International Journal for Uncertainty Quantification, 1(1):1–17, 2011

MARGINALIZATION OF UNINTERESTING DISTRIBUTED PARAMETERS IN INVERSE PROBLEMS —APPLICATION TO DIFFUSE OPTICAL TOMOGRAPHY

Ville Kolehmainen,[1] Tanja Tarvainen,[1] Simon R. Arridge,[2] & Jari P. Kaipio[1,3,*]

[1]*Department of Physics and Mathematics, University of Eastern Finland, P.O. Box 1627, 70211 Kuopio, Finland*

[2]*Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK*

[3]*Department of Mathematics, University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand*

Original Manuscript Submitted: 01/10/2010; Final Draft Received: 04/04/2010

With inverse problems there are often several unknown distributed parameters of which only one may be of interest. Since assigning incorrect fixed values to the uninteresting parameters usually leads to a severely erroneous model, one is forced to estimate all distributed parameters simultaneously. This may increase the computational complexity of the problem significantly. In the Bayesian framework, all unknowns are generally treated as random variables and estimated simultaneously and all uncertainties can be modeled systematically. Recently, the approximation error approach has been proposed for handling uncertainty and model-reduction-related errors in the models. In this approach approximate marginalization of these errors is carried out before the estimation of the interesting variables. In this paper we discuss the adaptation of the approximation error approach to the marginalization of uninteresting distributed parameters. As an example, we consider the marginalization of scattering coefficient in diffuse optical tomography.

KEY WORDS: *inverse problems, Bayesian inference, parameter estimation, spatial uncertainty, diffuse optical tomography*

1. INTRODUCTION

There are several inverse problems in which there are many unknown distributed parameters. Often, only one or a subset of the unknowns is of main interest. For example, in hydrogeophysics the unknown distributed parameters may include permittivity, capillarity, and diffusivity [1, 2]. In diffuse optical tomography (DOT), the most important distributed parameters are the scattering and absorption coefficients [3]. Of these, at least in biomedical applications, the absorption coefficient is the one of interest since it is related to the oxygenization level of tissues [4]. In these applications, the scattering coefficient is considered as a nuisance parameter. The scattering coefficient, however, has to be estimated simultaneously due to the so-called crosstalk of the coefficients [3].

The simultaneous estimation of two distributed parameters is naturally a more unstable problem than estimating either of these if the other were known. In addition, with nonlinear problems the convergence of algorithms is a further problem. With applications which are eventually meant to be almost real time ones, such as biomedical optical tomog-

*Correspond to Jari P. Kaipio, E-mail: jari@math.auckland.ac.nz, URL: http://www.math.auckland.ac.nz/ kaipio/

2152 -5080/11/$35.00

raphy, it is of major interest to reduce the computational times as much as possible. Thus, in addition to estimating the interesting parameters only, there is usually pressure to also use otherwise heavily reduced computational models.

The approximation error approach was introduced in [5, 6] originally to handle pure model reduction errors. For example, in electrical impedance (resistance) tomography (EIT, ERT) and deconvolution problems, it was shown that significant model reduction is possible without essentially sacrificing the quality of estimates. With EIT, for example, this means that very low dimensional finite element approximations can be used. Later, the approach was also applied to handle other kinds of approximation and modeling errors as well as other inverse problems. Model reduction, domain truncation, and unknown anisotropy structures in diffuse optical tomography were treated in [7–10]. Missing boundary data in the case of image processing and geophysical ERT/EIT were considered in [11] and [12], respectively. In [13–15] the problem of recovery from simultaneous geometry errors and model reduction was found to be possible.

The approximation error approach was extended to nonstationary inverse problems in [16] in which linear nonstationary (heat transfer) problems were considered, and in [17] and [18] in which nonlinear problems and state space identification problems were considered, respectively. The earliest similar but partial treatment within the framework of nonstationary inverse problems was considered in [19], in which the the boundary data that is related to stochastic convection diffusion models was partially unknown. A modification in which the approximation error statistics can be updated with accumulating information was proposed in [20] and an application to hydrogeophysical monitoring in [21].

From pure model reduction and unknown (nondistributed) parameters or boundary data, a step forward was recently considered in [22] in which the physical forward model itself was replaced with a (computationally) much simpler model. In [22], the radiative transfer model (Boltzmann transfer equation), which is considered to be the most accurate model for light transfer in (turbid) media, was replaced with the diffusion approximation. It was found that also in this kind of case, the statistical structure of the approximation errors enabled the use of a significantly less complex model, again simultaneously with significant model reduction for the diffusion approximation. But also here, both the absorption and scattering coefficients were estimated simultaneously.

The approximation error approach relies on the Bayesian framework of inverse problems, in which all unknowns are explicitly modeled as random variables [5, 23, 24]. The uncertainty in the unknowns is given in the models and measurements is reflected in the posterior (probability) distribution. In the Bayesian framework, all unknowns are subject to inference simultaneously, which often results in excessively heavy computational loads. Generally, Markov chain Monte Carlo algorithms have to be used to obtain a representative set of samples from the posterior distribution. Then, after a set of samples has been computed, marginalization over the uninteresting unknowns is trivial. Only in a few special but important cases, such as the additive error model, some of the uninteresting unknowns can be eliminated before inference. We refer to such elimination as premarginalization.

In the present paper we consider the approximation error approach in the context of approximate premarginalization of uninteresting distributed parameters. Furthermore, we also consider the simultaneous treatment of the errors that are related to model reduction. As a computational example, we consider the approximate premarginalization of the scattering coefficient in diffuse optical tomography. This example shows that at least in this case it is possible to premarginalize over one distributed parameter and successfully estimate another.

The rest of the paper is structured as follows. In Section 2 we give a brief account of the approximation error approach and its formulation for the case of several distributed parameters. In Section 3 we describe the diffuse optical tomography problem. In Section 4, numerical examples of reconstructing the scattering coefficient in optical tomography with different degrees of severity are treated.

2. APPROXIMATION ERROR APPROACH

In the Bayesian framework for inverse problems, all unknowns are treated and modeled as random variables [5, 23, 24]. Once the probabilistic models for the unknowns and the measurement process have been constructed, the *posterior distribution* $\pi(x \mid y)$ is accessed, which reflects the uncertainty of the interesting unknowns x given the measurements y. This distribution can then be explored to answer all questions which can be expressed in terms of probabilities. For general discussion of Bayesian inference (see, for example, [25, 26]).

Bayesian inverse problems are a special class of problems in Bayesian inference. Usually, the dimension of a feasible representation of the unknowns is significantly larger than the number of measurements. Thus, for example, a maximum likelihood estimate is impossible to compute. Even in cases in which the number of unknowns would be significantly smaller than the number of measurements, the structure of the forward problem is such that maximum likelihood estimates would still be unstable. In addition to the instability, the variances of the likelihood model are almost invariably much smaller than the variances of the prior models. The posterior density is often extremely narrow and, in addition, may be a nonlinear manifold.

2.1 Marginalization Over Additive Errors

In the approximation error approach, the modeling and other errors are treated as additive errors. Therefore, we review briefly how the additive errors are formally premarginalized [5]. Let the observation model be

$$y = \bar{A}(x) + e \tag{1}$$

where e are the additive errors and $x \mapsto \bar{A}(x)$ is the deterministic forward model. With deterministic we mean that the model $\bar{A}$ does not contain any uncertainties or other model errors. Let the joint *prior distribution* of the unknowns x and e be $\pi(x, e)$. Using the Bayes' theorem repeatedly, we can decompose the joint distribution of all associated random variables as

$$\begin{aligned} \pi(y, x, e) &= \pi(y \,|\, x, e)\pi(e \,|\, x)\pi(x) && (2)\\ &= \pi(y, e \,|\, x)\pi(x) && (3) \end{aligned}$$

In the case of the additive model (1), the conditional distribution $\pi(y \,|\, x, e)$ is formally given by

$$\pi(y \,|\, x, e) = \delta(y - \bar{A}(x) - e)$$

which yields the *likelihood distribution*

$$\begin{aligned} \pi(y \,|\, x) &= \int \pi(y, e \,|\, x)\,\mathrm{d}e = && (4)\\ &= \int \delta(y - \bar{A}(x) - e)\pi(e \,|\, x)\,\mathrm{d}e && (5)\\ &= \pi_{e \,|\, x}(y - \bar{A}(x) \,|\, x) && (6) \end{aligned}$$

and further,[1] noting that once the measurements have been obtained, $\pi(y) > 0$ is a fixed normalization constant, we have the posterior distribution

$$\begin{aligned} \pi(x \,|\, y) &\propto \pi(y \,|\, x)\pi(x) && (7)\\ &= \pi_{e \,|\, x}(y - \bar{A}(x) \,|\, x)\pi(x) && (8) \end{aligned}$$

In the quite common case of mutually independent x and e, we have $\pi_{e \,|\, x}(e \,|\, x) = \pi_e(e)$. Furthermore, if e and x are normal, we can write $\pi(e) = \mathcal{N}(e_*, \Gamma_e)$ and $\pi(x) = \mathcal{N}(x_*, \Gamma_x)$ and we have the familiar form

$$\pi(x \,|\, y) \propto \exp\left(-\frac{1}{2}\left(\|L_e(y - \bar{A}(x) - e_*)\|^2 + \|L_x(x - x_*)\|^2\right)\right) \tag{9}$$

where $L_e^{\mathrm{T}} L_e = \Gamma_e^{-1}$ and $L_x^{\mathrm{T}} L_x = \Gamma_x^{-1}$, for the posterior distribution. In the above, the unknown (uninteresting) additive error e was *premarginalized*, that is, marginalized before the inference procedure, and is not present in (8) or (9).

[1] The subscripts, such as $e \,|\, x$ in $\pi_{e \,|\, x}$, are used to determine the actual probability density function. If the arguments, however, coincide with the density, we drop the subscripts. For example, we write $\pi_x(x) = \pi(x)$ and $\pi_{e \,|\, x}(e \,|\, x) = \pi(e \,|\, x)$, but retain the subscript in $\pi_{e \,|\, x}(y - \bar{A}(x) \,|\, x)$. Furthermore, we use the terms density and distribution interchangeably.

2.2 Approximate Premarginalization Over Model Reduction Related Errors and Other Uncertainties

The problem that is generally related to uninteresting auxiliary unknowns ξ is that we usually cannot perform premarginalization such as in Eqs. (5) and (6). In most cases we have to estimate both x and ξ, which may be a considerably more demanding undertaking than estimating just x when ξ were known. For example, if a Markov chain Monte Carlo (MCMC) approach were used, the marginalization over ξ can only be done after running the chain for both x and ξ. Once this is carried out, however, the marginalization over ξ is trivial. For MCMC methods in general (see, for example, [27, 28]). For MCMC and inverse problems, see [29–31] for applications to EIT. In this section we discuss the computational procedure in more detail in the case in which there are two distributed parameters of which premarginalization over the other one is to be carried out.

Now let the unknowns be (x, z, ξ, e), where again e represents additive errors and ξ represents auxiliary uncertainties such as unknown boundary data, and (x, z) are two distributed parameters of which x is of interest only. The accurate forward model

$$(x, z, \xi) \mapsto \bar{A}(x, z, \xi) \tag{10}$$

is usually a nonlinear one. The uncertainties ξ can sometimes be modeled to be mutually dependent with (x, z), especially when ξ represents boundary data on the computational domain boundary and (x, z) are modeled as random fields. On the other hand, if ξ represents an unknown boundary shape, ξ might be modeled as mutually independent with (x, z). In the following we consider the case in which the noise e is additive and the unknowns (x, z, ξ) are not necessarily mutually independent.

Let

$$y = \bar{A}(\bar{x}, z, \xi) + e \in \mathbb{R}^m$$

denote an accurate model for the relation between the measurements and the unknowns,[2] and let e be mutually independent with (x, z, ξ).

In the following we approximate the accurate representation of the primary unknown $\bar{x}$ by $x = P\bar{x}$, where P is typically a projection operator. Let $\pi(x, z, \xi, e)$ be a feasible model for the joint distribution of the unknowns. We identify $x = P\bar{x}$ with its coordinates in the associated basis when applicable.

In the approximation error approach, we proceed as follows. Instead of using the accurate forward model $(\bar{x}, z, \xi) \mapsto \bar{A}(\bar{x}, z, \xi)$ with $(\bar{x}, z, \xi)$ as the unknowns, we fix the random variables $(z, \xi) \leftarrow (z_0, \xi_0)$ and use a computationally (possibly drastically reduced) approximative model

$$x \mapsto A(x, z_0, \xi_0)$$

Thus, we write the measurement model in the form

$$\begin{aligned} y &= \bar{A}(\bar{x}, z, \xi) + e && (11)\\ &= A(x, z_0, \xi_0) + \big[\bar{A}(\bar{x}, z, \xi) - A(x, z_0, \xi_0)\big] + e && (12)\\ &= A(x, z_0, \xi_0) + \varepsilon + e && (13) \end{aligned}$$

where we define the *approximation error* $\varepsilon = \varphi(\bar{x}, z, \xi) = \bar{A}(\bar{x}, z, \xi) - A(x, z_0, \xi_0)$. Thus, the approximation error is the discrepancy of predictions of the measurements (given the unknowns) when using the accurate model $\bar{A}(\bar{x}, z, \xi)$ and the approximate model $A(x, z_0, \xi_0)$. Note that (13) is exact.

Formally, after the models $\bar{A}$ and A are fixed, we have $\pi(\varepsilon \,|\, \bar{x}, z, \xi) = \delta[\varepsilon - \varphi(\bar{x}, z, \xi)]$. We will later, however, employ approximative joint distributions and therefore consider $\pi(\varepsilon, \bar{x}, z, \xi)$ without any special structure. As the first approximation, we approximate $\varphi(\bar{x}, z, \xi) \approx \varphi(Px, z, \xi)$ and thus $\pi(\varepsilon \,|\, \bar{x}, z, \xi) \approx \pi(\varepsilon \,|\, Px, z, \xi)$. This means that we assume that the model predictions and thus the approximation error is essentially the same for $\bar{x}$ as $x = P\bar{x}$. This assumption holds for inverse problems in general and for such projections in particular.

[2] If there are no additive errors, we write $e = 0$ and consider the other types of errors to be included in ξ.

Proceeding as in Section 2.1, we use the Bayes' formula repeatedly

$$\begin{aligned}\pi(y,x,z,\xi,e,\varepsilon) &= \pi(y\,|\,x,z,\xi,e,\varepsilon)\pi(x,z,\xi,e,\varepsilon) = \delta[y-A(x,z_0,\xi_0)-e-\varepsilon]\pi(e,\varepsilon\,|\,x,z,\xi)\pi(z,\xi\,|\,x)\pi(x)\\ &= \pi(y,z,\xi,e,\varepsilon\,|\,x)\pi(x)\end{aligned}$$

Hence,

$$\begin{aligned}\pi(y\,|\,x) &= \iiiint \pi(y,z,\xi,e,\varepsilon\,|\,x)de\,d\varepsilon\,dz\,d\xi = \iint \delta[y-A(x,z_0,\xi_0)-e-\varepsilon]\\ &\cdot \left[\iint \pi(e,\varepsilon\,|\,x,z,\xi)\pi(z,\xi\,|\,x)dz\,d\xi\right] de\,d\varepsilon = \iint \delta[y-A(x,z_0,\xi_0)-e-\varepsilon]\pi(e,\varepsilon|x)de\,d\varepsilon\\ &= \int \pi_e[y-A(x,z_0,\xi_0)-\varepsilon]\pi_{\varepsilon|x}(\varepsilon\,|\,x)\,d\varepsilon \qquad (14)\end{aligned}$$

since e and x are mutually independent, and (14) is a convolution integral with respect to ε. In particular, since e is mutually independent with (x,z,ξ), e and ε are also mutually independent.

At this stage, in the approximation error approach, both π_e and $\pi_{\varepsilon|x}$ are approximated with normal distributions. Let the normal approximation for the joint density $\pi(\varepsilon,x)$ be

$$\pi(\varepsilon,x) \propto \exp\left\{-\frac{1}{2}\begin{pmatrix}\varepsilon-\varepsilon_*\\ x-x_*\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}\Gamma_{\varepsilon\varepsilon} & \Gamma_{\varepsilon x}\\ \Gamma_{x\varepsilon} & \Gamma_{xx}\end{pmatrix}^{-1}\begin{pmatrix}\varepsilon-\varepsilon_*\\ x-x_*\end{pmatrix}\right\} \qquad (15)$$

Thus we write

$$e \sim \mathcal{N}(e_*,\Gamma_e), \qquad \varepsilon\,|\,x \sim \mathcal{N}(\varepsilon_{*,x},\Gamma_{\varepsilon|x})$$

where

$$\varepsilon_{*,x} = \varepsilon_* + \Gamma_{\varepsilon x}\Gamma_{xx}^{-1}(x-x_*) \qquad (16)$$

$$\Gamma_{\varepsilon|x} = \Gamma_{\varepsilon\varepsilon} - \Gamma_{\varepsilon x}\Gamma_{xx}^{-1}\Gamma_{x\varepsilon} \qquad (17)$$

Define the normal random variable ν so that[3] $\nu\,|\,x = e+\varepsilon\,|\,x$

$$\nu\,|\,x \sim \mathcal{N}(\nu_{*|x},\Gamma_{\nu|x})$$

where

$$\nu_{*|x} = e_* + \varepsilon_* + \Gamma_{\varepsilon x}\Gamma_x^{-1}(x-x_*) \qquad (18)$$

$$\Gamma_{\nu|x} = \Gamma_e + \Gamma_\varepsilon - \Gamma_{\varepsilon x}\Gamma_x^{-1}\Gamma_{x\varepsilon} \qquad (19)$$

Thus, we obtain for the approximate likelihood distribution

$$y\,|\,x \sim \mathcal{N}[y-A(x,z_0,\xi_0)-\nu_{*|x},\Gamma_{\nu|x}]$$

Since we are after computational efficiency, a normal approximation for the prior model is also conventionally used:

$$x \sim \mathcal{N}(x_*,\Gamma_x)$$

Thus, we obtain the approximation for the posterior distribution

$$\pi(x\,|\,y) \propto \pi(y\,|\,x)\pi(x) \propto \exp\left[-\frac{1}{2}V(x)\right]$$

[3]With autocovariances, we may notate $\Gamma_{xx}=\Gamma_x$ below.

where $V(x)$ is the posterior potential

$$V(x) = [y - A(x, z_0, \xi_0) - \nu_{*|x}]^{\mathrm{T}} \Gamma_{\nu|x}^{-1} [y - A(x, z_0, \xi_0) - \nu_{*|x}] + (x - x_*)^{\mathrm{T}} \Gamma_x^{-1} (x - x_*) \quad (20)$$

$$= \|L_{\nu|x}[y - A(x, z_0, \xi_0) - \nu_{*|x}]\|^2 + \|L_x(x - x_*)\|^2 \quad (21)$$

where $\Gamma_{\nu\,|\,x}^{-1} = L_{\nu|x}^{\mathrm{T}} L_{\nu|x}$ and $\Gamma_x^{-1} = L_x^{\mathrm{T}} L_x$.

2.3 Computational Considerations

In Section 2.2, we wrote the normal approximation (15) for the joint distribution of (x, ε). Generally, this approximation is done to make an efficient computation of the maximum a posteriori (MAP) estimate feasible. If the actual prior model is normal, the marginal distribution of x induced by (15) coincides with the actual prior model. The prior model $\pi(\bar{x}, z, \xi)$ does not, however, have to be jointly normal and neither, in particular, does the marginal prior model $\pi(\bar{x})$. In practice, whatever the prior model $\pi(\bar{x}, z, \xi)$ is, a set of samples $(\bar{x}^{(\ell)}, z^{(\ell)}, \xi^{(\ell)})$ is usually to be drawn and the approximation errors

$$\varepsilon^{(\ell)} = \varphi(\bar{x}^{(\ell)}, z^{(\ell)}, \xi^{(\ell)}) = \bar{A}(x^{(\ell)}, z^{(\ell)}, \xi^{(\ell)}) - \bar{A}(x^{(\ell)}, z_0, \xi_0)\,, \quad \ell = 1 \ldots n_{\mathrm{samp}}$$

are then to be computed, where n_{samp} is the number of draws. The normal approximation for $\pi(\varepsilon, x)$ is then formed by setting $x^{(\ell)} = P\bar{x}^{(\ell)}$ and computing the mean and joint covariance as sample averages over the ensemble.

In the *enhanced error model*, one neglects the cross covariance and sets $\Gamma_{\varepsilon x} = 0$ (see, for example, [5]). With the enhanced error model and nonlinear forward problems, we need to estimate the covariance Γ_ε practically always by simulations and sample averages. If the prior model $\pi(\bar{x})$ is Gaussian, however, the covariance $\Gamma_{\bar{x}}$ is available in the first place and in principle, would, not have be computed as a sample average.

Irrespective of what the (original) prior model for the primary unknown is, we note the following: When the cross covariances $\Gamma_{\varepsilon x}$ are employed, the sample average has to be used in practice also for Γ_x. Although the prior model covariance $\Gamma_{\bar{x}\bar{x}}$ would yield a "better estimate" for the covariance of the reduced order covariance Γ_x than a sample covariance, we might be forced to use the latter because when the sample covariance is computed, the term $\Gamma_{\varepsilon\varepsilon} - \Gamma_{\varepsilon x}\Gamma_x^{-1}\Gamma_{x\varepsilon}$ is guaranteed to be non-negative definite. If, in addition, Γ_e has full rank, $\Gamma_{\nu|x}$ is guaranteed to be positive definite and the Cholesky factor $L_{\nu|x}$ exists. But if the prior model covariance $\Gamma_x = P\Gamma_{\bar{x}\bar{x}}P^{\mathrm{T}}$ is used, this condition is not generally met. By the law of large numbers, the condition is met asymptotically but it is impossible to specify a safe sample size. From the point of view of numerical stability, this is a problem especially when the approximation errors clearly dominate the additive errors, that is, the case for which the approximation error approach is targeted in the first place. It is thus advised to use the sample covariance estimate also for Γ_x.

It is difficult to predict how many draws are needed to compute the mean and joint covariance for (x, ε). Loosely speaking, this depends generally on (the covariances of) the model $\pi(x, z, \xi)$ and the degree of nonlinearity of $\bar{A}$. With relatively small covariances, few draws seem to be enough (see, for example, [12]). In the approximation error approach, the bottleneck is the computation of the solutions of the full accurate forward model $\bar{A}(\bar{x}, z, \xi)$. As for using the full accurate forward model in the inversion with nonlinear problems, we, of course, have to compute this model along the iteration, but also typically compute the related Jacobian mapping. Thus, the overhead that is related to the computation of the approximation error statistics often corresponds to the computation of a few MAP estimates with the full accurate model.

3. DIFFUSE OPTICAL TOMOGRAPHY

Diffuse optical tomography (DOT) is a noninvasive imaging modality in which images of the optical absorption and scattering within turbid media are derived based on measurements of near-infrared light on the surface of the body (for reviews see [3, 32]). The DOT problem is an exceptionally challenging inverse problem due to the diffuse nature of the forward model, and also since the measurements can span 10 orders of magnitude. Furthermore, there are several unknown distributed parameters involved, of which the *absorption coefficient* and the *(reduced) scattering coefficient* are usually reconstructed. Both coefficients are usually measured in mm^{-1}.

The applications of DOT include the detection and classification of tumors from breast tissue, monitoring of infant brain tissue oxygenation level, and functional brain activation studies. For reviews on clinical applications (see [4, 33]). The absorption coefficient, which is related to the oxygenation level of blood, is usually the interesting parameter. In most applications, the scattering coefficient is considered a nuisance parameter. In this paper, our task is thus to employ the approximation error approach for approximate premarginalization over the inhomogeneous scattering coefficient $z(\vec{r})$ and reconstruct only the inhomogeneous absorption coefficient $x(\vec{r})$.

In the numerical examples below, we consider cases in which there are no auxiliary unknowns ξ of any type. Thus, we only have the uninteresting distributed parameter z and additive measurement noise to deal with.

3.1 Measurements in Optical Tomography

In the experimental setup of DOT, m_s optical fibers are placed on the source positions (surface patches) $\partial\Omega_{\mathrm{s},k} \subset \partial\Omega$ on the boundary of the body Ω. The measurements are obtained through m_d optical fibers that are placed in the detector positions (surface patches) $\partial\Omega_{\mathrm{m},i} \subset \partial\Omega$. A collection of measurements is formed by turning on the sources one at a time and measuring the light intensity at all measurement locations (for each source). The measurements in DOT may consist of direct intensity measurements, frequency-modulated amplitude and phase shift measurements, or time-resolved impulse response measurements.

In this paper, we consider the frequency domain measurement system. In the frequency domain measurements light from a sinusoidally modulated laser source is guided via the optic fibers to one of the source locations $\partial\Omega_{\mathrm{s},k}$ at a time, and the amplitudes and phase shifts of the transmitted light are measured on all the detector locations $\partial\Omega_{\mathrm{m},i}, i = 1, \ldots, m_d$. The measurement vector is thus $y \in \mathbb{R}^m$ with $m = m_s m_d$. The inverse problem in DOT is to estimate a pair of functions (x, z), representing the (spatially inhomogeneous) optical absorption coefficient $x(\vec{r})$ and the scattering coefficients $z(\vec{r})$ of the tissues in Ω, given the measurements y and the forward model for the measurement process and noise.

3.2 Forward Model

In the context of inverse problems, the model $x \mapsto A(x)$ is referred to as the *forward model*. We consider DOT simulations in a diffusive regime where the body Ω consists of turbid, scattering dominated media. In such cases the light transport is commonly modeled with the diffusion approximation (DA) of the radiative transfer equation (RTE) [3]. The diffusion approximation is also used as transport model in this paper. For further details on the derivation and properties of the transport models and boundary conditions (see [3, 34, 35]).

We consider the frequency domain system below. Let the light source $\partial\Omega_{\mathrm{s},k}$ be on and $\Phi_k(\vec{r}, \omega)$ be the induced photon density at $\vec{r}$, where ω is the modulation frequency of the light source. The frequency domain version of the diffusion approximation with the Robin boundary condition is of the form [3, 35]

$$-\nabla \cdot D(\vec{r})\nabla\Phi_k(\vec{r}, \omega) + x(\vec{r})\Phi_k(\vec{r}, \omega) + \frac{\mathrm{i}\omega}{c}\Phi_k(\vec{r}, \omega) = 0, \ \vec{r} \in \Omega \tag{22}$$

$$\Phi_k(\vec{r}, \omega) + 2\zeta D(\vec{r})\frac{\partial\Phi_k(\vec{r}, \omega)}{\partial\nu} = g_k(\vec{r}, \omega) \quad \vec{r} \in \partial\Omega \tag{23}$$

where

$$D(\vec{r}) = \left\{3[x(\vec{r}) + z(\vec{r})]\right\}^{-1}$$

where D (units $[\mathrm{mm}]$) is the diffusion coefficient, c is the speed of light in the medium, ζ is a parameter that describes reflection on the boundary, ν is the outward normal vector at $\partial\Omega$, and $g_k(\vec{r}, \omega)$ is the boundary source term for source at $\partial\Omega_{\mathrm{s},k}$,

$$g_k(\vec{r}, \omega) = \begin{cases} I & \text{on} \quad \partial\Omega_{\mathrm{s},k} \\ 0 & \text{on} \quad \partial\Omega \setminus \partial\Omega_{\mathrm{s},k} \end{cases} \tag{24}$$

where I is the intensity of the source. The complex-valued flux $\rho_{i,k}(\omega)$ at the measurement site $\partial\Omega_{\mathrm{m},i}$ can be written as the surface integral

$$\rho_{i,k}(\omega) = \int_{\partial\Omega_{\mathrm{m},i}} \frac{1}{2\zeta}\Phi_k(\vec{r},\omega)\, dS \tag{25}$$

The collection $\{\rho_{i,k}\},\ k = 1,\ldots,m_s,\ i = 1,\ldots,m_d$ is the *raw* data for the experiment. For numerical reasons, primarily for the range of measurements, transformed data and the correspondingly transformed forward model are typically used. Furthermore, the measurement systems are constructed according to these transformations. The experimental systems for frequency domain optical tomography export the log-amplitude and phase shift of the complex valued (demodulated) raw data. Thus, the measurements can be written

$$y = \begin{pmatrix} \mathrm{Re}\,\log(\rho) \\ \mathrm{Im}\,\log(\rho) \end{pmatrix} \in \mathbb{R}^{2m} \tag{26}$$

and the forward model is transformed accordingly. For different end uses, data with different frequencies ω may be acquired.

For the numerical realization of the diffusion approximation model (22)–(24), the finite element method (FEM) is typically used (see, for example, [3, 7]). In the FEM approximation, photon density is approximated in a finite dimensional basis as

$$\Phi^h(\vec{r}) = \sum_{i=1}^{N_n} \alpha_i \varphi_i(\vec{r}) \tag{27}$$

where $\varphi_i(\vec{r})$ are the nodal basis functions of the finite element mesh and N_n is the number of nodes in the FEM mesh, and h is mesh element size parameter.

The absorption and scattering coefficients are written as finite dimensional approximations

$$x(\vec{r}) = \sum_{j=1}^{n_p} x_j \chi_j(\vec{r})\ , \quad z(\vec{r}) = \sum_{j=1}^{n_p} z_j \chi_j(\vec{r}) \tag{28}$$

where χ_j denotes characteristic functions of disjoint elements in the reconstruction mesh. In the following, we identify the absorption and scattering coefficients and their representations as the coordinates

$$x = (x_1,\ldots,x_{n_p})^{\mathrm{T}} \in \mathbb{R}^{n_p}\ , \quad z = (z_1,\ldots,z_{n_p})^{\mathrm{T}} \in \mathbb{R}^{n_p}$$

Thus, the solution of the forward problem amounts to the solution of m_s complex valued $N_n \times N_n$ systems of equations for one DOT experiment. The FEM-based forward model is thus of the form

$$y = A_h(x,z) \tag{29}$$

where h refers to the discretization level parameter in (27).

3.3 Gaussian MRF Prior Model for Scattering and Absorption Coefficients

In the following, we model (x,z) as mutually independent. As the prior model for both $\pi(x)$ and $\pi(z)$, we used a proper Gaussian smoothness prior, constructed similarly as in [5, 7, 8]. In this construction, the distributed parameter, say x, is considered in the form

$$x(\vec{r}) = x_{\mathrm{in}}(\vec{r}) + x_{\mathrm{bg}}(\vec{r})$$

where $x_{\mathrm{in}}(\vec{r})$ is a spatially inhomogeneous (absorption) coefficient[4] with zero mean, and $x_{\mathrm{bg}}(\vec{r})$ is a spatially homogeneous (background) absorption coefficient with nonzero mean. For the latter, we can write $x_{\mathrm{bg}}(\vec{r}) = q\mathbb{I}$, where $\mathbb{I}$ is a vector of ones and q is a scalar random variable with distribution $q \sim \mathcal{N}(x_*, \sigma^2_{\mathrm{bg},x})$. With respect to the basis for x, we have the coordinates $x_{\mathrm{in}} \in \mathbb{R}^{n_p}$, $\mathbb{I} \in \mathbb{R}^{n_p}$, and set $x_{\mathrm{in}} \sim \mathcal{N}(0,\Gamma_{\mathrm{in}})$. We model the spatial distributions x_{in} and

[4] In the sequel, "in" refers to inhomogeneous, "bg" to background.

$q\mathbb{I}$ as mutually independent, that is, the background is mutually independent with the inhomogeneities. An equivalent construction for z was considered.

Thus, we have the $\Gamma_x = \Gamma_{\mathrm{in},x} + \sigma^2_{\mathrm{bg},x}\mathbb{I}\mathbb{I}^{\mathrm{T}}$, $\Gamma_z = \Gamma_{\mathrm{in},z} + \sigma^2_{\mathrm{bg},z}\mathbb{I}\mathbb{I}^{\mathrm{T}}$ and

$$\pi(x) = \mathcal{N}(x_*\mathbb{I}, \Gamma_x)\,, \quad \pi(z) = \mathcal{N}(z_*\mathbb{I}, \Gamma_z)$$

In the construction of $\Gamma_{\mathrm{in},x}$ and $\Gamma_{\mathrm{in},z}$, the approximate correlation lengths can be adjusted to match the size of the expected inhomogeneities. (See [5, 7, 8] for details.)

This prior model is a proper distribution, that is, the covariance exists. Traditional smoothness prior models are improper and samples cannot be drawn from such distributions. The approximation error approach, on the other hand, is based on computing the statistics of ε over the prior distribution. This is not possible with a prior of unbounded variances.

In [7] it was found that such a specific construction for the prior model for scattering and absorption coefficients was exceptionally suitable. This was the case even without the variable background. In [8], the prior described above with the variable background was shown to be feasible also for real data.

4. NUMERICAL STUDIES

We evaluate the approximate premarginalization by the approximation error approach with three two-dimensional numerical examples. In the first one, we study only the errors that are related to marginalization over the scattering coefficient using an otherwise accurate forward model $\bar{A}(x, z_0)$. Thus, numerical model reduction errors are not present. The second example is similar, but the prior model for the scattering coefficient is somewhat off in the sense that the actual background of scattering differs from the modeled one in $\bar{A}(x, z_0)$. In the third example, the numerical model reduction is included, that is, we use the model $A(x, z_0)$.

In the following we explain the common details for the numerical examples.

4.1 Computational Forward Models and Prior Model

In the numerical studies, the domain $\Omega \subset \mathbb{R}^2$ is a circle with radius 25 mm. The measurement setup consists of $m_s = 32$ sources and $m_d = 32$ detectors, located at equispaced intervals on the boundary $\partial\Omega$. With this setup, the number of log-amplitude and phase measurements is $2m = 2048$.

The simulated measurement data is computed with the FEM approximation of the diffusion approximation model in a mesh $\mathcal{M}_1$ which is dense enough to consider the solution as the solution of the problem (22)–(24). Two other finite element meshes and models are constructed, the (relatively) accurate $[\mathcal{M}_2, \bar{A}(x, z)]$ and the radically reduced one $[\mathcal{M}_3, A(x, z)]$ (see Table 1). The actual simulated measurement data was obtained by adding mutually independent (non-identically distributed) noise $\sigma_{e,j} = \delta|y_{*,j}|/100$ with $\delta = 0.5$, that is, the error level was 0.5% of the respective noiseless measurement. Thus, the (instrument) noise is additive and independent, identically distributed, but the individual variances are not equal, that is, the covariance Γ_e is diagonal but the diagonal entries are not equal.

For the reconstruction basis in (28) for the coefficients (x, z) in the inverse problem, we divide the domain Ω into $n_p = 1904$ square pixels for both the accurate and approximate models. Thus we have $P = I$, $\bar{x} = Px = x$, and the parameter vectors $\bar{x}, z \in \mathbb{R}^{1904}$. Having $P : \mathbb{R}^n \mapsto \mathbb{R}^m$ where $m \ll n$ has an impact mainly when m is very small,

TABLE 1: The FEM meshes used in the simulations: N_n is the number of nodes and N_e is the number of triangular elements in the mesh.

Mesh	Use	N_n	N_e
$\mathcal{M}_1$	Simulation of measurements	12,853	25,326
$\mathcal{M}_2$	Accurate forward model	11,329	22,302
$\mathcal{M}_3$	Reduced forward model	703	1326

and less with having $m \ll n$. An exception is poor (initial) modeling, that is, if the accurate representation $\bar{x}$ is a poor approximation for the reality.

The prior model was constructed as in Section 3.3. The parameters for the prior distribution are given in Table 2. Elementwise, this means that the values of absorption and scatter coefficients are expected to lie within the two standard deviation intervals $x_* \pm 2\sigma_x$ and $z_* \pm 2\sigma_z$ with probability of 95%, with respect to the marginal prior distributions.

Two draws from the prior models $\pi(x)$ and $\pi(z)$ are shown in Fig. 1. In the draws, the variation of the background that was part of the construction of the prior model is clearly visible.

4.2 Estimates and Approximation Error Statistics

We compute MAP estimates only by minimizing the respective posterior potentials. The following particular estimates are computed in the three test cases:

TABLE 2: The parameters for the prior model: means and the standard deviations σ for the homogeneous background and inhomogeneities. The correlation length for both coefficients was set as 8 mm.

x_*	$10.0 \cdot 10^{-3}$ mm^{-1}
$\sigma_{\mathrm{bg},x}$	$1.2 \cdot 10^{-3}$ mm^{-1}
$\sigma_{\mathrm{in},x}$	$3.0 \cdot 10^{-3}$ mm^{-1}
z_*	$1000 \cdot 10^{-3}$ mm^{-1}
$\sigma_{\mathrm{bg},z}$	$120 \cdot 10^{-3}$ mm^{-1}
$\sigma_{\mathrm{in},z}$	$300 \cdot 10^{-3}$ mm^{-1}

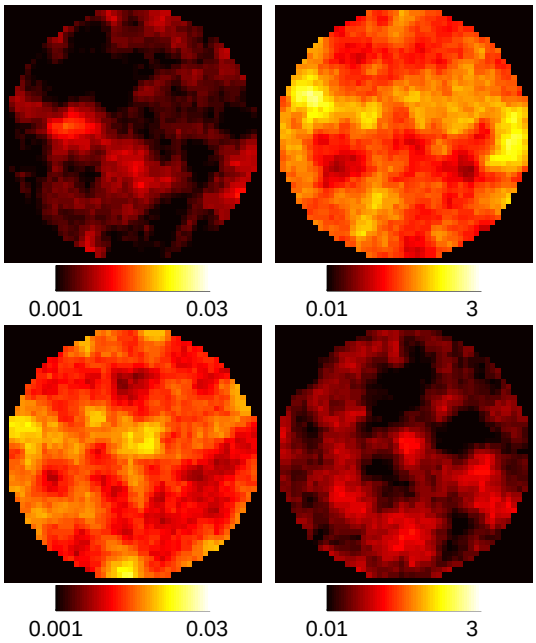

FIG. 1: Two draws from the prior distribution: (left) absorption coefficients $x(\vec{r})$ and (right) scattering coefficient $z(\vec{r})$.

1. MAP-REF—Maximum a posteriori estimate for both parameters (x, z) with the conventional error model $y = \bar{A}(x, z) + e$. This estimate is obtained by computing

$$\min_{x,z} \left\{ \|L_e[y - \bar{A}(x, z) - e_*]\|^2 + \|L_x(x - x_*)\|^2 + \|L_z(z - z_*)\|^2 \right\} \tag{30}$$

   and can be considered as a reference estimate in which both distributed parameters are estimated simultaneously.

2. MAP-CEM—Maximum a posteriori estimate for the primary unknown x using fixed $z = z_*$ and conventional error model $y = \bar{A}(x, z_*) + e$, corresponding to

$$\min_{x} \left\{ \|L_e[y - \bar{A}(x, z_*) - e_*]\|^2 + \|L_x(x - x_*)\|^2 \right\} \tag{31}$$

3. MAP-AEM—Maximum a posteriori estimate for x using fixed $z = z_*$ and the approximation error model, corresponding to

$$\min_{x} \left\{ \|L_{\nu|x}[y - \bar{A}(x, z_*) - \nu_{*|x}]\|^2 + \|L_x(x - x_*)\|^2 \right\} \tag{32}$$

In the above functionals, when the reduced-order model is used, the model $\bar{A}(x, \cdot)$ is to be substituted by $A(x, \cdot)$ [see the posterior potential $V(x)$ in (21)].

In the following numerical examples, the realization z_* is the mean of the prior model $\pi(z)$. The estimates (30)–(32) are computed with the Gauss–Newton optimization method with an explicit line search [36].

For the construction of the approximation error statistics, we proceed as follows. The statistics were used only with MAP with the approximation error model (MAP–AEM) and were computed to correspond to the employed forward model. The means and covariances for (x, ε) in the approximation error model (15) were estimated by Monte Carlo simulation. For this, we draw the sets of samples $\{x^{(\ell)}, \ell = 1, \ldots, n_{\text{samp}}\}$ and $\{z^{(\ell)}, \ell = 1, \ldots, n_{\text{samp}}\}$ from the prior models $\pi(x)$ and $\pi(z)$, respectively. Using the sets of samples, the realizations of the approximation error are computed as

$$\varepsilon^{(\ell)} = \bar{A}(x^{(\ell)}, z^{(\ell)}) - \bar{A}(x^{(\ell)}, z_0) \,, \quad \ell = 1 \ldots n_{\text{samp}}$$

for the case where ε is due to using fixed $z = z_0$ only, that is, no model reduction errors are present, and

$$\varepsilon^{(\ell)} = \bar{A}(x^{(\ell)}, z^{(\ell)}) - A(x^{(\ell)}, z_0) \,, \quad \ell = 1 \ldots n_{\text{samp}}$$

for the cases in which both errors from model reduction and using fixed $z = z_0$ are present. The means x_* and $\nu_{*|x}$ and the covariances Γ_x and $\Gamma_{\nu|x}$ are then estimated as sample averages using the samples $\{x^{(\ell)}, \varepsilon^{(\ell)}, \ell = 1 \ldots n_{\text{samp}}\}$. In the following examples, we use sample size $n_{\text{samp}} = 20,000$.

4.3 Reconstructions

We have three different parameters z: $z_{\text{bg,actual}}$ is the actual (unknown distributed) parameter, z_* is the mean of the modeled prior distribution, and z_0 is the fixed value of z used in the posterior model. Both z_0 and z_* can be chosen separately and we do not need to have $z_0 \neq z_*$. Technically, it is possible to optimize z_0, but this is not considered here.

Three reconstruction cases are considered: case 1—marginalization over z only with $z_0 = z_*$, and furthermore, $z_0 = z_{\text{bg,actual}}$; case 2—as in Case 1 but with $z_0 \neq z_{\text{bg,actual}}$ to assess the robustness toward the poor choice of z_0; and case 3—both marginalization over z and numerical model reduction errors are present. The actual spatial distributions for x and z are blocky targets in a homogeneous background. The probability of the actual x and z with respect to the prior models is relatively low since they are discontinuous (see top row of any of Figs. 2–4). This is one of the ways to check the robustness of the approximation error approach against prior model design.

Case 1: Modeling errors caused by using a fixed value for z.—The results for case 1 are shown in Fig. 2. The actual coefficients (x, z) are shown in the top row. The background values of the target distributions coincide with the prior means. In particular, $z_{\text{bg,actual}} = z_0$.

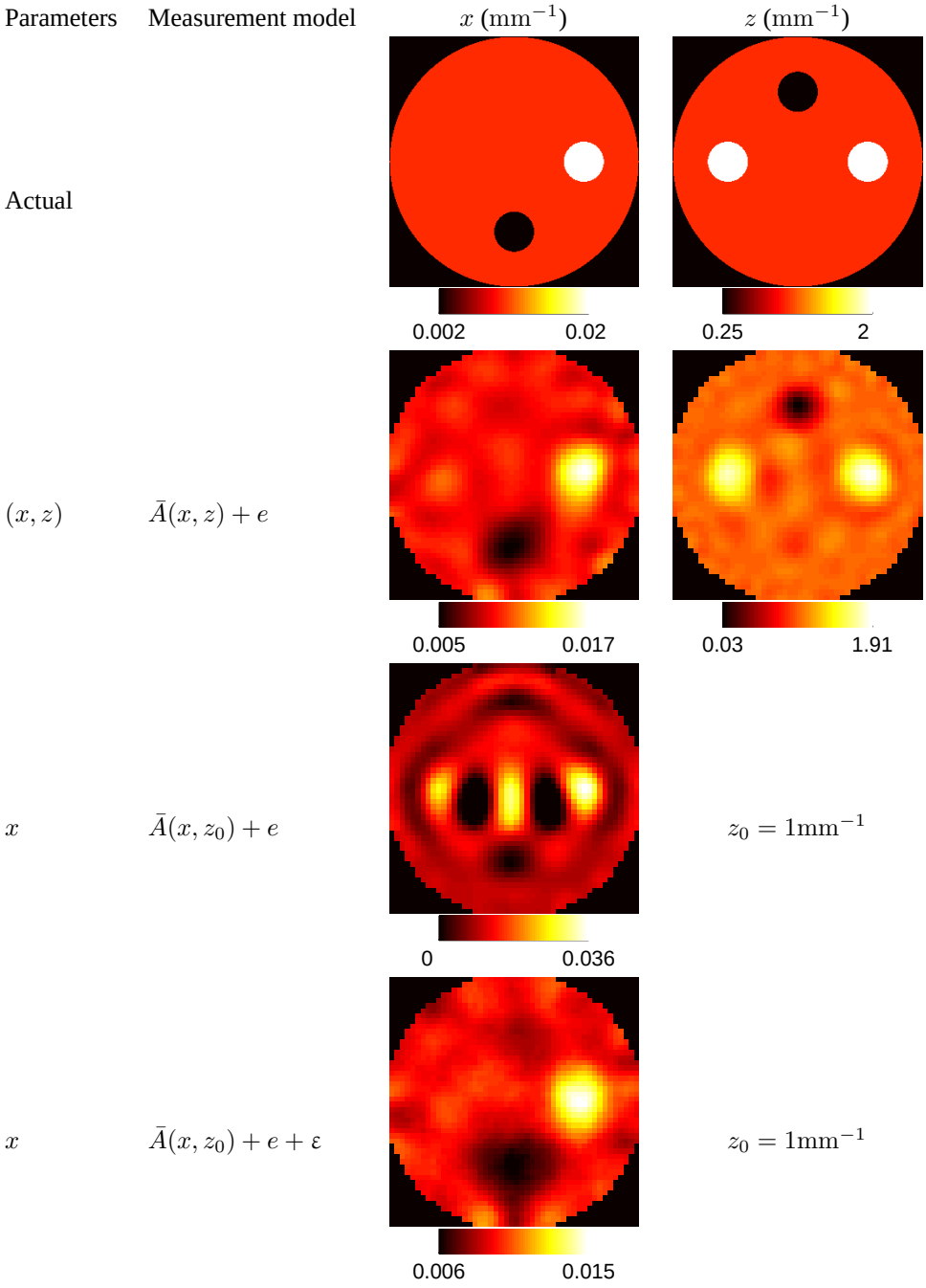

FIG. 2: Case 1. Rows from top to bottom: actual distributions, MAP-REF estimate (30) for (x, z), MAP-CEM estimate (31) for x using fixed $z_0 = z_*$ and conventional noise model, and MAP-AEM estimate (32) using fixed $z_0 = z_*$ and the approximation error model. The modeling error in MAP-CEM and MAP-AEM is caused solely by using the (incorrect) fixed value $z_0 = z_*$. Here $z_{\mathrm{bg,actual}} = z_* = z_0$.

The reference estimates MAP-REF for (x, z), computed by minimization of (30), are shown in the second row. The estimates are computed using the accurate forward model $\bar{A}(x, z)$, meaning that there are no discretization-related errors present. The MAP estimate with the conventional error model (MAP-CEM) estimate (31) for x, using the fixed

value $z = z_*$ for the scattering and the conventional noise model $y = \bar{A}(x, z_*) + e$, is shown in the third row. As can be seen, the use of incorrect value z_* (which is here equivalent to the background scatter) has effectively destroyed the reconstruction of the absorption coefficient x, although the actual z differs from the modeled z_0 only in three small subdomains.

