US Department of Energy, Office of Science
Office of Biological and Environmental Research (BER)
Subsurface Biogeochemical Research (SBR)
FY11 Fourth Quarter Performance Measure

(Based in part on a paper submitted to Advances in Water Resources)
Yoram Rubin, UC Berkeley ©, August 29, 2011.

Stochastic Framework for Goal-oriented, Hydrogeological Site Characterization

1. Introduction

This third quarter FY11 SBR overall Performance Assessment Rating Tool (PART) measure is focused on the topic: advanced computational methods for assimilating multi-scale field data sets and for estimating reactive transport model parameters. This milestone is focused on research being performed at the Hanford Integrated Field Research Challenge (IFRC) site, located in the 300 Area at the Hanford Site in southeastern Washington State.

Previous milestones reported on the on the complexity of the transport models that need to be addressed (See Lichtner, First Quarter 2011 Milestone, and Scheibe, Second Quarter 2011 Milestone). Inverse modeling is the quantitative process of identifying the values of parameters of these models. Data assimilation in inverse modeling is a particular approach for inverse modeling that utilizes multiple and complimentary types of data as sources of information relevant to these parameters.

The success of inverse modeling could be measured in different ways. If success means being able to accurately quantify uncertainty, then this could be accomplished through a comprehensive theory that would account for all sources of uncertainty without bias. If success means reducing uncertainty or increasing confidence in predictions, then inverse modeling should task-oriented, which means being driven by prediction goals. Prediction goals could be defined quite broadly, departing from the consensus of improved characterization of target variables such as conductivity and porosity, and looking at variables such as risk (to the environment or to humans). Over the last few decades, significant efforts have been invested in the former, but only recently a concerted effort has been invested towards the later, shifting the focus from statistical fundamentals to a comprehensive approaches which attempt to build a statistically-consistent framework which also rationalize field sampling campaigns and align them with prediction goals. Significant part of that effort carried out within the Hanford IFRC program. The departure from the former approach towards the latter is motivated by the need to accomplish a comprehensive UQ within limited budgetary confines.

The focus of this report is on the data acquisition component of inverse modeling, with a specific focus on methods developed at the Hanford 300 IFRC project. We will review the evolution of the thought on

2. Background

Inverse modeling in general is made complicated by the large natural variability of parameters such as the hydraulic conductivity, porosity and chemical reaction parameters, and by the scarcity of adequate data. The 300 Area IFRC site is no exception in terms of spatial variability.

The site is unique in that extraordinary efforts have been invested in data acquisition. Measurements were taken along boreholes, and geophysical surveys were taken, both surface and cross-borehole. Pumping tests were conducted at multiple spatial scales, ranging from large-scale (tens of meters in diameter) constant-rate injection tests to small-scale (a few meters in diameter) EBF profiling. Tracer experiments were conducted to better understand solute transport processes. These tests are intended to provide the foundation for sound interpretation of various research hypotheses related to the transport and fate of uranium in the subsurface. But the road map for achieving that foundation is not easy to define. One could assume that such a large variety of data types would eliminate the need to perform detailed design of the parameters of the data acquisition campaign (i.e., what to collect, where, how many, what frequency, etc). However, this is not the case, and there are still open questions regarding the adequacy of the characterization efforts. The reason for that is quite obvious: hydrogeological data acquisition is like throwing darts in the dark, only much worse, because the outcome is more consequential.

The development of a sound rationale for designing field sampling campaigns was addressed in several studies over the last 2 decades, and so there is an evolution of thinking on this topic that is worth reviewing. These studies commonly try to make data acquisition better than shooting darts. However, this thinking has not translated yet into practical guidelines, primarily because it lags the rapid evolution of field sampling techniques. What had been endeavor that is based primarily on pumping tests, includes now a wide range of multiple type and multiple scale techniques (Hubbard and Rubin, 2000; Hubbard et al., 2001; Rubin and Hubbard, 2005; Rubin et al., 2006). The rapid growth in the variety of data acquisition techniques is demonstrated in Figure 2 (from Hubbard et al., 2001). This figure makes the point that multi-scale, multi-type data acquisition alternatives are available, and hence different strategies of data acquisition could be pursued, depending on the scale of the inquiry. However, this is only a partial view, because processes that may need low-resolution (coarse) characterization at one time, may require a much finer, high-resolution characterization at a later time. This point is discussed and demonstrated in Figure XXX+1 (from Hubbard and Rubin, 2001). So what's needed is a rational method for designing field sampling campaigns that would allow the user to consider multiple data types, and at the same time consider the uncertainty that is involved in "throwing darts in the dark". The next few sections provide a review of ideas and the evolution of thinking in this direction.

2. Maxwell et al., 1999

Maxwell et al. (1999) is among the first to make significant progress towards task-oriented inverse modeling. It showed how uncertainty in hydrologic variables could be related to health-risk assessment to individuals utilizing contaminated household water produced from groundwater sources. Although the focus of that work is on health-risk, the ideas pursued in that study could be implemented for a variety of tasks whose success depend on data hydrogeological data acquisition.

The methodology presented in Maxwell et al. (1999) relates between variability and uncertainty in the hydrogeology and physiology, on one hand, and their compound effects on the probability of enhanced cancer risk, on the other. On the hydrogeology side, variability refers to spatial variability in the hydrological parameters and its effect on contaminant transport (cf., Rubin et al., 1994), whereas on the physiology side, it refers to variations in the physiological responses between individuals. The hydrogeological and physiological models could be subject to parametric uncertainty and to conceptual model uncertainty. Parametric uncertainty refers to estimating the parameters of a particular model due to data scarcity, whereas conceptual model uncertainty refers to the potential existence of multiple and competing models that could be used to explain the data. These distinctions are important in the context of inverse modeling because they lead to differences in characterization needs and in inverse modeling strategies. Maxwell et al., (1999) dealt extensively with variability and parametric uncertainty, and ignored conceptual model uncertainty.

