$t$ setting $\alpha$ affects both mean and noise covariances, which further complicates inference and prediction schemes. Fortunately, input-dependent (log) noise can be learned via a latent GP in exactly the same way as in `hetGP`, leading to an effective `hetTP` formulation. In

fact, it is remarkable that, at least from an implementation perspective, no further modifications are required.

Although one could potentially entertain a Student-$t$ process on the noises, we have not found any practical value for such a setup within our own experimentation. Based on expressions given in [55], closed-form derivatives for the log-likelihood are available, similar to [8]. As these represent a substantial component of our contribution, we provide such expressions in detail in Appendix A. An implementation for this `hetTP` is provided as an alternative in our `hetGP` package on CRAN [7]. To the best of our knowledge, ours is the first application of input-dependent, and simultaneously leptokurtic, noise in GP regression.

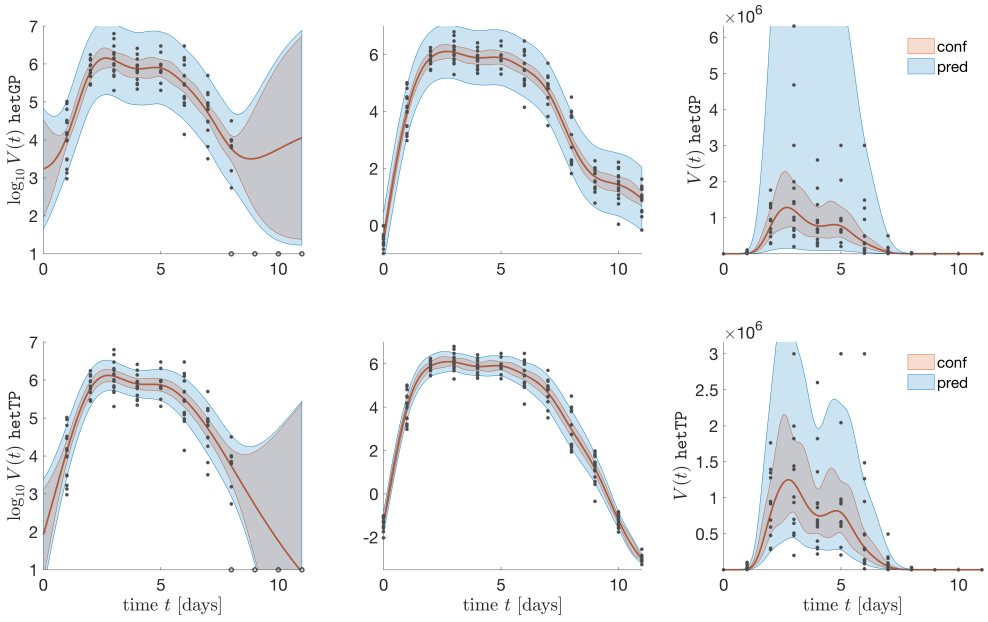


FIG. 4. *Initial and final models. Top:* `hetGP` *models. Bottom:* `hetTP` *models. Left: Before data augmentation,* $\log_{10}$ *scale. Center: After data augmentation,* $\log_{10}$ *scale. Right: After data augmentation, original scale. 95% confidence and prediction intervals are given as shaded areas, respectively, and the mean is represented as a red curve. (Color available online.)*

Figure 4 provides an illustration on our motivating influenza virus infection data. The left panel shows a fit to the completely observed portion of that data on the $\log_{10}$ scale, using `hetGP` (top) and `hetTP` (bottom). Observe the narrower intervals provided by the latter, which attributes a larger portion of the noisy observations to outlying events. The middle panel incorporates our data augmentation scheme (described below) for handling censored observations at the extremes of time. Observe that the `hetTP` variation (bottom) provides a "tighter" distribution on those censored values and offers a more consistent decline as time passes from 9–11 d pi. The right panel in Figure 4 shows the same pair of plots on a linear scale. In this view, the advantage of treating the largest values, from times 3–6 d pi, as outliers is more readily apparent than it is on the $\log_{10}$ scale. It is the appropriate handling of these output extremes, as noise rather than as signal on the $y$-axis, that can lead to the favorable properties of the output at the beginning and end (i.e., rapid increase at the start of the viral

titer dynamics, and the rapid decrease at the end).

**3.3. Censored observations at the extremes.** Other features present in our motivating influenza example are the censored observations at latter times, and a prior for decreasing functions as the data fall into this censoring regime, $t \to 0$ and $t \to \infty$. It is known that the virus count ultimately decays to exactly zero [63]; however, this presents challenges in $\log_{10}$ space as it corresponds to negative infinity. Therefore, even when the crude option of replacing the censored values with a choice of $\varepsilon \ll 1$ is possible, it creates stochastic modeling challenges, i.e., via GPs (for ordinary, `hetGP`, and `hetTP`). Subsequent optimization of ODE solutions, based on surrogate data coming out of such fits, exhibits bias in the parameter estimates because the resulting trajectories (e.g., in viral load over time) struggle to reproduce the plateaus that arise.

Therefore, we recommend a softer approach: encouraging a decreasing mean in the GP predictions as $t$ decreases to 0 d pi, starting at 1 d pi, and as $t$ increases to $\infty$, starting at 8 d pi. A posteriori, such a prior is easy to implement via a rejecting sampling-based data augmentation scheme when obtaining draws from the Gaussian (or Student-$t$) predictive distribution. In what follows, we describe our scheme for data augmentation for censored values as $t$ gets large. The scheme for dealing with $t \to 0$ proceeds similarly.

The first step involves estimating hyperparameters of the GP covariance kernel, $k(\cdot, \cdot)$, which we obtain via the complete data log-likelihood. Conditional on those settings and on the complete data, we may draw from the predictive equations (9). Two examples are shown in the left column of Figure 4. Using those equations, we developed an imputation scheme that proceeds iteratively from the smallest time index with a censored observation, say index $j$ at time $t_j$. In our influenza example in Figure 2, this is time $t_j = 8$ d. We then repeatedly draw from the *noise-free* predictive equations (e.g., using $v(\cdot) = 0$ in (9), using a set of inputs $\mathcal{T} = \{t_i : t_i \leq t_j\}$) and stop when a draw is obtained that is monotonically decreasing in all censored time indices. At this point, only checking at $t_j$ and $t_{j-1}$ is required. Then, we take a number of draws equaling the number of censored observations from the appropriate noise distribution, conditional on those values being below the censoring threshold. For our motivating influenza problem, this value is $\log_{10}$ 201 $\text{TCID}_{50}$. In the case of `hetGP`, the value is chosen from the Gaussian distribution with variance $\hat{v}(t_j)$. For `hetTP`, a Student-$t$ with degrees of freedom $\alpha_N$ is used and multiplied by $\sqrt{\hat{v}(t_j)}$. Finally, we treat the sampled draws at $t_j$ as actual data values and repeat, moving on to $t_{j+1}$.