The fourth row shows the MAP-AEM estimate (32) for x using the same fixed $z = z_*$ and the approximation error model $y = \bar{A}(x, z_*) + \varepsilon + e$. The estimate for the absorption is similar to the MAP-REF estimate, but the circular targets are slightly more blurred and the values of the inhomogeneous targets are slightly more off the actual values than with the MAP-REF. This is unavoidable in most cases, since the inclusion of the statistics of the approximation errors increases the variances of their likelihood, which in turn drives the estimates toward the mean of the prior model.

Case 2: Tolerance against poor choice of z_0.— The results for case 2 are shown in Fig. 3. The estimates are computed and arranged as in Fig. 2, the only difference being that in Fig. 3 the actual background value of the scattering target in $z_{\mathrm{bg,actual}} = 1.5z_0$, the discrepancy of which corresponds to $1.5\sigma_z$ with respect to the prior model.

The decrease of the quality of the MAP-REF is due to using a poor model for the prior (background) mean only. This is still a reasonable estimate showing all five objects of the actual spatial distributions. The MAP-CEM estimate is completely useless.

The reconstruction quality in the MAP-AEM estimate is similar to that in Fig. 2. This means that a $1.5\sigma_z$ underestimation of the scattering coefficient in the model $A(x, z_0)$ is tolerated well. We also tested a case where the background of the scattering target was $-1.5\sigma_z$ and $\pm 2\sigma_z$ away from the mean z_0. The MAP-AEM estimate of x remained similar to the estimate shown in bottom row in Fig. 3, except in the case $z_{\mathrm{bg,actual}} = z_0 - 2\sigma_z$. In this case, the quality of the MAP-AEM estimate decreased significantly, producing a nearly useless estimate. Such misspecification of the background is not expected in practice, since the estimation of the best spatially homogeneous estimates for x and z can be done readily if the approximate background values are not known (see [8]). Based on these numerical tests, the approximate premarginalization with respect to the unknown scattering parameter z is reasonably tolerant against a discrepancy between the prior model and the fixed $z = z_0$.

Case 3: Combined approximation error caused by fixed $z = z_*$ and model reduction.— The results for case 3 are shown in Fig. 4. The estimates are computed from the same data that was used in Fig. 2. The only difference between case 1 and case 3 is that in case 3 the estimates are computed using the reduced forward model $A(x, z_0)$, with the number of nodes in the FEM approximation being $N_n = 1326$, instead of $N_n = 22,302$ in the accurate model $\bar{A}(x, z_0)$.

The reference estimate MAP-REF for both unknowns (x, z) was destroyed by the unaccounted for approximation error caused by use of the reduced FEM model. As with case 2, the MAP-CEM estimate is again useless. On the other hand, the MAP-AEM estimate with the proposed approach remained similar to case 1, showing that simultaneous recovery from the use of an incorrect fixed value of z and model-reduction-related errors is feasible with the proposed approach.

We did not consider any auxiliary uncertainties ξ in the above example. In DOT, the principal candidate for ξ is poorly known exterior geometry, but there are also other topics. In clinical applications of optical tomography in particular, the actual optode locations might be slightly off the modeled locations. Furthermore, the channel amplifications can only be calibrated up to a constant with, which constant is difficult to estimate.

5. CONCLUSIONS

In this paper we applied the recently proposed approximation error approach for approximate premarginalization of uninteresting distributed parameters. The approximation error approach is based on the Bayesian framework for inverse problems. The approximation error approach has earlier been shown to be able to deal with diverse types of approximation and modeling errors and uncertainties, such as those related to pure model reduction, unknown boundary data, mismodeled geometry, and use of approximative physical models.

We considered the special case of approximate premarginalization over the inhomogeneous scattering coefficient in diffuse optical tomography. This is an example of a problem in which there are two or more unknown distributed parameters, of which only one is of interest. The results for this example problem suggest that the approximation error

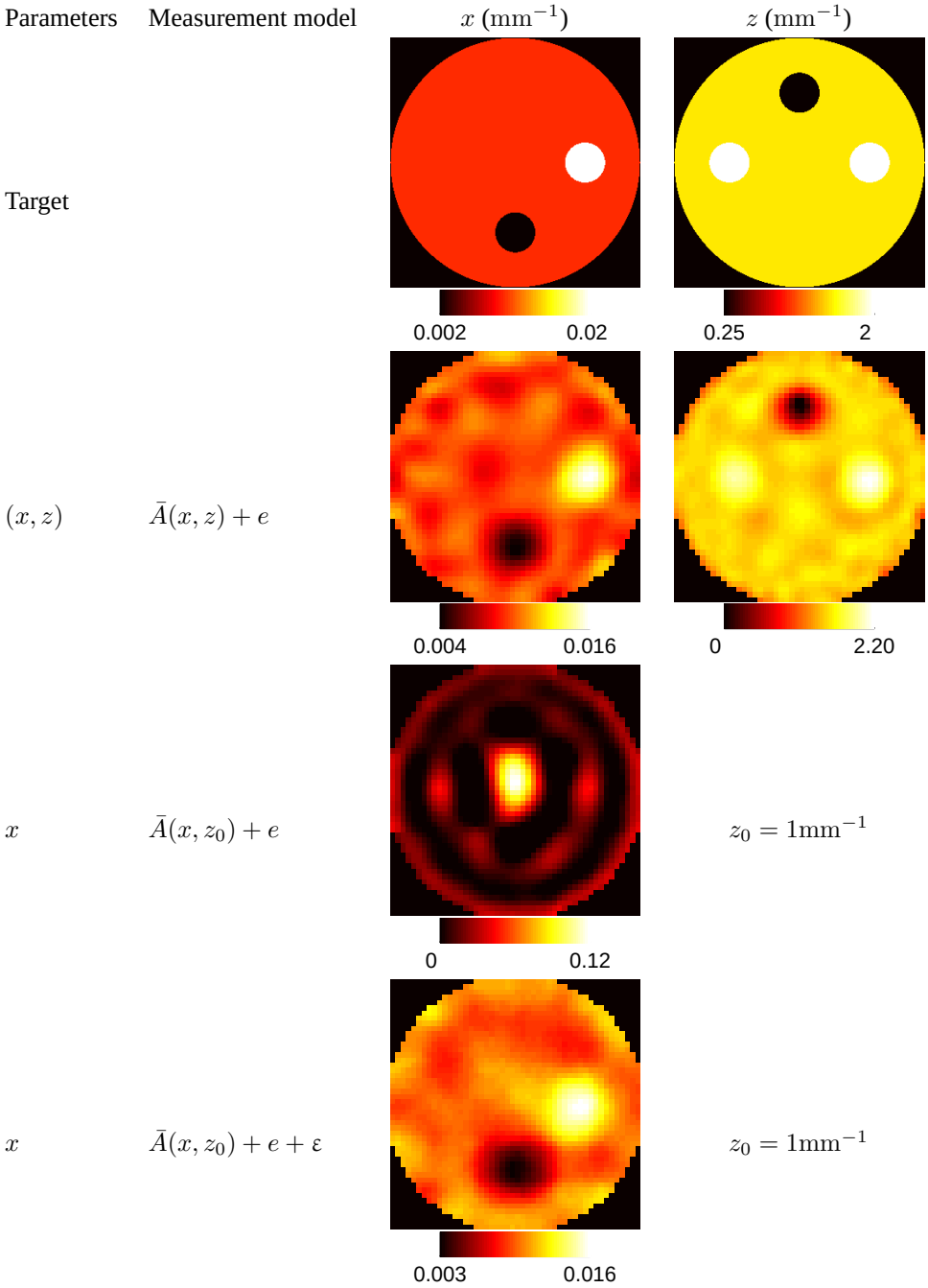

FIG. 3: Case 2. Misspecification of z_0. Rows as in Fig. 2. The modeling error in MAP-CEM and MAP-AEM is caused solely by using a fixed value $z = z_0$. Here $z_{\text{bg,actual}} = z_* + 1.5\sigma_z = z_0 + 1.5\sigma_z$.

approach is able to compensate for using an incorrect fixed value for the uninteresting distributed parameter. In this particular example, the premarginalization was carried over the scattering coefficient, as well as the errors related to simultaneous reduction of the computational forward model. It was also shown that at least in the studied cases, the approach is tolerant to a reasonable misspecification of the fixed scattering coefficient.

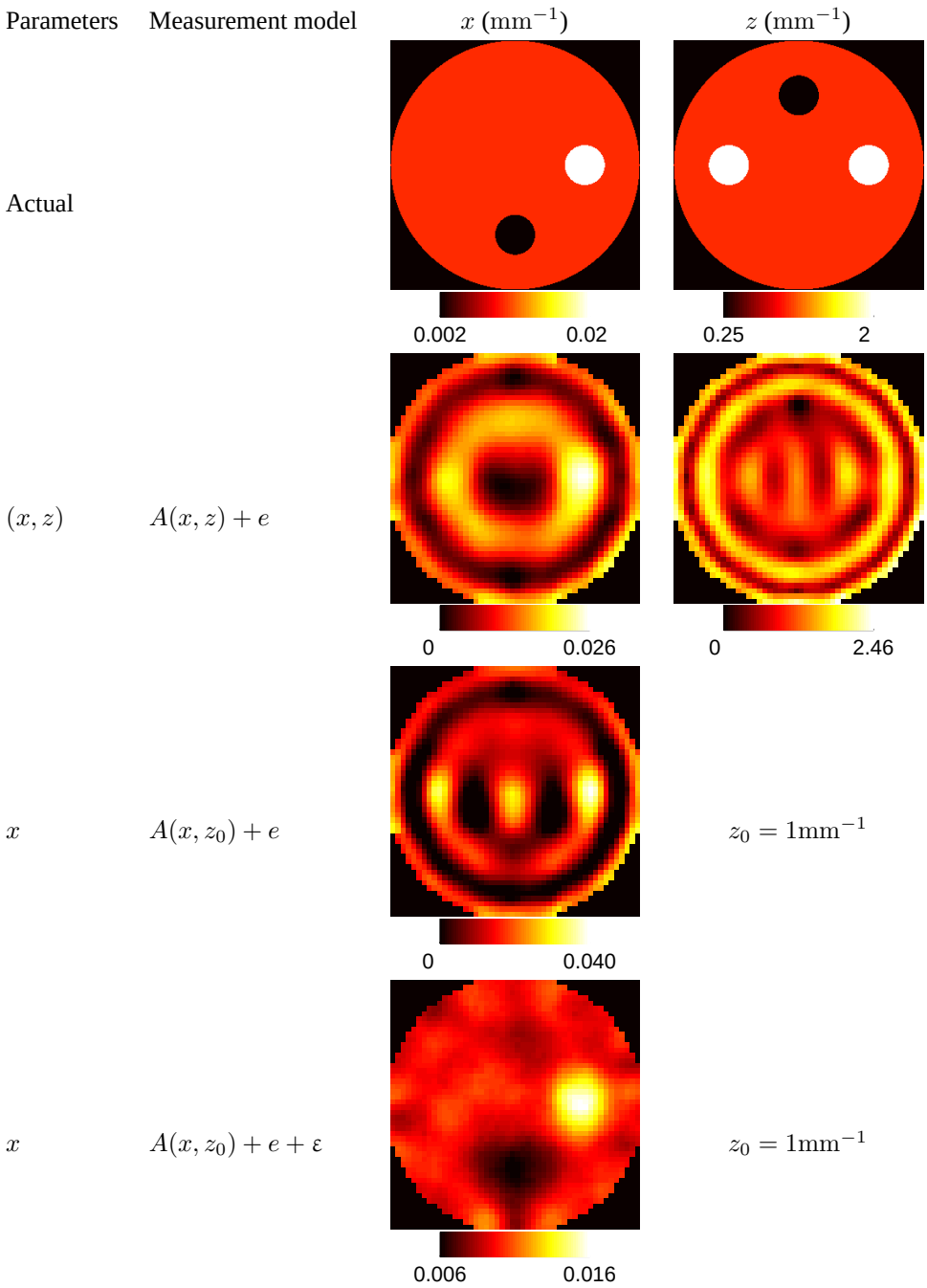

FIG. 4: Case 3. Rows as in Fig. 2. The modeling error in MAP-CEM and MAP-AEM is caused by using a fixed value $z = z_0$ and the reduced-order model $A(x, z_0)$ for the forward problem. Here $z_{\mathrm{bg,actual}} = z_* = z_0$.

ACKNOWLEDGMENTS

This work was supported by the Academy of Finland, projects 119270, 122499, 218183, 140731 and 213476, the Finnish Centre of Excellence in Inverse Problems Research 2006-2011, and by EPSRC project EP/E034950/1.

REFERENCES

1. Daily, W. D., Ramirez, A. L., LaBrecque, D. J., and Nitao, J., Electrical resistivity tomography of vadose water movement, *Water Resour. Res.*, 28:1429–1442, 1992.
2. Daily, W. D. Ramirez, A. L., Electrical resistivity tomography, *The Leading Edge*, 23(1):438–442, 2004.
3. Arridge, S. R., Optical tomography in medical imaging, *Inverse Probl.*, 15:R41–R93, 1999.
4. Gibson, A. P., Hebden, J. C., and Arridge, S. R., Recent advances in diffuse optical imaging, *Phys. Med. Biol.*, 50:R1–R43, 2005.
5. Kaipio, J. and Somersalo, E., *Statistical and Computational Inverse Problems*, Springer, New York, 2005.
6. Kaipio, J. and Somersalo, E., Statistical inverse problems: Discretization, model reduction and inverse crimes, *J. Comput. Appl. Math.*, 198:493–504, 2007.
7. Arridge, S., Kaipio, J., Kolehmainen, V., Schweiger, M., Somersalo, E., Tarvainen, T., and Vauhkonen, M., Approximation errors and model reduction with an application in optical diffusion tomography, *Inverse Probl.*, 22:175–195, 2006.
8. Kolehmainen, V., Schweoger, M., Nissilä, I., Tarvainen, T., Arridge, S., and Kaipio, J., Approximation errors and model reduction in three-dimensional optical tomography, *J. Opt. Soc. Am. A*, 26:2257–2268, 2009.
9. Heino, J. and Somersalo, E., A modelling error approach for the estimation of optical absorption in the presence of anisotropies, *Phys. Med. Biol.*, 49:4785–4798, 2004.
10. Heino, J., Somersalo, E., and Kaipio, J., Compensation for geometric mismodelling by anisotropies in optical tomography, *Opt. Express*, 13(1):296–308, 2005.
11. Calvetti, D., Kaipio, J. P., and Somersalo, E., Aristotelian prior boundary conditions, *Int. J. Math.*, 1:63–81, 2006.
12. Lehikoinen, A., Finsterle, S., Voutilainen, A., Heikkinen, L., Vauhkonen, M., and Kaipio, J., Approximation errors and truncation of computational domains with application to geophysical tomography, *Inverse Probl. Imag.*, 1:371–389, 2007.
13. Nissinen, A., Heikkinen, L., and Kaipio, J. P., Approximation errors in electrical impedance tomography—An experimental study, *Meas. Sci. Technol.*, 19:015501, 2008.
14. Nissinen, A., Heikkinen, L., Kolehmainen, V., and Kaipio, J. P., Compensation of errors due to discretization, domain truncation and unknown contact impedances in electrical impedance tomography, *Meas. Sci. Technol.*, 20:105504, 2009.
15. Nissinen, A., Kolehmainen, V., and Kaipio, J. P., Compensation of modelling errors due to unknown domain boundary in electrical impedance tomography, *IEEE Trans. Med. Imag.*, 2010, in press.
16. Huttunen, J. and Kaipio, J., Approximation errors in nostationary inverse problems, *Inverse Probl. Imag.*, 1(1):77–93, 2007.
17. Huttunen, J. and Kaipio, J., Approximation error analysis in nonlinear state estimation with an application to state-space identification, *Inverse Probl.*, 23:2141–2157, 2007.
18. Huttunen, J. and Kaipio, J., Model reduction in state identification problems with an application to determination of thermal parameters, *Appl. Numer. Math.*, 59:877–890, 2009.
19. Seppänen, A., Vauhkonen, M., Vauhkonen, P., Somersalo, E., and Kaipio, J., State estimation with fluid dynamical evolution models in process tomography — An application to impedance tomography, *Inverse Probl.*, 17:467–484, 2001.
20. Huttunen, J., Lehikoinen, A., Hämäläinen, J., and Kaipio, J., Importance filtering approach for the nonstationary approximation error method, *Inverse Probl.*, 2010, in press.
21. Lehikoinen, A., Huttunen, J., Voutilainen, A., Finsterle, S., Kowalsky, M., and Kaipio, J. P., Dynamic inversion for hydrological process monitoring with electrical resistance tomography under model uncertainties, *Water Resour. Res.*, 46:W04513, 2010.
22. Tarvainen, T., Kolehmainen, V., Pulkkinen, A., Vauhkonen, M., Schweiger, M., Arridge, S. R., and Kaipio, J. P., Approximation error approach for compensating for modelling errors between the radiative transfer equation and the diffusion approximation in diffuse optical tomography, *Inverse Probl.*, 26:015005, 2010.
23. Tarantola, A., *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, 2004.
24. Calvetti, D. and Somersalo, E., *An Introduction to Bayesian Scientific Computing, Ten Lectures on Subjective Computing*, ISBN 978-0-387-73393-7, Springer, 2007.
25. Berger, J., *Statistical Decision Theory and Bayesian Analysis*, Springer, 2006.
26. Robert, C., *The Bayesian Choice*, Springer, 2007.

27. Chen, M.-H., Shao, Q.-M., and Ibrahim, J., *Monte Carlo Methods in Bayesian Computation*, Springer, 2000.
28. J. Liu, *Monte Carlo Strategies in Scientific Computing*, ISBN 0-387-95230-6 in Springer Series in Statistics, Springer, 2005.
29. Somersalo, E., Kaipio, J., Vauhkonen, M., Baroudi, D., and Järvenpää, S., Impedance Imaging and Markov Chain Monte Carlo Methods, In R. Barbour, M. Carvlin, and M. Fiddy (Eds.), *Proc SPIE's 42nd Annual Meeting, Computational, Experimental and Numerical Methods for Solving Ill-Posed Inverse Imaging Problems: Medical and Nonmedical Applications*, pp. 175–185, San Diego, USA, June 27-Aug 1, 1997.
30. Fox, C. and Nicholls, G., Sampling conductivity images via MCMC, In K. V. Mardia, C. A. Gill, and R. G. Aykroyd (Eds.), *The Art and Science of Bayesian Image Analysis, Proceedings of the Leeds Annual Statistics Research Workshop*, pp. 91–100, Leeds University Press, Leeds, UK, 1–4 July 1997.
31. Kaipio, J. P., Kolehmainen, V., Somersalo, E., and Vauhkonen, M., Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography, *Inverse Probl.*, 16:1487–1522, 2000.
32. Arridge, S. and Schotland, J., Optical tomography: Forward and inverse problems, *Inverse Probl.*, 25:123010, 2009.
33. Tromberg, B. J., Pogue, B. W., Paulsen, K. D., Yodh, A. G., Boas, D. A., and Cerussi, A. E., Assessing the future of diffuse imaging technologies for breast cancer management, *Med. Phys.*, 35:2443–2451, 2008.
34. Schweiger, M., Arridge, S. R., Hiraoka, M., and Delpy, D. T., The finite element model for the propagation of light in scattering media: Boundary and source conditions, *Med. Phys.*, 22(11):1779–1792, 1995.
35. Heino, J. and Somersalo, E., Estimation of optical absorption in anisotropic background, *Inverse Probl.*, 18:559–573, 2002.
36. Nocedal, J. and Wright, S. J., *Numerical Optimization*, 2nd ed., Springer, New York, 2006.

International Journal for Uncertainty Quantification, 1(1):19–33, 2011

ASSESSMENT OF COLLOCATION AND GALERKIN APPROACHES TO LINEAR DIFFUSION EQUATIONS WITH RANDOM DATA

Howard C. Elman,[1,*] Christopher W. Miller,[2] Eric T. Phipps,[3] & Raymond S. Tuminaro[4]

[1]*Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA*

[2]*Department of Applied Mathematics and Scientific Computation, University of Maryland, College Park, MD 20742, USA*

[3]*Sandia National Laboratories, PO Box 5800, MS 1318, Albuquerque, NM 87185, USA*

[4]*Sandia National Laboratories, PO Box 969, MS 9159, Livermore, CA 94551*

Original Manuscript Submitted: 03/22/2010; Final Draft Received: 07/07/2010

We compare the performance of two methods, the stochastic Galerkin method and the stochastic collocation method, for solving partial differential equations (PDEs) with random data. The stochastic Galerkin method requires the solution of a single linear system that is several orders larger than linear systems associated with deterministic PDEs. The stochastic collocation method requires many solves of deterministic PDEs, which allows the use of existing software. However, the total number of degrees of freedom in the stochastic collocation method can be considerably larger than the number of degrees of freedom in the stochastic Galerkin system. We implement both methods using the Trilinos software package and we assess their cost and performance. The implementations in Trilinos are known to be efficient, which allows for a realistic assessment of the computational complexity of the methods. We also develop a cost model for both methods which allows us to examine asymptotic behavior.

KEY WORDS: *uncertainty quantification, stochastic partial differential equations, polynomial chaos, stochastic Galerkin method, stochastic sparse grid collocation, Karhunen-Loève expansion*

1. PROBLEM STATEMENT

We investigate the linear elliptic diffusion equation with zero Dirichlet boundary conditions, where diffusivity is given by a random field. If D is an open subset of $\mathbb{R}^n$ and (Ω, Σ, P) is a complete probability space, then this can be written as

$$\begin{aligned} -\nabla \cdot [a(\mathbf{x}, \omega)\nabla u(\mathbf{x}, \omega)] &= f(\mathbf{x}, \omega) && (\mathbf{x}, \omega) \in D \times \Omega \\ u(x, \omega) &= 0 && (\mathbf{x}, \omega) \in \partial D \times \Omega. \end{aligned} \tag{1}$$

The random input field is often given as a truncated Karhunen–Loève (KL) expansion [1] or by a polynomial chaos (PC) expansion [2]. The truncated KL expansion is given by

*Correspond to Howard C. Elman, E-mail: elman@cs.umd.edu

2152 -5080/11/$35.00

$$a(\mathbf{x},\omega) \approx \hat{a}_M(\mathbf{x},\boldsymbol{\xi}(\omega)) = a_0(\mathbf{x}) + \sum_{k=1}^{M} \sqrt{\lambda_k}\xi_k(\omega)a_k(\mathbf{x}), \tag{2}$$

where (λ_i, a_i) are solutions to the integral equation

$$\int_D C(\mathbf{x}_1,\mathbf{x}_2)a_i(\mathbf{x}_2)d\mathbf{x}_2 = \lambda_i a_i(\mathbf{x}_1), \tag{3}$$

and C is the covariance kernel of the random field. That is, (λ_i, a_i) are eigenvalues and eigenfunctions of the covariance operator $\mathcal{C}$ defined by

$$[\mathcal{C}(\alpha)](x_1) = \int_D C(x_1,x_2)\alpha(x_2)\,dx_2. \tag{4}$$

The random variables are uncorrelated, mean zero, and are given by

$$\xi_k(\omega) = \frac{1}{\sqrt{\lambda_k}}\int_D [a(\mathbf{x},\omega) - a_0(\mathbf{x})]\,a_k(\mathbf{x})\,d\mathbf{x}. \tag{5}$$

We make the further modeling assumption that the random variables $\{\xi_k\}$ are independent and admit a joint probability density of the form $\rho(\boldsymbol{\xi}) = \prod_{k=1}^{M}\rho_k(\xi_k)$. The covariance kernel is positive semidefinite and its eigenvalues can be ordered so that $\lambda_1 \geq \lambda_2 \geq ... \geq 0$. To ensure the existence of a unique solution to (1) it is necessary to assume that the diffusion is uniformly bounded away from zero; we assume that there exist constants $a_{\min}$ and $a_{\max}$ such that

$$0 < a_{\min} \leq \hat{a}_M(\mathbf{x},\boldsymbol{\xi}) \leq a_{\max} < \infty, \tag{6}$$

almost everywhere P-almost surely, $\hat{a}_M(\cdot,\boldsymbol{\xi}) \in L_2(D)$ P-almost surely, and $\hat{f}_M \in L_2(\Omega)\otimes L_2(D)$.

The goal of this paper is to model the computational costs and compare the performance of the stochastic Galerkin method [3–7] and the sparse grid collocation method [8–10] for computing the solution of (1) (cf. [11] for related work). Section 2 outlines the stochastic Galerkin method. Section 3 outlines the sparse grid collocation method. Section 4 presents our model of the computational costs of the two methods. Section 5 explores the performance of the methods applied to several numerical examples using the *Trilinos* software package [12]. Finally, in Section 6 we draw some conclusions.

2. STOCHASTIC GALERKIN METHOD

Define $\Gamma = \times_{k=1}^{M}\Gamma_k = \times_{k=1}^{M} Im(\xi_k)$ and let

$$\langle u, v\rangle = \int_\Gamma u(\boldsymbol{\xi})v(\boldsymbol{\xi})\rho(\boldsymbol{\xi})\,d\boldsymbol{\xi} = \int_\Omega u[\boldsymbol{\xi}(\omega)]v[\boldsymbol{\xi}(\omega)]\,dP \tag{7}$$

be the inner product over the space $L_2(\Gamma) = \{v(\boldsymbol{\xi}) : ||v||^2_{L_2(\Gamma)} = \langle v^2\rangle < \infty\}$. We can define a variational form of (1) in the stochastic domain by the following: For all $\mathbf{x} \in D$, find $u(\mathbf{x},\boldsymbol{\xi}) \in L_2(\Gamma)$ such that

$$-\langle \nabla\cdot(a\nabla u), v\rangle = \langle f, v\rangle \tag{8}$$

for all $v \in L_2(\Gamma)$. This leads to a set of coupled second-order linear partial differential equations (PDEs) in the spatial dimension. It is common in the literature to combine (8) with a variational formulation of the spatial component of the problem, which after discretization of both the spatial and stochastic components, leads to the stochastic finite element method. A variant of this approach, which we use, is to discretize in space by finite differences. Details are as follows.

Define S_p to be the space of multivariate polynomials in $\boldsymbol{\xi}$ of total degree at most p. This space has dimension $N_\xi = [(M+p)!]/(M!p!)$. Let $\{\Psi_k\}_{k=0}^{N_\xi-1}$ be a basis for S_p orthonormal with respect to the inner product (7). Substituting KL-expansions for $a(\mathbf{x},\omega)$ and $f(\mathbf{x},\omega)$ and restricting (8) to $v \in S_p$ gives

$$-\int_\Gamma \nabla\cdot\left[\hat{a}_M(\mathbf{x},\boldsymbol{\xi})\left(\sum_{i=0}^{N_\xi-1}\nabla u_i(\mathbf{x})\Psi_i\right)\right]\Psi_j\, d\boldsymbol{\xi} = \int_\Gamma \hat{f}_M\Psi_j\, d\boldsymbol{\xi},\ \ \forall\, j=0:N_\xi-1. \tag{9}$$

This is a set of coupled second-order differential equations for the unknown functions $u_i(\mathbf{x})$ defined on D, which can then be discretized using finite differences. This gives rise to a global linear system of the form

$$A\vec{u} = \vec{f}. \tag{10}$$

In practice the random variables appearing in the KL expansion of $a(\mathbf{x},\omega)$ and $f(\mathbf{x},\omega)$ would be different since the diffusivity and loading terms would typically have different correlation structures. In this case one would expand a, f, and u as

$$a(\mathbf{x},\omega) \approx \hat{a}_M(\mathbf{x},\boldsymbol{\xi}) = a_0(\mathbf{x}) + \sum_{k=1}^{M}\sqrt{\lambda_k}\xi_k a_k(\mathbf{x}) \tag{11}$$

$$f(\mathbf{x},\omega) \approx \hat{f}_M(\mathbf{x},\tilde{\boldsymbol{\xi}}) = f_0(\mathbf{x}) + \sum_{k=1}^{M}\sqrt{\tilde{\lambda}_k}\tilde{\xi}_k f_k(\mathbf{x}) \tag{12}$$

$$u(\mathbf{x},\omega) \approx u(\mathbf{x},\xi_1,...,\xi_M,\tilde{\xi}_1,...,\tilde{\xi}_k), \tag{13}$$

where $\tilde{\lambda}_k$ and $\tilde{\xi}_k$ are the eigenvalues and random variables appearing in the KL expansion of f. For the sake of simplicity we choose to ignore this issue and proceed as if the random variables appearing in the KL expansion of f and a are the same.

With orderings of $\vec{u}$ and $\vec{f}$ (equivalently, the columns and rows of A, respectively) corresponding to a blocking by spatial degrees of freedom, $\left(\vec{u}^T = \left[u_1^T, u_2^T, \cdots, u_{N_\xi}^T\right]\right)$, the coefficient matrix and right-hand side have the tensor product structure

$$A = \sum_{k=0}^{M} G_k \otimes A_k, \quad \vec{f} = \sum_{k=0}^{M}\vec{g_k}\otimes\vec{f_k}. \tag{14}$$

The matrices $\{G_k\}$ and the vectors $\{g_k\}$ depend only on the stochastic basis,

$$\begin{aligned} G_0(i,j) &= \langle\Psi_i\Psi_j\rangle,\ g_0(i) = \langle\Psi_i\rangle = \delta_{i0}, \\ G_k(i,j) &= \langle\xi_k\Psi_i\Psi_j\rangle,\ g_k(i) = \langle\xi_k\Psi_i\rangle,\ (k>0). \end{aligned} \tag{15}$$

The matrices $\{A_k\}$ correspond to a standard five-point operator for $-\nabla\cdot(a_k\nabla u)$, and $\{f_k\}$ are the associated right-hand side vectors. In the two-dimensional examples we explore below, we use a uniform mesh of width h. The discrete difference operators are formed by using the following five-point stencil

$$\begin{bmatrix} & a_k\left(x, y+\frac{h}{2}\right) & \\ a_k\left(x-\frac{h}{2}, y\right) & a_k(x,y) & a_k\left(x+\frac{h}{2}, y\right) \\ & a_k\left(x, y-\frac{h}{2}\right) & \end{bmatrix}. \tag{16}$$

The matrix A_k is symmetric for all k and A is positive-definite by (6). Since the random variables appearing in (5) are mean-zero, it also follows from (6) that A_0 is positive-definite.

The matrix A is of order $N_x N_\xi$, where N_x is the number of degrees of freedom used in the spatial discretization. It is also sparse in the block sense due to the orthogonality of the stochastic basis functions. Specifically, since the random variables $\{\xi_k\}$ are assumed to be independent, we can construct the stochastic basis functions $\{\Psi_i\}$ by taking tensor products of univariate polynomials satisfying the orthogonality condition

$$\langle \psi_i(\xi_k), \psi_j(\xi_k)\rangle_k = \int_{\Gamma_k} \psi_i(\xi_k)\psi_j(\xi_k)\rho_k(\xi_k)\, d\xi_k = \delta_{ij}. \tag{17}$$

This basis is referred to as the generalized polynomial chaos of order p. The use of this basis for representing random fields is discussed extensively in [4] and [7]. The univariate polynomials appearing in the tensor product can be expressed via the familiar three-term recurrence

$$\psi_{i+1}(\xi_k) = (\xi_k - \alpha_i)\psi_i(\xi_k) - \beta_i \psi_{i-1}(\xi_k), \tag{18}$$

where $\psi_0 = 1,\ \psi_{-1} = 0$. It follows that

$$G_0(i,j) = \langle \Psi_i, \Psi_j\rangle = \prod_{k=1}^{M} \langle \psi_{i_k}(\xi_k), \psi_{j_k}(\xi_k)\rangle_k = \prod_{k=1}^{M} \delta_{i_k j_k} = \delta_{ij}, \tag{19}$$

and for $k > 0$ the entries in G_k are

$$\begin{aligned} G_k(i,j) &= \langle \xi_k \Psi_i, \Psi_j\rangle = \langle \xi_k \psi_{i_k}, \psi_{j_k}\rangle_k \prod_{l=1, l\neq k}^{M} \langle \psi_{i_l}, \psi_{j_l}\rangle_l \\ &= (\langle \psi_{i_k+1}, \psi_{j_k}\rangle_k + \alpha_{i_k}\langle \psi_{i_k}, \psi_{j_k}\rangle_k + \beta_{i_k}\langle \psi_{i_k-1}, \psi_j\rangle_k) \prod_{l=1, l\neq k}^{M} \langle \psi_{i_l}, \psi_{j_l}\rangle_l. \end{aligned} \tag{20}$$

Thus G_0 is diagonal and G_k has at most three entries per row for $k > 0$. Furthermore, if the density functions ρ_k are symmetric with respect to the origin, i.e., $\rho_k(\xi_k) = \rho_k(-\xi_k)$, then the coefficients α_i in the three-term recurrence are all zero and G_k then has at most two non-zeros per row.

The stochastic Galerkin method requires the solution to the large linear system (10). Once the solution to (10) is obtained, statistical quantities such as moments or a probability distribution associated with the solution process can be obtained cheaply [4]. Although the Galerkin linear system is large, there are techniques available by which this task can be performed efficiently. We elect to directly solve the large symmetric and positive-definite Galerkin system using the conjugate gradient (CG) method. CG only requires the evaluation of matrix–vector products so that it is unnecessary to store the assembled matrix A. The matrix–vector products can be performed implicitly following a procedure described in [13]. Each matrix A_k is assembled and the matrix–vector product is expressed as $(Au)_j = \sum_{i=0}^{N_\xi - 1} \sum_{k=0}^{M} \langle \xi_k \Psi_i \Psi_j\rangle (A_k u_i)$. The terms $A_k u_i$ are precomputed and then scaled as needed. This approach is efficient since most of the terms $\langle \xi_k \Psi_i \Psi_j\rangle$ are zero. The cost of performing the matrix–vector product in this manner is essentially determined by the computation of $A_k u_i$ for $0 \le k \le M$ and $0 \le i \le N_\xi - 1$, which entails $(M+1)N_\xi$ sparse matrix–vector products by matrices $\{A_k\}$ of order N_x. The implicit matrix–vector product also only requires the assembly of $M+1$ order-N_x stiffness matrices and the assembly of the components $\langle \xi_k \Psi_i \Psi_j\rangle$ of $\{G_k\}$. Alternatively, one could assemble the entire Galerkin matrix and perform the block matrix–vector product in the obvious way. This is, of course, less efficient in terms of memory usage since it requires the assembly and storage of many matrices of the form $\langle \xi_k \Psi_i \Psi_j\rangle (A_k u_i)$. It is also shown in [13] that performing the matrix–vector products in this way is less efficient in terms of memory bandwidth.

To obtain fast convergence, we will also use a preconditioner. In particular, it has been shown in [14] that an effective choice is an approximation to $A_0^{-1} \otimes G_0^{-1}$, where A_0 is the mean stiffness matrix. Since the stochastic basis functions are orthonormal, G_0 is the identity matrix. The preconditioner then entails the approximate action of N_ξ uncoupled copies of A_0^{-1}. For this we use a single iteration of an algebraic multigrid solver provided by [15].

3. SPARSE GRID COLLOCATION

An alternative to the Galerkin scheme is the collocation method, which samples the input operator at a predetermined set of points $\Theta = \left\{\boldsymbol{\xi}^{(1)}, ..., \boldsymbol{\xi}^{(n)}\right\}$ and constructs a high-order polynomial approximation to the solution function using discrete solutions to the deterministic PDEs

$$-\nabla \cdot \left[\hat{a}_M\left(\mathbf{x}, \boldsymbol{\xi}^{(l)}\right) \nabla u\left(\mathbf{x}, \boldsymbol{\xi}^{(l)}\right)\right] = \hat{f}_M\left(\mathbf{x}, \boldsymbol{\xi}^{(l)}\right), \tag{21}$$

where the diffusion coefficients are evaluated at the sample points. Once the polynomial approximation to u is constructed, statistical information can be obtained at low cost [10], as for the stochastic Galerkin method.

For simplicity of presentation, we first discuss a collocation method using the full tensor product of one-dimensional point sets. Let $\{\psi_i\}$ be the set of polynomials orthogonal with respect to the measure ρ_k. Let $\theta_i = \{\xi : \psi_i(\xi) = 0\} := \left\{\xi_{i,k}^{(j)}\right\}_{j=1}^{i+1}$ for $i = 1, 2, ...,$ and $j = 1, 2, ..., i$. These are the abscissas for an (i)-point Gauss quadrature rule with respect to the measure ρ_k. A one-dimensional (i)-point interpolation operator is given by

$$U^i(u)(\xi) = \sum_{j=1}^{i} u\left(\xi_i^{(j)}\right) l_i^{(j)}(\xi),\ l_i^{(j)}(\xi) = \prod_{n=1, n\neq j}^{i} \frac{\xi - \xi_i^{(n)}}{\xi_i^{(j)} - \xi_i^{(n)}}. \tag{22}$$

These can be used to construct an approximation to the M-dimensional random function $u(\mathbf{x}, \boldsymbol{\xi})$ by defining a tensor interpolation operator

$$U^{i_1} \otimes \cdots \otimes U^{i_M}(u)(\boldsymbol{\xi}) = \sum_{j_1=1}^{i_1} \cdots \sum_{j_M=1}^{i_M} u\left(\xi_{i_1}^{(j_1)}, \cdots, \xi_{i_M}^{(j_M)}\right)\left(l_{i_1}^{(j_1)} \otimes \cdots l_{i_M}^{(j_M)}\right). \tag{23}$$

The evaluation of this operator requires the solution of a collection of deterministic PDEs (21), one for each sample point in $\Theta_{\text{tensor}} = \times_{j=1}^{M} \theta_{i_j}$.

This method suffers from the so-called curse of dimensionality, since the number of sample points $|\Theta_{\text{tensor}}| = \prod_{j=1}^{M} |\theta_{i_j}| = \prod_{j=1}^{M} (i_j)$ grows exponentially with the dimension of the problem. This makes tensor-product collocation inappropriate for problems where the stochastic dimension is moderate or large. This cost can be significantly reduced using sparse grid methods [10].

Sparse grid collocation methods are based on the Smolyak approximation formula. The Smolyak operator $\mathcal{A}(p, M)$ is a linear combination of the product formulas in (23). Let $Y(p, M) = \{\boldsymbol{i} \in \mathbb{N}^M : p + 1 \leq |\mathbf{i}|_1 \leq p + M\}$. Then the Smolyak formula is given by

$$\mathcal{A}(p, M)(u) = \sum_{\mathbf{i} \in Y(p,M)} (-1)^{p+M-|\mathbf{i}|_1} \binom{M-1}{p+M-|\mathbf{i}|_1} (U^{i_1} \otimes \cdots \otimes U^{i_M}). \tag{24}$$

The evaluation of the Smolyak formula requires the solution of deterministic PDEs (21) for $\boldsymbol{\xi}^{(l)}$ in the set of points

$$\Theta_{p,M} = \bigcup_{\mathbf{i} \in Y(p,M)} (\theta_{i_1} \times \cdots \times \theta_{i_M}). \tag{25}$$

For moderate or large values of M, $|\Theta_{p,M}| \ll |\Theta_{\text{tensor}}|$.

If Gaussian abscissas are used in the definition of θ_i and if u is an M-variate polynomial of total degree p in $\boldsymbol{\xi}$, then $u = \mathcal{A}(p, M)u$ [11]; that is, the Smolyak interpolant exactly reproduces such polynomials.[1] We refer to the

[1] An alternative choice of sparse grid points is to use the Clenshaw–Curtis abscissas with $|\theta_1| = 1$ and $|\theta_i| = 2^{i-1} + 1$ for $i > 1$, which produces nested sparse grids [9, 10, 16]. The choice used here, non-nested Gaussian abscissas with a linear growth rate, $|\theta_i| = i$, produces grid sets of cardinalities comparable to those for the nested Clenshaw–Curtis grids, i.e., $|\Theta_{p,M}^{\text{Gaussian}}| \approx |\Theta_{p,M}^{\text{Clenshaw-Curtis}}|$.

parameter p in $\mathcal{A}(p, M)$ as the sparse grid level. It is shown in [9] that sampling the differential operator on the sparse grid $\Theta_{p,M}$ will produce $\mathcal{A}(p, M)(u) = u_p$, where u_p is an approximate solution to (1) of similar accuracy to the solution obtained using an order p stochastic Galerkin scheme. The sparse grid will have on the order of 2^p more points than there are stochastic degrees of freedom in the Galerkin scheme, $|\Theta| \approx 2^p N_\xi$ for $M \gg 1$ [10].

For a fully nonintrusive collocation method, the diffusion coefficients of (21) would be sampled at the points in the sparse grid, and for each sample the deterministic stiffness matrix would be constructed for the PDE,

$$-\nabla \cdot \left[\hat{a}_M\left(\mathbf{x}, \boldsymbol{\xi}^{(l)}\right) \nabla u\left(\mathbf{x}, \boldsymbol{\xi}^{(l)}\right)\right] = \hat{f}_M\left(\mathbf{x}, \boldsymbol{\xi}^{(l)}\right). \tag{26}$$

This repeated assembly can be very expensive. We elect in our implementations to take advantage of the fact that the stiffness matrix at a given value of the random variable is a scaled sum of the stiffness matrices appearing in (14). For a given value of $\boldsymbol{\xi}$, the deterministic stiffness matrix can be expressed as

$$A(\boldsymbol{\xi}) = A_0 + \sum_{k=1}^{M} \xi_i A_k. \tag{27}$$

In our implementation we assemble the matrices $\{A_k\}$ first and then compute the scaled sum (27) at each collocation point. This is somewhat intrusive in that this method may not be compatible with with existing deterministic solvers; however, it greatly reduces the amount of time required to perform assembly in the collocation method.

One could construct a separate multigrid preconditioner for each of the deterministic systems. This can become very expensive, as the cost of constructing an algebraic multigrid (AMG) preconditioner can often be of the same order as the iterative solution. This repeated cost can be eliminated if one simply builds an algebraic preconditioner for the mean problem A_0^{-1} and applies this preconditioner to all of the deterministic systems. If the variance of the operator is small, then the mean-based AMG preconditioner is nearly as effective as doing AMG on each subproblem and saves time in setup costs. Other techniques for developing preconditioners balancing performance with the cost of repeated construction are considered in [16].

4. MODELING COMPUTATIONAL COSTS

From an implementation perspective, collocation is quite advantageous in that it requires only a modest modification to existing deterministic PDE applications. Collocation samples the stochastic domain at a discrete set of points and requires the solution of uncoupled deterministic problems. This can be accomplished by repeatedly invoking a deterministic application with different input parameters determined by the collocation point-sampling method. A Galerkin method, on the other hand, is much more intrusive as it requires the solution of a system of equations with a large coefficient matrix which has been discretized in both spatial and stochastic dimensions. To better understand the relationship between these two methods, we develop a model for the computational costs.

We begin by stating in more detail some of the computational differences between the two methods. The Galerkin method requires the computation of the matrices $G_0 = \langle \Psi_i \Psi_j \rangle$ and $G_k = \langle \xi_k \Psi_i \Psi_j \rangle$ associated with the stochastic basis functions, the assembly of the right-hand side vector and the spatial stiffness matrices $\{A_k\}$, and finally, the solution to the large coupled system of equations. Collocation requires the construction of a sparse grid and the derivation of an associated sparse grid quadrature rule, and the assembly/solution of a series of deterministic subproblems. Further, as observed above, the number of sample points needed for collocation tends to be much larger than the dimension of the Galerkin system required to achieve comparable accuracy.

In this study we examine only methods which are isotropic in the stochastic dimension, allocating an equal number of degrees of freedom to each stochastic direction. Anisotropic versions of both the sparse grid collocation method and the stochastic Galerkin could be implemented by weighting the maximum degree of the approximation space in each direction. This has been explored in the case of sparse grid collocation [17]. We expect a cost comparison for an anisotropic stochastic Galerkin method and the anisotropic sparse grid collocation method to be comparable to that of their isotropic counterparts. Additional modifications to the stochastic collocation for adaptively dealing with very high dimensional problems are considered in [18, 19]. We do not consider these methods here.

For a fixed M, p, let Z_G be the number of preconditioned conjugate gradient (PCG) iterations required to solve the Galerkin system, let $N_\xi \alpha$ be the cost of applying the mean-based preconditioner during a single iteration of the stochastic Galerkin method, and let $N_\xi \gamma$ be the cost of a single matrix–vector product for (10), where α and γ are constants. Note in particular that α is constant because of the optimality of the multigrid computation. Then the total cost of the Galerkin method can be modeled by

$$\text{Galerkin cost} = N_\xi Z_G(\alpha + \gamma). \tag{28}$$

The parameter γ can be thought of as the number of order-N_x matrix–vector products required per block row in the stochastic Galerkin matrix. When implementing the implicit matrix–vector product, γ is equal to $M + 1$.

We can model the costs of the collocation method with the mean-based multigrid preconditioner by

$$\text{Collocation cost} = Z_C 2^p N_\xi(\alpha + 1), \tag{29}$$

where p is the Smolyak grid level, N_ξ is the number of degrees of freedom needed by an order p Galerkin system, Z_C is the average number of PCG iterations needed to solve a single deterministic system, and $\alpha + 1$ is the cost of the preconditioning operation and a single order-N_x matrix–vector product. The factor of 2^p derives from the relation between the number of degrees of freedom for the stochastic Galerkin and sparse grid collocation methods for large M.

In our application, we fix the multigrid parameters as follows: One V-cycle is performed at each iteration and within each V-cycle one symmetric Gauss–Seidel iteration is used for both presmoothing and postsmoothing. The coarsest grid is assumed coarse enough so that a direct solver can be used without affecting the cost per iteration; in our implementations we use a 1×1 grid. These parameters were chosen to optimize the run time of a single deterministic solve. The cost to apply a single multigrid iteration is roughly equivalent to 5–6 matrix products (two matrix–vector products for fine-level presmoothing, another two for fine-level postsmoothing, and one matrix–vector product for a fine-level residual calculation). Thus, α can be assumed to be 5 or 6 after accounting for computational overhead.

The relative costs of the two methods depend on the parameter values. In particular,

$$\frac{\text{Collocation cost}}{\text{Galerkin cost}} = \left(\frac{Z_C}{Z_G}\right) 2^p \frac{(\alpha + \gamma)}{(\alpha + 1)}. \tag{30}$$

If, for example, the ratio of iteration counts (Z_G/Z_C) is close to 1 and the preconditioning costs dominate the matrix vector costs (i.e., $\alpha \gg \gamma$), then we can expect the stochastic Galerkin method to outperform the sparse grid collocation method because of the factor 2^p. Alternatively, if γ is comparable compared to α, the preconditioning cost, then collocation is more attractive. The cost of the two methods is identical when (29) and (28) are equal. After canceling terms this gives $2^p \alpha \approx (Z_{SG}/Z_C)(\alpha + \gamma)$. Table 1 gives values of N_ξ and $|\Theta|$ for various values of M and p. One can observe that $2^p N_\xi \approx |\Theta|$ is a slight overestimation, but it improves as M grows larger. For reference, the number of points used by a full tensor product grid is also shown.

In the remainder of this paper, we explore the model and assess the validity of assumptions. In particular, we compare the accuracy of a level-p Smolyak grid with a degree-p polynomial approximation in the Galerkin approach. We also investigate the cost of matrix–vector products and the convergence behavior of mean-based preconditioning.

5. EXPERIMENTAL RESULTS AND MODEL VALIDATION

In this section we present the results of numerical experiments with both discretization methods, with the aims of comparing their accuracy and solution costs and validating the model developed in the previous section. First, we investigate a problem with a known solution to verify that both methods are converging to the correct solution and to examine the convergence of the PCG iteration. Second, we examine two problems where the diffusion coefficient is defined using a known covariance function, and we measure the computational effort required by each method.

