A COMPUTATIONAL FRAMEWORK FOR INFINITE-DIMENSIONAL BAYESIAN INVERSE PROBLEMS, PART II: STOCHASTIC NEWTON MCMC WITH APPLICATION TO ICE SHEET FLOW INVERSE PROBLEMS*

NOEMI PETRA†, JAMES MARTIN†, GEORG STADLER†, AND OMAR GHATTAS‡

Abstract. We address the numerical solution of infinite-dimensional inverse problems in the framework of Bayesian inference. In Part I of this paper [T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler, *SIAM J. Sci. Comput.*, 35 (2013), pp. A2494–A2523] we considered the linearized infinite-dimensional inverse problem. In Part II, we relax the linearization assumption and consider the fully nonlinear infinite-dimensional inverse problem using a Markov chain Monte Carlo (MCMC) sampling method. To address the challenges of sampling high-dimensional probability density functions (pdfs) arising upon discretization of Bayesian inverse problems governed by PDEs, we build upon the stochastic Newton MCMC method. This method exploits problem structure by taking as a proposal density a local Gaussian approximation of the posterior pdf, whose covariance operator is given by the inverse of the local Hessian of the negative log posterior pdf. The construction of the covariance is made tractable by invoking a low-rank approximation of the data misfit component of the Hessian. Here we introduce an approximation of the stochastic Newton proposal in which we compute the low-rank-based Hessian at just the maximum a posteriori (MAP) point, and then reuse this Hessian at each MCMC step. We compare the performance of the proposed method to the original stochastic Newton MCMC method and to an independence sampler. The comparison of the three methods is conducted on a synthetic ice sheet inverse problem. For this problem, the stochastic Newton MCMC method with a MAP-based Hessian converges at least as rapidly as the original stochastic Newton MCMC method, but is far cheaper since it avoids recomputing the Hessian at each step. On the other hand, it is more expensive per sample than the independence sampler; however, its convergence is significantly more rapid, and thus overall it is much cheaper. Finally, we present extensive analysis and interpretation of the posterior distribution and classify directions in parameter space based on the extent to which they are informed by the prior or the observations.

Key words. Bayesian inference, infinite-dimensional inverse problems, uncertainty quantification, MCMC, stochastic Newton, low-rank approximation, ice sheet dynamics

AMS subject classifications. 35Q62, 62F15, 35R30, 35Q93, 65C40, 65C60, 49M15, 86A40

DOI. 10.1137/130934805

1. Introduction and background. We consider the problem of estimating the uncertainty in the solution of infinite-dimensional inverse problems within the framework of Bayesian inference [33, 53, 55]. Namely, given observational data and their uncertainties, a (possibly stochastic) forward model that maps model parameters to observations, and a prior probability distribution on model parameters that encodes any prior knowledge or assumptions about the parameters, find the *posterior probability distribution* of the parameters conditioned on the observational data. This

†Institute for Computational Engineering & Sciences, The University of Texas at Austin, Austin, TX 78712 (noemi@ices.utexas.edu, jmartin@ices.utexas.edu, georgest@ices.utexas.edu).

‡Institute for Computational Engineering & Sciences, Department of Mechanical Engineering, and Department of Geological Sciences, The University of Texas at Austin, Austin, TX 78712 (omar@ices.utexas.edu).

probability density function (pdf) is defined as the Bayesian solution of the inverse problem. The posterior distribution assigns to any candidate set of parameter fields our belief (expressed as a probability) that a member of this candidate set is the "true" parameter field that gave rise to the observed data.

The standard approach to exploring the posterior distribution is based on sampling using a Markov chain Monte Carlo (MCMC) method. However, the use of conventional MCMC methods becomes intractable for large-scale inverse problems, which arise upon discretization of infinite-dimensional inverse problems. This is due to the twin difficulties of high dimensionality of the uncertain parameters and computationally expensive forward models.

A number of methods have emerged to address Bayesian inverse problems governed by PDEs (we give a representative recent reference in each case, which can be consulted for additional references to historical work; further references can be found in the recent survey [19]): replacing the forward problem with a reduced order model in both parameter and state space [35]; approximating the parameter-to-observable map [30] or the posterior [10] with a Gaussian process response surface; employing a polynomial chaos approximation of the forward problem [38]; using a two-stage "delayed acceptance" MCMC method in which the first stage employs an approximate forward model [14]; employing gradient information (of the negative log posterior) to accelerate sampling, as in Langevin methods [16, 51, 54] and their preconditioned variants [2]; exploiting Riemannian geometry of parameter space to accelerate sampling [22]; and creating an MCMC proposal that uses local gradient and low-rank Hessian information of the negative log posterior to construct a local Gaussian approximation [37].

Here we focus on the last of these methods, the so-called *stochastic Newton MCMC method*. This method employs local Hessian-based Gaussian proposals that exploit the structure of the underlying posterior to guide the sampler to regions with higher acceptance probability. In particular, such proposals capture the highly stretched contours of the posterior that are typical for ill-posed inverse problems, in which the data inform the model parameters very well in some directions in parameter space, and poorly in others. One of the challenges in employing the Hessian is that its explicit construction entails solution of as many forward problems as there are parameters, which is out of the question for large-scale forward problems. These difficulties are addressed by introducing low-rank approximations of the Hessian, motivated by the compact nature of the Hessian operator for many inverse problems [6, 7, 8, 9, 11, 18, 37]. This delivers accurate approximation of the Hessian at a cost that is independent of the parameter dimension (when the parameter represents a discretized field), leading to solution of Bayesian inverse problems with nontrivial dimensions [37]. Other work employing Hessian-based proposals includes the tailored chain approach [21], and, specifically, in the context of nonlinear filtering [20]; the Hessian-based Metropolis–Hastings (HMH) algorithm with a learning rate to influence step size [48]; a position-specific preconditioned Metropolis adjusted Langevin algorithm (PSP-MALA) implemented with a block Metropolis–Hastings algorithm [29]; function-space MCMC proposals for which the prior is invariant, and thus insensitive to mesh refinement [34]; and, finally, the random maximized likelihood (RML) algorithm which generates samples as the solutions of related deterministic inverse problems [42].

Despite the low-rank approximation, stochastic Newton MCMC (and any method that uses local Hessian information) is computationally expensive for large-scale problems, since at every proposed sample point the gradient and a low-rank approximation

of the Hessian are computed, which requires multiple forward and adjoint PDE solves having the same linear operator (or its adjoint). When these PDE solves are done iteratively, there is little opportunity to exploit the fact that the linear operators are the same, beyond amortizing the cost of preconditioner construction over the solves.

To alleviate this computational cost, here we propose a modified stochastic Newton MCMC that uses proposals based on local gradient information as well as on Hessian information computed initially at the maximum a posteriori (MAP) point and then reused at every sample point. We call this the *stochastic Newton MCMC method with MAP-based Hessian*. We compare this proposed method with the original stochastic Newton MCMC method (with dynamically computed Hessian) as well as with an independence sampler that uses a Gaussian proposal centered at the MAP point, using the Hessian computed at the MAP as the covariance [42]. This independence sampler is computationally attractive since (like the proposed stochastic Newton MCMC method with MAP-based Hessian), the Hessian is computed just once, but (unlike the new method), the gradient is used only to determine the MAP. Because the proposed stochastic Newton MCMC method with MAP-based Hessian uses local (gradient) information, we expect it will outperform the independence sampler; because it freezes the Hessian at the MAP point, it will be significantly cheaper per sample than the original stochastic Newton MCMC method.

The stochastic Newton MCMC method with MAP-based Hessian can be derived as a particular variant of a preconditioned Metropolis-adjusted Langevin algorithm using preconditioning based on the Hessian at the MAP point. Note that all of the above methods attempt to exploit problem structure—in particular the local curvature of the posterior—by making use of Hessian information to one degree or another. Note also that all three of these Hessian-based methods reduce to the same method when the target inverse problem is linear and the prior and noise pdfs are Gaussian (in which case the posterior is also Gaussian). For non-Gaussian posteriors, however, the three methods take distinct steps.

Beyond this new, more efficient variant of stochastic Newton MCMC, this article extends our previous work on methods for large-scale Bayesian inverse problems [11, 37] in several directions. In [11], we presented a computational framework for linearized infinite-dimensional Bayesian inverse problems, building on the infinite-dimensional formulation of Stuart [53]. Here, we extend our computational framework to nonlinear inverse problems, for which the posteriors are non-Gaussian, requiring MCMC sampling. To this end, we extend the finite-dimensional stochastic Newton MCMC method presented in [37] to be consistent with the infinite-dimensional setting. This requires care in discretizing the prior and likelihood and establishing finite-dimensional inner products, which arise in multiple steps of stochastic Newton.

We study the efficiency of the proposed method in the context of an ice sheet flow Bayesian inverse problem, in which a basal boundary condition parameter field is inferred from surface velocity observations. Here, the parameter-to-observable map involves the solution of a nonlinear Stokes equation describing viscous, creeping, incompressible, non-Newtonian ice flow. This extends recent research on ice sheet inverse problems, which focused on deterministic inversion or the computation of the MAP solution [24, 40, 45, 46, 47, 49]. We apply the full Bayesian inference framework and study the performance of the three Hessian-based methods described above in exploring the posterior pdf. Convergence of the three methods is studied using various diagnostics to assess MCMC chain convergence. We also compare our method to the delayed rejection adaptive metropolis (DRAM) sampler [27], which, similar

to stochastic Newton, attempts to capture the curvature of the posterior, but without relying on gradient or Hessian information. The results reveal that, among the Hessian-based methods, the stochastic Newton MCMC method with MAP-based Hessian yields the fastest convergence in terms of both the number of samples and the computational work. In comparison, DRAM is incapable of making progress on this problem.

Finally, we study and interpret visually the solution of the Bayesian inverse problem with respect to the information contained in the data and in the prior and the effect they have on the posterior in high dimensions. This can be challenging in high dimensions, but we demonstrate that it can be made tractable by exploiting information contained within the spectral structure of the Hessian of the log likelihood evaluated at the MAP point as well as the prior covariance. Because this structure is common to many Bayesian inverse problems, we expect that these strategies for visualization will be of general value beyond the specific application.

The remaining sections of this paper are organized as follows. We begin by providing in section 2.1 an overview of the framework for infinite-dimensional Bayesian inverse problems following [11, 53]. Next, in section 2.2 we present a consistent discretization of the infinite-dimensional inverse problem. Section 3 presents the proposed stochastic Newton MCMC method with MAP-based Hessian, while section 3.5 describes our low rank-based Hessian approximation. Section 4 introduces a Bayesian formulation of an ice sheet flow inverse problem and gives expressions for adjoint-based gradient and Hessian-vector products (of the negative log posterior). In section 5, we discuss the performance of the three sampling methods. Finally, in section 6 we interpret the posterior distribution by visualizing marginals with respect to the eigenvectors of the covariance operator. This provides insight into the ability of the observations to infer model parameters. Section 7 provides concluding remarks.

2. Background on the infinite-dimensional Bayesian inverse problem, its consistent discretization, and characterization of the posterior. Formulating and solving the Bayesian inverse problem for an infinite-dimensional parameter field presents difficulties. First, the usual notion of a pdf is not defined since there is no Lebesgue measure in infinite dimensions. Second, the prior measure must be chosen appropriately to lead to a well-posed inverse problem and facilitate computation of the posterior. Third, the choice of discretization must be consistent with the infinite-dimensional structure of the problem. Finally, exploring the posterior that arises upon discretization via an MCMC method is typically prohibitive due to the resulting high dimensionality of the parameter space.

In this section we formulate the Bayesian inverse problem in infinite dimensions (section 2.1) in the framework of [53], which uses the Radon–Nikodým derivative and an appropriately chosen Gaussian prior that employs as covariance operator the inverse of an elliptic differential operator. In section 2.2, we describe the discretization of this infinite-dimensional inverse problem in a way that is consistent with the underlying infinite-dimensional function spaces. This leads to nonstandard definitions of operator adjoints. When the posterior is nearly Gaussian, its mean and covariance can be approximated by the MAP point and the inverse of the Hessian evaluated at the MAP. Inversion of the Hessian is intractable in high dimensions; section 3.5 presents a low-rank approximation of the Hessian of the data misfit in order to make these Hessian computations tractable. When the posterior is not approximately Gaussian, the method of choice is often to sample it with an MCMC method and then compute sample statistics; section 2.3 gives an overview of MCMC methods for sampling

posteriors.

2.1. Bayesian formulation of infinite-dimensional inverse problems. In an inverse problem, we seek to infer the unknown (or uncertain) input parameters to a mathematical model from observations of the outputs of the model. For ill-posed inverse problems, the uncertain parameter $m \in \mathcal{H}$ is often a heterogeneous field over a domain Ω, and $\mathcal{H}$ is typically a subset of $L^2(\Omega)$. The mathematical model is characterized by the parameter-to-observable map $\boldsymbol{f} : \mathcal{H} \to \mathbb{R}^q$, which predicts observables $\boldsymbol{y} \in \mathbb{R}^q$ corresponding to a given parameter m. Note that this map involves solution of the forward problem, typically a system of PDEs, followed by an application of an observation operator. We assume here that the observables $\boldsymbol{y}$ are finite-dimensional. Given observation data $\boldsymbol{d}^{\mathrm{obs}} \in \mathbb{R}^q$, the solution to the inverse problem seeks parameters m such that

$$\boldsymbol{f}(m) \approx \boldsymbol{d}^{\mathrm{obs}}$$

in a sense made precise by the Bayesian formulation described next.

The Bayesian formulation poses the inverse problem as a problem of statistical inference over parameter space. The solution of the resulting *Bayesian inverse problem* is a probability distribution that represents our belief about the correct value of the parameter. Solving the inverse problem using Bayes' approach requires specification of a *prior model*, which describes our beliefs about the parameter before any data are considered, and a *likelihood model*, which quantifies the relative probability that a candidate parameter m could have produced the observed data $\boldsymbol{d}^{\mathrm{obs}}$.

Here we present a summary of the discussion in [11]. The prior is taken to be the Gaussian measure $\mu_0 = \mathcal{N}(m_0, \mathcal{C}_0)$ on $L^2(\Omega)$, where $m_0 \in \mathcal{H}$, and $\mathcal{C}_0$ is an appropriate covariance operator $\mathcal{C}_0$; in particular, $\mathcal{C}_0$ must be symmetric, positive, and of traceclass [53]. We choose the covariance operator to be the inverse of an elliptic differential operator $\mathcal{A}$ that is of sufficiently high order to guarantee a well-posed Bayesian inverse problem [53]. We choose $\mathcal{A}$ to be second order differential operator[1] expressed in weak form: for $s \in L^2(\Omega)$, the solution $m = \mathcal{A}^{-1}s$ satisfies

$$(2.1) \qquad \int_\Omega [a\nabla m \cdot \nabla p + bmp]\, d\boldsymbol{x} = \int_\Omega sp\, d\boldsymbol{x} \quad \text{for all } p \in H^1(\Omega),$$

with $a, b > 0$. These coefficients control the correlation length and the variance in the covariance operator $\mathcal{A}^{-1}$. Choosing for spatially dependent coefficients a and b or a tensor coefficient a allows the incorporation of further problem-specific knowledge, such as spatially varying or anisotropic correlations, in the covariance operator $\mathcal{A}^{-1}$ [11].

