A hypothesis-driven approach to optimize field campaigns

Wolfgang Nowak,[1] Yoram Rubin,[2] and Felipe P. J. de Barros[3]

Received 7 June 2011; revised 27 March 2012; accepted 18 April 2012; published 8 June 2012.

[1] Most field campaigns aim at helping in specified scientific or practical tasks, such as modeling, prediction, optimization, or management. Often these tasks involve binary decisions or seek answers to yes/no questions under uncertainty, e.g., Is a model adequate? Will contamination exceed a critical level? In this context, the information needs of hydro(geo)logical modeling should be satisfied with efficient and rational field campaigns, e.g., because budgets are limited. We propose a new framework to optimize field campaigns that defines the quest for defensible decisions as the ultimate goal. The key steps are to formulate yes/no questions under uncertainty as Bayesian hypothesis tests, and then use the expected failure probability of hypothesis testing as objective function. Our formalism is unique in that it optimizes field campaigns for maximum confidence in decisions on model choice, binary engineering or management decisions, or questions concerning compliance with environmental performance metrics. It is goal oriented, recognizing that different models, questions, or metrics deserve different treatment. We use a formal Bayesian scheme called PreDIA, which is free of linearization, and can handle arbitrary data types, scientific tasks, and sources of uncertainty (e.g., conceptual, physical, (geo)statistical, measurement errors). This reduces the bias due to possibly subjective assumptions prior to data collection and improves the chances of successful field campaigns even under conditions of model uncertainty. We illustrate our approach on two instructive examples from stochastic hydrogeology with increasing complexity.

Citation: Nowak, W., Y. Rubin, and F. P. J. de Barros (2012), A hypothesis-driven approach to optimize field campaigns, *Water Resour. Res.*, *48*, W06509, doi:10.1029/2011WR011016.

1. Introduction

[2] Uncertainty quantification and reduction are fundamental challenges in the environmental and hydrological sciences. Uncertainties arise due to data scarcity, limited observability, and our incapacity to fully resolve spatial variability and dynamics, or to define correctly all the physical, chemical, and biological processes involved with their boundary and initial conditions and forcing terms [e.g., *Christakos*, 1992; *Rubin*, 2003; *Oreskes et al.*, 1994]. As an outcome, full validation of model concepts and perfect model calibration is an almost impossible task [*Oreskes et al.*, 1994].

[3] In the hydro(geo)logical sciences, we often use models to predict and address scientific hypotheses or challenges in engineering and management under uncertainty:

[4] Should we shut down a drinking water well before contamination arrives [e.g., *Frind et al.*, 2006; *Enzenhöfer et al.*, 2012]?

[5] How large is the risk emanating from a contaminated site [e.g., *Troldborg et al.*, 2010]?

[6] Is natural attenuation occurring [e.g., *Schwede and Cirpka*, 2010; *Cvetkovic*, 2011]?

[7] Is a proposed remediation design safe [e.g., *Cirpka et al.*, 2004; *Bolster et al.*, 2009]?

[8] Is a high-level radioactive waste site safe [e.g., *Andricevic and Cvetkovic*, 1996]?

[9] Is a CO_2 injection site safe [e.g., *Oladyshkin et al.*, 2011a, 2011b]?

[10] Is human health risk above a critical value or not [e.g., *de Barros and Rubin*, 2008; *de Barros et al.*, 2009, 2011]?

[11] Is a proposed model adequate to answer these questions [e.g., *Neuman*, 2003; *Refsgaard et al.*, 2006]?

[12] All of the above examples include scientific hypotheses, binary decisions, or binary questions of compliance with environmental performance metrics such as human health risk or maximum contaminant levels [see *de Barros et al.*, 2012]. *Pappenberger and Beven* [2006] provide a list of more studies where binary decisions were taken under uncertainty. They also report the fact that practitioners often complain about the discrepancy between soft uncertainty bounds and the binary character of decisions.

[13] One way of dealing with binary questions in the face of uncertainty is to formalize them as hypothesis tests.

[1]Institute for Modelling Hydraulic and Environmental Systems (LH^2)/SimTech, University of Stuttgart, Stuttgart, Germany.
[2]Department of Civil and Environmental Engineering, University of California, Berkeley, California, USA.
[3]Department of Geotechnical Engineering and Geosciences, Technical University of Catalonia (UPC), Catalonia, Spain.

Corresponding author: W. Nowak, Institute for Modelling Hydraulic, Environmental Systems (LH^2)/SimTech, University of Stuttgart, Pfaffenwaldring 61 70569 Stuttgart, Germany. (wolfgang.nowak@iws.uni-stuttgart.de)

0043-1397/12/2011WR011016

This makes it possible to systematically and rigorously test assumptions, models, predictions, or decisions. *Beven* [2002] summarizes strong arguments that models and theories in the environmental sciences are nothing else but hypotheses. A prominent example is the work by *Luis and McLaughlin* [1992], who indeed approach model validation via formal statistical hypothesis testing. Consequently, modelers and scientists should admit the hypothesis-like character of models and their underlying theories, conceptualizations, assumptions, parameterizations, and parameter values. We propose that the same approach should be taken to support any type of decisions that modelers, engineers, scientists, and managers need to take under uncertainty. One should treat model predictions and derived conclusions and decisions as hypotheses, and one should continuously try to assess and test their validity.

[14] The credibility of and confidence in any answer or decision under uncertainty increases with well-selected additional data that help to better test, support, and calibrate the involved assumptions, models, and parameters. However, data must be collected in a rational and goal-oriented manner because field campaigns and laboratory analysis are expensive while budgets are limited [e.g., *James and Gorelick*, 1994].

[15] This is where optimal design and geostatistical optimal design [e.g., *Pukelsheim*, 2006; *Ucinski*, 2005; *Christakos*, 1992] come into play. Optimal design (OD) gets the maximum gain of information from limited sampling, field campaigns, or experimentation. It optimizes the projected trade-offs between the costs spent on additional data versus the higher level of information. It can be used to optimize (1) what types of data (e.g., material parameters, state variables) to collect, (2) where to sample (e.g., the spatial layout and time schedule of observation networks), and (3) how to best excite the system to observe an informative response (e.g., designing tracer injections or hydraulic tests). Many applications in groundwater hydrology can be found in the literature [e.g., *James and Gorelick*, 1994; *Reed et al.*, 2000a; *Herrera and Pinder*, 2005; *Nowak et al.*, 2010; *Leube et al.*, 2012].

[16] Classical OD theory is based on utility theory [e.g., *Fishburn*, 1970]. It optimizes the utility of sampling [e.g., *Pukelsheim*, 2006], which is traditionally defined as increased information [*Bernardo*, 1979] or reduced uncertainty (measured by variances, covariances, or entropies [e.g., *Pukelsheim*, 2006; *Nowak*, 2010; *Abellan and Noetinger*, 2010]). Unfortunately, these are only surrogate (approximate) measures for the actual utility, rather than ultimate measures [e.g., *Loaiciga et al.*, 1992]. The use of surrogates may corrupt the optimality of designs for the originally intended purpose.

[17] Goal-oriented approaches define the utility of sampling via ultimate measures, i.e., measures defined in the context of a given management application [e.g., *Ben-Zvi et al.*, 1988; *James and Gorelick*, 1994; *Feyen and Gorelick*, 2005; *Bhattacharjya et al.*, 2010; *Li*, 2010]. Thus, optimal sampling and field campaign strategies can adapt to the interplay between the actual information needs of the goal at hand, the available measurement and investigation techniques, and the specific composition of uncertainty [e.g., *Maxwell et al.*, 1999; *de Barros et al.*, 2009; *Nowak et al.*, 2010]. For instance, *de Barros et al.* [2012] showed how the utility of data depends on the considered environmental performance metric (e.g., maximum concentration levels, travel times, or human health risk).

[18] Our work focuses, for now, on situations where the objective for field campaigns or experimentation is to support binary decision problems with maximum confidence in a one-time field effort. By proper formulation of corresponding hypothesis tests, we cast the binary decision problem into the context of Bayesian decision theory [e.g., *Berger*, 1985] and Bayesian hypothesis testing [e.g., *Press*, 2003]. The latter differs from classical testing in two respects: (1) it can absorb prior knowledge on the likelihood of hypotheses, and (2) it can assess the probabilities of all possible decision errors. Then, we optimize field campaigns in a goal-oriented manner such that the total probability of decision error is minimized. The outcome of our approach are data collected rationally toward the specific hypothesis, question, or decision at hand, providing the desired confidence at minimal costs.

[19] There is a substantial body of work in the literature that optimizes field campaigns via Bayesian decision theory, maximizing the expected data worth [e.g., *Massmann and Freeze*, 1987; *James and Gorelick*, 1994]. The data worth concept follows classical ideas from utility theory and decision theory [e.g., *Fishburn*, 1970; *Ben-Zvi et al.*, 1988; *Raiffa et al.*, 1995]. It assigns monetary units for utility, and then weighs up expected benefits against the costs of field campaigns. The practical difference between our suggested approach and classical data worth studies is twofold. First, our approach encompasses arbitrary hypothesis tests or binary questions, whereas data worth studies are restricted to management tasks that provide a context for monetizing data utility. Second, we do not maximize the monetary worth of data collection but use the error probability of binary decisions or conclusions derived from the data as objective function to minimize. In contrast to classical data worth analysis, this avoids commensuration, i.e., does not require to have a common (monetary) scale for different values such as sampling costs versus improved scientific confidence or reduced health risk.

[20] An alternative approach to avoid commensuration is multiobjective optimization (MOO) [e.g., *Sawaragi et al.*, 1985; *Marler and Arora*, 2004]. MOO provides a suite of Pareto-optimal candidate solutions, i.e., solutions that cannot improve in any aspect without degrading in at least one other aspect. The final decision is found by inspecting and discussing the trade-offs in the different aspects between the suggested Pareto optima [e.g., *Reed and Minsker*, 2004; *Kollat and Reed*, 2007]. Thus, the problem of commensuration or preference articulation is postponed to after the optimization, where more information on the trade-offs and their consequences are available. A second reason to use MOO is that there may be a multitude of competing or (seemingly) incompatible objectives by different stakeholders. Such objectives may as well evolve and change over time, especially in the design of long-term groundwater monitoring networks [e.g., *Loaiciga et al.*, 1992; *Reed and Kollat*, 2012].

[21] Looking at binary decisions leads to classical single-objective optimization, just like most of optimal design theory or data worth concepts. We restrict our current work to the single-objective context, for now looking at the case where a planned field campaign should chiefly support a

single decision problem in a one-time field effort. Still, nothing restricts our approach from integration into MOO approaches in future work, e.g., if other objectives coexist, or for a detailed trade-off analysis between improved decision confidence and costs of the field campaign.

[22] The hypotheses or decision problems that can be supported with our approach include, e.g., model validity, model choice for geostatistical, physical, chemical, or biological model assumptions, parameterization forms or closure assumptions, compliance with environmental performance metrics or other model predictions that are related to binary decisions, and reliability questions in optimization and management. In a synthetic test case for the sake of illustration, we feature the prediction of compliance with maximum contaminant levels as an environmental performance metric, looking at contaminant transport to an ecologically sensitive location through a two-dimensional heterogeneous aquifer with uncertain covariance function and uncertain boundary conditions.

2. General Approach and Mathematical Formulation

[23] Assume that either a modeler, scientist, or a manager is asked to provide a yes/no decision on a proposed statement. Due to the inherent uncertainty of the problem-at-hand, the answer can only be found at a limited confidence level. The outline of our approach for such situations is as follows:

[24] 1. To cast the corresponding yes/no question or binary decision into a hypothesis test (section 2.1).

[25] 2. To insert the hypothesis test into the Bayesian framework, which yields the probability of making a false decision (section 2.2).

[26] 3. To analyze the expected reduction of error probability through planned field campaigns as criterion for optimal design (section 2.3).

[27] 4. To minimize the error criterion by optimizing the field campaign (section 2.3 and section A3).

[28] The obtained sampling schemes allow the proposed hypotheses (and the final decision) to be affirmed or refuted at the desired confidence and at minimum costs for the field campaign. In the following we use a most generic formulation. We will illustrate our methodology based on one scenario with two different levels of complexity in sections 3, 4, and 5.

2.1. Classical Hypothesis Testing

[29] This section summarizes the key steps of classical hypothesis testing and introduces the notation, before we move on to Bayesian hypothesis testing and optimal design. The well-known individual steps of hypothesis testing are [e.g., *Stone*, 1996; *Casella and Berger*, 2002]:

[30] 1. Identify the null hypothesis H_0 and the alternative hypothesis H_1.

