AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2015WR017295

Parameter estimation and prediction for groundwater contamination based on measure theory

S. A. Mattis[1], T. D. Butler[2], C. N. Dawson[1], D. Estep[3], and V. V. Vesselinov[4]

[1]Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, Texas, USA, [2]Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, Colorado, USA, [3]Department of Statistics, Colorado State University, Fort Collins, Colorado, USA, [4]Computational Earth Science Group, Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

Key Points:
- Apply measure-theoretic framework to inverse problem for contaminant transport
- Identify and estimate both high- and low-probability parameter events
- Quantify uncertainties in predictions for remediation strategies

Correspondence to:
S. A. Mattis,
steven@ices.utexas.edu

Citation:
Mattis, S. A., T. D. Butler, C. N. Dawson, D. Estep, and V. V. Vesselinov (2015), Parameter estimation and prediction for groundwater contamination based on measure theory, *Water Resour. Res.*, *51*, 7608–7629, doi:10.1002/2015WR017295.

Received 24 MAR 2015
Accepted 24 AUG 2015
Accepted article online 28 AUG 2015
Published online 18 SEP 2015

Abstract The problem of groundwater contamination in an aquifer is one with many uncertainties. Properly quantifying these uncertainties is essential in order to make reliable probabilistic-based predictions and decisions regarding remediation strategies. In this work, a measure-theoretic framework is employed to quantify uncertainties in a simplified groundwater contamination transport model. Given uncertain data from observation wells, the stochastic inverse problem is solved numerically to obtain a probability measure on the space of unknown model parameters characterizing groundwater flow and contaminant transport in an aquifer, as well as unknown model boundary or source terms such as the contaminant source release into the environment. This probability measure is used to make predictions of future contaminant concentrations and to analyze possible remediation techniques. The ability to identify regions of small but nonzero probability using this method is illustrated.

1. Introduction

Contamination of groundwater aquifers is a significant problem in many locations around the world [*Rügner et al*., 2006; *National Research Council*, 2013]. In many instances, there may be various remediation strategies for reducing contamination levels that are economically viable and technically sound. However, determination of an optimal remediation strategy is complicated by many factors. Most remediation strategies are subject to governmental budget constraints [*National Research Council*, 1999, 2013], and the actual outcomes of various plausible remediation strategies are often uncertain for a variety of reasons, e.g., due to the limited information in characterizing the past, current, and future nature and extent of contaminants in the environment. This causes substantial uncertainties in future concentrations predicted from models under different remediation scenarios. There are many additional uncertainties related to the contaminant transport and fate including the identification of the magnitude, size, and duration of the contaminant release in the environment. Further complicating the problem of making accurate predictions are inherent uncertainties in the geological, hydrological, and biogeochemical conditions of the affected aquifer due to limited characterization data. Consequently, in order to make useful predictions of concentrations of a contaminant, its effect on the environment, and to best plan for remediation, quantifying uncertainties impacting the contaminant transport and fate is essential. The process for quantifying uncertainties affecting real-world decisions regarding complex physical processes requires the fusion of data-driven and model-driven techniques to incorporate what little information is known.

There has been a great deal of work regarding the development of theoretical and computational frameworks for contaminant remediation incorporating uncertainties in the parameters and models [see *Bolster et al*., 2009; *Agostini et al*., 2009a, 2009b; *Argent et al*., 2009; *Tartakovsky*, 2007; *Jordan and Abdaal*, 2013; *National Research Council*, 1999]. Often, there is uncertainty in the underlying flow and transport models and model inadequacy studies can improve results [*Ye et al*., 2004; *Beven and Westerberg*, 2011]. There have been many attempts to incorporate model and prediction uncertainty into the decision-making process for remediation [*Caselton and Luo*, 1992; *Hipel and Ben-Haim*, 1999; *Bolster et al*., 2009; *Reeves et al*., 2010; *Harp and Vesselinov*, 2013; *O'Malley and Vesselinov*, 2014]. It is common to characterize the uncertainties in groundwater contamination models probabilistically [*Delhomme*, 1979; *Dagan*, 1982; *Wagner*

and Gorelick, 1987; *Abbaspour et al*., 1997; *Keating et al*., 2010], where a probability distribution is associated with each parameter separately. Representing probabilities in high dimensions can be challenging, but some recent methods have increased computational efficiency [*Tonkin and Doherty*, 2009; *Laloy and Vrugt*, 2012].

We present a recently developed measure-theoretic framework for the formulation and solution to a physically meaningful stochastic inverse problem for quantifying uncertainties in a physics-based model [*Breidt et al*., 2011; *Butler et al*., 2012; *Butler and Estep*, 2013; *Butler et al*., 2014; T. Butler et al., Solving stochastic inverse problems using sigma-algebras on contour maps, 1407.3851, 2014]. A computational algorithm is described in detail and applied to a mathematical model for groundwater contamination. The inverse solution is a probability measure that is consistent with the unique solution in the space of parameter equivalence classes defined by the physics-based map from parameters to data. By consistent, we mean that propagation of the probability measure through the computational model exactly reproduces the probability measure on the data space. These equivalence classes can be identified as *generalized contours* in the original parameter space. We demonstrate how the probability measure can be used to identify and approximate the probability of failure events for various remediation strategies in terms of reducing future contaminant levels below the maximum concentration limit (MCL).

The use of generalized contours in the formulation and solution of the stochastic inverse problem is quite different from other approaches that have appeared in the literature. A complete literature review is infeasible, so we limit the comparison below to popular approaches focusing on similar applications. Bayesian and Generalized Likelihood Uncertainty Estimation (GLUE) approaches have proven to be among the most popular approaches for quantifying uncertainties in hydrologic models and environmental problems [see *Freer and Beven*, 1996; *Leube et al*., 2012; *Nowak et al*., 2010; *Troldborg et al*., 2010; *Vrugt et al*., 2008; *Beven and Freer*, 2001]. The Bayesian and GLUE approaches replace the map from uncertain parameters to observations defined by the physics-based model with a statistical map called the likelihood function. After specifying a prior distribution on the parameters, the objective is to interrogate a posterior distribution defined in terms of the likelihood and prior distributions often by using a Markov Chain Monte Carlo (MCMC) sampling. In other words, the inverse problem is formulated in terms of a statistical problem involving the fit of model outputs to data, and typical objectives are either to produce a set of samples generated from the posterior distribution or to determine the parameters of maximum likelihood. The maximum likelihood can often be reinterpreted as the minimum of a misfit functional which defines a so-called "regularized" solution to the inverse problem (e.g., see the sequence of papers [*Carrera and Neuman*, 1986a,b,c] for the methods, algorithms, and applications of this idea). In regularization, one seeks to minimize a misfit functional defined using both residual errors along with a "penalty" term. The covariances of the likelihood and prior distributions in a Bayesian formulation can often be used to define the norms used for the residual errors and penalty term, respectively. Since the regularized solution is the solution to an optimization problem, interpretation of the obtained results often involve some type of sensitivity analysis [see *Carrera et al*., 2005; *Carrera and Neuman*, 1986a; *Mayer et al*., 2002]. By design, the misfit functionals in regularization approaches and the related posterior distributions in Bayesian approaches have completely different contour structures than the generalized contours of the physics-based map between model parameters and observational data. The implication is that the solutions to the regularization and Bayesian approaches solve a different inverse problem and have completely different interpretations than what is obtained by the measure-theoretic formulation and solution of the inverse problem.

The paper is organized as follows. In section 2, we describe the contaminant transport model, physical domain, and model parameters that are the focus of this work. In section 3, we summarize the measure-theoretic framework and computational algorithm used in the rigorous probabilistic analysis of uncertainties. In section 4, we provide numerical results demonstrating this probabilistic analysis on the contaminant transport model. A low-dimensional problem is solved to illustrate the effect of the underlying geometry defined by the choice of quantities of interest on the solution to the stochastic inverse problem. A higher-dimensional stochastic inverse problem is then solved and used to quantify uncertainties in both model and source parameters. The solution of the higher-dimensional stochastic inverse problem is used to make predictions about contaminant plumes when no remediation is performed and for a variety of remediation strategies. We focus the analysis of predictions specifically on reducing contaminant levels below the MCL. Concluding remarks follow in section 5.

2. Contaminant Transport Model and Parameters

Contaminant transport in groundwater aquifers is a complex process and its modeling is associated with a great deal of underlying uncertainty. We demonstrate the benefit of quantifying uncertainties using the measure-theoretic framework in section 4.4 by comparing the resulting probabilistic predictions to those obtained from an uninformed predictive analysis. In order to explore the entire set of possible parameters and identify regions of high probability in an uninformed way, typically, a large number of forward model solutions are computed. In order to do this efficiently, we use a relatively simple analytical model with many sources of uncertainty [*Wang and Wu*, 2009] as a forward model. We note that using a more complicated model presents no theoretical difficulties in the measure-theoretic approach while the practical approximation issues can be addressed directly using adaptive sampling based on computable error estimates/bounds (T. Butler et al., Solving stochastic inverse problems using sigma-algebras on contour maps, 1407.3851, 2014), but this is beyond the scope of this work. In section 3.3, we provide a general description of the convergence and error analysis in the measure-theoretic approach and compare to the more traditional sampling methods using posterior distributions obtained via a Bayesian formulation.

Some assumptions are made about the model domain and setup in order to simplify the model as shown in Figure 1. We assume that the domain is infinite in the horizontal (x-y) plane and semiinfinite in the vertical (z) direction, and that the region of contaminant release (the contamination source) is a rectangular box, and the volumetric mass released per unit time (the mass flux) is uniform within the source at a given time. The groundwater flow field is steady and uniform in the horizontal direction. The contaminant undergoes first-order decay, accounting for a variety of geochemical processes that might be reducing the contaminant concentration (e.g., first-order chemical reactions) [*Petrucci et al.*, 1993]. We assume there is no contaminant in the domain at the initial time, and the aquifer hydraulic conductivity and porosity are homogeneous throughout the domain. The resulting contaminant transport is modeled using a three-dimensional advection-dispersion-reaction equation:

$$\frac{\partial C}{\partial t}+u\frac{\partial C}{\partial x}-u\alpha_x\frac{\partial^2 C}{\partial x^2}-u\alpha_y\frac{\partial^2 C}{\partial y^2}-u\alpha_z\frac{\partial^2 C}{\partial z^2}+\delta C=I/n, \tag{1}$$

where $C\ [ML^{-3}]$ is the contaminant concentration. The model parameters α_x, α_y, and $\alpha_z\ [L]$ are the dispersivities in the x, y, and z directions, respectively, $u\ [LT^{-1}]$ is the pore groundwater flow velocity along the x axis, $\delta\ [T^{-1}]$ is the first-order constant for decay, $n\ [L^3/L^3]$ is the porosity of the aquifer, and $I\ [ML^{-3}T^{-1}]$ is the contaminant mass flux. The source I is defined by

$$I(x',y',z';t)=\begin{cases} I_0, & x-\frac{1}{2}x_s<x'<x+\frac{1}{2}x_s, y-\frac{1}{2}y_s<y'<y+\frac{1}{2}y_s, z_0<z'<z_1, \\ & t_0<t<t_1 \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where I_0 is a given constant. The source parameters defining I include the location of the center of the contaminant source at x (m), y (m), and z_0 (m) and the size of the source x_s (m), y_s (m), and z_1 (m) (in the x, y, and z directions, respectively). The remaining source parameters are t_0 (year) the initial time of the source, t_1 (year) the final time of the source, and f (kg/year) the contaminant flux within the source. Under the given assumptions, there exists an analytical solution:

$$\begin{aligned} C(x,y,z;t)=\frac{1}{8n}\int_0^t \Bigg(& I(t-\tau)\exp(-\delta\tau) \\ & \times\left[\operatorname{erfc}\left(\frac{x-\frac{1}{2}x_s-ut}{\sqrt{2\alpha_x u\tau^\kappa}}\right)-\operatorname{erfc}\left(\frac{x-\frac{1}{2}x_s-ut}{\sqrt{2\alpha_x u\tau^\kappa}}\right)\right] \\ & \times\left[\operatorname{erfc}\left(\frac{y-\frac{1}{2}y_s}{\sqrt{2\alpha_y u\tau^\kappa}}\right)-\operatorname{erfc}\left(\frac{y-\frac{1}{2}y_s}{\sqrt{2\alpha_y u\tau^\kappa}}\right)\right] \\ & \times\left[\operatorname{erfc}\left(\frac{z-z_0}{\sqrt{2\alpha_z u\tau^\kappa}}\right)-\operatorname{erfc}\left(\frac{z-z_1}{\sqrt{2\alpha_z u\tau^\kappa}}\right)\right]\Bigg) d\tau. \end{aligned} \tag{3}$$

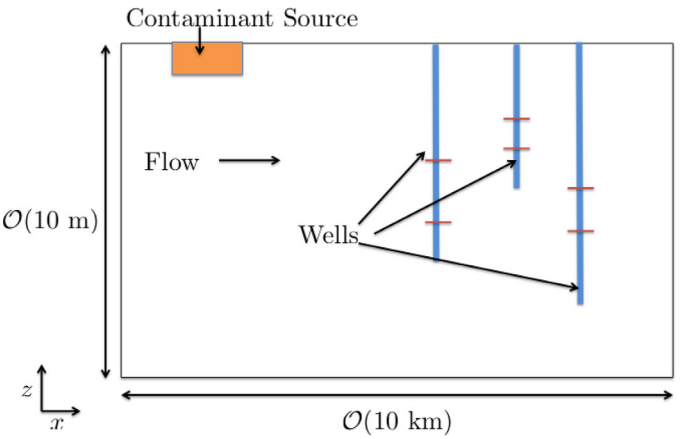

Figure 1. Configuration of contaminant transport model.

where κ is a time scaling factor for dispersivity. If $\kappa = 1$, there is no time scaling for dispersivity. If $\kappa > 1$, the dispersivity scales with the travel time. Scaling of the dispersivity for the case of transport in porous geologic media has been frequently observed in field hydrogeologic studies [*Neuman*, 1990; *Gelhar et al*., 1992; *Schulze-Makuch*, 2005].

