Stochastic inverse problems in groundwater modeling
J.E. Capilla, J.J. Gómez, A. Sahuquillo &
H.J.W.M. Hendricks Franssen
School of Civil Engineering, Technical University of Valencia, Camino
de Vera s/n, E-46071 Valencia, Spain.

Abstract

The inverse problem is, by its own definition, an ill-posed problem that requires assumptions about the structure of the unknown parameters. Yet, it has been traditionally addressed in a somehow deterministic way, searching for an optimum or best estimate solution that reproduces the available data. The impossibility of measuring hydrodynamic parameters and variables all over the aquifer makes groundwater-modeling results uncertain at any given conceivable situation. This fact has been the motivation to the development of stochastic hydrogeology. In this framework, hydrodynamic parameters are treated as realizations of a stochastic process conditioned to available data and characterized by a random function model inferred from the own data. In practical cases, the difficulty to solve the stochastic partial differential equations resulting from this approach leads to proceed simulating multiple equally likely conditional parameter fields, solving the classic deterministic equations, and then analyzing the ensemble of results. In order to obtain more realistic results, the stochastic conditional parameter fields have to honor the available variable measurements (usually of head and concentrations). The implementation of new methodologies to reach this goal has defined a new field generally referred to as stochastic inverse modeling.

In this paper we present some of the most advance state-of-the-art methods showing their application to a real case of risk assessment. It corresponds to a low pervious dolomite formation that might constitute the fastest pathway to the biosphere for radionuclides escaped from a nuclear waste repository.
1 Introduction

The inverse problem arises in many fields of sciences and engineering where the variables of interest are related, through a mathematical model, to physical parameters. The estimation of the model parameters, given a set of variable measurements, constitutes the inverse problem. These parameters usually describe a physical property that exhibits a certain spatial distribution. In the field of mathematical modeling in groundwater are, typically, hydraulic conductivity, transmissivity, storativity and dispersivity. However, any other parameter might also be considered as, for instance, rates in chemical reactions or sorption distribution coefficients. Even external stresses, boundary conditions or initial conditions are often treated as "parameters" to be determined given a set of field data (variable/s measurements). Field data use to be measurements of piezometric head, chemical species concentrations and flux discharges.

The determination of the hydraulic conductivity field (K) is the problem that has drawn more attention from hydrogeologists. The reason is that it is a key parameter to describe the flow field and may be highly heterogeneous. The available measurements to honor use to be piezometric data and, to a less extend, concentration data. The determination of external stresses and boundary flux or head values has an easier mathematical treatment than estimating K although presents similar conceptual problems. The same happens with other parameters whose estimation has rarely been addressed.

The inverse problem is inherently ill-posed because we want to determine the value at every point of a spatially distributed parameter (undetermined number of unknowns) given a finite set of measurements. Even when the formation domain is discretized in blocks/elements to apply a numerical method (as, for instance, Finite Differences or Finite Elements) the number of unknowns and equations will generally diverge and, besides, the temporal and spatial distribution of measurements may create many undesirable numerical problems. The general approach to overcome this difficulty is to parametrize the unknown K field (or any other parameter to be determined), very often using a geostatistical description of its spatial distribution. A model like this renders the parameter at every location (discretization block/element) given the values at a reduce set of locations/blocks/elements. The latter are computed based on obtaining the best possible approximation of data by means of least squares or maximum likelihood criteria. For a near up-to-date review of inverse methods see McLaughlin and Townley [1].

The weakness in most of classical inverse methods is that they have been conceived to yield parameter fields that exhibit a spatial variability far smoother than what can be observed in real formations. The same happens to fields obtained by geostatistical estimation techniques. Yet, in mass transport problems, reproducing realistic patterns of spatial variability may be a critical issue. The current approach to address heterogeneity reproduction is to consider physical parameters as realizations of a stochastic process. The flow and transport equations become stochastic partial differential equations and so the variables of interest (head, flows, concentrations, etc.). Although a lot of research
has been made in this field by many authors, the solution of these equations in real situations is far to be obtained. Thus the only available way to deal with heterogeneity is to simulate multiple equally probable stochastic fields of parameters, solve the equations for each of them and analyze the ensemble of multiple results. This allows to consider the heterogeneity and uncertainty of the parameters and, though in a "numerical format", renders a model of uncertainty of results.

