WATER RESOURCES RESEARCH, VOL. 34, NO. 6, PAGES 1373–1413, JUNE 1998

A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow

D. A. Zimmerman,[1] G. de Marsily,[2] C. A. Gotway,[3] M. G. Marietta,[4] C. L. Axness,[4] R. L. Beauheim,[4] R. L. Bras,[5] J. Carrera,[6] G. Dagan,[7] P. B. Davies,[4] D. P. Gallegos,[4] A. Galli,[8] J. Gómez-Hernández,[9] P. Grindrod,[10] A. L. Gutjahr,[11] P. K. Kitanidis,[12] A. M. Lavenue,[13] D. McLaughlin,[5] S. P. Neuman,[14] B. S. RamaRao,[13] C. Ravenne,[15] and Y. Rubin[16]

Abstract. This paper describes the first major attempt to compare seven different inverse approaches for identifying aquifer transmissivity. The ultimate objective was to determine which of several geostatistical inverse techniques is better suited for making probabilistic forecasts of the potential transport of solutes in an aquifer where spatial variability and uncertainty in hydrogeologic properties are significant. Seven geostatistical methods (fast Fourier transform (FF), fractal simulation (FS), linearized cokriging (LC), linearized semianalytical (LS), maximum likelihood (ML), pilot point (PP), and sequential self-calibration (SS)) were compared on four synthetic data sets. Each data set had specific features meeting (or not) classical assumptions about stationarity, amenability to a geostatistical description, etc. The comparison of the outcome of the methods is based on the prediction of travel times and travel paths taken by conservative solutes migrating in the aquifer for a distance of 5 km. Four of the methods, LS, ML, PP, and SS, were identified as being approximately equivalent for the specific problems considered. The magnitude of the variance of the transmissivity fields, which went as high as 10 times the generally accepted range for linearized approaches, was not a problem for the linearized methods when applied to stationary fields; that is, their inverse solutions and travel time predictions were as accurate as those of the nonlinear methods. Nonstationarity of the "true" transmissivity field, or the presence of "anomalies" such as high-permeability fracture zones was, however, more of a problem for the linearized methods. The importance of the proper selection of the semivariogram of the $\log_{10}(T)$ field (or the ability of the method to optimize this variogram iteratively) was found to have a significant impact on the accuracy and precision of the travel time predictions. Use of additional transient information from pumping tests did not result in major changes in the outcome. While the methods differ in their underlying theory, and the codes developed to implement the theories were limited to varying degrees, the most important factor for achieving a successful solution was the time and experience devoted by the user of the method.

1. Introduction

1.1. Background

For many practical problems of groundwater hydrology, such as aquifer development, contaminated aquifer remediation, or performance assessment of planned waste disposal projects, it is no longer enough to determine the "best estimate" of the distribution in space of the aquifer parameters. A measure of the uncertainty associated with this estimation is also needed. Geostatistical techniques are ideally suited to filling this role. Basically, geostatistics fits a "structural model" to the data, reflecting their spatial variability. Then, both "best estimates" (by kriging) and the variance of the estimation error can be developed. Geostatistical techniques can also produce "conditional simulations" that honor the data at measurement points and, through multiple realizations, display the uncertainty in the spatial distribution of the parameters. These conditional simulations can then be used in a Monte Carlo analysis (e.g., as input to groundwater flow and transport models) to display the uncertainty in the final outcome of the study (flow rates, concentrations, travel times, etc). In some cases the probability distribution function (pdf) of the final outcome can be directly predicted analytically from the "structural" characteristics of the data. Nongeostatistical approaches such as weighted least squares optimization followed by sensitivity studies to assess parameter uncertainty have also been used. Examples of such simulations have been given by *Delhomme*

[1]GRAM, Inc., Albuquerque, New Mexico.
[2]Université Paris IV, Paris, France.
[3]Centers for Disease Control and Prevention, Atlanta, Georgia.
[4]Sandia National Laboratories, Albuquerque, New Mexico.
[5]Massachusetts Institute of Technology, Cambridge.
[6]Universitat Politècnica de Cataluña, Barcelona, Spain.
[7]Tel Aviv University, Tel Aviv, Israel.
[8]Ecole de Mines de Paris, Fontainebleau, France.
[9]Universidad Politécnica de Valencia, Valencia, Spain.
[10]QuantiSci, Ltd., Henley-on-Thames, England, United Kingdom.
[11]New Mexico Institute of Mining and Technology, Socorro.
[12]Stanford University, Stanford, California.
[13]Duke Engineering and Services, Inc., Austin, Texas.
[14]University of Arizona, Tucson.
[15]Institut Français du Pétrole, Rueil-Malmaison, France.
[16]University of California, Berkeley.

Paper number 98WR00003.
0043-1397/98/98WR-00003$09.00

[1979], *Dagan* [1985, 1989], *Rubin and Dagan* [1987, 1992], *Rubin* [1991a, b], *Rubin and Journel* [1991], *Desbarats and Srivastava* [1991], *Robin et al*. [1993], *Gutjahr et al*. [1994], *Harvey and Gorelick* [1995], and *Koltermann and Gorelick* [1996], among others.

In groundwater hydrology, data may come from at least four sources: (1) transmissivity (or permeability) measurements; (2) hydraulic head measurements; (3) tracer concentrations in wells from tracer tests; and (4) geologic information on the nature and characteristics of the formation. The incorporation of geologic information is generally made by zoning the parameter field or by including a trend or a piecewise-varying identification of the structural model. For example, the inclusion of geophysical information has been presented by *Rubin et al*. [1992], *Copty et al*. [1993], and *Hyndman et al*. [1994]. Because it is difficult to incorporate these types of information in an "inverse" approach simultaneously, it has been common engineering practice to calibrate models by "trial and error," sometimes using sensitivity analyses and optimization subroutines to accelerate the fitting [e.g., *Dettinger and Wilson*, 1981; *Peck et al*., 1988]. Such approaches are, however, limited to producing a "best estimate" and can only assess a residual uncertainty (i.e., an estimate of the confidence interval of each parameter after calibration) by a postcalibration sensitivity study. This approach is insufficient to characterize the uncertainty after calibration. Therefore a large number of geostatistically-based inverse techniques have been developed for handling both head and transmissivity data. In general, these techniques follow these steps: (1) Calibrate a "structural model" of the spatial variability using either the transmissivity data only or the transmissivity and the head data; (2) determine the cross covariance between the transmissivity and the head; (3) use an optimization procedure to estimate the transmissivity based on autocovariances and cross covariances. Alternatively, the estimation can be replaced by simulations of alternative realizations of the transmissivity fields.

A good review and comparison of a number of these approaches has been very recently prepared by *McLaughlin and Townley* [1996], who not only described the approaches but also presented them in a unified framework, with a discussion of their respective theoretical merits. Prior to that, several surveys had been presented by *Kuiper* [1986], *Yeh* [1986], *Carrera and Neuman* [1986a], *Carrera* [1988], *Ginn and Cushman* [1990], *Keidser and Rosbjerg* [1991], *Ahmed and de Marsily* [1993], and *Sun* [1994], among others.

Although we believe that this comparison is to date the largest effort undertaken to evaluate inverse approaches objectively, it is worth mentioning here the comparison of four inverse techniques (the pilot point, pure zoning, a combination of zoning and kriging, and a version of linear cokriging) published by *Keidser and Rosbjerg* [1991] on four different data sets that included both hydraulic and contaminant data. They compared the precision and robustness of the approaches and concluded that pure zonation (without any geostatistical assumptions) was superior to the other approaches when data are scarce or when measurement errors exist. Pilot point performed best for reproducing large-scale heterogeneities, the combination of zoning and kriging was robust and flexible, and linear cokriging was found to be very sensitive to the reliability of the T data. Pure zoning did not perform well in the case of fairly complex aquifers. *Rubin and Dagan* [1987] used the data on the Avra Valley presented by *Clifton and Neuman* [1982] in an inverse method different from that of these authors (a linear semianalytical method for the former, a maximum likelihood estimate using zoning for the latter). They concluded that reasonably similar results had been obtained by the two approaches. *Carrera and Glorioso* [1991] also compared linear and nonlinear approaches and obtained similar conclusions, except for large variances of $\ln(T)$, for large head measurement errors, or in the presence of sink/source terms. Under any of these conditions they found that nonlinear approaches performed much better than the linear ones. The reader is also referred to special issues of *Advances in Water Resources* (volume 14, numbers 2 and 5, 1991), in which a large number of inverse approaches have been presented and analysed. Finally, it should be noted that this paper represents the completion of the in-progress comparison study presented by *Zimmerman et al*. [1996].

1.2. Motivation for This Study

A comparison of inverse approaches was undertaken by Sandia National Laboratories (SNL) in conjunction with the performance assessment (PA) of the Waste Isolation Pilot Plant (WIPP) site. The WIPP is a U.S. Department of Energy (DOE) facility currently being evaluated to assess its suitability for isolating transuranic wastes generated by the defense programs in the United States. It should be noted that this work was not performed in accordance with the SNL WIPP quality assurance (QA) program and that none of these results are to be referenced for any work performed under the SNL WIPP QA program.

The proposed repository is located within the bedded salt of the Salado Formation at a depth of about 650 m. A description of the WIPP site and of the first application of an inverse technique to this site is given by *Lappin* [1988], *LaVenue and Pickens* [1992], and *LaVenue et al*. [1995].

The Culebra Dolomite, a 7-m-thick member of the 120-m-thick Rustler Formation located at a depth of about 250 m, has been characterized as the most transmissive, laterally continuous hydrogeologic unit above the repository and is considered a potentially important transport pathway for off-site radionuclide migration within the subsurface. This transport could occur if, in the future, a well drilled for exploration purposes created an artificial connection between the waste storage rooms and the Culebra, allowing radionuclides to leak into the Culebra. Such a scenario is part of a probabilistic PA that the U.S. Environmental Protection Agency (EPA) requires DOE to perform to demonstrate compliance of the repository system with regulations governing disposal of radioactive wastes [*EPA*, 1985; *Sandia National Laboratories*, 1992]. The data base for modeling the Culebra is available from *Cauffman et al*. [1990]. Because the EPA regulation is probabilistic, the PA must adequately reflect the variability and uncertainty within all factors that contribute to the simulation of the repository performance for isolating wastes.

For performance assessment of a nuclear waste repository a hydrologist must provide not just a transmissivity field or a series of transmissivity fields but the probability density function (pdf) of the outcome of the flow simulation (the travel time). Thus the inverse problem serves only as the means for estimating this pdf, conditioned on the available data. The current probabilistic approach to PA [*Sandia National Laboratories*, 1991] accommodates parameter correlations, including spatial correlations, and conditioning on sample data. For a contaminant transport problem such as radionuclide migration in the Culebra at the WIPP, the focus is on adequately

Table 1. The Seven Inverse Methods Compared

Inverse Method	Symbol	First Author	Affiliation
Fast Fourier transform	FF	A. Gutjahr	New Mexico Institute of Mining and Technology
Fractal simulation	FS	P. Grindrod	QuantiSci, United Kingdom
Linearized cokriging	LC	P. Kitanidis	Stanford University
Linearized semianalytical	LS	Y. Rubin	University of California, Berkeley
Maximum likelihood	ML	J. Carrera	Universitat Politècnica de Cataluña, Spain
Pilot point method	PP	B. S. RamaRao	Duke Engineering and Services, Inc.
Sequential self-calibration	SS	J. Gómez-Hernández	Universidad Politècnica de Valencia, Spain

characterizing the hydraulic properties of the medium and their uncertainty. However, the real quantity of interest is the conditional pdf of the PA outcome. Thus solving the inverse problem is not an objective per se but is just a means to generate adequate intermediate parameter fields to be used in the PA simulations.

1.3. Objectives

In this study we compare seven inverse approaches, outline their differences, and discuss their potential strengths and weaknesses. The results point to areas of research that may be useful for improving the inverse techniques. This paper addresses the following issues by comparing the different inverse approaches on three different test problems: (1) How different are the inverse techniques considered in this paper? (2) How effective are they for solving practical problems? (3) How dependent are they on the various assumptions that are made to derive the algorithm, for example, statistical homogeneity, Gaussian distributions, and small magnitude of the log transmissivity ($\log_{10}(T)$) variance? The problem sets are artificial in order to be able to compare the approaches to one another and also with a synthetic "truth." Each test was also designed to be comparable with that of an actual site and to address the validity of the underlying assumptions inherent in the different approaches.

We have chosen to consider the advective groundwater travel time (GWTT) of a conservative tracer as a surrogate for the more complex solute transport problem. We will therefore generate pdf's of GWTT (as the PA outcome) and evaluate inverse approaches on their ability to reflect the uncertainty in aquifer parameters adequately as described by these conditional GWTT pdf's. Our objective is to reveal how the estimate of the conditional pdf's of GWTT can be affected by either the differences in the principles and coding of the inverse methods or the manner in which a given method was applied by the person who ran it.

1.4. Geostatistical Approaches To Be Compared

Seven inverse methods were selected for comparison. The selected methods were to estimate the transmissivity field from measurements of transmissivity and head and produce an ensemble of simulated transmissivity fields conditioned on all the available data on transmissivity and head. These simulated transmissivity fields should reflect the uncertainty in the transmissivity estimate after calibration and would be the input T fields in the Monte Carlo simulations of flow through the system. There should be as many different T fields as Monte Carlo simulations (about 100), all considered as having an equal probability of occurrence.

The geostatistical inverse approaches are listed in alphabetical order in Table 1. In Appendix B we give a short summary of the description of each method, with references to the major publications where the methods were presented and applied. For clarity and brevity we will refer to the methods by their two-letter symbols (see Table 1). These seven methods are by no means an exhaustive sampling of all the methods that have been published in the literature. Among the most prominent "absences" are the approaches proposed by *Cooley* [1977, 1979, 1982, 1983], *Townley and Wilson* [1985], and *Sun and Yeh* [1992], who unfortunately could not participate.

The seven approaches can be categorized as being either linearized or nonlinear. While the groundwater flow equation for confined aquifers is always linear for the head, this same equation is nonlinear for the relation of T to head. The linearized approaches are generally based upon simplifying assumptions about the flow field (e.g., a uniform hydraulic head gradient, a small $\ln(T)$ variance, etc.), that lead to a linearized relation between T and head using a perturbation expansion of the head and transmissivity fields. This equation can then be solved analytically or numerically. The nonlinear approaches have no such restrictions placed on them and can, in principle, handle more complex flow fields or larger $\ln(T)$ variances. Methods FF, LC, and LS fall into the linearized category, while methods FS, ML, PP, and SS fall into the nonlinear one.

The LS method is able to calculate the GWTT cumulative distribution functions (CDFs) directly, so this method did not produce transmissivity fields. T fields could have been produced by this method, but these fields would then not have been linked to a particular travel path or travel time and so were not calculated.

1.5. Overview of the Test Problem Exercise

The test problem exercise was conceived and performed by group of participants referred to by SNL as the Geostatistical Expert Group (GXG). A listing of the participants is given in Appendix A. Four test problems were developed in secrecy from the participants who would receive the data and run the inverse models. The test problems were designed to be "WIPP-like," meaning that the hydrogeologic characteristics and the complexity of the problems, as well as the type of data and their spatial distribution, should be relatively similar to that of the WIPP site. The synthetic transmissivity fields should also have properties similar to those observed at WIPP or believed to exist at the WIPP on the basis of inference from geological and hydrological data. Four different T fields were generated. Synthetic hydraulic head data were obtained by solving the two-dimensional flow equations with prescribed boundary conditions using these synthetic T fields. A limited number of observations of head and transmissivity obtained from the exhaustive (synthetic) data sets would then be provided to the participants. Additionally, particle-tracking calculations were performed to compute advective travel times and travel paths

of a conservative solute for the synthetic data sets. Particles were released in a number of locations and the "true" groundwater travel times were calculated but not given to the participants.

For each test problem the participants would analyze the sampled T and head data (about 40 observations of each) and use their inverse procedure to generate the ensemble of conditional transmissivity fields and corresponding head fields (in general between 50 and 100) that were given to the GXG coordinator. The coordinator would then calculate the travel times and travel paths for the same release points as those in the "true" field, using the same particle-tracking code as the one used for the true field but using the T values, the grid size, and the boundary conditions specified by the participants as a result of their efforts. Throughout this paper the term "GWTT" is defined as the time it takes for a particle to reach a radial distance of 5 km from the release point. The calculated GWTTs taken across all realizations produced by a method were used to construct a GWTT CDF which was compared to the "true GWTT." This is referred to as the "fixed well approach," described in more detail below. In a second set of analyses (the "random well approach," also described in detail below) the GWTTs from an ensemble of release points contained within a localized area were used to construct the "true GWTT CDF," which was then compared with the calculated GWTT CDFs for the same release points from each of the methods. In the case of the linearized semianalytical method, only a particle travel time CDF was requested because this method does not require generation of a transmissivity field to estimate this CDF.

In the real world it is clear that parameters other than transmissivity are variable and uncertain in the system. For example, porosity, aquifer thickness, dispersivity, sorptive properties, etc., are all variable, and the GXG made suggestions on how to incorporate these uncertainties into the PA. However, for the present intercomparison, only the transmissivity is involved, and all other parameters are given uniform values.

2. Description of the Four Test Problems

The test problems (TPs) were developed as a series of independent synthetic data sets that were intended to span the range of possible conceptual models of the Culebra transmissivity distribution at the WIPP site. Estimates of transmissivity at 41 boreholes at the actual WIPP site have been obtained through slug tests, local pumping tests, and three regional-scale pumping tests lasting from 1 to 3 months [*Beauheim*, 1991]. The T values obtained from these tests span 7 orders of magnitude. Analyses of these data indicate that it is likely that the spatial distribution of heterogeneity is not random, but made of specific zones of high and low values. Transmissivity is strongly impacted by the presence or absence of open fractures. Large-scale pumping tests indeed suggest that narrow, relatively conductive fracture zones are possible in some areas. Whether these fractures form a connected network or are isolated from each other by low-transmissivity zones is not clear. In other areas local well tests have indicated the existence of rather low-permeability zones. There are lithologic indicators of high or low transmissivities such as the presence or absence of gypsum filling in fractures, although these indicators are not strict. An attempt has been made in the test problems to represent the presence or absence of such features.

Although the PA calculations to date have assumed a perfectly confined, two-dimensional flow system for the Culebra, there may be vertical flow into or out of the Culebra. Vertical leakage is therefore reflected in some of the test cases. There is also a known salinity gradient in the Culebra which was not considered in the test problems, as most inverse approaches assume constant density.

Hydraulic heads obtained prior to the WIPP site characterization activities when the system was in a quasi–steady state condition were available at 32 locations. Transmissivity estimates were available at 41 locations. Thus the test problems were developed as steady state systems, and the sample data were limited to, at most, 41 observations of head and transmissivity (at the same locations). The spatial distribution of the boreholes (i.e., density, pattern) in the TPs were kept similar to that present at the WIPP. Three large-scale pumping tests were also simulated in TPs 3 and 4. In the real world these data are all subject to measurement errors. However, none was considered in these calculations because the objective of the comparison was not to assess the robustness of an approach to the magnitude of measurement errors, but for a given set of data, to determine the residual uncertainty on the transport properties of the domain as evaluated by each approach. Adding a measurement error would only increase this uncertainty and decrease the ability to distinguish between the approaches. Another reason is that the synthetic data were generated on a very small grid (20–40 m) and the participants were given the grid values at the sampled locations. Thus the small-scale variability of the synthetic log (T) fields can be viewed as measurement error, compared to a "measured value" which could have been provided by averaging over the larger domain such as that which an actual pumping test would have produced.

Boundary conditions for flow in the vicinity of the WIPP site are not well constrained. Thus the boundary conditions were not defined for the participants. Given the 41 head measurements in the domain, they were asked to select the boundary conditions they felt appropriate.

In test problems 1 and 2, the synthetic T fields were generated as unconditional random fields using the two-dimensional random field generator TUBA (*Zimmerman and Wilson*, [1990]; see also work by *Mantoglou and Wilson* [1982] and *Matheron* [1973]). In test problems 3 and 4 the initial field was also generated using TUBA, but additional discrete modifications were made to each. For all test problems, Dirichlet boundary conditions (different for each test problem) were developed for calculating the synthetic heads by generating a stationary random field and adding that to a trend surface. These dense synthetic data sets comprised from one to three million nodes. In all test cases a uniform mesh was used and the true head-field solution was obtained via a multigrid solver (finite difference method) provided by Pacific Northwest Laboratory (see acknowledgments). The size of the area over which the observation data are distributed is 20 km $\times$ 20 km for TPs 1, 2, and 3 and approximately 30 km $\times$ 30 km for TP 4. This is of similar scale to the area where data are available at the actual WIPP site.

The exact correlation structure of each synthetic data set was determined via semivariogram analysis using GSLIB routines [*Deutsch and Journel*, 1992]. Over 3600 randomly located sample points were used in the computation of the exhaustive data set semivariograms in order to obtain enough pairs for stable semivariogram estimates from a single realization (the "true" field). An exponential semivariogram model was then fit to each empirical semivariogram via nonlinear regression in or-

der to report a correlation length parameter and a variance. Exponential semivariogram models have been fit to the WIPP site $\log_{10}(T)$ data and an exponential semivariogram model was used to generate the $\log_{10}(T)$ fields for TPs 1 and 2. The main features of each test problem, including means, variances, correlation lengths, etc., are summarized in Tables 2a and 2b.

2.1. Test Problem 1

TP 1 was the simplest conceptual model. It was developed using a model of the Culebra transmissivities that was based on a geostatistical analysis of the real WIPP site data. The $\log_{10}(T)$ field (T in m^2/s) was modeled as an isotropic process having a mean of -5.5, a variance of 1.5, and an exponential covariance structure with correlation length $\lambda = 3905$ m, close to the values of the real WIPP site. A map of the synthetic $\log_{10}(T)$ field with the location of the observation points is shown in Figure 1. A large regional field (40 km $\times$ 40 km) was generated on a 1793 $\times$ 1793 size grid (over 3.2 million unknowns) with each grid block being 22.5 m on a side. However, the observation data were located in the central 20 km $\times$ 20 km area which is the portion of the field that is shown in Figure 1. The mean and variance of the exhaustive $\log_{10}(T)$ data for the inner region are -5.84 and 1.56, respectively. The sample data consisted of 41 transmissivity and 32 head measurements taken from the exhaustive synthetic data set.

Boundary conditions were generated using a combination of a linear trend surface and spatially correlated noise. The trend surface, given by $Z = 890 + 0.36X + 1.28Y$, with X and Y in km, was derived from an analysis of the WIPP site data to provide a similar head gradient. An anisotropic Gaussian process with zero nugget, a sill of 50 m^2, and ranges of 5 and 15 km in the north-south and east-west directions, respectively, was used to model the hydraulic head variability for generating the boundary values.

2.2. Test Problem 2

The second test problem data set was generated specifically to examine how well the linearized techniques could handle high-variance cases. The model of spatial variability of TP 2 is identical to TP 1; only the mean and variance of $\log_{10}(T)$ were changed. In fact, the pattern of spatial variability remains exactly the same except that the field is rotated counterclockwise by 90°. The mean of $\log_{10}(T)$ was increased to -1.26, resulting in faster travel times, and the $\log_{10}(T)$ variance was increased to 2.14. The boundary values remained the same, albeit rotated by 90°. The same number and similar configuration of observation data as for TP 1 were provided to the participants. The sample $\log_{10}(T)$ data have a mean of -0.52 and a variance of 2.39. The $\log_{10}(T)$ field and observation points are shown in Figure 2.

Table 2a. $\text{Log}_{10}(T)$ Field Exhaustive Data Set and Sample Data Statistics

TP	Covariance Model	Exhaustive Data μ	Exhaustive Data σ^2	Sample Data μ	Sample Data σ^2	Observations Head	Observations $\log_{10}(T)$
1	exponential	−5.84	1.56	−5.30	1.84	32	41
2	exponential	−1.26	2.14	−0.52	2.39	32	41
3	Telis	−5.64	1.38	−5.70	1.82	41	41
4	bessel	−5.32	1.93	−5.32	1.89	41	41

T in m^2/s.

Table 2b. Additional Characteristics of Each Test Problem Data Set

TP	True Field Correlation Length, m	Recharge Included?	Transient Pumping?	Well log "Geology"?
1	2808	no	no	no
2	2808	no	no	no
3	425	yes	yes	yes
4	2063	yes	yes	no

2.3. Test Problem 3

The intent of test problem 3 was to incorporate some of the more complex geohydrologic characteristics of the WIPP site. Several high-transmissivity fracture zones approximately 1–3 km apart have been inferred from pumping tests in the northwest and southeast areas of the WIPP site and in other areas of the site; aquifer tests conducted at several wells have resulted in very low transmissivity values.

The transmissivity field of represents a possible nonstationary conceptual model of the WIPP site transmissivity distribution that includes, within a background medium of variable transmissivity, disconnected high-transmissivity "channels" that represent fracture zones, local low-transmissivity subregions representing tight zones, and a large low-transmissivity zone in the southwest corner of the field. Information on "the type of geology encountered in each borehole" was provided to the participants (descriptors such as "porous," "fractured," and "tightly cemented" were used to relate to the "background," the high-T "channels," or the low-T subregions, respectively). Such information would, of course, be available at any real site.

The $\log_{10}(T)$ distribution is shown in Figure 3. The map is

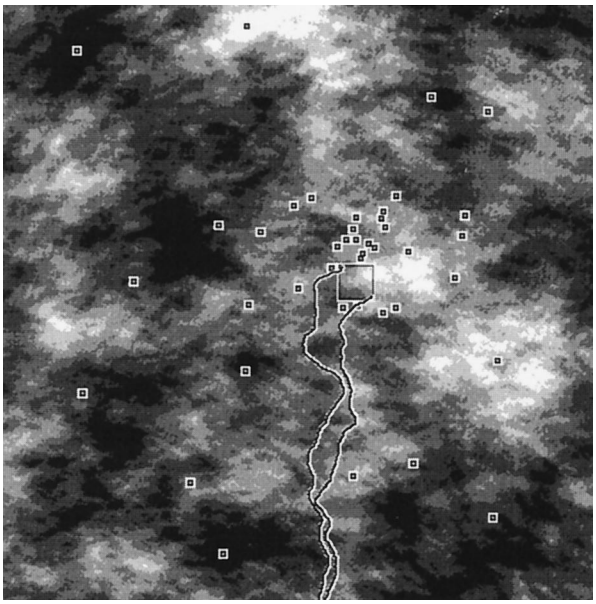


Figure 1. Test problem 1 true $\log(T)$ field (20 km $\times$ 20 km). Squares are assumed waste disposal areas. The flow lines originating from these squares display the flow direction from the disposal area to the boundaries. The six gray shades are $\log_{10}(T)$ intervals of 10^{-7}–10^{-6}, 10^{-6}–10^{-5}, 10^{-5}–10^{-4}, 10^{-4}–10^{-3}, and $>10^{-3}$ (lightest).

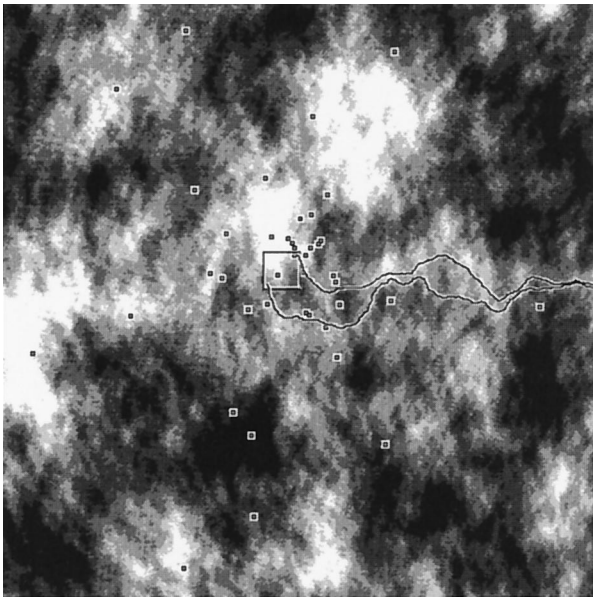


Figure 2. Test problem 2 true log (T) field (20 km × 20 km). Squares are assumed waste disposal areas. The flow lines originating from these squares display the flow direction from the disposal area to the boundaries. The six gray shades are $\log_{10}$ (T) intervals of 10^{-7}–10^{-6}, 10^{-6}–10^{-5}, 10^{-5}–10^{-4}, 10^{-4}–10^{-3}, and 10^{-3}–10^{-1} (lightest).

best thought of as several "geologic overlays." The underlying "geostatistical background field" (the T field without the low-T subregions and high-T fractures) was generated as a stationary field having a $\log_{10}$ (T) mean and variance of −5.5 and 0.8, respectively, and an anisotropic Telis covariance with correlation lengths of 1.025 km and 0.512 km in the east-west and north-south directions, respectively. The high-T "channels" were generated with a $\log_{10}$ (T) mean of −2.5, a variance of 0.1, and an exponential covariance with a correlation length of 6667 m.

