AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR016795

On the optimal design of experiments for conceptual and predictive discrimination of hydrologic system models

C. P. Kikuchi[1,2], T. P. A. Ferré[1], and J. A. Vrugt[3,4]

[1]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA, [2]E. Montgomery and Associates, Tucson, Arizona, USA, [3]Department of Civil and Environmental Engineering, University of California Irvine, Irvine, California, USA, [4]Department of Earth System Science, University of California Irvine, Irvine, California, USA

Key Points:
- We present a metric to quantify conceptual and predictive discrimination
- The expected discrimination of prospective data points are calculated
- Optimal data points depend on between-model variance and information redundancy

Correspondence to:
C. P. Kikuchi,
ckikuchi@elmontgomery.com

Citation:
Kikuchi, C. P., T. P. A. Ferré, and J. A. Vrugt (2015), On the optimal design of experiments for conceptual and predictive discrimination of hydrologic system models, *Water Resour. Res.*, *51*, 4454–4481, doi:10.1002/2014WR016795.

Received 10 DEC 2014
Accepted 16 MAY 2015
Accepted article online 20 MAY 2015
Published online 21 JUN 2015

Abstract Experimental design and data collection constitute two main steps of the iterative research cycle (aka the scientific method). To help evaluate competing hypotheses, it is critical to ensure that the experimental design is appropriate and maximizes information retrieval from the system of interest. Scientific hypothesis testing is implemented by comparing plausible model structures (conceptual discrimination) and sets of predictions (predictive discrimination). This research presents a new Discrimination-Inference (DI) methodology to identify prospective data sets highly suitable for either conceptual or predictive discrimination. The DI methodology uses preposterior estimation techniques to evaluate the expected change in the conceptual or predictive probabilities, as measured by the Kullback-Leibler divergence. We present two case studies with increasing complexity to illustrate implementation of the DI for maximizing information withdrawal from a system of interest. The case studies show that highly informative data sets for conceptual discrimination are in general those for which between-model (conceptual) uncertainty is large relative to the within-model (parameter) uncertainty, and the redundancy between individual measurements in the set is minimized. The optimal data set differs if predictive, rather than conceptual, discrimination is the experimental design objective. Our results show that DI analyses highlight measurements that can be used to address critical uncertainties related to the prediction of interest. Finally, we find that the optimal data set for predictive discrimination is sensitive to the predictive grouping definition in ways that are not immediately apparent from inspection of the model structure and parameter values.

1. Introduction

To balance the many conflicting uses of water by society and the environment, hydrologists are called upon to make specific predictions about future hydrologic conditions that will form the basis for water management decisions. Frequently, these hydrologic predictions are guided by the outcome of hydrologic models. It is increasingly recognized, however, that multiple models often provide acceptable agreement with existing observations [e.g., *Neuman*, 2003; *Refsgaard et al.*, 2006; *Tsai and Li*, 2008; *Foglia et al.*, 2013; *Wöhling et al.*, 2015], and hence model conceptualization itself is often uncertain.

Conceptual uncertainty and predictive uncertainty [*Ye et al.*, 2010; *Wöhling et al.*, 2015], greatly complicate science-based decision making. Related to the problem of poor model identifiability is the problem of multimodal predictive distributions [e.g., *D'Odorico et al.*, 2000; *Milly*, 2001; *Ajami et al.*, 2008; *Wöhling and Vrugt*, 2008] for which probability mass may concentrate in multiple areas of the prediction space. To resolve the stated problems of conceptual and predictive identification requires an experimental design suited to achieving discrimination – among competing conceptualizations or competing prediction groups. Indeed, what is needed is an experimental design that identifies optimally informative data to be collected from complex hydrologic system of interest in order to address the related problems of conceptual and predictive discrimination.

This research introduces a novel approach to the identification of highly informative hydrologic data, which we refer to as Discrimination-Inference (DI). The DI methodology provides an information theoretic basis for experimental design with the objective of discriminating among competing system conceptualizations or prediction modes. This paper is organized as follows. Section 1 places DI in the context of previously published experimental design studies in hydrology. Section 2 develops key theoretical bases of DI; these

include the derivation of the proposed discrimination metric based on Bayesian model averaging (BMA) theory and the principle of predictive grouping. Section 3 presents two case studies demonstrating the use of DI. Section 3 also examines the characteristics of optimal designs as selected by different data utility functions. The paper concludes with a summary of findings in section 4.

1.1. Optimal Design of Monitoring Networks

Optimal design (OD) studies attempt to identify optimal or near-optimal measurement sets to maximize some data utility function [*Chaloner and Verdinelli*, 1995]. The first phase of an OD study is to define the physical situation and the experimental design objective. Objectives typically fall into one of three categories. The first category of studies focuses on reducing prediction uncertainty as the experimental design objective. These include, for example, uncertainty in model-predicted concentration [*Herrera and Pinder*, 2005], advective transport [*Hill et al*., 2013], contaminant arrival time [*Nowak et al*., 2010], or other environmental performance metrics [*de Barros et al*., 2012]. The second category of studies focuses on system parameter identification [e.g., *Vrugt et al*., 2002] and uncertainty reduction. Examples include identification of transport parameters [*Cleveland and Yeh*, 1990], reducing the uncertainty in log-permeability [*Lu et al*., 2012], and geostatistical parameter estimation [*Sun and Yeh*, 2007; Nowak *et al*., 2010; *Neuman et al*., 2012]. The third category of studies focuses on minimizing costs associated with management of a natural system. For example, the classic study of *James and Gorelick* [1994] considered the selection of monitoring locations to minimize expected costs of both contaminant remediation and data collection. More recently, *Liu et al*. [2012] presented a framework to determine the value of improved parameter information on expected contaminant remediation costs.

The majority of optimal design studies in subsurface hydrology have been developed in the context of groundwater contamination problems. Two additional objectives specific to this context are minimizing the probability of contaminant detection failure [*Dokou and Pinder*, 2009] and estimating the spatial and/or temporal moments of a contaminant plume [*Kollat et al*., 2008]. Conceptual model discrimination as a design objective has in general received less attention, and is further reviewed in section 1.2. Finally, it should be recognized that the data utility function need not be limited to a single design objective [*Kollat et al*., 2011]; indeed, multiobjective formulation will provide insight into trade-offs between different design objectives.

The second phase of OD studies entails the mathematical formulation of a data utility function used to quantify data worth. In general, the data utility function reflects the intended use of prospective hydrologic data toward one of the three categories listed above: reduced prediction uncertainty, parameter identification, or cost-based management. *Herrera and Pinder* [2005] used the trace of the concentration error covariance matrix as the data utility function; similarly, *Lu et al*. [2012] used the trace of the predicted permeability covariance matrix. The total variance approach used in these two studies considers uncertainty lumped across a spatial domain. A more common approach is to consider prediction uncertainty at one particular target location. For example, *Zhang et al*. [2005] used the coefficient of variation (CV) of contaminant concentration at a sensitive location as the data utility function, with the intent of reducing the prediction CV. A related statistic to the prediction CV is the entropy in risk as considered by *de Barros and Rubin* [2008]. It is also possible to evaluate the uncertainty associated with a binary prediction or indicator variable [*Nowak et al*., 2012] such as the exceedance of a legal concentration limit in groundwater. Alternately, some OD studies quantify the anticipated change in prediction uncertainty associated with one or more potential measurements, and then proceed with measurement selection to maximize this quantity [e.g., *de Barros et al*., 2012], yielding a unit value of uncertainty reduction associated with the hydrologic data. Some approaches translate the reduction in prediction uncertainty directly into monetary terms [e.g., *James and Gorelick*, 1994; *Feyen and Gorelick*, 2005], allowing for cost-benefit analysis of hydrologic data; this approach is more representative of problems aimed at minimizing expected cost.

The third phase of OD studies forecasts how currently unknown hydrologic data will impact the data utility function. Many studies that are focused on reducing prediction uncertainty linearize the error propogation between proposed candidate measurements and predictions at sensitive target locations through first-order, second moment (FOSM) approaches [*Glasgow et al*., 2003] such as the ensemble Kalman filter [e.g., *Herrera and Pinder*, 2005; *Nowak et al*., 2010] and OPR-PPR statistics [*Tonkin et al*., 2007]. These techniques provide a computationally efficient means of undertaking the preposterior analysis for linear or nearly linear

problems, but lead to less suitable monitoring network design for nonlinear problems [*Leube et al*., 2012]. The primary OD approach for nonlinear problems has been to condition the prediction variance upon a number of randomly sampled realizations of the unknown data values for candidate measurements [*Neuman et al*., 2012; *Leube et al*., 2012]; this is known as preposterior estimation. The drawback of this approach is that it is very computationally expensive. To reduce computational costs of the preposterior analysis for nonlinear problems, *Lu et al*. [2012] introduced three possible approximations. The approximations are based upon disregarding: parameter uncertainty; data uncertainty; or both. It has been recognized that of these three approximations, disregarding data uncertainty in particular may introduce large errors into preposterior estimates of prediction variance reduction [*Lu et al*., 2012].

The fourth and final phase of OD studies is to select optimal or near optimal measurement sets; that is, those measurement sets that maximize the data utility function. Measurement optimization algorithms that have been used in OD studies include the sequential exchange algorithm [e.g., *Nowak et al*., 2010; *Leube et al*., 2012], genetic algorithm [*Zhang et al*., 2005], and simulated annealing [*Nowak et al*., 2012]. Consideration of the number, timing, spatial coordinates, and type of candidate measurements as free design variables to be optimized leads to a very challenging problem in combinatorial optimization. This problem is compounded by the high dimensionality and computationally demanding nature of many hydrologic models. As a result, most OD studies reduce the degrees of freedom in candidate measurement selection. This may include, for example, limiting the number of measurement locations and times under consideration.

1.2. Model Discrimination Criteria and Data Utility Functions

The OD studies discussed above focus primarily on identifying optimal designs related to reducing either prediction or parameter uncertainty. Model discrimination – in other words, critically testing individual models, or sets of models – is another experimental design objective. In comparison with OD for reducing prediction uncertainty, however, OD for the objective of model discrimination has received relatively little attention in the hydrologic sciences literature.

The idea of model discrimination as an experimental design objective was first introduced by *Hunter and Reiner* [1965]. They defined model discrimination as testing rival conceptual models against each other using a likelihood ratio to quantify discrimination between two rival models. In this framework, discrimination is achieved by finding experimental conditions under which the models – each using respective maximum likelihood parameter estimates – differ to the greatest extent. *Hunter and Reiner* [1965] proposed a data utility function, the *S*-index, based on differences in measureable, model-predicted quantities.

Following this benchmark study, *Buzzi-Ferraris and Forzatti* [1983] proposed a modified data utility function, the *T*-index, also capable of accounting for estimated variance of the measurement errors, and variance of the predicted response. In contrast to the likelihood ratio advocated by *Hunter and Reiner* [1965], *Buzzi-Ferraris and Forzatti* [1983] recommended quantifying model discrimination as the rejection of poor models subsequent to performing classical statistical tests such as the *F* test on the residuals from each model under consideration. *Box and Hill* [1967] took a fundamentally different approach, defining model discrimination as the expected change in Shannon entropy before and after additional data collection. The Bayesian approach of *Box and Hill* [1967] makes explicit use of prior probabilities on the models. They developed a data utility function, the *D*-index, to quantify the maximum possible change in entropy due to the addition of a new experiment or data point.

Applications of discrimination criteria in hydrologic experimental and monitoring network design are presented by *Knopman and Voss* [1988] and *Usunoff et al*. [1992]. *Knopman and Voss* [1988] proposed four discrimination criteria as a basis for accepting or rejecting rival models: the magnitude of prediction errors; the presence of systematic error; changes in maximum likelihood parameter estimates; and measures of model fit before and after data collection. It should be noted that systematic errors consist of all sources of error, and are in practice difficult to disentangle. The proposed discrimination criteria were applied to determine the locations and times of groundwater sampling to identify boundary conditions and aquifer layering for a solute transport problem. This application is representative of field-scale hydrologic monitoring, for which sampling locations are the design variables to be optimized.

In contrast, *Usunoff et al*. [1992] investigated whether proposed column transport experiments – considering hydraulic boundary conditions and tracer pulse duration as the design variables – could discriminate among

competing conceptual models. *Usunoff et al.* [1992] quantified model discrimination as the difference in predicted values between model pairs for which the second of the two models had been calibrated to simulated concentration values of the first model. Mathematically this procedure consists of quantifying pairwise minimum differences over many possible model pairings. Discrimination is achieved when the minimum difference is sufficiently large; for example, greater than some threshold such as the level of experimental error.

The idea of conceptual discrimination, as discussed above, has been applied to testing system conceptual models against each other, with the ultimate goal of selecting a subset of more probable models. The fundamental idea underlying conceptual discrimination can be logically extended to the idea of predictive discrimination – that we may wish to test two or more predictions of future conditions against each other. This idea has received attention in the medical field [e.g., *Hanley and McNeil*, 1982; *Steyerberg et al.*, 2010], in particular when the prediction of interest is a binary or indicator variable such as the presence or absence of an illness. We are aware of only one study [*Nowak et al.*, 2012] that considers the question of predictive discrimination for water resources applications; that is, examining how data will change prediction probabilities. *Nowak et al.* [2012] evaluated the prospective impact of data on the probabilities of committing type I or type II errors in hypothesis testing, requiring that predictions be embedded in a binary or indicator variable.

The research presented here develops a DI framework for evaluating the suitability of prospective data sets for conceptual and predictive discrimination, which we refer to as Discimination-Inference (DI). The worth of potential data sets to achieve conceptual or predictive discrimination is evaluated as the expected Kullback-Leibler (KL) divergence between prior and posterior probability distributions, of either the conceptual models, or the predictive groups. We use numerical preposterior techniques to evaluate the data utility function for various prospective data sets. Finally, the preposterior analysis is embedded within a discrete optimization framework, thereby addressing measurement selection as an OD problem. The DI framework facilitates the comparison of optimal designs for both conceptual and predictive discrimination, providing a solid basis for analysis of water resources in the face of uncertainty.

2. Methods

The DI methodology quantifies the expected discrimination due to additional data collection as the distance between prior and posterior probability distributions. This metric can be applied to either conceptual (model) discrimination or predictive discrimination. The starting point for our analysis is a collection of simulations capable of predicting past, current, and future hydrologic conditions. Each simulation is based on a unique combination of underlying concept, mathematical formulation, system property parameters, and boundary conditions. We refer to the collection of simulations as the simulation ensemble, and to each individual simulation as an ensemble member.

2.1. Discrimination Metric

In the context of experimental and monitoring network design, we define discrimination as the extent to which the acquisition of a new data set causes a change in the ensemble member probabilities. Let us consider a simulation ensemble with N members comprising diverse conceptual-mathematical models, parameterizations, and boundary conditions. For notational convenience, boundary condition variability is included in the model conceptualization and parameterization. Let p_i^{u-1} denote the prior probability of the i^{th} ensemble member, with $i=\{1,2,\ldots,N\}$ conditioned on the existing data set $\mathbf{d}^{u-1}$. Similarly, let p_i^u denote the posterior probability of the i^{th} ensemble member after acquiring the additional measurements to form the u^{th} data set, $\mathbf{d}^u$. We now collect the prior and posterior probabilities into two $N\times 1$ vectors $\mathbf{p}^{u-1}$ and $\mathbf{p}^u$. We wish to quantify the distance between $\mathbf{p}^{u-1}$ and $\mathbf{p}^u$.

A natural choice is the Kullback-Leibler divergence, D_{KL} [*Kullback and Leibler*, 1951] which is used to measure the distance between two probability distributions. For notational convenience, we hereafter use the variable Φ to represent D_{KL}:

$$\Phi = D_{\mathrm{KL}} = \sum_i \ln\left(\frac{p_i^u}{p_i^{u-1}}\right) p_i^u \tag{1}$$

We adopt equation (1) to define discrimination; in other words, the value of Φ quantifies the extent to which the ensemble probabilities have changed due to the acquisition of new data $\mathbf{d}^u$.