After the censored observations have been replaced with "synthetic" samples, we obtain draws from the full predictive distribution for the use in the wider parameter estimation exercise (section 2). Specific illustrations are included below. In this way, the scheme is a variation on so-called data augmentation for Bayesian spatial modeling [26]. Figure 4 provides an illustration of the overall surfaces which arise under this scheme using Gaussian and Student-$t$ innovations. However, to reduce clutter no actual sample paths or latent samples are shown in the figure. A glimpse at those for `hetTP` is provided in Figure 5, whose panels "zoom in" on times 7–11 d pi. The leftmost panel shows three draws from the posterior predictive distribution, each of which is conditioned on a separate sample of latent values obtained via the scheme described above. The subsequent three panels in the figure redraw those sample paths separately, along with a visualization of the latent values which aided in its production.

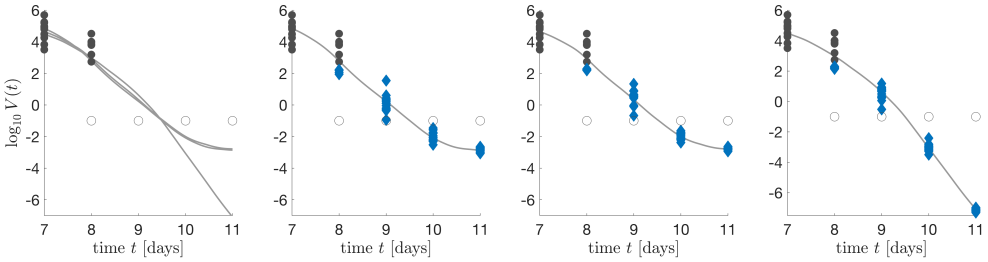


FIG. 5. *The leftmost panel shows three sample paths in gray, overlaid onto a zoomed-in portion of Figure 2 covering time $t \geq 7$. The subsequent three panels show each of those sample paths separately, accompanied by the latent samples (as blue diamond characters) generated for the censored values in the data. (Color available online.)*

These surrogate draws (gray lines in Figure 5) have a distribution characterized by the bottom-right panel in Figure 4, in the case of the `hetTP` model transformed back onto the original scale of the data. Each draw is a sample from a posterior (predictive) distribution. When treated as surrogate data in a "single shooting" least-squares search for parameters $\widehat{\mathbf{p}}_j$, we obtain a map from surrogate posterior to posterior over parameters to the ODE (3). These details are laid out algorithmically and illustrated empirically on a toy example and by our motivating influenza example in the following section.

**4. Numerical experiment.** Here we provide numerical illustrations of the scheme obtained by chaining together the methodological pieces detailed above. Algorithm 1 provides a skeleton for the overall procedure. The following discusses the algorithm's specifications and subroutines with reference to particular procedures and equations provided earlier.

**Algorithm 1** Parameter & Uncertainty Estimation via Stochastic Processes

**input:** data $\mathbf{d}$ and ODE model
1: use $\mathbf{d}$ to fit the stochastic process $\mathcal{G}$
2: **parallel for** $j = 1$ to $J$ **do**
3: generate sample $\mathbf{g}_j$ from $\mathcal{G}$
4: compute $\widehat{\mathbf{p}}_j$ from (2) using $\mathbf{g}_j$
5: **end parallel for**
**output:** $\{\widehat{\mathbf{p}}_j\}_{j=1}^{J}$

1. Inputs include observations $\mathbf{d}$ (e.g., data represented in Figure 2 in section 2.1) and model equations (e.g., (3) in section 2.2).
2. An appropriate stochastic process $\mathcal{G}$ needs to be determined, reflecting the data $\mathbf{d}$, and given prior knowledge on the dynamics of the system (line 1). Examples are provided in section 2.3 and sections 3.1–3.3, including variations on GPs (i.e., ordinary, `hetGP`, and `hetTP`).
3. Monte Carlo samples $\mathbf{g}_j$ are drawn from the posterior predictive distribution provided by $\mathcal{G}$ in line 3. This is typically a straightforward process; however, standard sampling methods (e.g., rejection sampling) may be required to obtain a sample $\mathbf{g}_j$ from $\mathcal{G}$. We follow the discretize-then-optimize approach, discretizing the $\mathcal{L}_2$ norm in (2) using an equidistant grid. One option

is to match the inputs with the data inputs, i.e., the times at which observations were collected. However, finer or coarser resolutions may be used. We choose a finer resolution, spanning the original range of times imposing the smoothness of the state variables.

4. Given the sample $\mathbf{g}_j$, an optimization scheme and a numerical ODE solver are required to compute point estimates $\widehat{\mathbf{p}}_j$ as discussed in section 1.1 and at the beginning of section 2. The optimization scheme requires an initial guess $\mathbf{p}_0$ and may be chosen differently for each $j$ if desired.
5. The algorithm returns a set $\{\widehat{\mathbf{p}}_j\}_{j=1}^{J}$ reflecting the posterior distribution of parameter estimates.
6. Although the pseudocode shows surrogate data generation and parameter estimation happening within the same (potentially parallel) **for** loop, those two steps need not be executed in tandem. After fitting $\mathcal{G}$, generating a collection of realizations $\mathbf{g}_j$, subsequent fitting of $\widehat{\mathbf{p}}_j \mid \mathbf{g}_j$ may even be performed offline or on an ad hoc basis. Fitting of $\mathcal{G}$ is the only point of contact between them regarding statistical inference, and therefore this step is independent of the model equations. Other models can be entertained ex post without revisiting the data or fitting of $\mathcal{G}$.

As a benchmark, we consider a Bayesian approach to parameter inference [64] in this setting as applied to our motivating influenza example. The essence of that scheme is a Gaussian likelihood measuring the distance between the data and single shooting paths derived from the ODE under parameters, $\mathbf{p}$. Specifically, $\pi(\mathbf{d}|\mathbf{p}) \propto -\exp(\frac{1}{2\sigma^2}\|\mathbf{s}(\mathbf{y}(\mathbf{p})) - \mathbf{d}\|_2^2)$. This is paired with independent priors on the individual parameters, often chosen to be uniform in an appropriate range. Our particular choices are application-dependent and are detailed below. Inference is facilitated by a Metropolis MCMC [36]. This approach is beneficial in terms of simplicity and is straightforward to implement. However, tuning the Metropolis proposals to obtain adequate mixing of the Markov chains can be highly application-dependent, and it is not easily parallelizable. Although such challenges are surmountable, a deficiency that remains is that it is not straightforward to incorporate prior information on the regularity of the observation trajectory, such as smoothness or (local) monotonicity. As we show, this results in a far more diffuse posterior distribution compared to our proposed method.

We turn now to two numerical investigations of our proposed method. For validation, we first consider a simulation study. Then we attack our motivating influenza problem (sections 2.1–2.2).

