ALGORITHMICALLY-DEFINED RANDOM FUNCTION MODELS

CLAYTON V. DEUTSCH
Exxon Production Research Co., P.O. Box 2189, Houston, TX 77252

Three criteria are proposed to select an appropriate stochastic simulation technique: 1) the implementation of the technique must allow realizations to be generated with a reasonable amount of human-involvement and CPU time, 2) all relevant prior information must be honored, and 3) the realizations should generate the largest space of uncertainty. These criteria are independent of the mechanism or random function (RF) underlying the stochastic simulation technique. This paper compares algorithmically-defined random function models, such as those based on simulated annealing, to analytically defined random function models, such as the multiGaussian model.

Stochastic simulation techniques based on these different random function models are presented. An example is presented which illustrates the relative advantages of algorithmically-defined and analytically defined techniques. The mathematical cleanliness and internal consistency of analytical models must be balanced against the flexibility of algorithmically-defined models.

INTRODUCTION

The generation of alternative numerical models or images of spatially-varying attributes that account for the known aspects of the spatial distribution is generally referred to as *stochastic simulation* or *stochastic imaging*. For practical problems, the stochastic realizations must be further processed by a *transfer function*, e.g., a program that simulates the mining operation or the fluid flow in an aquifer or petroleum reservoir. Generating the stochastic realizations is the first step; it is the performance of the models that matters. This paper will focus on the petroleum reservoir context; many of the conclusions and examples, however, apply to other areas of application.

In practice, the spatial distribution of porosity and permeability must be modeled using information from many sources: a limited number of good quality well data, a greater number of indirect seismic data, knowledge of the geological setting, and interpretations from a limited number of well tests. The reservoir management problem is to assess the performance of a number of alternative production scenarios.

Given sparse sampling and uncertainty in the available data there should not be a unique model of the spatial distributions of porosity and permeability. The idea behind stochastic reservoir modeling is to generate a number of alternative numerical