[31] Null hypothesis H_0: A fallback assumption H_0 on some target variable q holds.

[32] Alternative hypothesis H_1: A (desirable) assumption H_1 on q is true.

[33] H_0 is the hypothesis that is accepted for the time being, while the burden of proof is on H_1, which is the hypothesis one desires to prove [e.g., *Shi and Tao*, 2008]. Per definition, falsely accepting H_1 is the more critical type of error. This calls for sufficient evidence (i.e., statistically significant data) before accepting H_1 over H_0, and coincides well with the precautionary principle in policy, environmental sciences, and the public health sector [e.g., *Kriebel et al.*, 2001].

[34] 2. Choose a level of significance $\alpha \in [0\%, 100\%]$. Typically, α should be small (e.g., $\alpha \leq 10\%$). It is the probability of falsely accepting H_1 although H_0 is in fact true, and so controls the worse type of test failure. The quantity $\beta = 1 - \alpha$ is often called the power of the test.

[35] 3. Decide what type of data are appropriate to judge H_0 and H_1, and define the relevant test statistic T that can be computed from the data.

[36] 4. Derive the distribution $p(T)$ of the test statistic, while considering all statistical assumptions (independence, distribution, etc.) of the data. More accurately, $p(T)$ should be denoted as $p(T|H_0)$ because it is the distribution of T that can be observed if H_0 was in fact true. Since T is a sample statistic, $p(T|H_0)$ will depend, among other things, on the sample size N.

[37] 5. The significance level α and the character of the test (one-sided, two-sided) partitions the distribution $p(T|H_0)$ into the critical or rejection region (reject H_0), and the acceptance region (accept H_0 due to lacking statistical evidence against it).

[38] 6. Evaluate the observed value T_{obs} of the test statistic from the data.

[39] 7. Depending on the value of T_{obs}, decide on the two hypotheses:

[40] Decision D_0: Accept H_0 if T_{obs} is outside the critical region of $p(T|H_0)$.

[41] Decision D_1: Else, reject H_0 in favor of H_1.

[42] Either of the decisions D_0 and D_1 could be false, leading to the corresponding error events:

[43] α error E_α: Decision D_1 was taken (H_1 accepted) although hypothesis H_0 is in fact true (called *false positive* of H_1).

[44] β error E_β: Decision D_0 was taken (H_0 accepted) although hypothesis H_1 is in fact true (called *false negative* of H_1).

[45] Within classical hypothesis tests, these errors can only be quantified via their *conditional* probabilities:

$$Pr[D_1|H_0] \equiv P_\alpha, \qquad Pr[D_0|H_1] \equiv P_\beta. \tag{1}$$

The significance level α is the maximum acceptable conditional probability P_α for the α error. A total (unconditional) error probability cannot be specified for two reasons. First, P_β can only be assessed if the alternative hypothesis H_1 is sufficient to infer a sampling distribution $p(T|H_1)$. For example, using an inequality in H_1 would be insufficient. Second, one would need prior probabilities $Pr[H_0]$ and $Pr[H_1]$ of H_0 or H_1 being true to remove the conditions on H_0 and H_1 in equation (1).

2.2. Bayesian Hypothesis Testing

[46] With Bayesian hypothesis testing [e.g., *Press*, 2003], one can assess, for all possible hypotheses and the resulting decisions, the probability of being false both

before and after considering the data. If desired, one could even test more than two competing hypotheses at once.

[47] Given the prior probabilities $Pr[H_0]$ and $Pr[H_1]$ of H_0 or H_1 being true, respectively, and a formulation of the hypotheses that defines the sampling distribution of any data vector $\mathbf{y}$ under each hypothesis, Bayes rule yields for the conditional probabilities of the hypotheses:

$$P_i = Pr[H_i|\mathbf{y}] = p(\mathbf{y}|H_i) \cdot Pr[H_i]/p(\mathbf{y}), \qquad i = \{0, 1\}. \tag{2}$$

[48] Then, one can use a Bayesian decision rule [e.g., *Berger*, 1985] and infer the total probability of committing any of the two errors:

[49] Decision D_0: Accept H_0 if $Pr[H_0|\mathbf{y}] \geq \alpha$.

[50] Decision D_1: Else, reject H_0 in favor of H_1.

[51] Individual risks: The overall risks R_α and R_β to commit an α or β error (E_α, E_β) are

$$\begin{aligned} R_\alpha &= Pr[E_\alpha|\mathbf{y}] = Pr[D_1|H_0, \mathbf{y}]Pr[H_0|\mathbf{y}] \equiv P_\alpha|\mathbf{y} \cdot P_0|\mathbf{y}, \\ R_\beta &= Pr[E_\beta|\mathbf{y}] = Pr[D_0|H_1, \mathbf{y}]Pr[H_1|\mathbf{y}] \equiv P_\beta|\mathbf{y} \cdot P_1|\mathbf{y}. \end{aligned} \tag{3}$$

[52] Total risk R: the total risk R of committing any of the errors (sometimes called the Bayes risk) is

$$\begin{aligned} R &= R_\alpha + R_\beta = P_\alpha P_0 + P_\beta P_1, \\ R(\mathbf{y}) &= R_\alpha(\mathbf{y}) + R_\beta(\mathbf{y}) = P_\alpha|\mathbf{y} \cdot P_0|\mathbf{y} + P_\beta|\mathbf{y} \cdot P_1|\mathbf{y}, \end{aligned} \tag{4}$$

where the first and second lines are the prior and conditional versions, respectively. A weighted version is given by

$$R = w_\alpha R_\alpha + w_\beta R_\beta\,, \tag{5}$$

where the weights w_α and w_β reflect the severeness of the α error versus the β error. The link to utility theory [e.g., *Fishburn*, 1970] would be to assign weights that quantify the respective losses, e.g., in monetary units. The limiting cases for different weighting are $R = R_\alpha$ and $R = R_\beta$, reflecting the preference to never falsely assume safety, or to never sound a wrong alarm.

[53] In Bayesian hypothesis testing, the significance level α is derived from different considerations than in classical tests. Typically, Bayesian decision rules [e.g., *Berger*, 1985] find a significance level α such that, for the given weighting or loss functions in equation (5), the total Bayes risk according to equation (5) is minimal. This decision rule is called the Bayes decision criterion by *Jaynes* [2003]. In the simple case that no weighting is applied, a significance level of $\alpha = 50\%$ is optimal. Thus, in absence of recommendations through regulations or legislation, $\alpha = 50\%$ is the value we recommend, because it encodes an unbiased search for "the truth," without articulating any specific preferences between the alpha or beta error.

2.3. Hypothesis-Driven Optimal Design

[54] We will now derive the framework that optimizes field campaigns to support Bayesian decision (hypothesis testing) problems with data for maximum confidence. To this end we minimize the total risk R of false decision (see equation (5)). We assume that the decision rule (here the value of α) is fixed a priori. Therefore we consider the impact of additional planned data on the total risk R by taking the difference between the first and the second line of equation (4).

[55] Assume that the initial Bayesian risk R_0 of false decision (prior to additional sampling) is too large for the task or purpose of the modeler, scientist, or manager. The maximum acceptable risk shall be $R_{\max}$. In such cases it is only natural to try and improve one's knowledge by collecting data. Additional data will have the power to narrow down all relevant distributions of parameters, predictions, and of the relevant sampling distributions upon which the decision between H_0 and H_1 is based. This can make the final decision between H_0 and H_1 more defensible, as illustrated schematically in Figure 1, up to the point where $R \leq R_{\max}$.

[56] In classical univariate hypothesis testing, the sampling distribution $p(T|H_0)$ of the test statistic T depends on sample size. In the (geo)statistical inverse context, distributions and distribution shapes also depend on many other factors [e.g., *Rubin*, 2003]. These include sampling locations, data types, the nonlinearity of governing equations

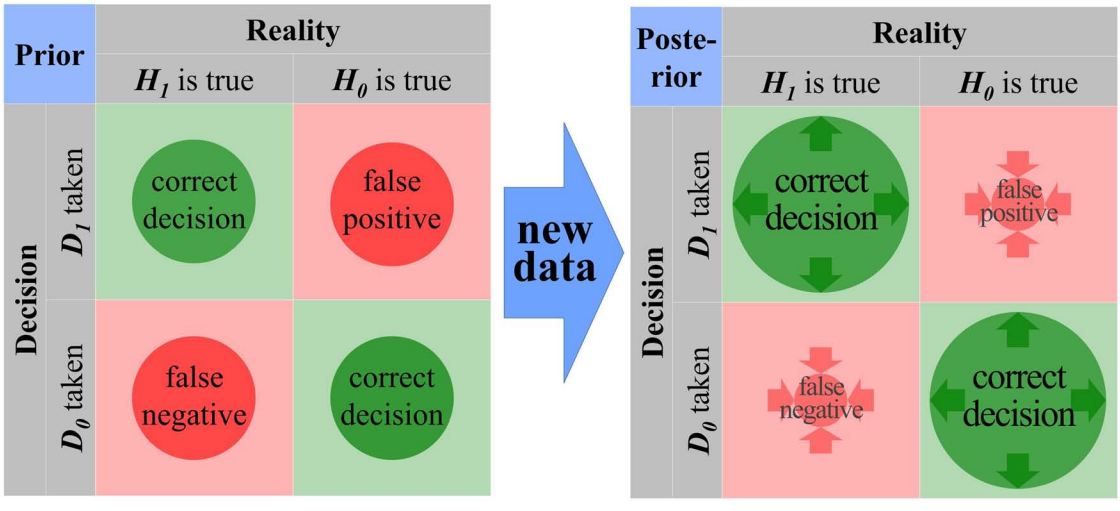

Figure 1. The objective of hypothesis-driven field campaigns.

and of the inversion, and the circumstances of data collection, e.g., the design of tracer chemistry and injection setup, or the strength and type of any other system excitation versus passive system observation.

[57] The entire strategy, layout, and schedule of the field campaign is typically termed an experimental design. All corresponding design parameters are lumped together in a formal vector $\mathbf{d}$. This vector contains, e.g., a list of sampling locations and data types to be collected, as well as parameters that control the circumstances of data acquisition.

[58] When collecting data according to some design $\mathbf{d}$, one obtains a data vector $\mathbf{y}_\mathbf{d}^*$. The planning phase prior to sampling is often called the preposterior stage [e.g., *Ben-Zvi et al.*, 1988; *James and Gorelick*, 1994; *Raiffa et al.*, 1995]. During that phase the yet unmeasured data values are conceptualized as random values drawn from a distribution $p(\mathbf{y}_\mathbf{d})$. Knowing $p(\mathbf{y}_\mathbf{d})$ requires an underlying model with statistical prediction capability for potential data, e.g., based solely on prior statistics, or on conditional statistics that honor all data already available.

[59] Given a specific data vector $\mathbf{y}_\mathbf{d}^*$, one could condition all involved distributions on $\mathbf{y}_\mathbf{d}^*$: the distribution $p(q)$ for the target quantity could be updated to $p(q|\mathbf{y}_\mathbf{d}^*)$, and the probabilities P_0 and P_1 (of H_0 and H_1 being true, respectively) would also change. Thus, following equations (3) and (4), the total risk R becomes a function of $\mathbf{y}_\mathbf{d}^*$:

$$R(\mathbf{y}_\mathbf{d}^*) = P_\alpha|\mathbf{y}_\mathbf{d}^* \cdot P_0|\mathbf{y}_\mathbf{d}^* + P_\beta|\mathbf{y}_\mathbf{d}^* \cdot P_1|\mathbf{y}_\mathbf{d}^*\,, \tag{6}$$

where $P_{(\cdot)}|\mathbf{y}_\mathbf{d}^*$ is the conditional version of probability $P_{(\cdot)}$. Following decision-theoretic ideas, we define the decision utility ϕ^* of the data set $\mathbf{y}_\mathbf{d}^*$ as the reduction of total risk:

$$\phi^*(\mathbf{y}_\mathbf{d}^*) = R_0 - R|\mathbf{y}_\mathbf{d}^*, \tag{7}$$

where R_0 is the initial decision risk in absence of additional data evaluated according to the first line of equation (4). The data set $\mathbf{y}_\mathbf{d}^*$ is informative if the conditional risk is smaller than the initial risk R_0, and it is sufficient if $R|\mathbf{y}_\mathbf{d}^* \le R_{\max}$.