For the likelihood model, we assume that observational uncertainty (i.e., uncertainty in $\boldsymbol{d}^{\mathrm{obs}}$ related to measurement error) and model uncertainty (i.e., uncertainty in $\boldsymbol{f}(m)$ due to inadequacy of the forward model) are each centered, additive, and Gaussian. We combine these into a single *noise model*,

$$\boldsymbol{f}(m) = \boldsymbol{y} + \boldsymbol{\eta}, \qquad \text{with} \qquad \boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}_{\mathrm{noise}}),$$

---

[1]The necessary order of $\mathcal{A}$ to lead to a valid covariance operator depends on the spatial dimension of the domain Ω [53]. In the example considered in section 4, the inversion parameter is a one-dimensional field, and a second order differential operator is sufficient to guarantee that $\mathcal{C}_0$ is a valid covariance operator. While there is no distinction in one dimension between ordinary and partial derivatives, we choose to express $\mathcal{A}$ in the language of PDEs for notational consistency with the development in [11].

where $\boldsymbol{\eta} \in \mathbb{R}^q$ is a random variable representing noise, and $\boldsymbol{\Gamma}_{\text{noise}} \in \mathbb{R}^{q \times q}$ is the noise covariance matrix. We can then express the pdf for the likelihood model explicitly as

$$(2.2) \qquad \pi_{\text{like}}(\boldsymbol{d}^{\text{obs}}|m) \propto \exp\left[-\frac{1}{2}(\boldsymbol{f}(m) - \boldsymbol{d}^{\text{obs}})^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{f}(m) - \boldsymbol{d}^{\text{obs}})\right].$$

Bayes' theorem in infinite dimensions is expressed using the Radon–Nikodým derivative $\frac{d\mu^y}{d\mu_0}$ of the posterior measure μ^y with respect to the prior measure μ_0,

$$(2.3) \qquad \frac{d\mu^y}{d\mu_0} = \frac{1}{Z}\pi_{\text{like}}(\boldsymbol{d}^{\text{obs}}|m),$$

where $Z = \int_X \pi_{\text{like}}(\boldsymbol{d}^{\text{obs}}|m)\,d\mu_0$ is a normalization constant. For technical conditions under which the posterior measure is well defined, and a discussion of the Bayes rule for probability measures on function spaces, we refer the reader to [12, 13, 53].

2.2. Discretization of the Bayesian inverse problem. In this section, we present a brief discussion of the finite-dimensional approximations of the prior and the posterior distributions; a lengthier discussion can be found in [11]. We start with a finite-dimensional subspace V_h of $L^2(\Omega)$ originating from a finite element discretization with continuous Lagrange basis functions $\{\phi_j\}_{j=1}^n$ [1, 52]. The approximation of the inversion parameter function $m \in L^2(\Omega)$ is then $m_h = \sum_{j=1}^n m_j \phi_j \in V_h$, where the vector of the n inversion parameters is $\boldsymbol{m} = (m_1, \ldots, m_n)^T \in \mathbb{R}^n$.

Since we postulate the prior Gaussian measure on $L^2(\Omega)$, the finite-dimensional space V_h inherits the L^2-inner product. Thus, inner products between nodal coefficient vectors must be weighted by a mass matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ to approximate the infinite-dimensional L^2-inner product. This M-weighted inner product is denoted by $\langle \cdot, \cdot \rangle_{\boldsymbol{M}}$, where $\langle \boldsymbol{y}, \boldsymbol{z} \rangle_{\boldsymbol{M}} = \boldsymbol{y}^T \boldsymbol{M} \boldsymbol{z}$ and $\boldsymbol{M}$ is the (symmetric positive definite) mass matrix

$$M_{ij} = \int_\Omega \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})\,d\boldsymbol{x}\,, \quad i, j = 1, \ldots, n.$$

To distinguish $\mathbb{R}^n$ equipped with the M-weighted inner product with the usual Euclidean space $\mathbb{R}^n$, we denote it by $\mathbb{R}_{\boldsymbol{M}}^n$.

When using the M-weighted inner product, there is a critical distinction that must be made between the matrix adjoint and the matrix transpose. For an operator $\boldsymbol{B} : \mathbb{R}_{\boldsymbol{M}}^n \to \mathbb{R}_{\boldsymbol{M}}^n$, we denote the matrix transpose by $\boldsymbol{B}^T$ with entries $(B^T)_{ij} = B_{ji}$. In contrast, the M-weighted inner product adjoint $\boldsymbol{B}^*$ satisfies, for $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$,

$$\langle \boldsymbol{B}\boldsymbol{y}, \boldsymbol{z} \rangle_{\boldsymbol{M}} = \langle \boldsymbol{y}, \boldsymbol{B}^*\boldsymbol{z} \rangle_{\boldsymbol{M}},$$

which implies that $\boldsymbol{B}^*$ is given by

$$(2.4) \qquad \boldsymbol{B}^* = \boldsymbol{M}^{-1}\boldsymbol{B}^T\boldsymbol{M}.$$

In the following, we also need the adjoint $\boldsymbol{V}^\diamond$ of $\boldsymbol{V} : \mathbb{R}^r \to \mathbb{R}_{\boldsymbol{M}}^n$ (for some r), where $\mathbb{R}^r$ is endowed with the Euclidean inner product. In this case, we have

$$(2.5) \qquad \boldsymbol{V}^\diamond = \boldsymbol{V}^T\boldsymbol{M},$$

since $\langle \boldsymbol{V}\boldsymbol{y}, \boldsymbol{z} \rangle_{\boldsymbol{M}} = \langle \boldsymbol{y}, \boldsymbol{V}^\diamond \boldsymbol{z} \rangle$. With these definitions, the matrix representation of the elliptic PDE operator $\mathcal{A}$ defined by (2.1) is given by $\boldsymbol{A} = \boldsymbol{M}^{-1}\boldsymbol{K} \in \mathbb{R}^{n \times n}$ [11], where $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ is the stiffness matrix

$$K_{ij} = \int_\Omega [a\nabla\phi_i(\boldsymbol{x}) \cdot \nabla\phi_j(\boldsymbol{x}) + b\phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})]\,d\boldsymbol{x}, \quad i, j \in \{1, \ldots, n\}.$$

Then the finite-dimensional approximation μ_0^h of the prior Gaussian measure μ_0 is the more familiar multivariate Gaussian with density

$$(2.6) \qquad \pi_{\text{prior}}(\boldsymbol{m}) \propto \exp\left[-\frac{1}{2}\left\langle \boldsymbol{m} - \boldsymbol{m}_0, \boldsymbol{A}(\boldsymbol{m} - \boldsymbol{m}_0)\right\rangle_{\boldsymbol{M}}\right],$$

where $\boldsymbol{m}_0 \in \mathbb{R}^n$ is the discretization of the prior mean m_0. The finite-dimensional Bayes' formula, i.e.,

$$(2.7) \qquad \pi_{\text{post}}(\boldsymbol{m}) := \pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d}^{\text{obs}}) \propto \pi_{\text{prior}}(\boldsymbol{m})\pi_{\text{like}}(\boldsymbol{d}^{\text{obs}}|\boldsymbol{m}),$$

where $\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{d}^{\text{obs}})$ is the density of the finite-dimensional approximation $\mu^{y,h}$ of the posterior measure μ^y, and π_{like} is the likelihood (2.2), gives the finite-dimensional posterior density explicitly as

$$(2.8) \quad \pi_{\text{post}}(\boldsymbol{m}) \propto \exp\left[-\frac{1}{2}\left\|\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}^{\text{obs}}\right\|_{\boldsymbol{\Gamma}_{\text{noise}}^{-1}}^2 - \frac{1}{2}\left\langle \boldsymbol{m} - \boldsymbol{m}_0, \boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{m} - \boldsymbol{m}_0)\right\rangle_{\boldsymbol{M}}\right],$$

where $\boldsymbol{\Gamma}_{\text{prior}} = \boldsymbol{A}^{-1}$. Note that in (2.8) and the remainder of this paper we denote by $\boldsymbol{f}(\boldsymbol{m})$ the parameter-to-observable map evaluated at the finite element function corresponding to the parameter vector $\boldsymbol{m}$. The Bayesian solution of the inverse problem is then given by (2.8). Unfortunately, for inverse problems governed by expensive forward models and for high-dimensional parameter spaces, exploring the posterior density $\pi_{\text{post}}(\boldsymbol{m})$ is extremely challenging, since evaluation of this density at any point in parameter space requires the solution of the forward model $\boldsymbol{f}(\boldsymbol{m})$ for the given $\boldsymbol{m}$, and a very large number of such evaluations will be required in high dimensions. Methods for exploring $\pi_{\text{post}}(\boldsymbol{m})$ that do not exploit its structure are thus impractical.

We observe that the negative log posterior density is analogous to the least squares functional that is minimized in the solution of a deterministic inverse problem. That is,

$$(2.9) \qquad -\log \pi_{\text{post}}(\boldsymbol{m}) = J(\boldsymbol{m}) + \text{ const.},$$

where

$$(2.10) \qquad J(\boldsymbol{m}) := \frac{1}{2}\|\boldsymbol{f}(\boldsymbol{m}) - \boldsymbol{d}^{\text{obs}}\|_{\boldsymbol{\Gamma}_{\text{noise}}^{-1}}^2 + \frac{1}{2}\left\langle \boldsymbol{m} - \boldsymbol{m}_0, \boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{m} - \boldsymbol{m}_0)\right\rangle_{\boldsymbol{M}}.$$

In the context of deterministic inversion, the first term in (2.10) is the data misfit term, weighted by $\boldsymbol{\Gamma}_{\text{noise}}^{-1}$, and the second term plays the role of Tikhonov regularization, which is chosen to make the inverse problem well-posed. This connection between the negative log posterior and the deterministic inverse problem cost function in (2.10) is often exploited to find an approximation of the mean of the posterior pdf by finding the point that maximizes the posterior $\pi_{\text{post}}(\boldsymbol{m})$, or equivalently minimizes the cost function $J(\boldsymbol{m})$. This so-called maximum a posteriori (MAP) point is equal to the mean when the parameter-to-observable map $\boldsymbol{f}(\boldsymbol{m})$ is linear in the parameters $\boldsymbol{m}$ and the noise and prior models are Gaussian. When the Gaussian-linear conditions are not satisfied, obviously the MAP point only approximates the mean, the quality of this approximation depending on the degree of nonlinearity. Moreover, under these Gaussian-linear conditions, the posterior $\pi_{\text{post}}(\boldsymbol{m})$ is Gaussian with mean given by the MAP point, and covariance given by the inverse of the Hessian matrix of the cost function $J(\boldsymbol{m})$ [11, 55].

Algorithm 1. Metropolis–Hastings MCMC algorithm to sample the pdf π.

Choose initial parameters $\boldsymbol{m}_0$
Compute $\pi(\boldsymbol{m}_0)$
for $k = 0, \ldots, N-1$ **do**
  Draw sample $\boldsymbol{y}$ from the proposal density $q(\boldsymbol{m}_k, \cdot)$
  Compute $\pi(\boldsymbol{y})$
  Compute $\alpha_k(\boldsymbol{y}) = \min\left\{1, \frac{\pi(\boldsymbol{y})q(\boldsymbol{y},\boldsymbol{m}_k)}{\pi(\boldsymbol{m}_k)q(\boldsymbol{m}_k,\boldsymbol{y})}\right\}$
  Draw $u \sim \mathcal{U}([0,1])$
  if $u < \alpha_k(\boldsymbol{y})$ **then**
    Accept: Set $\boldsymbol{m}_{k+1} = \boldsymbol{y}$
  else
    Reject: Set $\boldsymbol{m}_{k+1} = \boldsymbol{m}_k$
  end if
end for

2.3. Exploring the posterior. As implied above, when the parameter-to-observable map is nonlinear, the posterior $\pi_{\text{post}}(\boldsymbol{m})$ generally is non-Gaussian, and cannot be represented by its mean and covariance. Thus it must be characterized by other means. This can be extremely challenging for PDE-based inverse problems, since evaluating the posterior (2.8) at any point in parameter space involves solving the forward PDEs, and many such evaluations are anticipated for the high-dimensional parameter spaces that stem from discretization of infinite-dimensional inverse problems.

The method of choice for exploring the posterior pdf is the Metropolis–Hastings (M-H) MCMC method [28, 39, 50, 56], which employs a given proposal probability density $q(\boldsymbol{m}_k, \boldsymbol{y})$ at each sample point $\boldsymbol{m}_k$ in parameter space to generate a proposed sample point $\boldsymbol{y} \in \mathbb{R}^n$. Once generated, the M-H criterion chooses to either accept or reject the proposed sample point and repeats from the new point, thereby generating a chain of samples $\{\boldsymbol{m}_k\}_{k=1,\ldots}$ from the posterior density $\pi_{\text{post}}(\boldsymbol{m})$. Algorithm 1 presents pseudocode for the M-H MCMC method.

Critical to the success of M-H MCMC is the choice of the proposal density $q(\boldsymbol{m}_k, \boldsymbol{y})$. Observe that if $q(\boldsymbol{m}_k, \boldsymbol{y}) = \pi_{\text{post}}(\boldsymbol{y})$, the M-H algorithm would accept every sample with probability 1; however, this defeats the purpose, because we would not know how to sample from this choice of proposal: the whole point of appealing to MCMC is that we cannot draw a sample directly from $\pi_{\text{post}}(\boldsymbol{y})$.

Instead, a common choice for the proposal is the isotropic Gaussian,

$$q^{\text{RWMH}}(\boldsymbol{m}_k, \boldsymbol{y}) = \tfrac{1}{(2\pi)^{n/2}} \exp[-\tfrac{1}{2}(\|\boldsymbol{m}_k - \boldsymbol{y}\|)^2].$$

The resulting method is known as random walk Metropolis–Hastings (RWMH). This proposal density is easy to sample, but it can lead to poor MCMC performance due to the mismatch between the proposal and posterior densities. The challenge is to come up with a proposal that at least locally reflects the behavior of the target posterior density and at the same time is easy to sample. Satisfying these two requirements becomes increasingly difficult with increasing parameter dimension. This will be the subject of the next section.

3. A modified stochastic Newton MCMC method. In [37], we introduced a so-called stochastic Newton MCMC method that featured a Gaussian proposal constructed from the local gradient vector and local Hessian matrix (of the negative

log posterior). To make the construction of the proposal tractable, we employed adjoint-based methods to compute the gradient and Hessian, which amount to a pair of forward/adjoint PDE solves for the gradient and for each column of the Hessian. Moreover, to make the Hessian computation scalable with respect to parameter dimension, we use matrix-free methods to construct low-rank approximations of the data misfit component of the Hessian, which often has a rapidly decaying spectrum reflecting the ill-posedness of the inverse problem [18, 37]. With these features, the stochastic Newton method is able to handle inverse problems with hundreds to thousands of parameters; its efficiency increases with decreasing nonlinearity of the parameter-to-observable map and with decreasing information content of the data. We denote this original form of the stochastic Newton MCMC method (i.e., with dynamically changing Hessian) as *SN*.

Unfortunately, SN becomes prohibitive for very large-scale problems, because it requires recomputation of the Hessian at each sample point. Despite the use of efficient adjoint-based matrix-free Hessian-vector products to find the low-rank approximations of the data misfit component of the Hessian, we still need $O(2r)$ linearized forward/adjoint PDE solves to compute it, where r is the effective rank. When r is large—as is the case for high-dimensional problems, for which the observations are highly informative about the parameters and, hence, the data misfit Hessian has a high effective rank—we must find alternatives to computing the Hessian at each sample point.