This correlation structure applies only along narrow zones described as "fractures," as shown in Figure 3. The low-T zones, shown as dark, nonuniform ovoid regions, had a mean $\log_{10}$ (T) of −7.5, a variance of 0.5, and an isotropic exponential covariance model with a correlation length of 3417 m. In the lower left-hand corner of the field is an area in which there is a trend of decreasing transmissivity toward the corner of the field. After overlaying the low-T subregions and linear features, a 3 × 3 moving window block filter (averaging on the $\log_{10}$ (T)) was passed over the field to help smooth the transition between these zones in order to avoid potential numerical-convergence problems. This did not significantly reduce the variance of the final $\log_{10}$ (T) field, which had a mean and variance of −5.64 and 1.38, respectively. The mean and variance of the sample $\log_{10}$ (T) data are −5.70 and 1.82, respectively (for the 41 sample points).

Vertical recharge was applied uniformly over the northwestern portion of the model domain; the recharge rate was 6.5×10^{-9} m^3/s. The recharge distributed over this region accounts for approximately 10% of the regional flow through the system. Such recharge could be inferred by the participants from the observed heads in this area, which showed a localized piezometric mound, but no information on recharge was given to the participants.

Boundary conditions were generated in a similar fashion as those for TPs 1 and 2, using a combination of linear trend surface and spatially correlated noise. The trend surface was based on an analysis of the WIPP site data, but with the x direction of the trend reversed. The trend model is given by $Z = 890 + 0.36(41 - X) + 1.28Y$, where X and Y are given in km. An anisotropic exponential covariance model having zero nugget, a sill of 50 m^2, and X and Y correlation lengths of 15 and 5 km, respectively, was used to model the head spatial variability for generating the boundary values.

In addition to the steady state hydraulic head data and transmissivity values, transient information was provided to the participants in the form of three independent aquifer tests. Pumping from the aquifer was simulated numerically in three different wells (one at a time) and drawdown data in the surrounding wells and the pumping rates were given to the participants. A uniform storativity value of 5×10^{-6} was assigned to the system (but not made known to the participants). These tests were loosely modeled after the H-3, H-11, and WIPP-13 large-scale pumping tests conducted at the WIPP site [*Beauheim*, 1991]. Details of the three aquifer tests, including drawdowns at each observation well and estimates of transmissivity and storativity based on conventional well-test analysis, were given to the participants.

2.4. Test Problem 4

TP 4 is a complex, nonstationary conceptual model of the transmissivity distribution reflecting large-scale connectivity of fracture zones (contrary to) that have been shown to exist in some areas of the Culebra. The features of the conceptual model included the following: (1) well-connected high-T channels, (2) a variation in transmissivity of 5–6 orders of magnitude, (3) a small trend in $\log_{10}$ (T) (1–2 orders of magnitude for T across the entire field), (4) a local recharge area correlated with the high-T zones, and (5) some high-T zones that

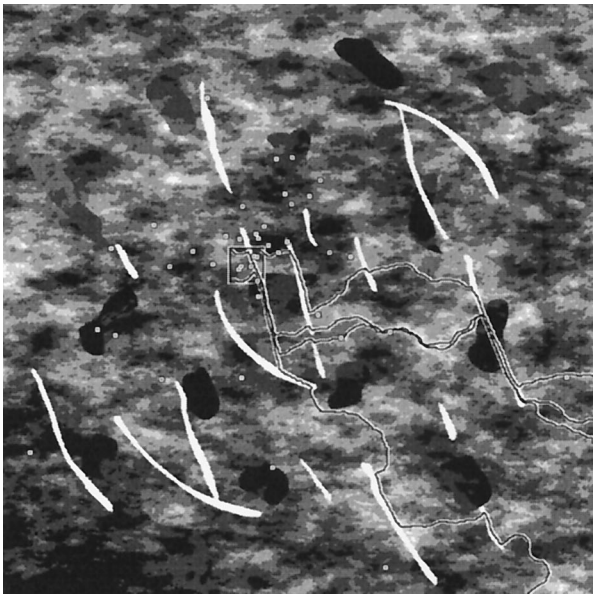


Figure 3. Test problem 3 true log (T) field (20 km × 20 km). Squares are assumed waste disposal areas. The flow lines originating from these squares display the flow direction from the disposal area to the boundaries. The six gray shades are $\log_{10}$ (T) intervals of 10^{-7}–10^{-6}, 10^{-6}–10^{-5}, 10^{-5}–10^{-4}, 10^{-4}–10^{-3}, and $>10^{-3}$ (lightest).

are well-identified while others are missed by the observation wells.

The field was generated as follows: Initially, an unconditioned field having an anisotropic Bessel covariance structure was generated with correlation lengths of 2.05 and 1.025 km in the east-west and north-south directions, respectively. Through a series of repeated kriging exercises, a network of connected high-T channels was developed. Each time the kriging was performed, a number of "fake conditioning points" was added to develop the high-T channels iteratively in the kriged "true field." The final conditionally simulated field was generated via the classical method of conditional simulation described by *Journel and Huijbregts* [1978], being the sum of the kriged true field and the perturbations resulting from the difference between the unconditioned field and the kriged unconditioned field. The final $\log_{10}(T)$ field had a mean of -5.32 and a variance of 1.93; the field is shown in Figure 4 along with the 41 observations points. The field was generated on a 1025×1025 grid with 40-m grid blocks. The $\log_{10}(T)$ mean and variance of the 41 sample observations are -5.32 and 1.89, respectively.

Boundary values were obtained by generating a random head field using a generalized isotropic covariance $C(h) = h^5$ using the TUBA code. The field so generated had a southwest-to-northeast trend diagonally across the field and through the high-T channels. This field was scaled to provide a head difference of 94 m along that diagonal ($\sim$58 km).

Areal recharge was applied to the southern portion of the field where the transmissivity is generally somewhat higher than average; this also helped to direct the flow through the high-T channels without causing any mounding. The recharge (leakage) was applied nonuniformly, being highly correlated with the transmissivity distribution in this area of the field. Because it occurs at the margin of the field, no observation points are located within this region. The recharge amounted to approximately 6% of the regional flow moving through the system, but no recharge information was given to the participants.

As in TP 3, three independent numerical pumping tests were performed to provide transient information for those techniques that could use it. A detailed description of the three tests and conventional analyses of the results were given to the participants.

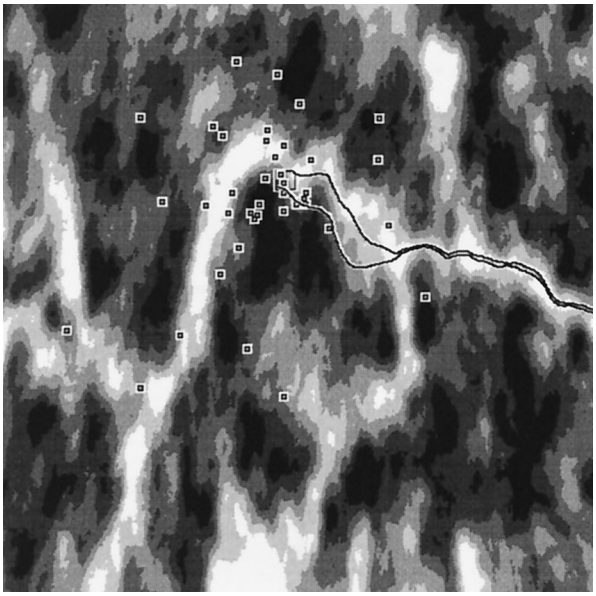


Figure 4. Test problem 1 true log (T) field (30 km $\times$ 30 km). Squares are assumed waste disposal areas. The flow lines originating from these squares display the flow direction from the disposal area to the boundaries. The six gray shades are $\log_{10}(T)$ intervals of 10^{-7}–10^{-6}, 10^{-6}–10^{-5}, 10^{-5}–10^{-4}, 10^{-4}–10^{-3}, and $>10^{-3}$ (lightest).

3. Qualitative Results

In this section we present the results of the groundwater travel time (GWTT) distributions and the transmissivity maps produced by the different approaches along with a statistical analysis of all the evaluation measures.

3.1. Comparison of GWTT CDFs

It is assumed that radionuclides can reach the aquifer when at some future time an exploratory well is drilled through the repository. Because this hypothetical future drilling location is unknown, it is reasonable to treat the unknown location as a random variable within the repository (also referred to as "the waste panel area"). The objective in this analysis is to compute GWTT CDFs for both the true fields and the fields produced by the inverse methods and to compare them. These CDFs represent the uncertainty in GWTT resulting from an intrusion borehole whose location is unknown but which lies somewhere within the waste panel area. What we want to investigate is if the CDFs produced by each approach, for each test problem, are reasonably close to the true CDFs.

Hereinafter, the designations "true field," "true travel time," and "true GWTT CDF" refer to quantities computed using the exhaustive (synthetic) data set.

To construct the true GWTT CDF, a hypothetical repository of similar scale to the waste panel area at the real WIPP site (1.1 km $\times$ 1.1 km) is located within the study area. The repository is located in the zone where the density of the observation data is greatest. Particle tracking is performed for each of 100 particles, distributed uniformly over the waste panel area, out to a radial distance of 5 km. These GWTTs are then used to construct the "true CDF."

To construct the GWTT CDFs for each of the approaches, the procedure was similar, but the uncertainty will of course be larger, since knowledge of the true transmissivity distribution is not perfect. The transmissivity fields used to calculate the travel times are those derived via the inverse procedures. These CDFs, however, will be conditioned on the available data. That is, the velocity fields were computed by solving the forward problem where the boundary conditions, source terms, and discretization were assigned individually by each participant (i.e., they were different for each inverse approach). The GWTT CDFs for each approach were constructed as follows: For each of the 100 release points within the waste panel area, a GWTT CDF was constructed from the ensemble of GWTTs obtained across all realizations. Hence 100 GWTT CDFs are obtained, each CDF being conditional on a particular release point location. The mean CDF for the entire waste panel area was computed as

$$\text{mean CDF} = \int_{\text{waste panel}} \text{CDF}(\mathbf{x})\, f(\mathbf{x})\, d\mathbf{x} \qquad (1)$$

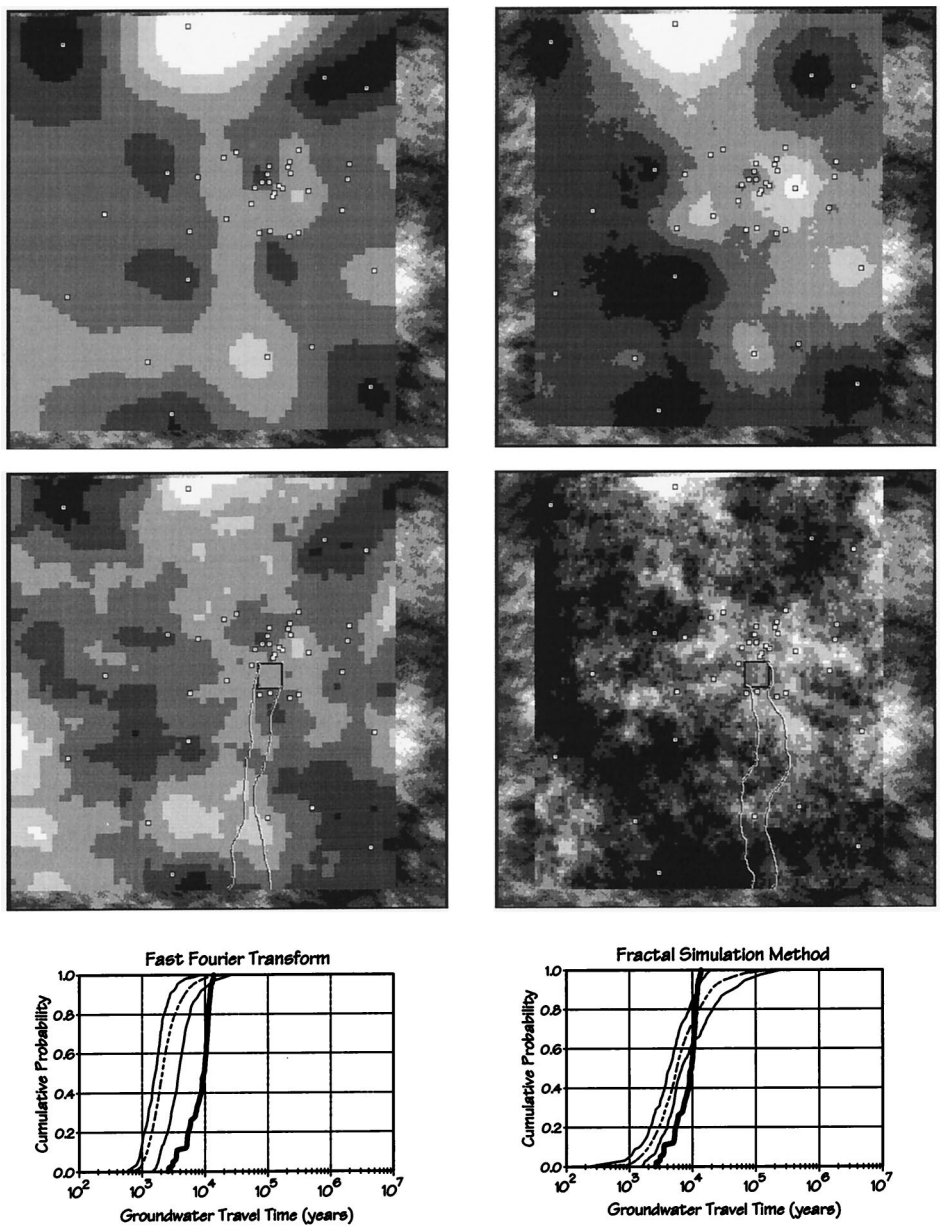


Figure 5a. Test problem 1 mean $\log_{10}(T)$ field, randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, dashed line is mean CDF, and other curves bound the inner 95% of all conditional CDFs.

where $f(\mathbf{x})$ is the probability density function for the borehole location (which was treated as uniform; hence the release points are equally weighted). Figures 1–4 show the position of the repository for each TP, with envelopes of all the true-field particle paths originating from the edges of the waste panel area. Both the true GWTT CDFs (thick line) for each TP and the median GWTT CDF for each method (dashed line) are plotted at the bottom of Figures 5a–8c. The GWTT CDFs for method LS (which did not need to produce T fields) are shown in Figure 9. In addition, on these plots a bounding envelope containing the inner 95% of the CDF curves at each travel time value was constructed. These $\mathrm{GWTT}_{0.025}$ and $\mathrm{GWTT}_{0.975}$ bounding curves reflect the degree of variability in GWTT within the repository area from realization to realization.

This type of analysis allows for the fact that conceptually, the distributions of properties could be identical and yet the underlying T fields different. This is because an identical GWTT may be obtained for very different random well locations in the simulated and true fields. The test only compares the distributions and does not consider if the short or long GWTTs originate from the same locations.

3.2. Comparison of the Transmissivity Fields and Their Semivariograms

As an additional means of comparison, some of the T fields produced by each approach for each TP are shown in Figures 5a–8c. In each figure we have chosen to show both the average of the simulated $\log_{10}(T)$ fields (50–100 simulations) and one individual realization, selected at random. Both maps are "embedded" in the true T fields in order to reveal the area the participants decided to model, the grid orientation, and the level of discretization they used. Also shown in these figures

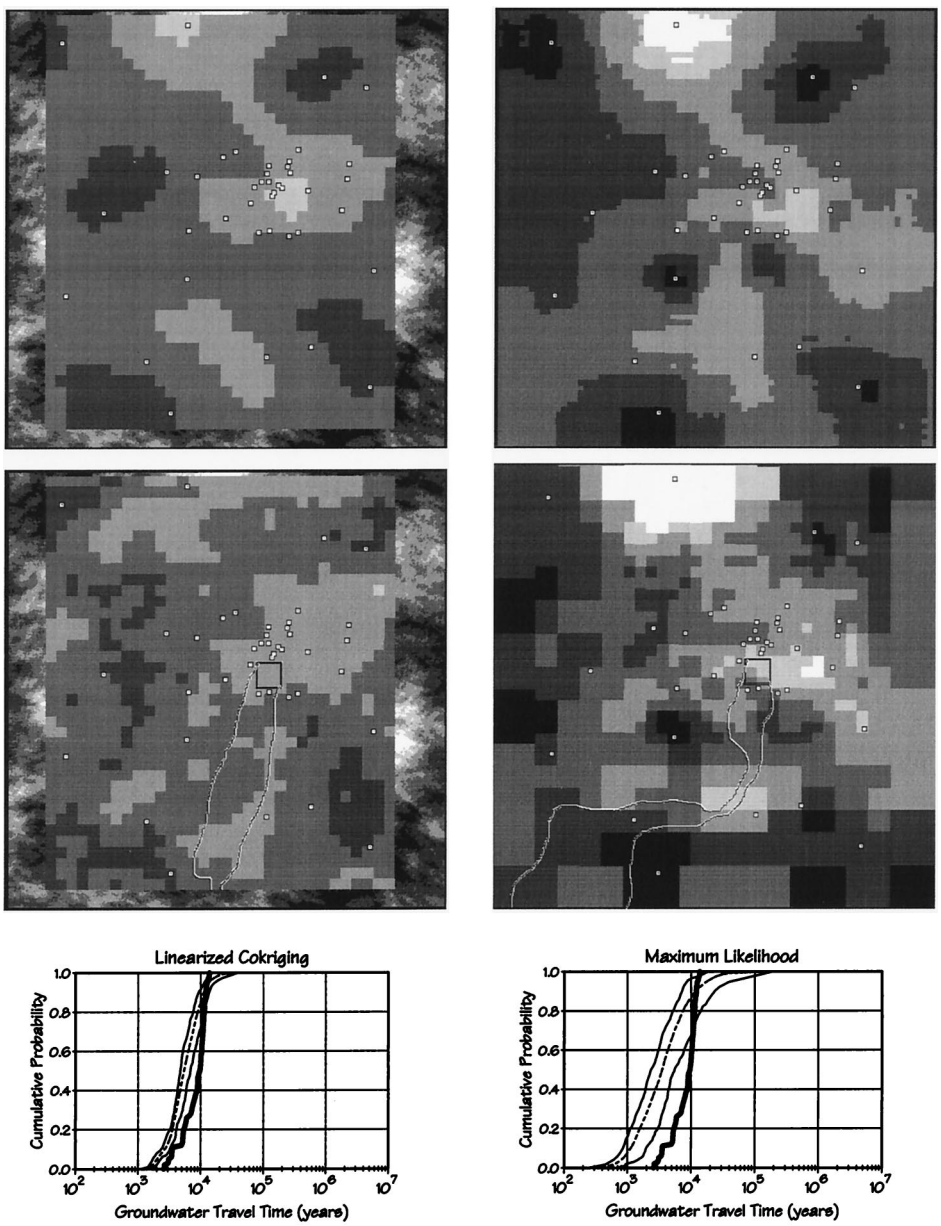


Figure 5b. Test problem 1 mean $\log_{10}(T)$ field (top), randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, dashed line is mean CDF, and other curves bound the inner 95% of all conditional CDFs.

are the envelopes of the travel paths from the edges of the waste panel area. Six gray shades are used; each represents an order-of-magnitude change in the value of transmissivity.

The ability of an inverse method to reproduce the correlation structure of the T field in the realizations correctly was considered an important feature for predicting contaminant transport and spreading. Therefore semivariogram estimates of the simulated $\log_{10}(T)$ fields were computed for each realization of a method, using the GAMV2M routine from the GSLIB software package [*Deutsch and Journel*, 1992]. On the order of 600–1000 randomly placed sampling points were used in the estimation of each semivariogram. Then the average semivariogram was computed across the ensemble of realizations for each TP. For method LS the participant gave directly the parameters of the exponential variogram he had selected. For the other approaches, estimates of the parameters of an exponential semivariogram model fit to each of the average empirical semivariograms (one for each TP) were made via nonlinear regression. The same analysis was performed on each of the true $\log_{10}(T)$ field realizations; approximately 3600 sample values were used for the semivariogram estimates in each of the true-field exhaustive data sets.

3.3. Qualitative Comparison Observations

First, a visual comparison of the mean CDFs with the true CDF reveals large differences among the approaches. Second, no one particular approach is obviously superior to all others, for all TPs.

A third observation is that the mean CDF of each method spans a broader range than the true one: The uncertainty linked to the position of the intrusion borehole is significantly increased when additional uncertainty is introduced by only

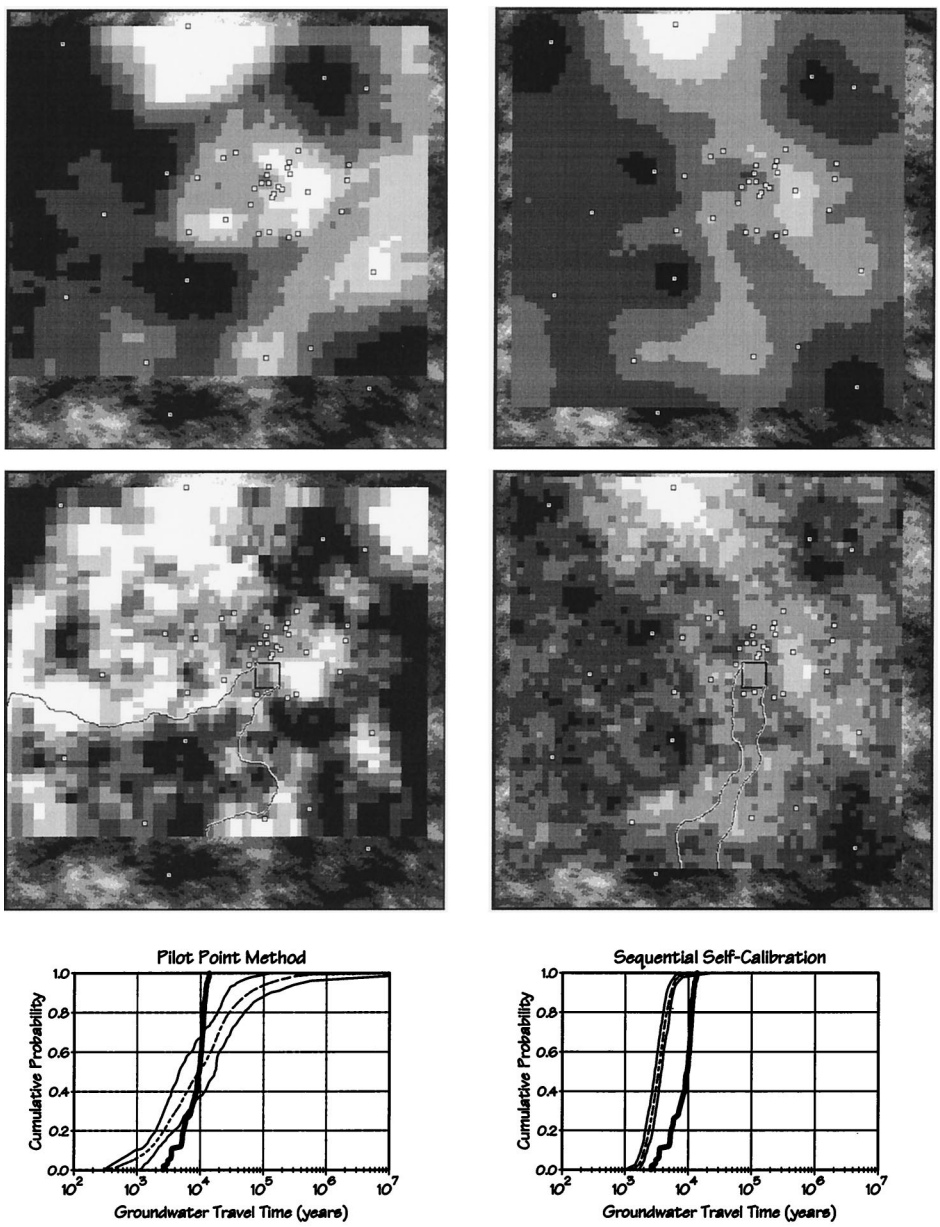


Figure 5c. Test problem 1 mean $\log_{10}(T)$ field, randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, and other curves bound the inner 95% of all conditional CDFs.

incomplete knowledge of the parameters. To give some perspective to these conditional CDFs, we have crudely estimated what would have been the unconditional CDFs, if we had not used any inverse and had directly sampled uncertain parameters in "traditional" Monte Carlo simulations. For this, we assume that we know only, for each TP, the pdf of T from the 41-sample T data and the average head gradient from the head observation data, and we take the same porosity as in all our calculations (16%). For simplicity, we assume that the unconditional T field is uniform over the whole domain and that its value is sampled from a lognormal distribution defined by the mean and variance of the 41 $\log_{10}(T)$ sample data. These four unconditional CDFs are shown in Figure 10. It is clear from this figure that conditioning drastically reduced the uncertainty, which otherwise would have spanned an interval several orders of magnitude larger than the true one.

Fourth, all approaches do relatively well for TPs 1 and 2; their mean CDFs are reasonably close to the true ones, the error is small (less than half an order of magnitude), and the overall uncertainty range is relatively small. The results for TP 1 are generally conservative, and the uncertainties are generally higher in TP 2. But for TPs 3 and 4 the results are in general rather poor. The error can reach several orders of magnitudes, and in general the methods are systematically biased. Thus the predicted GWTT is in general longer than the true one.