R. Dimitrakopoulos (ed.), Geostatistics for the Next Century, 422–435.

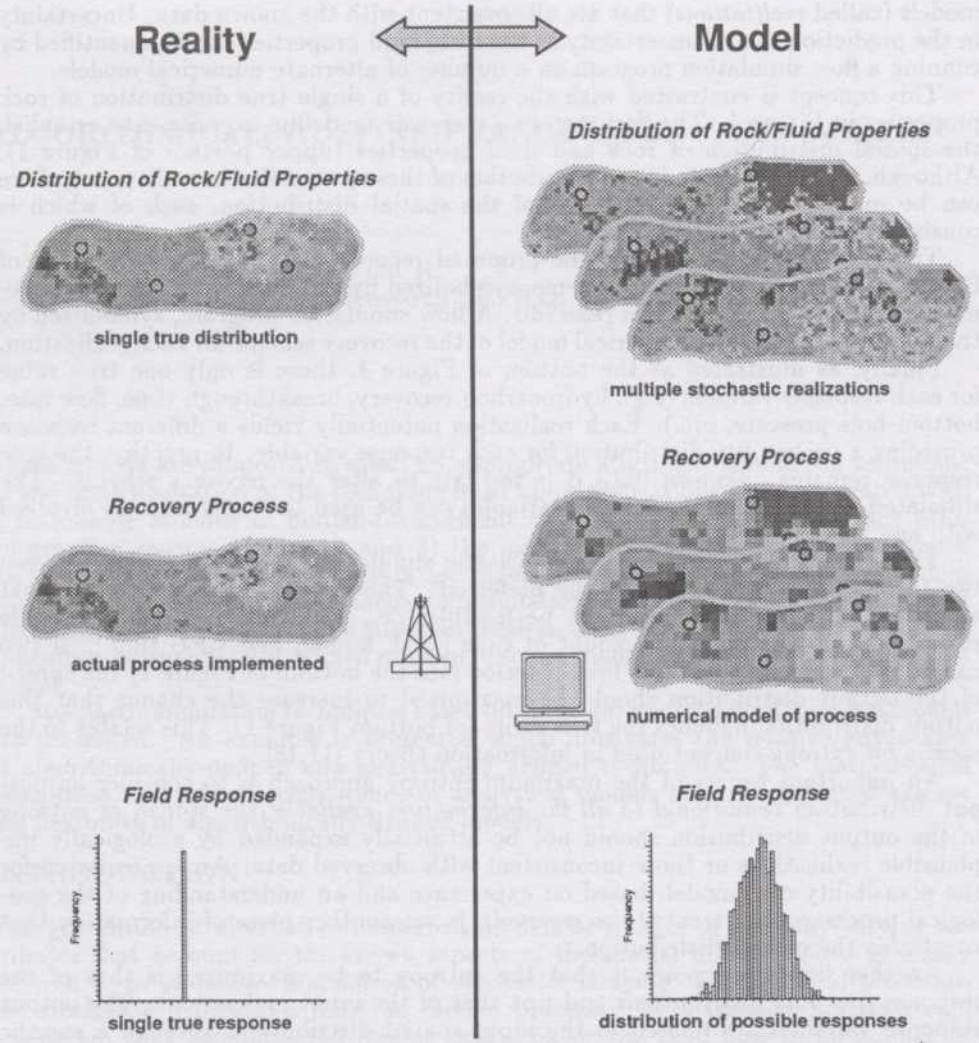

Figure 1: A schematic illustration of stochastic reservoir modeling. The first step consists of establishing the spatial distribution of rock and fluid properties. In reality, there is only one true distribution of these properties, yet, there can be many stochastic realizations of that distribution. The next step is the implementation of the recovery scheme. In reality, the recovery scheme may be implemented only once in the field. A flow simulator provides a numerical model of the recovery scheme using each alternate input model. Finally, there is only one true reservoir response, but there is a distribution of possible responses given the alternate stochastic realizations which can be generated.

models (called *realizations*) that are all consistent with the known data. Uncertainty in the prediction due to uncertainty in the rock/fluid properties can be quantified by running a flow simulation program on a number of alternate numerical models.

This concept is contrasted with the reality of a single true distribution of rock properties in Figure 1. The first step in a reservoir modeling exercise is to establish the spatial distribution of rock and fluid properties (upper portion of Figure 1). Although, there is only one true distribution of these properties in reality, yet, there can be many stochastic realizations of the spatial distribution, each of which is consistent with the available data.

The next step is to consider the proposed recovery scheme (central portion of Figure 1). The actual recovery scheme, symbolized by the drilling rig, can be implemented only once in the actual reservoir. A flow simulation program, symbolized by the computer, provides a numerical model of the recovery scheme for each realization.

Finally, as illustrated at the bottom of Figure 1, there is only one true value for each response variable (e.g., hydrocarbon recovery, breakthrough time, flow rate, bottom hole pressure, etc.). Each realization potentially yields a different response providing a probability distribution for each response variable. In practice, the true response remains unknown until it is too late to alter the recovery scheme. The simulated distributions of response variables can be used to assess the risk involved with any particular recovery scheme.

There are many techniques for stochastic simulation. Given a choice between two techniques which one should be preferred? The first practical criterion is that all potentially *good* methods must be feasible, i.e., they must generate plausible realizations in a reasonable amount of time (both human and CPU time). If two candidate techniques pass this first criterion (see the bottom of Figure 1) the spread of the output distribution should be maximized to increase the chance that this output distribution includes the true value (of bottom Figure 1). This relates to the *maximum entropy* concept used in information theory [11,18].

An important aspect of the maximum entropy approach is to consider an output distribution *conditional to all the information available*; the spread or entropy of the output distribution should not be artificially expanded by geologically implausible realizations or those inconsistent with observed data. An appreciation for the plausibility of a model, based on experience and an understanding of the geological processes that created the reservoir, is yet another piece of information that constrains the output distribution.

Another important point is that the entropy to be maximized is that of the response (output) distributions and **not** that of the input realizations. The output response variables are related to the input spatial distributions through a specific transfer function (flow simulator); however, that transfer function is usually very complex and non-linear. Even though spatial entropy of the input realizations can be defined and predicted (see [15]), in general, its relation to the entropy of the response or output distribution is not known a priori.

The spread of a response distribution is sometimes referred to as a *space of uncertainty*. The extent of that space, i.e., the uncertainty about the output is necessarily related to the model for input uncertainty. To summarize, a good technique:

1. Must generate plausible realizations in a reasonable amount of time. The time refers to the human and the CPU time required for the initial set up and the repeated application of the technique.

2. Allows the maximum prior information to be accommodated. This is the only

direct way to ensure that the output distribution is as accurate as possible.

3. Explores the largest space of uncertainty for the output, i.e., one that generates a maximum entropy distribution of response variables.

These criteria provide the basis for comparison of different simulation techniques and RF models.