**4.1. Simulation study.** In our simulation study we assume that observations of predator and prey are provided by a Lotka–Volterra system,

$$\begin{aligned} y_1' &= -y_1 + \alpha_1 y_1 y_2, \\ y_2' &= y_2 - \alpha_2 y_1 y_2, \end{aligned} \tag{10}$$

with $\boldsymbol{\alpha}_{\text{true}} = [1,1]^\top$ and initial condition $\mathbf{y}_{\text{true}}(0) = [y_1(0), y_2(0)]^\top = [2, 1/2]^\top$ on the interval $t \in [0, 10]$ at 20 equidistant time points $t_j = \frac{10}{19}(j-1)$, $j = 1, \ldots, 20$.

We assume that for each state we are given *five* repeated samples at times $t_j$, where the samples are subject to additive noise. Here, $\mathbf{y}_{\text{true}}(t_j) + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{10}\mathbf{I}_2\right)$. Hence, $\mathbf{d} \in \mathbb{R}^{200}$ (see Figure 6). Given these observations, we seek to estimate model parameters $\boldsymbol{\alpha}_{\text{true}}$ and the initial condition $\mathbf{y}_{\text{true}}(0)$. Here, $\mathbf{p} = [\alpha_1, \alpha_2, y_1(0), y_2(0)]^\top$.

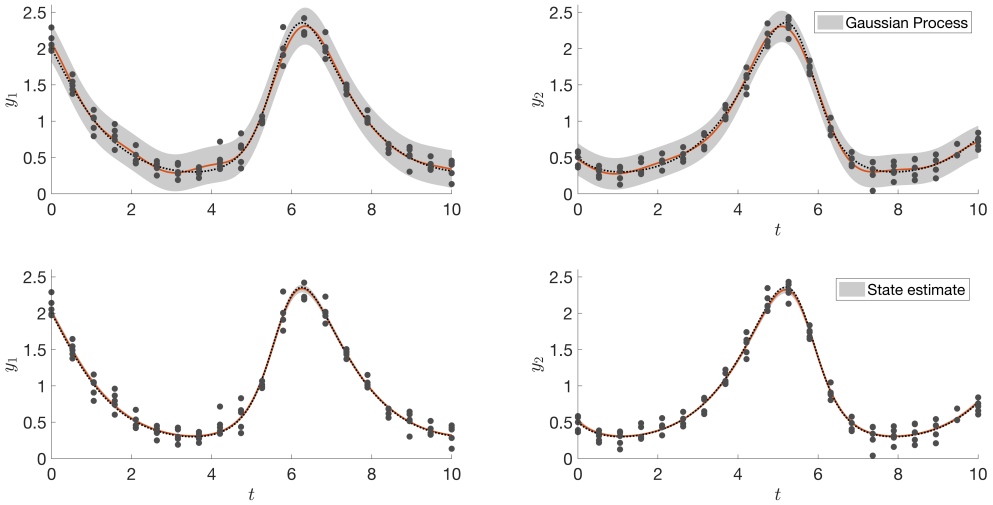


FIG. 6. *The top panel shows the dynamics of the true predator and prey system* $\mathbf{y}_{\text{true}}(t)$ *in dotted black. Simulated data* $\mathbf{d}$ *are depicted as black dots. The generated GP is represented by its mean* $\mu$ *(red) and central* 95% *confidence region depicted in gray shade. The lower panel shows state estimates after using our framework, again with the true predator prey state in dotted black lines. The red line gives the* 50*th percentile with the barely visible central* 95% *interval shaded in gray. (Color available online.)*

The first step is to train the surrogate stochastic process, $\mathcal{G}$. In this controlled experiment, there is no need to consider heteroskedasticity or censored observations. Thus, we entertain an ordinary GP and use straightforward likelihood-based optimization methods to infer the unknown hyperparameters to a Gaussian covariance kernel. Figure 6 provides the mean of the GP predictive surface in red, while the gray shaded area reflects the variances. The true curves generating the data are shown as dotted black lines.

Conditional on that fit, we generate 100,000 samples $\{\mathbf{g}_j(t)\}_{j=1}^{100,000}$ from the predictive equations corresponding to the fitted $\mathcal{G}$ (i.e., we follow the unique-$n$ variation on (5)). Numerically, we discretize the sample processes $\mathbf{g}_j(t)$ at 201 equidistant times in the interval $[0, 10]$ to solve the optimization problem (2). For simplicity, we utilize a single shooting method via a direct search method optimizing (2), while the dynamical system (10) is solved via an explicit Runge–Kutta 4 method. Optimization problem (2) is solved 100,000 times resulting in a set of parameter estimates $\{\widehat{\mathbf{p}}_j\}_{j=1}^{100,000}$. This set $\{\widehat{\mathbf{p}}_j\}_{j=1}^{100,000}$ defines a distribution of estimates of the underlying true parameter values $\mathbf{p}_{\text{true}}$. The lower panel of Figure 6 illustrates the uncertainty estimates of the states $y_1$ and $y_2$. Note the narrow uncertainty margins and that the true solution lies completely within the error estimates.

Figure 7 displays the projected 1D densities for each of the four model parameters $\alpha_1$, $\alpha_2$, $y_1(0)$, and $y_2(0)$ (panels 1–4). Highlighted in blue are the densities generated by our new GP based approach, while the results for the MCMC alternative [64] are overlaid in red. For that method, we chose a uniform prior in the domain $0 \leq p_i \leq 10$ for $i = 1, \ldots, 4$. We initialized the Markov chain at $\mathbf{p}_0 = [1, 1, 2, 1/2]^\top$ and used random-walk Metropolis proposals as $\mathcal{N}(\tilde{\mathbf{p}}_j, 1/5\,\mathbf{I}_4)$, where $\tilde{\mathbf{p}}_j$ is the previous posterior sample. Mixing in the chain was reasonably good. We determined it to have converged after one million samples and generated another million thereafter

to save as posterior draws: $\{\tilde{\mathbf{p}}_j\}_{j=1}^{1,000,000}$. Furthermore, Figure 7 also includes the marginal GP prior densities in yellow. The GP prior has been computed by solving (2) replacing the posterior samples $\mathbf{g}_j$ by GP prior samples. Note that we only used nonnegative GP prior samples due to the nonnegative nature of the Lotka–Volterra system. By comparing the marginal GP prior and GP posterior densities in Figure 7 (yellow and blue, respectively), we observe that the GP posterior does not just evolve from a more informative prior. The dotted lines in Figure 7 represent the true parameter values $\mathbf{p}_{\text{true}}$ (dotted line). The maximum a posteriori (MAP) $\mathbf{p}_{\text{MAP}} \approx [1.005, 0.993, 2.023, 0.507]^\top$ (dashed line), respectively. Although our new method generally agrees with the MAP obtained via the MCMC, the densities generated by the GPs are narrower and more tightly sandwich the MAP. While the MCMC method generates adequate posterior densities, the tighter densities of our method are due to the regularization implicitly induced by the GP prior, which imposes smoothness and decreases curvature in the state solution $\mathbf{y}(t)$.