We can now examine these results in more detail to reveal some differences among the approaches. Remember that three of the approaches are "linearized" (LC, FF, and LS) and therefore should be sensitive to the magnitude of the variance of the $\log_{10}(T)$ fields. The only difference between TPs 1 and 2 is an increase in the $\log_{10}(T)$ variance (from 1.56 to 2.14, or, for

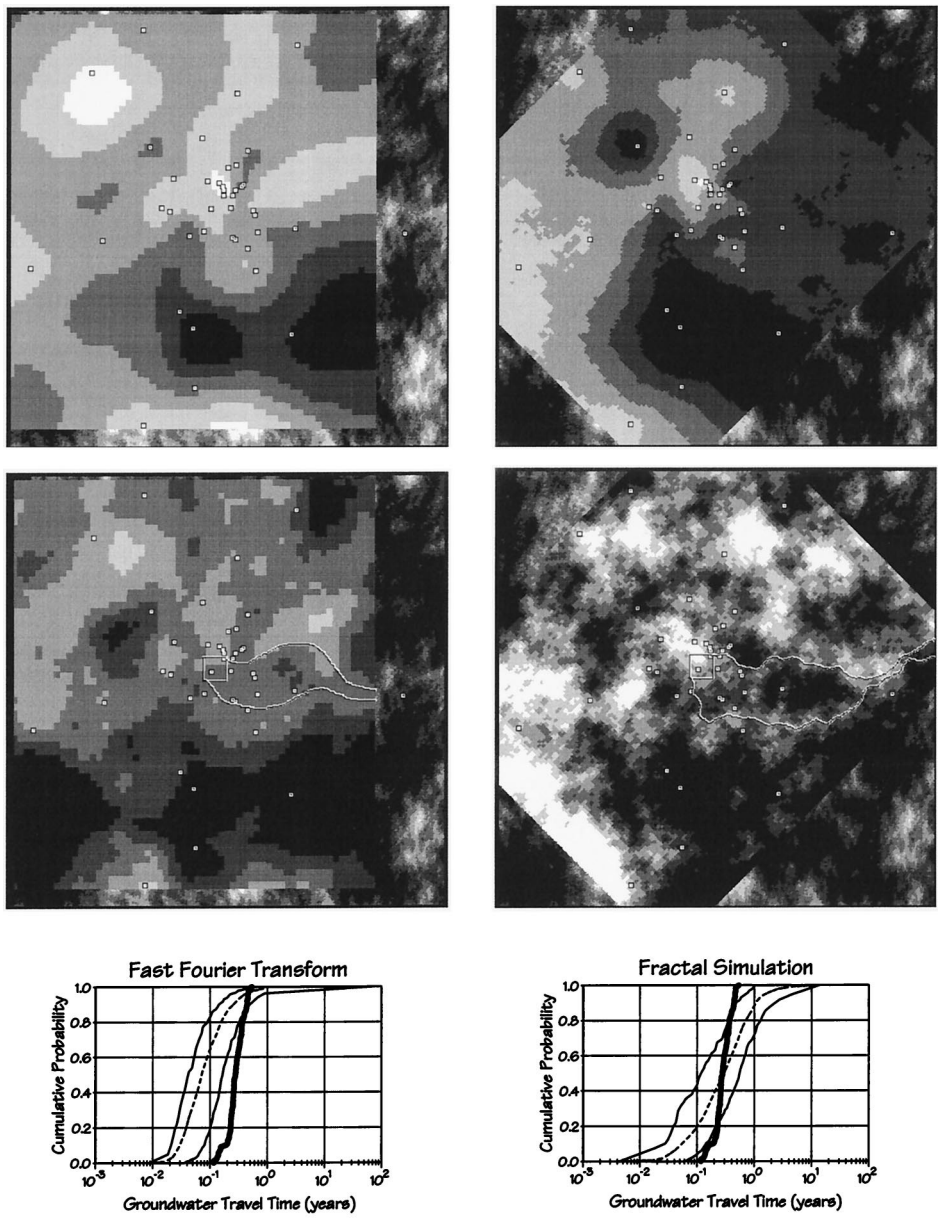


Figure 6a. Test problem 2 mean $\log_{10}(T)$ field, randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, dashed line is mean CDF, and other curves bound the inner 95% of all conditional CDFs.

$\ln(T)$, from 8.25 to 11.32). It is clear, when comparing Figures 5a–9, that the increase in variance did not affect any of the methods, neither the linearized nor the nonlinear ones. The magnitude of the variance of $\log_{10}(T)$ is thus apparently not a critical issue. Linearization is generally assumed valid for $\ln(T)$ variances on the order of 1; the $\ln(T)$ variances in these test problems are well beyond this range. When we look at TPs 3 and 4, it becomes clear that some of the nonlinear approaches systematically perform better than the linearized ones: SS, ML, and PP have mean CDFs substantially closer to the true CDFs than FF and LC. FS, although nonlinear, does not do as well.

Let us now turn to the bounding curves of the waste panel CDFs. These curves reveal how different the CDFs can be, for a given approach, from simulation to simulation. If the bounding curves are very near the mean CDF, it means that the T fields are relatively well known from the available data, by calibration, and that the residual uncertainty is small. This would be desirable only if the mean CDF was very close to the true one. Otherwise, the method can be said to be "overconfident." In PA, overconfidence can be regarded as an unacceptable "sin." This is because the decision on whether or not to license a waste repository, that is, to declare it "safe" with regard to isolating the wastes, would then be based on overly optimistic predictions of the repository's performance. If application of the inverse methods was to result in overconfidence, then their use in PA should be questioned.

Another test for evaluating PA methodology was conducted in the United Kingdom [*Mackay*, 1993] to see how the "VANDAL" PA approach, developed by Her Majesty's Inspectorate of Pollution, would perform in a synthetic case, as the amount of information made available to the modeler

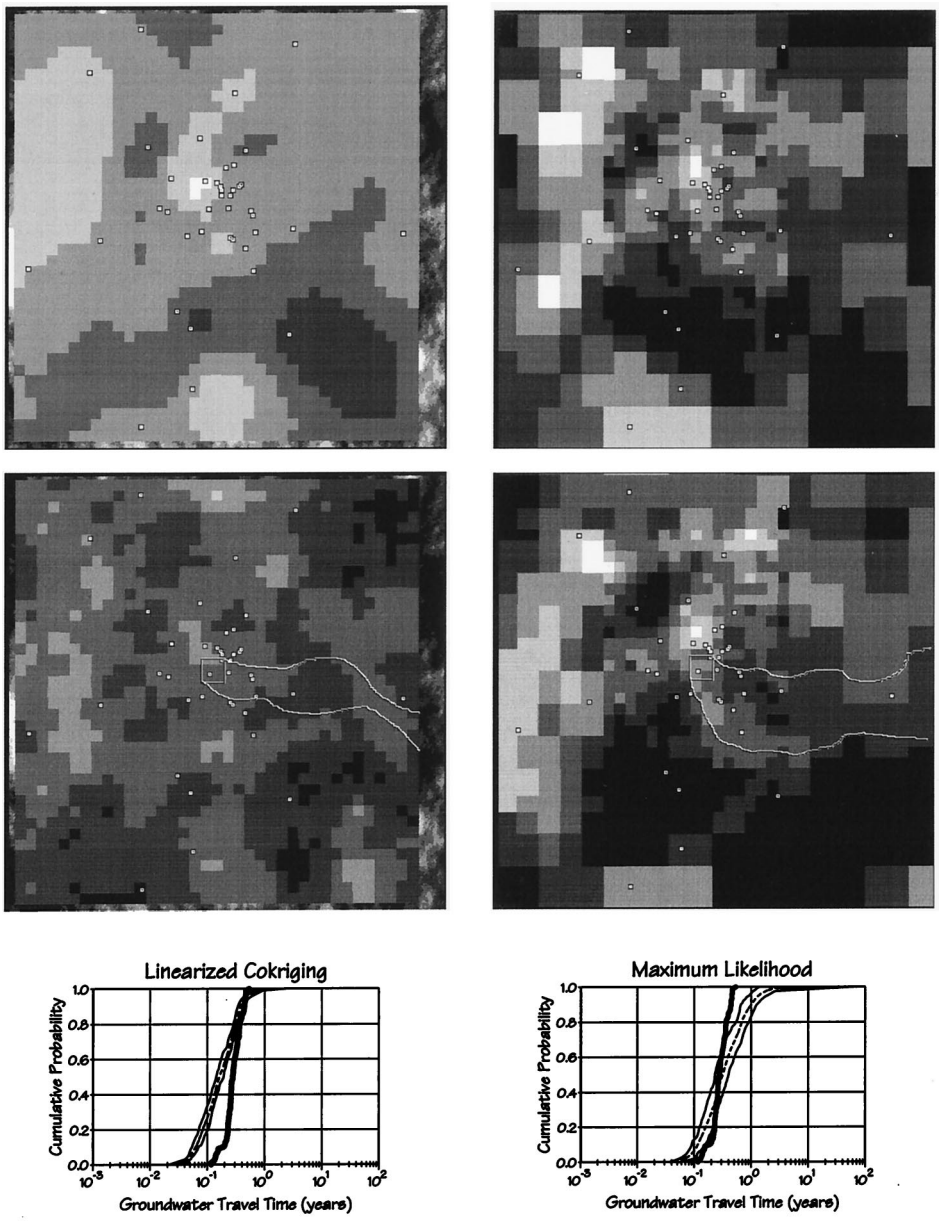


Figure 6b. Test problem 2 mean $\log_{10}(T)$ field, randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, dashed line is mean CDF, and other curves bound the inner 95% of all conditional CDFs.

would increase through reconnaissance. Although calibration was done by hand, this exercise clearly showed that the methodology employed in this case resulted in "overconfidence."

The decision on whether or not to declare the repository safe may be affected by the degree of overconfidence associated with an approach. If an approach produces results indicating the site will adequately contain the waste, but the approach is deemed to result in too much confidence, this could sway the decision maker to reject the application.

In our case the distances between bounding curves predicted by the methods are small for TPs 1 and 2 and wider for TPs 3 and 4. This is satisfactory, as it shows that the methods account for more uncertainty in the more complex TPs. If we look further, we see that method SS is overconfident for TP 1, but not so for the other TPs. Method LS produces almost systematically the largest range, and LC the smallest. The range for PP does not vary a great deal between TPs. The range for ML is, in general, the most appropriate over all the TPs; PP and SS come next.

We now briefly examine the T fields (Figures 5a–8c). For TPs 1 and 2, visually, the major high-T zones seem to be reasonably well captured by PP, SS, and ML, in that order, and a little less so by the others. For TP 4 it is evident that SS looks closer to the true field than the others.

It is interesting to know that for TP 3, the participant realized that the $\log_{10}(T)$ sample had a bimodal distribution (because of the different geology of the "features" in the aquifer) and decided that the multi-Gaussian assumption would not be appropriate for this distribution. He therefore used the indicator kriging approach, optional in his code, with two populations, to account for this bimodal distribution. One population had high-T values; since the transient head data indi-

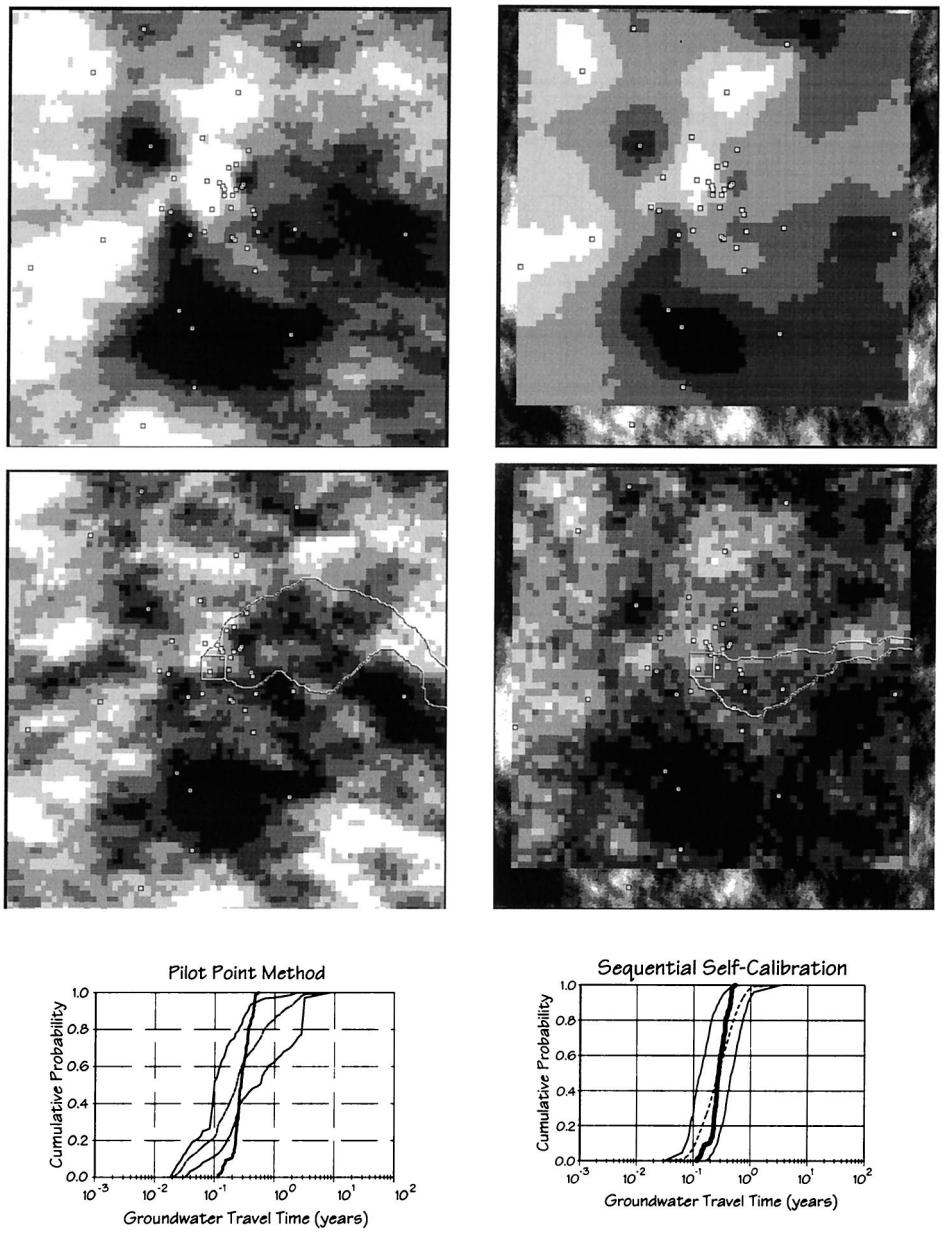


Figure 6c. Test problem 2 mean $\log_{10}(T)$ field, randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, dashed line is mean CDF, and other curves bound the inner 95% of all conditional CDFs.

cated that this population should be well connected in the $-45°$ direction, he decided to use an anisotropic indicator variogram for that population, consistent with this observation. These high-T "features" in the $-45°$ direction can easily be seen in Figure 7c in the given realization, and because of conditioning, some of these features are included in all realizations so that they appear in the mean T field. The fact that method SS did not rank first in the global ranking for may reflect that the "features" in the true T field were slightly more complex than accounted for by the anisotropic variogram.

The other approaches did not really identify the disconnected high-T channels. The darker zone at the lower left corner (low-T zone) was captured by SS, LC, FS, and, although not exactly in place, by PP. For TP 4 it is interesting to see that all approaches were more or less able to identify the connected high-T "channel" present across the domain.

One issue of interest in comparing inverse approaches is parameterization [e.g., *McLaughlin and Townley*, 1996]. The way each method parameterizes its T field is described in Appendix B. We tried in several ways to relate parameterization to the present results but did not find any real clues. One reason is that most methods parameterized the T field with a relatively similar number of unknowns (on the order of 50) and used geostatistics to interpolate the values; they furthermore constrained the unknowns in predefined ranges so that in the end, the specifics of the parametrization of each approach did not seem to make a large difference. We will return to this issue in the discussion, together with that of uniqueness.

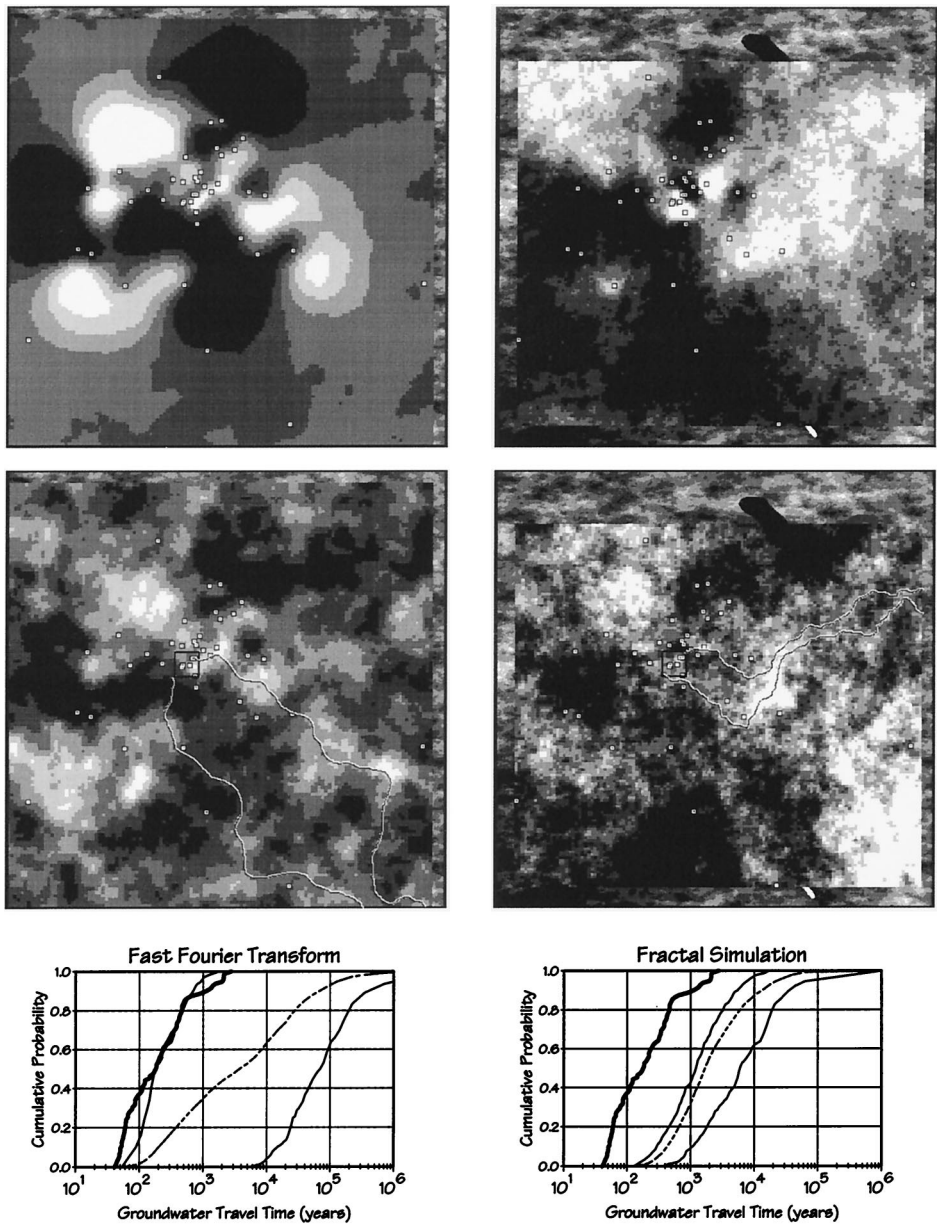


Figure 7a. Test problem 3 mean $\log_{10}(T)$ field, randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, dashed line is mean CDF, and other curves bound the inner 95% of all conditional CDFs.

Examination of the semivariograms (Figures 11–14) shows large differences among the approaches. Furthermore, it is clear that these differences are not systematic. To illustrate what can be learned from these figures, we will look for instance at the results of method SS. In TP 1, SS underestimates the variability: the sill of the variogram is approximately 2/3 that of the true field. As a result, the CDF of SS for TP 1 is "overconfident," as we have seen. In TP 2, SS has the variogram which is the closest to the true-field semivariogram, and, as a result, SS does very well on the CDF and its bounds. In TP 3, SS does the best job for short distances, even if it overestimates the sill. Since for this problem, the short-scale spatial variability dominates the T fields (because of the presence of the channels), SS again does very well. In TP 4, SS (as well as LC, ML, and PP) is very close to the true-field semivariogram and also produces accurate flow results. In TP 1, PP used a generalized covariance model with no sill (see insets of Figures 11–14). After the completion of TP 1 the PP approach was rerun using an exponential covariance model whose sill matches the true-field sill.

To summarize the initial findings so far: (1) There are significant differences between the methods and the way each approach is implemented (e.g., grid discretization and orientation). (2) The use of any of the inverse methods to condition the CDFs of GWTT on transmissivity and head data drastically reduces the uncertainty in these GWTTs, compared with the unconditional CDFs. (3) For "simple" (classical geostatistical) problems, all the approaches do a reasonably good job; the errors in GWTT are within half an order of magnitude, and in general, with only few exceptions, the inverse methods do not build "overconfidence." (4) For "complex" cases the nonlinear approaches do better in general. The LS method is an excep-

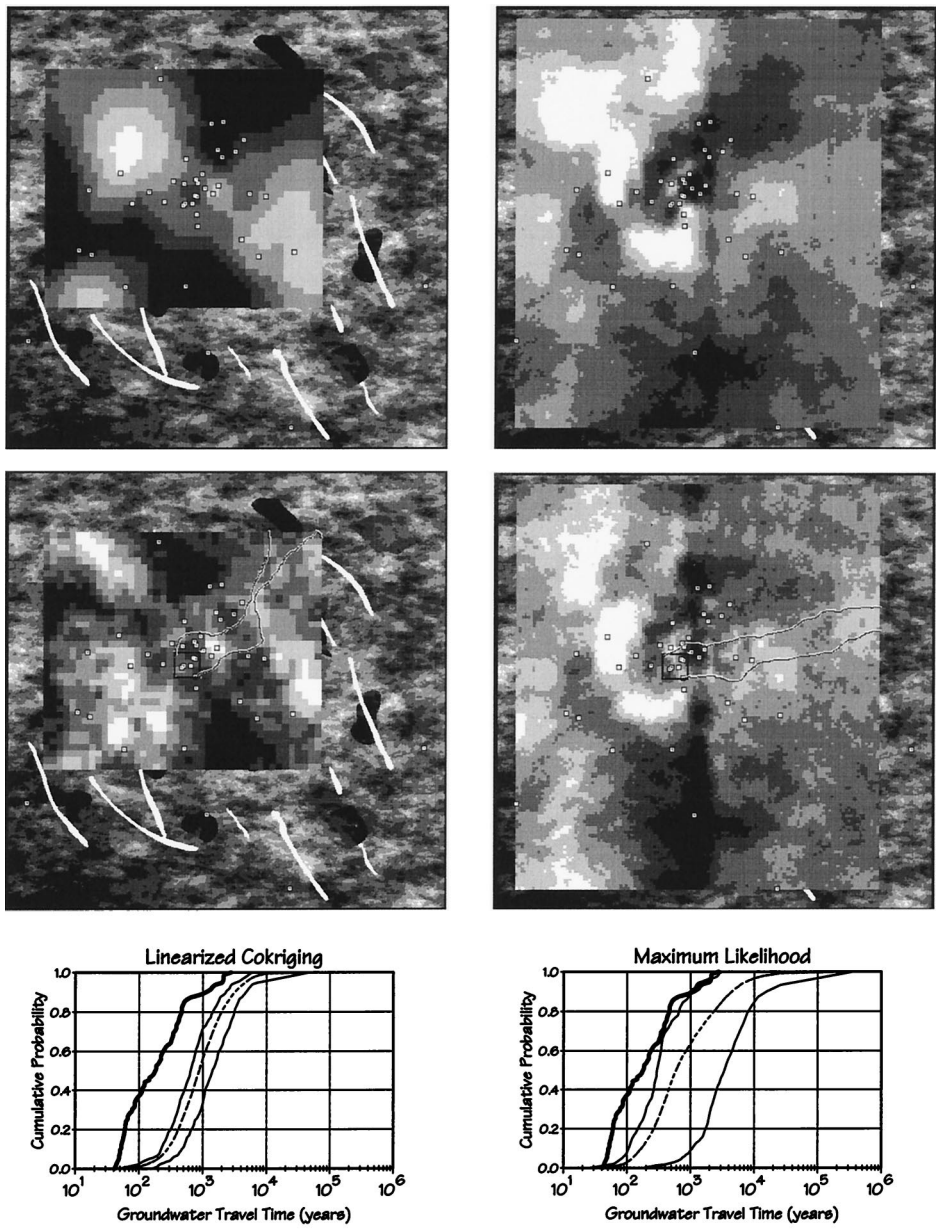


Figure 7b. Test problem 3 mean $\log_{10}(T)$ field, randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, dashed line is mean CDF, and other curves bound the inner 95% of all conditional CDFs.

tion in this respect, as it does better than the other linearized ones, for reasons that we will investigate later. There is, however, a tendency for all approaches to be overconfident and to overestimate the GWTT, meaning not erring on the side of greater safety. (5) The magnitude of the variance of $\ln(T)$, up to 11 in our case, does not seem to be a problem, even for the linearized approaches. However, nonstationarity (and departure from a true "geostatistical" distribution) is obviously more difficult to handle for the linearized approaches than for the nonlinear ones. (6) Unconnected channels are poorly identified by most inverse methods; an exception is, however, the SS method using the multiple-population approach. If the presence and average direction of such channels can be identified by external data (in this case, the transient head response to the aquifer tests), then this approach can be geared to generate such channels in the selected direction(s). An alternative, which was attempted by the participant of method FF, is to introduce such features "by hand." In both cases, if the features (or position) of these channels have been correctly identified, this will of course improve the results. (7) A good selection of the variogram of the true T field seems to improve the results of the inverse.

This first series of findings was purposely based on qualitative "subjective" judgments without any attempts to quantify the results. In the next section we build a number of objective evaluation measures and analyze their results statistically.

4. Quantitative Comparisons

The GXG decided that the comparison of the methods should be primarily based on quantifiable measures that can be directly related to the ability of the model to predict transport.

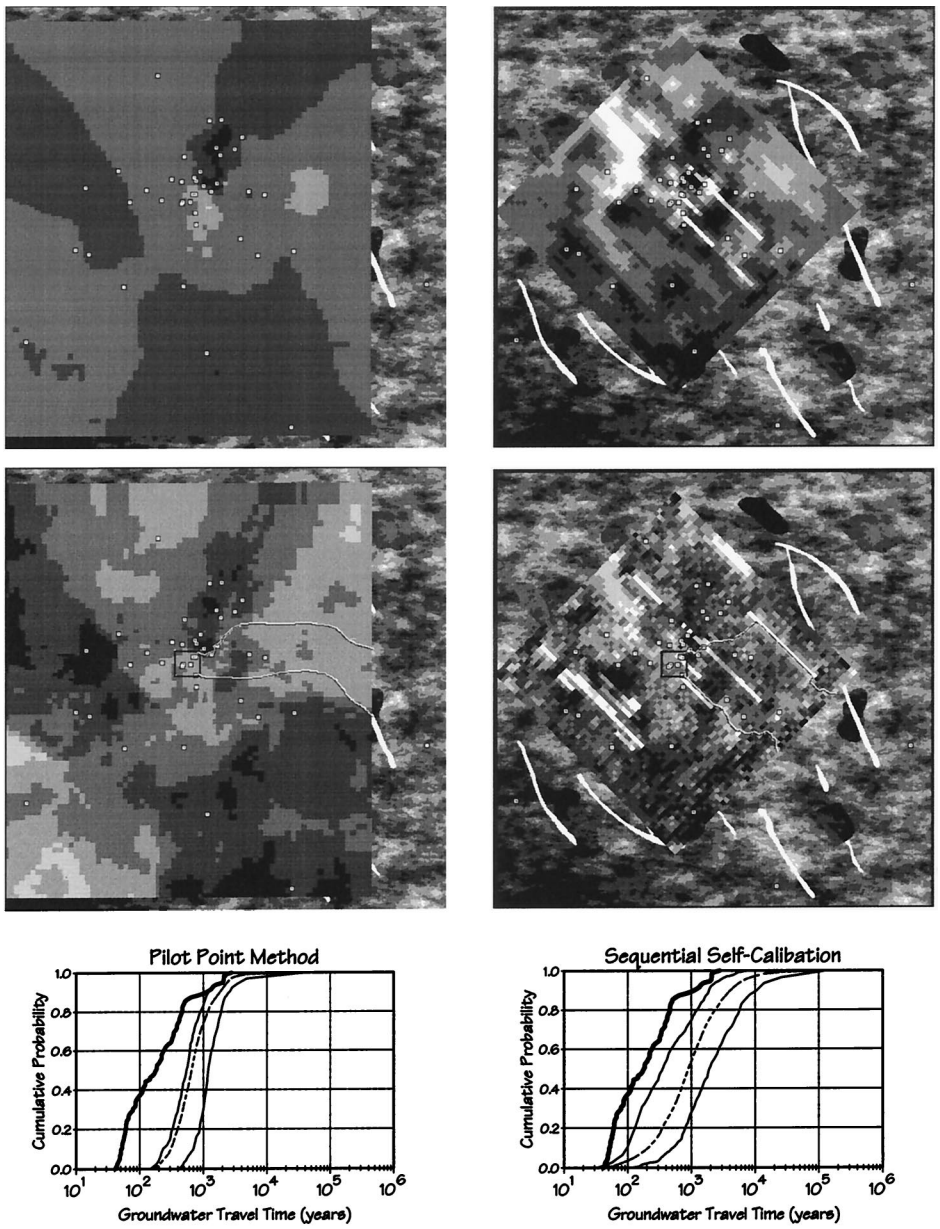


Figure 7c. Test problem 3 mean $\log_{10}(T)$ field, randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, dashed line is mean CDF, and other curves bound the inner 95% of all conditional CDFs.

