A Hierarchical Spatiotemporal Statistical Model Motivated by Glaciology

Giri GOPALAN, Birgir HRAFNKELSSON, Christopher K. WIKLE, Håvard RUE, Guðfinna AÐALGEIRSDÓTTIR, Alexander H. JAROSCH, and Finnur PÁLSSON

In this paper, we extend and analyze a Bayesian hierarchical spatiotemporal model for physical systems. A novelty is to model the discrepancy between the output of a computer simulator for a physical process and the actual process values with a multivariate random walk. For computational efficiency, linear algebra for bandwidth limited matrices is utilized, and first-order emulator inference allows for the fast emulation of a numerical partial differential equation (PDE) solver. A test scenario from a physical system motivated by glaciology is used to examine the speed and accuracy of the computational methods used, in addition to the viability of modeling assumptions. We conclude by discussing how the model and associated methodology can be applied in other physical contexts besides glaciology.

Key Words: Model discrepancy; Uncertainty quantification; Emulation.

1. INTRODUCTION

Scientists and engineers often study a physical system with the goal of making spatiotemporal predictions (e.g., temperature or glacier thickness) and inferring unknown quantities governing the system (e.g., atmospheric density or ice viscosity). This system's dynamics can often be phrased in terms of spatiotemporal partial differential equations (PDEs) that are based on approximations. The scientist or engineer may also be able to simulate the physical system with a computer simulator, such as a numerical PDE solver, which is subject to imperfections (e.g., numerical error). Moreover, the scientific constants entering into the system's dynamical equations such as density, friction, or viscosity may not be known precisely, but their range can be constrained to some set of plausible values. Additionally field data, though potentially scarce and noisy, can be incorporated into the analysis.

G. Gopalan (✉), B. Hrafnkelsson, G. Aðalgeirsdóttir, F. Pálsson, University of Iceland, 101 Reykjavík, Iceland (E-mail: *gopalan88@gmail.com*). C. K. Wikle, University of Missouri, Columbia, MO 65211, USA. H. Rue, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia. A. H. Jarosch, University of Innsbruck, 6020 Innsbruck, Austria.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 24, Number 4, Pages 669–692
https://doi.org/10.1007/s13253-019-00367-1

Such scenarios can be modeled with a variant of a Bayesian hierarchical spatiotemporal model that was introduced in Gopalan et al. (2018) for glacial dynamics, if considered more generally. We delineate three methods to make posterior inference efficient: The first is to utilize bandwidth limited linear algebraic routines for likelihood evaluation (Rue 2001), the second is to utilize an embarrassingly parallel approximation to the likelihood, and the third is to use first-order emulators (Hooten et al. 2011) for speeding up computer simulators. Though our modeling and numerical results are still within a glaciology context, we conclude with a discussion of how the model can be applied to other physical scenarios. Before introducing the Bayesian hierarchical model and associated methodology for computationally efficient posterior inference, it is appropriate to summarize relevant statistical literature developed over the last two decades.

Bayesian hierarchical modeling for geophysical problems was introduced in Berliner (1996) and Wikle et al. (1998), and summarized in Berliner (2003), Cressie and Wikle (2011), and Wikle (2016). In this modeling approach, prior distributions are specified for physical parameters of interest, a physical process is modeled at the intermediary, latent level (conditional on the physical parameters), and the data collection process is modeled conditional on the latent physical process values. Both numerical error and model uncertainty can be incorporated at the process level, while measurement errors can be modeled at the data level. This approach has been applied in a variety of scientific contexts, including the study of ozone concentrations (Berrocal et al. 2014), sediment loads at the Great Barrier Reef (Pagendam et al. 2014), precipitation in Iceland (Sigurdarson and Hrafnkelsson 2016), Antarctic contributions to sea level rise (Zammit-Mangion et al. 2014), and tropical ocean surface winds (Wikle et al. 2001) (among many others). In Gopalan et al. (2018), the motivating example for the work in this paper, a Bayesian hierarchical model for shallow glaciers based on the shallow ice approximation (SIA) PDE was developed and evaluated.

Kennedy and O'Hagan (2001) suggest constructing Bayesian statistical models that incorporate the output of a computer simulator of a physical process, such as a numerical solver for the underlying system of PDEs. Fundamental to their approach is the inclusion of a specific term that represents the deviation between the output of a computer simulator and the actual process values, known as *model discrepancy* or *model inadequacy*. This framework is developed in Higdon et al. (2004), Higdon et al. (2008), and Brynjarsdóttir and O'Hagan (2014). In particular, Higdon et al. (2008) use a Bayesian model along with a principal component-based approach for reducing the computational overhead of running a computer simulation with high-dimensional output multiple times (an approach termed as *emulation*). Brynjarsdóttir and O'Hagan (2014) note that the prior for model discrepancy must be chosen carefully to mitigate bias of physical parameters and predictions. In particular, as more prior information is incorporated into a model discrepancy term through a constrained Gaussian process (GP) prior over a space of functions, the less biased inferences and predictions tend to become. The notions of an emulator, a computer simulator, and model discrepancy enter naturally into the aforementioned Bayesian hierarchical framework. Conditional on physical parameters coupled with initial and/or boundary conditions, the physical process values at the latent level can be written as the sum of a computer simulator or emulator term and a model discrepancy term.

To be precise, let us assume that the physical process $\boldsymbol{S}$ can be indexed through time, i.e., as $\boldsymbol{S}_j$, and $\boldsymbol{S}_j$ is a vector where each element corresponds to a distinct spatial location. One can specify the process-level conditional on physical parameter $\boldsymbol{\theta}$ as

$$\boldsymbol{S}_j = \boldsymbol{f}(\boldsymbol{\theta}, j) + \boldsymbol{\delta}(j), \tag{1}$$

where $\boldsymbol{\delta}(.)$ is a vector-valued model discrepancy function that is independent of $\boldsymbol{\theta}$, and $\boldsymbol{f}(\boldsymbol{\theta}, j)$ is the output of a computer simulation or emulator for physical parameter $\boldsymbol{\theta}$ at time index j. If, for instance, at each time point j an observation $\boldsymbol{Y}_j$ of $\boldsymbol{S}_j$ is made with associated measurement error $\boldsymbol{\eta}_j$, then observations can be written as

$$\boldsymbol{Y}_j = \boldsymbol{f}(\boldsymbol{\theta}, j) + \boldsymbol{\delta}(j) + \boldsymbol{\eta}_j, \tag{2}$$

which is analogous to Eq. 5 of Kennedy and O'Hagan (2001).

In Kennedy and O'Hagan (2001), $\boldsymbol{\delta}(.)$ is a fixed but unknown function independent of $\boldsymbol{\theta}$ that is learned with a GP prior distribution. Similarly, $\boldsymbol{\delta}(.)$ has a constrained GP prior in Brynjarsdóttir and O'Hagan (2014). The approach in this paper instead assumes a temporally indexed stochastic process (with spatial correlation) that follows a multivariate random walk, rather than a deterministic function. Additionally, in Liu and West (2009), the authors frame a computer emulator of time series run under multiple inputs as a dynamic linear model (DLM). As part of their approach, they allow for time-varying auto-regressive coefficients that follow a random walk process, to embed non-stationarity into the model.

While the approach taken in this paper most closely follows the above literature (i.e., Bayesian hierarchical modeling, model discrepancy, and emulation), we briefly review the literature in probabilistic numerics and Bayesian numerical analysis; the emphasis in Bayesian numerical analysis is to use probabilistic methods to solve numerical problems, whereas, in the Bayesian hierarchical setup, one is also interested in inference of scientifically relevant parameters and predictions of the physical process. In Conrad et al. (2017), a probabilistic ordinary differential equation (ODE) solver is developed that adds stochasticity at each iteration; conditions for the convergence of this method to the ODE solution are given. Chkrebtii et al. (2016) utilize GPs for solving ODEs; moreover, Calderhead et al. (2008) use a GP regression-based method to avoid explicitly solving nonlinear ODEs when performing inference for parameters that provides computational speedups; additionally, Owhadi and Scovel (2017) present a gamblet-based solver that comes with provably computationally efficient solutions to PDEs. The approach is derived from a game theoretic and stochastic PDE framework.

In the spatiotemporal model described in this paper, stochasticity is induced with an error-correcting process that is separated from the numerical solution. In general, another way to achieve this is to define a stochastic process by equating a PDE to a white noise term—that is, the solution $\boldsymbol{X}$ to a stochastic partial differential equation (SPDE) $L[\boldsymbol{X}] = \boldsymbol{W}$, where $\boldsymbol{L}$ is a differential operator and $\boldsymbol{W}$ is a white noise process (indexed by spatiotemporal coordinates). For instance, a fractional Laplacian operator yields the Matérn covariance function (Whittle 1954, 1963; Lindgren et al. 2011). We employ the former approach mainly because it is difficult to derive exact covariance functions for arbitrary differential equations (e.g., in

the presence of non-linearities), though we highlight the utility of the latter approach in situations where an analytical covariance function can be derived exactly.

A major feature of this work is to represent the discrepancy between real physical process values and the output of a computer simulator for these physical process values as a multivariate random walk; typically, model discrepancy is endowed with a GP prior or a constrained GP prior over a space of functions as in Kennedy and O'Hagan (2001) and Brynjarsdóttir and O'Hagan (2014). Along with this model is the development of two ways for making computations faster: The first is harnessing first-order emulator inference (Hooten et al. 2011) for speeding up the computation of a numerical solver and the second is the use of bandwidth limited numerical linear algebra (Rue 2001) for computing the likelihood efficiently. Moreover, in the regime of a high signal-to-noise ratio, an embarrassingly parallel approximation to the likelihood can be employed. Finally, methodology to fit a spatial Gaussian field for the log of the scale of numerical errors is discussed.

We must also be clear about what distinguishes this work from its predecessor (Gopalan et al. (2018)). This includes the use of emulators, probing higher-order random walks besides order 1, derivation of sparsity and computational complexity of log-likelihood evaluation, empirical run-time results, and methodology to fit an error-correcting process when little prior information is available. The structure of this paper is as follows: First a test system from glaciology is described. Then the statistical model that is the focus of this work is presented in detail (in the context of the glaciology test case), followed by the exact and approximate likelihood. Then this model is analyzed in terms of computational run-time and accuracy of inference, based on the test system from glaciology; moreover, the random walk error-correcting process is assessed with residual analysis. Afterward, we discuss how the model and associated methodology can be applied to other physical scenarios, and conclude by summarizing the model, method, and limitations of the approach.

