A Hierarchical Spatio-Temporal Statistical Model for Physical Systems

Giri Gopalan¹, Birgir Hrafnkelsson¹, Christopher K. Wikle², Håvard Rue³, Guðfinna Aðalgeirsdóttir¹, Alexander H. Jarosch⁴ and Finnur Pálsson¹

¹University of Iceland. Reykjavik, Iceland. 101.
²University of Missouri. Columbia, Missouri. 65211.
³King Abdullah University of Science and Technology. Thuwal, Saudi Arabia. 23955.
⁴University of Innsbruck. Innsbruck, Austria. 6020.

Abstract

In this paper, we extend and analyze a Bayesian hierarchical spatio-temporal 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.

Keywords: model discrepancy, uncertainty quantification, emulation.
1 Introduction

Scientists and engineers are often immersed in the following scenario: they would like to study a physical system to make spatio-temporal predictions (e.g., temperature or glacier thickness) or infer unknown quantities governing the system (e.g., atmospheric density or ice viscosity). This system’s dynamics can often be phrased in terms of spatio-temporal partial differential equations (PDEs). Because such PDEs can be based on approximations, they are often imperfect. The scientist or engineer may also be able to simulate the physical system with a computer simulator or program, such as a numerical PDE solver, which is subject to imperfections (e.g., numerical error). Moreover, the fundamental 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.

Such scenarios can be modeled with a variant of a Bayesian hierarchical spatio-temporal 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.

1.1 Literature review

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 Hafnkelsson 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.
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 components 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 in order 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 $S$ can be indexed through time, i.e., as $S_j$, and $S_j$ is a vector where each element corresponds to a distinct spatial location. One can specify the process level conditional on physical parameter $\theta$ as

$$S_j = f(\theta, j) + \delta(j)$$

(1)

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

$$Y_j = f(\theta, j) + \delta(j) + \eta_j,$$

(2)

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

In Kennedy and O’Hagan (2001), $\delta(.)$ is a fixed but unknown function independent of $\theta$ that is endowed
with a GP prior distribution. Similarly, $\delta(.)$ 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. 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 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 speed ups; 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 spatio-temporal 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. For instance, a fractional Laplacian operator yields the Matérn covariance function (Whittle 1954, Whittle 1963, and 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 nonlinearities), though we highlight the utility of the latter approach in situations where an analytical covariance function can be derived exactly.

1.2 Contributions

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
for speeding up the computation of a numerical solver, and the second is the use of bandwidth limited numerical linear algebra \cite{Rue2001} 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.

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.

\cite{Bueler2005} introduce analytical solutions for the SIA PDE, a commonly used model for the dynamics of glaciers (\cite{Fowler1978}, \cite{Hutter1982}, \cite{Flowers2005}, \cite{Cuffey2010}, \cite{van_der_Veen2017}, \cite{Brinkerhoff2016}, \cite{Guan2016}, and \cite{Gopalan2018}). 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).

As shown in \cite{Gopalan2018}, the SIA PDE, written in terms of the thickness of the glacier, \(H(x, y, t)\), is

\[
H_t = -[\bar{u}H]_x - [\bar{v}H]_y + b, \tag{3}
\]
\[
-\bar{u}H_x = -[-C_0\gamma(-\rho g[H + R]_x)H + \frac{2B}{n + 2}(\rho g \alpha)^{n-1}H^{n+1}(-\rho g[H + R]_x)_x], \tag{4}
\]
\[
-\bar{v}H_y = -[-C_0\gamma(-\rho g[H + R]_y)H + \frac{2B}{n + 2}(\rho g \alpha)^{n-1}H^{n+1}(-\rho g[H + R]_y)_y], \tag{5}
\]
\[
\alpha = \sqrt{[H + R]_x^2 + [H + R]_y^2}. \tag{6}
\]

While an explanation of all of the terms in this PDE is given in \cite{Gopalan2018}, 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 and time (in polar coordinates), is

\[
H(r,t) = H_s(r) + P(r,t),
\]

(7)

\[
P(r,t) = C_p \sin(2\pi t/T_p) \cos^2\left(\frac{\pi(r - 0.6L)}{6L}\right) \text{; if } 0.3L < r < 0.9L,
\]

(8)

\[
P(r,t) = 0; \text{if } 0 \leq r \leq 0.3L \text{ or if } r \geq 0.9L.
\]