By equation (1), the source I and the solution $C(x, y, z; t)$, there are a large number of parameters that determine the concentration of the contaminant at a given point in the space-time domain even for this simplified model. They can be divided into two categories: source and transport parameters. Source parameters characterize the contaminant release into the aquifer. Transport parameters represent the underlying transport processes in the aquifer. The measure-theoretic framework for quantifying uncertainties formulates the stochastic inverse problem with respect to a parameter domain containing all of these model and source parameters. Generally these parameters are not known exactly or even estimated. In practice, the available information consists of observations of contaminant concentrations at specific points in space-time. However, under reasonable assumptions about the source and the properties of the aquifer, intervals in which the parameter values are contained may be determined or estimated.

3. Measure-Theoretic Framework for Uncertainty Quantification

This section summarizes the mathematical methodology and numerical methods for a measure-theoretic framework for stochastic inverse problems for physics-based maps as formulated by *Breidt et al*. [2011], *Butler et al*. [2012], *Butler and Estep* [2013], *Butler et al*. [2014], and T. Butler et al. (Solving stochastic inverse problems using sigma-algebras on contour maps, 1407.3851, 2014). Below, we emphasize the core deterministic forward and inverse maps at the heart of all the computations involving uncertainties modeled with probability measures. In order to make these ideas less abstract, we connect the notation and concepts to the model described above.

We let Λ denote the space of possible (input) parameters for the model (including source parameters) and Q be a map from the parameter space Λ to the output space $\mathcal{D} := Q(\Lambda)$. Here $\mathcal{D}$ represents contaminant concentrations given by the map Q defined by evaluating $C(x, y, z; t)$ at specific points in space-time that depend implicitly on the choice of parameters. The components of Q are referred to as "quantities of interest" (QoI) and Q is called the QoI map. In an idealized case, Q would be a bijection (i.e., a one-to-one and onto map) between Λ and $\mathcal{D}$. This idealized situation is rarely the case even when Λ and $\mathcal{D}$ have the same dimension. Typically, there is a set of points in Λ that are mapped to the same point in $\mathcal{D}$ by Q, i.e., the map Q^{-1} is *set-valued*. For example, increasing the porosity while decreasing the contaminant mass flux may result in no difference in observed contaminant concentrations. Moreover, the dimension of $\mathcal{D}$ is often less than that of Λ, which implies that the set-valued inverses are defined by a collection of lower-dimensional manifolds embedded in Λ that we call generalized contours. In either case, it is impossible to assign a distinct parameter to a distinct output datum. We illustrate this idea in the left plot in Figure 2. Putting this in the stochastic setting where parameters and/or data are random variables does not change the fundamental issue that the map Q^{-1} is set-valued. Solving the stochastic inverse problem involves computing a probability measure on the original parameter space Λ, given a probability measure on $\mathcal{D}$ and the set-valued map Q^{-1} that maps to generalized contours in Λ.

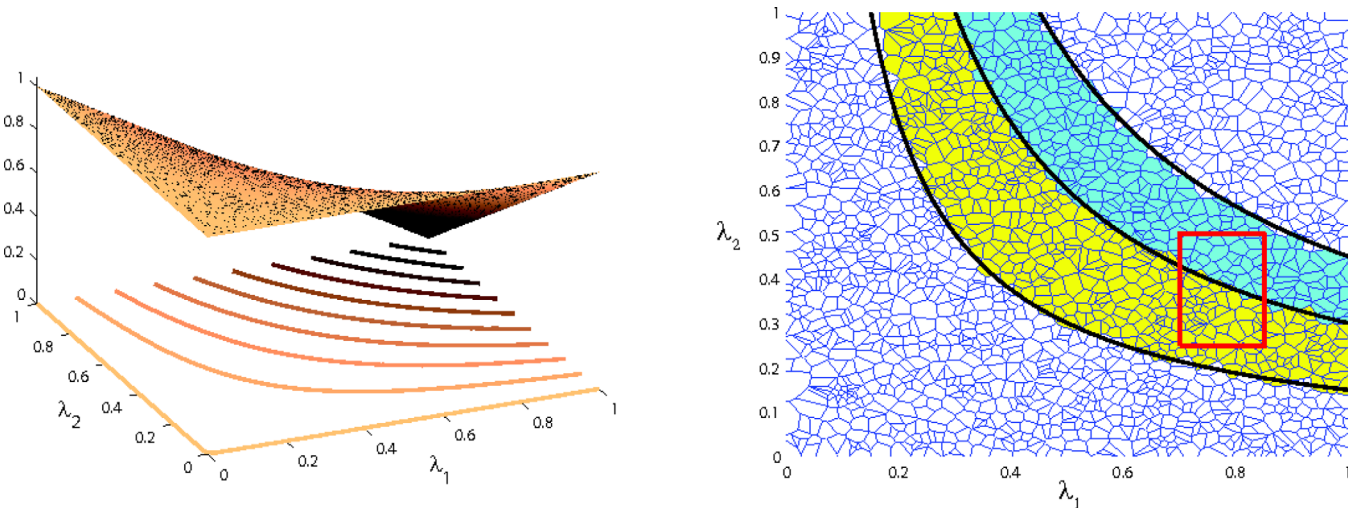

Figure 2. (left) A simple two-to-one QoI map with two parameters λ_1 and λ_2 with several contours illustrated identifying specific set-valued inverses. (right) The regions between the black curves correspond to two distinct adjacent output intervals for the QoI map. The highlighted Voronoi cells in yellow and blue denote the approximations to these contour regions. The red rectangle indicates a possible event A for which an approximate probability may be computed.

3.1. Formulation of the Stochastic Inverse Problem

For the contaminant transport model, the inverse of any particular observable contaminant vector of concentrations is defined by the set of all transport and source parameters that can possibly produce these contaminant concentrations. This implies that the map Q^{-1} maps to sets instead of points in Λ. We call any particular inverse set a generalized contour. We denote by $\mathcal{L}$ the collection of unique generalized contours in Λ.

We consider the case when Q is a map between $\Lambda \subset \mathbf{R}^n$ and $\mathcal{D} \subset \mathbf{R}^d$ with $n > d \geq 1$. This is the more challenging case than when $n = d$ although that case is easily handled within this framework and the resulting algorithm. It is possible to uniquely represent and index each generalized contour using a representation of $\mathcal{L}$ as a (piecewise-defined) d-dimensional manifold embedded in Λ. For the contaminant transport model, this means that there is a set of transport and source parameters such that the mapping is one-to-one between this set and all observable contaminant levels.

Suppose that a probability measure $P_{\mathcal{D}}$ with density $\rho_{\mathcal{D}}$ is defined on $\mathcal{D}$ (e.g., representing uncertainty in observed contaminant concentration levels), so that for any event B in $\mathcal{D}$, the probability of B is $P_{\mathcal{D}}(B) = \int_B \rho_{\mathcal{D}}$. Since $Q_{\mathcal{L}} : \mathcal{L} \rightarrow \mathcal{D}$ is a bijection, there is a unique inverse probability measure $P_{\mathcal{L}}$ with density $\rho_{\mathcal{L}}$ induced on $\mathcal{L}$ by

$$P_{\mathcal{L}}(A) = \int_A \rho_{\mathcal{L}} = \int_{Q_{\mathcal{L}}(A)} \rho_{\mathcal{D}} = P_{\mathcal{D}}(Q_{\mathcal{L}}(A)), \tag{4}$$

for any event A in $\mathcal{L}$. We emphasize that the solution of the stochastic inverse problem in $\mathcal{L}$ is uniquely specified within this measure-theoretic framework.

Note that $\rho_{\mathcal{L}}$ provides a method of computing probabilities of events in $\mathcal{L}$ which correspond to generalized contour events in Λ. The objective is to calculate the probability of more arbitrary events in Λ such as generalized rectangles or ellipsoids which are unlikely to be approximated well by generalized contour events. We make use of the Disintegration Theorem [*Butler et al.*, 2014; *Dellacherie and Meyer*, 1978] and a standard ansatz to compute a probability measure P_Λ in terms of an approximate density ρ_Λ that is consistent with $P_{\mathcal{D}}$. By saying P_Λ is consistent with $P_{\mathcal{D}}$, we mean that if B is any event in $\mathcal{D}$ (so $Q^{-1}(B)$ is a generalized contour event in Λ), then

$$\int_{Q^{-1}(B)} \rho_\Lambda(\lambda) d\mu_\Lambda = P_\Lambda(Q^{-1}(B)) = P_{\mathcal{D}}(B) = \int_B \rho_{\mathcal{D}} d\mu_{\mathcal{D}}. \tag{5}$$

We summarize the stochastic inverse problem as the following: given a parameter space Λ, a data space $\mathcal{D}$, a map $Q : \Lambda \rightarrow \mathcal{D}$, and a probability measure $P_{\mathcal{D}}$ on $\mathcal{D}$, determine a probability measure defined on Λ satisfying equation (5). For the contaminant transport problem, this means the problem is to determine a probability

measure on the transport and source parameters such that mapping this probability measure through the model produces the same probability measure as the one defined on the observable contaminant concentrations.

3.2. Numerical Solution to the Stochastic Inverse Problem

Suppose the probability density function $\rho_{\mathcal{D}}$ associated with $P_{\mathcal{D}}$ is known. Let $\{D_k\}_{k=1}^{M}$ be a partition of $\mathcal{D}$, and let p_k be the probability of D_k. Let A be an event in Λ. We thus have the approximation

$$\int_{Q(A)} \rho_{\mathcal{D}} d\mu_{\mathcal{D}} \approx \sum_{D_k \subset Q(A)} p_k. \tag{6}$$

Let $\{\lambda^{(j)}\}_{j=1}^{N}$ be a collection of N points in Λ, which implicitly defines an n-dimensional Voronoi tessellation $\{\mathcal{V}_j\}_{j=1}^{N}$ associated with the points. For example, given $\lambda \in \Lambda$, $\lambda \in \mathcal{V}_k$ if $\lambda^{(k)}$ is the closest $\lambda^{(j)}$ to λ. Using the implicitly defined Voronoi tessellation and the partitioning of $\mathcal{D}$, Algorithm 1 can be used to calculate the probability $\hat{\rho}_{\Lambda,j}$ associated with the implicitly defined Voronoi cell $\mathcal{V}_j$. We illustrate the steps of Algorithm 1 using the simple quadratic QoI map from a 2-D parameter space ($n = 2$) to a 1-D data space ($d = 1$) shown in Figure 2. In practice, we only construct the samples in step 1 of Algorithm 1 and the Voronoi tessellation of step 2 is implicitly defined by the set of samples from step 1. However, for pedagogical reasons, it is easier to visualize the remaining steps of the algorithm using the explicit Voronoi tessellation as shown in the right plot of Figure 2. Step 3 of the algorithm simply defines which data value is associated with each Voronoi cell. Step 4 defines a partitioning of the data space which corresponds to a unique collection of generalized contour events in Λ. In the right plot of Figure 2, the regions between the black curves represent the exact contour events corresponding to two subintervals from a possible partition of $\mathcal{D}$. The probabilities associated with these regions are determined by $\rho_{\mathcal{D}}$ as indicated in step 5 of Algorithm 1. Steps 6 and 7 are used to identify which Voronoi cells are used to approximate the exact contour events implicitly defined by step 4. In the right plot of Figure 2, we use yellow and blue highlighting to distinguish which collections of Voronoi cells are identified as approximating the two contour events with boundaries given by the black curves. Note that steps 8 and 9 of Algorithm 1 determine the probability of each individual Voronoi cell according to the ansatz discussed above, where the volume is determined in step 8 and the ansatz is applied in step 9. Note that in steps 5 and 8 in Algorithm 1, we must approximate integrals. The computation of step 8 uses Monte Carlo integration in high-dimensional spaces. We then approximate the probability of any event $A \subset \Lambda$ using a counting measure

$$P_{\Lambda}(A) \approx \sum_{\lambda^{(j)} \in A} \hat{\rho}_{\Lambda,j}. \tag{7}$$

Thus, we obtain an approximation to the probability measure on Λ that solves the stochastic inverse problem as described above and satisfies equation (5). In the right plot of Figure 2, we consider A defined as the region interior to the red rectangle and the probabilities of the Voronoi cells have all been determined. By simply identifying the samples in A and summing the probabilities associated with these samples to approximate $P(A)$, we can reinterpret this counting measure approximation as a type of Monte Carlo approximation to $P_{\Lambda}(A)$.