[60] Unfortunately, for nonlinear statistical inference, the shape of conditional distributions (including their critical tails) depends on the actual values of data, which are yet unknown at the stage of planning the design. For each possible data set $\mathbf{y}_\mathbf{d}^*$ from $p(\mathbf{y}_\mathbf{d})$ for a given design $\mathbf{d}$, a different risk reduction $\phi^*(\mathbf{y}_\mathbf{d}^*)$ may result, yielding an entire distribution $p(\phi|\mathbf{d})$ for the design's utility. We therefore invoke the concept of expected utility [*Schoemaker*, 1982; *Raiffa et al.*, 1995] by marginalizing equation (7) over all possible data values $\mathbf{y}_\mathbf{d} \sim p(\mathbf{y}_\mathbf{d})$, and obtain

$$\phi(\mathbf{d}) = \int_{\mathbf{y}_\mathbf{d}} \phi^*(\mathbf{y}_\mathbf{d}) p(\mathbf{y}_\mathbf{d})\, d\mathbf{y}_\mathbf{d}, \tag{8}$$

where $\phi(\mathbf{d})$ is the expected utility of the design. The conditional expected risk is then

$$E[R|\mathbf{d}] = R_0 - \phi(\mathbf{d})\,. \tag{9}$$

Evaluating equation (9) requires estimation of the possible conditional distributions of the target quantity q for all possible data sets $\mathbf{y}_\mathbf{d}$, and evaluation of the decision risk in each case. This task calls for nonlinear Bayesian inference schemes such as PreDIA [*Leube et al.*, 2012] (see also sections 4.3 and Ax).

[61] When using equation (8) as objective function for optimizing the design $\mathbf{d}$, the resulting design will be optimal (in the expected value sense) in supporting a most confident decision on the proposed hypotheses. The reliability of a design to actually deliver the promised (expected) utility and related possible modifications of equation (8) for improved robustness as well as the relation between α and the possible posterior risk values are discussed in section 5.3.

[62] Optimal design theory and its geostatistical branch know an entire list of optimality criteria that mostly work on the conditional covariance matrix of model parameters [e.g., *Pukelsheim*, 2006; *Müller*, 2007; *Nowak*, 2010], often called the optimality alphabet [e.g., *Box*, 1982]. To distinguish our criterion from the existing ones, we call it the R criterion. We denote the extreme end-members that consider only R_α or R_β (see equation (5)) as R_α optimality and R_β optimality, and the case without weighting as $R_{\alpha,\beta}$ optimality.

[63] The formal optimization task is

$$\mathbf{d}_{opt}^{(R)} = \underset{\mathbf{d}\in\Omega_\mathbf{d}}{\arg\max}\ \phi(\mathbf{d})\,, \tag{10}$$

with the space of allowable designs $\Omega_\mathbf{d}$. The superscript (R) in the expression above can be either R_α, R_β, or $R_{\alpha,\beta}$. For any fixed cost of the design, the choice of data types and configurations can be optimized. Alternatively, for fixed $R_{\max}$, the cheapest sufficient design can be found. Several possible optimization algorithms are provided in section A3.

[64] In the presence of an already existing data set $\mathbf{y}_0$, all prior probabilities are simply exchanged for probabilities conditional on $\mathbf{y}_0$, without otherwise changing our proposed framework.

2.4. Bayesian (Geo)statistics and Model Averaging

[65] The resulting design will be optimal, conditional on all prior probabilities and assumptions that enter equations (3)–(5). The challenge is to provide a sufficient envelope for the uncertainties and errors that plague real field campaigns and hydro(geo)logical modeling efforts. This is a substantial challenge, especially if complex statistical assumptions in combination with hydro(geo)logical simulation models serve to provide the priors. In such situations it is advisable to explicitly account for uncertainties in model choice and parameters since this will provide designs that are robust to variations in the prior assumptions. This calls for approaches such as Bayesian model averaging [e.g., *Hoeting et al.*, 1999; *Neuman*, 2003] or Bayesian geostatistics [e.g., *Kitanidis*, 1986].

[66] The importance of considering parametric uncertainty and uncertain model choice in the context of geostatistical optimal design has recently been pointed out by *Nowak et al.* [2010] for geostatistical inverse problems, while *Diggle and Lophaven* [2006] and *Neuman et al.* [2012] performed similar studies restricted to the kriging-like context. If desired, our hypothesis-driven framework can also help selecting between different mathematical or conceptual model structures. This can be achieved by formulating the model choice in terms of hypotheses.

3. Simplistic Illustration: Contaminant Arrival Times

[67] In the following we will first illustrate our concept of *R*-optimal designs on a simplistic version of a synthetic case in order to discuss the principles without complication. A more complex version of this test case will follow in sections 4 and 5.

3.1. Simplistic Setup

[68] Consider a contaminant source within a 2-D homogeneous aquifer. We assume that the effective porosity n_e and the regional hydraulic gradient J are known and uniform, whereas the spatially constant value of log-transmissivity $Y = \log K$ is unknown. Authorities are concerned that the contaminant arrives at some sensitive location at a distance L downstream within the aquifer faster than a specified arrival time τ_0, and demand 95% confidence for rejecting their concern. Under the given circumstances, the arrival time τ is a so-called environmental performance metric (see section 3.2 of *de Barros et al.* [2012]) of interest and is given by

$$\tau = \frac{Ln_e}{KJ}\,, \tag{11}$$

where $K = \exp(Y)$ [e.g., *Rubin*, 2003]. We can rearrange this for the limiting value $Y = Y_0$ which leads to $\tau = \tau_0$. Using the data shown in Table 1, we obtain $Y_0 = 0.9$ (with K in units of m d^{-1}).

[69] Next, a site investigator is entrusted with the task to provide a suitable data acquisition strategy, such that one can determine with sufficient confidence (here 95%), whether or not the arrival time is smaller than τ_0. For simplicity we limit this example to collecting aquifer cores, assuming that they can provide conductivity data with uncorrelated Gaussian measurement error of variance σ_ε^2.

3.2. Application of Hypothesis-Driven Design

[70] First, we set up a hypothesis test and perform all steps as outlined in sections 2.1 and 2.2: Null hypothesis H_0: $Y \geq Y_0$: unsafe situation with $\tau \leq \tau_0$. Alternative hypothesis H_1: $Y < Y_0$: safe situation with $\tau > \tau_0$.

[71] Since human health is at risk, the challenge is to prove the safe situation (H_1), whereas the null hypothesis H_0 represents the more conservative choice. In accordance with the requested confidence of 95% typical for regulations by the US EPA [e.g., *USEPA*, 1989, 1991, 2001], we choose a significance level of $\alpha = 5\%$.

[72] When taking n_s samples of $Y = \log K$ we can use the sample mean m_Y as an estimate of Y and as the test statistic T. Because we face independent and normally distributed samples with known variance σ_ε^2, the distribution $p(T|H_0)$ of the test statistic $T = m_Y$ under H_0 is normal with mean Y_0 and sample variance $n_s^{-1}\sigma_\varepsilon^2$ [e.g., *Stone*, 1996]. Under these conditions we can simplify the procedure outlined in section A2 by analytical means.

[73] In order to drive the hypothesis test in a Bayesian manner, we need prior probabilities for H_0 and H_1. This can be achieved by assuming a distribution for Y, for example from world-wide data bases, expert elicitation or minimum relative entropy considerations [e.g., *Woodbury and Ulrych*, 1993; *Woodbury and Rubin*, 2000; *Hou and Rubin*, 2005]. Here we assume that $p(Y)$ is Gaussian with prior mean μ_Y and variance σ_Y^2. In our case, σ_Y^2 denotes uncertainty in the constant mean value of $Y = \log K$, not spatial variability. The data to be collected will have independent zero-mean Gaussian measurement errors $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. The values for μ_Y, σ_Y^2, and σ_ε^2 and several scenario variations for discussion are provided in Table 2. The data from Tables 1 and 2 yield prior probabilities of $Pr[H_0] = Pr[Y \geq Y_0] \approx 34\%$ and $Pr[H_1] = Pr[Y < Y_0] \approx 66\%$. For the base case with $\alpha = 5\%$, we decide a priori in favor of H_0, because $Pr[H_0] > 5\%$. This immediately yields that the initial risk of wrong decision according to equations (3) and (4) is $R_0 = P_\beta = 66\%$.

[74] Under the conditions adopted for this example, optimal design amounts to finding the minimal number n_s of core samples that need to be collected in order to achieve the requested confidence. Thus, the design vector $\mathbf{d}$ simplifies to $\mathbf{d} = [n_s]$ [e.g., *Raiffa et al.*, 1995; *Rasch et al.*, 2011]. Next, we will use the algorithm explained in section A2 with different sample sizes ($n_s = 1, \ldots, 50$) to assess the dependence of R on n_s, and perform this for all scenario variations listed in Table 2. Results are provided in section 3.3.

3.3. Results

[75] Figure 2 looks at the resulting expected reduction ϕ of total risk $R = R_\alpha + R_\beta$ as a function of sample size n_s. It indicates what n_s is necessary such that we can expect to support or refute the hypotheses with an error probability of $R \leq R_{\max}$, e.g., $n_s = 10$ for $R_{\max} = 5\%$ when $\sigma_\varepsilon^2 = 1/16$, $\sigma_Y^2 = 1$, $\mu_Y = 0.5$, and $\alpha = 5\%$ (see darkest line in Figure 2(a)). Seen from the viewpoint of multiobjective optimization, all plots in Figure 2 resemble Pareto fronts. They visualize the trade-off between the costs and benefits of sampling, i.e., between n_s and the reduced risk $R_0 - R(n_s)$.

[76] In general, the initial risk R_0 and $R(n_s)$ depend nonlinearly on the distance between the sampling statistic T and its critical value T_{crit}. This distance results from the relations between prior mean, prior variance, measurement precision, limiting value Y_0 and significance level α, as discussed in the following.

[77] 1. Measurement error: Figure 2(a) shows that, quite intuitively, more precise measurements satisfy the information needs faster. For accurate samples, a smaller number suffices to acquire the desired confidence.

Table 1. Data Used for Hypothesis Testing Example in Section 3

Name	Symbol	Value	Units
Gradient	J	0.04	–
Porosity	n_e	0.2	–
Error	σ_ε^2	1/4	–
Travel distance	L	500	m
Critical arrival time	τ_0	1000	day

Table 2. Scenario Variations for the Simplistic Example Used in Section 3[a]

Name	Symbol	Values	Units
Prior mean of Y	μ_Y	−0.5, 0, **0.5**, 1, 1.5	–
Prior variance of Y	σ_Y^2	1/4, 1/2, **1**, 2, 4	–
Measurement error	σ_ε^2	1/16, 1/8, **1/4**, 1/2, 1	–
Significance level	α	1, **5**, 50, 95, 99	%

[a]Bold numbers refer to the base case scenario.

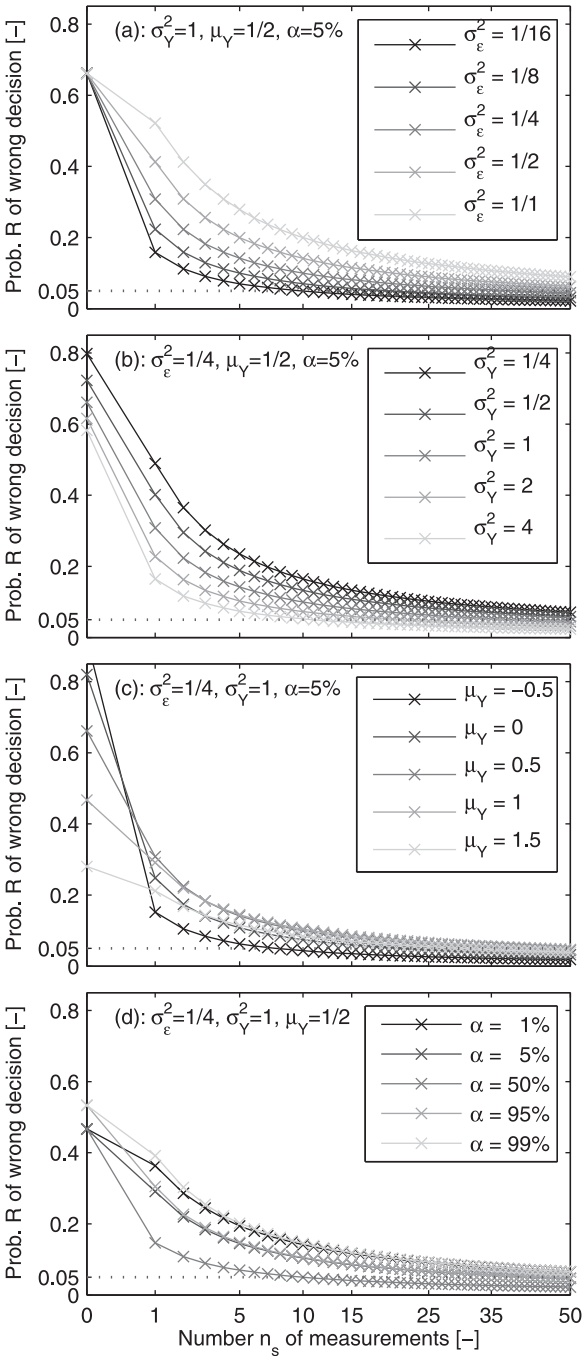

Figure 2. Probability R of wrong decision (both α and β error, equation (4)) versus the number n_s of measurements for the scenario defined in section 3.2 with data from Tables 1 and 2. Plot (a) represents different measurement precision; plot (b) describes the dependency on different prior variances; plot (c) illustrates the behavior for different prior mean values, and (d) shows variations with the significance level α.

