SIAM J. SCI. COMPUT.
Vol. 35, No. 6, pp. A2494–A2523
© 2013 Society for Industrial and Applied Mathematics

A COMPUTATIONAL FRAMEWORK FOR INFINITE-DIMENSIONAL BAYESIAN INVERSE PROBLEMS PART I: THE LINEARIZED CASE, WITH APPLICATION TO GLOBAL SEISMIC INVERSION*

TAN BUI-THANH†, OMAR GHATTAS‡, JAMES MARTIN†, AND GEORG STADLER†

Abstract. We present a computational framework for estimating the uncertainty in the numerical solution of linearized infinite-dimensional statistical inverse problems. We adopt the Bayesian inference formulation: given observational data and their uncertainty, the governing forward problem and its uncertainty, and a prior probability distribution describing uncertainty in the parameter field, find the posterior probability distribution over the parameter field. The prior must be chosen appropriately in order to guarantee well-posedness of the infinite-dimensional inverse problem and facilitate computation of the posterior. Furthermore, straightforward discretizations may not lead to convergent approximations of the infinite-dimensional problem. And finally, solution of the discretized inverse problem via explicit construction of the covariance matrix is prohibitive due to the need to solve the forward problem as many times as there are parameters. Our computational framework builds on the infinite-dimensional formulation proposed by Stuart [*Acta Numer.*, 19 (2010), pp. 451–559] and incorporates a number of components aimed at ensuring a convergent discretization of the underlying infinite-dimensional inverse problem. The framework additionally incorporates algorithms for manipulating the prior, constructing a low rank approximation of the data-informed component of the posterior covariance operator, and exploring the posterior that together ensure scalability of the entire framework to very high parameter dimensions. We demonstrate this computational framework on the Bayesian solution of an inverse problem in three-dimensional global seismic wave propagation with hundreds of thousands of parameters.

Key words. Bayesian inference, infinite-dimensional inverse problems, uncertainty quantification, scalable algorithms, low rank approximation, seismic wave propagation

AMS subject classifications. 35Q62, 62F15, 35R30, 35Q93, 65C60, 35L05

DOI. 10.1137/12089586X

1. Introduction.

We present a scalable computational framework for the quantification of uncertainty in large-scale *inverse problems*; that is, we seek to estimate probability densities for uncertain parameters,[1] given noisy observations or measurements and a model that maps parameters to output observables. The forward problem—which, without loss of generality, we take to be governed by PDEs—is usually well-posed (the solution exists, is unique, and is stable to perturbations in inputs), causal (later-time solutions depend only on earlier time solutions), and local (the forward operator includes derivatives that couple nearby solutions in space and time).

*Submitted to the journal's Methods and Algorithms for Scientific Computing section October 22, 2012; accepted for publication (in revised form) July 29, 2013; published electronically November 12, 2013. This research was partially supported by AFOSR grant FA9550-09-1-0608; DOE grants DE-FC02-11ER26052, DE-FG02-09ER25914, DE-FG02-08ER25860, DE-FC52-08NA28615, and DE-SC0009286; and NSF grants CMS-1028889, OPP-0941678, and DMS-0724746.
http://www.siam.org/journals/sisc/35-6/89586.html
†Institute for Computational Engineering & Sciences, The University of Texas at Austin, Austin, TX 78712 (tanbui@ices.utexas.edu, jmartin@ices.utexas.edu, georgst@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).
[1]We use the term *parameters* broadly to describe general model inputs that may be subject to uncertainty, which might include model parameters, boundary conditions, initial conditions, sources, geometry, and so on.

The inverse problem, on the other hand, reverses this relationship by seeking to estimate uncertain parameters from measurements or observations. The great challenge of solving inverse problems lies in the fact that they are usually ill-posed, noncausal, and nonlocal: Many different sets of parameter values may be consistent with the data, and the inverse operator couples solution values across space and time.

Nonuniqueness stems in part from the sparsity of data and the uncertainty in both measurements and the PDE model itself, and in part from nonconvexity of the parameter-to-observable map. The popular approach to obtaining a unique "solution" to the inverse problem is to formulate it as an optimization problem: Minimize the misfit between observed and predicted outputs in an appropriate norm while also minimizing a *regularization* term that penalizes unwanted features of the parameters. Estimation of parameters using the regularization approach to inverse problems as described above will yield an estimate of the "best" parameter values that simultaneously fit the data and minimize the regularization penalty term. However, we are interested not just in point estimates of the best-fit parameters, but a *complete statistical description* of the parameters values that is consistent with the data. The *Bayesian* approach [29, 44] does this by reformulating the inverse problem as a problem in *statistical inference*, incorporating uncertainties in the observations, the parameter-to-observable map, and prior information on the parameters. The solution of this inverse problem is the *posterior* probability distribution of the parameters, which reflects the degree of confidence in their values. Thus we are able to quantify the resulting uncertainty in the parameters, taking into account uncertainties in the data, model, and prior information.

The inverse problems we target here are characterized by *infinite-dimensional* parameter fields. This presents multiple difficulties, including proper choice of prior to guarantee well-posedness of the infinite-dimensional inverse problem, proper discretization to assure convergence to solutions of the infinite-dimensional problem, and algorithms for constructing and manipulating the posterior covariance matrix that ensure scalability to very large parameter dimensions. The approach we adopt in this paper follows [42], which seeks to first fully specify the statistical inverse problem on the infinite-dimensional parameter space. In order to accomplish this goal, we postulate the prior distribution as a Gaussian random field with covariance operator given by the square of the inverse of an elliptic PDE. This choice ensures that samples of the parameter field are (almost surely) continuous as functions, and that the statistical inverse problem is well-posed. To achieve a finite-dimensional approximation to the infinite-dimensional solution, we carefully construct a function-space-aware discretization of the parameter space.

The remaining challenge presented by infinite-dimensional statistical inverse problems is in computing statistics of the (discretized) posterior distribution. This is notoriously challenging for inverse problems governed by expensive-to-solve forward problems and high-dimensional parameter spaces (as in our application to global seismic wave propagation in section 6). The difficulty stems from the fact that evaluation of the probability of each point in parameter space requires solution of the forward PDE problem (which can take many hours on a large supercomputer), and many such evaluations (millions or more) are required to adequately sample the (discretized) posterior density in high dimensions by conventional Markov-chain Monte Carlo (MCMC) methods. In complementary work [36], we are developing methods that accelerate MCMC sampling of the posterior by employing a local Gaussian approximation of the posterior as a proposal density, which is computed from the Hessian of the negative log posterior. Here, as an alternative, we consider the case of the lin-

earized inverse problem; by linearization we mean that the parameter-to-observable map is linearized about the point that maximizes the posterior, which is known as the maximum a posteriori (MAP) point. With this linearization, the posterior becomes Gaussian, and its mean is given by the MAP point; this can be found by solving an appropriately weighted regularized nonlinear least squares optimization problem. Furthermore, the posterior covariance matrix is given by the Hessian of the negative log posterior evaluated at the MAP point.

Unfortunately, straightforward computation of the—nominally dense—Hessian is prohibitive, requiring as many forward-like solves as there are uncertain parameters (which in our example problem in section 6 are hundreds of thousands). However, the data are typically informative about a low-dimensional subspace of the parameter field: that is, the Hessian of the data misfit term is a compact operator that is sparse with respect to some basis. We exploit this fact to construct a low rank approximation of the (prior-preconditioned) data misfit Hessian using matrix-free Lanczos iterations [22, 36], which we observe to require a dimension-independent number of iterations. Each iteration requires a Hessian-vector product, which amounts to just a pair of forward/adjoint PDE solves, as well as a prior covariance operator application. Since we take the prior covariance in the form of the inverse of an elliptic differential operator, its application can be computed scalably via multigrid. The Sherman–Morrison–Woodbury formula is then invoked to express the covariance of the posterior. Finally, we show that the resulting expressions necessary for visualization and interrogation of the posterior distribution require just elliptic PDE solves and vector sums and inner products. In particular, the corresponding dense operators are never formed or stored. Solving the statistical inverse problem thus reduces to solving a fixed number of forward and adjoint PDE problems as well as an elliptic PDE representing the action of the prior. Thus, when the forward PDE problem can be solved in a scalable manner (as it is for our seismic wave propagation example in section 6), the entire computational framework is scalable with respect to forward problem dimension, uncertain parameter field dimension, and data dimension.

The computational framework presented here is applied to a sequence of realistic large-scale three-dimensional (3D) Bayesian inverse problems in global seismology, in which the acoustic wave speed of an unknown heterogeneous medium is to be inferred from noisy waveforms recorded at sparsely located receivers. Numerical results are presented for several problems with the number of unknown parameters up to 431,000. We have employed a similar approach for problems with more than one million parameters in related work [8].

In the following sections, we provide an overview of the framework for infinite-dimensional Bayesian inverse problems following [42] (section 2), present a consistent discretization scheme (section 3) for the infinite-dimensional problem, summarize a method for computing the MAP point (section 4), describe our low rank–based covariance approximation (section 5), and present results of the application of our framework to the Bayesian solution of an inverse problem in 3D global seismic wave propagation (section 6).

2. Bayesian framework for infinite-dimensional inverse problems.

2.1. Overview. In the Bayesian formulation, we state the inverse problem as a problem of *statistical inference* over the space of parameters. The solution of the resulting statistical inverse problem is a posterior probability distribution that reflects our degree of confidence that any set of candidate parameters might contain the actual values that gave rise to the data via the model and were consistent with the prior

information. Bayes' formula, presented in its infinite-dimensional form in section 2.2, defines this posterior probability distribution by combining a prior probability distribution with a likelihood model.

The inversion parameter is a function assumed to be defined over an open, bounded, and sufficiently regular set $\Omega \subset \mathbb{R}^3$. The statistical inverse problem is, therefore, naturally posed in an appropriate function space setting. Here, we adopt the infinite-dimensional framework developed in [42]. In particular, we choose a prior that ensures sufficient regularity of the parameter as required for the statistical inverse problem to be well-posed. We will represent the prior as a Gaussian random field whose covariance operator is the inverse of an elliptic differential operator. For certain problems, non-Gaussian priors can be important, but the use of non-Gaussian priors in statistical inverse problems is still subject to active research, in particular for infinite-dimensional parameters. Thus, here we restrict ourselves to priors given by Gaussian random fields. Let us motivate the choice of the covariance operator as inverse of an elliptic differential operator by considering two alternatives. A common choice for covariance operators in statistical inverse problems with a moderate number of parameters is to specify the covariance function, which gives the covariance of the parameter field between any two points. This necessitates either construction and "inversion" of a dense covariance matrix or expansion in a truncated Karhunen–Loéve (KL) basis. In the large-scale setting, inversion of a dense covariance matrix is clearly intractable, and the truncated KL approach can be impractical since it may require many terms to prevent biasing of the solution toward the strong prior modes. On the contrary, specifying the covariance as the inverse of an elliptic differential operator enables us to build on existing fast solvers for elliptic operators without constructing the dense operator. Discretizations of elliptic operators often satisfy a conditional independence property, which relates them to Gaussian Markov random fields and allows for statistical interpretation [6, 41]. Even if a Gaussian Markov random field is not based on an elliptic differential operator, this Markov property permits the use of fast, sparsity-exploiting algorithms, for instance, for taking samples from the distribution [40]. Our implementation employs multigrid as solver for the discretized elliptic systems.

A useful prior distribution must have bounded variance and have meaningful realizations. In our infinite-dimensional setting, we require samples to be pointwise well-defined, for instance, continuous. Furthermore, it is convenient to have the ability to apply the square root of the covariance operator, e.g., this is used to compute samples from a Gaussian distribution. We consider a Gaussian random field m on a domain $\Omega \subset \mathbb{R}^3$ with mean m_0 and covariance function $c(\boldsymbol{x}, \boldsymbol{y})$ describing the covariance between $m(\boldsymbol{x})$ and $m(\boldsymbol{y})$,

$$c(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}\left[(m(\boldsymbol{x}) - m_0(\boldsymbol{x}))(m(\boldsymbol{y}) - m_0(\boldsymbol{y}))\right] \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in \Omega. \tag{2.1}$$

The corresponding covariance operator $\mathcal{C}_0$ is

$$\left(\mathcal{C}_0 \phi\right)(\boldsymbol{x}) = \int_\Omega c\left(\boldsymbol{x}, \boldsymbol{y}\right) \phi\left(\boldsymbol{y}\right) \, d\boldsymbol{y} \tag{2.2}$$

for sufficiently regular functions ϕ defined over Ω. Thus, if the covariance operator is given by the solution operator of an elliptic PDE, the covariance function is the corresponding Green's function. Thus, Green's function properties have direct implications for properties of the random field m. For instance, since Green's functions of the Laplacian in one spatial dimension are bounded, the random field with the

Laplacian as covariance operator is of bounded variance. However, in two and three space dimensions, Green's functions $c(\boldsymbol{x}, \boldsymbol{y})$ of the Laplacian are singular along the diagonal, and thus the corresponding distribution has unbounded variance. Thus, intuitively the PDE solution operator used as covariance operator $\mathcal{C}_0$ has to be sufficiently smoothing and have bounded Green's functions. Indeed, this is necessary for the well-posedness of the infinite-dimensional Bayesian formulation [42]. The biharmonic operator, for example, has bounded Green's functions in two and three space dimensions. We choose $\mathcal{C}_0 = \mathcal{A}^{-2}$, where $\mathcal{A}$ is a Laplacian-like operator specified in section 2.3. This provides the desired simple and fast-to-apply square root operator $\mathcal{C}_0^{1/2} = \mathcal{A}^{-1}$ and allows a straightforward discretization.

An approach to extract information from the posterior distribution is to find the maximum a posterior (MAP) point, which amounts to the solution of an optimization problem as summarized in section 2.4. Finally, in section 2.5, we introduce a linearization of the parameter-to-observable map. This results in a Gaussian approximation of the posterior, which is the main focus of this paper.