Here we propose a modified stochastic Newton MCMC method that employs a (low-rank approximation-based) Hessian that is computed once and for all at the MAP point and reused for each proposal. This modification, which we refer to as *stochastic Newton MCMC with MAP-based Hessian (SNMAP)*, employs a locally computed gradient in the Gaussian proposal, but evaluates the Hessian in that Gaussian at the MAP point. Before describing SNMAP, we begin with a brief summary of the proposal construction for the original stochastic Newton MCMC method.

3.1. Stochastic Newton MCMC with dynamically changing Hessian (SN). The stochastic Newton MCMC method employs a local Gaussian approximation of the target posterior pdf. This is done by constructing, about a given point $\boldsymbol{m}_k$, a local quadratic approximation $\tilde{J}_k(\boldsymbol{m})$ of the negative log posterior $J(\boldsymbol{m})$ (given in (2.10)), i.e.,

$$(3.1) \qquad \tilde{J}_k(\boldsymbol{m}) := J(\boldsymbol{m}_k) + \langle \boldsymbol{g}_k, \boldsymbol{m} - \boldsymbol{m}_k \rangle_{\boldsymbol{M}} + \frac{1}{2} \langle \boldsymbol{m} - \boldsymbol{m}_k, \boldsymbol{H}_k(\boldsymbol{m} - \boldsymbol{m}_k) \rangle_{\boldsymbol{M}}.$$

Here, $\boldsymbol{g}$ and $\boldsymbol{H}$ are the gradient vector and Hessian matrix of $J(\boldsymbol{m})$, respectively, and $\boldsymbol{g}_k := \boldsymbol{g}(\boldsymbol{m}_k) \in \mathbb{R}^n$ and $\boldsymbol{H}_k := \boldsymbol{H}(\boldsymbol{m}_k) \in \mathbb{R}^{n \times n}$. Rearranging terms,

$$\tilde{J}_k(\boldsymbol{m}) = \frac{1}{2} \left\langle \boldsymbol{m} - \boldsymbol{m}_k + \boldsymbol{H}_k^{-1}\boldsymbol{g}_k, \boldsymbol{H}_k(\boldsymbol{m} - \boldsymbol{m}_k + \boldsymbol{H}_k^{-1}\boldsymbol{g}_k) \right\rangle_{\boldsymbol{M}} + \text{const.}$$

To obtain the proposal density q^{SN} for stochastic Newton MCMC (with dynamically changing Hessian), we take the exponential of the negative of $\tilde{J}_k(\boldsymbol{m})$ and compute the scaling factor to make it a proper pdf. This leads to
(3.2)

$$q^{\text{SN}}(\boldsymbol{m}_k, \boldsymbol{y}) = \frac{\det \boldsymbol{H}_k^{1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \left\langle \boldsymbol{y} - \boldsymbol{m}_k + \boldsymbol{H}_k^{-1}\boldsymbol{g}_k, \boldsymbol{H}_k(\boldsymbol{y} - \boldsymbol{m}_k + \boldsymbol{H}_k^{-1}\boldsymbol{g}_k) \right\rangle_{\boldsymbol{M}} \right),$$

which is a Gaussian with mean $\boldsymbol{m}_k - \boldsymbol{H}_k^{-1}\boldsymbol{g}_k$ and covariance matrix $\boldsymbol{H}_k^{-1}$. Note that at a local minimum, $\boldsymbol{H}_k$ is positive semidefinite and at an arbitrary point $\boldsymbol{y}$, $\boldsymbol{H}_k$ can be
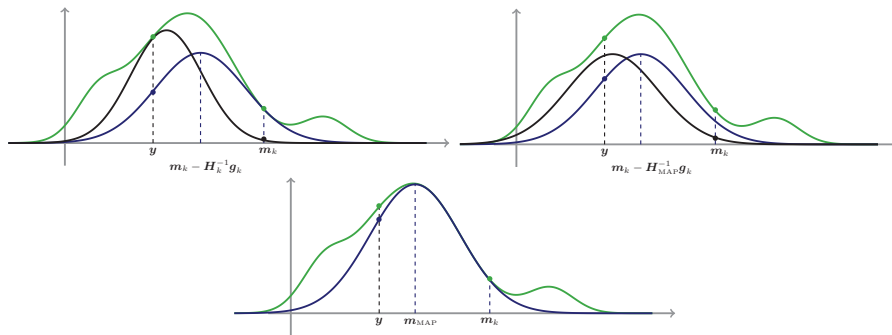
FIG. 1. *Illustration of proposals for the three Hessian-based methods: stochastic Newton MCMC with dynamically computed Hessian (top left); stochastic Newton MCMC with MAP-based Hessian (top right); and independence sampler with MAP-based Hessian (bottom). The green curve depicts the true posterior density, $\pi_{\mathrm{post}}(\boldsymbol{m})$; the blue curve displays the forward proposal density, $q(\boldsymbol{m}_k, \boldsymbol{y})$; and the black curve shows the backward proposal density, $q(\boldsymbol{y}, \boldsymbol{m}_k)$. The green, blue, and black dotted lines indicate the points at which the posterior, the backward, and the forward proposals, respectively, are evaluated.*

indefinite. To ensure that (3.2) defines a proper pdf, we discard negative eigenvalues of the data misfit component of $\boldsymbol{H}_k$ and hence replace $\boldsymbol{H}_k$ with a modified positive definite Hessian. We also note that the backward proposal $q^{\mathrm{SN}}(\boldsymbol{y}, \boldsymbol{m}_k)$, needed for the M-H acceptance probability α_k, is computed using the Hessian and gradient evaluated at $\boldsymbol{y}$. In summary, the SN step at each MCMC iteration draws a proposed sample $\boldsymbol{y}$ from the proposal $q^{\mathrm{SN}}(\boldsymbol{m}_k, \boldsymbol{y})$, which is then subject to the accept/reject framework of the M-H MCMC algorithm. The SN proposal is illustrated in Figure 1 (top left).

3.2. Stochastic Newton MCMC with MAP-based Hessian (SNMAP). As stated above, the original form of the stochastic Newton MCMC method becomes prohibitive for very large-scale problems, because it requires recomputation of the Hessian of $J(\boldsymbol{m})$ (whose inverse is needed to construct the Gaussian proposal) at each sample point. Therefore, we avoid recomputing this Hessian by the following modification: we first find the MAP point and compute the Hessian there, and then we use this MAP-based Hessian for all proposals. The gradient is still computed at each sample point. Hence, the proposal $q^{\mathrm{SNMAP}}(\boldsymbol{m}_k, \boldsymbol{y})$ for SNMAP is given by replacing the Hessian in (3.2) with the Hessian evaluated at the MAP point, $\boldsymbol{H}_{\mathrm{MAP}}$. This leads to

(3.3)
$$q^{\mathrm{SNMAP}}(\boldsymbol{m}_k, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\left\langle \boldsymbol{y} - \boldsymbol{m}_k + \boldsymbol{H}_{\mathrm{MAP}}^{-1}\boldsymbol{g}_k, \boldsymbol{H}_{\mathrm{MAP}}(\boldsymbol{y} - \boldsymbol{m}_k + \boldsymbol{H}_{\mathrm{MAP}}^{-1}\boldsymbol{g}_k)\right\rangle_{\boldsymbol{M}}\right),$$

which is a Gaussian with mean $\boldsymbol{m}_k - \boldsymbol{H}_{\mathrm{MAP}}^{-1}\boldsymbol{g}_k$ and covariance matrix $\boldsymbol{H}_{\mathrm{MAP}}^{-1}$. Note that the scaling factor is not necessary in (3.3), since for proposals with MAP-based Hessians, the scaling factors are constant and thus they cancel when computing the acceptance probability α_k in Algorithm 1. The SNMAP proposal is illustrated in Figure 1 (top right). Note that the SNMAP proposal (3.3) can also be understood as a preconditioned Langevin MCMC proposal [54] with a MAP-based Hessian preconditioner.

Avoiding SN's Hessian recomputation at each sample point results in substantial computational savings, since, as will be made explicit in section 3.5, computing the Hessian typically requires a number of forward/adjoint PDE solves on the order of the

effective rank of (a properly preconditioned) Hessian of the data misfit term in the negative log posterior. Once SNMAP has computed the Hessian at the MAP point, the only cost per sample is a pair of forward/adjoint PDE solves to compute the gradient. However, this may result in a deterioration in the acceptance rate, since the Gaussian proposal employs local gradient information but a global Hessian, and thus may not fully capture local curvature information of the posterior if the curvature is changing rapidly (as may happen in a highly nonlinear parameter-to-observable map). However, the fact that the proposal q^{SNMAP} changes less from sample to sample compared to q^{SN} can also have a positive effect on the acceptance probability and the chain convergence. In section 5 we conduct numerical experiments on a specific Bayesian inverse problem to assess whether this trade-off is profitable.

3.3. Independence sampling with a MAP point-based Gaussian proposal (ISMAP). As seen in the above SNMAP modification of the stochastic New-ton MCMC method, freezing the Hessian at the MAP point avoids Hessian recomputation and results in substantial savings. However, the gradient is still recomputed at each sample point, motivated by the desire to construct a Gaussian proposal that captures some local information, as well as the fact that the gradient is far cheaper to compute than (a low-rank approximation-based) Hessian.

One can go one step further and shed the need to compute local gradient information by defining an independence sampler that takes the proposal to be a Gaussian centered at the MAP point, using the Hessian computed at the MAP as the inverse covariance, and neglecting the gradient (since it vanishes at the MAP). This method has been suggested previously in the subsurface flow inversion literature [41, 42]; here, we refer to it as *ISMAP*. Since ISMAP, like SN and SNMAP, makes use of Hessian information, a fair assessment of SNMAP vis-á-vis SN should include comparisons to ISMAP as well. Therefore, we next provide a description of the ISMAP proposal as well.

The proposal density q^{ISMAP} is obtained by taking $\boldsymbol{m}_k$ in (3.2) as the MAP point $\boldsymbol{m}_{\text{MAP}}$ (which means that $\boldsymbol{g}_k$ is zero) and, as with SNMAP, replacing the Hessian at $\boldsymbol{m}_k$ with the Hessian evaluated at the MAP point, $\boldsymbol{H}_{\text{MAP}}$. This leads to

$$(3.4) \qquad q^{\text{ISMAP}}(\boldsymbol{m}_{\text{MAP}}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2} \left\langle \boldsymbol{y} - \boldsymbol{m}_{\text{MAP}}, \boldsymbol{H}_{\text{MAP}}(\boldsymbol{y} - \boldsymbol{m}_{\text{MAP}}) \right\rangle_{\boldsymbol{M}} \right),$$

which is a Gaussian with mean $\boldsymbol{m}_{\text{MAP}}$ and covariance matrix $\boldsymbol{H}_{\text{MAP}}^{-1}$. We note that the proposal q^{ISMAP} is independent of the current sample point, and thus does not change during the sampling process. The ISMAP proposal is illustrated in Figure 1 (bottom).

We note that ISMAP not only avoids Hessian recomputation at each sample point (as with SNMAP) but also avoids computing the gradient; thus, its cost—once the MAP-based Hessian is determined—is a forward PDE solve at each sample point. However, this additional approximation over SNMAP has the potential to lead to additional deterioration of the acceptance rate. Note that one advantage of ISMAP is that, since the proposal is constant, the samples can all be precomputed offline or in parallel, after which they can be subjected (sequentially) to the M-H accept/reject criterion in Algorithm 1.

Finally, we remark that if the posterior itself is a Gaussian, the three Hessian-based methods described above collapse to the same method. As such, they all sample from the true posterior with probability 1 at every step, resulting in an acceptance rate of 100% and posterior samples that are independent [37].

3.4. Relation to Newton's method for optimization. Recall that the stochastic Newton MCMC method (in particular SN) uses, as a proposal, the local quadratic approximation $\tilde{J}_k(\boldsymbol{m})$ of the negative log posterior $J(\boldsymbol{m})$ about the current sample point $\boldsymbol{m}_k$. The minimizer of $\tilde{J}_k(\boldsymbol{m})$ is given by $\boldsymbol{m}_k - \boldsymbol{H}_k^{-1}\boldsymbol{g}_k$, where $\boldsymbol{H}_k$ and $\boldsymbol{g}_k$ are the Hessian and the gradient of $J(\boldsymbol{m})$ evaluated at $\boldsymbol{m}_k$, respectively. Note that $-\boldsymbol{H}_k^{-1}\boldsymbol{g}_k$ is the classical Newton optimization step. A proposal point drawn from the local Gaussian approximation of the posterior with mean $\boldsymbol{m}_k - \boldsymbol{H}_k^{-1}\boldsymbol{g}_k$ and covariance $\boldsymbol{H}_k^{-1}$ is thus

$$(3.5) \qquad \boldsymbol{y} = \boldsymbol{m}_k - \boldsymbol{H}_k^{-1}\boldsymbol{g}_k + \boldsymbol{H}_k^{-1/2}\tilde{\boldsymbol{n}},$$

where $\tilde{\boldsymbol{n}} = \boldsymbol{M}^{-1/2}\boldsymbol{n}$ is a random sample from a Gaussian with zero mean and identity covariance matrix in $\mathbb{R}_M^n$, and $\boldsymbol{n} \in \mathbb{R}^n$ is a random sample from the standard normal density in $\mathbb{R}^n$. Iterating the stochastic Newton MCMC method without the random term amounts to the classical Newton method from nonlinear optimization, which converges to the MAP point (or another stationary point of $J(\cdot)$).

Since SNMAP reuses the Hessian at the MAP point (i.e., it is held constant throughout the sampling process), proposal points are computed as in (3.5), but with $\boldsymbol{H}_k$ replaced by the Hessian at the MAP point $\boldsymbol{H}_{\mathrm{MAP}}$, i.e.,

$$(3.6) \qquad \boldsymbol{y} = \boldsymbol{m}_k - \boldsymbol{H}_{\mathrm{MAP}}^{-1}\boldsymbol{g}_k + \boldsymbol{H}_{\mathrm{MAP}}^{-1/2}\tilde{\boldsymbol{n}},$$

with $\tilde{\boldsymbol{n}}$ as above. We note that if the random term is neglected, SNMAP reduces to an $\boldsymbol{H}_{\mathrm{MAP}}$-preconditioned steepest descent method.

For completeness, let us show how proposals from the independence sampler with MAP-based Gaussian, ISMAP, are computed. With $\tilde{\boldsymbol{n}}$ defined as above, the proposed point is found as

$$(3.7) \qquad \boldsymbol{y} = \boldsymbol{m}_{\mathrm{MAP}} + \boldsymbol{H}_{\mathrm{MAP}}^{-1/2}\tilde{\boldsymbol{n}}.$$

Note that the right-hand side in (3.7) is independent of $\boldsymbol{m}_k$, and hence the designation "independence sampler."

3.5. Efficient operations with the Hessian via low-rank approximation. Up to this point, we have described the three Hessian-based MCMC methods (SN, SNMAP, and ISMAP) in terms of the Hessian matrix of the negative log posterior. Indeed, examination of the form of the three proposal densities (3.2), (3.3), and (3.4), as well as the expressions for the samples from the proposals (3.5), (3.6), and (3.7), reveals that the following operations with the Hessian are required: action of the Hessian on a vector; action of the inverse Hessian on a vector; action of the inverse of the square root of the Hessian on a vector; and determinant of the square root of the Hessian (the determinant is required only for SN).

