

Application of geostatistical inverse modeling to contaminant source identification at Dover AFB, Delaware

Application du modèle inverse en géostatistique pour l'identification de source de pollutions à Douvres AFB, Delaware

ANNA M. MICHALAK, *Visiting Scientist, Climate Monitoring and Diagnostics Laboratory, National Oceanic and Atmospheric Administration (NOAA), Boulder, CO 80305-3328, USA (author for correspondence)*

PETER K. KITANIDIS, *Professor, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA 94305-4020, USA*

ABSTRACT

Analysis of subsurface soil cores from the site of a field-scale groundwater remediation experiment at Dover Air Force Base, Delaware, has revealed that tetrachloroethene and trichloroethene contamination extends into an aquitard underlying a groundwater aquifer. Geostatistical inverse modeling is used to make inferences regarding the historical concentration conditions in the overlying aquifer. Because geostatistical inverse modeling is a stochastic approach, it treats parameters as jointly distributed random fields. Therefore, this approach is used to compute confidence intervals in addition to best estimates. This framework is also used to compute large numbers of conditional realizations, which are equally probable solutions given the data, and which allow for a better understanding of the form of the unknown function. Finally, a Markov Chain Monte Carlo method combined with the application of Lagrange multipliers is used to enforce concentration non-negativity.

RÉSUMÉ

L'analyse d'échantillons (carottes) du site de traitement expérimental des eaux souterraines à l'échelle de la parcelle sur la base aérienne de Douvres, Delaware, a indiqué que la pollution en tetrachloroethene et trichloroethene pénètre dans un aquitard en correspondance avec un aquifère qui surmonte l'aquitard. Un modèle géostatistique inverse est employé pour remonter aux causes, en s'appuyant sur l'historique des concentrations relevées dans l'aquifère sus-jacent. Le modèle géostatistique inverse procède par approche stochastique, et traite des paramètres en tant que champs aléatoires conjointement distribués. Par conséquent, cette approche est employée pour calculer les intervalles de confiance en plus des meilleures estimations des paramètres. Ce cadre est également utilisé pour calculer un grand nombre de solutions conditionnelles qui sont sous-tendues par une égale probabilité en regard des données de départ, et qui permettent une meilleure compréhension de la forme de la fonction inconnue. Enfin, une méthode de Monte Carlo Markov combinée avec l'application des multiplicateurs de Lagrange est employée pour imposer la non-négativité de la concentration.

Keywords: Stochastic inverse modeling, contaminant source identification, inference under constraints, Markov Chain Monte Carlo (MCMC), Metropolis–Hastings algorithm, Bayesian inference.

1 Introduction

In many cases of environmental contamination, the origin or history of contamination is initially unknown. The ability to conclusively identify the source of observed contamination can not only help in the remediation process, but can be critical to the identification of responsible parties, and therefore to the apportionment of any liability associated with a given site.

Methods that are currently available for contaminant source identification differ in their ranges of applicability in terms of contaminants and media, the level of confidence associated with their results, and a variety of other factors. These methods can, however, be subdivided into three categories. The first category is compositional analysis, which determines the source

by analyzing differences in the molecular or isotopic compositions of contaminants among potential sources. The second category is the use of either naturally occurring or introduced tracers discharged with a contaminant at a given potential source. The third category encompasses methods based on conclusions made from the contamination distribution itself, once the transport behavior of the contaminant and the medium in which it is being transported have been quantified. A review of these methods is available in Michalak (2001).

One set of methods based on analyzing the contamination distribution is inverse methods. These methods use modeling and statistical tools to determine either the prior location of observed contamination or the release history from a known source. Reviews of such methods are presented in

Snodgrass and Kitanidis (1997) and Liu and Ball (1999), among others.

One subset of work uses a function estimate to characterize the source location or release history. In this case, the source characteristics are not limited to a set number of parameters, but are instead free to vary in space and in time. This category includes methods that use a deterministic approach and others that offer a stochastic approach to the problem.

The assumption that the model parameters are deterministic but unknown differs from stochastic approaches, where parameters are viewed as jointly distributed random fields. Because there will always be uncertainty in contaminant concentration estimates, release history and release location, it makes sense to treat these quantities as random functions that can be described by their statistical properties. In this framework, estimation uncertainty is recognized and its importance can sometimes be determined.

One of the stochastic methods proposed in the past for the estimation of the release history of a contaminant is the use of the geostatistical approach to inverse modeling. Snodgrass and Kitanidis (1997) estimated the release history of a conservative solute being transported in a one-dimensional (1-D) homogeneous domain, given point concentration measurements at some time after the release.

In this article, we present the first application of the geostatistical approach to contaminant source identification to the interpretation of field data. This work is also the first demonstration of the applicability of this approach to a physically non-uniform domain. Finally, we develop a Markov Chain Monte Carlo (MCMC) method for enforcing concentration non-negativity while maintaining the statistical rigor of the geostatistical methodology.

Aquitard cores taken from the Dover Air Force Base (DAFB) in Delaware are analyzed to infer the contamination history in the overlying aquifer. These data sets have previously been examined by Ball *et al.* (1997) and Liu and Ball (1999). Ball *et al.* (1997) assumed that the history was made up of one-step and two-step constant concentrations at the aquifer/aquitard interface and the times of step concentration changes were estimated from the data. Liu and Ball (1999) applied Tikhonov regularization to obtain a function estimate of the concentration history. However, whereas Tikhonov regularization is a deterministic technique that identifies a single estimate of the source function, a geostatistical approach allows for more in-depth analysis. This method results in a best estimate that is the median of all possible contamination histories, as well as confidence intervals about that best estimate. Furthermore, conditional realizations can be generated which allow for better visualization of the unknown process. Finally, structural parameters that describe the continuity of the contamination-history function and the data measurement error are optimized using the data themselves.

2 Site and data description

The research site is located at DAFB. At the site, an unconfined sand aquifer is underlain by an aquitard, which consists of two

Table 1 Summary of parameters in two-layer aquitard

Physical definition	Parameter	Units	Layer 1 (OSCL)	Layer 2 (DGSL)
Effective diffusivity	D (PCE)	m ² /s	4.2×10^{-10}	4.2×10^{-10}
	D (TCE)	m ² /s	4.9×10^{-10}	4.9×10^{-10}
Retardation factor	R (PCE)	–	2	45
	R (TCE)	–	1.4	20
Porosity	η	–	0.53	0.56
Bulk density	ρ_b	kg/l	1.22	1.15

layers of distinctly different characteristics: an upper layer of orange silty clay loam (OSCL) and a bottom layer of dark gray silt loam (DGSL). Tetrachloroethene (PCE) and trichloroethene (TCE) are two principal chemical contaminants of the overlying aquifer contaminant plume, and concentration profiles for both these chemicals have been obtained in the underlying aquitard at several locations. A detailed description of the site geology and hydrogeology can be found in Ball *et al.* (1997) and Mackay *et al.* (1997). A description of the sampling at the site is available in Liu and Ball (1999). The data sets used for the analysis presented in this work are at locations referred to as PPC11 and PPC13.

The soil core samples were also used to independently determine the sorption properties and porosity of the two aquitard layers (Ball *et al.* 1997). The physical parameters as used by Ball *et al.* (1997) are presented in Table 1. Identical values were used in the current work, in order to facilitate a direct comparison between the two methods.