2. DESCRIPTION OF A TEST SYSTEM FROM GLACIOLOGY

Before delving into the specifics of the Bayesian hierarchical model and computational subtleties, we begin with a brief discussion of glaciology. Glaciology is the study of physical systems consisting mostly of ice and snow. This broad definition includes the study of the crystalline nature of ice, the transformation and compaction of snow into ice, the dynamics of the flow of ice and water in a glacier, the relationships between fundamental quantities like viscosity, temperature, and pressure, the relationships between precipitation and meteorology with said ice systems, the interaction of ice systems with other geological systems such as volcanoes and bedrock, and so on. As such, glaciology synthesizes elements from a multitude of scientific disciplines including continuum mechanics, fluid mechanics, hydraulics, chemistry, and meteorology.

Bueler et al. (2005) introduce analytical solutions for the SIA PDE, a commonly used model for the dynamics of glaciers (Fowler and Larson 1978; Hutter 1982; Flowers et al. 2005; Cuffey and Paterson 2010; van der Veen 2013; Brinkerhoff et al. 2016; Guan et al. 2016; Gopalan et al. 2018). Based on the principle of conservation of mass, the SIA dictates that glacier flow is in the direction of the (negative) gradient of the glacier surface and is

due to gravity and basal sliding (also referred to as friction or drag if in the direction of the positive gradient). While an explanation of the SIA PDE is given in Gopalan et al. (2018), our focus is on ice viscosity, B. Intuitively, this parameter controls the softness of the ice. The other main physical parameter, which is not the subject of this paper, is $C_0\gamma$. This controls basal sliding or friction.

For the analysis that follows, we focus on a periodic solution to the SIA in which the thickness of the glacier oscillates through time; $H(r,t)$, the thickness of the glacier as a function of two-dimensional space (in polar coordinates) and time, is

$$H(r,t) = H_s(r) + P(r,t), \tag{3}$$

$$P(r,t) = C_p \sin(2\pi t/T_p)\cos^2\left[\frac{\pi(r-0.6L)}{.6L}\right]; \text{ if } 0.3L < r < .9L, \tag{4}$$

$$P(r,t) = 0; \text{ if } 0 \le r \le 0.3L \text{ or if } r \ge 0.9L. \tag{5}$$

In Eq. 3, H_s is a static initial profile of the glacier (i.e., a dome as in Eq. 21 of Bueler et al. (2005)), P is a perturbation (e.g., precipitation) function, L is the margin length, C_p is the magnitude of the periodic perturbation, and T_p is the period of the perturbation. Bueler et al. (2005) derive a mass balance function that achieves this periodic solution for the SIA PDE. Qualitatively, this test case appears like a dome with a periodic oscillation in thickness around an annulus defined by $0.3L < r < .9L$. In Fig. 1, an illustration of the oscillations of glacier thickness through time is displayed.

The value of each surface elevation measurement is the value of the exact analytical solution above added to a zero-mean Gaussian random variable with standard deviation of 1 m, larger than errors of the digital GPS instruments employed by the UI-IES. We use the same values of parameters as in Bueler et al. (2005) to make for easier comparison with that work and the EISMINT experiment. In particular, $H_0 = 3600$ m, $L = 750$ km, $C_p = 200$ m, and $T_p = 5000$ years.

Employing the same setup as Gopalan et al. (2018), glacial surface elevation measurements are assumed to be collected for 20 years, twice a year, and at 25 fixed spatial locations across the glacier, to emulate how the glaciology team at the University of Iceland Institute of Earth Sciences (UI-IES) collects data at Icelandic glaciers (e.g., see Fig. 2 illustrating Langjökull and the mass balance measurement sites).

3. THE HIERARCHICAL SPATIOTEMPORAL MODEL AND ITS PROPERTIES

Now that we have acquainted the reader with some facts about glaciology and the particular test case used for the analysis in this paper, we next delineate the hierarchical spatiotemporal model that is the focus of this work, by specifying its variables, parameters, and properties, including efficient computation of the likelihood and connections to other modeling frameworks. For the sake of specificity of the presentation, the glaciology example is referenced, similar to the setup in Gopalan et al. (2018). We assume that n spatial locations are modeled at the latent level, and m of those locations are observed, where m is typically

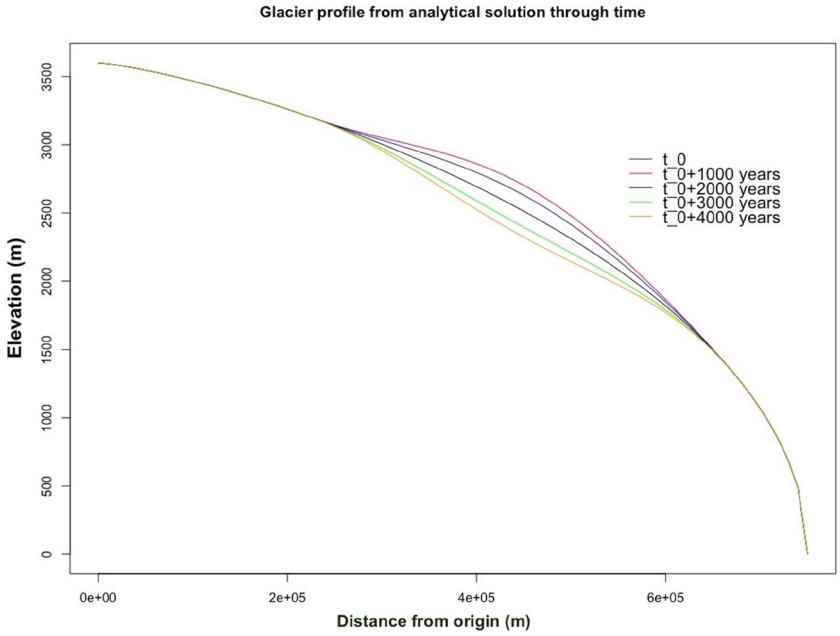

Figure 1. An illustration of the periodic oscillatory exact solution to the SIA PDE that is used for the analysis. Since the solution is radially symmetric, only a radial cross section is illustrated. This solution is stationary except for an annulus defined by $0.3L < r < .9L$, where L is 750 km; in the annulus, the glacier thickness vibrates back and forth periodically, as illustrated..

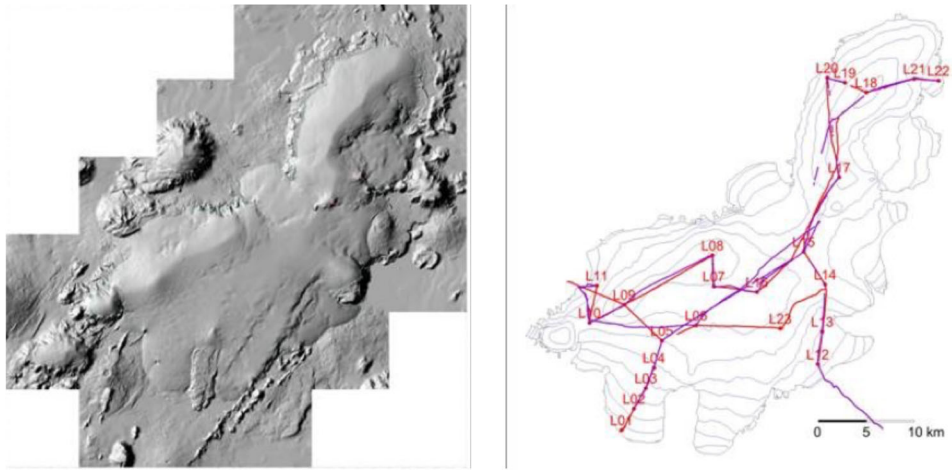

Figure 2. A digital elevation map of Langjökull along with measurement sites demarcated on the right, provided by the University of Iceland Institute of Earth Sciences (UI-IES). Langjökull is Iceland's second largest glacier by area, at 900 sq. km, and its mean thickness is 210 m above sea level (Björnsson and Pálsson 2008), so Langjökull is shallow. .

much smaller than n. We use the index j to refer to time indices and i to refer to spatial indices; while space and time are discretized, the differences between successive time and spatial points can be made as small as desired depending on the context of the application and computational resources available. Throughout, we use bolded notation for vectors and uppercase, unbolded, and non-italic notation for matrices. All other mathematical symbols are scalars.

We introduce the Bayesian hierarchical model in the parameter, process, data-level framework of Berliner (1996). We denote the physical parameters as $\boldsymbol{\theta}$ and initial and/or boundary conditions for the physical process as $\boldsymbol{\phi}$. At the parameter level, one possibility is to use a truncated normal distribution for $\boldsymbol{\theta}$ if the support of the parameter value can be constrained, as was done in Gopalan et al. (2018), where $\boldsymbol{\theta}$ represented ice viscosity. However, more generally, the distribution can be specified based on domain knowledge or expertise. We denote the output of a computer simulator, which could be either a numerical solver or an emulator, at time j with the notation $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, j)$, which, in full generality, is an element of $\mathbb{R}^n$. While some values could be negative (e.g., temperature), in many cases the computer simulator output can be restricted to the nonnegative real numbers. For a specific example, in Appendix A of Gopalan et al. (2018), $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, j)$ is a second-order finite difference solver for glacier thickness, which is constrained to be nonnegative based on a boundary condition. Evidence for a nonnegative support for the physical process, in glaciology, can be found in Gopalan et al. (2018). Particularly, this is evident in Fig. 6 of that paper, which shows the process (i.e., glacier thickness) predictions across the glacier, and the distributions are all greater than zero. Specifically, the minimum of the smallest box-plot is more than 750 m. Nonetheless, the reader is suggested to think carefully about whether a negligible amount of probability mass is below zero in different applications (e.g., temperature models).