Scaling in transport is an important issue in this context, as noted in Hubbard and Rubin (2000). The relationship between the various length scales of the transport problem is known to impact the uncertainty associated with it (cf., Dagan, 1991; Rubin et al., 1999 and 2003; de Barros and Rubin, 2011, de Barros et l., 2011). For example, under ergodic conditions (Dagan, 1991), and assuming no uncertainty in the parameters of the geostatistical models, macrodispersion coefficients and the spatial moments of the solute body are not subject to uncertainty. They can be modeled through the use of effective parameters and they do not require to model the spatial variability in detail. This has profound implications on the site characterization needs, depending on the nature of the environmental performance metrics of the transport problem (e.g., spatial averages vs. point concentration).

Maxwell et al. (1999) imply that hydrological characterization needs could and should be linked to the prediction goals that could be defied in terms of environmental performance metrics (EPMs) such as concentrations or travel times, or other, directly or indirectly related metrics, such as health risks, and that the characteristic length scales of the environmental performance metrics (EPMs) must be recognized. For example, the health risk to individuals from groundwater resources is affected by spatially-averaged concentrations and not by point concentrations, with obvious and significant implications with regard to site characterization needs. Pumping groundwater in large quantities and

over time means that average concentrations, and not point concentrations, are the critical EPMs. Predicting average concentrations could reduce or even eliminate the need for detailed mapping of variables such as the conductivities, shifting the focus away from detailed characterization and towards estimating geostatistical models. The challenges associated with these concepts will be addressed in the next sections.

3. Goal/task-oriented Inverse Approach (de Barros et al., 2008, 2009)

The multiple sources of uncertainty considered in Maxwell et al. (1999), and their impact on risk uncertainty, suggest that it would be beneficial to relate the uncertainty level in the EPM with characterization needs. This raises a couple of questions. First, how to relate EPM uncertainty with characterization needs? The second question is how to allocate resources between the hydrologeological and physiological aspects of the problem, and between the data acquisition options that exist for each? These are challenging questions because, firstly, while in principle one could expect to reduce EPM uncertainty by investing in any of these aspects, there could be significant differences in the rewards in terms of reduced risk uncertainty. And secondly, the rewards are difficult to predict because of the inherent uncertainty. We end up with a problem of allocating resources under uncertainty, and it is a vicious circle because the benefits of characterization could not be predicted accurately a-priori, yet decisions need to be made about what measurements to take etc.

De Barros and Rubin (2008) presented a probabilistic framework for addressing the challenges of relating health risk assessment to uncertainty in the physiology and in the hydrogeology. In their approach, a probabilistic health risk model in the form of a cumulative distribution function is coupled with entropy-based information measures and is employed to relate the expected reduction in hydrological and physiological uncertainty due to data acquisition, to potential reduction in risk uncertainty. To discuss their ideas, let us consider a hydrological model with a vector of parameters $\boldsymbol{\theta}_H$. Such vector may assume different forms. For example, it can include geostatistical parameters such as mean, variance and correlation lengths. We could also define a physiological model, such as the USEPA high carcinogenic risk model, with a corresponding vector of parameters $\boldsymbol{\theta}_P$, which includes parameters such as the metabolized cancer potency factor (USEPA, 1989). These models are subject to uncertainty, both parametric and conceptual. This uncertainty could be quantified in different ways. For example, following Christakos (1992), we could assess uncertainty using the entropy of these vectors, E_P and E_H. The entropies E_P and E_H could be obtained from the joint statistical distributions of $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_H$, respectively. These distributions could be inferred from analysis of data, possibly in the form of inverse modeling (cf., Rubin et al., 2010), leading to $\boldsymbol{\theta}_P = \boldsymbol{\theta}_P(I_P)$ and $\boldsymbol{\theta}_H = \boldsymbol{\theta}_H(I_H)$, where I_P and I_H denote the relevant and available physiological and hydrological information.

Let us now consider $I_P = \{I_P^{(1)},..,I_P^{(N)}\}$, with each of the terms in brackets representing a set of data out of *N* data sets, which could be different data types or same type of data only collected at different times and/or different locations). Similarly let us also define $I_H = \{I_H^{(1)},..,I_H^{(M)}\}$. The challenge in site characterization is to allocate resources between the various types of data $I_P^{(i)}$

and $I_H^{(j)}$, such that we could obtain estimates of risk that are optimal in some sense. This challenge can be viewed as a problem of dynamic resources allocation, for example, given that $I_P = \{I_P^{(1)},..,I_P^{(N-1)}\}$ and $I_H = \{I_H^{(1)},..,I_H^{(M-1)}\}$ represent data sets that are already available, should one invest resources in obtaining $I_P^{(N)}$ or rather, $I_H^{(M)}$.

To simplify notation, let us define z_P and z_H as the data sets currently available for physiological and hydrogeological characterization, respectively. Similarly, $\boldsymbol{y}_P$ and $\boldsymbol{y}_H$ will be used to denote the physiological and hydrogeological data not yet available but under consideration. The entropy corresponding to $\boldsymbol{\theta}_H$ is defined from the conditional distributions $f(\boldsymbol{\theta}_H|z_H)$, or from $f(\boldsymbol{\theta}_H|z_H,y_H)$, once $\boldsymbol{y}_H$ becomes available. The entropy E_H given $\boldsymbol{z}_H$ is given by:

$$E_H^{(1)} = -\int f(\boldsymbol{\theta}_H|z_H)\ln[f(\boldsymbol{\theta}_H|z_H)]d\boldsymbol{\theta}_H \quad (1)$$

and the one corresponding to $\boldsymbol{z}_H$ and $\boldsymbol{y}_H$ is given by

$$E_H^{(2)} = -\int f(\boldsymbol{\theta}_H|z_H,y_H)\ln[f(\boldsymbol{\theta}_H|z_H,y_H)]d\boldsymbol{\theta}_H\,. \quad (2)$$

For a vector $\boldsymbol{\theta}_H$ of order *p*, $d\boldsymbol{\theta}_H = d\boldsymbol{\theta}_1..d\boldsymbol{\theta}_p$. Similar expressions could be defined for physiology entropies, $E_P^{(1)}$ and $E_P^{(2)}$.