Algorithm 1. Numerical Approximation of Inverse Probability Density

1. Choose points $\{\lambda^{(j)}\}_{j=1}^{N} \in \Lambda$.
2. Denote the associated Voronoi tessellation $\{\mathcal{V}_j\}_{j=1}^{N} \subset \Lambda$.
3. Evaluate $Q_j = Q(\lambda^{(j)})$ for all $\lambda^{(j)}$, $j = 1, \ldots, N$.
4. Choose a partitioning of $\mathcal{D}$, $\{D_k\}_{k=1}^{M} \subset \mathcal{D}$.
5. Compute $p_k \approx \int_{D_k} \rho_{\mathcal{D}} d\mu_{\mathcal{D}}$, for $k = 1, \ldots, M$.
6. Let $\mathcal{C}_k = \{j | Q_j \in D_k\}$ for $k = 1, \ldots, M$.
7. Let $\mathcal{O}_j = \{k | Q_j \in D_k\}$, for $j = 1, \ldots, N$.
8. Let V_j be the approximate measure of $\mathcal{V}_j$, i.e., $V_j \approx \int_{\mathcal{V}_j} d\mu(\mathcal{V}_j)$ for $j = 1, \ldots, N$.
9. Set $\hat{\rho}_{\Lambda,j} = (V_j / \sum_{i \in \mathcal{C}_{\mathcal{O}_j}} V_i) p_{\mathcal{O}_j}$, $j = 1, \ldots, N$.

3.3. Computational Complexity and Error Analysis

As described in section 1, the inverse problems formulated in the measure-theoretic framework and Bayesian frameworks are fundamentally different, which complicates any direct comparisons between the solution methods. However, if we assume there are no hyperparameters used in the definitions of the prior distributions for the Bayesian formulation, then the solutions to either problem formulation are (different) probability measures on the (same) parameter space. If the goal of generating samples from the Bayesian posterior distribution is to approximate probabilities (i.e., measures) of events, then it may at least be possible to compare the cost of approximating the solutions to within some specified accuracy with sample-based approaches. A rigorous analysis of the computational cost for numerical methods for either problem formulation is not the focus of this work. Here we summarize the basic components of a convergence and error analysis to describe how such comparisons of the computational cost associated with using Algorithm 1 can be made to the computational cost of using a standard Monte Carlo sampling scheme to numerically approximate posterior distributions using a Bayesian framework. For a more thorough explanation of the theory and proofs related to convergence and error analysis of Algorithm 1, we direct the interested reader to *Butler et al.* [2014] and T. Butler et al. (Solving stochastic inverse problems using sigma-algebras on contour maps, 1407.3851, 2014).

Using a standard Monte Carlo sampling scheme in a Bayesian framework to estimate the posterior ensures independence and can easily be used to provide estimates of probabilities of events written as expected values. The convergence of the estimates of probabilities of events is then subject to the well-known Central Limit Theorem. As described in *Butler et al*. [2014], Algorithm 1 can be interpreted as a Monte Carlo approximation to probabilities of events. The key to the convergence results in *Butler et al*. [2014] is to understand the approximation of events using results from stochastic geometry that rely on the Strong Law of Large Numbers, which is a key result used in proving the Central Limit Theorem. Thus, it may be possible to compare the rates of convergence for either method using similar statistical tools.

It may also be possible to compare the cost of using an MCMC approach to sample the posterior distribution for a Bayesian formulation to an importance sampling approach used to place more samples in regions of high probability within Algorithm 1. This is a key interpretation that would lead to what we refer to as "adaptive sampling" algorithms within the measure-theoretic approach and is the subject of future work.

There remains the issue of the effect of numerical error on the computed solution. The effect of numerical error in solving a computational model pollutes the QoI samples in both the measure-theoretic and Bayesian framework. However, the measure-theoretic framework allows for a straightforward error analysis where the effect of this deterministic error on the estimated probabilities of events can be quantified. The key to this part of the error analysis is to use computable a posteriori error estimates (e.g., using an adjoint to the original model) as a way to identify possible "mischaracterization" of an output sample (i.e., incorrectly determining $\mathcal{O}_j$ in step 7 of Algorithm 1). This type of error analysis for a fixed computational model is unusual in a Bayesian framework, and to the authors knowledge, has not been carried out.

Addressing these issues in more detail and within the numerics would require a detailed description of adaptive sampling approaches, adjoint models, and a posteriori error estimates, which is beyond the scope of this work. We avoid issues of error in the numerics below by using the analytical solution and a sufficiently large number of samples so that all asymptotic bounds on statistical error can be considered negligible.

4. Numerical Results

Given a set of source and transport parameters, equation (3) can be used to calculate the concentration of the contaminant at points within the domain of interest. We use the open-source software package Model Analysis and Decision Support (MADS) (http://mads.lanl.gov; https://gitlab.org/monty/MADS) developed at Los Alamos National Laboratory, which makes such calculations computationally efficient. Accurate approximations to the transport solutions are obtained using GNU Science Library (GSL) subroutines. The open-source software package called BET (https://github.com/UT-CHG/BET) developed by the authors is used to compute the probability measure.

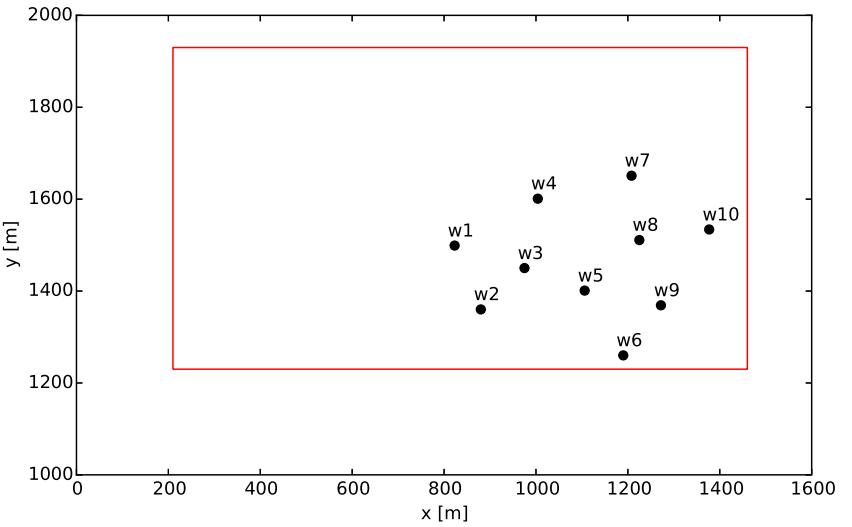

Figure 3. The domain of interest for the model problem with wells w1–w10 shown. The red rectangle indicates the potential locations for the center of the source.

When $Q : \Lambda \subset \mathbf{R}^n \rightarrow \mathcal{D} \subset \mathbf{R}^d$ is piecewise differentiable, the implicit function theorem implies that the rank of the Jacobian of Q is at most $\min\{n, d\}$. This motivates the definition of *geometrically distinct (GD) QoI* defined as a set of QoI whose Jacobian is full rank [*Butler et al.*, 2014]. Clearly, there exists at most n GD QoI. Given $m \geq n$ possible QoI, we may be able to choose several $d \leq n$ subsets of the possible QoI that satisfy the property of being GD. Given two distinct sets of QoI that are GD does not imply that the solutions to the inverse problem computed from either set will be the same. The quality of a solution to the inverse problem is affected by how the map Q skews events mapped between $\mathcal{L}$ and $\mathcal{D}$. This is similar to the condition number of a square matrix. A full analysis on the effect of the choice of QoI and the skewness on the probability measure P_Λ is beyond the scope of this work, and we direct the interested reader to *Butler et al.* [2015] for a more thorough discussion. However, we demonstrate in section 4.2 how the choice of GD QoI can have a large influence on the structure of the inverse probability measure for a low-dimensional parameter example where the number of possible QoI is larger than the number of parameters. In the higher-dimensional numerical example of section 4.3, the number of available QoI from the experiment is strictly less than the number of uncertain parameters, so we simply use all of the available QoI. The problem of how to choose an optimal GD subset from a proposed set of QoI by quantifying the skewness is the subject of on-going research. In section 4.3, we demonstrate how to analyze a probability measure in high dimensions. In section 4.4, we use the probability measure on the higher-dimensional parameter space to make predictive inferences and analyze various remediation strategies.

Table 1. Spatial Locations of Wells and Observed Contaminant Concentrations C at $t=1$ year Taken From Studies in Harp and Vesselinov [2012][a]

Well Name	x (m)	y (m)	C (mg/kg) at $t=1$ year
w1	823	1499	7.01
w2	880	1360	180.55
w3	975	1450	339.08
w4	1004	1601	4.97
w5	1106	1401	294.84
w6	1190	1260	3.47×10^{-4}
w7	1208	1651	1.05×10^{-5}
w8	1225	1511	n/a
w9	1272	1369	n/a
w10	1377	1534	n/a

[a]"n/a" = unknown. There is no measurement.

4.1. Configuration

The region of interest for the solution to equation (1) is shown in Figure 3. The region contains 10 wells labeled w1 through w10. The *x-y* coordinates of these wells are shown in Table 1. Wells w1–w7 are monitoring wells from which measurements of the concentration of the contaminant are taken. Wells w8, w9, and w10 are *points of compliance (PoC)*. Contaminant concentrations in the PoC are required to be below a threshold known as the MCL (EPA, http://water.epa.gov/drink/contaminants). For this study, the MCL is assumed to be 25 mg/kg. The measured concentrations of the contaminant are known at $t=1$ year in the monitoring wells w1–w7 and are listed in Table 1. This model setup and data

Table 2. Ranges for Unknown Parameters and Fixed Values for Known Parameters for Low-Dimensional Inverse Stochastic Sensitivity Analysis

Unknown Parameters	Min.	Max.
Source coordinate x (m)	210	1460
Source coordinate y (m)	1230	1930
Contaminant source flux f (kg/year)	0.1	100
Known Parameters	**Value**	
Porosity n	0.1	
Flow angle θ (°)	0	
Pore velocity u (m/year)	16	
Source size x_s (m)	250	
Source size y_s (m)	250	
Source location z_0 (m)	0	
Source size z_1 (m)	1	
Reactive decay λ (1/year)	0.01	
Longitudinal dispersivity α_x (m)	70	
Transverse horizontal dispersivity α_y (m)	$\alpha_x/10$	
Transverse vertical dispersivity α_z (m)	$\alpha_x/100$	
Dispersivity time-scaling factor κ	1.2	
Initial source time t_0 (year)	0	
End source time t_1 (year)	20	

are taken from studies in *Harp and Vesselinov* [2012]. The contaminants in the PoC are assumed unknown at time $t=1$ year.

In the numerical results below, we define a probability measure $P_{\mathcal{D}}$ on $\mathcal{D}$ in terms of a probability density $\rho_{\mathcal{D}}$ in the following way. We assume that any measured datum is subject to uncertainty with known error bounds for each component of the QoI map such that each component of the true datum is within these bounds. With no additional assumptions on the structure of the uncertainty within $\mathcal{D}$, this implies that the true vector-valued datum exists within a hyperbox centered on this measured datum and contained in $\mathcal{D}$. In each example, we specify the size of the hyperbox relative to the size of a circumscribing hyperbox $\mathcal{D}$. In order not to bias results too closely to the specific contour mapping to the measured datum, we take $\rho_{\mathcal{D}}$ to be the uniform density on this hyperbox.