3 Model

3.1 Physical model

Solute transport in this two-layer aquitard is mainly controlled by a diffusive process. If we assume that each layer is locally homogeneous and that there were no small-scale differences in the forcing at the top of the aquitard, the diffusion can be mathematically described by the following 1-D differential equation:

$$R_1 \frac{\partial c_1^{\text{aq}}}{\partial t} = D_1 \frac{\partial^2 c_1^{\text{aq}}}{\partial x^2} \quad 0 < x < L$$

$$R_2 \frac{\partial c_2^{\text{aq}}}{\partial t} = D_2 \frac{\partial^2 c_2^{\text{aq}}}{\partial x^2} \quad L < x < +\infty$$

where c_1^{aq} and c_2^{aq} are aqueous concentrations, R_1 and R_2 are retardation factors, D_1 and D_2 are effective diffusion coefficients in layer 1 (OSCL) and layer 2 (DGSL), respectively, x is the depth within the aquitard, L is the thickness of the first layer (OSCL) and is 0.74 m for location PPC11 and 0.91 m for location PPC13.

3.2 Inverse model

The objective is to estimate an unknown function. The standard estimation problem may be expressed in the following form:

$$\mathbf{z} = \mathbf{h}(\mathbf{s}, \mathbf{r}) + \varepsilon$$

where \mathbf{z} is an $n \times 1$ vector of observations and \mathbf{s} is an $m \times 1$ "state vector" obtained from the discretization of the unknown function that we wish to estimate. The vector \mathbf{r} contains other parameters needed by the model function $\mathbf{h}(\mathbf{s}, \mathbf{r})$ (e.g. diffusivity). The measurement error is represented by the vector ε . Following geostatistical methodology, \mathbf{s} and ε are represented as random vectors. In this case, the function $\mathbf{h}(\mathbf{s}, \mathbf{r})$ is linear in \mathbf{s} such that

$$\mathbf{h}(\mathbf{s}, \mathbf{r}) = \mathbf{H}\mathbf{s}$$

where \mathbf{H} is a known matrix.

In the 1-D case we have an analytical solution for the forward problem (Liu and Ball, 1999):

$$c(x, T) = \int_0^T s(t) f(x, T - t) dt$$

where c is the total concentration (aqueous and sorbed) and T is measurement time. The source is a function of time and is expressed by $s(t)$. The transfer function $f(x, T - t)$ applies the appropriate weight to the source function:

$$\begin{aligned} f(x, T - t) &= \frac{\eta_1 R_1}{\rho_{b1}} \left(\frac{R_1}{D_1} \right)^{1/2} \\ &\times \sum_{i=0}^{\infty} \vartheta^i \left(\frac{2iL + x}{2[\pi(T - t)^3]^{1/2}} \exp\left(-\frac{R_1(2iL + x)^2}{4D_1(T - t)}\right) \right. \\ &\quad \left. - \vartheta \frac{(2i + 2)L - x}{2[\pi(T - t)^3]^{1/2}} \exp\left(-\frac{R_1[(2i + 2)L - x]^2}{4D_1(T - t)}\right) \right) \\ &0 < x < L \end{aligned}$$

$$\begin{aligned} f(x, T - t) &= \frac{\eta_2 R_2}{\rho_{b2}} \left(\frac{R_1}{D_1} \right)^{1/2} \\ &\times (1 - \vartheta) \sum_{i=0}^{\infty} \vartheta^i \frac{\gamma_i}{2[\pi(T - t)^3]^{1/2}} \exp\left(-\frac{R_1 \gamma_i^2}{4D_1(T - t)}\right) \\ &L < x < \infty \end{aligned}$$

where

$$\begin{aligned} \vartheta &= \frac{\eta_2(D_2 R_2)^{1/2} - \eta_1(D_1 R_1)^{1/2}}{\eta_2(D_2 R_2)^{1/2} + \eta_1(D_1 R_1)^{1/2}} \\ \gamma_i &= (2i + 1)L + \left(\frac{D_1 R_2}{D_2 R_1} \right)^{1/2} (x - L) \end{aligned}$$

where the subscripts 1 and 2 refer to parameter values in the upper and lower aquitard layers, respectively.

Let $x_i, i = 1, \dots, n$ be the n depths at which the measurements are taken, and let us discretize the time domain into m temporal points $t_j, j = 1, \dots, m$, with a time step $\Delta t = (t_m - t_1)/(m - 1)$. All measurements are taken at time $t = T$. In this case, the sensitivity matrix is:

$$\mathbf{H} = \Delta t \begin{bmatrix} f(x_1, T - t_1) & \cdots & f(x_1, T - t_m) \\ f(x_2, T - t_1) & \cdots & f(x_2, T - t_m) \\ \vdots & \ddots & \vdots \\ f(x_n, T - t_1) & \cdots & f(x_n, T - t_m) \end{bmatrix}$$

We shall assume that ε has zero mean and known covariance matrix \mathbf{R} . The covariance of the measurement errors used is

$$\mathbf{R} = \sigma_R^2 \mathbf{I}$$

where σ_R^2 is the variance of the measurement error, and \mathbf{I} is an $n \times n$ identity matrix.

Furthermore, we will model \mathbf{s} , the unknown, as a random vector with expected value

$$E[\mathbf{s}] = \mathbf{Y}\beta$$

where \mathbf{Y} is a known $m \times p$ matrix and β are p unknown drift coefficients. For this problem, a linear but unknown trend in the contaminant release concentration was assumed. Thus

$$\mathbf{Y} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix}$$

The covariance function of \mathbf{s} is

$$\mathbf{Q}(\theta) = E[(\mathbf{s} - \mathbf{Y}\beta)(\mathbf{s} - \mathbf{Y}\beta)^T]$$

where $\mathbf{Q}(\theta)$ is a known function of unknown parameters θ . The generalized covariance matrix \mathbf{Q} was assumed to have cubic form:

$$Q(t_i, t_j | \theta) = \theta |t_i - t_j|^3$$

where $|t_i - t_j|$ is the separation distance (in units of time) and $|$ means "given". Such a covariance results in smooth estimates, in the sense that the second derivatives are minimized.

The approach used to obtain the structural parameters is detailed by Kitanidis (1995). In short, the parameters, in this case θ and σ_R^2 , are estimated by maximizing the probability of the measurements given these parameters:

$$p(\mathbf{z} | \theta, \sigma_R^2) \propto |\Sigma|^{-1/2} |\mathbf{Y}^T \mathbf{H}^T \Sigma^{-1} \mathbf{H} \mathbf{Y}|^{-1/2} \exp\left[-\frac{1}{2} \mathbf{z}^T \Xi \mathbf{z}\right]$$

where

$$\Sigma = \mathbf{H} \mathbf{Q} \mathbf{H}^T + \mathbf{R}$$

$$\Xi = \Sigma^{-1} - \Sigma^{-1} \mathbf{H} \mathbf{Y} (\mathbf{Y}^T \mathbf{H}^T \Sigma^{-1} \mathbf{H} \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{H}^T \Sigma^{-1}$$

and $| |$ denotes matrix determinant.

