1 **Bayesian Experimental Design for Identification of Model Propositions and Conceptual**
2 **Model Uncertainty Reduction**

3
4 Hai V. Pham and Frank T.-C. Tsai*

5
6 Department of Civil and Environmental Engineering, Louisiana State University, 3418G Patrick
7 F. Taylor Hall, Baton Rouge, Louisiana, 70803, USA, Phone: 1-225-5784246, Email:
8 hpham28@lsu.edu; ftsai@lsu.edu

9
10 *Corresponding author

Abstract

The lack of hydrogeological data and knowledge often results in different propositions (or alternatives) to represent uncertain model components and creates many candidate groundwater models using the same data. Uncertainty of groundwater head prediction may become unnecessarily high. This study introduces an experimental design to identify propositions in each uncertain model component and decrease the prediction uncertainty by reducing conceptual model uncertainty. A discrimination criterion is developed based on posterior model probability that directly uses data to evaluate model importance. Bayesian model averaging (BMA) is used to predict future observation data. The experimental design aims to find the optimal number and location of future observations and the number of sampling rounds such that the desired discrimination criterion is met. Hierarchical Bayesian model averaging (HBMA) is adopted to assess if highly probable propositions can be identified and the conceptual model uncertainty can be reduced by the experimental design. The experimental design is implemented to a groundwater study in the Baton Rouge area, Louisiana. We design a new groundwater head observation network based on existing USGS observation wells. The sources of uncertainty that create multiple groundwater models are geological architecture, boundary condition, and fault permeability architecture. All possible design solutions are enumerated using a multi-core supercomputer. Several design solutions are found to achieve an 80%-identifiable groundwater model in five years by using six or more existing USGS wells. The HBMA result shows that each highly probable proposition can be identified for each uncertain model component once the discrimination criterion is achieved. The variances of groundwater head predictions are significantly decreased by reducing posterior model probabilities of unimportant propositions.

Keywords: Bayesian model averaging, experimental design, groundwater, model discrimination, observation network design, uncertainty.

1. Introduction

Identification of a reliable groundwater model for future applications is the ultimate goal of model development. This goal, however, remains challenging due to high uncertainty in groundwater systems. Over the past several decades, many studies have focused on understanding, quantifying, and reducing head prediction uncertainty arising from model parameter uncertainty given a conceptual model [1,2]. Considering only one model may lead to statistical bias and underestimation of uncertainty because groundwater systems are often complex and have multiple interpretations [3–5].

In recent years, conceptual model uncertainty has received much attention in groundwater applications [e.g., 6–14]. Many studies have shown that contribution of conceptual model uncertainty to predictive uncertainty is significantly larger than that of model parameter uncertainty [3–5, 7–10, 14–17]. This motivates the research to consider multi-model methods that seek to obtain multi-model ensemble predictions and analyze uncertainty from a set of candidate groundwater models.

Bayesian model averaging (BMA) [18–21] is often employed to conduct multi-model prediction studies because BMA employs probabilistic techniques to derive consensus predictions from a set of candidate models based on their corresponding posterior model probabilities. Averaged predictions from BMA are less biased than predictions obtained from individual models [22–24]. Moreover, BMA is able to study uncertainty propagation from model parameter uncertainty and model structure uncertainty to model prediction uncertainty, thereby distinguishing prediction uncertainty arising from individual models, between models, and between methods [13,25,26].

The lack of hydrogeological data and knowledge often results in different propositions (or alternatives) to represent uncertain model components and creates many candidate groundwater models using the same data. For example, geological architecture can be one uncertain model component in groundwater modeling. Many hydrostratigraphy modeling techniques may be proposed to construct different geological architectures (propositions) potentially leading to an overwhelming number of models with non-dominant posterior model probabilities. Conducting prediction and uncertainty analysis using a great deal of computationally intensive groundwater models can become intractable. By incorporating many conceptual models, prediction results using BMA can become useless when the prediction uncertainty is very high. This concern highlights the importance of conducting an experimental design to discriminate groundwater model propositions, identify highly probable models, and in turn, reduce conceptual model uncertainty (model choice uncertainty) in prediction.

In the field of groundwater hydrology, many studies have conducted experimental designs to improve parameter estimation [e.g., 13, 27–32], reduce model prediction uncertainty [e.g., 33–36] and minimize decision errors in hypothesis tests [37]. Some studies have used experimental designs to discriminate candidate models and identify the “true” model [38–42]. Knopman and Voss [39] investigated the theoretical discrimination power of designs and suggested sample locations where predictions of candidate models were the most different. They proved the efficiency of the method in a sampling design for a solute transport study at Cape Cod, Massachusetts in a subsequent paper [41]. Usunoff et al. [38] hypothesized that the “true” model is close to one of the candidate models and no admissible parameter sets of remaining candidate models could produce similar predictions. They used an experimental design to

discriminate models based on the distance of model predictions. Yakirevich et al. [42] recently applied the Kulback-Leibler information to discriminate flow and transport models.

Significant efforts have been invested in the development of different discrimination criteria for experimental designs in different fields [42–54], but only a few criteria have been applied to groundwater modeling, e.g., the differences between model outputs [38,39,41,55] and the Kulback-Leibler information [42].

Model discrimination criteria should serve not only for model discrimination, but also for model identification. As discussed in Box and Hill [43], the ultimate goal of maximizing information from the system should aim to make the posterior probability of one model to be 1 and others to be zero. In line with Box and Hill, this study introduces a new experimental design to identify the most probable model under the BMA framework. A discrimination criterion is developed based on maximizing the maximum posterior model probability across all candidate models such that a unique and highly probable groundwater model can be identified. Posterior model probability is a straightforward quantitative indicator for model discrimination and is suitable for the experimental design since it represents model importance based on data evidence. The experimental design is implemented to a groundwater study at Baton Rouge, southeastern Louisiana. The purpose of the experimental design in this case study is to find the optimal number and location of groundwater observation wells and the number of sampling rounds such that a groundwater model meets the desired value of the discrimination criterion.

2. Discrimination Criterion

Unlike prior studies that use differences between model outputs [38,39,41,55] or Kulback-Leibler information [42] to set up a discrimination criterion, this study uses the

posterior model probability to discriminate one model from other candidate models. Let $\mathbf{M} = \{M_p; p = 1, 2, \dots, P\}$ be a set of P candidate models developed by propositions from different sources of model structure uncertainty. The posterior model probability for a model M_p is $\Pr(M_p | \Delta^{obs})$, where Δ^{obs} is existing observation data (e.g., groundwater heads, concentration, fluxes, etc. in aquifers). From Bayes' theorem, the sum of the posterior model probabilities is unity, i.e., $\sum_{p=1}^P \Pr(M_p | \Delta^{obs}) = 1$. A model is said to be γ -identifiable from the rest of the candidate models if its posterior model probability is superior to others and is over a probability threshold, γ . The probability threshold, γ , should be much larger than 50% (e.g., 80%) to ensure a unique and highly probable model. To find the maximum posterior model probability and check if the γ -identifiable model can be identified, we define the discrimination criterion as:

$$\max \left\{ \Pr(M_p | \Delta^{obs}), p = 1, 2, \dots, P \right\} \geq \gamma, \quad (1)$$

where the γ -identifiable model is the best model among the candidate models in terms of the highest posterior model probability.

3. Experimental Design using BMA Prediction

If none of the candidate models meets the probability threshold, the existing observation data Δ^{obs} is insufficient to identify a highly probable model. An experimental design is conducted to seek potential locations and time to collect future observation data through an observation network such that the γ -identifiable model can be identified. The experimental

design is to maximize the maximum posterior model probability across all the candidate models and is defined as follows:

$$D^* = \arg \left[\max_D \max \left\{ \Pr \left(M_p \mid \Delta^{obs}, \Delta_D^{new} \right), p = 1, 2, \dots, P \right\} \geq \gamma \right], \quad (2)$$

where D^* is the optimal design to determine potential locations and time of observations that produces the highest posterior model probability; Δ_D^{new} is the future observation data at potential observation locations and time using a design D ; $\Delta = \{ \Delta^{obs}, \Delta_D^{new} \}$ are the total data. The optimal design should have the maximum posterior model probability $\Pr \left(M_p \mid \Delta^{obs}, \Delta_{D^*}^{new} \right) \geq \gamma$.