2.2. Bayes' formula in infinite dimensions. To define Bayes' formula, we require a likelihood function that defines, for a given parameter field m, the distribution of observations $\boldsymbol{y}^{\text{obs}}$. Here, we assume a finite-dimensional vector $\boldsymbol{y}^{\text{obs}} \in \mathbb{R}^q$ of such observations. We introduce the *parameter-to-observable map* $\boldsymbol{f} : X := L^2(\Omega) \to \mathbb{R}^q$ as a deterministic function mapping a parameter field m to so-called observables $\boldsymbol{y} \in \mathbb{R}^q$, which are predictions of the observations. For the problems targeted here, an evaluation of $\boldsymbol{f}(m)$ requires a PDE solve followed by the application of an observation operator to extract $\boldsymbol{y}$ from the PDE solution. Even when the parameter m coincides with the "true" parameter, the observables $\boldsymbol{y}$ may still differ from the measurements $\boldsymbol{y}^{\text{obs}}$ due to measurement noise and inadequacy (i.e., the lack of fidelity of the governing PDEs with respect to reality) of the parameter-to-observable map $\boldsymbol{f}$. As is common practice, we assume the discrepancy between $\boldsymbol{y}$ and $\boldsymbol{y}^{\text{obs}}$ to be described by a Gaussian additive noise $\boldsymbol{\eta} \sim \mu_{\text{noise}} = \mathcal{N}(0, \boldsymbol{\Gamma}_{\text{noise}})$, independent of m. In particular, we have

$$\boldsymbol{y}^{\text{obs}} = \boldsymbol{f}(m) + \boldsymbol{\eta}, \tag{2.3}$$

which allows us to write the likelihood probability density function (pdf) as

$$\pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|m) \propto \exp\left(-\frac{1}{2}(\boldsymbol{f}(m) - \boldsymbol{y}^{\text{obs}})^T \boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{f}(m) - \boldsymbol{y}^{\text{obs}})\right). \tag{2.4}$$

The Bayesian solution to the infinite-dimensional inverse problem is then defined as follows: Given the likelihood π_{like} and the prior measure μ_0, find the conditional measure μ^y of m that satisfies the Bayes' formula

$$\frac{d\mu^y}{d\mu_0} = \frac{1}{Z}\pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|m), \tag{2.5}$$

where $Z = \int_X \pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|m)\, d\mu_0$ is a normalization constant. The formula (2.5) is understood as the Radon–Nikodym derivative of the posterior probability measure μ^y with respect to the prior measure μ_0. In order for (2.5) to be well-defined, $\boldsymbol{f} : X \to \mathbb{R}^q$ is assumed to be locally Lipschitz and quadratically bounded in the sense of Assumption 2.7 in [42]. While the Bayes' formula (2.5) is valid in finite and infinite dimensions, a more intuitive form of Bayes' formula that uses Lebesgue measures and thus only holds in finite dimensions is given in section 3.5.

2.3. Parameter space and the prior. As discussed in the introduction of section 2, we use a squared inverse elliptic operator as covariance operator $\mathcal{C}_0$ in (2.1), i.e., $\mathcal{C}_0 = \mathcal{A}^{-2}$. We first specify the elliptic PDE corresponding to $\mathcal{A}$ in weak form. For $s \in L^2(\Omega)$, the solution $m = \mathcal{A}^{-1}s$ satisfies

$$\alpha \int_\Omega (\boldsymbol{\Theta}\nabla m) \cdot \nabla p + mp \, d\boldsymbol{x} = \int_\Omega sp \, d\boldsymbol{x} \quad \text{for all } p \in H^1(\Omega), \tag{2.6}$$

with $\alpha > 0$, and $\boldsymbol{\Theta}(\boldsymbol{x}) \in \mathbb{R}^{3\times 3}$ is symmetric, uniformly bounded, and positive definite. Note that for $s \in L^2(\Omega)$, there exists a unique solution $m \in H^1(\Omega)$ by the Lax–Milgram theorem. Since $s \in L^2(\Omega)$ in (2.6), regularity results, e.g., [5,18], show that, in fact, $m \in H^2(\Omega)$ provided $\partial\Omega$ is sufficiently smooth, e.g., Ω is a $C^{1,1}$ domain. In this case, (m, s) satisfies the elliptic differential equation

$$-\alpha\nabla \cdot (\boldsymbol{\Theta}\nabla m) + \alpha m = s \quad \text{in } \Omega, \tag{2.7a}$$

$$\alpha(\boldsymbol{\Theta}\nabla m) \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega, \tag{2.7b}$$

where $\boldsymbol{n}$ denotes the outward unit normal on $\partial\Omega$.

Let us denote by $\mathcal{A}$ the differential operator together with its domain of definition specified by (2.7); hence $\mathcal{A}$ is a densely defined operator on $L^2(\Omega)$ with the domain

$$D(\mathcal{A}) := \left\{ m \in H^2(\Omega) : \alpha\boldsymbol{\Theta}\nabla m \cdot \mathbf{n} = 0 \right\}.$$

The operator $\mathcal{A}$ is assumed to be "Laplacian-like" in the sense of Assumption 2.9 in [42]. Briefly, this assumption requires $\mathcal{A}$ to be positive definite, self-adjoint, invertible, and have eigenfunctions that form an orthonormal basis of $L^2(\Omega)$. Additionally, certain growth conditions on the eigenvalues and $L^\infty(\Omega)$ norms of the eigenfunctions are enforced.[2]

To summarize, we consider m as a Gaussian random field whose distribution law is a Gaussian measure $\mu_0 := \mathcal{N}(m_0, \mathcal{C}_0)$ on $\boldsymbol{L}^2(\Omega)$, with mean $m_0 \in D(\mathcal{A})$ and covariance operator $\mathcal{C}_0 := \mathcal{A}^{-2}$. The definition of the Gaussian prior measure is meaningful since $\mathcal{A}^{-2}$ is a trace class operator on $L^2(\Omega)$ [42], which guarantees bounded variance and almost surely pointwise well-defined samples since $\mu_0(X) = 1$ holds, where $X := C(\Omega)$ denotes the space of continuous functions defined on Ω (see [42, Lemma 6.25]).

2.4. The MAP point. As a first step in exploring the solution of the statistical inverse problem, we determine the MAP estimate of the posterior measure. In a finite-dimensional setting, the MAP estimate is the point in parameter space that maximizes the posterior pdf. This notion does not generalize directly to the infinite-dimensional setting, but we can still define the MAP estimate m_{MAP} as the point m in parameter space that asymptotically maximizes the measure of a ball with radius ε centered at m, in the limit as $\varepsilon \to 0$. We recall that the Cameron–Martin space E equipped with the inner product $(\cdot, \cdot)_E := (\mathcal{C}_0^{-1/2}\cdot, \mathcal{C}_0^{-1/2}\cdot)$ associated with $\mathcal{C}_0$ is the range of $\mathcal{C}_0^{1/2}$ [27], and hence coincides with $D(\mathcal{A})$. Using variational arguments, it can be shown (see [42]) that m_{MAP} is given by solving the optimization problem

$$\min_{m\in E} \mathcal{J}(m), \tag{2.8}$$

[2] We note that this growth condition on the eigenfunctions may not be straightforward to demonstrate (or may not even hold) for a nonrectangular domain Ω and nonconstant coefficient $\boldsymbol{\Theta}$. In these cases, we expect that alternative proofs of the results in [42] can be accessed via regularity properties of the covariance function for the prior distribution. See, for example, [1, 35].

where

$$\mathcal{J}(m) := \frac{1}{2}\left\|\boldsymbol{f}(m) - \boldsymbol{y}^{\text{obs}}\right\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \frac{1}{2}\left\|\mathcal{A}(m - m_0)\right\|^2_{L^2(\Omega)}. \tag{2.9}$$

The well-posedness of the optimization problem (2.8) is guaranteed by the assumptions on $\boldsymbol{f}(m)$ in section 2.2.

2.5. A linearized Bayesian formulation. Once we have obtained the MAP estimate m_{MAP}, we approximate the parameter-to-observable map $\boldsymbol{f}(m)$ by its linearization about m_{MAP}, which ultimately results in a Gaussian approximation to the posterior distribution, as shown below. When the parameter-to-observable map is nearly linear this is a reasonable approximation; moreover, there are other scenarios in which the linearization, and the resulting Gaussian approximation, may be useful. Of particular interest here are the limits of small data noise and many observations. In the small noise case, the parameter-to-obvervable map can be nearly linear as a mapping into the subset of the observable space on which the likelihood distribution is nonnegligible—even when $\boldsymbol{f}(m)$ is significantly nonlinear. The asymptotic normality discussions in [24, 33] suggest that under certain conditions, the many observations case can lead to a Gaussian posterior. Finally, even if this approximation fails to describe the posterior distribution adequately, the linearization is still useful in building an initial step for the rejection sampling approach or a Gaussian proposal distribution for the Metropolis–Hastings algorithm [36, 39]. These methods are related to the sampling algorithm in [25], which also employs derivative information to respect the local structure of the parameter space.

Assuming that the parameter-to-observable map $\boldsymbol{f}$ is Fréchet differentiable, we linearize the right-hand side of (2.3) around m_{MAP} to obtain

$$\boldsymbol{y}^{\text{obs}} \approx \boldsymbol{f}(m_{\text{MAP}}) + \boldsymbol{F}(m - m_{\text{MAP}}) + \boldsymbol{\eta},$$

where $\boldsymbol{F}$ is the Fréchet derivative of $\boldsymbol{f}(m)$ evaluated at m_{MAP}. Consequently, the posterior distribution μ^y of m conditional on $\boldsymbol{y}^{\text{obs}}$ is a Gaussian measure $\mathcal{N}(m_{\text{MAP}}, \mathcal{C}_{\text{post}})$ with mean m_{MAP} and covariance operator $\mathcal{C}_{\text{post}}$ defined by [42]:

$$\mathcal{C}_{\text{post}} = (\boldsymbol{F}^{\natural}\boldsymbol{\Gamma}^{-1}_{\text{noise}}\boldsymbol{F} + \mathcal{C}_0^{-1})^{-1}, \tag{2.10}$$

with $\boldsymbol{F}^{\natural}$ denoting the adjoint of $\boldsymbol{F}$, an operator from the space of observations $\mathbb{R}^q$ to $L^2(\Omega)$. In principle, a local Gaussian approximation of the posterior at the MAP point can also be found for non-Gaussian priors and when the noise in the observables is not additive and Gaussian as in (2.3). In these cases, however, even for a linear parameter-to-observable map the local Gaussian approximation might only be a reasonable approximation to the true posterior distribution in a small neighborhood around the MAP point.

3. Discretization of the Bayesian inverse problem.

3.1. Overview. Next, we present a numerical discretization of the infinite-dimensional Bayesian statistical inverse problem described in section 2.2. The discretized parameter space is inherently high-dimensional (with dimension dependent upon the mesh size). If discretization is not performed carefully at each step, it is unlikely that the discrete solutions will converge to the desired infinite-dimensional solution in a meaningful way [32, 42].

In the following, and particularly in section 3.3, we choose a mass matrix-weighted vector product instead of the standard Euclidean vector product. While this is a natural choice in finite element discretizations [7, 46], this does lead to a few complications, for instance, the use of covariance operators that are not symmetric in the conventional sense (they are self-adjoint, however). This choice is much better suited for proper discretization of the infinite-dimensional expressions given in this paper, and the resulting numerical expressions for computation will more closely resemble their infinite-dimensional counterparts in section 2. By contrast, the correct corresponding expressions in the Euclidean inner product are significantly less intuitive in our opinion, and ultimately more cumbersome to manipulate and interpret than the development we give here.

We provide finite-dimensional approximations of the prior and the posterior distributions in sections 3.4 and 3.5, respectively. To study and visualize the uncertainty in Gaussian random fields, such as the prior and posterior distributions, we generate realizations (i.e., samples) and compute pointwise variance fields. This must be done carefully in light of the mass-weighted inner products due to the finite element discretization introduced in section 3.3. We present explicit expressions for computing these quantities for the prior in sections 3.6 and 3.7. The fast generation of samples and the pointwise variance field from the Gaussian approximation of the posterior exploits the low rank ideas presented in section 5. Thus, the presentation of the corresponding expressions is postponed to section 5.3.

3.2. Finite-dimensional parameter space. We consider a finite-dimensional subspace V_h of $L^2(\Omega)$ originating from a finite element discretization with continuous Lagrange basis functions $\{\phi_j\}_{j=1}^n$, which correspond to the nodal points $\{\boldsymbol{x}_j\}_{j=1}^n$, such that

$$\phi_j(\boldsymbol{x}_i) = \delta_{ij} \qquad \text{for } i,j \in \{1,\dots,n\}.$$

Instead of statistically inferring parameter functions $m \in L^2(\Omega)$, we perform this task on the approximation $m_h = \sum_{j=1}^n m_j \phi_j \in V_h$. Consequently, the coefficients $(m_1,\dots,m_n)^T \in \mathbb{R}^n$ are the actual parameters to be inferred. For simplicity of notation, we shall use the boldface symbol $\boldsymbol{m} = (m_1,\dots,m_n)^T$ to denote the nodal vector of a function m_h in V_h.