4.2. QoI Influence on the Inverse Probability Measure

We first consider the case where there are only three unknown parameters: the x coordinate of the source x, the y coordinate of the source y, and the contaminant source flux f. Probabilities can be visualized effectively for such a low-dimensional problem. Table 2 shows the possible intervals of the unknown parameters and the fixed values of the other parameters that are assumed known. The specified intervals for x, y, and f define the parameter domain $\Lambda=[210,1460]\times[1230,1930]\times[0.1,100]$ where we let $\lambda=(x,y,f)$ (i.e., $\lambda_1=x$, $\lambda_2=y$, and $\lambda_3=f$). Suppose that the QoI are the concentrations of the contaminant in wells w1–w7 after 1 year of contamination, and let $q_i(\lambda)$ be the concentration in well wi for $i=1,2,\ldots,7$ for a given value of $\lambda\in\Lambda$. Let $\hat{q}_i$ denote the measured concentration in well wi for $i=1,2,\ldots,7$ shown in Table 1. To more clearly demonstrate the influence of the choice of QoI on P_Λ, we focus on using different sets of two GD QoI and present some representative examples below.

Suppose q_i and q_j are the chosen pair of GD QoI so that $Q(\lambda)=[q_i(\lambda),q_j(\lambda)]^T$ for all $\lambda\in\Lambda$ and $\mathcal{D}=Q(\Lambda)$. We define a probability measure on $\mathcal{D}$ following the steps described above. We first define a hyperbox (in this

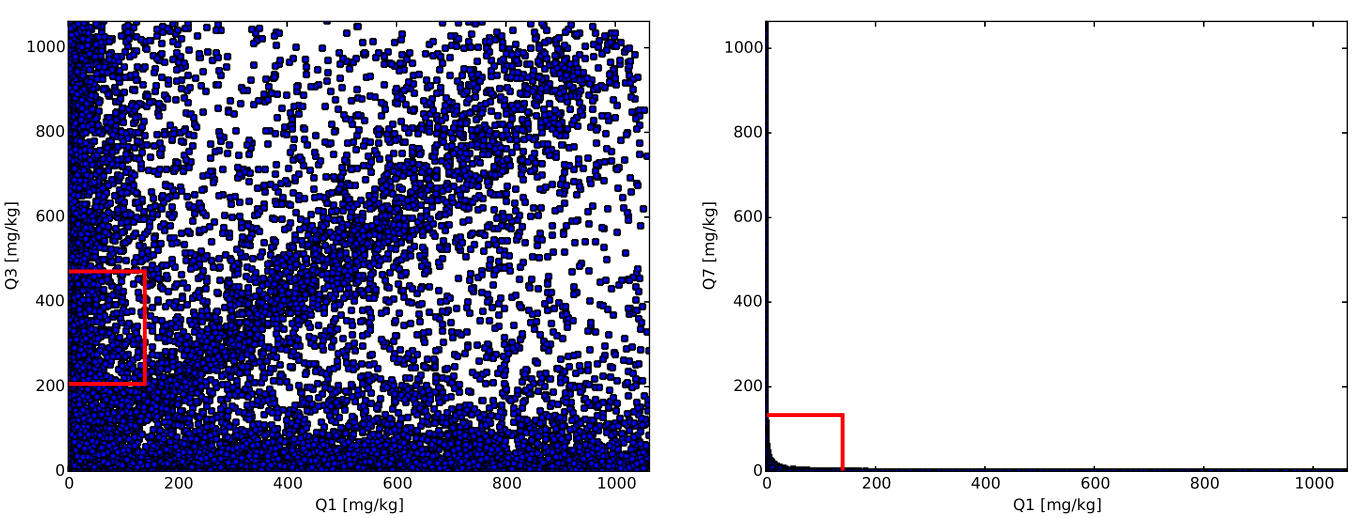

Figure 4. Data points in $\mathcal{D}$ associated with the low-dimensional parameter estimation problem. The data points correspond to the $Q(\lambda)=(q_1(\lambda),q_j(\lambda))$ for (left) $j=3$ and (right) $j=7$. (In this case, q_1, q_2, and q_3 are model predicted concentrations at wells w1, w3, and w7, respectively.) The set of data points for each scatterplot correspond to the mapping of the same i.i.d. uniform samples in Λ. The red box is the region where the probability is assumed to be uniform and nonzero.

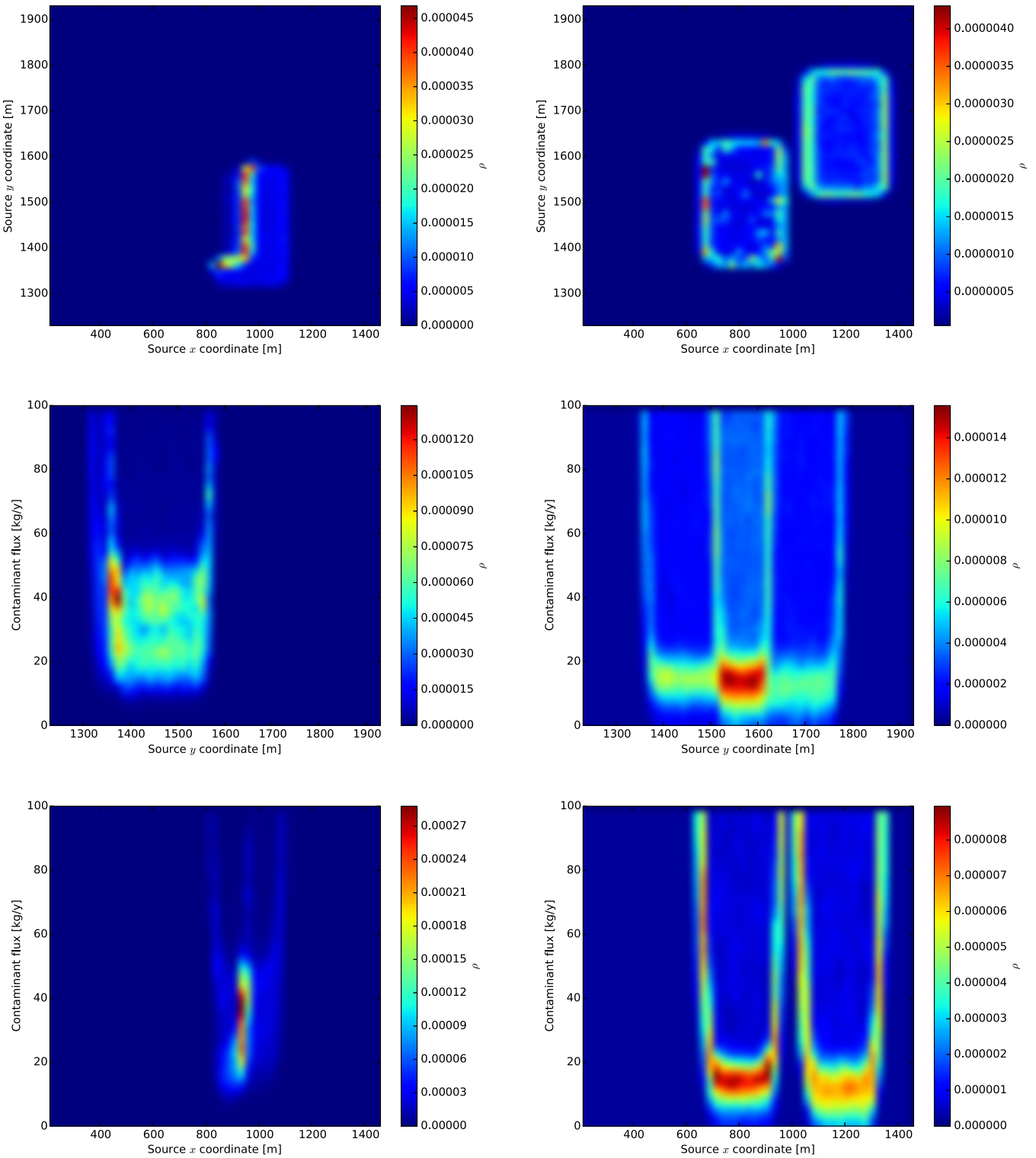

Figure 5. Computed 2-D marginal probabilities for the (top row) *x* and *y* coordinates of the source, (middle row) *y* coordinate of the source and contaminant source flux *f*, and (bottom row) *x* coordinate of the source and contaminant source flux *f* obtained by solving the stochastic inverse problem using $Q(\lambda)=(q_1(\lambda), q_j(\lambda))$ for (left column) $j=3$ and (right column) $j=7$.

case, a rectangle since $\mathcal{D} \subset \mathbf{R}^2$) centered at $[\hat{q}_i, \hat{q}_j]$ of dimensions $(0.3q_i(\Lambda))\times(0.3q_j(\Lambda))$ (see Figure 4 for some representative plots). We apply $N=5\times10^4$ independent identically distributed (i.i.d.; in this case, they are uniform) samples $\left\{\lambda^{(j)}\right\}_{j=1}^{N}$ in Algorithm 1 to approximate P_Λ.

Table 3. Ranges for Unknown Parameters and Fixed Values for Known Parameters for High-Dimensional Inverse Stochastic Sensitivity Analysis[a]

Unknown Parameters	Min.	Max.
Source coordinate x (m)	210	1460
Source coordinate y (m)	1230	1930
Contaminant source flux f (kg/year)	0.1	100
Porosity n	0.05	0.15
Flow angle θ (°)	−30	30
Pore velocity u (m/year)	6	60
Longitudinal dispersivity α_x (m)	10	140
Dispersivity time scaling factor κ	1	2
Known Parameters	**Value**	
Source size x_s (m)	250	
Source size y_s (m)	250	
Source location z_0 (m)	0	
Source size z_1 (m)	1	
Reactive decay λ (1/year)	0.01	
Transverse horizontal dispersivity α_y (m)	$\alpha_x/10$	
Transverse vertical dispersivity α_z (m)	$\alpha_x/100$	
Initial source time t_0 (year)	0	
End source time t_1 (year)	20	

[a]Note that the transverse dispersivities α_y and α_z are tied to the unknown longitudinal dispersivity α_x.

For low-dimensional QoI spaces such as these, it is possible to analyze scatter plots of the QoI to infer the geometric relationship of the data with respect to the parameters and determine which QoI map skews events mapped between Λ and $\mathcal{D}$ the most. We show some representative QoI scatter plots in Figure 4 for $(q_1(\lambda), q_3(\lambda))$ and $(q_1(\lambda), q_7(\lambda))$ (observed concentrations at w1 versus w3, and w1 versus w7, respectively). Note that the scatter plot for $(q_1(\lambda), q_3(\lambda))$ indicates that the QoI map defined by these components maps the rectangular box defining Λ onto the Cartesian product $q_1(\Lambda)\times q_3(\Lambda)$. This is in contrast to the scatter plot for $(q_1(\lambda), q_7(\lambda))$ where the specification of one of the components severely limits the possible values for the other component of this QoI map and Λ is skewed severely into a nonconvex set that fills only a fraction of the area of $q_1(\Lambda)\times q_7(\Lambda)$.

Figure 5 shows the marginal probabilities for P_Λ corresponding to the different pairs of QoI discussed above. All plots of marginal probabilities in this work are computed by defining a regular grid of the shown parameter (or pairs of parameters) and using the computed P_Λ to approximate the probabilities of the given intervals (or rectangles) and normalizing by the length (or area) to obtain a simple function approximation to the density. In order to better illustrate any geometric information imparted to P_Λ by the chosen QoI, we avoid methods such as kernel density estimation that smooth plots such as these to provide a better visual estimate of the density function. We see that the geometric information and degree of skewness inherent in the QoI map has a significant effect on the probabilities in the parameter space. Clearly, using the QoI with better skewness properties produces probability measures that can identify small regions in parameter space with high probability. This is useful when making predictions since fewer samples, and thus model simulations, are required.

4.3. Parameter Estimation in a Higher-Dimensional Problem

Suppose that there are now three unknown source parameters and five unknown transport parameters for a total of eight unknown parameters: the x coordinate of the source x, the y coordinate of the source y, the contaminant source flux f, the porosity n, the flow angle θ, the pore velocity u, the dispersivity in the flow direction α_x, and the dispersivity time scaling factor κ. Table 3 shows the possible intervals of the unknown parameters and the fixed values of the other parameters that are assumed known. As before, the parameter space Λ is defined by the Cartesian product of the intervals for the unknown parameters. Since there are seven wells where the concentration is known at $t=1$ year and the parameter space is eight dimensional, we use the entire set of measurements in wells w1–w7 as the QoI, i.e., $Q(\lambda)=[q_1(\lambda), q_2(\lambda), q_3(\lambda), q_4(\lambda), q_5(\lambda), q_6(\lambda), q_7(\lambda)]^T$ for all $\lambda \in \Lambda$ and $\mathcal{D}=Q(\Lambda)$.