The advective groundwater travel time of a conservative solute in the aquifer was selected as the most significant outcome of a calibrated model on which the evaluation of the methods should be based. This is a performance measure related to the ability of a waste disposal site to isolate waste yet is simpler than prediction of solute concentration. In addition, it was decided that GWTT alone was insufficient and that the groundwater flow paths should also be examined. It was reasoned that calculations resulting in accurate GWTT but very inaccurate groundwater flow paths would probably not be defendable, even though there is no regulatory requirement pertaining specifically to contaminant migration paths. To this end, particle-tracking calculations were performed; particles were released at a number of selected locations and the travel path and groundwater travel time to reach a radial distance of 5 km from the release point were calculated. In addition, the orientation of the flow path from the release point to the crossing point at the 5-km radial boundary was determined. For brevity, the name PATH will be used to refer to the particle pathline analyses.

Ten quantitative evaluation measures were tested and applied to the results of the test problems. These measures will be described under three headings: the fixed well approach, the random well approach, and the field variables measures.

4.1. The Fixed Well Approach

A selection of 10–30 particle release points was used for each test problem. The points were randomly located but more or less uniformly distributed over the total area of the studied field (as opposed to just within the waste panel area). The flow lines originating from these points were calculated by particle tracking (Figure 15). For each release point the distribution of

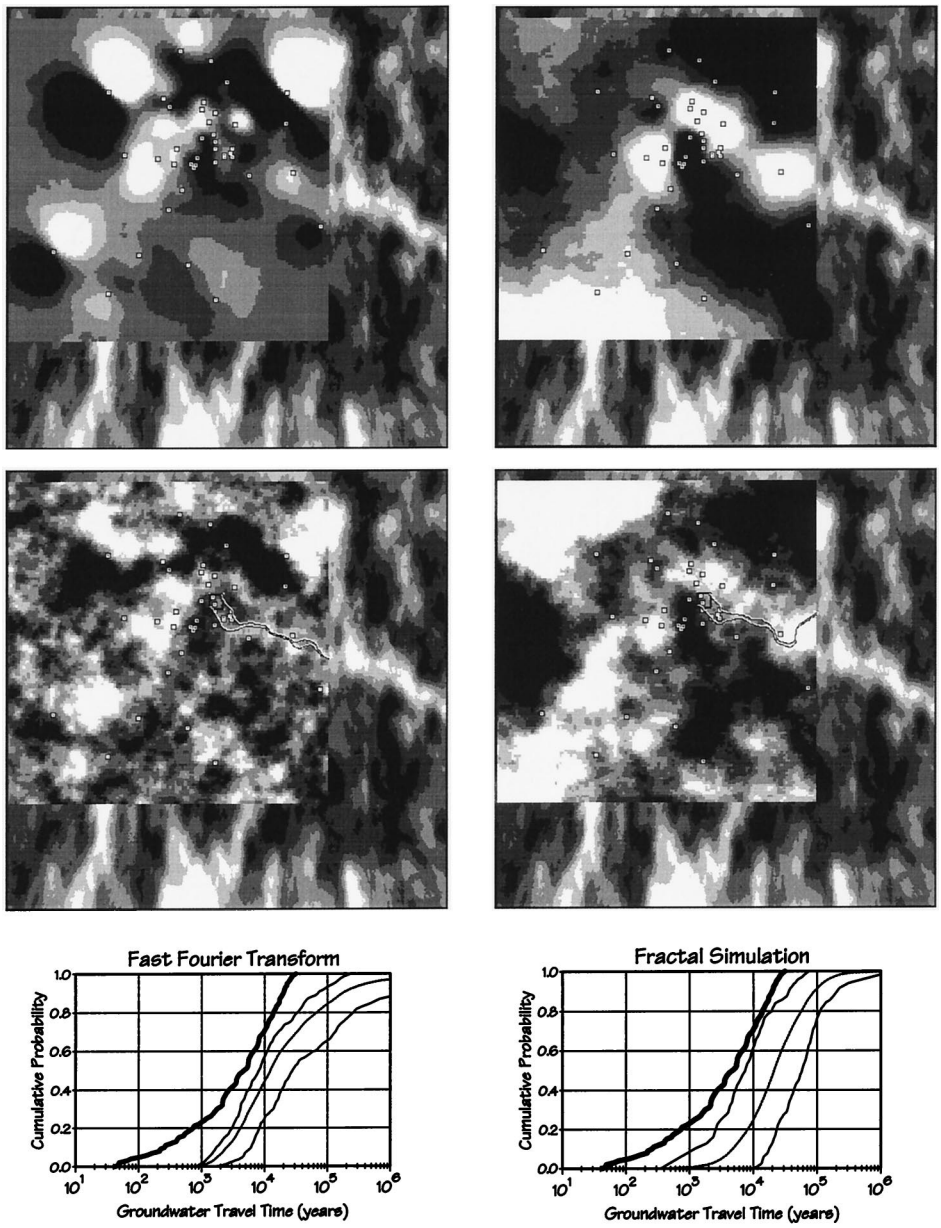


Figure 8a. Test problem 4 mean $\log_{10}(T)$ field, randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, dashed line is mean CDF, and other curves bound the inner 95% of all conditional CDFs.

GWTTs estimated from the ensemble of simulated fields is compared with the GWTT value from the true field. The analysis thus involves comparing a distribution of GWTTs to a single value (the true-field GWTT), contrary to the random well approach (described below), where we compare two GWTT distributions. The path lines predicted by each approach are also compared with those of the true fields (path line calculations are denoted "PATH"). Combining all four test problems, 88 particle paths were analyzed for each approach for the fixed well release points case. Given that there are two CDFs for each release point (one for GWTT, one for PATH) and seven approaches, this results in more than 1200 CDFs. Consequently, only a few samples are shown here (Figure 16).

The aim of this analysis is also to evaluate the inverse approaches on their ability to predict advective transport, but this time the locations of the release points are distributed over the whole domain. Because the analysis involves several independent measures and numerous release points, it will be possible to conduct statistical tests to assess differences in performance. Five evaluation measures were defined within the framework of the fixed well approach, the details of which are given in Appendix C.

Evaluation measure (EM) 1: The GWTT error (denoted "Error" in the tables) compares the median of the simulated GWTTs with the true GWTT.

EM 2: The GWTT "degree of caution" (denoted "Caut") measures the propensity of an approach (if any) to underestimate rather than overestimate the GWTT. This is a PA-specific concept where it is considered better to err on the side of greater safety and protection of public health which equates to predicting faster travel times. In the parlance of PA termi-

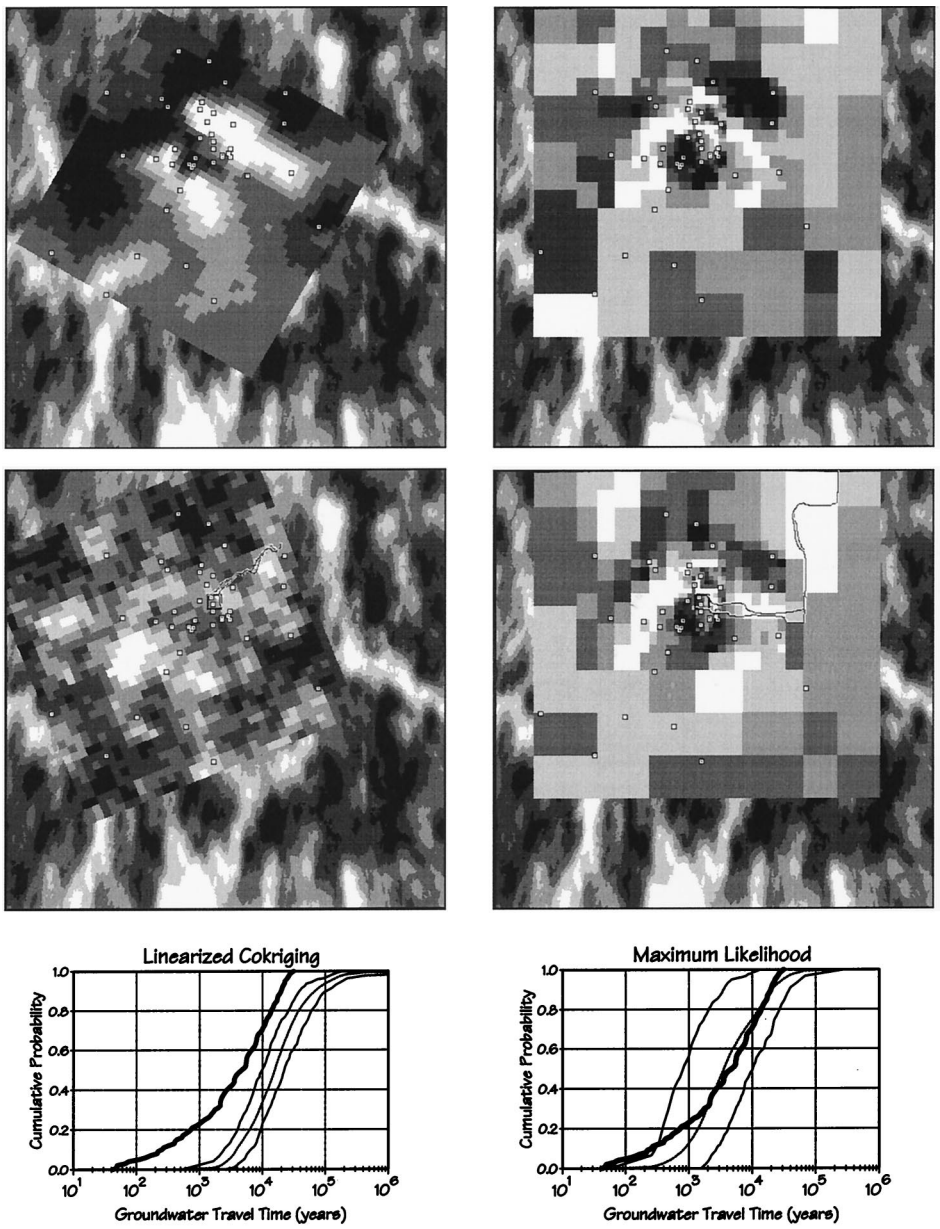


Figure 8b. Test problem 4 mean $\log_{10}(T)$ field, randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, dashed line is mean CDF, and other curves bound the inner 95% of all conditional CDFs.

nology, such predictions are referred to as "conservative" predictions.

EM 3a: The GWTT spread (denoted "Sprd") measures the width of the GWTT CDF produced by a method, i.e., the uncertainty that an approach associates with its GWTT predictions.

EM 3b: The GWTT robustness (denoted "Boot") measures the number of times the true GWTT falls within the inner 95% of the simulated GWTT distribution (the bootstrap test).

The Sprd and Boot measures must be considered simultaneously. For instance, an approach which predicts a small Sprd (a small uncertainty) but which fails the bootstrap test clearly underestimates the uncertainty. Similarly, an approach resulting in a large Sprd and a good Boot measure may be overpredicting the uncertainty. The goal is to have the smallest Sprd with a good Boot. Together the Sprd and Boot measures reflect whether an approach is "self-consistent," that is, whether it over or underpredicts the uncertainty. For ranking purposes a single index grouping the two, called the normalized self-consistency measure (NSC), was also constructed (see Appendix C).

EM 4: The PATH error quantifies the absolute deviation (in degrees) between the median path direction angle and the direction of the true path. The orientation of the path is defined by the angle from the release point to the point where the path crosses a circle of radius 5 km (centered at the release point).

EM 5a: The PATH spread measure quantifies the spread in the distribution of path angles.

EM 5b: The PATH bootstrap (robustness) measure, as for GWTT, measures the robustness of the path line calculations.

It should be noted that all these measures are computed for

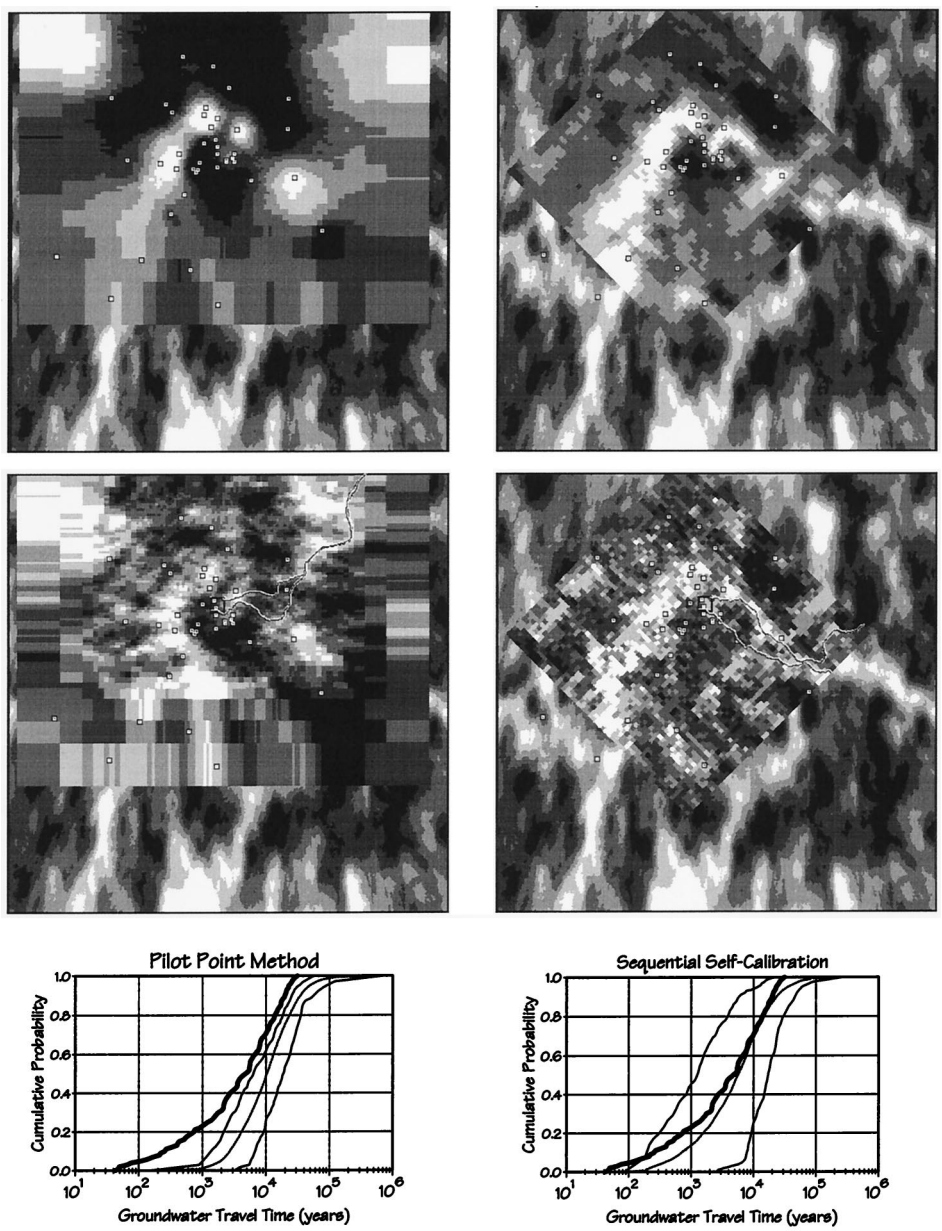


Figure 8c. Test problem 4 mean $\log_{10}(T)$ field, randomly selected single realization overlain with bounding pathlines for that realization, and the GWTT CDF curves from all realizations. Thick line is true CDF, dashed line is mean CDF, and other curves bound the inner 95% of all conditional CDFs.

each release point and are averaged over all release points, where the averaging is made with nonuniform weights. Additionally, each release point was given a weight reflecting how close this point was from the true observation data. It was reasoned that those points surrounded by measurements should be better predicted by the methods than those far away from any measurements so they are assigned relatively higher weights. The way in which these weights were derived is explained in Appendix C.

4.2. The Random Well Approach

This approach involves a comparison of the estimated GWTT CDF with the true GWTT CDF. The CDFs to be compared are the ones generated from the 100 release points within the waste panel area shown in Figures 5a–8c. Three evaluation measures have been defined. Figure 17 and Appendix C clarify how these measures are constructed.

EM 6: The GWTT error which is a measure of the disparity between the median GWTT CDF and the true CDF.

EM 7a: The GWTT spread measures the area between the 95% bounding envelopes of all waste panel GWTT CDFs.

EM 7b: The GWTT robustness measure quantifies what proportion of the true CDF is contained within the 95% bounding envelope.

4.3. The Field Variables

The final three evaluation measures (detailed in Appendix C) compare the simulated T fields and head fields with the true fields, and the semivariograms of the $\log_{10}(T)$ fields.

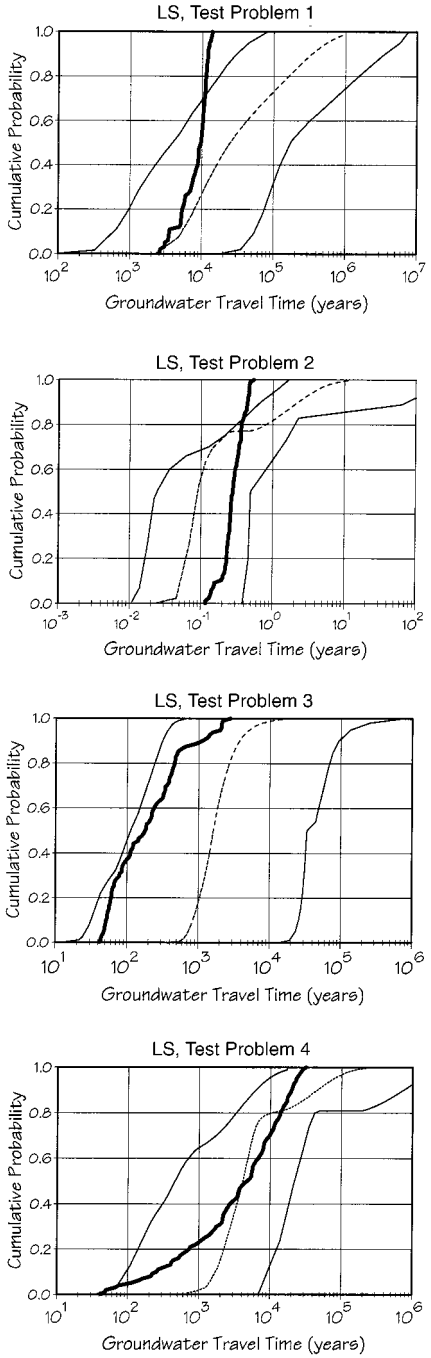


Figure 9. Waste panel GWTT CDFs for the linearized semi-analytical (LS) method.

EM 8: The $\log_{10}(T)$ error measures the difference between the ensemble of T fields and the true T field.

EM 9: The head error similarly measures the difference between the ensemble of head fields and the true-field head solution.

EM 10: The semivariogram error measures the difference between the average of the semivariograms of each simulated $\log_{10}(T)$ fields and that of the true one (Table 3).

4.4. Analysis of the Results

The 10 evaluation measures described in the previous section were computed for each of the approaches in each test problem and the "raw" (untransformed) evaluation scores are listed in Table 4. For several of the measures, the scores vary only within the range [0, 1]. For the other measures, however, the scores are only bounded by zero and can range beyond one. Because these "raw" measures are not all computed in the same units, they cannot be meaningfully averaged and are not directly comparable. Transformation of the evaluation measure scores to consistent units was performed in two ways, by converting them to standardized variables and via rank transformation. Statistical analyses were performed on both the standardized and the rank-transformed variables. However, analyses of the rank-transformed variables were considered more powerful from a statistical viewpoint, and therefore only the rank-transformed results are presented here. For method

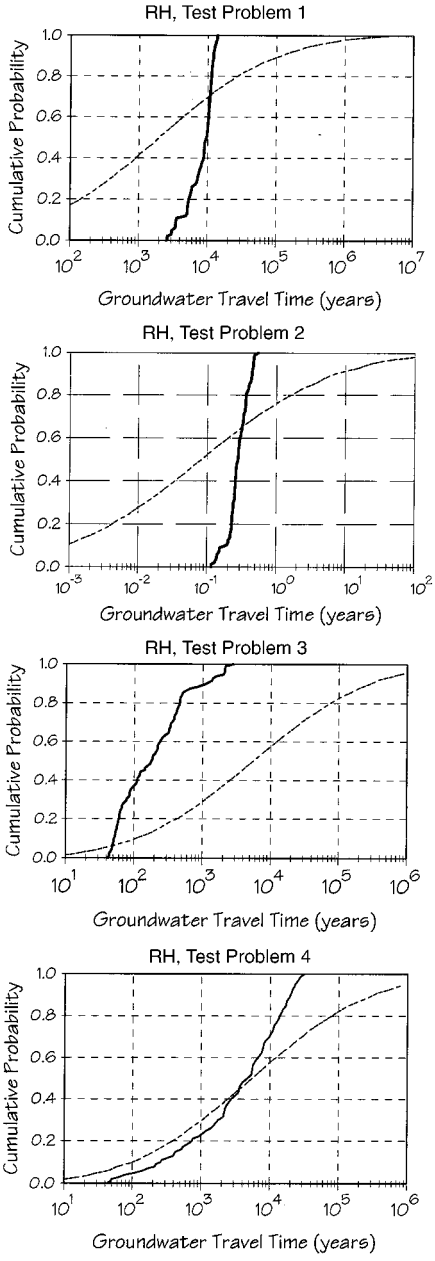


Figure 10. Unconditional waste panel GWTT CDFs from the random homogeneous (RH) case. Note that the ranges of GWTT in these plots span several orders of magnitude more than those from the inverse methods (Figures 5a–8c).

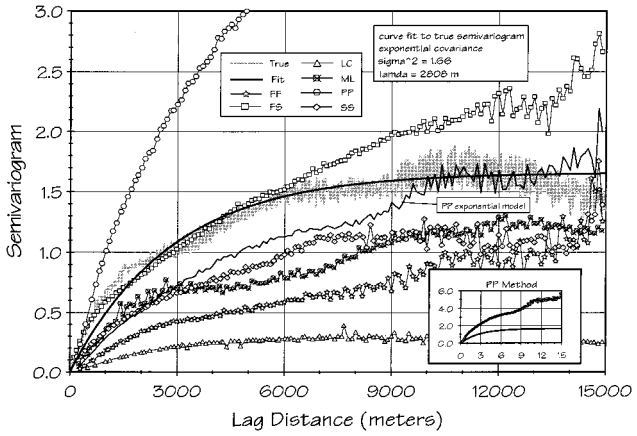


Figure 11. Average semivariograms for test problem 1.

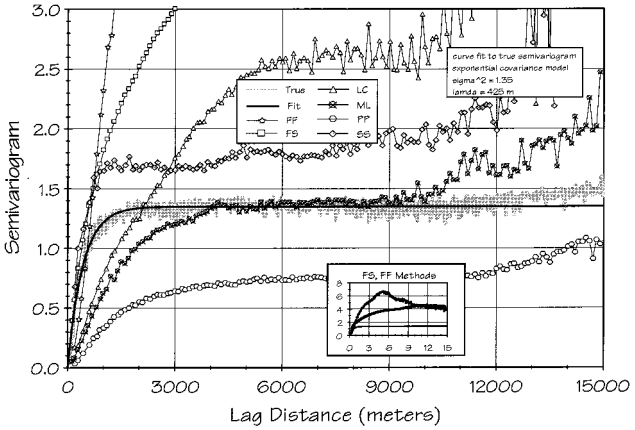


Figure 13. Average semivariograms for test problem 3.

LS, which has not produced transmissivity and head fields, the comparison was done excluding the head and $\log_{10}(T)$ field error measures.

4.4.1. Evaluation measures overview and ranking. The rank-transformed scores for the 10 evaluation measures for each test problem and method are listed in Table 5. The average rank across all measures in a test problem is shown in the far right column and the average score for each measure across all four test problems is shown down the columns.

It was of interest to determine if the two sets of GWTT analyses (the random well and the fixed well approaches) would lead to different conclusions. To compare these measures, we averaged the five evaluation measures for the fixed well and the two evaluation measures for the random well approaches from Table 5, for each approach over all test problems.

Table 6 shows that on average, similar rankings are obtained for both analysis approaches except for LC, which does slightly better for the random well case. The comparison of distributions with the single-valued "truth" used in the fixed well approach, although involving more evaluation measures, is thus a relatively robust indicator of the performance of an approach and compares favorably with the better founded comparison of distributions. This result also shows that the relative performance of methods does not change depending on whether the analysis is conducted in the vicinity of the highest data density (random well case) or throughout the region with more sparse data (fixed well case). This may be expected because of the kriging variance-based weighting used in the fixed well case.

Thus far we have not paid much attention to the correct orientation of the flow lines predicted by each method. On average, it appears that method LS performs somewhat better than any other method to predict the correct path. The spread of the PATH CDFs for this approach is also, in general, wider than for the others, as was the case for GWTT. It is also worth mentioning that LS scores best for Caut, as a result of a conscious decision of the participant to "fine tune" his approach to meet this criterion.

The head and $\log_{10}(T)$ errors are highly correlated. The head and $\log_{10}(T)$ errors listed in Table 4 are plotted in Figure 18 to show this correlation. The closer the T fields are to the true T field, the smaller, in general (save for TP 3), are the head errors. Linear regression performed on the results from TPs 1, 2, and 4 (dashed line in Figure 18) has a coefficient of determination of 0.70. In TP 3 the average magnitude of the head error (for all approaches) is approximately half an order of magnitude larger than for the other TPs (see Table 4). The exception is method PP, which does much better on head than any other approach. Note, however, that its performance in the $\log_{10}(T)$ error in is not that much better than SS or ML. One likely reason is that the flexibility of this approach in optimally choosing the parametrization (optimal selection of the location

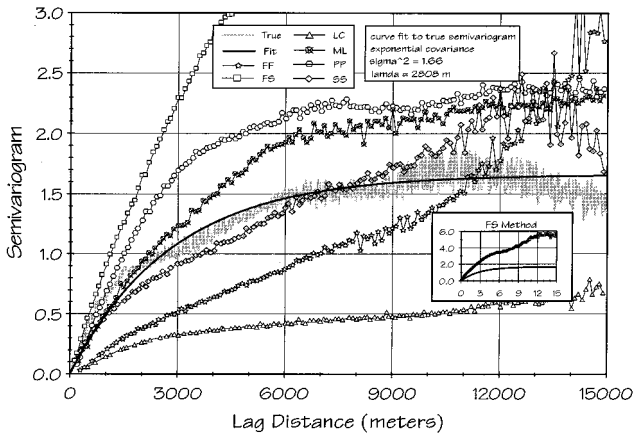


Figure 12. Average semivariograms for test problem 2.