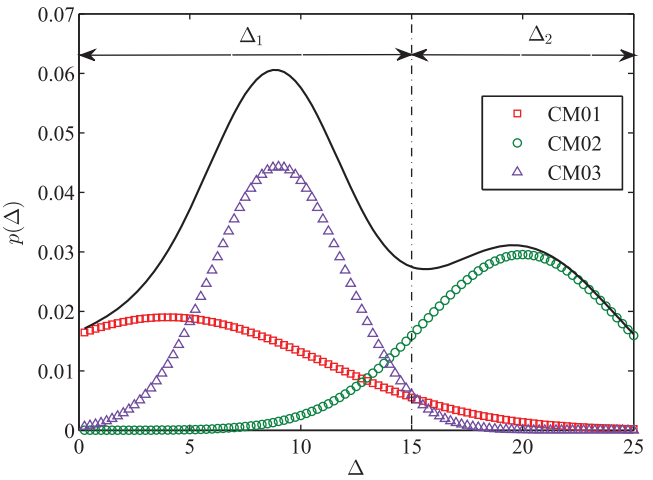

Figure 1. Illustrative example showing predictive distributions for three conceptual models (colored symbols), the expected value over the three models (solid black line), the predictive grouping threshold (at approximately 15 here), and the definition of two predictive groups.

The KL-divergence is one member from the family of *f*-divergences [*Ali and Silvey*, 1966] that quantifies differences between two probability distributions. The KL-divergence has traditionally been used as the data utility function in Bayesian experimental design [*Chaloner and Verdinelli*, 1995], and is used in this study to quantify discrimination achieved by collection of new data. However, the framework presented here is entirely general and readily accomodates alternate summary statistics including other *f*-divergences.

2.1.1. Simulation Grouping Schemes

The focus of this work is to quantify the degree of change among conceptual model structures or predictive groups. To achieve this, we introduce a scheme to calculate the probabilities of different ensemble groupings, defined specifically to accommodate different experimental design objectives. Conceptual discrimination is targeted by grouping simulations sharing the same underlying conceptual-mathematical representation of the system of interest. Suppose that the ensemble has been generated from K underlying conceptual-mathematical models. Then, the KL-divergence for the conceptual models, Φ_{cm}, is evaluated over p_k^{u-1} and p_k^u, $k=\{1,\ldots,K\}$:

$$\Phi_{cm}=\sum_k \ln\left(\frac{p_k^u}{p_k^{u-1}}\right)p_k^u \tag{2}$$

Equation (2) quantifies the conceptual discrimination that can be attributed to the u^{th} data set. This definition differs substantially from discrimination metrics used by previous studies [e.g., *Box and Hill*, 1967; *Buzzi-Ferraris and Forzatti*, 1983; *Knopman and Voss*, 1988; Usunoff *et al*., 1992] and to our knowledge has not previously been used as a data utility function for optimal design studies.

Predictive discrimination is targeted by grouping simulations that produce similar predictions of interest for water management decisions. Figure 1 illustrates an example distribution on the prediction Δ; the solid line represents the weighted average of predictive distributions over three conceptual models (CM-01, CM-02, CM-03). A predictive group consists of a set of simulations yielding values of future predicted quantities that fall within a specified range. In Figure 1, the vertical dashed line represents a threshold delineating the predictive groups, Δ_1 and Δ_2. The boundary between these groups is user-defined and reflects the intended application of the predictions. Each predictive group will typically comprise ensemble members from more than one conceptual model. Consequently, predictive discrimination is distinct from conceptual discrimination.

Once the predictive groups are defined, the analysis is similar to that described for conceptual discrimination. First, W predictive groups are defined. Figure 1 illustrates $W=2$ predictive groups; however, there are no restrictions on the number of predictive groups. The KL-divergence for discriminating among predictive groups is now evaluated over p_w^{u-1} and p_w^u, $w=\{1,\ldots,W\}$:

$$\Phi_{pr}=\sum_w \ln\left(\frac{p_w^u}{p_w^{u-1}}\right)p_w^u \tag{3}$$

Equation (3) quantifies the predictive discrimination that can be attributed to the u^{th} data set.

2.1.2. Groupwise Probabilities

To calculate the conceptual and predictive discrimination metrics, Φ_{cm} and Φ_{pr}, requires evaluation of the posterior groupwise conceptual and predictive probabilities. The posterior probabilities $p(M_k|\mathbf{d}^u)$ of the K discrete conceptual models, conditioned on data $\mathbf{d}^u$, follow *Hoeting et al*. [1999]:

$$p(M_k|\mathbf{d}^u) = \frac{p(\mathbf{d}^u|M_k)p(M_k)}{\sum_{j=1}^{K} p(\mathbf{d}^u|M_j)p(M_j)} \tag{4}$$

$$p(\mathbf{d}^u|M_k) = \int p(\mathbf{d}^u|\boldsymbol{\beta}_k, M_k)p(\boldsymbol{\beta}_k|M_k)d\boldsymbol{\beta}_k \tag{5}$$

In equation (5) $\boldsymbol{\beta}_k$ denotes the parameters of the k^{th} model; the integral is referred to as the Bayesian model evidence (BME). The first probability under the integral in equation (5) is the likelihood function, which quantifies in probabilistic terms the size of the model-data mismatch, including bias, uncertainty, correlation, etc. The second probability under the integral is the prior on the parameters for the k^{th} conceptual model.

Evaluating equation (5) is quite challenging in practice, especially if the parameter dimensionality is large [*Schöniger et al*., 2014]. The use of Markov Chain Monte Carlo (MCMC) simulation (section 2.2), renders equation (5) more tractable. Specialized estimators have been developed to calculate the BME from MCMC sampling of the posterior distribution; we adopt the use of a Laplace-Metropolis estimator of the BME [*Lewis and Raftery*, 1997]:

$$\text{BME} \approx (2\pi)^{n_k/2}|\boldsymbol{\Sigma}_k|^{1/2}p(\boldsymbol{\beta}_k^*)p(\mathbf{d}^{u-1}|\boldsymbol{\beta}_k^*) \tag{6}$$

In equation (6), n_k is the number of parameters in the k^{th} conceptual model, $|\cdot|$ is the determinant operator, $\boldsymbol{\Sigma}_k$ is the covariance matrix of posterior MCMC samples from the k^{th} conceptual model, and $\boldsymbol{\beta}_k^*$ is the maximum likelihood parameter set corresponding to the k^{th} conceptual model.

The groupwise predictive probabilities, $p(\Delta_w)$, are defined as the probability that the predicted quantity of interest, Δ, lies in the region between a, b. Mathematically, this is expressed as $p(\Delta_w) = p(\Delta \in \{a, b\})$. The value of $p(\Delta_w)$ is evaluated by integrating over the region between a, b:

$$p(\Delta_w) = \int_a^b p(\Delta)d\Delta \tag{7}$$

In equation (7), $p(\Delta) = p(\Delta|\mathbf{d}^u)$, and

$$p(\Delta|\mathbf{d}^u) = \sum_{k=1}^{K} p(\Delta|M_k, \mathbf{d}^u)p(M_k|\mathbf{d}^u) \tag{8}$$

Equation (8) follows *Hoeting et al*. [1999], and the model probabilities $p(M_k|\mathbf{d}^u)$ can be calculated with equations (4) and (5).

2.2. Sampling the Feasible Model Space

We consider N ensemble members, developed to explain and make predictions related to a hydrologic system and based on a broad range of K plausible model conceptualizations (based on variable geologic structures, choice of governing equations, boundary and initial conditions, etc.), and parameterizations $\boldsymbol{\beta}$. Let $\mathbf{k}$ represent the $K \times 1$ vector of discrete conceptual models. We then populate a vector of simulation inputs $\mathbf{s}$ of size N, with the i^{th} entry, s_i, drawn at random from the joint pdf of $\mathbf{k}$, $\boldsymbol{\beta}$ conditioned upon existing data $\mathbf{d}^{u-1}$:

$$s_i \sim p^{u-1}(\mathbf{k}, \boldsymbol{\beta}|\mathbf{d}^{u-1})p(\mathbf{d}^{u-1}) \tag{9}$$

The ensemble inputs corresponding to the k^{th} conceptual model are propogated through the corresponding model operator, f_k, to populate a matrix of predictions, $\hat{\mathbf{Y}} = f_k(\mathbf{s})$, with dimensions $N \times R$, where R is the number of predictions. Prospective candidate measurements $\hat{\mathbf{d}}^u$ such as groundwater levels or streamflow, and future predicted quantities relevant to decision-making $\hat{\boldsymbol{\Delta}}$, such as contaminant fluxes or streamflow depletion, may be included in the list of predictions.

The distribution of $\hat{\mathbf{Y}}$ depends directly on the distribution of $\mathbf{s}$; consequently, the accuracy of the predictive moments depends on the degree to which the posterior density on $\mathbf{s}$ has been sampled. MCMC sampling is the most reliable and efficient approach for populating the input vector $\mathbf{s}$; the analyses presented here use MCMC to sample the posterior parameter densities on $\boldsymbol{\beta}_k$. MCMC fully samples the posterior density on $\boldsymbol{\beta}_k$,

revisiting with high frequency regions of the parameter space leading to more probable model simulations, while still retaining less probable realizations. Upon convergence, the input sample produced by MCMC simulation can be used directly to populate the model inputs **s**.

The development of efficient and accurate MCMC algorithms for hydrologic problems has been the topic of extensive research. The DiffeRential Evolution Adaptive Metropolis [*Vrugt et al*., 2009] (DREAM) algorithm represents the state of the art in MCMC sampling, and is used in this study. Briefly, DREAM runs multiple Markov Chains concurrently, and multivariate proposals are created on the fly using differential evolution [*Storn and Price*, 1997], each of which moves through the parameter space, converging on those regions associated with higher density. This methodology is easy to implement in practice and exhibits a high sampling efficiency. For the DI methodology considered here, ensembles based on more than one underlying conceptualization require that MCMC sampling be conducted for each different conceptualization.

2.3. Preposterior Estimation of the Discrimination Metric

Given the values of two data sets, $\mathbf{d}^{u-1}$ and $\mathbf{d}^{u}$, the conceptual or predictive discrimination can be calculated using equations (2) and (3), respectively. Our objective in this work, however, is to determine the suitability of prospective data sets for which $\mathbf{d}^{u}$ is unknown. That is, we aim to predict the discriminatory capabilities of data before they are collected. To solve this problem, we turn to preposterior estimation techniques [e.g., *Reichard and Evans*, 1989; *James and Gorelick*, 1994; *Feyen and Gorelick*, 2005; *Leube et al*., 2012; *Neuman et al*., 2012], which estimate the expected value of some data utility function by calculating the average over many ensemble-generated realizations of the prospective data set.

The preposterior estimation procedure used in this study consists of four steps. First, we generate a set of M data realizations. To accomplish this, we independently populate the $M\times1$ vector of model inputs to be used for data realizations, $\mathbf{s}^{\mathrm{drz}}$ from the model input space according to equation (9). Then, we pass those inputs through the model operator to predict the numerical values corresponding to the u^{th} prospective data set, $\hat{\mathbf{d}}^{u}_{\mathrm{drz}}$. We then conduct a Bayesian updating procedure for each of the M data realizations. For each data realization, the prospective data set values $\hat{\mathbf{d}}^{u}_{\mathrm{drz}}$ are then compared to the ensemble-predicted equivalents, $\hat{\mathbf{d}}^{u}_{\mathrm{ens}}$. The difference between the predicted measurement values under the i^{th} ensemble member $\hat{\mathbf{d}}^{u}_{i}$ and the j^{th} data realization $\hat{\mathbf{d}}^{u}_{j}$ is used to evaluate the likelihood function. The case studies considered in this paper use a Gaussian likelihood function:

$$p\left(\hat{\mathbf{d}}^{u}_{i}|\hat{\mathbf{d}}^{u}_{j}\right)=\frac{1}{\left[(2\pi)^{n_u}|2\boldsymbol{\Sigma}_{\epsilon}|\right]^{1/2}}\exp\left[-\frac{1}{2}\mathbf{e}^{T}(2\boldsymbol{\Sigma}_{\epsilon})^{-1}\mathbf{e}\right] \tag{10}$$

In equation (10), n_u is the size of the u^{th} data set, $\boldsymbol{\Sigma}_{\epsilon}$ is the error covariance matrix, and **e** is the difference between $\hat{\mathbf{d}}^{u}_{i}$ and $\hat{\mathbf{d}}^{u}_{j}$. This formulation uses noise-free ensemble members and data realizations, but doubles the error covariance matrix, according to the marginalization derived by *Leube et al*. [2012]. Equation (10) is substituted directly into equation (5), from which the conceptual and predictive probabilities are calculated. It should be noted that although Gaussian likelihood functions are shown in the following examples, the DI framework can handle non-Gaussian likelihood functions.

Having calculated the posterior probabilities, we now evaluate the discrimination metric (Φ^{cm}_{j} or Φ^{pr}_{j}) for the j^{th} realization using equation (2) or (3). For each realization, the values of the discrimination metric are stored in the $M\times1$ vector $\boldsymbol{\Phi}$. In the final step, we calculate the expected value of the discrimination metric for each proposed set of observations by averaging over the data realizations:

$$\mathrm{E}[\Phi]=\frac{1}{M}\sum_{j=1}^{M}\Phi_{j} \tag{11}$$

The preposterior estimation procedure described above may be sensitive to both the ensemble size, N, and the number of data realizations M. For small values of N, conditioning on $\hat{\mathbf{d}}^{u}$ may concentrate the probability mass in a relatively small number of ensemble members. This problem is known as filter degeneracy [*Doucet and Johansen*, 2008], and may introduce error into the calculation of the groupwise posterior probabilities. Selecting an adequately large value for N is problem-specific, and can be guided by recording and evaluating the effective sample size (ESS) [*Liu*, 2008], which measures sample diversity.

$$\mathrm{ESS} = \left(\sum_{i=1}^{N} p_i^2 \right)^{-1} \tag{12}$$

Averaging the ESS over the M data realizations provides an average effective sample size (AESS) [*Leube et al.*, 2012]. The ESS quantifies the approximate number of perfect samples drawn from the distribution of interest [*Doucet and Johansen*, 2008]. The required ESS for inference is application dependent. Previous studies of preposterior data worth estimation [e.g., *Leube et al.*, 2012] determined that ESS values of 500 were adequate for the purposes of evaluating the worth of prospective data.

Obtaining a representative value of $\mathrm{E}[\Phi]$ requires a sufficiently large number of data realizations, M, to smooth out the variability in the conditioning data due to model input uncertainty. To evaluate the reliability of the preposterior discrimination metric for each case study, we conduct benchmark calculations to determine the sensitivity of $\mathrm{E}[\Phi]$ to M, enabling selection of sufficiently large M. Similarly, we determine appropriate values of N applicable to problems of varying complexity. These results are presented in section 3.1.4.

2.4. Optimization Algorithm

The DI procedure, as detailed above, may be used to estimate the ability of candidate data sets to achieve conceptual or predictive discrimination between ensemble members. Specifically, we wish to maximize $\mathrm{E}[\Phi]$; formally, this can be written:

$$\mathbf{u}_{\mathrm{opt}} = \max_{\mathbf{u} \in \mathbf{D}} \left\{ \mathrm{E}[\Phi] = \frac{1}{M} \sum_{j=1}^{M} \Phi_j \right\} \tag{13}$$

In equation (13), **D** is the space containing all admissible experimental designs satisfying logistical constraints (e.g., total cost, available observations). The vector **u** is a subset of **D** representing experimental design settings such as measurement types, times, and locations. High-dimensional **D** renders infeasible the exhaustive evaluation and comparison of all prospective data sets. Therefore, it will generally be necessary to solve equation (13) using optimization techniques. The prospective data set is based on design variables, which may be either discrete or continuous. The optimization algorithm must be capable of handling both variable types. This research uses the Nonlinear Mesh Adaptive Direct search (NOMAD) algorithm [*Le Digabel*, 2011], as implemented in the OPTimization Interface (OPTI) toolbox [Currie and Wilson, 2012] for MATLAB software.