The process level of the model, conditional on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, can be written as:

$$\boldsymbol{X}_j = \boldsymbol{X}_{j-1} + \boldsymbol{\epsilon}_j, \tag{6}$$

$$\boldsymbol{S}_j = \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, j) + \boldsymbol{X}_j, \tag{7}$$

where $\boldsymbol{X}_0$ is a vector of zeros.

In the above expressions, $\boldsymbol{\epsilon}_j$ is MVN$(0, \Sigma)$ and independent of $\boldsymbol{\epsilon}_l$ for $j \neq l$. Furthermore, $\boldsymbol{X}_j$, $\boldsymbol{\epsilon}_j$, $\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, j)$, and (consequently) $\boldsymbol{S}_j$ are members of $\mathbb{R}^n$. In Gopalan et al. (2018), $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \}$ was referred to as an error-correcting process because it was meant to represent the difference between the numerical solver and the exact solution to the SIA PDE. Note that in Gopalan et al. (2018), $\boldsymbol{S}_j$ referred to glacial thickness at a particular time point, where each component referred to the glacial thickness at a particular grid point. In more generality, the error-correcting statistical process can be a random walk of higher order; a multivariate RW process of order q $(RW(q))$ is given by:

$$\boldsymbol{X}_j + \sum_{p=1}^{q} (-1)^p \binom{q}{p} \boldsymbol{X}_{j-p} = \boldsymbol{\epsilon}_j, \tag{8}$$

where $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_q$ are independent and marginally MVN$(0, \Sigma)$. This form of a higher-order random walk is a multivariate extension of the integrated auto-regressive process given in

Chapter 5.6 of Madsen (2007). For $q = 2$, this corresponds to RW(2) of Rue and Held (2005).

At the data level, it is assumed that data are regularly sampled at every kth time point, so that one observes $\boldsymbol{Y}_k, \boldsymbol{Y}_{2k}, \ldots, \boldsymbol{Y}_{Nk} \in \mathbb{R}^m$; in the glaciology test case, the variables $\boldsymbol{Y}$ referred to glacial surface elevation measurements, and k was set to 5, to represent the fact that the glaciologists take a set of measurements in the summer and winter, or twice a year. The corresponding observation errors $\boldsymbol{\eta}_k, \boldsymbol{\eta}_{2k}, \ldots, \boldsymbol{\eta}_{Nk}$ are IID MVN$(0, \sigma^2 \mathrm{I})$ and represent digital GPS measurement errors in the glaciology example. We define the matrix $\mathrm{A} \in \mathbb{R}^{m \times n}$ to be such that its rows are unit basis vectors (i.e., an incidence matrix as in Cressie and Wikle (2011)). That is, $\mathrm{A}_{ab} = 1$ if and only if the bth index of the process-level vector $\boldsymbol{S}$ has been observed, and $\mathrm{A}_{ab} = 0$ for all other entries. Then the data-level model, conditional on the process $\boldsymbol{S}$, is

$$\boldsymbol{Y}_{ck} = \mathrm{A}\boldsymbol{S}_{ck} + \boldsymbol{\eta}_{ck}, \tag{9}$$

where we assume that $j \in \{1, 2, \ldots, T\}$ and $c \in \{1, 2, \ldots, N\}$, so there are N total observed spatial vectors, observed with a period of length k.

Conditional on $\boldsymbol{\theta}, \boldsymbol{\phi}$, and a computer simulator, the model can be thought of as a hidden Markov model (HMM) (Baum and Petrie 1966); the latent physical process evolves according to a RW(1) process added to a numerical solution, and it is observed indirectly with Gaussian noise. It can also be thought of as a conditional general state space model. This is because, conditioning on $\boldsymbol{\theta}, \boldsymbol{\phi}$, and a computer simulator, one can write:

$$\boldsymbol{S}_j = \boldsymbol{S}_{j-1} + [-\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, j-1) + \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, j)] + \boldsymbol{\epsilon}_j, \tag{10}$$

$$\boldsymbol{Y}_{ck} = \mathrm{A}\boldsymbol{S}_{ck} + \boldsymbol{\eta}_{ck}. \tag{11}$$

Here, the state evolves linearly with a time-dependent offset term: $[-\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, j-1) + \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, j)]$. The notation ck is used in Eq. 11 to indicate that observations of the process are only observed every kth time point, whereas the latent process evolves at every time step j. The reader who is interested in further understanding the connection between Gaussian processes and state space models may consult Solin and Särkkä (2014).

3.1. Exact Likelihood

An advantage of using this model is that the likelihood, $p(\boldsymbol{Y}_k, \boldsymbol{Y}_{2k}, \ldots, \boldsymbol{Y}_{Nk} | \boldsymbol{\theta}, \boldsymbol{\phi})$, can be computed exactly in an efficient manner. It can also be approximated in a way that leads to embarrassingly parallel computation when the signal-to-noise ratio is high. The next several sections provide more details for these considerations. The likelihood of the model, $L(\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{Y}_k, \boldsymbol{Y}_{2k}, \ldots, \boldsymbol{Y}_{Nk} | \boldsymbol{\theta}, \boldsymbol{\phi})$, has a multivariate normal PDF form:

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{(2\pi)^{(mN)/2} |\Sigma_l|^{1/2}} \exp^{-(\boldsymbol{Y} - \boldsymbol{\mu}_l)^T \Sigma_l^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}_l)/2}, \tag{12}$$

where the mean is:

$$\boldsymbol{\mu}_l = (\mathrm{A}\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, k), \ldots, \mathrm{A}\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, Nk)), \tag{13}$$

and the covariance matrix is:

$$\Sigma_l = \mathrm{U} \otimes \mathrm{V} + \sigma^2 \mathrm{I}, \tag{14}$$

where $\mathrm{U}_{ab} = k \min(a, b)$ with $\mathrm{U} \in \mathbb{R}^{N \times N}$, and $\mathrm{V} = \mathrm{A}\Sigma\mathrm{A}^\intercal$. Also, the symbol $\otimes$ stands for the Kronecker product. $\boldsymbol{Y}_{ck}$ is multivariate normal (conditioning on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$) as a direct consequence of Eqs. 7 and 9, noting that $\boldsymbol{X}_{ck}$ and $\boldsymbol{\eta}_{ck}$ are independent conditional on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Moreover, the linearity property of expectations can be used to show that the mean of $\boldsymbol{Y}_{ck}$ is $E[\mathrm{A}\boldsymbol{S}_{ck} + \boldsymbol{\eta}_{ck}] = E[\mathrm{A}\boldsymbol{S}_{ck}] + E[\boldsymbol{\eta}_{ck}] = E[\mathrm{A}\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, ck) + \mathrm{A}\boldsymbol{X}_{ck}] + E[\boldsymbol{\eta}_{ck}] = E[\mathrm{A}\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, ck)] + E[\mathrm{A}\boldsymbol{X}_{ck}] + E[\boldsymbol{\eta}_{ck}] = \mathrm{A}\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, ck) + \mathbf{0} + \mathbf{0}$ (again, conditional on fixed $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ fixed). Appendix A contains more details of the covariance matrix.

Since evaluating Eq. 12 requires the calculation of the inverse of the matrix Σ_l and its determinant, these must be calculated efficiently. (Generally this takes $O(N^3 m^3)$ operations, which can grow very quickly with more space and time observations.) Since U^{-1} is tridiagonal, the bandwidth of U^{-1} is 1, and the band-limited nature of U^{-1} allows us to compute Σ_l^{-1} and $|\Sigma_l|$ in $O(Nm^3)$ time (Rue 2001; Golub and Van Loan 2012). More details for this derivation are given in Appendix A. While using band-limited linear algebra routines can improve computation, in the next subsection we derive an approximation to the likelihood that is embarrassingly parallel and can therefore accelerate computation even more.

3.2. An Approximate Likelihood

Here we show how to approximate the likelihood in a way that leads to embarrassingly parallel computation. The likelihood $p(\boldsymbol{Y_k}, \ldots, \boldsymbol{Y_{Nk}} | \boldsymbol{\theta}, \boldsymbol{\phi})$ can be equivalently written as $p(\boldsymbol{Y_k} | \boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{Y_{2k}} | \boldsymbol{Y_k}, \boldsymbol{\theta}, \boldsymbol{\phi}) \ldots p(\boldsymbol{Y_{Nk}} | \boldsymbol{Y_k}, \ldots, \boldsymbol{Y_{(N-1)k}}, \boldsymbol{\theta}, \boldsymbol{\phi})$. First note that:

$$\boldsymbol{Y}_k = \mathrm{A}\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, k) + \boldsymbol{\eta}_k + \sum_{j=1}^{k} \mathrm{A}\boldsymbol{\epsilon}_j. \tag{15}$$

Hence, $p(\boldsymbol{Y}_k | \boldsymbol{\theta}, \boldsymbol{\phi})$ is multivariate normal with mean $\mathrm{A}\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, k)$ and covariance matrix $\mathrm{A}(k\Sigma)\mathrm{A}^\intercal + \sigma^2 \mathrm{I}$. More generally, we have the relationship:

$$\begin{aligned} \boldsymbol{Y}_{ck} = \boldsymbol{Y}_{(c-1)k} + \mathrm{A}[\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, ck) - \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, (c-1)k)] + \boldsymbol{\eta}_{ck} - \boldsymbol{\eta}_{(c-1)k} \\ + \sum_{j=(c-1)k+1}^{ck} \mathrm{A}\boldsymbol{\epsilon}_j. \end{aligned} \tag{16}$$

Thus, we can approximate $p(\boldsymbol{Y}_{ck} | \boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_{(c-1)k}, \boldsymbol{\theta}, \boldsymbol{\phi})$ as a MVN distribution with mean $\boldsymbol{Y}_{(c-1)k} + \mathrm{A}[\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, ck) - \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, (c-1)k)]$ and covariance matrix $\mathrm{A}(k\Sigma)\mathrm{A}^\intercal + 2\sigma^2 \mathrm{I}$. Nonetheless, to clarify, $p(\boldsymbol{Y}_{ck} | \boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_{(c-1)k}, \boldsymbol{\theta}, \boldsymbol{\phi})$ is not exactly a MVN with mean

$\boldsymbol{Y}_{(c-1)k} + \mathrm{A}[\boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, ck) - \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, (c-1)k)]$ and covariance matrix $\mathrm{A}(k\Sigma)\mathrm{A}^\intercal + 2\sigma^2\mathrm{I}$ because $\boldsymbol{Y}_{(c-1)k}$ and $\boldsymbol{\eta}_{(c-1)k}$ are dependent. However, when the magnitude of the observation error $\boldsymbol{\eta}_{(c-1)k}$ is much smaller in comparison with the magnitude of the observation $\boldsymbol{Y}_{(c-1)k}$, and for $\boldsymbol{Z} \sim \mathrm{MVN}(0, \sigma^2\mathrm{I})$ with $\boldsymbol{Z}$ independent of $\boldsymbol{Y}_{(c-1)k}$, $\boldsymbol{Y}_{(c-1)k} - \boldsymbol{\eta}_{(c-1)k} \approx \boldsymbol{Y}_{(c-1)k} - \boldsymbol{Z}$.