RANDOM FUNCTION MODELS

A random function (RF), defined over some field of interest, e.g., $\{Z(\mathbf{u}), \mathbf{u} \in$ study area $A\}$, is characterized by the set of all its K-variate cdf's for any number K and any choice of the K locations $\mathbf{u}_k, k = 1, \ldots, K$:

$$F(\mathbf{u}_1, \ldots, \mathbf{u}_k; z_1, \ldots, z_K) = Prob\{Z(\mathbf{u}_1) \leq z_1, \ldots, Z(\mathbf{u}_K) \leq z_K\} \quad (1)$$

Just as a univariate probability distribution (cdf) may be used to characterize uncertainty about a single value $z(\mathbf{u})$, the multivariate cdf (1) may be used to characterize joint uncertainty about the K values $z(\mathbf{u}_1), \ldots, z(\mathbf{u}_K)$.

Stochastic simulation is the process of drawing alternative, equally probable, joint realizations from a RF model. The (usually gridded) realizations $\{z^{(l)}(\mathbf{u}), \mathbf{u} \in A\}$ $l = 1, \ldots, L$ represent L possible images of the spatial distribution of the attribute values $z(\mathbf{u})$ over the field A. Each realization reflects the properties imposed on the RF model $Z(\mathbf{u})$. As mentioned above, the more properties that are inferred from the sample data and incorporated into the RF model $Z(\mathbf{u})$, the better that RF model.

An analytical RF is one where the multivariate distribution is known analytically and may be written in a mathematically concise expression. An algorithmically-defined RF model is one where the multivariate probability distribution is observed by generating alternative realizations. In the case of algorithmically-defined RF models, there is no need for a mathematical definition of the distribution; the only requirement is a repeatable algorithm to generate stochastic realizations that honor the conditioning data and spatial statistics. The obvious advantage of analytically-defined RF models is that they may be studied theoretically; the mathematical consistency allows proofs, theorems, and limit (stationary and ergodic) properties to be evaluated a priori.

The best example of an analytically-defined RF model is the Gaussian RF model. Most analytically-defined RF models are related to the Gaussian model in some way.

The Gaussian RF Model

The Gaussian RF model is unique in statistics for its analytical simplicity and for being the limit distribution of many analytical theorems known as "central limit theorems" [2,12]. Some *characteristic* properties of the multivariate Gaussian (normal) RF model $Y(\mathbf{u})$ are:

- all subsets $\{Y(\mathbf{u}), \mathbf{u} \in B \subset A\}$ are also multivariate normal.
- all linear combinations of the components of $Y(\mathbf{u})$ are (univariate) normally distributed, e.g.,

$$X = \sum_{\alpha=1}^{n} \omega_\alpha Y(\mathbf{u}_\alpha) \text{ is normally distributed,}$$

$\forall$ the weights ω_α, as long as $\mathbf{u}_\alpha \in A$.

- zero covariance (or correlation) entails full independence: If $Cov\{Y(\mathbf{u}), Y(\mathbf{u}')\} = 0$, the two RV's $Y(\mathbf{u})$ and $Y(\mathbf{u}')$ are not only uncorrelated, they are independent.

- all conditional distributions of any subset of the RF $Y(\mathbf{u})$ are (multivariate) normal.

The Gaussian RF is the only analytically-defined RF model considered in this paper. The following discusses two different algorithmically-defined RF models. The indicator RF model, discussed first, has a better theoretical pedigree [13,16] than the annealing RF model, discussed last.

The Indicator RF Model

Indicator kriging of a continuous variable is *not* aimed at estimating the indicator transform

$$i(\mathbf{u}; z_k) = \begin{cases} 1, & \text{if } z(\mathbf{u}) \leq z_k \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Indicator kriging provides a least-squares estimate of the conditional cumulative distribution function (ccdf) at cutoff z_k:

$$\begin{aligned} [i(\mathbf{u}; z_k)]^* &= E\{I(\mathbf{u}; z_k|(n)\}^* \\ &= Prob^*\{Z(\mathbf{u}) \leq z_k|(n)\} \end{aligned} \tag{3}$$

where (n) represents the conditioning information available in the neighborhood of location $\mathbf{u}$.

The IK process is repeated for a series of K cutoff values $z_k, k = 1, \ldots, K$, which discretize the interval of variability of the *continuous* attribute z. The conditional cdf, built from assembling the K indicator kriging estimates represents a probabilistic (RF) model for the uncertainty about the unsampled value $z(\mathbf{u})$.

If $z(\mathbf{u})$ is a continuous variable, then the optimum selection of the cutoff values z_k at which indicator kriging takes place is essential: too many cutoff values and the inference and computation becomes needlessly tedious and expensive; too few, and the details of the distribution are lost.