Since long several authors have contributed with different stochastic simulation methods, see for instance Deutsch and Journel [2] and Goovaerst [3]. These methods are able to reproduce a given spatial structure and parameter measurements but not to honor the available data, which are related with the parameters through the flow and transport equations. The implementation of stochastic simulation methods that honor measurements has become recently a very active field of research and is yielding promising results. In fact, this constitutes what is referred to as "stochastic inverse modeling" or "stochastic inverse simulation". This terminology was first introduce in the field of Earth Sciences by Thorson and Claerbout [4] in relation with the inversion of data collected by seismic surveys. In the beginning of the 90's the idea started to be present in several hydrogeology papers although the first to propose a methodology in the sense that stochastic inversion has been introduced here, and able to deal with fields of high variance, were Sahuquillo et al. [5]. Since them this approach has become a very active research field in hydrogeology and also in petroleum engineering. Zimmerman et al. [6] provide a review of the state-of-the-art of conditional simulation in the context of the inverse problem along with an intercomparison of seven methods.

In this paper we briefly describe two of the most advanced state-of-the-art methods in stochastic inverse simulation, the Self-Calibrated method and the Conditional Probabilities method, introducing further future developments as well as an illustration of why stochastic modeling, compared to deterministic modeling, should be consider as a far superior tool to obtain flow and mass transport predictions. The case study corresponds to a low permeability dolomite formation, known as Culebra formation, that might constitute the fastest pathway to the biosphere in the hypothetical scenario of an escape from an existing nuclear waste repository.

2 Inverse Stochastic Modeling

2.1 The Self-Calibrated Method

Among the existing stochastic inverse simulation methods, according to the results of the intercomparison exercise carried out by the US Sandia National Laboratories Geostatistical Experts Group, see Zimmerman et al. [6], the Self-Calibrated (SC) method gives one of the best performance and, to the best of our knowledge, it was the first method developed to deal with high variance fields. Initially proposed by Sahuquillo et al. [5], is implemented and tested by Gómez-Hernández et al. [6] and Capilla et al. [7, 8].
Consider an aquifer domain discretized in a finite grid of $N (i=1, ..., N)$ square blocks. The SC method is based on perturbing stochastic logtransmissivity ($\log T$) fields, previously generated, and called seed fields, that reproduce transmissivity measurements and the geostatistical structure inferred from measured data. Being $Y^0$ a seed field, a perturbation $\delta Y$ is determined such that the $\log T$ field given by $Y^0 + \delta Y$ honors piezometric data. The perturbation field, $\delta Y$, is parameterized by means of ordinary kriging. Let it be $Y^0_i$ ($i=1, ..., N$) the seed $Y=\log T$ value at every block and $\delta Y_i$ ($i=1, ..., N$) the corresponding perturbation. Considering $m$ master locations ($j=1, ..., m$) where the perturbation has to be determined, then the perturbation is given by Eqn (1),

$$\delta Y_i = \sum_{j=1}^{m} \lambda_{ij} \delta Y_j$$

(1)

where $\lambda_{ij}$ are the kriging weights obtained by ordinary kriging using the $\log T$ variogram retain for the $\log T$ measurements, $\gamma_T$. $\delta Y$ preserves transmissivity or modifies its value in agreement with previously estimated standard deviation errors. The computation of $\delta Y_i$ at master locations is performed minimizing a penalty function that penalizes the deviations among piezometric data and the piezometric head field obtained for the perturbed transmissivity field. In order to build it, the flow equation is linearized on $\delta Y_j$ ($j=1, ..., m$), thus leading to a quadratic programming problem with 2$m$ constraints, all at master blocks. The constraints are obtained using the kriging estimates of $\log T$, $Y_{j, krig}$, and its standard deviations, $\sigma_{j, krig}$. The final $\log T$ value ($Y^0_j + \delta Y_j$) has to be in the range given by $Y_{j, krig} \pm 2\sigma_{j, krig}$. Due to the linearizations involved in the perturbation procedure it is necessary to iterate applying a relaxation factor, $\alpha$, to the accumulation of the successive perturbations. For every iteration $k$, a perturbation $\delta Y^k$ ($\delta Y^k_i$, $i=1, ..., N$) is obtained and them the perturbed field is $Y^k + \alpha \delta Y^k$ ($Y^k_i + \alpha \delta Y^k_i$, $i=1, ..., N$). The penalty function minimized in every iteration is given by eqn (2),

$$F = (h - h^*)^T W_h (h - h^*)$$

(2)

in which $h$ is a vector whose components are the head values at piezometric measurement locations (for field $Y^k+\delta Y^k$), $h^*$ is a vector containing the piezometric measurements at the latter locations and $W_h$ is a weighting matrix. Capilla et al. [8] also consider the simultaneous simulation of the $K$ field and constant head boundary conditions to honor piezometric data.

The SC method has been successfully applied in several practical situations both in 2-dimensional and 3-dimensional domains with transmissivity fields of very high variance. It has been also implemented to be used in the simulation of permeability fields that honor production data in oil reservoir modeling, see Wen et al. [10]. The possibilities of the method have been considerably expanded with its extension to transient problems where both the $K$ field and the storativity field are simulated to honor transient head data, see Hendricks Franssen et al. [11].

