GEOSTATISTICAL SOLUTION TO THE INVERSE PROBLEM USING SURROGATE FUNCTIONS FOR REMEDIATION OF SHALLOW AQUIFERS

DOMENICO BAÚ[1], ALEX S. MAYER[1]

[1]Dept. of Geological and Mining Engineering & Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton MI 49931, USA.

1. INTRODUCTION

Pump-and-treat (PAT) techniques are often applied to the remediation of dissolved chemicals from shallow aquifers. A related management problem typically consists of the selection of the pumping strategy and the most appropriate treatment method, in order to minimize the total cleanup cost while meeting a set of technical, economic and social constraints. Contaminant flow and transport (FT) models combined with optimization algorithms are used to tackle the management problem. Due to scarcity of information about the hydrogeological system, stochastic modeling approaches are often appropriate. Of primary concern is the inherent spatial variability of hydraulic conductivity.

In general, since the implementation of a remediation strategy assessed based on uncertain hydrogeological parameters leads to a decision involving the risk of constraint violations, operations may be structured into a stochastic optimal control framework, in which the pumping schedules are sequentially updated based upon new measurements collected during the actual cleanup process. The framework requires the implementation of an inverse simulation model to estimate the uncertain hydrogeological parameters based on a set of potential measurements.

In this work, we follow a geostatistical conceptual model where the spatial distribution of hydraulic conductivity is thought of as a realization of a log-normally distributed stationary process, characterized by an exponential covariance function. Using the maximum likelihood (ML) method [12], the parameter estimation problem is solved by determining the set of geostatistical parameters (GP's) – average, variance, and correlation scales. Available data may include direct measurements of hydraulic conductivity, water table elevation at a number of monitoring wells, and contaminant mass extracted from active remediation wells.

A rigorous solution to this optimization problem would require a stochastic FT model to be included in the optimization loop to calculate the expected values and the covariance matrix of the available measurements as functions of the decision variables (DV's). Because of the overwhelming computational effort involved, a surrogate model or response surface is introduced to approximate the objective function (OF). The surrogate model is developed using a multidimensional kriging interpolation over a set of data points obtained from stochastic FT simulations for combinations of GP's determined while simultaneously solving the optimization problem.

2. METHODOLOGY

The framework presented in this work consists of three primary components: a FT model; a stochastic inverse model for parameter estimation, formulated as a single objective optimization problem; and a search algorithm for optimizing the actual OF by surrogates.

2.1. **Flow and transport model.** Groundwater flow and contaminant transport are simulated with a fully 3-D finite element (FE) unsaturated flow code along with a random walk (RW) particle tracking transport code.

Isothermal fluid flow in variably saturated 3-D porous media is governed by the partial differential equation (PDE) obtained by substituting Darcy's law into the continuity equation [11]. Using the pressure head, ψ, as the dependent variable, the PDE reads:

$$\frac{\partial}{\partial x_i}\left[K_{ij}\cdot K_{rw}\left[S_w(\psi)\right]\cdot\left(\frac{\partial\psi}{\partial x_j}+\eta_j\right)\right]=\left[S_w(\psi)\cdot S_s+\phi\cdot\frac{dS_w}{d\psi}\right]\cdot\frac{\partial\psi}{\partial t}-q \tag{1}$$

where: x_i represents the reference system coordinates ($x_1\equiv x; x_2\equiv y; x_3\equiv z$) (L), t is time (T); K_{ij} is the saturated hydraulic conductivity tensor (L/T); K_{rw} indicates the (dimensionless) relative conductivity with respect to water phase; $S_w(\psi)$ represents the saturation of water (dimensionless); S_s is the specific elastic storage (1/L); ϕ is the porosity (dimensionless); q represents a source/sink (per unit volume) term (1/T); and $\eta_1=\eta_2=0$, $\eta_3=1$ represents the gravity term. PDE (1) is solved using a Galerkin FE discretization in space and a finite difference discretization in time. From Darcy's law, the component of pore velocity along x_i is given by:

$$v_i=-\frac{K_{ij}\cdot K_{rw}}{\theta}\cdot\left(\frac{\partial\psi}{\partial x_j}+\eta_j\right) \tag{2}$$

where $\theta=\phi\cdot S_w$ is the soil moisture content.