3.3. Discrete inner product. 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 $\boldsymbol{M}$ to approximate the infinite-dimensional L^2-inner product. We denote this weighted inner product by $(\cdot,\cdot)_{\boldsymbol{M}}$ and assume that $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is symmetric and positive definite. To distinguish $\mathbb{R}^n$ with the $\boldsymbol{M}$-weighted inner product from the usual Euclidean space $\mathbb{R}^n$, we denote it by $\mathbb{R}^n_{\boldsymbol{M}}$. For any $m_1, m_2 \in L^2(\Omega)$, observe that $(m_1,m_2)_{L^2(\Omega)} \approx (m_{1h},m_{2h})_{L^2(\Omega)} = (\boldsymbol{m}_1,\boldsymbol{m}_2)_{\boldsymbol{M}} = \boldsymbol{m}_1^T\boldsymbol{M}\boldsymbol{m}_2$, which motivates the choice of $\boldsymbol{M}$ as the finite element mass matrix defined by

$$M_{ij} = \int_\Omega \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x})d\boldsymbol{x}\ , \quad i,j \in \{1,\dots,n\}\,. \tag{3.1}$$

When using the $\boldsymbol{M}$-inner product, there is a critical distinction that must be made between the matrix adjoint and the matrix transpose. For an operator $\boldsymbol{B} : \mathbb{R}^n_{\boldsymbol{M}} \to \mathbb{R}^n_{\boldsymbol{M}}$, we denote the matrix transpose by $\boldsymbol{B}^T$ with entries $[B^T]_{ij} = B_{ji}$. The

adjoint $\boldsymbol{B}^*$ of $\boldsymbol{B}$, however, must satisfy

$$(\boldsymbol{B}^*\boldsymbol{m}_1, \boldsymbol{m}_2)_{\boldsymbol{M}} = (\boldsymbol{m}_1, \boldsymbol{B}\boldsymbol{m}_2)_{\boldsymbol{M}} \quad \text{for all } \boldsymbol{m}_1, \boldsymbol{m}_2 \in \mathbb{R}^n_{\boldsymbol{M}}. \tag{3.2}$$

This implies that

$$\boldsymbol{B}^* = \boldsymbol{M}^{-1}\boldsymbol{B}^T\boldsymbol{M}. \tag{3.3}$$

In the following, we also need the adjoints $\boldsymbol{F}^\natural$ of $\boldsymbol{F} : \mathbb{R}^n_{\boldsymbol{M}} \to \mathbb{R}^q$ and $\boldsymbol{V}^\diamond$ of $\boldsymbol{V} : \mathbb{R}^r \to \mathbb{R}^n_{\boldsymbol{M}}$ (for some r), where $\mathbb{R}^q$ and $\mathbb{R}^r$ are endowed with the Euclidean inner product. The desired adjoints can be expressed as

$$\boldsymbol{F}^\natural = \boldsymbol{M}^{-1}\boldsymbol{F}^T, \tag{3.4}$$

$$\boldsymbol{V}^\diamond = \boldsymbol{V}^T\boldsymbol{M}. \tag{3.5}$$

Next, let P_h be the projection from $L^2(\Omega)$ to V_h. Then, the matrix representation $\boldsymbol{B} : \mathbb{R}^n_{\boldsymbol{M}} \to \mathbb{R}^n_{\boldsymbol{M}}$ for the operator $\mathcal{B}_h := P_h\mathcal{B}P_h'$, where $\mathcal{B} : L^2(\Omega) \to L^2(\Omega)$ and $P_h' : V_h \to L^2(\Omega)$, is implicitly given with respect to the Lagrange basis $\{\phi_i\}_{i=1}^n$ in V_h by

$$\int_\Omega \phi_i \mathcal{B}\phi_j \, dx = (\boldsymbol{e}_i, \boldsymbol{B}\boldsymbol{e}_j)_{\boldsymbol{M}},$$

where $\boldsymbol{e}_i$ is the coordinate vector corresponding to the basis function ϕ_i. As a result, one can write $\boldsymbol{B}$ explicitly as

$$\boldsymbol{B} = \boldsymbol{M}^{-1}\boldsymbol{K}, \tag{3.6}$$

where $\boldsymbol{K}$ is given by

$$K_{ij} = \int_\Omega \phi_i \mathcal{B}\phi_j \, dx, \quad i, j \in \{1, \ldots, n\}.$$

3.4. Finite-dimensional approximation of the prior. Next, we derive the finite-dimensional representation of the prior. The matrix representation of the operator $\mathcal{A}$ defined in section 2.3 is given by the stiffness matrix $\boldsymbol{K}$ with entries

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

It follows that both $\boldsymbol{A} = \boldsymbol{M}^{-1}\boldsymbol{K}$ and $\boldsymbol{A}^{-1} = \boldsymbol{K}^{-1}\boldsymbol{M}$ are self-adjoint operators in the sense of (3.3).

We are now in a position to define the finite-dimensional Gaussian prior measure μ_0^h specified by the following density (with respect to the Lebesgue measure):

$$\pi_{\text{prior}}(\boldsymbol{m}) \propto \exp\left[-\frac{1}{2}\|\boldsymbol{A}(\boldsymbol{m} - \boldsymbol{m}_0)\|^2_{\boldsymbol{M}}\right]. \tag{3.7}$$

This definition implies that $\boldsymbol{\Gamma}_{\text{prior}} = \boldsymbol{A}^{-2}$.

3.5. Finite-dimensional approximation of the posterior. In infinite dimensions, the Bayes' formula (2.5) has to be expressed in terms of the Radon–Nikodym derivative since the prior and posterior distributions do not have density functions with respect to the Lebesgue measure. Since we approximate the prior measure μ_0 by μ_0^h, it is natural to define a finite-dimensional approximation $\mu^{y,h}$ of the posterior measure μ^y such that

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

where $Z^h = \int_X \pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|\boldsymbol{m})\, d\mu_0^h$, and π_{like} is the likelihood (2.4) evaluated at m_h. If we define $\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{y}^{\text{obs}})$ as the density of $\mu^{y,h}$, again with respect to the Lebesgue measure, we recover the familiar finite-dimensional Bayes' formula

$$\pi_{\text{post}}(\boldsymbol{m}|\boldsymbol{y}^{\text{obs}}) \propto \pi_{\text{prior}}(\boldsymbol{m})\pi_{\text{like}}(\boldsymbol{y}^{\text{obs}}|m_h), \tag{3.8}$$

where the normalization constant $1/Z^h$, which does not depend on $\boldsymbol{m}$, is omitted. Finally, we can express the posterior pdf explicitly as

$$\pi_{\text{post}}(\boldsymbol{m}) \propto \exp\left(-\frac{1}{2}\left\|\boldsymbol{f}(m_h) - \boldsymbol{y}^{\text{obs}}\right\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} - \frac{1}{2}\left\|\boldsymbol{A}(\boldsymbol{m}-\boldsymbol{m}_0)\right\|^2_{\boldsymbol{M}}\right), \tag{3.9}$$

where, to recall our notation, $m_h = \sum_{j=1}^n m_j\phi_j \in V_h$ and $\boldsymbol{m} = (m_1, \ldots, m_n)^T$. We observe that the negative log of the right-hand side of (3.9) is the finite-dimensional approximation of the objective functional in (2.8).

As a finite-dimensional counterpart of section 2.5, we linearize the parameter-to-observable map $\boldsymbol{f}$ at the MAP point, but now consider it as a function of the coefficient vector $\boldsymbol{m}$. Let $\boldsymbol{\Gamma}_{\text{post}}$ be the posterior covariance matrix in the $\boldsymbol{M}$-inner product. Using (2.10), we obtain

$$\boldsymbol{\Gamma}_{\text{post}} = \left(\boldsymbol{F}^\natural\boldsymbol{\Gamma}^{-1}_{\text{noise}}\boldsymbol{F} + \boldsymbol{\Gamma}^{-1}_{\text{prior}}\right)^{-1}, \tag{3.10}$$

with $\boldsymbol{F}^\natural = \boldsymbol{M}^{-1}\boldsymbol{F}^T$ as defined in (3.4). Note that $\boldsymbol{\Gamma}_{\text{post}}$ is self-adjoint, i.e., $\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{\Gamma}^*_{\text{post}}$ in the sense of (3.3).

Since the posterior covariance matrix $\boldsymbol{\Gamma}_{\text{post}}$ is typically dense, we wish to avoid explicitly storing it, especially when the parameter dimension n is large. Even if we are able to do so, it is prohibitively expensive to construct. The reason is that the Jacobian of the parameter-to-observable map, $\boldsymbol{F}$, is generally a dense matrix, and its construction typically requires n forward PDE solves. This is clearly intractable when n is large and solving the PDEs is expensive. However, one can exploit the structure of the inverse problem, to approximate the posterior covariance matrix with desired accuracy, as we shall show in section 5.

3.6. Sample generation in a finite element discretization. We begin by developing expressions for a general Gaussian distribution with mean $\bar{\boldsymbol{m}}$ and covariance matrix $\boldsymbol{\Gamma}$. Then, they are specified for the Gaussian prior with $(\boldsymbol{m}_0, \boldsymbol{\Gamma}_{\text{prior}})$. Realizations of a finite-dimensional Gaussian random variable with mean $\bar{\boldsymbol{m}}$ and covariance matrix $\boldsymbol{\Gamma}$ can be found by choosing a vector $\boldsymbol{n}$ containing independent and identically distributed (i.i.d.) standard normal random values and computing

$$\boldsymbol{m} = \bar{\boldsymbol{m}} + \boldsymbol{L}\boldsymbol{n}, \tag{3.11}$$

where $\boldsymbol{L}$ is a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n_{\boldsymbol{M}}$ such that $\boldsymbol{\Gamma} = \boldsymbol{L}\boldsymbol{L}^{\diamond}$, in which the adjoint $\boldsymbol{L}^{\diamond} = \boldsymbol{L}^T\boldsymbol{M}$ (see also (3.5)). Note that $\boldsymbol{M}^{-1/2}\boldsymbol{n}$ is a sample from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ in the mass-weighted inner product.

In particular, for $\bar{\boldsymbol{m}} = \boldsymbol{m}_0$ and $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}_{\text{prior}}$, we have $\boldsymbol{L}_{\text{prior}} = \boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{M}^{-1/2} = \boldsymbol{K}^{-1}\boldsymbol{M}^{1/2}$ (see the appendix) and samples from the prior are computed as $\boldsymbol{m} = \boldsymbol{m}_0 + \boldsymbol{K}^{-1}\boldsymbol{M}^{1/2}\boldsymbol{n}$. Samples from the Gaussian approximation to the posterior use the low rank representation introduced in section 5, and the corresponding expressions are given in (5.8) and (5.9).

3.7. The pointwise variance field in a finite element discretization. Let us approximate the covariance function in V_h for a generic Gaussian measure with covariance operator $\mathcal{C}$. Recall from section 2.3 that the covariance function $c(\boldsymbol{x}, \boldsymbol{y})$ corresponding to the covariance operator $\mathcal{C}$ is the Green's function of $\mathcal{C}^{-1}$, i.e.,

$$\mathcal{C}^{-1}c(\boldsymbol{x}, \boldsymbol{y}) := \delta_{\boldsymbol{y}}(\boldsymbol{x}) \quad \text{for } \boldsymbol{x} \in \Omega,$$

where $\delta_{\boldsymbol{y}}$ denotes the Dirac delta function concentrated at $\boldsymbol{y} \in \Omega$. We approximate $c(\boldsymbol{x}, \boldsymbol{y})$ in the finite element space V_h by $c_h(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n c_i(\boldsymbol{y})\,\phi_i(\boldsymbol{x})$ with coefficient vector $\boldsymbol{c}(\boldsymbol{y}) = [c_1(\boldsymbol{y}), \ldots, c_n(\boldsymbol{y})]^T$. Using the Galerkin finite element method to obtain a finite element approximation of the preceding equation results in

$$\mathbf{C}^{-1}\boldsymbol{c}(\boldsymbol{y}) = \boldsymbol{\Phi}(\boldsymbol{y}) \quad \text{with} \quad \boldsymbol{\Phi}(\boldsymbol{y}) = [\phi_1(\boldsymbol{y}), \ldots, \phi_n(\boldsymbol{y})]^T$$

and the entries of the matrix $\mathbf{C}^{-1}$ are given by $C_{ij}^{-1} = \left(\phi_i, \mathcal{C}^{-1}\phi_j\right)_{L^2(\Omega)}$. It follows that $\boldsymbol{c}(\boldsymbol{y}) = \mathbf{C}\boldsymbol{\Phi}(\boldsymbol{y})$ and

$$c_h(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n c_i(\boldsymbol{y})\Phi_i(\boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{x})^T\,\mathbf{C}\boldsymbol{\Phi}(\boldsymbol{y}) \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in \Omega.$$

Let us denote by $\boldsymbol{\Gamma}^{-1}$ the representation of $P_h\mathcal{C}^{-1}P_h'$ in V_h; then, using (3.6) yields that $\mathbf{C} = \boldsymbol{\Gamma}\boldsymbol{M}^{-1}$. Consequently, the discretized covariance function for the covariance operator $\mathcal{C}$ now becomes

$$c_h(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Phi}(\boldsymbol{x})^T\,\boldsymbol{\Gamma}\boldsymbol{M}^{-1}\boldsymbol{\Phi}(\boldsymbol{y}). \tag{3.12}$$

Let us now apply (3.12) to compute the prior variance field. As discussed in section 3.4, $\boldsymbol{\Gamma}_{\text{prior}} = \boldsymbol{A}^{-2} = \boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{K}^{-1}\boldsymbol{M}$. This results in the discretized prior covariance function

$$c_h^{\text{prior}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Phi}(\boldsymbol{x})^T\,\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{K}^{-1}\boldsymbol{\Phi}(\boldsymbol{y}).$$

By taking $\boldsymbol{y} = \boldsymbol{x}$, the prior variance field at an arbitrary point $\boldsymbol{x} \in \Omega$ reads

$$c_h^{\text{prior}}(\boldsymbol{x}, \boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{x})^T\,\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{K}^{-1}\boldsymbol{\Phi}(\boldsymbol{x}).$$

The pointwise variance field of the posterior distribution builds on the low rank representation introduced in section 5. The resulting expression, which requires the prior variance field, is given in (5.7).

4. Finding the MAP point. Section 2.4 introduced the idea of the MAP point as a first step in exploring the solution of the statistical inverse problem. To find the MAP point, one needs to solve a discrete approximation (using the discretizations of section 3) of the optimization problem (2.8), which amounts to a large-scale nonlinear least squares numerical optimization problem. In this section, we provide just a brief summary of a scalable method we use for solving this problem, and refer the reader to our earlier work for details, in particular the work on inverse wave propagation [2, 3, 17]. We use an inexact matrix-free Newton–conjugate gradient (CG) method in which only Hessian-vector products are required. These Hessian-vector products are computed by solving a linearized forward-like and an adjoint-like PDE problem, and thus the Hessian matrix is never constructed explicitly. Inner CG iterations are terminated prematurely when sufficient reduction is made in the norm of the gradient, or when a direction of negative curvature is encountered. The prior operator is used to precondition the CG iterations. Globalization is through an Armijo backtracking line search.