This approximation is embarrassingly parallel because each of the N terms in the product form of the likelihood $p(\boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_T|\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{Y}_k|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{Y}_{2k}|\boldsymbol{Y}_k, \boldsymbol{\theta}, \boldsymbol{\phi}) \ldots p(\boldsymbol{Y}_{Nk}|\boldsymbol{Y}_k, \ldots, \boldsymbol{Y}_{(N-1)k}, \boldsymbol{\theta}, \boldsymbol{\phi})$ (or sum, if computing the log-likelihood) can be evaluated independently of each other. Therefore, in parallel, the computation comes down to evaluating a multivariate normal PDF of dimension m – this can be done in $O(m^3)$.

3.3. COMPUTATIONAL COMPLEXITY SUMMARY

If no attention is paid to the structure of Σ_l, the cost of evaluating $L(\boldsymbol{\theta}, \boldsymbol{\phi})$ is limited by the evaluation of Σ_l^{-1} and $|\Sigma_l|$, which generally takes $O(N^3m^3)$ operations. However, the exact likelihood evaluation can be reduced to $O(Nm^3)$ using band-limited numerical linear algebra. The computational complexity of the approximation is also $O(Nm^3)$ (if no parallelism is used). While an exact likelihood is preferred to an approximation, a benefit of the approximation is that it is embarrassingly parallel—if parallelized, the time complexity is that of evaluating a multivariate normal PDF of dimension m, which is $O(m^3)$. Nonetheless, there also exist parallel versions of sparse Cholesky decomposition, for instance in Gupta and Kumar (1994). Empirical comparisons of the exact and approximate likelihood computations are presented in Sect. 4.

4. ANALYSIS OF THE MODEL AND ASSOCIATED METHODOLOGY

The purpose of this section is to motivate the various modeling choices introduced in this paper using the previously described test system from glaciology, in terms of both computational run-time and quality of inferences. In particular, we compare a posterior based on an emulator to a posterior based on a numerical PDE solver, motivate the use of the random walk error-correcting process with residual analysis, examine the impact of prior information encoded into the error-correcting process on the bias of posterior distributions for physical parameters, and compare the run-time and accuracy of the likelihood approximation versus the exact likelihood. The physical parameter of interest in these examples is ice viscosity, B, whose actual value is the same as Bueler et al. (2005), Payne et al. (2000), and Gopalan et al. (2018): 31.7×10^{-25} in units of $\mathrm{s}^{-1}\,\mathrm{Pa}^{-3}$.

Consistent with Gopalan et al. (2018) is the choice of settings for the numerical PDE solver: A 21 by 21 grid (so $n = 441$) is used with $\Delta_x = \Delta_y = 10^5$ m and $\Delta_t = .1$ years. Note that, consequently, the number of simulator runs (25) is much smaller than the dimensionality of the output of the solver (441).

Table 1. Summary statistics of 10^6 posterior samples of the ice viscosity parameter using an emulator for the SIA and a numerical solver for the SIA; qualitatively, these posterior samples are similar.

Test case	Min	1st quartile	Median	Mean	3rd quartile	Max
Emulator SIA solver	15.0	26.5	27.0	27.4	29.0	38.5
Numerical SIA solver	15.0	24.5	26.5	26.3	28.0	37.5

Units are in $10^{-25}\,s^{-1}\,Pa^{-3}$.

4.1. Posterior Inference of the Ice Viscosity Parameter with an Emulator Compared to a Numerical PDE Solver

In this section, we conduct an empirical study to examine how a first-order spatiotemporal emulator (i.e., an emulator based on the method in Appendix B) compares to a numerical solver of the PDE, in terms of both run-time of computations and posterior inference of ice viscosity. While the precise technical details for constructing a first-order spatiotemporal emulator are given in Appendix B, the idea is to approximate the numerical solver output for each time point that there is collected data. To do this, we train an emulator using the following values for ice viscosity: $\{10, 12.5, 15.0, \ldots, 70.0\}$ in units of $10^{-25}\,s^{-1}\,Pa^{-3}$, a grid of values that is intentionally coarser than the values used for posterior computation, since in this case the emulator must be used for parameter values not in the training set. We used the `rbenchmark` (Kusnierczyk 2012) package to benchmark the run-time of the log-likelihood of the model evaluated at the actual parameter value computed with a numerical solver versus a first-order spatiotemporal emulator, using a MacBook Pro early 2015 model with a 2.7 GHz Intel Core i5 processor and 8 GB 1867 MHz DDR3 memory. The emulator version performs 14.5 times faster (.354 s for the emulator-based log-likelihood versus 5.148 s for the numerical solver-based log-likelihood). We also generated samples from the posterior distribution of ice viscosity with grid sampling (grid [10,70] inclusive with grid width .50 in units of $10^{-25}\,s^{-1}\,Pa^{-3}$), using both the numerical PDE version and the emulated version. The summary statistics of 10^6 posterior samples for ice viscosity using both the emulator and numerical solver are given in Table 1. Qualitatively, the summary statistics are similar.

The principle behind choosing the ice viscosity parameter values in the training set is to fill the space of the support for ice viscosity, but not to choose a grid as fine as the one used for posterior sampling. (Such an approach would be circular, in that the emulator would just be generating predictions inside of the training set.) However, such a heuristic will not be feasible as the number of parameters grows beyond one parameter. (The number of design points would need to grow exponentially in the number of dimensions.) In such cases, we suggest using other space-filling designs: Notably, a latin hypercube design has been used extensively in the computer experiments literature, for instance in Higdon et al. (2008).

4.2. Assessing a Random Walk for Representing Model Discrepancy

The choice of using a random walk to correct for deviations between the output of a computer simulator and the actual physical process values has a few important motivations:

1. The inaccuracy of a spatiotemporal computer simulation is most likely going to increase as it is run further into the future. Conveniently, a random walk's variance increases with time—for example, a RW(1) has marginal variance $j\Sigma$ at time j.
2. As shown in Appendix A and Sect. 3.1, the likelihood involves band-limited matrices, for which there exist specialized numerical linear algebra routines. However, there is a trade-off between bandwidth and the order of the random walk utilized.
3. Spatial correlations in the inaccuracies of a computer simulation can be captured with the covariance matrix Σ.

In addition to these motivations, the purpose of this section is to empirically assess how a random walk model performs for correcting the output of a numerical SIA PDE solver. To do this, we use the analytical SIA solution as a gold standard. This is a simplification in the sense that the real glacial dynamics will not follow the SIA PDE and therefore the analytical SIA solution exactly, but nonetheless this is a way to check the veracity of the random walk error model in some capacity—at the very least, as a model for numerical error but not model uncertainty.

Figure 3 displays the differences between the analytical SIA PDE solution for glacial thickness and the numerical SIA PDE solution for glacial thickness at all of the glacier grid points, run forward for 5000 time steps (i.e., 500 years). More precisely, the points in blue are at the margin of the glacier, the points in red are at the interior, and the points filled in black are close to the top (also referred to as the dome) of the glacier. Recall from Fig. 1 that the glacier looks like a shallow ellipsoid sliced in half (in the x–y plane), and so, the panel on the right of this figure is a top view of the glacier grid points, which looks like a circle of radius 750 km projected onto the x–y plane. In comparison, the height is 3600 m.

The differences are all very smooth (i.e., continuous) functions of time, implying that the numerical SIA PDE solver is producing continuous output as well—we know that the analytical solution is continuous based on the functional form in Eqs. 3–5. Thus, it appears that a random walk of at least a few orders is necessary to represent these differences. Moreover, as expected from Bueler et al. (2005), the largest errors occur at the margin, whereas the interior and dome differences are less extreme.

To assess whether a random walk model is appropriate, for each time point j and for orders 1–7, we computed residuals, in other words, the left-hand side of Eq. 8, which should theoretically be distributed like $\boldsymbol{\epsilon}_j$ (i.e., independent MVN(0, Σ) random variables). To compute $\boldsymbol{X}_j$, we take the difference $\boldsymbol{S}_j - \boldsymbol{f}(\boldsymbol{\theta}, j)$, where $\boldsymbol{S}_j$ is the analytical glacial thickness solution to the SIA PDE at time j (i.e., the real physical process for the purpose of this analysis) and $\boldsymbol{f}(\boldsymbol{\theta}, j)$ is the numerical glacial thickness solution to the SIA PDE at time j. We examine the residuals for two randomly selected grid points of the glacier (one at the interior and one at the margin) in Figs. 4 and 5.

