Sensitivity analysis of prior model probabilities and the value of prior knowledge in the assessment of conceptual model uncertainty in groundwater modelling

Rodrigo Rojas[a*], Luc Feyen[b], Alain Dassargues[a,c]

[*] Corresponding author

[a] Applied geology and mineralogy, Department of Earth and Environmental Sciences,
Katholieke Universiteit Leuven
Celestijnenlaan 200 E, B-3001 Heverlee, Belgium
Tel.: +32 016 326449; fax: +32 016 326401.
E-mail address: Rodrigo.RojasMujica@geo.kuleuven.be

[b] European Commission - DG Joint Research Centre (JRC), Institute for Environment and Sustainability, Land management and natural hazards unit
TP261, Via E. Fermi 2749, 21027 Ispra (Va), Italy
Tel.: +39 0332 789258; fax: +39 0332 786653
E-mail address: luc.feyen@jrc.it

[c] Hydrogeology and environmental geology, Department of Architecture, Geology, Environment, and Constructions (ArGEnCo), Université de Liège
B.52/3 Sart-Tilman, B-4000 Liège, Belgium
Tel.: +32 4 3662376; fax: +32 4 3669520
E-mail address: alain.dassargues@geo.kuleuven.be

Abstract

A key point in the application of multi-model Bayesian averaging techniques to assess the predictive uncertainty in groundwater modelling applications is the definition of prior model probabilities, which reflect the prior perception about the plausibility of alternative models. In this article we analyze the influence of prior knowledge and prior model probabilities on posterior model probabilities, multi-model predictions and conceptual model uncertainty estimations. The sensitivity to prior model probabilities is assessed using an extensive numerical analysis in which the prior probability space of a set of plausible conceptualizations is discretised to obtain a large ensemble of possible combinations of prior model probabilities. Additionally, we assess the value of prior knowledge about alternative models in reducing conceptual model uncertainty by considering three example knowledge states, expressed as quantitative relations among the alternative models. A constrained maximum entropy approach is used to find the set of prior model probabilities that correspond to the different prior knowledge states. For illustrative purposes, we employ a 3-dimensional hypothetical setup approximated by 7 alternative conceptual models. Results show that posterior model probabilities, leading moments of the predictive distributions and estimations of conceptual model uncertainty are very sensitive to prior model probabilities, indicating the relevance of selecting proper prior probabilities. Additionally, including proper prior knowledge improves the predictive performance of the multi-model approach, expressed by reductions of the multi-model prediction variances up to 60%. However, the ratio between-model to total variance does not substantially decrease. This suggests that the contribution of conceptual model uncertainty to the total variance can not be further reduced based only on prior knowledge about the plausibility of alternative models. These results advocate including proper prior knowledge about alternative conceptualizations in combination with extra conditioning data to further reduce conceptual model uncertainty in groundwater modelling predictions.

Keywords

Multi-model prediction, uncertainty assessment, maximum entropy, prior knowledge, conceptual model uncertainty

1. Introduction and scope

Groundwater modelling has become an essential part of groundwater management and accurate model predictions are required to ensure an acceptable degree of confidence in model results. However, incomplete knowledge about the geological setting and scarce or prone to error information about model parameters, boundary conditions and input data, render the predictions of groundwater dynamics and pollutant transport uncertain. Practice, on the other hand, suggests that once a conceptual model is successfully calibrated its results are rarely questioned and the conceptual model is assumed to be correct (Bredehoeft, 2005; Hojberg and Refsgaard, 2005). However, a successful calibration does not guarantee the correctness of the conceptual model. Rather, many parameter sets together with different conceptual models may produce equally good results in a calibration process (Bredehoeft, 2003; Harrar *et al*., 2003; Carrera *et al*., 2005). In this sense, relying on a single hydrological concept will likely produce biased and under-dispersive predictions due to neglecting conceptual model uncertainty (Neuman, 2003).

In recent years, a number of multi-model methods have been proposed to address the problem of conceptual model uncertainty in hydrological modelling (Neuman, 2003; Poeter and Anderson, 2005, Ajami *et al*., 2005; Refsgaard *et al*., 2006). These methods seek to obtain consensus predictions from a set of plausible models by linearly combining individual model predictions. One such approach is Bayesian Model Averaging (BMA) (Draper, 1995; Hoeting *et al*., 1999), which weights the predictions of competing models by their corresponding posterior model probability, representing each model's relative skill to reproduce system behaviour in the training period. Hence, BMA weights are tied directly to individual model performance. Several studies applying the method to a range of different problems have demonstrated that BMA produces more accurate and reliable predictions than other existing multi-model techniques (e.g., Raftery and Zheng, 2003; Ye *et al*., 2004; Ajami *et al*., 2005).

In the field of groundwater hydrology, applications of BMA have been rare. Neuman (2003) proposed the Maximum Likelihood Bayesian Model Averaging (MLBMA) method, which is an approximation of BMA that relies on maximum likelihood parameter estimation and expanding around these values through Monte Carlo simulation. Ye *et al*., (2004) expanded upon the theoretical framework of MLBMA and applied it to model the log permeability in unsaturated fractured tuff using alternative variogram models.

Rojas *et al*., (2008) proposed a methodology to assess uncertainty in predictions of groundwater models arising from errors in the model structure, forcing data and parameter estimates by integrating the Generalized Likelihood Uncertainty Estimation (GLUE) (Beven and Binley, 1992) methodology with BMA. The methodology is based on the concept that there exist many good simulators of the system that may be located in different regions of the combined model, input and parameter space, given the data at hand. For a set of plausible system conceptualizations, input and parameter realizations are sampled from the joint prior input and parameter space. A likelihood measure is then calculated for each simulator based on its ability to reproduce system state variable observations. The integrated likelihood of each conceptual model is obtained by integrating the likelihood of the different simulators over the input and parameter space. The integrated likelihoods are consequently used in BMA to weight the model predictions to obtain ensemble predictions. Key advantages of this methodology are that: (i) there is no restriction on the diversity of conceptual models or on the level of uncertainty in the input data or parameters that can be included; (ii) it does not rely on a single optimum set of (calibrated) parameter values, hence, avoiding biased parameter estimates that compensate for errors in model structure, input data and measurement errors; (iii) it allows for different ways of expressing the likelihood of a simulator (including a formal Bayesian one), hence allowing different types of knowledge to be incorporated (quantitative as well as qualitative); and (iv) it is Bayesian in nature, which provides a formal framework to incorporate prior knowledge about the model structures and parameters, or to update the estimates should new information become available.

Rojas *et al*., (2008) applied the methodology by considering 7 alternative conceptualizations with increasing complexity to represent a 3-dimensional synthetic example consisting of 2 aquifers separated by an aquitard. An extensive numerical analysis showed that neglecting conceptual model uncertainty results in biased and overly conservative predictions. However, two important aspects concerning the application of the methodology remained unanswered; first, the sensitivity of posterior model probabilities, multi-model groundwater predictions, and conceptual model uncertainty estimations to different sets of prior model probabilities; and, second, the value of prior knowledge about the alternative conceptualizations to further reduce conceptual model uncertainty. We address these two points in this article.