In indicator kriging the K cutoff values z_k are usually chosen so that the corresponding indicator covariances $C_I(\mathbf{h}; z_k)$ are different one from another. There are cases, however, when the sample indicator covariances/variograms appear proportional to each other, i.e., the sample indicator correlograms are all similar in shape. The corresponding continuous RF model $Z(\mathbf{u})$ is the so-called "mosaic" model [14] such that:

$$\rho_Z(\mathbf{h}) = \rho_I(\mathbf{h}; z_k) = \rho_I(\mathbf{h}; z_k, z_{k'}), \ \forall z_k, z_{k'} \tag{4}$$

where $\rho_Z(\mathbf{h})$ and $\rho_I(\mathbf{h}; z_k, z_{k'})$ are the correlograms and cross correlograms of the continuous RF $Z(\mathbf{u})$ and its indicator transforms.

Indicator kriging under the mosaic model (4) is called "median indicator kriging" [13]. It is a particularly simple and fast procedure since it calls for a single easily

inferred variogram (often the median indicator variogram) that is used for all K cutoffs.

The Annealing RF Model

In the "annealing" approach to stochastic simulation there is no explicit random function model. Rather, the creation of a simulated realization is formulated as an optimization problem to be solved with a numerical optimization technique (seminal references for the application of these techniques to spatial problems include [8,9,17, 19]). The first requirement of this class of methods is the construction of an objective (or energy) function which is some measure of difference between the desired spatial characteristics and those of a candidate realization.

The global optimization technique most often used to obtain such realizations is based on an analogy with the metallurgical process of annealing. Annealing is the process by which a material undergoes extended heating and is slowly cooled. Thermal vibrations permit a reordering of the atoms/molecules to a structured lattice, i.e., a low energy state. In the context of 3-D numerical modeling, the *annealing* process may be simulated through the following steps:

1. An initial 3-D numerical model (analogous to the initial metal in true annealing) is created, for example, by assigning a random value at each grid node by drawing from the population distribution.

2. An energy function (analogous to the Gibbs free energy in true annealing) is defined as a measure of difference between desired spatial features and those of the realization. For example, the energy or objective function could be the sum of the squared difference between the variogram of the realization and a model variogram over a predefined set of lag distances.

3. The image is perturbed, for example, by swapping pairs or sets of values taken at random locations in the 3-D numerical model (this mimics the thermal vibrations in true annealing).

4. The perturbation (thermal vibration) is always accepted if the energy is decreased; it is accepted with a certain probability if the energy is increased (the Boltzmann probability distribution of true annealing). Technically, the name "simulated annealing" applies only when the acceptance probability is based on the Boltzmann distribution [1,17] Through common usage, however, the name "annealing" is used to describe the entire family of methods that are based on this optimization principle.

5. Continue the perturbation procedure while reducing the probability with which unfavorable swaps are accepted (lower the temperature parameter of the Boltzmann distribution) until a low energy state is achieved.

6. Low energy states correspond to plausible numerical models (realizations).

At first glance this approach appears terribly inefficient; millions of perturbations may be required to obtain an image having the desired spatial structure. These methods, however, are more efficient than they might seem as long as only a few arithmetic operations are required to update the objective function after each perturbation. Virtually all conventional spatial statistics (e.g., covariances/correlations)

may be updated locally (considering only a few locations) rather than recalculated globally (considering all locations).

The objective function is defined as some measure of difference between a set of reference properties and the corresponding properties of a candidate realization. The reference properties could consist of any *quantified* geological, statistical, or engineering property. Some examples include two-point transition probabilities [8, 5], seismic data [7], mutiple-point statistics [3,10], and well test-derived effective properties [4]. Traditional two-point variogram/covariance functions and correlation coefficients with a secondary attribute will be considered in this paper. The real advantage of annealing is this ability to integrate many disparate sources of data.

A modified version of the public domain `sasim` source code documented in [6] was used for the examples presented below.

AN EXAMPLE

A cross section bounded by two wells is shown at the top of Figure 2. The rock is a binary mixture of sandstone (1000md) and shale (0.01md). The gray-shaded profile next to each well represents the shale proportion for the corresponding vertical interval; white corresponds to 0% shale and black to 100% shale. The absolute permeability of each vertical interval was taken as the geometric average of the component sandstone and shale fractions. The histogram of shale proportion, 50 values from each well, is given on the lower-left of Figure 2. The variogram model of shale proportion (standardized to a unit sill) is shown on the lower-right of Figure 2. Note that the vertical extent of the cross-section is 10 meters (5 times the vertical variogram range) and the horizontal extent is 25 meters (2.5 times the horizontal variogram range).