Since we do not know the future observation data Δ_D^{new} , these data may be predicted deterministically or stochastically. It is understood that uncertainties in the future observation data may affect posterior model probability calculations, thereby affecting the experimental design results; however, this is not investigated in the current study. In the following sampling approach, we take the deterministic approach and propose BMA mean predictions as the future observation data in the experimental design because the BMA mean prediction presents unbiased estimation.

3.1 Time-Sequential Sampling Approach

Future observation data at different locations and times provide different information about model predictions and its ability to discriminate models. A time-sequential sampling (TSS) approach is used to sequentially collect spatiotemporal future observation data and update posterior model probabilities over time. To best predict future observation data, Δ^{new} , using all

candidate models, BMA [18–21] is adopted to obtain mean predictions. Given a number of observation locations, the experimental design with the TSS approach is as follows:

Step 1: For the first round of data collection at time t_1 , the future observation data presented by the BMA mean predictions are

$$E[\Delta_D^{new}(t_1) | \Delta^{obs}] = \sum_{p=1}^P E[\Delta_D^{new}(t_1) | \Delta^{obs}, M_p] \Pr(M_p | \Delta^{obs}), \quad (3)$$

where $E[\Delta_D^{new}(t_1) | \Delta^{obs}, M_p]$ are the means of the future observation data predicted by model p at time t_1 given the existing observation data Δ^{obs} and a design D , $E[\Delta_D^{new}(t_1) | \Delta^{obs}]$ are the BMA means at time t_1 , and $\Pr(M_p | \Delta^{obs})$ is the posterior model probability of M_p calculated using the existing observation data. This study adopts the large sample assumption of Draper [18], that $E[\Delta_D^{new}(t_1) | \Delta^{obs}, M_p]$ is approximated by $E[\Delta_D^{new}(t_1) | \Delta^{obs}, \hat{\beta}_p, M_p]$, where $\hat{\beta}_p$ is the maximum likelihood estimate of model parameters β_p of model M_p . The total data $\Delta_1 = \{\Delta^{obs}, E[\Delta_D^{new}(t_1) | \Delta^{obs}]\}$ are used to update the posterior model probabilities $\Pr(M_p | \Delta_1)$, which will be explained in the next section. Either an optimization or an enumeration approach can be used to find the best number and location of observations. If the design D satisfies equation (2), data collection is completed. Otherwise, we need a second round to collect additional data.

Step 2: For the second round of data collection at time t_2 , the future observation data presented by the BMA mean predictions are

$$E[\Delta_D^{new}(t_2) | \Delta^{obs}] = \sum_{p=1}^P E[\Delta_D^{new}(t_2) | \Delta^{obs}, M_p] \Pr(M_p | \Delta^{obs}), \quad (4)$$

164 where $E[\Delta_D^{new}(t_2) | \Delta^{obs}, M_p]$ are the means of the future observation data predicted by model p
165 at time t_2 given existing observation data Δ^{obs} and the design D , $E[\Delta_D^{new}(t_2) | \Delta^{obs}]$ are the BMA
166 mean predictions at time t_2 . The total data $\Delta_2 = \{\Delta_1, E[\Delta_D^{new}(t_2) | \Delta^{obs}]\}$ are used to update the
167 posterior model probabilities $\Pr(M_p | \Delta_2)$. If equation (2) is met, the design is completed.
168 Otherwise, the next round of data collection is needed.

169 Step 3: Repeat Step 2 until equation (2) is satisfied. For the k^{th} round of data collection at
170 time t_k , the future observation data presented by the BMA mean predictions are

$$171 \quad E[\Delta_D^{new}(t_k) | \Delta^{obs}] = \sum_{p=1}^P E[\Delta_D^{new}(t_k) | \Delta^{obs}, M_p] \Pr(M_p | \Delta^{obs}), \quad (5)$$

172 where $E[\Delta_D^{new}(t_k) | \Delta^{obs}, M_p]$ are the mean predictions of the future observation data predicted
173 by model p at time t_k given the existing observation data Δ^{obs} and the design D ,
174 $E[\Delta_D^{new}(t_k) | \Delta^{obs}]$ are the BMA mean predictions at time t_k . The total data
175 $\Delta_k = \{\Delta_{k-1}, E[\Delta_D^{new}(t_k) | \Delta^{obs}]\}$ are used to update the posterior model probabilities $\Pr(M_p | \Delta_k)$.

176 The decision variables of the experimental design are the number and location of future
177 observations and the number of sampling rounds. Different types of future observation data Δ_D^{new}
178 can be collected from existing and new observation locations. From an economic perspective,
179 existing observation locations should be considered first to design a new observation network
180 before a new observation location is introduced. In case the discrimination criterion cannot be
181 achieved, one may need to increase observation period, increase sampling frequency, seek new
182 observation locations, or apply all of these strategies. Moreover, if the future observation data

can be collected from the field between sampling rounds, all candidate models should be re-calibrated using both the existing data and the new data. Δ^{obs} and $\Pr(M_p | \Delta^{obs})$ should be updated by including the new data. This study does not conduct model re-calibration because the new observation data are not available. The BMA mean predictions cannot be used to re-calibrate the models.

3.2 Update Posterior Model Probabilities Using BMA Mean Predictions

From Bayes' theorem, the posterior model probability of a candidate model after the k^{th} round of data collection is

$$\Pr(M_p | \Delta_k) = \frac{\Pr(\Delta_k | M_p) \Pr(M_p)}{\sum_{p=1}^P \Pr(\Delta_k | M_p) \Pr(M_p)}, \quad (6)$$

where $\Pr(\Delta_k | M_p)$ is the marginal likelihood and $\Pr(M_p)$ is the prior model probability. The marginal likelihood is

$$\Pr(\Delta_k | M_p) = \int_{\beta_p} \Pr(\Delta_k | \beta_p, M_p) \Pr(\beta_p | M_p) d\beta_p, \quad (7)$$

where β_p is the model parameters of model M_p . Laplace approximation can be used to obtain an estimate of the marginal likelihood [56] as

$$\Pr(\Delta_k | M_p) \approx \exp\left[-\frac{1}{2} \text{BIC}_{p,k}\right], \quad (8)$$

where the Bayesian information criterion (BIC) at the k^{th} round of data collection at time t_k is

$$\text{BIC}_{p,k} = -2 \ln \Pr(\Delta_k | \hat{\beta}_p, M_p) + m_p \ln(N + kN_w), \quad (9)$$

where N is the number of existing observation data; k is the number of sampling rounds; N_w is the number of potential observation locations; and m_p is the number of unknown model parameters to be estimated.

Using equation (8), equal prior model probabilities and the variance window [25] to account for the effect of data size, the posterior model probability in equation (6) is estimated by

$$\Pr(\mathbf{M}_p | \Delta_k) = \frac{\exp\left\{-\frac{1}{2}\alpha\left[\text{BIC}_{p,k} - (\text{BIC}_{p,k})_{\min}\right]\right\}}{\sum_{p=1}^P \exp\left\{-\frac{1}{2}\alpha\left[\text{BIC}_{p,k} - (\text{BIC}_{p,k})_{\min}\right]\right\}}, \quad (10)$$

where α is the scaling factor for the variance window and $(\text{BIC}_{p,k})_{\min}$ is the minimum BIC value of the models. The scaling factor depends on (1) the choice of a BIC difference corresponding to a significance level in Occam's window and (2) the choice of a variance window size in terms of the standard deviation of a chi-square distribution of the BIC, denoted as σ_D . Tsai and Li [25] provide more details on the statistical meaning and selection of the scaling factor.