The transport of conservative solutes in partially saturated porous media relies on the PDE [2]:

$$\frac{\partial}{\partial x_i}\left[D_{ij}\cdot\frac{\partial C}{\partial x_j}\right]-\frac{\partial(v_i\cdot C)}{\partial x_i}=\frac{\partial C}{\partial t}-\frac{q\cdot C^*+f}{\theta} \tag{3}$$

where: C is the solute concentration (M/L^3); C^* is the concentration associated with q (M/L^3); and f represents the specific solute mass released with no fluid exchange (M/L^3/T). D_{ij}, the classic hydrodynamic dispersion tensor (L^2/T) [2]. The transport PDE (3) is solved with a particle tracking method based on a RW algorithm [14]. A detailed description of the FT model used in this study may be found in [1].

2.2. **Stochastic inverse model.** In this paper, we consider the problem of estimating hydrogeological parameters, in particular hydraulic conductivity, following a geostatistical approach. The typical heterogeneous distribution of hydraulic conductivity in groundwater systems is modeled as a single realization of a spatial stochastic process described by a probability density function (PDF). Under the assumption of ergodicity and stationarity [9], the joint PDF for conductivities at all points throughout the aquifer is invariant with respect to translation through space. In other words, the mean of the random field is constant in space, and the covariance at any two points depends only on the separation vector.

Throughout this analysis, the hydraulic conductivity tensor, K_{ij}, in Equation (1) is specified as an isotropic correlated random field following a normal, or Gaussian, distribution [6], along with an exponential covariance model in a log-transformed space [7]:

$$\text{(a)}\ \ K_{ij} = K \cdot \delta_{ij}\ \ ;\ \ \text{(b)}\ \ \log K = N\left(\mu_{\log K}, \sigma_{\log K}\right)\ \ ;\ \ \text{(c)}\ \ cov\left(\mathbf{d}\right) = \sigma^2_{\log K} \cdot \exp\sqrt{\sum_{i=1}^{3} \frac{d_i^2}{\lambda_i^2}} \tag{4}$$

where: δ_{ij} is the Kronecker delta function; $\mu_{\log K}$ and $\sigma_{\log K}$ are the average and the standard deviation of K in the log space, respectively; λ_i's are the spatial correlation scales in the coordinate directions; and d_i are the components of the distance vector $\mathbf{d}$. Under the assumptions of stationarity, lognormality, and exponential covariance, the log-K random field is statistically characterized by the five parameters $\mu_{\log K}$, $\sigma_{\log K}$, λ_1, λ_2 , and λ_3.

The stochastic inverse method used here is based upon the ML estimation [12] as previously introduced by [4, 8, 10]. By adopting the ML approach, the inverse problem is structured into an optimization framework where the negative log likelihood (NLL) is to be minimized as a function of a set of DV's, namely the GP's. Therefore, under the Gaussian assumption, the framework may be stated as:

$$\min\left[L\left(\mathbf{z}|\boldsymbol{\theta}\right) = \frac{N}{2} \cdot ln(2\pi) + \frac{1}{2} \cdot ln\left|\mathbf{Q}\right| + \frac{1}{2} \cdot \left(\mathbf{z} - \overline{\mathbf{z}}\right)^T \cdot \mathbf{Q}^{-1} \cdot \left(\mathbf{z} - \overline{\mathbf{z}}\right)\right] \tag{5}$$

where $L\left(\mathbf{z}|\boldsymbol{\theta}\right)$ is the NLL, that is, the OF; $\mathbf{z}$ is the vector of and N the number of available measurements; $|\ |$ denotes determinant; $\boldsymbol{\theta} \equiv (\mu_{\log K}, \sigma_{\log K}, \lambda_1, \lambda_2, \lambda_3)$ is the vector of GP's. $\overline{\mathbf{z}}$ and $\mathbf{Q}$ are the expected value and the covariance matrix of measurements:

$$\text{(a)}\ \ \overline{\mathbf{z}} = E\left[\mathbf{z}|\boldsymbol{\theta}\right]\ \ ;\ \ \text{(b)}\ \ \mathbf{Q} = E\left[\left(\mathbf{z} - \overline{\mathbf{z}}\right)^T \cdot \left(\mathbf{z} - \overline{\mathbf{z}}\right)|\boldsymbol{\theta}\right] \tag{6}$$