TABLE 1: Degrees of freedom for various methods.

$M = 2$	Level p sparse grid (Gaussian) $\lvert\Theta\rvert$	Galerkin N_ξ	Non-zero blocks per row in Galerkin matrix	Tensor grid
$p = 1$	5	3	2.33	4
$p = 2$	13	6	3.00	9
$p = 3$	29	10	3.40	16
$p = 4$	53	15	3.67	25
$M = 10$				
$p = 1$	21	11	2.82	1024
$p = 2$	221	66	4.33	59,049
$p = 3$	1581	286	5.62	1,048,576
$p = 4$	8761	1001	6.71	9,765,625
$M = 20$				
$p = 1$	41	21	2.90	1.04×10^6
$p = 2$	841	231	4.64	3.49×10^9
$p = 3$	11,561	1771	6.22	1.10×10^{12}

5.1 Behavior of the Preconditioned Conjugate Gradient Algorithm

For well-posed Poisson problems, PCG with a multigrid preconditioner converges rapidly. Since collocation entails the solution of multiple deterministic systems, we expect multigrid to behave well. For Galerkin systems, the performance of mean-based preconditioning is more complicated. To understand this we investigate the problem

$$-\nabla \cdot [a(\mathbf{x}, \xi)u(\mathbf{x}, \xi)] = f(\mathbf{x}, \xi) \tag{31}$$

in the domain $[-0.5, 0.5]^2$ with zero Dirichlet boundary conditions, where the diffusion coefficient given as a one-term KL expansion,

$$a(\mathbf{x}, \xi) = 1 + \sigma \frac{1}{\pi^2} \xi \cos\left[\frac{\pi}{2}(x^2 + y^2)\right]. \tag{32}$$

We choose the function

$$u = \exp(-|\xi|^2)16(x^2 - 0.25)(y^2 - 0.25) \tag{33}$$

as the exact solution, and the forcing term f is defined by applying (31) to u.

The diffusion coefficient must remain positive for the problem to remain well-posed. This is the case provided

$$\left|\sigma \frac{1}{\pi^2} \xi \cos\left(\frac{\pi}{2} r^2\right)\right| < 1, \tag{34}$$

which holds when $|\xi| < (\pi^2)/(\sigma)$. As a consequence of this, well-posedness cannot be guaranteed when ξ is unbounded. There are various ways this can be addressed. We assume here that the random variable in (32) has a *truncated Gaussian density*,

$$\rho(\xi) = \frac{1}{\int_{-c}^{c} \exp(-\frac{\xi^2}{2})\, d\xi} \exp\left(-\frac{\xi^2}{2}\right) \mathbf{1}_{[-c,c]}, \tag{35}$$

which corresponds to taking the diffusion coefficient from a screened sample where the screening value c is chosen to enforce the conditions (1.7) for ellipticity and boundedness. The cutoff parameter c is chosen to be equal to 2.575. For

this cutoff the area under a standard normal distribution between $\pm c$ is equal to 0.99. For this value of c, $|\xi| < 2.575$ and the problem is guaranteed to remain well posed provided that $\sigma < (\pi^2)/[\max(|\xi|)] = 3.8329$.

Polynomials orthogonal to a truncated Gaussian measure are referred to as Rys polynomials [20]. As the parameter c is increased, the measure approaches the standard Gaussian measure and the Rys polynomials are observed to approach the behavior of the Hermite polynomials. For our implementation of collocation, the sparse grids are based on the zeros of the Rys polynomials for the measure determined by (35). This leads to an efficient multidimensional quadrature rule using the Gaussian weights and abscissas.

The recurrence coefficients for orthogonal polynomials can be expressed explicitly as

$$\alpha_i = \frac{\int_\Gamma \xi\psi_i(\xi)^2\rho(\xi)\,d\xi}{\int_\Gamma \psi_i(\xi)^2\rho(\xi)\,d\xi},\ \beta_i = \frac{\int_\Gamma \psi_i(\xi)^2\rho(\xi)\,d\xi}{\int_\Gamma \psi_{i-1}(\xi)^2\rho(\xi)\,d\xi}. \tag{36}$$

In the case of Hermite polynomials there exist closed forms for the recurrence coefficients. No such closed form is known in general for the Rys polynomials so a numerical method must be employed. The generation of orthogonal polynomials by numerical methods is discussed extensively in [20] and the use of generalized polynomial chaos bases in the stochastic Galerkin method is discussed in [7]. We compute the coefficients $\{\alpha_i\}$ and $\{\beta_i\}$ via the discretized Steltjies procedure [21] where integrals in (36) are approximated by quadrature.

Testing for both the sparse grid collocation method and the stochastic Galerkin method was performed using the truncated Gaussian PDF and Rys polynomials for several values of σ. The linear solver in all cases was stopped when $(||r_k||_2)/(||b||_2) < 10^{-12}$, where $r_k = b - Ax_k$ is the linear residual and A and b are the coefficient matrix and right-hand side, respectively. We constructed the sparse grids using the *Dakota* software package [22].

Table 2 reports $||\langle e_p\rangle||_{l_\infty}$, the discrete l_∞-norm of the mean error $\langle e_p\rangle$ evaluated on the grid points. For problems in one random variable, the stochastic collocation and stochastic Galerkin methods produce identical results. Table 3 shows the average number of iterations required by each deterministic subproblem as a function of grid level and σ.

TABLE 2: Mean error in the discrete l_∞ norm for the stochastic collocation and stochastic Galerkin methods.

Level/p	σ				
	1	2	3	4	5
1	0.1856	0.1971	0.2175	0.2466	0.2807
2	0.0737	0.0811	0.0932	0.1095	0.1207
3	0.0245	0.0279	0.0331	0.0389	0.1195
4	0.0070	0.0082	0.0099	0.0121	DNC
5	0.0017	0.0021	0.0026	0.0029	DNC
6	3.7199e-4	4.6301e-4	5.7900e-4	6.7702e-4	DNC
7	7.2002e-5	9.1970e-5	1.1605e-4	4.1598e-4	DNC

TABLE 3: Iterations for the stochastic collocation (left) and stochastic Galerkin methods (right).

Level	σ					p	σ				
	1	2	3	4	5		1	2	3	4	5
1	10	10	10.5	11	11	1	13	15	16	18	21
2	10	10.33	10.67	11.33	12.67	2	13	17	22	28	38
3	10	10.5	11	12.25	22	3	14	19	26	39	140
4	10	10.6	11.2	13	DNC	4	14	20	29	53	DNC
5	10.17	10.5	11.33	13.83	DNC	5	14	21	31	69	DNC
6	10.14	10.43	11.43	15	DNC	6	15	21	33	94	DNC
7	10.13	10.63	11.38	16.75	DNC	7	15	21	34	136	DNC

Problems to the right of the double line do not satisfy (34), and some of the associated systems will be indefinite for a high enough grid level, as some of the collocation points will be placed in the region of ill-posedness. If the solver failed to converge for any of the individual subproblems, the method is reported as having failed using "DNC".

Table 3 shows the PCG iteration counts for both methods. Again, problems to the right of the double line are ill-posed, and the Galerkin linear system as well as a subset of the individual collocation systems are guaranteed to become indefinite as the degree of polynomial approximation p (for stochastic Galerkin) or sparse grid level (for collocation) increases [14]. Table 3 shows that the iteration counts are fairly well behaved when mean-based preconditioning is used. In general, iterations grow as the degree of polynomial approximation increases.

It is well known that bounds on convergence of the conjugate gradient method are determined by the condition number of the matrix. It is shown in [14] that if the diffusion coefficient is given by a stationary field, as in (32), then the eigenvalues of the preconditioned stochastic Galerkin system lie in the interval $[1-\tau, 1+\tau]$, where

$$\tau = C_{p+1}^{\max} \frac{\sigma}{\mu} \left(\sum_{k=1}^{M} \sqrt{\lambda_k} ||a_k(\mathbf{x})||_{L_\infty} \right), \tag{37}$$

and $C_{p+1}^{\max}$ is the magnitude of the largest zero of the degree $p+1$ orthogonal polynomial. Therefore the condition number is bounded by $\kappa(A) \leq [1+\tau]/[1-\tau]$. It is possible to bound the eigenvalues of a single system arising in collocation in a similar manner using the relation (27). The eigenvalues of the system arising from sampling (27) at $\boldsymbol{\xi}$ lie in the bounded interval $[1-\tilde{\tau}(\boldsymbol{\xi}), 1+\tilde{\tau}(\boldsymbol{\xi})]$ where

$$\tilde{\tau}(\boldsymbol{\xi}) = \frac{\sigma}{\mu} \left(\sum_{k=1}^{M} \sqrt{\lambda_k} ||a_l(\mathbf{x})||_{L_\infty} |\xi_k| \right). \tag{38}$$

Likewise, the condition number for a given preconditioned collocation system can be bounded by $\kappa[A(\boldsymbol{\xi})] \leq (1+\tilde{\tau})/(1-\tilde{\tau})$. For both methods, as σ increases relative to μ the associated systems may become ill-conditioned and will eventually become indefinite. Likewise, as p or the sparse grid level increases, $C_{p+1}^{\max}$ and $\max_{\Theta_{p,M}} |\boldsymbol{\xi}|$ increase and the problems may again become indefinite. However, if Γ is bounded then both $C_{p+1}^{\max}$ and $\max_{\Theta_{p,M}} |\boldsymbol{\xi}|$ are bounded for all choices of p and the sparse grid level and the systems are guaranteed to remain positive definite, provided σ is not too large.

The effect of these bounds can be seen in the above examples since as σ increases the iteration counts for both methods increase until finally for large choices of σ and large p or grid level the PCG iteration fails to converge. However, for smaller values of σ the PCG iteration converges in a reasonable number of iterations for all tested values of p and grid level.

5.2 Computational Cost Comparison

In this section we compare the performance of the two methods using both the model developed above and the implementations in *Trilinos*. For our numerical examples, we consider a problem where only the covariance of the diffusion field is given. We consider two problems of the form

$$-\nabla \cdot \left\{ \left[\mu + \sigma \sum_{k=1}^{M} \sqrt{\lambda_k} \xi_k f_k(x) \right] \nabla u \right\} = 1, \tag{39}$$

where values of M between 3 and 15 are explored and $\{\lambda_k, f_k\}$ are the eigenpairs associated with the covariance kernel

$$C(\mathbf{x}_1, \mathbf{x}_2) = \exp(-|x_1 - x_2| - |y_1 - y_2|). \tag{40}$$

The KL expansion of this kernel is investigated extensively in [4]. For the first problem, the random variables $\{\xi_k\}$ are chosen to be identically independently distributed uniform random variables on $[-1, 1]$. For the second problem, the

random variables $\{\xi_k\}$ are chosen to be identically independently distributed truncated Gaussian random variables as in the previous section. For the first, problem $\mu = 0.2$ and $\sigma = 0.1$. For the second problem, $\mu = 1$ and $\sigma = 0.25$. These parameters were chosen to ensure that the problem remains well posed. Table 4 shows approximate values for τ for both of the above problems. In the second case, where truncated Gaussian random variables are used, $1 - \tau$ becomes close to zero as the stochastic dimension of the problem increases. Thus this problem could be said to be nearly ill-posed. In terms of computational effort this should favor the sparse grid collocation method, since, as was seen in the previous section, iteration counts for the stochastic Galerkin method increased faster than those for the collocation method as the problem approaches ill-posedness. The spatial domain is discretized by a uniform mesh with discretization parameter $h = 1/32$. Note that the mean-based preconditioning eliminates the dependence on h of the conditioning of the problem [14], so we consider just a single value of the spatial mesh parameter.

Approximate solutions are used to measure the error since there is no analytic expression for the exact solution to either of the above problems. To measure the error for the Galerkin method the exact solution is approximated by a high order ($p = 10$) Galerkin scheme. For the collocation method we take the solution from a level-10 sparse grid approximation as an approximation to the exact solution. These two differed by an amount on the order of the machine precision. The error in the stochastic space is then estimated by computing the mean and variance of the approximate solutions and comparing it to the mean and variance of the order-10 (level-10) approximations. The linear solves for both methods stop when $(||r_k||_2)/(||b||_2) < 10^{-12}$. In measuring the time, setup costs are ignored. The times reported are nondimensionalized by the time required to perform a single deterministic matrix vector product and compared with the model developed above.

Figure 1 explores the accuracy obtained for the two discretizations for $M = 4$; the behavior was the same for $M = 3$ and $M = 5$. In particular, it can be seen that for both sample problems, the same value of p (corresponding to the polynomial space for the Galerkin method and the sparse grid level for the collocation method) in the two methods

TABLE 4: Approximate values of τ for model problems.

M	Uniform random variables $\Gamma_i = [-1, 1]$, $\sigma = 0.1$, $\mu = 0.2$	Truncated Gaussian random variables $\Gamma_i = [-2.576, 2.576]$, $\sigma = 0.25$, $\mu = 1$
3	0.533	0.686
4	0.549	0.708
5	0.566	0.729

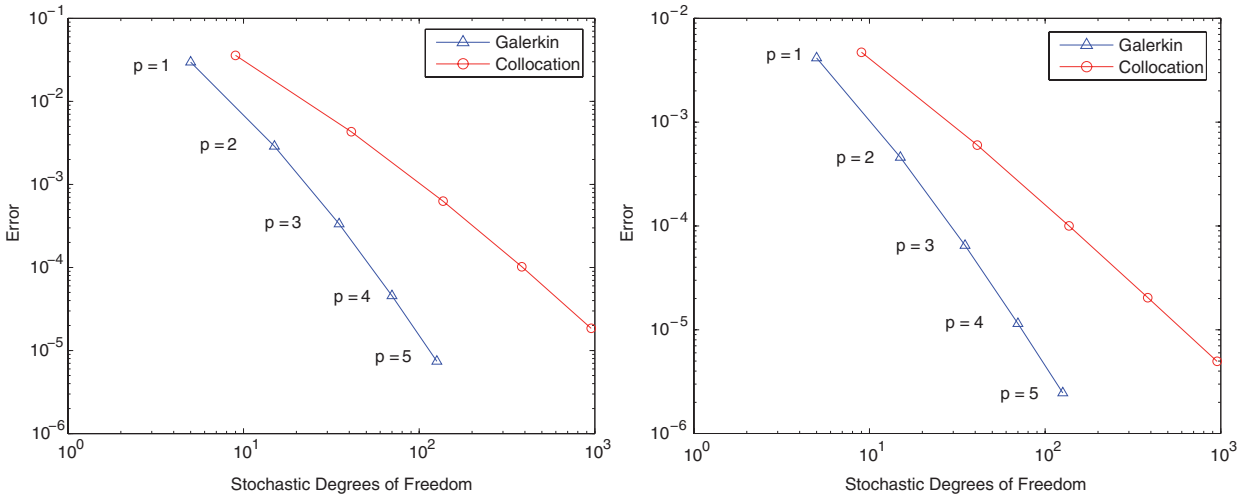

FIG. 1: Errors vs stochastic DOF for $M = 4$. Uniform random variables (left), and truncated Gaussian random variables (right).

produce solutions of comparable accuracy. Thus, the Galerkin method gives higher accuracy per stochastic degree of freedom. Since the unknowns in the Galerkin scheme are coupled, the cost per degree of freedom will be higher. In terms of computational effort, the question is whether or not the additional accuracy per degree of freedom will be worth the additional cost.

Figures 2 and 3 compare the costs incurred by the two methods, measured in CPU time, for obtaining solutions of comparable accuracy. The timings reflect time spent to execute the methods on an Intel Core 2 Duo machine running at 3.66 GHz with 6 Gb of RAM. In the figures these timings are nondimensionalized by dividing by the cost of a single sparse matrix–vector product with the (five-diagonal) nonzero structure of $\{A_k\}$. This cost is measured by dividing the total time used by the collocation method for matrix–vector products by the total number of CG iterations performed in the collocation method. This allows the times to be compared to the cost model (28) and (29), which in turn helps ensure that the implementations are of comparable efficiency. The model is somewhat less accurate for the collocation method, because for these relatively low-dimensional models the approximation $|\Theta_{p,M}| = 2^p N_\xi$ is an overestimate. For the values of M used for these results ($M = 3, 4$, and 5), it can be seen that the Galerkin method requires less CPU time than the collocation method to compute solutions of comparable accuracy, and that the gap widens as the

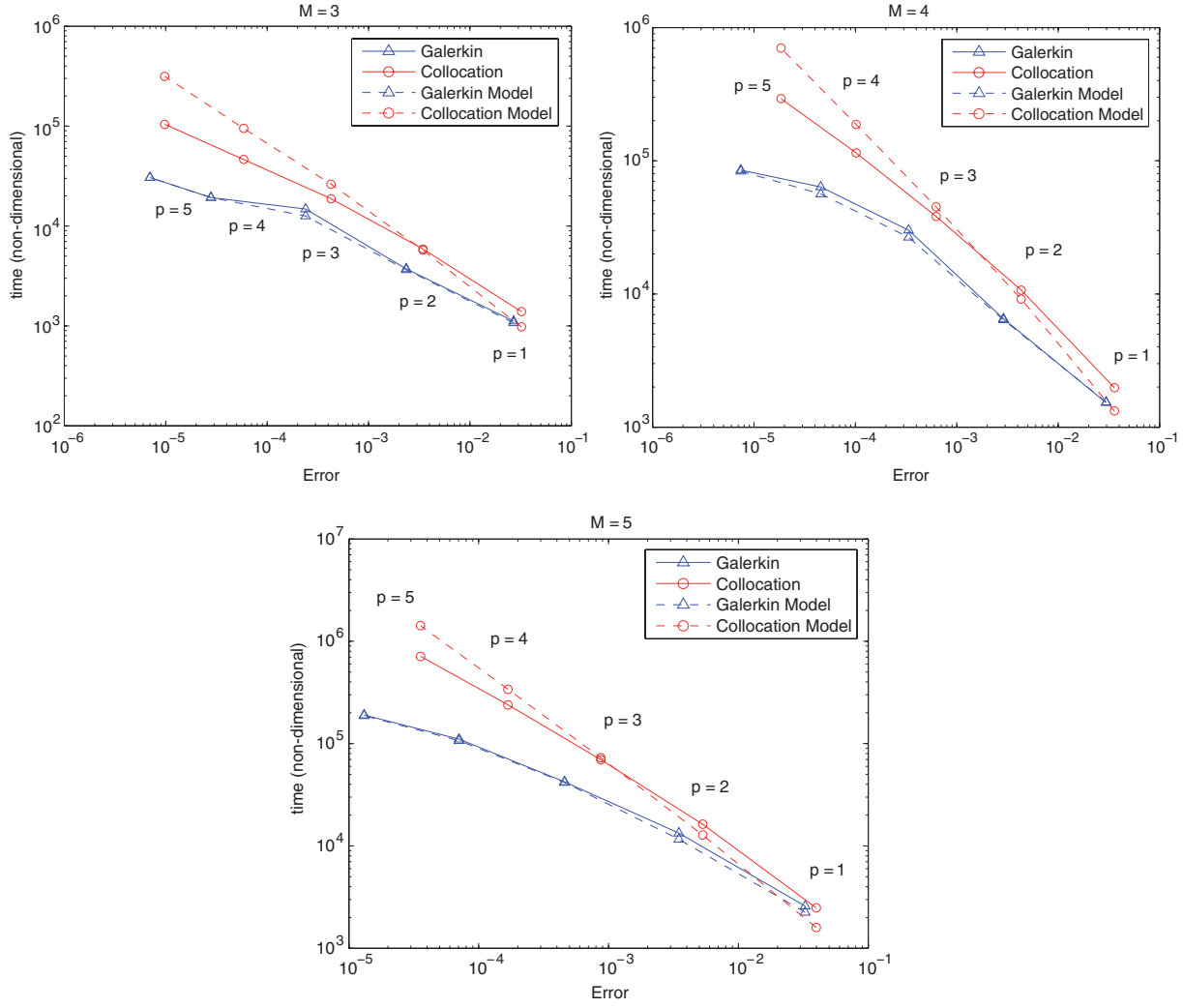

FIG. 2: Solution time vs error for $M = 3, 4, 5$. Uniform random variables.

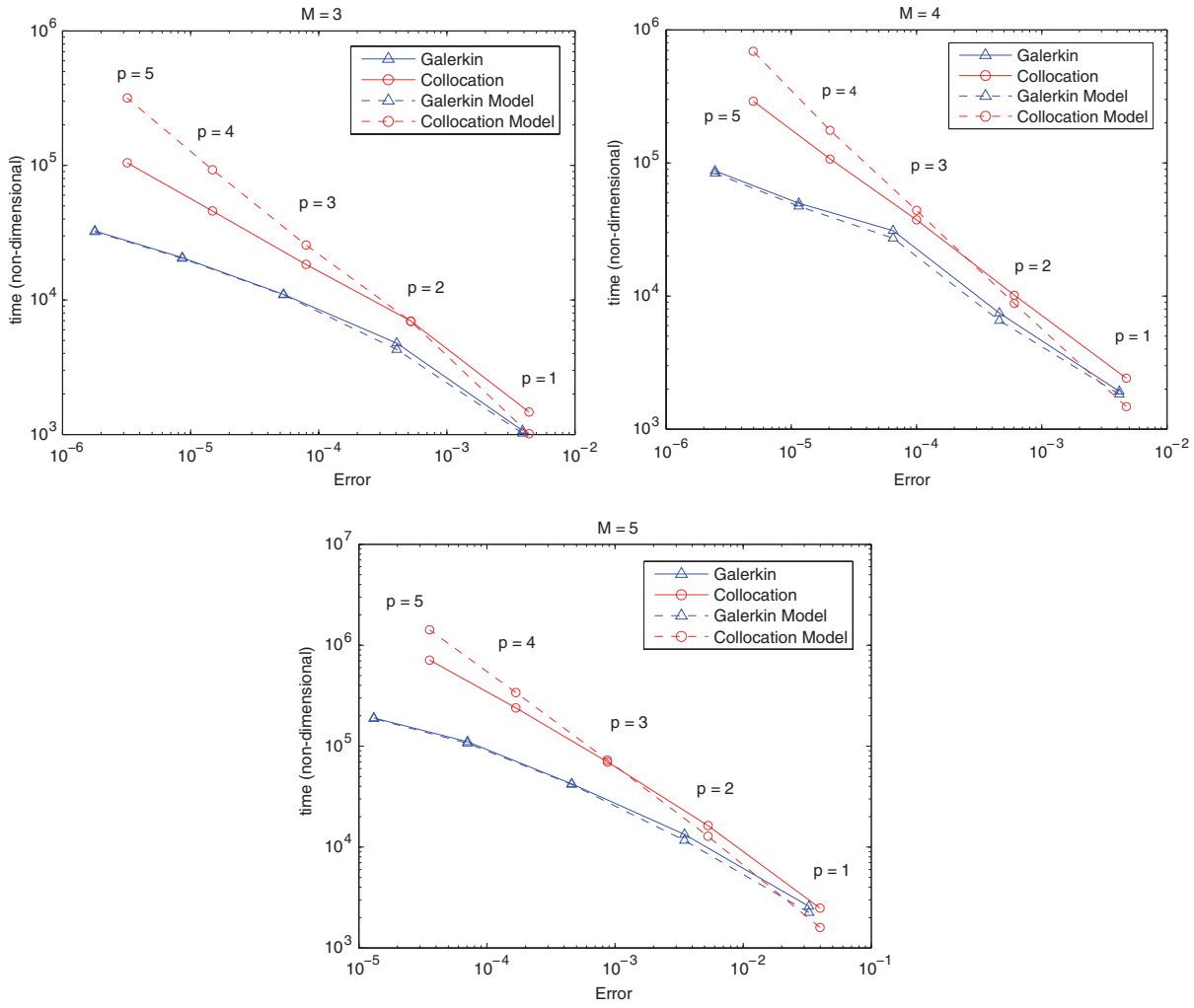

FIG. 3: Solution time vs error for $M = 3, 4, 5$. Truncated Gaussian random variables.

dimension of the space of random variables increases. Also, it is seen in Figs. 2 and 3 that the performance of each method is largely independent of the density functions used in defining the random variables ξ_k.

Table 5 expands on these results for larger values of M, based on our expectation that the same value of p (again, corresponding to the polynomial space for the Galerkin method or the level for the collocation method) yields solutions of comparable accuracy. The trends are comparable for all M and show that as the size of the approximation space increases, the overhead for collocation associated with the increased number of degrees of freedom becomes more significant.

6. CONCLUSION

In this study we have examined the costs of solving the linear systems of equations arising when either the stochastic Galerkin method or the stochastic collocation method is used to discretize the diffusion equation in which the diffusion coefficient is a random field modeled by (2). The results indicate that when mean-based preconditioners are coupled with the conjugate gradient method to solve the systems that arise, the stochastic Galerkin method is quite competitive with collocation. Indeed, the costs of the Galerkin method are typically lower than for collocation, and this differential

TABLE 5: Solution (preconditioning) time in seconds for second model problem.

	Stochastic Galerkin			Sparse grid collocation		
	$M = 5$	$M = 10$	$M = 15$	$M = 5$	$M = 10$	$M = 15$
Level/$p = 1$	0.058139 (0.026912)	0.147306 (0.051521)	0.320443 (0.085775)	0.068934 (0.036288)	0.163258 (0.078107)	0.285779 (0.123893)
2	0.269301 (0.119066)	1.20465 (0.0385744)	3.80461 (1.04111)	0.532407 (0.275829)	2.13126 (0.98289)	5.07825 (2.1247)
3	1.20353 (0.372013)	13.1382 (2.57246)	51.448 (7.40171)	2.41468 (1.20969)	16.9871 (7.54744)	57.9837 (23.1414)
4	3.50061 (1.1846)	53.786 (10.1633)	168.112 (41.325)	8.31068 (4.14521)	102.595 (44.0484)	493.042 (193.199)
5	6.510255 (2.89493)	117.729 (36.2012)		24.5645 (12.0362)	515.751 (221.546)	

becomes more pronounced as the number of terms in the truncated KL expansion increases. We have also developed a cost model for both methods that closely mirrors the complexity of the algorithms.

ACKNOWLEDGMENTS

Howard C. Elman was supported by the U.S. Department of Energy under grant DEFG0204ER25619, and by the U.S. National Science Foundation under grant CCF0726017. Christopher W. Miller was supported by the U.S. Department of Energy under grant DEFG0204ER25619. Eric T. Phipps was supported in part by the U.S. Department of Energy National Nuclear Securety Administration through its Advanced Simulation and Computing Program. Raymond S. Tuminaro was supported by the U.S. Department of Energy Office of Science ASCR Applied Math Research program. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy under contract DE-AC04-94AL85000.

REFERENCES

1. Loève, M., *Probability Theory*, Springer-Verlag, New York, 1978.
2. Weiner, N., The homogeneous chaos, *Am. J. Math.*, 60(4):897–936, 1938.
3. Babuška, I., Tempone, R., and Zouraris, G., Galerkin finite element approximations of stochastic elliptic partial differential equations, *SIAM J. Numer. Anal.*, 42:800–825, 2004.
4. Ghanem, R. and Spanos, P., *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York, 1991.
5. Sarkar, A. and Ghanem, R., Mid-frequency structural dynamics with parameter uncertainty, *Comput. Methods Appl. Mech. Eng.*, 191:5499–5513, 2002.
6. Xiu, D. and Shen, J., Efficient stochastic Galerkin methods for random diffusion equations, *J. Comput. Phys.*, 228:266–281, 2009.
7. Xiu, D. and Karniadakis, G., Modeling uncertainity of elliptic partial differential equations via generalized polynomial chaos, *Comput. Methods Appl. Mech. Eng.*, 191:4928–4948, 2002.
8. Babuška, I., Nobile, F., and Tempone, R., A stochastic collocation method for elliptic partial differential equations with random input data, *SIAM J. Numer. Anal.*, 45:1005–1034, 2007.
9. Nobile, F., Tempone, R., and Webster, C. G., A sparse grid stochastic collocation method for partial differential equations with random input data, *SIAM J. Numer. Anal.*, 46(5):2309–2345, 2008.
10. Xiu, D. and Hesthaven, J., High-order collocation methods for differential equations with random inputs, *SIAM J. Sci. Comput.*, 27(3):1118–1139, 2005.
11. Bäck, J., Nobile, F., Tamellini, L., and Tempone, R., Stochastic Galerkin and Collocation Methods for PDEs with Random Coefficients: A Numerical Comparison, Tech. Rep. 09-33, Institute for Computational Engineering and Sciences, Univer-

sity of Texas at Austin, 2009. To appear in *Proceedings of ICOSAHOM'09*, Lecture Notes in Computational Science and Engineering, Springer-Verlag, New York, 2009.

12. Heroux, M., Bartlett, R., Hoekstra, V. H. R., Hu, J., Kolda, T., Lehoucq, R., Long, K., Pawlowski, R., Phipps, E., Salinger, A., Thornquist, H., Tuminaro, R., Willenbring, J., and Williams, A., An Overview of Trilinos, Tech. Rep. SAND2003-2927, Sandia National Laboratories, 2003.
13. Ghanem, R. and Pellissetti, M., Iterative solution of systems of linear equations arising in the context of stochastic finite elements, *Adv. Eng. Software*, 31(8):607–616, 2000.
14. Powell, C. and Elman, H., Block-diagonal preconditioning for spectral stochastic finite element systems, *IMA J. Numer. Anal.*, 29:350–375, 2009.
15. Gee, M., Siefert, C., Hu, J., Tuminaro, R., and Sala, M., ML 5.0 Smoothed Aggregation User's Guide, Tech. Rep. SAND2006-2649, Sandia National Laboratories, 2006.
16. Gordon, A. and Powell, C., Solving Stochastic Collocation Systems with an Algebraic Multigrid, MIMS EPrint 2010.19, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, February 2010.
17. Nobile, F., Tempone, R., and Webster, C. G., An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data, *SIAM J. Numer. Anal.*, 46(5):2411–2442, 2008.
18. Ma, X. and Zabaras, N., An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations, *J. Comput. Phys.*, 228:3084–3113, 2009.
19. Ma, X. and Zabras, N., High-dimensional stochastic model representation technique for the solution of stochastic PDEs, *J. Comput. Phys.*, 229(10):3884–3915, 2010.
20. Gautschi, W., *Orthogonal Polynomials: Computation and Approximation*, Oxford University Press, Oxford, 2004.
21. Sagar, R. and Smith, V., On the calculation of Rys polynomials and quadratures, *Int. J. Quant. Chem.*, 43:827–836, 1992.
22. Eldred, M., Giunta, A., van Bloemen Waanders, B., Wojtkiewicz, S., Hart, W., and Alleva, M., Dakota, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis, Version 4.0 user's manual, Tech. Rep. SAND2006-6337, Sandia National Laboratories, October 2006.

International Journal for Uncertainty Quantification, 1(1):35–47, 2011

PROBABILISTIC PREDICTIONS OF INFILTRATION INTO HETEROGENEOUS MEDIA WITH UNCERTAIN HYDRAULIC PARAMETERS

Peng Wang & Daniel M. Tartakovsky[*]

Department of Mechanical and Aerospace Engineering, University of California, San Diego, 9500 Gilman Drive, Mail Code 0411, La Jolla, CA 92093-0411, USA

Original Manuscript Submitted: 06/04/2010; Final Draft Received: 16/07/2010

Soil heterogeneity and the lack of detailed site characterization are two ubiquitous factors that render predictions of flow and transport in the vadose zone inherently uncertain. We employ the Green–Ampt model of infiltration and the Dagan–Bresler statistical parameterization of soil properties to compute probability density functions (PDFs) of infiltration rate and infiltration depth. By going beyond uncertainty quantification approaches based on mean and variance of system states, these PDF solutions enable one to evaluate probabilities of rare events that are required for probabilistic risk assessment. We investigate the temporal evolution of the PDFs of infiltration depth and corresponding infiltration rate, the relative importance of uncertainty in various hydraulic parameters and their cross-correlation, and the impact of the choice of a functional form of the hydraulic function.

KEY WORDS: *Uncertainty quantification, stochastic, infiltration rate, Green–Ampt model*

1. INTRODUCTION

Soil heterogeneity and the lack of detailed site characterization are two ubiquitous factors that hamper one's ability to predict flow and transport in the vadose zone. The continuing progress in data acquisition notwithstanding, measurements of hydraulic properties of partially saturated media remain scarce and prone to measurement and interpretive errors. Consequently, spatial distributions of hydraulic parameters (saturated and relative hydraulic conductivities, and parameters in retention curves) are typically uncertain and their statistical properties are subject to considerable debate.

Despite some reservations, e.g., [1, 2], it has become common to treat saturated hydraulic conductivity $K_s(\mathbf{x})$ as a multivariate log-normal random field whose ensemble statistics (e.g., mean, variance, and correlation length) can be inferred from spatially distributed data by means of geostatistics. No such consensus exists about statistical distributions of various hydraulic parameters entering relative hydraulic conductivity and retention curves. For example, various data analyses concluded that spatial variability of a soil parameter $\alpha_G(\mathbf{x})$ in the Gardner model of relative conductivity, which is often referred to as the reciprocal of the macroscopic capillary length, exhibits either a normal [3] or log-normal [4] distribution and is either correlated [5] or uncorrelated [3] with K_s. We defer a more detailed review of the statistical properties of both $\alpha_G(\mathbf{x})$ and parameters in the van Genuchten model of relative conductivity until Section 2. Here, it suffices to say that any approach to uncertainty quantification for flow and transport in the vadose zone must be flexible enough to accommodate arbitrary statistical distributions of soil properties.

Statistical treatment of hydraulic parameters renders corresponding flow equation stochastic. Solutions of these equations are probability density functions (PDFs) of system states (water content, pressure, and macroscopic flow velocity) and can be used not only to predict flow in heterogeneous partially saturated porous media but also to quantify

[*]Correspond to Daniel M. Tartakovsky, E-mail: dmt@ucsd.edu

2152 -5080/11/$35.00

predictive uncertainty. Rather than computing PDFs of system states, standard practice in subsurface hydrology is to compute (analytically or numerically) the first two moments of system states, and to use their ensemble means as predictors of a system's behavior and variances (or standard deviations) as a measure of predictive uncertainty. A large body of literature employing this approach to solve the stochastic Richards equation includes [6–11], to name just a few. With the exception of solutions based on the Kirchhoff transformation [12–14], such analyses require one to linearize constitutive relations in the Richards equation, introducing errors that are hard to quantify a priori. More important, none of these solutions can be used to estimate the probability of rare events, which is of crucial importance for uncertainty quantification and risk assessment [15].

The Green–Ampt model described in some detail in Section 2 (see also [16, Section 5.2]) provides an alternative description of flow in partially saturated porous media. The relative simplicity of the Green–Ampt formulation makes it easier to solve than the Richards equation, which explains its prevalence in large numerical codes—e.g., SCS developed by U.S. Environmental Protection Agency (USEPA), DR3M developed by U.S. Geological Survey (USGS), and HIRO2 developed by U.S. Department of Agriculture (USDA)—that are routinely used to predict overland and channel flows. The first analysis of the impact of soil heterogeneity and parametric uncertainty on solutions of the Green–Ampt equations was carried out by Dagan and Bresler [17]. Saturated hydraulic conductivity—the sole source of uncertainty in their analysis—was treated as a two-dimensional *random field*, $K_s(x_1, x_2)$. This enables one to model vertical infiltration with a collection of one-dimensional (in the x_3 direction) solutions each of which corresponds to a different *random variable* K_s. The Dagan–Bresler statistical model [17], whose precise formulation is provided in Section 2, was found to yield accurate predictions of infiltration into heterogeneous soils [18, 19] and has been adopted in a number of subsequent investigations, e.g., [19–24]. These and other similar analyses aimed to derive effective (ensemble averaged) infiltration equations, and some of them quantified predictive uncertainty by computing variances of system states.

Driven by the needs of probabilistic risk assessment, we focus on the derivation of single-point PDFs (rather than the first two moments) of system states describing infiltration into heterogeneous soils with uncertain hydraulic parameters. Our analysis employs the Green–Ampt model of infiltration with the Dagan–Bresler parameterization, both of which are formulated in Section 2. This Section also contains an overview of experimentally observed statistical properties of the coefficients entering the Gardner and van Genuchten expressions of relative hydraulic conductivity K_r. A general framework for derivation of PDF solutions of the Green–Ampt model is presented in Section 3. In Section 4 we investigate the temporal evolution of the PDFs of a wetting front (Setion 4.1) and corresponding infiltration rate (Section 4.2), the relative importance of uncertainty in various hydraulic parameters (Section 4.3) and their cross-correlation (Section 4.4), and the impact of the choice of a functional form of K_r (Section 4.5). Concluding remarks are presented in Section 5.

2. PROBLEM FORMULATION

Consider infiltration into a heterogeneous soil with saturated hydraulic conductivity K_s, porosity ϕ, residual water content θ_r, and relative hydraulic conductivity $K_r(\psi; \boldsymbol{\alpha})$ that varies with pressure head ψ in accordance with a constitutive model and model parameters $\boldsymbol{\alpha}$. While the subsequent analysis can be applied to any constitutive relation, we focus on the Gardner model [16, Table 2.1]

$$K_r = \mathrm{e}^{\alpha_G \psi} \tag{1}$$

and the van Genuchten model (ibid)

$$K_r = \frac{[1 - \psi_d^{mn}(1 + \psi_d^n)^{-m}]^2}{(1 + \psi_d^n)^{m/2}}, \quad \psi_d \equiv \alpha_{vG}|\psi|. \tag{2}$$

The model parameters $\boldsymbol{\alpha}$ ($\boldsymbol{\alpha} \equiv \alpha_G$ and $\{\alpha_{vG}, n, m = 1 - 1/n\}$ for the Gardner and van Genuchten models, respectively) and the rest of the hydraulic properties mentioned above vary in space and are sparsely sampled. To quantify uncertainty about values of these properties at points $\mathbf{x} = (x_1, x_2, x_3)^T$ where measurements are unavailable, we treat them as random fields. Thus, a soil parameter $\mathcal{A}(\mathbf{x}, \omega)$ varies not only in the physical domain, $\mathbf{x} \in \mathcal{D}$, but also

in the probability space $\omega \in \Omega$. A probability density function $p_{\mathcal{A}}$, which describes the latter variability, is inferred from measurements of $\mathcal{A}$ by invoking ergodicity. Experimental evidence for the selection of PDFs $p_{\mathcal{A}}$ for various soil parameters $\mathcal{A}$ is reviewed in Section 2.1, and the Dagan–Bresler statistical model used in our analysis is formulated in Section 2.2.

The overreaching aim of the present analysis is to quantify the impact of this parametric uncertainty on predictions of both the dynamics of wetting fronts and infiltration rates. Uncertainty in the former may significantly affect the accuracy and reliability of field-scale measurements of soil saturation [25], while uncertainty in the latter is of fundamental importance to flood forecasting [23].

2.1 Statistics of Soil Parameters

Saturated hydraulic conductivity. In addition to the experimental studies reviewed in [12], the data analyses reported in [4, 24], etc., support our treatment of saturated hydraulic conductivity K_s as a log-normal random field.

Gardner's constitutive parameter. The (scarce) experimental evidence reviewed in [12] suggests that α_G, the reciprocal of the macroscopic capillary length, can be treated alternatively either as a Gaussian (normal) or as a log-normal random field. While the approach described below is capable of handling both distributions, in the subsequent computational examples we will treat α_G as a log-normal field, which is a model adopted in more recent computational investigations (e.g., [4, 10]).

Van Genuchten's constitutive parameters. The van Genuchten hydraulic function (2) is a two-parameter model obtained from its more general form by setting $m = 1 - 1/n$ and $l = 1/2$ (hence, the power $m/2$ in the denominator). We employ this form because of its widespread use [16, Table 2.1], but the approach described below can be readily applied to quantify uncertainty in more general formulations with arbitrary m and l. The experimental evidence presented in [4, 26, 27] shows that the coefficient of variation of α_{vG} is much larger than that of n. These data suggest that α_{vG} can be treated as a log-normal field and the shape factor n as a deterministic constant.

Correlations between hydraulic parameters. Experimental evidence presented in [4, 12] suggests that the coefficient of variation (CV) of K_s is generally much larger than that of either α_G or α_{vG}. These parameters were found to be either perfectly correlated or uncorrelated or anticorrelated (see also [28]). Our analysis allows for an arbitrary degree of correlation between K_s and either α_G or α_{vG}.

Finally, since the difference between the full and residual saturations $\Delta\theta = \phi - \theta_r$ typically exhibits lower spatial variability than both K_s and α_G (or α_{vG}), we treat it as a deterministic constant to simplify the presentation. Our approach can be adopted to quantify uncertainty in $\Delta\theta$ and the shape factor n in the van Genuchten hydraulic function, as discussed in Section 3.

2.2 Statistical Model for Soil Parameters

Following [17], we restrict our analysis to infiltration depths that do not exceed vertical correlation lengths l_v of (random) soil parameters $\mathcal{A}(\mathbf{x}, \omega)$. Then $\mathcal{A} = \mathcal{A}(x_1, x_2, \omega)$, so that a heterogeneous soil can be represented by a collection of one-dimensional (in the vertical direction x_3) homogeneous columns of length L_3, whose uncertain hydraulic properties are modeled as random variables (rather than random fields). The restriction $l_v > L_3$ formally renders the Dagan–Bresler parameterization [17] suitable for heterogeneous topsoils, and thus can be used to model surface response to rainfall events [23, 24] and transport phenomena in topsoil [21]. Yet it was also used to derive effective properties of the whole vadose zone [4, 28]. Rubin and Or [19] provide an additional justification for the Dagan–Bresler parameterization by noting that "the determination of soil hydraulic properties through field methods. . .homogenize the properties vertically, thus eliminating the variability in the vertical direction in a practical sense."

Consider a three-dimensional flow domain $\Omega = \Omega_h \times [0, L_3]$, where Ω_h represents its horizontal extent. A discretization of Ω_h into N elements represents Ω by an assemblage of N columns of length L_3 and facilitates the complete description of a random field $\mathcal{A}(x_1, x_2, \omega)$—in the analysis below, $\mathcal{A}$ stands for K_s, α_G, and α_{vG} but can also include other hydraulic properties and the ponding pressure head ψ_0 at the soil surface $x_3 = 0$—with a joint

probability function $p_{\boldsymbol{\mathcal{A}}}(A_1, \ldots, A_N)$. Probability density functions (PDFs) of hydraulic properties of the ith column are defined as marginal distributions,

$$p_{\mathcal{A}_i}(A_i) = \int p_{\boldsymbol{\mathcal{A}}}(A_1, \ldots, A_n) \mathrm{d}A_1 \ldots \mathrm{d}A_{i-1} \mathrm{d}A_{i+1} \ldots \mathrm{d}A_N. \tag{3}$$

Since statistical properties of soil parameters $\mathcal{A}$ are inferred from spatially distributed data by invoking ergodicity, the corresponding random fields (or their fluctuations obtained by data de-trending) must be stationary so that

$$p_{\mathcal{A}_i} = p_{\mathcal{A}} \quad \text{for} \quad i = 1, \ldots, N. \tag{4}$$

Furthermore, if such soil parameters (e.g., K_s and α_G) are correlated, their statistical description requires the knowledge of a joint distribution. For multivariate Gaussian $Y_1 = \ln K_s$ and $Y_2 = \ln \alpha_G$ (or $Y_2 = \ln \alpha_{vG}$), their joint PDF is given by

$$p_{Y_1,Y_2}(y_1, y_2) = \frac{1}{2\pi\sigma_{Y_1}\sigma_{Y_2}\sqrt{1-\rho^2}} \exp\left[-\frac{R}{2(1-\rho^2)}\right] \tag{5a}$$

where

$$R = \frac{(y_1 - \overline{Y}_1)^2}{\sigma_{Y_1}^2} - 2\rho\frac{y_1 - \overline{Y}_1}{\sigma_{Y_1}}\frac{y_2 - \overline{Y}_2}{\sigma_{Y_2}} + \frac{(y_2 - \overline{Y}_2)^2}{\sigma_{Y_2}^2}; \tag{5b}$$

$\overline{Y}_i$ and σ_{Y_i} denote the mean and standard deviation of Y_i ($i = 1, 2$), respectively; and $-1 \le \rho \le 1$ is the linear correlation coefficient between Y_1 and Y_2. The lack of correlation between Y_1 and Y_2 corresponds to setting $\rho = 0$ in (5).

2.3 Green–Ampt Model of Infiltration

During infiltration into topsoils, the Dagan–Bresler parameterization of soil heterogeneity can be supplemented with an assumption of vertical flow. The rationale for, and implications of, neglecting the horizontal component of flow velocity can be found in [17, 19, 20] and other studies reviewed in the Introduction.

This assumption obviates the need to solve a three-dimensional flow problem, replacing the latter with a collection of N one-dimensional flow problems to be solved in homogeneous soil columns with random but constant hydraulic parameters. Such a framework was used to predict mean (ensemble averaged) flow with either the Green–Ampt model [17, 20] or the steady-state Richards equation with the Gardner hydraulic function [19]. We employ the Green–Ampt description because it enables one to handle transient flow and to employ arbitrary hydraulic functions, without resorting to linearizing approximations [29].

Let $I(t)$ denote (uncertain) cumulative infiltration due to ponding water of height ψ_0 at the soil surface $x_3 = 0$. The Green–Ampt model of infiltration approximates an S-shaped wetting front with a sharp interface $x_{\mathrm{f}}(t)$ that separates fully saturated soil (saturation ϕ) from dry soil (saturation θ_r). The latter is also known as infiltration depth. If the x_3 coordinate is positive downward, Darcy's law defines macroscopic (Darcy's) flux q as (e.g., [16, Eq. (5.1)])

$$q = -K_s \frac{\psi_{\mathrm{f}} - x_{\mathrm{f}} - \psi_0}{x_{\mathrm{f}}}. \tag{6}$$

Pressure head ψ_{f} at the infiltration depth $x_{\mathrm{f}}(t)$ is empirically set to a "capillary drive",

$$\psi_{\mathrm{f}} = -\int_{\psi_{\mathrm{in}}}^{0} K_r(\psi)\, \mathrm{d}\psi, \tag{7}$$

where ψ_{in} is the initial pressure head in the dry soil.

Mass conservation requires that $I(t) = (\omega - \theta_r)\, x_{\rm f}(t)$ and the infiltration rate $i \equiv {\rm d}I/{\rm d}t$ equals q. The first condition yields

$$i = \Delta\theta \frac{{\rm d}x_{\rm f}}{{\rm d}t}, \qquad \Delta\theta = \omega - \theta_r, \tag{8}$$

which, combined with the second condition and (6), leads to a (stochastic) ordinary differential equation for the position of the wetting front,

$$\Delta\theta \frac{{\rm d}x_{\rm f}}{{\rm d}t} = K_s \frac{\psi_0 + x_{\rm f} - \psi_{\rm f}}{x_{\rm f}}, \quad x_{\rm f}(t=0) = 0. \tag{9}$$

Our goal is to relate uncertainty in hydraulic parameters K_s and α_G (or α_{vG}) to predictive uncertainty about the infiltration depth $x_{\rm f}(t)$ and the infiltration rate $i(t)$, i.e., to express the PDFs of the latter, $p_{\rm f}(x_{\rm f}; t)$ and $p_{\rm i}(i; t)$, in terms of the PDF of the former (5).