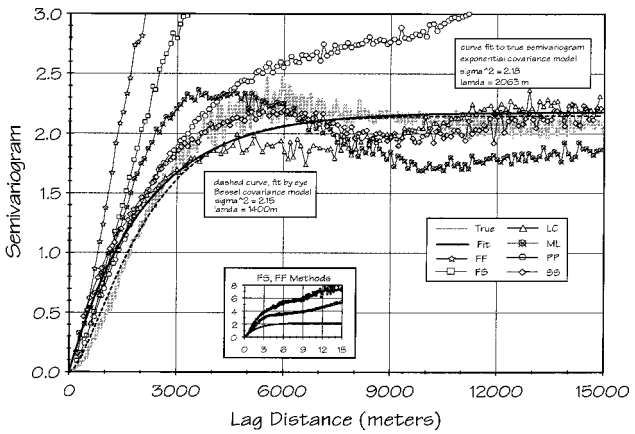
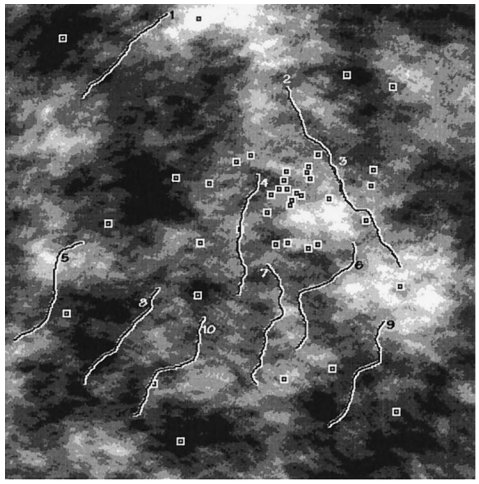
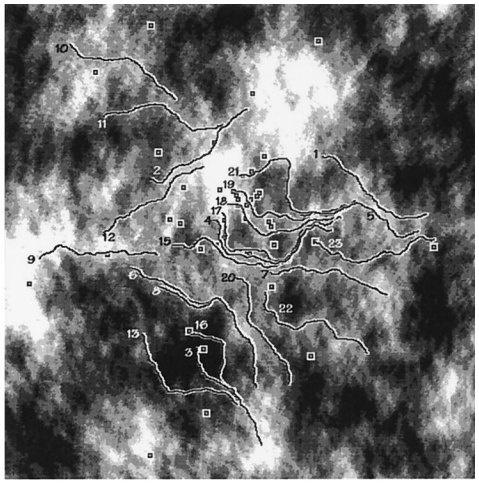


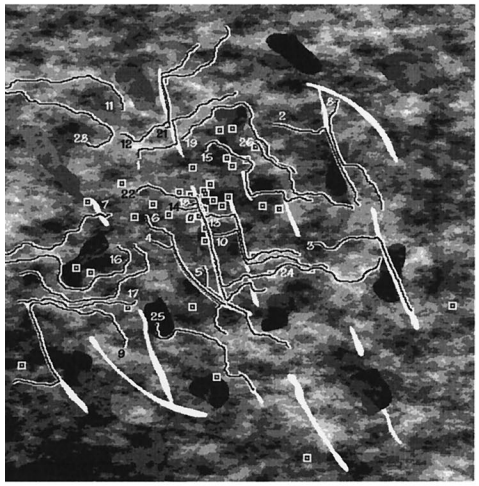
Figure 14. Average semivariograms for test problem 4.



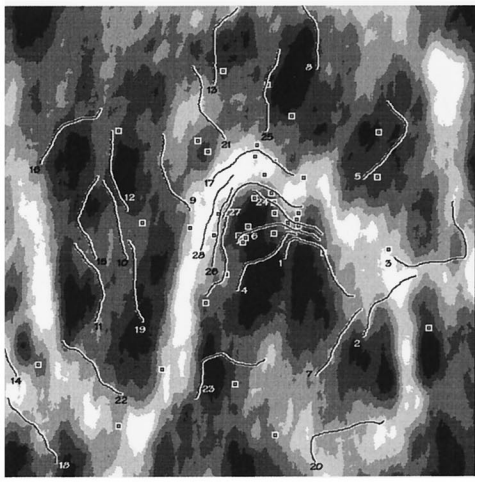
Test Problem No 1 (Fixed Well case)



Test Problem No 2 (Fixed Well case)



Test Problem No 3 (Fixed Well case)



Test Problem No 4 (Fixed Well case)

Figure 15. Fixed release points and pathlines on the true log (T) fields. Pathlines extend a radial distance of 5 km.

of the PP) makes it possible to fit the head better than other approaches. This may be especially true for complex flow systems, but perhaps at the expense of producing a T field which poorly corresponds to the true T field (see Figures 7a–7c, where it is clear that the PP T fields have a more continuous design pattern (zones of high and low Ts) than the ML and SS methods). This exemplifies the necessity to prescribe, in one form or another, a "plausibility" criterion on the T field in an inverse solution. As shown by *Carrera and Neuman* [1986a, b], the minimization of the head differences alone is insufficient.

4.4.2. Statistical analysis of the evaluation measures. The results presented in Table 5 can be used in statistical analyses to indicate if the performances of the seven methods (as quantified through the evaluation measures) are significantly different. In these analyses it is assumed that each evaluation measure (appropriately transformed to ranks or standardized) is an independent measure of the performance of the approach. A two-factor analysis of variance (ANOVA) was used to analyze the performance measure information. This approach considered the two factors, test problem and method, and their interaction to be potentially significant sources of variation in method performance. The validity of statistical tests obtained through the ANOVA on ranked data depends on the assumption that the performance values of each approach are independent measures of the same quantity which have a constant variance. Thus we are assuming that each evaluation measure is an independent measure of the performance of the method where each measure quantifies performance in a different way. Because the performance values are clearly not independent and the assumption of constant variance may be questionable, the results from the ANOVA tests were used only as an indicator that substantial differences may exist and to suggest a general ordering of the methods rather than to declare the results "statistically significant."

The null hypothesis for the two-way ANOVA is that there is no difference among the approaches or the test problems or the combination of the two. The computed p value is the probability that an F statistic greater than the one observed

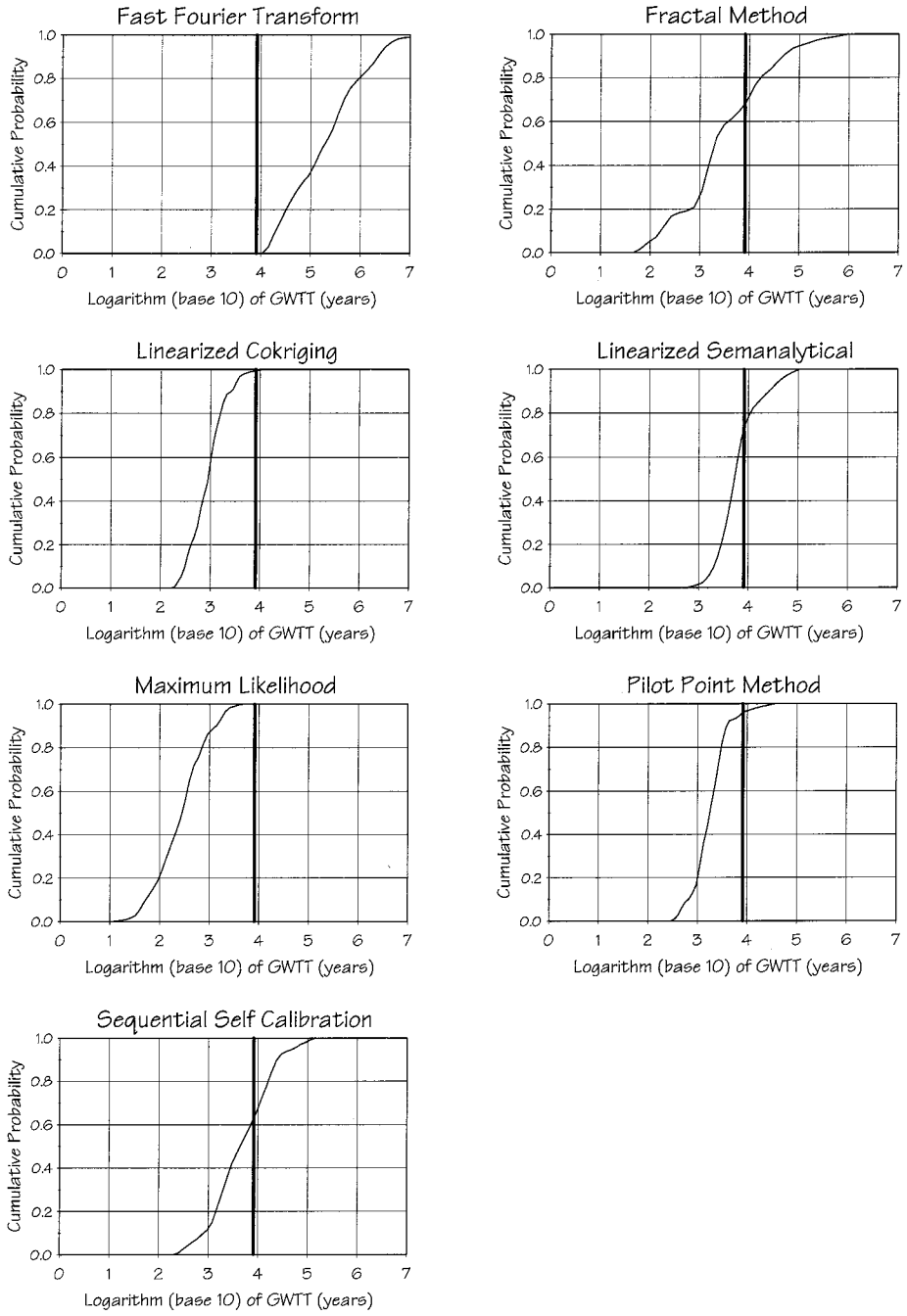


Figure 16. Examples of some GWTT CDFs for the fixed release points case (all plots are release point 16, TP 3). Vertical line is the true GWTT.

would be obtained if the null hypothesis were true. The p value for the method/test problem interaction was 0.0079, indicating a strong interaction between the approach and test problem, that is, that the performance of an approach tends to differ depending on the test problem. The nature of this interaction is illustrated in Figure 19, which is a plot of the average evaluation scores across the 10 measures in each test problem, for each approach. From this figure we can see that while the performance of some of the methods is relatively consistent across all test problems (such as SS and FF), the performance of other methods (such as LS and PP) tends to depend on the particular test problem considered. In fact, the performance of the LS method, and perhaps also that of the PP method, tends to improve across test problems. For method LS, the most likely explanation is that the participant changed the method of integration and could thus use finer time steps. For method PP the explanation could be that the participants took more care in the application of the method (e.g., selection of the semivariogram), or that the approach is more suited to more complex test problems, or both.

Apart from the dependence of method performance on the test problem, Figure 19 also indicates that the performance of

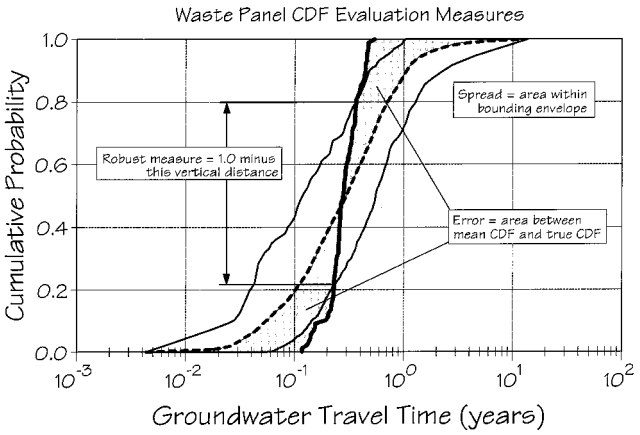


Figure 17. GWTT CDF evaluation measures used in the random well case. Thin solid lines are the 0.025th and 0.975th percentile CDFs, thick dashed line is the median CDF, and thick solid line is the "true" CDF.

the SS method is, three times out of four, superior to that of the other approaches. For certain test problem/method comparisons (such as the comparison of SS and FF on TP 4), the difference in performance is substantial; for other test problem/method comparisons (such as the SS and PP methods on TP 3), the difference is extremely small. Given this result and the test problem method interaction, it does not appear that there is sufficient evidence to conclude that the performance of one approach was consistently significantly superior to that of all other approaches.

These observations were reinforced by conducting a one-way ANOVA on the average of the rank-transformed evaluation measures scores taken across all 10 measures for each test problem. This gives four measures of performance for each approach, one for each test problem. Because the test problems were constructed, sampled, and analyzed independently of each other, the overall average performance measure scores should also be independent. The one-way ANOVA was used to test the hypothesis that the average performance of all the

Table 3. Parameters of the Exponential Semivariogram Model Fit to the Average Semivariogram Across All Realizations

Method	σ^2	λ, m	R^2	Cut, 1000 m	RMSE	J_{σ^2}	J_λ	J_γ	Rank
				Test Problem 1					
True	1.66	2808	0.9949	12	0.00	···	···	···	···
FF	0.79	4074	0.9976	9	0.72	0.34	0.31	0.32	3
FS	2.41	5503	0.9978	12	0.26	0.31	0.49	0.45	6
LC	0.29	2142	0.9947	15	1.13	0.45	0.19	0.26	1
LS	2.99	6900	···	···	···	0.44	0.59	0.55	7
ML	1.26	4730	0.9913	15	0.47	0.19	0.41	0.35	5
PP	1.52	3767	0.9998	9	1.52	0.60	0.25	0.34	4
SS	1.28	3995	0.9975	10	0.41	0.19	0.30	0.27	2
				Test Problem 2					
True	1.66	2808	0.9949	12	0.00	···	···	···	···
FF	2.37	12576	0.9989	10	0.49	0.30	0.78	0.66	7
FS	4.59	4395	0.9994	7	1.40	0.64	0.36	0.43	6
LC	0.54	3617	0.9971	12	0.94	0.40	0.22	0.27	3
LS	2.99	4000	···	···	···	0.44	0.30	0.34	5
ML	2.32	3776	0.9993	15	0.49	0.28	0.26	0.26	2
PP	2.34	2525	0.9993	15	0.66	0.29	0.09	0.14	1
SS	1.92	4687	0.9978	9	0.13	0.14	0.40	0.33	4
				Test Problem 3					
True	1.35	425	0.9986	12	0.00	···	···	···	···
FF	5.09	1134	0.9726	15	3.56	0.73	0.63	0.66	4
FS	4.29	2429	0.9978	15	2.43	0.69	0.83	0.79	7
LC	2.91	3099	0.9983	9	0.93	0.54	0.86	0.78	6
LS	0.75	2700	···	···	···	0.31	0.84	0.71	5
ML	1.41	1605	0.9992	9	0.25	0.04	0.74	0.56	2
PP	0.74	1562	0.9990	10	0.67	0.31	0.73	0.62	3
SS	1.76	384	0.9981	9	0.43	0.23	0.09	0.12	1
				Test Problem 4					
True	2.18	2063	0.9927	15	0.00	···	···	···	···
FF	6.28	3320	0.9986	9	2.56	0.65	0.38	0.45	7
FS	3.98	2715	0.9968	9	1.15	0.45	0.24	0.29	5
LC	2.12	2128	0.9967	15	0.23	0.03	0.03	0.03	1
LS	2.24	1600	···	···	···	0.03	0.18	0.14	2
ML	1.97	1046	0.9842	15	0.32	0.09	0.33	0.27	4
PP	2.99	3044	0.9993	9	0.35	0.27	0.32	0.31	6
SS	2.09	1497	0.9979	15	0.18	0.04	0.22	0.18	3

Parameters for LS method provided by participant. RMSE is calculated between the average semivariogram and the true-field semivariogram using equally spaced observation points and equal weights. "Cut" is the limiting distance used for the curve fitting and the RMSE calculations. Beyond cut (which was chosen subjectively) the semivariogram estimates become erratic and are likely to be very unreliable. J_λ and J_{σ^2} are the evaluation measure scores for the correlation length and sill, respectively. Rank is based on the overall correlation structure score, $J_\gamma = (3 \cdot J_\lambda + J_{\sigma^2})/4$.

Table 4. "Raw" Evaluation Measure Scores, Test Problem 1–4

		Fixed Release Points							Random Well			Field Variable		
Method	Npts	GWTT Error	GWTT Cnsv	GWTT Boot	GWTT Sprd	PATH Error	PATH Boot	PATH Sprd	GWTT Error	GWTT Sprd	GWTT Boot	Log (T) Error	Head Error	Semi-variogram
							Test Problem 1							
FF	10	0.24	0.25	0.26	1.19	0.28	0.16	38.8	5.73	0.38	0.97	0.543	2.41	0.32
FS	10	0.20	0.12	0.05	2.11	0.42	0.05	61.0	3.15	0.34	0.73	0.507	3.94	0.45
LC	9	0.23	0.46	0.06	1.20	0.32	0.06	37.7	1.20	0.16	0.82	0.334	2.28	0.26
LS	4	0.17	0.27	0.21	1.69	0.20	0.05	179.0	8.19	1.87	0.29	N/A	N/A	0.55
ML	10	0.17	0.25	0.05	1.35	0.23	0.05	48.9	4.49	0.44	0.74	0.739	4.19	0.35
PP	10	0.39	0.45	0.05	3.38	0.50	0.05	65.3	5.94	0.53	0.64	0.853	2.56	0.34
SS	8	0.15	0.40	0.05	0.90	0.17	0.21	29.0	3.93	0.11	0.99	0.483	2.51	0.27
							Test Problem 2							
FF	20	0.25	0.12	0.19	1.42	0.29	0.16	48.5	6.34	0.61	0.89	0.691	2.59	0.66
FS	22	0.27	0.25	0.05	2.71	0.51	0.05	106.2	3.99	0.71	0.42	1.169	5.99	0.43
LC	22	0.25	0.11	0.19	1.35	0.24	0.43	39.0	3.02	0.19	0.93	0.559	2.43	0.27
LS	11	0.45	0.09	0.14	1.62	0.29	0.05	49.2	6.37	1.41	0.29	N/A	N/A	0.34
ML	23	0.15	0.18	0.01	1.81	0.37	0.05	112.6	3.14	0.26	0.59	0.856	3.58	0.26
PP	23	0.16	0.14	0.01	2.15	0.35	0.04	59.3	7.07	0.53	0.57	0.708	3.01	0.14
SS	22	0.13	0.32	0.14	1.03	0.21	0.28	32.2	1.80	0.56	0.00	0.584	2.35	0.33
							Test Problem 3							
FF	27	0.50	0.57	0.42	2.19	0.80	0.53	44.0	14.03	2.54	0.58	1.202	11.5	0.66
FS	27	0.33	0.63	0.06	3.27	0.86	0.34	106.7	10.66	0.74	1.00	1.366	16.4	0.79
LC	19	0.37	0.45	0.39	2.14	1.36	0.94	44.6	7.01	0.36	1.00	1.321	19.3	0.78
LS	22	0.25	0.47	0.52	1.65	0.45	0.47	43.5	11.39	2.63	0.00	N/A	N/A	0.71
ML	27	0.35	0.15	0.14	2.22	0.75	0.30	84.2	6.43	1.11	0.91	0.907	12.8	0.56
PP	27	0.33	0.58	0.42	1.45	0.55	0.22	66.8	6.13	0.38	0.96	0.807	6.7	0.62
SS	21	0.18	0.34	0.00	1.91	0.79	0.45	58.7	7.19	0.75	0.99	0.859	15.7	0.12
							Test Problem 4							
FF	27	0.39	0.52	0.22	3.15	0.38	0.26	7E76.8	9.04	0.79	1.00	1.731	5.57	0.45
FS	27	0.31	0.48	0.06	3.01	0.37	0.05	110.9	8.84	0.94	1.00	1.768	6.33	0.29
LC	24	0.44	0.45	0.39	2.41	0.83	0.17	131.4	8.37	0.42	1.00	1.344	4.25	0.03
LS	23	0.32	0.29	0.14	2.27	0.24	0.01	138.0	4.72	1.78	0.01	N/A	N/A	0.14
ML	26	0.36	0.58	0.39	1.77	0.26	0.07	46.3	2.87	1.09	0.12	1.032	4.20	0.27
PP	27	0.26	0.33	0.03	2.66	0.31	0.05	79.1	6.03	0.50	1.00	1.194	6.09	0.31
SS	22	0.28	0.36	0.28	2.01	0.29	0.00	63.5	2.66	1.20	0.07	0.914	3.19	0.18

All measures were constructed such that the target value is zero. Npts is the number of release points used for the fixed release points GWTT and PATH analyses.

approaches is the same. The F test from the ANOVA indicated significant differences (p value = 0.0055); therefore Fisher's least significant difference (LSD) pairwise comparison procedure [*Steel and Torrie*, 1980] was used to determine which approaches differed. The results show a great deal of overlap in the performance of the approaches. This is expected because of averaging across test problems when significant interaction is present. Although no single approach performed significantly better than all other methods in all cases, we can roughly delineate three performance groups. The SS approach had the best overall performance, although its overall average performance may not be substantially better than that of a middle group comprising the ML, LS, and PP methods. The performance of the SS method may be significantly superior to all other approaches for a particular test problem, as indicated in Figures 19 and 20. Method SS performs significantly better than the third group containing the FS, LC, and FF methods. While there is a strong similarity between the performance of the FF and the LC approaches, the results of the ANOVA do not indicate a clear differentiation in performance between the linearized and nonlinear methods. This is in part because the results for the LS method were more similar to the nonlinear approaches and method LS did not perform as poorly as the other linearized approaches.

In addition to significance testing via ANOVA procedures, cluster analyses were performed using the average, across the four test problems, of each of the evaluation measure scores (except for the head and $\log_{10}$ (T) error measures because method LS did not produce head and log (T) fields). Cluster analysis is a statistical procedure for partitioning multivariate data into groups based on some measure of similarity. The correlation between the performance vectors was used as the measure of similarity, so that two approaches are deemed similar if the correlation between their performance vectors is high. Clustering began with each method as a separate cluster and was allowed to continue until all methods were combined into one cluster. Amalgamation of the approaches into clusters was performed using unweighted pair-group averaging [*Johnson and Wichern*, 1982]. The results are shown in Figure 21, where the clustering is stopped when there are two clusters remaining. These results appear to distinguish the behavior of the linear and nonlinear approaches, as the two remaining clusters fall into those categories.

5. Discussion of Results

Before drawing conclusions from this comparison, the reasons for some of the results observed should be clarified. In the following sections we discuss issues that are generic to inverse modeling, issues related to the assumptions used in the mod-

Table 5. Rank-Transformed Evaluation Measure Scores

		Fixed Release Points					Random Well		Field Variable			
Method	TP	GWTT Error	GWTT Cnsv	GWTT NSC	PATH Error	PATH NSC	GWTT Error	GWTT NSC	Log (T) Error	Head Error	Variogram	Average Score
FF	1	6.0	2.5	6.5	4.0	6.0	5.0	6.5	4.0	2.0	3.0	4.55
FF	2	4.5	3.0	7.0	3.5	5.0	5.0	7.0	3.0	3.0	7.0	4.80
FF	3	7.0	5.0	6.0	5.0	6.0	7.0	2.0	4.0	2.0	4.0	4.80
FF	4	6.0	6.0	5.0	6.0	6.5	7.0	6.0	5.0	4.0	7.0	5.85
FF Average		5.88	4.13	6.13	4.63	5.88	6.00	5.38	3.75	2.75	5.25	5.00
FS	1	4.0	1.0	3.0	6.0	2.0	2.0	2.5	3.0	5.0	6.0	3.45
FS	2	6.0	6.0	4.0	7.0	3.0	4.0	3.5	6.0	6.0	6.0	5.15
FS	3	3.5	7.0	2.0	6.0	4.0	5.0	7.0	6.0	5.0	7.0	5.25
FS	4	3.0	5.0	2.5	5.0	4.0	6.0	7.0	6.0	6.0	5.0	4.95
FS Average		4.13	4.75	2.88	6.00	3.25	4.25	5.00	5.25	5.50	6.00	4.70
LC	1	5.0	7.0	5.0	5.0	5.0	1.0	4.5	1.0	1.0	1.0	3.55
LC	2	4.5	2.0	6.0	2.0	1.0	2.0	6.0	1.0	2.0	5.0	3.15
LC	3	6.0	3.0	4.0	7.0	7.0	3.0	6.0	5.0	6.0	5.0	5.20
LC	4	7.0	4.0	7.0	7.0	6.5	5.0	4.0	4.0	3.0	2.0	4.95
LC Average		5.63	4.00	5.50	5.25	4.88	2.75	5.13	2.75	3.00	3.25	4.21
LS	1	2.5	4.0	6.5	2.0	4.0	7.0	1.0	N/A	N/A	7.0	4.25
LS	2	7.0	1.0	5.0	3.5	7.0	6.0	2.0	N/A	N/A	2.0	4.19
LS	3	2.0	4.0	7.0	1.0	4.0	6.0	1.0	N/A	N/A	2.0	3.38
LS	4	4.0	1.0	2.5	1.0	2.0	3.0	1.0	N/A	N/A	4.0	2.31
LS Average		3.88	2.50	5.25	1.88	4.25	5.55	1.25	N/A	N/A	3.75	3.53
ML	1	2.5	2.5	2.0	3.0	1.0	4.0	4.5	5.0	6.0	5.0	3.55
ML	2	2.0	5.0	1.0	6.0	4.0	3.0	5.0	5.0	5.0	3.0	3.90
ML	3	5.0	1.0	3.0	3.0	2.0	2.0	3.5	3.0	3.0	6.0	3.15
ML	4	5.0	7.0	6.0	2.0	5.0	2.0	3.0	2.0	2.0	1.0	3.50
ML Average		3.63	3.88	3.00	3.50	3.00	2.75	4.00	3.75	4.00	3.75	3.53
PP	1	7.0	6.0	4.0	7.0	3.0	6.0	2.5	6.0	4.0	4.0	4.95
PP	2	3.0	4.0	2.0	5.0	2.0	7.0	3.5	4.0	4.0	1.0	3.55
PP	3	3.5	6.0	5.0	2.0	1.0	1.0	3.5	1.0	1.0	3.0	2.70
PP	4	1.0	2.0	1.0	4.0	3.0	4.0	5.0	3.0	5.0	6.0	3.40
PP Average		3.63	4.50	3.00	4.50	2.25	4.50	3.63	3.50	3.50	3.50	3.65
SS	1	1.0	5.0	1.0	1.0	7.0	3.0	6.5	2.0	3.0	2.0	3.15
SS	2	1.0	7.0	3.0	1.0	6.0	1.0	1.0	2.0	1.0	4.0	2.70
SS	3	1.0	2.0	1.0	4.0	4.0	4.0	5.0	2.0	4.0	1.0	2.80
SS	4	2.0	3.0	4.0	3.0	1.0	1.0	2.0	1.0	1.0	3.0	2.10
SS Average		1.25	4.25	2.50	2.25	4.50	2.25	3.63	1.75	2.25	2.50	2.70

The "raw" Boot and Sprd scores were combined into the "NSC measure" as described in Appendix C. The lower the rank, the better the performance.

eling, issues related to the characteristics of the test problem data sets, the comparison exercise itself, and issues which are approach specific.

5.1. Uniqueness and Ill-posedness

The issue of uniqueness of the inverse solution is discussed by *McLaughlin and Townley* [1996], who describe conditions that must be met for an inverse problem to be well posed. *Dietrich and Newsam* [1990] show that the problem of estimating transmissivity from steady state head measurements is ill posed unless the flow system is forced by a known recharge or pumpage which is sufficiently large to produce closed head contours over the region of interest.

Although ill-posedness can be mitigated to some extent when head measurements are augmented by transmissivity measurements, as in the WIPP test problems, it is still possible, that the resulting problems do not have unique solutions. That is, many different transmissivity fields may yield equally good fits to the available measurements. Some of these may be fortuitously closer to the "true" transmissivity field than others, but all are equally consistent with the data presented in the TPs. Clearly, this complicates the process of comparing different inverse approaches, but this is the reality facing a modeler at any site.

Although it might have seemed reasonable to base an inverse comparison on TPs that were well posed, a conscious decision was made to model the TPs after the real WIPP problem, which is probably ill posed in the sense that it does not have a unique solution. This decision forced each partici-

Table 6. Comparison of Fixed Well Versus Random Well Evaluation Measure Scores

	Inverse Method						
	FF	FS	LC	LS	ML	PP	SS
Fixed well average	5.3	4.2	5.1	3.6	3.4	3.6	3.0
Random well case average	5.7	4.6	3.9	3.4	3.4	4.1	2.9
Absolute value of difference	0.4	0.4	1.2	0.2	0.0	0.5	0.1

pant to deal with the issue of ill-posedness in their own way, generally by constraining the set of possible transmissivity solutions. In this study, the constraints were conveyed primarily by the transmissivity parametrization, which specifies how transmissivity values must vary over space [*McLaughlin and Townley*, 1996]. Parametrizations were implemented by specifying a particular transmissivity variogram, a particular spatial block scheme, and/or a particular set of pilot points, depending on the approach used (see Appendix B). A properly designed parametrization should transform the original ill-posed problem into a well-posed problem with a unique solution.

The need to deal with ill-posedness was thus one of the intrinsic features of the comparison. Since the different inverse approaches constrained the original problem in different ways, it could be argued that these approaches ultimately solved different problems. This even led one participant to claim that since there was no proven unique "truth" and since any one solution of the inverse problem obtained by any method could have equally well been the "truth" (provided its values of transmissivity and head at the measurement points were sufficiently close to the sample data), then the best simulation technique should be the one having the widest spectrum encompassing all the possible "truths" produced by each approach. In performance assessment, widening the spread of the possible outcomes of simulations is termed "risk dilution," as it may diminish the probability of the high-consequence region. This criterion was not employed, and the GXG considered that there was only one "truth" and that one of the aims of the comparison was to determine if the approaches could come close to that one "truth" and no other one.