Generally, and depending the quality of $\boldsymbol{y}_H$, we should expect $E_H^{(2)}$ to be smaller than $E_H^{(1)}$, which would indicate a gain in information. However, $E_H^{(2)}$ cannot be defined until the measurements included in $\boldsymbol{y}_H$ are actually taken. Until then, $E_H^{(2)}$ could only be estimated, for example by taking its expected value over all possible combinations of $\boldsymbol{y}_H$, as follows:

$$\langle E_H^{(2)}|\boldsymbol{y}_H\rangle = \int f(\boldsymbol{\theta}_H|\boldsymbol{z}_H,\boldsymbol{y}_H)f(y_H|\boldsymbol{z}_H,\boldsymbol{\theta}_H)\ln[f(\boldsymbol{\theta}_H|\boldsymbol{z}_H,\boldsymbol{y}_H)]d\boldsymbol{y}_H \quad (3)$$

The expected value given in eq. (3) could provide an indication about the potential benefits of $\boldsymbol{y}_H$. It is a useful tool for design, because once could try alternative designs for $\boldsymbol{y}_H$ and focus on those that would lead to the smallest expected value.

Equation (3) is difficult to solve from the main reason that $f(\boldsymbol{\theta}_H|z_H,y_H)$ and for $f(y_H|z_H,\boldsymbol{\theta}_H)$ could actually vary with each $\boldsymbol{y}_H$, not only in value but also in form, in ways that are difficult to anticipate. Several assumptions have been employed to address this challenge, as outlined below.

One approach for solving (3) is to assume models with simple parameterization for $f(\boldsymbol{\theta}_H|z_H,y_H)$ and for $f(y_H|z_H,\boldsymbol{\theta}_H)$. For example, it is common to adopt a multivariate normal density function for $f(y_H|z_H,\boldsymbol{\theta}_H)$ when $\boldsymbol{z}_H$ and $\boldsymbol{y}_H$ represent measurements of logconductivity (cf., Rubin, 2003, Chapter 2). A multivariate normal density function was also assumed in situations where $\boldsymbol{z}_H$ and $\boldsymbol{y}_H$ represent measurements of logconductivity and hydraulic heads under certain limiting conditions (see Rubin and Dagan 1987a,b). Similar definitions could be written for $E_P^{(1)}$ and $E_P^{(2)}$, by changing the subscripts from "*H*" to "*P*" in both (2) and (3).

De Barros et al. (2008, 2009) looked at the dependence of the estimation variance of enhanced cancer risk, and related the expected changes in this variance due to the

acquisition of either $\boldsymbol{y}_H$ or $\boldsymbol{y}_P$, or both. The links between enhanced cancer risk on one hand and hydrology and physiology on the other were established through physical and physiological models. These models were expressed in terms of parameters that are defined by density functions that could be conditioned on measurements using the Bayesian formalism, as stated in (3).

The links between hydrology, physiology and enhanced cancer risk established in these two studies showed that data acquisition should be viewed in context. The priorities in data acquisition could shift from hydrology to physiology or in reverse depending on many factors. One trivial factor of course is the level of information available in $\boldsymbol{z}_H$ and $\boldsymbol{z}_P$ as reflected in $E_H^{(1)}$ and $E_P^{(1)}$. The relationship between these two could vary, as incremental data acquisitions are performed on both sides, possibly shifting the focus of data acquisition between the hydrogeology and physiology sides.

The gain in information on the hydrogeology side was shown to depend on the nature of the transport problem, and more specifically, on its characteristics length scales. The theoretical basis for this dependence was provided in Rubin et al., (1999), de Barros and Rubin (2011) and de Barros et al. (2011). All these studies analyzed solute transport in terms of the dimensions of the solute body and of the travel distance. Predicting the transport of small solute bodies or of large solute bodies poses different challenges. The trajectories of large solute bodies are relatively easy to predict compared to small solute bodies, and can be obtained with some confidence using macrodispersion coefficients. Small solute bodies, on the other hand, depend much more on local configurations of the hydraulic parameters (e.g., the conductivity), and as such, would require many point measurements of the conductivity in order to reach some reasonable level of confidence. Hence, there could be significant gain in information from local pump tests in the case of small solute bodies, making such investments worthwhile in pursuing, whereas for large solute bodies such tests would not yield significant gains in information.

To address the complex interplay between gains in information on the hydrogeology and physiology sides, de Barros and Rubin (2009) introduced a graphic tool called Comparative Information Yield Curve, or CIYC in short, which allows the user to assess at each stage of the data acquisition where the information yield would be the largest. The yield in that study was defined in terms of uncertainty reduction in the EPM (e.g., enhanced cancer risk). In these two studies the information yield was calculated for different levels of $E_H^{(2)}$ and $E_P^{(2)}$. In this way, the information yield could be assesses for various data acquisition scenarios, $\boldsymbol{y}_H$ and $\boldsymbol{y}_P$. Specifically, the information yield could be related to the difference between $E_P^{(1)}(\boldsymbol{z}_P)$ and $\left\langle E_P^{(2)} | \boldsymbol{z}_P, \boldsymbol{y}_P \right\rangle$, ΔE_P, on the one hand, and the difference between $E_H^{(1)}(\boldsymbol{z}_H)$ and $\left\langle E_H^{(2)} | \boldsymbol{z}_H, \boldsymbol{y}_H \right\rangle$, ΔE_H, on the other. By relating ΔE_P and ΔE_H to expenditures, the CIYC could relate information yield to investments in data acquisition, which would make it an effective management tool.