In Bayesian inference two basic interpretations can be given to prior probability distributions. First, in the *population* interpretation, a prior distribution represents a population of possible parameter values from which a potential candidate is to be drawn. Second, in the more subjective *state of knowledge* interpretation, the guiding principle is that knowledge (and uncertainty) about a given parameter must be expressed as if the value of that parameter could be thought of as a random realization from the prior probability distribution (Gelman *et al*., 2004, p. 39), i.e., prior probability distributions can be interpreted as a formal representation of knowledge (uncertainty) about a given parameter. More importantly, there is no unique prior probability distribution for representing this knowledge (uncertainty) (Kass and Wasserman, 1996).

In Bayesian literature, different methods to assign prior probability distributions to different classes of problems can be found. We do not wish to provide a complete overview of these methods but refer the reader to Kass and Wasserman (1996) for an excellent review.

A key point when adopting a prior probability distribution is the influence of this distribution, after updating, on the results. Two general courses of action can be mentioned to alleviate this influence. First, with increasing data availability, prior probability distributions are expected to have less influence on inferences about parameters and predicted variables (Kass and Wasserman, 1996). Thus, one strategy consists in collecting as much data as possible to overcome the influence of the prior probability distributions. For most groundwater modelling applications, however, obtaining enough data to overrule the effects of prior model probabilities may in many cases be cost prohibitive. Second, one can assign non-informative prior probability distributions, with the uniform distribution being the most common case, hoping that information contained in the data will dominate the form of the resulting posterior distribution. Consequently, reported multi-model methodologies used in groundwater modelling have employed, generally, a uniform prior model probability distribution reflecting no prior preference on the plausibility of alternative conceptual models (see, e.g., Meyer *et al*., 2007). This is also the approach followed by Rojas *et al*., (2008).

Panels of experts, prior elicitation, and theoretical or empirical grounds, on the other hand, can be helpful in defining suitable prior model probabilities based on expert knowledge (see, e.g., Ye *et al*., 2006). These prior model probabilities are inherently subjective, i.e., they reflect preference over a particular conceptualization and, probably, other group of experts will arrive to different prior model probabilities based on different grounds. In this context, we stick to the idea expressed by Ghosh *et al*., (2006, p. 55) who stated that whenever prior information is available, an attempt to use a prior probability distribution reflecting that prior knowledge should be used as far as possible.

Given that there is no unique way to express the prior knowledge about alternative conceptual models, due mainly to the subjective nature of the task, a procedure to select among potential sets of prior model probabilities is required. Ye *et al*., (2005) recently proposed an approach aimed to find a set of prior model probabilities that maximizes Shannon's entropy (Shannon, 1948) subject to a series of constraints. Hereby, the constraints reflect prior knowledge about the alternative conceptualizations. The key idea behind this approach is that uncertainty represents "potential information" in the sense that when a random variable takes on a value we gain information and lose uncertainty (Applebaum, 1996). In this sense, Shannon's entropy measures the amount of information contained in the set of prior model probabilities. Therefore, less informative sets will have a higher entropy compared with more informative sets since a larger amount of information can be gained in the first. For example, when the set of prior model probabilities corresponds to the uniform prior distribution, i.e., all alternative conceptual models have equal prior probabilities, we are at a state of maximum uncertainty and entropy is at its maximum. When a more informative set of prior model probabilities is available the entropy will be lower.

In the case that several sets of constraints reflecting different prior knowledge about the conceptual models are proposed, the problem translates into a min-max choice, i.e., to find the set of prior model probabilities that maximizes Shannon's entropy subject to the respective constraints, but which is minimum among different proposed sets. Solving this min-max problem, however, does not guarantee optimum predictive performance. To overcome this problem, Ye *et al*., (2005) propose to follow one of the following two approaches: (1) when enough data are available to perform a meaningful model (cross)validation, they advocate selecting a posteriori the set that outperforms based on suitable performance criteria (see, e.g., Liang *et al*., 2001 for examples on performance criteria); (2) when there is not enough data available to estimate meaningful posterior measures of model quality, they advocate selecting the min-max set that, additionally, maximizes the likelihood for the ensemble of alternative conceptual models.

In this work, we conduct a numerical experiment to analyze the sensitivity of the posterior model probabilities, the groundwater multi-model predictions, and the conceptual model uncertainty estimations to prior model probabilities. To this end, the prior probability space of the alternative conceptual models is discretised in equidistant intervals and all possible combinations of prior model probabilities for the set of conceptualizations are formed, given that the sum of the prior model probabilities for each combination equals 1.

Furthermore, we extend upon the work of Ye *et al*., (2005) and assess the value of prior knowledge about the plausibility of alternative conceptualizations in reducing conceptual model uncertainty. To this end we employ the constrained maximum entropy approach proposing (out of the ensemble of discrete sets of prior model probabilities) three sets of prior model probabilities that reflect the following knowledge states: (i) a non-informative case about the plausibility of alternative conceptualizations, i.e., alternative conceptual models have equal prior probabilities; (ii) relevant and proper prior knowledge about the plausibility of alternative conceptualizations, i.e., alternative conceptual models receive higher prior probabilities as they approach a “true” 3-dimensional hypothetical setup; and (iii) improper prior knowledge about the plausibility of alternative conceptualizations, i.e., alternative conceptual models receive prior probabilities that are inconsistent as they approach the “true” 3-dimensional hypothetical setup. Results obtained using the three optimized sets of prior model probabilities are compared to find the set that outperforms in terms of predictive capacity and to assess the value of this prior knowledge to further reduce conceptual model uncertainty.

The remainder of this paper is organized as follows. In section 2, we provide a condensed overview of the combined GLUE-BMA methodology. Section 3 details a 3-dimensional hypothetical aquifer system that is used to illustrate the methodology and to assess the sensitivity of the groundwater multi-model predictions. Implementation details are described in section 4. In this section, we elaborate on the different conceptual models, input and parameter uncertainty, the methodology to account for the sensitivity of the results due to different discrete sets of prior model probabilities and the constrained maximum entropy method to assess suitable sets of prior model probabilities in agreement with prior knowledge. Results are discussed in section 5 and a summary of conclusions is presented in section 6.

2. Materials and Methods

To render the article self-contained sections 2.1 and 2.2 elaborate on the basis of GLUE and BMA methodologies, respectively. For a detailed description the reader is referred to Rojas *et al*., (2008).

2.1. Generalized Likelihood Uncertainty Estimation (GLUE) methodology

GLUE is a Monte Carlo simulation technique based on the concept of equifinality (Beven and Freer, 2001). It rejects the idea of a single correct representation of a system in favour of many acceptable system representations (Beven, 2005). For each potential system simulator, sampled from a prior set of possible system representations, a likelihood measure is calculated which reflects its ability to simulate the system responses, given the available training data **D**. Simulators that perform below a subjectively defined rejection criterion are discarded from the further analysis and likelihood measures of retained simulators are rescaled so as to render the cumulative likelihood equal to 1. Ensemble predictions are based on the predictions of the retained set of simulators, weighted by their respective rescaled likelihood.