As an example, the relatively good fit on the head error obtained by method PP on TP 3, although its T field was not superior to those from other approaches, can be taken as an indication of a potential nonuniqueness of the solution of that problem. This example shows, however, that the type of parametrization chosen by method PP was not close enough to the optimum to unravel the type of T distribution of the "true" T field.

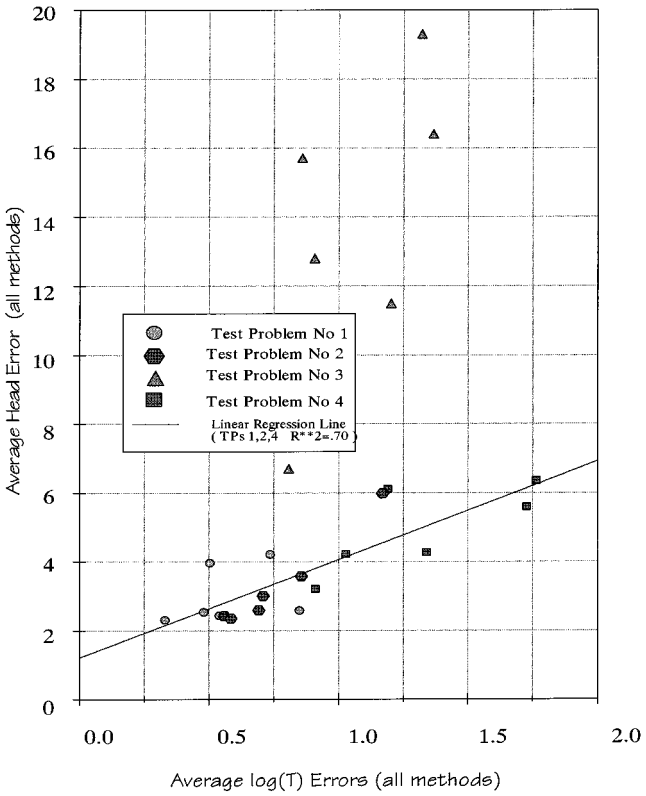


Figure 18. Correlation of head errors with errors in the log (T) field.

5.2. Choice of the Underlying Covariance Structure

For any geostatistical method a very important step is to determine the statistical structure of the field and to select the semivariogram of the random T field that is to be simulated. The seven approaches can be classified into two groups: those that use only the T data to select the semivariogram and those that use both T and head data simultaneously in the initial statistical inference step to develop the semivariogram of the T field. Only methods LS and LC strictly fall into the second category, using the maximum likelihood approach for this inference (see method descriptions in Appendix B). It was expected that the use of both head and T data would give these techniques an advantage over the other approaches, since incorrect assumptions about the semivariogram can have important consequences, as will be illustrated below. It is interesting to see that method LC ranked second for the semivariogram measure over all TPs, on average, even though the code did not allow for the selection of any variogram model other than exponential, which is a code limitation, not a method limitation. Method LS ranked fourth on this measure. However, the average score may be biased because LS ranked seventh on this measure for TP 1, perhaps because of insufficient initial attention of the participant to the importance of this selection. One advantage of the LS method (not used in the present exercise) is that it could be extended to use the transient head data to infer the semivariogram [*Dagan and Rubin*, 1988].

The average of the semivariogram measures for methods LC and LS over all TPs is lower (better) than the average of all other approaches excluding SS (in which the head information was used to guide the selection of the semivariogram model "manually"). Thus there is some evidence to support the contention that approaches that use both T and head data will in general do better than those using just the T data in selecting the semivariogram model. This is probably particularly true when there are many more head data than T data.

The importance of the selection of the semivariogram on the performance of an approach is illustrated by Figure 22, which shows the dependence of the average performance of each approach over all four TPs as a function of the average rank of the semivariogram measure. It is interesting to note that method SS, which in general ranks first across all measures, also ranks first in semivariogram selection (see Table 5). Method SS does not infer the semivariogram from both T and h data; it uses only T data. However, according to the participant, a great deal of attention was given to the fitting of the semivariogram model to the sample T data (careful analysis of the data, declustering, test of multiple population, elimination of outliers). The excellent exploratory data analysis to select the semivariogram (which is not method-specific, but participant-specific) is most likely one of the reasons for the success of the SS method. In addition, method SS also has the ability to recalibrate itself during the optimization, that is, to modify the semivariogram. In case the number of data is larger, the role of the selection of the semivariogram may not be as critical, because

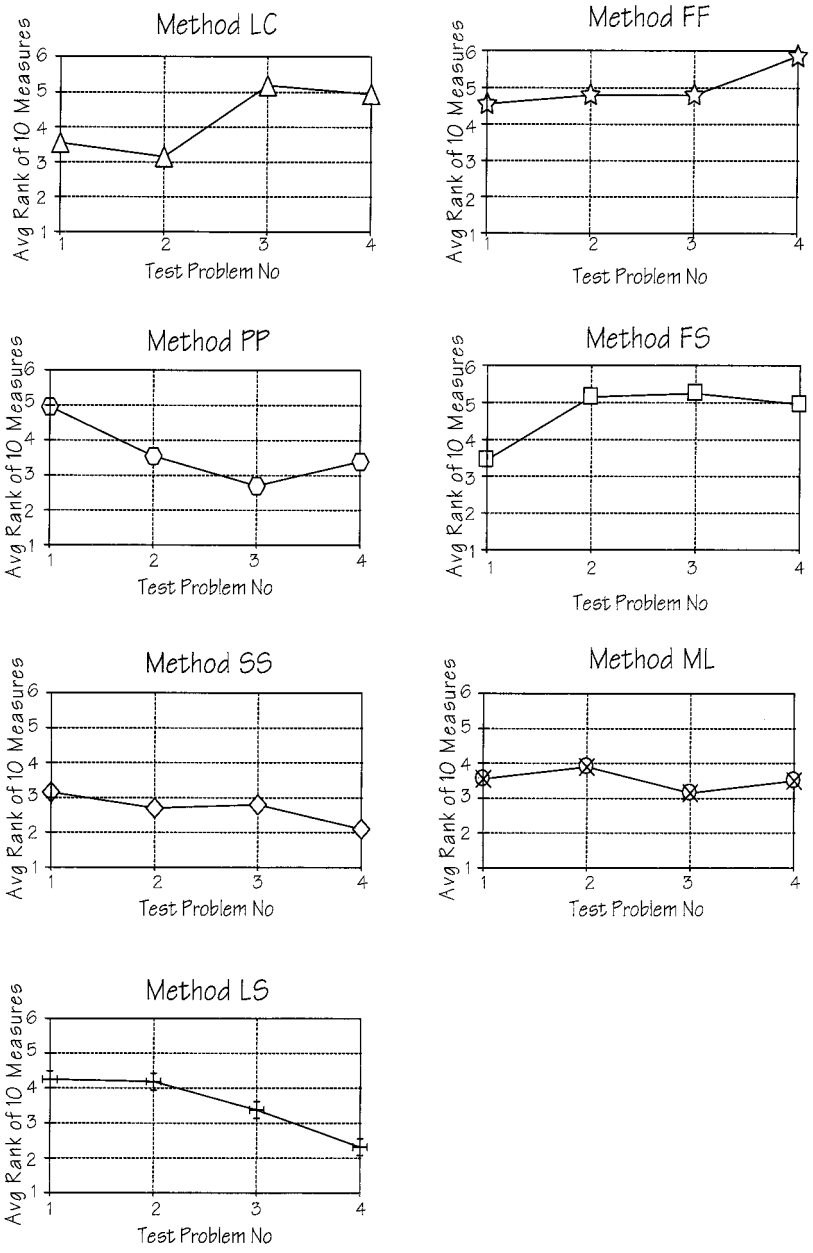


Figure 19. Comparison of method performance across test problems; the lower the average rank, the better the performance. Methods are indicated by the two-character abbreviation.

the conditioning on these data becomes dominant to structure the T field. Additional evidence of the importance of the choice of semivariogram model is provided in a section below.

5.3. The Multi-Gaussian Assumption

One of the methods, SS, has the advantage of being directly able to use the geostatistical "indicator" approach [*Journel and Huijbregts*, 1978] and permit any form of T distribution to be used, not just the multi-Gaussian one. The small sample size, however (41 data points), does not make it easy, in general, to detect the underlying type of transmissivity distribution. Only for TP 3, as we have seen, was this feature used in the exercise. In this case the method is able to generate values with a bimodal distribution and specific spatial correlation patterns for each population. The difference in results between method SS and the other methods is, however, small, and the multi-Gaussian assumption was not too erroneous. In real cases, however, it may happen that the underlying distribution of T is not lognormal or displays connectivity patterns at extreme threshold values inconsistent with a multi-Gaussian distribution. Therefore the ability of method SS to handle these characteristics would be quite valuable.

This intercomparison exercise therefore might not have been sufficient to evaluate the usefulness of this capability in an inverse method adequately. It can be stated, however, that the errors caused by making an erroneous choice will generally decrease as the number of conditioning data increases. Method ML can also use the indicator approach, and method PP was later adapted to include this capacity as a result of this comparison exercise.

5.4. The Assumption of Stationarity

TP 1 and TP 2 are, by construction, true stationary fields. TP 3 and TP 4 are not, as there are distinct local features and trends. Linearized methods assume the existence of a constant mean and random fluctuation around that mean and uniform flow on the average. Nonlinear methods do not depend on such assumptions. Tables 4 and Figure 19 clearly show that this had a significant effect on some of the linearized methods: LC in particular, as well as FF, show a systematic decrease of the performance between TP 1 and TP 2, and TP 3 and TP 4. The case of LS is quite interesting; the participant was able to detect from the sample data that the field did not look stationary and decided to apply a piecewise-linear approximation. He divided the domain into several subareas and assumed different means for each area. This made the results of LS for TP 3 and TP 4 much better than the results of the other linearized methods. Again, the skill of the participant to "tailor" the method to the particular features of the problem can improve the results significantly. The issue of nonstationarity is thus thought to be the primary reason why the cluster analysis (Figure 21) clearly makes a distinction between the linearized and the nonlinear methods.

5.5. Effect of Introducing Recharge

In and TP 3 and TP 4, recharge was introduced into the synthetic data sets, comprising 10% and 6%, respectively, of the total flow through the system. This information was not communicated to the participants, but it could have been inferred from the sample data for TP 3, which clearly showed a mound. In TP 4 the existence of recharge was not as evident from the sample data. Method PP was the only method to include recharge, which amounted to about 12% of the PP model system flux in TP 3. This may partially explain the better performance of method PP in TP 3.

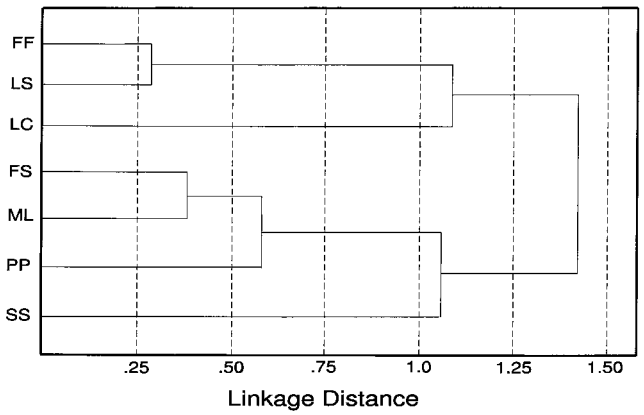


Figure 21. Cluster analysis tree diagram. Linkage distance is 1-Pearson correlation coefficient.

5.6. Effect of Grid Discretization

It is well known that particle tracking is very sensitive to grid size and to time steps. In order to minimize this effect, the same particle-tracking code and time-stepping scheme was used by the coordinator for all the T fields provided by the participants. But the grid used was the one provided by the participants. The degree to which the various discretization schemes affected the GWTT calculations was, unfortunately, not assessed.

(a)

T Grouping			Mean	Method
	A		5.000	FF
	A			
B	A		4.700	FS
B	A			
B	A	C	4.213	LC
B		C		
B		C	3.650	PP
	D	C		
	D	C	3.531	LS
	D	C		
	D	C	3.525	ML
	D			
	D		2.688	SS

(b)

FF	–						
FS	X	–					
LC	X	X	–				
LS	O	X	X	–			
ML	O	O	X	X	–		
PP	O	O	X	X	X	–	
SS	O	O	O	X	X	X	–
	FF	FS	LC	LS	ML	PP	SS

Figure 20. Results from the LSD pairwise comparison. (a) Means with the same letter are not significantly different; $a = 0.05$, critical value of $T = 2.08$, least significant difference = 1.12. (b) Results from the LSD pairwise comparison. "X" indicates the means are not significantly different, and "O" indicates the difference in the means is statistically significant at the $a = 0.05$ level.

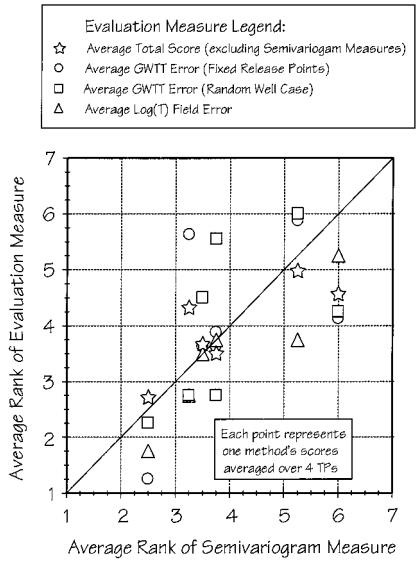


Figure 22. The sensitivity of the methods performance to the estimation of the covariance structure of the log (T) field. Figure shows how errors are correlated with the quality of the semivariogram estimates.

5.7. The Magnitude of the $\log_{10}$ (T) Variance

In the so-called linearized methods (FF, LC, and LS), the development of the inverse equations is based on the perturbation method, which assumes that the $\ln(T)$ field has a "small" variance. In the literature [e.g., *Dagan*, 1989] and depending on the problem at hand, it is generally assumed that such a linearization is valid for $\ln(T)$ variances smaller than 1. However, in some cases larger variances do not jeopardize linearized methods, while in other cases, variances larger than 0.1 could not be adequately handled by linearization [e.g., *Roth et al*., 1996]. The variances of the true (synthetic) $\log_{10}$ (T) fields ranged from 1.38 to 2.14 across all four TPs. This corresponds to $\ln(T)$ variances in the range of 7.30 to 11.32, far in excess of the variance typically considered valid for the linearized approach. We have shown, however, when qualitatively comparing the results of TP 1 and TP 2, where the only difference was the variance of the $\log_{10}$ (T) field, that the linearized methods did not have any difficulty with such large variances. This is also clear from a comparison of the average rank scores over all measures for each TP (see Figure 19). Thus, it seems that given the type of inverse problems dealt with in this comparison, and given the objectives of the exercise, the magnitude of the variance is not an important issue. This conclusion is most likely linked to the effect of conditioning, as the small variance assumption was initially formulated for unconditional cases. This is particularly true in this exercise, where the average distance between measurement points (in the central area) is much shorter than the correlation length of the $\log_{10}$ (T) fields in all test problems. If this had not been the case, the effect of conditioning might have been much smaller. Indeed, *Neuman and Orr* [1993] compared linear and nonlinear stochastic approximations of effective hydraulic conductivity in three-dimensional (3-D) infinite domains and concluded that nonlinearity becomes critical for $\ln(K)$ variances in excess of 2. Similar results were obtained by *Paleologos et al*. [1996] for bounded domains and by *Hsu et al*. [1996] for transport problems, for variances of $\ln(K)$ on the order of 1 to 2.

5.8. Connectivity of the High-T Zones

The major difference between TP 3 and TP 4 is that the high-T zones are discontinuous for TP 3 and connected for TP 4. The discontinuous high-T zones were very difficult for all methods in TP 3, and most raw evaluation measures show poor results, particularly for the head error. The ranking order for the top four methods in TP 3 (based on average scores across all evaluation measures) is PP, SS, ML, and LS. One may interpret this ranking on the basis of the intrinsic features of the methods. In the PP method the structure of the T field results from the selection of the "pilot points," where the transmissivity value is calibrated by the optimization algorithm and further used to krige (or simulate) the whole T field. But contrary to all the other methods, the location of the pilot points is also one of the unknowns of the problem. An optimum selection of the location of the pilot points is made iteratively, prior to optimizing the T value (see Appendix B). By contrast, SS selects a priori the position of the "master locations," which are to some extent equivalent to the pilot points, and ML also selects the zoning of the field a priori. For TP 3 method SS was able to generate unconnected high-T channels using the transient head data to determine the major direction of the channels. Other participants tried to prescribe channels "by hand," without a great deal of success. The PP T fields, while matching the heads better than the other methods in TP 3, do not match the T field values very well; we have already pointed out that this may reflect a nonuniqueness problem. The PP method may, perhaps, more easily detect local anomalies if they cannot be introduced a priori from geological or external knowledge.

By contrast, in the case of continuous high-T zones (TP 4), all methods performed reasonably well. The calculated T fields more or less correctly show a continuous high-T flow path, in general correctly located. Continuous high-T zones are thus more easily detected by inverse methods than discontinuous ones, at least in steady state and for problems similar to the ones examined.

5.9. Importance of Transient Data

It is difficult to see any significant differences between the two methods that could directly use the transient data in the formulation of the inverse problem (ML and PP) and the other methods. In TP 4, ML and PP obtain very similar results, but SS and LS, which do not directly use the transient information, perform significantly better. Similar results were reported by *Gonzalez et al*. [1997], where both stationary and transient data were shown to improve stability but did not lead to a better solution. In order to better understand the reason for this outcome, ML and PP were asked to rerun TPs 3 and 4 using only the steady state data and discarding the transient information. The outcome showed that both methods produced results very similar to those obtained with the transient information, particularly for TP 3. This is in contrast to the real WIPP site data, where the PP methodology was used in a preliminary PA [*LaVenue et al*., 1995] and where the use of the transient information proved to make a significant change in the outcome of the inverse calculations. The reason the additional transient information from the pumping tests did not result in major changes in the outcome is believed to be due to the limited areal influence of these tests. However, close examination and analysis of the TPs in the areas affected by the pumping was not performed. Thus we see that the evaluation measures did not provide enough information about the value of the transient information. We will not therefore be able to draw any conclusions on the value of transient information from this exercise.

5.10. Effect of Code Limitations

The methods that were compared were not all at the same stage of development. In particular, method LC was developed in 1983 for solving a specific problem and had not been significantly updated since. The available code could not handle more than about 1600 grid blocks, which forced the use of a coarse grid to represent a small domain, giving perhaps too much importance to the head boundary conditions and limiting its ability to reflect the desired correlation behavior adequately. The performance of this method in these test problems is therefore hindered by this constraint, which is specific to the code, not to the method itself.

5.11. Motivation of the Participants

The participants learned about the effectiveness of their method during the course of the comparison, as the "true" field and some preliminary evaluation measures were computed and released to the participants after each TP had been run, prior to starting the next one. Apart from method LC, which was run by D. Gallegos and C. Axness and not by the code developer, all other participants either ran their own

codes or were directly involved in the supervision of the exercise. Some participants treated the exercise as a competition and felt peer pressure to become "the winner," while others treated the exercise as more of a learning experience.

In some cases the participants made some improvements to the codes to accommodate difficulties encountered during the tests. It is clear, for instance, that method PP did very poorly on TP 1, for at least three reasons that were later understood: the grid was too coarse; the domain was too small, which gave too much importance to the boundary conditions; and the selected variogram was estimated without enough care. The choice of a linear variogram, without a sill, resulted in a too-large variance in the simulated fields. A more careful analysis of the sample data, as was performed by the participant of the SS method, could have shown that an exponential variogram would have been a better choice. This was verified by rerunning method PP for with an exponential variogram without changing the grid, which produced better evaluation measure scores than the linear variogram case. A comparison of the waste panel CDFs produced by method PP using both the linear and the exponential semivariogram models (Figure 23) shows significant improvement with the exponential model; the envelope of CDF curves for the exponential case more completely covers the true CDF, while the CDF for the linear semivariogram case covers a much broader range and has a very long tail (see also Figure 11). Therefore the results of the comparison reflect not only the intrinsic quality of a method but also the skill and experience of the team that ran it and occasionally improved it; these two effects cannot be easily distinguished.

5.12. Case of the Linearized Semianalytical Method

Among the seven methods that were compared, six use numerical techniques that discretize the domain and transform the problem into discrete grid blocks and solve the flow equation by finite differences or finite element techniques. The seventh method, LS, is "semianalytical" and solves the GWTT problem directly but without discretization, without defining boundaries (therefore without the need to specify boundary conditions), and without generating block values. The method directly calculates the movement of particles in the velocity field, which is conditioned on the T and head data via the geostatistical model. This method, although linear, produced very good results and compares favorably with the three nonlinear methods. As noted earlier, the method was used in a piecewise-linear fashion to account for the nonstationarity of the TP 3 and TP 4 fields. One of the major advantages of the LS method is its computational efficiency; it does not need the large computer resources required by many of the other methods. It should be noted that the method could be extended to produce simulated values of transmissivities on a grid, which could then be used as input for a numerical solver of the flow and transport equations. This method could also be extended to produce concentrations directly. It has been extended to 3-D to include transient data and uniform recharge. During the course of this comparison, however, resources were not available to evaluate the adequacy of the discretized transmissivities that the LS method could produce and to compare them with the transmissivities produced by the other methods (note that in Tables 4 and 5, the T and head-field evaluation measures are not available for method LS).

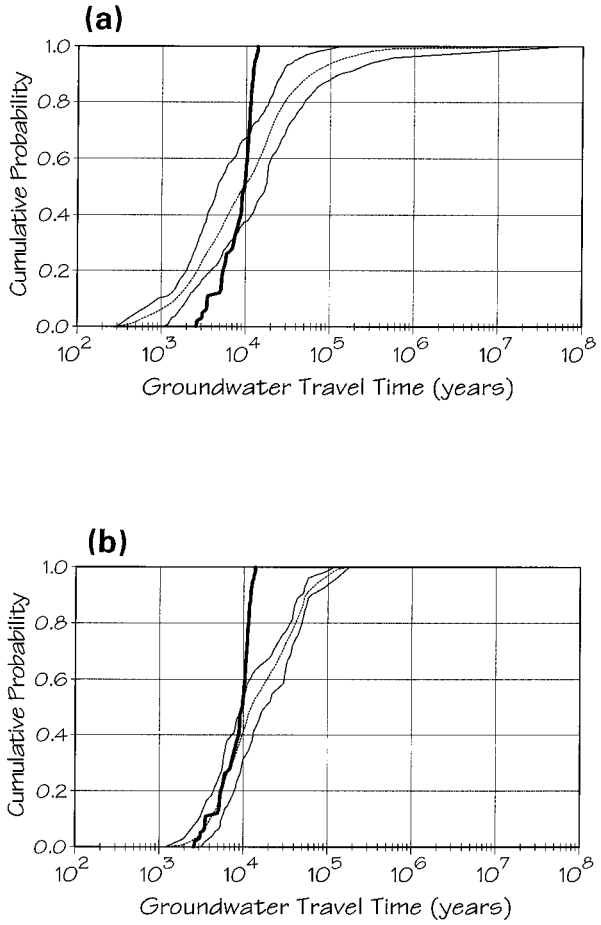


Figure 23. Waste panel CDF for method PP in test problem 1: (a) linear semivariogram model and (b) exponential semivariogram model.

5.13. The Case of the Fractal Simulation Method

The FS method performed relatively well for, but much less so for the other cases, which had either a larger $\ln(T)$ variance or nonstationary fields. It seems that this is due to the principles of the method. First of all, the FS method is not really an inverse algorithm (see Appendix B). Once a T field has been simulated (with a fractal underlying semivariogram and conditioning on the T data only), the conditioning to the head data is not done by altering the simulated T field but by optimizing the head boundary condition values (which, as specified earlier, are left to each participant to decide). If no constraints are applied to these head values, the results can be physically meaningless. If constraints such as continuity or ranges are added to these head values, the fitting of the head may be poor (as each T field is fixed). This is especially true for the more "complex" fields of TP 3 and TP 4, thus leading to a poor global performance. Therefore this method seems limited to stationary fields with rather small $\ln(T)$ variances. This method generally produced the largest GWTT spread (see Table 4), which means that in general, it overestimates the uncertainty. However, a reasonable fit of the sample T data could be obtained with a fractal semivariogram, at least for TPs 1, 2, and 4 even if the underlying semivariogram of the synthetic field was not fractal.

5.14. The Case of the Maximum Likelihood Method on TP 4

TP 4 was non–multi-Gaussian but was interpreted as multi-Gaussian by the participant for the ML method for TP 4. This is acceptable when the effect that this error has on head data is small, which is the case with steady state data. However, the effect of continuous channels is much more severe when transient head data are used. Transient head data are indeed able to identify high-T channels better than steady state head data. Since such channels cannot be generated in a stationary multi-Gaussian field, the only option left available to method ML was to increase the T away from the channels. Consequently, the resulting T fields have a higher mean than the true field. This is probably the reason for the poor performance of method ML on TP 4. It also explains why the T fields simulated with only the steady state head data were better than those simulated with both the steady state and transient data. This explanation is also consistent with the relatively better performance of method ML on TP 3, where the participant artificially increased T along "guessed" channels, based on a manual interpretation of the transient data.

5.15. The "Robustness" of a Method With Respect to the Type of Heterogeneity

It has been shown that some methods perform better for a given type of heterogeneity, while they would perform less well for another. In practice, it may be difficult to know in advance which type of heterogeneity is dominant for a given aquifer. At the WIPP site, for instance, it is not yet clear if the high-T zones in the aquifer are discontinuous or connected. It is therefore of interest to detect if there are methods that perform well on the average but may occasionally produce very poor results. The ANOVA and cluster analyses have shown that four methods seem to have a similar behavior: LS, ML, PP, and SS. The three others, FF, FS, and LC, fall into a second, less desirable category, probably because of their difficulty in dealing with nonstationary fields. Among the first four methods, PP behaved rather poorly for TP 1, but this was thought to be linked to insufficient discretization and poor selection of the variogram from the sample data. For the more complex TP 3 and TP 4, which are also the more realistic "WIPP-like" cases, the average scores (lower being better) are SS = 2.5, LS = 2.8, PP = 3.1, and ML = 3.3. Thus method SS appears to be the most robust, followed by LS, PP, and ML.

6. Conclusions

6.1. Importance of the Topologic Structure of the T Field

The results of the comparison exercise demonstrate the following:

1. In developing the inverse model, the greatest attention should be given to the selection of the semivariogram to be used in the inversion, as this appears to be very significant in achieving success. This selection must be made from the ensemble of transmissivity data available from the site, with careful declustering and checking of the distribution of the data and elimination of outliers. Using both the T and the heads in calculating the semivariogram can improve this selection significantly, particularly if the number of the T data is few compared to the number of head data.

2. Identifying the proper parametrization (topologic/geometric structure of the T field) can be more important than estimating parameter values. It has been shown that the calibration of the model (head matching) can be very good even with a T field that is not very representative of reality. Flexibility in this parametrization is thus an important factor for an inverse model.

3. The issue of the level of discretization did not receive sufficient attention. Many participants would agree that using a fine grid can be important but not necessarily the dominant factor. Method SS, for instance, used a coarse grid and did very well. On another hand, the choice of the grid is closely linked to the issue of upscaling, to the size of the domain which a measured value (e.g., pumping test) represents, and the degree to which the assumed correlation structure can be represented. Assigning the "measured" values to a given grid size in their mesh may have been a source of bias for some participants. But this could not be determined from the results.

4. Neglecting to consider recharge when it exists in the real problem does not seem to be of major importance, if this recharge remains on the order of 10% or less of the total flux through the aquifer system. But including recharge when it indeed is present in the real system seems to improve the calibration of the model, even when the quantity and distribution is unknown.

