Identification of hydraulic conductivity structure in sand and gravel aquifers: Cape Cod data set

J. R. Eggleston, S. A. Rojstaczer, and J. J. Peirce
Center for Hydrologic Science, Duke University, Durham, North Carolina

Abstract. This study evaluates commonly used geostatistical methods to assess reproduction of hydraulic conductivity (K) structure and sensitivity under limiting amounts of data. Extensive conductivity measurements from the Cape Cod sand and gravel aquifer are used to evaluate two geostatistical estimation methods, conditional mean as an estimate and ordinary kriging, and two stochastic simulation methods, simulated annealing and sequential Gaussian simulation. Our results indicate that for relatively homogeneous sand and gravel aquifers such as the Cape Cod aquifer, neither estimation methods nor stochastic simulation methods give highly accurate point predictions of hydraulic conductivity despite the high density of collected data. Although the stochastic simulation methods yielded higher errors than the estimation methods, the stochastic simulation methods yielded better reproduction of the measured ln (K) distribution and better reproduction of local contrasts in ln (K). The inability of kriging to reproduce high ln (K) values, as reaffirmed by this study, provides a strong instigation for choosing stochastic simulation methods to generate conductivity fields when performing fine-scale contaminant transport modeling. Results also indicate that estimation error is relatively insensitive to the number of hydraulic conductivity measurements so long as more than a threshold number of data are used to condition the realizations. This threshold occurs for the Cape Cod site when there are approximately three conductivity measurements per integral volume. The lack of improvement with additional data suggests that although fine-scale hydraulic conductivity structure is evident in the variogram, it is not accurately reproduced by geostatistical estimation methods. If the Cape Cod aquifer spatial conductivity characteristics are indicative of other sand and gravel deposits, then the results on predictive error versus data collection obtained here have significant practical consequences for site characterization. Heavily sampled sand and gravel aquifers, such as Cape Cod and Borden, may have large amounts of redundant data, while in more common real world settings, our results suggest that denser data collection will likely improve understanding of permeability structure.

Introduction

A short supply of hydrologic data prevents detailed description of most groundwater systems. For a typical groundwater modeling effort, hydrologic parameters of a large aquifer volume must be assigned based on just a few point measurements (a volume of 10 km^3 described by 40 measurements, for example). Even the most heavily sampled aquifers, such as the shallow aquifer at the Macrodispersion Experiment (MADE) site in Mississippi with over 2200 hydraulic conductivity measurements, have data describing no more than 1% of the total volume. The lack of detailed subsurface data adds significant uncertainty to groundwater simulation results. For example, *Rehfeldt et al.* [1992] estimated that hydraulic conductivity (K) measurements for 400,000 nodes would be needed to make an accurate deterministic model of groundwater flow and contaminant transport at the MADE site. The uncertainty of groundwater simulation results is often largely attributable to spatial variability in hydraulic conductivity. Hydraulic conductivity controls both advective transport and dispersive transport [*Neuman*, 1990] yet can vary over 6 orders of magnitude at a single site.

Efforts to overcome sparse hydrologic data often rely on geostatistical methods such as kriging. Kriging is designed to make estimates at unsampled locations and is now commonly used as a tool for expanding sparse spatial data. In addition, a variety of other geostatistical methods are currently employed in groundwater modeling efforts to assign hydraulic conductivity values. Methods such as ordinary kriging that produce just a single field of values are known as "estimation" methods, while methods that produce many alternate fields of values are commonly called "stochastic simulation" methods.

What is the relative value of using different geostatistical methods to predict unsampled hydrologic parameter values? This question has only recently begun to receive attention. *Ritzi et al.* [1994] evaluated the ability of three indicator-based geostatistical methods to predict the occurrence of high-hydraulic-conductivity facies in a glacially deposited aquifer. *Boman et al.* [1995] evaluated the ability of four geostatistical methods to produce hydraulic conductivity fields that, when used as input to a transport model, yielded breakthrough curves in agreement with tracer tests in a layered coastal plain aquifer. Both of these studies found that in comparison to estimation methods, stochastic simulation methods produced

Paper number 96WR00272.
0043-1397/96/96WR-00272$05.00

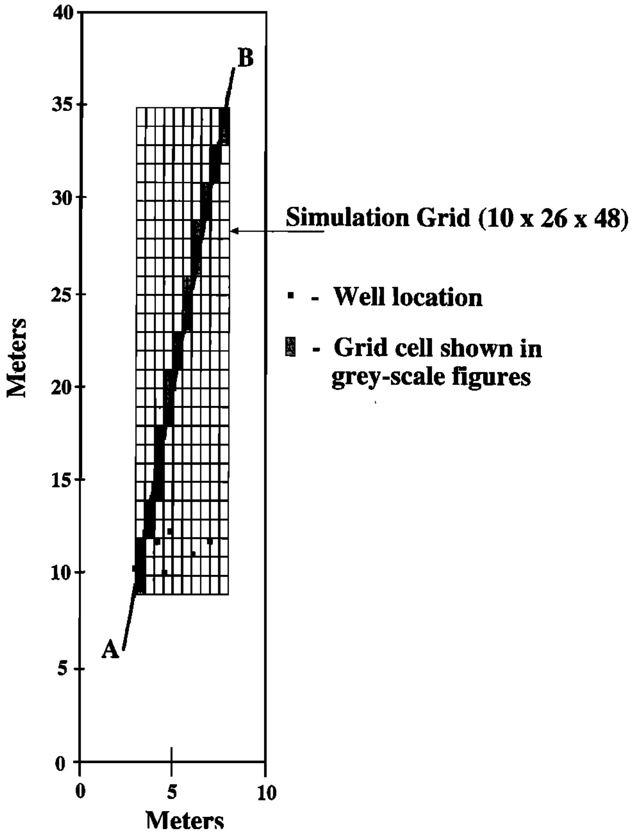
40
35
30
25
20
15
10
5
0
Meters
0
5
10
Meters
A
B
Simulation Grid (10 x 26 x 48)
▪ - Well location
▮ - Grid cell shown in grey-scale figures

Figure 1. Cape Cod site and simulation grid. Squares mark locations of long-screened wells used for flowmeter conductivity measurements. Position of simulation grid is shown for reference. Vertical section AB is used in later grey scale figures.

more realistic continuity structure and gave the additional advantage of allowing flow simulation to be stochastic. The sensitivity to conditioning data has also received attention. *Smith and Schwartz* [1981] examined how increasing amounts of hydraulic conductivity data affect the variability of contaminant arrival times in a synthetically generated two-dimensional aquifer. *Clifton and Neuman* [1982] examined how the prediction error in hydraulic head is affected by conditioning to spatial correlation of hydraulic conductivity measurements, measured flow rates, and measured heads. However, the influence of the amount of conditioning data on real-world hydraulic conductivity estimation has not been examined.

In this study we focus on a glacially deposited sand and gravel aquifer and examine the ability of some common geostatistical methods to identify known hydraulic conductivity structures. Our study is unique in that we directly test the accuracy of these methods to infer hydraulic conductivity in a real-world aquifer. We use detailed hydraulic conductivity measurements from the Cape Cod aquifer test site in Massachusetts [*Hess et al.*, 1992] as the ground truth data for our analyses.

The Cape Cod data set is used to address two principal questions: (1) to what degree can some common geostatistical methods based on the variogram accurately predict hydraulic conductivity, and (2) how much does predictive capability improve with increasing amounts of data? Our results indicate that for relatively homogeneous sand and gravel aquifers such as the Cape Cod aquifer, neither estimation nor stochastic simulation methods give highly accurate point predictions of hydraulic conductivity, although stochastic simulation methods do realistically simulate observed hydraulic conductivity structure. The stochastic simulation methods produced higher estimation error than the estimation methods by approximately 30%. Higher point errors for the stochastic methods were offset by better reproduction of the measured distribution of the natural logarithm of hydraulic conductivity, ln (K), and better reproduction of local contrasts in ln (K). Our results also indicate that estimation error is relatively insensitive to the number of conditioning data as long as more than a threshold number of hydraulic conductivity measurements are used to condition the realizations. This threshold occurs for the Cape Cod site when there are approximately three measurements per integral volume. The point ln (K) measurements, rather than model variogram parameters, appear to have primary control over estimation error.

The Cape Cod Site

Because high-hydraulic-conductivity sediment structures such as sand and gravel layers have been found to often control mass transport in sedimentary aquifers [*Anderson*, 1990; *Fogg*, 1986], we felt it was important to use data that contain natural conductivity patterns. The hydraulic conductivity measurements we used were taken in the upper 7 m of the saturated zone at the Cape Cod aquifer. The Cape Cod test site was developed beginning in the early 1980s for detailed analysis of groundwater movement and contaminant transport. The surficial aquifer is the principal aquifer at the site and is composed of permeable unconsolidated sediments about 100 m thick. The upper 30 m of the aquifer contains stratified sand and gravel outwash deposited during the last glacial retreat [*LeBlanc et al.*, 1991; *Hess et al.*, 1992].

The hydraulic conductivity measurements from Cape Cod were taken with a flowmeter in 16 long-screened wells adjacent to the main tracer test area. None of the wells are separated by more than 25 m, as is shown in Figure 1. The complete flowmeter data set consists of 668 hydraulic conductivity measurements. The downhole flowmeter is currently the best available technology for taking in situ hydraulic conductivity measurements in the saturated zone and has been found to yield reproducible hydraulic conductivity measurements [*Rehfeldt et al.*, 1989]. At the Cape Cod site, flowmeter measurements have a mean hydraulic conductivity of 0.11 cm/s. Measurements of hydraulic conductivity taken with a permeameter using core samples yielded a much lower mean hydraulic conductivity of 0.035 cm/s, possibly because core samples containing large gravels were excluded from testing [*Hess et al.*, 1992]. We decided not to use the permeameter measurements in this study because the higher hydraulic conductivity samples were excluded.