2.5. Computational Implementation

The Discrimination-Inference (DI) method is illustrated schematically in Figure 2, and is implemented as described by the following steps:

1. Propose and implement a set of K conceptual-mathematical models to describe the hydrologic system of interest. For each conceptual model, describe quantitatively the distribution of the model parameters, based on field samples, existing databases, etc.
2. If data suitable for model calibration are available, use those existing data to derive posterior parameter densities under each conceptual model. If data are not available, then parameter densities must be estimated on the basis of qualitative data, literature values, etc.
3. Based on the outcome of steps (1) and (2), use equation (9) to generate the $N \times 1$ vector $\mathbf{s}^{\mathrm{ens}}$ containing ensemble inputs. Then, propogate the simulation inputs through the model operator f_k to produce the $N \times R$ array of ensemble predictions. Repeat this procedure independently for each data realization, to generate the $M \times 1$ vector $\mathbf{s}^{\mathrm{drz}}$ containing inputs for generating the data realizations. Then, populate the $M \times R$ array of data realizations via the same model operator f_k.
4. Determine the admissible space, **D**, of measurement sets as dictated by logistical, time, and fiscal constraints, and subject to expert judgment, where possible.
5. Implement a discrete optimization algorithm to maximize the data utility function $\mathrm{E}[\Phi]$ subject to the constraints identified in step 4; each evaluation of the data utility function entails the following:
   i. Loop through the M data realizations; for the j^{th} realization, evaluate the Bayesian model evidence (equation (5)).

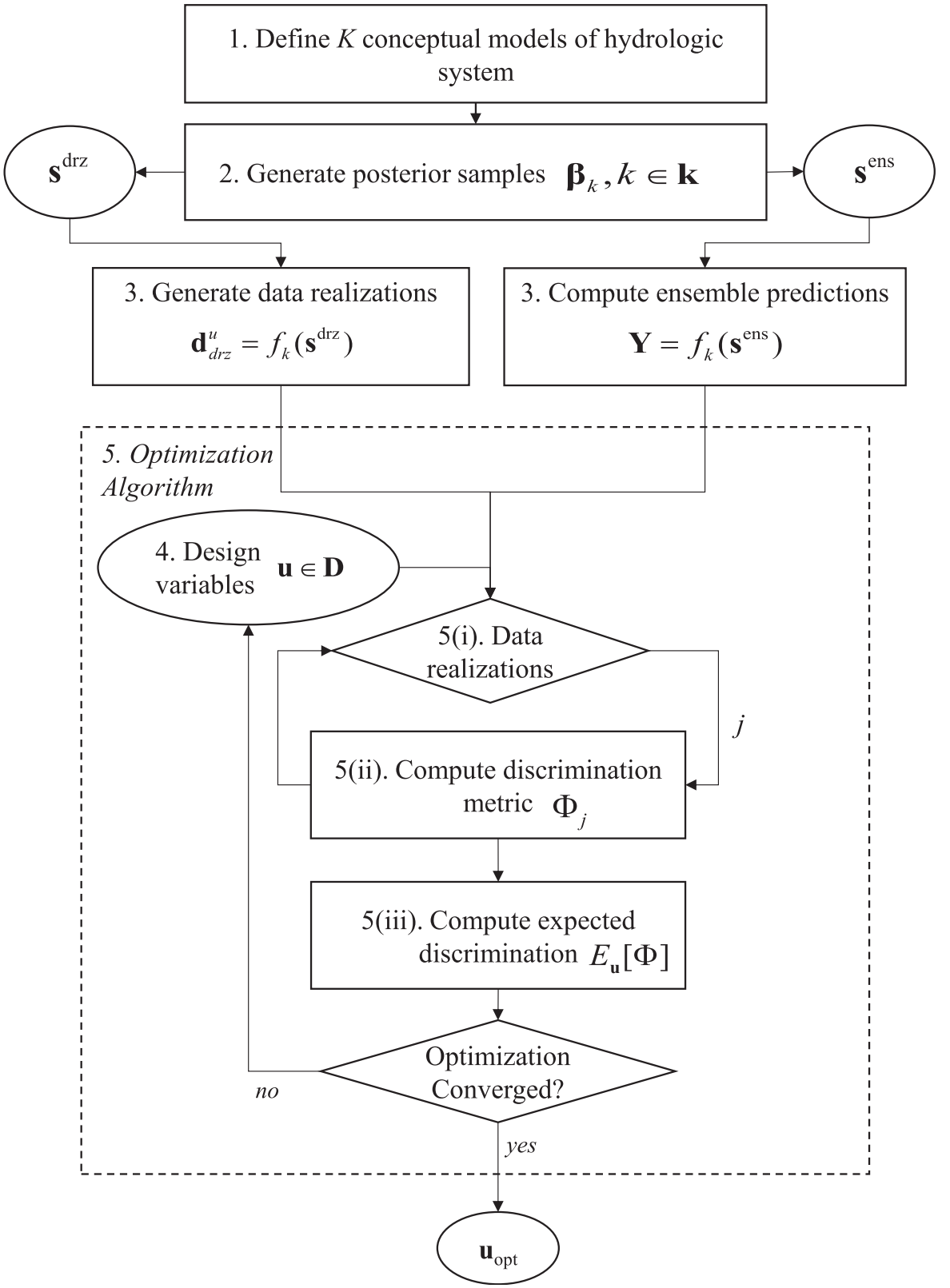

Figure 2. Schematic diagram showing workflow of the Discrimination-Inference framework, including optimal design analysis (within dashed line).

ii. Compute $p_j\left(M_k|\hat{\mathbf{d}}\right)$ for each group; then, compute either Φ_j^{cm} or Φ_j^{pr}, using equation (2) or (3), depending on the objective of the investigation (conceptual versus predictive discrimination).

iii. Store the data utility function of the j^{th} realization—either Φ_j^{cm} or Φ_j^{pr}—in the vector $\mathbf{\Phi}$. Once all of the data realizations have been processed, calculate $E[\Phi]$ using equation (11).

Steps 5(i–iii) are repeated within the optimization algorithm until the convergence criterion for globally optimally $E[\Phi]$ has been satisfied.

3. Case Studies

We present two case studies to illustrate the DI procedure for optimal selection of hydrologic measurements. The first case study concerns the selection of paired pressure head and water content measurements in soils, with the objective to discriminate among multiple soil hydraulic models. The second case study considers the long-term effects of groundwater pumping on spring discharge in a closed hydrologic basin. The objective in the second case study is to select a set of predevelopment measurements best suited to discriminate among predictions of spring depletion under postdevelopment conditions.

3.1. Case Study 1: Soil Water Characteristic Curve

The soil water characteristic (SWC) curve relates pressure head and water content in a porous medium. This relation is central to modeling processes such as infiltration, soil evaporation, root water uptake, groundwater recharge, and water redistribution. Several different models have been proposed that describe this relationship. As their application often involves important differences in the simulated soil moisture regime and associated hydrologic processes, it is important to select the most appropriate hydraulic model for a given soil and experimental data. The objective of the first case study is to identify two sets of paired measurements of water content, θ, and pressure head, ψ, that are optimally suited to discriminate among soil hydraulic models. The analysis presented here considers three soil hydraulic models: Mualem-van Genuchten [*Van Genuchten*, 1980], Brooks-Corey [*Brooks and Corey*, 1964], and Kosugi [*Kosugi*, 1996] models. Each of the three conceptual models described above contains four fitting parameters whose values are typically estimated from in situ (field) or laboratory $\theta-\psi$ data using nonlinear least-squares. The Mualem-van Genuchten (MVG) model of the SWC is given by:

$$\theta(\psi)=\theta_r+(\theta_s-\theta_r)[1+(\alpha|\psi|^n)]^{1/n-1} \tag{14}$$

The MVG model contains four fitting parameters whose values are soil dependent; θ_s is the saturated water content, θ_r is the residual water content, α is related to the inverse of the air-entry pressure, and n is related to the pore-size distribution. The Brooks-Corey (BC) model of the SWC is given by:

$$\theta(\psi)=\theta_r+(\theta_s-\theta_r)\left(\frac{\psi_b}{\psi}\right)^{\lambda} \tag{15}$$

In addition to θ_r and θ_s, the BC model contains two parameters - the bubbling capillary pressure, ψ_b, and the pore-size index, λ, that are related to the parameters of the MVG model by $\psi_b=\alpha^{-1}$ and $\lambda=n-1$ [*Rawls et al.*, 1993]. Finally, the Kosugi (KM) model of the SWC is given by:

$$\theta(\psi)=\theta_r+(\theta_s-\theta_r)\left\{\frac{1}{2}\mathrm{erfc}\left[\frac{\ln(\psi/a)}{\sqrt{2}n}\right]\right\} \tag{16}$$

The KM fitting parameters are a and n, in addition to θ_s and θ_r. The KM parameters a and n are not related the the MVG or BC parameters.

3.1.1. Generation of Simulation Ensemble and Data Realizations

Following the procedure outlined in section 2.5, we first populate a simulation ensemble comprising parameter realizations from each of the three conceptual models. The prior distributions on the parameters of the MVG and BC model are derived from calibrated MVG parameter values in the ROSETTA database [*Schaap et al.*, 2001]. Specifically, for a given soil texture – in this case, a sandy loam soil – we extract all corresponding MVG parameter estimates (481 total soil samples), and compute a sample mean and covariance over the set of parameter estimates. The sample mean and covariance of the KM parameters are not included in the ROSETTA database. Rather, we calculated them as follows: for each of the sandy loam soils for which

paired θ-ψ are listed, the KM parameters were estimated to minimize the model-data residuals over available soils from the ROSETTA database, producing a set of KM parameter estimates. Then, the sample mean and covariance were calculated over all the parameter sets.

For each conceptual model, the corresponding parameter mean and covariance matrix were used to construct a Gaussian prior. We evaluated the posterior probability density over each of the three conceptual models, and then calculated the conceptual model probabilities according to equation (4). We used the DREAM algorithm [*Vrugt et al*., 2009] to derive the posterior parameter distribution using a total of 100,000 function evaluations per conceptual model. The observational data $\mathbf{d}^{u-1}$ consist of six paired $\theta-\psi$ measurements for a sandy loam soil, taken from the ROSETTA database [*Schaap et al*., 2001]. We used a Gaussian likelihood function as described by equation (9), with parameters based on the assumed distribution of measurement errors. For this case, we assume uncorrelated measurement errors; the diagonal entries of Σ_ϵ contain the constant measurement error variance of 0.0001 ($m^3\ m^{-3}$) associated with a water content measurement error (treated as 95% confidence interval) of 0.02 ($m^3\ m^{-3}$). These values are typical for time-domain reflectrometry measurement of soil water content [*Topp et al*., 1980].

We monitored the convergence of each DREAM run based on the $\hat{R}$-statistic of *Gelman and Rubin* [1992]. Inspection of DREAM output showed that the $\hat{R}$-statistic typically dropped below 1.2 within the first 10,000 function evaluations for each chain. It was therefore determined that the algorithm had converged to the posterior density after this point. That is, parameter samples drawn from each chain after the first 10,000 function evaluations represent a full sample of the posterior parameter space. We extracted the last 50,000 parameter samples and associated log-likelihood function values for each conceptual model, and used the samples to evaluate equation (4) for each conceptual model, assuming equal (uniform) prior probabilities for each conceptual model. The posterior probabilities of the MVG, BC, and KM conceptual models after conditioning on $\mathbf{d}^{u-1}$ were 0.237, 0.233, and 0.530, respectively. The distribution on conceptual model probabilities exhibits a preference for the KM; however, the other two models are still plausible.

Populating both the simulation ensemble and the set of data realizations for preposterior analysis requires drawing two independent sets of model inputs, $\mathbf{s}^{ens}$ and $\mathbf{s}^{drz}$. For this example, the model inputs consist of parameter realizations under each conceptual model, and are readily available as the parameter output samples from the DREAM algorithm. We populated $\mathbf{s}^{ens}$ directly from the DREAM parameter sampling outputs, and populated $\mathbf{s}^{drz}$ from a second, independent DREAM run with 50,000 maximum function evaluations per conceptual model. Our preliminary analysis used $N = 150{,}000$ ensemble members divided equally among the conceptual models, and with $M = 1{,}500$ data realizations. We also conducted benchmark runs to evaluate the effect of N and M on the results of the DI procedure.

3.1.2. Simulation Ensemble Characteristics

The effect of data on the distribution of model predictions can be evaluated through BMA statistics – specifically, the decomposition of prediction variance to within-model variance and between-model variance [Hoeting *et al*., 1999]:

$$\mathrm{E}_K[\mathrm{var}(\Delta)] = \sum_k \mathrm{var}(\Delta|\mathbf{d})p(M_k|\mathbf{d}) \tag{17}$$

$$\mathrm{var}_K[\mathrm{E}(\Delta)] = \sum_k [\mathrm{E}(\Delta|\mathbf{d}, M_k) - \mathrm{E}(\Delta|\mathbf{d})]^2 p(M_k|\mathbf{d}) \tag{18}$$

Equation (17) defines the within-model variance as the expectation of the variance over the conceptual models, and quantifies prediction uncertainty associated with factors such as parameter variability and forcing data. Equation (18) defines the between model variance as the variance of expectation over the conceptual models, and quantifies prediction uncertainty due to conceptual, or model structural variability. The total prediction variance is equal to the sum of the within and between-model variance. Conditioning the parameters of each conceptual model on the initial data set $\mathbf{d}^{u-1}$ of six θ-ψ measurement points substantially reduced both within-model and between-model prediction variance in the pressure range corresponding to the existing data. Figure 3a compares the expectation of the predicted water content $\mathrm{E}[\theta_k|M_k]$, under each of the three conceptual models. The expectation of the predicted water content, $\mathrm{E}[\theta_k|M_k]$, calculated over all ensemble members under each of the soil hydraulic models, are represented by blue (MVG), red (BC), and green (KM) lines. The dark gray area in Figures 3a–3c represents the prediction uncertainty, as

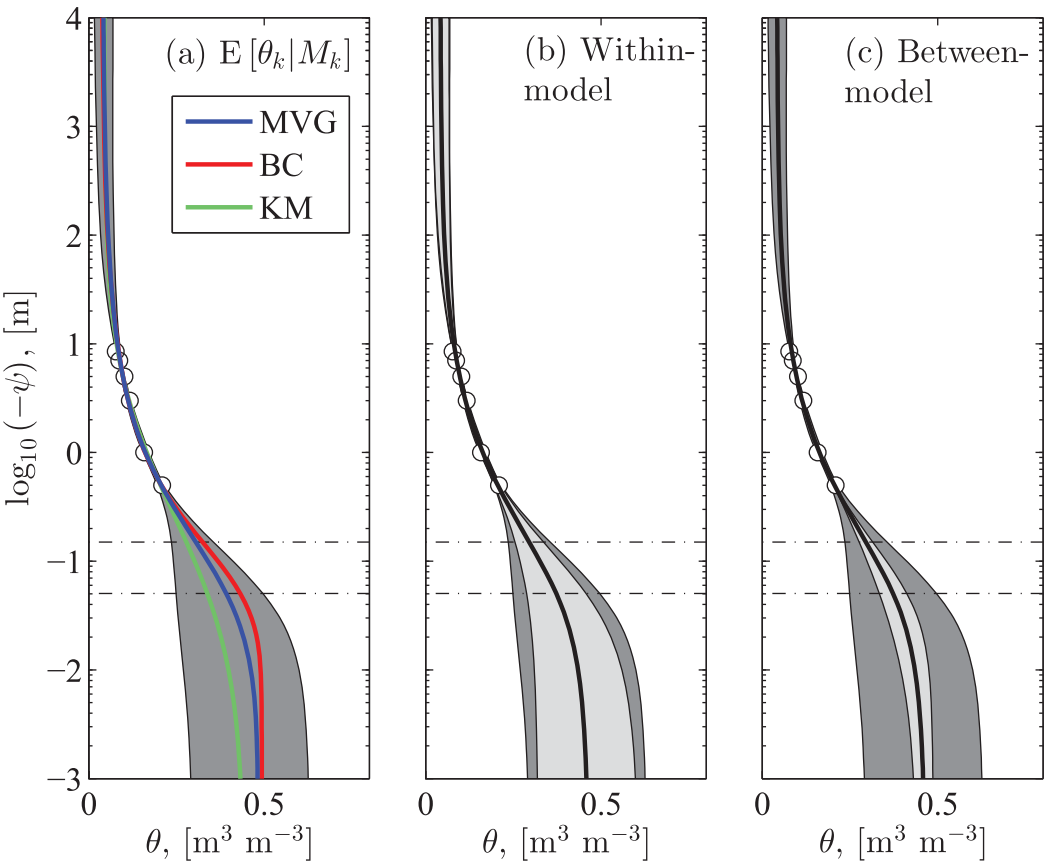

Figure 3. Posterior expectation and uncertainty of predicted volumetric water content based on three conceptual models. For all plots, the dark gray area represents total posterior standard deviation of the predicted water content. (a) Posterior expectation, $E[\theta_k|M_k, \mathbf{d}]$ for each conceptual model, (b) posterior expectation, $E[\theta|\mathbf{d}]$, over all three conceptual models; light gray area represents within-model standard deviation, (c) posterior expectation, $E[\theta|\mathbf{d}]$ over all three conceptual models; light gray area represents between-model standard deviation. Horizontal dashed lines indicate the optimal set of two θ-ψ measurement pairs that maximize conceptual model discrimination.

represented by the posterior BMA standard deviation of the predicted water content. The predictions diverge to the greatest extent at pressure heads closest to zero. Figures 3b and 3c illustrate the posterior weighted expectation over all three conceptual models underlain by the standard deviation of the predicted water content, representing prediction uncertainty. The light gray area in Figure 3b shows the within-model component of the total predictive uncertainty, and the light gray area in Figure 3c shows the between-model component of the total predictive uncertainty. To generalize, the prediction standard deviations are relatively tight in the vicinity of measurements corresponding to $\mathbf{d}^{u-1}$, but relatively wide at very low and very high ψ values. However, most of the uncertainty is attributed to within-model variations rather than between-model variations.