The stochastic simulation problem is to construct representative 2-D models of the shale proportion and convert them to elementary grid-block permeabilities (here using a geometric average of the component shale and sandstone).

Three candidate stochastic simulation techniques based on three different RF models were considered: multiGaussian (`sgsim` program in GSLIB [6]), indicator (`sisim` [6]), and simulated annealing (`sasim` [6]). The objective function in `sasim` was the variogram for 50 lags ($n_h = 50$) covering all directions:

$$O = \sum_{i=1}^{n_h} [\gamma^*(h_i) - \gamma(h_i)]^2 \tag{5}$$

where n_h is the number of variogram lags to be honored $\gamma^*(h_i)$ is the variogram of the realization for lag h_i, and $\gamma(h_i)$ is the model variogram.

One hundred realizations were generated by each technique. The CPU time requirements are shown on Table 1. The time to generate 100 realizations varies from 15 minutes for `sgsim` to 4 hours for `sisim` (10 cutoffs). Two realizations from the Gaussian, indicator, and annealing RF models are shown on Figure 3; the realizations appear quite different due to characteristics of the RF model beyond the bivariate (variogram) level. All of the methods, however, appear plausible and meet the first criterion, i.e., they generate plausible realizations in a reasonable amount of time.

A transfer function is required to judge the space of uncertainty that is sampled by each RF model. The response variable selected here is the effective absolute permeability in the horizontal direction. Many other flow-related response variables

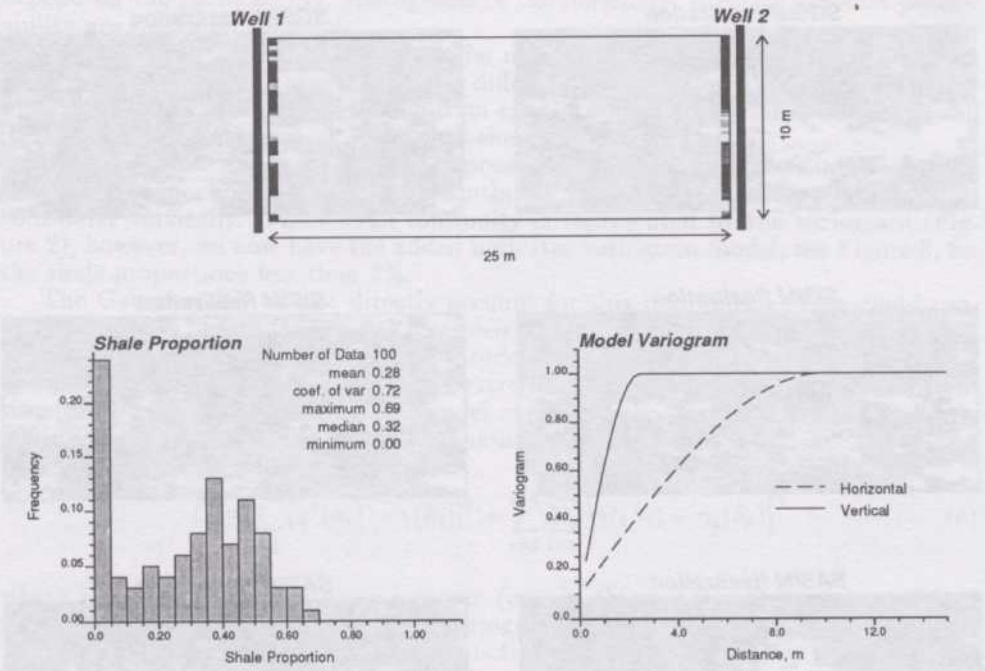

Figure 2: A gray scale plot of the two wells (at the far right and far left sides), a histogram of the 100 data and a model variogram.

Algorithm	Program	Total CPU Time(min)	Time (sec/node)	Relative Time
Gaussian	`sgsim`	15.4	0.0037	1.0
Indicator (median approx.)	`sisim`	30.0	0.0072	2.0
Indicator (10 cutoffs)	`sisim`	238.7	0.0573	15.5
Annealing (50 lags)	`sasim`	109.2	0.0262	7.1

Table 1: The CPU time, measured on a Silicon Graphics Crimson workstation, required to generate 100 simulations each with 2500 nodes. The time to simulate one node is also shown. The relative times are with respect to the `sgsim` program.