6.2. Improvements in the Inverse Methodologies

The results presented herein clearly show that there is much room for improvement in the inverse methodology. It is disturbing to see that the available methods still do not adequately assess the uncertainty of the prediction. The clear message of this exercise comes, we believe, from the results of TP 3. In this case the design committee tried to create an aquifer that was realistic in its complexity and not constructed to be "geostatistical," that is, not a realization of a stationary random function with a multi-Gaussian log (T) distribution with a simple semivariogram. In the past, researchers have perhaps focused too much on validating their inverse methods on too-simplistic synthetic T fields. What can be recommended, on the basis of the present study, is the following:

1. Gaussian geostatistically based inverse methods have a tendency to generate parameter fields with circular (or ellipsoidal) heterogeneities. This is due to a basic principle of Gaussian geostatistics, which assumes that the correlation of the parameter values in space is a regular function of the distance, with or without anisotropy, valid for all classes of transmissivities. In the case where the heterogeneity of the aquifer is made of linear features (such as faults, channels, etc.) of varying orientation imbedded into a different matrix, the multi-Gaussian geostatistical approach is probably inadequate. The indicator approach is then a better choice, as it can use different variograms for each class of transmissivities. If some variograms are taken as very anisotropic, then some channels or fractures can be represented with the orientation prescribed by the anisotropy of the variogram. This approach was taken by *Tsang* [1996], among others. Inverse methods based on conditional expectations, maximum a posteriori probability, maximum likelihood, minimum variance, or variants will not be able to produce natural features leading to discontinuities (such as fractures, paleochannels, dissolution channels, etc). These have to be incorporated explicitly in the model. Nonparametric geostatistics are more flexible in this respect. One alternative could be to generate such structures randomly, like Boolean objects used in the oil industry, while introducing fine-scale variability within each object or estimating the T through inverse procedures for each class of object.

This is similar to the approach taken by *McKenna and Poeter* [1995]. Fields that do not reproduce head data can then be eliminated. They could also possibly be introduced via some optimization procedure, provided uncertainty remains included.

2. For those cases where a geostatistical description of the heterogeneity is adequate, it is clear that the ability to use both the T and head data to identify the underlying log (T) semivariogram is very desirable, since it has been shown in this test that it improves the statistical inference. This is one of the best features of the LC and LS methods, but it could be added to others. The maximum likelihood inference part of the LS method, for instance, could be used as a front end to any other inverse. As an example of how this recommendation could be implemented slightly differently in an existing inverse, let us take the case of the PP method. The initial developer of this method [*de Marsily*, 1978; *de Marsily et al*., 1984] felt it to be a deficiency of the method that after calibration, the variogram of the calibrated field, calculated by using the measured T values and the calibrated T values at the pilot points, could be different from the initial prescribed one, based only on the measured T data. The validity of the PP was often questioned because of this potential "deficiency." Given the results of this comparison, it is clear that the evolution of the variogram before and after calibration in some way reflects the conditioning by the head data. One could therefore adopt a strategy where the PP inverse would be run once, to obtain additional T values at the pilot points, to better infer the semivariogram, and then rerun (or iterated) with the new semivariogram. Such a strategy is already imbedded in method SS.

3. Allowing for simultaneous calibration of the T field and the boundary conditions (in cases where they are not well defined), as is done in most methods, is certainly an important feature. But it is also necessary to impose reasonable (physically plausible) constraints on these boundary conditions during calibration.

4. If linearized methods are to be used, the main issue seems to be the stationarity of the field rather than the magnitude of the ln(T) variance. It would therefore be desirable to develop methods that could detect nonstationarities and optimally select zoning with piecewise stationary properties.

5. Using transient head data is in general a significant improvement of a method. Methods ML, PP, and LS were able to use transient data. Method SS was extended to it during the course of this exercise.

6.3. Effort Applied Toward Solution of the Inverse Problem

At this stage of the development of the inverse methodologies, it is not advisable to use them as robust "black boxes." The experience and skill of the modeler and the time and effort spent on the modeling of the problem have been shown to be essential components of success. Some participants consciously decided to use their methods with minimal intervention, to see precisely what the outcome would be. Their methods, in general, performed less well than the methods for which a substantial effort was applied. In this respect, method LS, which does not have a large number of options or parameter values to select, such as discretization, boundary conditions, time steps, etc., would most likely produce more reproducible results if used by different modelers.

6.4. Design of Intercomparison Studies

This intercomparison has highlighted some difficulties that may be of interest for those who want to perform a similar exercise. Among these are the following:

1. A design subcommittee separate from the participants is very desirable. If the objective is really to evaluate methodologies and not at the same time improve them, a series of tests should be given without the outcome of the first test being available before the next is run. The design subcommittee should not limit the synthetic fields to "classical" fields, but should try to imagine (based on geological knowledge and experience) what real fields might look like and incorporate them into the "true field" exhaustive data set.

2. The set of evaluation measures on which the methods will be compared needs to be specified from the start. This is not easy. In order to do so, the objectives of the comparison must be fully developed and stated explicitly from the outset, and the measures must be designed to achieve those objectives. In this exercise, the initially agreed upon set of measures proved to be inadequate and the final set was only decided when all the tests had been made. We do not believe that this has created biases in the results presented in this paper, but it certainly created a lot of discussion and confusion, as the methods could be adjusted to satisfy (or not) a given criterion.

6.5. Selection of Appropriate Inverse Approaches for PA

Four approaches have been identified as being approximately equivalent for use in performance assessment at sites such as WIPP; these methods are LS, ML, PP, and SS. With such methods, the uncertainty is very clearly reduced by the conditioning compared with unconditional simulations respecting only the pdf of the measured parameter. The outcome of the simulations (in our case the CDF of the advective travel time) is reasonably similar among these methods.

It should be noted that these approaches do not give identical results: the T fields and the predicted uncertainty, as given by the spread of the CDF, are significantly different among the methods. These differences stem from the differences between the techniques (e.g., parametrization, assumptions on stationarity) and thus reflect a fundamental uncertainty associated with the inverse problem because of its nonuniqueness. Each method is "conditioned" by its own assumptions to make the problem well posed, and the differences between the methods display the importance of these assumptions. The total uncertainty could therefore be better described by the results of the ensemble of several methods, as any one single method in general tends to underestimate the uncertainty.

This study has not addressed the question whether these differences between methods could lead to different conclusions, in terms of performance assessment, when contaminant transport is simulated and not just travel time. This would depend on how close the outcome of the simulations is from the performance target, but we believe that it would not be vastly different for the examples reported here.

It should be emphasized that the four approaches which were found to be approximately equivalent are not just "methods," but are at the same time codes and a manifestation of the manner in which the method was applied, reflecting the time, effort, and experience of the modeling team that worked on the problems. Those three factors are unequivocally imbedded in this comparison.

The other methods involved in this comparison were found to have either too stringent assumptions (e.g., stationarity),

coding constraints, or insufficient time and effort devoted by the modeling team to produce results of the same level as the previous ones, particularly for the more realistic TPs.

Appendix A: Geostatistical Expert Group Participants

(1) C. L. Axness, Sandia National Laboratories, Albuquerque, New Mexico; (2) R. L. Beauheim (TP design committee member), Sandia National Laboratories, Albuquerque, New Mexico; (3) R. L. Bras, Massachusetts Institute of Technology, Cambridge; (4) J. Carrera, Universitat Politècnica de Cataluña, Barcelona, Spain; (5) G. Dagan, Tel Aviv University, Tel Aviv, Israel; (6) P. B. Davis (TP design committee member), Sandia Laboratories, Albuquerque, New Mexico; (7) G. de Marsily (TP design committee member), Université Paris IV, Paris, France; (8) D. P. Gallegos, Sandia National Laboratories, Albuquerque, New Mexico; (9) A. Galli, Ecole de Mines de Paris, Fontainebleau, France; (10) J. Gómez-Hernández, Universidad Politècnica de Valencia, Valencia, Spain; (11) S. M. Gorelick (TP design committee member), Stanford University, Stanford, California; (12) C. A. Gotway (TP design committee member), University of Nebraska, Lincoln; (13) P. Grindrod, QuantiSci Ltd., Henley-on-Thames, England, United Kingdom; (14) A. L. Gutjahr, New Mexico Institute of Mining and Technology, Socorro; (15) P. K. Kitanidis, Stanford University, Stanford, California; (16) A. M. Lavenue, Duke Engineering and Services Inc., Austin, Texas; (17) M. G. Marietta (TP design committee member), Sandia National Laboratories, Albuquerque, New Mexico; (18) D. McLaughlin, Massachusetts Institute of Technology, Cambridge; (19) S. P. Neuman, University of Arizona, Tucson; (20) B. S. RamaRao, Duke Engineering and Services, Inc., Austin, Texas; (21) C. Ravenne, Institut Français du Pétrole, Rueil-Malmaison, France; (22) Y. Rubin, University of California, Berkeley; and (23) D. A. Zimmerman (TP design committee member), GRAM, Inc., Albuquerque, New Mexico.

G. de Marsily was chairman of the GXG. D. Zimmerman was the GXG coordinator. (He created the test problem data sets, distributed them to the participants, collected from them their results after calibration, performed the GWTT calculations, and conducted the comparative analyses.)

Appendix B: Brief Description of Each Inverse Method

B1. The Fast Fourier Transform Method (FF)

This technique was developed by A. Gutjahr at New Mexico Institute of Mining and Technology, Socorro [*Gutjahr and Wilson*, 1989; *Robin et al*., 1993; *Gutjahr et al*., 1994]. The method is implemented in the code CSIMFFT. This code solves 2-D, steady state groundwater flow problems with a fast Fourier transform technique for field generation. The log transmissivity field and the mean-removed head field were considered to be statistically homogeneous for this exercise. The newest version of the code is able to consider a ln$T$ trend. An iterative cokriging procedure is implemented to condition on transmissivity and head field measurements. The FFT technique is very efficient and is capable of generating many realizations with modest computing resources over times on the order of minutes. The procedure used in these tests could not handle recharge or time-dependent data.

B2. The Linearized Semianalytical Method (LS)

This technique is based on the conceptual and analytical tools developed by G. Dagan at Tel Aviv University and by Y. Rubin at Tel Aviv University and at the University of California, Berkeley [*Dagan*, 1985; *Rubin and Dagan*, 1987, 1992; *Dagan and Rubin*, 1988; *Rubin*, 1991a, b]. The procedure comprised two stages: first, the solution of the inverse problem and, second, the solution of the transport problem.

The solution of the inverse problem is achieved by adapting a stationary log transmissivity structure of an analytical form (e.g., exponential) that is fully characterized by a few unknown parameters, by using a first-order, linearized, solution for the head field to obtain analytical expressions for the head-ln$T$ cross covariance and the head covariance and by identifying the unknown parameters (mean head gradient, log transmissivity mean, variance, and integral scale) with the aid of measurements. This is done by a maximum-likelihood procedure applied concomitantly to both transmissivity and head measurements. The head and transmissivity fields can be generated subsequently at any point by conditioning on measurements (through cokriging). The method does not imply a lognormal distribution of transmissivity, though it is supposedly better suited to such distributions.

The solution of the transport problem is carried out by particle tracking. At each time step and along the trajectory of each particle, the velocity is generated directly by conditioning (cokriging) on head and transmissivity measurements by using first-order analytical solutions for the velocity log transmissivity and velocity head cross covariances.

To account for trends in log transmissivity which may by responsible for nonstationarity and large variances present when the entire domain is regarded as a single unit, the method was applied over separate subdomains in the final test problem. The method does not require numerical solutions of the flow equations and is free of discretization errors. The numerical computations pertain to the maximum-likelihood stage, to conditioning by cokriging, and to particle tracking. The main limitation is the first-order approximation, implying that conditioning on measurements extends the range of validity of the method. The method can be easily applied to 3-D simulations, and a 3-D code is available.

B3. The Linearized Cokriging Method (LC)

This technique was developed by P. Kitanidis, R. Hoeksema, E. Vomvoris, and R. Bras and is implemented in the GEOINVS code [*Kitanidis and Vomvoris*, 1983; *Hoeksema and Kitanidis*, 1984; *Kitanidis and Lane*, 1985]. The technique differs from the other linear techniques because it implements maximum-likelihood estimation of the structural parameters associated with the log transmissivity covariance based on both T and h data. The GEOINVS code implementation of this methodology is limited by the fact that an $N \times N$ (where N is the number of interior nodes in the flow model) matrix is inverted directly, so that calculation is restricted in the present version of the code to grids on the order of 40×40 for simulation on present-day workstations. An improved numerical implementation of this code is in development. An advantage of this technique is that it is very simple to implement and has not suffered from convergence problems.

B4. The Fractal Simulation Method (FS)

The selfaffine fractal technique was developed by P. Grindrod and M. D. Impey of Intera Information Technologies (now

QuantiSci) [*Grindrod and Impey*, 1991]. This technique depends upon the assumption that the spatial variability of the log transmissivity may be represented in the form

$$\Gamma_{\psi}(h) \equiv \langle |\psi(\mathbf{x} + \mathbf{h}) - \psi(\mathbf{x})|^2 \rangle \propto h^{2p}$$

where ψ is log transmissivity, h is the separation between two points in the field, p is the Hurst coefficient [*Mandelbrot*, 1983] and $\langle \quad \rangle$ denotes the ensemble average over all realizations. The method proceeds by calculating the experimental semivariogram of the log transmissivity data and then fitting the fractal scaling law to the data. The parameter a and the Hurst coefficient p are chosen to best fit the data using maximum likelihood estimation. A set of fractal fields is then generated using the fast Fourier transform method with randomly generated phase and amplitude coefficients. The conditioning of these fields to the transmissivity data is accomplished through a linear superposition of the unconditioned fields, where the difference between the variance of the final field and the observed data is minimized. The full flow equation is then solved using the T fields generated above. For each realization, the set of head measurements are in effect "fit" by calibrating the heads at the boundary and the head data do not affect the individual T fields.

B5. The Pilot Point Method (PP)

This technique was developed by B. S. RamaRao, A. M. LaVenue, and G. de Marsily [*RamaRao et al*., 1995; *LaVenue et al*., 1995] and begins by estimating the variogram using the T data and then generates unconditional simulations of the transmissivity field with this variogram using the turning bands method. These transmissivities are then conditioned to honor measured transmissivities by the addition of a simulated kriging error to the kriged field based on the measured data. An automated iterative calibration follows in which an objective function defined by a weighted sum of the squared deviations between the computed and the observed pressures over points in the spatial and temporal domains is minimized. Pilot points are synthetic transmissivity data points and are used as parameters of calibration. During calibration, pilot points are added to the measured transmissivity database to produce a revised conditional simulation. Coupled adjoint-sensitivity analysis and kriging are used to locate pilot points optimally, where their potential for reducing the objective function is the highest. Gradient search methods, subject to subsequent constraints, are used to derive optimal transmissivities at the pilot points. The pilot points are added to the transmissivity data base for purposes of kriging, but the simulated kriging error to be added for conditional simulation is based on the measured transmissivities only and, thus, remains the same across all iterations. At the end of an iteration, a revised transmissivity field and the corresponding pressure field are obtained. The test for convergence of iterations is based primarily on a prescribed minimum value for the objective function and a prescribed maximum number of pilot points. Each conditionally simulated transmissivity field is calibrated separately.

B6. The Maximum-Likelihood Method (ML)

This technique, implemented in the INVERT code, is a very general nonlinear technique that estimates the aquifer parameters (transmissivity, recharge, storage, leakage coefficients, prescribed boundary heads, or flow rates) using prior estimates of their values along with transient or steady state head measurements. It was developed by *Carrera and Neuman* [1986a, b]. Parameter estimation is performed using the maximum-likelihood theory, for which several optimization methods are available. The nonlinear flow equation is solved by the finite element method using a fully implicit lumped time integration. The flow domain can be 1-D, 2-D, 2-D radial, or quasi-3-D, where 1-D linear string elements may be used to represent vertical flow, fractures, well bore effects, etc. The INVERT code minimizes an objective function consisting of an error component associated with the measured head data and a weighted error component associated with the prior estimates of other hydrologic parameters. The weighting is a parameter that is varied manually in simulation. For several values of this weighting parameter the objective function is minimized. The INVERT code offers a number of gradient and Gauss-Newton methods for minimizing the objective function. In conjunction with this exercise, INVERT was used to estimate aquifer parameters simultaneously from three transient pumping tests using prior INVERT block transmissivity estimates computed from steady state transmissivity and head data. Some of the advantages of the INVERT implementation of the ML technique are that it is a fast, powerful, well-documented code that is being used extensively and is actively undergoing development. A more up-to-date description of the code's geostatistical formulation is given by *Carrera et al*. [1993]. When the exercise was started, the optimization algorithm CPU time was highly sensitive to the number of blocks (pixels) over which T is estimated. As a result of this exercise, a new optimization method, whose CPU cost is virtually independent of the number of blocks, was developed [*Carrera and Medina*, 1994]. However, zones for TPs had already been prepared with cost-reduction constraints in mind. That is, small zones were used where data were abundant, and large zones were used elsewhere. To simulate small-scale variability at each block, the following algorithm was used. First, starting from the simulations for each finite element block, assign the simulated values as measurements to 2×2 gauss points of each finite element grid block. Second, generate a simulation conditioned on these points on any desired grid (e.g., on a very fine regularly spaced grid).

B7. The Sequential Self-Calibration (SS)

This method was developed by the Department of Hydraulic and Environmental Engineering [*Sahuquillo et al.*, 1992; *Gómez-Hernández et al*., 1997; *Capilla et al.*, 1997]. SS is able to accommodate both multi-Gaussian and nonmulti-Gaussian random function models using the indicator kriging approach. The indicator kriging approach can be seen as a superset of the multi-Gaussian approach. That is, if the data are suitable to be modeled by a multi-Gaussian random function, the indicator approach will produce the same results. However, it can handle the histogram of the data as is, that is, normal, lognormal, or otherwise, and it can inject spatial patterns to certain transmissivity classes that could not be reproduced with multi-Gaussian models. The nonmulti-Gaussian model can be used to introduce multiple populations of transmissivities or fracture-like features in the simulations. The decision to use which model is taken after careful examination of the data. The transmissivity data are then kriged, and the kriging standard deviation is calculated at each grid block location. A grid oriented in the mean flow direction is constructed, and a seed transmissivity field, according to the random function model chosen and conditional to the available transmissivity data, is

generated. The next step is the computation of a transmissivity perturbation field so that forward simulation of flow in the seed field plus the perturbation reproduces the head data.

Determination of the seed field is done by optimization. The perturbation field is parameterized by a few values at selected master locations. Perturbation of the remaining cells is obtained by kriging interpolation of the master location values. The set of master locations always includes the transmissivity measurements (at which the perturbation is constrained by the transmissivity measurement error) and the transmissivity perturbation at the master cells never falls outside the interval of the kriging estimate plus or minus three kriging standard deviations.

Appendix C: Detailed Description of Evaluation Measures

Three classes of analyses were used to compare and rank the inverse methods: (1) GWTT analyses, (2) PATH analyses, and (3) field variable analyses. Within each class there are several "evaluation measures" that were used to quantify and characterize the performance of the methods. In the end, there were 10 evaluation measures used to characterize the performance of the methods. The final set of evaluation measures was arrived at through iteration and consensus of the GXG and test problem participants. The measures of method performance are not all necessarily independent.

For GWTT and PATH analysis, the measures were designed to quantify such performance characteristics as (1) the error between the estimated CDFs and the true value or true CDF, (2) the magnitude of the spread in the distributions, (3) the robustness and self-consistency of the method, and (4) for GWTT only, the bias toward cautious or noncautious estimates (i.e., underpredicting rather than overpredicting the GWTT). A method is considered robust if the true GWTT or PATH falls within the calculated range of the CDF. For example, over several release points, one would expect that sometimes the true GWTT is greater than the calculated median, sometimes less than the median, and that on the average, the true GWTT falls somewhere within the full range of the calculated GWTTs. If this is not the case, we can have the following situations:

1. There can be a systematic bias, for example, the true GWTT always or nearly always being greater (or less) than the median GWTT. If a bias exists and the tendency is to overestimate the true GWTT, then this should be noted, as the method will be considered "noncautious." In a PA context, producing cautious estimates is more desirable than producing noncautious ones because the error is on the side of greater protection of public health. The degree of bias is assessed by two numbers: the magnitude of the error and a measure of the degree of caution in the estimates.

2. Independent of any bias, the calculated range of GWTT can be much too wide or much too narrow. In the former case the true GWTT generally falls only in a limited portion of the estimated GWTT CDF. A method which consistently does this is said to overpredict the uncertainty. In the latter case the true GWTT does not, in general, fall within the calculated range; such a method is said to underpredict the uncertainty. If the method neither overpredicts nor underpredicts the uncertainty, it is considered to be "self-consistent."

To evaluate how well a method performs, we assess the magnitude of the error between the estimated CDF and the true value, quantify the degree of spread in the distribution, and perform a bootstrap confidence interval test to evaluate the robustness and self-consistency of the method. In addition, for GWTT, we compute a measure which quantifies the degree of conservatism in the estimates. For reasons described later on, the robustness and spread measures must be evaluated jointly, and are thus combined into a single measure, which is referred to as the "normalized self-consistency" measure. We refer to these calculations as performance measure calculations which are computed for the GWTT and PATH CDFs at each release point. Then we compute the evaluation measures by averaging the performance measures across all release points in each test problem.

Two approaches were used to evaluate the GWTT performance of the methods, the fixed well and the random well cases described in the text. The first compares the distribution of simulated GWTT values from each release point to the single, known value; the second involves a comparison of the estimated distribution for particles released from within a designated area to the true distribution of GWTTs for that area. Hereinafter, the designations "true field," "true travel time," or "true distribution" refer to quantities computed using the exhaustive (synthetic) data set based on the reference model. All of the evaluation measures were constructed such that the target value is zero, that is, the closer the measure value is to zero, the better the performance. The reason for this was to provide a consistent target value for each evaluation measure and to aid in transforming the computed measures into consistent units (e.g., ranks) for use in averaging across the measures. The measures are described below.

C1. GWTT Analyses: Fixed Well Approach

Because some of the release points were located close to observation points and others were not, the performance measures computed at each release point were weighted accordingly (i.e., release points placed in close proximity to observation points were assigned larger weights). Thus, in the formulations that follow, the evaluation measures are presented as a weighted average over the number of release points, of the performance measures. If a method did not produce a GWTT for a given release point (e.g., the particle reached the edge of the model domain at a distance less than 5 km), then this release point was not considered, and the weights of the remaining release points were adjusted accordingly such that they still summed to one. In other words, no penalty is applied for skipping a release point. In Table 4 the number of release points used by each method in each test problem is given. The rationale and formulations involved in the weighting of the release points are as follows.

C1.1. Weighting of release points. Each true groundwater flow path was discretized into as many points as there were grid cells intercepted by the path in the true-solution model (in general, on the order of 300 points). At each of these points, kriging was performed using the locations of all observation locations in the field, both transmissivity and head data, indistinctly. A linear variogram model, given by $\gamma(\xi) = \xi$, was used for the kriging. The mixture of head and T data in the kriging was inconsequential as only the kriging variances were kept; the kriged estimates were discarded. The arithmetic average of the kriging variances was computed for all the points falling along the path. The weighting factor assigned to each release point was then calculated as being proportional to the inverse

of this average kriging variance, normalized so that the sum of these weights equals one.

This weighting approach was selected because the kriging variance is only a function of the location of the data points and not of the actual measured values. Generally speaking, this variance increases when the point to be estimated becomes further away from the observation locations. The weights, being inversely proportional to the kriging variance, represent a "measure of the distance" between the particle path and the observation points. Mixing together the T and head data (i.e., using the information regarding their locations) represents the simplest assumption on the worth of the data, and using a linear variogram does not give a limited range of influence to any observation location (a variogram with a sill would have made all points beyond the range appear to be at the same distance). The weights were constructed as

$$W_i = s \sum_{j=1}^{m} \left(\frac{1}{\sigma_K^2} \right)_j \tag{2}$$

where σ_K^2 is the kriging variance in cell j of the true field, m is the number of cells intercepted by particle path i, and s is a scale factor chosen such that

$$\sum_{i=1}^{\text{nrp}} W_i = 1.0 \tag{3}$$

where nrp is the number of release points.

C1.2. GWTT CDF error measure. The GWTT error measure quantifies the discrepancy between the median GWTT and the true GWTT, in $\log_{10}$ space and in absolute value. For each TP and method, it is given by

$$\text{GWTT error} = \sum_{i=1}^{\text{nrp}} W_i \left[\frac{|\log(\text{median GWTT})_i - \log(\text{true GWTT})_i|}{\frac{1}{M}\sum_{j=1}^{M} [\log(\text{GWTT}_{0.975}) - \log(\text{GWTT}_{0.025})]_{ij}} \right] \tag{4}$$

where nrp is the number of release points, M is the number of methods being compared, and GWTT_x is the xth quantile of the log (GWTT) CDF. The evaluation measure was constructed so that the same error measure value will be computed regardless of whether the median and true GWTT values are, respectively, 1000 and 2000, 2000 and 1000, or 10,000 and 20,000. That is, the error measure is independent of where the true travel time falls on the timescale and of whether or not it is to the left or the right of the median GWTT. This allows the evaluation measures to be averaged over the release points (for each TP, each of which has a different true GWTT) and does not discriminate between underestimation and overestimation of the GWTT.

The denominator represents an average of the spread in the distributions across all methods for each TP. This normalization was chosen so that these error-measure values could be compared across release points and test problems where the reference GWTT can differ substantially. It also provides a constant divisor, common to all methods, that facilitates an objective comparison among them.

C1.3. GWTT CDF spread measure. The GWTT spread measure is given by

$$\text{GWTT spread} = \sum_{i=1}^{\text{nrp}} W_i [\log(\text{GWTT}_{0.975})_i - \log(\text{GWTT}_{0.025})_i \tag{5}$$

where nrp is the number of release points and W_i are the release point weights, as before.

C1.4. GWTT CDF robustness measure. As noted previously, the bootstrap test is a measure of the robustness of the method and, together with the spread measure, provides an indication of the self-consistency of the method. The GWTT bootstrap measure is defined as

$$\text{GWTT bootstrap} = \frac{\left| 0.95 - \frac{\text{nci}}{\text{nrp}} \right|}{0.95} \tag{6}$$

where nci is the number of times the true GWTT or PATH fell within the 0.025 and 0.975 quantiles of the GWTT or PATH CDFs and nrp is the number of release points (CDFs).

C1.5. GWTT CDF degree of caution measure. The GWTT degree of caution measure was considered important from a PA standpoint. It is a binary measure given by

$$\text{GWTT degree of caution} = \sum_{i=1}^{\text{nrp}} W_i \text{NC}_i \tag{7}$$

$$\text{NC}_i = \begin{cases} 1 & \text{if } Q_{\text{true}} < 0.20 \\ 0 & \text{otherwise} \end{cases}$$

where Q_{true} is the quantile of the estimated GWTT CDF corresponding to the true GWTT. If the CDF for a release point is noncautious (80% of the CDF exceeds the true GWTT), score a 1 (undesirable); if it is not noncautious, score a zero (desirable). The measure was formulated and the description was phrased in this manner to point out a subtle distinction: Caution by itself is not particularly desirable, but being noncautious is definitely undesirable.

If the weighted average of the NC scores is close to 1, it means the method generally produces noncautious solutions. When associated with GWTT error, it is a measure of the bias, but centered on the 0.20 quantile, not on the median. The selection of the 0.20 quantile is arbitrary, but reflects the belief that a method that overpredicts the GWTT 80% of the time is noncautious.