(9)

Here, \( H_s \) is a static initial profile of the glacier (i.e., a dome as in Eq. 21 of \( \text{Bueller et al. (2005)} \)), \( L \) is the margin length, \( C_p \) is the magnitude of the periodic perturbation, and \( T_p \) is the period of the perturbation. \( \text{Bueller et al. (2005)} \) 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 \( 0.3L < r < 0.9L \). In Figure 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 meter, larger than errors of the digital-GPS instruments employed by the UI-IES. We use the same values of parameters as in \( \text{Bueller et al. (2005)} \) to make for easier comparison to 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 set up as \( \text{Gopalan et al. (2018)} \), glacial surface elevation measurements are assumed to be collected for 20 years, twice a year, and at 25 fixed spatial locations across the glacier, in order 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 Figure 2 illustrating Langjökull and the mass balance measurement sites).

### 3 The hierarchical spatio-temporal 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 spatio-temporal 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, similarly to the set up in \( \text{Gopalan et al. (2018)} \). We assume that \( n \) spatial locations are modeled at the latent level, and \( m \) of those locations are observed, where \( m \) is typically much smaller than \( n \). We use the index \( j \) to refer to time indices and \( i \) to refer to spatial indices; while space
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 \(0.3L < r < 0.9L\), 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 meters above sea level (Björnsson and Pálsson 2008), so Langjökull is shallow.

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.

We introduce the Bayesian hierarchical model in the parameter, process, data level framework of Cressie and Wikle (2011). We denote the physical parameters as \(\theta\) and initial and/or boundary conditions for the physical process as \(\phi\). At the parameter level, one possibility is to use a truncated normal distribution for \(\theta\) if the support of the parameter value can be constrained, as was done in Gopalan et al. (2018), where \(\theta\) 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 $f(\theta, \phi, j)$, which is an element of $\mathbb{R}^n$. For a specific example, in Appendix A of Gopalan et al. (2018), $f(\theta, \phi, j)$ is a second-order finite difference solver for the SIA PDE.

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

\begin{align*}
X_j &= X_{j-1} + \epsilon_j, \\
S_j &= f(\theta, \phi, j) + X_j,
\end{align*}

where $X_0$ is a vector of zeros.

In the above expressions, $\epsilon_j$ is $MVN(0, \Sigma)$ and independent of $\epsilon_k$ for $k \neq j$. Furthermore, $X_j$, $\epsilon_j$, $f(\theta, \phi, j)$, and (consequently) $S_j$ are members of $\mathbb{R}^n$. In Gopalan et al. (2018), $\{X_1, X_2, \ldots, \}$ 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), $S_j$ 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:

\begin{equation}
X_j + \sum_{k=1}^{q} (-1)^k \binom{q}{k} X_{j-k} = \epsilon_j
\end{equation}

where $\epsilon_1, \ldots, \epsilon_q$ 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 $k$-th time point, so that one observes $Y_k, Y_{2k}, \ldots, Y_{Nk} \in \mathbb{R}^m$; in Gopalan et al. (2018), the variables $Y$ referred to glacial surface elevation measurements. The corresponding observation errors $\eta_k, \eta_{2k}, \ldots, \eta_{Nk}$ are i.i.d. $MVN(0, \sigma^2 I)$. If we define the matrix $A \in \mathbb{R}^{m \times n}$ to be such that its rows are unit basis vectors (i.e., an incidence matrix as in Cressie and Wikle (2011)), where nonzero elements correspond to the spatial indices that have been observed, then the data level model, conditional on the process $S$, is

\begin{equation}
Y_{ck} = AS_{ck} + \eta_{ck},
\end{equation}
where we assume that $j \in \{1, 2, ..., T\}$ and $c \in \{1, 2, ..., N\}$, so there are $N$ total observed spatial vectors, observed with a period of length $k$.

Conditional on $\theta$, $\phi$, and a computer simulator, the model can be thought of as a hidden Markov model (HMM); 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 $\theta$, $\phi$, and a computer simulator, one can write:

$$S_j = S_{j-1} + [-f(\theta, \phi, j-1) + f(\theta, \phi, j)] + \epsilon_j,$$

$$Y_{ck} = A S_{ck} + \eta_{ck}.$$  \hfill (14)  

Here, the state evolves linearly with a time dependent offset term: $[-f(\theta, \phi, j-1) + f(\theta, \phi, j)]$. The notation $ck$ is used in Eq. 15 to indicate that observations of the process are only observed every $k$th time point, whereas the latent process evolves at every time step $j$.