Measurements taken with a flowmeter contain significant error. This error originates from random errors in the flowmeter device as well as from incorrectly estimated aquifer and well loss parameters used in the calculations. *Rehfeldt et al.* [1989] calculated error from replicate flowmeter tests at the MADE site and found that while trends in the ln (K) profiles were well reproduced, random errors having a standard deviation of 0.21 occurred for measurements in the range of $-4 <$ ln $(K) < 0$. This is the same range as that found at Cape Cod.

Although the flowmeter can potentially yield hydraulic conductivity measurements at a fine scale, the aquifer volume described by each measurement at Cape Cod is relatively large. When calculating hydraulic conductivity from the flowmeter measurements, flow to the well is assumed to be horizontal, and the aquifer is divided into a series of horizontal layers. At Cape Cod the layers were assumed to be 15 cm in height. The pumping radius of influence, which is the horizontal scale of support for the flowmeter measurement, can be estimated with the following equation [*Rehfeldt et al.*, 1989]:

$$r = 1.5\sqrt{\frac{K_pHt}{S}} \qquad (1)$$

where

r radius at which drawdown is ~10% of drawdown at the well face;

K_p depth-averaged hydraulic conductivity for the whole aquifer;

H thickness of the aquifer;

S storage coefficient, or specific yield;

t time to steady state water levels.

The radius of influence is calculated to be 7.2 m when using a depth-averaged hydraulic conductivity of 0.11 cm/s, a thickness of 7 m, a specific yield of 0.1, and a time of 5 min, a typical time for water levels to reach quasi-steady state during pumping (K. M. Hess, personal communication, 1995). Because many of the wells at Cape Cod are separated by only 1–2 m, the flowmeter measurements have much horizontal overlap. Despite this overlap, which should cause neighboring values to be similar, the variance of the flowmeter measurements is higher than the variance of the laboratory permeameter measurements. The variance of the ln (K) measurements taken with the flowmeter is 0.24. In comparison, the variance for the permeameter measurements is 0.14, using samples 10 cm high and 5 cm in diameter.

The error associated with the flowmeter measurements and the relatively large aquifer volume described by each measurement probably mean that the measured data capture large-scale trends present at the site but do not accurately reflect smaller-scale heterogeneities. Vertical heterogeneity will be more accurately reflected by the flowmeter measurements than horizontal heterogeneity because aquifer flow during the tests is largely horizontal.

As an alternative to the flowmeter measurements, artificially generated hydraulic conductivity fields could be used as the ground-truth data upon which the hydraulic conductivity simulations are based. Other authors have followed the practice of using synthetic hydraulic conductivity data [e.g., *Scheibe and Freyberg*, 1990]. The main advantage of using synthetic data is that the known reference field can be closely controlled. The main disadvantage is that the relationship between synthetic data and a natural aquifer is ambiguous. Extrapolation of results to real-world settings is therefore difficult or impossible. A comparison of simulation methods can be made whether one is using measured or synthetic data. We chose to use the flowmeter measurements because their reproducibility and consistency of scale do allow some extrapolation of results to real-world settings. The extrapolation is limited by errors associated with the measurements and the relatively large aquifer volume described by the measurements. However, even with these limits, the measured data provide a more valuable test than do synthetic data.

Methods

The first step in a geostatistical analysis is to determine descriptive statistics and spatial correlation. A geostatistical analysis treats the variable under consideration, Z as a spatially continuous function with a continuous correlation structure. Once the functions describing the mean and spatial correlation have been determined, then the prediction of Z at unsampled locations can proceed. A field of estimated values generated by a stochastic method such as simulated annealing or sequential Gaussian simulation is known as a "realization" because the estimated field of values is just one of many possible guesses at the true field of values. The measured data used to constrain a realization are "conditioning data," and if the measured values are assigned at their measurement locations, then the realization is "conditional."

Partitioning of Data and Discretization of Site

We allocated 396 of the 668 conductivity measurements, about 60%, for conditioning the simulations. The other 40%, or 272, of the conductivity measurements were then available for checking accuracy of the ln (K) realizations. To gauge the worth of having more conditioning data, we used various fractions of the 60% to condition the realizations. The fractions of conditioning data that we used correspond to 5%, 10%, 20%, 30%, 40%, 50%, and 60% of the complete (668) conductivity measurements.

The 60% subset was chosen by randomly selecting from the complete data. Ten different 60% subsets were randomly generated, and the one subset with sample statistics closest to the complete (100%) data was retained for use. The 60% subset has the same declustered ln (K) variance as the complete data and a mean ln (K) that is 1.1% lower than for the 100% data. Since the statistics of the conditioning data are close to the statistics of the complete data set, the magnitude of estimation error that is directly the result of the geostatistical method is high. Subsets containing 5%, 10%, 20%, 30%, 40%, and 50% of the complete data were then randomly selected from the 60% subset to ensure that estimation error would be calculated at the same 272 points. In generating the 5%, 10%, 20%, 30%, 40%, and 50% subsets, no consideration was given to whether or not the sample statistics of the subset match those of the complete data set. The 60% subset was already chosen to reflect the complete data, and we did not want to further mask the statistical instability that comes with smaller sample sizes.

A fine three-dimensional grid was used to discretize the Cape Cod site, as is shown in Figure 1. The grid has 12,480 nodes (10 × 26 × 48) with x, y, and z coordinate spacing of 0.5 m, 1.0 m, and 0.15 m. The spacing is just fine enough so that only one measurement location falls into a cell. Each measurement was relocated to the center of its cell for the geostatistical simulations.

Spatial Correlation Determination and Variogram Model Fitting

Although there is a great deal of random variation, measurements of similar ln (K) tend to be located near one another. As is standard for geostatistical practice [*Deutsch and Journel*, 1992], we used the variogram to express spatial correlation. We calculated the experimental variogram using the traditional estimator [*de Marsily*, 1986].

$$\gamma(h) = \frac{1}{2n(h)}\sum_{i=1}^{n(h)}[z(x_i + h) - z(x_i)]^2 \qquad (2a)$$

where $\gamma(h)$ is variogram, h is the lag, $n(h)$ is the number of sample pairs separated by h, and $z(x_i)$ and $z(x_i + h)$ are two measured ln (K) values measured at x_i and $x_i + h$. Throughout this study we assumed second-order stationarity, meaning that the variogram is a function of h and direction only.

We followed the common practice in stochastic hydrology of using a negative exponential function to model the ln (K) experimental variograms [*Woodbury and Sudicky*, 1991; *Rehfeldt et al.*, 1992].

$$\gamma(h) = \gamma_0+(\sigma_Z^2-\gamma_0)\left(1-\exp\left[-\left(\frac{h_1^2}{\lambda_1^2}+\frac{h_2^2}{\lambda_2^2}+\frac{h_3^2}{\lambda_3^2}\right)^{1/2}\right]\right) \quad (2b)$$

where γ_0 is the nugget, σ_Z^2 is the sill, h_i is the lag or separation vector, and λ_i is the correlation length (with $i = 1, 2, 3$). It should be noted that λ is only one third of the practical range. At a separation of λ, $\gamma(h)$ is only 63% of the sill, while at a separation distance of 3 λ, the practical range, $\gamma(h)$ is 95% of the sill. It is at distances of more than 3 λ that one expects very little spatial correlation in Z.

To determine the variogram parameter values we followed the practice of *Rehfeldt et al.* [1992] and held the sill equal to the variance of the conditioning data while γ_0 and λ were varied to find the least squares difference between the experimental and model variograms. This simplified the fitting procedure and allowed us to use only values in the rising limb of the variogram when fitting the model nugget and correlation length. The sill is a random variable, and some authors have included the sill as a parameter in the least squares fit [*Hoeksema and Kitanidis*, 1985; *Woodbury and Sudicky*, 1991]. We determined the sill by finding the variance of the declustered data (which is equal to an average of the experimental variogram values for large lags). Model variograms were constructed for the complete (100%) data as well as for all subsets of the conditioning data. The nugget, sill, and correlation length values changed as different sets of conditioning data were used. Details on the fitted models are given in the Results section.

Fitting a variogram model is unavoidably a subjective exercise [*Woodbury and Sudicky*, 1991]. The processes of selecting a model, choosing lag increments, and deciding on variations of the least squares fit all depend on visual interpretation and intuition.

Geostatistical Simulation Methods

We compare two methods for estimating mean values of a random variable with two methods for generating stochastic realizations of a random variable. The first of the estimation methods was the simpler. The mean of the conditioning data was used as the ln (K) estimate for every cell that did not have a measured value. The other estimation method, ordinary kriging, and the two stochastic methods, sequential Gaussian simulation and simulated annealing, are all based on an assumption of second-order stationarity and rely on the variogram for inference of spatial correlation. Ordinary kriging is designed to produce estimates with low estimation errors, whereas the simulation methods, sequential Gaussian simulation and simulated annealing, intentionally introduce greater estimation error to allow multiple realizations. Alternate hydraulic conductivity realizations created by stochastic simulation are generally input to multiple groundwater simulations [e.g., *Poeter and Townsend*, 1994] to gauge uncertainty in contaminant transport. An extensive comparison of estimation and stochastic simulation methods is given by *Deutsch and Journel* [1992].

In selecting ordinary kriging, simulated annealing, and sequential Gaussian simulation for analysis we chose methods that can produce three-dimensional anisotropic fields, honor conditioning data, use the variogram as a basis for controlling spatial correlation, and have been successfully demonstrated and made available by previous investigations. There are many other methods that we could have analyzed and that deserve more attention than can be given here. These other methods include indicator kriging [*Johnson and Dreiss*, 1989; *Suro-Perez and Journel*, 1991] and Bayesian updating [*Woodbury*, 1989] as well as whole classes of methods that recreate fractal or multiscale structures [*Brannan and Haselow*, 1993; *Molz and Boman*, 1993], arrange sediment bodies of a prescribed shape [*Haldorsen and Damsleth*, 1990], or simulate sediment deposition [*Anderson*, 1989; *Webb*, 1994].