Likelihood measures used in GLUE must be seen in a much wider sense than the formal likelihood functions used in traditional statistical estimation theory (Binley and Beven, 2003). These likelihoods are a measure of the ability of a simulator to reproduce a given set of training data, therefore, they represent an expression of belief in the predictions of that particular simulator rather than a formal definition of probability. However, GLUE is fully coherent with a formal Bayesian approach when the use of a classical likelihood function is justifiable (see, e.g., Romanowicz *et al*., 1994).

In the work of Rojas *et al*., (2008) no significant differences were observed in the estimation of posterior model probabilities, predictive capacity and conceptual model uncertainty when using a Gaussian, a model efficiency or a Fuzzy-type likelihood function. The analysis in this work is therefore confined to a traditional Gaussian likelihood function $L\left(\mathrm{M}_k, \boldsymbol{\theta}_l, \mathbf{Y}_m \middle| \mathbf{D}\right)$, where M_k is the *k*-th conceptual model (or model structure) included in the finite and discrete ensemble of alternative conceptualizations $\mathbf{M}$, $\boldsymbol{\theta}_l$ is the *l*-th parameter vector, $\mathbf{Y}_m$ is the *m*-th input data vector, and $\mathbf{D}$ is the observed system variable vector.

2.2. Bayesian Model Averaging (BMA)

BMA provides a coherent framework for combining predictions from multiple conceptual models to attain a more realistic and reliable description of the total prediction uncertainty. It yields consensus predictions by weighing predictions from competing models based on their relative skill, with predictions from better performing models receiving higher weights than those of worse performing models. BMA avoids having to choose one model over the others, instead, observed data **D** give the competing models different weights (Wasserman, 2000).

Following the notation of Hoeting *et al*., (1999), if Δ is a quantity to be predicted, the BMA predictive distribution of Δ is given by

$$p\left(\Delta \mid \mathbf{D}\right)=\sum_{k=1}^{K} p\left(\Delta \mid \mathbf{D}, \mathrm{M}_k\right) p\left(\mathrm{M}_k \mid \mathbf{D}\right). \qquad (1)$$

Equation 1 is an average of the posterior distributions of Δ under each alternative conceptual model considered, $p(\Delta \mid \mathbf{D}, \mathrm{M}_k)$, weighted by their posterior model probability, $p(\mathrm{M}_k \mid \mathbf{D})$. This latter term reflects how well model *k* fits the observed data $\mathbf{D}$ and can be computed using Bayes' rule

$$p(\mathrm{M}_k \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \mathrm{M}_k)\, p(\mathrm{M}_k)}{\sum_{k'=1}^{K} p(\mathbf{D} \mid \mathrm{M}_{k'})\, p(\mathrm{M}_{k'})} \tag{2}$$

where $p(\mathrm{M}_k)$ is the prior probability of model *k*, and $p(\mathbf{D} | \mathrm{M}_k)$ is the integrated likelihood of the model *k*.

The leading moments of the BMA prediction of Δ are given by Draper (1995)

$$E[\Delta \mid \mathbf{D}] = \sum_{k=1}^{K} E[\Delta \mid \mathbf{D}, \mathrm{M}_k]\, p(\mathrm{M}_k \mid \mathbf{D}) \tag{3}$$

$$\begin{aligned} Var[\Delta \mid \mathbf{D}] = & \sum_{k=1}^{K} Var[\Delta \mid \mathbf{D}, \mathrm{M}_k]\, p(\mathrm{M}_k \mid \mathbf{D}) \\ & + \sum_{k=1}^{K} \left(E[\Delta \mid \mathbf{D}, \mathrm{M}_k] - E[\Delta \mid \mathbf{D}] \right)^2 p(\mathrm{M}_k \mid \mathbf{D}) \end{aligned} \tag{4}$$

From equations 1 and 2 it is seen that estimations of posterior model probabilities (weights) and, subsequently, estimations of the first two leading moments of the BMA predictive distribution (equations 3 and 4), are functions of the prior model probabilities assigned to the alternative conceptual models. From equation 4 it is seen that the variance of the BMA predictions consists of two terms; the first representing the within-model variance and the second representing the between-model variance (variance due to conceptual model uncertainty).

2.3. Combining GLUE and BMA

Combining GLUE and BMA involves the following sequence of steps

1. Based on prior and expert knowledge about the site, a suite of alternative conceptual models is proposed.
2. Realistic prior ranges are defined for the input and parameter vectors under each plausible model structure.
3. A likelihood measure and rejection criteria are defined.
4. For the suite of alternative conceptual models, input and parameter values are sampled from the prior ranges to generate possible simulators of the system.
5. A likelihood measure is calculated for each simulator based on the agreement between the simulated and observed system response.
6. Simulators that are not in agreement with the selected rejection criterion are discarded from the analysis by setting their likelihood to zero.
7. For each conceptual model M_k, a subset A_k of simulators with likelihood $p\left(\mathbf{D}\middle|\mathrm{M}_k,\boldsymbol{\theta}_l,\mathbf{Y}_m\right)=L\left(\mathrm{M}_k,\boldsymbol{\theta}_l,\mathbf{Y}_m\middle|\mathbf{D}\right)$ is retained. Steps 4-6 are repeated until the hyperspace of possible simulators is adequately sampled, i.e., when the conditional distributions of predicted state variables based on the likelihood weighted simulators in the subset A_k converge to a stable distribution for each of the conceptual models M_k.
8. The integrated likelihood of each conceptual model M_k is approximated by summing the likelihood weights of the retained simulators in subset A_k, or

$$p\left(\mathbf{D}\middle|\mathrm{M}_k\right)\approx\sum_{l,m\in A_k}L\left(\mathrm{M}_k,\boldsymbol{\theta}_l,\mathbf{Y}_m\middle|\mathbf{D}\right)\tag{5}$$

9. The posterior model probabilities are then obtained by normalizing the integrated model likelihoods such that they sum up to 1,

$$p\left(\mathrm{M}_k \middle| \mathbf{D}\right) \approx \frac{\sum_{A_k} L\left(\mathrm{M}_k, \boldsymbol{\theta}_l, \mathbf{Y}_m \middle| \mathbf{D}\right) p\left(\mathrm{M}_k\right)}{\sum_{j=1}^{K} \sum_{l,m \in A_j} L\left(\mathrm{M}_j, \boldsymbol{\theta}_l, \mathbf{Y}_m \middle| \mathbf{D}\right) p\left(\mathrm{M}_j\right)} \tag{6}$$

10. After normalization of the likelihood weighted predictions under each individual model (such that the cumulative likelihood under each model equals 1) a multi-model prediction is obtained with equation 1 using the weights obtained with equation 6.