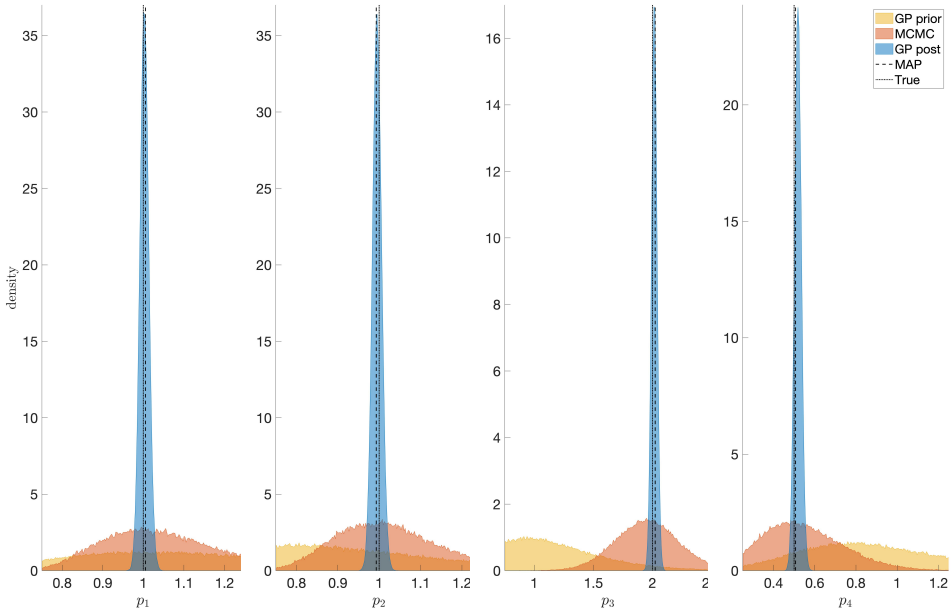


FIG. 7. *1D marginal posterior (and prior) density plots of the estimates* $\{\widehat{\mathbf{p}}_j\}_{j=1}^{100,000}$ *(GP posterior density in blue; prior in yellow) and* $\{\tilde{\mathbf{p}}_j\}_{j=1}^{1,000,000}$ *(MCMC posterior density in red). The true model parameters* $\mathbf{p}_{\text{true}} = [1, 1, 2, 1/2]^\top$ *are represented by a dotted line, while the joint maximum a posteriori estimate* $\mathbf{p}_{\text{MAP}}$ *is represented as a dashed line. (Color available online.)*

**4.2. Influenza.** We next discuss the influenza virus setup introduced in section 2.1, with data visualized in Figure 2 and associated mathematical model detailed by (3). The data include virus counts, but no data is available for the infected cells $I_1(t)$ and $I_2(t)$ or for the susceptible target cells $T(t)$.

Equation (3) is of the form $\mathbf{y}' = \mathbf{f}(t, \mathbf{y}, \mathbf{p})$, where the state variable is given by $\mathbf{y}(t) = [T(t), I_1(t), I_2(t), V(t)]^\top$. We assume that the parameters $\beta, \rho, c, \delta, K_d$ and the initial condition $T(0)$ are unknown, i.e., $\mathbf{p} = [\beta, \rho, c, \delta, K_d, T(0)]^\top$. We assume that

the other parameters and initial conditions in (3) are given. Here, we choose $\kappa = 4\ \mathrm{d}^{-1}$ and $I_1(0) = 10$ cells, $I_2(0) = 0.02$ cells, and $V(0) = 0.07\ \mathrm{TCID}_{50}$. Typically, these are not the initial conditions used in influenza modeling studies (e.g., as in [63]). In particular, there are no productively infected cells ($I_2$) at the time of infection. However, a positive value was necessary for the simulation. The values were chosen arbitrarily.

For the optimization, we use the same setup as in our simulation study of section 4.1. For 100,000 stochastic process realizations $\{\mathbf{g}_j(t)\}_{j=1}^{100,000}$ from data augmented `hetTP`s, we use a single shooting method with direct search optimization, and the ODE is solved via an explicit Runge–Kutta 4 method. Numerically, we discretized the residual $\mathbf{s}(\mathbf{y}(t)) - \mathbf{g}_j(t)$ at 3000 equidistant time points. Hence, we generate a set of samples $\{\widehat{\mathbf{p}}_j\}_{j=1}^{100,000}$ from the posterior distribution of the underlying true parameter values $\mathbf{p}_{\text{true}}$.

Figure 8 provides an example of the generated data fits. The top panel shows the (denoised) posterior predictive surface obtained from our fitted `hetTP` surrogate, $\mathcal{G}$. Sample paths $\mathbf{g}_j$ yielding $\widehat{\mathbf{p}}_j$ were used to generate a realization of the states derived from the system of differential equations, and the distribution of those curves is shown in the bottom panel of the figure. Notice that the top surface is not strictly unimodal like the bottom surface—as demanded by the ODE. In this way, the figure shows how least-squares calculations "filter" posterior inference into parameters of the system of equations via the predictive distribution, as exhibited by their resulting distribution of states.

For comparison we again applied an MCMC framework. We utilized a similar MCMC framework for this influenza data as described in the Master's thesis [70] and also discussed in [64]. Uniform priors were chosen in the range(s) $-1/\varepsilon < p_j < 1/\varepsilon$, where $\varepsilon$ is given by the machine precision, and random-walk Metropolis proposals were generated as $\mathcal{N}(\tilde{\mathbf{p}}_j, 1/5\,\mathbf{I}_6)$. The maximum likelihood estimate was used to generate a starting value. We determined the MCMC chain to have converged after one million samples, and the next one million were saved as posterior samples $\{\tilde{\mathbf{p}}_j\}_{j=1}^{1,000,000}$. The MAP estimate is computed as $\mathbf{p}_{\text{MAP}} \approx [2.9601 \cdot 10^{-5}, 4.4085 \cdot 10^4, 2.8540, 28.1280, 0.0436, 154.3949]^\top$.

Marginal 1D posterior densities of the parameters for our newly proposed method and the MCMC are depicted in Figure 9. Again, the densities utilizing the stochastic process generate a tight distribution, while the distributions generated through the MCMC chain give wide uncertainty estimates of the model parameters. The MAP $\mathbf{p}_{\text{MAP}}$ exhibits unrealistic estimates due to the ill-posedness in the optimization problem (see Figure 9) given as the black dashed line.