Unfortunately, explicitly computing the Hessian requires as many (linearized) forward PDE solves as there are parameters; for large-scale problems, these computations are prohibitive. Thus, we need efficient algorithms for the operations with the Hessian summarized above. In this section, we briefly describe previous work that employs low-rank approximations of the data misfit portion of the Hessian, preconditioned by the prior covariance, to execute all of the above operations with the Hessian at a cost (measured in forward PDE solves) that is independent of the parameter dimension [11, 18, 37]. The discussion below is in terms of a generic Hessian, $\boldsymbol{H}$; this can refer to the Hessian at any point in parameter space, including the MAP point.

The Hessian of the negative log posterior $J(\boldsymbol{m})$ in (2.10) can be written as the sum of the Hessian of the data misfit term, $\boldsymbol{H}_{\mathrm{misfit}}$, and the inverse of the prior covariance $\boldsymbol{\Gamma}_{\mathrm{prior}}^{-1}$. If we consider a decomposition of the prior such that $\boldsymbol{\Gamma}_{\mathrm{prior}} = \boldsymbol{L}\boldsymbol{L}^*$, then

$$(3.8) \qquad \boldsymbol{H} = \boldsymbol{H}_{\mathrm{misfit}} + \boldsymbol{\Gamma}_{\mathrm{prior}}^{-1} = \boldsymbol{H}_{\mathrm{misfit}} + \boldsymbol{L}^{-*}\boldsymbol{L}^{-1} = \boldsymbol{L}^{-*}(\boldsymbol{L}^*\boldsymbol{H}_{\mathrm{misfit}}\boldsymbol{L} + \boldsymbol{I})\boldsymbol{L}^{-1}.$$

Here, the data misfit Hessian $\boldsymbol{H}_{\mathrm{misfit}}$ is given by

$$\boldsymbol{H}_{\mathrm{misfit}} := \boldsymbol{F}^{\natural}\boldsymbol{\Gamma}_{\mathrm{noise}}^{-1}\boldsymbol{F} + \text{ second order terms,}$$

where $\boldsymbol{F}$ is the Jacobian matrix of the parameter-to-observable map $\boldsymbol{f}(\boldsymbol{m})$, $\boldsymbol{F}^{\natural} := \boldsymbol{M}^{-1}\boldsymbol{F}^T$ is its (properly weighted) adjoint, and the second order terms involve second derivatives of $\boldsymbol{f}(\boldsymbol{m})$ with respect to $\boldsymbol{m}$. Notwithstanding the form of the second order terms, the expression above suggests that the Hessian of the data misfit involves the solution of linearized forward and adjoint PDE problems. This will be seen explicitly for the target ice sheet inverse problem described in section 4.

We begin by describing the computation of the application of the inverse Hessian to a vector in order to compute the Newton step $\boldsymbol{H}^{-1}\boldsymbol{g}$. From (3.8), we obtain

$$(3.9) \qquad \boldsymbol{H}^{-1}\boldsymbol{g} = \boldsymbol{L}(\boldsymbol{L}^*\boldsymbol{H}_{\mathrm{misfit}}\boldsymbol{L} + \boldsymbol{I})^{-1}\boldsymbol{L}^*\boldsymbol{g},$$

and thus we require the inverse of $(\boldsymbol{L}^*\boldsymbol{H}_{\mathrm{misfit}}\boldsymbol{L} + \boldsymbol{I})$. Since for ill-posed inverse problems observations typically inform only a limited number of eigenvectors of the parameter field, the spectrum of the data misfit Hessian often decays rapidly (see, for example, [7, 8, 9] for the inverse scattering case). In addition, the prior is often smoothing, in which case left and right preconditioning of the data misfit Hessian by the square root of the prior, $\boldsymbol{L}$, enhances the decay of the eigenvalues. Thus, the prior-preconditioned data misfit Hessian, $\boldsymbol{L}^*\boldsymbol{H}_{\mathrm{misfit}}\boldsymbol{L}$, can typically be well approximated by a low-rank matrix, and this can be exploited to enable efficient computations with the Hessian. To construct the low-rank approximation of the prior-preconditioned data misfit Hessian, we seek a matrix-free method (since $\boldsymbol{H}$ cannot be formed explicitly) that requires just Hessian-vector products; crucially, the number of Hessian-vector products must be of the order of the effective rank, r, of the prior-preconditioned data misfit Hessian, as opposed to the parameter dimension, n. Note that each Hessian-vector product can be formed efficiently at the cost of a single pair of linearized forward/adjoint PDE solves (this will be seen explicitly for the ice sheet flow problem in section 4.5).

The Lanczos eigenvalue algorithm meets the requirements outlined above, and we use it to construct an r-dimensional low-rank approximation for the prior-preconditioned data misfit Hessian, i.e., $\boldsymbol{L}^*\boldsymbol{H}_{\mathrm{misfit}}\boldsymbol{L} \approx \boldsymbol{V}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^{\diamond}$, where $\boldsymbol{V}_r \in \mathbb{R}^{n \times r}$ contains r eigenvectors of the prior-preconditioned data misfit Hessian corresponding to the r largest eigenvalues $\lambda_i, i = 1, \dots, r$, $\boldsymbol{\Lambda}_r = \mathrm{diag}(\lambda_1, \dots, \lambda_r) \in \mathbb{R}^{r \times r}$, and $\boldsymbol{V}_r^{\diamond} \in \mathbb{R}^{r \times n}$ denotes the adjoint defined in (2.5). The rank-r approximation can typically be formed in a number of Hessian-vector products that is slightly larger than r, which amounts to approximately r forward/adjoint pairs of linearized PDE solves all containing the same PDE operator or its adjoint (this presents an opportunity to employ an effective PDE preconditioner, since it will be amortized over r PDE forward/adjoint solves.) Once the low-rank approximation has been constructed, the product of this approximate Hessian with a vector can then be formed by successively applying $\boldsymbol{V}_r$ and $\boldsymbol{V}_r^{\diamond}$ to vectors, each application amounting to r inner products. The cost of this linear algebra is negligible relative to the PDE solves needed to form the low-rank approximation.

Moreover, using the Sherman–Morrison–Woodbury formula [25] in combination with expressing the prior-preconditioned data misfit Hessian as the sum of a low-rank term and a remainder, we can write the inverse Hessian as

$$(3.10) \qquad (\boldsymbol{L}^*\boldsymbol{H}_{\mathrm{misfit}}\boldsymbol{L} + \boldsymbol{I})^{-1} = \boldsymbol{I} - \boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^\diamond + \mathcal{O}\left(\sum_{i=r+1}^n \frac{\lambda_i}{\lambda_i + 1}\right),$$

where $\boldsymbol{D}_r := \mathrm{diag}(\lambda_1/(\lambda_1+1),\ldots,\lambda_r/(\lambda_r+1)) \in \mathbb{R}^{r\times r}$. As can be seen from the form of the remainder term above, to obtain an accurate low-rank approximation of $\boldsymbol{H}^{-1}$, we can neglect eigenvectors corresponding to eigenvalues that are small compared to 1. Therefore,

$$(3.11) \quad \boldsymbol{H}^{-1}\boldsymbol{g} \approx \boldsymbol{L}(\boldsymbol{I} - \boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^\diamond)\boldsymbol{L}^*\boldsymbol{g} = \boldsymbol{L}\big\{\boldsymbol{V}_r\big[(\boldsymbol{\Lambda}_r + \boldsymbol{I}_r)^{-1} - \boldsymbol{I}_r\big]\boldsymbol{V}_r^\diamond + \boldsymbol{I}\big\}\boldsymbol{L}^*\boldsymbol{g}.$$

The expression on the right side of (3.11) can be used to efficiently apply the square-root inverse Hessian to a vector $\boldsymbol{x}$, as needed for drawing samples from a Gaussian distribution with covariance $\boldsymbol{H}^{-1}$. Namely,

$$(3.12) \qquad \boldsymbol{H}^{-1/2}\boldsymbol{x} \approx \boldsymbol{L}\big\{\boldsymbol{V}_r\big[(\boldsymbol{\Lambda}_r + \boldsymbol{I}_r)^{-1/2} - \boldsymbol{I}_r\big]\boldsymbol{V}_r^\diamond + \boldsymbol{I}\big\}\boldsymbol{x}.$$

By a direct computation using the adjoint definitions (2.4) and (2.5), it can be verified that $\boldsymbol{H}^{-1}\boldsymbol{x} = \boldsymbol{H}^{-1/2}(\boldsymbol{H}^{-1/2})^*\boldsymbol{x}$. Finally, the determinant of the square-root Hessian can be computed efficiently from

$$(3.13) \qquad \det(\boldsymbol{H}^{1/2}) = (\det \boldsymbol{L})^{-1}\prod_{i=1}^r(\lambda_i + 1)^{1/2}.$$

In summary, once the low-rank approximation of the data misfit Hessian has been constructed, all of the operations with the Hessian described above (and required by the three Hessian-based methods) can be carried out using only inner products and vector sums, without recourse to PDE solves. These linear algebra operations are negligible relative to the PDE solves needed for the low-rank approximation, and thus the dominant cost of these methods is $O(r)$ forward/adjoint PDE solves needed for the low-rank approximation. As mentioned above, for ill-posed inverse problems (including the ice sheet flow inverse problem studied below), the prior-preconditioned data misfit Hessian is a compact operator with rapidly decaying eigenvalues, so that $r \ll n$. Moreover, when the dominant eigenvectors of the prior-preconditioned data misfit Hessian are spatially smooth, r is independent of the parameter dimension n and the observation dimension q.

3.6. Comparison of computational cost of ISMAP, SNMAP, and SN. The stochastic Newton MCMC methods and the independence sampler with MAP-based Gaussian all use the low-rank approximation and fast operations with the Hessian described in the previous section. However, they differ markedly in how frequently they recompute the low-rank approximation of the prior-preconditioned data misfit Hessian, which as mentioned above is by far the dominant cost relative to the linear algebra.

Let us now characterize the cost per MCMC sample for each of the three Hessian-based methods described in section 3, measured in number of (forward or adjoint) PDE solves. The independence sampling method (ISMAP) requires just a single evaluation of the parameter-to-observable map per sample, which amounts to a single (nonlinear)

forward PDE solve per sample. The stochastic Newton MCMC method with dynamically changing Hessian (SN) requires for each sample a nonlinear forward PDE solve, a (linear) adjoint PDE solve for the gradient computation, and approximately $2r$ linearized PDE solves to construct the rank-r approximation of the prior-preconditioned data misfit Hessian. Finally, the cost per sample for stochastic Newton MCMC with MAP-based Hessian (SNMAP) is one nonlinear forward PDE solve and one adjoint PDE solve, since SNMAP recomputes the gradient at each sample point. Depending on whether the forward problem is linear or nonlinear and stationary or time dependent, and depending on whether the linearized PDEs are solved by direct factorization (which permits reuse of the factors within the low-rank approximation) or iteratively (which permits reuse of only the preconditioner), the number of PDE solves per sample translates differently into computational time per sample. Thus, the metric we use to compare the performance of these three Hessian-based methods to each other in section 5 is the number of linearized PDE solves required by each method.

4. Application to the inversion of basal boundary conditions in ice flow problems. In the remainder of this paper, we apply the methods discussed in section 3 to an inverse problem in ice dynamics, in which we seek to find a statistical description of the uncertain basal sliding coefficient field from pointwise velocity observations at the surface of the moving mass of ice. In this section, we summarize the physics describing the dynamics of ice flows, present the two-dimensional problem used to exercise our methods, and the prior distribution and the likelihood for the Bayesian inverse problem. We also give expressions of the gradient and the Hessian-vector product of the negative log posterior function using adjoint ice flow equations and describe the discretization of these equations.

4.1. The dynamics of ice flow. We model the flow of ice as a non-Newtonian, viscous, incompressible, isothermal fluid [26, 32, 36, 43]. The balance of mass and linear momentum in a domain $\Omega \subset \mathbb{R}^d$ of dimension $d = 2$ or $d = 3$ state that

$$\text{(4.1a)} \qquad \boldsymbol{\nabla} \cdot \boldsymbol{u} = 0 \quad \text{in } \Omega,$$

$$\text{(4.1b)} \qquad -\boldsymbol{\nabla} \cdot \boldsymbol{\sigma_u} = \rho \boldsymbol{g} \quad \text{in } \Omega,$$

where $\boldsymbol{u}$ denotes the velocity vector, $\boldsymbol{\sigma_u}$ the stress tensor, ρ the density of the ice, and $\boldsymbol{g}$ gravity. The stress, $\boldsymbol{\sigma_u}$, can be decomposed as $\boldsymbol{\sigma_u} = \boldsymbol{\tau_u} - \boldsymbol{I}p$, where $\boldsymbol{\tau_u}$ is the deviatoric stress tensor, p the pressure, and $\boldsymbol{I}$ the unit tensor. We employ a constitutive law for ice that relates stress and strain rate tensors by Glen's flow law [23],

$$\text{(4.1c)} \qquad \boldsymbol{\tau_u} = 2\eta(\boldsymbol{u})\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}, \text{ with } \eta(\boldsymbol{u}) = \frac{1}{2}A^{-\frac{1}{n}}\,\dot{\varepsilon}_{\mathrm{II}}^{\frac{1-n}{2n}},$$

where η is the effective viscosity, $\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} = \frac{1}{2}(\boldsymbol{\nabla}\boldsymbol{u} + \boldsymbol{\nabla}\boldsymbol{u}^T)$ the strain rate tensor, $\dot{\varepsilon}_{\mathrm{II}} = \frac{1}{2}\mathrm{tr}(\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}^2)$ its second invariant, $n \geq 1$ Glen's flow law exponent, and A the temperature-dependent flow rate factor (here taken as constant in isothermal ice).

At the base Γ_{b} of the ice sheet, one commonly assumes nonpenetrating normal boundary conditions and a linear sliding law for the tangential components, [43] i.e.,

$$\text{(4.1d)} \qquad \boldsymbol{u} \cdot \boldsymbol{n} = 0, \;\; \boldsymbol{T}\boldsymbol{\sigma_u}\boldsymbol{n} + \exp(\beta)\boldsymbol{T}\boldsymbol{u} = \boldsymbol{0},$$

where $\beta = \beta(\boldsymbol{x})$ is the log basal sliding coefficient field, and $\boldsymbol{T} := \boldsymbol{I} - \boldsymbol{n} \otimes \boldsymbol{n}$ the projection onto the tangential plane. Here, "$\otimes$" represents the tensor (or outer)
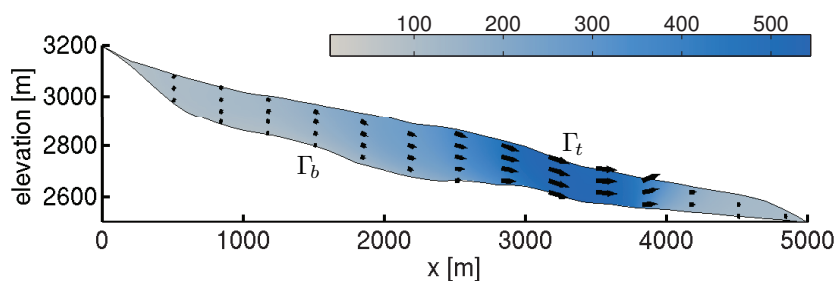
FIG. 2. *The longitudinal profile of Haut Glacier d'Arolla from the ISMIP-HOM benchmark collection* [44]. *This profile follows a flowline of* 5 *km length with a grid spacing of* 100 *m. The arrows represent the flow field obtained by solving* (4.1) *with the basal sliding coefficient field given by* (4.2).

product defined by $(\boldsymbol{a} \otimes \boldsymbol{b})\boldsymbol{c} = \boldsymbol{a}\boldsymbol{b}{\cdot}\boldsymbol{c}$, $\boldsymbol{n}$ is the outward normal vector, and $\boldsymbol{I}$ is the second order unit tensor. Together with appropriate boundary conditions on $\partial\Omega \setminus \Gamma_b$, (4.1) represents an accepted model for the flow of ice sheets and glaciers. Note that the Robin coefficient field $\exp(\beta)$, which relates tangential velocity to tangential traction, subsumes several complex physical phenomena such as the frictional behavior of the ice sheet, the roughness of the bedrock, and hydrological phenomena. It does not itself represent a physical parameter and is highly uncertain. Our target is to infer the log sliding coefficient field β, which in the following we simply refer to as sliding coefficient field, within a Bayesian inversion approach. In the next section, we specify the ice flow model problem used to study the efficiency of our algorithms and to interpret results of Bayesian inversions.