3. PDF SOLUTIONS

To simplify the presentation, we assume that the height of ponding water, ψ_0, does not change with t during the simulation time T. Then an implicit solution of (9) takes the form

$$x_{\rm f} - (\psi_0 - \psi_{\rm f}) \ln\left(1 + \frac{x_{\rm f}}{\psi_0 - \psi_{\rm f}}\right) = \frac{K_s}{\Delta\theta} t. \tag{10}$$

For small t, (10) can be approximated by an explicit relation [16, Eq. (5.12)]

$$x_{\rm f} \approx \sqrt{\frac{2(\psi_0 - \psi_{\rm f}) K_s t}{\Delta\theta}}. \tag{11}$$

For large t, flow becomes gravity dominated, $i \sim K_s$, and [16, p. 170]

$$x_{\rm f} \approx \frac{K_s}{\Delta\theta} t. \tag{12}$$

For intermediate t, various approximations, e.g., [30] and [16, p. 170], can be used to replace the implicit solution (10) with its explicit counterparts. We will use the implicit solution (10) to avoid unnecessary approximation errors.

Several of the simplifying assumptions made above can be easily relaxed. First, since K_s and $\Delta\theta$ enter the stochastic Eq. (9) and its implicit solution (10) as the ratio $K_s^\star = K_s/\Delta\theta$, one can easily incorporate uncertainty in (randomness of) $\Delta\theta$ by replacing the PDF of K_s with the PDF of $K_s^\star$. Second, the implicit relation $F(x_{\rm f}, K_s/\Delta\theta, \boldsymbol{\alpha}; t) = 0$ given by (10) and (7) allows one to express the PDF of $x_{\rm f}$ in terms of the PDFs of *any* number of hydraulic parameters by following the procedure described below. Third, uncertainty in, and temporal variability of, the height of ponding water $\psi_0(t)$ can be dealt with by replacing (10) with an appropriate solution of (9).

3.1 PDF of Infiltration Depth

Let $G_{\rm f}(x_{\rm f}^\star) = P(x_{\rm f} \le x_{\rm f}^\star)$ denote the cumulative distribution function of $x_{\rm f}$, i.e., the probability that the random position of the wetting front $x_{\rm f}$ takes on a value not larger than $x_{\rm f}^\star$. Since (10) provides an explicit dependence of random K_s on random $x_{\rm f}$ and α (where α stands for either α_G or α_{vG}), i.e.,

$$K_s(x_{\rm f}, \alpha) = \frac{\Delta\theta}{t}\left[x_{\rm f} - (\psi_0 - \psi_{\rm f}) \ln\left(1 + \frac{x_{\rm f}}{\psi_0 - \psi_{\rm f}}\right)\right], \tag{13}$$

it follows from the definition of a cumulative distribution function that

$$G_{\rm f}(x_{\rm f}^\star) = \int_0^\infty \int_0^{K_s(x_{\rm f}^\star, \alpha)} p_{Y_1,Y_2}(K_s, \alpha) \frac{{\rm d}K_s {\rm d}\alpha}{K_s \alpha}. \tag{14}$$

The denominator in (14) reflects the transition from (5), the joint Gaussian PDF for Y_1 and Y_2, to the log-normal variables $K_s = \exp(Y_1)$ and $\alpha = \exp(Y_2)$.

The PDF of the random (uncertain) infiltration depth $p_{\rm f}(x_{\rm f}^\star; t)$ can now be obtained as

$$p_{\rm f}(x_{\rm f}^\star; t) = \frac{{\rm d}G_{\rm f}(x_{\rm f}^\star; t)}{{\rm d}x_{\rm f}^\star}. \tag{15}$$

Using Leibnitz's rule to compute the derivative of the integral in (14) and (15), we obtain

$$p_{\rm f}(x_{\rm f}^\star; t) = \int_0^\infty \frac{p_{Y_1,Y_2}[K_s(x_{\rm f}^\star, \alpha), \alpha]}{\alpha K_s(x_{\rm f}^\star, \alpha)} \frac{\partial K_s(x_{\rm f}^\star, \alpha)}{\partial x_{\rm f}^\star} {\rm d}\alpha. \tag{16}$$

Equation (16) holds for an arbitrary implicit solution of the Green–Ampt equation, $F(x_{\rm f}, K_s/\Delta\theta, \boldsymbol{\alpha}; t) = 0$, and hence, the PDF solution (16) is applicable to a large class of infiltration regimes that are amenable to the Green–Ampt description. For the flow regime considered in the present analysis, $K_s(x_{\rm f}^\star, \alpha)$ is given by (13), and (16) takes the form

$$p_{\rm f}(x_{\rm f}^\star; t) = \frac{\Delta\theta}{t} \int_0^\infty \frac{p_{Y_1,Y_2}[K_s(x_{\rm f}^\star, \alpha), \alpha]}{\alpha K_s(x_{\rm f}^\star, \alpha)} \frac{x_{\rm f}^\star \, {\rm d}\alpha}{\psi_0 - \psi_{\rm f} + x_{\rm f}^\star}. \tag{17}$$

3.2 PDF of Infiltration Rate

Let $G_{\rm i}(i^\star) = P(i \le i^\star)$ denote the cumulative distribution function of i, i.e., the probability that the random infiltration rate i takes on a value not larger than $i^\star$. Since $q = i$, Eqs. (6) and (7) define a mapping $K_s = K_s(i, \alpha)$. This enables one to compute the cumulative distribution function $G_{\rm i}(i^\star)$ as

$$G_{\rm i}(i^\star) = \int_0^\infty \int_0^{K_s(i^\star, \alpha)} p_{Y_1,Y_2}(K_s, \alpha) \frac{{\rm d}K_s {\rm d}\alpha}{K_s \alpha} \tag{18}$$

and the PDF of infiltration rate, $p_{\rm i} = {\rm d}G_{\rm i}/{\rm d}i^\star$, as

$$p_{\rm i}(i^\star; t) = \int_0^\infty \frac{p_{Y_1,Y_2}[K_s(i^\star, \alpha), \alpha]}{\alpha K_s(i^\star, \alpha)} \frac{\partial K_s(i^\star, \alpha)}{\partial i^\star} {\rm d}\alpha. \tag{19}$$

The derivative $\partial K_s/\partial i^\star$ is computed from (6) as the inverse of

$$\frac{\partial i^\star}{\partial K_s} = 1 + \frac{\psi_0 - \psi_{\rm f}}{x_{\rm f}} \left(1 - \frac{K_s t}{\Delta\theta} \frac{x_{\rm f} - \psi_{\rm f} + \psi_0}{x_{\rm f}^2}\right). \tag{20}$$

3.3 Dimensionless Form of PDFs

To facilitate an analysis of the effects of various sources of parametric uncertainty on the PDF $p_{\rm f}(x_{\rm f}^\star; t)$ of the uncertain (random) infiltration depth $x_{\rm f}(t)$, given by the analytical solution (17), we introduce the following dimensionless quantities. Let the averaged quantities $(\overline{\alpha})^{-1}$ and $\overline{K}_s$ represent a characteristic length scale and a characteristic value of saturated hydraulic conductivity, respectively. Then a characteristic time scale τ can be defined as

$$\tau = (\overline{\alpha}\overline{K}_s)^{-1}, \tag{21}$$

and the following dimensionless quantities can be introduced,

$$t' = \frac{t}{\tau}, \quad \psi' = \overline{\alpha}\psi, \quad \alpha' = \frac{\alpha}{\overline{\alpha}}, \quad K_s' = \frac{K_s}{\overline{K}_s}. \tag{22}$$

This leads to a PDF solution for the dimensionless infiltration depth $x_{\rm f}' = \overline{\alpha} x_{\rm f}$,

$$p_{\rm f}(x_{\rm f}'; t') = \frac{\Delta\theta}{t'} \int_0^\infty \frac{p_{Y_1', Y_2'}[K_s'(x_{\rm f}', \alpha'), \alpha']}{\alpha' K_s'(x_{\rm f}', \alpha')} \frac{x_{\rm f}' \, {\rm d}\alpha'}{\psi_0' - \psi_{\rm f}' + x_{\rm f}'}. \tag{23}$$

Likewise, the PDF of the dimensionless infiltration rate $i' = i/\overline{K}_s$ takes the form

$$p_{\rm i}(i'; t') = \int_0^\infty \frac{p_{Y_1', Y_2'}[K_s'(i', \alpha'), \alpha']}{\alpha' K_s'(i', \alpha')} \frac{\partial K_s'(i', \alpha')}{\partial i'} {\rm d}\alpha'. \tag{24}$$

In the following, we drop the primes to simplify the notation.

4. RESULTS AND DISCUSSION

In this Section, we explore the impact of various aspects of parametric uncertainty on the uncertainty in predictions of infiltration rate $i(t)$ and infiltration depth $x_{\rm f}(t)$. Specifically, we investigate the temporal evolution of the PDFs of the wetting front (Section 4.1) and the infiltration rate (Section 4.2), the relative importance of uncertainty in K_s and α_i (Section 4.3), and the effects of cross-correlation between them (Section 4.4). This is done for the Gardner hydraulic function (1), in which case (7) results in the interfacial pressure head $\psi_{\rm f} = -\alpha_G^{-1}$. In Section 4.5, we explore how the choice of a functional form of the hydraulic function, i.e., the use of the van Genuchten model (2) instead of the Gardner relation (1), affects the predictive uncertainty.

Unless explicitly noted otherwise, the simulations reported below correspond to the dimensionless initial pressure head $\psi_{\rm in} = -9999.9$, the dimensionless height of ponding water $\psi_0 = 0.1$, $\Delta\theta = 0.45$, the coefficients of variation $CV_{\ln K} \equiv \sigma_{Y_1}/\overline{Y}_1 = 3.0$ and $CV_{\ln \alpha} \equiv \sigma_{Y_2}/\overline{Y}_2 = 0.5$ with the means $\overline{Y}_1 = 0.25$ and $\overline{Y}_2 = 0.1$, and the cross-correlation coefficient $\rho = 0$. (The use of the soil data in Table 1 of [26] in conjunction with these dimensionless parameters would result in the height of ponding water $\psi_0 = 0.6$ cm.)

4.1 PDF of Wetting Front

Since the initial position of the wetting front is assumed to be known, $x_{\rm f}(t = 0) = 0$, the PDF $p_{\rm f}(x_{\rm f}; 0) = \delta(x_{\rm f})$, where $\delta(\cdot)$ denotes the Dirac delta function. As the dimensionless time becomes large ($t \to \infty$), $p_{\rm f} \sim p_{K_s}$ in accordance with (12). The PDF $p_{\rm f}(x_{\rm f}; t)$ in (23) describes the temporal evolution of predictive uncertainty between these two asymptotes, with Fig. 1 providing snapshots at dimensionless times $t = 0.01$, 0.1, and 0.5. (For the soil parameters reported in Table 1 of [26], this corresponds to dimensional times 1.5, 15, and 75 min, respectively). The uncertainty in predictions of infiltration depth increases rapidly, as witnessed by wider distributions with longer tails.

4.2 PDF of Infiltration Rate

Figure 2 provides snapshots, at dimensionless times $t = 0.01$, 0.1, and 0.5, of the temporal evolution of the PDF of infiltration rate $p_{\rm i}(i; t)$ given by (24). Both the mean infiltration rate and the corresponding predictive uncertainty decrease with time. At later times (the dimensionless time $t = 5.0$, for the parameters used in these simulations), the PDF appears to become time invariant. This is to be expected on theoretical grounds, see (12), according to which $p_{\rm i}(i'; t') \to p_K(K_s')$ as $t' \to \infty$. The reduced χ^2 test confirmed this asymptotic behavior at dimensionless time $t = 100.0$.

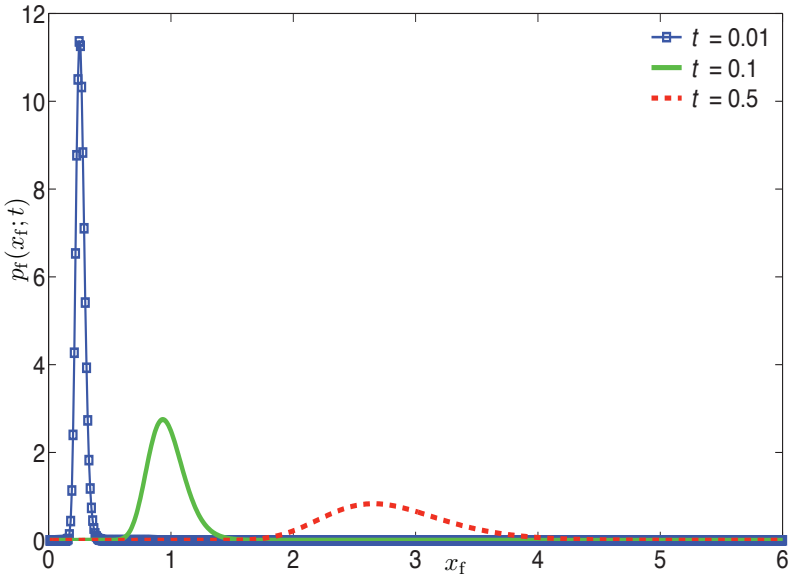

FIG. 1: Temporal evolution of the PDF of infiltration depth $p_{\mathrm{f}}(x_{\mathrm{f}};t)$.

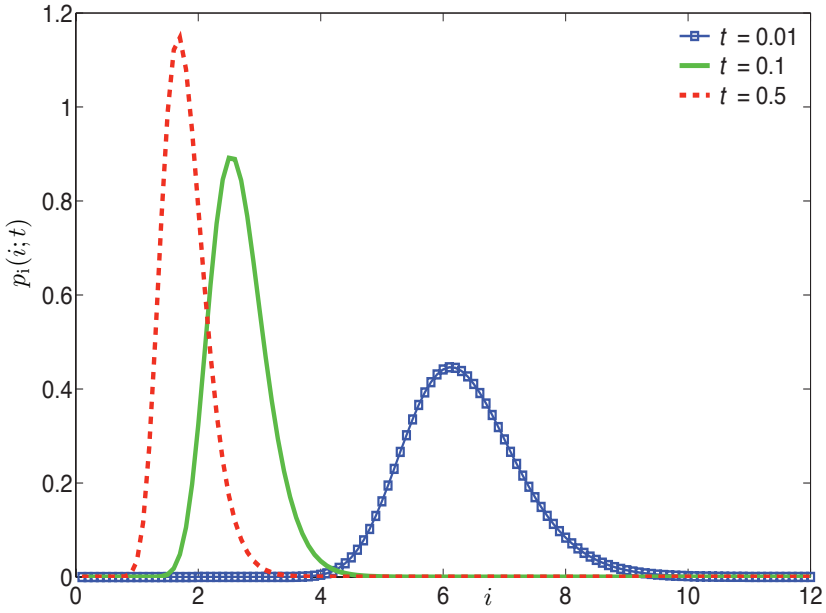

FIG. 2: Temporal evolution of the PDF of the infiltration rate $p_{\mathrm{i}}(i;t)$.

4.3 Effects of Parametric Uncertainty

The degree of uncertainty in hydraulic parameters $\ln K_s$ and $\ln \alpha_G$ is encapsulated in their coefficients of variation $CV_{\ln K}$ and CV_{α}, respectively. Figure 3 demonstrates the relative effects of these two sources of uncertainty upon the predictive uncertainty, as quantified by the infiltration depth PDF $p_{\mathrm{f}}(x_{\mathrm{f}};t)$, computed at $t = 0.1$. Uncertainty in saturated hydraulic conductivity K_s affects predictive uncertainty more than uncertainty in the Gardner parameter α_G does. Although not shown in Fig. 3, we found similar behavior at later times $t = 0.5$ and 1.0. These findings are in

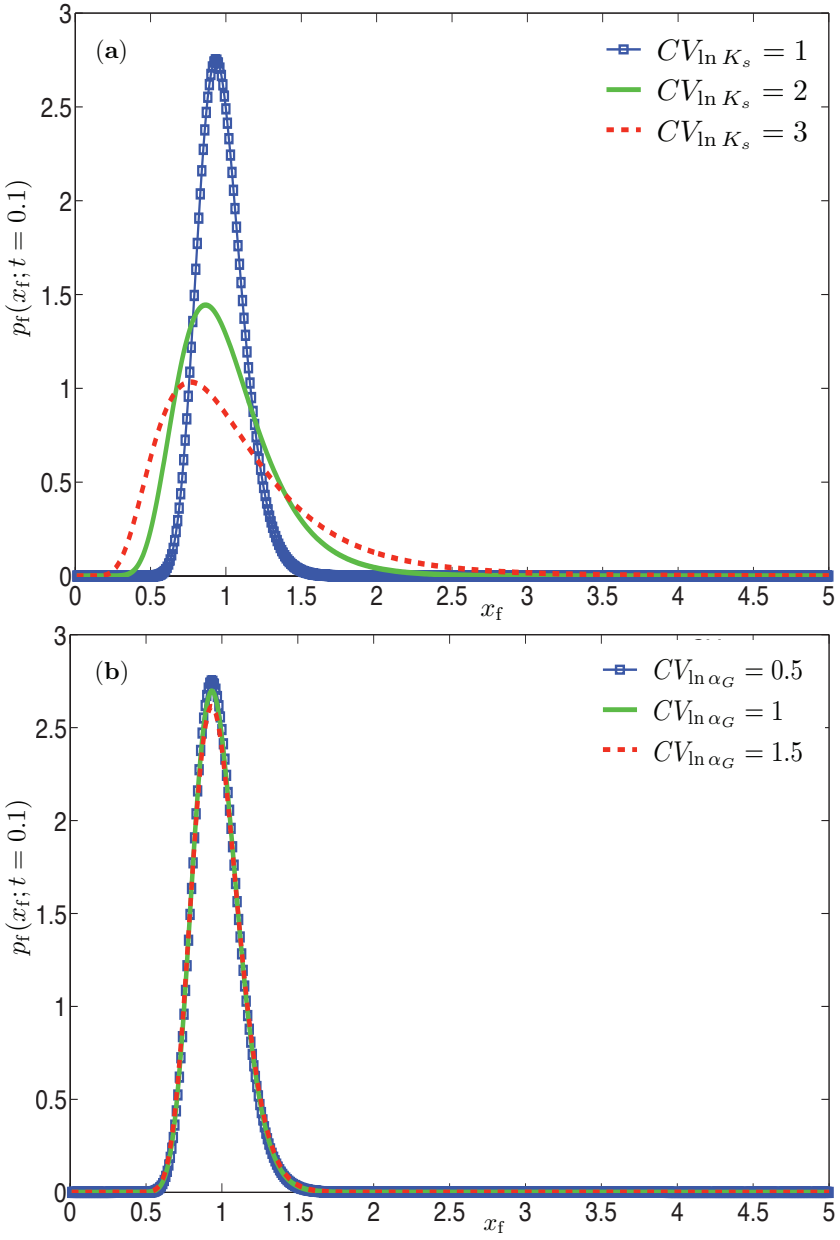

FIG. 3: The infiltration depth PDF $p_f(x_f; t = 0.1)$ for different levels of uncertainty in (a) saturated hydraulic conductivity K_s and (b) the Gardner parameter α_G.

agreement with those reported in [17, 31], wherein variances of state variables were used to conclude that uncertain saturated hydraulic conductivity K_s is the dominant factor affecting predictive uncertainty.

4.4 Effects of Cross-Correlation

The question of whether various hydraulic parameters are correlated with each other remains open, with different data sets supporting opposite conclusions (see Section 2.1). This suggests that the presence or absence of such cross-

correlations is likely to be site-specific rather than universal. The general PDF solution (23) enables us to investigate the impact of cross-correlations between saturated hydraulic conductivity K_s and the Gardner parameter α_G on predictive uncertainty. This is done by exploring the dependence of the PDF of the wetting front $p_{\mathrm{f}}(x_{\mathrm{f}}; t)$ on the correlation coefficient ρ. Figure 4 presents $p_{\mathrm{f}}(x_{\mathrm{f}}; t = 0.1)$ for $\rho = -0.99$, 0.0, and 0.99, which represent perfect anti-correlation, independence, and perfect correlation between K_s and α_G, respectively. The perfect correlation between K_s and α_G ($\rho = 0.99$) results in the minimum predictive uncertainty (the width of the distribution), while the perfect anticorrelation ($\rho = -0.99$) leads to the maximum predictive uncertainty. Predictive uncertainty resulting from the lack of correlation between K_s and α_G ($\rho = 0.0$) falls amid these two limits. The impact of cross-correlation between soil hydraulic parameters (a value of ρ) decreases with time, falling from the maximum difference of about 21% at $t = 0.01$ to about 3% at $t = 0.1$.

4.5 Effects of Selection of Hydraulic Function

Finally, we examine how the choice of a hydraulic function $K_r(\psi; \boldsymbol{\alpha})$ affects predictive uncertainty. Guided by the data analyses presented in Section 2.1, we treat α_{vG} as the only uncertain parameter in the van Genuchten hydraulic function with $n = 1.5$. To make a meaningful comparison between predictions based on the Gardner (1) and van Genuchten (2) relations, we select statistics of their respective parameters $\boldsymbol{\alpha}$ in a way that preserves the mean effective capillary drive defined by (7) [29, 32]. Specifically, we use the equivalence criteria to select the mean of $\ln \alpha_{vG}$ (-1.40, for the parameters used in these simulations) that maintains the same mean capillary drive as the Gardner model with $\overline{\ln \alpha_G} = 0.1$, and choose the variance of $\ln \alpha_{vG}$ as to maintain the original values of the coefficients of variation $CV_{\ln \alpha_{vG}} = CV_{\ln \alpha_G} = 0.5$. Figure 5 reveals that the choice between the van Genuchten and Gardner models has a significant effect on predictive uncertainty of the wetting front dynamics, although this influence diminishes with time. For example, the difference between the variances is 40% at $t = 0.01$ and 23% at $t = 0.1$.

5. CONCLUSION

We presented an approach for computing probability density functions (PDFs) of both infiltration rates and wetting fronts propagating through heterogeneous soils with uncertain (random) hydraulic parameters. Our analysis employs

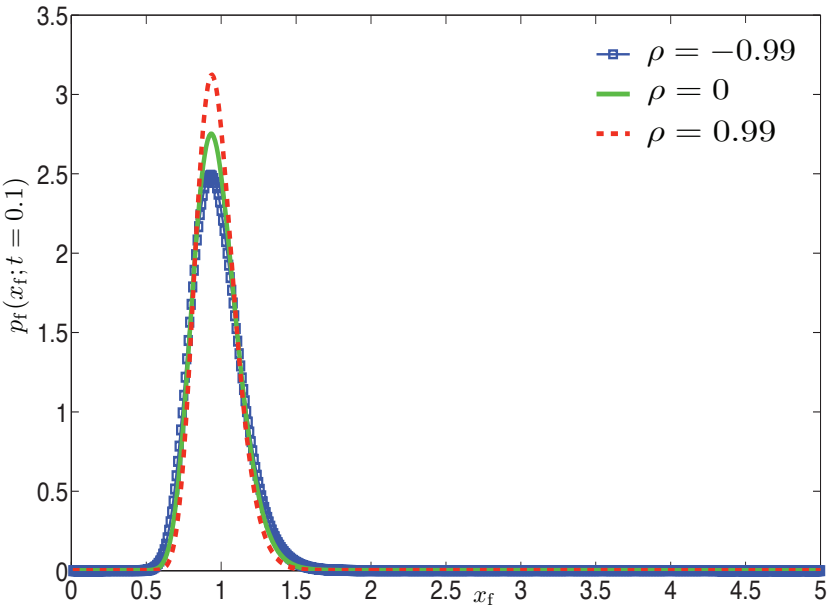

FIG. 4: The infiltration depth PDF $p_{\mathrm{f}}(x_{\mathrm{f}}; t = 0.1)$ for different levels of correlation ρ between hydraulic parameters K_s and α_G.

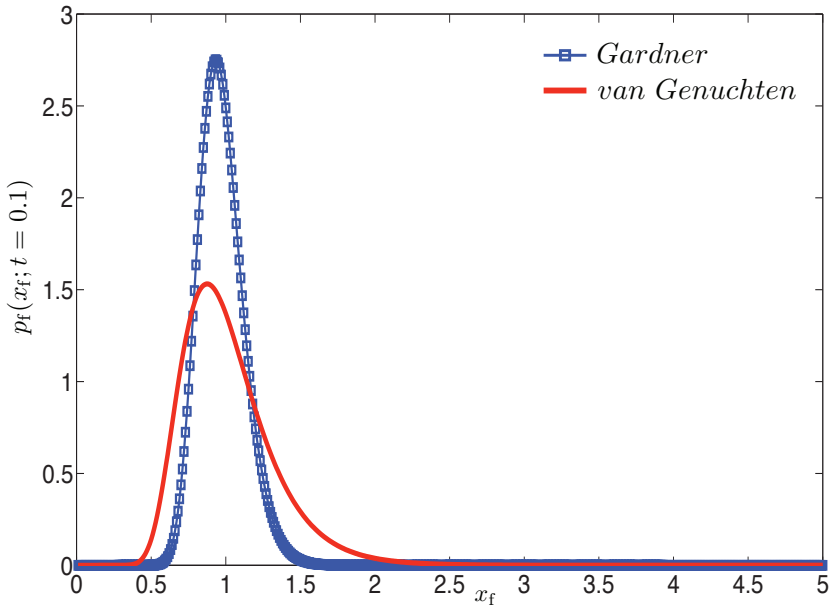

FIG. 5: The infiltration depth PDF $p_{\mathrm{f}}(x_{\mathrm{f}}; t = 0.1)$ resulting from use of the Gardner and van Genuchten hydraulic functions.

the Green–Ampt model of infiltration and the Dagan–Bresler statistical parameterization of soil properties. Our analysis leads to the following major conclusions.

1. The proposed approach goes beyond uncertainty quantification based on mean and variance of system states by computing their PDFs. This enables one to evaluate probabilities of rare events, which are necessary for probabilistic risk assessment.
2. Both the type and parameters of the PDF of a wetting front's depth change with time. As time increases, so does the width of the PDF, reflecting the increased predictive uncertainty.
3. Both the type and parameters of the PDF of infiltration rate change at early time. At large times, the PDF of infiltration rate coincides with the PDF of saturated hydraulic conductivity, which can serve as the lower bound of uncertainty associated with predictions of infiltration rate.
4. Predictive uncertainty is most sensitive to uncertainty in the saturated hydraulic conductivity K_s. Tripling the coefficient of variation of $\ln K_s$ significantly affects the shape of the infiltration depth PDF, while the effects of tripling the coefficient of variation of $\ln \alpha_G$ (a measure of uncertainty about the Gardner parameter α_G) are relatively insignificant.
5. The degree of correlation between the hydraulic parameters K_s and α_G has considerable influence on predictive uncertainty at early times and diminishes at later times.
6. The choice of a functional form of the hydraulic function (e.g., the Gardner model vs the van Genuchten model) has a significant effect on predictive uncertainty during early stages of infiltration. This effect diminishes with time.

ACKNOWLEDGMENT

This research was supported by the Department of Energy Office of Science Advanced Scientific Computing Research (ASCR) program in Applied Mathematical Sciences; and by research grant no. IS-4090-08R from BARD, the United States—Israel Binational Agricultural Research and Development Fund.

REFERENCES

1. Gómez-Hernández, J. J. and Wen, X., To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology, *Adv. Water Resour.*, 21(1):47–61, 1998.
2. Winter, C. L. and Tartakovsky, D. M., Groundwater flow in heterogeneous composite aquifers, *Water Resour. Res.*, 38(8):1148, 2002, doi:10.1029/2001WR000450.
3. Mualem, Y., A new model for predicting the hydraulic conductivity of unsaturated porous media, *Water Resour. Res.*, 12(3):513–522, 1976.
4. Zhu, J., Young, M. H., and van Genuchten, M. T., Upscaling schemes and relationships for the Gardner and van Genuchten hydraulic functions for heterogeneous soils, *Vadose Zone J.*, 6(1):186–195, 2007.
5. Russo, D., Russo, I., and Laufer, A., On the spatial variability of parameters of the unsaturated hydraulic conductivity, *Water Resour. Res.*, 33(5):947–956, 1997.
6. Yeh, T. J., Gelhar, L. W., and Gutjahr, A. L., Stochastic analysis of unsaturated flow in heterogeneous soils. 2. Statistically anisotropic media with variable α, *Water Resour. Res.*, 21(4):457–464, 1985.
7. Mantoglou, A. and Gelhar, L. W., Stochastic modeling of large-scale transient unsaturated flow system, *Water Resour. Res.*, 23(1):37–46, 1987.
8. Hu, X. and Cushman, J., Nonequilibrium statistical mechanical derivation of a nonlocal Darcy's Law for unsaturated/saturated flow, *Stoch. Hydrol. Hydraul.*, 8(2):109–116, 1994.
9. Russo, D., Stochastic analysis of the velocity covariance and the displacement covariance tensors in partially saturated heterogeneous anisotropic porous formations, *Water Resour. Res.*, 31(7):1647–1658, 1995.
10. Severino, G. and Santini, A., On the effective hydraulic conductivity in mean vertical unsaturated steady flows, *Adv. Water Resour.*, 28(9):964–974, 2005.
11. Russo, D. and Fiori, A., Stochastic analysis of transport in a combined heterogeneous vadose zone–groundwater flow system, *Water Resour. Res.*, 45:W03426, 2009 doi:10.1029/2008WR007157.
12. Tartakovsky, D. M., Neuman, S. P., and Lu, Z., Conditional stochastic averaging of steady state unsaturated flow by means of Kirchhoff transformation, *Water Resour. Res.*, 35(3):731–745, 1999.
13. Tartakovsky, A. M., Garcia-Naranjo, L., and Tartakovsky, D. M., Transient flow in a heterogeneous vadose zone with uncertain parameters, *Vadose Zone J.*, 3(1):154–163, 2004.
14. Lu, Z., Neuman, S. P., Guadagnini, A., and Tartakovsky, D. M., Conditional moment analysis of steady-state unsaturated flow in bounded, randomly heterogeneous soils, *Water Resour. Res.*, 38(4):1038, 2002, doi:10.1029/2001WR000278.
15. Tartakovsky, D. M., Probabilistic risk analysis in subsurface hydrology, *Geophys. Res. Lett.*, 34:L05404, 2007, doi:10.1029/2007GL029245.
16. Warrick, A. W., *Soil Water Dynamics*, Oxford University Press, 2003.
17. Dagan, G. and Bresler, E., Unsaturated flow in spatially variable fields, 1. Derivation of models of infiltration and redistribution, *Water Resour. Res.*, 19(2):413–420, 1983.
18. Bresler, E. and Dagan, G., Unsaturated flow in spatially variable fields, 2. Application of water flow models to various fields, *Water Resour. Res.*, 19(2):421–428, 1983.
19. Rubin, Y. and Or, D., Stochastic modeling of unsaturated flow in heterogeneous soils with water uptake by plant roots: The parallel columns model, *Water Resour. Res.*, 29(3):619–631, 1993.
20. Indelman, P., Touber-Yasur, I., Yaron, B., and Dagan, G., Stochastic analysis of water flow and pesticides transport in a field experiment, *J. Contam. Hydrol.*, 32(1–2):77–97, 1998.
21. Zeller, K. F. and Nikolov, N. T., Quantifying simultaneous fluxes of ozone, carbon dioxide and water vapor above a subalpine forest ecosystem, *Environ. Pollut.*, 107(1):1–20, 2000.
22. Zhu, J. and Mohanty, B. P., Soil hydraulic parameter upscaling for steady-state flow with root water uptake, *Vadose Zone J.*, 3(4):1464–1470, 2004.
23. Morbidelli, R., Corradini, C., and Govindaraju, R. S., A simplified model for estimating field-scale surface runoff hydrographs, *Hydrol. Process.*, 21(13):1772–1779, 2007.
24. Meng, H., Green, T. R., Salas, J. D., and Ahuja, L. R., Development and testing of a terrain-based hydrologic model for spatial

Hortonian infiltration and runoff/on, *Environ. Model. Soft.*, 23(6):794–812, 2008.

25. Gómez, S., Severino, G., Randazzo, L., Toraldo, G., and Otero, J. M., Identification of the hydraulic conductivity using a global optimization method, *Agric. Water Manage.*, 96:504–510, 2009.
26. Russo, D. and Bouton, M., Statistical analysis of spatial variability in unsaturated flow parameters, *Water Resour. Res.*, 28(7):1911–1925, 1992.
27. Saito, H., Seki, K., and Simunek, J., An alternative deterministic method for the spatial interpolation of water retention parameters, *Hydrol. Earth Syst. Sci.*, 13:453–465, 2009.
28. Zhu, J. and Mohanty, B. P., Spatial averaging of van Genuchten hydraulic parameters for steady-state flow in heterogeneous soils: A numerical study, *Vadose Zone J.*, 1(2):261–272, 2002.
29. Tartakovsky, D. M., Guadagnini, A., and Riva, M., Stochastic averaging of nonlinear flows in heterogeneous porous media, *J. Fluid Mech.*, 492:47–62, 2003.
30. Serrano, S. E., Improved decomposition solution to Green and Ampt equation, *J. Hydrol. Eng.*, 8(3):158–160, 2003.
31. Coppola, A., Basile, A., Comegna, A., and Lamaddalena, N., Monte Carlo analysis of field water flow comparing uni- and bimodal effective hydraulic parameters for structured soil, *J. Contam. Hydrol.*, 104(1–4):153–165, 2009.
32. Morel-Seytoux, H. J., Meyer, P. D., Nachabe, M., Tourna, J., van Genuchten, M. T., and Lenhard, R. J., Parameter equivalence for the Brooks-Corey and van Genuchten soil characteristics: Preserving the effective capillary drive, *Water Resour. Res.*, 32(5):1251–1258, 1996.

International Journal for Uncertainty Quantification, 1(1):49–76, 2011

ASSIMILATION OF COARSE-SCALE DATA USING THE ENSEMBLE KALMAN FILTER

S. Akella,[1] A. Datta-Gupta,[2] & Y. Efendiev[3,]*

[1]*Department of Earth & Planetary Sciences, The Johns Hopkins University, Baltimore, MD 21218, USA*

[2]*Department of Petroleum Engineering, Texas A&M University, College Station, TX 77843, USA*

[3]*Department of Mathematics, Texas A&M University, College Station, TX 77843, USA*

Original Manuscript Submitted: 04/22/2010; Final Draft Received: 06/13/2010

Reservoir data is usually scale dependent and exhibits multiscale features. In this paper we use the ensemble Kalman filter (EnKF) to integrate data at different spatial scales for estimating reservoir fine-scale characteristics. Relationships between the various scales is modeled via upscaling techniques. We propose two versions of the EnKF to assimilate the multiscale data, (i) where all the data are assimilated together and (ii) the data are assimilated sequentially in batches. Ensemble members obtained after assimilating one set of data are used as a prior to assimilate the next set of data. Both of these versions are easily implementable with any other upscaling which links the fine to the coarse scales. The numerical results with different methods are presented in a twin experiment setup using a two-dimensional, two-phase (oil and water) flow model. Results are shown with coarse-scale permeability and coarse-scale saturation data. They indicate that additional data provides better fine-scale estimates and fractional flow predictions. We observed that the two versions of the EnKF differed in their estimates when coarse-scale permeability is provided, whereas their results are similar when coarse-scale saturation is used. This behavior is thought to be due to the nonlinearity of the upscaling operator in the case of the former data. We also tested our procedures with various precisions of the coarse-scale data to account for the inexact relationship between the fine and coarse scale data. As expected, the results show that higher precision in the coarse-scale data yielded improved estimates. With better coarse-scale modeling and inversion techniques as more data at multiple coarse scales is made available, the proposed modification to the EnKF could be relevant in future studies.

KEY WORDS: *Kalman filter, reservoir engineering, spatial uncertainty, multiscale estimation, parameter estimation*

1. INTRODUCTION

Broadly speaking, the measured data used for description of reservoir porosity and permeability characterization consist of static and dynamic data. Static data such as well logs and core samples can resolve heterogeneity at a scale of a few inches or feet with high reliability. However, dynamic data such as fractional flow or water cut (neglecting any pre-existing mobile water in the reservoir, this could be defined as the ratio of the injection fluid to the total fluid produced at the production wells), pressure transient, and tracer test data typically scan the length scales comparable to the interwell distances. Additional dynamic data such as time-lapse seismic images [1] can provide improved spatial sampling but at a lower precision. The ensemble Kalman filter (EnKF) is now being used in a number of studies for reservoir history matching. Some of the recent studies are listed in Evensen [2]; also see Nævdal et al. [3], Wen and Chen [4], Gu and Oliver [5], and Jafarpour and McLaughlin [6]. In general, reservoir data is often scale-dependent and exhibits multiscale features, and integration of additional multiscale data could further reduce the uncertainty (see Lee

*Correspond to Y. Efendiev, E-mail: yalchinrefendiev@gmail.com

2152 -5080/11/$35.00

et al. [7], Efendiev et al. [8, 9] and references therein). Also, it is important to resolve fine-scale heterogeneity for various purposes such as enhanced oil recovery, environmental remediation, etc. With that perspective, integration of data at coarse and fine scales is an important objective. Computationally efficient assimilation of multiscale data using EnKF to estimate fine-scale fields for subsurface characterization is the main topic of this study. The main reason we used EnKF in this paper is because it requires fewer ensemble members than the particle filters (where, rather than updating the ensemble members model state, we update the probability assigned to each ensemble member based on model data misfit), e.g., see [10] and references therein for further details.

In this paper, apart from the water-cut data, we consider two kinds of coarse-scale measured data as well. The coarse-scale data are assumed to be permeability and/or saturation at some specified level of precision. The unknown variables (permeability, at the fine scale), are estimated using a modification to the EnKF algorithm, linking the data at different scales via upscaling (from the finest to the coarsest scales). The main idea behind upscaling is to obtain an *effective* coarse-scale permeability which yields the same average response as that of the underlying fine-scale field, locally. First we consider coarse-scale permeability data, which could be obtained either from geological considerations or coarse-scale inversion of dynamic, fractional flow data on a coarse grid [7, 9] or also using Markov Chain Monte Carlo (MCMC) techniques [8]. This coarse-scale, static data could be viewed as *prior* information regarding the permeability or in other words, a *constraint* which is to be satisfied up to the prescribed variance while obtaining the fine-scale estimates in every data assimilation cycle using the EnKF. Upscaling methods relate the solution at the fine scale to the coarse scale; therefore, in the Kalman filtering context, it amounts to modeling a nonlinear observation operator. In this paper we study two ways to assimilate the coarse-scale data using the EnKF. The standard EnKF [2] could be used for assimilating all the available data in one step, or alternatively, the measured data could be used in batches. For example, the estimate with one data becomes a prior while assimilating the other measured data; further details are given in Section 3.

The second kind of coarse-scale observed data we consider is dynamic and is motivated based on the increasing availability of time-lapse seismic images (or 4d seismic data). Integration of inverted 4d seismic data (at fine scale) using the EnKF has been addressed in Dong et al. [11] and Skjervheim et al. [12]. In this article we consider the seismic data, not to correspond to the finest scale but to a coarse scale, since time-lapse seismic data typically have a lower spatial resolution compared to the fine-scale geologic models [13]. Since the time-lapse seismic data is collected only at specific time intervals, we used coarse-scale fluid saturation as measured data to be available at a prescribed level of precision (which accounts for the inaccuracies involved in inversion of 4d seismic data) and only for certain assimilation cycles. Therefore, unlike the coarse-scale static permeability data considered earlier, the coarse-scale saturation data is assimilated only in certain assimilation cycles (see Section 4.3 for details).

Following is the plan of this paper. For the paper to be self-contained and for notational clarity, we briefly review the governing equations and sequential data assimilation using the EnKF in Section 2. This is followed by a description of the EnKF for assimilation of coarse-scale data in Section 3. For our numerical results in Section 4, we consider a five-spot pattern, with the injection well placed in the middle of a rectangular domain and four production wells located at the vertices of the rectangle. A reference case is used to provide *true* data, which is randomly perturbed to obtain synthetic measurements in a twin experiment setup. After presenting the assimilation results with both coarse-scale permeability and saturation data, we conclude with some directions for future work in Section 5.

2. PRELIMINARIES

2.1 Fine-Scale Model

In this paper we consider two-phase flow in a subsurface formation under the assumption that the displacement is dominated by viscous effects. For simplicity, we neglect the effects of gravity, compressibility, and capillary pressure, although our proposed approach is independent of the choice of physical mechanisms. Also, porosity is considered constant. The two phases are referred to as water and oil, designated by subscripts w and o, respectively. We write Darcy's law for each phase as follows:

$$\boldsymbol{v}_j = -\frac{k_{rj}(S)}{\mu_j}\boldsymbol{\kappa}_f \nabla pr, \quad \nabla \cdot [\lambda(S)\boldsymbol{\kappa}_f \nabla pr] = h \tag{1}$$

$$\lambda(S) = \frac{k_{rw}(S)}{\mu_w} + \frac{k_{ro}(S)}{\mu_o}, \quad f(S) = \frac{k_{rw}(S)/\mu_w}{k_{rw}(S)/\mu_w + k_{ro}(S)/\mu_o}$$

$$\boldsymbol{v} = \boldsymbol{v}_w + \boldsymbol{v}_o = -\lambda(S)\boldsymbol{\kappa}_f \cdot \nabla pr \tag{2a}$$

$$\phi \frac{\partial S}{\partial t} + \boldsymbol{v} \cdot \nabla S = 0 \tag{2b}$$

The above descriptions are henceforth referred to as the *fine-scale* model of the two-phase flow problem. Here $\boldsymbol{\kappa}_f$ is the (fine-scale) permeability of the medium, $\lambda(S)$ is the total mobility, μ_j denotes phase viscosity, pr is the pressure, h is the source term, and ϕ and S denote porosity and water saturation (volume fraction), respectively.

2.2 Sequential Estimation Using EnKF

Using dynamic measured data such as water cut, we can sequentially estimate the unknown parameters (permeability, porosity, etc.) and state variables such as pressure, water saturation (two-phase flow) and production data at well locations using the EnKF [3, 5, 6, 14]. The combined state-parameter to be estimated is given by $\boldsymbol{\Psi} = [\ln(\boldsymbol{\kappa}_f), \mathbf{pr}, \mathbf{S}, \mathbf{W}_c]^T$, where $\ln(\cdot)$ is natural logarithm of the permeability field, $\mathbf{W}_c$ denotes water cut, and porosity is assumed to be known.

The EnKF introduced by Evensen [15] is a sequential Monte Carlo method where an ensemble of model states evolves in state-space, with the mean as the best estimate and spread of the ensemble as the error covariance, as summarized in the following steps. Each of the ensemble members is forecasted independently,[1]

$$\boldsymbol{\Psi}^{(i)}_{n+1} = F[\boldsymbol{\Psi}^{(i)}_n] \tag{3}$$

where $F[\cdot]$ is the forecast operator [fine-scale model Eqs. (1) and (2b)], superscript (i) denotes the ith ensemble member; from this point we on drop the time subscript. The ensemble mean and covariance are defined as

$$\overline{\boldsymbol{\Psi}} = \frac{1}{N_e}\sum_{i=1}^{N_e} \boldsymbol{\Psi}^{(i)} \tag{4a}$$

$$\mathbf{P}^f \approx \frac{1}{N_e - 1}\mathbf{A}\,(\mathbf{A})^T \tag{4b}$$

where $\mathbf{A} = (\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \ldots, \mathbf{b}^{(N_e)})$, $\mathbf{b}^{(i)} = \boldsymbol{\Psi}^{(i)} - \overline{\boldsymbol{\Psi}}$, and N_e is the number of ensemble members.

In a twin experiment, the observed water cut $\mathbf{W}^o_c$ is related to the truth via $\mathbf{W}^o_c = \mathbf{H}[\boldsymbol{\Psi}^t]$, where $\mathbf{H}[\boldsymbol{\Psi}^t]$ is the true water cut. For each ensemble member, we randomly perturb $\mathbf{W}^o_c$ to generate observational samples,

$$\mathbf{y}^{(i)} = \underbrace{\mathbf{W}^o_c}_{=\mathbf{H}[\boldsymbol{\Psi}^t]} + \boldsymbol{\nu}^{(i)} \tag{5}$$

where $\boldsymbol{\nu}^{(i)}$ simulates observational error sampling, obtained as independent and identically distributed (*iid*) samples [16] from a normal distribution with zero mean and variance $\mathbf{R}$. We note that if only the water-cut data is being measured, the mapping from model to observational space $\mathbf{H}$ is trivially equal to $[\mathbf{0}\,\mathbf{0}\,\mathbf{0}\,\mathbf{I}]$, since $\boldsymbol{\Psi} = [\ln(\boldsymbol{\kappa}), \mathbf{pr}, \mathbf{S}, \mathbf{W}_c]^T$.

The forecasted ensemble Eq. (3) is updated by assimilating the observed data,

$$\widetilde{\boldsymbol{\Psi}}^{(i)} = \boldsymbol{\Psi}^{(i)} + \mathbf{K}(\mathbf{y}^{(i)} - \mathbf{H}[\boldsymbol{\Psi}^{(i)}]) \tag{6}$$

[1] In this work we focus primarily on assimilation of coarse-scale data using the EnKF, its feasibility, and impact on fine-scale estimates with different kinds of coarse-scale data (Section 3); hence, we neglect modeling errors, which will be addressed in the future.

where $\mathbf{K}$ is the Kalman gain, given by

$$\mathbf{K} = \mathbf{P}^f \mathbf{H}^T [\mathbf{H}\mathbf{P}^f \mathbf{H}^T + \mathbf{R}]^{-1}$$

Computationally efficient implementation of the EnKF is discussed, for example, in [2] and [17] (note that the ensemble error covariance, before or after assimilation, is not explicitly computed and we instead use the ensemble members for obtaining the covariance information). We use the above set of assimilated ensemble states, $\{\widetilde{\mathbf{\Psi}}^{(i)}\}_{i=1}^{N_e}$, in the fine-scale simulation model Eq. (3) for prediction until the next set of observational data is available.

3. COARSE-SCALE DATA ASSMILATION

The EnKF presented so far used only the dynamic production data (water cut) $\mathbf{y}$ with error $\boldsymbol{\nu} = \mathbf{y} - \mathbf{H}[\mathbf{\Psi}^t]$, $\boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ to update the ensemble Eq. (6). In addition to $\mathbf{y}$, we now consider another independently measured data with independent errors $\mathbf{z}$, which is *static* and is on a coarser scale compared to the fine-scale variables in Eqs. (1) and (2b). We assume that the corresponding measurement error is given by $\boldsymbol{\omega} = \mathbf{z} - \mathbf{U}[\mathbf{\Psi}^t]$, with zero mean and $\mathbf{Q}$ covariance, $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, and $\mathbf{U}$ is a mapping of fine-scale variables $\mathbf{\Psi}$ to coarse-scale data, $\mathbf{z}$, i.e., $\mathbf{U} : \mathbf{\Psi} \mapsto \mathbf{z}$. Then the likelihood of $\mathbf{z}$ is given by

$$p(\mathbf{z}|\mathbf{\Psi}) \propto \exp\left\{ \underbrace{-\frac{1}{2}(\mathbf{z} - \mathbf{U}[\mathbf{\Psi}])^T \mathbf{Q}^{-1} (\mathbf{z} - \mathbf{U}[\mathbf{\Psi}])}_{\mathcal{J}_{\boldsymbol{z}}} \right\} \tag{7}$$

If this static data $\mathbf{z}$ corresponds to coarse-scale permeability data (as considered in [9] and [7]), then $\mathbf{U} = [\mathcal{U}\,\mathbf{0}\,\mathbf{0}\,\mathbf{0}]$, where $\mathcal{U} : \boldsymbol{\kappa}_f \mapsto \boldsymbol{\kappa}_c$ is a nonlinear mapping that maps the fine-scale permeability field ($\boldsymbol{\kappa}_f$) to coarse-scale field ($\boldsymbol{\kappa}_c$) via an upscaling procedure (e.g., Durlofsky [18] and Durlofsky [19]). (Details are provided in Section 3.3.) Alternatively, if $\mathbf{z}$ corresponds to coarse-scale saturation inverted from 4d seismic data (as mentioned in the Introduction), then $\mathbf{U} = [\mathbf{0}\,\mathbf{0}\,\mathcal{A}\,\mathbf{0}]$, such that $\mathcal{A}$ is a mapping of fine-scale saturation S_f to coarse-scale saturation $S_c = \mathcal{A}S_f$. (Here we consider a simple volume averaging for $\mathcal{A}$; further details are provided in Section 4.3).