[78] 2. Prior uncertainty: Figure 2(b) may be surprising at first sight. A smaller prior variance σ_Y^2 leads to an increased risk of wrong decision for any given n_s. The reason is that μ_Y lies within the region of H_1. Thus, decreasing σ_Y^2 increases the prior probability of H_1 as we concentrate $p(Y)$ more and more within the region of H_1. Since H_1 is initially being rejected over the entire explored range of scenario conditions in Figure 2(b), this leads to an increase of decision risk.

[79] 3. Prior mean: Figure 2(c) shows that the initial risk R_0 is larger for smaller μ_Y. This is because smaller values of μ_Y move $p(Y)$ more into the region of H_1, with the same arguments following as above. For increasing n_s and different values of μ_Y in Figure 2(c), the different curves for $R(n_s)$ cross each other. The reason can be seen from Figure 3, which shows the dependence of decision risk on a standardized normal test statistic z: at the transition from some acceptance region of H_0 ($z < z_{\text{crit}}$) to some rejection region ($z \geq z_{\text{crit}}$), the decision risk jumps from $R = 1 - \alpha$ to $R = \alpha$, i.e., from 95% to 5% in the shown example. The closer the limiting value z_0 is to z_{crit} (but still below), the easier it is for additional data to move the observed value of z into the rejection region with $R \leq 5\%$. Thus, R decreases faster with n_s when the prior expected value μ_Y is just below the limiting value Y_0. This behavior is a direct outcome of the binary character of decisions in conjunction with the strong jump of error risk at the critical value $z = z_{\text{crit}}$. The jump that causes this effect disappears when choosing $\alpha = 50\%$, i.e., when the hypothesis test setup does not put the burden of proof on either H_0 or H_1.

[80] 4. Significance level: Figure 2(d) shows how $R(n_s)$ changes with the significance level α. The levels of $\alpha = 1\%, 5\%, 50\%$ yield a common initial risk of $R_0 = 46.67\%$ for the given example. For all three cases, R_0 is comprised solely from R_β. The prior decision is D_0 (that H_0 should be true), and $R_0 = 46.67\%$ is the probability P_1 that, in reality, H_1 is true. For $\alpha = 95\%$ and $\alpha = 99\%$, the decision in absence of data flips to D_1 (that H_1 should be true), and the common value $R_0 = 53.33\%$ is the probability $P_0 = 1 - P_1$ that H_0 is the actual truth. After several samples collected (in fact, already after the first), the risks in the different scenarios change their rank: the values $R(n_s \geq 1)$ for $\alpha = 1\%$ and $\alpha = 99\%$ are now the largest ones, the values for $\alpha = 50\%$ are the lowest, and the values for $\alpha = 5\%$ and $\alpha = 95\%$ fall in between. The reason is that $\alpha = 1\%$ and $\alpha = 99\%$ represent the most demanding decision rules, requiring very strong statistical evidence for H_0 or H_1, respectively. Contrary to that, the other values of α require only weaker statistical evidence. Therefore, the

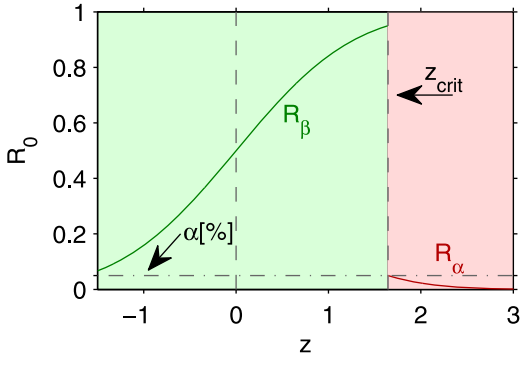

Figure 3. The initial decision risk R depends very nonlinearly on z (a generic standardized normal test statistic). Red shaded area: rejection region of H_0 for $\alpha = 5\%$. Green shaded area: acceptance region of H_0.

decision rule behind $\alpha = 50\%$ triggers information needs that are easiest to satisfy.

[81] Altogether, this example demonstrates that hypothesis-driven design can optimize field campaigns (here: simply find the number of samples to collect) to maximize the probability of finding the correct answer for a specified hypothesis. It does so by combining all aspects of proper hypothesis testing with Bayesian updating and Bayesian decision theory.

4. Test Case

[82] Now we illustrate the methodology on a more complex version of the example, featuring a significant amount of uncertain parameters. Again, the goal is to support predictions of arrival time τ with data in order to come below a maximum allowable risk of wrong decision. For the current scenario we use numerical simulations of contaminant arrival time (details given in section 4.1) within a Monte Carlo framework, and consider an extensive list of sources for uncertainty (details again in section 4.1), in accordance with the discussion in section 2.4. In section 4.2 we will describe how we applied our hypothesis-driven framework to this situation. We present and discuss the results in section 5.

4.1. Physical Formulation and Setup

[83] Consider a steady state, 2-D depth-averaged flow in a heterogeneous aquifer with locally isotropic transmissivity K [L^2/t]. We assume a scenario free of sinks and sources. Under these conditions, the flow equation is given by

$$\nabla \cdot [K(\mathbf{x})\nabla h(\mathbf{x})] = 0\,, \tag{12}$$

with h [L] denoting hydraulic head and $\mathbf{x} = (x_1, x_2)$. The domain is rectangular with $L_1 \times L_2 = 200\,\mathrm{m} \times 200\,\mathrm{m}$. Boundary conditions for flow will be discussed below. We wish to predict the bulk arrival time ($\tau \equiv t_{50}$) from a continuous-release line source along the upstream domain boundary $(x_1 = 0)$ to a sensitive location at $\mathbf{x}_S = (L, 100\,\mathrm{m})$. Top this end, we use the groundwater age equation described by *Goode* [1996]:

$$\begin{aligned} &\mathbf{v} \cdot \nabla\tau = \nabla \cdot [\mathbf{D}_d \nabla\tau] + 1, \\ &\mathbf{v} = \mathbf{q}/n_e\,, \end{aligned} \tag{13}$$

subject to $\tau = 0$ on the upstream boundary and zero-gradient boundaries everywhere else. Here $\tau(\mathbf{x})[t]$ is the arrival time at any point $\mathbf{x}$ within the domain, $\mathbf{v}$ [L/t] is the effective transport velocity, $\mathbf{q}$ [L/t] is the specific discharge given by Darcy's law, n_e[–] is the effective porosity, and $\mathbf{D}_d$ [L^2/t] is the local porescale dispersion tensor [*Scheidegger*, 1954].

[84] Following our arguments from section 2.4, we choose the Matérn covariance function [*Matérn*, 1986; *Handcock and Stein*, 1993] with uncertain covariance parameters to model the heterogeneity of log-transmissivity $Y = \ln K$. The advantage of the Matérn covariance is that it encompasses a family of covariance functions within a single expression that depends on its shape parameter κ. For specific values of κ, the Matérn function recovers the exponential, Whittle, and Gaussian covariance as special cases [e.g., *Handcock and Stein*, 1993]. This maps structural model uncertainty onto a set of uncertain parameters, which is called continuous Bayesian model averaging [*Nowak et al.*, 2010].

[85] We superimpose a trend model with uncertain linear trend components along the directions of the x_1 and x_2 coordinates. We also admit uncertainty in flow boundary conditions via a regional gradient with uncertain slope J and rotation angle α_h relative to the x_1 axis, and assign corresponding fixed-head conditions to all four boundaries. All relevant parameter values and their uncertainty are listed in Table 3.

4.2. Hypothesis-Driven Design

[86] Following the approach described in section 2 and illustrated in section 3, we will now formulate the hypotheses directly in terms of the target quantity q for decision, i.e., in terms of arrival time $\tau \equiv q$:

[87] Null hypothesis H_0: $\tau < \tau_0$.

[88] Alternative hypothesis H_1: $\tau \geq \tau_0$.

[89] In the current example arrival time is our test statistic $T \equiv \tau$. In the previous case (see section 3.2), we featured a globally constant value of $Y = \ln K$ as test statistic and as the only unknown and observable quantity. All data were repeated measurements of the very same quantity, and the test statistic was a simple arithmetic average of the collected data. In the current example we consider hydraulic heads $h(x)$ and log-transmissivities $Y(x) = \ln K(x)$ as possible data types. Arrival time and our two data types are three different physical quantities, and the latter vary in space due to heterogeneity and boundary conditions.

[90] Now, the distribution $p(\tau)$ and the related sampling distributions $p(\tau|H_0)$ and $p(\tau|H_1)$ have to be constructed from prior assumptions and auxiliary models. The conditional distributions $p(\tau|\mathbf{y})$ for possible data values $\mathbf{y}$ reflect the entire uncertainty involved in inferring the *true* value of τ from prior assumptions and from scarce, erroneous additional data. In section 3.2 the main uncertainty originated from measurement errors. In the current example we additionally feature overall system uncertainty and a target

Table 3. Parameter Values Used for the Synthetic Test Case[a]

	Numerical Domain		
Domain size	$[L_1, L_2]$	m	[200, 200]
Grid spacing	$[\Delta_1, \Delta_2]$	m	[0.4, 0.8]
	Transport Parameters		
Head gradient	J	–	$\mathcal{U}(0.005, 0.015)$
Head angle	α_h	%	$\mathcal{N}(0, 5)$
Effective porosity	n_e	–	0.35
Local dispersivities	$[\alpha_\ell, \alpha_t]$	m	[1, 0.1]
Diffusion coefficient	D_m	$\mathrm{m^2\,s^{-1}}$	10^{-9}
	Geostatistical Model Parameters		
Global mean	$\mu_Y = \ln K_g$	$\ln(\mathrm{m^2\,s^{-1}})$	$\mathcal{N}(-7.2, 1)$
Trend in x_1	β_1	$\ln(\mathrm{m^2\,s^{-1}})$	$\mathcal{N}(0, 0.5)$
Trend in x_2	β_2	$\ln(\mathrm{m^2\,s^{-1}})$	$\mathcal{N}(0, 0.5)$
Variance	σ_Y^2	$\ln^2(\mathrm{m^2\,s^{-1}})$	$\mathcal{N}(2, 0.5)$
Integral scales	$[\lambda_1, \lambda_2]$	m	$[\mathcal{N}(20, 5), \mathcal{N}(20, 5)]$
Matérn's shape parameter	κ	–	$\mathcal{N}(3, 0.75)$
	Measurement Error Standard Deviations		
$\ln T$	$\sigma_{r,T}$	–	0.5
Head ϕ	$\sigma_{r,\phi}$	m	0.05

[a] $\mathcal{U}$ is uniform distribution with lower and upper bound provided. $\mathcal{N}$ is normal distribution with mean and SD provided, truncated at zero for σ_Y^2, λ_1, λ_2, and κ (according to minimum relative entropy considerations).

quantity that is not observable, but needs to be inferred via Bayesian updating.

[91] H_1 will be rejected in favor of H_0 if the critical value $\tau_0 = 50$ d lies outside the acceptance region of $p(\tau)$ (or of $p(\tau|\mathbf{y})$, if additional data are available). The acceptance region for H_1 is again defined by the significance level α (taken again as $\alpha = 5\%$). The possible decision errors in the current situation are:

[92] α error E_α: Accepting that $\tau > \tau_0$ according to H_1, although $\tau \leq \tau_0$ (falsely assuming a safe situation).

[93] β error E_β: Accepting that $\tau \leq \tau_0$ according to H_0, although $\tau > \tau_0$ (falsely issuing an alert).

[94] Following our hypothesis-driven framework, the objective function for optimizing the design is the total risk $R = R_\alpha + R_\beta$ of committing any of the two decision errors (see equation (10)), again using a maximum acceptable risk of decision error $R_{\max} = 5\%$. For the sake of later discussion, we produce R-optimal sampling patterns for different travel distances $L = 100$, 120, 140, and 160 m between the upstream boundary and the sensitive location $\mathbf{x}_S = (L, 100\,\mathrm{m})$, and for different maximum numbers of samples ($n_s = 2, 4, 6, 8, 9, 10$) that may be placed within the aquifer.