Once these parameters have been estimated, the minimization problem for estimating the concentration history in the aquifer overlying the aquitard is:

$$J = (\mathbf{z} - \mathbf{H}\mathbf{s})^T \mathbf{R}^{-1} (\mathbf{z} - \mathbf{H}\mathbf{s}) + (\mathbf{s} - \mathbf{Y}\beta)^T \mathbf{Q}^{-1} (\mathbf{s} - \mathbf{Y}\beta)$$

In this case, the vector of observations \mathbf{z} and that of the unknown function \mathbf{s} are:

$$\mathbf{z} = \begin{bmatrix} z(x_1, T) \\ z(x_2, T) \\ \vdots \\ z(x_n, T) \end{bmatrix}, \quad \mathbf{s} = \begin{bmatrix} s(t_1) \\ s(t_2) \\ \vdots \\ s(t_m) \end{bmatrix}$$

The corresponding system that needs to be solved is:

$$\begin{bmatrix} \Sigma & \vdots & \mathbf{H}\mathbf{Y} \\ \cdots & \cdots & \cdots \\ (\mathbf{H}\mathbf{Y})^T & \vdots & 0 \end{bmatrix} \begin{bmatrix} \Lambda^T \\ \cdots \\ \mathbf{M} \end{bmatrix} = \begin{bmatrix} \mathbf{H}\mathbf{Q} \\ \cdots \\ \mathbf{Y}^T \end{bmatrix}$$

where Λ is a $m \times n$ matrix of coefficients and \mathbf{M} is a $p \times m$ matrix of multipliers. The best estimate of the function is

$$\hat{\mathbf{s}} = \Lambda \mathbf{z}$$

and its covariance is

$$\mathbf{V} = -\mathbf{Y}\mathbf{M} + \mathbf{Q} - \mathbf{Q}\mathbf{H}^T\Lambda^T$$

Using geostatistical methodology, it is also possible to generate realizations of the unknown contamination history that are conditional on all the observations. Viewing a number of conditional realizations can aid in visualizing the unknown function and the uncertainty about the best estimate. The procedure for generating conditional realizations is discussed by Gutjahr *et al.* (1994) and Kitanidis (1995). First, an unconditional unconstrained realization $\mathbf{s}_{uu,l}$ is generated. A realization of the error vector ε_l must also be independently generated with zero mean and covariance \mathbf{R} . Then, the conditional unconstrained realization $\mathbf{s}_{cu,l}$ may be found by minimizing

$$\begin{aligned} & (\mathbf{z} + \varepsilon_l - \mathbf{h}(\mathbf{s}_{cu,l}))^T \mathbf{R}^{-1} (\mathbf{z} + \varepsilon_l - \mathbf{h}(\mathbf{s}_{cu,l})) \\ & + (\mathbf{s}_{cu,l} - \mathbf{s}_{uu,l})^T \mathbf{G} (\mathbf{s}_{cu,l} - \mathbf{s}_{uu,l}) \end{aligned}$$

with respect to $\mathbf{s}_{cu,l}$. Here

$$\mathbf{G} = \mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{Y}(\mathbf{Y}^T\mathbf{Q}^{-1}\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{Q}^{-1}.$$

3.3 Enforcing concentration non-negativity

The advantage of using a stochastic approach to the inverse problem is that physically significant confidence intervals and conditional realizations can be obtained in addition to a best estimate of the unknown function. We wish to enforce concentration non-negativity without jeopardizing the statistical rigor of the methodology. The traditional approaches to enforcing constraints have some limitations. The use of Lagrange multipliers (Gill, 1986) is not sufficient to guarantee that the conditional realizations are equiprobable, whereas the use of data transformations (Kitanidis, 1997) may result in confidence intervals that are highly asymmetrical and absolute value dependent, and may exhibit convergence problems.

The proposed methodology involves generating conditional realizations using the cubic-variogram model, and constraining them to be non-negative. We still want these realizations to be equiprobable realizations from the original cubic-variogram-based posterior distribution, however. Therefore, a MCMC method was developed to obtain conditional realizations.

MCMC methods allow for the sampling of probability density functions in multiple dimensions with computational effort that is manageable relative to performing the multidimensional integrations that would otherwise be required. The dimensionality of the posterior pdf is equal to the number of points in the discretized unknown function, m , and can therefore easily be on the order of hundreds. Ensemble properties of conditional realizations can then be used to infer other statistics of the unknown function, such as a best estimate and confidence intervals. Although MCMC algorithms have traditionally been used to estimate model parameters, they are being applied here to the

estimation of the time-dependent boundary condition. Examples of past applications of MCMC methods in a Bayesian context can be found, for example, in Gelman *et al.* (1995), Gamerman (1997) and Carlin and Louis (2000).

One subset of these methods is methods based on the Metropolis–Hastings algorithm (Chib and Greenberg, 1995). These methods involve the use of a candidate-generating pdf to obtain candidate realizations of the unknown function, and the objective function is then used to accept or reject these realizations. The accepted realizations can be shown to be equally likely samples from the pdf of interest.

For the current application, we define $\mathbf{s}_{cc,c}$ as the candidate conditional constrained realization, $\mathbf{s}_{cc,l}$ as the l -th accepted conditional constrained realization, $q(\mathbf{s}_{cc,\cdot}|\mathbf{s}_{cc,\cdot})$ as the transition probability from one constrained conditional realization to another, $U(0,1)$ as a uniform distribution in the range $[0,1]$, and ζ as the acceptance probability of $\mathbf{s}_{cc,c}$. If we initialize with an arbitrary constrained conditional realization $\mathbf{s}_{cc,0}$, the Metropolis–Hastings algorithm proceeds as follows (see, e.g. Chib and Greenberg (1995) for a more detailed discussion):

- Repeat for $l = 1, 2, \dots, N$.
- Generate $\mathbf{s}_{cc,c}$ from $q(\mathbf{s}_{cc,c}|\mathbf{s}_{cc,l})$ and u from $U(0,1)$.
- If $u \leq \zeta(\mathbf{s}_{cc,c}|\mathbf{s}_{cc,l})$
 - set $\mathbf{s}_{cc,l+1} = \mathbf{s}_{cc,c}$
- Else
 - set $\mathbf{s}_{cc,l+1} = \mathbf{s}_{cc,l}$
- Return the values $\{\mathbf{s}_{cc,1}, \mathbf{s}_{cc,2}, \dots, \mathbf{s}_{cc,N}\}$.

The candidate constrained conditional realizations are accepted or rejected based on their posterior probability relative to that of the last accepted realization. The probability of acceptance is defined as:

$$\zeta(\mathbf{s}_{cc,c}|\mathbf{s}_{cc,l}) = \min \left\{ \frac{p''(\mathbf{s}_{cc,c}) q(\mathbf{s}_{cc,l}|\mathbf{s}_{cc,c})}{p''(\mathbf{s}_{cc,l}) q(\mathbf{s}_{cc,c}|\mathbf{s}_{cc,l})}, 1 \right\},$$

where $p''(\mathbf{s}_{cc,\cdot})$ is the posterior probability distribution.

In the current work, an MCMC method is used in combination with the application of Lagrange multipliers. Specifically, a Metropolis–Hastings algorithm is applied to unconstrained conditional realizations $\mathbf{s}_{cu,l}$ generated using linear geostatistical inverse modeling with a cubic variogram, constrained to be non-negative using Lagrange multipliers as barrier functions (Gill, 1986).