The parameter distributions shown in Figure 9 illustrate the improvement made by the stochastic process (SP) method compared with the MCMC method. In addition, the results from fitting via SP more closely reflect those obtained from using traditional global optimization methods (e.g., adaptive simulated annealing) [63], which have been shown to yield accurate estimates [60, 61]. Altering the data through censoring does skew the values of the parameters, which is expected. The effect is particularly evident in the value of $\delta$, which dictates the viral decay dynamics and is the most sensitive parameter [59, 58, 63]. Here, we assumed that the censored data decreased monotonically, which is unconventional in viral kinetic modeling. Most often, censored data is imputed as one half the LOD and without any dynamical restrictions. By imposing monotonicity across several times with repeated LOD measurements rather than truncating at the first instance (i.e., 8–11 d pi versus 8–9 d

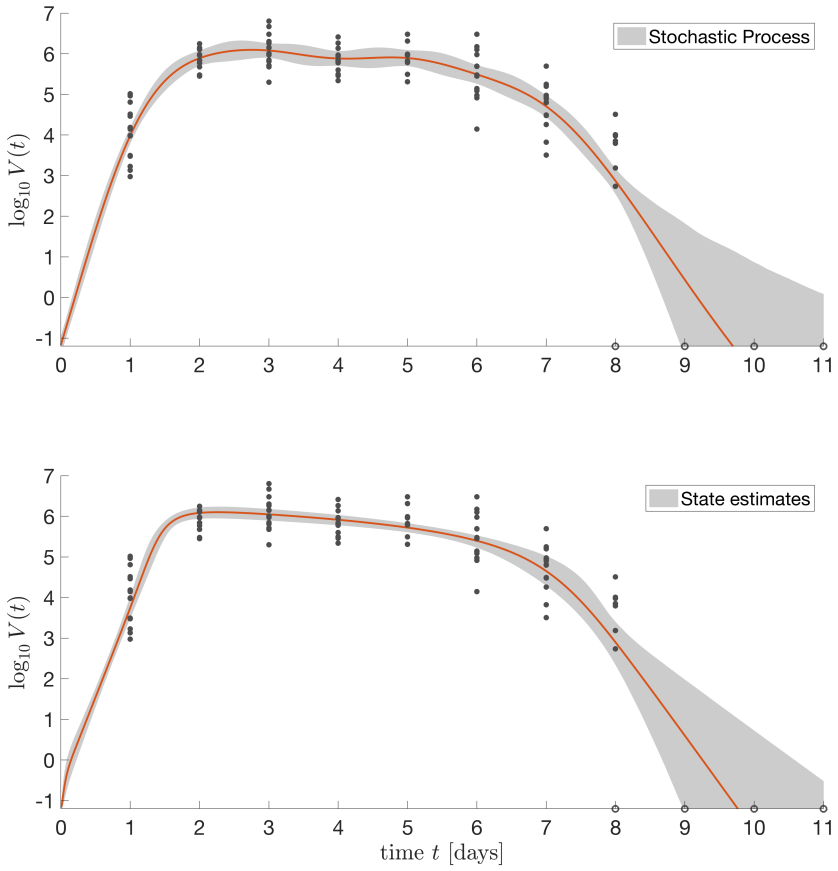


FIG. 8. *The top panel shows the statistics of the stochastic process. The lower panel shows the statistics of the reconstructed state solutions of $V$ for the estimated $\{\widehat{\mathbf{p}}\}_{j=1}^{100,000}$ reconstructions. The mean is represented by a red line, while the 95 percentiles are shaded in gray. Data are given as black dots. (Color available online.)*

pi), solutions inconsistent with experimental observations were produced (Figure 8). That is, the resulting model trajectories do not capture the rapid viral clearance in some mice between 7–8 d pi, where the values may indeed be true zeros. In addition, they suggest that animals may have viral loads above the LOD at 9 d pi and not clear the infection until 11 d pi, neither of which has been observed. Thus, one should use caution when censoring data as biological inference can be inhibited. However, the consistency in parameter distributions and behavior (e.g., correlation between $\rho$ and $c$) between the results here and the results in Smith et al. [63] supports the accuracy of the fitting method.

**5. Conclusion and discussion.** Our proposed method consists of two parts. In the first phase, we fit a surrogate stochastic process to given data. In the second phase, we use a set of samples from the posterior predictive distribution of that fitted stochastic process to generate surrogate data, which are then "passed through" a typ-

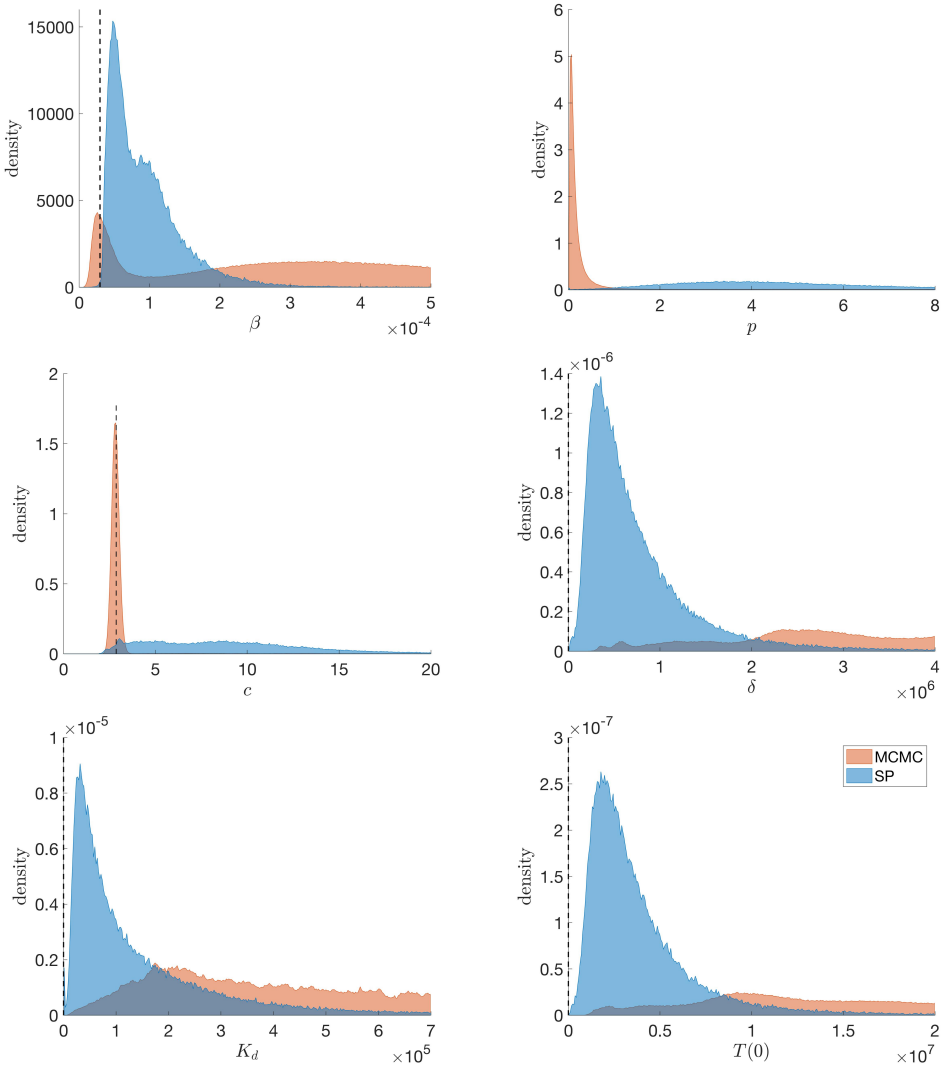


FIG. 9. *Projected density plots of the parameter estimates for the influenza model* (3) *and associated data depicted in Figure* 2. *Densities generated by the SP are given in blue, while density estimates of the posterior generated by an MCMC chain are in red. The dashed line represents the maximum likelihood estimates. The maximum likelihood estimate of* $p$ *in panel* 2 *is omitted because* $p_{\text{MLE}} \approx 4.4085 \cdot 10^4$. *(Color available online.)*

ical scheme used to tune a mathematical model's parameterization to the data. Such an approach represents a novel means of obtaining a posterior distribution on model parameters. A major advantage of this procedure is that we can induce prior knowledge (e.g., smoothness of the states) directly into the process samples and handle other nuances in the data like input-dependent noise, leptokurtic errors, and censoring. The result is a far more "focused" posterior distribution compared to other Bayesian alternatives. A further advantage is that our method provides uncertainty estimates but does not rely on the Markov property as in MCMC methods. Hence, this method is embarrassingly parallelizable. Depending on the nature of other, similar applications,

we envision many opportunities for extension via adaptations to the prior implied by the chosen family of surrogate stochastic processes, e.g., to deal with large amounts of data or additional known features in the data (e.g., symmetries).