In the previous derivations it was assumed that $f(\boldsymbol{\theta}_H|\cdot)$ or $f(\boldsymbol{\theta}_P|\cdot)$ are known up to parameter values, with uncertainty in the parameter values but not in the underlying models. This assumption allows the user to consider parametric uncertainty but it ignores the challenge of model uncertainty. This assumption is thus limiting in situations where alternative $f(\boldsymbol{\theta}_H|\cdot)$ or $f(\boldsymbol{\theta}_P|\cdot)$ could emerge with $\boldsymbol{y}_H$ or $\boldsymbol{y}_P$, respectively. One option for addressing the challenge of model uncertainty is to consider alternative models for

$f(\boldsymbol{\theta}_H|\cdot)$ or $f(\boldsymbol{\theta}_P|\cdot)$ and employ them within the framework of Bayesian model averaging (Neuman, 2003), as proposed in de Barros et al. (2009). In this approach, instead of using a single model, we could work with alternative models. For example, assuming *L* hydrogeological models, each with a different set of parameters $\boldsymbol{\theta}_{H,i}$, we would have a different density function for each, $f^{(i)}(\boldsymbol{\theta}_{H,i}|\cdot)$, *i*=1,..,*L*, and each of these models will be associated with a different degree of plausibility, given by a corresponding probabilities $\boldsymbol{\pi}_i$, i=1,..,*L*, such that $\sum_i \boldsymbol{\pi}_i = 1$. With this, the entropy in (3) would be replaced by:

$$\left\langle E_H^{(2)} | \boldsymbol{y}_H \right\rangle = \sum_i \boldsymbol{\pi}_i \int f^{(i)}(\boldsymbol{\theta}_H | \boldsymbol{z}_H, \boldsymbol{y}_H) f^{(i)}(y_H | \boldsymbol{z}_H, \boldsymbol{\theta}_{H,i}) \ln\left[f(\boldsymbol{\theta}_{H,i} | \boldsymbol{z}_H, \boldsymbol{y}_H) \right] d\boldsymbol{y}_H \qquad (4)$$

Equation (4) offers an attractive alternative to (3) in that it opens the door for alternative models that could possibly emerge with $\boldsymbol{y}_H$. However, it poses a few challenges that limit the application of (4). The first challenge is that the *L* alternative models need to be mutually independent models. This requirement is conceptually attractive because, presumably, it allows us to believe that the model selection process involves independent experts, representing independent points of views. There are several questions associated with this issue: What does that mean exactly, for models to be mutually independent, e.g., does that mean independently-derived? And if this is so, is this at all possible? How does one come up with mutually independent models and what test is there to ascertain the mutual independence of models? Can a single researcher or even a group of researchers come up with mutually independent models when looking at the same data and sharing the same professional background? The second challenge is how to define the set of model probabilities $\boldsymbol{\pi}_i$, *i*=1,..,*L*, before $\boldsymbol{y}_H$ is even acquired. And the third challenge is that new models, not foreseen at the initial model selection stage, could emerge as new data is acquired. These points suggest to us that the concept of Bayesian Model Averaging is viable only at a somewhat advanced stage of the investigation when only *L* plausible models could be defined and associated with probabilities, which also means that the emergence of additional alternatives could be safely ruled out.

4. Broad Spectrum Model Selection (Rubin et al., 2010; Nowak et al., 2010)

Entropy measures such as (4) must be related them with prediction goals such as estimating enhanced cancer risk or improving the accuracy of predicting concentrations at one or more environmentally-sensitive targets or travel times between the contamination source and a target. In the Hanford IFRC site, for example, there is a strong interest in both concentrations and travel times between various sources and targets such as the Columbia River.

Addressing the goal of planning data acquisition for specified prediction goals, and in light of parametric and conceptual model uncertainty, while relaxing the constraints discussed in the previous section is discussed in Rubin et al., (2010) and

Nowak et al., (2010). The main innovations of these papers is in translating (part of the) conceptual model uncertainty into parametric uncertainty, thus eliminating the limitations of working with a finite number of alternative, a-priori models, as well as the need to specify model probabilities π_i, i=1,..,L

Recalling equation (3), let us consider further the possibility that variations in θ could represent parametric uncertainty *as well* as modeling uncertainty. In other words, what if differences in parameter values could represent conceptually different models? We can demonstrate this idea the Matérn covariance function (Matérn, 1986; Rubin et al., 2010; Nowak et al., 2010, equation 10). Following the formulation used in Diggle and Ribeiro (2007, equation 3.6), this function is defined as follows:

$$C(h) = \frac{\sigma_Y^2}{2^{\kappa-1}\Gamma(\kappa)}(h)^{\kappa} K_{\kappa}(h) \ , \qquad (5)$$

with $h^2 = \sum_{i=1}^{m} (r_i / \lambda_i)^2$. The parameter m is the space dimensionality, whereas r_i and λ_i, i=1,..,m, denote the Euclidean lag distances and the corresponding length scales, respectively. K_{κ} is the modified Bessel function of order κ, Γ is the Gamma function and κ is a non-negative parameter. Specific values of κ lead to well-known models: with κ values of 0.5, 1 and ∞ , Equation (5) simplifies to the exponential, Whittle and Gaussian covariance functions, respectively. Combinations of κ and the length scales λ_i offer additional flexibility. For example, κ=1 and λ_i approaching infinity lead to the power law covariance. Hence it offers much flexibility of using simple parameterization in order to deal with conceptual model uncertainty.

The shape parameter κ is not limited to any of the values specified above. It can assume any non-negative value, and with the freedom to select a value from a broad range of values, we now have the flexibility to consider a huge number of alternative covariance models. Consider equation (3) with the vector of parameters θ_H including the parameters of the Matérn covariance function. By allowing κ and the rest of the parameters to vary over their support (e.g., between zero and infinity for κ), we could bring into consideration a large (infinite) number of alternative models, without the need to select a finite number of models and without the need to specify model probabilities, as required by (4).

The need to specify model probabilities in equation (4) is replaced by specifying the density function of the Matérn covariance function parameters. i.e., $f(\theta_H | \mathbf{z}_H, \mathbf{y}_H)$ in equation (3). However, there is an extensive body of literature that can be used for rational selection of this density function for any amount of information that is available. One approach is the minimum relative entropy (Woodbury and Rubin, 2000; Hou and Rubin, 2005). Such methods circumvent the need to select, somewhat speculatively, the model probabilities.