(6.a) and (6.b) are the unknown components of the NLL (5), which can be calculated as functions of $\boldsymbol{\theta}$ using a stochastic simulation approach, or Monte Carlo method. With this method, a large number, N_{MC}, of equally likely realizations of the hydraulic conductivity field $(\mathbf{K}_k; k = 1, 2, .., N_{MC})$ is generated accordingly with the prescribed GP's, $\boldsymbol{\theta}$. For each realization of the ensemble, the FT model is run to calculate the values correspondent to the available data at measurement locations and times, which are in turn used to obtain the components of $\overline{\mathbf{z}}$ and $\mathbf{Q}$:

$$\text{(a)}\ \ \overline{z}_i = \sum_{k=1}^{N_{MC}} \frac{z_i^{(k)}}{N_{MC}}\ \ ;\ \ \text{(b)}\ \ Q_{ij} = \sum_{k=1}^{N_{MC}} \frac{\left(z_i^{(k)} - \overline{z}_i\right) \cdot \left(z_j^{(k)} - \overline{z}_j\right)}{N_{MC} - 1}\ ;\ (i, j = 1, 2, .., N) \tag{7}$$

where z_i indicates the value of the simulated datum at the generic measurement location and time i. For PAT management problems, available data may include: direct measurements of hydraulic conductivity at prescribed locations; water table elevation at a number of monitoring wells and given times; and contaminant mass extracted from active remediation wells at prescribed times over the investigated remedial horizon. Hence vector $\overline{\mathbf{z}}$

and the matrix $\mathbf{Q}$ take on the following structures:

$$\text{(a)}\ \ \overline{\mathbf{z}} = \begin{bmatrix} \overline{\mathbf{z}}_{\log K} \\ \overline{\mathbf{z}}_{H} \\ \overline{\mathbf{z}}_{m_C} \end{bmatrix} \ \ ; \ \ \text{(b)}\ \ \mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{\log K,\log K} & \mathbf{Q}_{\log K,H} & \mathbf{Q}_{\log K,m_C} \\ \mathbf{Q}_{H,\log K} & \mathbf{Q}_{H,H} & \mathbf{Q}_{H,m_C} \\ \mathbf{Q}_{m_C,\log K} & \mathbf{Q}_{m_C,H} & \mathbf{Q}_{m_C,m_C} \end{bmatrix} \tag{8}$$

In (8.a), $\overline{\mathbf{z}}_{\log K}$, $\overline{\mathbf{z}}_{H}$, and $\overline{\mathbf{z}}_{m_C}$ are the subvectors including the expected values of measurements of log-hydraulic conductivity, water table elevation, and contaminant mass extracted from active wells, respectively. In (8.b), the submatrices, $\mathbf{Q}_{\log K,\log K}$, $\mathbf{Q}_{H,H}$, and $\mathbf{Q}_{m_C,m_C}$, along the main diagonal, represent the covariance of the three sets of measurements under consideration; whereas the extradiagonal submatrices, $\mathbf{Q}_{\log K,H}$, $\mathbf{Q}_{\log K,m_C}$, and $\mathbf{Q}_{H,m_C}$, are the cross-covariance matrices of measurements. The physics of groundwater flow and transport enters the ML estimation through $\overline{\mathbf{z}}$ and $\mathbf{Q}$. In particular, running the stochastic FT model is required to estimate the terms in Equations (7.a) and (7.b) involving measurements other than direct measurements of hydraulic conductivity.

Framework (5) constitutes a nonlinear optimization problem, which, in principle, might be solved by incorporating the stochastic FT model into an optimization algorithm. However, because of the large number of realizations required, the stochastic simulation approach entails an overwhelming computational burden, which can be eased by developing a less expensive surrogate of the NLL, $\widehat{L}\,(\mathbf{z}|\boldsymbol{\theta})$.

2.3. **Surrogate search algorithm.** Following the iterative process schematically represented in the flow chart of Figure 1, a surrogate form of the NLL is developed simultaneously with its minimization. The optimal set of GP's, $\boldsymbol{\theta}$, is obtained by:

(a) performing an initial series of stochastic FT simulations over pre-established sets of GP's;
(b) applying appropriate estimation techniques to define the NLL as a function over the range of the DV space;
(c) estimating the surrogate minimum using the current level of knowledge of the NLL;
(d) running a stochastic FT simulation using the GP's correspondent to the point of minimum calculated at the previous step;
(e) possibly updating the above set of NLL data points by including the point of minimum calculated in (d).

Steps (b) through (e) are to be iterated until the point of minimum converges to to the set of GP's that minimizes the actual NLL, unless a predefined maximum overall number of stochastic FT simulations is reached.

The procedure of Figure 1 has some similarity with that presented by [3], who developed a rigorous optimization method for solving single-objective optimization problems using surrogates, to apply when the evaluation of the OF requires in general an expensive computational analysis.