Although our problem setting has much in common with those typically tackled within the Kennedy and O'Hagan [39] (KOH) framework or related setups [37], there are several important reasons why those approaches are not well suited for our setting. One has to do with Bayesian computation. Inference in KOH settings requires MCMC with likelihood evaluations. This incurs a cubic cost (in the number of data points) for evaluation which is cumbersome in moderate data size settings and/or in parameter spaces of moderate size. When the data is large, the additional MCMC demands make inference all but impractical. Parallelization offers no respite due to the inherently serial nature of the Metropolis steps typically involved. Another contrast has to do with incorporating known dynamics into the prior on the surrogate stochastic process and related issues in handling censored observations. It is fairly easy to implement a rejection sampling scheme, such as the one described in section 3.3 for monotonicity, to generate appropriately constrained realizations from the posterior predictive distribution and subsequently map those (deterministically) to parameter values. It is quite another matter to approach the problem from the other direction by accepting or rejecting parameter settings that make such surfaces more or less likely under the posterior distribution. Because it was not obvious how we could accommodate these constraints in a KOH framework, we found it difficult to entertain it as a comparator in our empirical work.

The main focus of this work was to provide a proof of concept of our new methods. Many extensions, modifications, and analyses remain open and will be subject to future research. For instance, theoretical investigations in particular of linear problems are desirable. Further, the main computational burden of our proposed method is the repeated (but parallel) optimization procedure. Solving these ODE constrained optimization problems may be done more efficiently by using Newton-type optimization methods coupled with efficient ODE solvers and informed initial parameter guesses as proposed in [19]. Gaussian processes (GPs) provide a full covariance structure not fully utilized in this setup. Future research will be dedicated towards a weighted least squares framework acknowledging this structure. Even within an MCMC framework and combined with a randomize-than-optimize approach [6], our proposed method may provide computational advances, but so far a deliberate investigation remains open. A promising application and extension of our methods may lie within a data assimilation framework, where partial observations are used to predict future state dynamics. Here, our proposed method may, for instance, be integrated into a 3D-Var or 4D-Var framework [44, 45] to provide enhanced covariance and ultimately state and uncertainty estimates. Our methods can also be extended to optimal experimental design problems with underlying ODE systems [18]. Further investigations will also be directed towards loosening our stringent statistical assumption on the data: we assumed that the data comes from a single underlying true parameter. This is an oversimplification, and future research may target distributions of the true parameter.

**Appendix A. Derivatives of log-likelihood for Student-$t$ processes.** Recall that the full-$N$ log-likelihood is given by

$$\log(L) = -\frac{N}{2}\log((\alpha-2)\pi) - \frac{1}{2}\log|\mathbf{K}_N| + \log\left(\frac{\Gamma\left(\frac{\alpha+N}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\right) - \frac{(\alpha+N)}{2}\log\left(1+\frac{\beta}{\alpha-2}\right).$$

By taking into account savings from replicates, the reduced unique-$n$ log-likelihood is

$$\log(L) = -\frac{N}{2}\log((\alpha-2)\pi) - \frac{1}{2}\log|\tau^2\mathbf{C}_n + \mathbf{A}_n^{-1}\mathbf{\Sigma}_n| - \frac{1}{2}\sum_{i=1}^{n}[(a_i-1)\log\lambda_i + \log a_i] + \log\left(\frac{\Gamma\left(\frac{\alpha+N}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)}\right) - \frac{(\alpha+N)}{2}\log\left(1+\frac{\beta}{\alpha-2}\right),$$

with $\beta = \mathbf{d}^\top\mathbf{\Sigma}_N^{-1}\mathbf{d} - \bar{\mathbf{d}}^\top\mathbf{A}_n\mathbf{\Sigma}_n^{-1}\bar{\mathbf{d}} + \bar{\mathbf{d}}^\top(\tau^2\mathbf{C}_n + \mathbf{A}_n^{-1}\mathbf{\Sigma}_n)^{-1}\bar{\mathbf{d}}$.

For likelihood based optimization of the hyperparameters, derivatives become very useful. Shah, Wilson, and Ghahramani in [55] provide derivatives with respect to $\alpha$ and $\theta$ that we complement for our setup. The derivative with respect to $\alpha$ is

$$\frac{\partial}{\partial\alpha}\log L = -\frac{N}{2(\alpha-2)} + \frac{1}{2}\psi\left(\frac{\alpha+N}{2}\right) - \frac{1}{2}\psi\left(\frac{\alpha}{2}\right) - \frac{1}{2}\left(1+\frac{\beta}{\alpha-2}\right) + \frac{(\alpha+N)\beta}{2(\alpha-2)^2+2\beta(\alpha-2)},$$

with $\psi$ the digamma function.

For the other hyperparameters, denoting $\mathbf{\Upsilon}_n = \tau^2\mathbf{C}_n + \mathbf{A}_n^{-1}\mathbf{\Sigma}_n$,

$$\frac{\partial}{\partial\cdot}\log L = -\frac{1}{2}tr\left(\mathbf{\Upsilon}_n^{-1}\frac{\partial\mathbf{\Upsilon}_n}{\partial\cdot}\right) - \frac{\alpha+N}{2(\alpha+\beta-2)}\frac{\partial\beta}{\partial\cdot} - \frac{1}{2}\sum_{i=1}^{n}(a_i-1)\frac{\partial\log\lambda_i}{\partial\cdot},$$

in particular we get

$$\frac{\partial}{\partial\theta}\log L = -\frac{\tau^2}{2}tr\left(\mathbf{\Upsilon}_n^{-1}\frac{\partial\mathbf{C}_n}{\partial\theta}\right) + \frac{\alpha+N}{2(\alpha+\beta-2)}\tau^2\bar{\mathbf{d}}^\top\mathbf{\Upsilon}_n^{-1}\frac{\partial\mathbf{C}_n}{\partial\theta}\mathbf{\Upsilon}_n^{-1}\bar{\mathbf{d}},$$

$$\frac{\partial}{\partial\tau^2}\log L = -\frac{1}{2}tr\left(\mathbf{\Upsilon}_n^{-1}\mathbf{C}_n\right) + \frac{\alpha+N}{2(\alpha+\beta-2)}\bar{\mathbf{d}}^\top\mathbf{\Upsilon}_n^{-1}\mathbf{C}_n\mathbf{\Upsilon}_n^{-1}\bar{\mathbf{d}},$$

$$\frac{\partial}{\partial\mathbf{\Lambda}_n}\log L = -\frac{1}{2}\mathbf{A}_n^{-1}\operatorname{Diag}(\mathbf{\Upsilon}_n^{-1}) - \frac{\mathbf{A}_n-\mathbf{I}_n}{2}\mathbf{\Lambda}_n^{-1} + \frac{\alpha+N}{2(\alpha+\beta-2)}\mathbf{A}_n\mathbf{S}\mathbf{\Lambda}_n^{-2} + \mathbf{A}_n^{-1}\operatorname{Diag}(\mathbf{\Upsilon}_n^{-1}\bar{\mathbf{d}})^2.$$