4.2. The Arolla test problem. We use a two-dimensional test problem taken from the Ice Sheet Model Intercomparison Project for Higher-Order Ice Sheet Models (ISMIP-HOM) benchmark study [44]. The domain Ω, which is based on data from the Haut Glacier d'Arolla, is shown in Figure 2. Together with the basal boundary condition (4.1d), on the top boundary Γ_t we assume the traction-free condition

$$(4.1e) \qquad\qquad \boldsymbol{\sigma_u}\boldsymbol{n} = \boldsymbol{0} \text{ on } \Gamma_t.$$

The driving force in the Stokes equations (4.1) is the gravity $\rho\boldsymbol{g} = (0, -\rho g \cos\theta)$, where $\rho = 910$ kg/m^3 is the ice density, and $g = 9.81$ m/s^2 is the gravitational constant. The Glen's flow-law exponent parameter is $n = 3$, and the rate factor is assumed constant as $A = 10^{-16}$ Pa^{-n}a^{-1}, where "Pa" and "a" are units of Pascals and years, respectively [44].

As reference basal sliding coefficient field, which is also used to generate synthetic observations as described in the next section, we choose

$$(4.2) \qquad \beta_{\text{true}}(x) = \ln \begin{cases} 1000 + 1000 \sin\left(\frac{2\pi x}{5000}\right) & \text{if } 0 \leq x < 3750, \\ 1000\left(16 - \frac{x}{250}\right) & \text{if } 3750 \leq x < 4000, \\ 1000 & \text{if } 4000 \leq x < 5000. \end{cases}$$

The flow field corresponding to the basal sliding coefficient field (4.2) is shown in Figure 2.

4.3. The likelihood. The likelihood function expresses the probability that a candidate set of parameters reproduces the observations $\boldsymbol{d}^{\text{obs}}$. To specify the likelihood

function, we denote by $\boldsymbol{u}(\beta)$ the solution of the Stokes equation with basal sliding coefficient field $\beta(\boldsymbol{x})$, and by $\mathcal{B}$ the observation operator, which restricts the flow solution to ten measurement points on the right of the top surface Γ_t, i.e., lower part of the glacier, with x-coordinates uniformly distributed in $[2500, 5000]$. Thus, the parameter-to-observable map is $\boldsymbol{f}(\beta) = \mathcal{B}\boldsymbol{u}(\beta)$.

The observations $\boldsymbol{d}^{\mathrm{obs}}$ are synthetically generated by solving the ice flow Stokes equations (4.1) with basal sliding coefficient field β_{true} as specified in (4.2), restricting the resulting flow solution $\boldsymbol{u}(\beta_{\mathrm{true}})$ using the observation operator and adding additive Gaussian noise $\boldsymbol{\epsilon}$, i.e., $\boldsymbol{d}^{\mathrm{obs}} = \mathcal{B}\boldsymbol{u}(\beta_{\mathrm{true}}) + \boldsymbol{\epsilon}$. Each component of the noise vector $\boldsymbol{\epsilon}$ is independent and identically distributed (i.i.d.) with standard deviation $\bar{\sigma}_{\mathrm{noise}}$ for the horizontal flow components and with $\bar{\bar{\sigma}}_{\mathrm{noise}}$ for the vertical flow components. Adding noise mitigates the "inverse crime," which occurs when synthetic observations are used in an inversion and the same numerical method is employed both in the synthetization of the observations and in the inverse problem solution [33]. The likelihood function is then given by

$$(4.3) \qquad \pi_{\mathrm{like}}(\boldsymbol{d}^{\mathrm{obs}}|\boldsymbol{\beta}) \propto \exp\left[-\frac{1}{2}\left\|\mathcal{B}\boldsymbol{u}(\beta) - \boldsymbol{d}^{\mathrm{obs}}\right\|^2_{\boldsymbol{\Gamma}^{-1}_{\mathrm{noise}}}\right],$$

where the noise covariance matrix $\boldsymbol{\Gamma}_{\mathrm{noise}}$ is diagonal with the entries $\bar{\sigma}_{\mathrm{noise}}$ and $\bar{\bar{\sigma}}_{\mathrm{noise}}$ for the horizontal and vertical components, respectively. To understand the effect of the noise level on the performance of the three Hessian-based sampling methods and on the uncertainty in the reconstruction, we consider two problems based on the noise level in the observations:

- **Problem 1.** $\bar{\sigma}_{\mathrm{noise}} = 62$ for the horizontal flow components and with $\bar{\bar{\sigma}}_{\mathrm{noise}} = 10$ for the vertical flow components;
- **Problem 2.** $\bar{\sigma}_{\mathrm{noise}} = 18$ for the horizontal flow components and with $\bar{\bar{\sigma}}_{\mathrm{noise}} = 3$ for the vertical flow components.

4.4. The choice of prior. We specify the Gaussian prior by giving its mean β_0 and its covariance via the elliptic operator $\mathcal{A}$ discussed in section 2. Since the bottom surface of the Arolla geometry is a "curved" surface, the prior is defined in terms of the surface Laplacian (also called the Laplace–Beltrami operator). Using the projection $\boldsymbol{T}$ onto the tangential plane as defined above, $\boldsymbol{\nabla}_{\Gamma_b} = \boldsymbol{T}\boldsymbol{\nabla}$ is the tangential gradient, $\boldsymbol{\nabla}_{\Gamma_b}\cdot$ is the tangential divergence, and $\boldsymbol{\nabla}_{\Gamma_b}\cdot\boldsymbol{\nabla}_{\Gamma_b}$ is the Laplace–Beltrami operator [3, 15, 17]. Thus, we define $\mathcal{A}$ as the differential operator

$$(4.4\mathrm{a}) \qquad -\boldsymbol{\nabla}_{\Gamma_b} \cdot (a\boldsymbol{\nabla}_{\Gamma_b}\beta) + b\beta = s \quad \text{in } \Gamma_b,$$
$$(4.4\mathrm{b}) \qquad (a\boldsymbol{\nabla}_{\Gamma_b}\beta) \cdot \boldsymbol{\nu} = 0 \quad \text{on } \partial\Gamma_b,$$

where $\boldsymbol{\nu}$ denotes the outward unit normal on $\partial\Gamma_b$. The finite-dimensional representation of the prior inverse is $\boldsymbol{\Gamma}^{-1}_{\mathrm{prior}} = \boldsymbol{M}^{-1}\boldsymbol{K}$, where $\boldsymbol{M}$ and $\boldsymbol{K}$ are the corresponding surface mass and surface stiffness matrices, respectively. In our model problems, we use the parameters $a = 10^{-2}$ and $b = 10^2$. With these parameters, the standard deviation of the Green's function corresponding to the prior (i.e., the correlation length) is roughly 5% of the total length of the glacier.

4.5. Gradient and Hessian of the negative log posterior. The Hessian-based sampling methods presented in section 3 rely on the availability of gradients and Hessian-vector products of the negative log posterior. The derivation of these derivatives is complicated by the fact that the parameter-to-observable map involves

the solution of the ice flow equations. In this section, we give expressions for the efficient computation of gradients and Hessian-vector products using adjoint equations. For a more detailed presentation of derivative computation using adjoints, we refer the reader to the PDE-constrained optimization monographs [4, 31, 57] and to [45] for the ice flow dynamics setting.

The gradient of the negative log posterior can be found by requiring that variations of a Lagrangian function with respect to the forward velocity and pressure $(\boldsymbol{u}, p)$ and an adjoint velocity and pressure $(\boldsymbol{v}, q)$ vanish. Variations with respect to β then result in the following strong form of the gradient $\mathcal{G}$:

$$(4.5) \qquad \mathcal{G}(\beta) := \exp(\beta)\boldsymbol{Tu} \cdot \boldsymbol{Tv} + \mathcal{A}(\beta - \beta_0).$$

Here, the velocity $\boldsymbol{u}$ is obtained by solving the *forward Stokes problem* (4.1) for given β, and the adjoint velocity $\boldsymbol{v}$ is obtained by solving the following *adjoint Stokes problem* for given β and for $\boldsymbol{u}$ satisfying (4.1):

$$(4.6a) \qquad\qquad\qquad \boldsymbol{\nabla} \cdot \boldsymbol{v} = 0 \qquad\qquad \text{in } \Omega,$$

$$(4.6b) \qquad\qquad\qquad -\boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_v = \boldsymbol{0} \qquad\qquad \text{in } \Omega,$$

$$(4.6c) \qquad\qquad\qquad \boldsymbol{\sigma}_v \boldsymbol{n} = -\mathcal{B}^* \Gamma_{\text{noise}}^{-1} (\mathcal{B}\boldsymbol{u} - \boldsymbol{d}^{\text{obs}}) \qquad \text{on } \Gamma_t,$$

$$(4.6d) \qquad \boldsymbol{v} \cdot \boldsymbol{n} = 0, \ \boldsymbol{T}\boldsymbol{\sigma}_v\boldsymbol{n} + \exp(\beta)\boldsymbol{Tv} = \boldsymbol{0} \qquad \text{on } \Gamma_b,$$

where the adjoint stress $\boldsymbol{\sigma}_v$ is given by

$$\boldsymbol{\sigma}_v := 2\eta(\boldsymbol{u}) \left(\mathsf{I} + \frac{1-n}{n} \frac{\dot{\boldsymbol{\varepsilon}}_u \otimes \dot{\boldsymbol{\varepsilon}}_u}{\dot{\boldsymbol{\varepsilon}}_u : \dot{\boldsymbol{\varepsilon}}_u} \right) \dot{\boldsymbol{\varepsilon}}_v - \boldsymbol{I}q,$$

and I is the fourth order identity tensor.

The action of the Hessian operator evaluated at a sliding coefficient field β onto a direction $\hat{\beta}$ is given by

$$(4.7) \qquad \mathcal{H}(\beta)(\hat{\beta}) := \mathcal{A}\hat{\beta} + \exp(\beta)(\hat{\beta}\boldsymbol{Tu} \cdot \boldsymbol{Tv} + \boldsymbol{T\hat{u}} \cdot \boldsymbol{Tv} + \boldsymbol{Tu} \cdot \boldsymbol{T\hat{v}}),$$

where the *incremental forward velocity/pressure* $(\hat{\boldsymbol{u}}, \hat{p})$ satisfy the *incremental forward Stokes problem*,

$$(4.8a) \qquad\qquad\qquad \boldsymbol{\nabla} \cdot \hat{\boldsymbol{u}} = 0 \qquad\qquad \text{in } \Omega,$$

$$(4.8b) \qquad\qquad\qquad -\boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_{\hat{u}} = \boldsymbol{0} \qquad\qquad \text{in } \Omega,$$

$$(4.8c) \qquad\qquad\qquad \boldsymbol{\sigma}_{\hat{u}} \boldsymbol{n} = \boldsymbol{0} \qquad\qquad \text{on } \Gamma_t,$$

$$(4.8d) \qquad \hat{\boldsymbol{u}} \cdot \boldsymbol{n} = 0, \ \boldsymbol{T}\boldsymbol{\sigma}_{\hat{u}}\boldsymbol{n} + \exp(\beta)\boldsymbol{T\hat{u}} = -\hat{\beta}\exp(\beta)\boldsymbol{Tu} \qquad \text{on } \Gamma_b,$$

with $\boldsymbol{\sigma}_{\hat{u}} := 2\eta(\boldsymbol{u}) \left(\mathsf{I} + \frac{1-n}{n} \frac{\dot{\boldsymbol{\varepsilon}}_u \otimes \dot{\boldsymbol{\varepsilon}}_u}{\dot{\boldsymbol{\varepsilon}}_u : \dot{\boldsymbol{\varepsilon}}_u} \right) \dot{\boldsymbol{\varepsilon}}_{\hat{u}} - \boldsymbol{I}\hat{p}$, and the *incremental adjoint velocity/pressure* $(\hat{\boldsymbol{v}}, \hat{q})$ satisfy the *incremental adjoint Stokes problem*,

$$(4.9a) \qquad\qquad\qquad \boldsymbol{\nabla} \cdot \hat{\boldsymbol{v}} = 0 \qquad\qquad \text{in } \Omega,$$

$$(4.9b) \qquad\qquad\qquad -\boldsymbol{\nabla} \cdot \boldsymbol{\sigma}_{\hat{v}} = -\boldsymbol{\nabla} \cdot \tau_{\hat{u}} \qquad\qquad \text{in } \Omega,$$

$$(4.9c) \qquad\qquad\qquad \boldsymbol{\sigma}_{\hat{v}} \boldsymbol{n} = -\mathcal{B}^* \Gamma_{\text{noise}}^{-1} \mathcal{B}\hat{\boldsymbol{u}} - \tau_{\hat{u}} \qquad \text{on } \Gamma_t,$$

$$(4.9d) \qquad \hat{\boldsymbol{v}} \cdot \boldsymbol{n} = 0, \ \boldsymbol{T}\boldsymbol{\sigma}_{\hat{v}}\boldsymbol{n} + \exp(\beta)\boldsymbol{T\hat{v}} = -\boldsymbol{T}\tau_{\hat{u}}\boldsymbol{n} \qquad \text{on } \Gamma_b,$$

with $\boldsymbol{\sigma_{\hat{v}}} := 2\eta(\boldsymbol{u})\left(\mathsf{I} + \frac{1-n}{n}\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\otimes\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}:\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}\right)\dot{\boldsymbol{\varepsilon}}_{\hat{v}} - \boldsymbol{I}\hat{q}$, and $\tau_{\hat{u}} = 2\eta(\boldsymbol{u})\boldsymbol{\Psi}\dot{\boldsymbol{\varepsilon}}_{\hat{u}}$, where

$$\boldsymbol{\Psi} = \left(1 + \frac{1-n}{n}\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}} : \dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\right)\mathsf{I} + \frac{1-n}{n}\left[\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\otimes\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}:\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}} + 2\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\otimes\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{v}}}{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}:\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}} + \frac{1-3n}{n}\frac{\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}\otimes\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}}{(\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}:\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}})^2}\right].$$

In these expressions, $\dot{\boldsymbol{\varepsilon}}_{\hat{u}}$ and $\dot{\boldsymbol{\varepsilon}}_{\hat{v}}$ are defined analogously to $\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{u}}$ and $\dot{\boldsymbol{\varepsilon}}_{\boldsymbol{v}}$.