Considering $\Pr(\Delta_k | \hat{\mathbf{p}}_p, \mathbf{M}_p)$ to be a multivariate Gaussian distribution and assuming independent observation data for Δ_k , which includes the existing data and BMA mean prediction data, equation (9) becomes

$$\text{BIC}_{p,k} = Q_{p,k} + (N + kN_w) \ln(2\pi) + m_p \ln(N + kN_w), \quad (11)$$

where $Q_{p,k}$ is the sum of weighted squared errors. $Q_{p,k}$ is

$$Q_{p,k} = \sum_{i=1}^N \frac{(\Delta_i^{cal}(\hat{\mathbf{p}}_p) - \Delta_i^{obs})^2}{\sigma_i^2} + \sum_{r=1}^k \sum_{j=1}^{N_w} \frac{(\Delta_j^{cal}(t_r; \hat{\mathbf{p}}_p) - \mathbb{E}[\Delta_{D,j}^{new}(t_r) | \Delta^{obs}])^2}{\sigma_{j,r}^2}, \quad (12)$$

219 where Δ_i^{obs} and $\Delta_i^{cal}(\hat{\beta}_p)$ are the i^{th} observed value and calculated value using model p ,
 220 respectively. $\Delta_j^{cal}(t_r; \hat{\beta}_p)$ is the prediction at location j by model p at the r^{th} round of data
 221 collection; and $E[\Delta_{D,j}^{new}(t_r) | \Delta^{obs}]$ is the BMA mean prediction at location j and the r^{th} round of
 222 data collection given the existing observation data Δ^{obs} . The first term at the right side of
 223 equation (12) is the sum of weighted squared errors in the historical data space. The second term
 224 is the sum of weighted squared errors for k rounds of data collection in the prediction space. The
 225 variances σ_i^2 and $\sigma_{j,r}^2$ are calculated by Monte Carlo simulation, which samples a sufficient
 226 number of realizations of the model parameters using the following equations:

$$227 \quad \sigma_i^2 = \frac{1}{P \times S} \sum_{p=1}^P \sum_{s=1}^S \left(\Delta_i^{cal}(\beta_{p,s}) - \Delta_i^{obs} \right)^2, \quad (13)$$

$$228 \quad \sigma_{j,r}^2 = \frac{1}{P \times S} \sum_{p=1}^P \sum_{s=1}^S \left(\Delta_j^{cal}(t_r; \beta_{p,s}) - E[\Delta_{D,j}^{new}(t_r) | \Delta^{obs}] \right)^2, \quad (14)$$

229 where $\beta_{p,s}$ is the s^{th} realization of model parameters for model p , and S is the total number of
 230 realizations.

231

232 4. Proposition Discrimination and Uncertainty Analysis

233 This study adopts the hierarchical BMA (HBMA) method [57] to analyze proposition
 234 discrimination and prediction uncertainty reduction given the experimental design. The HBMA
 235 is different from the BMA in terms of providing more insights on model selection, model
 236 averaging and uncertainty propagation through a BMA tree. The HBMA is able to evaluate
 237 competing propositions of each uncertainty model component, to prioritize different sources of
 238 uncertainty in a hierarchical order, and to understand the uncertainty propagation. Each source of

239 uncertainty develops a level of the hierarchy and suggests a number of propositions to represent
 240 an uncertain model component. Combinations of propositions of all levels create candidate
 241 models $\mathbf{M} = \{M_p; p = 1, 2, \dots, P\}$, which are the base models at the base level of the hierarchy. A
 242 number of BMA models are developed bottom up from the base level to the top-most level by
 243 averaging models under the same propositions of their upper level. The top-most level is called
 244 the hierarch level. The BMA model at the hierarch level is called the hierarch model. For
 245 operational purposes, we define the base level of the hierarchy as level B. The level number
 246 decreases as the level goes from the base level B to the hierarch level. $\mathbf{M}_B$ are the base models
 247 at level B. $\mathbf{M}_{B-1}$ are the BMA models at level B-1, and so forth.

248 After the k^{th} round of data collection, the posterior model probabilities of the base models
 249 are $\Pr(\mathbf{M}_B | \Delta_k)$, which are $\Pr(M_p | \Delta_k)$, $p = 1, 2, \dots, P$ as discussed previously. The posterior
 250 model probabilities at level B-1 are the sums of the posterior model probabilities at level B under
 251 the same propositions of level B-1. That is

$$252 \quad \Pr(\mathbf{M}_{B-1} | \Delta_k) = \sum_i \Pr(\mathbf{M}_B^{(i)} | \Delta_k), \quad (15)$$

253 where $\mathbf{M}_B^{(i)}$ are the models at level B under the same propositions of level B-1. Therefore, the
 254 posterior model probabilities at any level n is

$$255 \quad \Pr(\mathbf{M}_n | \Delta_k) = \sum_i \Pr(\mathbf{M}_{n+1}^{(i)} | \Delta_k). \quad (16)$$

256 The conditional posterior model probabilities at level $n+1$ under a model at level n is

$$257 \quad \Pr(\mathbf{M}_{n+1}^{(i)} | \Delta_k, \mathbf{M}_n) = \frac{\Pr(\mathbf{M}_{n+1}^{(i)} | \Delta_k)}{\Pr(\mathbf{M}_n | \Delta_k)}. \quad (17)$$

Equations (16) and (17) develop a BMA tree of posterior model probabilities and conditional posterior model probabilities. We use equations (16) and (17) to assess if the γ -identifiable propositions can be determined for all uncertain model components by the experimental design.

We evaluate total prediction variances by BMA at any level n by

$$\begin{aligned} \text{Var}(\Delta | \Delta_k, \mathbf{M}_n) = & \sum_i \text{Var}(\Delta | \Delta_k, \mathbf{M}_{n+1}^{(i)}) \Pr(\mathbf{M}_{n+1}^{(i)} | \Delta_k, \mathbf{M}_n) \\ & + \sum_i \left[E(\Delta | \Delta_k, \mathbf{M}_{n+1}^{(i)}) - E(\Delta | \Delta_k, \mathbf{M}_n) \right]^2 \Pr(\mathbf{M}_{n+1}^{(i)} | \Delta_k, \mathbf{M}_n), \end{aligned} \quad (18)$$

where $\text{Var}(\Delta | \Delta_k, \mathbf{M}_n)$ are the total BMA prediction variances of Δ when models $\mathbf{M}_n$ are used.

The first term at the right side of equation (18) is the within-model variances of Δ using models at level $n+1$ and the second term is the between-model variances of Δ using models at level $n+1$. Calculations of both variances are referred to [26] and [57]. Variance reduction can be assessed by comparing the total variances before and after the experimental design.

Prioritization of the sources of model structure uncertainty that form the order of uncertain model components in the BMA tree is subject to the analyst's preference. However, prioritization will not affect the optimal experimental design since this study maximizes the maximum posterior model probability across all the base models, not the BMA models. Equation (16) shows the addition of the posterior model probabilities from the base level to the hierarchy level. Once the optimal experimental design achieves the desired discrimination criterion, γ -identifiable propositions for all uncertain model components are identified at the same time, which compose the best model.

5. Case Study

5.1 Baton Rouge Groundwater Models, Southeastern Louisiana

The methodology is tested using a real-world case study in the Baton Rouge area, in southeastern Louisiana, where we develop models to simulate groundwater flow in the “1,200-foot” sand, the “1,500-foot” sand, and the “1,700-foot” sand. These sand units were named according to their approximate depth below ground level of the Baton Rouge industrial district [58]. The goal of the case study is to design a new groundwater head observation network for discriminating among groundwater models. We use USGS observation wells as potential design locations since groundwater head is the most achievable future data in the study area.

The study area shown in Figure 1 is approximately 500 km^2 ($27.4 \text{ km} \times 18.6 \text{ km}$) and includes a major part of the Baton Rouge metropolitan area. The west-east trending Baton Rouge fault and the Denham Springs-Scotlandville fault cut across the study area. The recent hydrostratigraphy study of Elshall et al. [59] found that the aforementioned sand units are interconnected in the area between the two faults. To include these three sands, the vertical extent of the model domain is from 270 m to 560 m below the National Geodetic Vertical Datum of 1929 (NGVD29).