The conditional mean as a global estimate. Using the mean of the conditioning data as an estimate at all grid cells is a common practice in groundwater modeling when few measurements are available or when a full geostatistical analysis is not practicable. We refer to use of the conditional mean for a global estimate as the CM method.

Ordinary kriging. Ordinary kriging is the most widely used geostatistical estimation method and is often used to create hydraulic conductivity fields for input to groundwater flow and contaminant transport models. The kriging equations give not just an estimate of Z, they also calculate the model estimation error variance for each location, a theoretical uncertainty that does not necessarily reflect true estimation error. Ordinary kriging estimates are unbiased for the model random variable $(\hat{Z})$ and minimize the model estimation error variance $(\hat{\sigma}_R^2)$. However, because the model random variable is not the same as the true variable, the true mean (μ_Z) is not necessarily reproduced and the true estimation error (σ_R^2) is not necessarily minimized. A thorough discussion of ordinary kriging with detailed examples is given by *Isaaks and Srivastava* [1989].

Sequential Gaussian simulation. Sequential Gaussian simulation creates realizations of normal random variables. Assuming that Z is Gaussian, the kriging mean and variance can fully describe the distribution of Z at each point in space. The sequential Gaussian simulation method introduces estimation error to kriging, so that each realization created is different. Rather than using the kriged value at each point, a value is drawn from the normal distribution defined by the kriging mean and variance. In addition, previously simulated points are allowed to condition each new point that is simulated. The order in which nodes are considered is varied to further randomize the results. Before applying sequential Gaussian simulation, we normalized each conditioning data subset for ln (K) simulation using a Geostatistical Software Library (GSLIB) routine [*Deutsch and Journel*, 1992]. Variograms for the normalized ln (K) subsets were also constructed and fit with exponential models following the procedures described earlier. To regain ln (K) estimates, the normal simulated values were back-transformed as each realization was completed.

Simulated annealing. Simulated annealing is an optimization method first developed by *Metropolis et al.* [1953] and subsequently used in the field of image analysis [*Geman and Geman*, 1984]. As applied to random spatial variables [*Deutsch and Journel*, 1992], it is used as a stochastic simulation method for generating two- or three-dimensional fields of correlated values. An initial field is generated by randomly drawing values from a specified distribution; we used the nonparameterized ln (K) distribution defined by the conditioning data and added upper and lower tails to limits of −5.0 and 0.0. One randomly drawn value is assigned to each simulation node, and measured values are assigned at measurement locations. Pairs of nonconditional values are then switched according to an optimization function that describes the least squares error between the model variogram and the calculated variogram of the simulated field. A temperature function controls how fast the optimization function is reduced by allowing some switches that increase the optimization function. See *Deutsch and Journel* [1992] for further detail.

Assessment of the Conductivity Realizations

We gauged the accuracy of each realization by calculating ln (K) estimation error at the 272 measurement locations not used for conditioning data

$$r_i = \ln(\hat{K}_i) - \ln(K_i) \tag{3}$$

where r_i is estimation error at measurement location i, $\hat{K}_i$ is the estimated hydraulic conductivity, and K_i is measured hydraulic conductivity. Mean estimation error, mean absolute estimation error, and estimation error variance were calculated as follows:

$$\mu_r = \frac{1}{nN} \sum_{j=1}^{n} \sum_{i=1}^{N} r_{ij} \tag{4}$$

$$\mu_{|r|} = \frac{1}{nN} \sum_{j=1}^{n} \sum_{i=1}^{N} |r_{ij}| \tag{5}$$

$$\sigma_r^2 = \frac{1}{nN} \sum_{j=1}^{n} \sum_{i=1}^{N} (r_{ij} - \mu_r)^2 \tag{6}$$

where μ_r is mean estimation error, $\mu_{|r|}$ is mean absolute estimation error, σ_r^2 is estimation error variance, n is number of realizations, and N is number of points at which statistics are calculated (= 272).

We used mean absolute estimation error rather than estimation error variance to express the magnitude of error because mean error was generally not equal to zero. For the idealized case of a normally distributed random variable and mean error of 0 the following relation holds:

$$\mu_{|r|} = 0.675\sqrt{\sigma_r^2} \tag{7}$$

If $\mu_r = 0$, then

$$\sigma_r^2 = \sigma_{\ln(K)}^2 \tag{8}$$

Using the variance of the Cape Cod declustered ln (K) data (= 0.28) with equations (7) and (8), one arrives at a value of 0.35 for mean absolute error. The 0.35 value provided a reference against which we compared $\mu_{|r|}$ values from the simulation results. If ln (K) were normal and the mean error were zero, then the mean absolute estimation error with the CM method would be 0.35.

To check that mean absolute ln (K) error is an accurate and unbiased measure of estimation error, we also calculated median absolute ln (K) error and mean absolute hydraulic conductivity error for each simulation method. Because the results for median absolute ln (K) error and mean absolute hydraulic conductivity error were similar to results for mean absolute ln (K) error, we do not devote much attention to these alternate measures. A brief discussion of results for the different measures is given in the Results section.

The error statistics were recalculated as each new stochastic realization was produced. When additional realizations did not change any of the statistics by more than 0.1%, the simulations were stopped. It typically required 50–400 sequential Gaussian simulation realizations and 25–100 simulated annealing realizations to meet this convergence criterion.

Transport modeling is strongly influenced by the continuity of large-scale hydraulic conductivity structures because continuous high-hydraulic-conductivity sediments provide preferential flow paths that can dominate patterns of flow. We used visual comparison of ln (K) images to gauge the reproduction of continuity patterns. Other investigators [*Smith and Schwartz*, 1981; *Boman et al.*, 1995] have used the results of transport simulations or have devised measures of spatial continuity [*Fogg*, 1986] to gauge reproduction of ln (K) spatial continuity. However, at the Cape Cod site, tracer tests were not performed in the same area as flowmeter testing, so that there are no ground truth transport data that can be used in conjunction with the conductivity data. Also, the flowmeter data are too widely spaced to apply continuity measures. In addition, there is some question about using flow simulations to gauge accuracy of geostatistical simulations because dissimilar large-scale hydraulic conductivity patterns can produce similar flow characteristics. For example, contaminant travel times can be equal for a homogeneous, high-mean-hydraulic-conductivity, aquifer and for a low-mean-hydraulic-conductivity aquifer with a single high hydraulic conductivity inclusion, even though large-scale hydraulic conductivity patterns in the two aquifers are quite different. The ideal situation for evaluating the ability of geostatistical methods to simulated hydraulic conductivity would be to have both tracer test results and extensive hydraulic conductivity measurements from the same test site.

Ordinary kriging, sequential Gaussian simulation, and simulated annealing are all sensitive to the many parameters controlling their simulations. For instance, when the ln (K) distribution from which initial simulated annealing values were drawn was given an upper tail extending to ln $(K) = 5.0$ rather than to 0.0, the error variance increased by 30%. We did not tailor the geostatistical methods toward reproducing any particular aspect of the ln (K) measurements. The simulations were also not repeated numerous times for any one method with different parameter specifications (other than the ones discussed) because the goal of this study was to compare the methods under similar average conditions rather than to achieve lowest possible estimation errors.

Results

Normality and Spatial Trends in Hydraulic Conductivity

The natural log of the hydraulic conductivity measurements was taken to obtain a less skewed distribution (Figure 2). Mean ln (K) for the complete data set is relatively high at −2.18, and the variance of ln (K) is low at 0.24. As a comparison, the ln (K) measurements from the heterogeneous alluvial aquifer at the MADE site showed a mean of −5.2 and variance of 4.5 [*Rehfeldt et al.*, 1991]. Hydraulic conductivity statistics from other field sites are summarized by *Gelhar* [1993, p. 292]. All geostatistical simulations in this study were performed using ln (K) rather than K values. Although the ln (K) distribution is less skewed than the hydraulic conductivity distribution, it still has a negative skewness of −0.62. The ln (K) distribution for the complete set is nonnormal, failing a χ^2 test for normality at the 99% level. *Hess et al.* [1992] chose subsets of ln (K) data by discarding closely spaced data and reported that they passed a

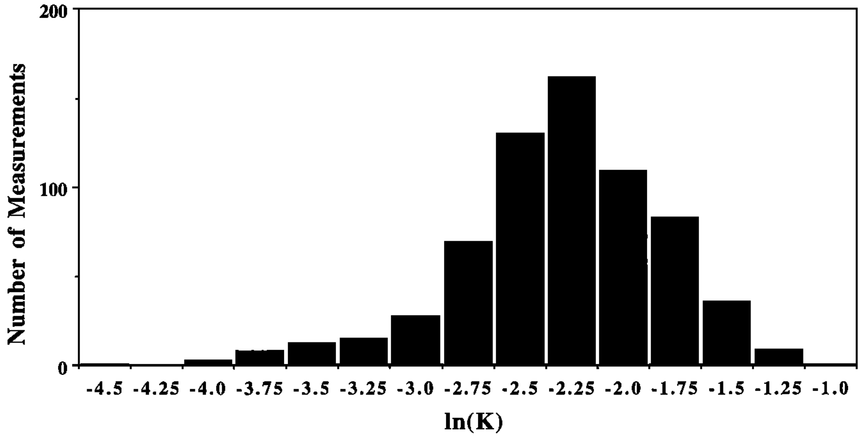

Figure 2. Distribution of Cape Cod measured ln (K): binned frequency distribution of ln (K) using all 668 flowmeter measurements.

χ^2 test for normality at the 95% level. However, given the close spacing of the wells, it is difficult to create uncorrelated subsets that are statistically significant representations of the site. This point is discussed further below.

Before applying geostatistical methods it is important to determine and remove any large-scale trends in the variable to be estimated (Z). Large-scale trends in either the mean or covariance of Z can cause errors in the estimations. For example, the presence of large-scale trends usually causes variance of the original data to be higher than variance of detrended data.