Details about the implementation of the methodology, applied to the 3-dimensional hypothetical setup described in the next section, are presented in Section 4.

3. Three-dimensional hypothetical setup

For illustrative purposes, we employ a 3-dimensional hypothetical setup for which the true conditions are known (Figure 1). Lateral dimensions are 5000 m (E-W) by 3000 m (N-S) discretised in 25 m by 25 m grid cells. The system extents over 60 m in the vertical direction, with undisturbed layer thicknesses of 35 m (upper aquifer), 5 m (middle aquitard) and 20 m (lower aquifer). We assume statistically homogeneous deposits with a constant mean hydraulic conductivity *K* (see Table 1). Smaller-scale variability is represented using the theory of random space functions, adopting isotropic exponential covariance functions for log *K* in all layers. The spatial distribution of the hydraulic conductivity in the layers of the example setup, as well as any other realization of the hydraulic conductivity field used in this work, is generated using the sequential Gaussian simulation (sGsim) algorithm of the Geostatistical Software Library (Deutsch and Journel, 1998). Parameters of the covariance function of log *K* for the different layers are presented in Table 1.

Simulation of steady-state flow is performed using Modflow-2000 (Harbaugh *et al*., 2000). At the north and south boundaries, as well as at the bottom of the lower layer, zero gradient conditions are imposed. A uniform recharge of 1.4 x 10^{-4} m d^{-1} is applied to the top layer. At the west boundary a constant head h = 46 m is defined. The east side of the domain is bounded by a 10 m-wide river with a constant stage of 25 m. The river bottom is at 20 m, defining a constant river water depth of 5 m. It is underlain by 5 m-thick sediments with a vertical hydraulic conductivity of 0.1 m d^{-1}. Five pumping wells are distributed in the area producing a total of 2450 m^3 d^{-1} from the lower aquifer (Figure 1). An evapotranspiration zone, delineated by the polygon in Figure 1, is defined with an evapotranspiration surface elevation at 43 m, an evapotranspiration rate of 1.37 x 10^{-3} m d^{-1} and an extinction depth of 5 m.

The resulting "true" groundwater head distribution for the top layer is presented as an overlay in Figure 1. The ambient background gradient from west to east is strongly influenced by the drawdown around pumping wells, the evapotranspiration zone as well as by local effects of spatially varying hydraulic conductivity. From the "true" groundwater head distribution for layer 1, values are selected at the 16 locations defined by the observation wells in Figure 1, which are used to estimate the likelihood weights in the evaluation of different simulators.

4. Implementation of the methodology and numerical analysis

4.1. Implementation of the GLUE-BMA approach

We consider 7 alternative conceptual models with increasing complexity to describe the 3-dimensional hypothetical setup described in section 3, namely: (1), (2) and (3) one-layer models with mean K and spatial correlation law of layer 1 (1Lhtg-L1), layer 2 (1Lhtg-L2) and layer 3 (1Lhtg-L3) of the hypothetical setup, respectively; (4) a one-layer model with average mean K and spatial correlation (1Lhtg-AVG); (5) a two-layer model with mean K and spatial correlation taken from layer 1 and layer 3 (2Lhtg); (6) a two-layer quasi-three dimensional model with mean K and spatial correlation taken from layer 1 and layer 3, and mean K of layer 2 used to define the aquitard (2LQ3Dhtg); and (7) a three-layer model based on the spatial K distributions of layer 1, layer 2 and layer 3 (3Lhtg). All conceptual models comprise a total aquifer thickness of 60 m and are forced by identical types of boundary conditions.

The dimensionality of the analysis is confined by considering uncertainty only in the input variables and parameters related to the evapotranspiration process, lateral boundary conditions, river description and recharge process, i.e., input variables and parameters that are common to all setups. Values are sampled from uniform prior distributions for the unknown inputs and parameters with ranges defined in Table 2. Unconditional realizations of the hydraulic conductivity field are generated with the same mean K and spatial correlation law as the respective layers in the hypothetical setup (Table 1). For the 1Lhtg-AVG conceptualization the average of these values is used.

For the simulation, parameter and input vectors sampled using a Latin Hypercube Sampling (LHS) scheme, are combined with unconditional hydraulic conductivity realizations and consequently evaluated under each conceptual model. Based on the evaluation of a set of initial runs, a rejection threshold is defined corresponding to a maximum allowable deviation of 5 m at any of the 16 observation wells depicted in Figure 1. A point rejection threshold rather than a global rejection threshold is chosen because under the latter criteria strong deviations at certain locations (typically in the vicinity of pumping wells) may be offset by small deviations at other wells. For each conceptual model, predictive distributions for the sixteen observation wells depicted in Figure 1 and different components of the groundwater budget (recharge inflows, groundwater inflows/outflows from the west boundary condition (WBC), river gains, and evapotranspiration (EVT) outflows) are obtained from the ensemble of likelihood weighted predictions. Sampling from the prior input and parameter space continued until the first and second moment of these predictive distributions stabilized.

4.2. Approach to assess sensitivity to prior model probabilities

To analyze the sensitivity to different values of prior model probabilities, the prior model probability space of the 7 alternative conceptualizations is discretised into 25 equidistant intervals of 4% probability each. To avoid extremely low model probabilities that *reject with high certainty* one of the proposed alternative conceptual models, the lowest probability intervals are discarded from the analysis (this implies that the highest probability of a model is 1-6*0.04 = 0.76). From the remaining 19 probability intervals the lowest value of each interval is retained, resulting in the following set of potential prior model probabilities for each of the 7 alternative conceptual models P = [0.04, 0.08, …, 0.76]. Subsequently, all combinations that fill the prior probability space conditional on $\mathbf{M}$, i.e., for which $\sum_{k=1}^{K} p\left(\mathrm{M}_k\right)=1$ (243 vectors of 7 elements), are formed. This yields a total of 132,861 potential discrete sets of prior model probabilities that are used to numerically analyze the sensitivity of the posterior model probabilities (weights in equation 1), multi-model predictions and conceptual model uncertainty estimation to prior model probabilities.

4.3. Constrained maximum entropy approach to assess value of prior knowledge

The value of prior knowledge about the plausibility of the 7 alternative conceptual models in assessing conceptual model uncertainty is evaluated following a constrained maximum entropy method (Ye *et al*., 2005). The method aims to find discrete sets of prior model probabilities that maximizes Shannon's entropy *H* (Shannon, 1948) given by

$$H=-\sum_{k=1}^{K} p\left(\mathrm{M}_k\right)\log p\left(\mathrm{M}_k\right) \tag{7}$$

and subject to

$$\begin{aligned} h_i &= 0 \qquad i = 1,\ldots,I \\ g_j &= 0 \qquad j = 1,\ldots,J \end{aligned} \tag{8}$$

where h_i and g_i represent quantitative relations that reflect prior knowledge about the plausibility of the alternative conceptual models. In equation (7) $p(\mathrm{M}_k)$ is the prior model probability of the k-th conceptual model contained in the ensemble $\mathbf{M}$ of dimension K. In the case that alternative sets of constraints (reflecting different knowledge states) are proposed and when not enough data are available to assess the quality of model results, Ye *et al.*, (2005) advocate selecting the set that: (i) maximizes entropy H, (ii) presents a minimum entropy among proposed sets and, (iii) maximizes the likelihood for $\mathbf{M}$ given by the normalizing term in equation 2.