To define $P_\mathcal{D}$, we scale each dimension of the circumscribing hyperbox of $\mathcal{D}$ by 0.2, and center the resulting scaled hyperbox at the measured datum $[\hat{q}_1, \hat{q}_2, \hat{q}_3, \hat{q}_4, \hat{q}_5, \hat{q}_6, \hat{q}_7]$. We take $P_\mathcal{D}$ to be a uniform probability measure within this scaled hyperbox. We use $N=10^6$ independent identically distributed (i.i.d.; uniform) samples $\left\{\lambda^{(j)}\right\}_{j=1}^{N}$ in Algorithm 1 to approximate P_Λ.

In this case, marginal probability plots provide limited insight into the structure of a probability measure in eight dimensions. For example, Figure 6 shows the 1-D marginal probabilities associated with each of the unknown parameters. Some of the marginal probabilities yield useful information in cases where the probability measure and structure of the generalized contours localize the probability to small ranges of certain parameter values. For example, the marginal densities for the x and y locations of the source have quite distinct peaks, indicating high probability for those locations in parameter space. It also appears to be probable that the contaminant flux is in the higher part of its interval; also the distribution has a binomial shape.

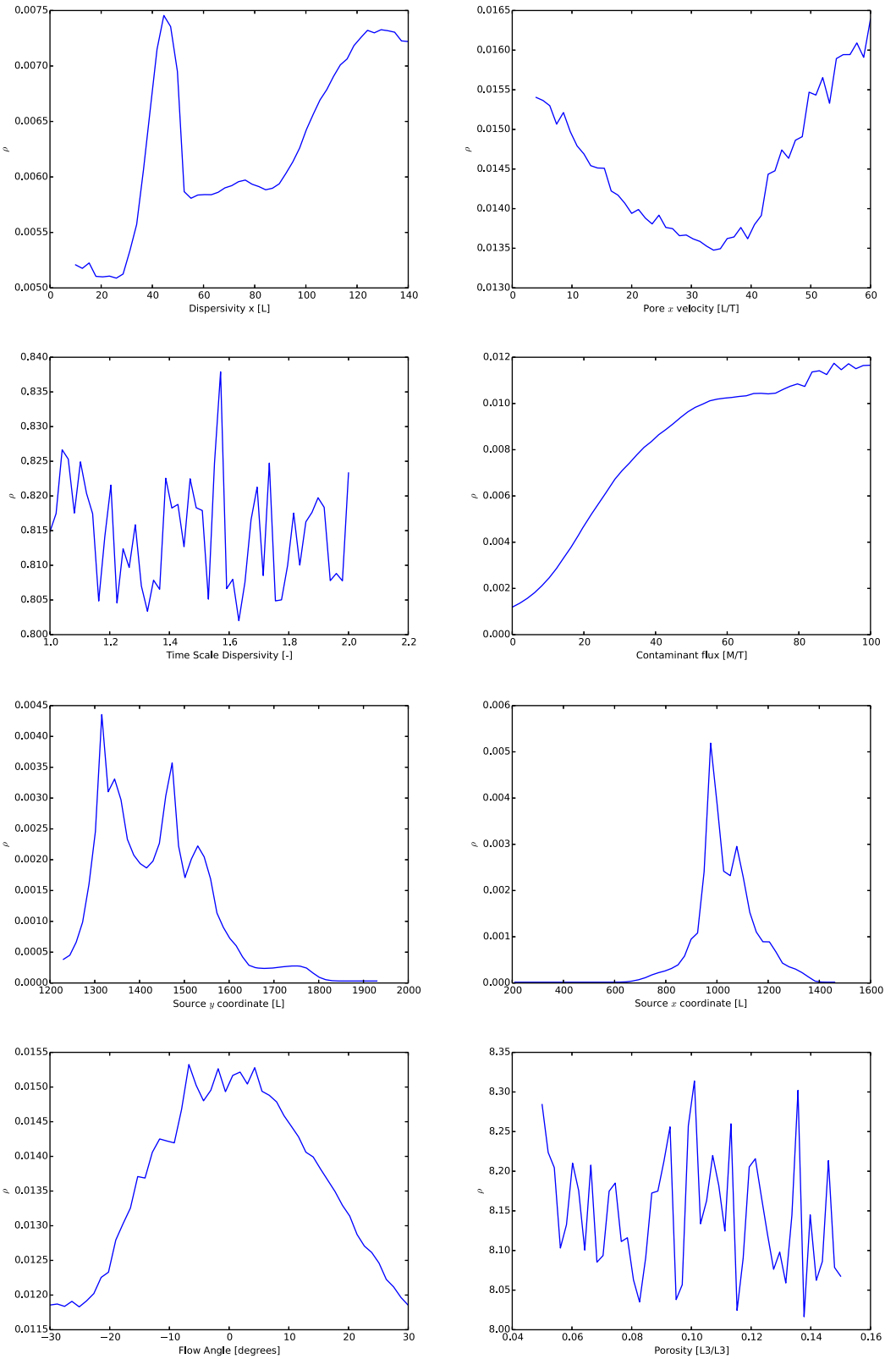

Figure 6. One-dimensional marginal probability density for each of the eight unknown parameters for the high-dimensional parameter estimation problem.

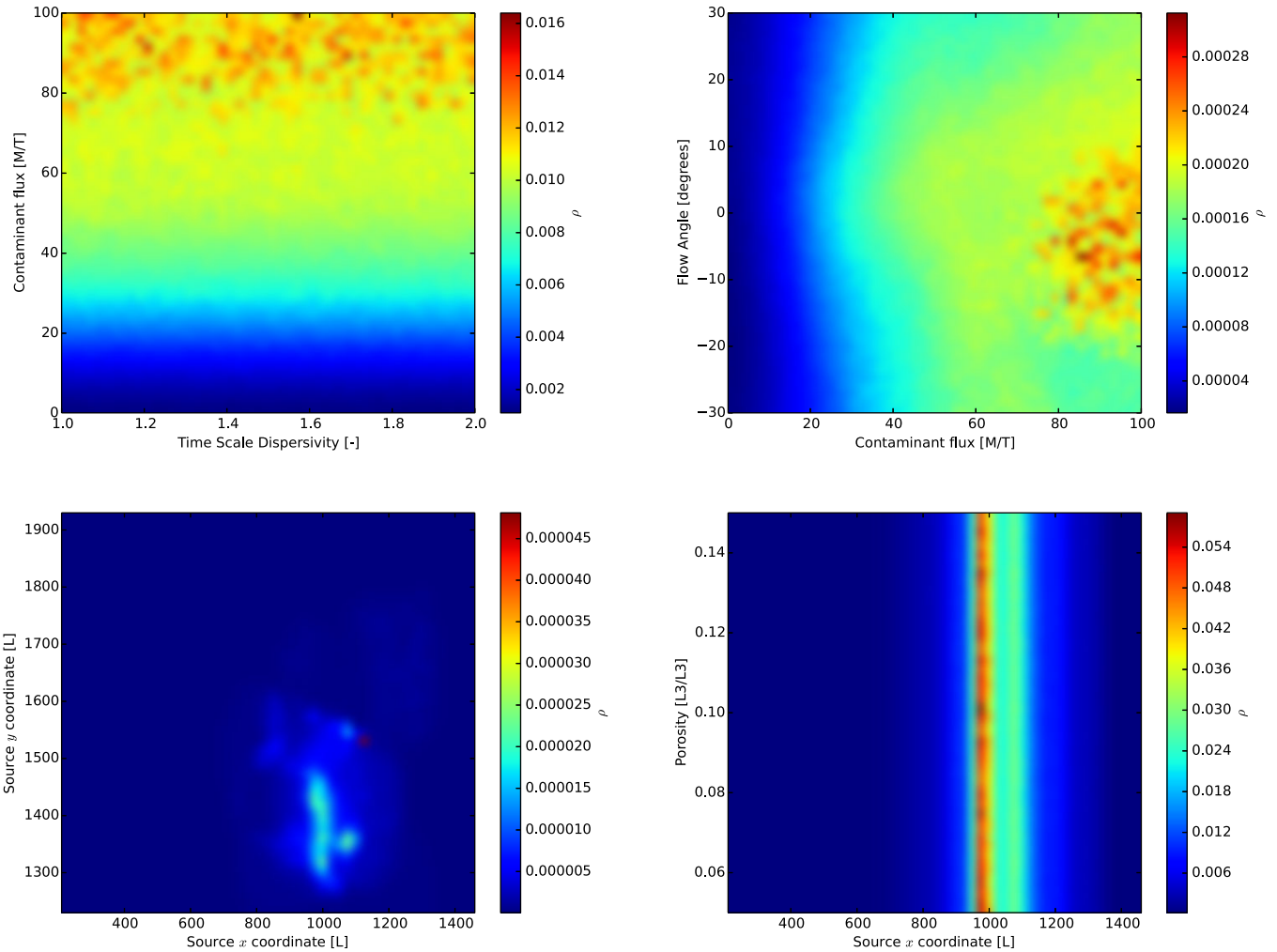

Figure 7. Two-dimensional marginal probability densities for selected pairs of unknown parameters for the high-dimensional stochastic inverse problem.

Figure 7 shows marginal probabilities for selected pairs of parameters. We note that the plots of Figures 5 and 7 show consistency among the low-dimensional and high-dimensional problems in the ability to identify small sets of high probability for the source locations (compare the top-left plot of Figure 5 to the bottom-left plot of Figure 7). However, other than a few interesting features, it appears that any type of inference or analysis based on these plots is of limited utility.

A useful way to interrogate a probability measure in high dimensions is to identify regions of high probability in the parameter space. The construction of the probability measure using Algorithm 1 lends itself to such an analysis quite naturally. Specifically, we may order by probability all of the implicitly defined Voronoi cells associated to each parameter sample. We may then identify the support of the probability measure, i.e., the event defined by all probable parameters. We can also determine the event of smallest volume containing certain percentages of the probability, i.e., the event is defined by identifying the smallest number of high-probability cells required to reach a certain probability threshold. Table 4 summarizes such an analysis. We obtain several interesting conclusions. First, the volume of the support of P_Λ (defined as the set in Λ where the probability density is positive) is roughly 53% of the entire parameter domain volume. However, we observe that 95% of the probability is contained in 11.6% of the volume of the parameter domain (which accounts for roughly 22% of the support). In other words, the cells accounting for the smallest 5% of probability define nearly 80% of the support. This implies that most of the probability is concentrated in an event of relatively small volume within Λ. Similarly, we observe that 90% of the probability is contained in 7.8% of the volume of the parameter domain (which is roughly 15% of the support). This

Table 4. Data for Regions of Highest Probability

$P(A \subset \Lambda)$	Number of Samples in A	$\mu(A \subset \Lambda)/\mu(\Lambda)$
100	442,346	0.53113
95	115,551	0.11616
90	81,081	0.07789
75	37,791	0.03519
50	10,071	0.00914
25	1,083	0.00103

suggests that when making predictions, it may be possible to ignore large portions of the support, and subsequently, of Λ, while maintaining accurate probabilistic predictions based on small events of high probability. This strategy is applied in the next section to analyze various alternative remediation strategies.