The Cape Cod ln (K) data show no trends in the horizontal direction that are consistent at all depths. There are mild trends in the vertical direction as seen in Figure 3, which shows mean ln (K) averaged over all 16 wells. Both the complete data and the reduced data show regions of higher conductivity near elevations of 7.5 and 12.5 m. The vertical trends are mild, having approximately the same magnitude as the random fluctuations in ln (K) seen at all depths.

Eleven of the sixteen wells fall nearly on line AB of Figure 1. A vertical section of the measured ln (K) from these wells is shown in Figure 4. The most obvious trends in Figure 4 are the somewhat continuous horizontal bands of higher hydraulic conductivity at elevations of 7.5 and 12.5 m.

Characteristics of the ln *(K)* Subsets

The vertical trends in ln (K) seen in the complete data are also seen in the subsets of the complete data (see Figure 3). Subsets having more data generally do a better job of reproducing the trends. The reproduction of the high ln (K) vertical trends at elevations of 7.5 and 12.5 m can be seen in Figure 3 for the 20% subset.

The mean and variance of the subsets as well as the best fit variogram parameters are given in Table 1. The subsets have ln (K) variance as much as 11% lower and 25% higher than ln (K) variance of the complete data. Subsets with less data generally exhibit greater differences in ln (K) variance as would be expected. Best fit variogram parameters for the data subsets deviate from best fit variogram parameters for the complete data. The best fit nugget, vertical correlation length,

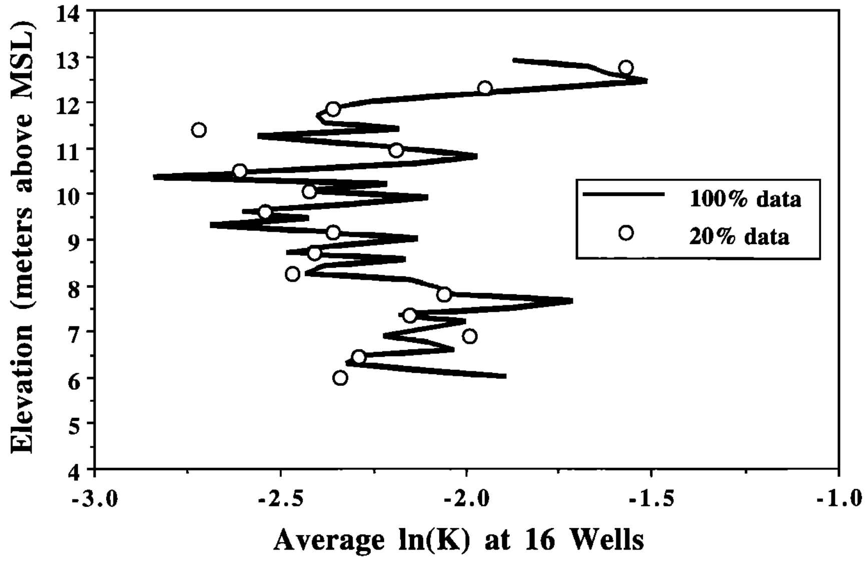

Figure 3. Average ln (K) with depth. The solid line is average ln (K) for all flowmeter data (K is in centimeters per second). Dots indicate average ln (K) for a reduced data set constructed by randomly selecting 20% of the total data.

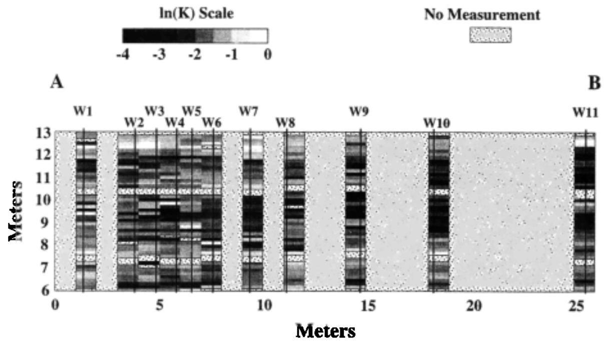

Figure 4. Vertical section of measured ln (K) (K is in centimeters per second). Each rectangular block represents one simulation grid cell along section AB. The cells are 0.15 m high and 1.0 m long. The vertical black lines are reflections of wells on AB.

and horizontal correlation length all tend to be smaller for the subsets than for the complete data.

While large-scale spatial trends can prevent effective use of geostatistical methods, the trends present at Cape Cod are of too small a spatial scale and too small a magnitude to have more than a minor influence on the ln (K) simulations. For example, we performed a test case comparing two simulations to see if the observed ln (K) trends significantly affected ln (K) estimation. One simulation removed trends in measured data; the other did not. For the detrended simulation a third-order least squares regression on ln (K) was performed with x, y, and z coordinates as the independent variables. The detrending removed all visible trends in ln (K) and reduced ln (K) variance from 0.24 to 0.19 (K in centimeters per second). Twenty percent of the ln (K) measurements were used for conditioning. It was found that detrending ln (K) did little to reduce ln (K) estimation error and in some cases actually increased it slightly. This was true for kriging as well as for sequential Gaussian simulation and simulated annealing. This test suggests that removing trends is not necessary for geostatistical simulation of the Cape Cod hydraulic conductivity field. In the subsequent work detailed below, we used ln (K) conditioning data that were not detrended.

Spatial Correlation and Variogram Models

Experimental variograms constructed in several directions for the complete ln (K) data indicate that the variogram is isotropic in the horizontal plane but that there is anisotropy between the horizontal and vertical directions. This result is in agreement with *Hess et al.* [1992]. Vertical and horizontal experimental variograms for the complete ln (K) data are shown in Figures 5a and 5b.

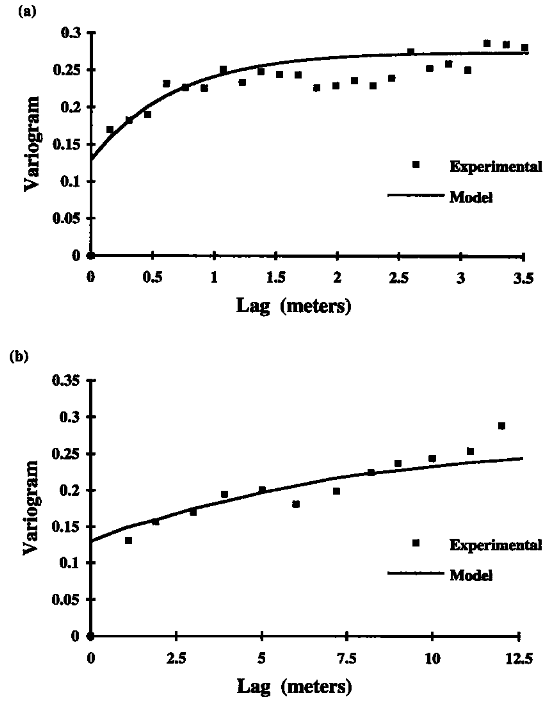

Figure 5. Variograms for the complete ln (K) data. Squares are experimental variogram points; solid lines are best fit exponential models. (a) Vertical variogram. Model parameters are sill = 0.276, nugget = 0.13, and λ = 0.69 m. (b) Horizontal variogram. Model parameters are sill = 0.276, nugget = 0.13, and λ = 8.08 m.

To construct points on the experimental variogram, discrete separation distances were chosen, and pairs separated by the chosen distances, plus or minus 50% of the lag increment, were binned together. We selected lag increments of 15 cm in the vertical direction and 1 m in the horizontal direction following *Hess et al.* [1992]. These lag increments yield relatively smooth variograms, are approximately 15% of the correlation scales, and correspond to the approximate minimum vertical and horizontal separation distance between measurements. When con-

Table 1. Statistical Description of Subsets of the Complete Data Used to Assess Estimation of ln (K) Under Limiting Amounts of Data

				Best Fit Exponential Variogram Parameters		
Subset of Complete Data, %	Number of Data	ln (K)			Correlation Length	
		Mean	Variance	Nugget	Vertical	Horizontal
100	668	−2.18	0.28	0.13	0.69	8.08
60	396	−2.21	0.28	0.05	0.51	4.80
50	335	−2.19	0.26	0.09	0.45	4.68
40	267	−2.21	0.26	0.02	0.30	2.43
30	199	−2.25	0.25	0.08	0.41	7.67
20	133	−2.27	0.35	0.00	0.53	4.43
10	66	−2.31	0.30	0.00	0.79	7.92
5	33	−2.27	0.29	0.14	0.39	3.89

Statistics are for declustered data.

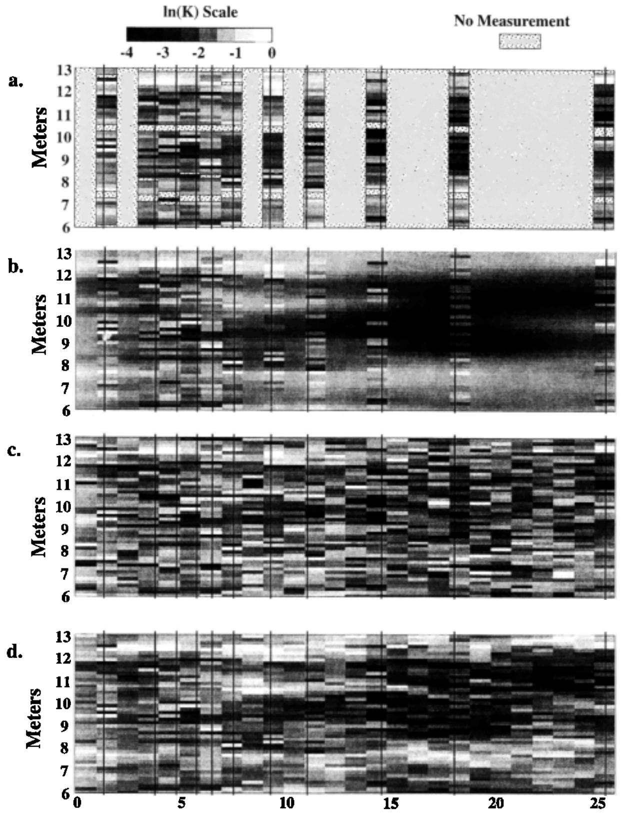

Figure 6. Examples of the generated ln (K) fields: (a) measured, (b) ordinary kriging, (c) sequential Gaussian simulation, and (d) simulated annealing. Each grey scale rectangle represents one simulation grid cell (0.15 m × 1.0 m). Well positions are indicated by vertical black lines.

structing the vertical variograms, only sample pairs from the same well were allowed. For the horizontal variograms, sample pairs were excluded by a 5° dip tolerance and a 1-m vertical deviation tolerance.