3.1.3. Implementation of DI Analysis

Having generated the simulation ensemble and the data realizations, we now compute the data utility function, $E[\Phi_{cm}]$, over the space of possible measurements. For this problem we consider the selection of two additional θ-ψ pairs, with $\psi \in \{-10^4, -10^{-3}\}$ [m]. We discretized the log-transformed pressure heads into 75 equally spaced points, resulting in $\binom{75}{2} = 2{,}775$ pairs of ψ coordinates. We augmented this candidate measurement set with 75 additional points, each constituting a single ψ coordinate, resulting in a total of 2,850 possible candidate measurement sets. For this particular example, the relatively small number of candidate measurement sets, coupled with the fast simulation time associated with equations (14–16) made possible a complete, exhaustive exploration of the entire candidate measurement space. In other words, we evaluated equation (11) a total of 2,850 times.

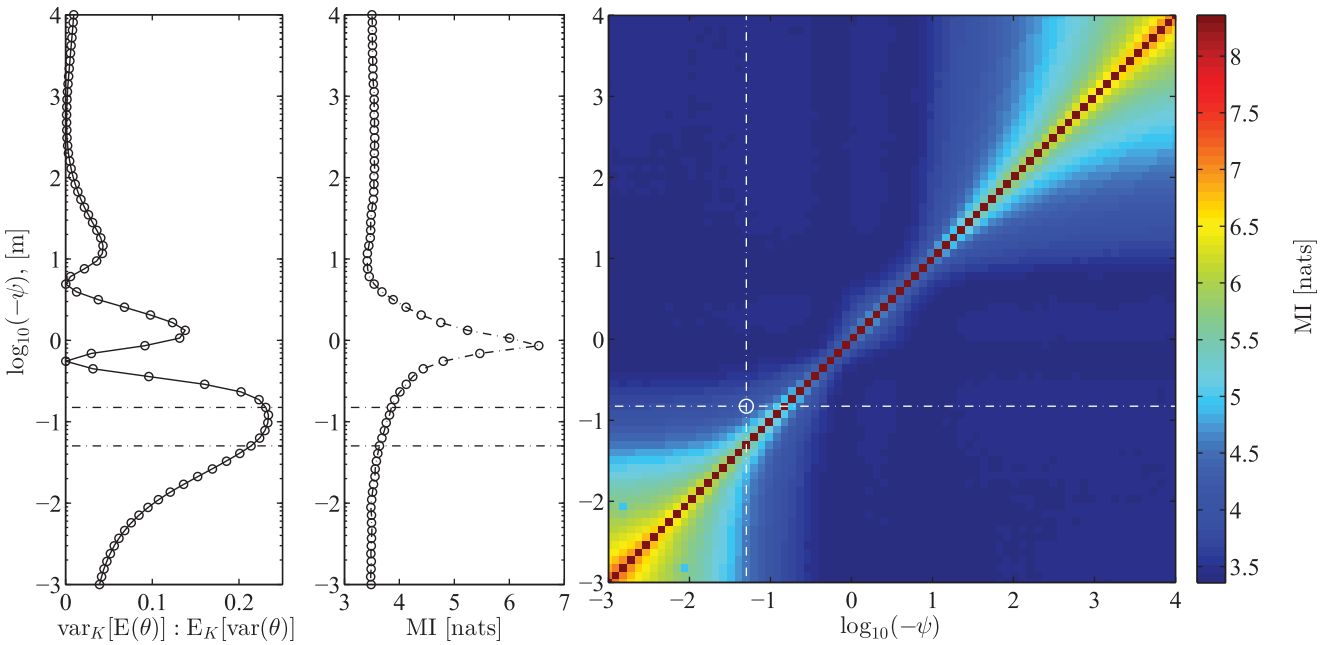

Figure 4. (a) The ratio of between-model to within-model variance, $\mathrm{var}_K[E(\theta)] : \mathrm{E}_K[\mathrm{var}(\theta)]$ evaluated over the range of feasible candidate pressure heads; dashed horizontal lines show the coordinates of the optimal measurement pair, ψ_1 and ψ_2. (b) Mutual information (MI) between possible candidate measurements and the set of existing $\theta-\psi$ measurement pairs. (c) Surface showing the MI evaluated for each possible candidate measurement pair; white circle identifies the location of ψ_1 and ψ_2.

3.1.4. Results of DI Analysis

The horizontal dashed lines in Figures 3 and 4 depict the pressure heads of the optimal measurement set for the purposes of conceptual model identification, as determined by the DI procedure described above. In other words, the pair of measurements identified by pressure heads of $\psi_1 = -10^{-1.6}$ and $\psi_2 = -10^{-0.95}$ m are those for which the value of $\mathrm{E}[\Phi_{\mathrm{cm}}]$ was largest. The optimal ψ-coordinates are surprising on first inspection, as Figure 3c indicates that the largest between model variation is associated with $\psi = -10^{-2}$ m. This result suggests that θ-ψ measurements in this area would support conceptual model discrimination. However, the optimal measurements are instead located at slightly more negative pressure heads. The somewhat counterintuitive selection of optimal observations can be explained by two key characteristics of optimal measurement set selection for conceptual model discrimination: the selected observations must target differences among the conceptual models in excess of the within-model uncertainty, and they must minimize the collection of redundant information.

The large within-model uncertainty—due exclusively to parameter uncertainty in this case—for pressure heads greater than $\psi = -10^{-1}$ m greatly diminishes the worth of water content measurements at pressures between 0 and -10^{-1} for conceptual discrimination. Figure 4a shows the ratio of the between-model variance to within-model variance, $\mathrm{var}_K[E(\theta)] : \mathrm{E}_K[\mathrm{var}(\theta)]$; this is effectively a signal-to-noise ratio, where model variability is the signal and parameter variability constitutes the noise. In other words, this ratio quantifies the extent to which between-model differences can be discerned given the degree of within-model variability. This ratio is highest in the pressure range of $\psi = -10^{-0.8}$ to $\psi = -10^{-1.8}$ m. The selected measurements, ψ_1 and ψ_2, lie within this range.

If we had been seeking one, rather than two, additional $\theta-\psi$ measurements, the data pair corresponding to the maximum value of $\mathrm{var}_K[E(\theta)] : \mathrm{E}_K[\mathrm{var}(\theta)]$ would have been selected. The optimal pressures for a set of two candidate measurements, however, must balance high discriminatory power - as represented by the between to within-model uncertainty—with the requirement to minimize the collection of redundant information. We use the mutual information to quantify the degree of redundancy of individual candidate measurements with the existing data set and with other individual measurements in the candidate measurement set. Briefly, $MI(\mathbf{X}, \mathbf{Y})$ describes how much information **X** contains regarding **Y**, and is suitable for measuring nonlinear relations among random variables. A relatively high MI value indicates a high

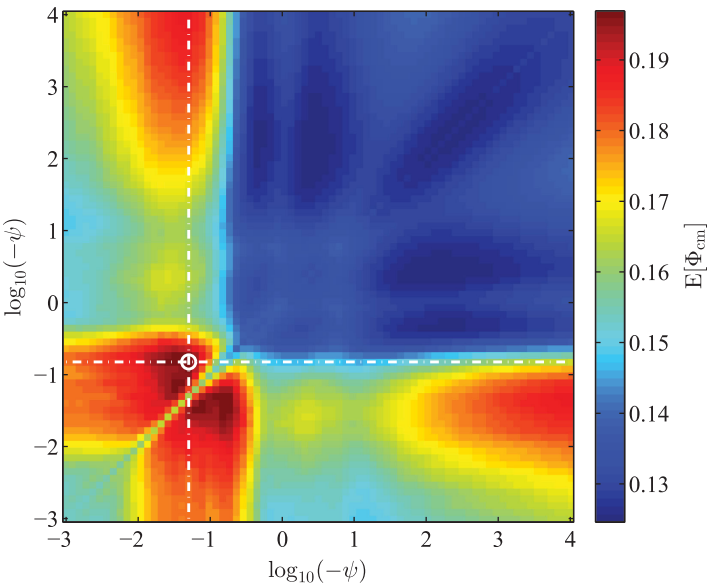

Figure 5. Two-dimensional map of $E[\Phi_{cm}]$ derived from preposterior estimation using all possible sets of of two $\theta-\psi$ measurement pairings. The white circle indicates the optimal measurement pair, associated with the highest possible value of $E[\Phi_{cm}]$.

degree of measurement information redundancy. We use the *k*-nearest neighbor estimator of *Kraskov et al*. [2004] to conduct two sets of MI calculations, populating the variables **X** and **Y** with ensemble-predicted water content values over all conceptual models and parameterizations.

Figure 4b shows the MI between existing data $\mathbf{d}^{u-1}$ and each possible ψ-θ pair. The highest values are found in the vicinity of the existing data, as expected. This result suggests that θ measurements in the range $-10^{-0.5} < \psi < -10^{0.2}$ m are redundant with the existing data. Figure 4c shows the MI calculated between possible pairs of candidate measurements *a* and *b* over all $(\psi_a, \psi_b) \in \mathbf{D}$. The highest MI values are found for pairs comprising identical ψ-coordinates, as expected. The white circle in Figure 4c shows the optimal measurements set as identified by the DI analysis. The MI value for the optimal ψ-coordinates is relatively small at these coordinates, indicating substantially lower redundancy than for other possible measurement pairs. To summarize, Figure 4 illustrates that a good measurement set for conceptual discrimination is one that maximizes between-model versus within-model prediction uncertainty and minimizes both the redundancy of the measurement set with existing data, and the internal redundancy of the measurement set. The relative importance for each of these three qualities cannot be known *a priori*, but the contribution of both is implicit in the preposterior estimation procedure.

Figure 5 illustrates the distribution of $E[\Phi_{cm}]$ over the feasible measurement space. It is worth noting that the $E[\Phi_{cm}]$ surface is symmetric; this is because the θ measurements are simultaneously, rather than sequentially, assimilated and processed. The surface is bisected by a 1:1 line that reflects measurement sets consisting of a single θ-ψ measurement pair; this line can be used to identify instances for which a single measurement may be just as informative or even more informative than two measurements for the purpose of conceptual discrimination. For example, a single θ-ψ measurement in the vicinity of $\psi = -10^{-1.5}$ m (coordinates of $\psi = -10^{-1.5}$ m for both ψ_1 and ψ_2) has a larger value of $E[\Phi_{cm}]$ than any pair of measurements for which both values of ψ lie between -10^{-1} and -10^{4} m. Note that the selected observations (shown as the white circle) are in an area of very high $E[\Phi_{cm}]$.

3.1.5. Reliability of the Preposterior Estimates

The reliability of the results shown in Figures 3–5 is contingent upon two important factors. First, the extent to which filter degeneracy was encountered during the preposterior updating of the simulation ensemble probabilities, and second, the convergence of $E[\Phi_{cm}]$ during the preposterior estimation procedure. We evaluate the magnitude of filter degeneracy by calculating the minimum, maximum, and median AESS over all 2,850 possible measurement sets. For the benchmark DI run with N=120,000, M=1,200, AESS values ranged from approximately 30,000 to 40,000. A generally accepted rule-of-thumb in particle filtering for state estimation is that ESS $\geq N/2$ is adequate to characterize the distribution of interest [*Doucet and Johansen*, 2008]; however, a similar rule of thumb has not been established for data worth applications. Therefore, we conducted benchmarking analyses to study the effect of the simulation ensemble size, *N*, on both the AESS and value of $E[\Phi_{cm}]$. Specifically, we repeated the DI analysis with a range of values for *N*, compiling the results in Table 1. We also conducted DI calculation runs for several values of *M* to provide preliminary guidance on the minimum value of *M* to ensure convergence of $E[\Phi_{cm}]$.

Table 1. Summary of DI Calculation Results Over a Range Values for Ensemble Size, *N*, and Number of Data Realizations, *M*[a]

Ensemble Size, *N*	Data Realizations, *M*	Min/Max Value AESS	Median Value AESS	Optimal Coordinates ψ	Wall-Time (min)
30,000	300	7,852/10,028	8,137	$\{-0.050, -0.186\}$	45
30,000	600	7,910/10,176	8,129	$\{-0.041, -0.078\}$	93
60,000	300	15,688/20,305	16,470	$\{-0.033, -0.050\}$	72
60,000	600	15,769/20,349	16,421	$\{-0.050, -0.121\}$	129
90,000	300	23,427/30,137	24,343	$\{-0.0126, -0.121\}$	106
90,000	600	23,419/30,174	24,369	$\{-0.033, -0.063\}$	198
120,000	1,200	31,280/40,261	32,621	$\{-0.050, -0.150\}$	456

[a]The benchmarking runs were conducted on a dual-core processor (2.40 GHz) with 24.0 GB RAM.

The optimal ψ coordinates do not appear to be any more sensitive to the ensemble size *N* than to *M*, over the range of values explored. We therefore conclude that the AESS was acceptably large. This result suggests that at least for low-dimensional problems such as this case study, a smaller ensemble size on the order of 30,000 ensemble members may be suitable. Increasing the ensemble size shifts upward the AESS range at the cost of additional computational time. Finally, the coordinates of the optimal measurement set are remarkably consistent over the range of *N* and *M* explored for this analysis. Increasing the number of data realizations averages out the effects of the underlying structural and parameter uncertainties. The results shown here suggest that setting *M* on the order of 500 is adequate for problems similar in complexity to the one considered here. Overall, the benchmarking runs summarized by Table 1 suggest that the preposterior estimation procedure yields reliable estimates of $E[\Phi_{cm}]$.

3.2. Case Study 2: Closed Groundwater Basin

The second case study considers groundwater development in a closed, or endorheic, groundwater basin. Groundwater discharge from a closed basin may occur as spring discharge or direct evaporation from saturated soil. Groundwater development in a closed basin will ultimately reduce the magnitude of the fluxes corresponding to both outflow mechanisms. The potential ecological consequences of reductions of spring discharge may be of great importance for the sustainable long-term management of closed groundwater basins. The objective of this case study is to identify measurements capable of predictive discrimination, where the prediction of interest is the long-term reduction in spring discharge rate due to groundwater pumping. The measurements are to be identified and conducted prior to groundwater development, and then used to inform ensemble predictions of long-term, postdevelopment changes in spring discharge.