This study uses MODFLOW-2005 [60] to develop groundwater models. The study area was discretized into 93 rows and 137 columns. The cell size is 200 m by 200 m. In the vertical direction, the model consists of 45 layers with thickness varying from 3.0 meters to 12.8 meters. Groundwater head data were collected from 20 USGS observation wells shown in Figure 1 from 1975 to 2010 and were used for model calibration. Predicted groundwater heads from 2011 to 2015 were used as future observation data in the experimental design problem. There are 432 monthly stress periods for the model calibration and 60 monthly stress periods for the

experimental design.

The study area includes 87 pumping wells and the connector well (EB-1293) shown in Figure 1. Monthly pumping data are available from the Louisiana Capital Area Ground Water Conservation Commission (CAGWCC) from 1975 to 2010. The initial groundwater head on 1/1/1975 and the time-varied head boundary condition were estimated by the measured heads in the USGS observation wells. The detailed approach to determine the initial and boundary conditions for the study area can be found in Tsai and Li [25]. The 20 USGS observation wells in Figure 1 are potential locations to be used in the experimental design.

5.2 Sources of Model Structure Uncertainty

To incorporate conceptual model uncertainty in the experimental design, this study considers three uncertainty sources in developing groundwater models. The first source of uncertainty is in the complex geological architecture, which was reconstructed by the indicator kriging (IK), the indicator zonation (IZ), and the generalized parameterization (GP) methods [59]. Figure 2 shows the different sand distributions for a layer using these methods.

The second source of uncertainty is from the boundary condition south of the Baton Rouge fault, where head boundary values are unclear due to limited head observation data. This study used EB-782A to estimate the head boundary values south of the fault for the “1,200-foot” sand and EB-780A to determine the head boundary values south of the fault for the “1,500-foot” sand and the “1,700-foot” sand. We denote this proposition as “ B_0 ” such that: (1) the head values at the south boundary are 4.5 m higher than those measured at EB-782A for the “1,200-foot” sand; and (2) the head values at the south boundary are 1.5 m higher than those measured at EB-780A for the “1,500-foot” sand and the “1,700-foot” sand. Another two propositions that adjust B_0 by ± 1.5 m are proposed and denoted as “ $B_{-1.5}$ ” and “ $B_{+1.5}$ ”, respectively.

The third source of uncertainty is from the fault permeability architecture. The Baton Rouge fault and the Denham Springs-Scotlandville fault are low-permeability faults and act as horizontal flow barriers. This study uses the hydraulic characteristic (HC) [61], which is barrier hydraulic conductivity divided by barrier width, to represent fault permeability. For simplicity, two propositions are proposed. The first proposition, denoted as “HC2”, considers homogeneous fault hydraulic characteristics for the Baton Rouge fault and for the Denham Springs-Scotlandville fault. Two fault hydraulic characteristic values, one for each fault, will be estimated. The second proposition, denoted as “HC4”, considers the fault hydraulic characteristic at the “1,200-foot” sand is different from that at the “1,500-foot” sand and the “1,700-foot” sand. Four fault hydraulic characteristic values, two for each fault, will be estimated.

Combinations of the propositions from three sources of uncertainty result in 18 groundwater models as the base models in the HBMA. Each groundwater model will be calibrated by estimating homogeneous model parameters, which are hydraulic conductivity [m/d], specific storage [1/m], and fault hydraulic characteristics [1/d] in the following section.

5.3 Model Calibration and Posterior Model Probabilities

To calibrate the 18 groundwater models and estimate model parameters, we used the covariance matrix adaptation-evolution strategy (CMA-ES) [62,63] to minimize the root mean square error (RMSE) between calculated and observed groundwater heads at the USGS observation wells from 1975 to 2010:

$$RMSE_p = \sqrt{\frac{1}{N} \sum_{i=1}^N (h_i^{cal}(\mathbf{p}_p) - h_i^{obs})^2}, \quad (19)$$

where $h_i^{cal}(\beta_p)$ is the calculated groundwater head i given model parameters β for model p , h_i^{obs} is the observed groundwater head i at a USGS observation well, and N is the total number of observed heads. To demonstrate the methodology we only used $N=539$ groundwater head values from seven USGS wells, EB-327, EB-413, EB-652, EB-782B, EB-807A, EB-996 and WBR-102A shown in Figure 1.

CMA-ES is a derivative-free optimization method and is capable of avoiding entrapment in a local optimum. In groundwater modeling, CMA-ES has been successfully applied to estimate model parameters [57,62,64,65] and optimize groundwater management problems [66]. Moreover, CMA-ES provides a full covariance matrix of the estimated parameters, which can later be used to assess head prediction uncertainty.

Another advantage of using CMA-ES is the ease with which it can be implemented in a core-based high performance computing system with little overhead. This study has successfully run CMA-ES to calibrate the 18 groundwater models in SuperMike-II, a supercomputer housed at Louisiana State University. SuperMike-II is a 146 TFlops Peak Performance 440 compute node cluster. Each node contains two 8-Core Sandy Bridge Xeon 64-bit processors operating at a core frequency of 2.6 GHz.

The results of the model calibration and posterior model probabilities are shown in Table 1. The RMSE of the 18 groundwater models ranges from 1.06 m to 1.71 m. The mean RMSE is 1.18 m and the standard deviation of the RMSE is 0.15 m. The good fittings to groundwater head observation data indicate the adequacy of considering homogenous model parameters. Using the complex geological architectures, we do not need to consider more complex model parameter fields for model calibration.

Using the estimated parameters and the full covariance matrix obtained from the CMA-ES for each model, realizations of the parameters are generated by Monte Carlo simulation and input into groundwater models to generate groundwater heads. Then, the variances of groundwater heads are calculated using equations (13) and (14). We choose a large window size based on 5% significance level and $4\sigma_D$. According to the formula in Tsai and Li [25], the scaling factor is $\alpha = 1.06/\sqrt{539} = 0.046$. Finally, the $Q_{p,k}$ values, the $BIC_{p,k}$ values, and the posterior model probabilities of the 18 groundwater models were obtained using equations (12), (11) and (10), respectively, and are listed in Table 1. Nine models have posterior model probabilities larger than 5% and have an influence on the BMA mean predictions and variances. The top three groundwater models (GP-B_{1.5}-HC2, GP-B_{1.5}-HC4 and GP-B₀-HC2) are deemed equally important, as indicated by comparable posterior model probabilities of 16.8%, 17.5% and 15.6%, respectively.

Figure 3 shows the BMA tree of posterior model probabilities and conditional posterior model probabilities for the three uncertain model components. The geological architecture obtained by the GP has the posterior model probability 62.0%, which implies a better geological architecture proposition than the two other alternatives. The B_{1.5} proposition has a posterior model probability of 49.9% which is better than the two other boundary propositions. The HC2 fault permeability architecture has the posterior model probability 55.6%, which is better than the HC4 fault permeability architecture. Nevertheless, none of propositions predominates in each uncertain model component.

Having so many influential groundwater models indicates that the groundwater head data used in the model calibration may not be sufficient. In what follows, the proposed methodology is implemented on the USGS observation wells to see if a model with a predominate posterior

model probability can be achieved.

6. Experimental Design Results

We apply the TSS approach to the 20 USGS observation wells shown in Figure 1 to see if an 80%-identifiable groundwater model can be determined in minimal time. The BMA mean predicted heads from 2011 to 2015 are used as future observation data at the USGS observation wells. WBR-102A was not considered in the design since the location is too close to the west boundary. Although observation locations can be different for different rounds of data collection, the experimental design problem may become intractable and impractical. In order to reduce this possibility, we restricted the design as follows: (1) observation wells are selected at the beginning of the observation, which is the first day of 2011; and (2) observation wells once selected, remain active until either the probability threshold or the maximum observation period is reached.