An interesting feature of the Cape Cod site is that the domain of K measurements is not much larger and perhaps smaller than the spatial correlation range of ln (K). Figures 5a and 5b show that the variogram is still increasing for the vertical direction at a lag of 3.5 m and still increasing for the horizontal direction at a lag of 12.5 m. The lags of 3.5 and 12.5 m correspond to 50% of the greatest separation distances between measurements, the usual limit to variogram reliability. This means that despite the relatively high degree of homogeneity, the Cape Cod domain is still not large enough to allow unambiguous determination of ln (K) statistics and correlation lengths.

In our analysis we had difficulty assigning a value for the variogram sill (or the variance of the conditioning data). As was noted above, with maximum well spacing of ~25 m and maximum vertical spacing of ~7 m, the magnitude of the Cape Cod domain is about the same as or perhaps smaller than the practical range of ln (K). It is therefore not possible to create a data subset containing only uncorrelated measurements that are still representative of the complete data. For example, when data subsets were selected randomly so that samples had minimum separation distances of 3 m vertically and 12 m horizontally (short estimates of the practical range), the subsets contain an average of less than 5 samples, which is certainly not representative of all 668 measurements. Our solution to this problem was to accept that the domain of the flowmeter measurements was too small for statistical homogeneity and to use ln (K) variance values from subsets that retain enough data to be representative of the complete data but that are only partially declustered. The minimum vertical and horizontal separation distances used in generating the subsets were 0.6 m and 5.0 m. Using these separation distances, 100 subsets of the complete data set were randomly generated that contained a mean of 37 data points, had an average $\mu_{\ln(K)}$ of -2.204, and an average $\sigma^2_{\ln(K)} = 0.276$. The value of 0.276 determined in this fashion is used as the sill value for the complete ln (K) variograms. The 0.276 value obtained after partial declustering is slightly larger than the 0.24 value for the complete nondeclustered data (which is the value of the variogram sill used by *Hess et al.* [1992]).

We used a nugget parameter in all variogram models because it significantly improved the least squares fit to the experimental variograms. Our nugget value of 0.13 for the complete ln (K) data is the same as that used by *Hess et al.* [1992].

The best fit vertical and horizontal correlation lengths based on the entire data set were 0.69 m and 8.08 m, giving practical ranges of 2.1 m and 24.2 m. Our correlation length values are slightly higher than the 0.38 and 8.0 m given by *Hess et al.* [1992] for the same experimental variograms. The evidence that the domain of the Cape Cod flowmeter measurements is too small to permit stationarity calls into question the estimation of macrodispersivity values for this site using the stochastic equations of *Gelhar and Axness* [1983]. We expect that correlation lengths might be even larger if more widely spaced measurements were available.

Comparison of Geostatistical Methods

Simulated fields of ln (K) were generated using conditional mean, ordinary kriging, simulated annealing, and sequential Gaussian simulation, each with seven different numbers of conditioning data (5%, 10%, 20%, 30%, 40%, 50%, and 60% of the complete data). Estimation error is generally lower for the estimation methods (CM and ordinary kriging) than for the stochastic simulation methods (sequential Gaussian simulation or simulated annealing), but the stochastic simulations methods are better at reproducing observed ln (K) structure. None of the methods produced ln (K) values that had the same distribution as the measured data: all methods failed a χ^2 test at the 99% level when simulated values were compared to measured values at the error checking locations. However, the stochastic methods did a better job of matching the right hand tail of the ln (K) distribution that contains the high conductivity values. An unexpected result was that mean absolute error was not sensitive to the number of conditioning data. Mean absolute error decreased by no more than 25% for any of the methods when the number of conditioning data was increased by a factor of 10. Point ln (K) conditioning values rather than model variogram parameters were found to have primary control over accuracy of the ln (K) estimation. Example ln (K) fields produced by ordinary kriging, sequential Gaussian simulation, and simulated annealing are shown in comparison to measured data in Figure 6. Table 2 gives a compilation of simulation results.

Conditional mean as estimate. The CM method is by far the easiest to implement because only the conditional mean must be calculated. Despite its simplicity, CM provided ln (K) fields with estimation error lower than sequential Gaussian

Table 2. Estimated ln (*K*) Statistics and Associated Error

Data Used to Condition, %	Simulated ln (*K*)		ln (*K*) Estimation Error		
	Average	Variance	Average	Average Absolute	Variance
		Conditional Mean as Estimate			
60	−2.21	0.00	0.07	0.38	0.25
50	−2.19	0.00	0.07	0.39	0.25
40	−2.21	0.00	0.08	0.39	0.25
30	−2.25	0.00	0.07	0.39	0.25
20	−2.27	0.00	0.12	0.40	0.26
10	−2.31	0.00	0.17	0.41	0.27
5	−2.27	0.00	0.16	0.41	0.27
		Kriging			
60	−2.22	0.09	0.07	0.34	0.20
50	−2.21	0.08	0.07	0.33	0.19
40	−2.23	0.10	0.08	0.34	0.21
30	−2.20	0.06	0.06	0.34	0.20
20	−2.27	0.11	0.13	0.38	0.24
10	−2.26	0.11	0.11	0.40	0.26
5	−2.32	0.05	0.17	0.41	0.24
		Sequential Gaussian Simulation			
60	−2.21	0.23	0.06	0.47	0.38
50	−2.21	0.26	0.07	0.49	0.42
40	−2.21	0.25	0.07	0.47	0.40
30	−2.20	0.30	0.06	0.51	0.48
20	−2.26	0.23	0.06	0.46	0.36
10	−2.24	0.34	0.09	0.50	0.50
5	−2.24	0.53	0.09	0.58	0.73
		Simulated Annealing			
60	−2.21	0.15	0.06	0.40	0.28
50	−2.19	0.20	0.05	0.44	0.33
40	−2.21	0.22	0.07	0.45	0.34
30	−2.17	0.18	0.03	0.43	0.31
20	−2.25	0.22	0.10	0.45	0.34
10	−2.24	0.20	0.10	0.45	0.35
5	−2.27	0.39	0.13	0.60	0.62

Geostatistical simulation results are shown for four methods using various numbers of conditioning data.

simulation and simulated annealing and only 0–16% higher than ordinary kriging. The magnitude of mean absolute error for each method is shown in Figure 7. With all seven numbers of conditioning data, mean absolute error for CM was about 0.4, slightly higher than the value of 0.35 derived for the case of normally distributed ln (*K*) and zero mean error. This deviation was expected, as was noted above, because the Cape Cod ln (*K*) data are not normally distributed and mean estimation error ranges from 0.07 to 0.17 as the number of conditioning data decreases.

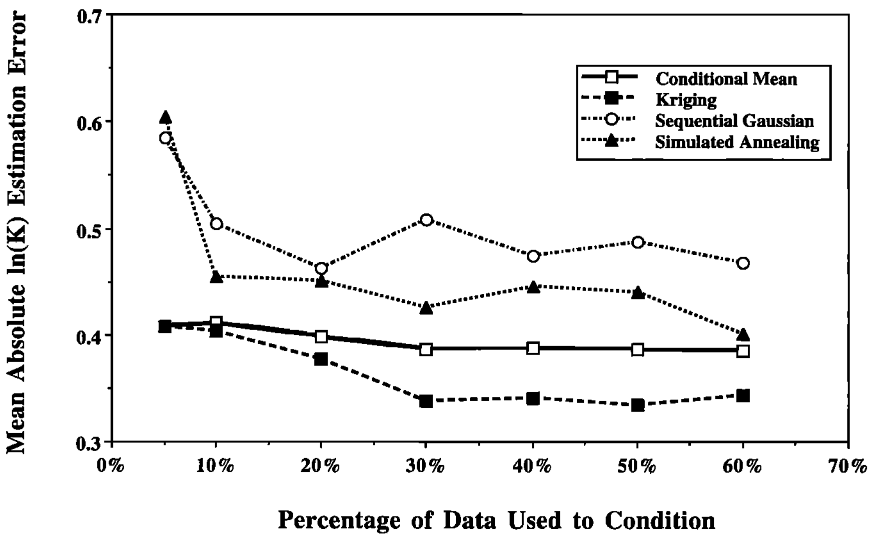

Figure 7. Mean absolute estimation error for the ln (*K*) simulations as a function of percentage of conditioning data (*K* is in centimeters per second).

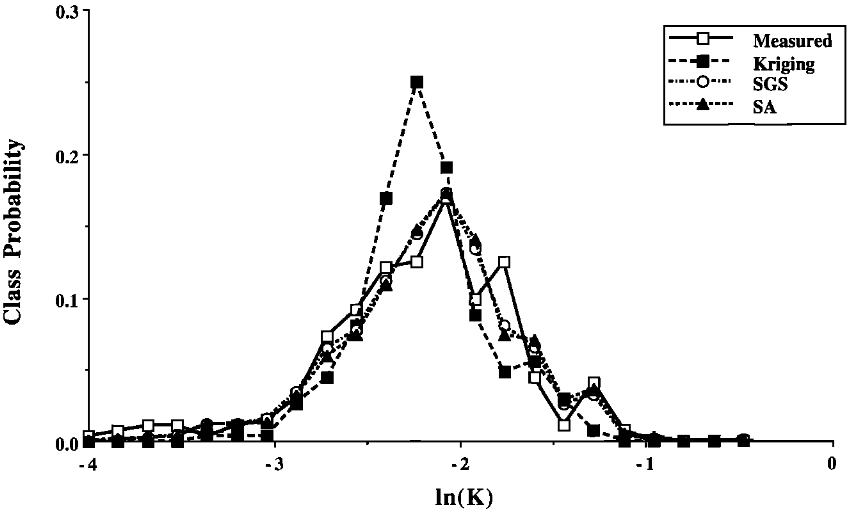

Figure 8. Distributions of measured and simulated ln (K). Data are from 272 measurement locations (K is in centimeters per second). Simulations used 20% of total data for conditioning.