4.3. Implementation

[95] The flow and travel time equations from section 4.1 are solved numerically using the code described by *Nowak et al.* [2008] and *Nowak et al.* [2010]. We obtain the prior arrival time probability distribution $p(\tau)$ via Monte Carlo simulation with $n_r = 40{,}000$ realizations, based on the parameter values listed in Table 3. To evaluate all the possible conditional distributions of travel time $p(\tau|\mathbf{y})$ for any possible data set $\mathbf{y}$, the required test statistics and the objective function (equation (10)) according to the scheme outlined in sections A2 and A1, we use the PreDIA framework by *Leube et al.* [2012].

[96] For optimizing the design we chose a simulated annealing (SA) algorithm [e.g., *Laarhoven and Aarts*, 1992] from the possible algorithms listed in section A3. The design parameters to be optimized were the number of samples, their data types, and their spatial coordinates. Because the optimization carried out here mainly serves to illustrate our hypothesis-driven approach with results, we leave out all further details on setup and implementation.

5. Results and Discussion

5.1. Optimal Measurement Locations

[97] Now we will discuss what design patterns are optimal to feed the information needs of the hypothesis test by relating the resulting patterns to the underlying flow and transport statistics. Figure 4 shows the resulting sampling pattern for the sensitive location placed at $\mathbf{x}_S = (140\,\mathrm{m}, 100\,\mathrm{m})$ and with $n_s = 10$ and $R = 5.00\%$.

[98] For this scenario variant, $n_s = 10$ is merely sufficient to achieve the desired maximum probability of wrong decision $R_{\max} = 5\%$. This can be seen in Figure 5, and will be discussed in more detail in section 5.2. When gradually stepping up from $n_s = 1$ to $n_s = 10$ samples, our framework first places six transmissivity samples, and then four head measurements. How these 10 measurements helped to reduce parametric uncertainty (for parameter definitions, see Table 3) is shown in Figure 6. Here uncertainty reduction is expressed as expected entropy of parameter groups relative to their prior entropy [cf. *Nowak et al.*, 2010], see also the information yield curves by *de Barros and Rubin* [2008] and *de Barros et al.* [2009].

[99] The six transmissivity samples are located between the upstream boundary and the sensitive location, helping to condition the transmissivity field locally. They are transversely more scattered near the upstream boundary, because the origin of solutes arriving at the sensitive location is uncertain. Both heterogeneity and the uncertain angle α_h of the regional gradient contribute to this effect. These six samples also help to identify the structural parameters of log-transmissivity (the mean μ_Y, the trend parameters β_1 and β_2, and the covariance parameters σ_Y^2, λ_1, λ_2, and κ, see Table 3). Due to their wide spacing, at typical lag distances of more than 30 m, they are practically uncorrelated. This is powerful for inferring the mean and the trend coefficients, as can be seen in Figure 6. They are also somewhat useful to infer the variance σ_Y^2. Due to the large lags, this design will not be very helpful to reliably infer values of the integral scales λ_1 and λ_2, but it will help to detect whether there is long-distance correlation or not. Certainly the design shown here does not support inference of the covariance shape parameter κ. This is consistent with the fact that μ_Y and σ_Y^2 are the dominant parameters in arrival time uncertainty [*Rubin and Dagan*, 1992a]. As a result, one can see in Figure 6 that the covariance parameter group is addressed less by this design.

[100] A total of four out of the ten locations are designated for measurements of hydraulic heads. These locations mainly address the uncertainty caused by the regional head gradient J. They are placed at maximum mutual distance in the respective corners of the domain, for two reasons: First, this arrangement forms a duplicated measurement of a longitudinal head gradient. Duplication helps to suppress measurement error. Second, their wide spacing protects the inference of the global head gradient J from the impact of mesoscale head fluctuations caused by heterogeneity in $Y = \ln K$. By matter of chance, this arrangement is also helpful to infer the angle α_h of the head gradient, although α_h has almost no impact on arrival times in our scenario. Overall, this leads to the quick reduction of uncertainty in the boundary condition (BC) parameter group visible in Figure 6 for the last samples placed.

5.2. Information Needs Change With Distance

[101] Figure 7 shows how the unconditional PDF (probability density function) of arrival time changes with distance to the inflow boundary. This PDF has been evaluated numerically from a Monte Carlo analysis using equations (12) and (13) with $n_r = 40{,}000$ realizations and the parameter values provided in Table 3. Our numerical results shown in Figure 7 indicate a PDF very close to a lognormal one. This is in agreement with results from the literature [e.g., *Rubin*, 2003; *Gotovac et al.*, 2010], although we consider an extended list of parametric uncertainty in boundary conditions and within the geostatistical model.

[102] Two properties of our test case setup modify the magnitude and character of information needs with travel distance to a sensitive location:

[103] 1. The variance of the arrival time PDF depends on travel distance [cf. *Rubin and Dagan*, 1992b; *Rubin*, 2003].

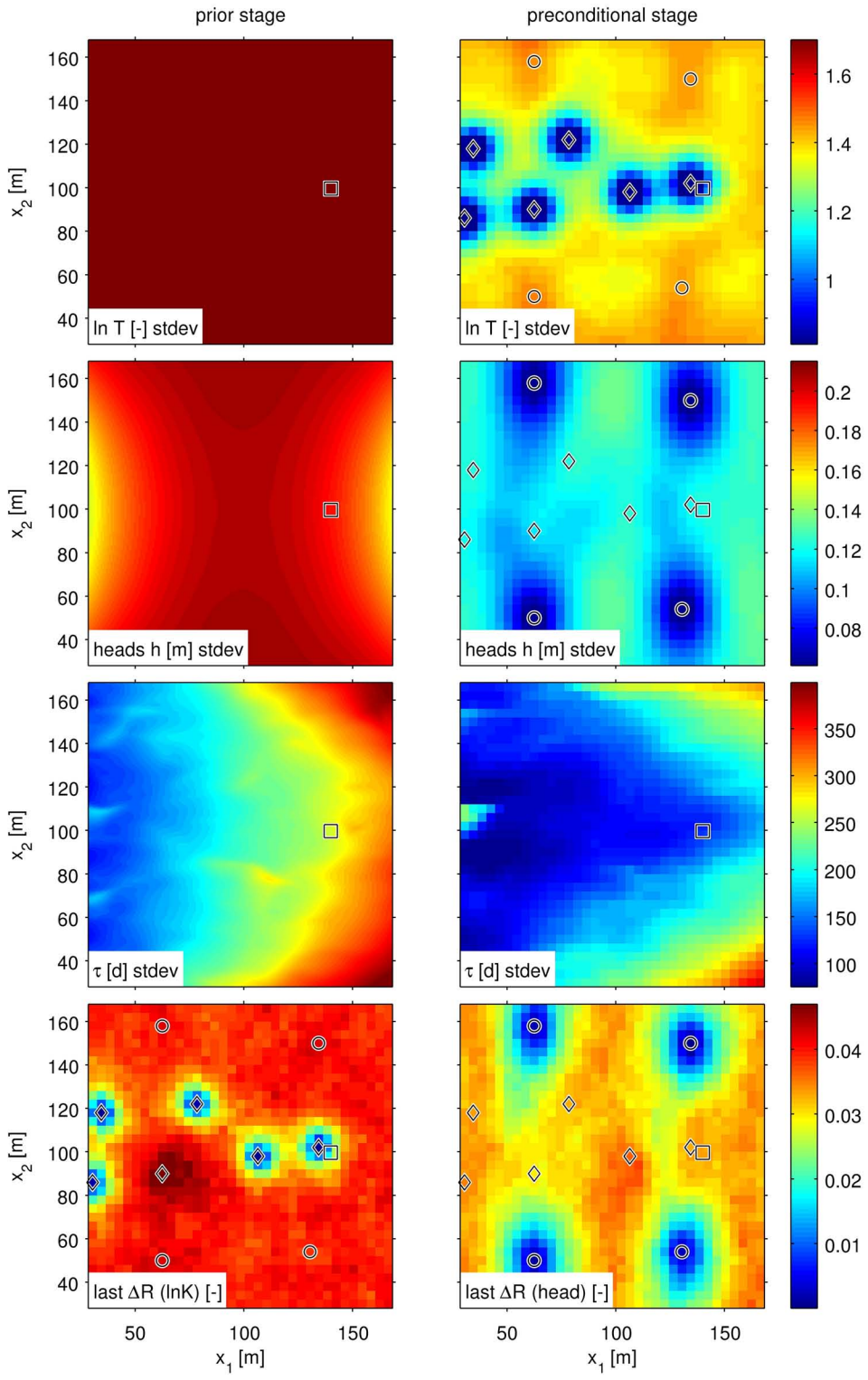

Figure 4. Statistics and sampling patterns for the scenario described in section 4 with the sensitive location at $x_1 = 140$ m. Parameter values are provided in Table 3. Circles: measurement locations of hydraulic heads; diamonds: measurement locations of $Y = \ln T$; square: sensitive location. Background maps show standard deviations of $Y = \ln T$, heads h and arrival time τ at the prior (left) and preposterior (right) stage. Bottom row shows reduction of decision risk R for each possible sampling location within the domain, as calculated for the final placement of the last measurement.

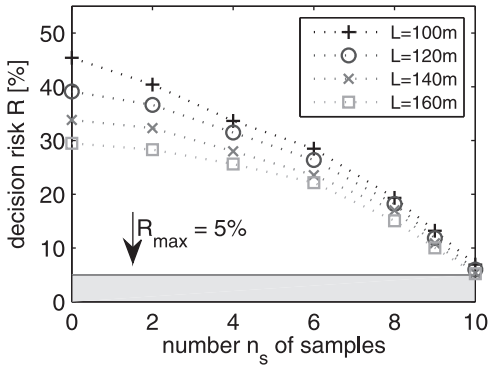

Figure 5. How the expected risk of wrong decision R decreases with number n_s of optimally placed samples for different travel distances L.

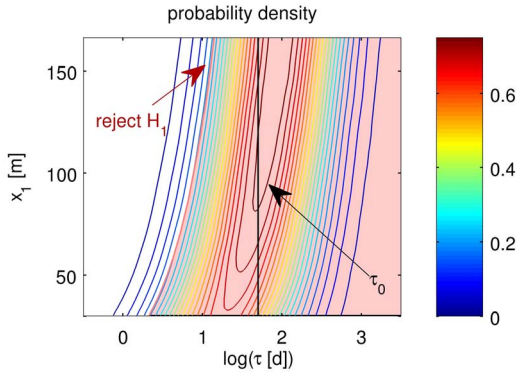

Figure 7. Prior probability distribution (PDF) of log-arrival time ($\log_{10}\tau$) as function of x_1 along the centerline of the domain. Thick black line: critical value $\log_{10}\tau = \log_{10}(50d)$. Red: rejection region of H_1. At the prior stage, H_1 is rejected everywhere within the domain, because the critical value of τ lies within the rejection region of H_1.

Specifically, the contribution of heterogeneity to arrival time variance fades with distance, in relative terms, against the contribution of parametric uncertainty (in the mean log conductivity μ_Y, the field variance σ_Y^2, the integral scales $\lambda_{1,2}$, the regional gradient J, and effective porosity n_e). Uncertainty in μ_Y, J, and n_e cause the arrival time variance to grow quadratically with travel distance (please be aware of the logarithmic scale in Figure 7), whereas the dispersion-related contribution of heterogeneity leads, asymptotically, only to a linear increase. The contribution of heterogeneity would even vanish in absolute terms, if we looked at averages of arrival time over large cross-sectional areas. Therefore, when predicting for distant and/or large sensitive locations, the reduction of parametric uncertainty by large-scale sampling patterns would be dominant, whereas local sampling for identification of heterogeneous patterns will dominate for close locations.

[104] 2. For sufficiently distant sensitive locations, late arrival (H_1) is almost sure. For very short travel distances, quick arrival (H_0) is almost sure. In both cases, only a few well-placed samples will suffice to achieve the desired confidence. The largest information needs will occur in tight cases, i.e., when the travel time distribution is concentrated tightly around the critical value (compare discussion in section 3.3).