For illustrative purposes we define 3 different prior knowledge states: (i) Prior Set 1, corresponding to a set of uniform prior model probabilities $p(\mathrm{M}_k) = 1/K$, reflecting a state of complete ignorance about the plausibility of the alternative conceptual models; (ii) Prior Set 2, corresponding to a set where alternative conceptual models receive higher prior probability as they approach the 3-dimensional hypothetical setup described in section 3 and, thus, reflecting relevant and proper prior knowledge about the alternative conceptualizations; and (iii) Prior Set 3, corresponding to a set where prior model probabilities are inconsistent with the degree of similarity between the alternative conceptual models and the 3-dimensional hypothetical setup and, thus, reflecting improper prior knowledge about the alternative conceptualizations.

We adopted the following set of constraints to reflect the information contained in the three proposed prior knowledge states

Set 1:

$$\max_{p(\mathrm{M}_k)} H = -\sum_{k=1}^{K} p(\mathrm{M}_k)\log p(\mathrm{M}_k)$$
$$h_1 = \sum_{k=1}^{K} p(\mathrm{M}_k) - 1$$
$$g_1 \ldots g_6 \ : p(\mathrm{M}_1) = p(\mathrm{M}_2) = p(\mathrm{M}_3) = p(\mathrm{M}_4) = p(\mathrm{M}_5) = p(\mathrm{M}_6) = p(\mathrm{M}_7)$$

Set 2:

$$\begin{aligned}
&\max_{p(\mathrm{M}_k)} H = -\sum_{k=1}^{K} p(\mathrm{M}_k)\log p(\mathrm{M}_k) \\
&h_1 = \sum_{k=1}^{K} p(\mathrm{M}_k) - 1 \\
&g_1: \quad p(\mathrm{M}_1) - p(\mathrm{M}_2) = 0 \\
&g_2: \quad p(\mathrm{M}_2) - p(\mathrm{M}_3) = 0 \\
&g_3: \quad p(\mathrm{M}_4) - 2.0p(\mathrm{M}_3) \geq 0 \\
&g_4: \quad p(\mathrm{M}_5) - 1.5p(\mathrm{M}_4) \geq 0 \\
&g_5: \quad p(\mathrm{M}_6) - 2.5p(\mathrm{M}_5) \geq 0 \\
&g_6: \quad p(\mathrm{M}_7) - 1.1p(\mathrm{M}_6) \geq 0
\end{aligned}$$

Set 3:

$$\begin{aligned}
&\max_{p(\mathrm{M}_k)} H = -\sum_{k=1}^{K} p(\mathrm{M}_k)\log p(\mathrm{M}_k) \\
&h_1 = \sum_{k=1}^{K} p(\mathrm{M}_k) - 1 \\
&g_1: \quad p(\mathrm{M}_1) - p(\mathrm{M}_2) = 0 \\
&g_2: \quad p(\mathrm{M}_2) - p(\mathrm{M}_3) = 0 \\
&g_3: \quad 0.5p(\mathrm{M}_3) - p(\mathrm{M}_4) \geq 0 \\
&g_4: \quad 0.8p(\mathrm{M}_4) - p(\mathrm{M}_5) \geq 0 \\
&g_5: \quad 0.5p(\mathrm{M}_5) - p(\mathrm{M}_6) \geq 0 \\
&g_6: \quad 0.5p(\mathrm{M}_6) - p(\mathrm{M}_7) \geq 0
\end{aligned}$$

For Prior Set 1 (uniform prior model distribution) the solution to the optimization problem is known to be $H = \log K = 1.95$ (see, e.g., Applebaum, 1996, p. 100) with $p(\mathrm{M}_k) = 1/7$. For Prior Set 2 and 3 the nonlinear optimization problem is solved numerically using a sequential equality constrained quadratic programming method implemented in an R interface (Tamura, 2007) for the code DONLP2 (Spellucci, 1998). The result of these optimization problems are three optimized sets of prior model probabilities for the 7 alternative conceptual models that are in agreement with the quantitative relations (constraints) expressing the prior knowledge states. The optimized values are presented in Table 3. These three sets of prior model probabilities are samples from the full range of possible prior probability combinations, approximated here by the ensemble of discrete sets. It is important to note that the values of the constants in the constraints for Prior Set 2 and 3 were set as an example. Other values for these constants would result in different prior model probabilities, however, still reflecting prior knowledge. Consequently, the present analysis is conditional on the proposed ensemble of alternative conceptual models, $\mathbf{M}$, and to the potential quantitative relations among them, i.e., h_i and g_i.

5. Results and discussion

In the numerical analysis, for the alternative conceptual models 1Lhtg-L1 and 1Lhtg-L2 none of the simulations were accepted, as all of them failed to meet the criterion of a maximum allowable departure of 5 m from the observed heads. This suggests that approximating the "true" 3-dimensional hypothetical setup using only information from layers 1 and 2 (see Table 1) is not supported by the training data **D** (i.e., observed head at 16 observation wells). Hence, the posterior probability of these conceptual models was set to zero and they were discarded from the posterior analysis.

5.1. Sensitivity of posterior model probabilities to prior model probabilities

The sensitivity of the posterior model probabilities to prior model probabilities for the 5 retained conceptual models is presented in Figure 2. In this figure, vertical columns represent posterior model probabilities (estimated using equation 6) corresponding to the 132,861 nonzero discrete sets of prior model probabilities described in section 4.2. It can be seen that the posterior model probabilities are sensitive to values of prior model probabilities for all the retained models. It should be noted that the increase of the posterior model probabilities for the 5 retained conceptual models, i.e., nearly all points lie above the bisector curve, is caused by the fact that 2 out of 7 alternative conceptual models were discarded from the posterior analysis based on the information contained in the training data **D**. As a consequence, the share in the prior probability space of the discarded conceptualizations is redistributed over the 5 retained conceptual models when filling the posterior probability space (i.e., sum of posterior probabilities should equal to 1). This explains why in most cases the posterior probability is larger than the prior probability for the retained models. Notwithstanding, for alternative conceptualizations 1Lhtg-L3 (Figure 2a) and 1Lhtg-AVG (Figure 2b) values of posterior model probabilities below the bisector curve can be found, suggesting that less weight is assigned a posteriori to these models. For alternative conceptual models 2Lhtg (Figure 2c), 2LQ3Dhtg (Figure 2d) and 3Lhtg (Figure 2e), on the other hand, posterior model probabilities are always higher than prior model probabilities, this being more noticeable for model 3Lhtg.