In this work, kriging interpolation is used to estimate the surrogate NLL (box (b) in Figure 1). Kriging is chosen because: its implementation is relatively parsimonious; it is an exact interpolator (it honors the true NLL at least for the data points used in its calibration); and it allows for a direct calculation of the variance of the interpolation error, which may in theory be considered to locate the initial set of data points that

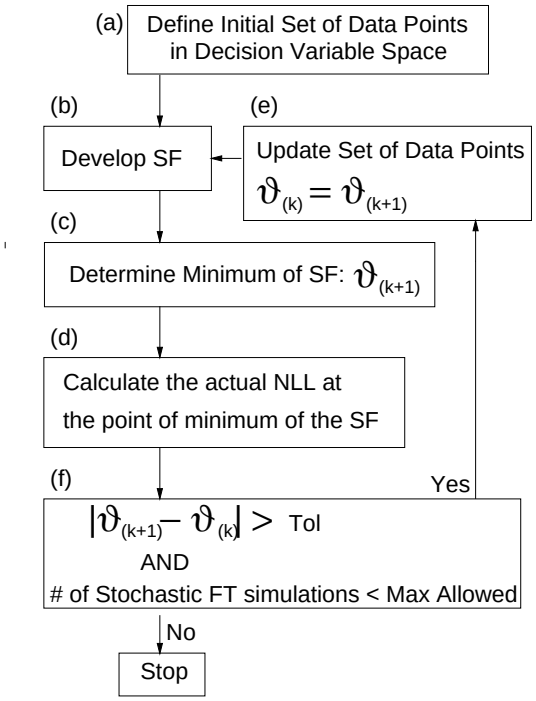

FIGURE 1. Flow chart for the procedure of calibration of the SF and simultaneous minimization of the NLL.

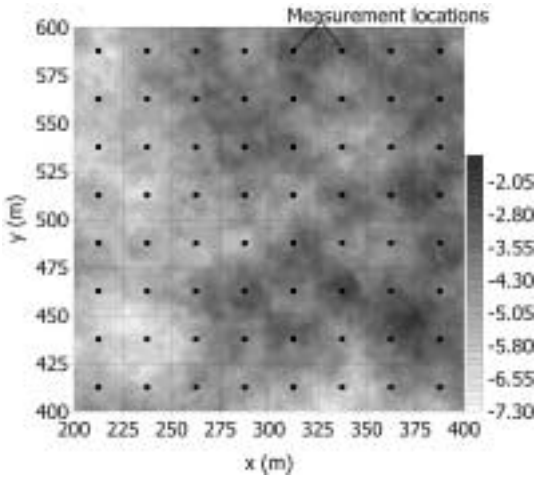

FIGURE 2. Spatial distribution of log-hydraulic conductivity at 10 m depth below the ground surface, within a 200×200 (m×m) region of the considered aquifer. The locations of boreholes where the aquifer has been sampled is also shown.

minimizes the max interpolation error (box (a) in Figure 1). With the kriging technique, for a given a number N_K of data points $[\boldsymbol{\theta}_i; L(\boldsymbol{\theta}_i)]$ $(i = 1, 2, .., N_K)$, the surrogate NLL, $\widehat{L}$, is evaluated with the following linear relationship:

$$\widehat{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N_K} \alpha_i^L [\boldsymbol{\theta}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, .., \boldsymbol{\theta}_{N_K}] \cdot L(\boldsymbol{\theta}_i) \tag{9}$$

Coefficients α_i^L's are obtained from the solution of a linear set of equations having the classical form described in [5]. The coefficients of the system matrix and the right-hand side term are calculated using the variogram function $\gamma = \gamma(\mathbf{d}')$, where $\mathbf{d}'$ is the generic distance vector in the DV space. The variogram represents the parameter that needs to be calibrated for a correct estimation of the considered function. In this study, a statistical inference analysis conducted on the available set of data indicates that the variogram is best fitted by an anisotropic power model:

$$\gamma(\mathbf{d}') = a_L \cdot [\mathbf{d}' \cdot \mathbf{s}]^{b_L} \tag{10}$$

where $\mathbf{s}$ is a scaling vector depending on the relative size of the DV space along the coordinate directions. The coefficients a_L and b_L are determined from a least-square linear regression over the dataset $\left\{ \log\left(\mathbf{d}'_{ij} \cdot \mathbf{s}\right); \log \frac{\sum_{k=1}^{N_{ij}} [L(\boldsymbol{\theta}_i) - L(\boldsymbol{\theta}_j)]^2}{2 \cdot N_{ij}} \right\}$ [5], where N_{ij} is the overall number of pairs of data points at a distance $\mathbf{d}'_{ij}$ from each other.