[105] As a consequence of these two mechanisms, the dependence of $R(n_s)$ on n_s changes with distance. This is shown in Figure 5. The CDF value (cumulative distribution value) of the critical value of τ in the distribution $p(\tau)$ is the relevant quantity in the current hypothesis test, and determines the risk of wrong decision. Because this CDF value changes with different travel distances L, the initial risk R_0 (before adding a single measurement) changes with L (property 2 explained above). Then, with an increasing number n_s of samples placed, it is easier to satisfy the information needs for predicting the travel time to a sensitive location with less travel distance (according to property 1 explained above). That is why the curves $R(n_s)$ for smaller L decay faster than those for larger L. By pure chance, both effects cancel out at about $n_s = 10$, which is the end point of our scenario analysis because $R(n_s = 10) \approx 5\% = R_{\max}$ for all analyzed values of L.

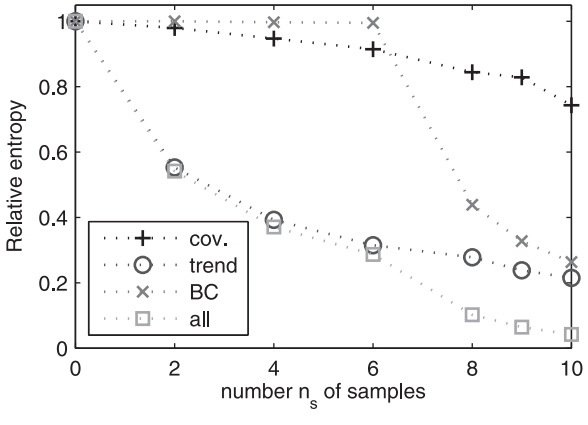

Figure 6. Reduction of parametric uncertainty during sequential placement of samples, expressed as relative entropy. Cov: covariance parameters σ_Y^2, λ_1, λ_2, and κ; trend: μ_Y, β_1, and β_2; BC: head gradient J and angle α_h; All: all nine parameters.

5.3. Preposterior Stage and Reliability of a Design

[106] In order to evaluate a given design candidate according to equation (10), hypothesis-driven design requires to assess the *expected* value of decision risk, i.e., to average over all possible data sets. The stage where data values are unknown and only random potential data values can be used is called the preposterior stage [e.g., *Ben-Zvi et al.*, 1988; *James and Gorelick*, 1994; *Raiffa et al.*, 1995; *Trainor-Guitton et al.*, 2011].

[107] Figure 8 shows potential cumulative distribution functions (CDFs) of arrival time at the preposterior stage for $n_s = 2$, 4 and 8 samples. One can clearly see how the CDFs become steeper with the increasing number of samples, reflecting higher levels of information. Still, the average over all possible conditional distributions will always yield the prior distribution:

$$E[p(\tau|\mathbf{y_d})] = p(\tau)\,, \tag{14}$$

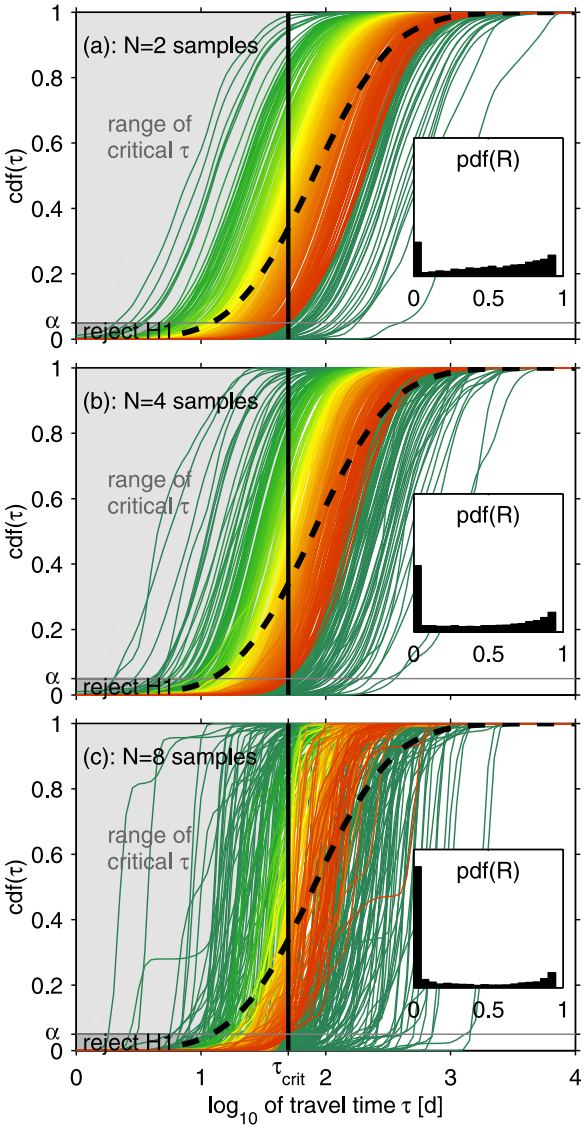

Figure 8. How hypothesis-driven design analyses potential data at the preposterior stage: Possible conditional distributions of arrival time given $n_s = 2$, 4, 8 samples at fixed (optimized) locations. Line color represents respective decision risk (red: 95%, green: 0%). Black dashed line: prior CDF. Inset histograms: statistics of decision risk for all the possible and yet unknown data values.

which means that the prior distribution is being conserved in the expected sense. The individual preposterior CDFs can become steeper only because their positions relative to each other become more and more spread. This is in direct correspondence to the well-known law of total variance:

$$\sigma_\tau^2 = E[\sigma_{\tau|\mathbf{y}}^2] + \mathrm{Var}[\mu_{\tau|\mathbf{y}}]\,, \tag{15}$$

which states that the prior variance is being conserved in the expected sense. Therefore, it is crucial that we first compute the individual decision risk values $R(\mathbf{y_d})$, and then average over all possible data sets $\mathbf{y_d} \sim p(\mathbf{y_d})$. If we evaluated the decision risk from the average of all preposterior CDFs (i.e., from the prior CDF), we would always obtain the prior value R_0.

[108] The color coding of the preposterior CDFs in Figure 8 illustrates the values R_i that would result from each CDF, if the respective random data values $\mathbf{y_{d,}}_i$ were the actual ones to be measured in future. The values R_i result directly from the cumulative probability values where each CDF crosses the $\tau = \tau_0$ line and from the respective decision taken by comparison with the significance level α, compare Figure 3. If a CDF crosses the $\tau = \tau_0$ line at cumulative probabilities $Pr[H_0 : \tau < \tau_0] > \alpha$, then $R_i = 1 - Pr[\tau < \tau_0]$. Conversely, if $Pr[H_0 : \tau < \tau_0] \leq \alpha$, then $R_i = Pr[\tau < \tau_0]$. Therefore, the largest possible risk value of $R_i = 1 - \alpha$ occurs, if the CDF value at τ_0 is just below α. Hence, α bounds the maximum possible posterior risk to $R_i = 1 - \alpha$, as desired by the design of the hypothesis test.

[109] The perfect situations that lead to expected risk values of $R = 0$ require that all CDFs are either zero or one at $\tau = \tau_0$. How fast they rise from zero to one elsewhere in the τ domain does not matter. That means, the actual information needs only require to know the statistics of the indicator quantity

$$I = \begin{cases} 1 & \text{if} \quad \tau \geq \tau_o \\ 0 & \text{if} \quad \tau < \tau_0 \end{cases},$$

and not to know the entire CDF of τ. These actual information needs would be represented only inaccurately by surrogate criteria such as the variance of τ. for example, *Leube et al.* [2012] demonstrate in a synthetic case study that optimal designs are significantly different when optimized to minimize the prediction variance of contaminant concentrations or to minimize the prediction variance of a corresponding indicator variable $I : c \geq c_0$.

[110] The inset histograms in Figure 8 illustrate the preposterior distribution $p\left(R(\mathbf{y_d})\right)$ of decision risk. These distributions can be used to assess the reliability of a proposed design, i.e., how reliable a design is to deliver the expected data utility [cf., e.g., *Trainor-Guitton et al.*, 2011]. Looking at these histograms, one can see that the more expensive design with $n_s = 8$ samples does not guarantee smaller risk values than the cheaper designs with $n_s = 2$ and $n_s = 4$. In that sense, field campaigns are a risky investment. However, the *probability* (reliability) of achieving smaller risk values does improve. This is not a property specific for hypothesis-driven optimal design, but is a general property of all preposterior analyses, i.e., a consequence of having to work with yet unknown data values in nonlinear optimal design.

[111] We could react to this insight by modifying the objective function of design: In future work it will be worth looking at designs that minimize some other statistic of risk than its expected value, such as its 90th or 95th percentile. This would lead to designs that guarantee to decrease the decision risk below a desired value, at a specified confidence level. Also, the expected risk and percentiles of risk could be combined in multiobjective optimization, in order to visualize the trade-offs between optimal expected performance and robustness. Such an analysis, however, is beyond the scope of the current study. Note that minimizing the *maximum* possible risk is not an option, because the maximum possible risk will always remain at $R_i = 1 - \alpha$.

6. Summary and Conclusions

[112] In this work, we developed a new framework to optimize field campaigns based on Bayesian hypothesis testing principles. The goal of our methodology is to support models, predictions, and derived conclusions or decisions with optimally collected data in order to achieve defensible, confident decisions. To this end, it optimizes field campaigns such that the probability of deriving false conclusions or decisions from the field campaign data is minimized, i.e., confidence is maximized. Specifically, our framework allows one to systematically test assumptions (e.g., related to conceptual models, parameter values, etc.), to test research hypotheses, or to test the compliance with environmental performance metrics. We have illustrated the applicability of our concept and methodology on two problems from stochastic hydrogeology with increasing complexity.

[113] Our approach consists mainly of three steps: (1) Formulating scientific tasks in terms of Bayesian hypothesis tests, (2) embedding the hypothesis tests into optimal design theory, and (3) using the expected probability of making wrong decisions in the hypothesis test as objective function to optimize field campaigns.

[114] The key feature that separates our approach from previous ones is the hypothesis-driven context. The hypothesis-driven context offers the following advantages (and features) over previous approaches:

[115] 1. Our framework assigns a confidence level for decision to the projected outcome of field campaigns, and then maximizes the decision confidence by finding the most informative field campaign or experimentation strategy. This supports taking maximum-confidence decisions in science, engineering, or management.

[116] 2. We are working with goal-oriented statistics that are directly related to the hypothesis, question, or environmental performance metric under concern [see *de Barros et al.*, 2012] and to the decision that needs to be made. We do not rely on generic statistical information measures (surrogate measures such as parameter variances or entropies) that could blur the character or extent of the true task-related information needs. Instead, we directly translate the risk of drawing wrong conclusions into an objective function for optimal design. This directly leads to a clear-cut and task-driven definition of information needs for the underlying hypothesis test, i.e., to a so-called ultimate measure of information.

[117] 3. The approach is flexible and can be applied to a variety of problems. This includes model choice, decision making in management, operation and maintenance, or robust engineering design, to name just a few. Examples include all the questions and studies enumerated at the beginning of section 1. As a specific example, when formulating competing models as hypotheses, the resulting designs deliver optimal data in the context of model identification and discrimination [e.g., *Luis and McLaughlin*, 1992; *Neuman and Ye*, 2009]. In applications to robust engineering design under uncertainty, one can formulate the compliance of the design with its specifications (i.e., remediation success in terms of percent removed mass or percent captured mass flux) as hypothesis test [e.g., *Cirpka et al.*, 2004].

[118] Hypothesis-driven optimal design requires to average the projected utility of a field campaign over a predicted distribution of possible (yet unmeasured) data values. In many cases, this requires brute-force Monte Carlo analysis of test statistics conditioned on random possible data. While this leads to high computer requirements for statistical analysis, the Monte Carlo approach has two highly welcome side effects:

[119] 1. Monte Carlo analyses are highly flexible and can account for arbitrary sources of uncertainty, such as model conceptual uncertainty, uncertainties in (geo)statistical model descriptions, uncertain boundary/initial conditions or forcing terms, and so forth.

[120] 2. The entire spectrum of possible outcomes, data utilities, and decision risk values for the yet unknown data values becomes available. This allows the reliability of designs to deliver the (average) promised utility to be assessed. In future work, the Bayesian decision-theoretic and hypothesis testing background could be extended, such that designs can guarantee (not only in the expected sense but on a higher reliability level) a desired level of performance.

[121] As a final remark for this paper, we were interested in minimizing the error probabilities of the decision being made. To do so, we invoked a single-objective formulation of the optimization problem, which arises from posing an individual "yes/no" type of question (see section 1). Multi-objective optimization (MOO) would be a valuable extension to our method since one could investigate trade-offs with competing criteria in situations where other objectives are also relevant [*Kollat et al.*, 2011]. The extension is straightforward since we provide a clear-cut objective function that can be plugged in as one of the competing criteria. MOO can also be valuable if the acceptable maximum risk of wrong decision for which the field campaign is being planned can only be found after analyzing the trade-offs with the costs of the planned field campaign.