The CM estimation error variance values were of approximately the same magnitude as measured ln (K) variance. The small differences are due to differences between mean ln (K) of the 100% data and mean ln (K) of the data subsets. There is no spatial variation or patterning in the CM realizations because every cell has the same estimated ln (K). The variance of the estimated ln (K) values was zero for the same reason.

Ordinary kriging. Ordinary kriging had the lowest estimation error of all methods, as one might expect from consideration of the kriging equations. However, considering the extra time and effort required for ordinary kriging as compared to CM, the difference in mean absolute error is small, only 16% at most. Because ordinary kriging take a moving average whereas CM takes a global average, the relative benefit of using ordinary kriging rather than CM would be greater if there were stronger trends in the hydraulic conductivity field.

The distributions of ln (K) values estimated by ordinary kriging for the error-checking locations differ significantly from the measured ln (K) distribution for the same locations. The ln (K) distributions produced by ordinary kriging were generally more peaked than the measured distribution (Figures 8 and 9), and variance of the estimated ln (K) values was consistently less than 40% of the measured variance. The kriged estimates generally failed to reproduce high ln (K) measurements. These high end values can be expected to significantly control groundwater movement, and their absence in the kriged estimates is likely a major shortcoming in the application of this method for transport models.

Reproduction of hydraulic conductivity structure by ordinary kriging was poor. It is a recognized problem that kriging produces fields that are smoothed and do not capture the discontinuities or sharp spatial changes of the true field [*Isaaks and Srivastava*, 1989]. This problem is especially pertinent to groundwater simulation because high and low K continuity patterns often control aquifer transport [*Fogg*, 1986]. The smoothness of the kriging fields can be seen in Figures 6 and 10. The local contrasts present in the measured ln (K) image were not reproduced by kriging. Large-scale patterns in ln (K), such as the low ln (K) structure around wells 9, 10, and 11 at elevations of 8–12 m, are visible but overly smoothed. The smoothing was more pronounced when fewer conditioning data were used.

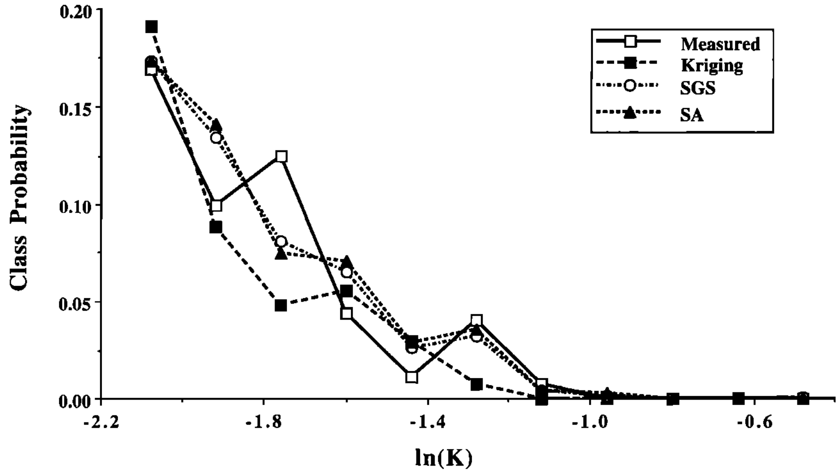

Figure 9. Detail of the high end portion of the distribution shown in Figure 8. Data are from 272 measurement locations (K is in centimeters per second). Simulations used 20% of total data for conditioning.

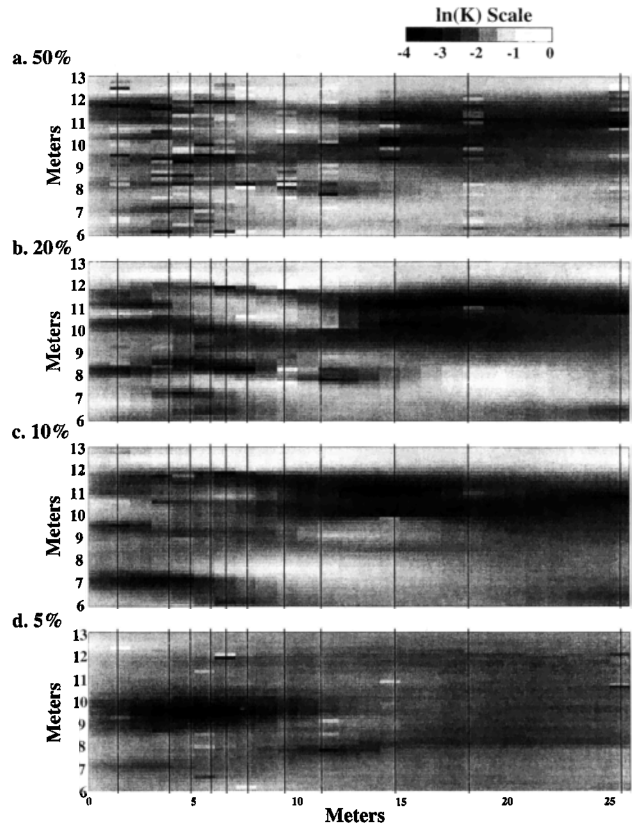

Figure 10. Examples of ln (K) fields produced by ordinary kriging using conditioning with (a) 50%, (b) 20%, (c) 10%, and (d) 5% of total data. Well positions are indicated by vertical black lines.

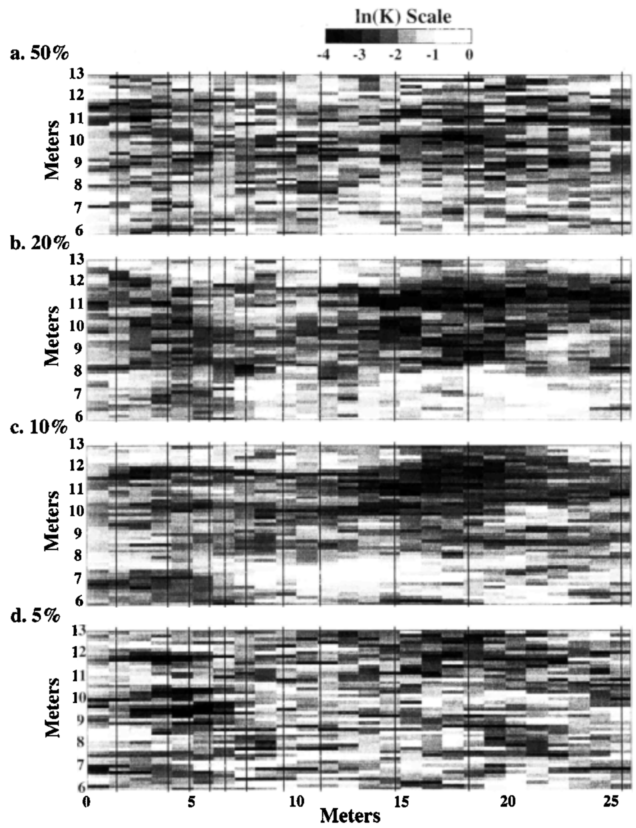

Figure 11. Examples of ln (K) fields produced by sequential Gaussian simulation using conditioning with (a) 50%, (b) 20%, (c) 10%, and (d) 5% of total data. Well positions are indicated by vertical black lines.

Sequential Gaussian simulation and simulated annealing. As might be expected, sequential Gaussian simulation and simulated annealing both had higher estimation error but better reproduction of hydraulic conductivity structure than CM or ordinary kriging. Mean absolute error was about 30% higher for sequential Gaussian simulation and simulated annealing than for CM and ordinary kriging. Sequential Gaussian simulation and simulated annealing are more sensitive to the number of conditioning data than are CM and ordinary kriging, especially when small amounts of data are used.

Sequential Gaussian simulation and simulated annealing did a good job of reproducing both the extreme measured ln (K) values and the overall ln (K) distribution. In Figures 8 and 9 it can be seen that both the upper and lower tails of the distribution were well reproduced. The reproduction of extreme ln (K) values was also reflected in the ln (K) variance of the simulated values being close to the measured ln (K) variance.

Simulated annealing had slightly lower estimation error than sequential Gaussian simulation and produced ln (K) continuity patterns with somewhat less local contrast in ln (K). The slightly smoother simulated annealing realizations are shown in comparison to sequential Gaussian simulation realizations in Figures 11 and 12. The large-scale and small-scale continuity patterns produced by Gaussian simulation and simulated annealing were visually similar to those seen in the measured data. The high ln (K) structures at elevations of 7.5 and 12.5 m are reproduced by sequential Gaussian simulation and simulated annealing except when number of the conditioning data falls below 10%. The sequential Gaussian simulation realizations were not as smooth as the simulated annealing realizations because sequential Gaussian simulation introduces error to the point value after taking account of neighboring values through the kriging process, whereas simulated annealing introduces error only by its initial random selection of values from a distribution.

Spatial patterns of error. The three variogram-based methods, ordinary kriging, sequential Gaussian simulation, and simulated annealing, had very similar spatial patterns of estimation error. Figure 13 shows mean absolute error at the 272 error-checking locations for realizations conditioned by the 20% subset. Error was generally highest in those areas having high local ln (K) contrast. For instance, Figure 13 shows that well 4 has generally low estimation error at the top and the bottom but high error at elevations of 8–12 m where high K values were immediately adjacent to low K values. The high estimation errors in areas of high local contrast were due to the assumptions of the geostatistical methods that both the random variable and the variogram are smoothly varying continuous functions.