Now, our goal is to obtain an estimate which is based on both water-cut and available coarse-scale data. The likelihood of water-cut data $\mathbf{y}$ is given by

$$p(\mathbf{y}|\mathbf{\Psi}) \propto \exp\left\{ \underbrace{-\frac{1}{2}(\mathbf{y} - \mathbf{H}[\mathbf{\Psi}])^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}[\mathbf{\Psi}])}_{\mathcal{J}_{\boldsymbol{y}}} \right\} \tag{8}$$

The probability distribution function (pdf) of the predicted ensemble,

$$p(\mathbf{\Psi}) \propto \exp\left\{ \underbrace{-\frac{1}{2}(\mathbf{\Psi} - \overline{\mathbf{\Psi}})^T (\mathbf{P}^f)^{-1} (\mathbf{\Psi} - \overline{\mathbf{\Psi}})}_{\mathcal{J}_f} \right\} \tag{9}$$

where $\overline{\mathbf{\Psi}}$ and $\mathbf{P}^f$ are the predicted ensemble mean and covariance, respectively Eqs. (4a) and (4b). Then, using Bayes theorem, we obtain

$$p(\mathbf{\Psi}|\mathbf{z},\mathbf{y}) = \frac{p(\mathbf{\Psi},\mathbf{z},\mathbf{y})}{p(\mathbf{z},\mathbf{y})} = \frac{p(\mathbf{z},\mathbf{y}|\mathbf{\Psi})\,p(\mathbf{\Psi})}{p(\mathbf{z},\mathbf{y})} \propto \underbrace{p(\mathbf{z},\mathbf{y}|\mathbf{\Psi})p(\mathbf{\Psi})}_{(\star)} = \overbrace{p(\mathbf{z}|\mathbf{\Psi})\,\underbrace{p(\mathbf{y}|\mathbf{\Psi})\,p(\mathbf{\Psi})}_{\propto\, p(\mathbf{\Psi}|\mathbf{y})}}^{(\dagger)}. \tag{10}$$

Based on the above equation, following the $(\star)$ term, all the available data ($\mathbf{z}$ and $\mathbf{y}$) could be assimilated in one step (details follow in Section 3.1), whereas based on the $(\dagger)$ term, the measured data $\mathbf{y}$ and $\mathbf{z}$ can be assimilated in a sequential manner. First assimilate the fractional flow ($\mathbf{y}$) to obtain an ensemble conditioned on $\mathbf{y}$, i.e., $p(\mathbf{\Psi}|\mathbf{y})$, which could then be used to assimilate the coarse-scale data $\mathbf{z}$ (further explained in the following Section 3.2).

3.1 Coarse-Scale Data Assmilation: In One-Step

All the available data, $\mathbf{y}$ and $\mathbf{z}$, could be assimilated in one assimilation step by a modification to the model-to-observation space operator, $\mathbf{H}$. If coarse-scale permeability data at a single coarse-scale is to be assimilated, $\mathbf{H} = [\mathcal{U}\,\mathbf{0}\,\mathbf{0}\,\mathbf{I}]$. Alternatively, if coarse-scale saturation data is available, $\mathbf{H} = [\mathbf{0}\,\mathbf{0}\,\mathcal{A}\,\mathbf{I}]$. The fine-scale estimated ensemble is obtained as in Section 2.2, with Eq. (6) modified to account for the additional coarse-scale data,

$$\widetilde{\mathbf{\Psi}}^{(i)} = \mathbf{\Psi}^{(i)} + \mathbf{K}([\mathbf{z}^{(i)}, \mathbf{y}^{(i)}]^T - \mathbf{H}[\mathbf{\Psi}^{(i)}]), \mathbf{K} = \mathbf{P}^f\mathbf{H}^T[\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}'],^{-1} \tag{11}$$

where $\mathbf{R}' = \begin{bmatrix} \mathbf{Q} & 0 \\ 0 & \mathbf{R} \end{bmatrix}$. From now we on refer to this one-step assimilation of coarse-scale data procedure as *reg EnKF*. A consequence of the particular form of $\mathbf{H}$ is that it introduces contributions for cross-correlations between upscaled variables and production data. From a computational point of view, it has been known that EnKF is not very efficient for assimilation of large amounts of data [20], which could arise in our case, in complex three-dimensional cases, and also if data at multiple coarse scales is to be assimilated. In such a situation, different kinds of data could be assimilated in batches [21], which is described in the Section 3.2.

3.2 Coarse-Scale Data Assmilation: In a Sequence

We obtain an intermediate ensemble by assimilating $\mathbf{y}$, denoted by $\{\widetilde{\mathbf{\Psi}}^{(i)}\}_{i=1}^{N_e}$,

$$p(\widetilde{\mathbf{\Psi}}) = p(\mathbf{\Psi}|\mathbf{y}) \propto \exp\{-(\mathcal{J}_f + \mathcal{J}_y)\} \tag{12}$$

as discussed in Section 2.2. This intermediate ensemble and likelihood in Eq. (7) can then be combined [† term in Eq. (10)] to obtain the final estimate $\{\widehat{\mathbf{\Psi}}^{(i)}\}_{i=1}^{N_e}$,

$$p(\widehat{\mathbf{\Psi}}) = p(\mathbf{\Psi}|\mathbf{z}, \mathbf{y}) \propto \exp\{-(\mathcal{J}_f + \mathcal{J}_y + \mathcal{J}_z)\} \tag{13}$$

Therefore, in a least-squared sense, the final estimate maximizes the posterior pdf $p(\mathbf{\Psi}|\mathbf{z}, \mathbf{y})$ and corresponds to the minimum of $\mathcal{J} = \mathcal{J}_z + \mathcal{J}_y + \mathcal{J}_f$. See Appendix A for further details (where we show that the solution $\widehat{\mathbf{\Psi}}^{(i)}$ corresponds to the minimum of $\mathcal{J}$, for any i^{th} ensemble member).

If coarse-scale data is available at only one coarse scale, then the fine-scale estimated ensemble is obtained by first assimilating production data followed by assimilation of the coarse-scale data,

$$\widetilde{\mathbf{\Psi}}^{(i)} = \mathbf{\Psi}^{(i)} + \mathbf{K}(\mathbf{y}^{(i)} - \mathbf{H}[\mathbf{\Psi}^{(i)}]), \mathbf{K} = \mathbf{P}^f\mathbf{H}^T[\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}]^{-1} \tag{14a}$$

$$\widehat{\mathbf{\Psi}}^{(i)} = \widetilde{\mathbf{\Psi}}^{(i)} + \widetilde{\mathbf{K}}(\mathbf{z}^{(i)} - \mathbf{U}[\widetilde{\mathbf{\Psi}}^{(i)}]), \widetilde{\mathbf{K}} = \widetilde{\mathbf{P}}^f\mathbf{U}^T[\mathbf{U}\widetilde{\mathbf{P}}^f\mathbf{U}^T + \mathbf{Q}]^{-1} \tag{14b}$$

$\widetilde{\mathbf{P}}^f$ is approximated using the intermediate ensemble $\widetilde{\mathbf{\Psi}}^{(i)}$; henceforth we refer to this sequential, coarse-scale EnKF data assimilation procedure as *cs-EnKF*. Note that data at multiple coarse scales can be sequentially assimilated by suitable repetition of Eq. (14b), with corresponding upscaling operators. For the coarse-scale saturation data, which may be available at only certain times, for only those assimilation cycles is Eq. (14b) applicable, whereas in the case of permeability data, considering it to be prior information regarding the fine-scale permeability, it is always to be honored; hence, both of the above steps (14a) and (14b) are to be always applied. The cs-EnKF algorithm is detailed in Appendix B, and a flow chart is given in Fig. 1. Implementation of this algorithm entails upscaling of each ensemble member at every assimilation step, i.e., N_e times the upscaling operator $\mathbf{U}[\cdot]$ needs to be applied. In addition, if the dimension of the coarse-scale grid is $N_c = n_c \times n_c$, then we need to perform an Singular Value Decomposition (SVD) of a rectangular matrix of size $N_c \times N_e$. Hence, the total computation expense involves N_e upscales and SVD of the $N_c \times N_e$ matrix. Note that a similar upscaling is involved in the case of the reg EnKF, but the size of the matrix to compute SVD is now $(N_c + N_{w_c}) \times N_e$, where N_{w_c} is the dimension of the water-cut data. In addition, if there are a number of coarse scales then the size of the matrix whose SVD is to be computed will grow for the reg EnKF, since all the data is assimilated in one step. Whereas for the cs-EnKF, coarse-scale data is assimilated in a sequence, the

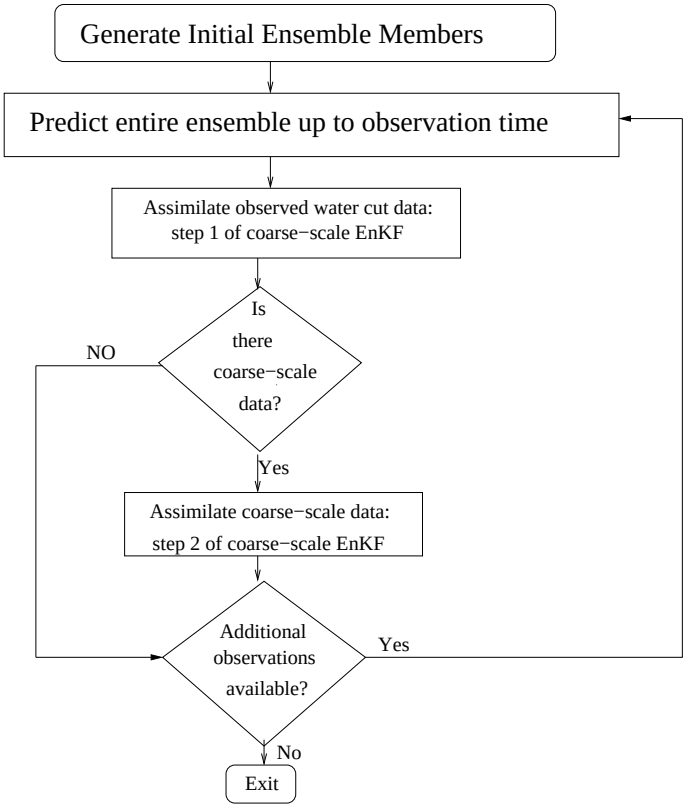

FIG. 1: Flowchart for assimilation of coarse-scale data using the EnKF.

estimate from one coarse scale being used as a prior for the next scale, the matrix size to compute SVD is always $N_c \times N_e$. (See also remark 3 of the coarse-scale EnKF algorithm in Appendix B.)

For nonlinear upscaling operators, such as $\mathcal{U}$, the final estimates from the reg EnKF would be different from those obtained using the cs-EnKF. When both the coarse-scale and water-cut data are assimilated together as in the above reg EnKF, it would imply fitting a multivariate normal likelihood to the different measured data together, whereas when the different kinds of data are assimilated one after another, as in the cs-EnKF, we fit each data separately, with a different pdf. (For further details on this topic, please see Dance [21] and references therein; also see Section 5.)

3.3 Upscaling Methods

In brief, the main idea behind upscaling of absolute fine-scale permeability is to obtain effective coarse-scale permeability for each coarse-grid block.Upscaling techniques in conjunction with the upscaling of absolute permeability have been used in groundwater applications (see, e.g., [19, 22, 23]). The link between the coarse- and the fine-scale permeability fields is usually nontrivial, because one needs to take into account the effects of all the scales present at the fine level. In the past simple arithmetic, harmonics, or power averages have been used to link properties at various scales. These averages can be reasonable for low heterogeneities or for volumetric properties such as porosity. For permeabilities, simple averaging can lead to inaccurate and misleading results. In this paper we use the flow-based upscaling methods.

Consider the fine-scale permeability that is defined on a domain with underlying fine grid as shown in Fig. 2. On the same plot a coarse-scale partitioning of the domain is also illustrated. To calculate the coarse-scale permeability field at this coarse level, we need to determine it for each coarse block, Ω_c. The coarse-scale permeability is computed

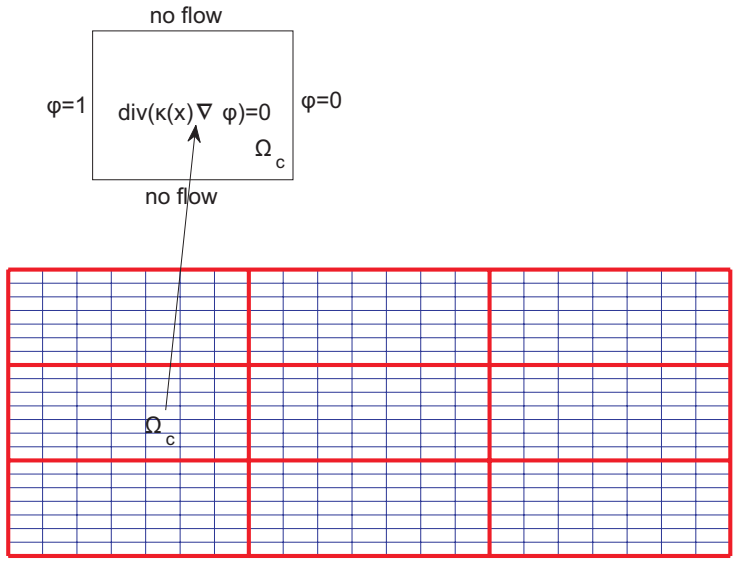

FIG. 2: Schematic illustration of upscaling (not to scale). Bold lines indicate a coarse-scale partitioning, while thin lines show a fine-scale partitioning within coarse-grid cells. In this paper we upscaled a 50×50 fine grid to a 5×5 coarse grid.

so that it delivers the same average response as that of the underlying fine-scale problem, locally. The calculation of the coarse-scale permeability based on local solutions is schematically shown in Fig. 2. For each coarse domain Ω_c we solve the local problems

$$\nabla \cdot [\boldsymbol{\kappa}_f(\mathbf{x}) \nabla \phi_j] = 0 \tag{15}$$

with some coarse-scale boundary conditions. An example of such boundary conditions is given by $\phi_j = 1$ and $\phi_j = 0$ on the opposite sides along the direction e_j and no flow boundary conditions on all other sides, alternatively, $\phi_j = x_j$ on $\partial\Omega_c$. For these boundary conditions the coarse-scale permeability ($\boldsymbol{\kappa}_c$) is given by

$$\boldsymbol{\kappa}_c \mathbf{e}_j \cdot \mathbf{e}_l = \frac{1}{|\Omega_c|} \int_{\Omega_c} \boldsymbol{\kappa}_f(\mathbf{x}) \nabla \phi_j \cdot \mathbf{e}_l \, dx \tag{16}$$

where ϕ_j is the solution of Eq. (15) with prescribed boundary conditions. Various boundary conditions such as periodic, Dirichlet, etc. can have some influence on the accuracy of the calculations. These issues have been discussed, e.g., in [24]. In particular, for determining the coarse-scale permeability field one can choose local domains that are larger than target coarse block, Ω_c, in Eq. (15). Furthermore Eq. (16) is used in the domain Ω_c, where ϕ_j are computed in the larger domains with correct scaling (see [24]). This way one reduces the effects of the artificial boundary conditions imposed on Ω_c (for details see [24]).

We denote by $\mathcal{U}$ the local operator that maps the local fine-scale permeability field $\boldsymbol{\kappa}_f$ onto $\boldsymbol{\kappa}_c$, defined on the coarse grid as in the above Eq. (16). For our computations we assume

$$\boldsymbol{\kappa}_c = \mathcal{U}(\boldsymbol{\kappa}_f) + \epsilon \tag{17}$$

where ϵ are some random fluctuations that represent inaccuracies in the coarse-scale permeability. One source of these fluctuations is the errors associated with solving inverse problems on the coarse grid. The other source of the inaccuracies include the fact that the inversion on the coarse grid does not take into account the adequate form of the coarse-scale models. Indeed, the inversion on the coarse grid for flow problems often involves the same flow equations as the underlying fine ones, for example, the same relative permeabilities are used for the coarse-scale problems as those for the fine-scale problems or the effects of macrodispersion are neglected. It is known that the flow equations at

the coarse level may have a different form than the underlying fine-scale equations [19, 25–27]. In general, this form depends on the detailed nature of the heterogeneities, which are very difficult to obtain in solving inverse problems. In our paper we use Gaussian errors in Eq. (17) and consider the impact of coarse-scale data precision (i.e., nature of ϵ) by varying the variance of ϵ (see Section 4.2 for more details).

4. NUMERICAL RESULTS

For our numerical tests, we use a 50×50 fine grid (dimensionless domain size 50×50) and two kinds of coarse-scale data in a twin experiment setup. First we consider coarse-scale permeability, which in reality, could be obtained by coarse-scale inversion of fractional flow data on a coarse grid [9, 28]. In this study we upscaled the reference fine-scale permeability (described below) to a 5×5 grid to obtain a coarse-scale permeability using flow based upscaling (Section 3.3). This coarse-scale field could be thought of as static data, which is to be honored as constraint (up to the prescribed measurement data variance) in Eq. (7); hence, we need to always assimilate it in every assimilation cycle. In reality we never know the reference field; therefore, this experimental setting is unrealistic. However, it allows us to compare and contrast a variety of test cases.

For the second set of results, a coarse-scale saturation is used which in practice could be obtained from inversion of 4d seismic measurements (see Section 1). Here, the coarse-grid saturation was obtained by volume averaging of true fine-scale saturation at some specific observation times (further details are given in Section 4.3). Therefore, unlike coarse-scale permeability, static data constraint, which is to be always satisfied, the coarse-scale saturation data is assumed to be available at only a few observation times. Following the flowchart in Fig. 1, we always have coarse-scale data if it is coarse permeability, and only at those few observation times for coarse-scale saturation data.

An initial ensemble with different permeability realizations was generated using the sequential Gaussian simulation (Deutsch and Journel [29]). We specified a Gaussian variogram model with a correlation length of 20 gridblocks in the x direction and 5 gridblocks in the y direction. One of the realizations is used as the "true" field (shown in Fig. 3) and was removed from the ensemble. Porosity (ϕ) is assumed to be equal to 0.15 for all grid blocks. For simplicity, relative permeabilities k_{rj} are assumed to be linear functions of water saturation (S): $k_{rw}(S) = S, k_{ro}(S) = 1 - S$. One injection well at the center of the field (injection rate: 71.4 m^3/day) and four producing wells at the four corners (all with equal rate of 17.85 m^3/day) were considered. The fine-scale model Eqs. (1)–(2b) are solved with no flow boundary conditions, zero initial water saturation, and by discretizing the transport equation using the first-order upwind finite volume method. In the top panel of Fig. 4 we provide the predicted fractional flow for 256 initial ensemble members along with the true fractional flow (obtained from the true permeability field).

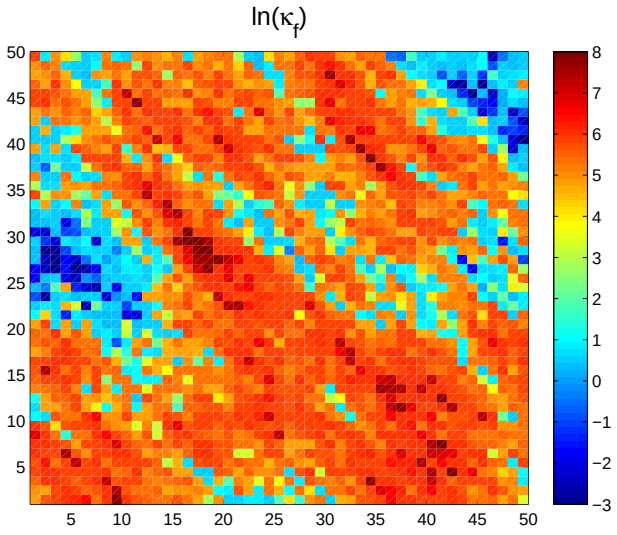

FIG. 3: Natural logarithm of 50×50 "true" permeability field.

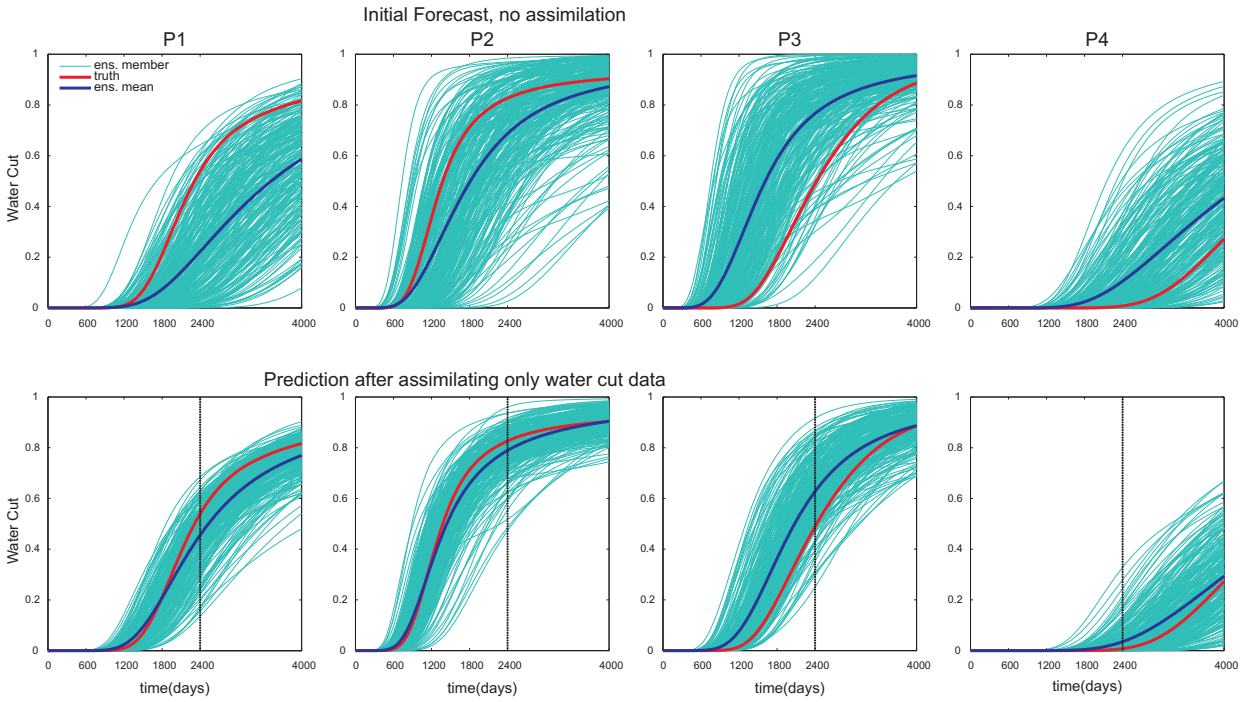

FIG. 4: Top panel: Water-cut prediction with 256 initial ensemble members (no data assimilation); ensemble members (cyan), ensemble mean (blue) compared with true water cut data (red). Bottom panel: Same as top panel but after assimilating only water-cut data as described in Section 4.1.

To compare and contrast our results using coarse-scale data and different versions of EnKF, we use the following *mean L_2-norm error*. Since we know the true (fine- and coarse-scale) field for our synthetic problem, denoting the true permeability field by $\boldsymbol{\kappa}^{\text{true}}$, the error for any ensemble member is given by

$$\mathbf{e}^{(i)} = ln(\boldsymbol{\kappa}^{(i)}) - ln(\boldsymbol{\kappa}^{\text{true}}),\ i = 1, 2, \ldots, N_e$$

Consider the L_2 norm of the error for each member, $\|\mathbf{e}^{(i)}\|_2 = \sqrt{\sum_j \left[\mathbf{e}_j^{(i)}\right]^2}$, by which we define the mean L_2 error as

$$\bar{\mathbf{e}} = \frac{1}{N_e} \sum_{i=1}^{N_e} \|\mathbf{e}^{(i)}\|_2 \tag{18}$$

so that $\bar{\mathbf{e}}$ gives us an indication of the *distance* of the entire ensemble from the true solution $\boldsymbol{\kappa}^{\text{true}}$. Since after assimilating any observation we updated all the ensemble members, we can monitor the variation of $\bar{\mathbf{e}}$ over the time of assimilation; the success of assimilation can therefore be related to the decrease in $\bar{\mathbf{e}}$.

4.1 EnKF with Water Cut Data Only

We start with a presentation of results obtained with assimilation of water-cut data only. Next we discuss results with coarse-scale data.

The water-cut data from the reference field is assumed to be available every 200 days, with mean zero and standard deviation of 0.01 (therefore $\mathbf{R}^{1/2} = 0.01\mathbf{I}_4$, where $\mathbf{I}_4$ is the unit matrix of size 4×4, since there are four producing wells). The observed data is assumed to be available up to 2400 days; hence, we performed assimilation between 200 and 2400 days. A prediction beyond the interval of data assimilation, up to 4000 days, is also provided.

The choice of ensemble size (N_e) is very important for successful data assimilation using EnKF. This is because a finite size ensemble prediction is used to estimate the prior error covariance $\mathbf{P}^f$ Eq. (4b). For small sample sizes, sampling errors in the covariance estimates result in insufficient variance for $\mathbf{P}^f$, so that observations which lie outside the small ensemble spread are completely ignored [17, 30]. (We are trying to sample a covariance matrix for unknown variables: $\ln(\boldsymbol{\kappa}), \mathbf{pr}, \mathbf{S}, \mathbf{W}_c$, i.e., an unknown of size 3×2500 plus four fractional flow data in this case, using sample sizes that are far lesser, resulting in severely reduced rank covariance matrices.) Different approaches such as covariance inflation and localization have been proposed to alleviate this problem of ensemble *inbreeding*, which is discussed elsewhere (see [31–34] and references therein for further details). An ensemble with *sufficiently large* number of members needs to be selected so that the assimilation system would not severely suffer from the above-described problem.[2] Here we present our data assimilation results with $N_e = 256$, and in Sections 4.2 and 4.3 we briefly discuss some important characteristics of the error covariance matrix such as variance and eigenspectrum in the context of coarse-scale data assimilation. The issue of coarse-scale data assimilation with smaller ensemble sizes would be tackled in the future.

In the bottom panel of Fig. 4 we plot the ensemble and true water-cut data after assimilation of only water-cut data with the EnKF. Comparing with the initial forecast (top panel), we observe that the assimilated ensemble better envelopes the true data. Also, the ensemble mean saturation field after 500, 1000, 2000, 3000, and 4000 days of simulation better compares with the true saturation than with no assimilation in Fig. 7. The final permeability field after assimilation for the ensemble mean and a few members is compared with the true field in Fig. 9. Note that the central, southeast–northwest channel is prominent, but the features at the southwest and northeast corners are not well captured, which is reflected in the plot of mean saturation (Fig. 7), where many fine-scale features present in the true saturation field are not present in the ensemble mean. Therefore, assimilation of only water-cut data helps in identifying some of the important features.

4.2 EnKF with Water-Cut and Coarse-Scale Permeability Data

In addition to water-cut production data, the coarse-scale permeability data, as described in Section 3.3 was used as additional measured data. Flow-based upscaling of the reference permeability field was used as a proxy for the inverted coarse field. Following our previous notation, this coarse-scale permeability data is denoted by $\mathbf{z}$ Eq. (7). The mapping between state variables (at fine-scale) and observations (at coarse-scale) as given by $\mathbf{U} = [\mathcal{U}\,\mathbf{0}\,\mathbf{0}\,\mathbf{0}]$, $\mathcal{U}$, denotes flow-based upscaling. For the reg EnKF, $\mathbf{H} = [\mathcal{U}\,\mathbf{0}\,\mathbf{0}\,\mathbf{I}]$ in Eq. (11) of Section 3.1.

Exactly as in the previous section, we prescribed the same frequency (of availability) and precision $\mathbf{R}$ for the water-cut data. Since we use coarse-scale permeability as additional data, it is to be assimilated whenever we assimilate water-cut data. For 5×5 coarse-scale data with mean zero and variance, $\mathbf{Q} = q\mathbf{I}_{25}$ (we present results with $q = 4, 2, 1,$ and 0.1), so that we can consider the impact of coarse-scale data precision. In the left panel of Fig. 5 we plot the variation of mean L_2 error $\bar{\mathbf{e}}$ Eq. (18) with observation time at the coarse scale for different values of q and using reg EnKF as well as cs-EnKF. In the right panel of the same figure we show the correlation between coarse-scale ensemble mean and true fields for $q = 1$. The values of correlation coefficients for different values of q are provided in Table 1. Note that as the precision of coarse-scale data is increased, i.e., for smaller value of variance, we observe a larger decrease in coarse-scale mean L_2 error and higher correlation with true coarse-scale field. This would be expected because smaller variance $\mathbf{Q}$ implies more strict coarse-scale data constraint in Eq. (7) and hence, the coarse-scale data is more accurately assimilated as it is made more precise. The water-cut data prediction using the final permeability field after assimilation for different coarse-scale data precisions is plotted in Fig. 6. (The nature of results with $q = 4$ is similar to those with $q = 2, 1, 0.5$; thus, we drop it.) Notice the improved fit of ensemble prediction to the true data for more precise coarse-scale data and also when compared to the assimilation of only water cut in Fig. 4, which is a consequence of the additional coarse-scale data being available. However, the water-cut prediction with the cs-EnKF compares better with the truth than that with the reg EnKFfor higher values of $q = 2$ and 1; with $q = 0.5$, the reg EnKF prediction is highly improved. In Fig. 7 we compare the ensemble mean with the true saturation. Once

[2]Our choice of $N_e = 256$ was based on observing the eigenspectrum and variance, discussed in Sections 4.2 and 4.3, by comparing the results for the 256 ensemble with a 1000-sized ensemble; the 256-sized ensemble did not suffer from the insufficient variance problem discussed above.

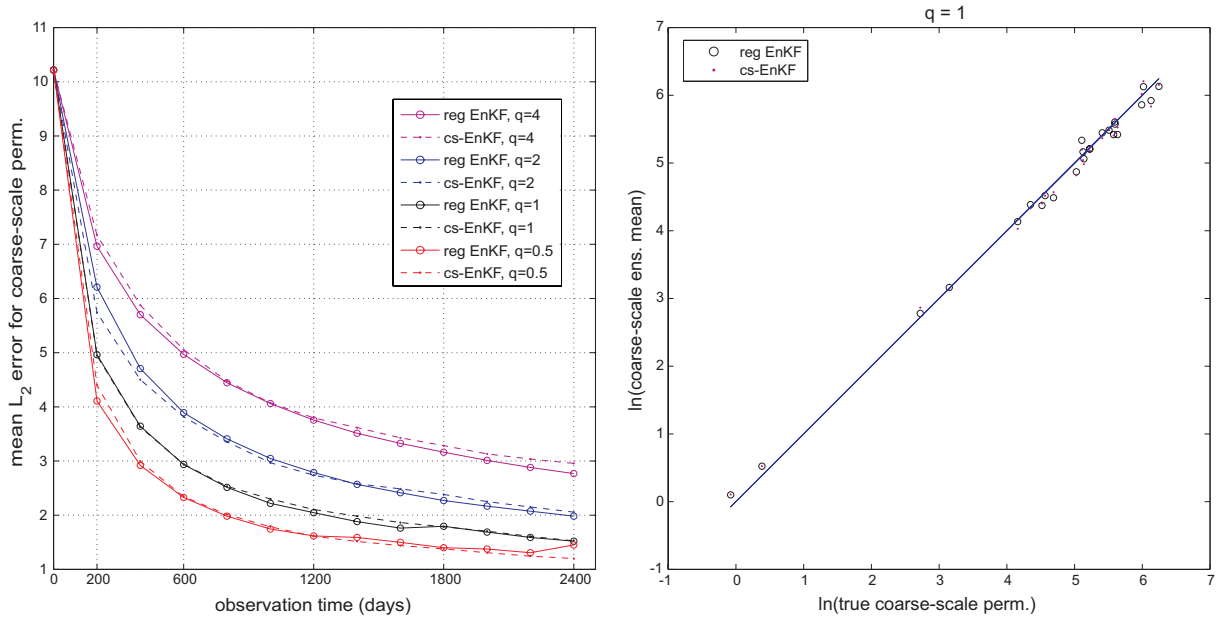

FIG. 5: Results with coarse-scale permeability and water-cut data assimilation. Left panel: Variation of mean L_2 norm error at coarse scale with assimilation time and precision of coarse-scale data $\mathbf{q}$, for both reg EnKF and cs-EnKF. Right panel: The correlation between ensemble mean permeability with the truth at coarse scale plotted for $\mathbf{q} = 1$.

TABLE 1: Correlation coefficient between ens. mean permeability $\ln(\overline{\boldsymbol{\kappa}})$ and true permeability $\ln(\boldsymbol{\kappa}^{\text{true}})$, at coarse as well as fine scales for different precisions $\mathbf{q}$ of coarse-scale permeability data. The coarse scale is denoted with subscript c and fine-scale with f. Results with both reg EnKF and cs-EnKF are given. For only water-cut data assimilation, $\text{corr}[\ln(\overline{\boldsymbol{\kappa}}_f), \ln(\boldsymbol{\kappa}_f^{\text{true}})] = 0.3074$.

	$\text{corr}[\ln(\overline{\boldsymbol{\kappa}}_c), \ln(\boldsymbol{\kappa}_c^{\text{true}})]$		$\text{corr}[\ln(\overline{\boldsymbol{\kappa}}_f), \ln(\boldsymbol{\kappa}_f^{\text{true}})]$	
$\mathbf{q}$	reg EnKF	cs-EnKF	reg EnKF	cs-EnKF
4	0.9887	0.9851	0.6484	0.6341
2	0.9963	0.9934	0.6573	0.6275
1	0.9974	0.9968	0.6546	0.6356
0.5	0.9971	0.9963	0.6096	0.6292

again, as the coarse-scale data constraint is more precisely imposed, the ensemble mean saturation captures most of the features in the true field. Also, notice that the reg EnKF saturation prediction improves more markedly as q is lowered when compared to the cs-EnKF, which could explain the better water-cut fit in Fig. 6 for $q = 0.5$ with the reg EnKF.

In the left panel of Fig. 8 we plot the fine-scale mean L_2 error for the permeability field with different values of q, and in the right panel of the same figure we plotted the correlation between the ensemble mean and the true fine-scale permeability for $q = 1$. The correlation coefficients are given in Table 1. (For assimilation of only water-cut data, we obtained a correlation coefficient equal to 0.3074.) Though the mean L_2 error is lower with the cs-EnKF, a slightly higher correlation is obtained with the reg EnKF. We observe that higher precision, i.e., lower q, does not necessarily imply highest correlation, whereas we obtained a lower mean L_2 error. The final permeability field after assimilation for the ensemble mean and a few members is shown in Fig. 9. We note that the low-permeability region at

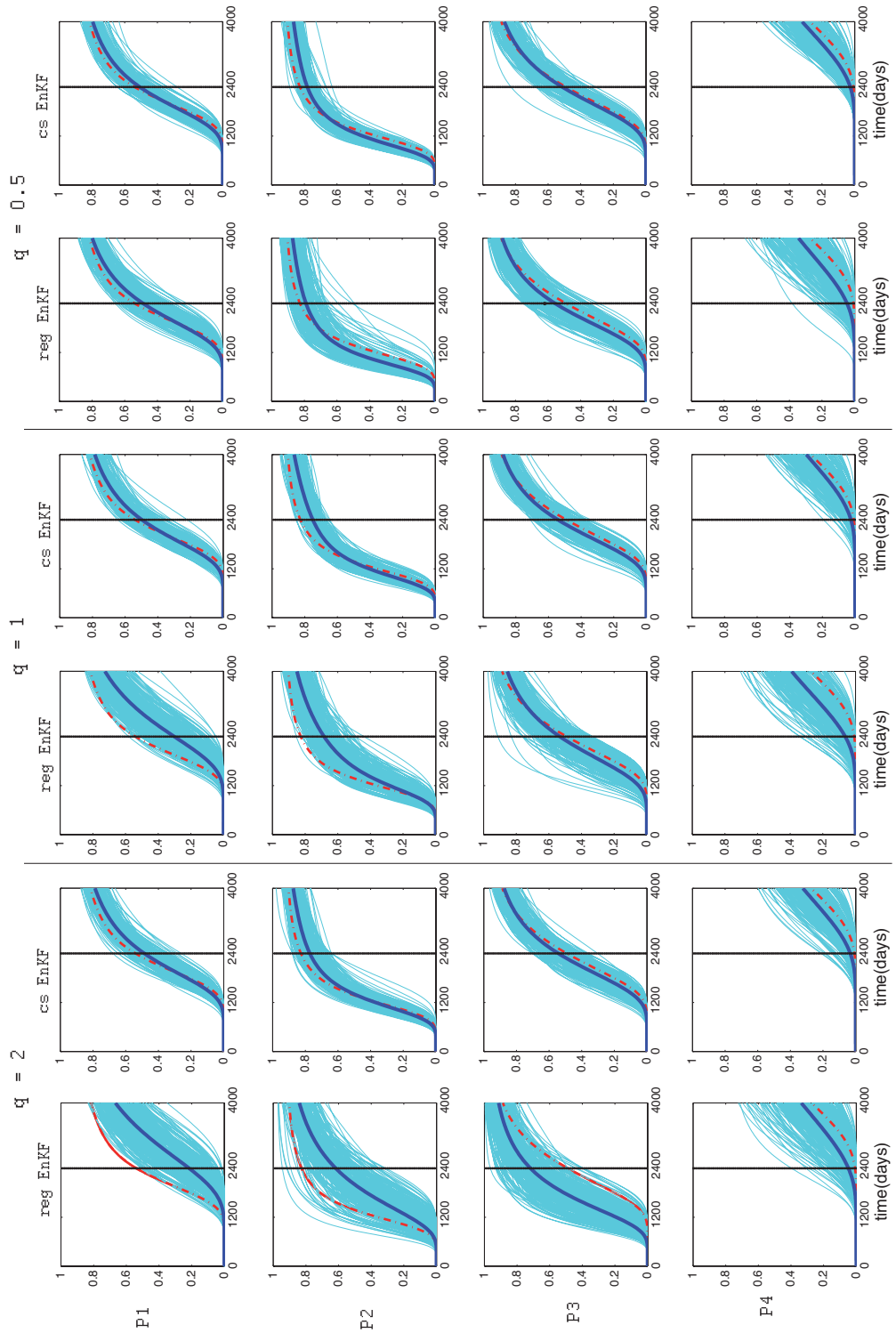

FIG. 6: Same as in Fig. 4, but for assimilation of both coarse-scale permeability and water-cut data. The coarse-scale data precision is varied, $\mathbf{q} = 2, 1, 0.5$. Results with $\mathbf{q} = 4$ had a trend similar to the ones plotted and hence, are not shown.

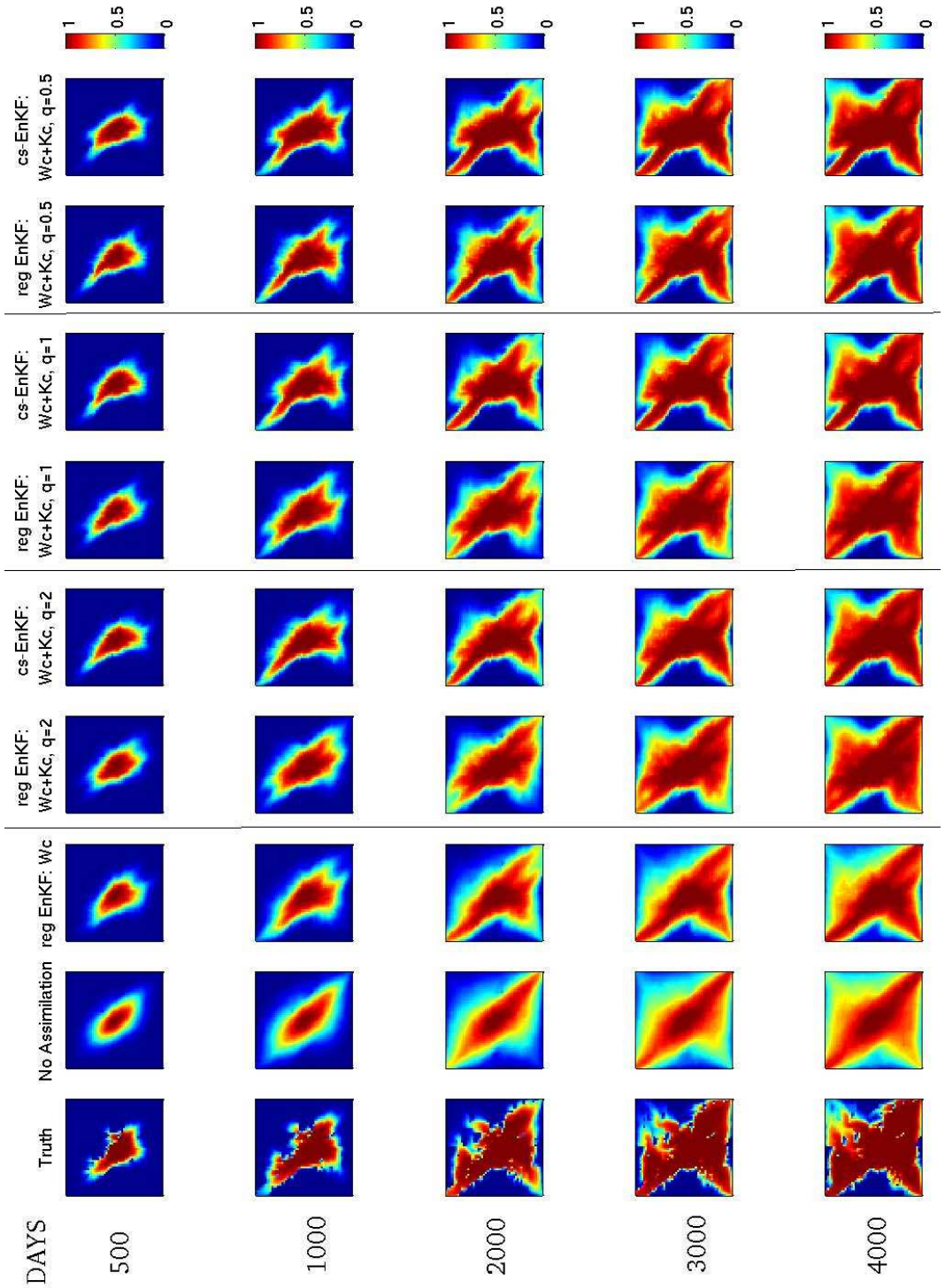

FIG. 7: Plot of the evolution of saturation field for the truth, ensemble mean: initial forecast (no assimilation), assimilation of only water cut, and assimilation of both water cut and coarse-scale permeability data with various precisions. We denote only water cut with "Wc", and water-cut and coarse-scale permeability data with "Wc+Kc".

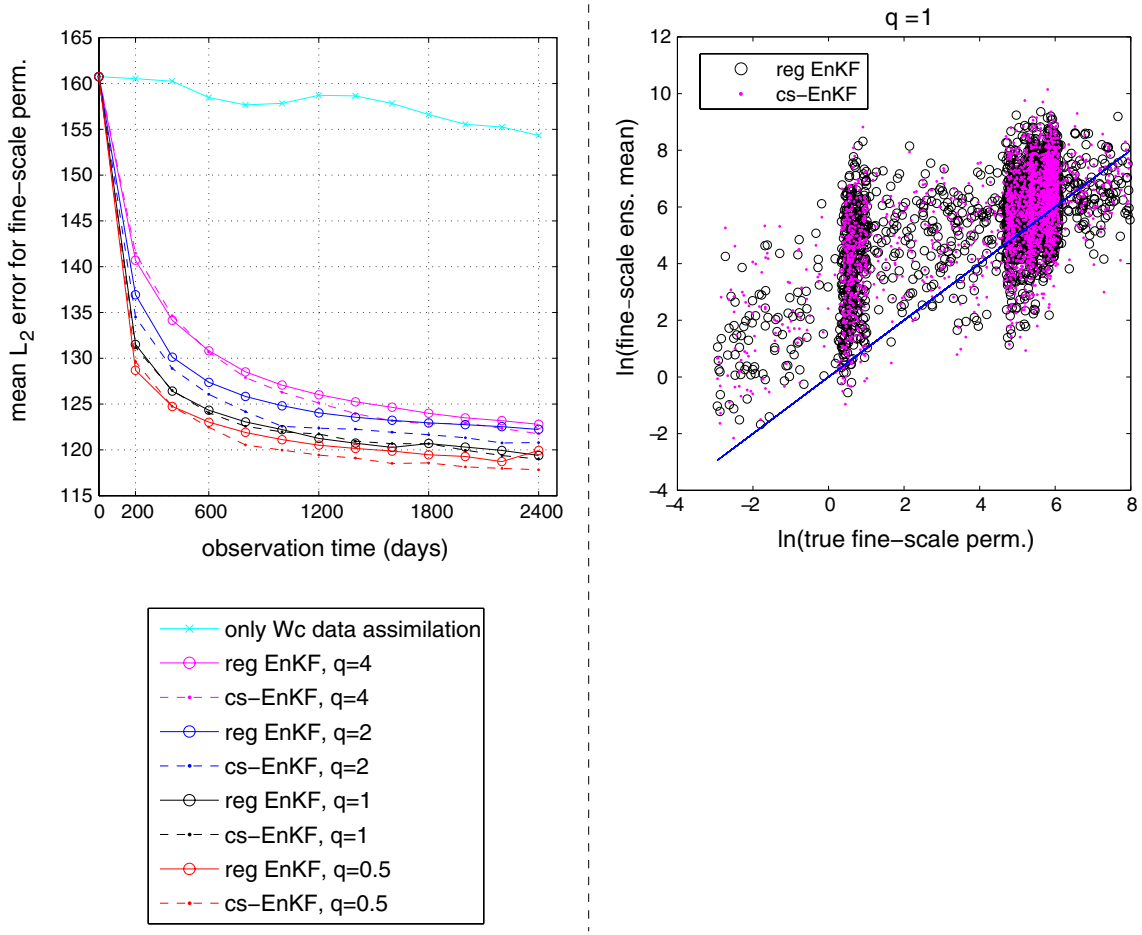

FIG. 8: Same as in Fig. 5, but for the fine scale.

the northeast and high permeability at the southwest corners are well captured. Also, the ensemble mean permeability with the cs-EnKF has some features not very well present, in that with the reg EnKF, particularly with $q = 2, 1,$ e.g., the low-permeability region to the left of the central southwest–northeast channel. As we noted in Section 3.1, the estimates obtained using the reg EnKF are expected to be different from those obtained with the cs-EnKF. We observed that for higher values of the coarse-scale data variance ($q = 2$ and 1), the cs-EnKF yielded better water cut, ensemble mean saturations, and permeability estimates when compared to the reg EnKF, whereas with $q = 0.5$, the results are similar (but not exactly same) with both versions of the EnKF. These results indicate that an optimal value for the coarse-scale data variance is important, particularly for the different versions of the EnKF (reg EnKF or cs-EnKF), and it could be obtained by a prior calculation of the uncertainty in the coarse-scale data which can be addressed in a future study. Also, more complex, studies in coarse-scale data at more than one scale and for three-dimensional models are needed to understand the merits and demerits of each version of the EnKF.