Because the major components of the method can be expressed as solving PDE-like systems, the method inherits the scalability (with respect to problem dimension) of the forward PDE solve. The remaining ingredient for overall scalability is that the optimization algorithm itself be scalable with increasing problem size. This is indeed the case: For a wide class of nonlinear inverse problems, the outer Newton iterations and the inner CG iterations are independent of the mesh size, as is found to be the case, for instance, for inverse wave propagation [3, 17]. This is a consequence of the use of a Newton solver, of the compactness of the Hessian of the data misfit term (i.e., the first term) in (2.9), and of the use of preconditioning by $\mathbf{\Gamma}_{\text{prior}}$, so that the resulting preconditioned Hessian is a compact perturbation of the identity, for which CG exhibits mesh-independent iterations.

5. Low rank approximation of the Hessian matrix.

5.1. Overview. As stated in section 2.5, linearizing the parameter-to-observable map results in the posterior covariance matrix being given by the inverse of the Hessian of the negative log posterior. Explicitly computing this Hessian matrix requires a (linearized) forward PDE problem for each of its columns, and thus as many (linearized) forward PDE solves are required as there are parameters. For inverse problems in which one seeks to infer an unknown parameter field, discretization results in a very large number of parameters; explicitly computing the Hessian—and hence the covariance matrix—is thus out of the question. As a remedy, we exploit the structure of the problem to find an approximation of the Hessian that can be constructed and dealt with efficiently.

When the linearized parameter-to-observable map is used in $\mathcal{J}(m)$ (as defined in (2.9)) and second derivatives of the resulting functional are computed, one obtains the Gauss–Newton portion of the Hessian of $\mathcal{J}(m)$. Both the full Hessian matrix as well as its Gauss–Newton portion are positive definite at the MAP point and they only differ in terms that involve the adjoint variable. Since the adjoint system is driven only by the data misfit (see, for instance, the adjoint wave (6.5)), the adjoint variable is expected to be small when the data misfit is small, which occurs provided the model and observational errors are not too large. The Gauss–Newton portion of the Hessian is thus often a good approximation of the full Hessian of $\mathcal{J}(m)$.

For conciseness and convenience of the notation, we focus on computing a low rank approximation of the Gauss–Newton portion of the (misfit) Hessian in section 5.2. The same approach also applies to the computation of a low rank approximation of the full

Hessian, whose inverse might be a better approximation for the covariance matrix if the data is very noisy and the data misfit at the MAP point cannot be neglected. The low rank construction of the misfit Hessian is based on the Lanczos method and thus only requires Hessian-vector products. Using the Sherman–Morrison–Woodbury formula, this approximation translates into an approximation of the posterior covariance matrix.

In section 5.3, we present low rank–exploiting methods for sample generation from the Gaussian approximation of the posterior, as well as methods for the efficient computation of the pointwise variance field. Finally, in section 5.4, we discuss the overall scalability of our approach.

5.2. Low rank covariance approximation. For many ill-posed inverse problems, the Gauss–Newton portion of the Hessian matrix (called the Gauss–Newton Hessian for short) of the data misfit term in (2.9) evaluated at any $\mathbf{m}$,

$$\boldsymbol{H}_{\text{misfit}} := \boldsymbol{F}^{\natural}\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{F}, \tag{5.1}$$

behaves like (the discretization of) a compact operator (see, e.g., [45, p. 17]). The intuitive reason for this is that only parameter modes that strongly influence the observations through the linearized parameter-to-observable map $\boldsymbol{F}$ will be present in the dominant spectrum of the Hessian (5.1). For typical inverse problems, observations are sparse, and hence the dimension of the observable space is much smaller than that of the parameter space. Furthermore, highly oscillatory perturbations in the parameter field often have a negligible effect on the output of the parameter-to-observable map. In [10, 11], we have shown that the Gauss–Newton Hessian of the data misfit is a compact operator, and that for smooth media its eigenvalues decay exponentially to zero. Thus, the range space of the Gauss–Newton Hessian is effectively finite-dimensional even before discretization, i.e., it is independent of the mesh. We can exploit the compact nature of the data misfit Hessian to construct scalable algorithms for approximating the inverse of the Hessian [22, 36].

A simple manipulation of (3.10) yields the following expression for the posterior covariance matrix, which will prove convenient for our purposes:

$$\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{H}_{\text{misfit}}\boldsymbol{\Gamma}_{\text{prior}}^{1/2} + \boldsymbol{I}\right)^{-1}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}. \tag{5.2}$$

We now present a fast method for approximating $\boldsymbol{\Gamma}_{\text{post}}$ with controllable accuracy by making a low rank approximation of the so-called *prior-preconditioned Hessian of the data misfit*, namely,

$$\tilde{\boldsymbol{H}}_{\text{misfit}} := \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{H}_{\text{misfit}}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}. \tag{5.3}$$

Let $(\lambda_i, \boldsymbol{v}_i), i = 1, \ldots, n$, be the eigenpairs of $\tilde{\boldsymbol{H}}_{\text{misfit}}$, and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^{n\times n}$. Define $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ such that its columns are the eigenvectors $\boldsymbol{v}_i$ of $\tilde{\boldsymbol{H}}_{\text{misfit}}$. Replacing $\tilde{\boldsymbol{H}}_{\text{misfit}}$ by its spectral decomposition (recall that $\boldsymbol{V}^{\diamond}$ is the adjoint of $\boldsymbol{V}$ as defined in (3.5)),

$$\left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{H}_{\text{misfit}}\boldsymbol{\Gamma}_{\text{prior}}^{1/2} + \boldsymbol{I}\right)^{-1} = (\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\diamond} + \boldsymbol{I})^{-1}.$$

When the eigenvalues of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ decay rapidly we can construct a low rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ by computing only the r largest eigenvalues, i.e.,

$$\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{H}_{\text{misfit}}\boldsymbol{\Gamma}_{\text{prior}}^{1/2} = \boldsymbol{V}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^{\diamond} + \mathcal{O}\left(\sum_{i=r+1}^{n}\lambda_i\right),$$

where $\boldsymbol{V}_r \in \mathbb{R}^{n\times r}$ contains r eigenvectors of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ corresponding to the r largest eigenvalues $\lambda_i, i = 1,\dots,r$, and $\boldsymbol{\Lambda}_r = \operatorname{diag}(\lambda_1,\dots,\lambda_r) \in \mathbb{R}^{r\times r}$. We can then invert the approximate Hessian using the Sherman–Morrison–Woodbury formula to obtain

$$\left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{H}_{\text{misfit}}\boldsymbol{\Gamma}_{\text{prior}}^{1/2} + \boldsymbol{I}\right)^{-1} = \boldsymbol{I} - \boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^{\diamond} + \mathcal{O}\left(\sum_{i=r+1}^{n}\frac{\lambda_i}{\lambda_i+1}\right), \tag{5.4}$$

where $\boldsymbol{D}_r := \operatorname{diag}(\lambda_1/(\lambda_1+1),\dots,\lambda_r/(\lambda_r+1)) \in \mathbb{R}^{r\times r}$. Equation (5.4) shows the truncation error due to the low rank approximation based on the first r eigenvalues. To obtain an accurate approximation of $\boldsymbol{\Gamma}_{\text{post}}$, only eigenvectors corresponding to eigenvalues that are small compared to 1 can be neglected. With such a low rank approximation, the final expression for the approximate posterior covariance is given by

$$\boldsymbol{\Gamma}_{\text{post}} \approx \boldsymbol{\Gamma}_{\text{prior}} - \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^{\diamond}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}. \tag{5.5}$$

Note that (5.5) expresses the posterior uncertainty (in terms of the covariance matrix) as the prior uncertainty less any information gained from the data. Due to the square root of the prior in the rightmost term in (5.5), the information gained from the data is filtered through the prior; i.e., only information consistent with the prior can reduce the posterior uncertainty.

5.3. Fast generation of samples and the pointwise variance field. Properties of the last term in (5.5), such as its diagonal (which provides the reduction in variance due to the knowledge acquired from the data) can be obtained numerically through just r applications of the square root of the prior covariance matrix to r columns of $\boldsymbol{V}_r$. Let us define

$$\tilde{\boldsymbol{V}}_r = \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{V}_r;$$

then (5.5) becomes

$$\boldsymbol{\Gamma}_{\text{post}} \approx \boldsymbol{\Gamma}_{\text{prior}} - \tilde{\boldsymbol{V}}_r\boldsymbol{D}_r\tilde{\boldsymbol{V}}_r^{\diamond}, \tag{5.6}$$

with $\tilde{\boldsymbol{V}}_r^{\diamond} = \boldsymbol{V}_r^{\diamond}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$.

The linearized posterior is a Gaussian distribution with known mean, namely the MAP point, and low rank–based covariance (5.6). Thus, the pointwise variance field and samples can be generated as in sections 3.7 and 3.6, respectively. The variance field can be computed as

$$c_h^{\text{post}}(\boldsymbol{x},\boldsymbol{x}) = c_h^{\text{prior}}(\boldsymbol{x},\boldsymbol{x}) - \sum_{k=1}^{r} d_k\left(\tilde{\boldsymbol{v}}_{kh}(\boldsymbol{x})\right)^2, \tag{5.7}$$

where $\tilde{\boldsymbol{v}}_{kh}(\boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{x})^T\tilde{\boldsymbol{v}}_k$, with $\tilde{\boldsymbol{v}}_k$ denoting the kth column of $\tilde{\boldsymbol{V}}_r$, is the function in V_h corresponding to the nodal vector $\tilde{\boldsymbol{v}}_k$.

Now, we can compute samples from the posterior provided that we have a factorization $\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{L}\boldsymbol{L}^{\diamond}$. One possibility for $\boldsymbol{L}$ (see the appendix for the detailed derivation) reads

$$\boldsymbol{L} := \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\left(\boldsymbol{V}_r\boldsymbol{P}_r\boldsymbol{V}_r^{\diamond} + \boldsymbol{I}\right)\boldsymbol{M}^{-1/2} \tag{5.8}$$

with $\boldsymbol{P}_r = \text{diag}\left(1/\sqrt{\lambda_1 + 1} - 1, \ldots, 1/\sqrt{\lambda_r + 1} - 1\right) \in \mathbb{R}^{r\times r}$, $\boldsymbol{L}$ as a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n_M$, and $\boldsymbol{I}$ as the identity map in both $\mathbb{R}^n$ and $\mathbb{R}^n_{\boldsymbol{M}}$. As discussed in section 3.6, samples can be then computed as

$$\boldsymbol{\nu}^{\text{post}} = \boldsymbol{m}_{\text{MAP}} + \boldsymbol{L}\boldsymbol{n}, \tag{5.9}$$

where $\boldsymbol{n}$ is an i.i.d. standard normal random vector.

5.4. Scalability. We now discuss the scalability of the above low rank construction of the posterior covariance matrix in (5.5). The dominant task is the computation of the dominant spectrum of the prior-preconditioned Hessian of the data misfit, $\tilde{\boldsymbol{H}}_{\text{misfit}}$, given by (5.3). Computing the spectrum by a matrix-free eigensolver such as Lanczos means that we need only form actions of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ with a vector. As argued at the end of section 3.5, the linearized parameter-to-observable map $\boldsymbol{F}$ is too costly to be constructed explicitly since it requires n linearized forward PDE solves. However, its action on a vector can be computed by solving a single linearized forward PDE (which we term the *incremental forward problem*), regardless of the number of parameters n and observations q. Similarly, the action of $\boldsymbol{F}^\natural$ on a vector can be found by solving a single linearized adjoint PDE (which we term the *incremental adjoint problem*). Explicit expressions for the incremental forward and incremental adjoint PDEs in the context of inverse acoustic wave propagation will be given in section 6. Solvers for the incremental forward and adjoint problems, of course, inherit the scalability of the forward PDE solver. The other major cost in computing the action of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ on a vector is the application of the square root of the prior, $\boldsymbol{\Gamma}^{1/2}_{\text{prior}}$, to a vector. As discussed in section 2.3, this amounts to solving a Laplacian-like problem. Using a scalable elliptic solver such as multigrid renders this component scalable as well. Therefore, the scalability of the application of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ to a vector—which is the basic operation of a matrix-free eigenvalue solver such as Lanczos—is assured, and the cost is independent of the parameter dimension.

The remaining requirement for independence of parameter dimension in the construction of the low rank–based representation of the posterior covariance in (5.5) is that the number of dominant eigenvalues of $\boldsymbol{H}_{\text{misfit}}$ be independent of the dimension of the discretized parameter. This is the case when $\boldsymbol{H}_{\text{misfit}}$ and $\boldsymbol{\Gamma}_{\text{prior}}$ in (5.1) are discretizations of a compact and a continuous operator, respectively. The continuity of $\mathcal{C}_0$ is a direct consequence of the prior Gaussian measure μ_0; in fact, $\mathcal{C}_0$, the infinite-dimensional counterpart of $\boldsymbol{\Gamma}_{\text{prior}}$, is also a compact operator. Compactness of the data misfit Hessian $\boldsymbol{H}_{\text{misfit}}$ for inverse wave propagation problems has long been observed (e.g., [15]) and, as mentioned above, has been proved for frequency-domain acoustic inverse scattering for both continuous and pointwise observation operators [10, 11]. Specifically, we have shown that the data misfit Hessian is a compact operator at any point in the parameter domain. We also quantify the decay of the data misfit Hessian eigenvalues in terms of the smoothness of the medium, i.e., the smoother it is the faster the decay rate. For an analytic target medium, the rate can be shown to be exponential. That is, the data misfit Hessian can be approximated well with a small number of its dominant eigenvectors and eigenvalues.

As a result, the number of Lanczos iterations required to obtain a low rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ is independent of the dimension of the discretized parameter field. Once the low rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ is constructed, no additional forward or adjoint PDE solves are required. Any action of $\boldsymbol{\Gamma}_{\text{post}}$ in (5.5) on a vector (which is required to generate samples from the posterior distribution and to compute the diagonal of the covariance) is now dominated by the action of $\boldsymbol{\Gamma}_{\text{prior}}$ on a vector.