The BMA is applied to obtain the averaged head values of the eighteen groundwater models, which will be used as “future observation” in the experimental design. We first execute the prediction models using the optimal estimated parameters to predict groundwater heads at 19 USGS observation wells. The assumptions for prediction models are: (1) the groundwater heads at four boundaries in the next 5 years (2011 to 2015) have the same trend as the groundwater heads over the last three years in the model calibration period (2008 to 2010); and (2) the last 5-year (2006-2010) monthly pumping rates are used for 2011-2015. Then, we use equation (5) to calculate the groundwater heads at selected USGS observation wells of a given experimental design. The posterior model probability for each groundwater model is recalculated using equation (10). It is noted that the purpose of our experimental design is to use the existing data

and BMA mean predictions to determine the best observation locations and required sampling rounds such that the discrimination criterion is met before actual data are collected. Therefore, the model parameters will remain the same in the design period. Models should be re-calibrated once real future observation data become available, which is beyond the scope of this study.

There are 524,287 possible experimental design solutions given the 19 USGS observation wells. An optimization approach is not guaranteed to find the best design. Instead, this study developed a code to implement the “embarrassingly parallel” algorithm to the SuperMike-II supercomputer to enumerate all possible design solutions. The total possible design solutions were split equally to 320 processors to be evaluated. Evaluation of one design took 82 seconds. Therefore, the computation time required for all designs was reduced from 1.36 years (estimated runtime by using one processor) to 1.55 days (actual runtime by using 320 processors). Using parallel computing, the result from the enumeration approach guarantees the goal of reaching the optimal solution.

6.1 Evolving Posterior Model Probabilities by Adding Future Data

Figure 4(a) shows the evolution of the posterior model probabilities if the seven USGS observation wells used in the model calibration continue to be used to collect the BMA mean predictions from 2011 to 2015. The figure shows that the posterior model probability of the model GP-B-1.5-HC2 monotonically increases from 16.8% before the design to 45.3% at the end of the 60-month observation. Other models may have their posterior model probabilities increase or decrease over time. By the end of 2015, there are six groundwater models with posterior model probabilities larger than 1%. Three groundwater models have influential posterior model probabilities. The model GP-B_{1.5}-HC2 has 45.3% posterior model probability, which is not dominant. The models GP-B_{1.5}-HC4 and IK-B₀-HC2 have posterior model probabilities 33.6%

and 14.1%, respectively. Using the seven USGS observation wells used in the model calibration cannot meet the discrimination criterion in 5 years of monthly data collection.

To understand the influence of individual observation wells, Figure 4(b) shows the posterior model probabilities after 60-month head observation if only one USGS observation well is used. Using EB-413 or EB-782A for observation, the posterior model probability of the model GP-B_{1.5}-HC2 is more than 20% higher than that of model GP-B_{1.5}-HC4. BMA mean predicted heads at EB-780A and EB-789B show strong disagreement with the model GP-B_{1.5}-HC2. The posterior model probability difference between the best and second best models is less than 2% when EB-291, EB-301, EB-327, EB-652, EB-946, EB-996 or WBR-101 is used for observation. Using one USGS observation well cannot meet the discrimination criterion in 5 years of monthly data collection.

Given that the number of observation wells is fixed, the enumeration method selects the best USGS observation wells that maximize the maximum posterior model probability across the 18 groundwater models. Figure 4(c) presents the result of posterior model probabilities after 60-month groundwater head observation. All design solutions support model GP-B_{1.5}-HC2 to be the best groundwater model with posterior model probability of 35.5%. Using only one USGS observation well, the best observation well location is at EB-782A, which is located close to the Baton Rouge fault. The second best model (IK-B₀-HC4) has posterior model probability of 11.8%. When using two observation wells, EB-168 and EB-782A are the best well locations that increase the highest posterior model probability to 50.3%. The second highest posterior model probability decreases to 7.4%. Figure 4(c) shows that at least six observation wells are needed in order to achieve the 80%-identifiable groundwater model in five years of groundwater head observation. These six USGS observation wells are EB-146, EB-168, EB-392, EB-413, EB-

782A and WBR-101. The highest posterior model probability is 82% by using the best 10 USGS observation wells (EB-146, EB-168, EB-301, EB-392, EB-413, EB-782A, EB-946, EB-996, WBR-5 and WBR-101). Using more than 10 USGS observation wells will decrease the maximum posterior model probability because the experimental design is restricted to the USGS observation wells and adding more USGS observation wells increases posterior model probabilities of other models. The maximum posterior model probability of the best model is lower than 80% when 16 or more USGS observation wells are used. Using 18 or 19 USGS observation wells significantly reduces the maximum posterior model probability of the best model because the two wells EB-780A and EB-789B that strongly disagree with the best model GP-B_{1.5}-HC2 are added. In other words, using less than 6 USGS observation wells or more than 15 USGS observation wells needs more than five years of monthly observation or never meets the discrimination criterion.

6.2 Non-Unique Experimental Design Solutions

The least time to reach the discrimination criterion 80% of posterior model probability given the best design solutions under different number of observation wells is shown in Figure 5. Ten experimental designs satisfy the discrimination criterion within 60 months and are therefore recommended. The fewest USGS observation wells needed to meet the discrimination criterion is six. The shortest time to meet the discrimination criterion is 48 months by using 13 USGS observation wells. When using more than 13 USGS observation wells, the least observation time increases. Again, this is because the experimental design is restricted to the USGS observation wells and adding more than 13 USGS observation wells increases posterior model probabilities of other models. Therefore, it requires longer time to achieve the discrimination criterion. When using more than 15 USGS observation wells, the discrimination criterion cannot be satisfied in

60 months.

The frequency of the USGS wells selected for the 10 experimental design solutions that meet the discrimination criterion is shown in Figure 6. The most popular wells are EB-168, WBR-101, EB-146, EB-392, EB-413, and EB-782A, which are selected 10 times, following by EB-946 with 9 times selected. The least selected wells are EB-291 and EB-918, which have only 1-2 times. EB-782B, EB-917 and EB-780A were not selected for any design.

6.3 Using Least Number of USGS Wells (Six Wells)

The temporal change of posterior model probabilities using the best six USGS wells is shown in Figure 7. The posterior model probability of the best model GP-B_{1.5}-HC4 before the design is 17.5% and decreases to 1.8% at the end of the 60-month observation. The posterior model probability of the second best model GP-B_{1.5}-HC2 before the design is 16.8% and monotonically increases to 80.0% at the end of the 60-month observation. Other than model GP-B_{1.5}-HC2, the posterior model probabilities of individual models may increase or decrease over time. The model GP-B_{1.5}-HC2 becomes the best model and meets the discrimination criterion.

The BMA tree of the updated posterior model probabilities and conditional posterior model probabilities is shown in Figure 8, using the best six USGS wells, where the BMA mean predicted heads from 2011-2015 are used as future data. The 80%-identifiable propositions for all levels are achieved. The result implies that the geological architecture obtained by the GP, which has conditional posterior model probability 84.8%, should be kept, and therefore the two other geological architectures may be discarded. Under the GP geological architecture proposition, the B_{1.5} proposition has conditional posterior model probability 89.4%. The two other boundary propositions may be discarded. Under the GP geological architecture proposition and the B_{1.5} proposition, the HC2 fault permeability architecture has conditional posterior model

probability 90.8%. Therefore, the HC4 fault permeability architecture can be discarded. As discussed previously, the posterior model probability of the best groundwater model is 80% shown at the base level, which is much greater than 1.8% of the second best groundwater model. Figure 9 shows the variance reduction of predicted groundwater heads at the 19 USGS wells at the end of the experimental design period (12/31/2015) using the 6-well experimental design for the best model (GP-B_{1.5}-HC2), GP-B_{1.5} model, the GP model and the hierarch model. Both between-model variances and within-model variances are significantly reduced by discriminating unimportant propositions through the experimental design.

It is noted that the conceptual model uncertainty is reduced as presented in Figure 9 due to the reduction in the posterior probabilities of unimportant model propositions. However, whether predictive performance is improved will depend on the future information (e.g., pumping data) in 2011-2015 that constructs the groundwater models, which cannot be assessed during the experimental design. Once future data are collected after implementing the design recommendations, the 18 groundwater models should be re-calibrated and their posterior model probabilities should be re-assessed.