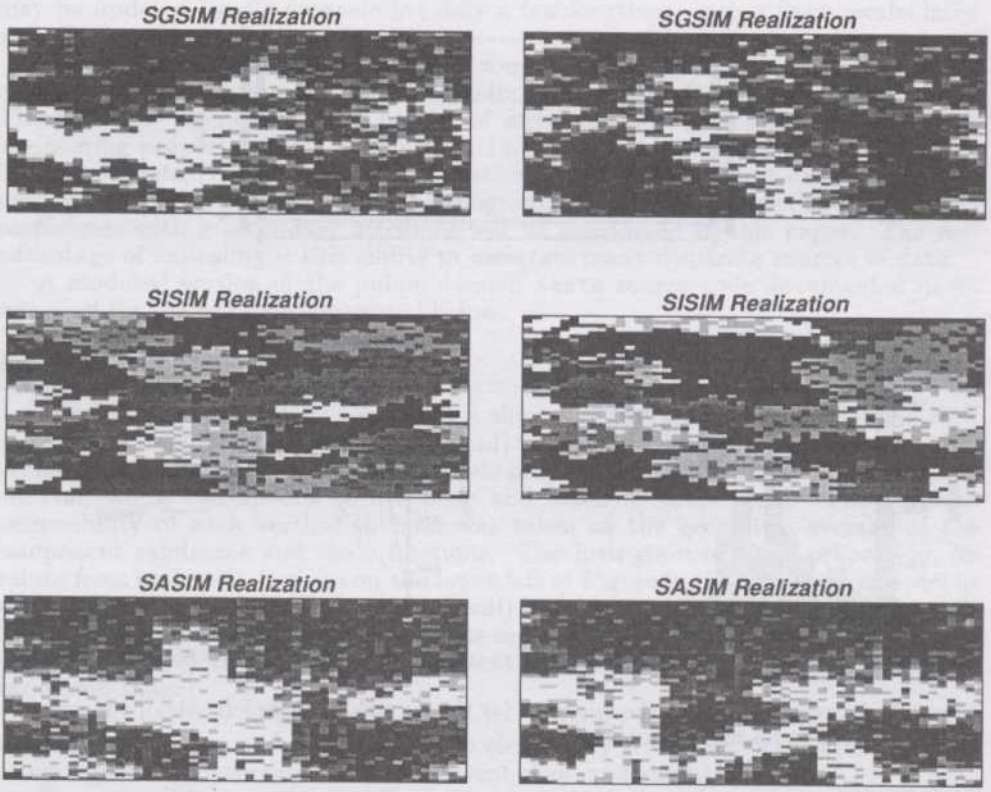

Figure 3: Two realizations of the multiGaussian RF (top), the indicator RF (center), and the annealing RF (bottom).

depend on the permeability. Histograms of the horizontal effective absolute permeability are shown on Figure 4.

The center of these distributions (the most likely response) and the spread (a measure of uncertainty) are significantly different for each simulation technique. The measures of the spread of the distribution are the standard deviation and the width of the 95% probability interval (shown below each histogram).

To extend this example further, suppose that we knew that the sandstone units (low shale proportion) had a greater continuity in the horizontal direction and less continuity vertically. The overall continuity is represented by the variogram (Figure 2), however, we now have the added indicator variogram model, see Figure 5, for the shale proportions less than 5%.

The Gaussian RF cannot directly account for this information. One could consider a mixture of two Gaussian RF models. The indicator RF model can integrate this information. However, the typical order relations corrections make it difficult to honor an indicator variogram for an extreme threshold and the overall variogram simultaneously. The annealing RF model can integrate this information by simply adding a component to the objective function:

$$O = \sum_{i=1}^{n_h} [\gamma^*(h_i) - \gamma(h_i)]^2 + \sum_{j=1}^{n_c} \sum_{i=1}^{n_h} \left[\gamma_j^*(h_i) - \gamma_j(h_i)\right]^2 \qquad (6)$$

where there are n_c indicator variograms (one in this case). The number of lags $n_h = 50$ will be constant for both the variogram and the single indicator variogram. Two realizations using this objective function are shown on Figure 6. Notice the increased continuity of the sand (white).

One hundred realizations were generated and the horizontal effective permeability was computed, see Figure 7. Notice that the center of the distribution has changed significantly; it has more than doubled from the previous runs (Figure 4). Further, the spread of this distribution has increased. The uncertainty should decrease as more information becomes available. The implication is that the uncertainty shown on Figure 4 is optimistic. It is a fairly general observation, that conventional measures of uncertainty used in geostatistics tend to increase rather than decrease as more information becomes available.