4.4. Predictions

We consider five different types of contaminant remediation strategies. In Strategy 1, the termination time of the source release is set to 1 year, i.e., $t_1 = 1$ year. This models a case where we are able to terminate the contaminant source after a year of contamination, e.g., by removal of the contaminant source. Similarly, in Strategy 2, the source is terminated after 5 years, i.e., $t_1 = 5$ year. In Strategy 3, the first-order reactive decay constant is increased from 0.01/year to 0.1/year. In Strategy 4, the first-order reactive decay constant is increased an additional order of

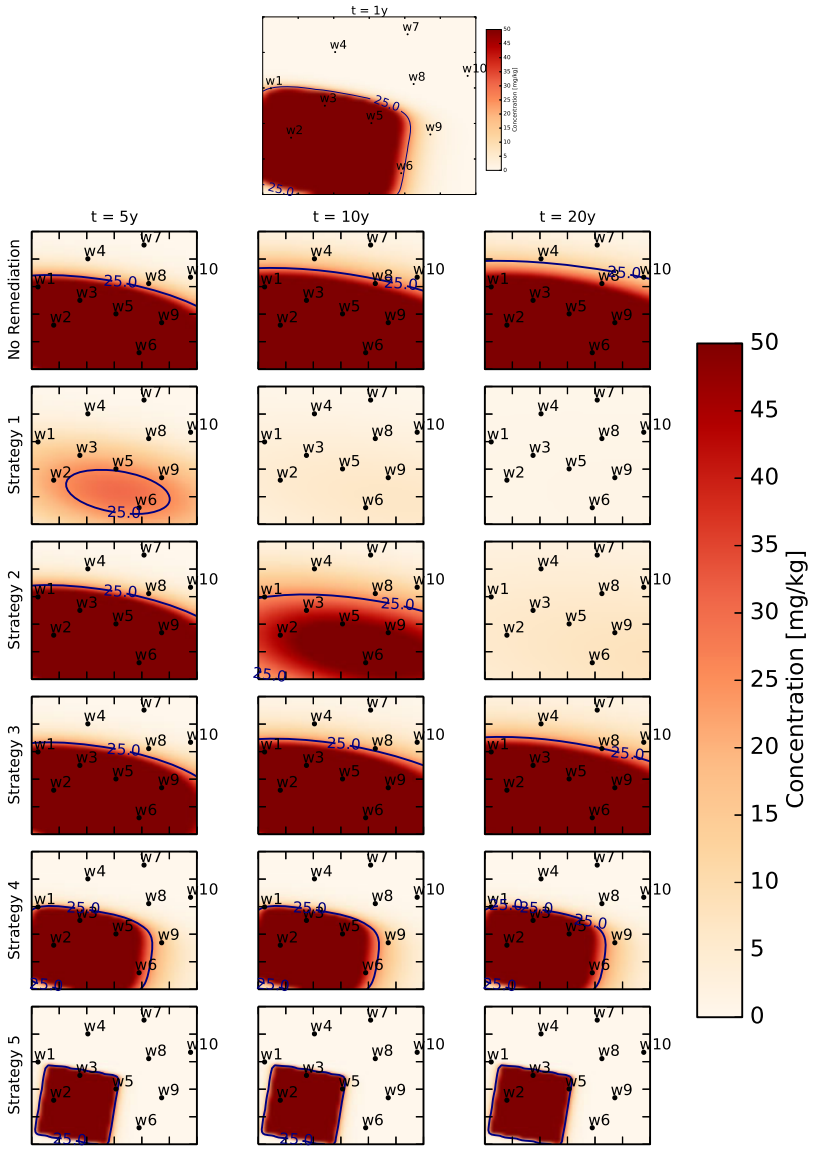

Figure 8. Effect of different remediation strategies for a parameter set randomly selected from the highest probability parameter cell.

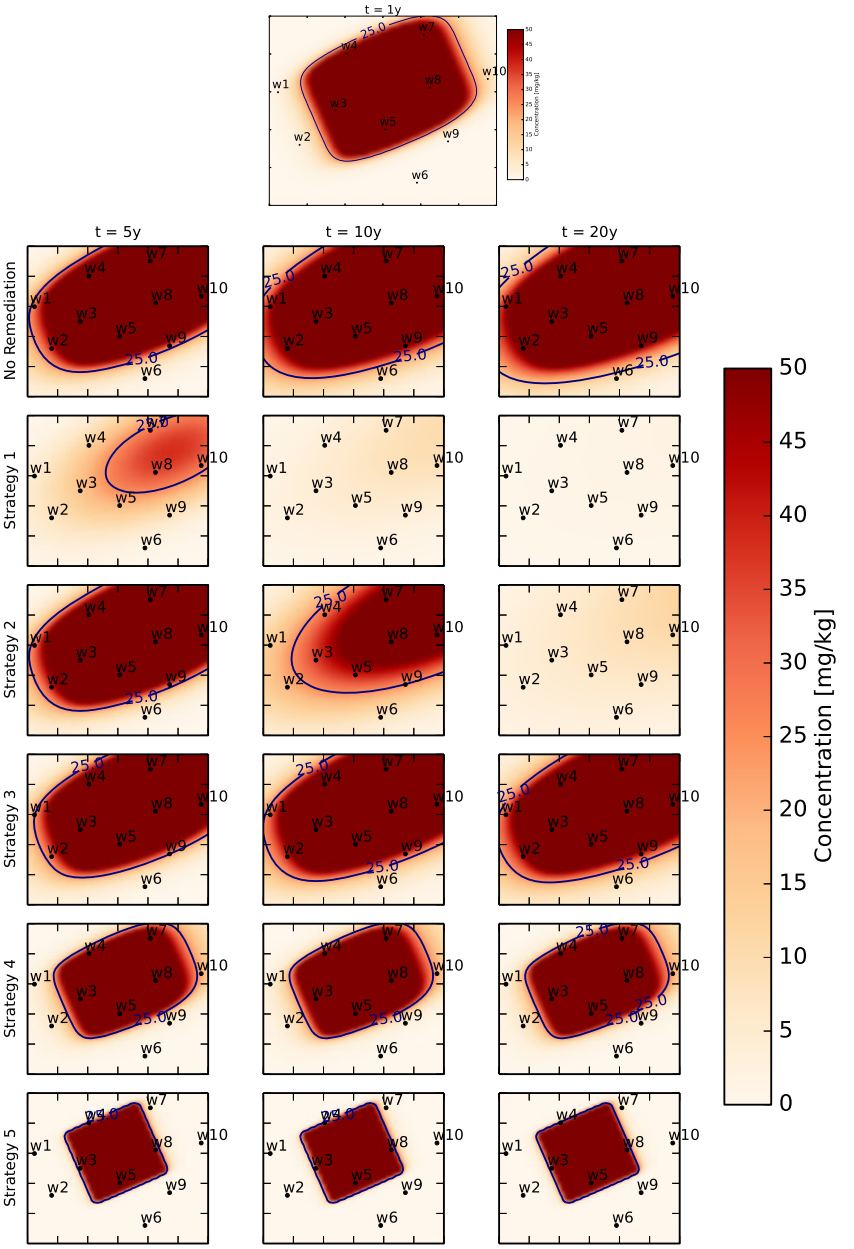

Figure 9. Effect of different remediation strategies for a parameter set randomly selected from third highest probability parameter cell.

magnitude to 1.0 /year. In Strategy 5, the first-order reactive decay constant is increased further to 10.0 /year. In Strategies 3–5, the changes in the decay constant represent artificially stimulated attenuation of the contaminants through some induced in situ biogeochemical processes.

Table 5. Mean Concentrations (mg/kg) in Wells With Different Remediation Strategies Calculated by Using 100% of the Computed Probability Measure

		Remediation Strategy					
Well	Time (Year)	None	R1	R2	R3	R4	R5
w8	5	296.32	22.39	296.32	279.47	121.49	22.02
	10	359.12	8.56	62.80	326.78	121.61	22.02
w9	5	207.62	20.72	207.62	193.92	73.59	12.27
	10	268.95	8.62	61.34	240.08	73.70	12.27
w10	5	70.43	12.50	70.43	63.88	13.36	0.96
	10	113.26	6.53	42.83	96.00	13.44	0.96

Table 6. Standard Deviation (mg/kg) of Concentrations in Wells With Different Remediation Strategies Calculated by Using 100% of the Computed Probability Measure

		Remediation Strategy					
Well	Time (Year)	None	R1	R2	R3	R4	R5
w8	5	555.21	34.71	555.21	527.94	256.73	50.31
	10	631.62	14.29	98.83	584.61	256.86	50.31
w9	5	403.53	25.99	403.53	384.45	193.44	38.74
	10	454.97	11.87	78.06	422.08	193.51	38.74
w10	5	178.03	22.68	178.03	165.17	59.60	10.40
	10	236.19	11.11	72.60	207.75	59.67	10.40

Individual predictions can vary significantly both qualitatively and quantitatively. We are primarily interested in predictions within the subdomain shown in the plots of Figures 8 and 9 containing the contaminant plume focused on the monitoring wells and PoC. This region is where we are interested in making predictive inferences and studying the effect of remediation strategies. For example, Figures 8 and 9 demonstrate the contaminant levels at $t=5$, 10, and 20 years for each remediation strategy (including the case of no remediation) for a parameter set taken randomly from the highest and third highest probability Voronoi cells, respectively. Despite the probabilities of the Voronoi cells containing these parameters being high relative to other cells, the orientation and magnitude of the contaminant plume at the production wells (w8–w10) vary significantly between these predictions. This demonstrates the fundamental problem of basing predictive inferences on any single parameter value, e.g., a mean or median value. Alternatively, one may propagate the probability measure on Λ to the space of predictions to compute statistics on the prediction space, e.g., mean predictions and standard deviations of the predictions. In Tables 5 and 6, we observe the mean prediction and standard deviation of this prediction for the contaminant level in each production well subject to each remediation strategy. Based on the means and standard deviations of the predictions, it appears that either remediation Strategy 1 or 5 does the best job of reducing the contaminant level below the MCL at the PoC. However, there are two obvious issues with using this information for prediction. First, the probability measures are clearly nonparametric with regions of high probability that are often nonconvex, defined by disconnected events, and/or are multimodal, so quantitative statistics such as the mean and variance have limited ability to describe the distribution. Second, while the standard deviation is smaller for the remediation strategies 1 and 5, the values are still quite large relative to the mean, i.e., the coefficient of variation is quite large. This indicates that the probability distribution is not concentrated near the mean value. Figures 10 and 11 which show probability density functions also display this tendency.

A far more robust analysis, i.e., an analysis that is less sensitive to outliers, uses the computed probability measure and identified events of high probability in parameter space to compute probabilities of exceeding the MCL. We restrict the scope of the analysis to the simplest type of computations involving the direct solution of all parameter samples in the various events of high probability listed in Table 4, and we discuss some more computationally efficient alternatives in the conclusions. In Table 7, the probabilities of exceeding the MCL based on propagating a uniform distribution on Λ to the prediction space. We may consider this an "uninformed predictive analysis." This "uninformed" analysis suggests between a 70% and 80% probability that performing no remediation will result in a reduction of contaminant levels below the MCL while in every case remediation Strategy 1 appears to outperform remediation Strategy 5.

In Tables (8–12), we show the probabilities that concentrations in the production wells exceed the MCL at $t_1=5$ and 10 year for the various remediation strategies conditioned on various events of high probability in the parameter space. Specifically, Table 8 shows the probabilities of concentrations exceeding the MCL computed using the entire probability distribution. Note that the probabilities of concentrations exceeding the MCL computed using 95% of the probability measure produces exactly the same probabilities of concentrations exceeding the MCL up to the number of digits reported in this table. Thus, we may obtain these predicted probabilities by either solving the model on the support of P_Λ or on the smaller event defining the most probable 95% of parameters. If we solve the model for the samples approximating these different events, then this implies that we can use roughly one fourth the number of samples to obtain the same prediction as indicated by Table 4.