A synthetic, closed basin is used in this case study; Table 2 presents relevant characteristics of the basin. The hydrogeologic structure of the basin is intended to reflect horst and graben geologic environments, for which down-dropped basin fill sediments constitute a productive unconfined aquifer bounded by largely impermeable bedrock. The hydraulic properties of the hydrogeologic units are assumed to be unknown, but internally homogeneous; this assumption is frequently employed for the purposes of basin-scale groundwater modeling applications. Figure 6 illustrates the hydrogeologic structure of the basin in plan-view and cross section, including unit boundaries and the location of pertinent hydrologic features. Inflows to the basin consist of distributed in-place recharge applied over the entire basin area, and three distinct zones of mountain front recharge. Outflows to the basin consist of spring discharge at two locations, and soil evapotranspiration in the basin center. The locations of available hydraulic head observations and proposed pumping wells are shown in Figure 6. Both of the pumping wells are screened between −100 to −150 m below land surface, and are each assumed to pump at 500 $m^3\ d^{-1}$ under postdevelopment conditions.

Table 2. Hydrogeologic and Numerical Model Specifications Applicable to All Five Conceptual Models of Groundwater Flow in the Closed Hydrologic Basin

Basin Dimensions	
Basin Area	37.5 km^2
Topographic high	10 m
Topographic low	0 m
Hydrogeologic Characteristics	
Ratio of horizontal to vertical hydraulic conductivity	10
Maximum sediment thickness	160 m
Elevation of soil evaporation area	0 m
Area of soil evaporation area	0.74 km^2
Maximum soil evaporation rate	5.0 mm d^{-1}
Evaporation extinction depth	2.0 m
Elevations of springs	$\{0.5, 0.5\}$
Area of mountain front recharge zones $\{A, B, C\}$	$\{0.2, 0.2, 0.2\}$
Elevations of mountain front recharge zones $\{A, B, C\}$	$\{10, 10, 10\}$

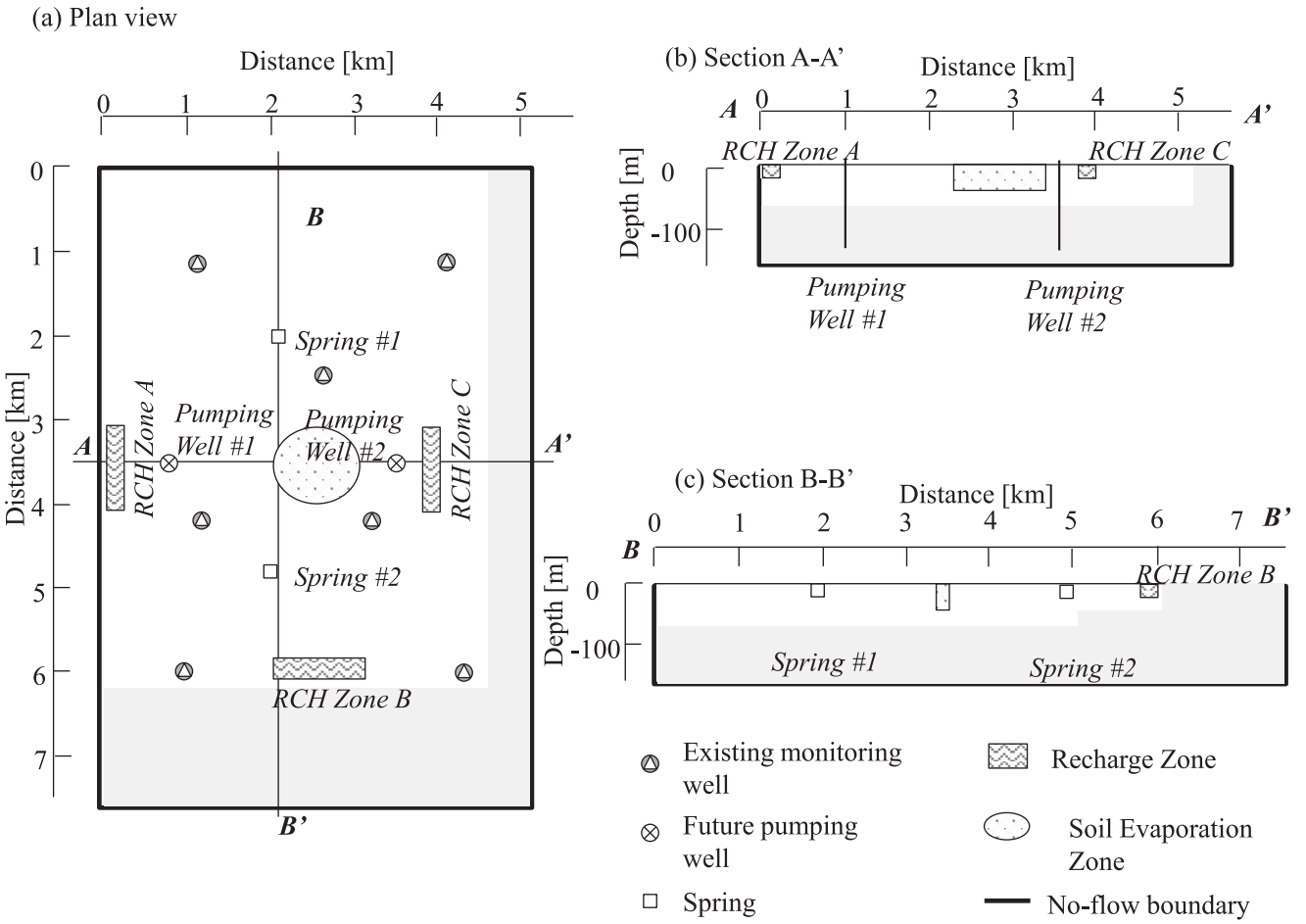

Figure 6. Hydrogeologic structure and hydrologic features of closed basin; gray and white areas represent hydraulic conductivity zones 1 and 2, respectively. Depth of the bottom of model layer 1 is variable. Depths of the bottoms of model layers 2 and 3 are at −100 and −150 m below land surface, respectively.

3.2.1. Alternate Hydrogeologic Conceptual Models

We assume that initial characterization of the basin has led to a set of four alternate conceptual models in addition to the conceptual model described above. The alternate conceptual models are based on plausible hydrogeologic structural uncertainties encountered in basin-scale groundwater investigations. Figure 7 illustrates the distinctive conceptual model characteristics. Conceptual model #1 (CM-01) consists of the basin as described in section 3.2, above. Conceptual model #2 (CM-02) includes the presence of two lenses, for which the hydraulic properties are considered be substantially different from adjacent aquifer units. Figure 7a illustrates the areal extent of the lenses; the minimum and maximum elevations of lens material are 0 and −50 m for lens A, and −25 and −100m for lens B. Conceptual model #3 (CM-03) includes an extensive lens to the west and northwest of the soil evaporation zone, as illustrated in Figure 7b. Conceptual model #4 (CM-04) represents mountain front recharge as occurring through discrete stream features rather than as a continuous line parallel to the mountain front. Finally, conceptual model #5 (CM-05) considers the possibility of subsurface zone in the northern portion of the basin, through which water may be transmitted as underflow to an adjacent basin.

3.2.2. Groundwater Simulation Approach

Our objective is to identify sets of measurements – including hydraulic heads, spring discharge, and recharge flux – that are optimal for predictive discrimination. We therefore require for each ensemble member a list of predictions for the candidate measurements under predevelopment conditions. We also need a list of the predicted spring discharge values under postdevelopment conditions. We are specifically interested in the long-term changes to spring discharge under postdevelopment conditions. The required

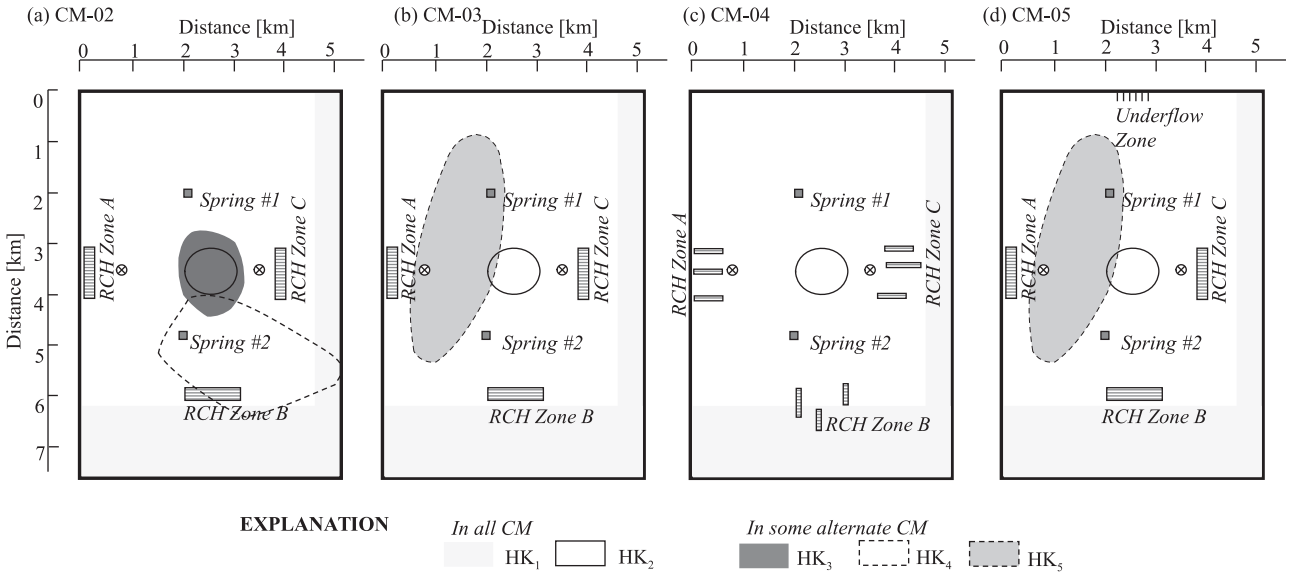

Figure 7. Differences in hydraulic property zonation and locations of hydrologic features between (a) CM-01, (b) CM-02, (c) CM-03, (d) CM-04. Dashed outline for hydraulic property zones indicates buried features.

predictions can be calculated by solving the steady state groundwater flow equation first under predevelopment, then under postdevelopment conditions:

$$\frac{\partial}{\partial x}\left(K_x\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_y\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_z\frac{\partial h}{\partial z}\right)+S=0 \tag{19}$$

In equation (19), h and K are the hydraulic head and hydraulic conductivity, respectively. The S term in equation (19) represents internal sources and sinks to the aquifer. For predevelopment conditions, this term includes recharge, soil evaporation, and spring discharge. For postdevelopment conditions, this term includes the mechanisms listed above and groundwater pumping. Equation (19) is solved numerically for each of the five conceptual models using MODFLOW-NWT [*Niswonger et al.*, 2011]. The model grid is rectilinear, consisting of three layers; each layer includes 75 rows and 50 columns. Constant saturated thickness is specified for layers 2 and 3, and the saturated thickness in layer 1 is allowed to vary as a function of hydraulic head, representing an unconfined aquifer. Cell sizes are uniform – 100×100 m – in the horizontal, and of variable thickness in the vertical, as shown in Figures 6b and 6c. Predicted values of the prospective candidate measurements, $\hat{\mathbf{d}}$, are calculated by evaluating equation (19) under predevelopment conditions. On the other hand, predicted values of the spring-flows are calculated by evaluating equation (19) under both pre and postdevelopment conditions to determine spring depletion due to groundwater pumping. Consequently, equation (19) is evaluated twice for each ensemble member.

3.2.3. Generation of Simulation Ensemble and Data Realizations

The simulation ensemble for this case study is populated using parameter realizations from each of the five conceptual models described above. The uncertain parameters of each conceptual model are summarized in Table 3, and consist of both hydraulic conductivity and recharge parameters. We used the DREAM [*Vrugt et al.*, 2009] algorithm with 150,000 total function evaluations per conceptual model to evaluate the posterior density under each conceptual model for predevelopment conditions. The observational data $\mathbf{d}^{u-1}$ consist of seven hydraulic head measurements generated using a randomly selected conceptual model and parameterization that were not contained within the simulation ensemble. The simulated data were then contaminated with random measurement noise following the error distribution described in section 3.2.5. We adopt a perfectly uninformative prior, with equal probability weights assigned to the conceptual models, and uniform distribution on the parameters for each conceptual model. The multinormal log-likelihood function described by equation (10) was used together with the prior to calculate the posterior density.

Table 3. Uncertain Parameters Considered Over All Conceptual Models[a]

Parameter	Description	Minimum/ Maximum Values	Conceptual Models With Parameter
HK_1	Horizontal hydraulic conductivity in zone 1	0.001/100	all
HK_2	Horizontal hydraulic conductivity in zone 2	0.001/100	all
HK_3	Horizontal hydraulic conductivity in zone 3	0.001/100	CM-02
HK_4	Horizontal hydraulic conductivity in zone 4	0.001/100	CM-02
HK_5	Horizontal hydraulic conductivity in zone 5	0.001/100	CM-03, CM-05
RCH_A	Recharge in zone A	0.00001/0.05	all
RCH_B	Recharge in zone B	0.00001/0.05	all
RCH_C	Recharge in zone C	0.00001/0.05	all
RCH_{base}	Recharge in zone D	$1.0\times10^{-8}/1.0\times10^{-5}$	all

[a]Units on all parameters are meters per day.

3.2.4. Simulation Ensemble Characteristics

Inspection of the *Gelman and Rubin* [1992] $\hat{R}$-statistic showed that for each conceptual model, the DREAM algorithm converged after approximately 25,000 function evaluations. To ensure the use of postconvergence samples, we extracted from each conceptual model all samples generated by DREAM after 50,000 function evaluations. We split the resulting parameter samples into two groups: the simulation ensemble inputs $\mathbf{s}^{ens}$, and the data realization inputs $\mathbf{s}^{drz}$. We used 30,000 and 10,000 realizations per conceptual model, respectively, resulting in 150,000 total ensemble members, and 50,000 data realizations.

Figure 8 shows BMA statistics for the simulated hydraulic heads in model layer 1. Figure 8a illustrates the BMA expectation of the predicted heads; the direction of groundwater flow is generally from the recharge zones and basin boundaries toward the spring and soil evaporation zones. Figures 8b and 8c illustrate the within and between-model variance, respectively. The within-model variance is pronounced in the vicinity of mountain-front recharge zone A; this result reflects the fact that specified recharge rates were included as uncertain parameters within each conceptual model. The between-model variance is largest near recharge zones A and B; this result can be attributed to different spatial distributions of mountain front recharge between CM-04 and the other conceptual models. The relatively large values of $\mathrm{var}_K[\mathrm{E}(h)]$ in the vicinity of recharge zone B obscure the spatial patterns over most of the domain. Figure 8d illustrates the ratio of the between-model to within-model variance, $\mathrm{var}_K[\mathrm{E}(h)] : \mathrm{E}_K[\mathrm{var}(h)]$, of ensemble predicted hydraulic heads. The spatial patterns of higher values in Figure 8d correspond to areas of the model domain where conceptual model uncertainty is particularly pronounced. For example, high values of $\mathrm{var}_K[\mathrm{E}(h)] : \mathrm{E}_K[\mathrm{var}(h)]$ are observed in the central part of the model domain, likely resulting from the different locations of lenses under CM-02 and CM-03.