But as discussed above, this amounts to an elliptic solve and can be readily carried out in a scalable manner. Since r is independent of the dimension of the discretized parameter field, estimating the posterior covariance matrix requires a constant number of forward/adjoint PDE solves, independent of the number of parameters, observations, and state variables. Moreover, since the dominant cost is that of solving forward and adjoint PDEs as well as elliptic problems representing the prior, scalability of the overall uncertainty quantification method follows when the forward and adjoint PDE solvers are scalable.

6. Application to global seismic statistical inversion. In this section, we apply the computational framework developed in the previous sections to the statistical inverse problem of global seismic inversion, in which we seek to reconstruct the heterogeneous compressional (acoustic) wave speed from observed seismograms, i.e., seismic waveforms recorded at points on the earth's surface. With the rapid advances in observational capabilities, exponential growth in supercomputing, and maturation of forward seismic wave propagation solvers, there is great interest in solving the global seismic inverse problem governed by the full acoustic or elastic wave equations [19,37]. Already, successful deterministic inversions have been carried out at regional scales; for example, see [20, 21, 34, 43, 48].

We consider global seismic model problems in which the seismic source is taken as a simple point source. Sections 6.1 and 6.2 define the prior mean and covariance operator for the wave speed and its discretization. Section 6.3 presents the parameter-to-observable map $\boldsymbol{f}(m)$ (which involves solution of the acoustic wave equation) and the likelihood model. We next provide the expressions for the gradient and application of the Hessian of the negative log-likelihood in section 6.4. Then, we discuss the discretization of the forward and adjoint wave equations and implementation details in section 6.5. Section 6.6 provides the inverse problem setup, while numerical results and discussion are provided in sections 6.7 and 6.8.

6.1. Parameter space for seismic inversion. We are interested in inferring the heterogeneous compressional acoustic wave speed in the earth. In order to do this, we represent the earth as a sphere of radius 6,371km. We employ two earth models, i.e., two representations of the compressional wave speed and density in the earth. We suppose that our current knowledge of the earth is given by the spherically symmetric preliminary reference earth model (PREM) [16], which is depicted in Figure 6.1.

The PREM is used as the mean of the prior distribution, and as the starting point for the determination of the MAP point by the optimization solver. Then, we presume that the real earth behaves according to the S20RTS velocity model [38], which superposes lateral wave speed variations on the laterally homogeneous PREM. S20RTS is used to generate waveforms used as synthetic observational data for inversion; we refer to it as the "ground truth" earth model. The inverse problem then aims to reconstruct the S20RTS ground truth model from the (noisy) synthetic data and from prior knowledge of the PREM, and quantify the uncertainty in doing so.

The parameter field m of interest for the inverse problem is the deviation or anomaly from the PREM, and hence it is sensible to choose the zero function as the prior mean. Owing to the prior covariance operator specified in section 2.3, the deviation is smooth; in fact it is continuous almost surely. The wave speed parameter space is discretized using continuous isoparametric trilinear finite elements on a hexahedral octree-based mesh. To generate the mesh, we partition the earth into 3 layers described by 13 mapped cubes. The first layer consists of a single cube surrounded

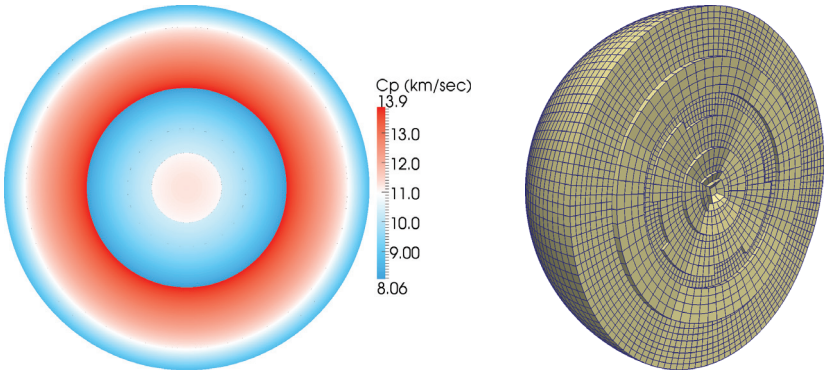

FIG. 6.1. *Left image: Cross section through the spherically symmetric PREM, which is the prior mean in the inversion. Right image: Mesh used for both wave speed parameters (discretized with $N = 1$) and wave propagation unknowns ($N = 3$). The mesh is tailored to the local wavelengths.*

by two layers of six mapped cubes. The resulting mesh is aligned with the interface between the outer core and the mantle, where the wave speed has a significant discontinuity (see Figure 6.1). It is also aligned with several known weaker discontinuities between layers.

The parameter mesh coincides with the mesh used to solve the wave equation described in section 6.3. The mesh is locally refined to resolve the local seismic wavelength resulting from a given frequency of interest for the PREM. We choose a conservative number of grid points per wavelength to permit the same mesh to be used for anticipated variations in the earth model across the iterations needed to determine the MAP point. For the parallel mesh generation and its distributed storage, we use fast forest-of-octree algorithms for scalable adaptive mesh refinement from the `p4est` library [12, 13].

6.2. The choice of prior. Since the prior is a Gaussian measure, it is completely specified by a mean function and a covariance operator. As discussed in section 6.1, the prior distribution for the anomaly (the deviation of the acoustic wave speed from that described by the PREM model) is naturally chosen to have zero mean. The choice of covariance operator for the prior distribution has to encode several important features. Recall that we specify the covariance operator via the precision operator $\mathcal{A}$ in section 2.3. Therefore, the size of the variance about the zero mean is set by α, while the product $\alpha\Theta$ determines the correlation length of the prior Gaussian random field. We next specify the scalar α and the tensor Θ based on the following observations of models for the local wave speeds in the earth:

- Smoothness. The parameter field describes the *effective* local wave speed, which, for a finite source frequency, depends on the local average of material parameters within a neighborhood of each point in space. This makes the effective wave speed mostly a smooth field. Note that the S20RTS-based target wave speed model (see [38]) is smooth.
- Prior variance. The deviation in this effective wave speed from the PREM model is believed to be within a few percent. Thus, we select α such that the prior standard deviation is about 3.5%. The S20RTS target model has a maximal deviation from PREM of 7%.
- Anisotropy in the mantle. We further incorporate the prior belief that the

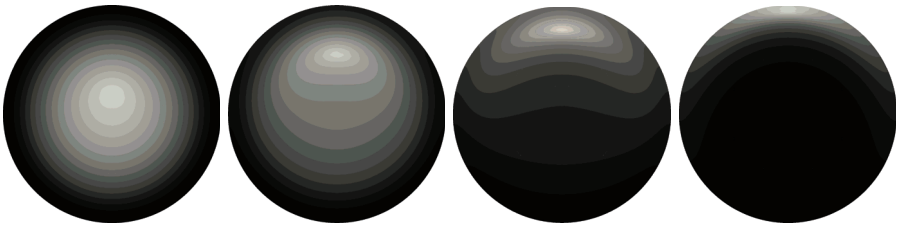

FIG. 6.2. *Contours of Green's functions at points in different depths for the precision operator of our prior $\mathcal{A}^2$. The contours are shown in slices through the earth that contain the points, and larger values of the Green's functions correspond to brighter shades of gray. These Green's functions correspond directly to the covariance function $c(\boldsymbol{x}, \boldsymbol{y})$ as discussed in section 2. Note the anisotropy for points closer to the surface.*

compressional wave speed has a stronger variation in depth than in the lateral directions. We encode this anisotropic variation through Θ. In particular, we select Θ such that the anisotropy is strongest near the surface, and gradually becomes weaker with higher overall correlation length at larger depths. We observe that the S20RTS target model also obeys a similar anisotropy.

From the preceding observations and discussion, we choose $\alpha = 1.5 \cdot 10^{-2}$, while Θ is chosen to have the following form:

$$\Theta = \beta \left(\boldsymbol{I}_3 - \theta(\boldsymbol{x}) \boldsymbol{x} \boldsymbol{x}^T \right) \quad \text{with } \theta(\boldsymbol{x}) := \begin{cases} \frac{1-\theta}{r\|\boldsymbol{x}\|^2} \left(2\|\boldsymbol{x}\| - \frac{1}{r}\|\boldsymbol{x}\|^2 \right) & \text{if } \|\boldsymbol{x}\| \neq 0, \\ 0 & \text{if } \|\boldsymbol{x}\| = 0, \end{cases} \tag{6.1}$$

where $\boldsymbol{I}_3$ is the 3×3 identity matrix, $r = 6,371$km is the earth radius, $\beta = 7.5 \cdot 10^{-3} r^2$, and $\theta = 4 \cdot 10^{-2}$. The choice $0 < \theta < 1$ introduces anisotropy in Θ such that the prior assumes longer correlation lengths in tangential than in radial directions, and the anisotropy decreases smoothly toward the center of the sphere. In Figure 6.2 we show several Green's functions for the precision operator $\mathcal{A}^2$, which illustrate this anisotropy. Figure 6.3 shows a slice through the $\pm 2\sigma$ fields, through samples from the prior and through the ground truth model, which is used to generate the synthetic seismograms. Note that close to the boundary, the standard deviation of the prior becomes larger. This is partly a result of the anisotropy in the differential operator used in the construction of the prior, but mainly an effect of the homogeneous Neumann boundary condition used in the construction of the square root of the prior. This larger variance close to the boundary is also reflected in the prior samples, which have their largest values close to the boundary. Note that these samples have a larger correlation length in tangential than in normal directions, as intended by the choice of the anisotropy in (6.1). The ground truth model, which is also shown in Figure 6.3, is comparable to realizations of the prior in terms of magnitude as well as correlation.

6.3. The likelihood. In this section, we construct the likelihood (2.4) for the inverse acoustic wave problem. In order to do this, we need to construct the parameter-to-observable map $\boldsymbol{f}(m)$ and the observations $\boldsymbol{y}^{\text{obs}}$. Let us start by considering the acoustic wave equation written in the first order form,

$$\rho \frac{\partial \boldsymbol{v}}{\partial t} - \nabla(\rho c^2 e) = \boldsymbol{g}, \tag{6.2a}$$

$$\frac{\partial e}{\partial t} - \nabla \cdot \boldsymbol{v} = 0, \tag{6.2b}$$

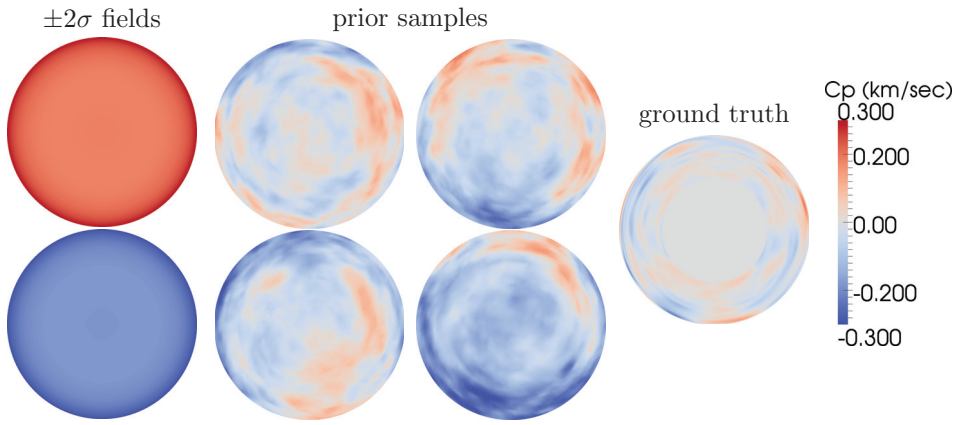

FIG. 6.3. *For illustration, we visualize several depictions of the prior using a common color scale. The images on the far left show slices through the pointwise positive and negative 2σ-deviation fields, which bound the pointwise 95% credible interval. The second and third columns show samples drawn from the prior distribution, while the fourth column depicts the "ground truth" parameter field. The prior has been chosen so that samples display similar qualitative features to the "ground truth" medium; they exhibit anisotropy in the outer layers of the mantle with larger correlation lengths in the lateral directions, and become more isotropic with higher overall correlation at depth.*

where $\rho = \rho(\boldsymbol{x})$ denotes the mass density, $c = c(\boldsymbol{x})$ the local acoustic wave speed, $\boldsymbol{g}(\boldsymbol{x}, t)$ a (smoothed) point source $\boldsymbol{x} \in \Omega$, $\boldsymbol{v}(\boldsymbol{x}, t)$ the velocity, and $e(\boldsymbol{x}, t)$ the trace of the strain tensor, i.e., the dilatation. We equip (6.2) with the initial conditions

$$e(\boldsymbol{x}, 0) = e_0(\boldsymbol{x}) \quad \text{and} \quad \boldsymbol{v}(\boldsymbol{x}, 0) = \boldsymbol{v}_0(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega, \tag{6.2c}$$

together with the boundary condition

$$e(\boldsymbol{x}, t) = 0, \quad \boldsymbol{x} \in \Gamma = \partial\Omega, t \in (0, T). \tag{6.2d}$$

Here, the acoustic wave initial-boundary value problem (6.2) is a simplified mathematical model for seismic waves propagation in the earth [4]. The choice of strain dilatation e together with velocity $\boldsymbol{v}$ in the first order system formulation is motivated from the strain-velocity formulation for the elastic wave equation used in [47].

Our goal is to quantify the uncertainty in inferring the spatially varying wave speed $m = c(\boldsymbol{x})$ from waveforms observed at receiver locations. In order to define the parameter-to-observable map for a given wave speed $c(\boldsymbol{x})$, we first solve the acoustic wave equation (6.2) given c, and record the velocity $\boldsymbol{v}$ at a finite number of receivers in the time interval $(0, T)$. Finally, we compute the Fourier coefficients of the seismograms and truncate them; the truncated coefficients are the observables in the map $\boldsymbol{f}(m)$. A similar procedure is used to generate synthetic seismograms to define $\boldsymbol{y}^{\text{obs}}$. The noise covariance matrix $\boldsymbol{\Gamma}_{\text{noise}}^{-1}$ is prescribed as a diagonal matrix with constant variance.