6.4 Discussions

The fundamental idea of the experimental design is to collect data in order to reduce conceptual model uncertainty. We know that observation data are not commonly available on many observation times. For example, in our study area, most of the USGS wells have less than four head observation data in a year because it is expensive to collect the head data for deep aquifers like ours. There are only a few wells being sampled frequently. Minimizing the number of observation time is one logical way to constraint the budget and to achieve the objective of the experimental design within a reasonable time. We understand that many factors will affect the

design results. While we believe that our methodology is valid, real future data should be collected according to the recommendations in order to carefully assess the design results. New real data should be used to re-calibrate the models and re-do the experimental design. However, this post-analysis is beyond the scope of this study because (1) USGS do not have monthly head data for the selected observation wells. For most of the wells, the USGS only measure groundwater level one or two times per year and (2) we do not have actual pumping data for 2011-2014.

Head data (as well as other types of data, e.g. fluxes) at different locations can have varying impact in discriminating models. We only consider groundwater head data in the illustrative example since groundwater head is the most achievable/economical data in the study area to be used to discriminate among groundwater models. In addition, the new observations are being treated as uncorrelated. In fact, there are some spatial correlation among the different locations. We are aware of the issue of correlated head observations. However, the full covariance matrix of head observations is very tedious to obtain in the experimental design. In this study, we assume that data collected from different wells and from different sampling rounds are independent.

The predictive uncertainty in the BMA mean prediction will definitely affect the posterior model probability calculations and experimental design results. The future data Δ_D^{new} in equation (2) are not deterministic and have a PDF. The posterior model probability for each model should have a PDF. In this study, to make the problem more tractable, we take the deterministic approach and consider future data Δ_D^{new} to be determined by the BMA mean prediction. We do not investigate this type of uncertainty in this work and do not know the impact on the design

548 results. We think it will need a much extensive computational resource using the Monte Carlo
549 simulation approach than what have been incurred in this study. We hope to do this in the future
550 work.

551 The improvement of predictive performance depends on future information in the
552 simulation models. For example, we did not know and guessed the pumping data and boundary
553 head values for 2011-2015. Therefore, the individual model predictions and the BMA mean
554 prediction for 2011- 2015 may not be close to the future real data because of incorrect future
555 information that constructs the forecast models. However, this issue is unrelated to the proposed
556 methodology. We think that adding new data will in some extent reduce uncertainty since the
557 new data always bring some additional information about the system. However, what types of
558 uncertainty and how effective the uncertainty can be reduced are dependent on what types of
559 data to collect and how you use the data, which is beyond the scope of this work. Our study is to
560 collect spatiotemporal groundwater head data to reduce conceptual model uncertainty

561 The hierarchical structure is subjected to the analyst's preference and prioritization of the
562 sources of uncertainty. If one is more interested in one specific source of uncertainty than others,
563 that specific source of uncertainty should be placed at the first level. However, the optimal
564 design will be the same regardless the order of the sources of uncertainty because we maximize
565 the posterior probability across all the simulation models (base models), not the BMA models at
566 the first level.

567 The results showed that the best model based on the existing data is not necessary the
568 best model after the experimental design using the BMA prediction data. The future data
569 collected from different locations and sampling rounds will decide which model has highest
570 potential to be the best model based on their posterior model probabilities. There is no guarantee

which model will be the best model based on existing data without the experimental design.

7. Conclusions

Using posterior model probability as a discrimination criterion is a straightforward approach to identify a unique and highly probable model from a pool of candidate models. Since representing model importance based on data evidence, the posterior model probability is a direct quantitative indicator for the experimental design to guide data collection through an observation network. The time-sequential sampling approach presents a practical strategy for data collection, by which future data are gradually collected and used along with historical data to achieve the unique and highly probable model.

The HBMA method is found to be useful to illustrate the results of the experimental design for proposition discrimination and prediction uncertainty reduction. First, using the BMA mean prediction as future observation data to guide the time sequential sampling data collection strategy is an unbiased approach since all candidate models contribute to prediction according to their up-to-date posterior model probabilities. Second, once the design goal of achieving the γ - identifiable model is met, the unique and highly probable proposition for each uncertain model component is identified through the BMA tree of posterior model probabilities. By reducing posterior model probabilities of unimportant propositions, the prediction variance owing to different sources of uncertainty is significantly reduced.

The experimental design illustrated by the real groundwater case study is practical since the study only uses the USGS observation wells. Implementing the enumeration approach to a multi-core supercomputer is viable and assures finding the best number and location of the USGS wells for the design. The study finds several experimental designs that can achieve the

80%-identifiable groundwater model for modeling the “1,200-foot” sand, the “1,500-foot” sand, and the “1,700-foot” sand of the Baton Rouge area, Louisiana. There is a tradeoff between the number of the USGS wells sampled and the number of sampling rounds. Through the HBMA framework, the highly probable propositions are identified for the geological architecture, the head boundary values south of the Baton Rouge fault, and the fault permeability. The HBMA shows significant reduction in head prediction variance at all USGS observation wells in the study area.

Using BMA mean prediction to re-calibrate simulation models in the time-sequential sampling approach is not suggested since the BMA predicted quantities are not real data. In addition, the posterior model probabilities calculated using BMA mean predictions only present a potential for a model to be the best model. Which model actually wins is only visible upon real data collection and model re-calibration. The experimental design might be repeated after some time when real future data are collected, the models are re-calibrated, and the posterior model probabilities are updated.

Acknowledgements: The work was partially supported by the LSU Economic Development Assistantship and by the National Science Foundation under Grant No. 1045064. The authors acknowledge the LSU High Performance Computing for providing computing resources and technical assistance. The authors also acknowledge constructive comments from Wolfgang Nowak and six anonymous reviewers.

References

[1] Sun N-Z. Inverse Problems in Groundwater Modeling. 1994th ed. Springer; 1994.

- 617 [2] Carrera J, Alcolea A, Medina A, Hidalgo J, Slooten LJ. Inverse problem in hydrogeology.
618 Hydrogeol J 2005;13:206–22. doi:10.1007/s10040-004-0404-7.
- 619 [3] Neuman SP. Maximum likelihood Bayesian averaging of uncertain model predictions.
620 Stoch Environ Res Risk Assess 2003;17:291–305. doi:10.1007/s00477-003-0151-7.
- 621 [4] Poeter E, Anderson D. Multimodel ranking and inference in ground water modeling.
622 Ground Water 2005;43:597–605. doi:10.1111/j.1745-6584.2005.0061.x.
- 623 [5] Refsgaard JC, van der Sluijs JP, Brown J, van der Keur P. A framework for dealing with
624 uncertainty due to model structure error. Adv Water Resour 2006;29:1586–97.
625 doi:10.1016/j.advwatres.2005.11.013.
- 626 [6] Beven K, Freer J. Equifinality, data assimilation, and uncertainty estimation in mechanistic
627 modelling of complex environmental systems using the GLUE methodology. J Hydrol
628 2001;249:11–29. doi:10.1016/S0022-1694(01)00421-8.
- 629 [7] Neuman SP, Wierenga PJ. A Comprehensive Strategy of Hydrogeologic Modeling and
630 Uncertainty Analysis for Nuclear Facilities and Sites. Washington, DC: Division of
631 Systems Analysis and Regulatory Effectiveness, Office of Nuclear Regulatory Research,
632 U.S. Nuclear Regulatory Commission; 2003.
- 633 [8] Trolldborg L, Refsgaard JC, Jensen KH, Engesgaard P. The importance of alternative
634 conceptual models for simulation of concentrations in a multi-aquifer system. Hydrogeol J
635 2007;15:843–60. doi:10.1007/s10040-007-0192-y.
- 636 [9] Rojas R, Feyen L, Dassargues A. Conceptual model uncertainty in groundwater modeling:
637 Combining generalized likelihood uncertainty estimation and Bayesian model averaging.
638 Water Resour Res 2008;44:W12418. doi:10.1029/2008WR006908.
- 639 [10] Ye M, Pohlmann KF, Chapman JB, Pohl GM, Reeves DM. A model-averaging method
640 for assessing groundwater conceptual model uncertainty. Ground Water 2010;48:716–28.
641 doi:10.1111/j.1745-6584.2009.00633.x.
- 642 [11] Rojas R, Feyen L, Batelaan O, Dassargues A. On the value of conditioning data to reduce
643 conceptual model uncertainty in groundwater modeling. Water Resour Res
644 2010;46:W08520. doi:10.1029/2009WR008822.
- 645 [12] Singh A, Mishra S, Ruskauff G. Model averaging techniques for quantifying conceptual
646 model uncertainty. Ground Water 2010;48:701–15. doi:10.1111/j.1745-6584.2009.00642.x.
- 647 [13] Hsu N-S, Yeh WW-G. Optimum experimental design for parameter identification in
648 groundwater hydrology. Water Resour Res 1989;25:1025–40.
649 doi:10.1029/WR025i005p01025.