3. EXAMPLE APPLICATION

The robustness of the search algorithm described in the previous section is tested with a simple example described in the following.

Figure 2 shows the spatial distribution of saturated log-hydraulic conductivity within a 200×200 (m×m) window from a 30-m thick shallow aquifer characterized by the GP's reported in the first row of Table 1. The distribution is synthetically created using a turning band random generator [13] on a regular 3-D grid with a 1×1×0.15 (m×m×m) gridblock size. This region of the aquifer is considered for a hypothetical sampling of the hydraulic conductivity field at the locations corresponding to the nodes centered in the 25×25×3.75 (m×m×m) gridblocks of the subgrid obtained by subdiving the considered portion of aquifer into 64 areal locations, with 8 equally spaced depths (Figure 2).

No measurements of water table elevation and contaminant mass extracted from pumping wells are supposed to be available. As a consequence, the inverse problem is essentially reduced to fitting the sampled hydraulic conductivities into a log-normally distributed stationary process characterized by an exponential covariance function, and no stochastic FT simulation is required to compute the NLL (box (d) in Figure 1).

By using the available set of N=64×8=512 direct measurements of hydraulic conductivity, the algorithm of Figure 1 is applied to determine the vector of GP's, $\boldsymbol{\theta}_{opt}$, minimizing the NLL (5).

In this case, we consider reasonable: a) to search for $\boldsymbol{\theta}_{opt}$ into a bounded domain, Ω, of the DV space, the limits of which are given in the second row of Table 1; b) to solve the surrogate framework (box (c) in Figure 1) following a discrete optimization approach. Therefore, in order to make sure no local minimum may be found when applying integer programming algorithms, the minimum of the kriged NLL is found by complete enumeration of possible combinations of GP's. If n_s is the resolution index, that is, the number of subdivions of the search domain, Ω, along the generic coordinate direction, the complete enumeration of sets of GP's amounts to $(n_s + 1)^5$ combinations. The use of the complete enumeration approach is obviously feasible and justified if the capability of available computer processors admits the choice of sufficiently high values of n_s.

Preliminary numerical tests reveal that the eventual convergence of the search algorithm is strictly connected to the number and locations of data points chosen to initially krig the NLL (box (a) in Figure 1). The most convenient choice appears to be that of 32 points arranged as the vertices of a hyperbox completely contained in the search domain Ω. The first and the last of these initial data points are shown in the upper rows of Table 2. Table 2 shows also the succession of points of minimum of the NLL (5) determined applying the search algorithm with a resolution index n_s=8. The estimation of the GP's, whose actual values are reported in Table 1, is fairly precise and could be further improved by reducing the size of the search domain around the estimated minimum.

The method described herein is indisputably advantageous as it allows for assessing the GP's with a relatively low number of stochastic FT runs. However, we must point out that the procedure is based on the calibration of the SF using points of minimum determined using the SF itself. This is an aggressive technique and, in general, there exists the possibility that the algorithm may be converging to sub-optimal minima of

$\mu_{\log K}$ (-)	$\sigma_{\log K}$ (-)	λ_1 (m)	λ_2 (m)	λ_3 (m)
-4.3	1	50	50	7.5
$[-1.6, -7]$	$[0.5, 2.5]$	$[20, 140]$	$[20, 140]$	$[3, 21]$

TABLE 1. Geostatistical parameters characterizing the actual spatial distribution of hydraulic conductivity, and limits of the bounded domain Ω wherein the solution of framework (5) is searched for.

#	$\mu_{\log K}$ (-)	$\sigma_{\log K}$ (-)	λ_1 (m)	λ_2 (m)	λ_3 (m)	$L(\boldsymbol{\theta})$ (-)	$\widehat{L}(\boldsymbol{\theta})$ (-)
1	-6.00	0.50	30	30	5.00	852.0675	852.0675
..	..	..	..	..	..	..	..
32	-3.00	2.00	100	100	15.00	574.8741	574.8741
33	-4.30	2.25	80	80	12.00	644.3362	236.2225
34	-4.30	1.25	50	50	7.50	535.8565	-201.8550
35	-7.00	2.50	140	140	21.00	592.6189	-176.5655
36	-2.95	2.50	110	125	18.75	615.7436	401.0913
37	-7.00	2.50	110	95	16.50	638.9171	449.4191
38	-4.30	1.25	35	35	5.25	570.4357	516.9619
39	-5.65	2.50	110	140	14.25	623.2350	526.4645
40	-4.30	1.50	50	50	7.50	579.8497	527.8028
41	-4.30	1.25	50	50	7.50	535.8565	535.8565

TABLE 2. Initial set of 32 data points chosen to initially krig the NLL, and succession of points of minimum of $\widehat{L}$ obtained with the surrogate search algorithm.

the SF instead of the absolute optimum of the NLL. The investigation on the possible occurance of such "anomalies" is part of our ongoing research.