Effects of More Conditioning Data on Estimation Error

Very little decrease in mean absolute estimation error was seen for any of the methods as the number of conditioning data was increased (Figure 7). The only substantial decrease in mean absolute estimation error was for sequential Gaussian simulation and simulated annealing as the amount of condi-

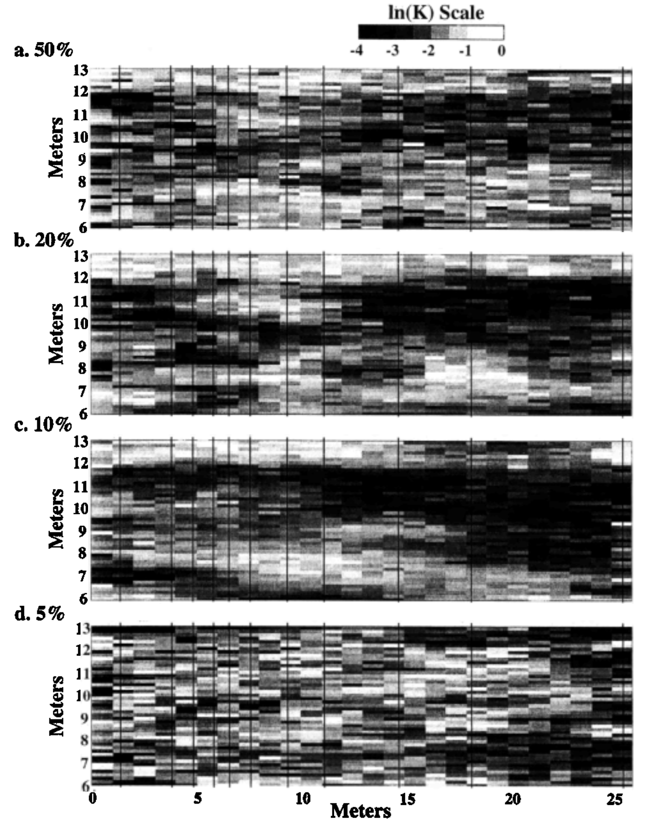

Figure 12. Examples of ln (*K*) fields produced by simulated annealing using conditioning with (a) 50%, (b) 20%, (c) 10%, and (d) 5% of total data. Well positions are indicated by vertical black lines.

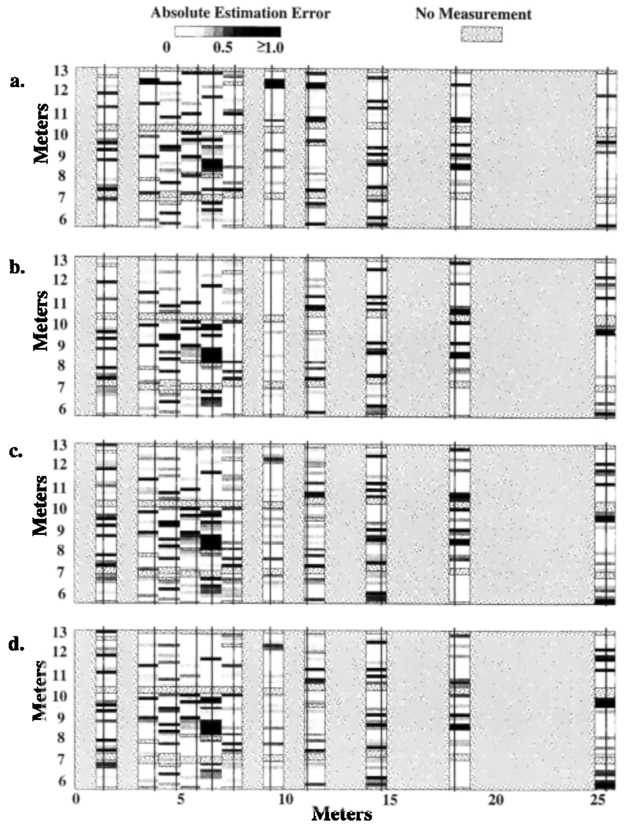

Figure 13. Spatial distribution of absolute estimation error with 20% of the total data used for conditioning: (a) conditional mean, (b) ordinary kriging, (c) sequential Gaussian simulation, and (d) simulated annealing. For ordinary kriging and CM, absolute error at each location is from a single ln (*K*) field. For sequential Gaussian simulation, mean absolute error is from 144 realizations, and for simulated annealing, mean from 39 realizations. Well positions are indicated by vertical black lines.

tioning data was increased from 5% to 10%, where mean absolute estimation error dropped by 14% for sequential Gaussian simulation and by 25% for simulated annealing. When the number of conditioning data was further increased to 60%, the additional decrease in mean absolute estimation error was only 7% for sequential Gaussian simulation and only 12% for CM. A threshold controlling the estimation accuracy of sequential Gaussian simulation and simulated annealing apparently exists between 5% and 10% of conditioning data. Below this threshold the realizations produced by sequential Gaussian simulation and simulated annealing can be improved by increasing the number of conditioning data. Above the threshold it takes a large number of additional conditioning data to reduce mean absolute estimation error. A similar threshold may exist for ordinary kriging, but it apparently is below 5% of the conditioning data.

When median absolute estimation error rather than mean absolute estimation error was used as the basis for comparing methods, the same threshold occurred. The only change to Figure 7 caused by substituting median absolute estimation error was a reduction of all values by about 25%. The relative position of the four curves and the relative changes with increasing amounts of conditioning data remained the same.

Because simulated ln (*K*) values would be transformed to *K* values before being used as input to a groundwater flow model, it is of interest to see how *K* estimation error behaves. To investigate this, we ran simulations with the same input parameters as before but transformed all ln (*K*) values to *K* values before calculating estimation error. The results showed the same error threshold between 5% and 10% of the conditioning data, and all four methods kept the same relative positions as in Figure 7. The ln (*K*) to *K* transformation expands the right-hand tail of the distribution, so errors associated with large *K* values were much larger than errors associated with smaller *K* values. Because the stochastic methods are designed to occasionally select an abnormally high value, the mean absolute *K* error for the stochastic methods was larger and more variable relative to the estimation methods than it was with mean absolute ln (*K*) error. This increase in error was particularly large for sequential Gaussian simulation, which tends to occasionally produce extremely high values due to the large error it introduces.

The threshold number of conditioning data most likely depends on a number of factors including spatial correlation, large-scale spatial trends, the magnitude of ln (*K*) variation, and the spatial scale of the hydraulic conductivity measurements. To normalize for spatial relation in ln (*K*), the number of ln (*K*) measurements per integral volume can be used. Taking ln (*K*) correlation lengths in x, y, z of 8.08 m, 8.08 m, and 0.69 m, and taking the domain of the Cape Cod *K* measurements to be (4.5 m × 25 m × 7.05 m), there are approximately 18 integral volumes in the domain. The threshold number of conditioning data thus corresponds to about three data per integral volume.

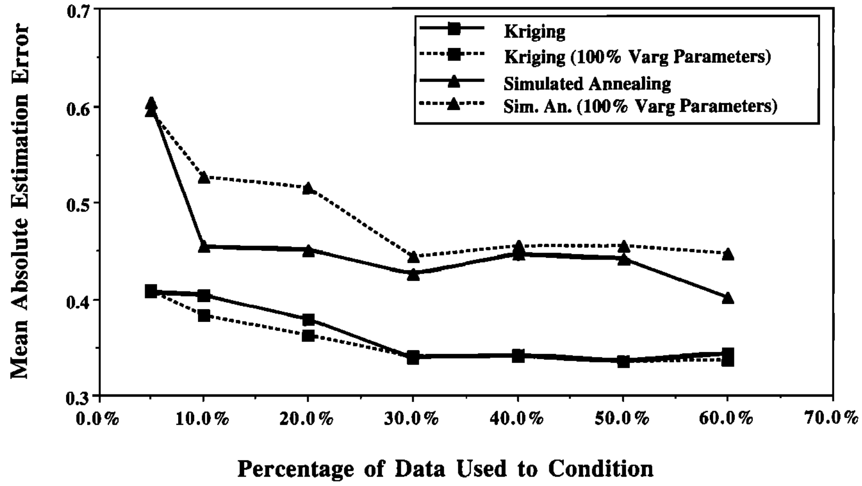

Figure 14. Effects of model variogram on mean absolute error. Solid lines represent geostatistical simulations using variograms constructed for data subsets. Dotted lines represent geostatistical simulations using variograms for complete data.

Effects of Variogram Model on Estimation Error

The ln (K) fields produced by ordinary kriging, sequential Gaussian simulation, and simulated annealing are conditioned in two ways: by the conditioning point ln (K) values and by the model variogram. We ran a simple test case to gauge the relative importance of point measurements and model variogram parameters in reducing estimation error. We retained the same conditioning point data subsets used in our previously discussed realizations but changed the model variogram parameters. In each case we assigned the best fit variogram parameters for the complete (100%) data set. The expectation was that estimation error would decrease because the 100% best fit parameters give the most complete description of ln (K) spatial correlation. We did not perform the test case for sequential Gaussian simulation because the best fit variogram parameters for the 100% data can be used only with the 100% simulations due to the normalization procedures.

Figure 14 shows results of changing the variogram parameters. Surprisingly, using variogram parameters for the 100% data actually increases estimation error for simulated annealing and makes almost no change for ordinary kriging. Even for the 20% subset, which has best fit variogram parameters the most different from the complete (100%) data, mean absolute estimation error is not decreased by using the 100% variogram parameters. This indicates that point ln (K) values, not variogram parameters, are the strongest controls on the accuracy of the realizations.

Conclusions

This study examined several commonly used geostatistical methods for accuracy of point ln (K) estimates and assessed their reproduction of hydraulic conductivity patterns. Extensive hydraulic conductivity measurements from the Cape Cod sand and gravel aquifer gave us a unique opportunity to evaluate the ability to infer hydraulic conductivity structure in the presence of limiting data.