6.4. Gradient and Hessian of the negative log posterior. Our proposed method for uncertainty quantification in section 3 requires the computation of derivatives of the negative log posterior, which, in turn, requires the gradient and Hessian of the likelihood and the prior. These derivatives can be computed efficiently using an adjoint method, as we now show. For clarity, we derive the gradient and action of the Hessian in an infinite-dimensional setting. Let us begin by denoting $\boldsymbol{v}(c)$ as the

space-time solution of the wave equation given the wave speed $c = c(\boldsymbol{x})$, and $\mathcal{B}$ as the observation operator. The parameter-to-observable map $\boldsymbol{f}(c)$ can be written as $\mathcal{B}\boldsymbol{v}(c)$. Thus, the negative log posterior is (compare with (2.9))

$$\mathcal{J}(c) := \frac{1}{2}\left\|\mathcal{B}\boldsymbol{v}(c) - \boldsymbol{y}^{\text{obs}}\right\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \frac{1}{2}\left\|\mathcal{A}(c - c_0)\right\|^2_{L^2(\Omega)}, \tag{6.3}$$

where $\boldsymbol{\Gamma}_{\text{noise}}$ is specified as in section 6.6. The dependence on the wave speed c of the velocity $\boldsymbol{v}$ and dilatation e is given by the solution of the forward wave propagation equation (6.2). The adjoint approach [17] allows us to write $\mathcal{G}(c)$, the gradient of $\mathcal{J}$ at a point c in parameter space, as

$$\mathcal{G}(c) := 2\rho c \int_0^T e(\nabla \cdot \boldsymbol{w})\, dt + \mathcal{A}^2(c - c_0), \tag{6.4}$$

where the adjoint velocity $\boldsymbol{w}$ and adjoint strain dilatation d satisfy the *adjoint wave propagation terminal-boundary value problem*

$$-\rho\frac{\partial \boldsymbol{w}}{\partial t} + \boldsymbol{\nabla}(c^2\rho d) = -\mathcal{B}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}(\mathcal{B}\boldsymbol{v} - \boldsymbol{y}^{\text{obs}}) \quad \text{in } \Omega \times (0,T), \tag{6.5a}$$

$$-\frac{\partial d}{\partial t} + \nabla \cdot \boldsymbol{w} = 0 \quad \text{in } \Omega \times (0,T), \tag{6.5b}$$

$$\rho\boldsymbol{w} = \boldsymbol{0}, d = 0 \quad \text{in } \Omega \times \{t = T\}, \tag{6.5c}$$

$$d = 0 \quad \text{on } \Gamma \times (0,T). \tag{6.5d}$$

Here, $\mathcal{B}^*$, an operator from $\mathbb{R}^q$ to the space-time cylinder $\Omega \times (0,T)$, is the adjoint of $\mathcal{B}$. Note that the adjoint wave equations must be solved backward in time (due to final time data) and have the data misfit as a source term, but otherwise resemble the forward wave equations.

Similar to the computation of the gradient, the Hessian operator of $\mathcal{J}$ at c acting on an arbitrary variation $\tilde{c}$ is given by

$$\mathcal{H}(c)\tilde{c} := 2\rho \int_0^T ce(\nabla \cdot \tilde{\boldsymbol{w}}) + c\tilde{e}(\nabla \cdot \boldsymbol{w}) + \tilde{c}e(\nabla \cdot \boldsymbol{w})\, dt + \mathcal{A}^2\tilde{c}, \tag{6.6}$$

where $\tilde{\boldsymbol{v}}$ and $\tilde{e}$ satisfy the *incremental forward wave propagation initial-boundary value problem*

$$\rho\frac{\partial \tilde{\boldsymbol{v}}}{\partial t} - \nabla(\rho c^2\tilde{e}) = \nabla(2\rho c\tilde{c}e) \quad \text{in } \Omega \times (0,T),$$

$$\frac{\partial \tilde{e}}{\partial t} - \nabla \cdot \tilde{\boldsymbol{v}} = 0 \quad \text{in } \Omega \times (0,T),$$

$$\rho\tilde{\boldsymbol{v}} = \boldsymbol{0}, \tilde{e} = 0 \quad \text{in } \Omega \times \{t = 0\},$$

$$\tilde{e} = 0 \quad \text{on } \Gamma \times (0,T).$$

On the other hand, $\tilde{\boldsymbol{w}}$ and $\tilde{d}$ satisfy the *incremental adjoint wave propagation terminal-boundary value problem*

$$-\rho\frac{\partial \tilde{\boldsymbol{w}}}{\partial t} + \nabla(c^2\rho\tilde{d}) = -\nabla(2\tilde{c}c\rho d) - \mathcal{B}^*\boldsymbol{\Gamma}^{-1}_{\text{noise}}\mathcal{B}\tilde{\boldsymbol{v}} \quad \text{in } \Omega \times (0,T),$$

$$-\frac{\partial \tilde{d}}{\partial t} + \nabla \cdot \tilde{\boldsymbol{w}} = 0 \quad \text{in } \Omega \times (0,T),$$

$$\rho\tilde{\boldsymbol{w}} = \boldsymbol{0}, \tilde{d} = 0 \quad \text{in } \Omega \times \{t = T\},$$

$$\tilde{d} = 0 \quad \text{on } \Gamma \times (0,T).$$

As can be seen, the incremental forward and incremental adjoint wave equations are linearizations of their forward and adjoint counterparts, and thus differ only in the source terms. Moreover, we observe that the computation of the gradient and the Hessian action amounts to solving a pair of forward/adjoint and a pair of incremental forward/incremental adjoint wave equations, respectively. For our computations, we use the Gauss–Newton approximation of the Hessian, which is guaranteed to be positive. This amounts to neglecting the terms that contain $\nabla \cdot \boldsymbol{w}$ in (6.6), and neglecting the term that includes d in the incremental adjoint wave equations.

6.5. Discretization of the wave equation and implementation details. We use the same hexahedral mesh to compute the wave solution $(\boldsymbol{v}, e)$ as is used for the parameter c. While the parameter is discretized using trilinear finite elements, the wave equation, and its three variants (the adjoint, the incremental forward, and the incremental adjoint), are solved using a high order discontinuous Galerkin (dG) method. The method, for which details are provided in [9, 47], supports hp-nonconforming discretization, but only h-nonconformity is used in our implementation. For efficiency and scalability, a tensor product of Lagrange polynomials of degree N (we use $N \in \{2, 3, 4\}$ for the examples in the next section) is employed together with a collocation method based on Legendre–Gauss–Lobatto (LGL) nodes. As a result, the mass matrix is diagonal, which facilitates time integration using the classical four-stage fourth order Runge–Kutta method. We equip our dG method with exact Riemann numerical fluxes at element faces. To treat the nonconformity, we use the mortar approach of Kopriva and coworkers [30, 31] to replace nonconforming faces by mortars that connect pairs of contributing elements. The actual computations are performed on the mortars instead of the nonconforming faces, and the results are then projected onto the contributing element faces. The method has been shown to be consistent, stable, and convergent with optimal order, and highly scalable [9, 47].

It should be pointed out that the discretizations of the gradient and Hessian action given in section 6.4 are not consistent with the discrete gradient and Hessian-vector product obtained by first discretizing the negative log posterior and then differentiating it. Here, inconsistency means that the former are equivalent to the latter only in the limit as the mesh size approaches zero (see also [26, 28]). The reason is that additional jump terms due to numerical fluxes at element interfaces are introduced in the dG discretization of the wave equation. In our implementation, we include these terms to ensure consistency, and this is verified by comparing the discretized gradient and Hessian action expressions with their finite difference approximations.

Moreover, since we use a continuous Galerkin finite element method for the parameter, but a dG method for the wave solution, it is necessary to prolongate the parameter to the solution space before solving the forward wave equation, and its variants (adjoint, incremental state, incremental adjoint). Conversely, the gradient and the Hessian-vector application are computed in the wave solution space, and then restricted to the parameter space to provide the correct derivatives for the optimization solver. To ensure the symmetry of the Hessian, we construct these restriction and prolongation operations such that they are adjoints of each other.

Our discretization approach for the Bayesian inverse problem in section 3 requires the repeated application of $\boldsymbol{A}^{-1}$, each amounting to an elliptic PDE solve on the finite-dimensional parameter space. To accomplish this task efficiently, we use the parallel algebraic multigrid (AMG) solver *ML* from the Trilinos project [23]. The cost of this elliptic solve is negligible compared to that of solving the time-dependent seismic wave equations, which employ high order discretization in contrast to the trilinear

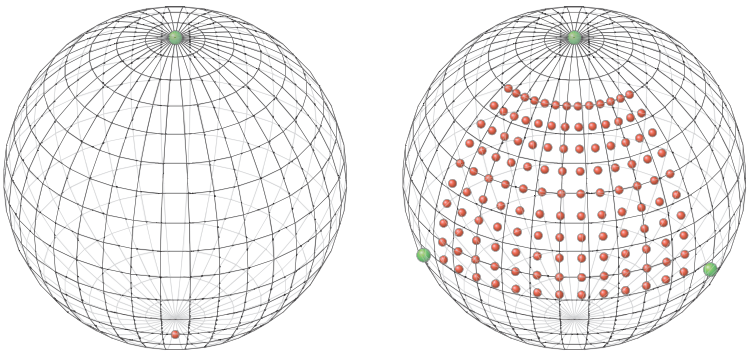

FIG. 6.4. *Location of sources (green) and receivers (red) for Problem* I *(left) and Problem* II *(right).*

discretization of the anisotropic Poisson operator, $\boldsymbol{A}$.

The adjoint equation has to be solved backwards-in-time (as shown in section 6.4); computation of the gradient (6.4) requires combinations of the state and adjoint solutions corresponding to the same time. Thus the gradient computation requires the complete time history of the forward solve, which cannot be stored due to the large-scale nature of our problem. A similar, but slightly more challenging storage problem occurs in the Hessian-vector application. Here, solving the incremental state equation requires the solution of the state equation, and the incremental adjoint solution requires the solution of the incremental state equation. We avoid storage of the time history of these wavefields by using a checkpointing method as employed in [17]. This scheme reduces the necessary storage at the expense of increasing the number of wave propagation solves.

Between 1,200 and 4,096 processor cores[3] for 10–20 hours are needed to solve the seismic inverse problems presented in the next section. The vast majority of the runtime is spent on computing solutions of the forward, adjoint, and incremental wave equations either for the computation of the MAP point (see section 4) or the Lanczos iterations for computing the low rank approximation of the misfit Hessian (see section 6.7). Due to the large number of required wave propagation solves, good strong scalability of the wave propagation solver is important for rapid turnaround. We refer to the discussion in [8] on the scalability of the wave propagation solver, as well to the overall scalability of our Bayesian inversion approach applied to seismic inverse problems of up to one million parameters.

6.6. Setup of model problems. Synthetic observations $\boldsymbol{y}^{\text{obs}}$ are generated from solution of the wave equation using the S20RTS earth model. To mitigate the inverse crime [29], the local wave speed on LGL nodes of the wave propagation mesh is used to generate the observations, which implies that a higher order approximation of the earth model is used to generate the synthetic data, but the inversion is carried out on a lower order mesh. Both sources and receivers are located at 10km depth from the earth surface. For the source term $\boldsymbol{g}$ in (6.2), we use a delta function point source in the z-direction convolved with a narrow Gaussian in space. In time, we employ a Gaussian with standard deviation of $\sigma = 20$s centered at 60s. The wave

[3]These computations were performed on the Texas Advanced Computing Center's Lonestar 4 system, which has 22,656 Westmere processor cores with 2GB memory per core.

propagation mesh (i.e., the discretization of velocity and dilatation) is chosen fine enough to accurately resolve frequencies below 0.05Hz. We Fourier transform the (synthetic) observed velocity waveforms at each receiver location and retain only the first 101 Fourier modes to define the observations $\boldsymbol{y}^{\text{obs}}$. In our problems, the Fourier coefficients $\boldsymbol{y}^{\text{obs}}$ vary between 10^{-5} and 10^{-1}, and we choose for the noise covariance a diagonal matrix with a standard deviation of 0.002.

We consider the following two model problems:

- **Problem I:** The first problem has a single source at the North Pole and a single receiver at 45° south of the equator, as illustrated in the left image of Figure 6.4. The wave propagation time is 1,800s. The wave speed (i.e., unknown material parameter) field is discretized on a mesh of trilinear hexahedra with 78,558 nodes, representing the unknowns in the inverse problem. The forward problem is discretized on the same mesh with third order dG elements, resulting in about 21.4 million spatial wave propagation unknowns, and in 2,100 four-stage, fourth order Runge–Kutta time steps.
- **Problem II:** The second problem uses 130 receivers distributed on a quarter of the Northern Hemisphere along zonal lines with 7.5° spacing and three simultaneous sources as shown on the right of Figure 6.4. The wave propagation time is 1,200s. The wave speed is discretized on three different trilinear hexahedral meshes with 40,842, 67,770, and 431,749 wave speed parameters, which represent the unknowns in the inverse problem. These meshes correspond to discretizations with fourth, third, and second order discontinuous elements for the wave propagation variables (velocity and dilatation). The results in the next section were computed with 67,770 wave speed parameters and the third order dG discretization for velocity and dilatation. This amounts to 18.7 million spatial wave propagation unknowns, and 1,248 Runge–Kutta time steps.

6.7. Low rank approximation of the prior-preconditioned misfit Hessian. Before discussing the results for the quantification of the uncertainty in the solution of our inverse problems, we numerically study the spectrum of the prior-preconditioned misfit Hessian. In Figure 6.5, we show the dominant spectrum of the prior-preconditioned Hessian evaluated at the MAP estimate for Problem I (left) and Problem II (right). As can be observed, the eigenvalues decay faster in the former than in the latter. That is, the former is more ill-posed than the latter. The reason is that the three simultaneous sources and 130 receivers of Problem II provide more information on the earth model. This implies that retaining more eigenvalues is necessary to accurately approximate the prior-preconditioned Hessian of the data misfit for Problem II compared to Problem I. In particular, we need at least 700 eigenvalues for Problem II as compared to about 40 for Problem I to obtain a sensible low rank approximation of the Hessian, and this constitutes the bulk of computation time (since each Hessian-vector product in the Lanczos solver requires incremental forward and adjoint wave propagation solutions). These numbers compare with a total number of parameters of 78,558 (Problem I) and 67,770 (Problem II), which amounts to a reduction of between two and three orders of magnitude. This directly translates into 2–3 orders of magnitude reduction in cost of solving the statistical inverse problem.