A new candidate conditional constrained realization is obtained by first obtaining an unconditional realization \mathbf{u}_c with mean zero and covariance \mathbf{Q} in the standard linear geostatistical manner. The unconditional realizations used to obtain the candidate conditional realizations are sequentially correlated:

$$\mathbf{s}_{uu,c} = \phi \mathbf{s}_{uu,l} + \alpha \mathbf{u}_c$$

where $\mathbf{s}_{uu,l}$ is the unconditional realization used in the generation of the last accepted realization, and

$$0 < \phi < 1, \quad \alpha = \sqrt{1 - \phi^2}$$

This method can be shown to yield unconditional unconstrained realizations $\mathbf{s}_{uu,c}$ that are equally likely realizations from the

covariance matrix \mathbf{Q} (see Appendix). A conditional unconstrained realization $\mathbf{s}_{cu,c}$ is generated from $\mathbf{s}_{uu,c}$ using the geostatistical procedure described earlier, and the candidate conditional constrained realization, $\mathbf{s}_{cc,c}$, is then obtained by applying the method of Lagrange multipliers to $\mathbf{s}_{cu,c}$. The candidate constrained conditional realizations generated in this fashion are accepted or rejected based on their posterior probability relative to that of the last accepted realization $\mathbf{s}_{cc,l}$, according to $\zeta(\mathbf{s}_{cc,c}|\mathbf{s}_{cc,l})$, defined earlier. In our case, the posterior probability distribution is:

$$p''(\mathbf{s}_{cc,\cdot}) \propto \exp \left[-\frac{1}{2}(\mathbf{z} - \mathbf{H}\mathbf{s}_{cc,\cdot})^T \mathbf{R}^{-1}(\mathbf{z} - \mathbf{H}\mathbf{s}_{cc,\cdot}) - \frac{1}{2}(\mathbf{s}_{cc,\cdot} - \mathbf{Y}\beta)^T \mathbf{Q}^{-1}(\mathbf{s}_{cc,\cdot} - \mathbf{Y}\beta) \right]$$

The transition probability $q(\mathbf{s}_{cc,\cdot}|\mathbf{s}_{cc,\cdot})$ defined earlier is approximated by $q(\mathbf{s}_{uu,\cdot}|\mathbf{s}_{uu,\cdot})$, the transition probability from one unconditional realization to the other:

$$q(\mathbf{s}_{uu,c}|\mathbf{s}_{uu,l}) \propto \exp \left[-\frac{1}{2}(\mathbf{s}_{uu,c} - \phi\mathbf{s}_{uu,l})^T \frac{\mathbf{Q}^{-1}}{\alpha^2}(\mathbf{s}_{uu,c} - \phi\mathbf{s}_{uu,l}) \right]$$

The chain is run until the probability space has been appropriately sampled. The chain exhibits stationary properties in terms of the posterior probability of its members from the start, and does not require a wind-up period for which all resulting realizations would be ultimately discarded.

4 Results and discussion

For all the data sets, the structural parameters θ in the generalized covariance function and the variance of the measurement error σ_R^2 were optimized using the method described for the unconstrained geostatistical approach. The optimal parameters are presented in Table 2, and were used for both the unconstrained and constrained solutions. Although these parameters may appear to be quite different for the PCE and TCE data sets, these differences are caused by the difference in the magnitudes of the concentrations of TCE versus PCE in the aquitard as well as in the estimated boundary concentrations. Time was discretized at 1-month intervals. The constrained best estimates and confidence intervals were calculated from chains of 100,000 conditional realizations ($\phi = 0.99$). The efficiency of the algorithm depends strongly on the number of iterations needed to constrain the candidate realizations using the method of Lagrange multipliers. As an example, approximately 5000 realizations of the PCE boundary concentration history at location PPC11 can be generated in an hour on a 2.0 GHz machine.

Table 2 Optimal structural parameter values

Structural parameter	PCE		TCE	
	PPC11	PPC13	PPC11	PPC13
θ [$(\mu\text{g}/\text{l})^2\text{month}^{-3}$]	7.9×10^{-1}	3.4×10^{-2}	16.7×10^2	6.3×10^2
σ_R^2 [$(\mu\text{g}/\text{kg})^2$]	6.1	4.4	2.1×10^4	1.7×10^4

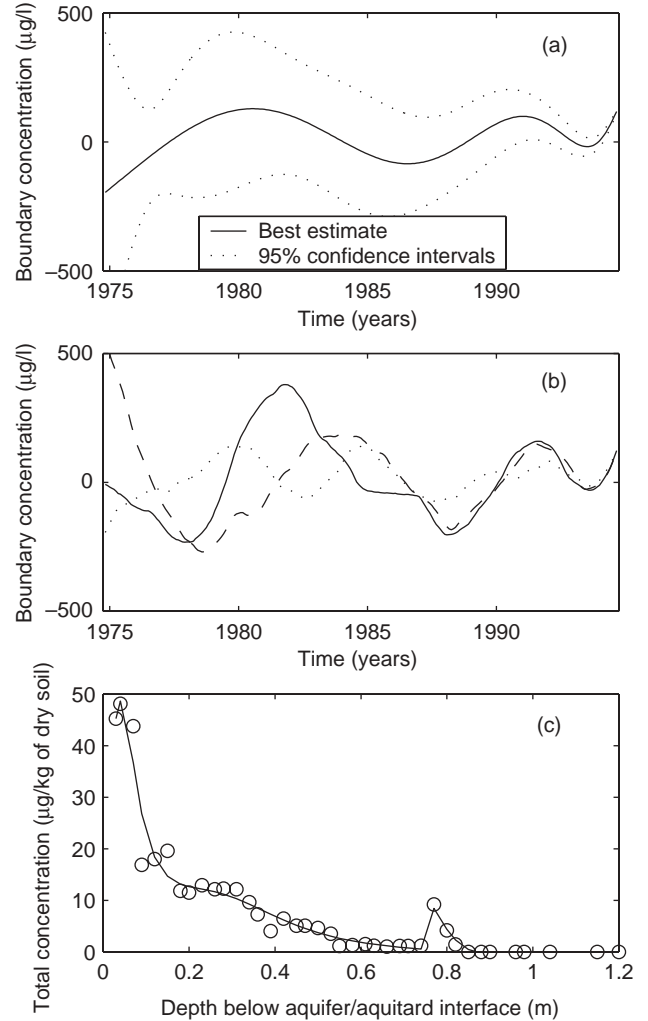


Figure 1 Results of source estimation from PCE data at location PPC11. (a) Estimated time variation of boundary concentration at the interface between the aquifer and aquitard. The end time represents the sampling date (October 27, 1994). (b) Conditional realizations of boundary concentrations. (c) Measurement data and fitted concentrations resulting from the estimated boundary conditions.

4.1 PCE

The results presented in Figs 1 and 2 were generated by applying the unconstrained methodology to the PCE concentration profiles measured at PPC11 and PPC13, respectively. Figures 1(a) and 2(a) show the estimated boundary concentration with 95% confidence intervals, Figs 1(b) and 2(b) show sample conditional realizations, and Figs 1(c) and 2(c) show the actual concentration profiles in the aquitard cores taken at these locations along with the fitted concentrations resulting from the best estimate of the boundary concentration. Figures 3 and 4 show the same information for the case where concentration non-negativity is enforced. In addition, these figures show the results obtained by Liu and Ball (1999) using Tikhonov regularization. As can be seen in Table 3, the fitted concentrations match the measured concentration in the aquitard to a degree consistent with the estimated measurement error, σ_R^2 , for both algorithms.