- 650 [14] Gupta HV, Clark MP, Vrugt JA, Abramowitz G, Ye M. Towards a comprehensive
651 assessment of model structural adequacy. *Water Resour Res* 2012;48.
652 doi:10.1029/2011WR011044.
- 653 [15] Bredehoeft JD. From models to performance assessment: The conceptualization problem.
654 *Ground Water* 2003;41:571–7. doi:10.1111/j.1745-6584.2003.tb02395.x.
- 655 [16] Bredehoeft J. The conceptualization model problem-surprise. *Hydrogeol J* 2005;13:37–46.
656 doi:10.1007/s10040-004-0430-5.
- 657 [17] Neuman SP, Xue L, Ye M, Lu D. Bayesian analysis of data-worth considering model and
658 parameter uncertainties. *Adv Water Resour* 2012;36:75–85.
659 doi:10.1016/j.advwatres.2011.02.007.
- 660 [18] Draper D. Assessment and propagation of model uncertainty. *J R Stat Soc Ser B Methodol*
661 1995;45–97.
- 662 [19] Kass RE, Raftery AE. Bayes Factors. *J Am Stat Assoc* 1995;90:773–95.
663 doi:10.1080/01621459.1995.10476572.
- 664 [20] Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression
665 models. *J Am Stat Assoc* 1997;92:179–91. doi:10.1080/01621459.1997.10473615.
- 666 [21] Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial.
667 *Stat Sci* 1999:382–401.
- 668 [22] Raftery AE, Zheng Y. Discussion: Performance of Bayesian model averaging. *J Am Stat*
669 *Assoc* 2003;98:931–8. doi:10.1198/016214503000000891.
- 670 [23] Ye M, Neuman SP, Meyer PD. Maximum likelihood Bayesian averaging of spatial
671 variability models in unsaturated fractured tuff. *Water Resour Res* 2004;40:W05113.
672 doi:10.1029/2003WR002557.
- 673 [24] Ajami NK, Duan Q, Gao X, Sorooshian S. Multimodel combination techniques for
674 analysis of hydrological simulations: Application to distributed model intercomparison
675 project results. *J Hydrometeorol* 2006;7:755–68. doi:10.1175/JHM519.1.
- 676 [25] Tsai FT-C, Li X. Inverse groundwater modeling for hydraulic conductivity estimation
677 using Bayesian model averaging and variance window. *Water Resour Res*
678 2008;44:W09434. doi:200810.1029/2007WR006576.
- 679 [26] Li X, Tsai FT-C. Bayesian model averaging for groundwater head prediction and
680 uncertainty analysis using multimodel and multimethod. *Water Resour Res*
681 2009;45:W09403. doi:200910.1029/2008WR007488.

- 682 [27] Cleveland T, Yeh W. Sampling network design for transport parameter identification. J
683 Water Resour Plan Manag 1990;116:764–83. doi:10.1061/(ASCE)0733-
684 9496(1990)116:6(764).
- 685 [28] Sun N-Z, Yeh WW-G. Coupled inverse problems in groundwater modeling: 2.
686 Identifiability and experimental design. Water Resour Res 1990;26:2527–40.
687 doi:10.1029/WR026i010p02527.
- 688 [29] Sciortino A, Harmon TC, Yeh WW-G. Experimental design and model parameter
689 estimation for locating a dissolving dense nonaqueous phase liquid pool in groundwater.
690 Water Resour Res 2002;38:15–1 – 15–9. doi:10.1029/2000WR000134.
- 691 [30] Chang L-F, Sun N-Z, Yeh WW-G. Optimal observation network design for parameter
692 structure identification in groundwater modeling. Water Resour Res 2005;41:W03002.
693 doi:10.1029/2004WR003514.
- 694 [31] McPhee J, Yeh WW-G. Experimental design for groundwater modeling and management.
695 Water Resour Res 2006;42:W02408. doi:10.1029/2005WR003997.
- 696 [32] Altmann-Dieses AE, Schlöder JP, Bock HG, Richter O. Optimal experimental design for
697 parameter estimation in column outflow experiments. Water Resour Res 2002;38:1186.
698 doi:10.1029/2001WR000358.
- 699 [33] McKinney DC, Loucks DP. Network design for predicting groundwater contamination.
700 Water Resour Res 1992;28:133–47. doi:10.1029/91WR02397.
- 701 [34] Wagner BJ. Sampling design methods for groundwater modeling under uncertainty. Water
702 Resour Res 1995;31:2581–91. doi:10.1029/95WR02107.
- 703 [35] Janssen GMCM, Valstar JR, Zee SEATM van der. Measurement network design including
704 traveltime determinations to minimize model prediction uncertainty. Water Resour Res
705 2008;44:W02405. doi:10.1029/2006WR005462.
- 706 [36] Nowak W, de Barros FPJ, Rubin Y. Bayesian geostatistical design: Task-driven optimal
707 site investigation when the geostatistical model is uncertain. Water Resour Res
708 2010;46:W03535. doi:10.1029/2009WR008312.
- 709 [37] Nowak W, Rubin Y, de Barros FPJ. A hypothesis-driven approach to optimize field
710 campaigns. Water Resour Res 2012;48:W06509. doi:10.1029/2011WR011016.
- 711 [38] Usunoff E, Carrera J, Mousavi SF. An approach to the design of experiments for
712 discriminating among alternative conceptual models. Adv Water Resour 1992;15:199–214.
713 doi:10.1016/0309-1708(92)90024-V.

- 714 [39] Knopman DS, Voss CI. Discrimination among one-dimensional models of solute transport
715 in porous media: Implications for sampling design. *Water Resour Res* 1988;24:1859–76.
716 doi:10.1029/WR024i011p01859.
- 717 [40] Knopman DS, Voss CI. Multiobjective sampling design for parameter estimation and
718 model discrimination in groundwater solute transport. *Water Resour Res* 1989;25:2245–58.
719 doi:10.1029/WR025i010p02245.
- 720 [41] Knopman DS, Voss CI, Garabedian SP. Sampling design for groundwater solute transport:
721 Tests of methods and analysis of Cape Cod tracer test data. *Water Resour Res*
722 1991;27:925–49. doi:10.1029/90WR02657.
- 723 [42] Yakirevich A, Pachepsky YA, Gish TJ, Guber AK, Kuznetsov MY, Cady RE, et al.
724 Augmentation of groundwater monitoring networks using information theory and ensemble
725 modeling with pedotransfer functions. *J Hydrol* 2013;501:13–24.
726 doi:10.1016/j.jhydrol.2013.07.032.
- 727 [43] Box GEP, Hill WJ. Discrimination among mechanistic models. *Technometrics* 1967;9:57–
728 71. doi:10.1080/00401706.1967.10490441.
- 729 [44] Hunter WG, Reiner AM. Designs for discriminating between two rival models.
730 *Technometrics* 1965;7:307–23. doi:10.2307/1266591.
- 731 [45] Atkinson AC, Fedorov VV. Optimal design : Experiments for discriminating between
732 several models. *Biometrika* 1975;62:289–303. doi:10.1093/biomet/62.2.289.
- 733 [46] Hill PDH. A review of experimental design procedures for regression model
734 discrimination. *Technometrics* 1978;20:15–21. doi:10.2307/1268155.
- 735 [47] Guido Buzzi-Ferraris PF. An improved version of a sequential design criterion for
736 discriminating among rival multiresponse models. *Chem Eng Sci* 1990;45:477–81.
737 doi:10.1016/0009-2509(90)87034-P.
- 738 [48] Leon ACP de, Atkinson AC. Optimum experimental design for discriminating between
739 two rival models in the presence of prior information. *Biometrika* 1991;78:601–8.
740 doi:10.2307/2337029.
- 741 [49] Schwaab M, Luiz Monteiro J, Carlos Pinto J. Sequential experimental design for model
742 discrimination: Taking into account the posterior covariance matrix of differences between
743 model predictions. *Chem Eng Sci* 2008;63:2408–19. doi:10.1016/j.ces.2008.01.032.
- 744 [50] Alberton AL, Schwaab M, Lobão MWN, Pinto JC. Design of experiments for
745 discrimination of rival models based on the expected number of eliminated models. *Chem*
746 *Eng Sci* 2012;75:120–31. doi:10.1016/j.ces.2012.03.010.