The idea of capturing model uncertainty through parameter uncertainty was referred to by Rubin at al (2010) as broad spectrum model selection, because a single parameter, when allowed to vary over a finite or infinite range, could be used to represent a broad spectrum of models. Broad spectrum model selection can be used as an alternative to the discrete model selection suggested by equation (4) in situations where alternative geostatistical models are being considered. The broad spectrum model

selection presented thus needs to be expanded to other aspects of model selection, such as the probability model for the hydrologic variables (e.g., the conductivity etc).

Nowak et al. (2010) employed this concept in the context of task-oriented data acquisition. They provided an important improvement compared to previous studies by allowing automatic adjustment of potential measurement locations. Whereas the abovementioned studies employed manual adjustment of measurement locations, here an algorithm was employed that could adjust measurement locations such as to optimize an a-priori defined goal. "Optimize" here implies placement that could produce optimal predictions (in the expected value sense) related to the task and not to hydrogeological variables. The optimal (or near-optimal, in fact, because global optimum is assumed, but not guaranteed) placement of measurement locations is driven by two objectives: the first objective is targets a modeling goal (could be enhanced cancer risk, as described previously, or others, as will be shown below), whereas the second objective focuses on reducing the uncertainty about the geostatistical model parameters (e.g., mean, variance, length scales).

The method developed in Nowak et al. (2010) was demonstrated in a couple of striking examples that are presented briefly below. In both cases the primary goal was to improve the accuracy of prediction of environmental performance metrics (EPM). The two EPM considered were the concentration at a given location at a given time (EPM1), and the second EPM was the mean arrival time between a source and an environmentally-sensitive target (EPM2). Results for EPM1 and EPM2 are provided in Figures 3 and 4, respectively.

In Figure 3 we note that the proposed measurement locations can be grouped in two groups: measurements in Group 1 sample the boundaries of the subdomain where the contaminants that are expected to hit the target could move, whereas the measurements in Group 2 surround the source. Group 1 can be explained by noting that the variability of the concentration (expressed in terms of the coefficient of variation) is largest at the plume boundaries (Rubin, 1991a,b; Rubin, 2003; de Barros et al., 2011). Hence, at early travel times, during the initial stages of plume expansion, we have the proposed measurement locations placed at increasing distances from the centerline (the mean trajectory). As we get closer to the target, it becomes more important to sample those streamlines that are expected to converge to the target, and hence the proposed measurement locations converge towards the centerline.

Group 2 focuses on source characterization. With integral scale of the order of 15 meters, we could conclude that source characterization is important even when the target is located at a distance of about 10 integral scales downstream. Why is source characterization so important? First, with good source characterization we could get strong indication about the plume's behavior and from that deduce clear guidelines on placement of measurements. For example, if the source discharge is narrowly-focused, we could expect the plume to disperse laterally very little (Dagan, 1984; Rubin et al., 1999) and that would focus Group 2 measurements along the expected trajectory, whereas for a solute discharge that is somewhat uniformly-distributed over the entire length of the source, Group 2 measurements would shift away from the mean trajectory. Second, because dispersion occurs primarily along the solute body's edges, where the concentration gradients are the largest, narrower solute bodies (focused discharge) will experience faster reduction in maximum concentration values, which would put a cap on

the concentrations at the target, with obvious implications on predictability of the concentrations at the target.

A somewhat different picture emerges in the case of EPM2 (see Figure 4). The first difference is the absence of grouping of measurements into two groups as noted in EPM1. The second difference is that measurements are placed along a narrow band linking the source with the target, and this is because a large fraction of the solute mass is expected to travel and reach the target through this narrow band. A few of the measurements are scattered outside of this band, and this scatter is intended to improve the identification of the geostatistical model parameters.

If most of the solute mass is expected to travel along this narrow band, why then most the measurements for EPM1, unlike EPM2, are placed outside of that band (as shown in Figure 3? This could be explained by the halo effect shown in Figure 5 (from de Barros et al., 2011). This effect was described in Rubin (1991). This figure shows the halo effect of the concentration variance: the variability is largest along a ring surrounding the plume's centroid and is relatively small at the plume's center. This halo expands in its radius over time. The halo will appear only after the maximum concentration becomes smaller than $C_0/2$. We see in Figure 5(b) that collecting measurements along the center trajectory does not reduce/eliminate the halo effect (although it does reduce the variance at the centroid) and hence there is no point in collecting measurements along the mean trajectory (i.e., the narrow band discussed earlier) for the purpose of improving the predictability of the concentration, because the largest contributors to uncertainty are at the halo. However, measurements collected along the narrow band would go a long way towards improving the predictability of travel times (Rubin and Dagan, 1995, de Barros et al., 2011).

References

Beven, K. (2002), Towards a coherent philosophy for modelling the environment, Proc. R. Soc. Lond. A, 458, 2465–2484.

Dagan, G., Solute transport in heterogeneous porous formations, J. Fluid Mech., 1101(115), 1 –177, 1984.

Dagan, G., Dispersion of a passive solute in non-ergodic transport by steady velocity fields in heterogeneous formations, Journal of Fluid Mechanics, 233: 197-210, DOI: 10.1017/S0022112091000459, 1991.

de Barros, F., and Y. Rubin, A Risk-Driven Approach for Subsurface Site Characterization, Water Resources Research, 44, W01414, doi:10.1029/2007WR006081, 2008.

de Barros, F.P.J., Y. Rubin, R.M. Maxwell, The concept of comparative information yield curves and its application to risk-based site characterization, Water Resour. Res., 45, W06401, doi:10.1029/2008WR007324, 2009.

de Barros FPJ, Y. Rubin and S. Ezzedine, Impact of hydrogeological data on measures of uncertainty, site characterization and environmental performance metrics. Adv Water Resour, doi:10.1016, 2011.

Diggle. P.J., Ribeiro, P.J., Model-based geostatistics, Springer, 2007.