A few important observations should be emphasized based on the empirical analysis displayed in these figures. The first is that a single-order random walk substantially filters the discrepancy; for the interior grid point, it is reduced from the order of 10 m to the order of .01 m (1000 times reduction in magnitude), and for the margin grid point from the order of 100–.05 m (more than 1000 times reduction). Additionally, for both the interior and margin

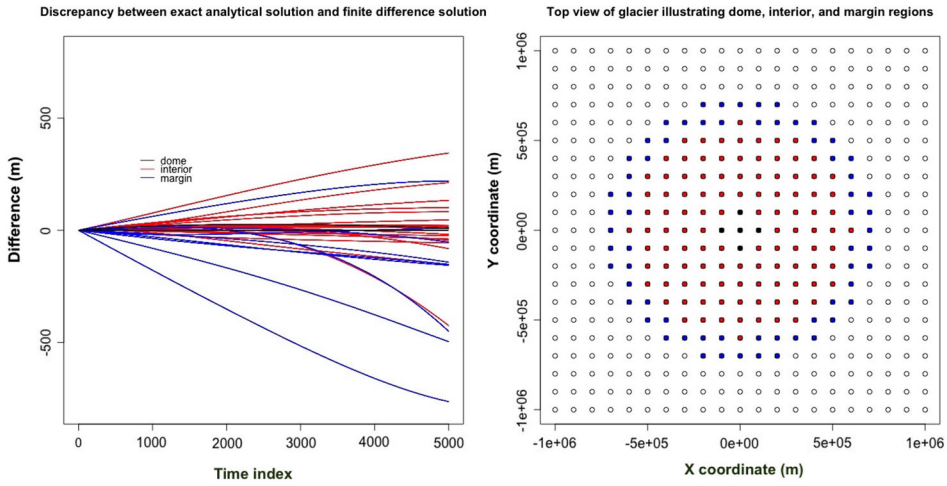

Figure 3. An illustration of the difference between the exact analytical solution and the numerical solution for the SIA PDE. On the *right* panel is a top view of the glacier, whose shape looks like a dome, and therefore, the projection on to the x–y plane is a circle. The blue points signify the margin of the glacier (where it drops down to zero thickness), the red points are at the interior of the glacier, and the black points are toward the top of the glacier. The points that are not filled in signify the border of the glacier, where there is no ice thickness. On the *left* panel, the discrepancies between the analytical SIA PDE solution and the numerical SIA PDE solution for all grid points are shown. Specifically, the color of each path corresponds to the grid points on the right panel. Additionally, the paths are shown for 500 years, or 5000 time steps. .

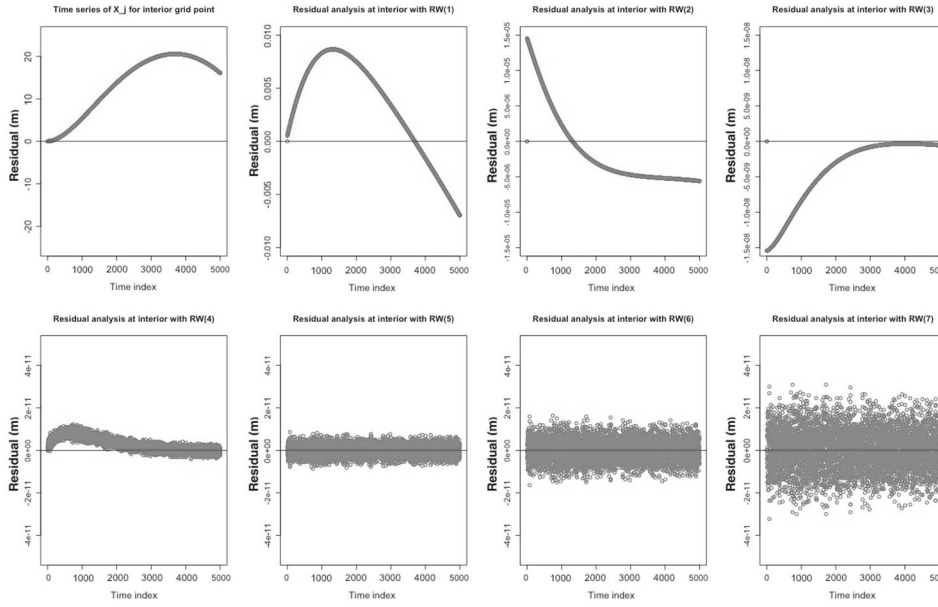

Figure 4. This figure displays residuals in units of meters (i.e., the term ϵ_j in Eq. 8) for RW(q) of orders 1–7 for a randomly selected interior grid point. The first four panels display values on different scaled *y* axes to better show the shapes, whereas the bottom four panels have the same scaling for the *y* axis to be able to compare across the figures. RW(5) and above look like white noise processes, though RW(5) has the smallest variance..

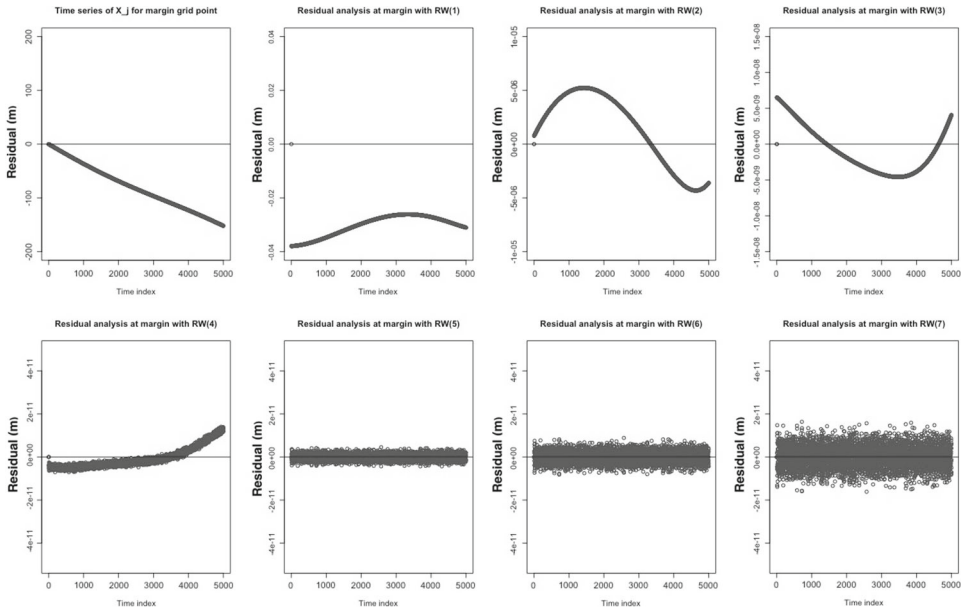

Figure 5. This figure also displays residuals in units of meters (i.e., the term ϵ_j in Eq. 8) for RW(q) of orders 1–7 for a randomly selected margin grid point. Just like the previous figure, the first four panels display values on different scaled y axes to better show the shapes, whereas the bottom four panels have the same scaling for the y axis to be able to compare across the figures. Just as in the previous figure, RW(5) and above look like white noise processes, though RW(5) has the smallest variance. .

grid points, it appears that RW(5) is optimal in the sense that the residuals closely resemble a white noise process and have the smallest variance. While the residuals from higher-order RW processes also resemble white noise, the magnitude of the noise is larger. Nonetheless, we believe that real physical processes will not always be as smooth as the analytical SIA PDE solutions, and hence, it is likely that a lower-order RW process will be preferred for real scenarios.

4.3. REDUCING BIAS FOR THE POSTERIOR DISTRIBUTION OF θ

In Brynjarsdóttir and O'Hagan (2014), when prior information about the model discrepancy term is introduced in a simple physical system (i.e., a constrained GP over a space of functions), the bias of a posterior distribution of a physically relevant parameter reduces. We have found that a very similar phenomenon occurs in the glaciology test case, a result that was pointed out in Gopalan et al. (2018). Specifically, in Bueler et al. (2005), it is shown that there is large spatial variation in the scale of deviations between the exact solution to the SIA and a numerical finite difference solver of the SIA. Specifically, there is spatial variation between the dome, interior, and margin of a glacier, with deviations at the margin being markedly larger than at the interior and dome. To investigate the effect of such prior information, we choose the matrix Σ to be such that it is block diagonal with 3 blocks: Σ_{int}, Σ_{dome}, and Σ_{margin}. Each of these blocks is derived from a square exponential covariance kernel with the same length scale parameter $\boldsymbol{\phi} = 70\,\text{km}$, but differing variance parameters

σ^2_{int}, σ^2_{dome}, and σ^2_{margin}. If we ignore prior information from Bueler et al. (2005), we assume that there is an equal prior probability that each of σ^2_{int}, σ^2_{dome}, and σ^2_{margin} is in the set $\{.1, 1, 10, 100\}$ in units of m^2. If we use prior information from Bueler et al. (2005), we instead assume equal prior probability on $\{.1, 1\}$ for σ^2_{int}, $\{1, 10\}$ for σ^2_{dome}, and $\{10, 100\}$ for σ^2_{margin} (again all units are m^2). As shown in Gopalan et al. (2018), the posterior for ice viscosity is less biased in the case that incorporates prior information for the scale of errors; this phenomenon is explored again in the next section.

While in the above discussion we have not been precise about the term bias, the following ought to make this notion more rigorous. Let $\boldsymbol{\theta}_0$ be the true parameter and $\hat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}_0$. The frequentist definition of bias is usually $E[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0]$, where the expectation (i.e., average) is taken over the sampling distribution, $p(\boldsymbol{Y}|\boldsymbol{\theta}_0)$. The Bayesian notion of bias used informally in the preceding paragraph (and essentially the same notion as in Brynjarsdóttir and O'Hagan (2014)) is $b(\boldsymbol{Y}, \boldsymbol{\theta}_0) = E[\boldsymbol{\theta} - \boldsymbol{\theta}_0]$, where the expectation (i.e., average) is taken with respect to the posterior distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\boldsymbol{Y})$. Consider $E[b(\boldsymbol{Y}, \boldsymbol{\theta}_0)]$, where the (outer) expectation is taken with respect to the sampling distribution. Then $E[b(\boldsymbol{Y}, \boldsymbol{\theta}_0)] = E[E[\boldsymbol{\theta} - \boldsymbol{\theta}_0]] = E[E[\boldsymbol{\theta}] - \boldsymbol{\theta}_0] = E[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0]$, which is the frequentist bias. In other words, the frequentist bias is equivalent to the average of $b(\boldsymbol{Y}, \boldsymbol{\theta}_0)$ over the sampling distribution, if the posterior mean is chosen as an estimator. In the glaciology test case, we have (informally) not noticed much variability in the posterior for ice viscosity over repeated sampling of the data, and hence, the distinction between Bayesian bias and frequentist bias is not significant.