Figures 9a–9c summarize the predicted effects of groundwater pumping on discharge from spring #2; results for spring #1 (not shown here) were very similar. Figure 9a shows the distribution of predevelopment spring discharge at spring #2; a relatively small number of ensemble members predict zero predevelopment spring flows, and the expected value of the predevelopment spring flow is approximately 350 $m^3\ d^{-1}$. Figure 9b shows the distribution of postdevelopment spring discharge at spring #2. Under postdevelopment conditions, a greater number of ensemble members predict the elimination of spring flows; overall, the histogram indicates a shift toward reduced spring flows. Figure 9c shows the distribution of the spring depletion, calculated as the difference between pre and postdevelopment spring flows over all of the ensemble members. The distribution in Figure 9c is distinctly bimodal, with modes centered on $-100\ m^3\ d^{-1}$ and $-250\ m^3\ d^{-1}$, representing 10% and 25% of the total groundwater pumping, respectively. We illustrate the selection of meauresments that could discriminate predictions of critical spring flow depletion (between -300 and $-150\ m^3\ d^{-1}$) from less dramatic reductions (between 0 and $-150\ m^3\ d^{-1}$).

3.2.5. Implementation of DI Analysis

For this case study, prospective data sets are ranked by the expected KL-divergence of predictive groups, $\mathrm{E}[\Phi_{pr}]$, before and after data collection. Our initial analyses (sections 3.2.6 and 3.2.7) explore the prospective worth of individual measurements, within each measurement type; in other words, we consider a single head, or spring flow measurement. For these initial analyses, we evaluate $\mathrm{E}[\Phi_{pr}]$ for simulated heads at each node in the model domain, for springs #1 and #2, and recharge zones A-C. Following this preliminary assessment, we evaluate $\mathrm{E}[\Phi_{pr}]$ over groups of prospective measurements, considering simultaneously all of the measurement types. This second analysis (section 3.2.8) requires the use of optimization, as it is infeasible to evaluate all possible combination of measurements. Based on practical considerations of the

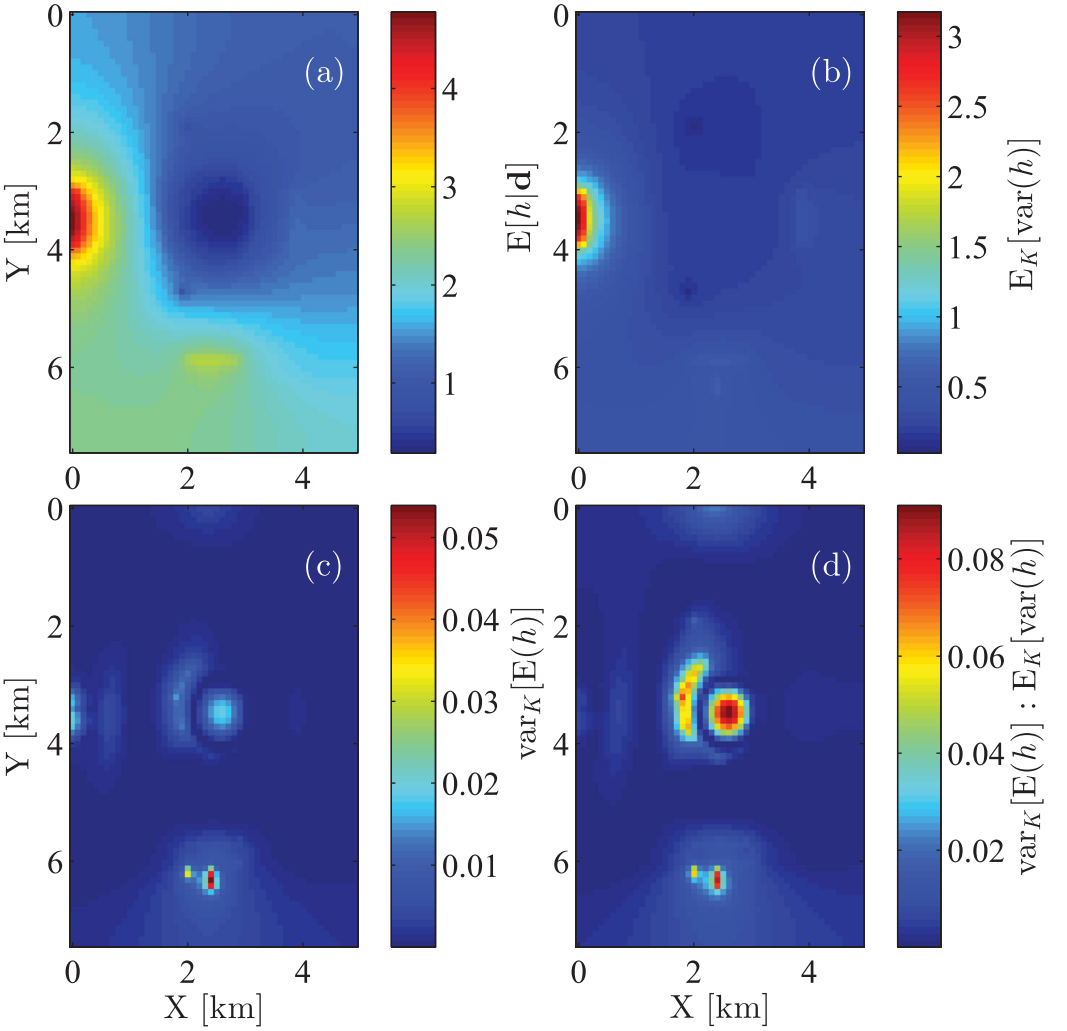

Figure 8. Bayesian model average statistics computed for ensemble-predicted predevelopment heads, in meters, in model layer 1. (a) Posterior expectation, (b) within-model variance, (c) between-model variance, and (d) ratio of between to within-model variance.

computational burden imposed by the number of possible measurement combinations, for the second analysis we use a subset of the candidate head measurements, imposing 500 m spacing in the y-direction between candidate measurements; this reduces the number of possible head measurements from 11,250 to 2250. Both analyses use $M = 500$ realizations of the candidate measurements.

One of the measurement types – mountain front recharge – is also a model parameter; therefore, measured values cannot enter directly into the likelihood function in equation (10), as is done with heads and spring discharge predicted by the ensemble members. Instead, realizations of the mountain front recharge measurements are used to update the parameter prior. Specifically, the zonal mountain front recharge parameter of interest for the j^{th} data realization is treated as a candidate measurement, with measurement error variance assigned as discussed below. Then, the parameter prior over the simulation ensemble for the mountain front recharge parameter of interest is recentered on the measured value, with variance defined by the measurement error variance, and the prior probabilities of the ensemble members are updated accordingly. The ultimate effect on the predictive groupwise probabilities is similar to the effect of recalculating the likelihood function based on model predicted values, but the Bayesian updating mechanism is fundamentally different.

Measurement errors for hydraulic heads are assumed to be homoscedastic, with errors of ± 0.1 m; treating the errors as the 95% confidence interval, this corresponds to a measurement error standard deviation of approximately 0.05 m. Measurement errors for spring flows are assumed to be heteroscedastic, as is typical

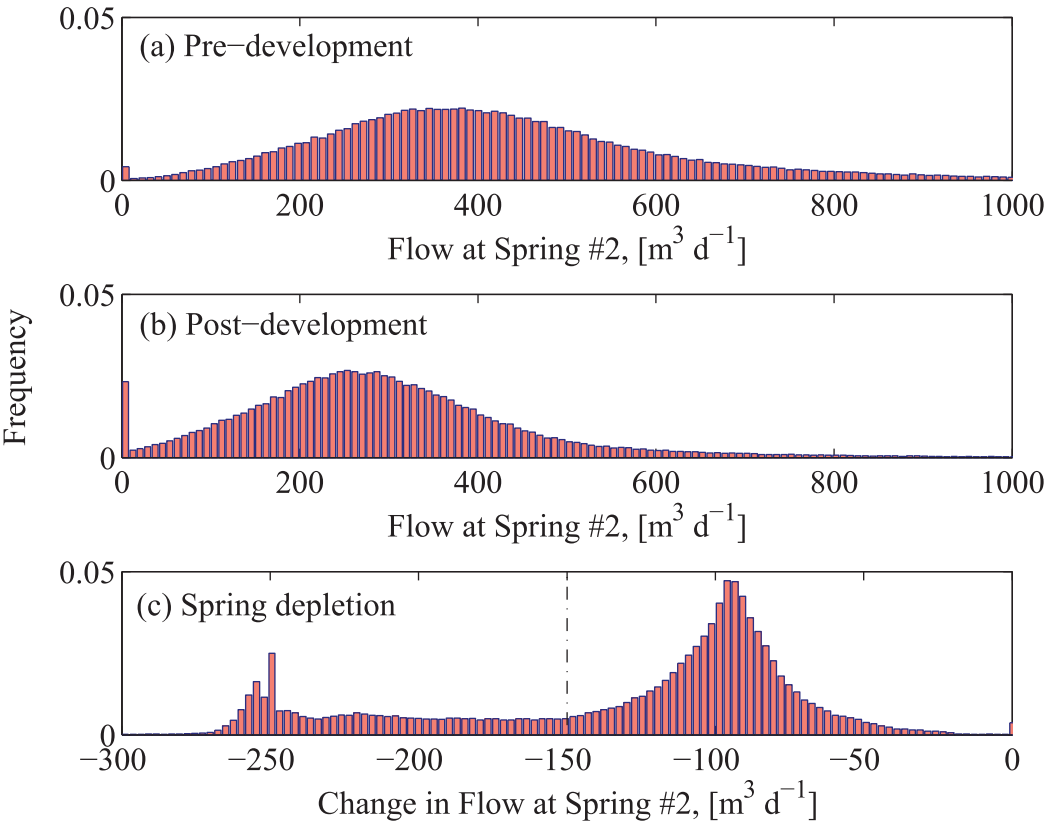

Figure 9. Histograms showing ensemble-predicted discharge at spring #2 under (a) predevelopment conditions, (b) postdevelopment conditions. (c) Spring depletion, equal to the change in discharge between predevelopment and postdevelopment conditions. Vertical dashed line in Figure 9c represents the predictive grouping threshold used for optimal design analyses in section 3.2.8.

of surface water discharge measurement. We assume that the measurement error equals 2% of the measured value. Consequently, the measurement error standard deviation is described as a linear function of the measured flow, $\sigma_\epsilon = 0.0102 Q_{spring}$, where Q_{spring} is the spring discharge.

Measurement techniques for the above-mentioned prospective measurements are standard, with well characterized error models. In contrast, direct measurement of mountain front recharge is much more challenging, and the uncertainty on such direct measurements is in general poorly constrained. Of the many techniques available for quantifying mountain front recharge [*Wilson and Guan*, 2004], one of the most widely accepted is the chloride mass balance technique [*Dettinger*, 1989]. We are unaware of any proposed measurement error model for the chloride mass balance technique – or indeed, for any measurement of mountain front recharge. However, the measurement error can be quantified given the assumed error on the data inputs required for a chloride mass balance calculation—namely, the annual precipitation and basin yield, and mean chloride concentration in precipitation, runoff, and groundwater. We adopt conservative error estimates of 20% on precipitation and basin yield, and 5% on chloride concentration, as was done by *Aishlin and McNamara* [2011]. Treating these error levels as 95% confidence intervals, the variance on the recharge flux is equal to approximately 0.005 m d^{-1}. We assume homoscedastic measurement error on the recharge measurement. Finally, we assume uncorrelated measurement errors among the prospective measurements.

3.2.6. Results of DI Analysis for Individual Head Measurements

Initially, we evaluated $E[\Phi_{pr}]$ over individual measurements within each measurement type. It is important to recognize that changes in groupwise probabilities depend in large part on the threshold delineating predictive groups, which we hereafter refer to as the predictive grouping threshold (PGT). Predictive histograms (Figure 9) suggest that a logical threshold between predictive groups would be located between −150 and −200 $m^3 d^{-1}$. However, we wish to explore the effect of threshold choice on the value of $E[\Phi_{pr}]$. Specifically we evaluate the effect on the distribution of $E[\Phi_{pr}]$ within each measurement type by

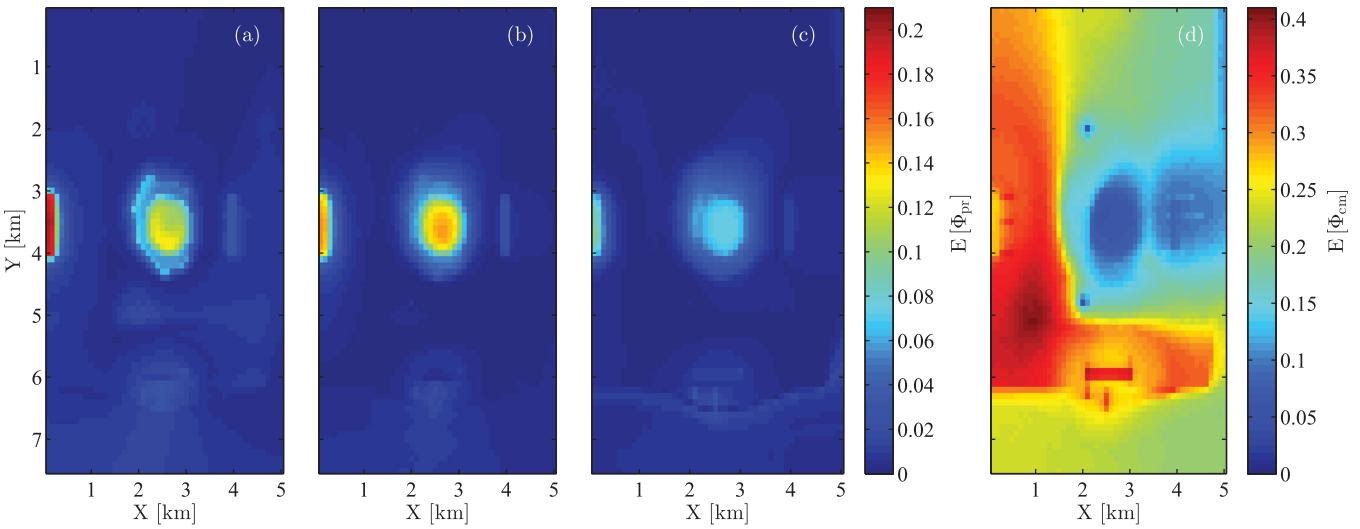

Figure 10. Expectation of the KL-divergence between predictive groups, $E[\Phi_{pr}]$, computed based on the hypothetical addition of candidate head measurements from model layer 1. Plots show changes in the distribution of $E[\Phi_{pr}]$ under different predictive group thresholds: (a) -100 m^3 d^{-1}, (b) -150 m^3 d^{-1}, and (c) -200 m^3 d^{-1}. Figure 10d shows the expected conceptual discrimination, $E[\Phi_{cm}]$, associated with the hypothetical addition of a single head measurement in model layer 1.

systematically varying the PGT. This analysis demonstrates the impact of user-defined PGT on the selection of optimal measurements, and furthermore provides an initial comparison of data worth among the three measurement types considered.

We first consider $E[\Phi_{pr}]$ for individual hydraulic head measurements. Maps of $E[\Phi_{pr}]$ over the nodes in layer 1 of the groundwater model with PGT set at -100, -150, and -200 m^3 d^{-1} are shown in Figure 10a–10c, respectively. A common color scale has been adopted across these three plots to facilitate intercomparison. The magnitudes of $E[\Phi_{pr}]$ calculated for head measurements in layer 1 differ with changes in the PGT value. For example, the maximum value of $E[\Phi_{pr}]$ is much higher in Figure 10a than in Figures 10b and 10c. This result may suggest that head measurements have higher value value for predictive discrimination if the probability mass between the predictive groups is more equally distributed. Such diagnostic exercises can be used to guide general inferences regarding data worth for a specific application.

A striking difference between Figures 10a and 10b is the location of the most informative head measurements. These measurements are located in and surrounding recharge zone A for the lower PGT, emphasizing basin inflows. In contrast, optimal single head measurements located in and surrounding the soil evaporation zone are more important for higher PGT values, balancing improved quantification of basin inflows and outflows, respectively. The relatively high value of head measurements in the soil evaporation zone for Figure 10b likely reflects differences among ensemble members predicting moderate spring depletion in the range of -150 m^3 d^{-1}. For ensemble members in this predictive range, predictive discrimination can be achieved by more accurately partitioning the outflows between soil evaporation and spring discharge. It is also worth noting the distinctive appearance of the soil evaporation zone across all predictive threshold values in Figures 10a–10c. We evaluated the mutual information between heads in the upper aquifer (model layer 1) and the postdevelopment spring depletion; the results showed that the mutual information between layer 1 heads and spring depletion is strongest in the vicinity of the soil evaporation zone; this result is independent of the PGT and instead reflects the dependence of springflow response to groundwater pumping on the partitioning of basin outflows.