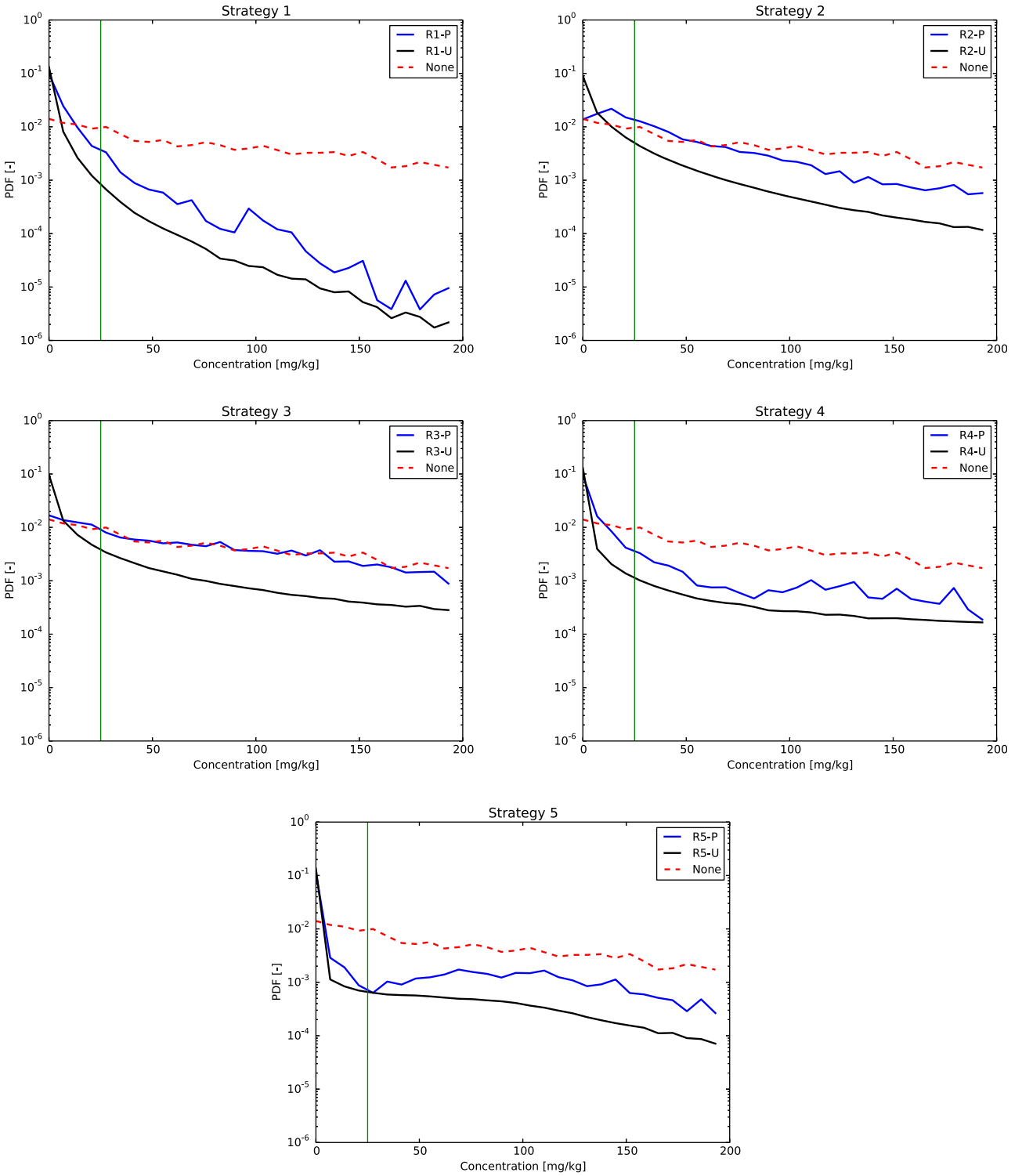

Figure 10. Probability densities for contaminant in well w8 at $t = 10$ years with different remediation strategies. Probabilities are computed using the solution to the stochastic inverse problem with remediation (blue), assuming a uniform distribution on Λ with remediation (black), and using the solution to the stochastic inverse problem with no remediation (red).

Tables (9–12) show that conditioning predictions on even smaller events of high probability shown in Table 4 generally lead to similar probabilistic predictions of contaminant levels. For example, solving the model for the parameter samples associated with the event containing 75% of the probability overestimates the

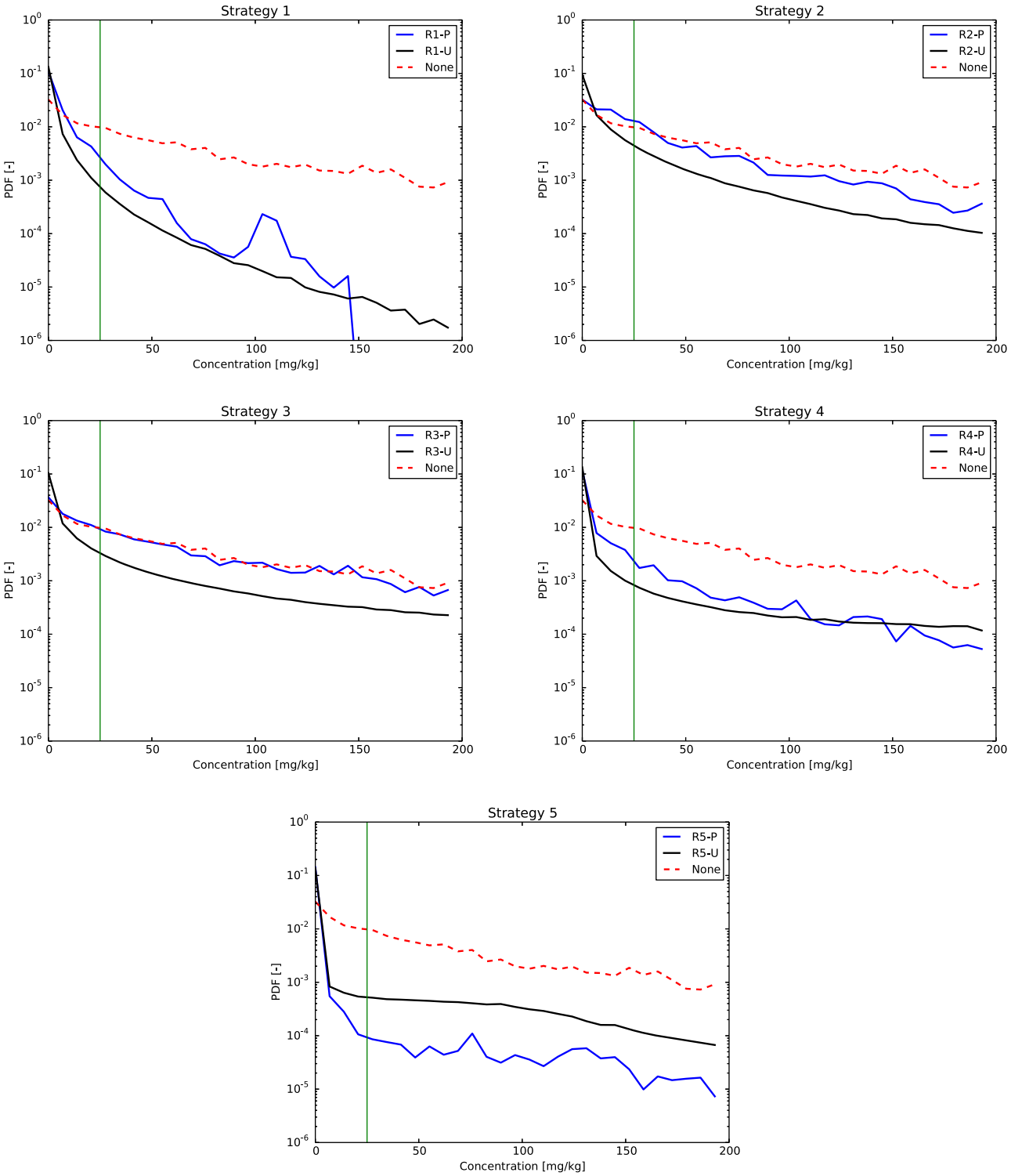

Figure 11. Probability densities for contaminant in well w10 at $t=10$ years with different remediation strategies. Probabilities are computed using the solution to the stochastic inverse problem with remediation (blue), assuming a uniform distribution on Λ with remediation (black), and using the solution to the stochastic inverse problem with no remediation (red).

probability of the contaminant levels exceeding the MCL by only a few percentage points in most cases while requiring approximately one twelfth the number of model solves. In almost all cases, comparing to the results from Table 7, we observe that the uninformed predictive analysis generally underestimated the probabilities of exceeding the MCL for every remediation strategy (with relative errors that vary from 100%

Table 7. Probability That Concentrations in Wells Will be Above 25 mg/kg With Different Remediation Strategies Sampling Uniformly From the Parameter Space

		Remediation Strategy					
Well	Time (Year)	None	R1	R2	R3	R4	R5
w8	5	0.2121	0.0562	0.2121	0.1926	0.1090	0.0619
	10	0.2966	0.0165	0.1737	0.2416	0.1091	0.0619
w9	5	0.1806	0.0522	0.1806	0.1672	0.1044	0.0615
	10	0.2421	0.0153	0.1449	0.2026	0.1045	0.0615
w10	5	0.1733	0.0498	0.1733	0.1567	0.0879	0.0521
	10	0.2520	0.0151	0.1550	0.2020	0.0880	0.0521

Table 8. Probability That Concentrations in Wells Will be Above 25 mg/kg With Different Remediation Strategies Calculated by Using Either 100% or 95% of the Computed Probability Measure

		Remediation Strategy					
Well	Time (Year)	None	R1	R2	R3	R4	R5
w8	5	0.68192	0.26152	0.68192	0.66377	0.33233	0.19593
	10	0.82678	0.07403	0.60490	0.79628	0.33331	0.19593
w9	5	0.74751	0.26849	0.74751	0.71932	0.25730	0.10659
	10	0.88352	0.07220	0.65598	0.85937	0.25796	0.10659
w10	5	0.40410	0.14373	0.40410	0.37875	0.09492	0.00663
	10	0.61853	0.04822	0.44968	0.56106	0.09524	0.00663

Table 9. Probability That Concentrations in Wells Will be Above 25 mg/kg With Different Remediation Strategies Calculated by Using 90% the Computed Probability Measure

		Remediation Strategy					
Well	Time (Year)	None	R1	R2	R3	R4	R5
w8	5	0.69270	0.26374	0.69270	0.67452	0.34033	0.20375
	10	0.83749	0.07437	0.61105	0.80739	0.34133	0.20375
w9	5	0.77087	0.27944	0.77087	0.74283	0.27089	0.11250
	10	0.89951	0.07492	0.66994	0.87768	0.27158	0.11250
w10	5	0.41027	0.14657	0.41027	0.38500	0.09622	0.00503
	10	0.62476	0.04898	0.45271	0.56671	0.09654	0.00503

to 300%) with a few notable exceptions. For Strategy 5, the probability that production well w10 has contaminant levels exceeding the MCL in years 5 and 10 is overestimated by the uninformed analysis by an order of magnitude.

The discrepancies between the uninformed predictive analysis and the predictive analysis using the computed probability measure on Λ are best explained by the discrepancies in the regions of small but nonzero probability located in the high-concentration portion in the space of predictions. To highlight the differences in these regions of the predicted densities, we show some representative semilog plots of these densities for the production wells w8 and w10 at year 10 in Figures 10 and 11 (the conditional

Table 10. Probability That Concentrations in Wells Will be Above 25 mg/kg With Different Remediation Strategies Calculated by Using 75% the Computed Probability Measure

		Remediation Strategy					
Well	Time (Year)	None	R1	R2	R3	R4	R5
w8	5	0.72164	0.27929	0.72164	0.70504	0.36144	0.22074
	10	0.86160	0.07883	0.63397	0.83458	0.36258	0.22074
w9	5	0.77676	0.26058	0.77676	0.74662	0.23232	0.08483
	10	0.90402	0.07141	0.66351	0.88336	0.23297	0.08483
w10	5	0.43297	0.15829	0.43297	0.40670	0.10446	0.00549
	10	0.64835	0.05211	0.46967	0.59101	0.10483	0.00549

Table 11. Probability That Concentrations in Wells Will be Above 25 mg/kg With Different Remediation Strategies Calculated by Using 50% the Computed Probability Measure

		Remediation Strategy					
Well	Time (Year)	None	R1	R2	R3	R4	R5
w8	5	0.78843	0.33309	0.78843	0.77527	0.44397	0.29045
	10	0.90346	0.09366	0.68062	0.88682	0.44534	0.29045
w9	5	0.86213	0.28780	0.86213	0.83038	0.25296	0.08209
	10	0.96227	0.07914	0.71466	0.95335	0.25356	0.08209
w10	5	0.50825	0.20336	0.50825	0.48242	0.14061	0.00807
	10	0.70626	0.06577	0.52442	0.65307	0.14112	0.00807

predicted densities have similar qualitative and quantitative features so are omitted). The blue curves are the predicted densities of the contaminant concentration computed using the computed P_Λ. The black curves are the predicted densities using the uninformed predictive analysis. To make it easier to compare these densities across the various remediation strategies, a reference density showing the predicted density with no remediation using the computed probability measure on Λ is included in each plot as a red dashed curve. We clearly observe a systematic bias of approximately an order of magnitude in the regions of low probability of these densities in almost all cases when using such an uninformed predictive analysis.

5. Conclusions and Future Work

A measure-theoretic framework has been employed to develop and solve the stochastic inverse problem for groundwater contamination. Measured contaminant concentration data from wells are used in posing the stochastic inverse problem. It has been shown that the choice of GD QoI from measurement data used in this process can have a significant effect on the solution. Events of high probability in the parameter space can be identified and analyzed. Furthermore, the probability measure on the parameter space can be used to predict the probabilities of other events, for instance, if the concentration of the contaminant in a well will be above the MCL. It can also be used to calculate probability distributions of model outputs and accurately estimating probabilities of remediation failure that the uninformed predictive analysis inaccurately estimates. This is critical when making decisions under uncertainty in model parameters.