Our results for location PPC11 are consistent with those presented by Liu and Ball (1999), indicating an increase in PCE concentrations in the aquifer in recent times. The double peak

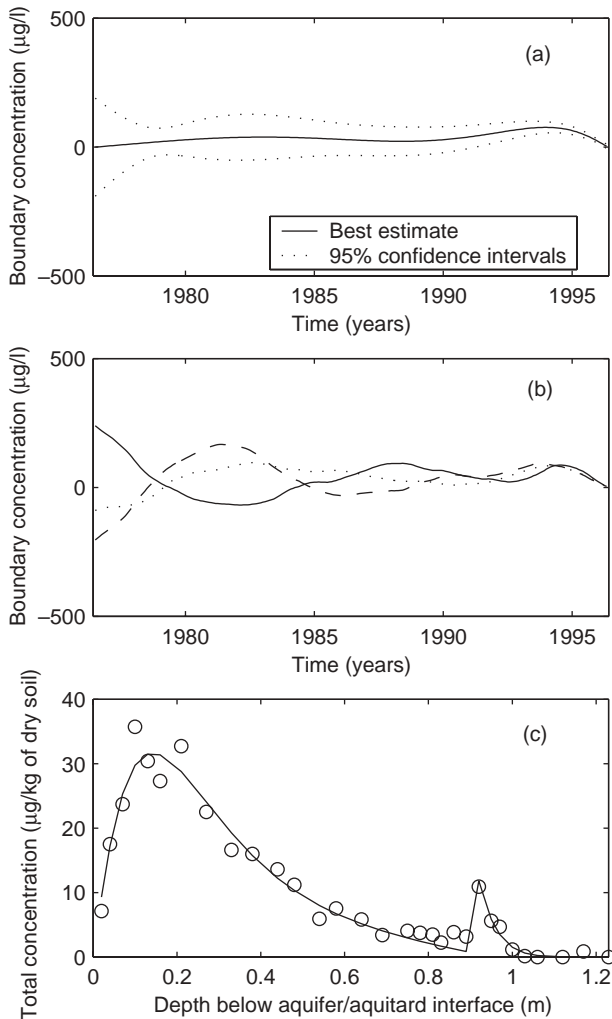


Figure 2 Results of source estimation from PCE data at location PPC13. (a) Estimated time variation of boundary concentration at the interface between the aquifer and aquitard. The end time represents the sampling date (June 6, 1996). (b) Conditional realizations of boundary concentrations. (c) Measurement data and fitted concentrations resulting from the estimated boundary conditions.

found by Liu and Ball (1999) is quite pronounced in the current analysis at location PPC11, but is very mild at location PPC13. However, as can be seen from the conditional realizations presented in Figs 1(a) and 3(a), a double peaked distribution is not the only possible explanation for the observed data.

As noted by Liu and Ball (1999), the boundary conditions estimated using Tikhonov regularization and those estimated using a two-step approach were quite different, and yet reproduced the observed concentration profiles equally well. This point emphasizes the advantages of a stochastic approach, because the uncertainty associated with the estimated boundary conditions can be quantified. The ability to generate confidence intervals and conditional realizations greatly improves the ability to interpret obtained results.

Overall, current results indicate that the diffusive process that led to the contamination of the aquitard, combined with the significant concentration measurement error, result in relatively wide confidence intervals about the estimated contamination history in the overlying aquifer. However, the introduction of additional

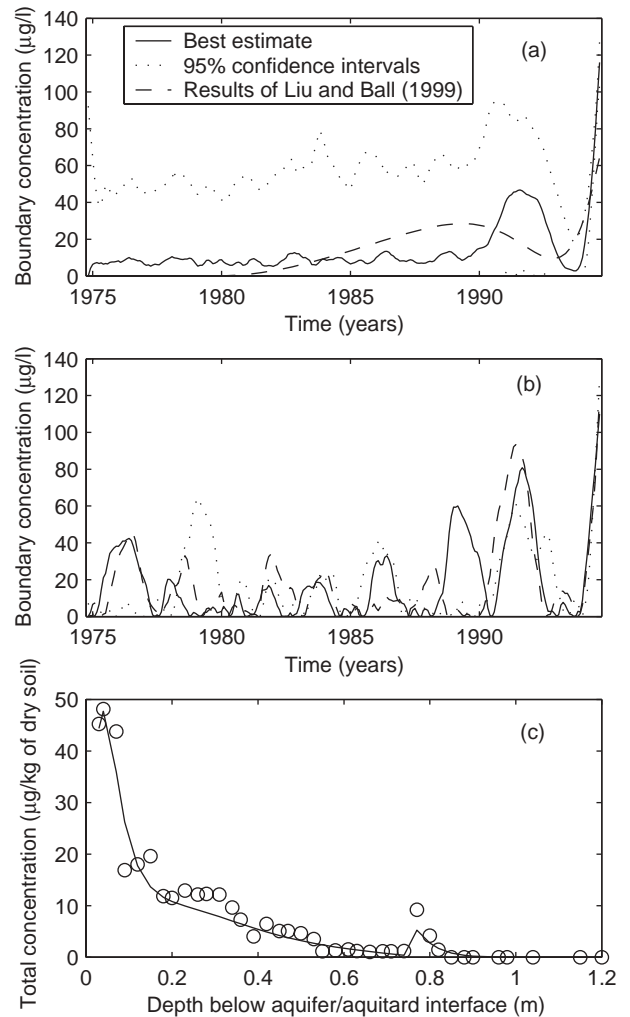


Figure 3 Results of source estimation from PCE data at location PPC11 with non-negativity constraint. (a) Estimated time variation of boundary concentration at the interface between the aquifer and aquitard. The end time represents the sampling date (October 27, 1994). (b) Conditional realizations of boundary concentrations. (c) Measurement data and fitted concentrations resulting from the estimated boundary conditions.

information into the system in the form of a non-negativity constraint greatly reduced the width of the confidence intervals. The results obtained by Liu and Ball (1999) fall within the obtained confidence intervals, but the conditional realizations presented in Figs 3 and 4 show the variety of concentration histories that may have led to the observed aquitard concentration profiles.

4.2 TCE

The results presented in Figs 5 and 6 were generated by applying the unconstrained methodology to the TCE concentrations profiles measured at PPC11 and PPC13, respectively. Figures 5(a) and 6(a) show the estimated boundary concentration with confidence intervals, Figs 5(b) and 6(b) show sample realizations, and Figs 5(c) and 6(c) show the actual concentration profiles in the aquitard cores taken at these locations along with the fitted concentrations resulting from the best estimate of the boundary concentration. Figures 7 and 8 show the same information for the case where concentration non-negativity is enforced, with the addition of results previously obtained by Liu and Ball