4. CONCLUSIONS

In this work, we presented a stochastic inverse simulation model to estimate the GP's of the hydraulic conductivity distribution in groundwater systems subject to PAT remediation. The inverse problem is organized into a nonlinear, single-objective, optimization framework where the problem is to determine the GP's that minimize the NLL of available measurements. In order to avert the use of a computationally expensive stochastic FT model in conjunction with an optimization algorithm, a surrogate form of the NLL is developed and simultaneously used in the search of the GP's that, according to the ML criterium, meet the available measurements.

As initial stage in the process of testing the robustness of the surrogate search algorithm, we considered the case when only a, yet consistent, number of direct measurements of hydraulic conductivity is available.

In general, the method turns out to be computationally efficient and produces results that well agree with the actual geostatistical distribution of hydraulic conductivity. However, numerical experiments show that the effectiveness of the search algorithm depends

strictly on the number and locations of data points chosen to initialize the NLL. In particular, since the the SF is progressively calibrated using points of NLL minimum determined using the SF itself, the search algorithm might in fact converge to sub-optimal minima of the SF. We are currently investigating on the conditions under which this may occur.

The stochastic inverse model presented in this work is in fact one of the key components of a stochastic optimal control framework being developed for the management of the PAT remediation of contaminated aquifers. Following this framework, the predefined remedial horizon is subdivided into a number of intervals of time, in which the pumping patterns are sequentially updated based on new data collected during the actual cleanup process. Within the same framework, the stochastic inverse model is also used to study the worth of acquisition of further data, so as to reduce the cost of the cleanup process and/or the risk of constraint violations.

References

1. D. A. Baú and A. S. Mayer, *Stochastic management of pump-and-treat strategies using surrogate functions*, Adv. Water Resources (2006), In press.
2. J. Bear, *Hydraulics of groundwater*, McGraw-Hill, New York, 1979.
3. A. J. Booker, J. E. Dennis, P. D. Frank, D. B. Serafini, V. Torczon, and M. W. Trosset, *A rigorous framework for optimization of expensive functions by surrogates*, Structural Optimization **17** (1999), 1–13.
4. G. Dagan, *Stochastic modeling of groundwater flow by conditional and unconditional probabilities*, Water Resour. Res. **23** (1985), no. 1, 65–72.
5. G. de Marsily, *Quantitative hydrology: Groundwater hydrology for engineers*, Academic Press, Inc., New York, 1986.
6. R. A. Freeze, *A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogeneous media*, Water Resour. Res. **11** (1975), no. 5, 725–741.
7. R. J. Hoeksema and P. K. Kitanidis, *Analysis of the spatial structure of properties of selected aquifers*, Water Resour. Res. **21** (1985a), no. 4, 563–572.
8. ———, *Comparison of gaussian conditional mean and kriging estimation in the geostatistical solution of the inverse problem*, Water Resour. Res. **21** (1985b), no. 6, 825–836.
9. A. G. Journel and Ch. J. Huijbregts, *Mining Geostatistics*, Academic, San Diego, CA, 1978.
10. P. K. Kitanidis and E. G. Vomvoris, *A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations*, Water Resour. Res. **19** (1983), no. 3, 677–690.
11. J. R. Philip, *Theory of infiltration*, Adv. Hydrosci. **5** (1969), 215–296.
12. S. C. Schweppe, *Model identification*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
13. A. F. B. Tompson, R. Ababou, and L. W. Gelhar, *Implementation of the three-dimensional turning bands random field generator*, Water Resour. Res. **25** (1989), no. 10, 2227–2243.
14. A. F. B. Tompson, E. G. Vomvoris, and L. W. Gelhar, *Numerical simulation of solute transport in randomly heterogeneous porous media: motivation, model development, and application*, Rep. UCID-21281, Lawrence Livermore Natl. Lab., Livermore, Calif., 1987.