For comparison, we show in Figure 10d the distribution of the expected conceptual model discrimination, $E[\Phi_{cm}]$, based on individual head measurements in model layer 1. The $E[\Phi_{cm}]$ surface is strikingly different from the $E[\Phi_{pr}]$ surfaces; conceptual model discrimination emphasizes head measurements in the vicinity of recharge zone B to a much greater extent. Even more pronounced is the importance of head measurements in the west-southwestern quadrant between recharge zones A and B at approximately $\{x, y\} = \{1, 5\}$ km. In contrast to the result from case study 1, the $E[\Phi_{cm}]$ surface for the present case study is not closely related to the ratio of between-model to within-model uncertainty (Figure 10d).

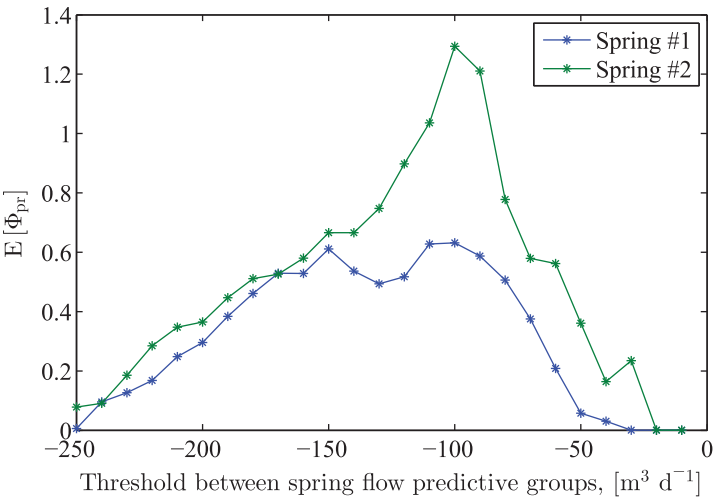

Figure 11. Expected value of the KL-divergence between predictive groups, $E[\Phi_{pr}]$, computed based on the hypothetical addition of candidate spring discharge measurements, evaluated over a range of predictive group thresholds between predictive groups.

To investigate this result, we evaluated the contribution of each conceptual model to the KL-divergence (equation (2)), and consequently, $E[\Phi_{cm}]$. Under the prior distribution (updated following MCMC sampling), CM-02 is assigned approximately 93% of the probability mass, whereas CM-01 is assigned only 0.1% of the probability mass. Decomposition of the KL-divergence—not shown here in the interest of brevity—reveals that the conceptual discrimination $E[\Phi_{cm}]$, as shown in Figure 9d is dominated by changes to the probability of CM-01. Specifically, conceptual discrimination from a single head measurement occurs primarily through redistributing probability mass to CM-01, which had been strongly discredited following the first round of data collection. Head measurements in the west-southwestern quadrant of the basin have a very high potential to test the viability of CM-01 and are therefore expected to provide the greatest degree of conceptual discrimination. Comparison among the alternate conceptual models (Figure 7) supports this conclusion, as several of the hydrogeologic lenses considered under CM-02, CM-03, and CM-05 terminate in the west-southwestern quadrant. The results of this analysis show broadly the importance of measuring heads in areas for which hydrogeologic structures vary among the conceptual models. An important distinction, however, is that important measurements cannot be identified from decomposition of the BMA variance alone. Instead, there are additional interactions within the simulation ensemble that can be captured only through the preposterior analysis. Furthermore, it should be stressed that the results of preposterior analysis may vary depending on the definition of the experimental design objective (e.g., conceptual or predictive discrimination). In fact, the ability to analyze both objectives with the same suite of simulations is one of the strengths of the DI methodology for experimental design.

3.2.7. Results of DI Analysis for Individual Spring Flow and Recharge Measurements

We conducted similar sensitivity analyses between $E[\Phi_{pr}]$ and the predictive threshold for both predevelopment spring discharge and mountain front recharge measurements. Figure 11 illustrates the results of the sensitivity analysis for spring discharge measurements. The numerical value of $E[\Phi_{pr}]$ is much greater for spring discharge measurements than for hydraulic head measurements. This result is consistent with intuition, because the magnitude of postdevelopment spring depletion depends directly on the predevelopment spring discharge. The distribution of $E[\Phi_{pr}]$ diverges most noticeably for springs #1 and #2 for predictive threshold values between -50 to -150 m^3 d^{-1}; within this range, the value of $E[\Phi_{pr}]$ for spring #2 greatly surpasses $E[\Phi_{pr}]$ for spring #1.

The values of $E[\Phi_{pr}]$ for recharge measurements are roughly two orders of magnitude lower than the values $E[\Phi_{pr}]$ for spring discharge measurements (Figure 12). Furthermore, the PGT value has a consistent impact on the discriminatory value of data in all of the recharge zones. This result suggests that there is no clear reason to select observations in a specific recharge zone; this is a somewhat surprising finding given the clear preference for measuring heads near recharge zone A (Figure 10). For all three recharge zones, the highest predictive discrimination for recharge measurement is achieved when the PGT is -220 m^3 s^{-1}. This reflects the characteristics of ensemble members predicting relatively high rates of spring depletion due to groundwater development. These ensemble members are associated with values of mountain front recharge toward the lower end of the feasible parameter space. It therefore stands to reason that

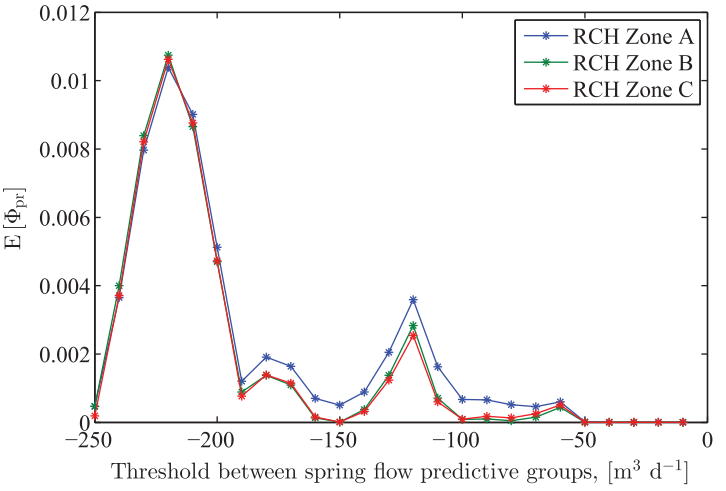

Figure 12. Expected value of the KL-divergence between predictive groups, $E[\Phi_{pr}]$, computed based on the hypothetical addition of candidate recharge measurements, evaluated over a range of predictive group thresholds.

identifying ensemble members associated with especially low recharge values would critically test this predictive group, resulting in high predictive discrimination.

3.2.8. Results of DI Analysis Over Prospective Measurement Sets

Having examined the distribution of $E[\Phi_{pr}]$ for individual measurements within a given measurement type, we now turn our attention to a more practical, and substantially more complex question. Given a fixed number of possible measurements, which collection of measurement types, and locations within each given type, are best suited to acheive predictive discrimination? This presents a nontrivial problem in combinatorial optimization. As a starting point for our analysis, we consider the space of admissible design variables, **D**, to consist of 2,250 possible head measurements, 2 spring discharge measurements, and 3 mountain front recharge measurements. Considering combination with repetition, we now have 2.5 million possible combinations for a set of 2 measurements, and 1.9 billion possible combinations for a set of 3 measurements. Even for a reduced number of measurements, design optimization is clearly needed.

We used the NOMAD algorithm [*Le Digabel*, 2011] to solve equation (13), thereby selecting the optimal measurement set for predictive discrimination given a fixed number of measurements. In general, discrete optimization routines such as NOMAD require selecting the best of several suboptimal solutions [*Christakos*, 1992]. We found that optimal solutions would vary among NOMAD runs; therefore, we used multiple starting points in the parameter space for NOMAD runs, then selected the design associated with the highest value of $E[\Phi_{pr}]$ in order to derive a robust estimate of the globally optimal measurement set. Finally, we repeated the multitry NOMAD optimization runs while systematically varying the number of measurements from 1 to 5. The goal of this procedure was to determine the relation between the size of the candidate measurement set and $E[\Phi_{pr}]$, to evaluate the marginal information gain associated with adding to the number of candidate measurements. Table 4 displays the results of the analysis including the values of $E[\Phi_{pr}]$, measurement coordinates, and computational time.

The results of the optimization analysis show that the value of $E[\Phi_{pr}]$ continues to rise with the number of measurements in the candidate data set, demonstrating that even the fifth measurement contributes unique information. However, the marginal information gain associated with adding measurement points decreases with the number of measurements. That is, increasing the size of the data set exhibits diminishing marginal changes in predictive discrimination. For example, the value of $E[\Phi_{pr}]$ increases by 0.119

Table 4. Results From Multitry NOMAD Optimization Runs, Including the Best $E[\Phi_{pr}]$ Achieved, and Measurements Corresponding to the Best $E[\Phi_{pr}]$ Value[a]

Number of Candidate Measurements	Best $E[\Phi_{pr}]$	Wall-Time (h)	Head Coordinates $\{x, y, z\}$ (m)	Spring Discharge Observations	Recharge Zone Observations
1	0.358	6.54		Spring #2	
2	0.477	18.5		Spring #1, #2	
3	0.505	35.95	{50, 3450, −125}	Spring #1, #2	
4	0.528	51.7	{50, 3450, −125}, {2650, 6950, −125}	Spring #1, #2	
5	0.546	81.5	{50, 3450, −125}, {2650, 6950, −125}, {2550, 5950, −125}	Spring #1, #2	

[a]Each row is based on the best $E[\Phi_{pr}]$ result from 25 individual NOMAD runs.

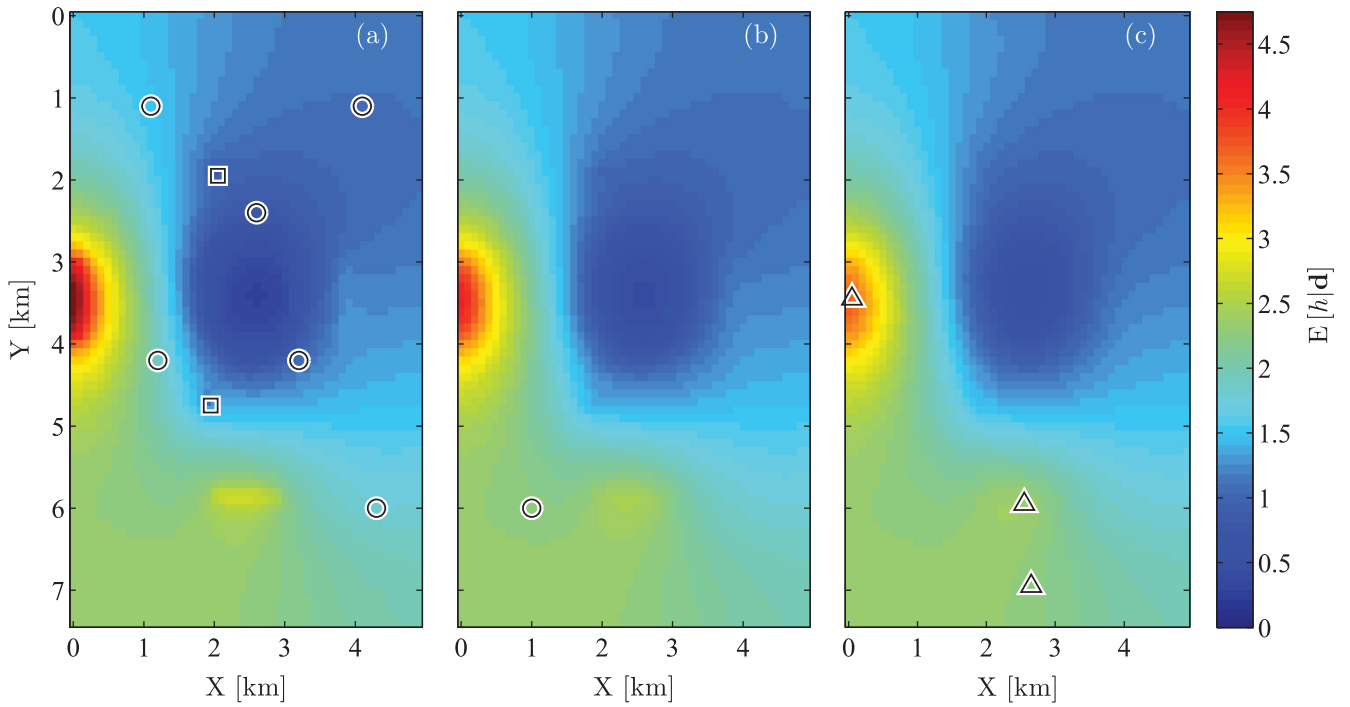

Figure 13. Bayesian model average expectation of the hydraulic heads, and existing proposed measurements in (a) model layer 1, (b) model layer 2, and (c) model layer 3. Circles represent existing head measurements, triangles represent candidate head measurements in the optimal measurement set, and squares represent spring discharge measurements in the optimal measurement set.

when considering two, rather than one candidate measurements, but only by 0.018 when considering five, rather than four candidate measurements. Evaluating the marginal information gain with size of the candidate measurement set provides an opportunity to weigh the expected information benefits from additional data points against the cost of additional data collection.

All of the measurement sets include measurement of predevelopment discharge from spring #2. Intuitively, this is a logical result, as the experimental design objective is to discriminate between predictive groups of flow depletion at spring #2; this result is furthermore consistent with the preliminary sensitivity analyses (Figure 11). Flow measurements at spring #1 are included for sets comprising at least two candidate measurements. Coupled measurements of flow at spring #1 and #2 greatly constrain the partitioning of basin outflows between springs and soil evaporation, as discussed in greater detail below.

The highest value of $E[\Phi_{pr}]$ was obtained for a set of five prospective measurements; Figure 13 shows the locations of the measurements in the optimal set. None of the optimal measurement sets includes recharge measurements, which we found to be surprising and counterintuitive given the great efforts typically made to quantify recharge. From a water balance perspective, recharge is the only inflow to the basin, and should therefore dictate to a large extent the expected severity of spring depletions due to groundwater pumping. However, basin outflow is divided (with the exception of CM-05) between two outflow mechanisms: spring discharge, and soil evaporation from the basin center. The simulated basin outflow from soil evaporation is, on average, five times larger than the simulated outflow from springs. Therefore, the partitioning of basin outflows between soil evaporation and spring discharge is also an important consideration in predicting postdevelopment spring depletion. This is complicated by the fact that the evaporation flux is defined as a function of head in the soil evaporation zone. Therefore, one could imagine a situation in which groundwater pumping might lower the head in the soil evaporation zone, reducing the evaporation outflows and therefore reducing the magnitude of postdevelopment spring depletion. These dynamics cannot be assessed based solely on water budget considerations, as is discussed more generally by *Bredehoft* [2002].

Set against this backdrop, the selected optimal measurement set is quite informative. The partitioning of basin outflows between spring discharge and soil evaporation under postdevelopment conditions strongly

controls the predicted spring depletion. Therefore, predevelopment measurements of groundwater flow patterns related to the outflow partitioning, and possibly influenced by groundwater development, should be particularly informative. Direct measurement of mountain front recharge may constrain the magnitude of inflows to the groundwater basin, but not the outflow partitioning. In contrast, hydraulic head measurements along flow paths from the recharge area to the soil evaporation zone provide important insight into outflow partitioning. This case study demonstrates the importance of experimental design analysis before collecting data and beyond its specific use in identifying measurement points. Indeed, improved understanding of which data achieve highest discrimination provides insight into the hydrologic function of the system in the context of a specific question (e.g., predicted spring depletion).