The method is currently being extended to simulate $K$ fields that honor head, concentration data and other possible parameter measurements. In this case the penalty function is as shown in eqn (3),
where \( \mathbf{c} \) and \( \mathbf{p} \) are vectors with concentrations and other parameters computed at measurement locations and measurements are in vectors \( \mathbf{c^*} \) and \( \mathbf{p^*} \); \( \mathbf{W}_c \) and \( \mathbf{W}_p \) are weighting matrices, and \( \Theta_c \) and \( \Theta_p \) are trade-off coefficients. The integration in time extends over a period that includes the available measurements.

### 2.2 The Conditional Probabilities method.

It is common to have a wealth of information coming from geophysical surveys (secondary data), or even from the expert knowledge (soft data), that properly used might help to obtain more realistic model predictions, see for instance Capilla et al. [12]. The importance of accounting for as much information as possible, including soft and secondary information, has been stressed by many authors and, for that purpose, several techniques have been proposed, see Deutsch and Journel [2]. These methods include, for instance, simulated annealing and probability fields simulation, and are capable to reproduce given patterns of continuity of extreme values that can be critical to determine early arrivals in mass transport predictions.

The Conditional Probabilities (CP) method, proposed by Capilla et al. [13], has been developed to simulate fields that, apart of fulfilling the requisites of fields obtained by the SC method, incorporate soft and secondary data without the need of assuming a multigaussian model. In essence, the CP method is based in defining an a priori random function model of the hydraulic conductivity (or transmissivity) field based on the conditional cumulative density functions (ccdf) at every block and the geostatistical structure of the probability fields. Then, the CP method proceeds basically as the SC although parametrizing the \( \delta \mathbf{Y} \) field differently. Eqn (1) now becomes eqn (4),

\[
\delta \mathbf{Y}_i^k = \left( \frac{dF_i}{dY} \right)^{-1} \sum_{j=1}^{m} \left( \frac{dF_j}{dY} \right) \lambda_{ij} \delta \mathbf{Y}_j
\]

where \( F_i(Y_i^k) \) is the ccdf for block \( i \), and \( \lambda_{ij} \) are the kriging weights obtained for the selected set of master points \( (j=1,\ldots,m) \) using the variogram of the probability field.

The CP method is computationally expensive compared to the SC method. However offers very important advantages because allows the integration of any type of data, preserves the sample histogram (even when it is asymmetric) and reproduces patterns of continuity observed in extreme values. So far the method has been applied to model hydraulic conductivity fields but its implementation for other variables of interest is straightforward. The computation of the ccdf's has to be perform just once before starting the simulation and so the structural analysis to retain a variogram for the probability fields. Besides, the ccdf's might be defined in a parametric or in a discrete way. The latter requires a complete definition of interpolations to sample every \( F_i(Y_i^k) \) and \( F_i^{-1}(p_i) \), being \( p_i \) the probability at block \( i \).
3 Application: Deterministic versus Stochastic analysis

The Waste Isolation Pilot Plant (WIPP) is a US Department of Energy facility to host a radioactive waste repository. It consists of an underground repository mined from a thick-bedded salt unit and the associated facilities at land surface. It is located in the Southeast of New Mexico, in an evaporite-bearing sedimentary basin and will be used for the permanent disposal of approximately 170,000 cubic meters of transuranic waste. A key component in the security assessment of the long term performance of the facility is evaluating the potential for radionuclide transport from the underground repository to the accessible environment by groundwater.

The WIPP repository is located approximately 650 meters below the land surface and among the formations above the site there are three units bearing laterally persistent water, the most transmissive being the Culebra dolomite formation. This is considered to be the principal pathway for transport of radionuclides to the biosphere in case of an inadvertent human intrusion would drill through the repository (future hypothetical scenario). In order to assess the security of the repository, the possible migration of radionuclides to the biosphere is modeled in the formation, see Capilla et al. [9]. The model covers a rectangular area known as WIPP2 of 22500 m x 30500 m, with its longer side oriented N-S, and approximately centered over the WIPP site.