The reader may wonder why a fixed $\boldsymbol{\theta}_0$ was assumed in the preceding paragraph, despite that a Bayesian model has been presented in this paper. In fact, it is typical to assume that the actual value of a parameter is fixed, despite ascribing a probability distribution to it in the form of a prior or posterior. Conceptually, such a probability distribution is a representation of a modeler's uncertainty regarding the fixed, unknown value of the parameter. For more on this interpretation of Bayesian statistics, the reader can consult results of statistical decision theory (e.g., on admissibility) in Lehmann and Casella (2003) and Robert (2007). This viewpoint is also taken in Bayesian asymptotic analysis, such as the Bernstein–von Mises theorem (van der Vaart 2000; Shen and Wasserman 2001).

4.4. INFERRING Σ

The covariance matrix Σ, first introduced after Eq. 7 in Sect. 3, determines the spatial correlation inherent in the error-correcting process, $\boldsymbol{X}$. Since spatial correlation in the error-correcting process is important to model (which is particularly evident in the glaciology example of Bueler et al. (2005)), we need to discuss how Σ ought to be specified. Choosing Σ can be difficult if no or little prior information is available, and in such a case, we suggest:

$$\Sigma = \text{diag}(\boldsymbol{v})\text{R}\,\text{diag}(\boldsymbol{v}),$$

where $\log(\boldsymbol{v}) \sim \text{MVN}(\boldsymbol{\mu}_v, \Sigma_v)$, Σ_v is derived from a GP kernel such as squared exponential or Matérn kernel, and R is a correlation matrix also derived from a GP kernel. To avoid non-identifiability and complexity of inference, it is suggested to pre-specify the parameters of these GP kernels. This approach is similar to the modeling strategy employed in Geirsson

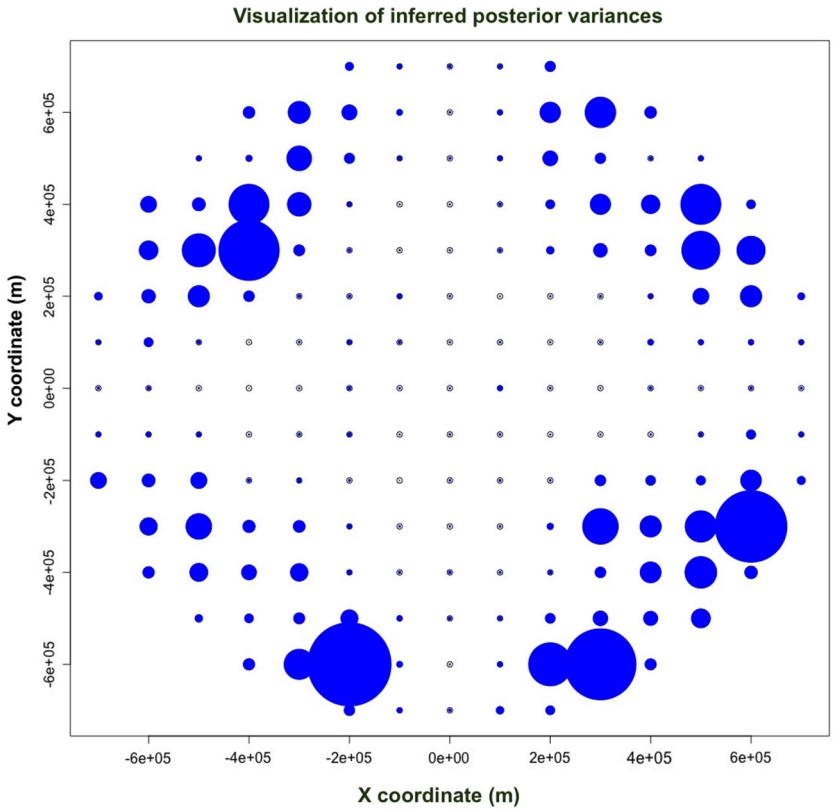

Figure 6. Inferred posterior variance field of the error-correcting process, where the area of each circle is proportional to the variance at the grid point centered at the circle. Qualitatively, this field behaves as one would expect from the work of Bueler et al. (2005), where the authors demonstrate that numerical inaccuracies for the SIA PDE are greatest toward the margin, but much smaller at the interior of the glacier..

et al. (2015). The intuition behind this approach is that the term $\boldsymbol{v}$ encodes spatial variability in the scale of deviations between the output of a computer simulator and the true physical process, and spatial correlation in these deviations is strongly enforced with non-diagonal terms in both Σ_v and R.

Figure 6 illustrates a map of the mean posterior field for the variances of the error-correcting process, where the area of each circle is proportional to the inferred posterior mean of variance; due to a multivariate normal prior on $\log(\boldsymbol{v})$, elliptical slice sampling is used as the method for posterior sampling (Murray et al. 2010). Consistent with Bueler et al. (2005), the variances tend to increase at the margins and are smaller at the interior. Additionally, the scaled differences between the analytical solution and numerical solver at the final time point the simulator is run (where scaling is inverse of the posterior mean of standard deviation) should theoretically approach a mean zero normal distribution according to the model. The p-value for an Anderson–Darling test is .436, suggesting that the scaled differences between the analytical solution and numerical solver are consistent with a normal distribution. Moreover, the sample mean for these scaled differences is .079 and the sample standard deviation is .409.

Table 2. Summary statistics of 10^6 posterior samples of the ice viscosity parameter under three versions of Σ.

Test case	Min	1st quartile	Median	Mean	3rd quartile	Max
Σ with GP field	10.0	21.0	36.0	35.7	50.5	70.0
Σ with strong prior information	18.0	25.0	26.5	26.6	28.0	35.5
Σ with weak prior information	12.5	18.5	19.5	19.5	20.5	26.5

While the weakly informative case leads to a very biased posterior, the biases for the ice viscosity posterior in the first two Σ matrices are comparable. Nonetheless, the posterior variance is much less in the case with prior information from Bueler et al. (2005).

As is discussed in the previous subsection, prior information for Σ has an effect on the inference of physical parameters (i.e., ice viscosity), and in particular, a lack of prior information can lead to a very biased posterior distribution for physical parameters. To compare the fitted Σ using a GP field against the Σ matrices discussed in the previous section, we show in Table 2 a comparison of posterior inference for the ice viscosity parameter for three choices of Σ. The first choice of Σ is the posterior mean of samples assuming the structure $\Sigma = \text{diag}(\boldsymbol{v})\text{Rdiag}(\boldsymbol{v})$, with $\log(\boldsymbol{v}) \sim \text{MVN}(\boldsymbol{\mu_v}, \Sigma_v)$. In the second and third scenarios, Σ is block diagonal with three variance parameters for each of the three blocks. A weakly informative case assumes that $\sigma^2_{\text{int}} = \sigma^2_{\text{dome}} = \sigma^2_{\text{margin}} = .1$, whereas a more informative case (using prior information from Bueler et al. (2005)) has $\sigma^2_{\text{int}} = \sigma^2_{\text{dome}} = .1$ and $\sigma^2_{\text{margin}} = 10$ (all units are m^2). The scenario for weak prior information for Σ results in a very biased posterior distribution whose support does not cover the actual parameter value (31.7×10^{-25} in units of $\text{s}^{-1}\,\text{Pa}^{-3}$)—the maximum in this case is 26.5×10^{-25} in units of $\text{s}^{-1}\,\text{Pa}^{-3}$. While the (absolute) biases of the posterior for ice viscosity for the GP field version compared to the prior information from Bueler et al. (2005) are comparable (5.09 versus 4.01 in units of $10^{-25}\,\text{s}^{-1}\,\text{Pa}^{-3}$), the posterior variance is markedly larger in the former case. This result suggests that prior knowledge from a domain expert is likely to be useful in determining Σ, though in a case when that does not exist, the methodology described in this section is an adequate alternative.

4.5. EXACT VERSUS APPROXIMATE LIKELIHOOD

In Sect. 3.1, we showed an exact way to calculate the model likelihood as well as an approximation in Sect. 3.2. In this subsection, our purpose is to compare these two methods of likelihood computation in terms of run-time and posterior inference. Using a MacBook Pro early 2015 model with a 2.7 GHz Intel Core i5 processor and 8 GB 1867 MHz DDR3 memory (as before), one component of the log-likelihood approximation (which can be computed in an embarrassingly parallel fashion with the other components of the sum) takes .0179 s, whereas the full log-likelihood calculation, as in Sect. 3.1, is .354 s (in both cases, using a first-order emulator). The results of comparing posterior samples for the ice viscosity parameter are given in Table 3—thus, while the mean, median, first, and third quartiles are comparable, the approximate version has larger posterior uncertainty than the exact version as is evidenced by the wider tails. These results suggest that, while there is

Table 3. Summary statistics of 10^6 posterior samples of the ice viscosity parameter using an exact likelihood and a likelihood approximation (units are in 10^{-25} s^{-1} Pa^{-3}).

Test case	Min	1st quartile	Median	Mean	3rd quartile	Max
Exact likelihood	15.0	26.5	27.0	27.4	29.0	39.5
Likelihood approximation	10.0	26.0	28.0	27.7	31.0	52.5

While the 1st quartile, median, mean, and 3rd quartile are similar, the tails in the approximation are much wider.

likely a computational speedup afforded by using the approximation (i.e., at least an order of magnitude), the price to pay is increased posterior uncertainty.

5. GENERALITY OF THE MODEL AND METHODOLOGY

Though we have tested the model and methodology in the previous sections in the context of a glaciology example, it should be noted that they can be used in other physical systems with similar components. In essence, this modeling and methodology can be applied in scenarios where:

1. A computer program (i.e., *computer simulator*) is available to simulate a continuous physical process through space and time, but there is a deviation between the output of the computer simulator and the actual physical process.
2. The deviations between the computer simulator output and the actual physical process values tend to grow with time and exhibit spatial correlation structure.
3. Measurements of the physical process are available, but they are potentially scarce in both space and time.
4. Physical parameters governing the physical process are uncertain but can be constrained with domain knowledge for the random walk error covariance (i.e., Σ).