### 3.1 Exact likelihood

An advantage of using this model is that the likelihood, $p(Y_{k}, Y_{2k}, ..., Y_{Nk}|\theta, \phi)$, 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.

As was shown in Gopalan et al. (2018) and is elaborated more upon in Appendix A, the likelihood of the model, $L(\theta, \phi) = p(Y_{k}, Y_{2k}, ..., Y_{Nk}|\theta, \phi)$, has a multivariate normal p.d.f form:

$$L(\theta, \phi) = \frac{1}{(2\pi)^{mN/2}|\Sigma_l|^{1/2}} \exp\left[-(Y-\mu_l)^T \Sigma_l^{-1} (Y-\mu_l)/2\right],$$  \hfill (16)  

where the mean is:

$$\mu_l = (Af(\theta, \phi, k), ..., Af(\theta, \phi, Nk)),$$  \hfill (17)  

and the covariance matrix is:

$$\Sigma_l = U \otimes V + \sigma^2 I,$$  \hfill (18)  

where $U_{ij} = k \min(i, j)$ with $U \in \mathbb{R}^{N \times N}$, and $V = A \Sigma A^T$. Also, the symbol $\otimes$ stands for the Kronecker product.
Since evaluating Eq. 16 requires the calculation of the inverse of the matrix $\Sigma_l$ 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 $\Sigma_l^{-1}$ and $|\Sigma_l|$ in $O(N m^3)$ time \cite{Rue2001, Golub2012}. 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 $p(Y_k, ..., Y_N|\theta, \phi)$ can be equivalently written as $p(Y_k|\theta, \phi)p(Y_{2k}|Y_k, \theta, \phi)\ldots p(Y_{Nk}|Y_k, ..., Y_{(N-1)k}, \theta, \phi)$. First note that:

$$Y_k = Af(\theta, \phi, k) + \eta_k + \sum_{j=1}^{k} A\epsilon_j. \quad (19)$$

Hence, $p(Y_k|\theta, \phi)$ is multivariate normal with mean $Af(\theta, \phi, k)$ and covariance matrix $A(k \Sigma) A^\top + \sigma^2 I$. More generally, we have the relationship:

$$Y_{ck} = Y_{(c-1)k} + A[f(\theta, \phi, ck) - f(\theta, \phi, (c-1)k)] + \eta_{ck} - \eta_{(c-1)k} + \sum_{j=(c-1)k+1}^{ck} A\epsilon_j. \quad (20)$$

Thus we can approximate $p(Y_{ck}|Y_k, ..., Y_{(c-1)k}, \theta, \phi)$ as a MVN distribution with mean $Y_{(c-1)k} + A[f(\theta, \phi, ck) - f(\theta, \phi, (c-1)k)]$ and covariance matrix $A(k \Sigma) A^\top + 2\sigma^2 I$. Nonetheless, to clarify, $p(Y_{ck}|Y_k, ..., Y_{(c-1)k}, \theta, \phi)$ is not exactly a MVN with mean $Y_{(c-1)k} + A[f(\theta, \phi, ck) - f(\theta, \phi, (c-1)k)]$ and covariance matrix $A(k \Sigma) A^\top + 2\sigma^2 I$ because $Y_{(c-1)k}$ and $\eta_{(c-1)k}$ are dependent. However, when the magnitude of the observation error $\eta_{(c-1)k}$ is much smaller in comparison to the magnitude of the observation $Y_{(c-1)k}$ (i.e., in a relatively high signal-to-noise regime), this dependence will intuitively not have much of an effect.

This approximation is embarrassingly parallel because each of the $N$ terms in the product form of the likelihood $p(Y_k, ..., Y_T|\theta, \phi) = p(Y_k|\theta, \phi)p(Y_{2k}|Y_k, \theta, \phi)\ldots p(Y_{Nk}|Y_k, ..., Y_{(N-1)k}, \theta, \phi)$ (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 p.d.f of dimension $m$ – this can be done in $O(m^3)$. 