Regardless of the version of EnKF being used, after every assimilation cycle we can obtain an estimate of the analysis error covariance matrix for the ensemble fine-scale permeability. Using the fine-scale assimilated permeability fields $\boldsymbol{\kappa}_f^{(i)}$ we define the following ensemble mean permeability and error covariance,

$$\overline{\boldsymbol{\kappa}}_f = \frac{1}{N_e} \sum_{i=1}^{N_e} \boldsymbol{\kappa}_f^{(i)} \tag{19a}$$

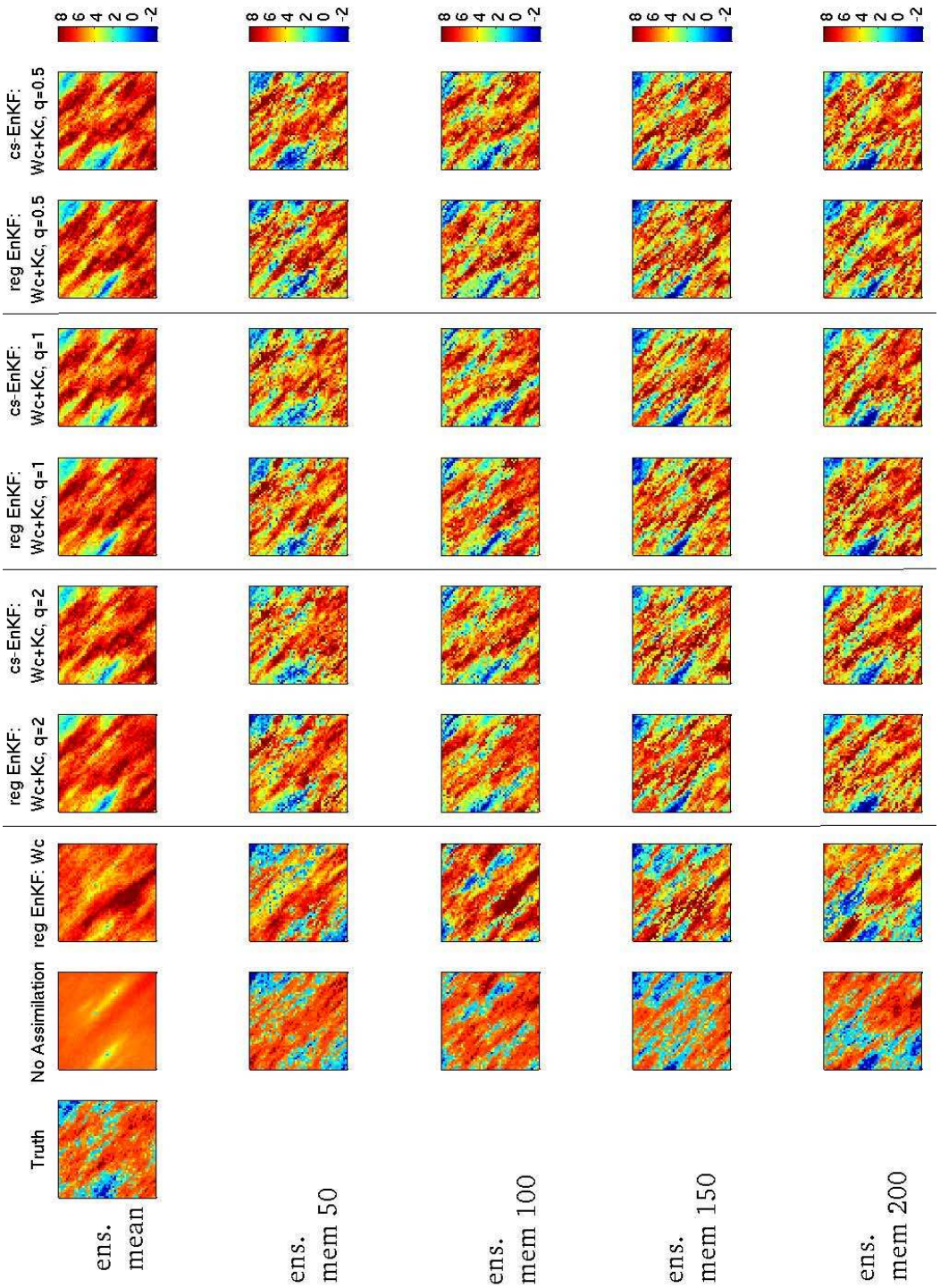

FIG. 9: Plot of the fine-scale permeability field for the truth, ensemble mean: initial forecast (no assimilation), assimilation of only water-cut, and assimilation of both water-cut and coarse-scale permeability data with various precisions in the top row. For the second column onward we show selected ensemble members: $50, 100, 150,$ and 200, respectively. We denote only water cut with "Wc", and water-cut and coarse-scale permeability data with "Wc+Kc".

$$\mathbf{P}_{\kappa_f,\kappa_f} \approx \frac{1}{N_e - 1} \sum_{i=1}^{N_e} [\boldsymbol{\kappa}_f^{(i)} - \overline{\boldsymbol{\kappa}}_f] \, [\boldsymbol{\kappa}_f^{(i)} - \overline{\boldsymbol{\kappa}}_f]^T \tag{19b}$$

Minimization of variance of the covariance matrix $\mathbf{P}_{\kappa_f,\kappa_f}$ is desired in various Kalman filtering applications [2] as it provides a measure of uncertainity. In the top row (left panel) of Fig. 10 we plotted the normalized variance of $\mathbf{P}_{\kappa_f,\kappa_f}$ during assimilation with various precisions of coarse-scale data and for both reg EnKF and cs-EnKF, and also for assimilation of only fractional flow data (we normalized using the variance from the initial ensemble). We observed that a higher reduction in variance is obtained as coarse-scale data precision is increased, and the cs-EnKF obtains more reduction than the reg EnKF. This trend is seen even for an ensemble of much larger size, for example, with 1000 members as shown in the bottom row (left panel) of Fig. 10.

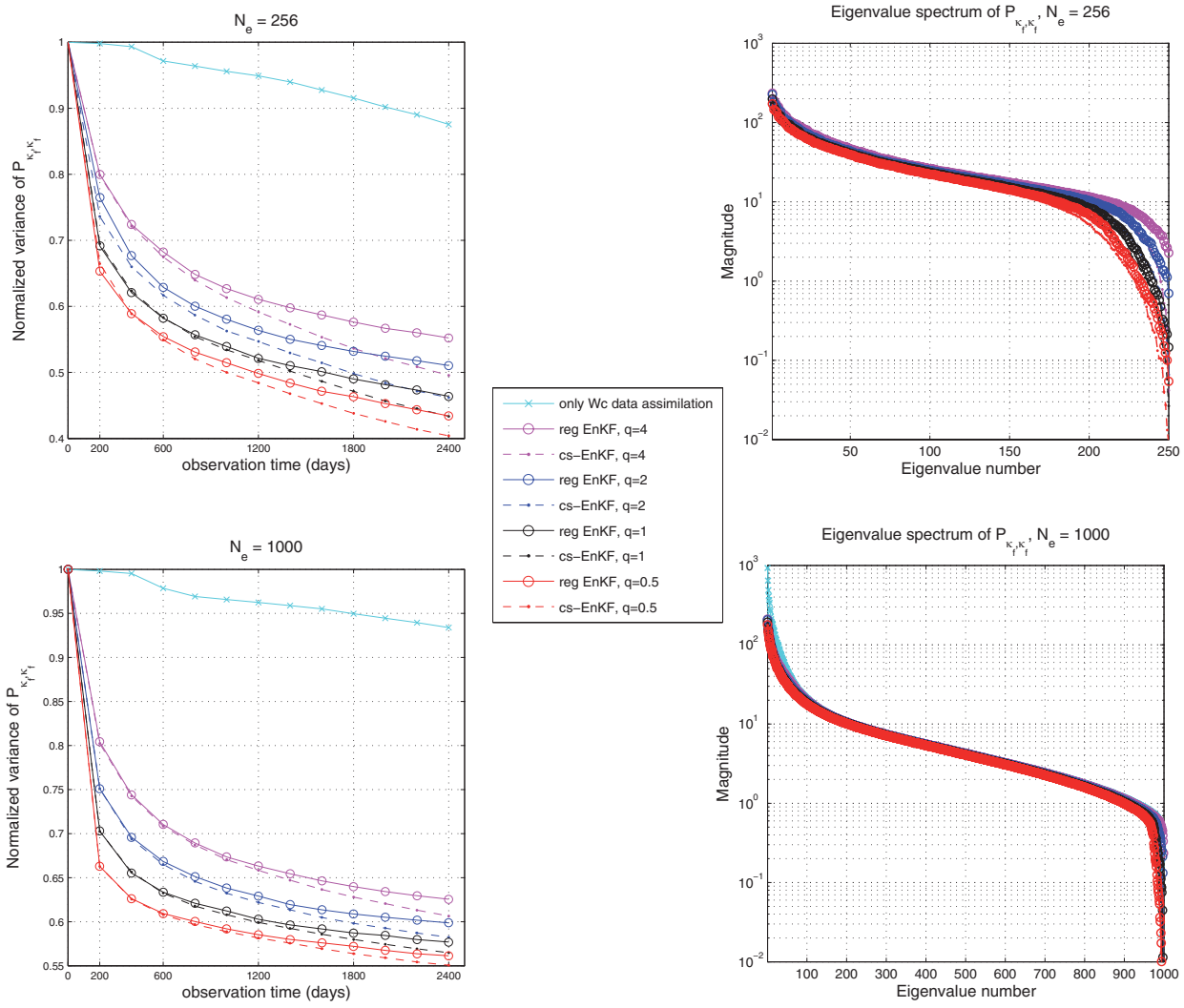

FIG. 10: Top row (left panel): Normalized variance of the error covariance matrix (Section 4.2) within assimilation window (normalized with respect to variance before assimilation, using the initial ensemble), for assimilation of only water-cut, and assimilation of both water cut and coarse-scale permeability data with various precisions. Right panel: Eigenspectrum of the error covariance matrix for the last assimilation cycle after 2400 days with 256 ensemble members. Bottom row: Same as top row but with 1000 ensemble members.

Another important property of interest, particularly for ensemble data assimilation is the eigenspectrum of $\mathbf{P}_{\kappa_f,\kappa_f}$, which is plotted in the right panels of Fig. 10; the top row represents 256 ensemble members, whereas the bottom row represents 1000 members. This eigenspectrum is from the last assimilation cycle at the end of 2400 days; for other assimilation cycles, the spectrum was similar. A *smooth* distribution of eigenvalues is desired, and indeed, we observe that, but the tails of the spectrum are steep for only water-cut data assimilation and cs-EnKF. As the coarse-scale data precision is improved, the trailing eigenspectrum gets smoothed out, which may indicate that the eigenvectors associated with the small eigenvalues are able to resolve some small-scale correlation features. Also, the larger variance associated with the leading eigenvectors as obtained for only water-cut data assimilation seems not to be the case with coarse-scale data assimilation (some of the desirable properties regarding eigenspectrum of covariance matrices are given, e.g., by Hamill et al. [32]). We note that for an ensemble size of $N_e = 1000$, the spectra are almost identical for both versions of the EnKF. In Fig. 11 we compare side-by-side the spectra with $N_e = 256$ for the reg EnKF and cs-EnKF, with the 1000 ensemble member eigenspectrum. (Since the spectra are similar for the 1000 ensemble member case, we plot and compared only with the reg EnKF.) As noted in Fig. 10, with higher precision of coarse-scale data, the tail of the spectrum gets smoothened, but also the magnitude of trailing eigenvalues is decreasing, which suggests that for smaller ensemble sizes there could be issues with loss of rank of the error covariance matrix. This problem seems to be slightly more aggrevated for the cs-EnKF. When compared to the 1000 member case, the leading eigenvalues seem to be slightly larger, as observed in [32]. In any case, in this study we did not apply any covariance inflation or localization (and always used an ensemble of size 256, which seems to preserve the rank of ensemble up to about 245 eigenvalues; see Fig. 11). These topics in the context of coarse-scale data assimilation will be covered in a future study.

4.3 EnKF with Water-Cut and Coarse-Scale Saturation Data

As mentioned in the Introduction, by coarse-scale inversion of 4d-seismic data we could obtain dynamic data such as coarse-scale pressure and saturation. In this section we attempt to assimilate such a coarse-scale saturation in a twin experiment along with the fractional flow data using the EnKF algorithms discussed in Sections 3.2 and 3.1. To this end, the saturation obtained by using the reference permeability is saved at three different times: $200, 1200$, and 2400 days, which respectively correspond to the start, middle, and end of the time window of data assimilation. This saved fine-scale saturation field is then upscaled (see Section 3.3) by volume averaging to a 5×5 coarse-scale grid and used as observed coarse-scale saturation data. If we denote the volume averaging by operator $\mathcal{A}$, acting on fine-scale saturation S_f, to give coarse-scale saturation $S_c = \mathcal{A}S_f$, then the mapping between state variables at fine scale and measured data at coarse-scale is given by $\mathbf{U} = [\mathbf{0}\,\mathbf{0}\,\mathcal{A}\,\mathbf{0}]$. Therefore for the reg EnKF (Section 3.1), the measured data is related to the fine-scale variables via $\mathbf{H} = [\mathbf{0}\,\mathbf{0}\,\mathcal{A}\,\mathbf{I}]$ in Eq. (11). For the cs-EnKFwe use above $\mathbf{U}$ operator to compute the misfit: $\mathbf{z} - \mathbf{U}[\mathbf{\Psi}]$ in Eq. (14b) (i.e., steps 2.1 and 2.4 of the cs-EnKF algorithm in Appendix B). Unlike the coarse-scale permeability data which was taken into account at every assimilation step, by construction, the coarse-scale saturation data is available only at a few assimilation steps, in this particular case, assimilation after $200, 1200$, and 2400 days.

To be consistent with our previous results, the frequency (of availability) and precision $\mathbf{R}$ for the water-cut data has been kept the same. For the coarse-scale saturation data we prescribed zero mean and variance, $\mathbf{Q} = q_s\mathbf{I}_{25}$, with $q_s = 0.1, 0.01$, such that the precision is varied from low to high. Since the saturation ranges between 0 and 1, and the fractional flow data is usually more accurately measured than 4d-seismic data, we picked q_s to be always larger than the variance in fractional flow data. In the left panel of Fig. 12 we plotted the variation of mean L_2 error for the coarse-scale saturation (while assimilating) vs observation time. (For our test case, we had assumed zero initial water saturation; therefore, the water saturation increases in time and hence, the inherent, increasing trend in this figure.) Note that whenever the coarse-scale saturation is assimilated ($200, 1200$, and 2400 days) the error decreases for both values of q_s considered. The water-cut data prediction using the assimilated ensemble members is given in Fig. 13; the fit of ensemble water-cut with the truth gets better as the coarse-scale saturation is prescibed higher precision.

We discuss the fine-scale results starting with fine-scale saturation and then the fine-scale permeability. The ensemble mean saturation is compared to the true field at certain times in Fig. 14. By assimilating the coarse-scale saturation data we are able to capture many of the subtle features that are present in the true saturation field, such

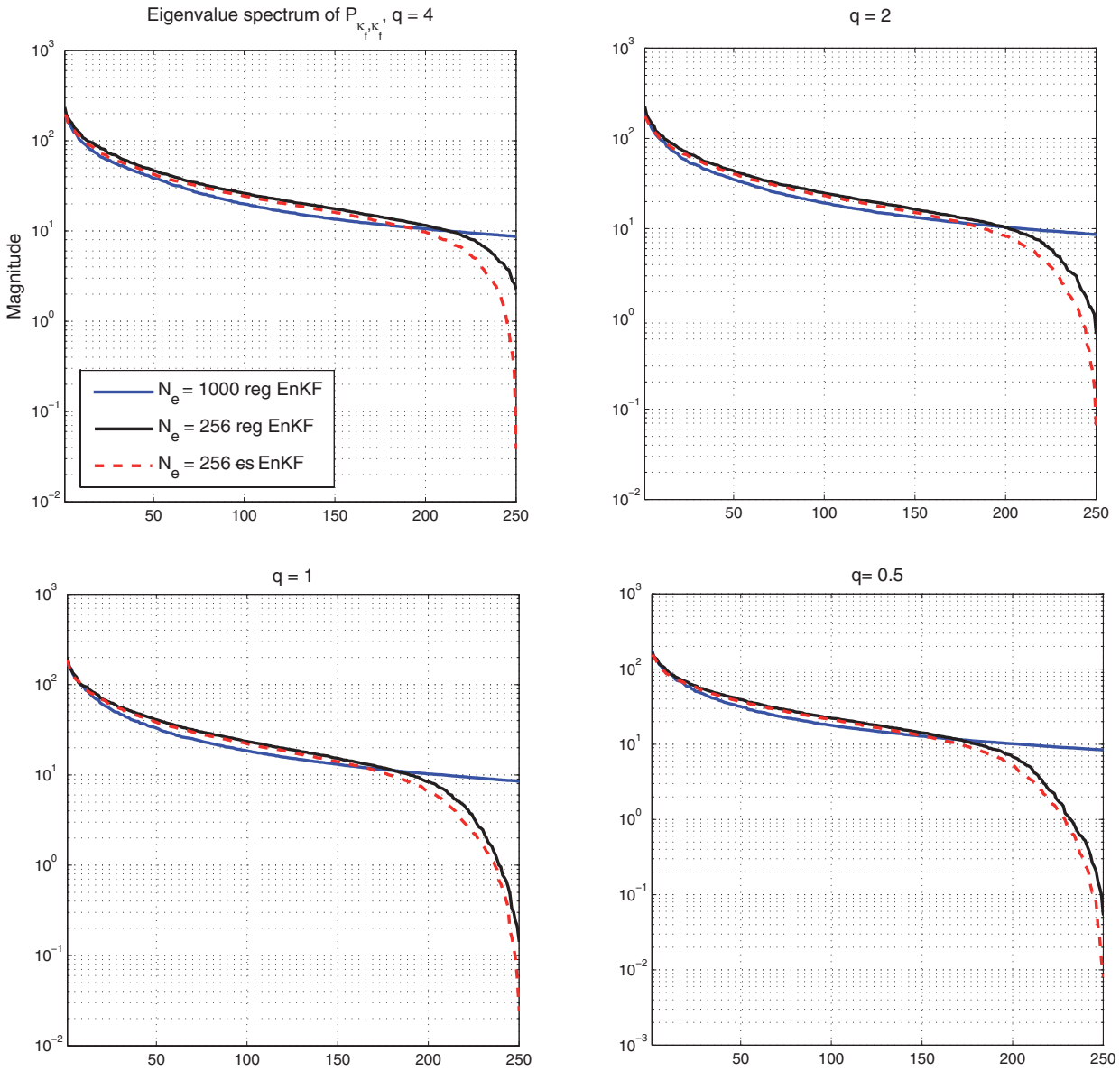

FIG. 11: Comparison of the eigenspectrum of the error covariance matrix for 256 and 1000 ensemble members for different precisions (q) of the coarse-scale data. As in Fig. 10, the spectrum corresponds to the last assimilation cycle. Plot of the spectrum for cs-EnKF and reg EnKF for 1000 ensemble members was very similar (bottom row of Fig. 10) and hence, is not plotted.

as the fingers that develop off the center toward the northeast corner and sharp contrast between different levels of saturation, throughout the entire time interval (up to 4000 days) considered. Regarding the fine-scale permeability, a comparison of the mean L_2 error is shown in the right panel of Fig. 12. Note the marked reduction in error when coarse-scale saturation data is assimilated. Also, the correlation of the ensemble mean permeability with the truth is improved as q_s is decreased, as shown in Table 2. The fine-scale permeability fields for a few ensemble members and the mean are shown in Fig. 15. Based on the above results, we observed that unlike in the case of coarse-scale permeability data assimilation, coarse-scale saturation data assimilation does not yield significant reduction in error

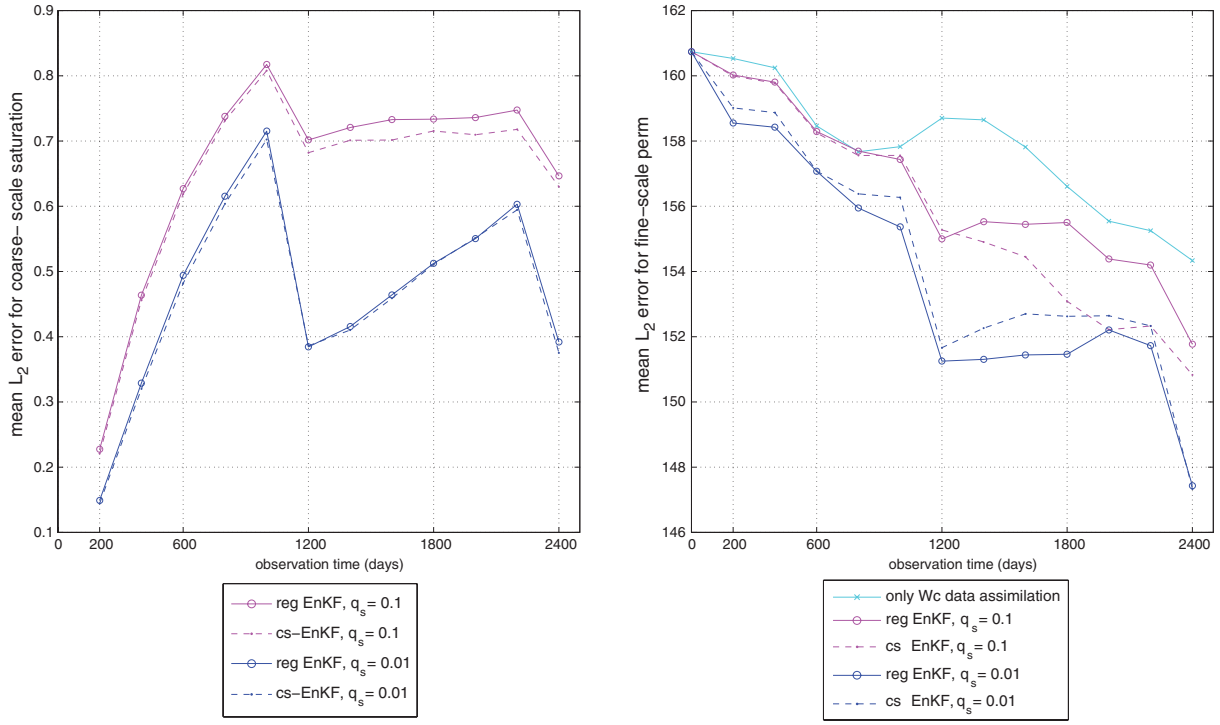

FIG. 12: Assimilation of coarse-scale saturation and water cut. Left panel: Variation of mean L_2 norm error for coarse-scale saturation, at different precisions of coarse-scale data $\mathbf{q}_s$, for both reg EnKF and cs-EnKF with coarse-scale saturation data. Right panel: Same as in left, but for the fine-scale permeability field.

(comparing the left panel in Fig. 8 and the right panel in Fig. 12); neither are the correlations of fine-scale ensemble mean with the truth (second column of Tables 1 and 2). The results (ensemble mean saturation and permeability) are improved when compared to assimilation of only water-cut data, particularly with $q_s = 0.01$, but again, not as much improved as with coarse-scale permeability data assimilation. This could be anticipated, since the fine-scale permeability is more correlated to coarse-scale permeability than to the coarse-scale saturation. Regarding the two versions of the EnKF considered for assimilating the coarse-scale saturation data, results are very similar to each other, even those for the variance of error covariance matrix $\mathbf{P}_{\kappa_f,\kappa_f}$ and its eigenspectrum at the end of 2400 days (Fig. 16). Note the steepness of the trailing eigenvalues with $q_s = 0.1$ (for both reg EnKF and cs-EnKF), which is similar to the assimilation of only fractional flow data; however, for $q_s = 0.01$, this undesirable effect has been smoothned out. This is similar to the result obtained with coarse-scale permeabilty data assimilation in Fig. 10. Our observation that both versions of EnKF performed similarly could be due to the linearity of the fine-scale to coarse-scale saturation mapping $\mathcal{A}$; however, identical results would not be possible due to the different analysis equations and sampling of errors used in Eqs. (14b) and (11).

5. CONCLUSIONS

The EnKF is increasingly being used for subsurface characterization in various geological and groundwater applications to identify fine-scale state and parameters. Recently, dynamic data other than production data has been considered in the EnKF context [11, 12]; nevertheless, the observed data to be assimilated was assumed to be at the finest scale. For a number of reasons, it is widely recognized that usage of additional multiscale data could further reduce the uncertainty at the fine scale. Also, it is often important to preserve large-scale features of the permeability field. These are coarse-scale features that can typically represent connectivity of the media. For example, facies consisting of high

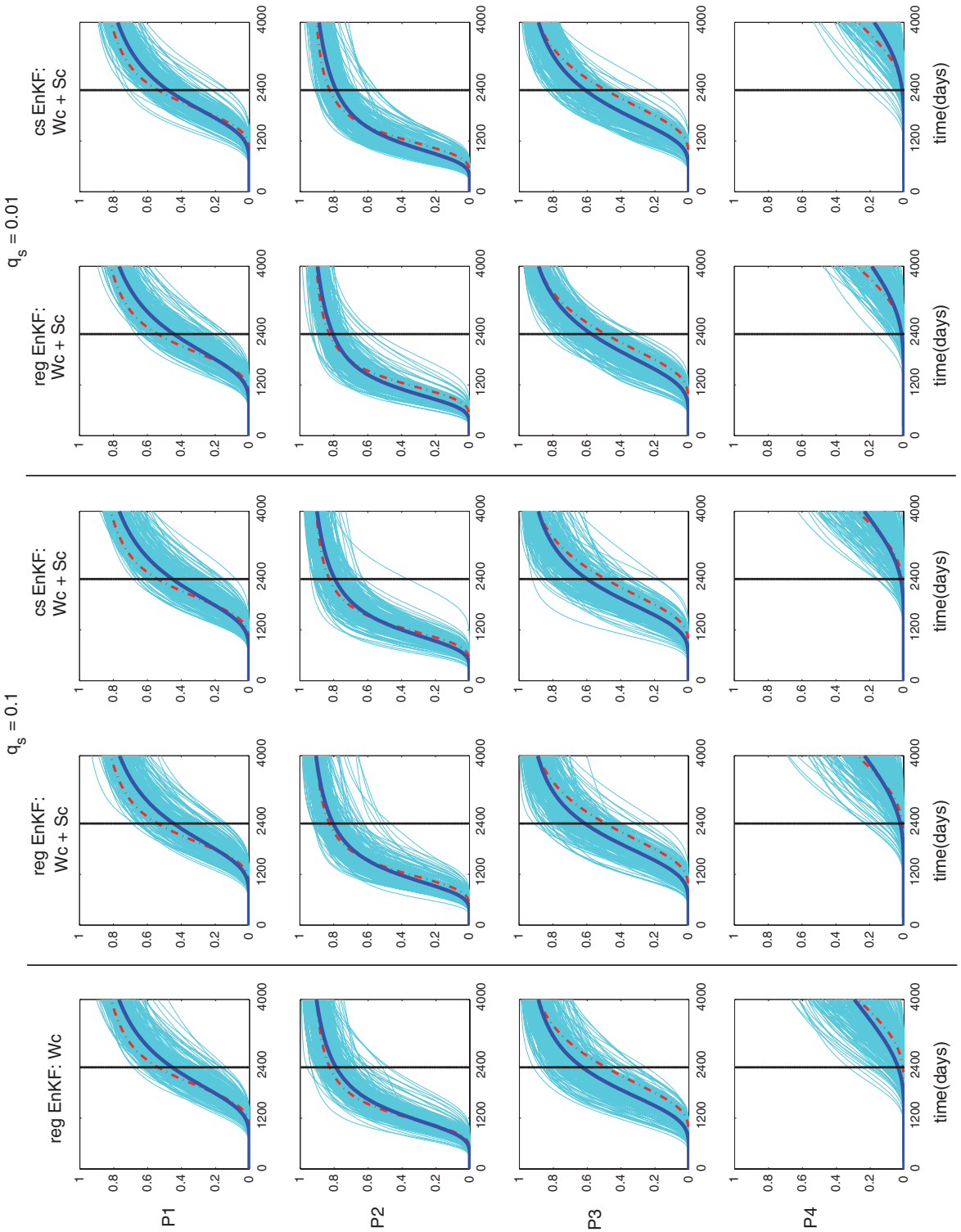

FIG. 13: Similar to Fig. 6, but assimilation of water-cut and coarse-scale saturation at different precisions. We denote only water cut with "Wc", and water-cut and coarse-scale saturation data with "Wc+Sc".

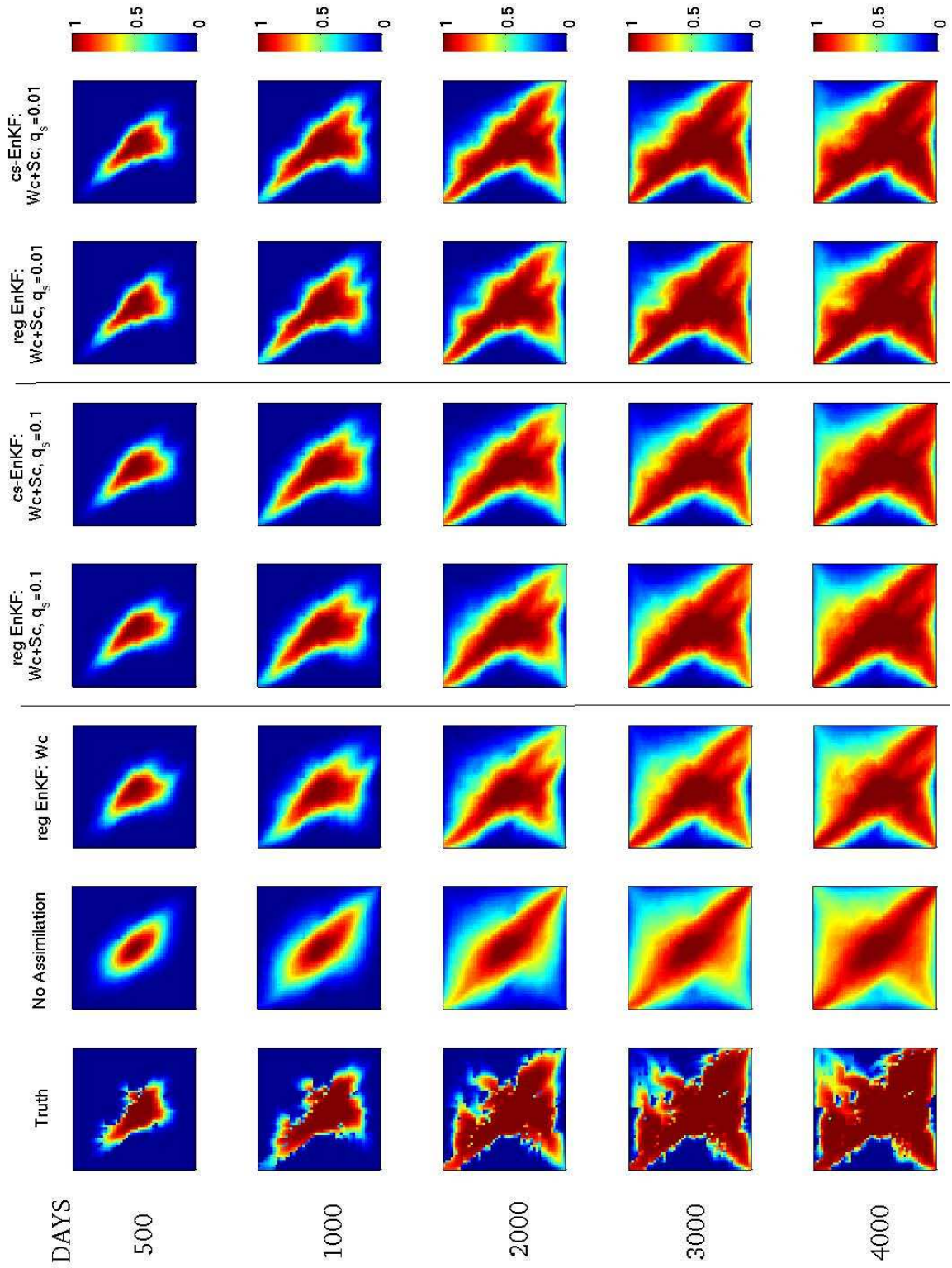

FIG. 14: Same as in Fig. 7, but for the assimilation of coarse-scale saturation and water cut.

TABLE 2: Same as Table 1, but for assimilation of coarse-scale saturation data.

	$\mathrm{corr}[\ln(\overline{\boldsymbol{\kappa}}_f), \ln(\boldsymbol{\kappa}_f^{\mathrm{true}})]$	
$\mathbf{q}_s$	reg EnKF	cs-EnKF
0.1	0.2861	0.2746
0.01	0.3149	0.3280

permeable regions are important for flow and transport. On the other hand, these features alone are not sufficient to match water-cut data. By performing inversion on the coarse grid, we can capture these large-scale features with more accuracy and certainty, and in turn, the coarse-scale inverted permeability could be used as a prior constraint in EnKF data assimilation of the fine-scale fields.

Here we proposed assimilation of coarse-scale data along with water-cut production data. We showed that the modifications to the EnKF for multiscale data assimilation are completely recursive and easily implementable. The relation between fine and coarse scales was modeled via flow-based upscaling, which could be thought of as a nonlinear observation operator linking the coarse-scale data to the unknown fine-scale variables. In addition, the proposed methodology could be used in any other sequential data assimilation method, as well to assimilate data at multiple coarse scales. Two versions of EnKF were suggested: (*i*) reg EnKF, where all the data (coarse-scale and water-cut) were assimilated together, and (*ii*) cs-EnKF, where the data were assimilated sequentially in batches. Ensemble members obtained after assimilating water-cut data are used as a prior to assimilate coarse-scale data. Though in our current paper we used only one coarse scale, the proposed method can be easily implemented to integrate as many coarse scales as required by the available data. Also the methodology is independent of the upscaling operator.

The assimilation setup was tested and compared for a two-dimensional synthetic 50×50 heterogenous true field. Two kinds of coarse-scale data were considered. In the first implementation, coarse-scale permeability data was considered and in the second, coarse-scale saturation. In our twin experiment setup, both of these data were derived from the reference field and, in both cases a 5×5 coarse grid was used. The coarse-scale data variance was varied from low to high in order to study its impact on fine-scale assimilated fields and water-cut predictions. In all cases we observed that the assimilated, ensemble mean coarse-scale (permeability/saturation) field for all variances was highly correlated to the true coarse-scale field. In addition, lower variance in the coarse-scale data yielded higher correlation. The water-cut data was better honored, for higher precision of coarse data. When assimilating coarse-scale permeability we observed that the cs-EnKF gave better fit with the true saturation, water-cut, and fine-scale permeability than the reg EnKF. It also yielded less error in an averaged L_2 norm error taken with regard to the reference field, whereas both versions of EnKF performed similarly when assimilating coarse-scale saturation. As shown in Appendix A, for a linear observation operator, assimilation of coarse-scale data in batches (i.e., as in cs-EnKF) or in shot (as in reg EnKF) would yield the same estimate. Then the cause for the difference in the performance of the two versions of EnKF for assimilation of coarse-scale permeability could be either due to ensemble size or linearity/nonlinearity of the upscaling (observation) operator. This issue has been outlined by Dance [21] and references therein. As far as the number of ensembles is concerned, we observed that even with a larger ensemble size of 1000 members, we noticed different performance of the two versions of EnKF (Fig. 10, reduction in normalized variance of $\mathbf{P}_{\kappa_f,\kappa_f}$ while having similar ranks in terms of the eigenspectrum). Certainly further study with different upscaling operators and coarse-scale data at multiple levels would be needed to study this aspect of the two versions of EnKF considered here. Over all, inclusion of coarse-scale data replicated many of the subtle features present in the fine-scale permeability and saturation fields which were not present after assimilating only water-cut data.

As our results indicate that the inclusion of coarse-scale data enhances identification of the multi-scale reservoir characteristics, it is important to study methods to obtain coarse-scale data as well as its precision. In a realistic scenario, coarse-scale inversion [7, 28], and in the future perhaps with more computing resources, MCMC methods [8], could be used to obtain such data. The coarse-scale saturation obtained using inversion has been shown to yield improved estimates in a three-dimensional reservoir case by Devegowda et al. [35]. Our current and future work is directed toward obtaining coarse-scale data with higher precision and its assimilation using ensembles of smaller size for complex three-dimensional cases.

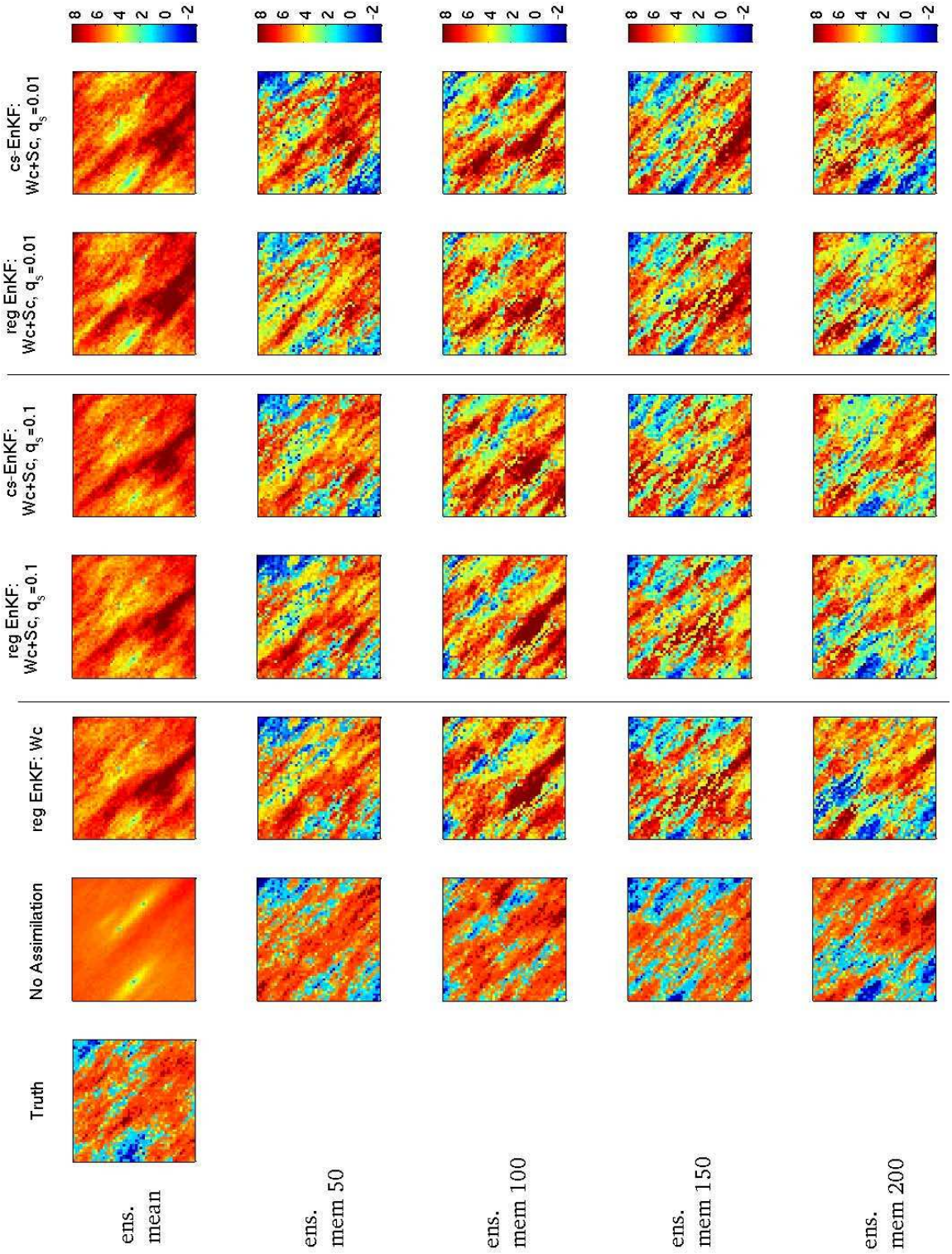

FIG. 15: Same as in Fig. 9, but for the assimilation of coarse-scale saturation and water cut.

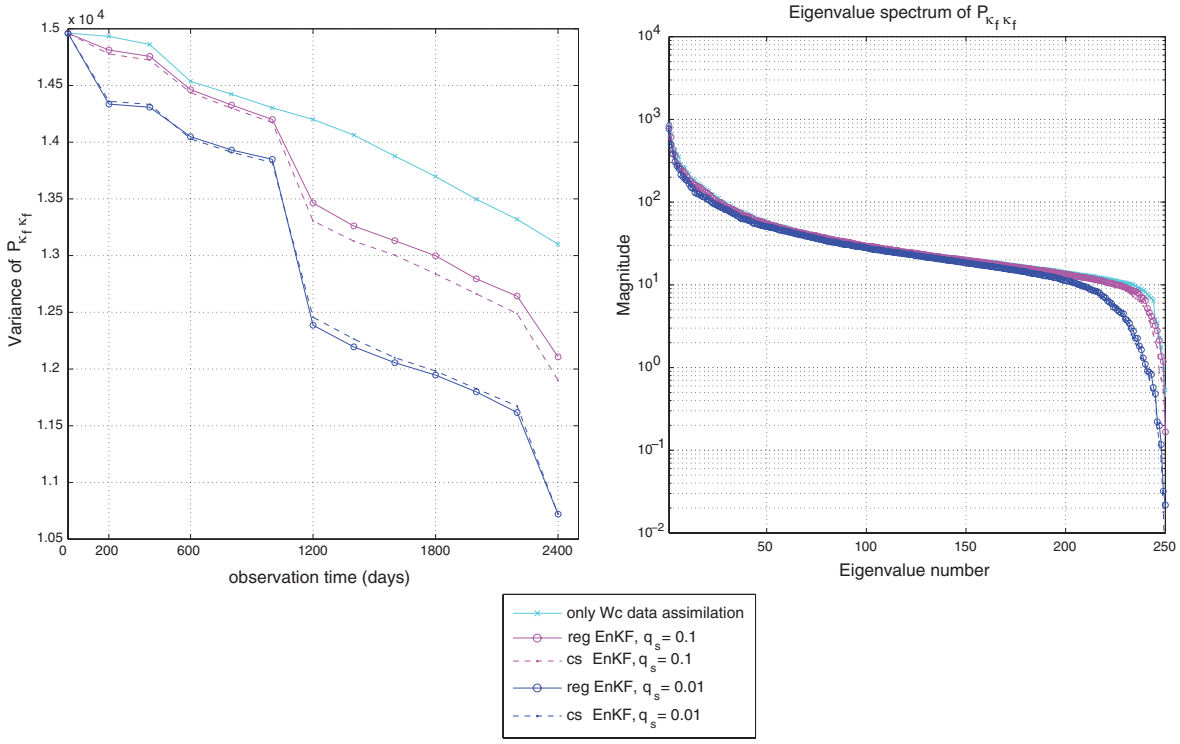

FIG. 16: Same as in Fig. 10 (top row), but for the assimilation of coarse-scale saturation and water cut.

6. ACKNOWLEDGMENTS

The work of Akhil Datta-Gupta and Yalchin Efendiev is partially supported by DOE (DE-FG03-00ER15034), NSF CMG 0724704, and KAUST award number KUS-C1-016-04.

REFERENCES

1. Lumley, D. E., Time-lapse seismic reservoir monitoring, *Geophysics*, 66:50-53, 2001.
2. Evensen, G., *Data Assimilation: The Ensemble Kalman Filter*, Springer, 2006.
3. Nævdal, G., Johnson, L., Aanonsen, S., and Vefring, E., Reservoir monitoring and continuous model updating using ensemble Kalman filter, *SPE J.*, 10:66-74, 2005.
4. Wen, X.-H. and Chen, W. H., Real-time reservoir model updating using ensemble Kalman filter, *SPE Reservoir Simulation Symposium*, SPE 92991, 2005.
5. Gu, Y. Q. and Oliver, D. S., The ensemble kalman filter for continuous updating of reservoir simulation models, *J. Energy Resour. Technol.–Trans. ASME*, 128:79–87, 2006.
6. Jafarpour, B. and McLaughlin, D. B., History matching with an ensemble Kalman filter and discrete cosine parametrization, *Comput. Geosci.*, 12:227–244, 2008.
7. Lee, S. H., Malallah, A., Datta-Gupta, A., and Higdon, D., Multiscale data integration using Markov random fields, *SPE Reservoir Eval. Eng.*, 5(1):68–78, 2002.
8. Efendiev, Y., Datta-Gupta, A., Ginting, V., Ma, X., and Mallick, B., An efficient two-stage Markov chain Monte Carlo method for dynamic data integration, *Water Resour. Res.*, 41:W12423, 2005.
9. Efendiev, Y., Datta-Gupta, A., Osako, I., and Mallick, B., Multiscale data integration using coarse-scale models, *Adv. Water Resour.*, 28:303-314, 2005.