The Culebra Dolomite has been intensively studied for years. The model calibrated by LaVenue et al. [14] has been the main source of data in later investigations. The Culebra formation thickness at the WIPP site region ranges from 5.5 to 11.3 m, with a mean of 7.7 m. The transmissivity values have been obtained from drill stem tests, slug tests and pumping tests at 37 locations and extend over a range of seven orders of magnitude. The mean is estimated as -5.5 log_{10} m^2/s with a standard deviation of 1.9 log_{10} m^2/s. An additional difficulty to model flow in the area is that density measurements show values over 1.1 g/l, mainly at the East of the WIPP area. Piezometric head measurements are available at 33 locations and we assume here an effective porosity of 0.16 all over the model area. The flow and transport equations are solved over a grid of 61 x 43 blocks of 500 x 500 m^2 size, accounting for the variable density in the case of flow, and assuming the simplification of modeling only the convective mechanism of transport. Figure 1 shows the log-transmissivity field obtained by ordinary kriging and the corresponding piezometric head field, with the location of measurements posted on them. On Figure 1(a) the trajectories of particles released from three given locations are represented. The locations chosen are known, from West to East, as H-2a, ALSHAFT and H-15. The travel times to the South boundary are 165, 149 and 338 Kyrs respectively.

The transport predictions given by the kriged logT field do not account for all the available data. Using the SC method, and the kriged logT as seed field, a new logT field was obtained that honors piezometric measurements. Modeling transport for this field, releasing particles from the same above locations, yields travel times to the South boundary of 82, 85 and 112 Kyrs which are noticeable lower than the above ones. Although this is still a deterministic prediction it makes clear the importance of accounting for all the available data.
Figure 1: (a) LogT field obtained by ordinary kriging, T measurement locations and trajectories of particles released from locations H-2a, ALSHAFT and H-15, and (b) corresponding piezometric head fields with piezometric measurement locations posted on it.

Figure 2: (a) Deterministic logT field honoring piezometric data and trajectories of particles released from locations H-2a, ALSHAFT and H-15, and (b) corresponding piezometric head fields.
In order to assess the uncertainty of flow and transport results the SC method was used to simulate 300 logT fields honoring piezometric data. In every case the reproduction of these data was as good as in the field shown in Figure 2(a). Figure 3 shows one of the simulated fields. As expected, it exhibits a spatial variability far higher than the deterministic fields and the trajectories tend to concentrate along preferential flowpaths. The average of the ensemble of the 300 logT fields looks very similar to the field in Figure 2(a). However, we are not interested in estimations of travel times that may represent average or median values. Instead, we want to determine how feasible or probable are logT fields leading to early arrivals to the South boundary. The analysis of the ensemble of results for the 300 equally likely fields is the only way to address this question.

Figure 3: Stochastic LogT field honoring piezometric data and trajectories of particles released from a W-E line, in the WIPP area, every 250 m.

To understand and interpret the huge volume of data provided by the 300 flow and transport problem results, we worked out different statistics and graphical representations. Figure 4 shows the mass transport results for particles released from the same points considered for the deterministic fields. With the available data the main trends in transmissivity and so in the piezometric head fields are well determined. Thus, the trajectories are all more or less localized within an area that includes those shown on Figure 2(a). However, the high heterogeneity of the stochastic fields leads to important variations in the resulting travel times. Travel time histograms on Figure 4 reveal how inaccurate and unreliable might be the result yielded by the deterministic field. If we look, for instance, at location H-2a we realize that a travel time of 13 Kyrs, the minimum value in the histogram, is as likely as the travel time 81 Kyrs given by the deterministic field. However, the histogram obtained by stochastic inverse modeling provides very useful information about the travel time, its most remarkable features being a markedly asymmetry, long upper tail, median of 86 Kyrs close to the deterministic value, lower quartile of 60 Kyrs and minimum of 13 Kyrs. The histograms for the other two releasing locations show similar
shapes and demonstrate the erratic nature of the deterministic value. In case ALSHAFT the deterministic travel time, 85 Kyrs, is again close to the median, 87 Kyrs, and in case H-15 the deterministic value, 112 Kyrs, is below the lower quartile, 130 Kyrs. The mean travel time in the histograms is always greater than the median due to the long upper tail.

![Histograms](image)

(a) H-2a, (b) ALSHAFT and (c) H-15.

**4 Conclusions**

Stochastic Inverse Modeling provides the only available way to assess the uncertainty of flow and mass transport models results in most practical situations. The estimations obtained by traditional deterministic modeling may be erratic when compared with the probability density functions that stochastic simulation, combined with the MonteCarlo method, may provide. This is due not only to the non-uniqueness of the inverse problem but to the necessity of reproducing the real patterns of spatial variability that hydrodynamic parameters use to exhibit in real formations, i.e., heterogeneity.

The current development of stochastic inverse simulation methods already allows to confront problems where hydraulic conductivity fields are not Gaussian, to integrate soft and secondary data and deal with the continuity of extreme values.

**Acknowledgements**

Thanks for partially funding this research are due to the Spanish Interministerial Comission for Science and Technology (CICYT, Spain), grant HID99-0481.
References