Hubbard, S., Rubin, Y., Hydrogeological parameter estimation using geophysical data: A review of selected techniques, J. Contam. Hydrology, 45, 3-34, 2000.

Hubbard, S., Chen, J., Peterson, J., ,Ernest L. Majer, Kenneth H. Williams, Donald J. Swift, Brian Mailloux and Y. Rubin, Hydrogeological characterization of the South Oyster Bacterial transport site using geophysical data, Water Resour. Res., 37(10), 2431-2456, 2001.

Matérn, B., Spatial variation, Springer, Berlin, Germany, 1986.

Maxwell, R., Kastenberg, W., and Rubin, Y., Hydrogeological site characterization and its implication on human exposure risk assessment, Water Resour. Res., 35(9), 2841-2855, 1999.

Nowak, W., F. P. J. de Barros, and Y. Rubin, Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain, Water Resour. Res., 46, W03535, doi:10.1029/2009WR008312, 2010.

Nowak, W., Y. Rubin and F. P.J. de Barros, A Hypothesis-Driven Approach to Optimal Site Investigation, Submitted for Review, Water Resour. Res., (2011)

Popper, K. R., The Logic of Scientific Discovery, 2nd ed., Routledge, London, 2002.

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

Rubin, Y. and Dagan, G., Stochastic identification of transmissivity and effective recharge in steady groundwater flow: 2. Case study, Water Resour. Research, 23 (7), 1193-1200, 1987.

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

Rubin, Y., Transport in heterogeneous porous media - prediction and uncertainty, Water Resour. Res., 27(7), 1723-1738, 1991

Rubin, Y., M.A. Cushey and A. Bellin, Modelinig of transport in groundwater for environmental risk assessment, Stochastic Hydrology and Hydraulics, 8, 57-77, 1994.

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

Rubin, Y., Sun, A., Maxwell, R., Bellin, A., The concept of block effective macrodispersion, J. Fluid Mech, 395, 161-180, 1999.

Rubin, Y., Applied Stochastic Hydrogeology, Oxford University Press, New-York, 2003.

Rubin, Y., A. Bellin and A. Lawrence, On the use of block-effective macrodispersion for numerical simulation of transport in heterogeneous formations, Water Resour. Res., 39(9), 1242, doi:10.1029/2002WR001727, 2003.

Rubin, Y., and S.S. Hubbard, Hydrogeophsics, Springer, 2005

Rubin, Y., I. Lunt, J. Bridge, Spatial variability in river sediments and its link with river channel geometry, Water Resour. Res., 42(6), W06D16, doi:10.1029/2005WR004853, 2006

Rubin, Y., X. Chen, H. Murakami, and M. Hahn, A Bayesian approach for inverse modeling, data assimilation, and conditional simulation of spatial random fields, Water Resour. Res., 46, W10523, doi:10.1029/2009WR008799, 2010.

USEPA, Risk Assessment Guidance for Superfund Volume 1: Human Health Manual (Part A), Dec. 1989, Rep.EPA/540/1-89/002, 1989.

FIGURES

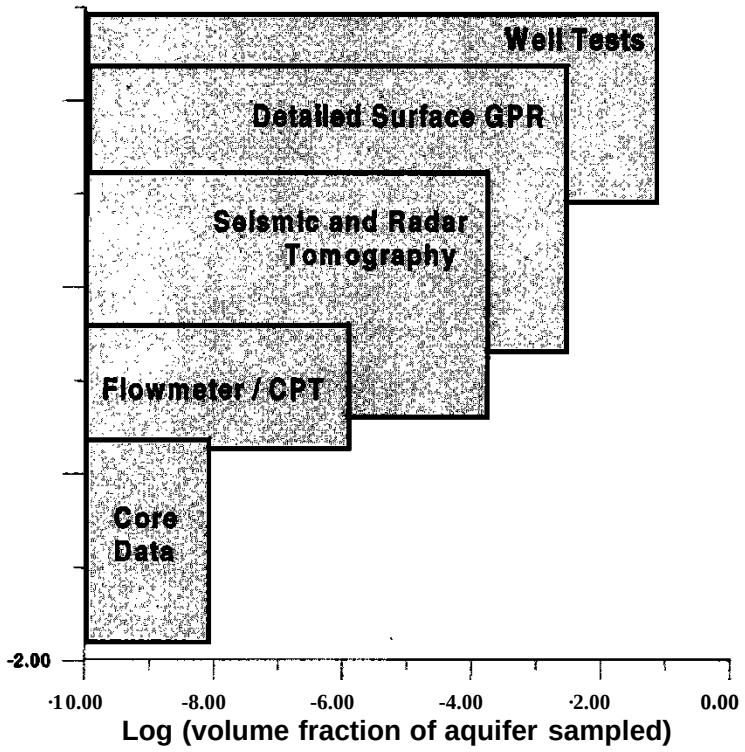

Figure 1: Comparison of resolution and fraction of aquifer volume sampled using different characterization tools at the Narrow Channel Focus Area. Geophysical data help to bridge the information gap in terms of both resolution and fraction of aquifer volume sampled between the more conventional hydrological sampling techniques of core analysis and well tests (From Hubbard et al., 2001).