From Figure 2 it is also seen that the uncertainty in the estimation of posterior model probabilities (expressed by the range of the vertical columns) is maximum when there is no clear preference a priori for a given conceptual model. On the contrary, the range of potential values for posterior model probabilities is reduced when an alternative conceptual model is preferred over the others.

Results for the three example sets of optimised prior model probabilities are also included in Figure 2 and are summarized in Table 3. Results confirm that posterior model probabilities, $p(\mathrm{M}_k|\mathbf{D})$ are largely influenced by the selection of a set of prior model probabilities. For Prior Sets 1 and 2, all retained models received more weight after conditioning. For Prior Set 2, on the other hand, the posterior probability of the two retained one-layer models was smaller than their respective prior probability, whereas the other 3 retained models received more weight after conditioning. However, for all 3 sets, the relative increase of the posterior probability compared to the prior probability is larger for the models approaching the true setup.

5.2. Sensitivity of the prior entropy, likelihood ratio and posterior entropy to prior model probabilities

The sensitivity of the prior entropy, likelihood ratio (with respect to the non-informative case) and posterior entropy (calculated using equation 7 with $p(\mathrm{M}_k|\mathbf{D})$ instead of $p(\mathrm{M}_k)$) is presented in Figure 3 for model 3Lhtg. It is seen in this figure that prior and posterior entropy decreased when prior model probabilities of model 3Lhtg increased. Moreover, the likelihood ratio (with respect to the non-informative case) tends to be maximized (Figure 3b) for a maximum probability of model 3Lhtg. Consider, for example, a prior model probability of 0.76 for model 3Lhtg and, consequently, 0.04 for the 6 remaining models. Clearly, this set of prior model probabilities is optimum (globally) in the sense that it minimizes posterior entropy and it maximizes the likelihood ratio.

For the 3 example sets, the smallest maximum prior entropy, the smallest posterior entropy, which can be interpreted as a measure of residual uncertainty after conditioning on the training data **D** (Ye *et al.*, 2005), and the largest likelihood ratio (1.34 times that of Prior Set 1) are obtained for Prior Set 2. On the contrary, the lowest likelihood ratio is observed for Prior Set 3, which suggests that this set is not in agreement with the information contained in the data and that it constitutes an improper expression of prior knowledge about the alternative conceptual models. Hence, for the problem at hand, a reasonable choice for a discrete set of prior model probabilities is to assign increasing probabilities in function of proximity to the 3-dimensional hypothetical setup, i.e., Prior Set 2.

5.3. Sensitivity of multi-model predictions and conceptual model uncertainty estimations

The sensitivity of the leading moments (estimated using equations 3 and 4) for model output river gains and for three alternative conceptual models (1Lhtg-L3, 2Lhtg and 3Lhtg) is presented in Figure 4. This figure shows that the posterior moments (plates a-f) of the predictive distribution for river gains are rather sensitivity to prior model probabilities. It is also seen that uncertainty in the estimation of the leading moments (expressed as the range of the vertical columns) increased when the corresponding prior model probabilities decreased. Additionally, when prior model probabilities for each alternative model increased, the leading moments converged to different values. The latter suggests that when a model is preferred over the others, i.e., relying only on a single conceptual model, predictions and uncertainty estimations tend to be biased. Moreover, estimation of the leading moments tends to be markedly more biased when prior model probabilities of simpler model 1Lhtg-L3 increased.

Plates g, h and i of Figure 4 show between-model variances for models 1Lhtg-L3, 2Lhtg and 3Lhtg, respectively, which are an expression of the conceptual model uncertainty. In general, the contribution of conceptual model uncertainty to the total spread is sensitive to prior model probabilities. Uncertainty in the estimation of between-model variance (expressed as the range of the vertical columns) increased when prior model probabilities decreased. Moreover, for the alternative conceptual models, between-model variances converged to different values when corresponding prior model probabilities increased. It should be noted, however, that for models 2Lhtg and 3Lhtg the converged values of between-model variances (2.1×10^3 and 2.8×10^{3} $[m^3\ d^{-1}]^2$, respectively) were rather similar for a maximum prior model probability of 0.76. However, the ratio between-model to total variance was somewhat different (7% and 18%, for 2Lhtg and 3Lhtg, respectively) due to the difference in the estimation of total variance for these models.

Figure 5 shows contour plots of the total variance and between-model variance (expressed as a percentage of the total variance) for model outputs west boundary condition (WBC) inflows, river gains and EVT outflows in the prior model probability space of 1Lhtg-L3 (simpler model) and 3Lhtg (model closer to the 3-dimensional hypothetical setup) when the 3 remaining alternative conceptual models approach a value near the uniform case (0.16). As consequence, only 52% of the prior model probability space is left to be distributed in the plates of Figure 5. More important than the actual values of the contour lines (which are approximations since the true uniform case has a value of 0.143) is the shape of the surface defined in the prior model probability space.

Plates a, b and c of Figure 5 show that the rate of change of the total variance (a measure of sensitivity) is much larger in the prior space of model 1Lhtg-L3 (x-axis) compared to the prior space of model 3Lhtg (y-axis). Hence, a more important reduction of the total variance would be expected when prior model probabilities of 1Lhtg-L3 decrease. This suggests that, for the problem at hand, to obtain more accurate multi-model predictions, simpler models should receive less prior weight compared to more elaborated models. In addition, it is seen from plates d, e and f that between-model variances does not fall below 5%, 20% and 12% of the total variance and, on the other hand, they can reach values as large as 12%, 30% and 18% of the total variance for WBC inflows, river gains and EVT outflows, respectively. Furthermore, the maximum contribution of between-model variances to total variances tends to be located around the middle area of the figures, which is contrasting with the fact that the non-informative case (uniform prior model probabilities) is not located in this area.

Overall, Figure 5 suggests that when a conceptual model tends to be preferred over the others, between-model variance tends to be minimum. This is in agreement with previous statements about under-dispersive properties of uncertainty estimations based on a single model. On the contrary, between-model variance tends to be maximum when there is no clear preference for a given conceptual model, suggesting that uncertainty estimations based on a suite of alternative models are more spread. This seems logic since including alternative conceptual models provides a more conservative assessment of uncertainty due to including conceptual model uncertainty.

Figures 4 and 5 also include values for the three optimized sets of prior model probabilities. Although posterior moments converged to different values for different conceptual models in Figure 4, convergence was in agreement with the values obtained using Prior Set 2 when models approached the “true” 3-dimensional hypothetical setup (see, e.g., plates c, f and i). This supports the idea stated before that Prior Set 2 is a suitable choice to assign prior model probabilities. This is also supported by the evidence provided by the data, which gave slightly higher integrated model likelihood values to model 3Lhtg. It is also seen from Figures 4 and 5 that the posterior variance, with respect to the non-informative case (Prior Set 1), significantly decreased when proper prior knowledge (Prior Set 2) was included in the analysis. On the contrary, in the case of improper prior knowledge (Prior Set 3) a significant increase of the total variance was observed. More importantly, between-model variances (plates g, h and i of Figure 4) significantly decreased with respect to the non-informative case (Prior Set 1) when proper prior knowledge (Prior Set 2) was included in the analysis, indicating the value of prior knowledge in reducing conceptual model uncertainty.