Appendix A: Suggested Methods and Algorithms

A1. Conditional Simulation and Bayesian Inference

[122] Within the hypothesis-driven context, all employed tools must be able to adequately capture the magnitudes and shapes of extreme value tails for all involved probability distributions. This holds, in particular, for all conditional distributions (via Bayesian updating [e.g., *Smith and Gelfand*, 1992]) of parameters and model predictions that arise when conditioning on hypothetical data from proposed designs, often calling for (geo)statistical inversion tools. For mildly nonlinear cases, we recommend the quasi-linear geostatistical approach [*Kitanidis*, 1995] and its upgrades [e.g., *Nowak and Cirpka*, 2004], ensemble Kalman filters [*Zhang et al.*, 2005; *Nowak*, 2009], and bias-aware modifications of it [e.g., *Drécourt et al.*, 2006; *Kollat et al.*, 2011], or other conditional Monte Carlo methods based on successive linearization compared in *Franssen et al.* [2009]. In some situations, analytical solutions and linearized approaches are inappropriate. For such reasons we recommend fully nonlinear and fully Bayesian evaluation schemes based on well-designed Monte Carlo analysis, such as the PreDIA framework [e.g., *Leube et al.*, 2012] based on bootstrap or particle filtering [e.g., *Efron*, 1982;

Gordon et al., 1993] or the method of anchored distributions [e.g., *Rubin et al.*, 2010].

[123] Unconditional and conditional Monte Carlo methods often pose high computational demands, but they can assess the sampling distributions of arbitrary test statistics whenever their (un)conditional distribution shapes are unknown, or when analytical solutions for the required sampling distributions are unavailable [*Wilks*, 1995]. A second advantage is that, this way, one can achieve full freedom in the data types (direct, indirect, linear, nonlinear) considered for conditioning and design optimization. One can also accommodate arbitrary prediction models (linear, nonlinear) and account for all relevant sources of uncertainty (heterogeneity, model structure, presence of physical processes, boundary/initial conditions, etc.), in accordance with the discussion in section 2.4. The PreDIA framework offers all of these capabilities, and is outlined in more detail in section A2.

A2. Evaluating The Expected Decision Risk

[124] The procedure to evaluate the decision utility $\phi(\mathbf{d})$ of some design $\mathbf{d}$ according to equation (8) is:

[125] 1. Generate many realizations $\mathbf{r}_i$, $i = 1, \ldots, n_r$ from the joint prior distribution of all model choices, parameters, and hypotheses. Each realization represents a physically and statistically plausible reality.

[126] 2. Based on adequate simulation models, generate synthetic data sets $\mathbf{y}^*_{\mathbf{d},j}$, $j = 1, \ldots, n_y$ from $p(\mathbf{y}_\mathbf{d})$, which include independent realizations of measurement error ϵ_j. Each data set $\mathbf{y}^*_{\mathbf{d},j}$ represents a physically plausible outcome of a sampling campaign.

[127] 3. For each data set $\mathbf{y}^*_{\mathbf{d},j}$, compute the conditional probabilities $Pr[H_0|\mathbf{y}^*_{\mathbf{d},j}]$ and $Pr[H_1|\mathbf{y}^*_{\mathbf{d},j}]$ by applying any of the conditioning methods listed in section A1 to the realizations $\mathbf{r}_i$.

[128] 4. Evaluate the decisions $D_{0,j}$ and $D_{1,j}$ for each case by comparing the conditional probabilities $Pr[H_0|\mathbf{y}^*_{\mathbf{d},j}]$ to the selected significance level α.

[129] 5. Assess for each possible data set and the derived decisions the probability that the decision was wrong, using $R|\mathbf{y}^*_{\mathbf{d},j} = D_{0,j}Pr[H_1|\mathbf{y}^*_{\mathbf{d},j}] + D_{1,j}Pr[H_0|\mathbf{y}^*_{\mathbf{d},j}]$.

[130] 6. Obtain the expected probability of decision error by averaging all j values of $R|\mathbf{y}^*_{\mathbf{d},j}$.

[131] 7. Obtain the expected utility $R(\mathbf{d})$ (equation (8)) of the design by comparing with the initial risk obtained without additional data.

[132] When using the PreDIA framework [*Leube et al.*, 2012] to assess the expected utility, the core of the above procedure can be perceived as a large $n_y \times n_r$ matrix of likelihoods $p(\mathbf{y}^*_{\mathbf{d},j}|\mathbf{r}_i)$. The column headings are "what if we observed the data set $\mathbf{y}^*_{\mathbf{d},j}$ from realization $\mathbf{r}_j$ ...", and the row headings are "... and how would that data set act on realization $\mathbf{r}_i$"? After normalizing each column to sum up to unity, the elements reflect proper weights. This reflects steps one and two in the above enumeration.

[133] For step three, resort all rows and columns according to whether H_0 or H_1 is true in the underlying realizations, so that all truthfully allotted weights for H_0 form the upper left block, all truthfully allotted weights for H_1 form the lower right block, and all falsely allotted weights form the off-diagonal blocks. Within each block, the column sums are the conditional probabilities of the two hypotheses. The two upper blocks (where H_0 generated the synthetic data sets) resemble the sampling statistic $p(\mathbf{y}_\mathbf{d}|H_0)$, and the two lower blocks resemble $p(\mathbf{y}_\mathbf{d}|H_1)$. Any good data set will deliver significant statistical evidence toward the hypothesis under which it was generated. Therefore, if the design is informative on the competing hypotheses, the diagonal bocks will contain large weights and the off-diagonal blocks will contain small weights.

[134] The remaining steps mean to take the decision $D_{0,j}$ or $D_{1,j}$, followed by deleting all weights that do not correspond to the respective decision, summing up the remaining values columnwise (reflecting the conditional probabilities of wrong decision), and then averaging the remaining probabilities along the rows (yielding the expected decision risk).

[135] The appealing advantage of the PreDIA framework is that it draws the necessary realizations $\mathbf{r}_i$ and $\mathbf{y}^*_{\mathbf{d},j}$ for steps one and two from an overall pool of realizations, which avoids the appearance of two nested Monte Carlo loops. Also, it performs an analytical marginalization over the possible values of synthetic measurement errors ε_j. Together, this leads to substantial computational speedups and faster convergence of the Monte Carlo analysis.

A3. Optimization Algorithms

[136] Once the data utility can be assessed for arbitrary proposed designs, high-dimensional nonlinear optimization algorithms are employed to find the optimal design according to equation (10).

[137] Typical options are simplistic but efficient choices such as greedy search or sequential exchange [e.g., *Christakos*, 1992], classical stochastic search algorithms such as genetic algorithms [e.g., *Reed et al.*, 2000a, 2000b], or simulated annealing [e.g., *Laarhoven and Aarts*, 1992], or more modern versions such as the CMA-ES evolutionary search [*Hansen et al.*, 2003]. A promising recent alternative is to combine different global and local search strategies, such as in the AMALGAM general-purpose optimization algorithm by *Vrugt et al.* [2009].

[138] When moving toward multiobjective optimization, we refer to the discussions and algorithms reviewed and developed by *Kollat et al.* [2008] and *Shah and Reed* [2011].

[139] **Acknowledgments.** The authors acknowledge the reviewers for their constructive comments, the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart, the Spanish Ministry of Science for support within the "Juan de la Cierva" program, and the U.S. Department of Energy (DOE) Office of Biological and Environmental Research, Environmental Remediation Science Program (ERSP), for their support through DOE-ERSP grant DE-FG02-06ER06-16 as part of the Hanford 300 Area Integrated Research Challenge Project.

References

Abellan, A., and B. Noetinger (2010), Optimizing subsurface field data acquisition using information theory, *Math. Geosci.*, *42*(6), 603–630, doi:10.1007/s11004-010-9285-6.

Andricevic, R., and V. Cvetkovic (1996), Evaluation of risk from contaminants migrating by groundwater, *Water Resour. Res.*, *32*(3), 611–621.

Ben-Zvi, M., B. Berkowitz, and S. Kesler (1988), Pre-posterior analysis as a tool for data evaluation: Application to aquifer contamination, *Water Resour. Manage.*, *2*(1), 11–20.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer, New York.

Bernardo, J. M. (1979), Expected information as expected utility, *Ann. Stat.*, *7*(3) 686–690.

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

Bhattacharjya, D., J. Eidsvik, and T. Mukerji (2010), The value of information in spatial decision making, *Math. Geosci.*, *42*(2), 141–163.

Bolster, D., M. Barahona, M. Dentz, D. Fernandez-Garcia, X. Sanchez-Vila, P. Trinchero, C. Valhondo, and D. M. Tartakovsky (2009), Probabilistic risk analysis of groundwater remediation strategies, *Water Resour. Res*, *45*, W06413.

Box, G. E. P. (1982), Choice of response surface design and alphabetic optimality, *Utilitas Math.*, *21*, 11–55.

Casella, G., and R. L. Berger (2002), *Statistical Inference*, Duxbury Pacific Grove, CA.

Christakos, G. (1992), *Random Field Models in Earth Sciences*, 4th ed., Dover, New York.

Cirpka, O. A., C. M. Bürger, W. Nowak, and M. Finkel (2004), Uncertainty and data worth analysis for the hydraulic design of funnel-and-gate systems in heterogeneous aquifers, *Water Resour. Res.*, *40*, W11502, doi:10.1029/2004WR003352.

Cvetkovic, V. (2011), Tracer attenuation in groundwater, *Water Resour. Res.*, *47*, W12541, doi:10.1029/2011WR010999.

de Barros, F. P. J., and Y. Rubin (2008), A risk-driven approach for subsurface site characterization, *Water Resour. Res.*, *44*, W01414, doi:10.1029/2007WR006081.

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

de Barros, F. P. J., D. Bolster, X. Sanchez-Vila, and W. Nowak (2011), A divide and conquer approach to cope with uncertainty, human health risk and decision making in contaminant hydrology, *Water Resour. Res.*, *47*, W05508, doi:10.1029/2010WR009954.

de Barros, F. P. J., S. Ezzedine, and Y. Rubin (2012), Impact of hydrogeological data on measures of uncertainty, site characterization and environmental performance metrics, *Adv. Water Resour.*, *36*, 51–63, doi:10.1016/j.advwatres.2011.05.004.

Diggle, P., and S. Lophaven (2006), Bayesian geostatistical design, *Scand. J. Statist.*, *33*, 53–64, doi:10.1111/j.1467-9469.2005.00,469.x.

Drécourt, J.-P., H. Madsen, and D. Rosbjerg (2006), Bias aware kalman filters: Comparison and improvements, *Adv. Water Res.*, *29*(5), 707–718, doi:10.1016/j.advwatres.2005.07.006.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*, Vol. 1, 1st ed., 92 pp., Society for Industrial Mathematics, Philadelphia, Pa.

Enzenhöfer, R., W. Nowak, and R. Helmig (2012), Probabilistic exposure risk assessment with advective-dispersive well vulnerability criteria, *Adv. Water Res.*, *36*, 121–232.

Feyen, L., and S. M. Gorelick (2005), Framework to evaluate the worth of hydraulic conductivity data for optimal groundwater resources management in ecologically sensitive areas, *Water Resour. Res.*, *41*, W03019, doi:10.1029/2003WR002901.

Fishburn, P. C. (1970), *Utility Theory for Decision Making*, Robert E. Krieger Publishing Co., Huntington, NY.

Franssen, H. J. H., A. Alcolea, M. Riva, M. Bakr, N. van de Wiel, F. Stauffer, and A. Guadagnini (2009), A comparison of seven methods for the inverse modelling of groundwater flow. Application to the characterisation of well catchments, *Adv. Water Res.*, *32*(6), 851–872, doi:10.1016/j.advwatres.2009.02.011.

Frind, E. O., J. W. Molson, and D. L. Rudolph (2006), Well vulnerability: A quantitative approach for source water protection, *Ground Water*, *44*(5), 732–742.

Goode, D. J. (1996), Direct simulation of groundwater age, *Water Resour. Res.*, *32*(2), 289–296.

Gordon, N., D. Salmond, and A. Smith (1993), Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEE Proc. F*, *140*(2), 107–113.

Gotovac, H., V. Cvetkovic, and R. Andricevic (2010), Significance of higher order moments to the complete characterization of the travel time PDF in heterogeneous porous media using the maximum entropy principle, *Water Resour. Res*, *64*, W05502.