REMARKS AND CONCLUSIONS

The multivariate probability law of RF models implicit to most simulation algorithms, aside from the multiGaussian RF, is usually too complex to be defined and understood analytically. The main advantages of an analytically-defined RF, mathematical cleanliness and internal consistency, are of no benefit when the response variable is obtained from a non-linear transfer function, such as a flow simulator. The output distributions of uncertainty, in general, are obtainable only by generating multiple realizations.

RF models which are inferred by generating a number of realizations and observing the multivariate probability law are refered to as algorithmically-defined RF models. The main advantage of these RF models is the flexibility to integrate additional data from various sources.

The RF models discussed in this paper are a fair sample of commonly used techniques. There are many other RF models that could have been discussed, e.g.,

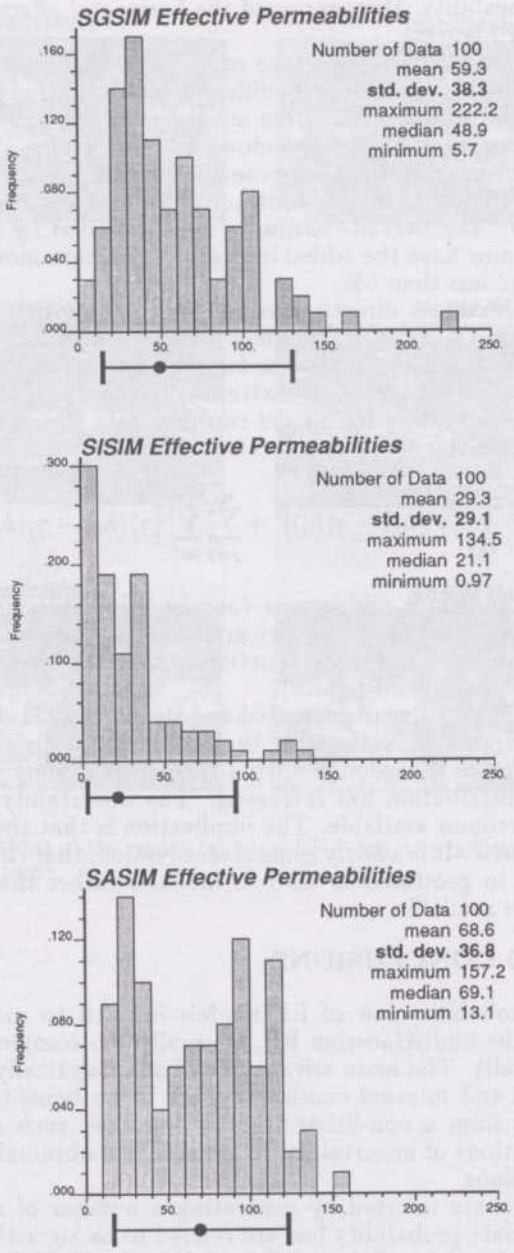

Figure 4: Histograms of the horizontal effective permeability for 100 realizations of the tiGaussian RF (top), the indicator RF (center), and the annealing RF (bottom). The 95% probability interval and the median is shown below each histogram.

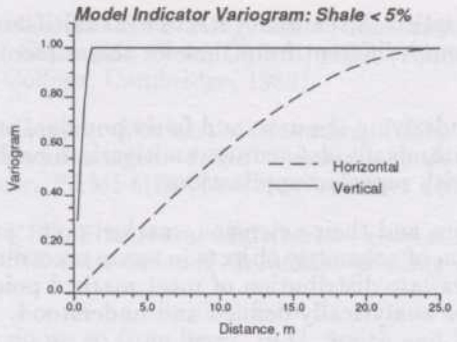

Figure 5: The model indicator variogram for the sandstone (low shale proportion).

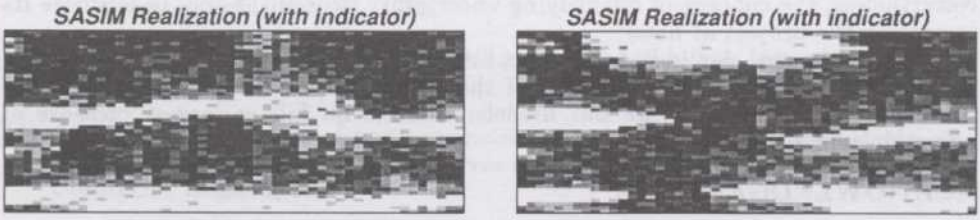

Figure 6: Two realizations of the annealing RF honoring the variogram and an indicator variogram.