To summarize, the computational cost (measured in the number of linearized Stokes solves, which represent the dominant cost) of the gradient evaluation is n_{ls} forward linearized Stokes solves for the nonlinear forward problem (4.1) (where n_{ls} is the number of Newton iterations required by the nonlinear solver to converge), and one linear adjoint solve for (4.6). Each computation of the Hessian-vector product (4.7) requires two linearized Stokes solves, namely, the solutions of (4.8) and (4.9).

4.6. Discretization and solvers. We discretize the domain Ω with 260 triangular mesh elements and use Taylor–Hood finite elements (i.e., linear elements for pressure and quadratic elements for the velocity components, which leads to 4714 degrees of freedom for the velocity field and 659 for the pressure) for the forward and adjoint Stokes problems as well as their incremental counterparts. The uncertain sliding coefficient field β is discretized using linear elements with 139 unknowns, i.e., parameters for the inverse problem. We ensure that the state and parameter fields are sufficiently resolved by comparing the solutions computed on different meshes. All Stokes systems are solved using a direct factorization method. The cost of the Stokes matrix factorization is amortized across the adjoint solve and the incremental forward and adjoint solves in all CG iterations needed in each Newton iteration; when the factorization is available, only triangular solutions are required at each CG iteration, gradient computation, or the application of the Hessian to a vector.

5. Performance of algorithms. The primary goal of this section is to compare the performance of the sampling methods presented in section 3 for the Bayesian inverse problem described in section 4. We start with a discussion on the computation of the MAP point in section 5.1 and study the approximation of prior-preconditioned data misfit Hessians—and thus covariance matrices—using low-rank ideas (section 5.2). In section 5.3, we present a systematic comparison of the three Hessian-based sampling methods (ISMAP, SN, and SNMAP) presented in section 3.

5.1. Computation of the MAP point. For the computation of the MAP point, we apply an adjoint-based inexact Newton method to solve the nonlinear least-squares optimization problem (2.9). Starting with an initial guess for the basal sliding coefficient field β (we use the prior mean $\beta_0 \equiv \ln(1000)$), Newton's method iteratively updates this parameter based on successive quadratic approximations of the negative log posterior functional $J(\cdot)$, using the expressions for the first and second derivatives presented in section 4.5. Since the CG method is used to solve the Newton linearization, the method does not require assembled Hessian matrices but only Hessian-vector products. For a more complete presentation of this optimization method to compute MAP points for ice sheet model problems, we refer the reader to [45].

We discuss the performance of the optimization algorithm for the computation of the MAP point for Problem 2 as defined in section 4.3. On the right in Figure 3, we show the "truth" sliding coefficient field, which is used to generate the synthetic surface velocity observations. Also shown is the MAP point, i.e., the solution of (2.9). In the upper part of the glacier the MAP point follows the prior mean since observations are available only in the lower half of the glacier (i.e., the right part of the domain).
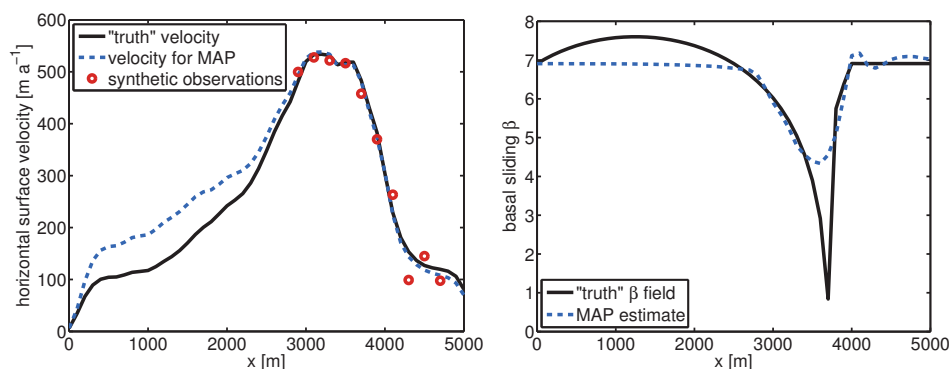
FIG. 3. *Left: The horizontal surface velocity obtained by solving the forward problem using the "truth" sliding coefficient field (solid line) and the synthetic pointwise observations (circles), generated by adding 1.5% Gaussian random noise to this surface velocity. The horizontal velocity corresponding to the MAP point is shown by the dashed line. Right: "Truth" sliding coefficient field (solid line) and MAP point (dashed line).*

To compute the MAP point, 8 (outer) Newton iterations were necessary to decrease the nonlinear residual by a factor of 10^5. In each of these outer Newton iterations, the nonlinear Stokes equation has to be solved, for which we use an (inner) Newton method. These inner Newton solves are also terminated after the residual is decreased by a factor of 10^5, which takes an average of 12 iterations, each amounting to a linearized Stokes solve. In addition to the nonlinear Stokes solve, each (outer) Newton iteration requires computation of the gradient and of several Hessian-vector products. Summing over all 8 (outer) Newton iterations, 32 conjugate gradient iterations—and thus 32 Hessian-vector products—are required. In total, the computation of the MAP point amounts to 208 linear(ized) Stokes solves.

5.2. Low-rank approximation of the prior-preconditioned data misfit Hessian. The computational feasibility of Hessian-based sampling for large-scale Bayesian inverse problems critically relies on low-rank approximations for the data misfit Hessian. Thus, we study the numerical rank of the prior-preconditioned data misfit Hessian for various points in the parameter space. Figure 4 shows a logarithmic plot of the spectra of the prior-preconditioned data misfit Hessians at the 21 MCMC chain starting points discussed in section 5.3. Note that all spectra decay rapidly. As seen in (3.10), an accurate low-rank approximation of the inverse Hessian can be obtained by neglecting eigenvalues that are small compared to 1. Thus, retaining 15–20 eigenvectors appears to be sufficient for any point from the posterior distribution.

In our sampling runs, we thus use $r = 20$ eigenvectors for the low-rank approximation of the prior-preconditioned data misfit Hessian. We note that the cost of obtaining this low-rank approximation, measured in the number of Stokes solves, is $2(r+l)$, where $r+l$ is the number of Lanczos iterations. Here, $l \geq 0$ iterations are used to ensure the accurate computation of the most significant eigenvalues/eigenvectors (we use $l = 5$). We discard any negative eigenvalues to guarantee that the low-rank approximation is positive semidefinite.

5.3. Performance of proposed stochastic Newton MCMC method with MAP-based Hessian. In this section, we compare the performance of the proposed SNMAP method (stochastic Newton MCMC with MAP-based Hessian) with SN (the
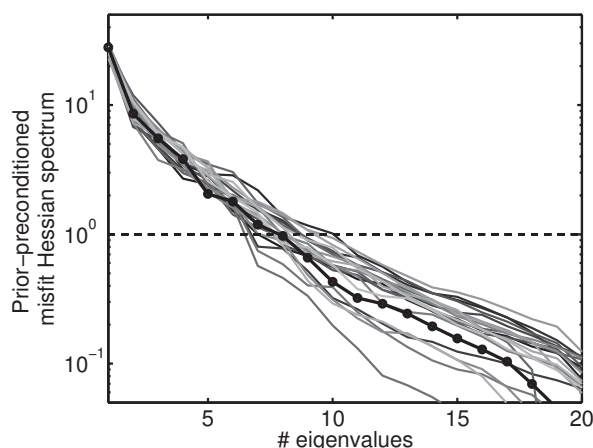
FIG. 4. *Logarithmic plot of the spectra of prior-preconditioned data misfit Hessians computed at the MAP point (black line with dots) and at* 21 *points distributed over the support of the posterior (gray lines). The horizontal line for* $\lambda = 1$ *shows the reference value for the truncation of the spectrum of the prior-preconditioned data misfit Hessian.*

original stochastic Newton MCMC method with dynamically computed Hessian) and with ISMAP (independence sampler with MAP-based Gaussian) for both ice flow inverse problems introduced in section 4.3. For each method, 21 MCMC chains are computed using a common set of 21 initial points. These points are selected from an MCMC chain with 25,000 samples initialized at the MAP point. From this chain, these 21 initial points are chosen to approximately maximize the minimum pairwise distances between points, so that the resulting set is distributed quasi-uniformly over the support of the posterior distribution. This ensures that the initial points are overdispersed with respect to the posterior, which is important for the convergence diagnostics used to compare the different MCMC methods.

In Table 1, we summarize convergence diagnostics and MCMC chain statistics averaged over 21 chains (excluding the MPSRF, which is a multichain diagnostic). To compare the different MCMC methods, in the second column we provide the multivariate potential scale reduction factor (MPSRF) diagnostic [5]. This diagnostic compares averaged properties of individual sample chains with properties of the pooled sample chain. When these properties are similar, we infer that each of the individual sample chains has converged. The closer the MPSRF is to 1, the better converged the individual sample chains are.

It is well known in Monte Carlo methods that the variance in the estimate decays as $1/N$ when averaging over N i.i.d. samples. However, MCMC samples are not independent, and in general we observe that averaging over N samples from an MCMC chain reduces the variance in the estimate by a factor of only τ/N, where $\tau > 1$ is the integrated autocorrelation time (IAT) [50], given by

$$(5.1) \qquad \tau = 1 + 2 \sum_{s=1}^{\infty} \rho(s).$$

Here, $\rho(s)$ is the usual autocorrelation function for a lag $s > 0$. In practice, $\rho(s)$ is noisy when computed from a finite length sample chain, and thus we estimate τ by the maximum value obtained by truncating the summation in (5.1). The autocorrelation

TABLE 1

*Multivariate potential scale reduction factor (*MPSRF*), integrated autocorrelation time (*IAT*), effective sample sample size (*ESS*), mean squared jump distance (*MSJ*), acceptance rate (*AR*), number of (linearized) Stokes solves per independent sample (*SPIS*), and the average wallclock time per independent sample (*TPIS*). We compare the performance obtained with the independence sampler with MAP-based Gaussian (*ISMAP*), the stochastic Newton MCMC method with MAP-based Hessian (*SNMAP*), and the stochastic Newton MCMC method with dynamic Hessian (*SN*) for two problems with different noise levels (e.g., $\bar{\sigma}_{noise} = 62$ and $\bar{\bar{\sigma}}_{noise} = 10$ for Problem 1, and $\bar{\sigma}_{noise} = 18$ and $\bar{\bar{\sigma}}_{noise} = 3$ for Problem 2). We use 21 MCMC chains, each with 25,000 samples, hence the total number of samples is 525,000. The dimension of the discretized basal sliding coefficient field, i.e., the number of parameters, is 139.*

	MPSRF	IAT	ESS	MSJ	AR (%)	SPIS	TPIS (s)
Problem 1							
ISMAP	1.210	253	2075	1456	41	2783	139
SNMAP	1.001	6	84004	1390	40	72	4
SN	1.073	125	4032	565	17	1375	69
Problem 2							
ISMAP	1.507	435	1207	280	9	4350	218
SNMAP	1.045	80	6563	190	6	960	48
SN	1.348	600	875	64	2	8400	420

is defined for a scalar quantity, and we report in column three the IAT corresponding to the sliding coefficient field at the point $x = 3450$. In the fourth column, we report the effective sample size (ESS) defined as N/τ, the number of independent samples that would be required for the same variance reduction as obtained from the MCMC chain.

The fifth column shows the mean squared jump distance (MSJ), which provides an indication of how well the MCMC chain is mixing. This metric is defined for a single MCMC chain with samples $\boldsymbol{m}_0, \ldots, \boldsymbol{m}_N$ as

$$(5.2) \qquad \text{MSJ} := \frac{1}{N} \sum_{k=0}^{N-1} \|\boldsymbol{m}_{k+1} - \boldsymbol{m}_k\|_{\boldsymbol{M}}^2.$$

In general, a larger mean square jump distance indicates faster mixing of the MCMC chain and tends to result in better chain convergence.

Finally, we address the question of greatest interest with regard to computational efficiency: Given an MCMC algorithm, how much computational work is required to obtain an independent sample? Column seven reports the total number of linearized Stokes solves required to obtain a single independent sample, and column eight reports the total wallclock time for these solves.

We summarize the following observations from Table 1:
- The number of independent samples is about one order of magnitude larger for Problem 1 than for Problem 2, suggesting that the posterior distribution for Problem 2 is more difficult to sample.
- SNMAP leads to the best MPSRF values for both problems, suggesting the fastest convergence with respect to the number of samples. As a consequence, the largest effective sample size is achieved using SNMAP. Note that this holds even though ISMAP yields larger acceptance rates and mean squared jump distances.
- SNMAP also requires the smallest number of forward solves per independent sample. For Problem 1, SNMAP is more efficient than SN by a factor of about

20, and than ISMAP by a factor of almost 40. For Problem 2, SNMAP is more efficient than SN and ISMAP by factors of about 10 and 5, respectively.
- Surprisingly, SN performs worse than SNMAP, even with respect to the number of samples. This is despite the fact that it uses a better local approximation of the posterior. We attribute this to the mismatch in the local Hessians at different points, which increases the asymmetry between the forward and backward proposals $q(\boldsymbol{m}_k, \boldsymbol{y})$ and $q(\boldsymbol{y}, \boldsymbol{m}_k)$, thus increasing the variability of the acceptance probability $\alpha_k(\boldsymbol{y})$ (see Algorithm 1).

We have also applied delayed rejection adaptive metropolis (DRAM) sampling [27] to explore the posterior distribution. We found it to be far from convergence after 1,000,000 samples. We attribute this to the high-dimensional parameter space and the lack of information about the problem structure in the sampling process. In the next section, we focus on visualization and interpretation of the posterior distribution.

6. Analysis and interpretation of the solution of the Bayesian inverse problem. Visualization and interpretation for high-dimensional posterior distributions is a difficult task. In this section, we highlight techniques motivated by the structure of the Bayesian inverse problem to guide our analysis. First, in section 6.1, we present visualizations of the posterior in the physical coordinate basis, which provides intuition about the solution at particular points or regions of the domain. Then, in section 6.2, we shift our perspective to eigenvectors of the posterior covariance (approximated using the Hessian at the MAP point), which can be classified into groups according to their contributions from the observation data and the prior. The qualitative features of each group provide insight into the posterior distribution. Finally, in section 6.3, we visualize one- and two-dimensional marginal distributions of the full posterior distribution with respect to these eigenvectors.

The results discussed in this section are for Problem 2 as defined in section 4.3, i.e., the problem with smaller data noise. The approximation of the posterior pdf is based on samples generated by the SNMAP method, and kernel density estimation is used to visualize the one- and two-dimensional marginal probability density functions.

6.1. Point marginals and samples from the posterior. In Figure 5, we present (one-dimensional) marginals of the prior and posterior distributions with respect to physical points in the domain. We refer to these marginals as point marginals. The probability density for each point marginal is visualized in grayscale along a vertical column at each point, with higher probability density indicated by darker shading. Because each point marginal is computed independently, the point marginal density values at neighboring points are not necessarily related, and thus any spatial correlation structure present in the distribution is neglected by this visualization. For this reason, we overlay a few samples from each distribution to provide some indication of the spatial correlation structure.