The results demonstrate how each of the geostatistical methods reproduces naturally occurring patterns of K and how simulations are affected by the number of K measurements. None of the methods did a very good job of predicting point ln (K) values. Ordinary kriging had the lowest ln (K) estimation errors but still had ln (K) estimation error variances that were about 75% of the measured ln (K) variance. The two estimation methods, conditional mean as estimate and ordinary kriging, had slightly smaller estimation errors than the stochastic simulation methods. The stochastic simulation methods, sequential Gaussian simulation and simulated annealing, provided the following advantages: multiple realizations, simulated ln (K) distributions close to measured ln (K) distributions, and good reproduction of both local ln (K) contrast and large-scale ln (K) patterns. The inability of kriging to reproduce high ln (K) values, as reaffirmed by this study, provides a strong motivation to choose stochastic simulation methods over estimation methods for use with contaminant transport models. This is especially true when performing fine-scale deterministic simulations of contaminant transport to estimate macrodispersion [*Scheibe and Cole*, 1994; *Davis*, 1986]. Ordinary kriging produced ln (K) fields with about 75% less variation than was seen in the measured data and failed to reproduce measured high ln (K) values, which are generally understood to control aquifer transport. Sequential Gaussian simulation and simulated annealing did a much better job of reproducing the extreme ln (K) values.

It was found that all the methods were relatively insensitive to the number of conditioning data and to model variogram parameters. This indicates that the effort of obtaining high-density K measurements provides relatively little improvement in the geostatistical estimates of ln (K) at Cape Cod. Estimation error for simulated annealing and sequential Gaussian simulation appeared to pass a threshold value as the number of conditioning data was increased from 5% to 10% of the complete data. When more than 10% of the data were used to condition the simulated annealing and sequential Gaussian simulation simulations, almost no decrease was seen in estimation error. In this aquifer the 10% data level is apparently a threshold, above which prediction of measured ln (K) improves very slowly. This threshold corresponds to approximately three conductivity measurements per integral volume. A threshold was not observed for ordinary kriging, probably because it occurs at less than 1.5 measurements per integral volume, the smallest number of conditioning data considered here.

The results of this study were strongly affected by the characteristics of the Cape Cod aquifer. The absence of strong heterogeneity and spatial trends in hydraulic conductivity at Cape Cod may explain why additional conditioning data did little to improve the simulations. The measurement error associated with the hydraulic conductivity data and the relatively large aquifer volume described by the flowmeter measurements (a 7-m radius around the well) as compared to the size of the test site (5 m × 25 m) probably also contribute to the redundancy of the additional data. Apparently, additional data did little to reveal significant spatial trends in hydraulic conductivity.

Although the horizontal and vertical variograms indicate that there is some fine-scale structure in hydraulic conductivity, geostatistical methods guided by the variograms do not accurately reproduce this structure. As measured by errors in the prediction of hydraulic conductivity, the fine-scale measurements at this site only provide redundant information on the conductivity structure. If the Cape Cod aquifer spatial conductivity characteristics are indicative of other sand and gravel deposits, then the results on predictive error versus data collection obtained here have significant practical consequences for site characterization. For the Cape Cod aquifer there is little benefit to be gained from sampling more than about 50 randomly spaced conductivity measurements over a domain of 5 m × 25 m × 7 m. Such a sampling effort, however, is still greater than is typical in real-world site characterization. While the Cape Cod data set contains redundant information concerning hydraulic conductivity, real-world site characterization has a spatial resolution of hydraulic conductivity data that is typically orders of magnitude coarser. Hence in real-world settings with sand and gravel aquifers, our results suggest that more data collection than is typically performed will likely improve understanding of permeability structure.

Acknowledgments. Our thanks to Kathy Hess of the USGS for providing access to the Cape Cod data. We also thank the reviewers for their insightful and constructive suggestions. Center for Hydrologic Science, Duke University, paper 95-22.

References

Anderson, M. P., Hydrogeologic facies to delineate large-scale spatial trends in glacial and glaciofluvial sediments, *Geol. Soc. Am. Bull.*, *101*, 501–511, 1989.

Anderson, M. P., Aquifer heterogeneity—A geological perspective, in *Fifth Canadian/American Conference on Hydrogeology*, edited by S. Bachu, pp. 3–22, Natl. Well Water Assoc., Dublin, Ohio, 1990.

Boman, G. K., F. J. Molz, and O. Guven, An evaluation of interpolation methodologies for generating three-dimensional hydraulic property distributions from measured data, *Ground Water*, *33*(2), 247–258, 1995.

Brannan, J. R., and J. S. Haselow, Compound random field models of multiple scale hydraulic conductivity, *Water Resour. Res.*, *29*(2), 365–372, 1993.

Clifton, P. M., and S. P. Neuman, Effects of kriging and inverse modeling on conditional simulation of the Avra Valley Aquifer in southern Arizona, *Water Resour. Res.*, *18*(4), 1215–1234, 1982.

Davis, A. D., Deterministic modeling of dispersion in heterogeneous permeable media, *Ground Water*, *24*(5), 609–615, 1986.

de Marsily, G., *Quantitative Hydrogeology*, Academic, San Diego, Calif., 1986.

Deutsch, C. V., and A. G. Journel, *Geostatistical Software Library and User's Guide*, Oxford Univ. Press, New York, 1992.

Fogg, G. E., Groundwater flow and sand body interconnectedness in a thick multiple aquifer system, *Water Resour. Res.*, *22*(5), 679–694, 1986.

Gelhar, L. W., *Stochastic Subsurface Hydrology*, Prentice-Hall, Englewood Cliffs, N. J., 1993.

Gelhar, L. W., and C. L. Axness, Three-dimensional stochastic analysis of macrodispersion in aquifers, *Water Resour. Res.*, *19*(1), 161–180, 1983.

Geman, S., and D. Geman, Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images, *IEEE Trans. Pattern Analysis Mach. Intel.*, *PAMI-6*(6), 721–741, 1984.

Haldorsen, H. H., and E. Damsleth, Stochastic modeling, *J. Pet. Technol.*, *42*(4), 404–412, 1990.

Hess, K. M., S. H. Wolf, and M. A. Celia, Large-scale natural gradient tracer test in sand and gravel, Cape Cod, Massachusetts, 3, Hydraulic conductivity variability and calculated macrodispersivities, *Water Resour. Res.*, *28*(8), 2011–2027, 1992.

Hoeksema, R. J., and P. K. Kitanidis, Analysis of the spatial structure of properties of selected aquifers, *Water Resour. Res.*, *21*(44), 563–572, 1985.

Isaaks, E. H., and R. M. Srivastava, *An Introduction to Applied Geostatistics*, Oxford Univ. Press, New York, 1989.

Johnson, N. M., and S. J. Dreiss, Hydrostratigraphic interpretation using indicator geostatistics, *Water Resour. Res.*, *25*(12), 2501–2510, 1989.

LeBlanc, D. R., S. P. Garabedian, K. M. Hess, L. W. Gelhar, R. D. Quadri, K. G. Stollenwerk, and W. W. Wood, Large-scale natural gradient tracer test in sand and gravel, Cape Cod, Massachusetts, 1, Experimental design and observed tracer movement, *Water Resour. Res.*, *27*(5), 895–910, 1991.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, *J. Chem. Phys.*, *21*(6), 1087–1092, 1953.

Molz, F. J., and G. K. Boman, A fractal-based stochastic interpolation scheme in subsurface hydrology, *Water Resour. Res.*, *29*(11), 3769–3774, 1993.

Neuman, S. P., Universal scaling of hydraulic conductivities and dispersivities in geologic media, *Water Resour. Res.*, *26*(8), 1749–1758, 1990.

Poeter, E., and P. Townsend, Assessment of critical flow path for improved remediation management, *Ground Water*, *32*(3), 439–447, 1994.

Rehfeldt, K. R., P. Hufschmied, L. W. Gelhar, and M. E. Schaefer, Measuring hydraulic conductivity with the borehole flowmeter, *EPRI Top. Rep. EN-6511*, Electr. Power Res. Inst., Palo Alto, Calif., 1989.

Rehfeldt, K. R., J. M. Boggs, and L. W. Gelhar, Field study of dispersion in a heterogeneous aquifer, 3, Geostatistical analysis of hydraulic conductivity, *Water Resour. Res.*, *28*(12), 3309–3324, 1992.

Ritzi, R. W., D. F. Jayne, A. J. Zahradnik, A. A. Field, and G. E. Fogg, Geostatistical modeling of heterogeneity in glaciofluvial, buried-valley aquifers, *Ground Water*, *32*(4), 666–674, 1994.

Scheibe, T. D., and C. R. Cole, Non-Gaussian particle tracking: Application to scaling of transport processes in heterogeneous porous media, *Water Resour. Res.*, *30*(7), 2027–2039, 1994.

Scheibe, T. D., and D. L. Freyberg, Impacts of geological structure on transport: Creating a data base, in *Fifth Canadian/American Conference on Hydrogeology*, edited by S. Bachu, pp. 56–71, Natl. Water Well Assoc., Dublin, Ohio, 1990.

Smith, L., and F. W. Schwartz, Mass transport, 3, Role of hydraulic conductivity data in prediction, *Water Resour. Res.*, *17*(5), 1463–1479, 1981.

Suro-Perez, V., and A. Journel, Indicator principal component kriging, *Math. Geol.*, *23*(5), 758–788, 1991.

Webb, E. K., Simulating the three-dimensional distribution of sediment units in braided-stream deposits, *J. Sediment. Res.*, *Sect. B*, *64*(2), 219–231, 1994.

Woodbury, A. D., Bayesian updating revisited, *Math. Geol.*, *21*(3), 285–308, 1989.

Woodbury, A. D., and E. A. Sudicky, The geostatistical characteristics of the Borden aquifer, *Water Resour. Res.*, *27*(4), 533–546, 1991.

J. R. Eggleston, J. J. Peirce, and S. A. Rojstaczer, Center for Hydrologic Science, Box 90230, Duke University, Durham, NC 27708.

(Received December 1, 1995; revised January 26, 1996;
accepted January 26, 1996.)