10
3.3 Computational complexity summary

If no attention is paid to the structure of $\Sigma_l$, the cost of evaluating $L(\theta, \phi)$ is limited by the evaluation of $\Sigma_l^{-1}$ 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 p.d.f 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 Section 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, both in terms of computational run time and quality of inferences. In particular, we compare an emulator based posterior inference to finite difference method based posterior inference, 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): $31.7 \times 10^{-25}$ in units of $s^{-1}Pa^{-3}$.

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).

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 spatio-temporal emulator (i.e., an emulator based on the method in Appendix B) compares to a numerical solver of the PDE, both in terms of run-time of computations and posterior inference of ice viscosity. To do this, we train an emulator using the following values for ice viscosity: $\{10, 12.5, 15.0, ..., 70.0\}$ 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 \texttt{rbenchmark} \cite{Kusnierczyk2012} 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 spatio-temporal 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 seconds for the emulator based log-likelihood versus 5.148 seconds 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, implying that conclusions derived from either posterior are not going to differ very much.

| 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 |

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

### 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 spatio-temporal 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 Section 3.1, the likelihood involves band-limited matrices, for which there exist specialized numerical linear algebra routines. However, there is a trade-off in 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 $\Sigma$.

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 in filled in black are close to the top (also referred to as the dome) of the glacier. Recall from Figure 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. 7-9. 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 if 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. 12, which should theoretically be distributed like \( \epsilon_j \) (i.e., independent \( MVN(0, \Sigma) \) random variables). To compute \( X_j \), we take the difference \( S_j - f(\theta,j) \), where \( S_j \) 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 $f(\theta, j)$ 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 Figures 4 and 5.

Figure 4: This figure displays residuals in units of meters (i.e., the term $\epsilon_j$ in Eq. 12) 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 $\epsilon_j$ in Eq. 12) 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.

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 m to .05 m (more than 1000 times reduction). Additionally, for both the interior and margin 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 $\theta$

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 $\Sigma$ to be such that it is block diagonal with 3 blocks, $\Sigma_{\text{int}}$, $\Sigma_{\text{dome}}$, and $\Sigma_{\text{margin}}$. Each of these blocks is derived from a square exponential covariance kernel with the same length-scale parameter $\phi = 70$ km, but differing variance parameters $\sigma^2_{\text{int}}$, $\sigma^2_{\text{dome}}$, and $\sigma^2_{\text{margin}}$. If we ignore prior information from [Bueler et al. (2005)], we assume that there is an equal prior probability that each of $\sigma^2_{\text{int}}$, $\sigma^2_{\text{dome}}$, and $\sigma^2_{\text{margin}}$ 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 $\sigma^2_{\text{int}}$, $\{1, 10\}$ for $\sigma^2_{\text{dome}}$, and $\{10, 100\}$ for $\sigma^2_{\text{margin}}$ (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.

4.4 Inferring $\Sigma$

Specifying $\Sigma$ can be difficult if no or little prior information is available, and in such a case, we suggest:

$$\Sigma = \text{diag}(v)R \text{diag}(v),$$
where \( \log(v) \sim MVN(\mu_v, \Sigma_v) \), \( \Sigma_v \) 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 et al. (2015). The intuition behind this approach is that the term \( v \) 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 \( \Sigma_v \) 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 \( \log(v) \), 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.

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 the work of Bueler et al. (2005), where the authors demonstrate that numerical inaccuracies for the SIA PDE are greatest towards the margin, but much smaller at the interior of the glacier.
As is discussed in the previous subsection, prior information for $\Sigma$ 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 $\Sigma$ using a GP field against the $\Sigma$ 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 $\Sigma$. The first choice of $\Sigma$ is the posterior mean of samples assuming the structure $\Sigma = \text{diag}(v)R\text{diag}(v)$, with $\log(v) \sim \text{MVN}(\mu_v, \Sigma_v)$. In the second and third scenarios, $\Sigma$ 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 $\Sigma$ results in a very biased posterior distribution whose support does not cover the actual parameter value ($31.7 \times 10^{-25}$ in units of $s^{-1}Pa^{-3}$) – the maximum in this case is $26.5 \times 10^{-25}$ in units of $s^{-1}Pa^{-3}$. 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 $10^{-25}s^{-1}Pa^{-3}$), 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 $\Sigma$, though in a case when that does not exist, the methodology described in this section is an adequate alternative.