Figure 6.5 presents the spectra for Problem II for three different discretization of the wave speed parameter field. The figure suggests that the dominant spectrum is essentially mesh-independent and that all three parameter meshes are sufficiently fine to resolve the dominant eigenvectors of the prior-preconditioned Hessian. Consequently,

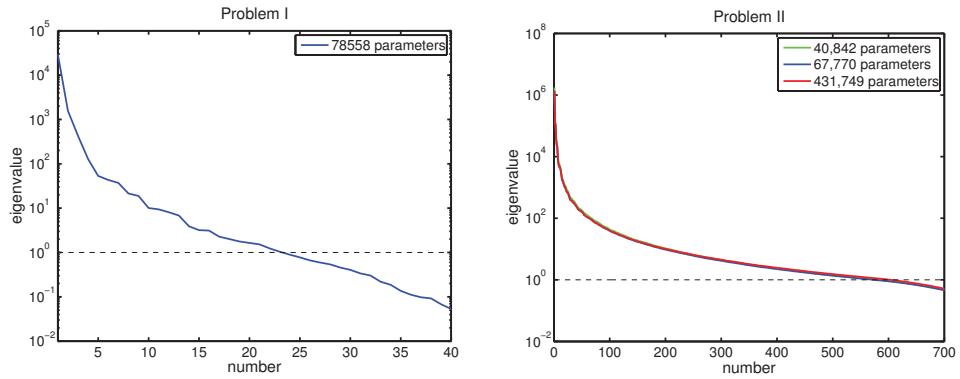

FIG. 6.5. *Logarithmic plot of the spectrum of prior-preconditioned data misfit Hessian for Problem* I *(left) and Problem* II *(right). The computation of the approximate spectrum for Problem* I *uses a discretization with* $78,558$ *wave speed parameters, third order dG finite elements for the wave propagation solution, and* 50 *Lanczos iterations. The spectrum for Problem* II *is computed on different discretizations of the parameter mesh using* 900 *Lanczos iterations. The eigenvalues for the three discretizations essentially lie on top of each other, which illustrates that the underlying infinite-dimensional statistical inverse problem is properly approximated. The horizontal line* $\lambda = 1$ *shows the reference value for the truncation of the spectrum of the misfit Hessian. For an accurate approximation of the posterior covariance matrix (i.e., the inverse of the Hessian), eigenvalues that are small compared to* 1 *can be neglected as discussed in section* 5*, and in particular as shown in* (5.4).

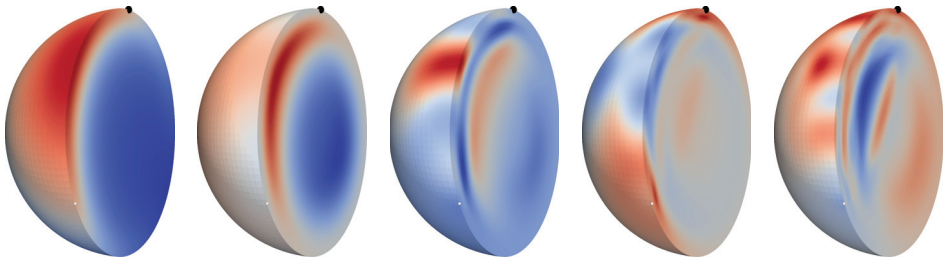

FIG. 6.6. *Problem* I: *Eigenvectors of the prior-preconditioned misfit Hessian corresponding to the first (i.e., the largest), the third, the fifth, eighth, and thirteenth eigenvalues (from left to right). The visualization employs a slice through the source and receiver locations.*

the Hessian low rank approximation, particularly the number of Lanczos iterations, is independent of the number of discrete parameters. Thus, in this example, the number of wave propagation solutions required by the low rank approximation does not depend on the parameter dimension.

Figures 6.6 and 6.7 show several eigenvectors of the prior-preconditioned data misfit Hessian (5.1) (corresponding to several dominant eigenvalues) for Problems I and II. Eigenvectors corresponding to dominant eigenvalues represent the earth modes that are "most observable" from the data, given the configuration of sources and receivers. As can be seen in these figures, the largest eigenvalues produce the smoothest modes, and as the eigenvalues decrease, the associated eigenvectors become more oscillatory, due to the reduced ability to infer smaller length scales from the observations.

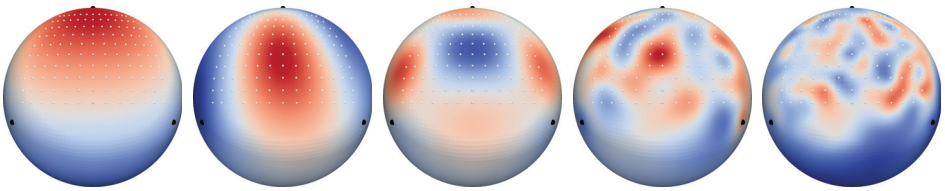

FIG. 6.7. *Problem* II: *Eigenvectors of the misfit Hessian corresponding to eigenvalues* 1, 5, 20, 100, *and* 350, *respectively. Note that the lower modes are smoothest and become more oscillatory with increasing mode number.*

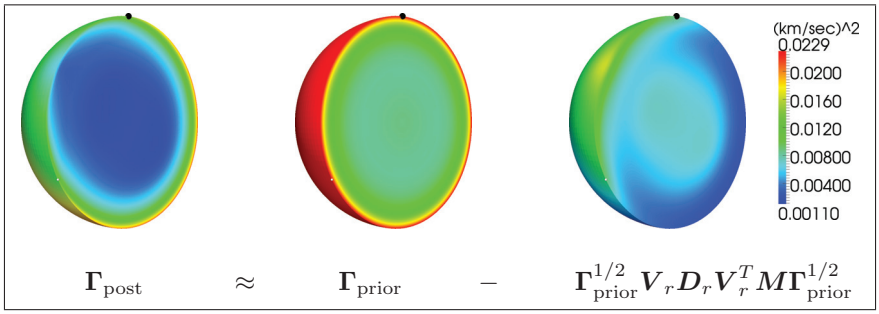

FIG. 6.8. *Problem* I: *The left image depicts the pointwise posterior variance field, which is represented as the difference between the original prior variance field (middle), and the reduction in variance due to data (right). The locations of the single source and single receiver are shown by the black and white dot, respectively. Color scale is common to all three images.*

6.8. Interpretation of the uncertainty in the solution of the inverse problem. We first study Problem I, i.e., the single source, single receiver problem. Since the data are very sparse, it is expected that we can reconstruct only very limited information from the truth solution; this is reflected in the smoothness of the dominant eigenmodes shown in Figure 6.6. To assess the uncertainty, Figure 6.8 shows prior variance, knowledge gained from the data (i.e., reduction in the variance), and posterior variance, which are computed from (5.7). As discussed in section 5, the posterior is the combination of the prior information and the knowledge gained from the data, so that the posterior uncertainty is decreased relative to the prior uncertainty. That is, the inference has less uncertainty in regions for which the data are more informative. In particular, the region of lowest uncertainty is at the surface half-way between source and receiver, as Figure 6.8 shows. Note that the data are also informative about the core-mantle boundary, where the strong material contrast results in stronger reflected energy back to the surface receivers, allowing greater confidence in the properties of that interface.

Next, we study the results for Problem II. The comparison between the MAP estimate and the ground truth earth model (S20RTS) at different depths is displayed in Figure 6.9. As can be seen, we are able to recover accurately the wave speed in the portion of the Northern Hemisphere covered by sources and receivers. We plot the prior and posterior pointwise standard deviations in Figure 6.10. One observes that the uncertainty reduction is greatest along the wave paths between sources and receivers, particularly in the quarter of the Northern Hemisphere surface where the receivers are distributed.

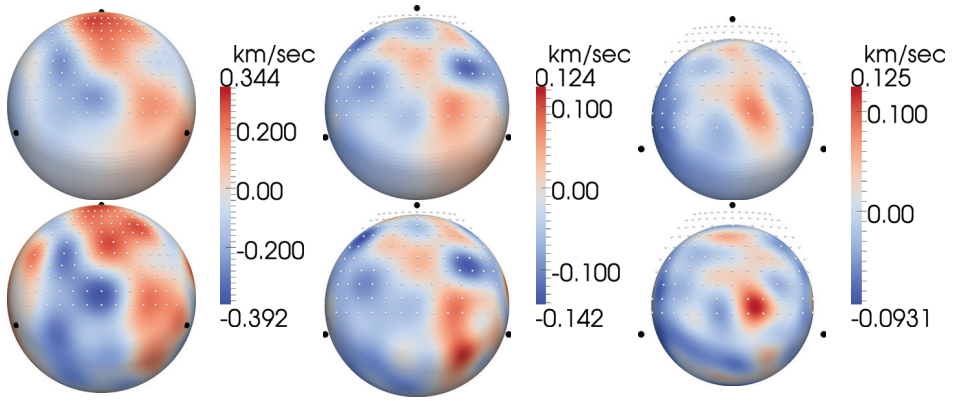

FIG. 6.9. *Comparison of MAP of posterior pdf (upper row) with the "truth" earth model (lower row) in a depth of* 67*km (left image),* 670*km (middle image), and* 1,340*km (right image). The colormap varies with depth, but is held constant between the MAP and "truth" images at each depth.*

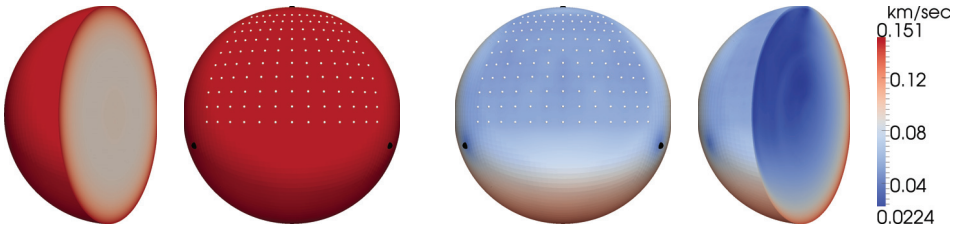

FIG. 6.10. *This figure compares the pointwise standard deviation for the prior (left) and posterior (right) distributions at a depth of* 67*km. The color indicates one standard deviation, and the scale is common to both prior and posterior images. We observe that the most reduction in variance due to data occurs in the region near sources and receivers, whereas the least reduction occurs on the opposite side of the earth.*

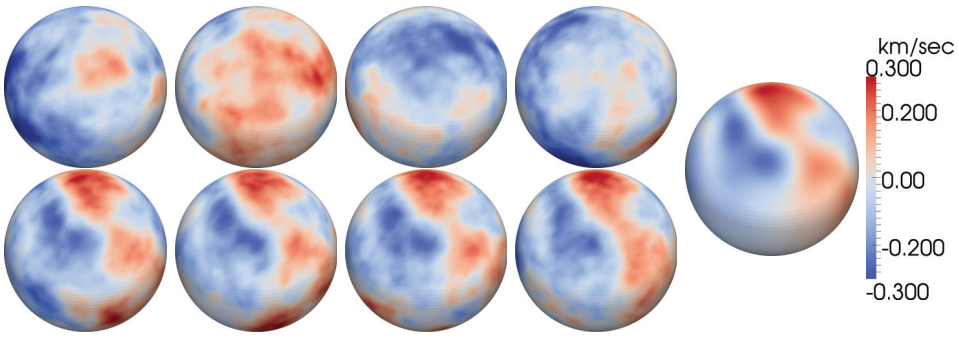

FIG. 6.11. *Samples from the prior (top row) and posterior (bottom row) distributions. The prior scaling was chosen such that the "ground truth" S*20*RTS would be a qualitatively reasonable sample from the prior distribution. For comparison purposes, the MAP estimate is shown on the far right.*

In Figure 6.11, we show a comparison between samples from the prior distribution and from the posterior. We observe that in the quarter of the Northern Hemisphere where the data are more informative about the medium, we have a higher degree of confidence about the wave speed, which is manifested in the common large-scale features across the posterior samples. The fine-scale features (about which the data are least informative) are qualitatively similar to those of the prior distribution, and vary from sample to sample in the posterior. We note that the samples shown here are computed by approximating $\boldsymbol{M}^{-1/2}$ in expression (5.8) using the (diagonal) lumped mass matrix to avoid computing a factorization of $\boldsymbol{M}$. If desired, this mass lumping can be avoided by applying $\boldsymbol{M}^{-1/2}$ to a vector using an iterative scheme based on polynomial approximations to the matrix function $f(t) = t^{-1/2}$, as in [14].

7. Conclusions. A computational framework for estimating the uncertainty in the numerical solution of linearized infinite-dimensional statistical inverse problems is presented. We adopt the Bayesian inference formulation: Given observational data and their uncertainty, the governing forward problem and its uncertainty, and a prior probability distribution describing uncertainty in the parameter field, find the posterior probability distribution over the parameter field. The framework, which builds on the infinite-dimensional formulation proposed by Stuart [42], incorporates a number of components aimed at ensuring a convergent discretization of the underlying infinite-dimensional inverse problem. It additionally incorporates algorithms for manipulating the prior, constructing a low rank approximation of the data-informed component of the posterior covariance operator, and exploring the posterior, that together ensure scalability of the entire framework to very high parameter dimensions. Since the data are typically informative about only a low-dimensional subspace of the parameter space, the Hessian is sparse with respect to some basis. We have exploited this fact to construct a low rank approximation of the Hessian and its inverse using a parallel matrix-free Lanczos method. Overall, our method requires a dimension-independent number of forward PDE solves to approximate the local covariance. Uncertainty quantification for the linearized inverse problem thus reduces to solving a fixed number of forward and adjoint PDEs (which resemble the original forward problem), independent of the problem dimension. The entire process is thus scalable with respect to the forward problem dimension, uncertain parameter dimension, and observational data dimension. We applied this method to the Bayesian solution of an inverse problem in 3D global seismic wave propagation with up to 430,000 parameters, for which we observe 2–3 orders of magnitude dimension reduction, making uncertainty quantification for large-scale inverse problems tractable.