All three of the hydraulic head measurements in the optimal set of five candidate measurements are located in the lower aquifer unit (model layer 3). These head measurements contain information about the magnitude of the flux along a regional flow path to the basin outflow areas, for which a large quantity would potentially be intercepted by the pumping wells. One of the head measurements is directly beneath recharge zone A; it should be noted that the expected value of recharge flux in zone A is the highest of the three mountain front zones. This result is consistent with the PGT sensitivity analysis (Figure 10), which identified head measurements near recharge zone A as especially informative. The remaining two head measurements are located in the vicinity of recharge zone B. These measurements provide information about the magnitude of the regional flux toward the basin outflows (springs and soil evaporation zone) through the lower aquifer unit.

For this case study, the selection of hydraulic head measurements is related to information redundancy among data points, as shown for case study #1. In a separate analysis not shown here, we evaluated the mutual information between head measurements and predevelopment spring flow measurements. The head measurements in model layer 3, in the southern part of the basin, exhibit relatively minor mutual information with the predevelopment spring flows. Head measurements in the southern part of the basin quantify the deeper groundwater flux to the basin outflows, which is potentially subject to capture by the pumping wells. Furthermore, these head measurements share minimal information with other data points in the optimal measurement set. Based on these considerations, the selected head measurements have clear value for predictive discrimination, yet are somewhat counterintuitive and could easily be overlooked in the absence of a systematic analysis during the planning stages of a field investigation.

4. Discussion and Conclusions

To date, efforts to guide optimally informative experimental and monitoring network designs in hydrology have focused primarily on objectives relating to parameter identification or prediction uncertainty reduction. State-of-the-art experimental design approaches—namely Monte Carlo simulation [e.g., *Leube et al*., 2012] and data assimilation [*Kollat et al*., 2011]—have generally been adapted to target these particular kinds of objectives. This research proposes a novel objective driving the collection of new data sets to be the discrimination achieved among competing model structures and predictive groupings. The DI methodology presented here uses a data utility function based on the distance between prior and posterior probability distributions to assess the discriminatory capabilities of candidate data sets. The probability distributions can either consist of conceptual probabilities in the case of conceptual discrimination, or predictive probabilities in the case of predictive discrimination.

From a practical standpoint, implementation of the DI methodology requires that the user specify the simulation ensemble size, N, and the number of data realizations, M. As part of this research, we evaluated for case study #1 the effect of changing both the ensemble size and number of data realizations on the convergence of the preposterior metric. We found that, in general, a relatively small number of data realizations may be suitable to obtain robust estimates of the optimal measurement set. However, this should be reconsidered for more complex problems. The choice of likelihood function is another key consideration, and should reflect any correlation or heteroscedasticity present in model and measurement errors [*Schoups and Vrugt*, 2010]. Finally, the BMA statistics are implicitly conditional on the set of conceptual models used in the analysis [*Hoeting et al*., 1999]. Therefore, the potential for conceptual discrimination depends directly on the choice of the K conceptual models and indeed, the DI framework will be most informative when a comprehensive set of conceptual models is considered.

The composition of optimal measurement sets for conceptual and predictive discrimination can be explained in part by characteristics of the underlying simulation ensemble. Case study #1 demonstrates that the ratio of between-model to within-model variance is one such important characteristic. Another key aspect, when multiple measurements are considered, is the minimization of redundant information. These general rules of thumb are useful as an initial screening for evaluating candidate data sets during the planning stages of a hydrologic investigation. However, the results of this study show that additional interaction among the processes of measurement selection, Bayesian updating, and discrimination, are best handled in a preposterior framework.

We also investigated the sensitivity of expected predictive discrimination, $E[\Phi_{pr}]$, to the specification of predictive groups. This kind of analysis demonstrates that the selection of predictive grouping threshold may strongly influence the relative importance of different measurements. On the other hand, certain measurement types—such as spring flow measurements in case study #2—are substantially more informative than other measurement types, regardless of the predictive grouping threshold. We did not undertake a systematic comparison of optimal data sets for conceptual versus predictive discrimination in this study; however, initial analyses, shown in Figure 10, indicate that the composition of the optimal measurement set differs between conceptual and predictive discrimination. In fact, it is likely that the optimal data set will be unique for each prediction or set of predictions of interest, underlining the importance of experimental design analyses that are tailored to each investigation.

The optimization algorithm as implemented here requires a fixed number of measurements to be specified. A logical next step would be to extend the optimization procedure to include the number of measurements as one of the decision variables, thereby allowing for an assessment of diminishing returns in additional data collection. Alternately, cost minimization may be specified as one of the objectives addressed by optimization. Numerous techniques previously developed for optimal design of experiments in hydrology specify multiple objectives driving data collection. Expanding the preposterior framework to jointly consider multiple objectives including discrimination, parameter identification, predictive uncertainty reduction, and cost minimization will provide further insight into the characteristics of optimally informative hydrologic data sets.

Acknowledgments

We gratefully acknowledge comments from the anonymous reviewers, which significantly improved the quality of this manuscript. This research benefitted from the public availability of the MILCA, http://www.ucl.ac.uk/ion/departments/sobell/Research/RLemon/MILCA/MILCA, tools to calculate mutual information both case studies. This work was funded under National Institutes for Water Resources grant 2010AZ412G. No data were used in producing this manuscript.

References

Aishlin, P., and J. P. McNamara (2011), Bedrock infiltration and mountain block recharge accounting using chloride mass balance, *Hydrol. Processes*, *25*(12), 1934–1948, doi:10.1002/hyp.7950.

Ajami, N. K., G. M. Hornberger, and D. L. Sunding (2008), Sustainable water resource management under hydrological uncertainty, *Water Resour. Res.*, *44*, W11406, doi:10.1029/2007WR006736.

Ali, S. M., and S. D. Silvey (1966), A general class of coefficients of divergence of one distribution from another, *J. R. Stat. Soc., Ser. B*, *28*(1), 131–142.

Box, G. E. P., and W. J. Hill (1967), Discrimination among mechanistic models, *Technometrics*, *9*(1), 57–71.

Bredehoft, J. D. (2002), The water budget myth revisited: Why hydrogeologists model, *Ground Water*, *40*(4), 340–345.

Buzzi-Ferraris, G., and P. Forzatti (1983), A new sequential experimental design procedure for discriminating among rival models, *Chem. Eng. Sci.*, *38*(2), 225–232.

Brooks, R. H., and A. T. Corey (1964), *Hydraulic properties of porous media: Hydrology Papers*, Colorado State University, 24 pp., Fort Collins, Colo.

Chaloner, K., and I. Verdinelli (1995), Bayesian experimental design: A review, *Stat. Sci.*, *10*(3), 273–304.

Christakos (1992), *Random Field Models in Earth Sciences*, 2nd ed., 474 pp., Dover, Mineola, N. Y.

Cleveland, T. G., and W. W-G. Yeh (1990), Sampling network design for transport parameter identification, *J. Water Resour. Plann. Manage.*, *116*(6), 764–783.

Currie, J., and D. I. Wilson (2012), OPTI: Lowering the Barrier Between Open Source Optimizers and the Industrial MATLAB User, Foundations of Computer-Aided Process Operations, Georgia, USA.

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., 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.

Dettinger, M. D. (1989), Reconnaissance estimates of natural recharge to desert basins in Nevada, U.S.A., by using chloride-balance calculations, *J. Hydrol.*, *106*, 55–78.

D'Odorico, P., L. Ridolfi, A. Porporato, and I. Rodriguez-Iturbe (2000), Preferential states of seasonal soil moisture: The impact of climate fluctuations, *Water Resour. Res.*, *36*(8), 2209–2219.

Dokou, Z., and G. Pinder (2009), Optimal search strategy for the definition of a DNAPL source, *J. Hydrol.*, *376*(3-4), 542–556.

Doucet, A., and A. M. Johansen (2008), A tutorial on particle filtering and smoothing: Fifteen years later, in *Handbook of Nonlinear Filtering*, edited by D. Crisian and B. Rozovsky, Oxford Univ. Press.

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.*, W03019, doi:10.1029/2003WR002901.

Foglia, L., S. W. Mehl, M. C. Hill, and P. Burlando (2013), Evaluating model structure adequacy: The case of the Maggia Valley groundwater system, southern Switzerland, *Water Resour Res.*, *49*, 260–282, doi:10.1029/2011WR011779.

Gelman, A. and D. B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Stat. Sci.*, *7*(4), 457–511.

Glasgow, H. S., M. D. Fortney, J. Lee, A. J. Graettinger, and H. W. Reeves (2003), MODFLOW-2000 head uncertainty, a first-order second moment method, *Ground Water*, *41*(3), 342–350.
Hanley, J. A., and B. J. McNeil (1982), The meaning and use of the area under a Receiver receiver Operating Characteristic (ROC) Curve, *Radiology*, *143*(1), 29–36.
Herrera, G. S., and G. F. Pinder (2005), Space-time optimization of groundwater quality sampling networks, *Water Resour. Res.*, *41*, W12407, doi:10.1029/2004/WR003626.
Hill, M. C., C. C. Faunt, W. R. Belcher, D. S. Sweetkind, C. R. Tiedeman, and D. Kavetski (2013), Knowledge, transparency, and refutability in groundwater models, and example from the Death Valley regional groundwater flow system, *Phys. Chem. Earth*, *64*, 105–116.
Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*(4), 382–417.
Hunter, W. G., and A. M. Reiner (1965), Designs for discriminating between two rival models, *Technometrics*, *7*, 307–323.
James, B. R., and S. M. Gorelick (1994), When enough is enough: The worth of monitoring data in aquifer remediation design, *Water Resour. Res.*, *30*(12), 3499–3513.
Knopman, D. S., and C. I. Voss (1988), Discrimination among one-dimensional models of solute transport in porous media: Implications for sampling design, *Water Resour. Res.*, *24*(11), 1859–1876.
Kollat, J. B., P. M. Reed, and J. R. Kasprzyk (2008), A new epsilon-dominated hierarchical Bayesian optimization algorithm for large multiobjective monitoring network design problems, *Adv. Water Resour.*, *31*, 828–845.
Kollat, J. B., 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, doi:10.1029/2010WR009194.
Kosugi, K. (1996), Lognormal distribution model for unsaturated soil hydraulic properties, *Water Resour. Res.*, *32*(9), 2967–2703.
Kraskov, A., H. Stögbauer, and P. Grassberger (2004), Estimating mutual information, *Phys. Rev. E.*, *32*(9), 2967–2703.
Kullback, S., and R. A. Leibler (1951), On Information and Sufficiency, *Ann. Math. Stat.*, *22*(1), 79–86.
Le Digabel, S. (2011), Algorithm 909: NOMAD: Nonlinear Optimization with the MADS Algorithm, *ACM Trans. Math. Software*, *37*(4), 1–15.
Leube, P. C., 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.
Lewis, S. M., and A. E. Raftery (1997), Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator, *J. Am. Stat. Assoc.*, *92*(438), 648–655.
Liu, J. S. (2008), *Monte Carlo Strategies in Scientific Computing*, 360 pp., Springer, N. Y.
Liu, X., J. Lee, P. K. Kitandis, J. Parker, and U. Kim (2012), Value of information as a context-specific measure of uncertainty in groundwater remediation, *Water Resour. Manage.*, *26*, 1513–1535.
Lu, D., M. Ye, S. P. Neuman, and L. Xue (2012), Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs, *Adv. Water Resour.*, *35*, 69–82.
Milly, P. C. D. (2001), A minimalist probabilistic description of root zone soil water, *Water Resour. Res.*, *37*(3), 457–463.
Neuman, S. P. (2003), Maximum likelihood Bayesian model averaging of alternative conceptual-mathematical models, *Stochastic Environ. Res. Risk Assess.*, *24*, 863–880, doi:10.1007/s00477-003-0151-7.
Neuman, S. P., L. Xue, M. Ye, and D. Lu (2012), Bayesian analysis of data-worth considering model and parameter uncertainties *Adv. Water Resour.*, *36*, 75–85.
Niswonger, R. G., S. Panday, and M. Ibaraki (2011), MODFLOW-NWT, A Newton formulation for MODFLOW-2005, *U.S. Geol. Surv. Tech. Methods*, 6-A37, 44 pp.
Nowak, W., F. P. J. de Barros, and Y. Rubin (2010), Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain, *Water Resour. Res.*, *46*, W03535, doi:10.1029/2009WR008312.
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.
Rawls, W. J., R. A. Lajpat, D. L. Brakensie, and A. Shirmohammadi (1993), Infiltration and soil water movement, in *Handbook of Hydrology*, edited by D. R. Maidment, pp. 5.1–5.8, McGraw-Hill, Inc.
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.
Reichard, E. G., and J. S. Evans (1989), Assessing the value of hydrogeologic information for risk-based remedial action decisions, *Water Resour. Res.*, *25*(7), 1451–1460.
Schaap, M. G., F. J. Leij, and M. Th. van Genuchten (2001), ROSETTA: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions, *J. Hydrol.*, *251* 163–176.
Schöniger, A., T. Wöhling, L. Samaniego, and W. Nowak (2014), Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence, *Water Resour. Res.*, *50*, 9484–9513, doi:10.1002/2014WR016062.
Schoups, G., and J. A. Vrugt (2010), A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors, *Water Resour. Res.*, *46*, w10531, doi:10.1029/2009WR008933.
Steyerberg, E. W., A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan (2010), Assessing the performance of prediction models: A framework for traditional and novel measures, *Epidemiology*, *21*(1), 128–138.
Storn, R., and K. Price (1997), Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces, *J. Global Optim.*, *1*, 341–359.
Sun, N. Z., and W. W. G. Yeh (2007), Development of objective-oriented groundwater models: 2. Robust experimental design, *Water Resour. Res.*, *43*, W02421, doi:10.1029/2006WR004888.
Tonkin, M. J., C. R. Tiedeman, E. D. Matthew, and M. C. Hill (2007), OPR-PPR, a computer program for assessing data importance to model predictions using linear Statistics, *U.S. Geol. Surv. Tech. Methods*, TM–6E2, 115 pp.
Topp, G. C., J. L. Davis, and A. P. Annan (1980), Electromagnetic determination of soil water content: Measurements in coaxial transmission lines, *Water Resour. Res.*, *16*(3), 574–582.
Tsai, F. T-C., and X. Li (2008), Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window, *Water Resour. Res.*, *44*, W09434, doi:10.1029/2007WR006576.
Usunoff, E., J. Carrera, and S. F. Mousavi (1992), An approach to the design of experiments for discriminating among alternative conceptual models, *Adv. Water Resour.*, *15*, 199–214.
Van Genuchten, M. Th. (1980), A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, *Soil Sci. Soc. Am. J.*, *44*, 892–898.
Vrugt, J. A., W. Bouten, H. V. Gupta, and S. Sorooshian (2002), Toward improved identifiability of hydrologic model parameters: The information content of experimental data, *Water Resour. Res.*, *38* (12), 1312, doi:10.1029/2001WR001118.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon (2009), Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, *10*(3), 273–290.

Wilson, J. L., and H. Guan (2004), Mountain-block hydrology and mountain-front recharge, in *Groundwater Recharge in a Desert Environment: The Southwestern United States*, edited by J. F. Hogan, F. M. Phillips and B. R. Scanlon, AGU, Washington, D. C., doi:10.1029/009WSA08.

Wöhling, T., and J. A. Vrugt (2008), Combining multiobjective optimization and Bayesian model averaging to calibrate forecase ensembles of soil hydraulic models, *Water Resour. Res.*, *44*, W12432, doi:10.1029/2008WR007154.

Wöhling, T., A. Schöniger, S. Gayler, and W. Nowak (2015), Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction, *Water Resour. Res.*, *51*, 2825–2846, doi:10.1002/2014WR016292.

Ye, M., K. F. Pohlmann, J. B. Chapman, G. M. Pohll, and D. M. Reeves (2010), A model-averaging method for assessing groundwater conceptual model uncertainty, *Ground Water*, *48*(5), 716–728.

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