- 747 [51] Donckels BMR, De Pauw DJW, Vanrolleghem PA, De Baets B. Performance assessment
748 of the anticipatory approach to optimal experimental design for model discrimination.
749 Chemom Intell Lab Syst 2012;110:20–31. doi:10.1016/j.chemolab.2011.06.008.
- 750 [52] Michalik C, Stuckert M, Marquardt W. Optimal experimental design for discriminating
751 numerous model candidates: The AWDC criterion. Ind Eng Chem Res 2010;49:913–9.
752 doi:10.1021/ie900903u.
- 753 [53] Pukelsheim F, Rosenberger JL. Experimental designs for model discrimination. J Am Stat
754 Assoc 1993;88:642–9. doi:10.2307/2290346.
- 755 [54] Atkinson AC. A comparison of two criteria for the design of experiments for
756 discriminating between models. Technometrics 1981;23:301–5. doi:10.2307/1267795.
- 757 [55] Nordqvist R, Voss CI. A simulation-based approach for designing effective field-sampling
758 programs to evaluate contamination risk of groundwater supplies. Hydrogeol J 1996;4:23–
759 39. doi:10.1007/s100400050081.
- 760 [56] Raftery AE. Bayesian model selection in social research. Sociol Methodol 1995;25:111–
761 64.
- 762 [57] Tsai FT-C, Elshall AS. Hierarchical Bayesian model averaging for hydrostratigraphic
763 modeling: Uncertainty segregation and comparative evaluation. Water Resour Res
764 2013;49:5520–36. doi:10.1002/wrcr.20428.
- 765 [58] Meyer RR, Turcan AN. Geology and Ground-Water Resources of the Baton Rouge Area,
766 Louisiana. US Government Printing Office; 1955.
- 767 [59] Elshall AS, Tsai FT-C, Hanor JS. Indicator geostatistics for reconstructing Baton Rouge
768 aquifer-fault hydrostratigraphy, Louisiana, USA. Hydrogeol J 2013;21:1731–47.
769 doi:10.1007/s10040-013-1037-5.
- 770 [60] Harbaugh AW. MODFLOW-2005, The U.S. Geological Survey Modular Ground-Water
771 Model - the Ground-Water Flow Process. Book 6 Model. Tech. Sect. Ground Water,
772 Virginia: US Dept. of the Interior, US Geological Survey; 2005.
- 773 [61] Hsieh PA, Freckleton JR. Documentation of a Computer Program to Simulate Horizontal-
774 Flow Barriers Using the U.S. Geological Survey's Modular Three-Dimensional Finite-
775 Difference Ground-Water Flow Model. United States Geological Survey; 1993.
- 776 [62] Hansen N, Ostermeier A. Completely derandomized self-adaptation in evolution strategies.
777 Evol Comput 2001;9:159–95.

- [63] Hansen N, Müller SD, Koumoutsakos P. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evol Comput* 2003;11:1–18.
- [64] Keating EH, Doherty J, Vrugt JA, Kang Q. Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality. *Water Resour Res* 2010;46:W10517. doi:10.1029/2009WR008584.
- [65] Elshall AS, Tsai FT-C. Constructive epistemic modeling of groundwater flow with geological structure and boundary condition uncertainty under the Bayesian paradigm. *J Hydrol* 2014;517:105–19. doi:10.1016/j.jhydrol.2014.05.027.
- [66] Bayer P, de Paly M, Bürger CM. Optimization of high-reliability-based hydrological design problems by robust automatic sampling of critical model realizations. *Water Resour Res* 2010;46:W05504. doi:10.1029/2009WR008081.

Table 1. RMSE, weighted sum of squared errors (Q), BIC values, and posterior model probabilities for the 18 groundwater models using the variance window with 5% significance level and $4\sigma_D$.

Model	RMSE (m)	Q	BIC	Δ BIC	Posterior Model Probabilities (%)
GP-B _{-1.5} -HC2	1.10	330.7	1359	2	16.8
GP-B _{-1.5} -HC4	1.19	316.3	1357	0	17.5
GP-B ₀ -HC2	1.11	334.1	1362	5	15.6
GP-B ₀ -HC4	1.34	348.9	1390	33	8.3
GP-B _{+1.5} -HC2	1.07	395.1	1423	66	3.9
GP-B _{+1.5} -HC4	1.28	854.9	1896	539	0.0
IK-B _{-1.5} -HC2	1.06	494.9	1523	166	0.4
IK-B _{-1.5} -HC4	1.11	341.6	1383	25	9.8
IK-B ₀ -HC2	1.71	367.3	1396	38	7.3
IK-B ₀ -HC4	1.14	435.5	1476	119	1.2
IK-B _{+1.5} -HC2	1.10	445.2	1474	116	1.2
IK-B _{+1.5} -HC4	1.19	366.8	1408	50	5.5
IZ-B _{-1.5} -HC2	1.06	382.5	1411	54	5.1
IZ-B _{-1.5} -HC4	1.17	504.9	1546	189	0.2
IZ-B ₀ -HC2	1.09	378.2	1407	49	5.7
IZ-B ₀ -HC4	1.12	479.0	1520	163	0.4
IZ-B _{+1.5} -HC2	1.23	511.0	1539	182	0.3
IZ-B _{+1.5} -HC4	1.22	451.6	1492	135	0.8

Figure Captions

Figure 1. Map of the study area in the Universal Transverse Mercator (UTM) coordinate system, which includes location of pumping wells, USGS observation wells, Baton Rouge Fault and Denham Springs -Scotlandville Fault.

Figure 2. Sand distribution at layer 36 using (a) IZ, (b) GP and (c) IK methods, showing uncertainty in the geological architecture.

Figure 3. The BMA tree of posterior model probabilities and conditional posterior model probabilities before observation network design. Posterior model probabilities and conditional posterior model probabilities less than 1% are not shown.

Figure 4. (a) Changes of posterior probabilities of 18 groundwater models by monthly groundwater head observation using the seven USGS observation wells used in the model calibration. Posterior probabilities of 18 groundwater models after 60-month head observation (b) using one USGS observation well at different locations and (c) using different number of best USGS observation wells.

Figure 5. The fewest months to reach the discrimination criterion of 80% posterior model probability given a number of best selected USGS observation wells.

Figure 6. The frequency of the USGS observation wells to be selected for the best observation network designs given the criterion of 80% posterior model probability.

Figure 7. Changes of posterior model probabilities of 18 groundwater models over time based on the best six USGS observation wells, EB-146, EB-413, EB-652, EB-789B, EB-946, and EB-996.

Figure 8. The BMA tree of posterior model probabilities and conditional posterior model probabilities after the observation network design with the best six USGS observation wells. Posterior model probabilities and conditional posterior model probabilities less than 1% are not shown.

Figure 9. Comparisons of the head variances given by the best model GP-B_{1.5}-HC2 and the BMA models (a) GP-B-1.5 model, (b) GP model and (c) hierarch model for 19 USGS observation wells on 12/31/2015. The experimental design uses the best six USGS observation wells.

Figure 1

Figure 2

(a) Indicator zonation

(b) Generalized parameterization

(c) Indicator kriging

Figure 3

Figure 4

Figure 5

Figure 6

Figure 7

Figure 9