| Test Case                      | Min   | 1st Quartile | Median | Mean  | 3rd Quartile | Max   |
|-------------------------------|-------|--------------|--------|-------|--------------|-------|
| $\Sigma$ with GP field       | 10.0  | 21.0         | 36.0   | 35.7  | 50.5         | 70.0  |
| $\Sigma$ with strong prior information | 18.0  | 25.0         | 26.5   | 26.6  | 28.0         | 35.5  |
| $\Sigma$ with weak prior information | 12.5  | 18.5         | 19.5   | 19.5  | 20.5         | 26.5  |

Table 2: Summary statistics of $10^6$ posterior samples of the ice viscosity parameter under three versions of $\Sigma$. While the weakly-informative case leads to a very biased posterior, the biases for the ice viscosity posterior in the first two $\Sigma$ matrices are comparable. Nonetheless, the posterior variance is much less in the case with prior information from [Bueler et al. (2005)].

4.5 Exact versus approximate likelihood

In Section 3.1, we showed an exact way to calculate the model likelihood as well as an approximation in Section 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 Section 3.1, is .354 seconds (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 likely a computational speed-up afforded by using the approximation (i.e., at least an order of magnitude), the price to pay is increased posterior uncertainty.

| 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 |

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 $10^{-25} \text{s}^{-1} \text{Pa}^{-3}$). While the 1st quartile, median, mean, and 3rd quartile are similar, the tails in the approximation are much wider.

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 both in space and time.

4. Physical parameters governing the physical process are uncertain but can be constrained with domain knowledge.

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

$$S_j = f(\theta, \phi, j) + X_j.$$ (21)
To apply the same setup to another physical scenario, a different version of \( f(\cdot, \cdot, \cdot) \), such as a numerical PDE solver for another system of spatio-temporal PDEs besides the SIA, can be used. However, while \( f(\cdot, \cdot, \cdot) \) will need to be tailored to another physical scenario based on a different numerical scheme or physical model, the \( X_j \) 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 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 spatio-temporal 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 spatio-temporal 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 further and further 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 \( \Sigma \)), the bias for the posterior distribution of \( \theta \) 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.

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. This is part of a Ph.D.
project entitled Statistical Models for Glaciology, supported by RANNIS, grant number 152457-051.

7 Acknowledgment

In addition to input from all authors of this paper, we appreciate discussions with Dr. Khachik Sargsyan and members of the Statistical Science Department at the University of Exeter.

A Derivation of the exact likelihood and computational simplifications

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

\[
U^{-1} = k^{-1} \begin{bmatrix}
2 & -1 & 0 & \ldots \\
-1 & 2 & -1 & 0 & \ldots \\
0 & -1 & 2 & -1 & 0 & \ldots \\
0 & 0 & \ddots & \ddots & \ddots \\
0 & \ldots & -1 & 2 & -1 \\
0 & \ldots & -1 & 1
\end{bmatrix}.
\] (22)

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

\[
(U \otimes V)^{-1} = U^{-1} \otimes V^{-1} = k^{-1} \begin{bmatrix}
2V^{-1} & -V^{-1} & 0 & \ldots \\
-V^{-1} & 2V^{-1} & -V^{-1} & 0 & \ldots \\
0 & -V^{-1} & 2V^{-1} & -V^{-1} & 0 & \ldots \\
0 & 0 & \ddots & \ddots & \ddots \\
0 & \ldots & -V^{-1} & 2V^{-1} & -V^{-1} \\
0 & \ldots & -V^{-1} & V^{-1}
\end{bmatrix},
\] (23)

whose bandwidth is \(O(m)\).

Let us denote \( U \otimes V \) as \( W \). By the matrix inversion lemma, it follows that \((\sigma^2 I + W)^{-1} = \sigma^{-2} I - \)

\[
\begin{bmatrix}
2V^{-1} & -V^{-1} & 0 & \ldots \\
-V^{-1} & 2V^{-1} & -V^{-1} & 0 & \ldots \\
0 & -V^{-1} & 2V^{-1} & -V^{-1} & 0 & \ldots \\
0 & 0 & \ddots & \ddots & \ddots \\
0 & \ldots & -V^{-1} & 2V^{-1} & -V^{-1} \\
0 & \ldots & -V^{-1} & V^{-1}
\end{bmatrix} \]