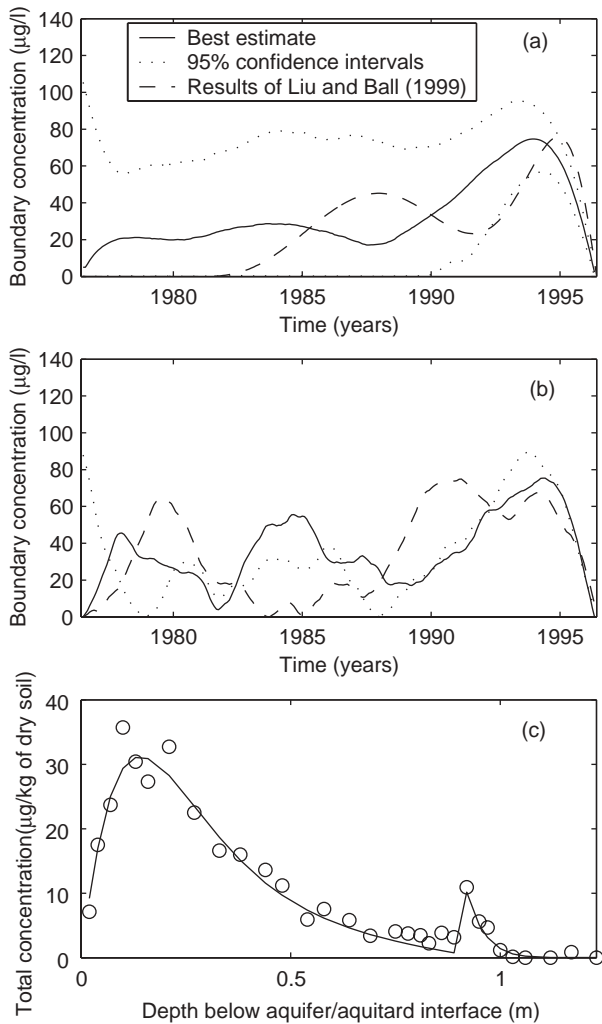


Figure 4 Results of source estimation from PCE data at location PPC13 with non-negativity constraint. (a) Estimated time variation of boundary concentration at the interface between the aquifer and aquitard. The end time represents the sampling date (June 6, 1996). (b) Conditional realizations of boundary concentrations. (c) Measurement data and fitted concentrations resulting from the estimated boundary conditions.

Table 3 Variance of measurement data reproduction error

Variance of data reproduction error σ^2 [(µg/kg) ²]	PCE		TCE	
	PPC11	PPC13	PPC11	PPC13
Linear method	5.3	3.8	1.8×10^4	1.3×10^4
Non-negative method	6.5	3.7	2.9×10^4	1.9×10^4

(1999). Again, the fitted concentrations match the measured concentrations in the aquitard to a degree consistent with the estimated measurement error, σ_R^2 , for both algorithms (see Table 3).

As with the PCE data, the addition of the non-negativity constraint greatly improves the results, as evidenced by the narrower confidence intervals. The results obtained by Liu and Ball (1999) fall within these intervals for location PPC13, but this earlier work had indicated a double peak for location PPC11. This second peak had been inconsistent with results at location PPC13, and does not appear in the current results.

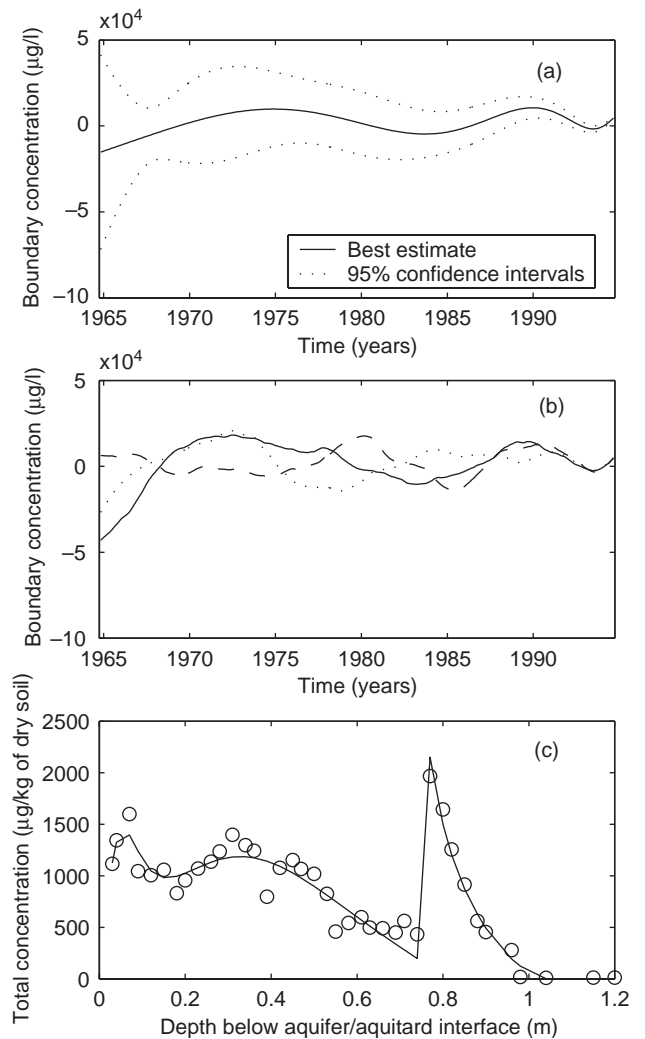


Figure 5 Results of source estimation from TCE data at location PPC11. (a) Estimated time variation of boundary concentration at the interface between the aquifer and aquitard. The end time represents the sampling date (October 27, 1994). (b) Conditional realizations of boundary concentrations. (c) Measurement data and fitted concentrations resulting from the estimated boundary conditions.

The addition of the non-negativity constraint also results in a much clearer signal for the concentration boundary condition. Results from both sampling locations suggest that the TCE concentration in the aquifer peaked around 1989. The magnitude of this peak is also similar for the two data sets.

5 Conclusions

This paper demonstrates the applicability to field data of a stochastic inverse modeling technique based on geostatistical principles. Furthermore, the robustness of a geostatistical approach when applied to a non-uniform domain is demonstrated. Finally, a new method for enforcing concentration non-negativity is developed using a Metropolis–Hastings MCMC algorithm combined with the application of Lagrange multipliers.

Results show that the history of contamination overlying the sampled aquitard at DAFB can be estimated with reasonable precision. This precision is greatly improved by the incorporation

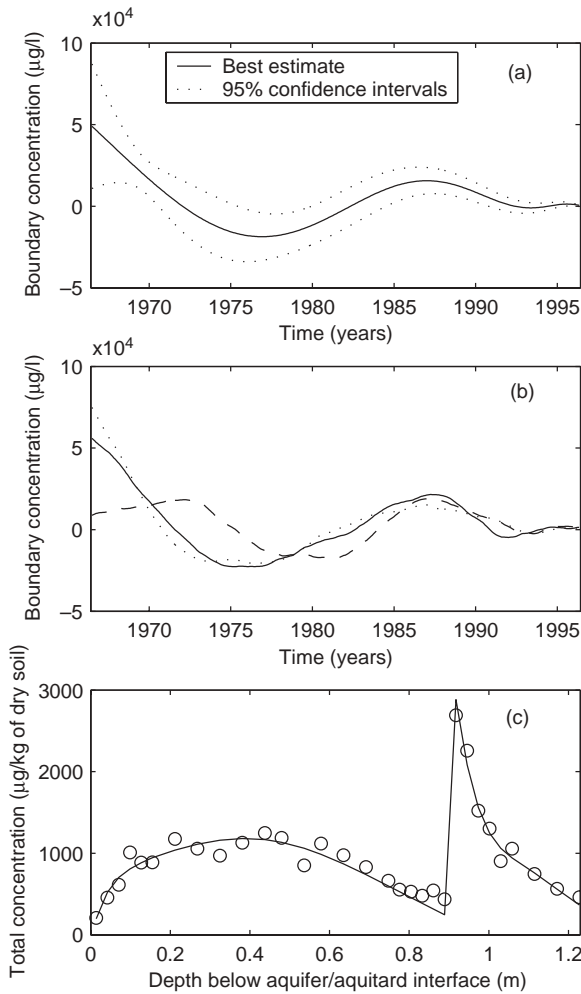


Figure 6 Results of source estimation from TCE data at location PPC13. (a) Estimated time variation of boundary concentration at the interface between the aquifer and aquitard. The end time represents the sampling date (June 6, 1996). (b) Conditional realizations of boundary concentrations. (c) Measurement data and fitted concentrations resulting from the estimated boundary conditions.

of the additional information provided by the non-negativity constraint.