## REFERENCES


[1] B. Ankenman, B. L. Nelson, and J. Staum, *Stochastic kriging for simulation metamodeling*, Oper. Res., 58 (2010), pp. 371–382.

[2] M. Asch, M. Bocquet, and M. Nodet, *Data Assimilation: Methods, Algorithms, and Applications*, Fundam. Algorithms 11, SIAM, Philadelphia, 2016, https://doi.org/10.1137/1.9781611974546.

[3] R. Aster, B. Borchers, and C. Thurber, *Parameter Estimation and Inverse Problems*, 2nd ed., Academic Press, Waltham, MA, 2012.

[4] E. Baake, M. Baake, H. G. Bock, and K. M. Briggs, *Fitting ordinary differential equations to chaotic data*, Phys. Rev. A, 45 (1992), 5524.

[5] S. Bandara, J. Schlöder, R. Eils, H. Bock, and T. Meyer, *Optimal experimental design for parameter estimation of a cell signaling model*, PLoS Comput. Biol., 5 (2009), e1000558.

[6] J. M. Bardsley, A. Solonen, H. Haario, and M. Laine, *Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems*, SIAM J. Sci. Comput., 36 (2014), pp. A1895–A1910, https://doi.org/10.1137/140964023.

[7] M. Binois and R. B. Gramacy, hetGP*: Heteroskedastic Gaussian Process Modeling and Design under Replication*, R package version 1.0.1, 2017.

[8] M. Binois, R. B. Gramacy, and M. Ludkovski, *Practical Heteroskedastic Gaussian Process Modeling for Large Simulation Experiments*, preprint, https://arxiv.org/abs/1611.05902, 2016.

[9] H. Bock and K. Plitt, *A multiple shooting algorithm for direct solution of optimal control problems*, in Proceedings of the 9th IFAC World Congress, Budapest, Hungary, 1984, pp. 243–247.

[10] H. G. Bock, *Recent advances in parameter identification techniques for O.D.E.*, in Numerical Treatment of Inverse Problems in Differential and Integral Equations, Progr. Sci. Comput. 2, Birkhäuser Boston, Boston, MA, 1983, pp. 95–121.

[11] A. Boukouvalas, D. Cornford, and M. Stehlik, *Optimal design for correlated processes with input-dependent noise*, Comput. Statist. Data Anal., 71 (2014), pp. 1088–1102.

[12] G. Burgers, P. Jan van Leeuwen, and G. Evensen, *Analysis scheme in the ensemble Kalman filter*, Monthly Weather Rev., 126 (1998), pp. 1719–1724.

[13] B. Calderhead, M. Girolami, and N. D. Lawrence, *Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes*, in Proceedings of the International Conference on Advances in Neural Information Processing Systems, 2009, pp. 217–224.

[14] D. Calvetti and E. Somersalo, *An Introduction to Bayesian Scientific Computing*, Springer-Verlag, New York, 2007.

[15] D. Calvetti and E. Somersalo, *Computational Mathematical Modeling: An Integrated Approach Across Scales*, Math Model. Comput. 17, SIAM, Philadelphia, 2012.

[16] M. Caracotsios and W. E. Stewart, *Sensitivity analysis of initial value problems with mixed ODEs and algebraic equations*, Comput. Chem. Engrg., 9 (1985), pp. 359–365.

[17] Y. Chen and D. S. Oliver, *Ensemble randomized maximum likelihood method as an iterative ensemble smoother*, Math. Geosci., 44 (2012), pp. 1–26.

[18] M. Chung and E. Haber, *Experimental design for biological systems*, SIAM J. Control Optim., 50 (2012), pp. 471–489, https://doi.org/10.1137/100791063.

[19] M. Chung, J. Krueger, and M. Pop, *Robust parameter estimation for biological systems: A study on the dynamics of microbial communities*, Math. Biosci., 294 (2017), pp. 71–84.

[20] M. Conrad, C. Hubold, B. Fischer, and A. Peters, *Modeling the hypothalamus-pituitary-adrenal system: Homeostasis by interacting positive and negative feedback*, J. Biolog. Phys., 35 (2009), pp. 149–162.

[21] E. M. Constantinescu, A. Sandu, T. Chai, and G. R. Carmichael, *Ensemble-based chemical data assimilation.* I: *General approach*, Quart. J. Roy. Meteorolog. Soc., 133 (2007), pp. 1229–1243.

[22] P. Courtier, J.-N. Thépaut, and A. Hollingsworth, *A strategy for operational implementation of 4D-var, using an incremental approach*, Quart. J. Roy. Meteorolog. Soc., 120 (1994), pp. 1367–1387.

[23] F. Dondelinger, D. Husmeier, S. Rogers, and M. Filippone, *ODE parameter inference using adaptive gradient matching with Gaussian processes*, in Proceedings of the 16th International Conferenceon Artificial Intelligence and Statistics, Scottsdale, AZ, 2013, pp. 216–228.

[24] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, CRC Press, Boca Raton, FL, 1994.

[25] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Springer, Cham, 1996.

[26] B. L. Fridley and P. Dixon, *Data augmentation for a Bayesian spatial model involving censored observations*, Environmetrics, 18 (2007), pp. 107–123.

[27] B. Göbel, K. M. Oltmanns, and M. Chung, *Linking neuronal brain activity to the glucose metabolism*, Theor. Biol. Med. Model., 10 (2013), 50.

[28] G. Goel, I.-C. Chou, and E. O. Voit, *System estimation from metabolic time-series data*, Bioinformatics, 24 (2008), pp. 2505–2511.

[29] P. W. Goldberg, C. K. Williams, and C. M. Bishop, *Regression with input-dependent noise: A Gaussian process treatment*, in Advances in Neural Information Processing Systems, Vol. 10, MIT Press, Cambridge, MA, 1998, pp. 493–499.