\[
U = k^{-1} \begin{bmatrix}
2 & -1 & 0 & \ldots \\
-1 & 2 & -1 & 0 & \ldots \\
0 & -1 & 2 & -1 & 0 & \ldots \\
0 & 0 & \ddots & \ddots & \ddots \\
0 & \ldots & -1 & 2 & -1 \\
0 & \ldots & -1 & 1
\end{bmatrix}.
\]
\[ \sigma^{-2}(W^{-1} + \sigma^{-2}I)^{-1}I\sigma^{-2}. \] The matrix \( W^{-1} + \sigma^{-2}I \) has bandwidth \( O(m) \) since \( W^{-1} \) has bandwidth \( O(m) \) as shown previously, so this expression can be computed in \( O(Nm^3) \) \cite{Rue2001,GolubVanLoan2012}.

Similarly, by the matrix determinant lemma, \( \log[\det(\sigma^2I + W)] \) is \( \log[\det(I + \sigma^2W^{-1})]\log[\det(W^{-1})] \). 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 \cite{Rue2001,GolubVanLoan2012}.

## B First-order spatio-temporal emulators

In the examples of this paper, the function \( f(.,.,.,.) \) (i.e., the computer simulator) can take one of 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 \cite{GopalanEtAl2018}, and the emulator is constructed based on the finite difference solver in a manner as suggested in \cite{HootenEtAl2011}, termed first-order emulation.

That is, we start with a set of plausible values for ice viscosity: \( \{\theta_1, \theta_2, ..., \theta_p\} \) and, for each time point there is collected data \( ck \), we store a matrix \( M_{ck} \), where the \( q \)-th column of matrix \( M_{ck} \) is the output of the numerical solver using parameter value \( \theta_q \) 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), \( U_{ck}D_{ck}V_{ck}^\top \). 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 \( U_{ck}D_{ck}v_{ck}(\theta^*) \). To find the \( q \)-th element of \( v_{ck} \), we train a random forest \cite{Breiman2001,LiawWiener2002} with \( (\theta_1, (V_{ck}^\top)_{q1}), (\theta_2, (V_{ck}^\top)_{q2}), ..., (\theta_p, (V_{ck}^\top)_{qp}) \) as training data, where \( (V_{ck}^\top)_{q1} \) is the first element of the \( q \)-th right singular vector, \( (V_{ck}^\top)_{q2} \) is the second element of the \( q \)-th 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 \cite{FriedmanEtAl2001} 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 non-negative, and the initial glacier profile (i.e., a dome) is known. In general, however, \( \phi \) may be incorporated into the
analysis above by considering \( \theta \) and \( \phi \) jointly. Additionally, a variant is to directly emulate the likelihood function. However, since there is flexibility in the choice of \( \Sigma \) (which enters into the likelihood), unless one is set on using a particular value of \( \Sigma \), it is sensible to emulate the numerical solver as opposed to retraining a likelihood emulator for each potential choice of \( \Sigma \).

References

Berliner, L. M. (1996), “Hierarchical Bayesian time series models,” in Hanson, K. and Silver, R. (editors), *Maximum entropy and Bayesian methods*, Springer, 15–22.

— (2003), “Physical-statistical modeling in geophysics,” *Journal of Geophysical Research: Atmospheres*, 108, n/a–n/a, URL http://dx.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,” *Mach. Learn.*, 45, 5–32, URL 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, URL 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, URL 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., URL 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, URL 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, 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, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2343.

Golub, G. H. and Van Loan, C. F. (2012), *Matrix computations*, volume 3, JHU 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, URL 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, URL [https://doi.org/10.1007/s13253-011-0073-7](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, URL [https://doi.org/10.1080/0309192820820913](https://doi.org/10.1080/0309192820820913).

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`, URL [https://CRAN.R-project.org/package=rbenchmark](https://CRAN.R-project.org/package=rbenchmark). R package version 1.0.0.

Liaw, A. and Wiener, M. (2002), “Classification and Regression by randomForest,” *R News*, 2, 18–22, URL [https://CRAN.R-project.org/doc/Rnews/](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 Anal.*, 4, 393–411, URL [https://doi.org/10.1214/09-BA415](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.

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.

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, URL https://ideas.repec.org/a/wly/envmet/v27y2016i1p27-41.html.

van der Veen, C. (2017), *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, URL 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.