Similar results were found for the other groundwater budget terms (Table 4). With respect to Prior Set 1, total variances decreased between 40 and 60 % when the more informative Prior Set 2 was used. On the contrary, total variances increased between 32 and 60% when improper prior knowledge was included (Prior Set 3). Between-model variances decreased for the informative Prior Set 2 by 50 up to 62% with respect to Prior Set 1. However, the relative contribution of between-model variance to the total variance did not substantially decrease. For example, for EVT outflows obtained using Prior Set 1, the contribution of between-model to total variance is 0.15 whereas for Prior Set 2 this ratio is 0.14. The largest reduction in the contribution of between-model to total variance for Prior Set 2 is observed for river gains; from 0.26 to 0.2. This suggests that the contribution of conceptual model uncertainty to total uncertainty can not be further reduced based only on prior knowledge about the plausibility of alternative conceptualizations. This indicates that other sources of information or conditioning data should be included to further reduce this component of total variance.

For Prior Set 3 the between-model variances for WBC inflows and river gains increased, whereas for recharge inflows, WBC outflows and EVT outflows, between-model variances decreased compared to Prior Set 1. This erratic behaviour in the between-model variances estimated using Prior Set 3 is explained by Figure 5.

5.4. Value of prior knowledge about alternative conceptualizations in the goodness of GLUE-BMA predictions

Summary statistics of the posterior predictive distributions for the groundwater budget terms as a function of the optimized sets of prior model probabilities are presented in Figure 6. In this figure maximum values are truncated to enhance visual comparison. Observed values for the groundwater budget terms, obtained from the 3-dimensional hypothetical setup, are captured by the inter-quartile range of Prior Set 1 and Prior Set 2. On the contrary, observed values for WBC outflows, river gains and EVT outflows are not captured by the inter-quartile range of Prior Set 3. Comparing the optimized sets for each plate in Figure 6, Prior Set 2 outperforms the other sets since the median values are closer to the observed values and its inter-quartile range is more concentrated, indicating less residual uncertainty after observing data **D** and incorporating prior knowledge. Hence, this suggests that multi-model predictions obtained using the GLUE-BMA approach in combination with proper prior knowledge (Prior Set 2) outperforms multi-model predictions obtained using sets reflecting a non-informative case (Prior Set 1) and improper prior knowledge (Prior Set 3).

GLUE-BMA predictions for groundwater heads at the locations depicted in Figure 1 are presented in Figure 7. The predictive mean and standard deviation are estimated using equations 3 and 4, respectively. The more pronounced differences in the mean predicted head are observed for observation wells Obs-8, Obs-13, Obs-14, Obs-15 and Obs-16. It is interesting to note that, for these observation wells, observed heads are captured by the interval (± 1 standard deviation) defined around the predicted mean value using Prior Set 2. On the contrary, observed heads are not captured by the interval defined using Prior Set 1 and Prior Set 3. The exception to this is observation well Obs-2, in which none of the optimized sets was able to capture the observed head. It is also shown in Figure 7 that for some observation wells the standard deviations obtained using Prior Set 3 are slightly smaller compared to those obtained with the other optimized sets. However, this gain in accuracy is irrelevant since observed heads are not captured by the intervals defined using Prior Set 3 in 7 out of 16 observation wells. Therefore, an over-confident and biased prediction of the observed heads is obtained when improper prior knowledge (Prior Set 3) is used.

These results confirm that, for the problem at hand, when relevant and proper prior knowledge about the plausibility of alternative conceptual models is included in an analysis following the GLUE-BMA approach, the predictive capacity of the approach is substantially improved.

6. Conclusions

We investigated the influence of prior knowledge and prior model probability definition in a multi-model Bayesian averaging methodology which follows Bayesian formalism and that is used to assess uncertainty in the predictions of groundwater models arising from errors in the model structure, input (forcing) data and parameter estimates. The sensitivity analysis was based on the partitioning of the prior model probability space into discrete equidistant intervals of fixed probability. Subsequently, potential combinatorial sets were permuted to obtain sets of prior model probabilities for 7 alternative conceptualizations. The discrete sets were used to numerically analyze the sensitivity of posterior model probabilities and the leading moments of multi-model predictions of groundwater budget terms.

Additionally, the value of prior knowledge about alternative conceptual models in reducing conceptual model uncertainty was assessed using three illustrative sets of prior model probabilities. The three sets represented knowledge states expressing a non-informative case, proper prior knowledge, and improper prior knowledge about the plausibility of alternative conceptual models. For each of the sets a nonlinear optimization problem was solved in the form of linear (in)equalities expressing quantitative relationships among the alternative conceptualizations. This resulted in three optimized sets of prior model probabilities in agreement with the prior knowledge at hand.

For illustrative purposes a 3-dimensional hypothetical setup consisting of 2 aquifers separated by an aquitard, in which the flow field was considerably affected by pumping wells and spatially variable hydraulic conductivity, was used. Seven alternative conceptualizations with increasing complexity were adopted to describe the 3-dimensional hypothetical setup. Two of the simpler one-layer models were discarded from further analysis based on the evidence provided by the data.

Posterior model probabilities and leading moments of the multi-model predictive distributions showed to be very sensitive to different sets of prior model probabilities. This sensitivity clearly states the relevance of selecting proper prior probabilities in the context of the multi-model approach proposed by Rojas *et al*., (2008). In addition, increasing the prior model probability of a given alternative conceptual model over the other conceptualizations yielded biased leading moments and under-dispersive uncertainty estimations.

We showed that an optimized set of prior model probabilities in agreement with proper prior knowledge outperformed the non-informative and improper prior knowledge cases. Reductions between 40 and 60% (with respect to the non-informative case) for the total variances in model predictions were observed when proper prior knowledge was included in the analysis. On the contrary, total variances increased between 32 and 60% respect to the non-informative case when improper prior knowledge was included. Between-model variances, on the other hand, decreased between 50 and 62% when proper prior knowledge was included. Although in absolute terms, between-model and total variances considerably decreased with respect to the non-informative case when proper prior knowledge was included, for the problem at hand, the ratio between-model variance to total variance, within each optimized set, was not substantially modified. This suggests that the contribution of conceptual model uncertainty to total uncertainty can not be further reduced based only on prior knowledge about the plausibility of alternative conceptual models. This implies that other sources of information or conditioning data should be included to further reduce this component of the total variance.

The results of this study advocate incorporating proper prior knowledge about alternative conceptual models whenever available. Using a 3-dimensional hypothetical setup and three optimized discrete sets of prior model probabilities, it was shown that the predictive performance of the multi-model methodology proposed by Rojas *et al*., (2008) could be largely improved when proper knowledge is included. It is expected that combining proper prior knowledge about alternative conceptual models with other qualitative or quantitative sources of conditioning data, such as conductivity data, transient groundwater head information or recharge estimates, will further reduce conceptual model uncertainty. These topics will be subject of future research.