[30] J. Gong, G. Wahba, D. R. Johnson, and J. Tribbia, *Adaptive tuning of numerical weather prediction models: Simultaneous estimation of weighting, smoothing, and physical parameters*, Monthly Weather Rev., 126 (1998), pp. 210–231.
[31] N. S. Gorbach, S. Bauer, and J. M. Buhmann, *Mean-Field Variational Inference for Gradient Matching with Gaussian Processes*, preprint, https://arxiv.org/abs/1610.06949, 2016.
[32] J. Hadamard, *Lectures on Cauchy's Problem in Linear Differential Equations*, Yale University Press, New Haven, CT, 1923.
[33] G. Hairer, S. Nørsett, and E. Wanner, *Solving Ordinary Differential Equations* I: *Nonstiff Problems*, 2nd ed., Springer, Berlin, 1993.
[34] G. Hairer and E. Wanner, *Solving Ordinary Differential Equations* II: *Stiff and Differential-Algebraic Problems*, 2nd ed., Springer, Berlin, 1996.
[35] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, Math. Model. Comput. 4, SIAM, Philadelphia, 1998, https://doi.org/10.1137/1.9780898719697.
[36] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.
[37] D. Higdon, M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne, *Combining field data and computer simulations for calibration and prediction*, SIAM J. Sci. Comput., 26 (2004), pp. 448–466, https://doi.org/10.1137/S1064827503426693.
[38] J. Kennedy and R. Eberhart, *Particle swarm optimization*, in Proceedings of the 1995 IEEE International Conference on Neural Networks, Vol. 4, IEEE, Washington, DC, 1995, pp. 1942–1948.
[39] M. Kennedy and A. O'Hagan, *Bayesian calibration of computer models*, J. R. Stat. Soc. Ser. B Stat. Methodol., 63 (2001), pp. 425–464.
[40] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard, *Most likely heteroscedastic Gaussian process regression*, in Proceedings of the International Conference on Machine Learning, ACM, New York, 2007, pp. 393–400.
[41] S. Kirkpatrick, C. Gelatt, Jr., and M. Vecchi, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.
[42] K. Law, A. Stuart, and K. Zygalakis, *Data Assimilation: A Mathematical Introduction*, Springer, Cham, 2015.
[43] B. A. J. Lawson, C. C. Drovandi, N. Cusimano, P. Burrage, B. Rodriguez, and K. Burrage, *Unlocking data sets by calibrating populations of models to data density: A study in atrial electrophysiology*, Sci. Adv., 4 (2018), e1701676.
[44] F.-X. Le Dimet and O. Talagrand, *Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects*, Tellus A Dynam. Meteorol. Oceanogr., 38 (1986), pp. 97–110.
[45] A. C. Lorenc, *The potential of the ensemble Kalman filter for NWP—a comparison with 4D-var*, Quart. J. Roy. Meteorolog. Soc., 129 (2003), pp. 3183–3203.
[46] S. Marino, I. B. Hogue, C. J. Ray, and D. E. Kirschner, *A methodology for performing global uncertainty and sensitivity analysis in systems biology*, J. Theoret. Biol., 254 (2008), pp. 178–196.
[47] D. E. Morris, J. E. Oakley, and J. A. Crowe, *A web-based tool for eliciting probability distributions from experts*, Environ. Model. Softw., 52 (2014), pp. 1–4.
[48] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer-Verlag, New York, 2006.
[49] V. Picheny and D. Ginsbourger, *A nonstationary space-time Gaussian process model for partially converged simulations*, SIAM/ASA J. Uncertain. Quantif., 1 (2013), pp. 57–78, https://doi.org/10.1137/120882834.
[50] J. Ramsay, *Principal differential analysis: Data reduction by differential operators*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 495–508.
[51] J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao, *Parameter estimation for differential equations: A generalized smoothing approach*, J. R. Stat. Soc. Ser. B Stat. Methodol., 69 (2007), pp. 741–796.
[52] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, 2006, http://www.gaussianprocess.org/gpml/.
[53] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global Sensitivity Analysis: The Primer*, John Wiley & Sons, Chichester, UK, 2008.
[54] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*, Springer, New York, 2013.

[55] A. Shah, A. Wilson, and Z. Ghahramani, *Student-t processes as alternatives to Gaussian processes*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, 2014, pp. 877–885.
[56] B. W. Silverman, *Some aspects of the spline smoothing approach to nonparametric curve fitting*, J. Roy. Statist. Soc. Ser. B, 47 (1985), pp. 1–52.
[57] D. Simon, *Evolutionary Optimization Algorithms*, John Wiley & Sons, Hoboken, NJ, 2013.
[58] A. Smith, F. Adler, J. McAuley, R. Gutenkunst, R. Ribeiro, J. McCullers, and A. Perelson, *Effect of* 1918 *PB*1-*F*2 *expression on influenza A virus infection kinetics*, PLoS Comput. Biol., 7 (2011), e1001081.
[59] A. Smith, F. Adler, and A. Perelson, *An accurate two-phase approximate solution to an acute viral infection model*, J. Math. Biol., 60 (2010), pp. 711–726.
[60] A. Smith, F. Adler, R. Ribeiro, R. Gutenkunst, J. McAuley, J. McCullers, and A. Perelson, *Kinetics of coinfection with influenza A virus and Streptococcus pneumoniae*, PLoS Pathog., 9 (2013), e1003238.
[61] A. Smith and A. Smith, *A critical, nonlinear threshold dictates bacterial invasion and initial kinetics during influenza*, Sci. Rep., 6 (2016), 38703.
[62] A. M. Smith, *Host-pathogen kinetics during influenza infection and coinfection: Insights from predictive modeling*, Immuno. Rev., 285 (2018), pp. 97–112.
[63] A. P. Smith, D. J. Moquin, V. Bernhauerova, and A. M. Smith, *Influenza virus infection model with density dependence supports biphasic viral decay*, Front. Microbiol., 9 (2018), 1554.
[64] R. C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*, Comput. Sci. Engrg. 12, SIAM, Philadelphia, 2013.
[65] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, 3rd ed., Springer, New York, 2002.
[66] D. S. Stoffer and K. D. Wall, *Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter*, J. Amer. Statist. Assoc., 86 (1991), pp. 1024–1033.
[67] J. Taubenberger and D. Morens, *The pathology of influenza virus infections*, Annu. Rev. Pathol., 3 (2008), pp. 499–522.
[68] L. Tenorio, *An Introduction to Data Analysis and Uncertainty Quantification for Inverse Problems*, Math. Industry 3, SIAM, Philadelphia, 2017, https://doi.org/10.1137/1.9781611974928.
[69] W. Thompson, D. Shay, E. Weintraub, L. Brammer, N. Bridges, C.B. Cox, and K. Fukuda, *Influenza-associated hospitalizations in the United States*, J. Amer. Med. Assoc., 292 (2004), pp. 1333–1340.
[70] R. Torrence, *Bayesian Parameter Estimation on Three Models of Influenza*, Master's thesis, Virgina Tech, Blacksburg, VA, 2017.
[71] C. R. Vogel, *Computational Methods for Inverse Problems*, Front. Appl. Math. 23, SIAM, Philadelphia, 2002, https://doi.org/10.1137/1.9780898717570.
[72] E. O. Voit, *A First Course in Systems Biology*, Garland Science, New York, 2013.