Handcock, M. S., and M. L. Stein (1993), A Bayesian analysis of kriging, *Technometrics*, *35*(4), 403–410.

Hansen, N., S. D. Müller, and P. Koumoutsakos (2003), Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), *Evolut. Comput.*, *11*(1), 1–18, doi:10.1162/106365603321828970.

Herrera, G. S., and G. F. Pinder (2005), Space-time optimiziation of groundwater quality sampling networks, *Water Resour. Res.*, *41*, W12407, doi:10.1029/2004WR003626.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Colinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*(4), 382–417.

Hou, Z., and Y. Rubin (2005), On minimum relative entropy concepts and prior compatibility issues in vadose zone inverse and forward modeling, *Water Resour. Res*, *41*(12), W12425.

James, B., and S. Gorelick (1994), When enough is enough: The worth of monitoring data in aquifer remediation design, *Water Resour. Res.*, *30*(12), 3499–3513.

Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*, Cambridge Univ. Press, Cambridge, U. K.

Kitanidis, P. K. (1986), Parameter uncertainty in estimation of spatial functions: Bayesian analysis, *Water Resour. Res.*, *22*(4), 499–507.

Kitanidis, P. K. (1995), Quasi-linear geostatistical theory for inversing, *Water Resour. Res.*, *31*(10), 2411–2419.

Kollat, J. B., and P. Reed (2007), A framework for visually interactive decision-making and design using evolutionary multi-objective optimization (VIDEO), *Environ. Model. Software*, *22*(12), 1691–1704, doi:10.1016/j.envsoft.2007.02.001.

Kollat, J. B., P. M. Reed, and J. R. Kasprzyk (2008), A new epsilon-dominance hierarchical Bayesian optimization algorithm for large multiobjective monitoring network design problems, *Adv. Water Res.*, *31*(5), 828–845, doi:10.1016/j.advwatres.2008.01.017.

Kollat, J. D., P. M. Reed, and R. M. Maxwell (2011), Many-objective groundwater monitoring network design using bias-aware ensemble Kalman filtering, evolutionary optimization, and visual analytics, *Water Resour. Res*, *47*, W02529.

Kriebel, D., J. Tickner, P. Epstein, J. Lemons, R. Levins, E. L. Loechler, M. Quinn, R. Rudel, T. Schettler, and M. Stoto (2001), The precautionary principle in environmental science, *Environ. Health Perspect.*, *109*(9), 871–876.

Laarhoven, P. J. M. V., and E. H. L. Aarts (1992), *Simulated Annealing: Theory and Applications*, Kluwer, Dordrecht, Netherlands.

Leube, P., A. Geiges, and W. Nowak (2012), Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design, *Water Resour. Res.*, *48*, W02501, doi:10.1029/2010WR010137.

Li, J. (2010), Application of copulas as a new geostatistical tool, Ph.D. thesis, Univ. of Stuttgart, Stuttgart, Germany.

Loaiciga, H. A., R. J. Charbeneau, L. G. Everett, G. E. Fogg, B. F. Hobbs, and S. Rouhani (1992), Review of ground-water quality monitoring network design, *J. Hydraul. Eng.*, *118*(1), 11–37, doi:10.1061/(ASCE)0733-9429.

Luis, S. J., and D. McLaughlin (1992), A stochastic approach to model validation, *Adv. Water Resour.*, *15*(1), 15–32.

Marler, R. T., and J. S. Arora (2004), Survey of multi-objective optimization methods for engineering, *Struct. Multidisciplinary Optim.*, *26*(6), 369–395, doi:10.1007/s00158-003-0368-6.

Massmann, J., and R. A. Freeze (1987), Groundwater contamination from waste management sites: The interaction between risk-based engineering design and regulatory policy. 1. Methodology, *Water Resour. Res*, *23*(2), 351–367.

Matérn, B. (1986), *Spatial Variation*, Springer, Berlin.

Maxwell, R., W. E. Kastenberg, and Y. Rubin (1999), A methodology to integrate site characterization information into groundwater-driven health risk assessment, *Water Resour. Res.*, *35*(9), 2841–2885.

Müller, W. G. (2007), *Collecting Spatial Data. Optimum Design of Experiments for Random Fields*, 3rd ed., Springer, Berlin.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stoch. Environ. Res. Risk Assess.*, *17*, 291–305, doi10.1007/s00477-003-0151-7.

Neuman, S. P., and M. Ye (2009), Assessing and optimizing the worth of information under model, parameters and data uncertainties, *Eos Trans. AGU*, *90*(52), Abstract H52E-04.

Neuman, S. P., L. Xue, M. Ye, and D. Lu (2012), Bayesian analysis of data-worth considering model and parameter uncertainties, *Adv. Water Res.*, *36*, 75–85.

Nowak, W. (2009), Best unbiased ensemble linearization and the quasi-linear Kalman ensemble generator, *Water Resour. Res.*, *45*, W04431, doi:10.1029/2008WR007328.

Nowak, W. (2010), Measures of parameter uncertainty in geostatistical estimation and design, *Math. Geosci.*, *42*(2), 199–221, doi:10.1007/s11004-009-9245-1.

Nowak, W., and O. A. Cirpka (2004), A modified Levenberg-Marquardt algorithm for quasi-linear geostatistical inversing, *Adv. Water Resour.*, *27*(7), 737–750.

Nowak, W., R. L. Schwede, O. A. Cirpka, and I. Neuweiler (2008), Probability density functions of hydraulic head and velocity in three-dimensional heterogeneous porous media, *Water Resour. Res.*, *44*, W08452, doi:10.1029/2007WR006383.

Nowak, W., F. P. J. de Barros, and Y. Rubin (2010), Bayesian geostatistical design: Task-driven optimal site investigation when geostatistical model is uncertain, *Water Resour. Res*, *46*, W03535.

Oladyshkin, S., H. Class, R. Helmig, and W. Nowak (2011a), An integrative approach to robust design and probabilistic risk assessment for CO_2 storage in geological formations, *Comput. Geosc.*, *15*(3), 565–577, doi:10.1007/s10596-011-9224-8.

Oladyshkin, S., H. Class, R. Helmig, and W. Nowak (2011b), A concept for data-driven probabilistic risk assessment and application to carbon dioxide storage in geological formations, *Adv. Water Res.*, *34*, 1508–1518, doi:10.1016/j.advwatres.2011.08.005.

Oreskes, N., K. Shrader-Frechette, and K. Belitz (1994), Verification, validation, and confirmation of numerical models in the earth sciences, *Science*, *263*(5147), 641.

Pappenberger, F., and K. J. Beven (2006), Ignorance is bliss: Or seven reasons not to use uncertainty analysis, *Water Resour. Res.*, *42*, W05302, doi:10.1029/2005WR004820.

Press, S. J. (2003), *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications, Wiley Series in Probability and Statistics*, 2nd ed., pp. 217–232, John Wiley, New York.

Pukelsheim, F. (2006), *Optimal Design of Experiments*, pp. 61–113, 135–157, Classics in Applied Mathematics, SIAM, Philadelphia.

Raiffa, H., R. Shlaifer, and J. Pratt (1995), *Introduction to Statistical Decision Theory*, MIT Press, Cambridge, Mass.

Rasch, D., J. Pilz, , and V. R. A. Gebhardt (2011), *Optimal Experimental Design With R*, Chapman and Hall, Englewood Cliffs, N. J.

Reed, P., B. Minsker, and A. J. Valocchi (2000a), Cost-effective long-term groundwater monitoring design using a genetic algorithm and global mass interpolation, *Water Resour. Res.*, *36*(12), 3731–3741.

Reed, P., B. Minsker, and D. E. Goldberg (2000b), Designing a competent simple genetic algorithm for search and optimization, *Water Resour. Res.*, *36*(12), 3757–3761.

Reed, P. M., and J. B. Kollat (2012), Save now, pay later? Multi-period many-objective groundwater monitoring design given systematic model errors and uncertainty, *Adv. Water Res.*, *35*, 55–68, doi:10.1016/j.advwatres.2011.10.011.

Reed, P. M., and B. S. Minsker (2004), Striking the balance: Long-term groundwater monitoring design for conflicting objectives, *J. Water Resour. Plann. Manage.*, *130*(2), 140–149, doi:10.1061(ASCE)0733-9496(2004)130:2(140).

Refsgaard, J. C., J. P. van der Sluijs, J. Brown, and P. van der Keur (2006), A framework for dealing with uncertainty due to model structure error, *Adv. Water Resour.*, *29*(11), 1586–1597.

Rubin, Y. (2003), *Applied Stochastic Hydrogeology*, Oxford University Press, Oxford.

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

Rubin, Y., and G. Dagan (1992b), A note on head an velocity covariances in three-dimensional flow through heterogeneous anisotropic porous media, *Water Resour. Res.*, *28*(5), 1463–1470.

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

Sawaragi, Y., H. Nakayama, and T. Tanino (1985), *Theory of Multiobjective Optimization*, Mathematics in Science and Engineering, Vol. 176, Academic, New York.

Scheidegger, A. E. (1954), Statistical hydrodynamics in porous media, *J. Appl. Phys.*, *25*, 994–1001.

Schoemaker, P. J. H. (1982), The expected utility model: its variants, purposes, evidence and limitations, *J. Econ. Lit.*, *20*(2), 529–563.

Schwede, R. L., and O. A. Cirpka (2010), Stochastic evaluation of mass discharge from pointlike concentration measurements, *J. Contam. Hydrol.*, *111*(1–4), 36–47,doi:10.1016/j.jconhyd.2009.10.011.

Shah, R., and P. M. Reed (2011), Comparative analysis of multiobjective evolutionary algorithms for random and correlated instances of multiobjective d-dimensional knapsack problems, *Eur. J. Oper. Res.*, *211*(3), 466–479, doi:10.1016/j.ejor.2011.01.030.

Shi, N.-Z., and J. Tao (2008), *Statistical Hypothesis Testing: Theory and Methods*, World Scientific, Singapore.

Smith, A. F. M., and A. E. Gelfand (1992), Bayesian statistics without tears: A sampling-resampling perspective, *Am. Stat.*, *46*(2), 84–88.

Stone, J. C. (1996), *A Course in Probability and Statistics*, Duxbury, Pacific Grove, Calif.

Trainor-Guitton, W. J., J. K. Caers, and T. Mukerji (2011), A methodology for establishing a data reliability measure for value of spatial information problems, *Math. Geosci.*, *43*, 929–949, doi:10.1007/s11,004-011-9367-0.

Troldborg, M., W. Nowak, N. Tuxen, P. L. Bjerg, R. Helmig, and P. Binning (2010), Uncertainty evaluation of mass discharge estimates from a contaminated site using a fully Bayesian framework, *Water Resour. Res.*, *46*, W12552, doi:10.1029/2010WR009227.

Ucínski, D. (2005), *Optimal Measurement Methods for Distributed Parameters System Identification*, CRC, Boca Raton, Fla.

USEPA (1989), Risk Assessment Guidance for Superfund Vol. 1: Human Health Manual (Part A), *Tech. Rep. Rep. EPA/540/1-89/002*, Off. Emerg. and Remedial Response, U.S. Environmental Protection Agency, Washington, D. C.

USEPA (1991), Risk Assessment Guidance for Superfund Vol. 1: Human Health Evaluation (Part B), *Tech. Rep. Rep. EPA/540/R-92/003*, Off. Emerg. and Remedial Response, U.S. Environmental Protection Agency, Washington, D. C.

USEPA (2001), Risk Assessment Guidance for Superfund: Vol. III—Part A, Process for Conducting Probabilistic Risk Assessment, *Tech. Rep. Rep. EPA 540/R-02/002*, Off. Emerg. and Remedial Response, U.S. Environmental Protection Agency, Washington, D. C.

Vrugt, J. A., B. A. Robinson, and J. M. Hyman (2009), Self-adaptive multimethod search for global optimization in real-parameter spaces, *IEEE Trans. Evol. Comput.*, *13*(2), 243–259, doi:10.1109/TEVC.2008.924428.

Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences: An Introduction*, Academic, San Diego.

Woodbury, A. D., and Y. Rubin (2000), A full-Bayesian approach to parameter inference from tracer travel time moments and investigation of scale effects at the Cape Cod experimental site, *Water Resour. Res.*, *36*(1), 159–171.

Woodbury, A. D., and T. J. Ulrych (1993), Minimum relative entropy: Forward probabilistic modeling, *Water Resour. Res.*, *29*(8), 2847–2860.

Zhang, Y., G. F. Pinder, and G. S. Herrera (2005), Least cost design of groundwater quality monitoring networks, *Water Resour. Res.*, *41*, W08412, doi:10.1029/2005WR003936.