This visualization provides some useful observations for our problem. In the unobserved part of the domain (the upper part of the glacier), the point marginals of the posterior are similar to those of the prior; our beliefs about the basal sliding coefficient field in this region are unchanged from the prior. On the contrary, in the region where observation data are available, we find the variance to be decreased significantly (i.e., we are more certain about the sliding coefficient field in this region), and in some regions most of the probability mass is shifted in the posterior compared to the prior; the evidence from the observation data has overwhelmed our prior beliefs in this region. Finally, while spatial correlation structure is difficult to infer from the limited number of overlaid samples, note that the average width of the variations
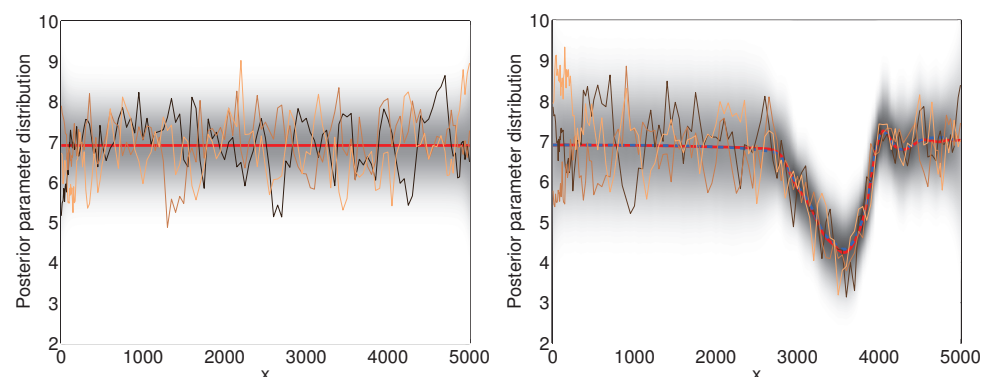
FIG. 5. *Shown in grayscale are the one-dimensional point marginals of the prior (left) and the posterior (right) probability distributions, with higher probability density indicated by darker shading. Point marginals are computed and plotted independently along a vertical line at each point, where the gray shaded area corresponds to a 95% confidence interval. To give an indication of spatial correlation, samples from the prior and the posterior are shown (in different shades of brown). Also shown are the prior and posterior mean (in red), and the MAP point of the posterior (in blue). We recall that the dimension of the discretized basal sliding coefficient field, i.e., the number of parameters, is* 139.

appears unchanged from the prior to the posterior in both, the parts of the glacier with and without observations. We interpret this as insufficient observational evidence to update our beliefs about the width of spatial variations.

6.2. Classification of posterior covariance eigenvectors. In this section, we classify the eigenvectors of the posterior covariance into groups according to their contributions from the observation data and prior, and subsequently use this classification to gain insight into the posterior distribution. While it is common to order eigenvectors by ascending or descending eigenvalues, this choice is poorly adapted to our purposes since it unpredictably interleaves data-influenced eigenvectors with prior-influenced eigenvectors. We therefore propose a general technique for sorting eigenvectors that groups them naturally.

To characterize the influence of the observations and prior on the eigenvectors, consider the Rayleigh quotients of the data misfit Hessian and of the inverse of the prior, i.e.,

$$(6.1) \qquad r_m^i = \frac{\langle \boldsymbol{v}_i, \boldsymbol{H}_{\mathrm{misfit}} \boldsymbol{v}_i \rangle_{\boldsymbol{M}}}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle_{\boldsymbol{M}}}, \quad r_p^i = \frac{\langle \boldsymbol{v}_i, \boldsymbol{\Gamma}_{\mathrm{prior}}^{-1} \boldsymbol{v}_i \rangle_{\boldsymbol{M}}}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle_{\boldsymbol{M}}}$$

for $i = 1, \ldots, n$, where $\boldsymbol{v}_i$ is the ith eigenvector of the inverse posterior covariance (approximated by the Hessian at the MAP point). Because the eigenvalue λ^i associated with $\boldsymbol{v}_i$ is simply the sum of r_m^i and r_p^i, these Rayleigh quotients quantify the contributions from the observation data and prior. We then order the eigenvectors according to the difference of the squared Rayleigh coefficients $d^i := (r_m^i)^2 - (r_p^i)^2$. Large positive values of d^i correspond to eigenvectors that are most informed by the data, whereas large negative values correspond to directions most informed by the prior. We note that there are several reasonable choices for d^i; we find that our choice best groups eigenvectors with similar qualitative features. The sorted Rayleigh quotients for the data misfit Hessian and for the inverse of the prior are presented in the
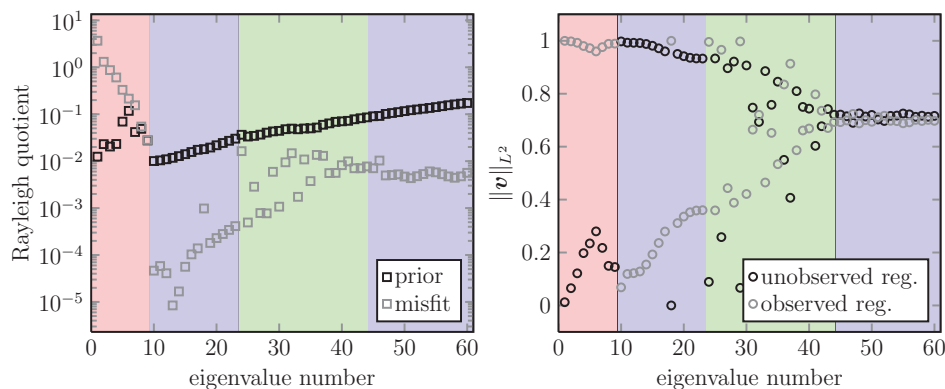
FIG. 6. *Left: Semilogarithmic plot of the Rayleigh quotients of the data misfit Hessian and of the inverse prior covariance as defined in* (6.1). *Right: Norm of the eigenvectors in the lower and upper parts of the glacier, i.e., in the region with and without observations, respectively.*
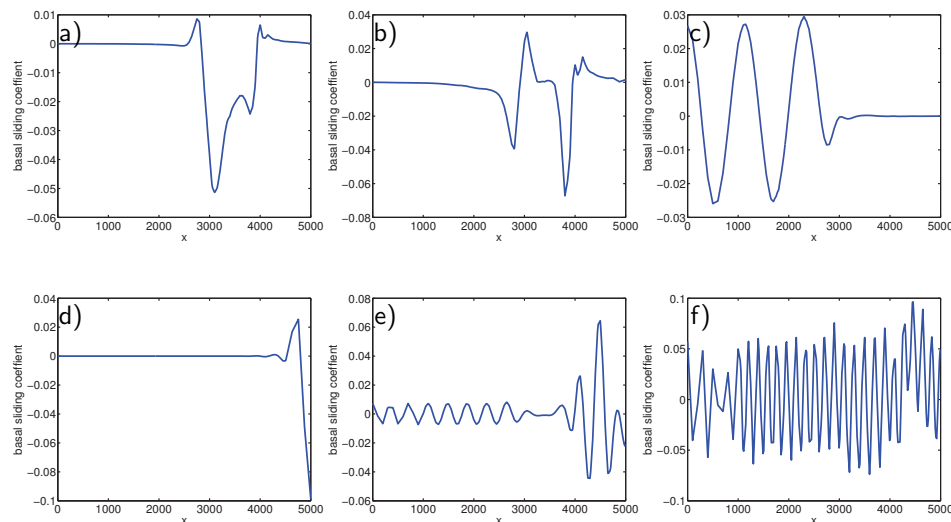

FIG. 7. *Eigenvectors of the Hessian corresponding to the* 1*st,* 3*rd,* 14*th,* 18*th,* 26*th, and* 55*th eigenvalues are shown in Figures (a)–(f), respectively. Note that different eigenvectors are concentrated in different parts of the domain and that eigenvectors corresponding to smaller eigenvalues are more oscillatory.*

left plot in Figure 6. A selection of these eigenvectors is shown in Figure 7.

Next, we study the qualitative features of these eigenvectors. Since the lower half of the glacier contains observation points and the upper half does not, we can also characterize these eigenvectors by determining whether the eigenvector is concentrated primarily in one half of the glacier. The right plot in Figure 6 studies these concentrations in each half of the domain using the corresponding L^2-norms. We can distinguish four groups of eigenvectors, highlighted by different colors in Figure 6, which we discuss next.

Data-informed eigenvectors. The first group (shown in red in Figure 6) contains eigenvectors for which d^i is positive; i.e., the information from the data dominates the information from prior. In the direction of these eigenvectors, the variance in the posterior is significantly reduced due to the observations (recall that the variance is $1/\lambda^i = 1/(r_m^i + r_p^i)$), and hence we say they have been informed by the data.

The eigenvectors in this group are primarily concentrated in the lower half of the glacier, where we have observations (see the right plot in Figure 6). They are relatively smooth (since r_p^i is not large) and qualitatively resemble the first nine Fourier modes in this region (see plots (a) and (b) in Figure 7 for eigenvectors 1 and 3). This last observation is powerful since it provides confidence that features of the MAP point that lie in the span of these first nine Fourier modes are indeed features of the true basal sliding coefficient field.

Shadowed eigenvectors. The next group contains eigenvectors for which the original prior variance was large, and yet the observations provide little information (they are not illuminated by the data, and thus "shadowed"). These eigenvectors are characterized by large ratios r_p^i/r_m^i, and the posterior is easy to characterize in these directions; it is similar to the prior. This group as well as the prior-tail group discussed below are shown in blue in Figure 6. Both groups are well characterized by the prior distribution, although for different reasons.

Shadowed eigenvectors generally concentrate in regions where the parameter-to-observable map is insensitive to the parameter. In our problem, the upper part of the glacier is far away from observation points, and the basal sliding coefficient field at a point only has significant influence on the ice velocity in a neighborhood of that point. Thus, the parameter-to-observable map is insensitive to the sliding coefficient in the upper part of the glacier. In Figure 6 we can see that indeed most of these eigenvectors are concentrated in the upper part of the glacier, and again resemble Fourier modes in the upper half of the glacier (see Figure 7(c)).

Parameter-to-observable map insensitivity also occurs at the very bottom edge of the glacier. Even though observations are available here, the flow in this region is determined primarily by the glacial boundary, preventing the basal sliding coefficient field from significantly influencing the surface velocity in this region. Exactly one shadowed eigenvector corresponds to this region, shown in Figure 7(d).

Mixed eigenvectors. The third group contains eigenvectors for which the observations and the prior both have a significant influence. In general it is not clear how this interaction will affect the posterior distribution, and as such it is perhaps too optimistic to make predictions based on this analysis, and we defer this discussion to section 6.3. Note that these eigenvectors seem to be generally characterized by a mixture of medium frequency Fourier-like modes on the upper and lower halves of the glacier, which is why we refer to them as "mixed" eigenvectors. One eigenvector from this group is shown in Figure 7(e).

Prior-tail eigenvectors. The remaining posterior eigenvectors represent directions in which the prior is very certain (i.e., the prior variance $1/r_p^i$ is small), and for which the observations do not provide sufficient evidence to either contradict or reinforce this assertion (i.e., as in the shadowed eigenvectors, the ratio r_p^i/r_m^i is large). In the continuous inverse problem, this final group contains an infinite number of eigenvectors, each behaving very similar to their prior counterparts, and we therefore refer to this as the "prior-tail." One eigenvector from this prior-tail group, which qualitatively resembles a high frequency Fourier mode, is shown in Figure 7(f).
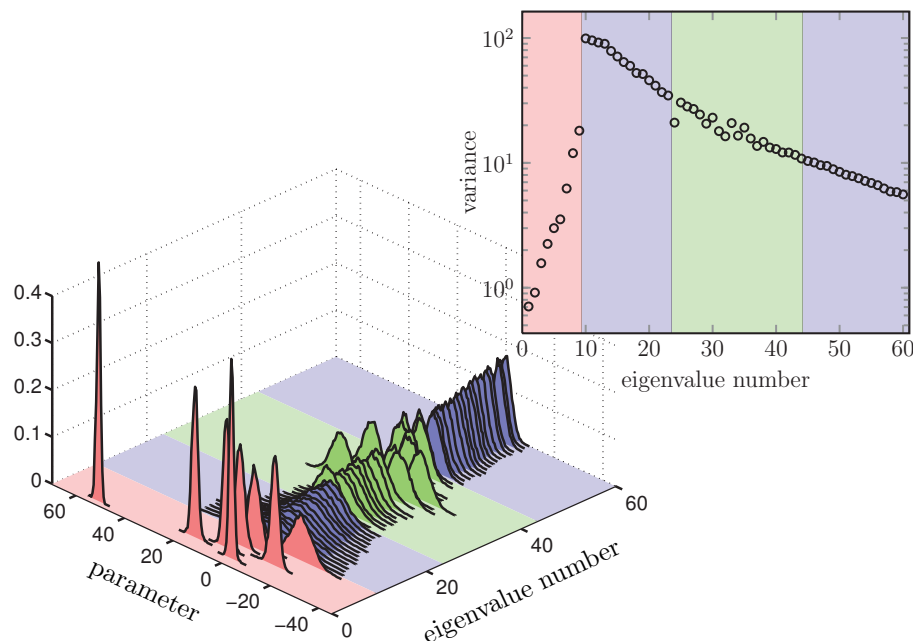
Fig. 8. *Marginals of the posterior distribution with respect to eigenvectors of the covariance (approximated using the Hessian at the MAP point). Shown are the marginals (left) and the corresponding sample variances (right). The eigenvectors are sorted with respect to qualitative features indicated by different background colors as described in the text.*

6.3. Marginals in the eigenvector directions. While the analysis of the previous section provides some insight into the posterior distribution, it has two important limitations. First, this analysis is predicated on the assumption that the posterior is completely characterized by its mean and covariance, and as such, any non-Gaussian behavior of the posterior is obscured. Second, the analysis makes use of the posterior covariance approximated at the MAP point, which may not reflect the behavior of the posterior away from this point. In this section, we make use of the insights gleaned from the above analysis, but return our focus to the full posterior distribution.

In Figure 8, we show the one-dimensional marginals and sample variances of the posterior distribution, with respect to eigenvectors of the posterior covariance using colors corresponding to the eigenvector groups discussed in section 6.2. Many features of these marginals are already anticipated: the data-informed eigenvectors (in red) have small variance and are most shifted with respect to the prior distribution. The shadowed eigenvectors (first blue group) have the largest variances, and the prior-tail eigenvectors (second blue group) have small variance and are essentially unchanged from the prior. To emphasize the departure of the posterior from the prior, all marginals are plotted with respect to the prior mean. Any shift of the marginal away from zero is due to observations.