An interesting consideration is that both this study and Liu and Ball (1999) use a regularization term that tends to minimize the second derivative of the function. Given the sparsity of the available measurements, this regularization term could have a significant impact on the boundary concentration estimates. A subsequent study (Michalak and Kitanidis, 2003), however, used a different regularization function that instead tended to minimize concentration differences over short separation times. The conclusions of that study were generally consistent with those found here, indicating that the measurements themselves adequately constrain the concentration histories at the aquifer/aquitard interface for most of the data sets.

Acknowledgments

This research was partially funded by the Natural and Accelerated Bioremediation Research (NABIR) program, Biological and Environmental Research (BER), U.S. Department of Energy

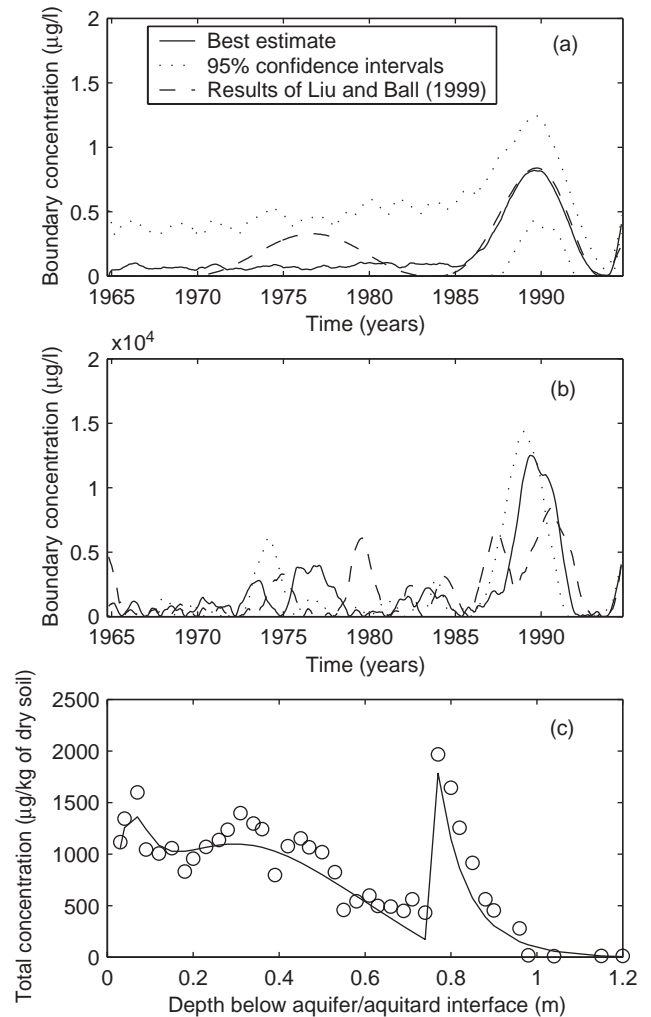


Figure 7 Results of source estimation from TCE data at location PPC11 with non-negativity constraint. (a) Estimated time variation of boundary concentration at the interface between the aquifer and aquitard. The end time represents the sampling date (October 27, 1994). (b) Conditional realizations of boundary concentrations. (c) Measurement data and fitted concentrations resulting from the estimated boundary conditions.

(grant # DE-FG03-00ER63046). We would like to thank Dr Chongxuan Liu for providing us with the data used in this study, and Prof. Peter Glynn for his valuable input.

Appendix

Consider that an unconditional realization $\mathbf{s}_{uu,l}$ has already been generated. We want to generate a new realization which will be close to the previous one. So, $\mathbf{s}_{uu,c}$ is correlated to $\mathbf{s}_{uu,l}$. We will use the following expression:

$$\mathbf{s}_{uu,c} = \phi \mathbf{s}_{uu,l} + \alpha \mathbf{u}_c$$

where ϕ is a real coefficient, for stability $0 \leq \phi < 1$; α is another real coefficient; and \mathbf{u}_c is an unconditional realization generated independently from the covariance matrix \mathbf{Q} . Considering it as a stochastic difference equation, we can compute moments:

$$E[\mathbf{s}_{uu,c}] = \phi E[\mathbf{s}_{uu,l}] + \alpha E[\mathbf{u}_c]$$

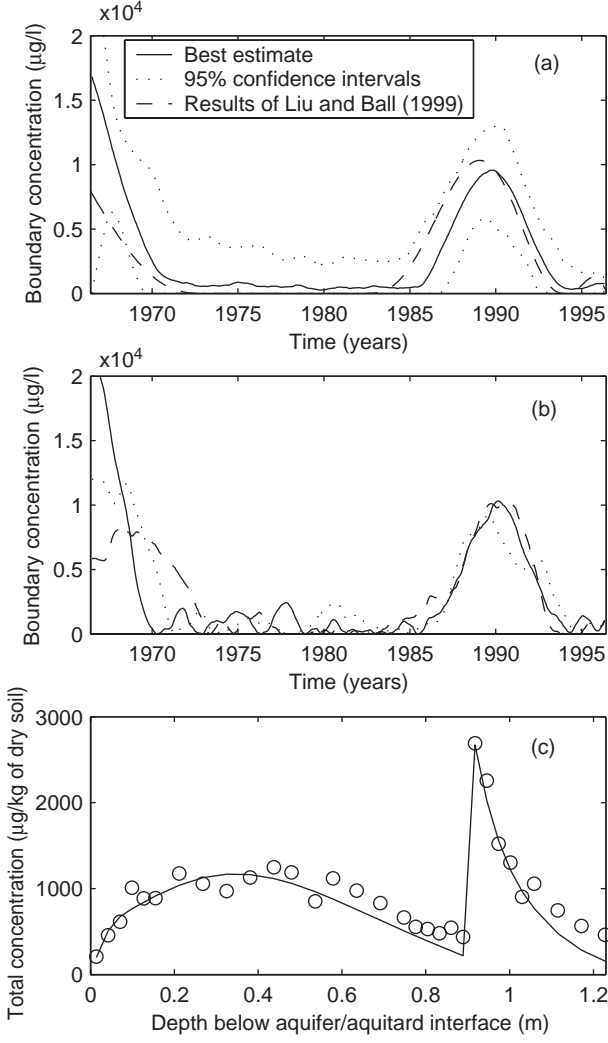


Figure 8 Results of source estimation from TCE data at location PPC13 with non-negativity constraint. (a) Estimated time variation of boundary concentration at the interface between the aquifer and aquitard. The end time represents the sampling date (June 6, 1996). (b) Conditional realizations of boundary concentrations. (c) Measurement data and fitted concentrations resulting from the estimated boundary conditions.

and

$$E[\mathbf{s}_{uu,c}\mathbf{s}_{uu,c}^T] = \phi^2 E[\mathbf{s}_{uu,l}\mathbf{s}_{uu,l}^T] + \alpha^2 E[\mathbf{u}_c\mathbf{u}_c^T] + \phi\alpha E[\mathbf{s}_{uu,l}\mathbf{u}_c^T + \mathbf{u}_c\mathbf{s}_{uu,l}^T]$$

which simplify when we consider the properties of \mathbf{u}_c to:

$$E[\mathbf{s}_{uu,c}] = \phi E[\mathbf{s}_{uu,l}]$$

and

$$E[\mathbf{s}_{uu,c}\mathbf{s}_{uu,c}^T] = \phi^2 E[\mathbf{s}_{uu,l}\mathbf{s}_{uu,l}^T] + \alpha^2 \mathbf{Q}$$