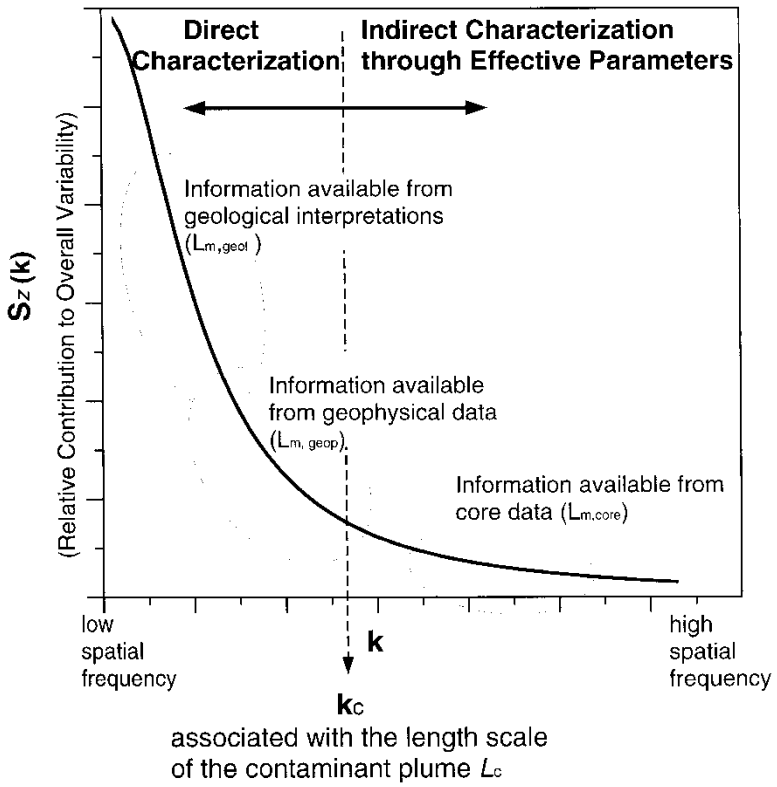

Figure 2

A spectral density function $Sz(\boldsymbol{k})$ illustrates the importance of scale in characterization and modeling efforts. Due to the measurement sampling interval and length of sampling window, different types of measurement can reside in different portions of the wave number $(\boldsymbol{k})$. range. The wave number range over which the measured data exist, compared to other existing wave number ranges _such as that associated with hydrological heterogeneity or with a contaminant plume. is important for determining the usefulness of the data for the characterization effort. In this example, $\boldsymbol{k}_c \sim (1/L_C)$ is the cut-off wave number associated with the scale of a hypothetical contaminant plume. This figure illustrates that high frequency $(\boldsymbol{k}>\boldsymbol{k}_c)$. borehole or outcrop information is most useful for determining the effective parameters rather than for actual mapping of Z, while lower wave number variability $Sz(\boldsymbol{k}<\boldsymbol{k}_c)$, such as those obtained from geophysical data, can be useful for direct characterization (From Hubbard and Rubin, 2000).

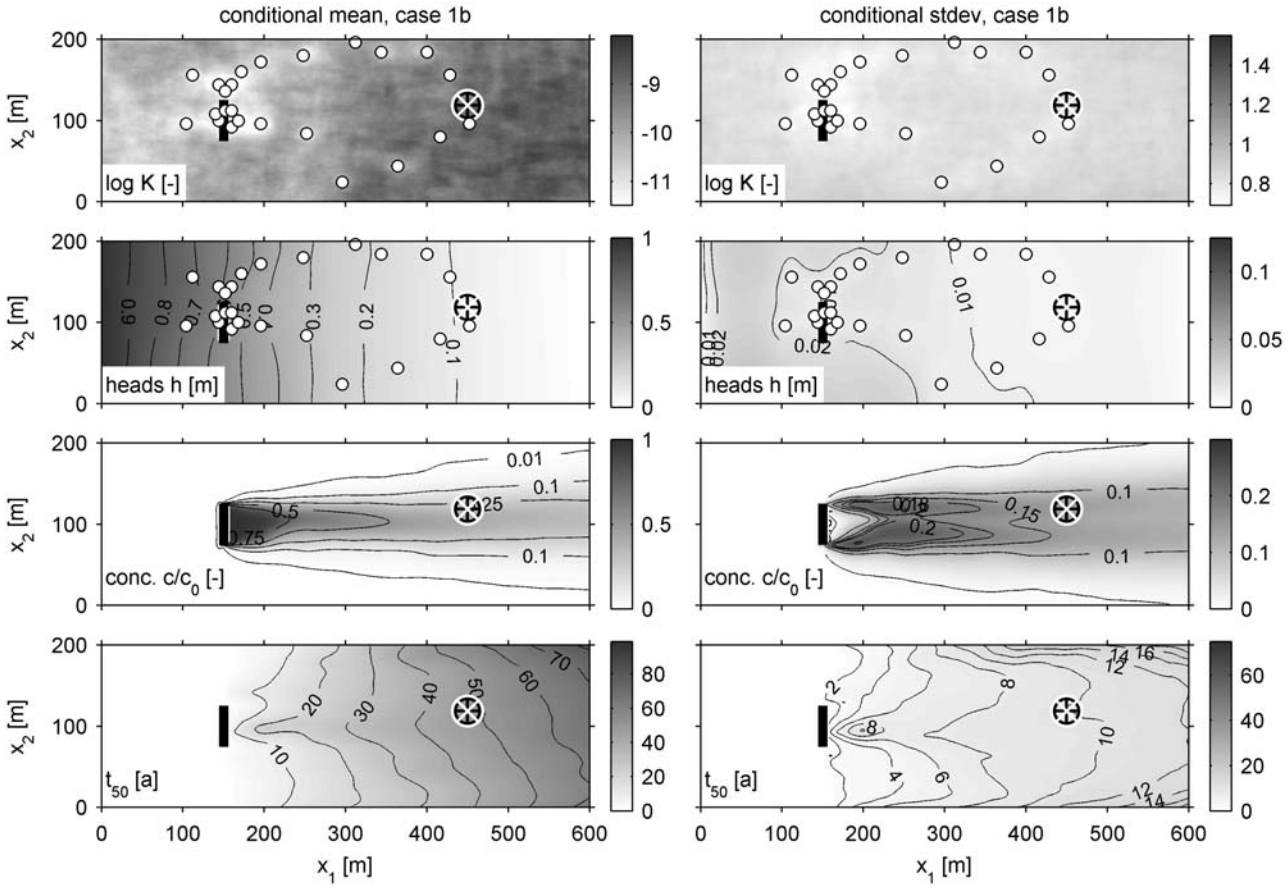

Figure 3 – Optimal sampling locations for EPM1 defined by the concentration at the crossed circle on the right-hand-side of the flow domain. The source is defined by the rectangle on the left-hand side. (left) Conditional mean for ln K, hydraulic heads h, steady state concentration and mean arrival time t_{50} of hypothetical plume. (right) Corresponding conditional standard deviations. Sensitive location (crossed circle). Near−optimal sampling locations (Y = ln K and head measurements) (solid white circles). From Nowak et al., 2010.