Appendix. In the following, we provide a constructive derivation of $\boldsymbol{L}$ in (5.8) such that it satisfies $\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{L}\boldsymbol{L}^{\diamond}$. Our goal is to draw a posterior Gaussian random sample with covariance matrix $\boldsymbol{\Gamma}_{\text{post}}$ in $\mathbb{R}^n_{\boldsymbol{M}}$. To accomplish this, a standard approach is first to find a factorization $\boldsymbol{\Gamma}_{\text{post}} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^{*}$, where $\tilde{\boldsymbol{L}}$ is a linear map from $\mathbb{R}^n_{\boldsymbol{M}}$ to $\mathbb{R}^n_{\boldsymbol{M}}$. Then, any random sample from the posterior can be written as

$$\boldsymbol{\nu}^{\text{post}} = \boldsymbol{m}_{\text{MAP}} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{n}}, \tag{A.1}$$

where $\tilde{\boldsymbol{n}}$ is a Gaussian random sample with zero mean and identity covariance matrix in $\mathbb{R}^n_{\boldsymbol{M}}$. It follows that

$$\tilde{\boldsymbol{n}} = \boldsymbol{M}^{-1/2}\boldsymbol{n},$$

where $\boldsymbol{n}$ is the standard Gaussian random sample with zero mean and identity covariance matrix in $\mathbb{R}^n$, i.e., $\boldsymbol{n} \sim \mathcal{N}(0, \boldsymbol{I})$, and $\boldsymbol{M}^{-1/2}$ a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n_{\boldsymbol{M}}$.

Therefore, what remains to be done is to construct $\tilde{\boldsymbol{L}}$. To begin the construction, we rewrite (5.6) as

$$\boldsymbol{\Gamma}_{\text{post}} \approx \boldsymbol{\Gamma}_{\text{prior}}^{1/2} \underbrace{\left(\boldsymbol{I} - \boldsymbol{V}_r \boldsymbol{D}_r \boldsymbol{V}_r^{\diamond}\right)}_{\boldsymbol{B}} \boldsymbol{\Gamma}_{\text{prior}}^{1/2}.$$

The simple structure of $\boldsymbol{B}$ allows us to write its spectral decomposition as

$$\boldsymbol{B} = \sum_{i=1}^{n} \boldsymbol{v}_i \boldsymbol{v}_i{}^{\diamond} - \sum_{i=1}^{r} \frac{\lambda_i}{\lambda_i + 1} \boldsymbol{v}_i \boldsymbol{v}_i{}^{\diamond} = \sum_{i=1}^{r} \frac{1}{\lambda_i + 1} \boldsymbol{v}_i \boldsymbol{v}_i{}^{\diamond} + \sum_{i=r+1}^{n} \boldsymbol{v}_i \boldsymbol{v}_i{}^{\diamond},$$

which, together with the standard definition of the square root of postive self-adjoint operators [5], immediately gives

$$\begin{aligned} \boldsymbol{B}^{1/2} &= \sum_{i=1}^{r} \frac{1}{\sqrt{\lambda_i + 1}} \boldsymbol{v}_i \boldsymbol{v}_i{}^{\diamond} + \sum_{i=r+1}^{n} \boldsymbol{v}_i \boldsymbol{v}_i{}^{\diamond} \\ &= \sum_{i=1}^{r} \left(\frac{1}{\sqrt{\lambda_i + 1}} - 1 \right) \boldsymbol{v}_i \boldsymbol{v}_i{}^{\diamond} + \sum_{i=1}^{n} \boldsymbol{v}_i \boldsymbol{v}_i{}^{\diamond} = \boldsymbol{V}_r \boldsymbol{P}_r \boldsymbol{V}_r^{\diamond} + \boldsymbol{I}, \end{aligned}$$

where $\boldsymbol{P}_r = \operatorname{diag}\left(1/\sqrt{\lambda_1 + 1} - 1, \ldots, 1/\sqrt{\lambda_r + 1} - 1\right) \in \mathbb{R}^{r \times r}$, and $\boldsymbol{B}^{1/2}$ is self-adjoint in $\mathbb{R}^n_{\boldsymbol{M}}$, namely, $\left(\boldsymbol{B}^{1/2}\right)^* = \boldsymbol{B}^{1/2}$. Now, we define

$$\tilde{\boldsymbol{L}} = \boldsymbol{\Gamma}_{\text{prior}}^{1/2} \boldsymbol{B}^{1/2},$$

which, by construction, is the desired matrix owing to the trivial identity

$$\tilde{\boldsymbol{L}} \tilde{\boldsymbol{L}}^* = \boldsymbol{\Gamma}_{\text{prior}}^{1/2} \boldsymbol{B}^{1/2} \left(\boldsymbol{B}^{1/2}\right)^* \left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\right)^* = \boldsymbol{\Gamma}_{\text{prior}}^{1/2} \boldsymbol{B} \boldsymbol{\Gamma}_{\text{prior}}^{1/2} = \boldsymbol{\Gamma}_{\text{post}},$$

where we have used the self-adjointness of $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ and $\boldsymbol{B}^{1/2}$ in $\mathbb{R}^n_{\boldsymbol{M}}$.

Finally, we can rewrite (A.1) in terms of $\boldsymbol{n}$ and $\boldsymbol{L} = \tilde{\boldsymbol{L}} \boldsymbol{M}^{-1/2}$, a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n_{\boldsymbol{M}}$, as

$$\boldsymbol{\nu}^{\text{post}} = \boldsymbol{m}_{\text{MAP}} + \boldsymbol{L}\boldsymbol{n},$$

where $\boldsymbol{L}$ satisfies the desired identity

$$\boldsymbol{L}\boldsymbol{L}^{\diamond} = \tilde{\boldsymbol{L}} \boldsymbol{M}^{-1/2} \left(\boldsymbol{M}^{-1/2}\right)^{\diamond} \tilde{\boldsymbol{L}}^* = \tilde{\boldsymbol{L}} \boldsymbol{M}^{-1/2} \boldsymbol{M}^{-1/2} \boldsymbol{M} \tilde{\boldsymbol{L}}^* = \tilde{\boldsymbol{L}} \tilde{\boldsymbol{L}}^* = \boldsymbol{\Gamma}_{\text{post}}.$$

REFERENCES

[1] P. Abrahamsen, *A Review of Gaussian Random Fields and Correlation Functions*, 2nd ed., Norwegian Computing Center, Oslo, Norway, 1997.

[2] V. Akçelik, J. Bielak, G. Biros, I. Epanomeritakis, A. Fernandez, O. Ghattas, E. J. Kim, J. Lopez, D. R. O'Hallaron, T. Tu, and J. Urbanic, *High resolution forward and inverse earthquake modeling on terascale computers*, in SC03: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, Phoenix, Arizona, ACM/IEEE, 2003.

[3] V. Akçelik, G. Biros, and O. Ghattas, *Parallel multiscale Gauss-Newton-Krylov methods for inverse wave propagation*, in SC02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing, Baltimore, MD, IEEE Computer Society Press, Los Alamitos, CA, 2002, pp. 1–15.

[4] K. Aki and P. G. Richards, *Quantitative Seismology*, 2nd ed., University Science Books, Sausalito, CA, 2002.
[5] T. Arbogast and J. L. Bona, *Methods of Applied Mathematics*, University of Texas at Austin, Austin, TX, 2008.
[6] J. M. Bardsley, *Gaussian Markov random field priors for inverse problems*, Inverse Probl. Imaging, 7 (2013), pp. 397–416.
[7] 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.
[8] 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, Salt Lake City, UT, 2012, pp. 1–11.
[9] T. Bui-Thanh and O. Ghattas, *Analysis of an hp-non-conforming discontinuous Galerkin spectral element method for wave propagation*, SIAM J. Numer. Anal., 50 (2012), pp. 1801–1826.
[10] 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.
[11] 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.
[12] C. Burstedde, O. Ghattas, M. Gurnis, T. Isaac, G. Stadler, T. Warburton, and L. C. Wilcox, *Extreme-scale AMR*, in SC10: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, ACM/IEEE, IEEE Computer Society Washington, DC, 2010, pp. 1–12.
[13] C. Burstedde, L. C. Wilcox, and O. Ghattas, *p4est: Scalable algorithms for parallel adaptive mesh refinement on forests of octrees*, SIAM J. Sci. Comput., 33 (2011), pp. 1103–1133.
[14] J. Chen, M. Anitescu, and Y. Saad, *Computing $f(a)b$ via least squares polynomial approximations*, SIAM J. Sci. Comput., 33 (2011), pp. 195–222.
[15] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering*, 2nd ed., Theory Appl. Math. Sci. 93, Springer-Verlag, Berlin, 1998.
[16] A. M. Dziewonski and D. L. Anderson, *Preliminary reference earth model*, Physics of the Earth and Planetary Interiors, 25 (1981), pp. 297–356.
[17] I. Epanomeritakis, V. Akçelik, O. Ghattas, and J. Bielak, *A Newton-CG method for large-scale three-dimensional elastic full-waveform seismic inversion*, Inverse Problems, 24 (2008), 034015.
[18] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, RI, 1998.
[19] A. Fichtner, *Full Seismic Waveform Modelling and Inversion*, Springer, Heidelberg, 2011.
[20] A. Fichtner, H. Igel, H.-P. Bunge, and B. L. N. Kennett, *Simulation and inversion of seismic wave propagation on continental scales based on a spectral-element method*, JNAIAM J. Numer. Anal. Ind. Appl. Math., 4 (2009), pp. 11–22.
[21] A. Fichtner, B. L. N. Kennett, H. Igel, and H.-P. Bunge, *Full waveform tomography for upper-mantle structure in the Australasian region using adjoint methods*, Geophysical Journal International, 179 (2009), pp. 1703–1725.
[22] 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.
[23] M. W. Gee, C. M. Siefert, J. J. Hu, R. S. Tuminaro, and M. G. Sala, *ML* 5.0 *Smoothed Aggregation User's Guide*, Technical report SAND2006-2649, Sandia National Laboratories, Albuquerque, NM, 2006.
[24] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis,* 2*nd ed.*, Chapman & Hall/CRC Texts in Statistical Science, Chapman and Hall/CRC, Boca Raton, FL, 2004.
[25] M. Girolami and B. Calderhead, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, J. R. Stat. Soc. Ser. B Stat. Methodol., 73 (2011), pp. 123–214.
[26] M. D. Gunzburger, *Perspectives in Flow Control and Optimization*, SIAM, Philadelphia, 2003.
[27] M. Hairer, *An introduction to stochastic PDEs*, preprint, http://arxiv.org/abs/0907.4178, 2009.
[28] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE Constraints*, Springer, New York, 2009.
[29] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Appl. Math.

Sci. 160, Springer-Verlag, New York, 2005.
[30] D. A. Kopriva, *A conservative staggered-grid Chebyshev multidomain method for compressible flows.* II. *A semi-structured method*, J. Comput. Phys., 128 (1996), pp. 475–488.
[31] D. A. Kopriva, S. L. Woodruff, and M. Y. Hussaini, *Computation of electromagnetic scattering with a non-conforming discontinuous spectral element method*, Internat. J. Numer. Methods Eng., 53 (2002), pp. 105–122.
[32] M. Lassas, E. Saksman, and S. Siltanen, *Discretization-invariant Bayesian inversion and Besov space priors*, Inverse Probl. Imaging, 3 (2009), pp. 87–122.
[33] L. LeCam, *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York, 1986.
[34] V. Lekić and B. Romanowicz, *Inferring upper-mantle structure by full waveform tomography with the spectral element method*, Geophysical Journal International, 185 (2011), pp. 799–831.
[35] F. Lindgren, H. Rue, and J. Lindström, *An explicit link between Gaussian fields and Gaussian markov random fields: The stochastic partial differential equation approach*, J. Roy. Stat. Soc. Ser. B Stat. Methodol., 73 (2011), pp. 423–498.
[36] 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.
[37] D. Peter, D. Komatitsch, Y. Luo, R. Martin, N. Le Goff, E. Casarotti, P. Le Loher, F. Magnoni, Q. Liu, C. Blitz, T. Nisson-Meyer, P. Basini, and J. Tromp, *Forward and adjoint simulations of seismic wave propagation on fully unstructured hexahedral meshes*, Geophysical Journal International, 186 (2011), pp. 721–739.
[38] J. Ritsema and J. Van Heijst, *Constraints on the correlation of p-and s-wave velocity heterogeneity in the mantle from p, pp, ppp and pkpab traveltimes*, Geophysical Journal International, 149 (2002), pp. 482–489.
[39] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd ed., Springer Texts Statist., Springer-Verlag, New York, 2004.
[40] H. Rue, *Fast sampling of Gaussian Markov random fields*, J. Roy. Stat. Soc. Ser. B Stat. Methodol., 63 (2001), pp. 325–338.
[41] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, Monogr. Statist. Appl. Probab. 104, Chapman & Hall, Boca Raton, FL, 2005.
[42] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.
[43] C. Tape, Q. Liu, A. Maggi, and J. Tromp, *Adjoint tomography of the southern California crust*, Science, 325 (2009), pp. 988–992.
[44] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, 2005.
[45] C. R. Vogel, *Computational Methods for Inverse Problems*, Frontiers Appl. Math. 23, SIAM, Philadelphia, 2002.
[46] J. Voss, *The effect of finite element discretization on the stationary distribution of SPDEs*, Commun. Math. Sci., 10 (2012), pp. 1143–1159.
[47] L. C. Wilcox, G. Stadler, C. Burstedde, and O. Ghattas, *A high-order discontinuous Galerkin method for wave propagation through coupled elastic-acoustic media*, J. Comput. Phys., 229 (2010), pp. 9373–9396.
[48] H. Zhu, E. Bozdapg, D. Peter, and J. Tromp, *Structure of the European upper mantle revealed by adjoint tomography*, Nature Geoscience, 5 (2012), pp. 493–498.