After many realizations, assuming that steady state is obtained, the mean and covariance will be the same at l and c ,

$$E[\mathbf{s}_{uu,c}] = \phi E[\mathbf{s}_{uu,l}] = 0$$

$$E[\mathbf{s}_{uu,c}\mathbf{s}_{uu,c}^T] = E[\mathbf{s}_{uu,l}\mathbf{s}_{uu,l}^T] = \mathbf{Q}$$

provided that we select

$$\alpha = \sqrt{1 - \phi^2}.$$

Notation

Roman symbols

- c = Total concentration (aqueous and sorbed) [$\mu\text{g/kg}$]
- c_1^{aq} = Aqueous concentration in upper aquitard layer [$\mu\text{g/l}$]
- c_2^{aq} = Aqueous concentration in lower aquitard layer [$\mu\text{g/l}$]
- D_1 = Effective diffusivity in upper aquitard layer [m^2/s]
- D_2 = Effective diffusivity in lower aquitard layer [m^2/s]
- \mathbf{H} = Sensitivity matrix ($n \times m$)
- \mathbf{I} = Identity matrix ($n \times n$)
- J = Objective function for linear geostatistical inverse problem
- L = Thickness of the upper aquitard layer [m]
- m = Number of points in discretized unknown function
- \mathbf{M} = Matrix of multipliers from solution of inverse problem ($p \times m$)
- n = Number of observations
- p = Number of drift coefficients
- $q(\cdot|\cdot)$ = Transition probability from one realization to the other
- \mathbf{Q} = Generalized prior covariance matrix of unknown function ($m \times m$)
- R_1 = Retardation factor in upper aquitard layer [-]
- R_2 = Retardation factor in lower aquitard layer [-]
- \mathbf{r} = Parameters needed by model function $\mathbf{h}(\mathbf{s}, \mathbf{r})$
- \mathbf{s} = Discretized unknown function ($m \times 1$)
- $\hat{\mathbf{s}}$ = Best estimate of unknown function ($m \times 1$)
- $\mathbf{s}_{uu,l}$ = l th unconditional unconstrained realization of unknown function \mathbf{s} ($m \times 1$)
- $\mathbf{s}_{uu,c}$ = Unconditional unconstrained realization used in generating candidate realization ($m \times 1$)
- $\mathbf{s}_{cu,l}$ = l th conditional unconstrained realization of unknown function \mathbf{s} ($m \times 1$)
- $\mathbf{s}_{cu,c}$ = Conditional unconstrained realization used in generating candidate realization ($m \times 1$)
- $\mathbf{s}_{cc,l}$ = l -th conditional constrained realization of unknown function \mathbf{s} ($m \times 1$)
- $\mathbf{s}_{cc,c}$ = Candidate conditional constrained realization of unknown function \mathbf{s} ($m \times 1$)
- t = Time
- T = Time at which measurements are taken
- u = Uniformly distributed random number sampled in the range $[0,1]$
- $U[0,1]$ = Uniform distribution in the range $[0,1]$
- \mathbf{u}_c = Independently generated unconditional unconstrained realization ($m \times 1$)
- \mathbf{V} = Posterior covariance of unknown function ($m \times m$)
- \mathbf{Y} = Matrix multiplying drift coefficients ($m \times p$)
- x = Depth below top of aquitard [m]
- \mathbf{z} = Vector of observations ($n \times 1$)

Greek symbols

- α = Multiplier for sequentially correlated unconditional realizations
- β = Vector of drift coefficients ($p \times 1$)
- ε = Measurement error vector ($n \times 1$)

ε_l = l th realization of measurement error vector ε ($n \times 1$)

ϕ = Multiplier for sequentially correlated unconditional realizations

η_1 = Porosity in upper aquitard layer [–]

η_2 = Porosity in lower aquitard layer [–]

Λ = Matrix of coefficients from solution of inverse problem ($m \times n$)

θ = Covariance matrix parameters [$(\mu\text{g}/l)^2(\text{month}^{-3})$]

ρ_{b1} = Bulk density in upper aquitard layer [kg/l]

ρ_{b2} = Bulk density in lower aquitard layer [kg/l]

σ_R^2 = Variance of measurement error [$(\mu\text{g}/\text{kg})^2$]

σ^2 = Variance of measurement reproduction error [$(\mu\text{g}/\text{kg})^2$]

$\zeta(\cdot|\cdot)$ = Probability of acceptance of candidate conditional constrained realization

References

1. BALL, W.P. *et al.* (1997). “A Diffusion-Based Interpretation of Tetrachloroethene and Trichloroethene Concentration Profiles in a Groundwater Aquitard.” *Water Resour. Res.* 33(12), 2741–2757.
2. CARLIN, B.P. and LOUIS, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. Chapman & Hall/CRC, Boca Raton, FL.
3. CHIB, S. and GREENBERG, E. (1995). “Understanding the Metropolis–Hastings Algorithm.” *Am. Statist.* 49(4), 327–335.
4. GAMERMAN, D. (1997). *Markov Chain Monte Carlo*. Chapman & Hall, London.
5. GELMAN, A. *et al.* (1995). *Bayesian Data Analysis*. Chapman & Hall, London.
6. GILL, P.E. (1986). *Practical Optimization*. Academic Press, San Diego.
7. GUTJAHR, A. *et al.* (1994). “Joint Conditional Simulations and the Spectral Approach for Flow Modeling.” *Stoch. Hydrol. Hydraul.* 8(1), 79–108.
8. KITANIDIS, P.K. (1995). “Quasi-Linear Geostatistical Theory for Inversing.” *Water Resour. Res.* 31(10), 2411–2519.
9. KITANIDIS, P.K. (1997). *Introduction to Geostatistics Applications in Hydrogeology*. Cambridge University Press, New York.
10. LIU, C. and BALL, W.P. (1999). “Application of Inverse Methods to Contaminant Source Identification from Aquitard Diffusion Profiles at Dover AFB, Delaware.” *Water Resour. Res.* 35(7), 1975–1985.
11. MACKAY, D.M. *et al.* (1997). Field and Laboratory Studies of Pulsed Pumping for Cleanup of Contaminated Aquifers. Final Rep. AL/EQ-TR-1997-0017, Armstrong Lab. Environics Dir., Tyndall AFB, Fla.
12. MICHALAK, A.M. (2001). “Feasibility of Contaminant Source Identification for Property Rights Enforcement.” In: ANDERSON T.L. and Hill P.J. *The Technology of Property Rights*. Lanham, MD, Rowman and Littlefield Publishers, Inc., pp. 123–145.
13. MICHALAK, A.M. and KITANIDIS, P.K. (2003). “A Method for Enforcing Parameter Nonnegativity in Bayesian Inverse Problems with an Application to Contaminant Source Identification.” *Water Resour. Res.* 39(2), 1033, doi:10.1029/2002WR001480.
14. SNODGRASS, M.F. and KITANIDIS, P.K. (1997). “A Geostatistical Approach to Contaminant Source Identification.” *Water Resour. Res.* 33(4), 537–546.