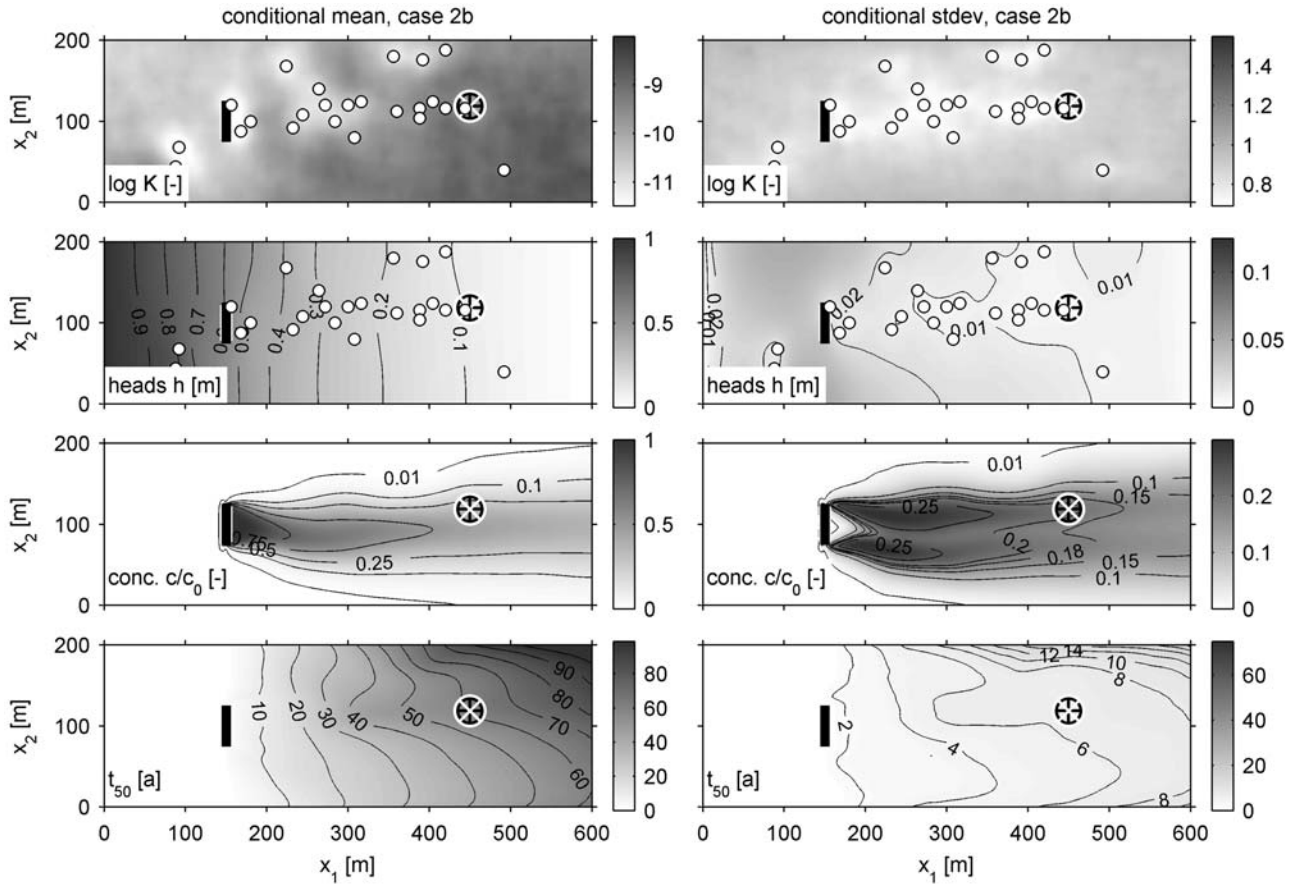

Figure 4 – Results for EPM2, defined as the mean arrival time between the source, defined by the rectangle on the left-hand side and the target defined by the crossed circle. (left) Conditional mean for ln K, hydraulic heads h, steady state concentration c and arrival time t_{50} of hypothetical plume. (right) Corresponding conditional standard deviations. Sensitive location (crossed circle). Near−optimal sampling locations (Y = ln K and head measurements) (solid white circles). From Nowak et al., (2010).

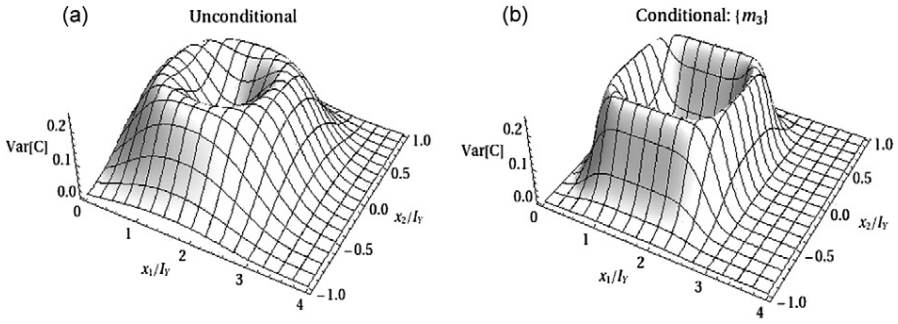

Figure 5 - Normalized concentration variances in 2D, $VAR(C) = \sigma_C^2 / C_0^2$, as a function of the spatial domain at dimensionless time tU/I_Y = 2.5. Mean flow direction is along the x_1 axis. U is the mean velocity and I_Y is the integral scale of the log-conductivity. C_0 is the initial concentration. The plume's centroid is at x_2=0. Plots shown for: (a) unconditional and (b) conditional on data comprised of measurements placed along x_2=0 (from de Barros et al., 2011).