Recall that at the process level, the model stipulates that:

$$\boldsymbol{S}_j = \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{\phi}, j) + \boldsymbol{X}_j. \tag{17}$$

To apply the same setup to another physical scenario, a different version of $\boldsymbol{f}(.,.,.)$, such as a numerical PDE solver for another system of spatiotemporal PDEs besides the SIA, can be used. However, while $\boldsymbol{f}(.,.,.)$ will need to be tailored to another physical scenario based on a different numerical scheme or physical model, the $\boldsymbol{X}_j$ term would be modeled in the same way (i.e., with a random walk).

6. CONCLUSION

The objective of this work has been to set forth a versatile physical statistical model in the Bayesian hierarchical framework that incorporates a computer simulator for a physical

process, such as a numerical solver for a system of PDEs. Posterior inference for physical parameters (and, consequently, posterior predictions of the physical process) can be computationally demanding within this model, since each evaluation of the likelihood requires a full PDE solve and computing the inverse and determinant of a large covariance matrix. Therefore, we have set forth two main ways to speed up the computation: First is the use of bandwidth limited linear algebra in a manner similar to Rue (2001) for quickly handling the covariance matrix in the likelihood and the second is the use of spatiotemporal emulation in a manner similar to Hooten et al. (2011) to emulate a PDE solver that is expensive to evaluate. An additional method for speeding up computation is to approximate the likelihood in a way that leads to embarrassingly parallel computation. The utility of this model and corresponding inference methodology is demonstrated with a test example from glaciology.

A unique feature of this work is how we represent the discrepancy between a computer simulator for a physical process and the real physical process values. One approach, as in Kennedy and O'Hagan (2001) and Brynjarsdóttir and O'Hagan (2014), is to assume that this is a fixed yet unknown function that can be learned with a GP (or constrained GP) prior distribution over a space of functions. Instead, we assume that this discrepancy is a spatiotemporal stochastic process (i.e., a random walk), which is motivated by the fact that a computer simulation is likely to become less accurate as it is run forward in time, as well as exhibit some degree of spatial correlation in inaccuracies. An interesting consequence of this modeling decision is that linear algebraic routines for band-limited matrices can be utilized for evaluating the likelihood of the model in an efficient manner. Another interesting artifact of this approach is that when prior information is used for the random walk's error term (i.e., in Σ), the bias for the posterior distribution of $\boldsymbol{\theta}$ is reduced. The same phenomenon is exhibited in the work of Brynjarsdóttir and O'Hagan (2014), where a constrained GP prior over a space of functions ends up reducing the bias of the physical parameter posterior distribution.

Despite that the model and methodology appear to perform well in the analysis of this paper, it is important to comment on some potential drawbacks of the approach, particularly when applied to other physical contexts. In this paper, emulation works adequately with a single parameter, though emulators do not always work well in other applications or higher-dimensional parameter spaces. For example, Salter et al. (2019) document some shortcomings of a principal component-based emulator in climate modeling. The second main computational advantages stem from log-likelihood evaluation speedups. The use of bandwidth limited matrix algebra for the exact log-likelihood can be used so long as the model holds, which may not always be the case (e.g., with a non-Gaussian data distribution). Additionally, the log-likelihood approximation holds when the measurement errors are small relative to the signal modeled, which depends on the measurement instruments used to collect the data. For instance, on common geophysical scales of thousands of meters, light detection and ranging (LIDAR) or digital GPS data have maximum errors on the order of a meter.

Additionally, if it is not possible to program the computer simulator to produce output at the data measurement locations, there are essentially two main ways to handle such a scenario. The first is to use spatial kriging to predict the value of the computer simulator at the spatial locations where data are collected, given the output of the computer simulator at the grid points. A simpler approach is to use inverse distance weighting of the simulator

output at the nearest neighbors; that is, take a weighted average of the four nearest grid points of the simulator, where the weights are proportional to the inverse of distance. Such an approach, for example, has been used in Geirsson et al. (2015).

Future research will include predicting Langjökull glacier surface elevation using the modeling and methodology within this paper, based on actual data collected by the UI-IES.

[Received December 2018. Revised June 2019. Accepted June 2019. Published Online June 2019.]

A DERIVATION OF THE EXACT LIKELIHOOD AND COMPUTATIONAL SIMPLIFICATIONS

As is shown in Appendix B of Gopalan et al. (2018), the covariance matrix for the observed data can be written as $\mathrm{U} \otimes \mathrm{V} + \sigma^2 \mathrm{I}$, where $\mathrm{U}_{ab} = k \min(a, b)$ with $\mathrm{U} \in \mathbb{R}^{N \times N}$, and $\mathrm{V} = \mathrm{A}(\Sigma)\mathrm{A}^\intercal$. It can be verified that U^{-1} is tridiagonal, so it has bandwidth one—more specifically:

$$\mathrm{U}^{-1} = k^{-1} \begin{bmatrix} 2 & -1 & 0 & \dots & & \\ -1 & 2 & -1 & 0 & \dots & \\ 0 & -1 & 2 & -1 & 0 & \dots \\ 0 & 0 & \ddots & \ddots & \ddots & \\ 0 & \dots & & -1 & 2 & -1 \\ 0 & \dots & & & -1 & 1 \end{bmatrix}. \tag{18}$$

One useful property of the Kronecker product is that $(\mathrm{U} \otimes \mathrm{V})^{-1} = \mathrm{U}^{-1} \otimes \mathrm{V}^{-1}$. Therefore:

$$(\mathrm{U} \otimes \mathrm{V})^{-1} = \mathrm{U}^{-1} \otimes \mathrm{V}^{-1} \tag{19}$$

$$= k^{-1} \begin{bmatrix} 2\mathrm{V}^{-1} & -\mathrm{V}^{-1} & 0 & \dots & & \\ -\mathrm{V}^{-1} & 2\mathrm{V}^{-1} & -\mathrm{V}^{-1} & 0 & \dots & \\ 0 & -\mathrm{V}^{-1} & 2V^{-1} & -\mathrm{V}^{-1} & 0 & \dots \\ 0 & 0 & \ddots & \ddots & \ddots & \\ 0 & \dots & & -\mathrm{V}^{-1} & 2\mathrm{V}^{-1} & -\mathrm{V}^{-1} \\ 0 & \dots & & & \mathrm{V}^{-1} & \mathrm{V}^{-1} \end{bmatrix}, \tag{20}$$

whose bandwidth is $O(m)$.

Let us denote $\mathrm{U} \otimes \mathrm{V}$ as W. By the matrix inversion lemma, it follows that $(\sigma^2\mathrm{I} + \mathrm{W})^{-1} = \sigma^{-2}\mathrm{I} - \sigma^{-2}(\mathrm{W}^{-1} + \sigma^{-2}\mathrm{I})^{-1}\mathrm{I}\sigma^{-2}$. The matrix $\mathrm{W}^{-1} + \sigma^{-2}\mathrm{I}$ has bandwidth $O(m)$ since W^{-1} has bandwidth $O(m)$ as shown previously, so this expression can be computed in $O(Nm^3)$ (Rue 2001; Golub and Van Loan 2012).

Similarly, by the matrix determinant lemma, $\log[\det(\sigma^2\mathrm{I} + \mathrm{W})]$ is $\log[\det(\mathrm{I} + \sigma^2\mathrm{W}^{-1})\det(\mathrm{W}^{-1})^{-1}] = \log[\det(\mathrm{I} + \sigma^2\mathrm{W}^{-1})] - \log[\det(\mathrm{W}^{-1})]$. Since both terms are log determinants of square matrices of dimension Nm and bandwidth $O(m)$, this can be calculated

in $O(Nm^3)$ due to the efficient Cholesky factorization of band-limited matrices (Rue 2001; Golub and Van Loan 2012).

B FIRST-ORDER SPATIOTEMPORAL EMULATORS

In the examples of this paper, the function $\boldsymbol{f}(.,.,.)$ (i.e., the computer simulator) can take one of the two forms: a numerical PDE solver for the SIA, or an emulator constructed from the numerical PDE solver for the SIA. The numerical method for solving the SIA PDE is as given in Gopalan et al. (2018), and the emulator is constructed based on the finite difference solver in a manner as suggested in Hooten et al. (2011), termed first-order emulation.

That is, we start with a set of plausible values for ice viscosity: $\{\theta_1, \theta_2, \ldots, \theta_p\}$ and, for each time point there is collected data ck, we store a matrix M_{ck}, where the qth column of matrix M_{ck} is the output of the numerical solver using parameter value θ_q after running for ck time steps forward. Thus, each matrix M_{ck} is of dimension n by p, and without essential loss of generality, we can assume that the number n is much larger than p, and each matrix M_{ck} is of rank p.

For each matrix, M_{ck}, we compute a singular value decomposition (SVD), $\mathrm{U}_{ck}\mathrm{D}_{ck}\mathrm{V}_{ck}^{\mathsf{T}}$. The goal is to find a (vector-valued) function $v_{ck}(\theta*)$ such that the emulated output at time ck for parameter value $\theta*$ is $\mathrm{U}_{ck}\mathrm{D}_{ck}v_{ck}(\theta*)$. To find the qth element of v_{ck}, we train a random forest (Breiman 2001; Liaw and Wiener 2002) with $(\theta_1, (V_{ck}^{\mathsf{T}})_{q1})$, $(\theta_2, (V_{ck}^{\mathsf{T}})_{q2}), \ldots, (\theta_p, (V_{ck}^{\mathsf{T}})_{qp})$ as training data, where $(V_{ck}^{\mathsf{T}})_{q1}$ is the first element of the qth right singular vector, $(V_{ck}^{\mathsf{T}})_{q2}$ is the second element of the qth right singular vector, and so on. Not all of the right singular vectors need be used in emulation, and a heuristic such as an elbow–scree plot or the randomization procedure of Friedman et al. (2001) can be used to determine the number of right singular vectors to keep. However, if the number of simulator runs (p) is much smaller than the dimensionality of the output (n), all of the right singular vectors can be utilized with computational savings, as is done in the experiments of this paper.

We have assumed the initial conditions and boundary conditions are known, since this is the case in the glaciology problems we have studied, where the boundary condition is that glacial thickness is nonnegative, and the initial glacier profile (i.e., a dome) is known. In general, however, $\boldsymbol{\phi}$ may be incorporated into the analysis above by considering θ and $\boldsymbol{\phi}$ jointly. Additionally, a variant is to directly emulate the likelihood function. However, since there is flexibility in the choice of Σ (which enters into the likelihood), unless one is set on using a particular value of Σ, it is sensible to emulate the numerical solver as opposed to retraining a likelihood emulator for each potential choice of Σ.