C1.6. PATH CDF error measure. The PATH error measure quantifies the absolute deviation between the median path direction angle and the direction of the true path. The orientation of the path is defined by the angle (α, in degrees, between $-180°$ and $+180°$, 0° being east) from the release point to the point where the path crosses a circle of radius 5 km (centered at the release point). As with GWTT, these errors are normalized by the average spread in the path direction errors across all methods in order to enable the magnitude of this measure to be comparable across test problems (the variance in the spread of the PATH CDFs will vary with the test problem due to the different types of hydrogeologic features and different degrees of heterogeneity represented in the different test problems). The PATH error measure is given for each TP by

$$\text{PATH error} = \sum_{i=1}^{\text{nrp}} W_i \left[\frac{\text{median } \theta_i}{\frac{1}{M}\sum_{j=1}^{M} [\theta_{0.975} - \theta_{0.025}]_{ij}} \right] \tag{8}$$

$$0 \le \theta \le 180$$

where θ is the magnitude of particle path direction error in degrees, θ_x is the xth quantile of the CDF of path direction errors, and M is the number of methods being compared. $\theta_i = |\alpha_i - \alpha_t|$, where α_i is the estimated particle path direction and α_t is the true path direction.

C1.7. PATH CDF spread measure. The path spread measure was formulated in terms of the actual angular orientation α rather than the magnitude of the error θ, as the latter depends on where the true α falls. Using θ for comparing the spread in the distributions could be misleading. For example, two distributions, each spanning 90°, will have different θ-based spread measures if true direction falls at the median of one distribution and the 0.01 quantile of the other distribution. This measure is designed to be independent of the error in the path direction.

$$\text{PATH spread} = \sum_{i=1}^{\text{nrp}} W_i[\alpha_{0.975} - \alpha_{0.025}]_i \qquad (9)$$

$$-180 \le \alpha \le +180$$

where α_x is the xth quantile of the CDF of angular groundwater flow path directions.

C1.8. PATH CDF robustness measure. The PATH CDF robustness measure was constructed identically to that for the GWTT CDFs:

$$\text{PATH bootstrap} = \frac{\left| 0.95 - \frac{\text{nci}}{\text{nrp}} \right|}{0.95} \qquad (10)$$

where nci refers to the number of times the true path direction falls within the 0.025 and 0.975 quantiles of the CDF of path directions.

C1.9. The "normalized self-consistency" measures. After careful examination of the evaluation measure scores, it was determined that the spread and bootstrap measures cannot be judged independently of each other. A distribution with a narrow spread, on the outset, appears desirable because it indicates little uncertainty. However, if the true GWTT rarely falls within the CDF bounds (a failure of the bootstrap test), the performance is deemed unsatisfactory. Conversely, if a method consistently satisfies the bootstrap interval bounds, but also consistently produces CDFs with a very large spread, this too is undesirable. From a PA viewpoint it is better to have a robust method where the range of estimates nearly always contains the true value than to produce narrow distributions which fail to capture the true value a significant portion of the time. This way of thinking was encoded by combining the spread and bootstrap measures into a single "normalized self-consistency" evaluation measure, denoted NSC, and formulated as

$$\text{NSC} = \frac{3 \cdot \text{Boot} + \text{Sprd}}{4} \qquad (11)$$

In this formula the bootstrap score (Boot) and the spread measure score (Sprd) are first converted to consistent units via rank transformation.

If the range of the transformed Boot and Sprd measures was [0, 1] (the rank-transform range is [1, 7]), then CDFs that were spikes always landing on the true value would yield NSC = 0, broad CDFs never containing the true value would lead to NSC = 1, and the overestimation and underestimation of uncertainty cases would lead to NSC = 0.25 and NSC = 0.75 respectively. This measure was constructed for both the GWTT and PATH analyses. The 3-to-1 weighting of the two measures was determined by successive trials of different weights along with comparisons and discussions of the merits of the CDFs produced in this test problem exercise.

C2. GWTT Analyses: Random Well Approach

For each TP and each method, in addition to computing the mean GWTT CDFs by averaging over the CDFs derived from each of the simulated flow fields, a bounding envelope containing the inner 95% of the CDF curves at each travel time value was constructed. These $\text{GWTT}_{0.025}$ and $\text{GWTT}_{0.975}$ bounding curves reflect the degree of variability in GWTT within the repository area from realization to realization. To distinguish these results from the fixed release points CDFs, we will refer to these CDFs as "the waste panel CDFs."

Comparison of pathlines was not deemed necessary for the random well case because, as noted earlier, the regulations do not mandate treatment of where the contamination occurs, only how much reaches the accessible environment. Also, the repository is located in the area of greatest data density so that there is much less uncertainty associated with capturing the general flow directions.

The evaluation measures attempt to quantify the deviation from the true CDF, the spread among the CDF estimates, and the robustness of the methods, as follows (see Figure 17). For GWTT error the area between the mean CDF and the true CDF was compared among the methods. Similarly, for the GWTT spread measure, the area between the two bounding envelopes was used to rank the methods. This area represents the degree of uncertainty in the estimate of the waste panel CDF. The GWTT robustness measure was computed by determining the proportion of the true CDF that lies within the bounding envelope of CDF curves and subtracting that from 1.0. As in the fixed-release points case, the spread and bootstrap measures are, after conversion to rank values, combined into a single normalized, centralized spread (NSC) measure as NSC = (3 · Boot + Sprd)/4.

C3. Field Variable Analyses

Inverse methods produce simulations of the entire transmissivity field conditioned on transmissivity and head data at a few observation points. T and head are related through the governing flow equation. The $\log_{10}(T)$ error measure was intended to quantify, in some average sense, how well the realizations reproduced the true transmissivity field. Because solute transport is a function of both $\log_{10}(T)$ variability and head gradients, a similar global measure of deviation from the true head field was also used as a performance measure. Because these measures are of such a global nature (a single scalar to quantify the degree of correspondence between true and estimated spatially variable quantities), they serve more as indicators of method performance than measures which can be compared to two decimal places. This is one of the reasons all of the measures were converted to rank values.

C3.1. Head and $\log_{10}(T)$ error measures. The global measure of error was computed as a weighted average of the absolute differences between the true field values and the values from the grid blocks of the participant's model, averaged across all realizations produced by the method. It is computed as

head or $\log_{10}(T)$ error

$$= \frac{1}{\mathrm{NR}\cdot\mathrm{NG}} \sum_{i=1}^{\mathrm{NR}} \sum_{j=1}^{\mathrm{NG}} W_j(\mathrm{TRUE}_j - \mathrm{MTHD}_{ij}) \tag{12}$$

where NR is the number of realizations produced by the method, NG is the number of grid blocks in the participant's grid, W_j is a weight associated with grid block j, TRUE_j is the value of $\log_{10}(T)$ or head from the true solution for the participant's grid block j, and MTHD_{ij} is the value of $\log_{10}(T)$ or head from the inverse solution of a method in grid block j of simulation i. Because the true solution was solved on a much finer grid than any of the participant's models, the true value, TRUE_j, corresponding to a grid block of the participant's model is given by an area-weighted average of the true-grid block values contained within the participant's grid block j,

$$\mathrm{TRUE}_j = \frac{1}{\sum_{k=1}^{\mathrm{NT}} A_k} \sum_{k=1}^{\mathrm{NT}} A_k \cdot \mathrm{TRUE}_k \tag{13}$$

where NT is the number of true-grid blocks having any portion overlapping with grid block j of the participant's model and A_k is the amount of area of true-grid block k which overlaps with the participant's grid block j. The weights, W_j in (12), were developed to account for the proximity of the observation points to the grid block where the error is being evaluated, in a similar way to that of the weights for the particle pathlines from the fixed release points case. The weights, identical for head and $\log_{10}(T)$, were computed as

$$W_j = \mathrm{Max}(\sigma_K) - \sigma_K(j) \tag{14}$$

where $\sigma_K(j)$ is the kriging standard deviation for cell j, using a linear semivariogram model and all data, both head and T, and $\mathrm{Max}(\sigma_K)$ is the maximum kriging standard deviation over all meshes j. There was no constraint that the sum of the weights be equal to unity.

C3.2. *T*-field correlation structure measure. Semivariogram estimates of the simulated $\log_{10}(T)$ fields were computed for each realization of a method using the GSLIB GAMV2M routine [*Deutsch and Journel*, 1992]. On the order of 600 to 1000 randomly placed sampling points were used in the estimation of each semivariogram. Then the average semivariogram was computed across the ensemble of realizations for each TP and each method. Finally, estimates of the parameters of an exponential semivariogram model fit to each of the average empirical semivariograms were made via nonlinear regression. The same analysis was performed on the single true $\log_{10}(T)$ field realization for each test problem.

The regression estimates of the sill, σ^2, and the correlation length parameter λ led to two geostatistical performance measures characterizing the correlation structure of the ensemble of fields produced by each method. The sill and correlation length evaluation measures, denoted J_{σ^2} and J_λ respectively, were constructed identically as

$$J_{\sigma^2} = 1 - \frac{\sigma_T^2}{\sigma_T^2 + |\sigma_i^2 - \sigma_T^2|} \tag{15}$$

$$J_\lambda = 1 - \frac{\lambda_T}{\lambda_T + |\lambda_i - \lambda_T|} \tag{16}$$

where subscript T denotes true field value and subscript i denotes the value from method i.

Just as the GWTT spread and bootstrap measures cannot be interpreted independently, neither can these two measures. It was decided that for PA purposes, it is more important to capture the correlation scale of the $\log_{10}(T)$ process than to match the variance in the distribution of true $\log_{10}(T)$ field values. This is particularly evident in TP 3, where the high-T channels lead to a very short correlation length. As a result of this thinking, the J_{σ^2} and J_λ evaluation measures were combined into a single geostatistical structure evaluation measure, J_γ, defined as

$$J_\gamma = \frac{3 \cdot J_\lambda + J_{\sigma^2}}{4}. \tag{17}$$

The ability of an inverse method to reproduce correctly the correlation structure of the T-field in the realizations was considered an important feature for predicting contaminant transport and spreading.

Acknowledgments. The test problem coordinator and first author would like to express gratitude for the help and constructive input provided by all the participants, design committee members, and others who attended the GXG meetings, and in particular to the following: Charlie Cole and Harlan Foote of Pacific Northwest Laboratory, who provided the multigrid solver used to obtain the solutions to the "true" fields; Anil Mishra of New Mexico Tech for helping with the numerical pumping in TP's 3 and 4; Randy Roberts and Paul Domski of Duke Engineering and Services, Inc. for their analyses of the aquifer test data; Hamilton Link of the University of Oregon who developed numerous codes used to carry out the comparison analyses; and Sean McKenna and Erik Webb of Sandia National Laboratories, Albuquerque, New Mexico for their extremely helpful review comments. Part of this paper was written while the second author, who also chaired the GXG, was spending a sabbatical at Stanford University in the Department of Geological and Environmental Science. The support of Stanford University is greatly appreciated. This work was funded by Sandia National Laboratories in the framework of research programs to assist in the development of the PA methodology for evaluation of the WIPP site on behalf of the Department of Energy.

References

Ahmed, S., and G. de Marsily, Cokriged estimation of aquifer transmissivity as an indirect solution of the inverse problem: A practical approach, *Water Resour. Res.*, *29*(2), 521–530, 1993.

Beaubeim, R. L., Identification of spatial variability and heterogeneity of the Culebra Dolomite at the Waste Isolation Pilot Plant site, in *Proceedings: NEA Workshop on Heterogeneity of Groundwater Flow and Site Evaluation*, *Paris*, *France*, *22–24 October 1990*, pp. 131–142, Nucl. Energy Agency, Org. for Econ. Coop. Dev., Paris, 1991.

Capilla, J. E., J. J. Gómez-Hernández, and A. Sahuquillo, Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, 2, Demonstration on a synthetic aquifer, *J. Hydrol.*, *203*, 175–188, 1997.

Carrera, J., State of the art of the inverse problem applied to the flow and solute transport equations, in *Groundwater Flow and Quality Modelling*, *NATO ASI Ser.*, vol. 224, pp. 549–585, Kluwer, Norwell, Mass., 1988.

Carrera, J., and L. Glorioso, On geostatistical formulations of the groundwater flow inverse problem, *Adv. Water Resour.*, *14*(5), 273–283, 1991.

Carrera, J., and A. Medina, An improved form of adjoint-state equations for transient problems, in *Computational Methods in Water Resources X*, pp. 199–206, Kluwer, Norwell, Mass., 1994.

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

Carrera, J., and S. P. Neuman, Estimation of aquifer parameters under

transient and steady state conditions, 2, Uniqueness, stability, and solution algorithms, *Water Resour. Res*., *22*(2), 211–227, 1986b.

Carrera, J., A. Medina, and G. Galarza, Groundwater inverse problem: Discussion on geostatistical formulations and validation, *Hydrogéologie*, *4*, 313–324, 1993.

Cauffman, T. L., A. M. LaVenue, and J. P. McCord, Ground-water flow modeling of the Culebra Dolomite, vol. II, Data base, *SAND89-7068/2*, Sandia Natl. Lab., Albuquerque, N. M., 1990.

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.

Cooley, R. L., A method of estimating parameters and assessing reliability for models of steady state groundwater flow, 1, Theory and numerical properties, *Water Resour. Res*., *13*(2), 318–324, 1977.

Cooley, R. L., A method of estimating parameters and assessing reliability for models of steady state groundwater flow, 2, Application of statistical analysis, *Water Resour. Res*., *15*(3), 603–617, 1979.

Cooley, R. L., Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 1, Theory, *Water Resour. Res*., *18*(4), 965–976, 1982.

Cooley, R. L., Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 2, Applications, *Water Resour. Res*., *19*(3), 662–676, 1983.

Copty, N., Y. Rubin, and G. Mavko, Geophysical-hydrological identification of field permeabilities through Bayesian updating, *Water Resour. Res*., *29*(8), 2813–2825, 1993.

Dagan, G., Stochastic modeling of groundwater flow by unconditional and conditional probabilities: The inverse problem, *Water Resour. Res*., *21*(1), 65–72, 1985.

Dagan, G., *Flow and Transport in Porous Formations*, 465 pp., Springer-Verlag, New York, 1989.

Dagan, G., and Y. Rubin, Stochastic identification of recharge, transmissivity and storativity in aquifer transient flow: A quasi-steady approach, *Water Resour. Res*., *24*(10), 1698–1710, 1988.

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

de Marsily, G., De l'identification des systèmes en hydrogeologiques (tome 1), Ph.D. thesis, pp. 58–130, L'Univ. Pierre et Marie Curie–Paris VI, Paris, 1978.

de Marsily, G., G. Lavedan, M. Boucher, and G. Fasanion, Interpretation of interference tests in a well field using geostatistical techniques to fit the permeability distribution in a reservoir model, in *Geostatistics for Natural Resources Characterization 2nd NATO Advanced Study Institute*, *South Lake Tahoe*, *CA*, *September 6–17, 1987*, part 2, edited by G. Verly et al., pp. 831–849, D. Reidel, Norwell, Mass., 1984.

Desbarats, A. J., and R. M. Srivastava, Geostatistical simulation of groundwater flow parameters in a simulated aquifer, *Water Resour. Res*., *27*(5), 687–698, 1991.

Dettinger, M. D., and J. L. Wilson, First order analysis of uncertainty in numerical models of groundwater flow, 1, Mathematical development, *Water Resour. Res*., *17*(1), 149–161, 1981.

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

Dietrich, C. R., and G. N. Newsam, Sufficient conditions for identifying transmissivity in a confined aquifer, *Inverse Prob*., *6*(3), L21–L28, 1990.

Ginn, T. R., and J. H. Cushman, Inverse methods for subsurface flow: A critical review of stochastic techniques, *Stochastic Hydrol. Hydraul*., *4*(1), 1–26, 1990.

Gómez-Hernández, J. J., and A. G. Journel, Joint sequential simulation of multi-gaussian fields, in *Geostatistics Troia '92*, vol. 1, edited by A. Soares, pp. 85–94, Kluwer Acad., Norwell, Mass., 1993.

Gómez-Hernández, J. J., A. Sahuquillo, and J. E. Capilla, Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, 1, Theory, *J. Hydrol.*, *203*, 162–174, 1997.

Gonzalez, R. V., M. Giudici, G. Ponzini, and G. Parravicini, The differential system method for the identification of transmissivity and storativity, *Transp. Porous Media*, *26*, 339–371, 1997.

Grindrod, P., and M. D. Impey, Fractal field simulations of tracer migration within the WIPP Culebra Dolomite, Intera Inf. Technol., Henley-upon-Thames, U. K., Dec. 1991.

Gutjahr, A. L., and J. R. Wilson, Co-kriging for stochastic flow models, *Transp. Porous Media*, *4*(6), 585–598, 1989.

Gutjahr, A., B. Bullard, S. Hatch, and L. Hughson, Joint conditional simulations and the spectral method approach for flow modeling, *Stochastic Hydrol. Hydraul*., *8*(1), 79–108, 1994.

Harvey, C. F., and S. M. Gorelick, Mapping hydraulic conductivity: Sequential conditioning with measurements of solute arrival time, hydraulic head and local conductivity, *Water Resour. Res*., *31*(7), 1615–1626, 1995.

Hoeksema, R. J., and P. K. Kitanidis, An application of the geostatistical approach to the inverse problem in two-dimensional groundwater modeling, *Water Resour. Res*., *20*(7), 1003–1020, 1984.

Hsu, K., D. Zhang, and S. P. Neuman, Higher-order effects on flow and transport in randomly heterogeneous porous media, *Water Resour. Res*., *32*(3), 571–582, 1996.

Hyndman, D. W., J. M. Harris, and S. M. Gorelick, Coupled seismic and tracer test inversion for aquifer property characterization, *Water Resour. Res*., *30*(7), 1965–1977, 1994.

Johnson, R. A., and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 594 pp., Prentice-Hall, Englewood Cliffs, N. J., 1982.

Journel, A. G., and C. J. Huijbregts, *Mining Geostatistics*, Academic, San Diego, Calif., 1978.

Keidser, A., and D. Rosbjerg, A comparison of four inverse approaches to groundwater flow and transport parameter identification, *Water Resour. Res*., *27*(9), 2219–2232, 1991.

Kitanidis, P. K., and R. W. Lane, Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss-Newton method, *Hydrol*., *79*(1–2), 53–71, 1985.

Kitanidis, P. K., and E. G. Vomvoris, A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations, *Water Resour. Res*., *19*(3), 677–690, 1983.

Koltermann, C. E., and S. M. Gorelick, Heterogeneity in sedimentary deposits: A review of structur-imitating, process-imitating, and descriptive approaches, *Water Resour. Res*., *32*(9), 2617–2658, 1996.

Kuiper, L. K., A comparison of several methods for the solution of the inverse problem in two-dimensional steady state groundwater flow modeling, *Water Resour. Res*., *22*(5), 705–714, 1986.

Lappin, A. R., Summary of site-characterization studies conducted from 1983 through 1987 at the Waste Isolation Pilot Plant (WIPP) site, southeastern New Mexico, *SAND88-0157*, Sandia Natl. Lab., Albuquerque, N. M., 1988.

LaVenue, A. M., and J. F. Pickens, Application of a coupled adjoint-sensitivity and kriging approach to calibrate a groundwater flow model, *Water Resour. Res*., *28*(6), 1543–1569, 1992.

LaVenue, A. M., B. S. RamaRao, G. de Marsily, and M. G. Marietta, Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields, 2, Application, *Water Resour. Res*., *31*(3), 495–516, 1995.

Mackay, R., A study of the effect of the extent of site investigation on the estimation of radiological performance: Overview, *DoE/HMIP/RR/93.053*, 28 pp., UK Dep. of the Environ., Her Majesty's Insp. of Pollut., London, 1993.

Mandelbrot, B. B., *The Fractal Geometry of Nature*, 468 pp., W. H. Freeman, New York, 1983.

Mantoglou, A., and J. L. Wilson, The turning bands method for simulation of random fields using line generation by a spectral method, *Water Resour. Res*., *18*(5), 1379–1394, 1982.

Matheron, G., The intrinsic random functions and their applications, *Adv. Appl. Prob*., *5*(3), 439–468, 1973.

McKenna, S. A., and E. P. Poeter, Field example of data fusion in site characterization, *Water Resour. Res*., *31*(12), 3229–3240, 1995.

McLaughlin, D., and L. R. Townley, A reassessment of the groundwater inverse problem, *Water Resour. Res*., *32*(5), 1131–1161, 1996.

Neuman, S. P., and S. Orr, Prediction of steady state flow in nonuniform geologic media by conditional moments: Exact nonlocal formalism, effective conductivities, and weak approximation, *Water Resour. Res*., *29*(2), 341–364, 1993.

Paleologos, E. K., S. P. Neuman, and D. Tartakovsky, Effective hydraulic conductivity of bounded, strongly heterogeneous porous media, *Water Resour. Res*., *32*(5), 1333–1341, 1996.

Peck, A., S. Gorelick, G. de Marsily, S. Foster, and V. Kovalevsky, Consequences of spatial variability in aquifer properties and data limitations for groundwater modelling practice, *IAHS Publ. 175*, 272 pp., 1988.

RamaRao, B. S., A. M. LaVenue, G. de Marsily, and M. G. Marietta, Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields, 1, Theory and computational experiments, *Water Resour. Res*., *31*(3), 475–493, 1995.

Robin, M. J. L., A. L. Gutjahr, E. A. Sudicky, and J. L. Wilson,

Cross-correlated random field generation with the direct Fourier transform method, *Water Resour. Res*., *29*(7), 2385–2397, 1993.

Roth, C., J. P. Chiles, and C. de Fouquet, Adapting geostatistical transmissivity simulations to finite difference flow simulators, *Water Resour. Res*., *32*(10), 3237–3242, 1996.

Rubin, Y., Prediction of tracer plume migration in disordered porous media by the method of conditional probabilities, *Water Resour. Res*., *27*(6), 1291–1308, 1991a.

Rubin, Y., Transport in heterogeneous porous media: Prediction and uncertainty, *Water Resour. Res*., *27*(7), 1723–1738, 1991b.

Rubin, Y., and G. Dagan, Stochastic identification of transmissivity and effective recharge in steady groundwater flow, 1, Theory, *Water Resour. Res*., *23*(7), 1185–1192, 1987.

Rubin, Y., and G. Dagan, Conditional estimation of solute travel time in heterogeneous formations: Impact of transmissivity measurements, *Water Resour. Res*., *28*(4), 1033–1040, 1992.

Rubin, Y., and A. J. Journel, Simulation of non-Gaussian space random functions for modeling transport in groundwater, *Water Resour. Res*., *27*(7), 1711–1721, 1991.

Rubin, Y., G. Mavko, and J. Harris, Mapping permeability in heterogeneous aquifers using hydrologic and seismic data, *Water Resour. Res*., *28*(7), 1809–1816, 1992.

Sahuquillo, A., J. E. Capilla, J. J. Gómez-Hernández, and J. Andreu, Conditional simulation of transmissivity fields honoring piezometric data, in *Hydraulic Engineering Software IV*, *Fluid Flow Modeling*, vol. 2, edited by Blain and Cabrera, pp. 201–214, Elsevier Sci., New York, 1992.

Sandia National Laboratories, Preliminary comparison with 40 CFR Part 191, Subpart B for the Waste Isolation Pilot Plant, December 1991, vol. 1, Methodology and results, *SAND91-0893/1*, Albuquerque, N. M., 1991.

Sandia National Laboratories, Preliminary performance assessment for the Waste Isolation Pilot Plant, December, 1992, vol. 1, Third comparison with 40 CFR 191, Subpart B, *SAND92-0700/1*, Albuquerque, N. M., 1992.

Steel, R. G. D., and J. H. Torrie, *Principles and Procedures of Statistics: A Biometrical Approach*, 2nd ed., 633 pp., McGraw-Hill, New York, 1980.

Sun, N. Z., *Inverse Problems in Groundwater Modeling*, 337 pp., Kluwer Acad., Norwell, Mass., 1994.

Sun, N. Z., and W. W. G. Yeh, A stochastic inverse solution for transient groundwater flow: Parameter identification and reliability analysis, *Water Resour. Res*., *28*(12), 3269–3280, 1992.

Townley, L. R., and J. L. Wilson, Computationally efficient algorithms for parameter estimation and uncertainty propagation in numerical models of groundwater flow, *Water Resour. Res*., *21*(12), 1851–1860, 1985.

Tsang, Y. Y. W., Stochastic continuum hydrological model of Äspö for the SITE-94 performance assessment project, *Rep. SKI-R-96-9*, 80 pp., Swed. Nucl. Power Insp., Stockholm, 1996.

U.S. Environmental Protection Agency, 40 CFR 191: Environmental standards for the management and disposal of spent nuclear fuel, high-level and transuranic radioactive wastes: Final rule, *Fed. Regist*., *50*(82), 38,066–38,089, 1985.

Yeh, W. W. G., Review of parameter identification procedures in groundwater hydrology: The inverse problem, *Water Resour. Res*., *22*(1), 95–108, 1986.

Zimmerman, D. A., and J. L. Wilson, Description of and user's manual for TUBA: A computer code for generating two-dimensional random fields via the turning bands method, GRAM, Inc., Albuquerque, N. M., 1990.

Zimmerman, D. A., C. L. Axness, G. de Marsily, M. G. Marietta, and C. A. Gotway, Some results from a comparison study of geostatistically-based inverse techniques, in *Parameter Identification and Inverse Problems in Hydrology, Geology and Ecology*, edited by J. Gottlieb and P. DuChateau, Kluwer Acad., Norwell, Mass., 1996.

C. L. Axness, R. Beauheim, P. B. Davies, D. P. Gallegos, and M. G. Marietta, Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185-5800. (e-mail: daxness@etseccpb.upc.es; rlbeauh@sandia.gov; dpgalle@sandia.gov; mgmarie@sandia.gov)

R. L. Bras, Department of Civil Engineering, Massachusetts Institute of Technology, Water Resources and Environmental Engineering Division, Room 48-311, Cambridge, MA 02139. (e-mail: rlbras@storm.mit.edu)

J. Carrera, Universitat Politècnica de Cataluña, E.T.S.I. Caminos, Jordi, Girona 31, E-08034 Barcelona, Spain. (e-mail: carrera@etseccpb.upc.es)

G. Dagan, Department of Fluid Mechanics and Heat Transfer, Tel Aviv University, P.O. Box 39040, Ramat Aviv, Tel Aviv 69978, Israel. (e-mail: dagan@eng.tau.ac.il)

G. de Marsily, Laboratoire de Géologie Appliquée, Université Paris VI, 4 place Jussieu, 75230 Paris Cedex 05, France. (e-mail: gdm@ccr.jussieu.fr)

A. Galli, Centre de Geostatistique, Ecole de Mines de Paris, 35 rue St. Honore, 77035 Fountainebleau, France.

J. J. Gómez-Hernández, Departmento de Ingenieria Hidraulica y Medio Ambiente, Universidad Politècnica de Valencia, Camino de Vera, S/N, 46071 Valencia, Spain. (e-mail: jaime@dihma.upv.es)

C. A. Gotway, National Center for Environmental Health, Centers for Disease Control and Prevention, MS F42, 1600 Clifton Rd. NE, Atlanta, GA 30333. (e-mail: cdg7@cdc.gov)

P. Grindrod, Quantisci Ltd., Chiltern House, 45 Station Road, Henley-on-Thames, Oxfordshire, RG9 1AT, UK. (e-mail: peterg@quantisci.co.uk)

A. Gutjahr, Department of Mathematics, New Mexico Institute of Mining and Technology, Socorro, NM 87801. (e-mail: agutjahr@nmt.edu)

P. K. Kitanidis, Department of Civil Engineering, Stanford University, Terman Engineering Center, Stanford, CA 94305-4020. (e-mail: pkk@cive.stanford.edu)

A. M. Lavenue and B. S. Rama Rao, Duke Engineering and Services, Inc., 9111 Research Blvd., Austin, TX 78758. (e-mail: amlavenu@duke-energy.com; bsramara@duke-energy.com)

D. McLaughlin, Massachusetts Institute of Technology, Room 48-209, Cambridge, MA 02139. (e-mail: dennism@mit.edu)

S. P. Neuman, College of Engineering and Mines, Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721. (e-mail: neuman@hwr.arizona.edu)

C. Ravenne, Institut Français du Pétrole, Rueil-Malmaison, France.

Y. Rubin, Department of Civil Engineering, University of California, Berkeley, CA 94270. (e-mail: rubin@arimor.ce.berkeley.edu)

D. A. Zimmerman, GRAM, Inc., 8500 Menaul Blvd. NE, B-335, Albuquerque, NM 87112. (e-mail: tonyz@graminc.com)

(Received May 14, 1997; revised December 22, 1997;
accepted December 29, 1997.)