10. Liu, Y. and Gupta, H. V., Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, 43:W07401, 2007.
11. Dong, Y., Gu, Y., and Oliver, D. S., Sequential assimilation of 4D seismic data for reservoir description using the ensemble Kalman filter, *J. Pet. Sci. Eng.*, 53:83–99, 2006.
12. Skjervheim, J.-A., Evensen, G., Aanonsen, S., Ruud, B., and Johansen, T., Incorporating 4D seismic data in reservoir simulation models using an ensemble Kalman filter, *SPE J.*, 12:282–292, 2007.
13. Gosselin, O., Aanonsen, S. I., Aavatsmark, I., Cominelli, A., Gonard, R., Kolasinski, M., Ferdinandi, F., Kovacic, L., and Neylon, K., History matching using time-lapse seismic (HUTS), *SPE Annual Technical Conference and Exhibition*, Denver, SPE 84464-MS, Oct 2003.
14. Gu, Y. and Oliver, D. S., History Matching of the PUNQ-S3 reservoir model using the ensemble Kalman filter, *SPE J.*, 10:217–224, 2005.
15. Evensen, G., Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Rese.*, 99(10):14310, 162, 1994.
16. Burgers, G., van Leeuwen, P. J., and Evensen, G., Analysis scheme in the ensemble Kalman filter, *Mon. Weather Rev.*, 126(6):1719–1724, 1998.
17. Chen, Y. and Zhang, D., Data assimilation for transient flow in geologic formations via ensemble Kalman filter, *Adv. Water Resour.*, 29:1107–1122, 2006.
18. Durlofsky, L. J., Numerical calculation of equivalent grid block permeability tensors for heterogeneous porous media, *Water Resour. Res.*, 27:699–708, 1991.
19. Durlofsky, L. J., Coarse scale models of two phase flow in heterogeneous reservoirs: Volume averaged equations and their relationship to the existing upscaling techniques, *Comput. Geosci.*, 2:73–92, 1998.
20. Lawniczak, W., Hanea, R., Hemmink, A., and McLaughlin, D., Multiscale ensemble filtering for reservoir engineering applications, *Comput. Geosci.*, 13:245–254, 2009.
21. Dance, S. L., Issues in high resolution limited area data assimilation for qualitative precipitation forecasting, *Physica D*, 196:1–27, 2004.
22. Durlofsky, L. J., Behrens, R. A., Jones, R. C., and Bernath, A., Scale up of heterogeneous three-dimensional reservoir descriptions, SPE Paper 30709, 1996.
23. Durlofsky, L. J., Jones, R. C., and Milliken, W. J., A nonuniform coarsening approach for the scale up of displacement processes in heterogeneous media, *Adv. Water Resour.*, 20:335–347, 1997.
24. Wu, X. H., Efendiev, Y. R., and Hou, T. Y., Analysis of upscaling absolute permeability, *Discrete Contin. Dyn. Syst., Ser. B*, 2:185–204, 2002.
25. Efendiev, Y. R. and Durlofsky, L. J., Numerical modeling of subgrid heterogeneity in two phase flow simulations, *Water Resour. Res.*, 38(8):W1128, 2002.
26. Efendiev, Y. and Durlofsky, L., Generalized convection-diffusion model for subgrid transport in porous media, *SIAM Multiscale Model. Simul.*, 1:504–526, 2003.
27. Efendiev, Y. and Durlofsky, L., Accurate subgrid models for two-phase flow in heterogeneous reservoirs, *SPE J.*, 9:219–226, 2004.
28. Yoon, S., Malallah, A. H., Datta-Gupta, A., Vasco, D. W., and Behrens, R. A., A multiscale approach to production-data integration using streamline models, *SPE J.*, 6:182–192, 2001.
29. Deutsch, C. V. and Journel, A., *GSLIB Geostatistical Software Library and User's Guide*, Oxford Univ. Press, 1992.
30. Hendricks Franssen, H. J. and Kinzelbach, W., Real-time groundwater flow modeling with the Ensemble Kalman Filter: Joint estimation of states and parameters and the filter inbreeding problem, *Water Resour. Res.*, 44:W09408, 2008.
31. Anderson, J. L., Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter, *Physica D*, 230(1-2):99–111, 2007.
32. Hamill. T., Whitaker, J., and Synder, C., Distance-dependent filtering of background error covariance estimates in ensemble Kalman filter, *Mont. Weather Rev.*, 129:2776–2790, 2001.
33. Arroyo-Negrete, E., Devegowda, D., Datta-Gupta, A., and Choe, J., Streamline-assisted ensemble Kalman filter for rapid and continuous reservoir model updating, *SPE Reservoir Eval. Eng.*, 11(6):1046–1060, 2008.

34. Sun, A. Y., Morrisa, A., and Mohantya S., Comparison of deterministic ensemble Kalman filters for assimilating hydrogeological data, *Adv. Water Resour.*, 32(2):280–292, 2009.

35. Devegowda, D., Akella, S., Datta-Gupta, A., and Efendiev, Y., Interpretation of partitioning interwell tracer tests using EnKF with coarse-scale constraints, *SPE Reservoir Simulation Symposium*, SPE 119125-MS, The Woodlands, TX, 2–4 Feb 2009.

APPENDIX A. TWO-STEP COARSE-SCALE CONSTRAINED KALMAN FILTER ESTIMATE

From Section 3,

$$\mathcal{J}_f = \frac{1}{2}(\mathbf{\Psi} - \overline{\mathbf{\Psi}})^T (\mathbf{P}^f)^{-1} (\mathbf{\Psi} - \overline{\mathbf{\Psi}})$$

and

$$\mathcal{J}_y = \frac{1}{2}(\mathbf{y} - \mathbf{H}[\mathbf{\Psi}])^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}[\mathbf{\Psi}])$$

For notational simplicity we denote $\boldsymbol{\mu}_{\mathbf{\Psi}}$ as $\boldsymbol{\mu}$ and denote $\mathbf{P}^f$ by $\mathbf{B}$.

Step 1 (minimize $\mathcal{J}_f + \mathcal{J}_y$):

First we minimize the sum, $\mathcal{J}_1 = \mathcal{J}_f + \mathcal{J}_y$. The gradient[3] of the above quadratic cost functional with respect to (w.r.t.) $\mathbf{\Psi}$ is given by

$$\nabla_{\mathbf{\Psi}} \mathcal{J}_1 = \mathbf{B}^{-1} (\mathbf{\Psi} - \boldsymbol{\mu}) - \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}[\mathbf{\Psi}])$$

Then the minimizer $\tilde{\boldsymbol{\mu}}$ of $\mathcal{J}_1$ satisfies (we assume $\mathbf{H}$ to be linear)

$$\mathbf{B}^{-1} (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}) - \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}\tilde{\boldsymbol{\mu}}) = 0$$

By rearranging the above equation we get

$$[\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1}\mathbf{H}]\tilde{\boldsymbol{\mu}} = \mathbf{B}^{-1}\boldsymbol{\mu} + \mathbf{H}^T \mathbf{R}^{-1}\mathbf{y} \tag{20}$$

Note that the Hessian of $\mathcal{J}_1$ w.r.t. $\mathbf{\Psi}$ is given by $\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1}\mathbf{H}$, and for linear quadratic cost functionals, the Hessian inverse is equal to the error covariance matrix. Therefore, the error covariance matrix $\widetilde{\mathbf{B}}$ for $\tilde{\boldsymbol{\mu}}$ is given by

$$\widetilde{\mathbf{B}} = [\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1}\mathbf{H}]^{-1} \tag{21}$$

Step 2 (minimize $\mathcal{J}_g + \mathcal{J}_z$):

We use $\tilde{\boldsymbol{\mu}}$, $\widetilde{\mathbf{B}}$ in

$$\mathcal{J}_g = \frac{1}{2}(\mathbf{\Psi} - \tilde{\boldsymbol{\mu}})^T (\widetilde{\mathbf{B}})^{-1} (\mathbf{\Psi} - \tilde{\boldsymbol{\mu}})$$

$$\mathcal{J}_z = \frac{1}{2}(\mathbf{z} - \mathbf{U}[\mathbf{\Psi}])^T \mathbf{Q}^{-1} (\mathbf{z} - \mathbf{U}[\mathbf{\Psi}])$$

Therefore, the minimum $\hat{\boldsymbol{\mu}}$ of $\mathcal{J}_g + \mathcal{J}_z$ satisfies

$$[(\widetilde{\mathbf{B}})^{-1} + \mathbf{U}^T \mathbf{Q}^{-1}\mathbf{U}]\hat{\boldsymbol{\mu}} = (\widetilde{\mathbf{B}})^{-1}\tilde{\boldsymbol{\mu}} + \mathbf{U}^T \mathbf{Q}^{-1}\mathbf{z}.$$

Using Eqs. (21) and (20) we can rewrite the above as

$$[\underbrace{\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1}\mathbf{H}}_{(\widetilde{\mathbf{B}})^{-1}} + \mathbf{U}^T \mathbf{Q}^{-1}\mathbf{U}]\hat{\boldsymbol{\mu}} = \underbrace{\mathbf{B}^{-1}\boldsymbol{\mu} + \mathbf{H}^T \mathbf{R}^{-1}\mathbf{y}}_{\text{r.h.s. of Eq. (20)}} + \mathbf{U}^T \mathbf{Q}^{-1}\mathbf{z}$$

It is trivial to show that $\hat{\boldsymbol{\mu}}$ also satisfies

$$\nabla_{\mathbf{\Psi}}[\mathcal{J}_f + \mathcal{J}_y + \mathcal{J}_z] = 0$$

Therefore, the two-step method to obtain the final estimate $\hat{\boldsymbol{\mu}}$ gives the same results as a one-shot approach of minimizing $\mathcal{J}_f + \mathcal{J}_y + \mathcal{J}_z$.

[3]We note in passing that $\mathbf{B}$ and $\mathbf{R}$ are covariance matrices and are positive definite by construction and hence, for our derivation purposes, are formally invertible.

APPENDIX B. THE COARSE-SCALE EnKF ALGORITHM

Algorithm 1. Coarse-scale EnKF algorithm

Run the simulation model up to a particular observation time for the entire ensemble to get predicted samples: $\{\mathbf{\Psi}^{(i)}\}_{i=1}^{N_e}$, $\mathbf{A} = \left(\mathbf{\Psi}^{(1)}, \mathbf{\Psi}^{(2)}, \ldots, \mathbf{\Psi}^{(N_e)}\right)$.

- Step 1: Using measured water-cut data $\mathbf{y}$ with variance $\mathbf{R}$, get the updated ensemble: $\left\{\widetilde{\mathbf{\Psi}}^{(i)}\right\}_{i=1}^{N_e}$.

  Step 1.1—Find ensemble mean [Eq. (4a)] $\overline{\mathbf{\Psi}}$.

  Step 1.2—Subtract deviation from the mean $\mathbf{A}' = \left(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \ldots, \mathbf{b}^{(N_e)}\right)$, $\mathbf{b}^{(i)} = \mathbf{\Psi}^{(i)} - \overline{\mathbf{\Psi}}$.

  Step 1.3—Apply $\mathbf{H}$ to each column of $\mathbf{A}'$ to get $\mathbf{S} = \mathbf{H}\,\mathbf{A}'$, i.e., simply pick the water-cut deviations in $\mathbf{A}'$.

  Step 1.4—*For* $i = 1, 2, \ldots, N_e$,

  sample $\boldsymbol{\nu}^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{R})$.
  $\mathbf{y}^{(i)} = \mathbf{y} + \boldsymbol{\nu}^{(i)}$,
  $\mathbf{R}^{1/2} = \left(\boldsymbol{\nu}^{(1)}, \boldsymbol{\nu}^{(2)}, \ldots, \boldsymbol{\nu}^{(N_e)}\right)$,
  $\mathbf{D} = \left(\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, \ldots, \mathbf{d}^{(N_e)}\right)$,
  $\mathbf{d}^{(i)} = \mathbf{y}^{(i)} - \mathbf{W}_c^{(i)}$; $\mathbf{W}_c^{(i)}$ is predicted water cut for each ensemble member.

  End for

  Step 1.5—Compute SVD $\left[\mathbf{S} + \mathbf{R}^{1/2}\right] = \mathbf{X}_L \Sigma \mathbf{X}_R$.

  Get $\hat{\Sigma}$ retaining the first few singular values which explain most variability in Σ, corresponding to the left singular vectors: $\hat{\mathbf{X}}_L$.

  Step 1.6—Update ensemble: Eq. (6), $\widetilde{\mathbf{A}} = \left(\widetilde{\mathbf{\Psi}}^{(1)}, \widetilde{\mathbf{\Psi}}^{(2)}, \ldots, \widetilde{\mathbf{\Psi}}^{(N_e)}\right)$,
  $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{A}'\mathbf{S}^T \hat{\mathbf{X}}_L \hat{\Sigma}^{-2} \hat{\mathbf{X}}_L^T \mathbf{D}$.

- Step 2: Using coarse-scale data $\mathbf{z}$ with variance $\mathbf{Q}$, get the updated ensemble: $\left\{\widehat{\mathbf{\Psi}}^{(i)}\right\}_{i=1}^{N_e}$.

  Step 2.1—Compute coarse-scale ensemble prediction: $\mathbf{u}^{(i)} = \mathbf{U}\widetilde{\mathbf{\Psi}}^{(i)}$, $i = 1, 2, \ldots, N_e$.

  Step 2.2—Coarse-scale mean: $\boldsymbol{\mu}' = \dfrac{1}{N_e}\displaystyle\sum_{i=1}^{N_e} \mathbf{u}^{(i)}$.

  Step 2.3—Coarse-scale deviations: $\mathbf{S}' = \left(\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \ldots, \mathbf{s}^{(N_e)}\right)$, $\mathbf{s}^{(i)} = \mathbf{u}^{(i)} - \boldsymbol{\mu}'$.

  Step 2.4—Repeat step 1.4 using coarse-scale measurement. *For* $i = 1, 2, \ldots, N_e$,

  sample $\boldsymbol{\omega}^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{Q})$.
  $\mathbf{z}^{(i)} = \mathbf{z} + \boldsymbol{\omega}^{(i)}$,
  $\mathbf{Q}^{1/2} = \left(\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}, \ldots, \boldsymbol{\omega}^{(N_e)}\right)$,
  $\mathbf{D}' = \left(\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, \ldots, \mathbf{d}^{(N_e)}\right)$, $\mathbf{d}^{(i)} = \mathbf{z}^{(i)} - \mathbf{u}^{(i)}$.

  End for

  Step 2.5—Compute SVD $[\mathbf{S}' + \mathbf{Q}^{1/2}] = \mathbf{X}_L \Sigma \mathbf{X}_R$. Get $\hat{\Sigma}$ and $\hat{\mathbf{X}}_L$ as in step 1.5.

  Step 2.6—Compute fine-scale mean: $\boldsymbol{\mu} = \dfrac{1}{N_e}\displaystyle\sum_{i=1}^{N_e} \widetilde{\mathbf{\Psi}}^{(i)}$.

  Step 2.7—Compute fine-scale deviations: $\mathbf{A}'' = \left(\mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \ldots, \mathbf{b}^{(N_e)}\right)$, $\mathbf{b}^{(i)} = \widetilde{\mathbf{\Psi}}^{(i)} - \boldsymbol{\mu}$.

Step 2.8—Update ensemble: $\widehat{\mathbf{A}} = \left(\widehat{\mathbf{\Psi}}^{(1)}, \widehat{\mathbf{\Psi}}^{(2)}, \ldots, \widehat{\mathbf{\Psi}}^{(N_e)}\right)$,

$$\widehat{\mathbf{A}} = \widetilde{\mathbf{A}} + (\mathbf{A}'')(\mathbf{S}')^T \hat{\mathbf{X}}_L \hat{\Sigma}^{-2} \hat{\mathbf{X}}_L^T \mathbf{D}'.$$

Remark 1:
Note that steps 2.6 and 2.7 in the above algorithm approximate the intermediate fine-scale error covariance,

$$\widetilde{\mathbf{P}^f} \approx \frac{1}{N_e - 1} \mathbf{A}'' \, (\mathbf{A}'')^T.$$

Remark 2:
Steps 2.1–2.3 accomplish[4]

$$\mathbf{S}' = \mathbf{U}\mathbf{A}''.$$

Note that the above algorithm is independent of the choice of upscaling procedure, and we can use the same algorithm for different kinds of coarse-scale observed data (if available).

Remark 3:
Note that the above coarse-scale constrained EnKF algorithm can be readily extended to incorporate data at multiple coarse scales with the appropriate upscaling procedure in $\mathbf{U}$. To elaborate, if we had other independent data at a scale different from $\mathbf{z}$, we could use the estimates $\left(\{\widehat{\mathbf{\Psi}}^{(i)}\}_{i=1}^{N_e}\right)$ obtained using $\mathbf{z}$. As an intermediate solution, repeating step 2 to assimilate the data at another scale.

[4] As noted in [2], this approach of accounting for the nonlinear observations operator $\mathbf{U}$ works well as long as $\mathbf{U}$ is weakly nonlinear and a monotonic function of model variables $\mathbf{\Psi}$.

International Journal for Uncertainty Quantification, 1(1):77–98, 2011

ON A POLYNOMIAL CHAOS METHOD FOR DIFFERENTIAL EQUATIONS WITH SINGULAR SOURCES

Jae-Hun Jung & Yunfei Song*

Department of Mathematics, State University of New York at Buffalo, Buffalo, NY 14260, USA

Original Manuscript Submitted: 05/13/2010; Final Draft Received: 08/10/2010

Singular source terms in the differential equation represented by the Dirac δ-function play a crucial role in determining the global solution. Due to the singular feature of the δ-function, physical parameters associated with the δ-function are highly sensitive to random and measurement errors, which makes the uncertainty analysis necessary. In this paper we use the generalized polynomial chaos method to derive the general solution of the differential equation under uncertainties associated with the δ-function. For simplicity, we assume the uniform distribution of the random variable and use the Legendre polynomials to expand the solution in the random space. A simple differential equation with the singular source term is considered. The polynomial chaos solution is derived. The Gibbs phenomenon and the convergence of high order moments are discussed. We also consider a direct collocation method which can avoid the Gibbs oscillations on the collocation points and enhance the accuracy accordingly.

KEY WORDS: *generalized polynomial chaos, stochastic Galerkin method, singular source, Dirac δ-function, Gibbs phenomenon*

1. INTRODUCTION

Differential equations with singular source terms are commonly found in various areas of applications [1–5]. Singular source terms are defined in a highly localized regime and play a crucial role in determining the global solution of the given differential equations. It is important to capture properly such small-scale phenomenon induced by the local singular source terms and understand the interaction between the small- and large-scale solution dynamics. Singular source terms are mathematically represented by the Dirac δ-function, $\delta(x-c)$, and its derivative(s) defined in the distribution sense with a function $f(x)$, which is defined at $x=c$ such that

$$\int_{-\infty}^{\infty}\delta(x-c)f(x)dx=f(c),\ \text{and}\ \int_{-\infty}^{\infty}\delta(x-c)dx=1. \tag{1}$$

The derivatives of the δ-function are also defined in a similar way for a function $f(x)$, whose derivatives are defined at $x=c$ such that

$$\int_{-\infty}^{\infty}\delta^{'}(x-c)f(x)dx=-f^{'}(c),\int_{-\infty}^{\infty}\delta^{''}(x-c)f(x)dx=f^{''}(c),\cdots \tag{2}$$

where the superscript $'$ denotes the derivative with respect to x.

Although singular sources are defined in a compact form mathematically, various uncertainties are easily involved to define them physically. For example, it is not easy to pinpoint the location of the singular source term. The detection of the singular object is based on the physical measurement, and such measurement has errors due to the locality of the singularity. Thus the realistic model of the singular source term should include the uncertainty of the location of

*Correspond to Jae-Hun Jung, E-mail: jaehun@buffalo.edu

2152 -5080/11/$35.00

the singular source, which can be introduced by a new random variable ξ in the physical space where the δ-function is defined such as

$$\delta(x-c) \to \delta(x-\xi),$$

where ξ is a random variable replacing c in $\delta(x-c)$. Another type of uncertainties can be introduced for the amplitude, for which one can rewrite the δ-function as the following form

$$\delta(x-c) \to \eta\delta(x-c),$$

where η is a random variable. If $\eta \to 0$, then the singular source term vanishes. One can consider other types of uncertainties for the δ-function besides the location and the amplitude. In this paper we consider the case where the uncertainty exists in the location.

In many cases solutions of differential equations with singular sources are nonsmooth, singular, or discontinuous. For example, in nonlinear optics, a defect in the optical media is modeled by the singular source, and such a singular source term plays a role as a potential around which the input signals yield nonlinear reflection and scattering phenomena. These nonlinear phenomena have been investigated using nonlinear partial differential equations (PDEs) including the sine-Gordon equation

$$u_{tt} - u_{xx} + \sin(u) = \epsilon\delta(x)\sin(u), \quad -\infty < x < \infty, t > 0, u : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}, \tag{3}$$

and the nonlinear Schrödinger equation

$$i\psi_t + \psi_{xx} + \kappa|\psi|^2\psi = \epsilon\delta(x), \quad -\infty < x < \infty, t > 0, \psi : \mathbb{C} \times \mathbb{R}^+ \to \mathbb{C}. \tag{4}$$

Previous research shows that the global solutions of PDEs as given above are sensitive to the singular potential term and that the mathematical structure of the solution dynamics is rich and complex [2, 3]. The sensitivity of the global solution to the singular source term is amplified if uncertainties are involved, which makes the global solution dynamics more complex. No significant research has been conducted for the uncertainty analysis for the solution of such singularly perturbed differential equations. In this paper, as a preliminary research, we use the polynomial chaos method to analyze the solution of differential equations with the singular source term.

The polynomial chaos method was introduced by Wiener [6] and was recently much further developed by Xiu and co-workers [7–14]. The polynomial chaos method with the spectral method approach has gained great popularity these days [15–17] (see Xiu's recent book and references therein [17]). The polynomial chaos method seeks the solution in a higher dimensional polynomial space by introducing a random variable associated with the uncertainty. Then the method expands the solution as a polynomial using the orthogonal polynomials [16, 17]. The orthogonal polynomials are determined by the distribution of the random variables considered. Different distributions and the corresponding orthogonal polynomials are given in Table 1 [11, 17]. In this paper we consider the uniform distribution and use the Legendre polynomials for simplicity.

This paper is composed of the following sections. In Section 2 we consider the simple differential equation with a singular source term. The uncertainty is in the location of the singular source term. The random variable has a uniform

TABLE 1: Continuous probability density functions and the associated orthogonal polynomials [11, 17].

Distribution (PDF)	Orthogonal polynomials	Support
$(1/2)\chi_{[-1,1]}$	$L_l(x)$, Legendre polynomials	$[-1.1]$
$(1/\sqrt{2\pi})\exp(-x^2/2)$	$H_l(x)$, Hermite polynomials	$(-\infty,\infty)$
$x^{k-1}\exp(-x/\theta)/\Gamma(k)\theta^k$	$L_l(x)$, Laguerre polynomials	$[0,\infty)$
$\{[\Gamma(\alpha+\beta)]/[\Gamma(\alpha)\Gamma(\beta)]\}x^{\alpha-1}(1-x)^{\beta-1}$	$P_l^{(\alpha,\beta)}$, Jacobi polynomials	$[-1,1]$

distribution. We derive the solution using the Galerkin method and provide some convergence results. We also consider the case that the uncertainty is confined in a local regime. This assumption yields the domain decomposition method. In Section 3 we discuss the Gibbs phenomenon which exists in the solutions obtained in Section 2. In Section 4 discussions on high-order moments are given. In Section 5 we consider the simple linear advection equation with the singular source term. A similar solution is obtained for the time-dependent problem. In Section 6 we consider the collocation method to solve the same time-dependent problem considered in Section 5. The singular source term is directly projected to the collocation space. As a result, the direct projection method removes the Gibbs phenomenon in the solution. In Section 7 we provide a brief summary and remark on future work.

2. A FIRST-ORDER DIFFERENTIAL EQUATION WITH UNIFORM DISTRIBUTION, $\xi \in [-1, 1]$

First we consider the following simple differential equation for the real-valued function $u(x)$,

$$\frac{du}{dx} = \delta(x), \quad x \in [-1, 1], \quad u(-1) = 0. \tag{5}$$

The exact solution is simply given by the Heaviside function $H(x)$ which is an integral of the right-hand side of Eq. (5), $\delta(x)$. The singularity is located at the origin for Eq. (5), but we assume that there is an uncertainty in the location of the δ-function and use ξ as the random variable to denote the uncertainty of the location. Then the given differential equation becomes

$$\frac{\partial u}{\partial x} = \delta(x - \xi), \quad x \in [-1, 1], \quad u(-1, \xi) = 0. \tag{6}$$

The solution u is now a function of both x and ξ. We also assume that ξ has the uniform distribution and is defined in the same interval of x, i.e., $\xi \in [-1, 1]$, with the probability density function (PDF) given by $(1/2)\chi_{[-1,1]}(\xi)$ where $\chi(\xi)$ is the characteristic function. The assumption of the uniform distribution yields that the associated orthogonal polynomials for ξ are the Legendre polynomials, as given in Table 1. The Legendre polynomials are defined by the solution to the Sturm–Liouville problem $\{(1 - x^2)[L_l(x)]'\}' + l(l + 1)L_l(x) = 0$ with $x \in [-1, 1]$ with the orthogonality condition $\int_{-1}^{1} L_l(x)L_{l'}(x)dx = [2/(2l + 1)]\delta_{ll'}$, where the superscript $'$ denotes the derivative with respect to x and δ is the Kronecker delta.

The solution of Eq. (6) is obvious,

$$u(x, \xi) = \begin{cases} 0 & \text{if } x < \xi \\ 1 & \text{if } x \geq \xi \end{cases}. \tag{7}$$

Let $E(u)$ denote the expectation value of u, and $Var(u)$ the variance of u. These two quantities, $E(u)$ and $Var(u)$, are all functions of x only. Let $f(x) = E(u)$ and $g(x) = Var(u)$. Then we have

$$f(x) = E[u(x, \xi)] = \int_{-1}^{1} u(x, \xi)\frac{1}{2}\chi_{[-1,1]}(\xi)d\xi = \frac{1}{2}(x + 1). \tag{8}$$

Similarly,

$$g(x) = Var[u(x, \xi)] = \int_{-1}^{1} u^2(x, \xi)\frac{1}{2}\chi_{[-1,1]}(\xi)d\xi - [E(u)]^2 = \frac{1}{4}(1 - x^2). \tag{9}$$

Definition: *Let $u^{(1)}(x)$, $u^{(2)}(x)$ and $u(x, \xi)$ be defined by the Galerkin solution of Eq. (5) in Legendre polynomials, the Galerkin projection of the exact solution of Eq. (5), $H(x)$ and the polynomial chaos solution of Eq. (6), respectively. The superscripts (1) and (2) denote that the associated quantity corresponds to $u^{(1)}(x)$ and $u^{(2)}(x)$. For example, $\hat{u}^{(1)}$ and $\hat{u}^{(2)}$ are the expansion coefficients of $u^{(1)}(x)$ and $u^{(2)}(x)$, respectively.*

First we consider the function $u^{(2)}(x) = \sum_{l=0}^{\infty} \hat{u}_l^{(2)} L_l(x)$, which is the direct projection of the exact solution of Eq. (5), that is, $\hat{u}_l^{(2)}$ are given by the following equation:

$$H(x) = \sum_{l=0}^{\infty} \hat{u}_l^{(2)} L_l(x). \tag{10}$$

Lemma 1. *The expansion coefficients $\hat{u}_l^{(2)}$ in Eq. (10) are given by*

$$\hat{u}_l^{(2)} = \begin{cases} \dfrac{1}{2} & l = 0 \\ 0 & l = \text{even} \\ -\dfrac{1}{2}[L_{l+1}(0) - L_{l-1}(0)] & l = \text{odd} \end{cases}. \tag{11}$$

Proof. By multiplying each side of Eq. (10) by $L_l(x)$ and using the orthogonality of the Legendre polynomials, the expansion coefficients are given by

$$\hat{u}_l^{(2)} = \frac{2l+1}{2} \int_{-1}^{1} H(x) L_l(x) dx = \frac{2l+1}{2} \int_{0}^{1} L_l(x) dx. \tag{12}$$

If $l = 0$, it is obvious that $\hat{u}_l^{(2)} = (1/2)$. For $l \neq 0$, we use the following property of the Legendre polynomials [18]:

$$(2l+1) L_l(x) = L_{l+1}'(x) - L_{l-1}'(x), \tag{13}$$

and

$$(2l+1) \int_{-1}^{x} L_l(x) = L_{l+1}(x) - L_{l-1}(x). \tag{14}$$

Since $\int_0^1 L_l(x)dx = \int_{-1}^0 L_l(x)dx$ for $l = \text{even}$ and $\int_0^1 L_l(x)dx = -\int_{-1}^0 L_l(x)dx$ for $l = \text{odd}$, the above relations yield

$$\hat{u}_l^{(2)} = \frac{(-1)^l}{2} [L_{l+1}(0) - L_{l-1}(0)]. \tag{15}$$

Since $L_l(0) = 0$ if $l = \text{odd}$, we obtain Eq. (11).

Next we consider the function $u^{(1)}(x) = \sum_{l=0}^{\infty} \hat{u}_l^{(1)} L_l(x)$, which is the Galerkin solution of the differential equation, Eq. (5). By plugging $u^{(1)}(x)$ into the differential equation and the initial condition, we have

$$\sum_{l=0}^{\infty} \hat{u}_l^{(1)} L_l'(x) = \delta(x), \tag{16}$$

$$\sum_{l=0}^{\infty} \hat{u}_l^{(1)} L_l(-1) = 0. \tag{17}$$

Lemma 2. *The coefficients $\hat{u}_l^{(1)}$, satisfying Eqs. (16) and (17), are given by*

$$\hat{u}_l^{(1)} = \begin{cases} \dfrac{1}{2} & l = 0 \\ 0 & l = \text{even} \\ -\dfrac{1}{2}[L_{l+1}(0) - L_{l-1}(0)] & l = \text{odd} \end{cases}. \tag{18}$$

Proof. From Eq. (13) we have

$$\begin{aligned} L_2^{'}(x) - L_0^{'}(x) &= (2\cdot 1+1)L_1(x), \\ L_4^{'}(x) - L_2^{'}(x) &= (2\cdot 3+1)L_3(x), \\ L_6^{'}(x) - L_4^{'}(x) &= (2\cdot 5+1)L_5(x), \\ &\cdots \\ L_n^{'}(x) - L_{n-2}^{'}(x) &= [2\cdot(n-1)+1]L_{n-1}(x). \end{aligned} \tag{19}$$

Adding both sides of Eq. (19) all together, we have

$$L_n^{'}(x) = 1 + 5L_2(x) + 9L_4(x) + 13L_6(x) + \cdots + (2n-1)L_{n-1}(x), \tag{20}$$

for $n = \text{even}$. Similarly, we have

$$L_n^{'}(x) = 3L_1(x) + 7L_3(x) + 11L_5(x) + \cdots + (2n-1)L_{n-1}(x), \tag{21}$$

for $n = \text{odd}$. From Eqs. (20) and (21), we know that $L_n^{'}(x)$ is a linear combination of all the previous odd (if n is odd) or even (if n is even) terms with coefficients $(2k+1)$ for $L_k(x)$. By plugging Eqs. (20) and (21) into Eq. (16), we have

$$\sum_{l=0}^{\infty} \hat{u}_l^{(1)}[\cdots + (2l-1)L_{l-1}(x)] = \delta(x). \tag{22}$$

In Eq. (22), $\cdots$ means $3L_1(x) + 7L_3(x) + 11L_5(x) + \cdots + (2l-3)L_{l-2}$ for even l or $1 + 5L_2(x) + 9L_4(x) + 13L_6(x) + \cdots + (2l-3)L_{l-2}$ for odd l. Multiplying each side of Eq. (22) by $L_k(x)$ yields

$$L_k(x)\sum_{l=0}^{\infty} \hat{u}_l^{(1)}[... + (2l-1)L_{l-1}(x)] = L_k(x)\delta(x). \tag{23}$$

We then integrate the above equation over x and switch the left and right sides to obtain

$$\begin{aligned} L_k(0) &= \hat{u}_{k+1}^{(1)}\int_{-1}^{1} L_k(x)(2k+1)L_k(x)dx + \hat{u}_{k+3}^{(1)}\int_{-1}^{1} L_k(x)(2k+1)L_k(x)dx + \hat{u}_{k+5}^{(1)}\int_{-1}^{1} L_k(x)(2k+1)L_k(x)dx \\ &+ \hat{u}_{k+7}^{(1)}\int_{-1}^{1} L_k(x)(2k+1)L_k(x)dx + ... = 2\left(\hat{u}_{k+1}^{(1)} + \hat{u}_{k+3}^{(1)} + \hat{u}_{k+5}^{(1)} + \hat{u}_{k+7}^{(1)} + ...\right), \end{aligned} \tag{24}$$

where we used the orthogonality condition of the Legendre polynomials. Eq. (24) also reads

$$L_{k+2}(0) = 2\left(\hat{u}_{k+3}^{(1)} + \hat{u}_{k+5}^{(1)} + \hat{u}_{k+7}^{(1)} + \hat{u}_{k+9}^{(1)}...\right). \tag{25}$$

Subtracting Eq. (24) from Eq. (25) yields

$$\hat{u}_{k+1}^{(1)} = -\frac{1}{2}[L_{k+2}(0) - L_k(0)]. \tag{26}$$

Now consider the boundary condition. Since $\hat{u}_{k+1}^{(1)}$ vanishes if k is odd, the boundary condition becomes

$$-\sum_{l=0,\text{odd}}^{\infty} \hat{u}_l^{(1)} = \lim_{k,\text{odd}\to\infty} \frac{1}{2}[L_{k+1}(0) - L_1(0)] = 0. \tag{27}$$

This completes the proof.

Finally we consider $u(x,\xi)$, which is the solution of the stochastic differential equation, Eq. (6),

$$u(x,\xi) = \sum_{l=0}^{\infty} \hat{u}_l(x) L_l(\xi), \tag{28}$$

where the expansion coefficients $\hat{u}_l$ are functions of x. Plugging $u(x,\xi)$ into the differential equation yields

$$\sum_{l=0}^{\infty} \hat{u}_l^{'}(x) L_l(\xi) = \delta(x-\xi), \tag{29}$$

where the superscript $'$ denotes the derivative with respect to x. To consider $u(x,\xi)$, let us first consider the general case where ξ is defined in the subinterval of x.

Domain decomposition: $\xi \in (-\epsilon, \epsilon)$. Assume that the location of the δ-function is confined in a small region $\xi \in (-\epsilon, \epsilon)$, $0 < \epsilon \ll 1$. The solution for Eq. (6) is then given by

$$u(x,\xi) = \begin{cases} 0 & \text{if } x < \xi \\ 1 & \text{if } x \geq \xi \end{cases}, \tag{30}$$

where $\xi \in (-\epsilon, \epsilon)$. In the interval $x \in [-\epsilon, \epsilon]$, the expectation value is

$$f(x) = E[u(x,\xi)] = \frac{1}{2\epsilon}(x+\epsilon). \tag{31}$$

Similarly, the variance is

$$g(x) = Var[u(x,\xi)] = \frac{1}{4\epsilon^2}(\epsilon^2 - x^2). \tag{32}$$

Thus, for a fixed ϵ, the expectation value and variance are given by

$$f(x) = \begin{cases} 0 & x \in [-1, -\epsilon) \\ \dfrac{1}{2\epsilon}(x+\epsilon) & x \in [-\epsilon, \epsilon] \\ 1 & x \in (\epsilon, 1] \end{cases}, \tag{33}$$

and

$$g(x) = \begin{cases} 0 & x \in [-1, -\epsilon) \\ \dfrac{1}{4\epsilon^2}(\epsilon^2 - x^2) & x \in [-\epsilon, \epsilon] \\ 0 & x \in (\epsilon, 1] \end{cases}. \tag{34}$$

For any ϵ, we have

$$|f(x)| = |E[u(x,\xi)]| = \left| \frac{1}{2\epsilon}(x+\epsilon) \right| \leq \frac{|x| + |\epsilon|}{2\epsilon} \leq 1, \tag{35}$$

$$|g(x)| = |Var[u(x,\xi)]| = \left| \frac{1}{4\epsilon^2}(\epsilon^2 - x^2) \right| \leq \frac{\epsilon^2}{4\epsilon^2} = \frac{1}{4}, \tag{36}$$

which shows the expectation value and variance are bounded although the PDF $[(1/2)\chi_{[-\epsilon,\epsilon]}(\xi) = (1/2\epsilon)]$ diverges as $\epsilon \to 0$. We know that $x \to 0$ as $\epsilon \to 0$, and we obtain the expectation value and variance at $x = 0$ by letting $\epsilon \to 0$. Also by Eqs. (35) and (36) we have

$$f(0) = E[u(0,\xi)] \equiv \frac{1}{2} \quad \text{and} \quad g(0) = Var[u(0,\xi)] \equiv \frac{1}{4}. \tag{37}$$

These values are the same as those for $\xi \in [-1,1]$. Thus, we know that if $x \to 0$, the expectation value and the variance are the same for any value of $\xi \in (-\epsilon, \epsilon)$.

The assumption that $\xi \in (-\epsilon, \epsilon)$ breaks the original differential equation into three equations in three regions, (1) $x \in I = [-1, -\epsilon]$, (2) $x \in II = (-\epsilon, \epsilon]$, and (3) $x \in III = (\epsilon, 1]$.

Interval I, $x \in [-1, -\epsilon]$: In this interval, the δ-function is absent and the differential equation and the boundary condition are given by

$$\frac{du}{dx} = 0, \quad u(-1) = 0,$$

and the solution is simply

$$u(x) = 0, \quad u(\epsilon) = 0. \tag{38}$$

Interval II, $x \in (-\epsilon, \epsilon)$: In this interval, the δ-function exists and the equation is given by

$$\frac{\partial u(x,\xi)}{\partial x} = \delta(x - \eta),$$

where $\eta \in (-\epsilon, \epsilon)$. Then we seek a solution $u(x, \eta)$ as

$$u(x,\eta) = \sum_{l=0}^{\infty} \hat{u}_l(x) L_l[\xi(\eta)], \tag{39}$$

where $\xi = (\eta/\epsilon)$ and $\xi \in [-1, 1]$.

Lemma 3. *The expansion coefficients $\hat{u}_l(x)$ in Eq. (39) are given by*

$$\hat{u}_l(x) = \begin{cases} \dfrac{1}{2\epsilon}(x+\epsilon) & l = 0 \\ \dfrac{1}{2}\left[L_{l+1}\left(\dfrac{x}{\epsilon}\right) - L_{l-1}\left(\dfrac{x}{\epsilon}\right)\right] & l \neq 0 \end{cases}. \tag{40}$$

Furthermore, the boundary value $u(x = \epsilon, \eta)$ is unity for any value of η, i.e.,

$$u(\epsilon, \eta) = 1. \tag{41}$$

Proof. By plugging Eq. (39) into the differential equation and using the orthogonality of $L_l(\xi)$ we obtain

$$\frac{2}{2k+1}\frac{d\hat{u}_k(x)}{dx} = \int_{-1}^{1} \delta(x - \epsilon\xi) L_k[\xi(\eta)] d\xi = \frac{1}{\epsilon} L_k\left(\frac{x}{\epsilon}\right). \tag{42}$$

The boundary condition at $x = -\epsilon$ is obtained by the solution at $x = -\epsilon$ in interval I,

$$\sum_{l=0}^{\infty} \hat{u}_l(-\epsilon) L_l[\xi(\eta)] = 0. \tag{43}$$

Thus, $\hat{u}_l(-\epsilon) = 0$ for all $l = 0, 1, 2, \cdots$. Using this boundary condition, we obtain

$$\hat{u}_0 = \frac{1}{2\epsilon}(x + \epsilon), \tag{44}$$

if $k = 0$. If $k \neq 0$, we have

$$\hat{u}_k(x) = \frac{2k+1}{2} \cdot \frac{1}{\epsilon} \int_{-\epsilon}^{x} L_k\left(\frac{y}{\epsilon}\right) dy = \frac{1}{2}\left[L_{k+1}\left(\frac{x}{\epsilon}\right) - L_{k-1}\left(\frac{x}{\epsilon}\right)\right], \tag{45}$$

where we used Eq. (14). The boundary value of $u(x, \eta)$ at $x = \epsilon$ is

$$u(\epsilon, \eta) = \frac{1}{2\epsilon} + \sum_{l=1}^{\infty} \frac{1}{2}[L_{l+1}(1) - L_{l-1}(1)] L_l\left(\frac{\eta}{\epsilon}\right) = 1. \tag{46}$$

From lemma 3, we know that the mean value of $u(x,\eta)$ in this interval is given by

$$E[u(x,\eta)] = \frac{1}{2\epsilon}(x+\epsilon), \tag{47}$$

which is $\hat{u}_0(x)$. If $x \to 0$, we confirm that

$$\lim_{x\to 0} E[u(x,\eta)] = \hat{u}(0) = \frac{1}{2}.$$

Interval III, $x \in (-\epsilon, 1]$: Since there is no δ-function in this interval, using the boundary value of $u(\epsilon,\eta) = 1$, the solution in $x \in (\epsilon, 1]$ is given by $u(x) = 1$. It is easy to show that if $\epsilon \to 0$, then we have

$$u(x,\eta) \to u^{(1)}(x), \text{ or } u^{(2)}(x). \tag{48}$$

Using lemma 3, we have the following corollary for $\xi[-1,1]$.

Corollary 4. *The expansion coefficients* $\hat{u}_l(x)$ *are given by*

$$\hat{u}_0(x) = \frac{1}{2}(x+1), \quad \hat{u}_k(x) = \frac{1}{2}[L_{k+1}(x) - L_{k-1}(x)], \tag{49}$$

and

$$u(x,0) = \frac{1}{2} - \sum_{l=1,\text{odd}}^{\infty} \frac{1}{2}[L_{l+1}(0) - L_{l-1}(0)]L_l(x). \tag{50}$$

Proof. From lemma 3, for $\epsilon \to 1$, we have Eqs. (49) and (50). Furthermore, by plugging $\xi = 0$ and equations in Eq. (49) into Eq. (28), we obtain

$$u(x,0) = \frac{1}{2}(x+1) + \sum_{l=2,\text{even}}^{\infty} \frac{1}{2}[L_{l+1}(x) - L_{l-1}(x)]L_l(0). \tag{51}$$

It is a simple exercise to show that Eq. (51) becomes

$$u(x,0) = \frac{1}{2} - \sum_{l=1,\text{odd}}^{\infty} \frac{1}{2}[L_{l+1}(0) - L_{l-1}(0)]L_l(x).$$

Here note that the first coefficient, $\hat{u}_0(x) = (1/2)(x+1)$, is the same as the expectation value of $u(x,\xi)$ in Eq. (8).

Remark: *Equations (50) and (51) are equivalent. They may, however, become different if they are truncated with the finite* N *in their given forms. For Eq. (51), since* $L_{l+1}(x) = L_{l-1}(x)$ *at* $x = \pm 1$ *if* l *is even, we know that at* $x = \pm 1$,

$$u(x,0) = \frac{1}{2}(x+1) = \begin{cases} 1 & x = 1 \\ 0 & x = -1 \end{cases},$$

which are the boundary values and they are determined regardless of how many terms are used in the series. For Eq. (50), since $u(x,0) = 1$ *at* $x = 1$, *and* $u(x,0) = 0$ *at* $x = -1$, *the following should be*

$$\sum_{l=1,\text{odd}}^{N} \frac{1}{2}[L_{l+1}(0) - L_{l-1}(0)]L_l(x) = \begin{cases} -\frac{1}{2} & x = 1 \\ \frac{1}{2} & x = -1 \end{cases},$$

which is only true if $N \to \infty$. *Thus we use Eq. (51) for the computation in the following sections.*

Figure 1 shows the expansion coefficients $u_l(x)$, Eq (49). The top figure shows $u_l(x)$ for $l = 0, \cdots, 9$ and the bottom figure for $l = 10, \cdots, 40$.

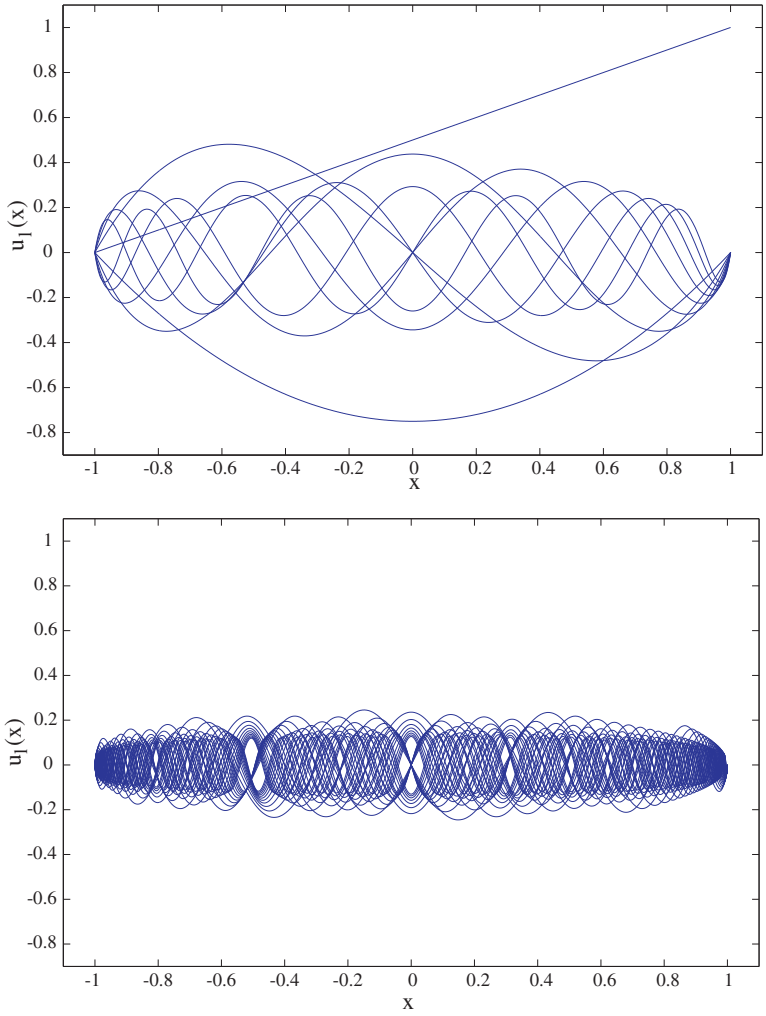

FIG. 1: Expansion coefficients $u_l(x)$. Top: $u_l(x)$ for $l = 0, \cdots, 9$. Bottom: $u_l(x)$ for $l = 10, \cdots, 40$.

Theorem 5.

$$u^{(1)}(x) = u^{(2)}(x) = u(x, \xi = 0).$$

Furthermore, $u_N(x, \xi)$ *converges to* $u(x, \xi)$ *at* $\xi = 0$*, i.e.,*

$$\lim_{N\to\infty} \|u(x, 0) - u_N(x, 0)\|_\infty = 0. \tag{52}$$

Proof. From lemmas 1 and 2 and corollary 4, we know that all the coefficients of $u^{(1)}(x)$, $u^{(2)}(x)$, and $u(x, \xi = 0)$ are the same.

Using the recurrence relation $(n + 1)L_{n+1}(x) = (2n + 1)xL_n(x) - nL_{n-1}(x)$, we have

$$(n + 2)L_{n+2}(0) = -(n + 1)L_n(0), \tag{53}$$

which yields

$$L_{2n}(0) = (-1)^n \frac{(2n)!}{4^n(n!)^2}. \tag{54}$$

Then we let $\hat{v}_k(x)$ be defined as

$$\hat{v}_{2k+1} = \frac{1}{2}[L_{2k}(0) - L_{2k+2}(0)] = \frac{1}{2}(-1)^k \frac{(2k)!}{4^k(k!)^2}\left[1 + \frac{(2k+2)(2k+1)}{4(k+1)^2}\right] \approx (-1)^k \frac{(2k)!}{4^k(k!)^2}. \tag{55}$$

Using the Stirling formula $n! \sim \sqrt{2\pi n}(n/e)^n$ and $(2n)! \sim \left[\sqrt{4\pi n}(2n/e)^{2n}\right]$, we have

$$\lim_{k\to\infty} \frac{(2k)!}{4^k(k!)^2} = \lim_{k\to\infty} \frac{\sqrt{4\pi k}\left(\frac{2k}{e}\right)^{2k}}{4^k\left[\sqrt{2\pi k}\left(\frac{k}{e}\right)^k\right]^2} = \lim_{k\to\infty} \frac{1}{\sqrt{4\pi k}} = 0. \tag{56}$$

Thus, the following series converges

$$\sum_{k=0}^{\infty} \hat{v}_{2k+1} = \sum_{k=0}^{\infty} (-1)^k \frac{(2k)!}{4^k(k!)^2}, \tag{57}$$

and we have

$$\|u(x,0) - u_N(x,0)\|_\infty = \left\| \sum_{k=N/2}^{\infty} \hat{v}_{2k+1} L_{2k+1}(x) \right\| \le \left\| \sum_{k=N/2}^{\infty} \hat{v}_{2k+1} \right\|_\infty \to 0 \text{ for } N \to \infty. \tag{58}$$

From theorem 5, we know that the stochastic solution $u(x,\xi)$ matches the deterministic solution well, particularly if the singularity is located at $x = 0$.