REFERENCES

Baum, L. E. and Petrie, T. (1966), "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *Annals of Mathematical Statistics*, 37, 1554–1563, https://doi.org/10.1214/aoms/1177699147.

Berliner, L. M. (1996), "Hierarchical Bayesian Time Series Models," in Hanson, K. M. and Silver, R. N. (editors), *Maximum Entropy and Bayesian Methods*, Dordrecht: Springer Netherlands.

— (2003), "Physical-statistical modeling in geophysics," *Journal of Geophysical Research: Atmospheres*, 108, n/a–n/a, https://doi.org/10.1029/2002JD002865. 8776.

Berrocal, V., Gelfand, A., and Holland, D. (2014), "Assessing exceedance of ozone standards: a space-time downscaler for fourth highest ozone concentrations," *Environmetrics*, 25, 279–291.

Björnsson, H. and Pálsson, F. (2008), "Icelandic glaciers," *Jökull*, 58, 365–386.

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32, https://doi.org/10.1023/A:1010933404324.

Brinkerhoff, D. J., Aschwanden, A., and Truffer, M. (2016), "Bayesian Inference of Subglacial Topography Using Mass Conservation," *Frontiers in Earth Science*, 4, 8, http://journal.frontiersin.org/article/10.3389/feart.2016.00008.

Brynjarsdóttir, J. and O'Hagan, A. (2014), "Learning about physical parameters: the importance of model discrepancy," *Inverse Problems*, 30, 114007, http://stacks.iop.org/0266-5611/30/i=11/a=114007.

Bueler, E., Lingle, C. S., Kallen-Brown, J. A., Covey, D. N., and Bowman, L. N. (2005), "Exact solutions and verification of numerical models for isothermal ice sheets," *Journal of Glaciology*, 51, 291–306.

Calderhead, B., Girolami, M., and Lawrence, N. D. (2008), "Accelerating Bayesian Inference over Nonlinear Differential Equations with Gaussian Processes," in *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'08, USA: Curran Associates Inc., http://dl.acm.org/citation.cfm?id=2981780.2981808.

Chkrebtii, O. A., Campbell, D. A., Calderhead, B., Girolami, M. A., et al. (2016), "Bayesian Solution Uncertainty Quantification for Differential Equations," *Bayesian Analysis*, 11, 1239–1267.

Conrad, P. R., Girolami, M., Särkkä, S., Stuart, A., and Zygalakis, K. (2017), "Statistical analysis of differential equations: introducing probability measures on numerical solutions," *Statistics and Computing*, 27, 1065–1082.

Cressie, N. and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, John Wiley & Sons.

Cuffey, K. M. and Paterson, W. (2010), *The Physics of Glaciers*, Academic Press, 4 edition.

Flowers, G. E., Marshall, S. J., Björnsson, H., and Clarke, G. K. (2005), "Sensitivity of Vatnajökull ice cap hydrology and dynamics to climate warming over the next 2 centuries," *Journal of Geophysical Research: Earth Surface*, 110.

Fowler, A. C. and Larson, D. A. (1978), "On the Flow of Polythermal Glaciers. I. Model and Preliminary Analysis," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 363, 217–242, http://www.jstor.org/stable/79748.

Friedman, J., Hastie, T., and Tibshirani, R. (2001), *The Elements of Statistical Learning*, volume 1, Springer series in statistics. New York, NY, NY, USA.

Geirsson, Ó. P., Hrafnkelsson, B., and Simpson, D. (2015), "Computationally efficient spatial modeling of annual maximum 24-h precipitation on a fine grid," *Environmetrics*, 26, 339–353, https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2343.

Golub, G. H. and Van Loan, C. F. (2012), *Matrix Computations*, volume 3, Johns Hopkins University Press.

Gopalan, G., Hrafnkelsson, B., Adalgeirsdóttir, G., Jarosch, A. H., and Pálsson, F. (2018), "A Bayesian hierarchical model for glacial dynamics based on the shallow ice approximation and its evaluation using analytical solutions," *The Cryosphere*, 12, 2229–2248.

Guan, Y., Haran, M., and Pollard, D. (2016), "Inferring Ice Thickness from a Glacier Dynamics Model and Multiple Surface Datasets," *ArXiv e-prints*.

Gupta, A. and Kumar, V. (1994), "A scalable parallel algorithm for sparse Cholesky factorization," in *Proceedings of the 1994 ACM/IEEE Conference on Supercomputing*, Supercomputing '94, Los Alamitos, CA, USA: IEEE Computer Society Press, http://dl.acm.org/citation.cfm?id=602770.602898.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008), "Computer Model Calibration Using High-Dimensional Output," *Journal of the American Statistical Association*, 103, 570–583.

Higdon, D., Kennedy, M., Cavendish, J. C., Cafeo, J. A., and Ryne, R. D. (2004), "Combining Field Data and Computer Simulations for Calibration and Prediction," *SIAM Journal on Scientific Computing*, 26, 448–466.

Hooten, M. B., Leeds, W. B., Fiechter, J., and Wikle, C. K. (2011), "Assessing First-Order Emulator Inference for Physical Parameters in Nonlinear Mechanistic Models," *Journal of Agricultural, Biological, and Environmental Statistics*, 16, 475–494, https://doi.org/10.1007/s13253-011-0073-7.

Hutter, K. (1982), "A mathematical model of polythermal glaciers and ice sheets," *Geophysical & Astrophysical Fluid Dynamics*, 21, 201–224, https://doi.org/10.1080/03091928208209013.

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

Kusnierczyk, W. (2012), *rbenchmark: Benchmarking routine for R*, https://CRAN.R-project.org/package=rbenchmark. R package version 1.0.0.

Lehmann, E. and Casella, G. (2003), *Theory of Point Estimation*, Springer Texts in Statistics, Springer New York, https://books.google.com/books?id=0q-Bt0Ar-sgC.

Liaw, A. and Wiener, M. (2002), "Classification and Regression by randomForest," *R News*, 2, 18–22, https://CRAN.R-project.org/doc/Rnews/.

Lindgren, F., Rue, H., and Lindström, J. (2011), "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 423–498.

Liu, F. and West, M. (2009), "A dynamic modelling strategy for Bayesian computer model emulation," *Bayesian Analysis*, 4, 393–411, https://doi.org/10.1214/09-BA415.

Madsen, H. (2007), *Time Series Analysis*, Chapman and Hall/CRC.

Murray, I., Adams, R. P., and MacKay, D. J. (2010), "Elliptical slice sampling", *Journal of Machine Learning Research W&CP*, 9, 541–548.

Owhadi, H. and Scovel, C. (2017), "Universal Scalable Robust Solvers from Computational Information Games and fast eigenspace adapted Multiresolution Analysis", ArXiv e-prints.

Pagendam, D., Kuhnert, P., Leeds, W., Wikle, C., Bartley, R., and Peterson, E. (2014), "Assimilating catchment processes with monitoring data to estimate sediment loads to the Great Barrier Reef," *Environmetrics*, 25, 214–229.

Payne, A. J., Huybrechts, P., Abe-Ouchi, A., Calov, R., Fastook, J. L., Greve, R., Marshall, S. J., Marsiat, I., Ritz, C., Tarasov, L., and Thomassen, M. P. A. (2000), "Results from the EISMINT model intercomparison: the effects of thermomechanical coupling," *Journal of Glaciology*, 46, 227–238.

Robert, C. (2007), *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Springer Texts in Statistics, Springer New York, https://books.google.com/books?id=NQ5KAAAAQBAJ.

Rue, H. (2001), "Fast sampling of Gaussian Markov random fields," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 325–338.

Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, CRC press.

Salter, J. M., Williamson, D. B., Scinocca, J., and Kharin, V. (2019), "Uncertainty Quantification for Computer Models With Spatial Output Using Calibration-Optimal Bases," *Journal of the American Statistical Association*, 0, 1–24, https://doi.org/10.1080/01621459.2018.1514306.

Shen, X. and Wasserman, L. (2001), "Rates of convergence of posterior distributions," *Annals of Statistics*, 29, 687–714, https://doi.org/10.1214/aos/1009210686.

Sigurdarson, A. N. and Hrafnkelsson, B. (2016), "Bayesian prediction of monthly precipitation on a fine grid using covariates based on a regional meteorological model," *Environmetrics*, 27, 27–41, https://ideas.repec.org/a/wly/envmet/v27y2016i1p27-41.html.

Solin, A. and Särkkä, S. (2014), "Explicit Link Between Periodic Covariance Functions and State Space Models," in Kaski, S. and Corander, J. (editors), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, Reykjavik, Iceland: PMLR, http://proceedings.mlr.press/v33/solin14.html.

van der Vaart, A. (2000), *Asymptotic Statistics*, Asymptotic Statistics, Cambridge University Press, https://books.google.com/books?id=UEuQEM5RjWgC.

van der Veen, C. (2013), *Fundamentals of Glacier Dynamics*, CRC Press, 2 edition.

Whittle, P. (1954), "ON STATIONARY PROCESSES IN THE PLANE," *Biometrika*, 434–449.

— (1963), "Stochastic processes in several dimensions," *Bulletin of the International Statistical Institute*, 40, 974–994.

Wikle, C. K. (2016), *Hierarchical Models for Uncertainty Quantification: An Overview*, Springer International Publishing, 1–26.

Wikle, C. K., Berliner, L. M., and Cressie, N. (1998), "Hierarchical Bayesian space-time models," *Environmental and Ecological Statistics*, 5, 117–154, https://doi.org/10.1023/A:1009662704779.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. (2001), "Spatiotemporal Hierarchical Bayesian Modeling Tropical Ocean Surface Winds," *Journal of the American Statistical Association*, 96, 382–397.

Zammit-Mangion, A., Rougier, J., Bamber, J., and Schön, N. (2014), "Resolving the Antarctic contribution to sea-level rise: a hierarchical modelling framework," *Environmetrics*, 25, 245–264.