Despite the nonlinearity of the parameter-to-observable map, we find that the posterior marginals all appear to be near-Gaussian. Since the noise and prior models are both Gaussian, it is reasonable to expect Gaussianity of the data-informed directions in the small-noise limit (the parameter-to-observable map is smooth and
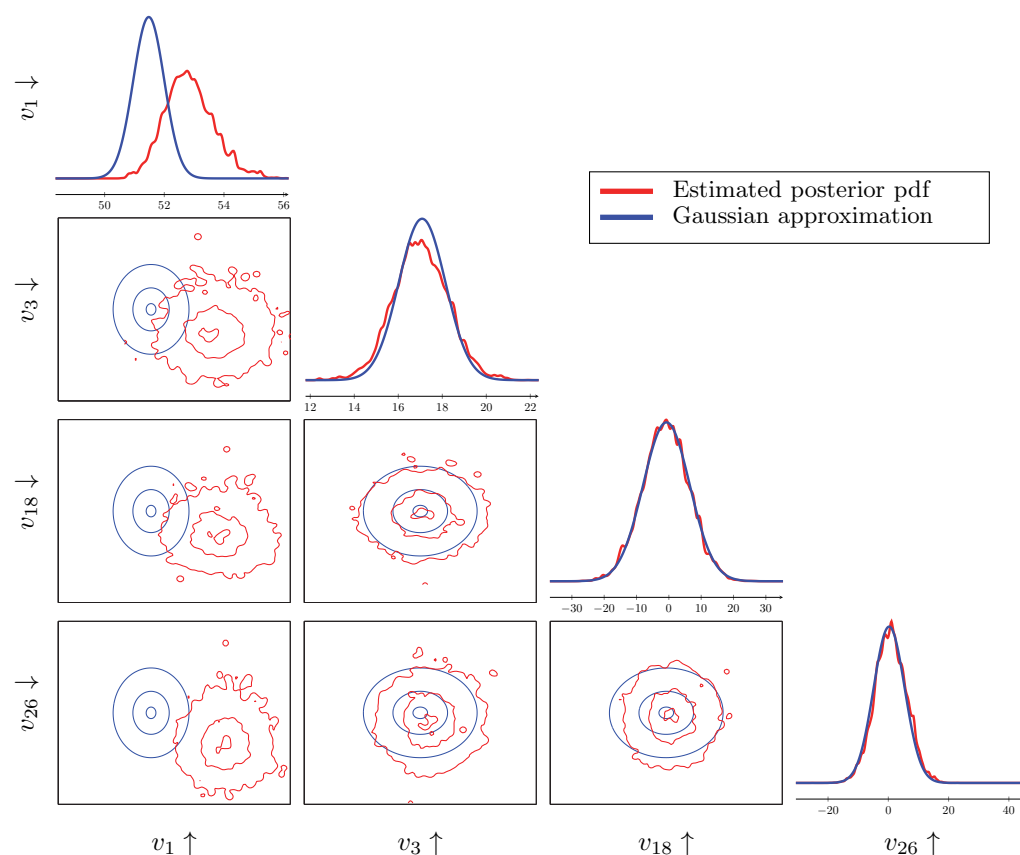
FIG. 9. *One- and two-dimensional marginals from the posterior (red) compared with marginals of the Gaussian approximation at the MAP point (blue). The two-dimensional plots show contour lines of the two-dimensional marginals, where the three contours are selected to contain 5%, 50%, and 95% of the density, respectively. The marginals are computed with respect to the eigenvectors v_1, v_3, v_{18}, and v_{26}, which are plotted in Figure 7(a), (b), (d), and (e), respectively. Kernel desity estimation is used to visualize the posterior pdf using the MCMC sample chain.*

thus nearly linear over a narrow range), and also in the directions where the prior is most influential, as the data do not update the prior distribution in these directions. We therefore anticipate that the most non-Gaussian behavior occurs in the mixed eigenvector directions (in green), as these are the directions with the largest variance (so that the parameter-to-observable map can deviate from a linear approximation) that are significantly influenced by the data.

Figure 9, depicts one- and two-dimensional marginals of the posterior distribution in selected eigenvector directions together with the Gaussian approximation of the posterior distribution at the MAP point. As in Figure 8, these marginals are plotted with respect to the prior mean. In all directions except for the first (the most data-informed eigenvector), we observe that the posterior marginal is close to the Gaussian approximation at the MAP point even in the mixed eigenvector direction (v_{26}). In the direction of the first eigenvector, there is a clear shift in the marginal

mean of the posterior distribution and its Gaussian approximation at the MAP point. Nevertheless, the corresponding posterior marginal looks Gaussian. To give a possible explanation for this behavior, consider a two-dimensional pdf with banana-shaped contours for which the MAP point is located along the banana ridge, but the mean is located at the banana's center of mass, in a region that itself may have low probability density. One-dimensional marginals of such a pdf are likely to have a similar discrepancy between the MAP point and the mean. Although, with respect to the other eigenvectors, the marginals of the posterior and the Gaussian approximation at the MAP point are close, this does not necessarily imply that the posterior is Gaussian.

7. Concluding remarks. We have addressed the problem of constructing efficient MCMC methods for exploring posterior distributions for uncertain parameter fields in infinite-dimensional Bayesian inverse problems governed by expensive forward models. The stochastic Newton MCMC method presented in [37] has been extended in several ways. First, the method is recast in a form that is consistent with the infinite-dimensional setting. In doing so, we have extended the work in [11] to nonlinear inverse problems. Second, the complexity of recomputing the Hessian at each sample point was addressed by investigating a modified stochastic Newton MCMC that reuses the Hessian evaluated at the MAP point.

The modified stochastic Newton MCMC method (with MAP-based Hessian) proposed in this paper is compared with the original stochastic Newton MCMC method (with dynamically changing Hessian) and with an independence sampling method based on a Gaussian proposal at the MAP point for an ice sheet flow inverse problem governed by a nonlinear Stokes equation. A performance comparison reveals that the proposed stochastic Newton MCMC method with a MAP-based Hessian proposal leads to the best convergence, both in terms of the number of samples as well as in terms of the number of PDE solves.

We also presented visualizations and interpretations of the posterior distribution in high dimensions. We showed point marginals of the posterior to provide intuition about the statistical solution at particular points or regions of the domain. The point marginals confirm the dependence of the variance on the availability of observations. We classified the eigenvectors of the covariance of the Gaussian approximation of the posterior at the MAP into groups depending on the extent to which they are influenced by the observational data versus the prior. This classification can be used to identify and exploit directions in parameter space in which the distribution is Gaussian (for directions that are not informed by the data and hence are dominated by a Gaussian prior) or non-Gaussian (for directions that are informed by the data and hence the nonlinearity of the parameter-to-observable map dominates).

REFERENCES

[1] E. B. BECKER, G. F. CAREY, AND J. T. ODEN, *Finite Elements: An Introduction, Vol.* I, Prentice–Hall, Englewood Cliffs, NJ, 1981.

[2] A. BESKOS, G. ROBERTS, AND A. STUART, *Optimal scalings for local Metropolis-Hastings chains on nonproduct targets in high dimensions*, Ann. Appl. Probab., 19 (2009), pp. 863–898.

[3] A. BONITO, R. H. NOCHETTO, AND M. S. PAULETTI, *Geometrically consistent mesh modification*, SIAM J. Numer. Anal., 48 (2010), pp. 1877–1899.

[4] A. BORZÌ AND V. SCHULZ, *Computational Optimization of Systems Governed by Partial Differential Equations*, SIAM, Philadelphia, 2012.

[5] S. Brooks and A. Gelman, *General methods for monitoring convergence of iterative simulations*, J. Comput. Graphical Statist., 7 (1998), pp. 434–456.

[6] T. Bui-Thanh, C. Burstedde, O. Ghattas, J. Martin, G. Stadler, and L. C. Wilcox, *Extreme-scale UQ for Bayesian inverse problems governed by PDEs*, in SC12: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2012.

[7] T. Bui-Thanh and O. Ghattas, *Analysis of the Hessian for inverse scattering problems. Part I: Inverse shape scattering of acoustic waves*, Inverse Problems, 28 (2012), 055001.

[8] T. Bui-Thanh and O. Ghattas, *Analysis of the Hessian for inverse scattering problems. Part II: Inverse medium scattering of acoustic waves*, Inverse Problems, 28 (2012), 055002.

[9] T. Bui-Thanh and O. Ghattas, *Analysis of the Hessian for inverse scattering problems. Part III: Inverse medium scattering of electromagnetic waves in three dimensions*, Inverse Probl. Imaging, 7 (2013), pp. 1139–1155.

[10] T. Bui-Thanh, O. Ghattas, and D. Higdon, *Adaptive Hessian-based nonstationary Gaussian process response surface method for probability density approximation with application to Bayesian solution of large-scale inverse problems*, SIAM J. Sci. Comput., 34 (2012), pp. A2837–A2871.

[11] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler, *A computational framework for infinite-dimensional Bayesian inverse problems Part I: The linearized case, with application to global seismic inversion*, SIAM J. Sci. Comput., 35 (2013), pp. A2494–A2523.

[12] S. L. Cotter, M. Dashti, J. C. Robinson, and A. M. Stuart, *Bayesian inverse problems for functions and applications to fluid mechanics*, Inverse Problems, 25 (2009), 115008.

[13] S. L. Cotter, M. Dashti, and A. M. Stuart, *Approximation of Bayesian inverse problems for PDEs*, SIAM J. Numer. Anal., 48 (2010), pp. 322–345.

[14] T. Cui, C. Fox, and M. J. O'Sullivan, *Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm*, Water Resour. Res., 47 (2011), W10521, doi:10.1029/2010WR010352.

[15] A. Demlow, *Higher-order finite element methods and pointwise error estimates for elliptic problems on surfaces*, SIAM J. Numer. Anal., 47 (2009), pp. 805–827.

[16] P. Dostert, Y. Efendiev, T. Y. Hou, and W. Luo, *Coarse-gradient Langevin algorithms for dynamic data integration and uncertainty quantification*, J. Comput. Phys., 217 (2006), pp. 123–142.

[17] G. Dziuk, *Finite elements for the Beltrami operator on arbitrary surfaces*, in Partial Differential Equations and Calculus of Variations, S. Hildebrandt and R. Leis, eds., Lecture Notes in Math. 1357, Springer, Berlin, Heidelberg, 1988, pp. 142–155.

[18] H. P. Flath, L. C. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas, *Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations*, SIAM J. Sci. Comput., 33 (2011), pp. 407–432.

[19] C. Fox, H. Haario, and J. Andrés Christen, *Inverse problems*, in Bayesian Volume in Honour of Sir Adrian F.M. Smith, Clarendon Press, 2012.

[20] J. Geweke and H. Tanizaki, *On Markov chain Monte Carlo methods for nonlinear and non-Gaussian state-space models*, Comm. Statist. Simulation Comput., 28 (1999), pp. 867–894.

[21] J. Geweke and H. Tanizaki, *Note on the Sampling Distribution for the Metropolis-Hastings Algorithm*, Comm. Statist. Theory Methods, 32 (2003), pp. 775–789.

[22] M. Girolami and B. Calderhead, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, J. Roy. Statist. Soc. Ser. B, 73 (2011), pp. 123–214.

[23] J. W. Glen, *The creep of polycrystalline ice*, Proc. R. Soc. Lond. Ser. A Math. Phys. Sci., 228 (1955), pp. 519–538.

[24] D. N. Goldberg and O. V. Sergienko, *Data assimilation using a hybrid ice flow model*, The Cryosphere, 5 (2011), pp. 315–327.

[25] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 1996.

[26] R. Greve and H. Blatter, *Dynamics of Ice Sheets and Glaciers*, Advances in Geophysical and Environmental Mechanics and Mathematics, Springer, New York, 2009.

[27] H. Haario, M. Laine, A. Miravete, and E. Saksman, *DRAM: Efficient adaptive MCMC*, Stat. Comput., 16 (2006), pp. 339–354.

[28] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.

[29] E. Herbst, *Gradient and Hessian-Based MCMC for DSGE Models*, unpublished manuscript, 2010.

[30] D. Higdon, J. Gattiker, B. Williams, and M. Rightley, *Computer model calibration using*

high-dimensional output, J. Amer. Statist. Assoc., 103 (2008), pp. 570–583.

[31] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Springer, New York, 2009.

[32] K. Hutter, *Theoretical Glaciology*, Mathematical Approaches to Geophysics, D. Reidel, 1983.

[33] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Appl. Math. Sci. 160, Springer-Verlag, New York, 2005.

[34] K. J. H. Law, *Proposals which speed up function-space MCMC*, J. Comput. Appl. Math., 262 (2014), pp. 127–138.

[35] C. Lieberman, K. Willcox, and O. Ghattas, *Parameter and state model reduction for large-scale statistical inverse problems*, SIAM J. Sci. Comput., 32 (2010), pp. 2523–2542.

[36] S. J. Marshall, *Recent advances in understanding ice sheet dynamics*, Earth Planetary Sci. Lett., 240 (2005), pp. 191–204.

[37] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas, *A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion*, SIAM J. Sci. Comput., 34 (2012), pp. A1460–A1487.

[38] Y. M. Marzouk and H. N. Najm, *Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems*, J. Comput. Phys., 228 (2009), pp. 1862–1902.

[39] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), pp. 1087–1092.

[40] M. Morlighem, E. Rignot, H. Seroussi, E. Larour, H. Ben Dhia, and D. Aubry, *Spatial patterns of basal drag inferred using control methods from a full-Stokes and simpler models for Pine Island Glacier, West Antarctica*, Geophys. Res. Lett., 37 (2010), L14502.

[41] D. S. Oliver, L. B. Cunha, and A. C. Reynolds, *Markov chain Monte Carlo methods for conditioning a permeability field to pressure data*, Math. Geo., 29 (1997), pp. 61–91.

[42] D. S. Oliver, A. C. Reynolds, and N. Liu, *Inverse Theory for Petroleum Reservoir Characterization and History Matching*, Cambridge University Press, Cambridge, UK, 2008.

[43] W. S. B. Paterson, *The Physics of Glaciers*, 3rd ed., Butterworth Heinemann, 1994.

[44] F. Pattyn, L. Perichon, A. Aschwanden, B. Breuer, B. de Smedt, O. Gagliardini, G. H. Gudmundsson, R. C. A. Hindmarsh, A. Hubbard, J. V. Johnson, T. Kleiner, Y. Konovalov, C. Martin, A. J. Payne, D. Pollard, S. Price, M. Ruckamp, F. Saito, O. Soucek, S. Sugiyama, and T. Zwinger, *Benchmark experiments for higher-order and full-Stokes ice sheet models (ISMIP-HOM)*, The Cryosphere, 2 (2008), pp. 95–108.

[45] N. Petra, H. Zhu, G. Stadler, T. J. R. Hughes, and O. Ghattas, *An inexact Gauss-Newton method for inversion of basal sliding and rheology parameters in a nonlinear Stokes ice sheet model*, J. Glaciology, 58 (2012), pp. 889–903.

[46] M. R. Pralong and G. H. Gudmundsson, *Bayesian estimation of basal conditions on Rutford ice stream, West Antarctica, from surface data*, J. Glaciology, 57 (2011), pp. 315–324.

[47] S. F. Price, A. J. Payne, I. M. Howat, and B. E. Smith, *Committed sea-level rise for the next century from Greenland ice sheet dynamics during the past decade*, Proc. Natl. Acad. Sci. USA, 108 (2011), pp. 8978–8983.

[48] Y. Qi and T. P. Minka, *Hessian-based Markov chain Monte-Carlo algorithms*, in First Cape Cod Workshop on Monte Carlo Methods, Cape Cod, MA, 2002.

[49] M. J. Raymond and G. H. Gudmundsson, *Estimating basal properties of ice streams from surface measurements: A non-linear Bayesian inverse approach applied to synthetic data*, The Cryosphere, 3 (2009), pp. 265–278.

[50] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer Texts in Statistics, Springer-Verlag, New York, Secaucus, NJ, 2005.

[51] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363.

[52] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Wellesley-Cambridge Press, Wellesley, MA, 1988.

[53] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[54] A. M. Stuart, J. Voss, and P. Wiberg, *Conditional path sampling of SDEs and the Langevin MCMC method*, Commun. Math. Sci., 2 (2004), pp. 685–697.

[55] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, 2005.

[56] L. Tierney, *Markov chains for exploring posterior distributions*, Ann. Statist., 22 (1994), pp. 1701–1762.

[57] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications*, Grad. Stud. Math. 112, AMS, Providence, RI, 2010.