We demonstrated the predictive analysis by propagating entire events of high probability in the parameter space requiring many model solves. Posing this problem in terms of simple integrals over the various events clearly indicates that more efficient Monte Carlo or quasi-Monte Carlo schemes may be used, where we *sample from* the identified events of high probability in parameter space rather than simply propagate every sample used in the construction of the inverse probability measure. This is the subject of on-going work as we implement additional features in the postprocessing subpackage of the BET Python package used for the computation and analysis of the inverse probability measure. Moreover, we are investigating adaptive sampling techniques for improved efficiency in the computation of the inverse probability measure, e.g., improving the placement of parameter samples in generalized contour events of highest probability. Finite element models for the contaminant transport equation are also being developed to solve the problem on more general physical domains. We are simultaneously developing adjoint codes to provide reliable a posteriori error estimates in computed functionals defining both the QoI used for the inverse problem and the

Table 12. Probability That Concentrations in Wells Will Be Above 25 mg/kg With Different Remediation Strategies Calculated by Using 25% the Computed Probability Measure

		Remediation Strategy					
Well	Time (Year)	None	R1	R2	R3	R4	R5
w8	5	0.74246	0.28503	0.74246	0.73050	0.38766	0.16865
	10	0.88842	0.07731	0.66533	0.87206	0.39009	0.16865
w9	5	0.86387	0.25580	0.86387	0.82038	0.14962	0.00167
	10	0.96744	0.07523	0.69289	0.95755	0.14970	0.00167
w10	5	0.46391	0.16014	0.46391	0.43755	0.08649	0.00000
	10	0.66208	0.04786	0.48942	0.60452	0.08662	0.00000

prediction functionals. Such error estimates have previously been used in other applications to produce reliable estimates of the error in the computed probability measure and can be used to guide adaptive sampling strategies (T. Butler et al., Solving stochastic inverse problems using sigma-algebras on contour maps, 1407.3851, 2014).

Effort has been made in the field of uncertainty quantification on using Bayesian and regularization approaches for performing decision analysis under uncertainty while simultaneously addressing sources of uncertainties in these approaches [see *Freeze et al.*, 1990; *Massmann et al.*, 1991; *O'Malley and Vesselinov*, 2014]. One source of uncertainty in the measure-theoretic approach is the choice of ansatz. We are exploring the use of information gap theory [*O'Malley and Vesselinov*, 2014] to quantify the effects of different choices of ansatz and the robustness of decisions made using the computed inverse probability measures under various choices of ansatz.

Acknowledgments
This material is based upon work supported by the U.S. Department of Energy Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under award DE-SC0009286 and DE-SC0009279 as part of the DiaMonD Multifaceted Mathematics Integrated Capability Center. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. URL: http://www.tacc.utexas.edu. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant ACI-1053575 under XSEDE grant TG-DMS080016N. This work is also supported in part by the National Science Foundation DMS-1228206. We note that there are no data sharing issues, since all of the numerical information is provided in the figures and tables produced by the methodology presented in the paper.

References

Abbaspour, K., M. T. van Genuchten, R. Schulin, and E. Schläppi (1997), A sequential uncertainty domain inverse procedure for estimating subsurface flow and transport parameters, *Water Resour. Res.*, *33*(8), 1879–1892.

Agostini, P., A. Critto, E. Semenzin, and A. Marcomini (2009a), Decision support systems for contaminated land management: A review, in *Decision Support Systems for Risk-Based Management of Contaminated Sites*, pp. 1–20, Springer, N. Y.

Agostini, P., G. W. Suter II, S. Gottardo, and E. Giubilato (2009b), Indicators and endpoints for risk-based decision processes with decision support systems, in *Decision Support Systems for Risk-Based Management of Contaminated Sites*, pp. 1–18, Springer, N. Y.

Argent, R. M., J.-M. Perraud, J. M. Rahman, R. B. Grayson, and G. Podger (2009), A new approach to water quality modelling and environmental decision support systems, *Environ. Modell. Software*, *24*(7), 809–818.

Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology, *J. Hydrol.*, *249*(14), 11–29, doi:10.1016/S0022-1694(01)00421-8.

Beven, K., and I. Westerberg (2011), On red herrings and real herrings: Disinformation and information in hydrological inference, *Hydrol. Processes*, *25*(10), 1676–1680.

Bolster, D., M. Barahona, M. Dentz, D. Fernandez-Garcia, X. Sanchez-Vila, P. Trinchero, C. Valhondo, and D. Tartakovsky (2009), Probabilistic risk analysis of groundwater remediation strategies, *Water Resour. Res.*, *45*, W06413, doi:10.1029/2008WR007551.

Breidt, J., T. Butler, and D. Estep (2011), A measure-theoretic computational method for inverse sensitivity problems I: Method and *analysis, SIAM J. Numer. Anal.*, *49*(5), 1836–1859.

Butler, T., and D. Estep (2013), A numerical method for solving a stochastic inverse problem for parameters, *Ann. Nucl. Energy*, *52*, 86–94.

Butler, T., D. Estep, and J. Sandelin (2012), A computational measure-theoretic approach to inverse sensitivity problems II: A posterior error *analysis, SIAM J. Numer. Anal.*, *50*(1), 22–45.

Butler, T., D. Estep, S. Tavener, C. Dawson, and J. Westerink (2014a), A measure-theoretic computational method for inverse sensitivity problems III: Multiple quantities of interest, *SIAM J. Uncertainty Quantification*, *2*(1), 174–202.

Butler, T., L. Graham, D. Estep, C. Dawson, and J. Westerink (2015), Definition and solution of a stochastic inverse problem for the manning's n parameter field in hydrodynamic models, *Adv. Water Resour.*, *78*, 60–79.

Carrera, J., and S. Neuman (1986a), Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information, *Water Resour. Res.*, *22*(2), 199–210.

Carrera, J., and S. Neuman (1986b), Estimation of aquifer parameters under transient and steady state conditions: 2. Uniqueness, stability, and solution algorithms, *Water Resour. Res.*, *22*(2), 211–227.

Carrera, J., and S. Neuman (1986c), Estimation of aquifer parameters under transient and steady state conditions: 3. Application to synthetic field data, *Water Resour. Res.*, *22*(2), 228–242.

Carrera, J., A. Alcolea, A. Medina, J. Hidalgo, and L. Slooten (2005), Inverse problem in hydrogeology, *Hydrogeol. J.*, *13*(1), 206–222, doi: 10.1007/s10040-004-0404-7.

Caselton, W. F., and W. Luo (1992), Decision making with imprecise probabilities: Dempster-Shafer theory and application, *Water Resour. Res.*, *28*(12), 3071–3083.

Dagan, G. (1982), Stochastic modeling of groundwater flow by unconditional and conditional probabilities: 1. Conditional simulation and the direct problem, *Water Resour. Res.*, *18*(4), 813–833.

Delhomme, J. (1979), Spatial variability and uncertainty in groundwater flow parameters: A geostatistical approach, *Water Resour. Res.*, *15*(2), 269–280.

Dellacherie, C., and P. Meyer (1978), *Probabilities and Potential*, North-Holland, Amsterdam, Netherlands.

Freer, J., and K. Beven (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resour. Res.*, *32*(7), 2161–2173.

Freeze, R. A., J. Massmann, L. Smith, T. Sperling, and B. James (1990), Hydrogeological decision analysis: 1. A framework, *Ground Water*, *28*(5), 738–766, doi:10.1111/j.1745-6584.1990.tb01989.x.

Gelhar, L. W., C. Welty, and K. R. Rehfeldt (1992), A critical review of data on field-scale dispersion in aquifers, *Water Resour. Res.*, *28*(7), 1955−1974.

Harp, D. R., and V. V. Vesselinov (2012), An agent-based approach to global uncertainty and sensitivity analysis, *Comput. Geosci.*, *40*, 19–27.

Harp, D. R., and V. V. Vesselinov (2013), Contaminant remediation decision analysis using information gap theory, *Stochastic Environ. Res. Risk Assess.*, *27*(1), 159–168.

Hipel, K. W., and Y. Ben-Haim (1999), Decision making in an uncertain world: Information-gap modeling in water resources management, *IEEE Trans. Syst. Man Cybern.*, *29*(4), 506–517.

Jordan, G., and A. Abdaal (2013), Decision support methods for the environmental assessment of contamination at mining sites, *Environ. Monit. Assess.*, *185*(9), 7809–7832.

Keating, E. H., J. Doherty, J. A. Vrugt, and Q. Kang (2010), Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality, *Water Resour. Res.*, *46*, W10517, doi:10.1029/2009WR008584.

Laloy, E., and J. A. Vrugt (2012), High-dimensional posterior exploration of hydrologic models using multiple-try dream (zs) and high-performance computing, *Water Resour. Res.*, *48*, W01526, doi:10.1029/2011WR010608.
Leube, P. C., A. Geiges, and W. Nowak (2012), Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design, *Water Resour. Res.*, *48*, W02501, doi:10.1029/2010WR010137.
Massmann, J., R. A. Freeze, L. Smith, T. Sperling, and B. James (1991), Hydrogeological decision analysis: 2. Applications to ground-water contamination, *Ground Water*, *29*(4), 536–548, doi:10.1111/j.1745-6584.1991.tb00545.x.
Mayer, A. S., C. Kelley, and C. T. Miller (2002), Optimal design for problems involving flow and transport phenomena in saturated subsurface systems, *Adv. Water Resour.*, *25*(812), 1233–1256, doi:10.1016/S0309-1708(02)00054-4.
National Research Council (1999), *An End State Methodology for Identifying Technology Needs for Environmental Management, With an Example From the Hanford Site Tanks*, The Natl. Acad. Press, Washington, D. C.
National Research Council (2013), *Alternatives for Managing the Nation's Complex Contaminated Groundwater Sites*, The Natl. Acad. Press, Washington, D. C.
Neuman, S. P. (1990), Universal scaling of hydraulic conductivities and dispersivities in geologic media, *Water Resour. Res.*, *26*(8), 1749–1758.
Nowak, W., F. P. J. de Barros, and Y. Rubin (2010), Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain, *Water Resour. Res.*, *46*, W03535, doi:10.1029/2009WR008312.
O'Malley, D., and V. Vesselinov (2014), Groundwater remediation using the information gap decision theory, *Water Resour. Res.*, *50*, 246–256, doi:10.1002/2013WR014718.
Petrucci, R. H., W. S. Harwood, F. G. Herring, and S. S. Perry (1993), *General Chemistry: Principles and Modern Applications*, Macmillan, N. Y.
Reeves, D. M., K. F. Pohlmann, G. M. Pohll, M. Ye, and J. B. Chapman (2010), Incorporation of conceptual and parametric uncertainty into radionuclide flux estimates from a fractured granite rock mass, *Stochastic Environ. Res. Risk Assess.*, *24*(6), 899–915.
Rügner, H., M. Finkel, A. Kaschl, and M. Bittens (2006), Application of monitored natural attenuation in contaminated land management: A review and recommended approach for Europe, *Environ. Sci. Policy*, *9*(6), 568–576.
Schulze-Makuch, D. (2005), Longitudinal dispersivity data and implications for scaling behavior, *Ground Water*, *43*(3), 443–456.
Tartakovsky, D. M. (2007), Probabilistic risk analysis in subsurface hydrology, *Geophys. Res. Lett.*, *34*, L05404, doi:10.1029/2007GL029245.
Tonkin, M., and J. Doherty (2009), Calibration-constrained Monte Carlo analysis of highly parameterized models using subspace techniques, *Water Resour. Res.*, *45*, W00B10, doi:10.1029/2007WR006678.
Troldborg, M., W. Nowak, N. Tuxen, P. L. Bjerg, R. Helmig, and P. J. Binning (2010), Uncertainty evaluation of mass discharge estimates from a contaminated site using a fully Bayesian framework, *Water Resour. Res.*, *46*, W12552, doi:10.1029/2010WR009227.
Vrugt, J., C. ter Braak, H. Gupta, and B. Robinson (2008), Equifinality of formal (dream) and informal (glue) Bayesian approaches in hydrologic modeling?, *Stochastic Environ. Res. Risk Assess.*, *23*(7), 1011–1026.
Wagner, B. J., and S. M. Gorelick (1987), Optimal groundwater quality management under parameter uncertainty, *Water Resour. Res.*, *23*(7), 1162–1174.
Wang, H., and H. Wu (2009), Analytical solutions of three-dimensional contaminant transport in uniform flow field in porous media: A library, *Frontiers Environ. Sci. Eng. China*, *3*(1), 112–128.
Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, *40*, W05113, doi:10.1029/2003WR002557.