Remark: *Theorem 5 can be extended to the more general case that*

$$\frac{du(x)}{dx} = \delta(x - c), \quad x \in [-1, 1],$$

where c is the real constant $c \in [-1, 1]$. Then solutions $u^{(1)}(x;c)$ and $u^{(2)}(x;c)$ are the same as $u(x,\xi)$ for any $\xi = c$.

This can be easily shown using the properties of the Legendre polynomials. First, for $H(x-c) = \sum_{l=0}^{N} \hat{u}_l^{(2)} L_l(x)$, the coefficients are given by

$$\hat{u}_l^{(2)} = \frac{2l+1}{2} \int_c^1 L_l(x)dx. \tag{59}$$

By the Galerkin projection, we get similar results as lemma 2,

$$\hat{u}_0^{(1)} = \frac{1-c}{2}, \quad \hat{u}_l^{(1)} = \frac{1}{2}[L_{l-1}(c) - L_{l+1}(c)], \tag{60}$$

for any $l > 0$. To prove Eqs. (59) and (60) are equal, we use the identity formula (14). Setting $x = 1$ and $x = c$ in Eq. (14), respectively, we have

$$(2l+1)\int_{-1}^{1} L_l(x) = L_{l+1}(1) - L_{l-1}(1) = 0, \tag{61}$$

$$(2l+1)\int_{-1}^{c} L_l(x) = L_{l+1}(c) - L_{l-1}(c). \tag{62}$$

Subtracting Eq. (62) from Eq. (61) yields the equation implying that Eqs. (59) and (60) are equal. Also, the boundary condition $\sum_{l=0}^{N} \hat{u}_l^{(1)} L_l(-1) = 0$ is obtained by plugging Eq. (60) into this formula. The coefficients from the polynomial chaos method for any c are obtained in as similar way as corollary 4 by just replacing 0 with c in Eq. (50). After some simple algebraic calculations, we can show that the coefficients by the polynomial chaos method are equal to those by the previous two methods.

Now we consider the convergence of $u(x,\xi)$ for any ξ. That is, we want to show

$$\lim_{N\to\infty} \|u(x,\xi) - u_N(x,\xi)\|_\infty = 0, \quad \forall \xi \in [-1,1]. \tag{63}$$

Using corollary 4 we have

$$\begin{aligned}\|u(x,\xi) - u_N(x,\xi)\|_\infty &= \left\| \sum_{l=N+1}^{\infty} \hat{u}_l(x) L_l(\xi) \right\|_\infty = \left\| \sum_{l=N+1}^{\infty} \frac{1}{2}(L_{l+1}(x) - L_{l-1}(x)) L_l(\xi) \right\|_\infty \\ &\le \frac{1}{2} \sum_{l=N+1}^{\infty} |L_{l+1}(x) - L_{l-1}(x)|_\infty ,\end{aligned} \tag{64}$$

where we used $|L_l(\xi)| \le 1$. Here we do not provide the convergence analytically, but instead we show the numerical result. Define $R_1(n)$,

$$R_1(n) = |L_{l+1}(x) - L_{l-1}(x)|_\infty ,$$

and the remainder

$$R_2(n, N_\infty) = \sum_{l=n+1}^{N_\infty} |L_{l+1}(x) - L_{l-1}(x)|_\infty .$$

For the numerical calculation of $R_1(n)$ and $R_2(n, N_\infty)$, we use $N_\infty = 6000$. Figure 2 shows the decay of $R_1(n)$ (blue solid line) and $R_2(n)$ (black solid line) with n in logarithmic scale. The figure shows that $R_1(n)$ decays with a rate of about $\sim n^{-4.95}$. The red line in the figure is a reference line which decays $\sim n^{-4.95}$. With this decay rate, we know that the series $\sum_{n=1}^{\infty} |L_{l+1}(x) - L_{l-1}(x)|_\infty$ will converge. Thus, the remainder $R_2(n, N_\infty)$ will decay as $n \to \infty$ and $N_\infty \to \infty$ for $n < N_\infty$, i.e.,

$$\lim_{n, N_\infty \to \infty} R_2(n, N_\infty) = 0.$$

The black solid line shows the decay of $R_2(n, N_\infty)$ with $N_\infty = 6000$. The figure implies that due to the decay property of $R_1(n)$, the remainder $R_2(n, \infty)$ will also decay to zero as $n \to \infty$, but the decay rate is only algebraic. That is, we know that $u_N(x,\xi)$ converges to $u(x,\xi)$, but convergence is slow because of the existence of the discontinuity at $x = \xi$.

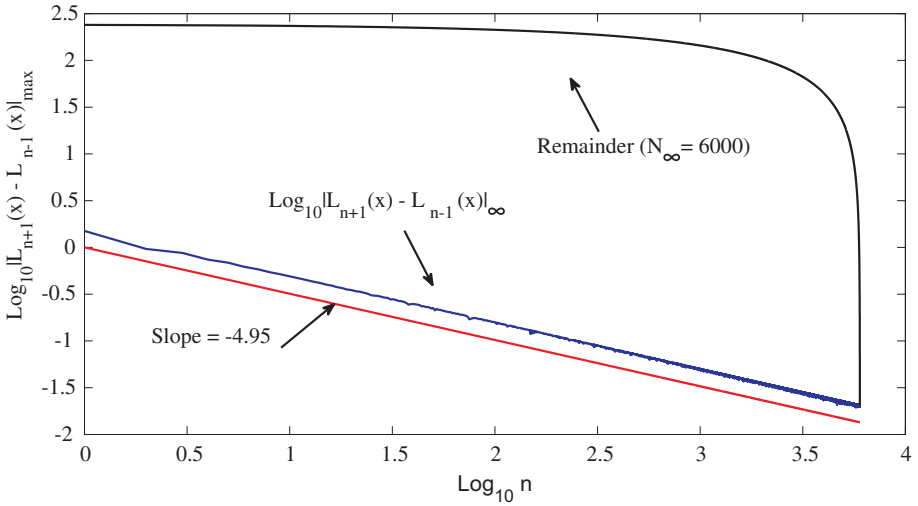

FIG. 2: Decay of $R_1(n)$ and $R_2(n, N_\infty)$ with $N_\infty = 6000$. The red solid line is the reference line which decays as $n^{-4.95}$.

3. GIBBS PHENOMENON

The solutions obtained in the previous section yield the Gibbs phenomenon. The Gibbs phenomenon is commonly found in high-order approximations of discontinuous functions with the spectral method [19, 20]. The exact solution of Eq. (5) is the Heaviside function with $H(0) = 1$ and $\lim_{x\to 0^-} H(x) = 0$. As we already saw, all solutions obtained in the previous section have the expectation value of $1/2$ at $x = 0$. Thus all solutions converge to $H(x)$ at every point x except $x = 0$. This appears as the Gibbs oscillations in the partial sum of each solution near $x = 0$.

Figure 3 shows the partial sum solution of $u(x, \xi = 0)$ (left) and $u(x, \xi)$ (right) for $N = 40$. The left figure shows the solution when $\xi = 0$. As shown in the figure, the solution is oscillatory near the discontinuity $x = 0$. The right figure shows the collection of solutions for every ξ and x. As shown in the figure, $u(x, \xi)$ are oscillatory near $x = \xi$. Figure 4 shows the variance and the mean of $u(x, \xi)$. The top figure shows the computed variance of $u(x, \xi)$ with 501 points of x and ξ for $N = 10$ (blue solid line), $N = 20$ (green), $N = 40$ (purple), and the theoretical variance of $u(x, \xi)$, $(1 - x^2)/4$ (red). As the figure shows, the variance approaches the exact variance as N increases, but the convergence is slow. The slow convergence is due to the fact that the variance is computed using every term in the series of the solution. As the series converges slowly, the variance also converges slowly. The bottom figure shows the error between the computed mean of $u(x, \xi)$ and the exact mean $(1 + x)/2$ in logarithmic scale using 5001 uniform points for $N = 4, 6, 10$. For the numerical integration, we used the Simpson's rule. The figure shows that the pointwise errors of the mean value are close to machine accuracy for the small value of N. This is because the first mode is the mean and the rest of the terms are canceled out. As N increases, the pointwise errors increase, which results from the incomplete numerical cancellations of high modes due to round-off errors.

4. HIGH-ORDER MOMENTS OF $U(X, \xi)$

With the uniform distribution, it is easy to show that the variance is given by

$$Var[u(x, \xi)] = \sum_{l=1}^{\infty} \frac{\hat{u}_l^2(x)}{2l + 1}, \tag{65}$$

where one should note that the index l runs from 1. In general, all the terms of $\hat{u}_l(x)$ are involved for the computation of the variance, as shown in the above equation and Fig. 4. It is, however, interesting to observe that the variance in our case is simply given by the second mode of $\hat{u}_l(x)$,

$$Var[u(x, \xi)] = -\frac{\hat{u}_1(x)}{3} = \frac{1}{4}(1 - x^2). \tag{66}$$

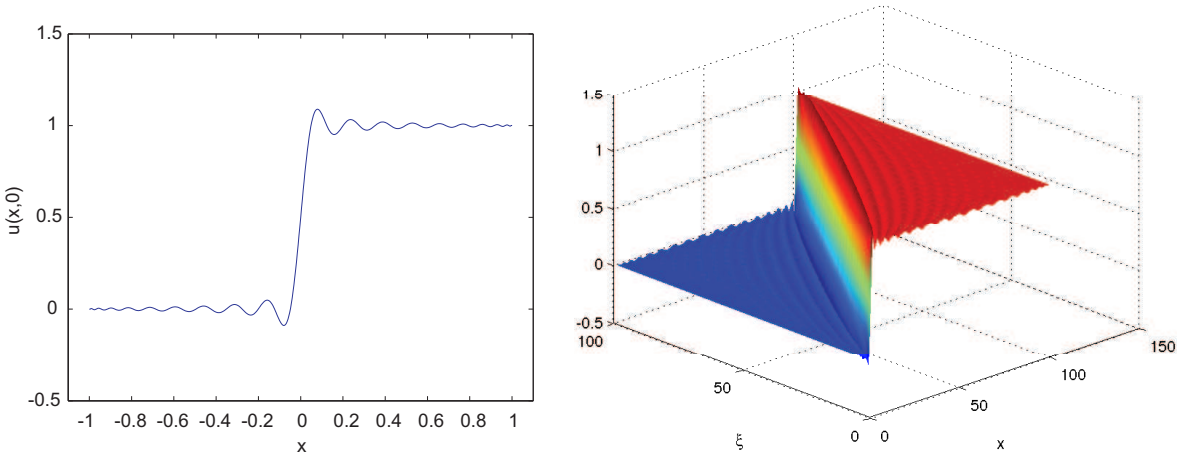

FIG. 3: Left: $u(x, \xi)$ for $\xi = 0$. Right: $u(x, \xi)$. For these figures, $N = 40$ is used.

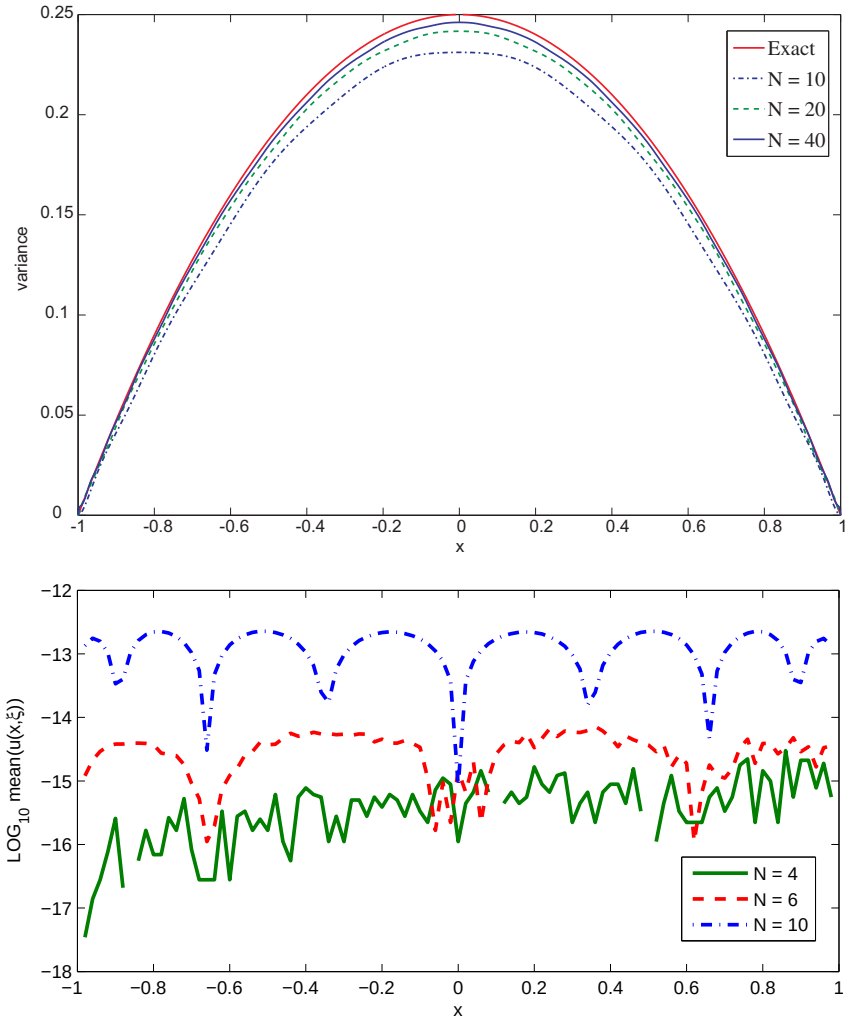

FIG. 4: Top: Variance of $u(x,\xi)$ with the exact variance $(1-x^2)/4$ (red), variances for $N=10$ (blue), $N=20$ (green), and $N=40$ (purple). Bottom: Pointwise errors between the exact mean $(x+1)/2$ and the computed mean for $N=4$ (green), 6 (red), 10 (blue).

That is, as the mean of $u(x,\xi)$, the variance can be determined exactly once $\hat{u}_1(x)$ is found. This implies that the slow convergence of the variance found in Fig. 4 can be resolved as the variance is obtained instantly. To understand this interesting aspect, we need to show the following:

$$\sum_{l=1}^{\infty} \frac{\hat{u}_l^2(x)}{2l+1} = \frac{1}{4}(1-x^2), \tag{67}$$

where $u_l(x) = (1/2)[L_{l+1}(x) - L_{l-1}(x)]$. To prove Eq. (67), first we plug Eq. (14) into the left-hand side (LHS) of Eq. (67). Then the LHS becomes

$$\text{LHS} = \frac{1}{4}\sum_{l=1}^{\infty}(2l+1)\left[\int_{-1}^{x} L_l(\mu)d\mu\right]^2.$$

For the proof we use the well-known property that the Legendre polynomials are complete, that is, $\left\{\sqrt{[(2l+1)/2]}\, L_l(x)\right\}_{l=0}^{\infty}$ are complete and orthonormal [21]. The completeness condition yields

$$\int_{-1}^{x} 1^2 d\mu = \sum_{l=0}^{\infty} \left[\int_{-1}^{x} 1 \cdot \sqrt{\frac{2l+1}{2}} L_l(\mu) d\mu\right]^2.$$

Thus,

$$x+1 = \sum_{l=0}^{\infty} \frac{2l+1}{2} \left[\int_{-1}^{x} L_l(\mu) d\mu\right]^2.$$

Using $L_0(x) = 1$ we obtain [21]

$$1-x^2 = \sum_{l=1}^{\infty} (2l+1) \left[\int_{-1}^{x} L_l(x) d\mu\right]^2.$$

This completes the proof. This special result is due to the following relation:

$$E[u^n(x,\xi)] = E[u(x,\xi)], \quad \text{for any} \quad n = 0, 1, \cdots. \tag{68}$$

Since the exact solution is $H(x-\xi)$, it is simple to show that

$$E[u^n(x,\xi)] = E[H^n(x,\xi)] = E[H(x,\xi)] = E(u).$$

The fact that the mean of any power of $u(x,\xi)$ is the same as the mean of $u(x,\xi)$ yields the following property. Let $E[u(x,\xi)] = \bar{u}$ and $v = -\bar{u}$. Then for $n = 0, 1, \cdots$, we have

$$E[(u-\bar{u})^n] = v^n(1+v) - v(1+v)^n. \tag{69}$$

It is easy to show Eq. (69),

$$\begin{aligned} E[(u-\bar{u})^n] &= E\left[\sum_{k=0}^{n} \binom{n}{k} u^k (-\bar{u})^{n-k}\right] = (-\bar{u})^n + \sum_{k=1}^{n} \binom{n}{k} E(u^k)(-\bar{u})^{n-k} = (-\bar{u})^n + (\bar{u}) \sum_{k=1}^{n} \binom{n}{k} (-\bar{u})^{n-k} \\ &= v^n - v \sum_{k=1}^{n} \binom{n}{k} (v)^{n-k} = v^n(1+v) - v(1+v)^n, \end{aligned} \tag{70}$$

where we used Eq. (68) and $v = -\bar{u}$. Equation (69) yields interesting results about the high-order moments of u. For example,

$$E[(u-\bar{u})^n] = (-1)^{n-1} \frac{u_{n-1}}{2(n-1)+1}, \quad n = 2, 3. \tag{71}$$

The second moment is the variance and the third moment is related to the skewness. Thus, we know that the first three moments (the mean, variance, and skewness) are obtained exactly by the first three modes of $u_l(x)$ for our case.

Figure 5 shows $E[(u-\bar{u})^n]$ with different $n = 1, \cdots, 50$. The left figure shows $E[(u-\bar{u})^n]$ for $n = 1, \cdots, 20$ and the right for $n = 21, \cdots, 50$. If $n = 1$, $E[(u-\bar{u})^n] = 0$. As n increases, the maximum value of $E[(u-\bar{u})^n]$ decreases in the figures. Note the different scale in the left and right figures.

5. TIME-DEPENDENT LINEAR ADVECTION EQUATION WITH UNCERTAINTY

We consider the time-dependent problem with a singular source term

$$\begin{aligned} &u_t + u_x = \delta(x), \quad u : [-1,1] \times \mathbb{R}^+ \to \mathbb{R}, \quad t > 0 \\ &u(x,0) = g(x), \quad t = 0 \\ &u(-1,t) = h(t) \quad t > 0. \end{aligned} \tag{72}$$

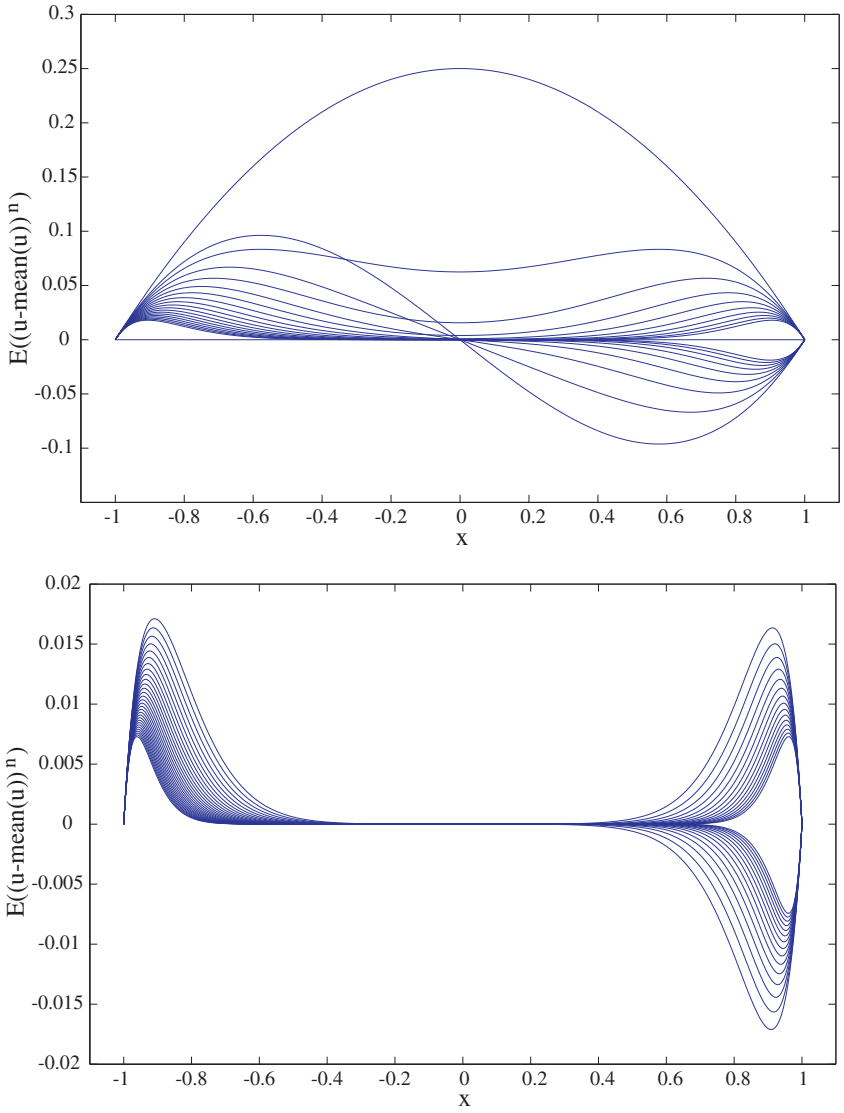

FIG. 5: Moments, $E[(u-\bar{u})^n]$. Top: $n = 1, \cdots, 20$. Bottom: $n = 21, \cdots, 50$.

If the boundary condition is homogeneous, i.e., $u(-1,t) = 0$, the solution $u(x,t)$ goes to the steady-state solution which is the Heaviside function, $H(x)$. We consider the case that the location of the singular source has an uncertainty ξ as in the previous sections, that is,

$$u_t + u_x = \delta(x-\xi), \tag{73}$$

where $u = u(x,t,\xi)$. We assume that $\xi \in (-1,1)$ with the uniform PDF and

$$u(x,t,\xi) = \sum_{l=0}^{\infty} \hat{u}_l(x,t) L_l(\xi). \tag{74}$$

By plugging Eq. (74) and using the orthogonality condition of the Legendre polynomials, we obtain

$$\frac{2}{2l+1}\left[\frac{\partial}{\partial t}\hat{u}_l(x,t) + \frac{\partial}{\partial x}\hat{u}_l(x,t)\right] = L_l(x). \tag{75}$$

If $l = 0$, we have

$$\frac{\partial}{\partial t}\hat{u}_0(x,t) + \frac{\partial}{\partial x}\hat{u}_0(x,t) = \frac{1}{2}. \tag{76}$$

Then the solution of Eq. (76) is given by

$$\hat{u}_0(x,t) = \hat{u}_0(x_0, t=0) + \frac{1}{2}x + C, \tag{77}$$

where C is the integration constant and $x_0 = x - t$. To determine the integration constant C, we use the given boundary and initial conditions. From the boundary condition $u(-1,t,\xi) = h(t)$, we have

$$\hat{u}_l(-1,t) \quad = \quad (2l+1)h(t)\delta_{l0} = \begin{cases} h(t) & l = 0 \\ 0 & l \neq 0 \end{cases}. \tag{78}$$

Similarly, using the given initial condition

$$\sum_{l=0}^{\infty} \hat{u}_l(x,0)L_l(\xi) = g(x),$$

we obtain

$$\hat{u}_l(x,0) \quad = \quad (2l+1)g(x)\delta_{l0} = \begin{cases} g(x) & l = 0 \\ 0 & l \neq 0 \end{cases}. \tag{79}$$

Using Eqs. (78) and (79), we obtain

$$\hat{u}_0(x,t) \quad = \quad g(x-t) + \frac{1}{2}x + C = g(x-t) + \frac{1}{2}x + h(t) + \frac{1}{2} - g(-1-t). \tag{80}$$

If $l \neq 0$, by using the orthogonality condition we obtain

$$\hat{u}_l(x,t) = \hat{u}_l(x_0, t=0) + \frac{2l+1}{2}\int_{-1}^{x} L_l(y)dy = \frac{1}{2}\left[L_{l+1}(x) - L_{l-1}(x)\right], \tag{81}$$

where we used Eqs. (18) and (79). Thus, the general solution of the stochastic equation (73) is given by

$$u(x,t,\xi) = g(x-t) - g(-1-t) + \frac{1}{2}(x+1) + h(t) + \sum_{l=1}^{\infty}\frac{1}{2}\left[L_{l+1}(x) - L_{l-1}(x)\right]L_l(\xi). \tag{82}$$

If $g(x) = 0 = h(t)$, then we obtain

$$u(x,t,\xi) = \frac{1}{2}(x+1) + \sum_{l=1}^{\infty}\frac{1}{2}\left[L_{l+1}(x) - L_{l-1}(x)\right]L_l(\xi).$$

This is the Legendre expansion of $H(x)$, and we know that $u(x,t) \to H(x)$ as $t \to \infty$.

To consider the numerical approximation of the solution, we use the Legendre polynomials both in x and ξ,

$$u(x,t,\xi) = \sum_{l=0}^{\infty}\hat{u}_l(x,t)L_l(\xi) = \sum_{l=0}^{\infty}\left[\sum_{k=0}^{\infty}\hat{v}_k^l(t)L_k(x)\right]L_l(\xi). \tag{83}$$

We seek the truncated sum of Eq. (83) for the numerical solution

$$u_{NM}(x,t,\xi) = \sum_{l=0}^{N}\left[\sum_{k=0}^{M}\hat{v}_k^l(t)L_k(x)\right]L_l(\xi). \tag{84}$$

For simplicity, we assume that $N = M$. Multiplying each side of Eq. (75) by $L_{l'}(x)$, $l' = 0, \cdots, N$ and using the integration by parts, we obtain

$$\frac{\partial}{\partial t}\int_{-1}^{1} \hat{u}_l L_{l'}(x)dx + \hat{u}_l(1,t) - \hat{u}_l(-1,t)L_{l'}(-1) - \int_{-1}^{1} \hat{u}_l L'_{l'}(x)dx = \delta_{ll'}, \tag{85}$$

where we use $L_{l'}(1) = 1, \forall l'$. We then plug the following relation into Eq. (85),

$$\hat{u}_l(x,t) = \sum_{k=0}^{N} \hat{v}_k^l(t) L_k(x).$$

For the given l, using the boundary condition and the properties of the Legendre polynomials, we obtain

$$\frac{2}{2l'+1}\frac{d\hat{v}_{l'}^l}{dt} + \sum_{k=0}^{N} \hat{v}_k^l(t) - h(t)\delta_{l0}(-1)^{l'} - \sum_{k=0}^{N} \hat{v}_k^l \int_{-1}^{1} L_k(x)L'_{l'}(x)dx = \delta_{ll'}. \tag{86}$$

Define the column vectors $\vec{v}^{\,l}$ and $\vec{b}_1$, whose ith elements are $\hat{v}_i^l$ and $(-1)^i$, and define the matrix $\mathbf{b}_2^l$, whose ith column has the element δ_{il} for $i = 0, \cdots, N$. Also, define the matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, whose ij elements are $A_{ij} = (2i+1)/2$, $B_{ij} = [2/(2i+1)]\delta_{ij}$, and $C_{ij} = \int_{-1}^{1} L'_i(x)L_j(x)dx$, for $i, j = 0, \cdots, N$, respectively. Then for given l, Eq. (86) becomes

$$\frac{d\vec{v}^{\,l}}{dt} = \left(\mathbf{B}^{-1}\mathbf{C} - \mathbf{A}\right)\vec{v}^{\,l} - h(t)\delta_{l0}\mathbf{B}^{-1}\vec{b}_1 + \mathbf{B}^{-1}\mathbf{b}_2^l. \tag{87}$$

Equation (87) is solved numerically using the initial condition

$$\hat{v}_k^l = \begin{cases} 0 & l \neq 0 \\ \dfrac{2k+1}{2}\displaystyle\int_{-1}^{1} g(x)L_k(x)dx & l = 0 \end{cases}. \tag{88}$$

For the numerical experiment, we use the following initial and boundary conditions:

$$u(x, t=0, \xi) = \sin(\pi x), \quad u(x=-1, t, \xi) = \sin[\pi(-1-t)].$$

With these conditions, the mean $f(x,t)$ and the variance $g(x,t)$ of the exact solution $u(x,t,\xi)$ are given by

$$f(x,t) = \sin[\pi(x-t)] + \frac{1}{2}(x+1), \quad g(x,t) = \frac{1}{4}(1-x^2). \tag{89}$$

The variance is the same as the variance of Eq. (9), which is because the homogeneous solution is independent of the random variable ξ,. For the time integration we use the third-order Runge-Kutta total variation diminishing (TVD) scheme [22]. The mean and the variance at t are computed by

$$\text{mean} = \sum_{k=0}^{N} v_k^0(t)L_k(x), \quad \text{variance} = \sum_{l=1}^{N} \frac{1}{2l+1}\left[\sum_{k=0}^{N} v_k^l(t)L_k(x)\right]^2.$$

Figure 6 shows the solution for $\xi = 0$ (left figure) and the variance (right) at $t = 10$. As shown in the right figure, convergence of variance is slow due to the Gibbs phenomenon.

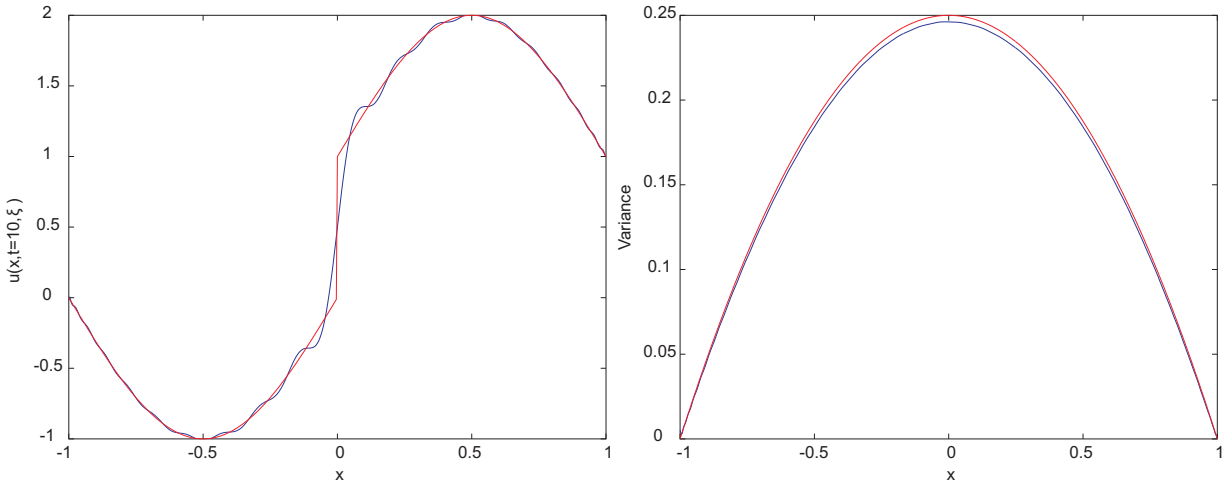

FIG. 6: Left: The numerical solution (oscillatory, blue) and the exact solution (red). Right: Variance. $\xi = 0$ at $t = 10$ with $N = 41$.

6. DIRECT PROJECTION COLLOCATION METHODS

In the previous sections, we used the Galerkin approach to obtain the solution of the differential equations with the random variable ξ. The Galerkin approach yields the Gibbs phenomenon, as shown in the previous sections. In this section, we solve the same equations using the collocation method based on the direct projection approach for the singular source term [23]. The direct projection approach uses the direct derivative of the Heaviside function for the singular source term on the collocation points. The direct collocation method was applied to several applications [23–25]. The main idea of the direct projection approach is to project the Heaviside function $H(x)$ to the collocation points using the spectral derivative matrix D_N, that is,

$$\delta_N(x) \longrightarrow D_N H_N(x),$$

where $\delta_N(x)$ is the spectral approximation of the δ-function on the collocation points with D_N the derivative matrix and $H_N(x)$ the Heaviside function on the collocation points. Several spectral derivative matrices related to the orthogonal polynomials can be found in [19].

6.1 A Simple First-Order Differential Equation

Consider the following differential equation with the random variable ξ,

$$\frac{du(x,\xi)}{dx} = \delta(x-\xi).$$

Let U_N be the approximation of u on $N+1$ collocation points for ξ, $\{\xi_l\}_{l=0}^N$. The collocation method yields the approximation $U_N(x,\xi)$ in the Legendre polynomials as in the previous sections,

$$U_N(x,\xi) = \sum_{l=0}^{N} \hat{u}_l(x) L_l(\xi). \tag{90}$$

Here we assume that we also seek $U_N(x,\xi)$ on the collocation points for x, $\{x_l\}_{l=0}^M$. That is, the spectral method is applied for both x and ξ directions, and the solution $U_N(x,\xi)$ is defined on the two-dimensional grid. By plugging $U_N(x,\xi)$ into the differential equation, we obtain

$$D_M U_N(x,\xi) = \delta(x-\xi), \tag{91}$$

where D_M is the spectral derivative matrix for the variable x on $M+1$ collocation points associated with some orthogonal polynomials such as Chebyshev or Legendre polynomials. For the singular source term in the right-hand side (RHS) of the above equation, the direct projection method uses

$$\delta(x-\xi) \longrightarrow D_M H_M(x-\xi),$$

where H_M is the Heaviside function on the collocation points which has the jump at $x=\xi$. Then Eq. (91) becomes

$$D_M U_N(x,\xi) = D_M H_M(x-\xi). \tag{92}$$

To solve the differential equation, we first use the boundary condition, which is

$$U_N(-1,\xi) = 0, \quad \forall \xi \in \{\xi_l\}_{l=0}^N. \tag{93}$$

From Eqs. (92) and (93) we obtain

$$\tilde{D}_M(U_N(x,\xi) - H_M(x-\xi)) = 0, \quad \forall \xi \in \{\xi_l\}_{l=0}^N,$$

where $\tilde{D}_M$ is the submatrix of D_M, which is obtained by subtracting the boundary column and row from D_M. The RHS 0 denotes a null vector, and $U_N(x,\xi)$ in the LHS is a solution vector for a certain ξ. Since $\tilde{D}_M$ is nonsingular [19], we obtain

$$U_M(x,\xi) = H_M(x-\xi), \tag{94}$$

which is the same as the exact solution, and we know that such solution is *Gibbs-free* on the collocation points.
Remark: *We note that the interpolation based on the solution at the collocation points yields the Gibbs oscillations, but the solution is Gibbs-free on the collocation points.*

6.2 A Simple Time-Dependent Problem

Now we consider the time-dependent problem with the collocation method

$$U_t + U_x = D_x H(x-\xi), \tag{95}$$

where $U = U(x,t,\xi)$ and D_x denotes the derivative operator with respect to x. U is defined in the same way,

$$U_N(x,t,\xi) = \sum_{l=0}^{N} \hat{u}_l(x,t) L_l(\xi). \tag{96}$$

For the steady-state problem, using the following boundary condition,

$$U(-1,t,\xi) = 0, \quad t > 0, \tag{97}$$

and we have the given differential equation which becomes as $t \to \infty$,

$$U_x = D_x H(x-\xi). \tag{98}$$

This steady-state solution becomes $U(x,t,\xi) \to H(x-\xi)$, as shown in the previous section.

For the numerical experiment we use the Chebyshev polynomials for x and the Legendre polynomials for ξ. As in the previous section, we use the third-order Runge-Kutta TVD scheme for the time integration [22]. For the initial and boundary conditions we use the following:

$$\begin{aligned} U^0 &= [\sin(\pi x_0), \sin(\pi x_1), \cdots, \sin(\pi x_{M-1}), \sin(\pi x_M)]^T \\ U^n(x_0) &= \sin[\pi(x_0 - t^n)], \quad \forall n = 1, 2, \cdots, \end{aligned} \tag{99}$$

where $x_i, i = 0, \cdots, M$ are the Chebyshev Gauss–Lobatto collocation points, $x_i = -\cos(\pi i/M), i = 0, \cdots, M$. With these initial and boundary conditions, the exact solution $u(x, \xi)$ is given by

$$u(x, t, \xi) = \sin[\pi(x - t)] + H(x - \xi), \tag{100}$$

where the first term is the homogeneous solution and the second term is the particular solution due to the singular source term.

Figure 7 shows the collocation solution for $u(x, \xi)$. Figure 7a shows the solution when $\xi = 0$, and the middle shows the collection of solutions with various ξ. Figure 7a shows that the solution is not affected by the Gibbs phenomenon without any oscillations on the collocation points. Figure 7b shows that along the line $x = \xi$, the jump of each solution is sharp, without any Gibbs oscillations. For these figures, we use $M = N+1$ and $N = 81$. Figure 7c shows the variance with $N = 41$ and $M = 21$. The variance from the numerical solution is the blue line with the $\square$ symbol. As shown in the figure, the variance is more accurately computed compared to the result in Fig. 6. The numerical results, however, show that the degree of accuracy is similar to that with the numerical simulation with the Galerkin approach, although the Gibbs oscillations are not seen on the collocation points. This result is somewhat different from what the authors expected, partly because the collocation approach has the ambiguity in defining the

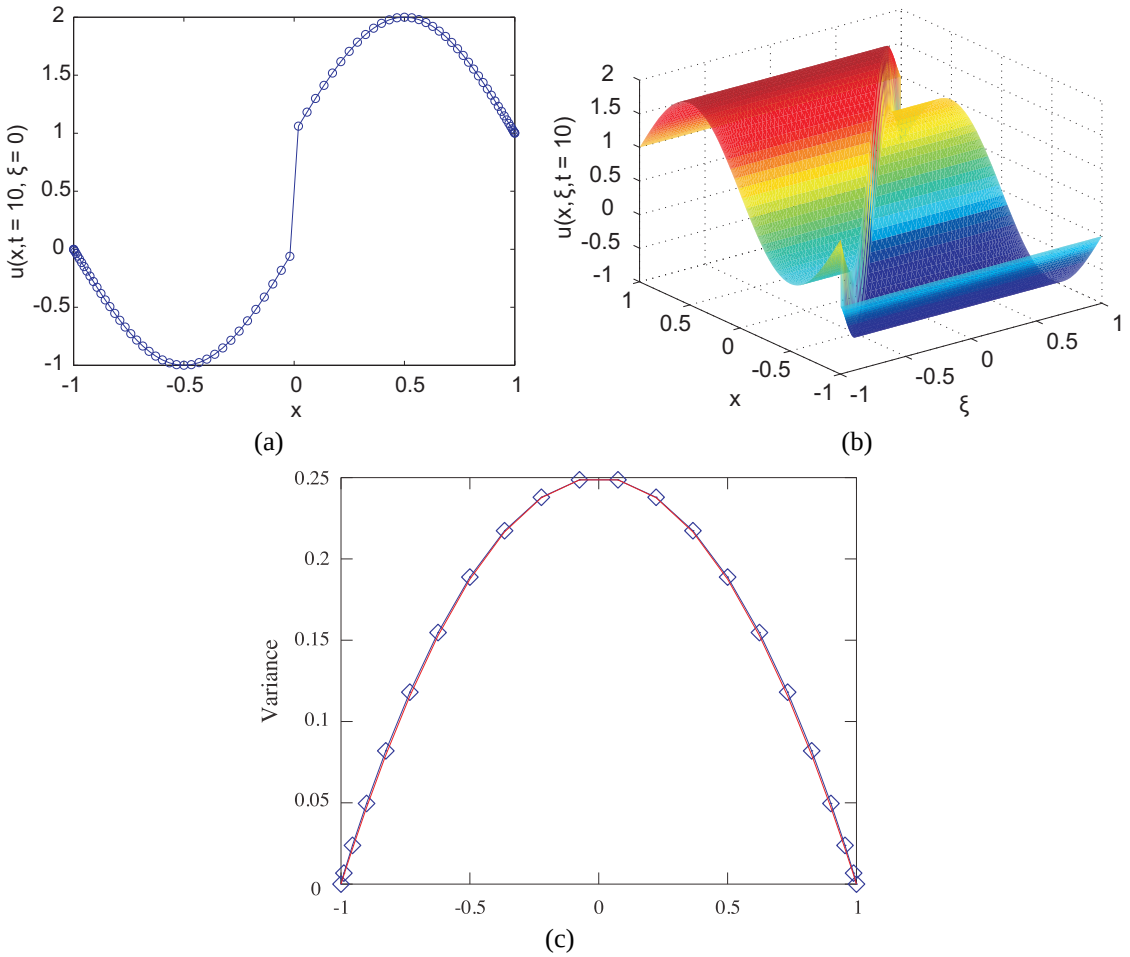

FIG. 7: (a) The solution at $\xi = 0$. (b) Polynomial chaos solutions for every ξ. The total number of grid points for x is $N = 81$. (c) The computed variance (blue line with square) and the exact variance (red line) with $M = 21$ and $N = 41$.

location of the δ-function and the Heaviside function. If the δ-function is located at a certain collocation point, the error does not decay at that point because the actual location of the δ-function with our collocation method exists off the collocation points. This issue will be further investigated in our future work.

7. CONCLUSION

In this paper we considered simple differential equations with a singular source term. For the singular source term, we used the Dirac δ-function. Due to the uncertainty of the location of the singular source term, we introduced a random variable and used the generalized polynomial chaos method to find the general solution of the differential equation under the uncertainty. For simplicity, we used the assumption that the uncertainty is associated with the uniform distribution. Based on this assumption, we derived the general solution of the differential equation in the Legendre polynomials using the Galerkin method, as well as the expectation value and variance of the solution. For this particular case, we show that the second- and third-order moments as well as the mean can be computed exactly using the first three expansion coefficients. The same technique was applied to the simple time-dependent problem. We showed that the Gibbs phenomenon appears in the polynomial chaos solution and consequently convergence is slow. As a preliminary work dealing with the Gibbs phenomenon in the solution, we considered the direct collocation method for the polynomial chaos solution. We showed that the direct collocation method can avoid the Gibbs phenomenon for the simple differential equations considered in this paper. Although the Gibbs oscillations are much reduced, the convergence of variance is about the same order as the Galerkin approach, which will be further investigated in our future work. The assumption of uniform distribution yields relatively easy analysis. In our future work we will consider more realistic cases with different distributions for more general types of differential equations with the singular source term. Thus, our future work will include the polynomial chaos method for more types of uncertainty variables associated with the singular source term and will further investigate the collocation method for the polynomial chaos solution and the Gibbs phenomenon with the singular source term.

ACKNOWLEDGMENTS

J.H.J. is supported by the National Science Foundation under grant no. DMS-0608844. The authors thank Dongbin Xiu for his help on the initial setup of the problem and his valuable feedback on our manuscript.

REFERENCES

1. Engquist, B., Tornberg, A.-K., and Tsai, R., Discretization of Dirac delta function in level set methods, *J. Comput. Phys.*, 207:28–51, 2005.
2. Fei, Z., Kivshar, Y. S., and Váquez, L., Resonant kink-impurity interactions in the sine-Gordon model, *Phys. Rev. A*, 45(8):6019–6030, 1992.
3. Goodman, R. H. and Haberman, R., Chaotic scattering and the n-bounce resonance in solitary-wave interactions, *Phys. Rev. Lett.*, 98(10):104103-1–104103-4, 2007.
4. Jacobs, G. and Don, W.-S., A high-order WENO-Z finite difference based particle-source-in-cell method for computation of particle-laden flows with shocks, *J. Comput. Phys.*, 228:1365–1379, 2008.
5. Lousto, C. O. and Price, R. H., Head-on collisions of black holes: The particle limit, *Phys. Rev. D*, 55:2124–2138, 1997.
6. Wiener, S., The homogeneous chaos, *Am. J. Math.*, 60:897–936, 1937.
7. Gottlieb, D. and Xiu, D., Galerkin method for wave equations with uncertain coefficients, *Comm. Comput. Phys.*, 3:505–518, 2008.
8. Xiu, D., Fast numerical methods for stochastic computations: A review, *Commun. Comput. Phys.*, 5(2–4):242–272, 2009.
9. Xiu, D. and Hesthaven, J., High-order collocation methods for differential equations with random inputs, *SIAM J. Sci. Comput.*, 27(3):1118–1139, 2005.
10. Xiu, D. and Karniadakis, G., Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos, *Comput. Meth. Appl. Mech. Eng.*, 191:4927–4948, 2002.

11. Xiu, D. and Karniadakis, G., The Wiener–Askey polynomial chaos for stochastic differential equations, *SIAM J. Sci. Comput.,* 24:619–644, 2002.
12. Xiu, D. and Karniadakis, G., Modeling uncertainty in flow simulations via generalized polynomial chaos, *J. Comput. Phys.,* 187:137–167, 2003.
13. Xiu, D. and Karniadakis, G., Supersensitivity due to uncertain boundary conditions, *Int. J. Numer. Meth. Eng.,* 61(12):2114–2138, 2004.
14. Xiu, D., Lucor, D., Su, C., and Karniadakis, G., Stochastic modeling of flow structure interactions using generalized polynomial chaos, *J. Fluids Eng.,* 124:51–59, 2002.
15. Ghanem, R. G. and Spanos, P., *Stochastic Finite Elements: A Spectral Approach*, Springer, New York, 1991.
16. Le Maître, O. P. and Knio, O. M., *Spectral Methods for Uncertainty Quantification*, Springer, New York, 2010.
17. Xiu, D., Numerical methods for stochastic computations: A spectral method approach, Princeton UP, Princeton, 2010.
18. Bateman, H., *Higher Transcendental Functions*, Vol. 2 (edited by A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi) McGraw-Hill, New York, 1953.
19. Hesthaven, J. S., Gottlieb, S., and Gottlieb, D., *Spectral Methods for Time-Dependent Problems*, Cambridge UP, Cambridge, 2007.
20. Gottlieb, D. and Orszag, S. A., *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, 1977.
21. Sansone, G., *Orthogonal Functions*, A. H. Diamond (Trans.), Robert E. Krieger Publishing Co., Huntington, NY, 1997.
22. Shu, C.-W., Total variation diminishing Runge-Kutta schemes, *SIAM J. Sci. Stat. Comp.,* 9:1079–1084, 1988.
23. Jung, J.-H., A note on the spectral collocation approximation of differential equations with singular sources in one dimension, *J. Sci. Comput.,* 39(1):49–66, 2009.
24. Jung, J.-H. and Don, W.-S., Collocation methods for hyperbolic partial differential equations with singular sources, *Adv. Appl. Math. Mech.,* 1(6):769–780, 2009.
25. Jung, J.-H., Khanna, G., and Nagle, I., A spectral collocation approximation for the radial-infall of a compact object into a Schwarzschild black hole, *Int. J. Mod. Phys. C,* 20(11):1827–1848, 2009.