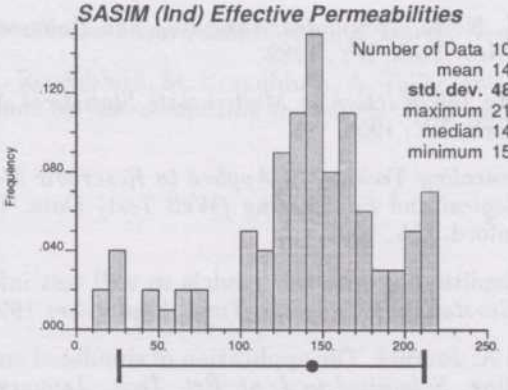

Figure 7: Histogram of the horizontal effective permeability for 100 realizations of the annealing RF honoring the variogram and an indicator variogram. The 95% probability interval and the median are shown below the histogram.

- The RF model implicit in the use of fractals is multiGaussian. The implementation is significantly different from that for the sequential algorithm adopted in GSLIB [6].
- The RF model underlying the new, and fairly popular, probability field simulation is also algorithmically-defined; its multivariate probability characteristics are understood with repeated applications.
- Boolean algorithms and their extension, marked point processes are generated by the distribution of geometric objects in space according to some probability laws. The multivariate distribution of most marked point processes is usually too complex to be analytically defined and understood.

Generating the stochastic realizations is the first step; it is the performance of the models that matter. Processing each stochastic realization with the a flow simulator allows the uncertainty to be quantified. Since reality is certain, it is difficult to validate our quantification of uncertainty; we cannot actually measure *real* uncertainty. Nevertheless, the concept of quantifying uncertainty is useful if only to evaluate its impact on the project at hand.

Perhaps the goal should be to obtain a limited number of realizations, constrained by **all** the data, to predict the center of the output response variable distribution. All models of uncertainty are but models; there is no full or largest measure of uncertainty in absolute.

ACKNOWLEDGEMENTS

The author would like to thank the management of Exxon Production Research Company for permission to publish this paper.

REFERENCES

[1] E. Aarts and J. Korst. *Simulated Annealing and Boltzmann Machines.* John Wiley & Sons, New York, NY, 1989.

[2] T. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York, NY, 1958.

[3] C. Deutsch. *Annealing Techniques Applied to Reservoir Modeling and the Integration of Geological and Engineering (Well Test) Data.* PhD thesis, Stanford University, Stanford, CA, 1992.

[4] C. Deutsch. Conditioning reservoir models to well test information. In *Fourth International Geostatistics Congress*, Troia, September 1992.

[5] C. Deutsch and A. Journel. The application of simulated annealing to stochastic reservoir modeling. *Submitted to J. of. Pet. Tech.*, January 1991.

[6] C. Deutsch and A. Journel. *GSLIB: Geostatistical Software Library and User's Guide.* Oxford University Press, New York, NY, 1992.

[7] P. Doyen and T. Guidish. Seismic discrimination of lithology: a Bayesian approach. In *Geostatistics Symposium*, Calgary, AB, May 1990.

[8] C. Farmer. Numerical rocks. In P. King, editor, *The Mathematical Generation of Reservoir Geology*, Clarendon Press, Oxford, 1992. (Proceedings of a conference held at Robinson College, Cambridge, 1989).

[9] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, November 1984.

[10] F. Guardiano and R. M. Srivastava. Multivariate geostatistics: beyond bivariate moments. In *Fourth International Geostatistics Congress*, Troia, September 1992.

[11] E. Jaynes. Where do we go from here? In C. Smith and W. Gandy Jr., editors, *Maximum Entropy and Bayesian Methods in Inverse Problems*, pages 21–58, Reidel, Dordrecht, Holland, 1985.

[12] R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis.* Prentice Hall, Englewood Cliffs, NJ, 1982.

[13] A. Journel. Non-parametric estimation of spatial distributions. *Math Geology*, 15(3):445–468, 1983.

[14] A. Journel. The place of non-parametric geostatistics. In G. Verly et al., editors, *Geostatistics for natural resources characterization*, pages 307–355, Reidel, Dordrecht, Holland, 1984.

[15] A. Journel and C. Deutsch. Entropy and spatial disorder. *Math Geology*, 1993.

[16] B. Kedem. *Binary Time Series.* Marcel Dekker, New York, NY, 1980.

[17] S. Kirkpatrick, C. Gelatt Jr., and M. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, May 1983.

[18] S. Kullback. *Information Theory and Statistics.* Dover, New York, NY, 1968.

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