Acknowledgments

The authors thank the financial support provided to the first author in the framework of the PhD IRO-scholarships of the Katholieke Universiteit Leuven (K.U. Leuven). Assistance provided by Jorge Gonzalez to implement the R scripts is also acknowledged.

References

Ajami N, Duan Q, Gao X, Sorooshian S. 2005. Multi-model combination techniques for hydrologic forecasting: application to distributed model intercomparison project results. *Journal of Hydrometeorology* **7**(4): 755-768.

Applebaum D. 1996. *Probability and information: an integrated approach*. Cambridge University Press: New York; 297.

Beven K, Binley A. 1992. The future of distributed models – model calibration and uncertainty prediction. *Hydrological Processes* **6**(3): 279-283.

Beven K, Freer J. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology* **249**(1-4): 11-29.

Beven K. 2005. A manifesto for the equifinality thesis. *Journal of Hydrology* **320**(1-2): 18-36.

Binley A, Beven K. 2003. Vadose zone flow model uncertainty as conditioned on geophysical data. *Ground Water* **41**(2): 119-127.

Bredehoeft J. 2003. From models to performance assessment: The conceptualization problem. *Ground Water* **41**(5): 571-577.

Bredehoeft J. 2005. The conceptualization model problem – surprise. *Hydrogeology Journal* **13**(1): 37-46.

Carrera J, Alcolea A, Medina A, Hidalgo J, Slooten L. 2005. Inverse problem in hydrogeology. *Hydrogeology Journal* **13**(1): 206-222.

Deutsch C, Journel A. 1998. *GSLIB Geostatistical software library and user's guide*. Oxford University Press: New York; 384.

Draper D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B (with discussion)* **57**(1): 45-97.

Gelman A, Carlin J, Stern H, Rubin D. 2004. *Bayesian data analysis*. Chapman & Hall/CRC: New York; 668.

Ghosh J, Mohan D, Tapas S. 2006. *An introduction to Bayesian analysis: Theory and methods*. Springer texts in Statistics: New York; 352.

Harbaugh A, Banta E, Hill M, McDonald M. 2000. *MODFLOW-2000, U.S. Geological Survey modular ground-water model-user guide to modularization concepts and the ground-water flow process*. U.S. Geological Survey. Open File Report, 00-92; 121.

Harrar W, Sonnenberg T, Henriksen H. 2003. Capture zone, travel time, and solute transport using inverse modelling and different geological models. *Hydrogeology Journal* **11**(5): 536-548.

Hoeting J, Madigan D, Raftery A, Volinsky C. 1999. Bayesian model averaging: A tutorial. *Statistical Science* **14**(4): 382-417.

Hojberg A, Refsgaard J. 2005. Model uncertainty – parameter uncertainty versus conceptual models. *Water Science & Technology* **52**(6): 177-186.

Kass R, Wasserman L. 1996. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**(435): 1343-1370.

Liang F, Truong Y, Wong W. 2001. Automatic Bayesian model averaging for linear regression and application in Bayesian curve fitting. *Statistica Sinica* **11**(4): 1005-1029.

Meyer P, Ye M, Rockhold S, Neuman S, Cantrell K. 2007. *Combined estimation of hydrogeologic conceptual model, parameter and scenario uncertainty with application to uranium transport at the Hanford site 300 area*. Report NUREG/CR-6940 PNNL-16396, US Nuclear Regulatory Commission, Washington, USA.

Neuman S. 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment* **17**(5): 291-305.

Poeter E, Anderson D. 2005. Multimodel ranking and inference in ground water modelling. *Ground Water* **43**(4): 597-605.

Raftery A, Zheng Y. 2003. Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association* **98**(464): 931-938.

Refsgaard J, van der Sluijs J, Brown J, van de Keur P. 2006. A framework for dealing with uncertainty due to model structure error. *Advances in Water Resources* **29**(11): 1586-1597.

Rojas R, Feyen L, Dassargues A. 2008. Conceptual model uncertainty in groundwater modeling: combining generalized likelihood uncertainty estimation and Bayesian model averaging, submitted to *Water Resources Research*.

Romanowicz R, Beven K, Tawn J. 1994. Evaluation of prediction uncertainty in non-linear hydrological models using a Bayesian approach. In *Statistics for the Environment II; Water Related Issues*, Barnett V, Turkman K (eds.). Wiley: New York; 297-317.

Rubin Y. 2003. *Applied stochastic hydrogeology*. Oxford University Press: New York; 391.

Shannon C. 1948. A mathematical theory of communication. *Bell System Technical Journal* **27**: 379-423, 623-656.

Spellucci P. 1998. An SQP method for general nonlinear programs using only equality constrained subproblems. *Mathematical Programming* **82**(3): 413-448.

Tamura R. 2007. *Rdonlp2: An R extension library to use Peter Spellucci's DONLP2 from R*. R package version 0.3-1. http://arumat.ner/Rdonlp2/.

Wasserman L. 2000. Bayesian model selection and model averaging. *Journal of Mathematical Psychology* **44**: 92-107.

Ye M, Neuman S, Meyer P. 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resources Research* **40**, W05113, doi:10.1029/2003WR002557.

Ye M, Neuman S, Meyer P, Pohlmann K. 2005. Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff. *Water Resources Research* **41**, W12429, doi:10.1029/2005WR004260.

Ye M, Pohlmann K, Chapman J, Shafer D. 2006. On evaluation of recharge model uncertainty: A priori and a posteriori. In *Proceedings of the International High –level Radioactive Waste Management Conference*, Las Vegas, Nevada; 12.

Figures captions

Figure 1: Three-dimensional hypothetical setup including (⊙) observation wells and (⊗) pumping wells overlain by the groundwater head distribution in the first layer.

Figure 2: Posterior model probabilities for alternative conceptual models: a) 1Lhtg-L3, b) 1Lhtg-AVG, c) 2Lhtg, d) 2LQ3Dhtg and e) 3Lhtg for various sets of discrete prior model probabilities. Symbols represent optimized values of Prior Set 1 (□), Prior Set 2 (◇) and Prior Set 3 (○) described in section 4.3.

Figure 3: Sensitivity analysis in function of prior model probabilities for alternative conceptual model 3Lhtg for: a) prior entropy, b) likelihood ratio (respect to the Prior Set 1) and c) posterior entropy. Symbols represent optimized values of Prior Set 1 (□), Prior Set 2 (◇) and Prior Set 3 (○) described in section 4.3.

Figure 4: Leading moments for the posterior predictive distribution of river gains as function of the prior model probabilities of three alternative conceptual models 1Lhtg-L3 (a-d-g), 2Lhtg, (b-e-h) and 3Lhtg (c-f-i). Symbols represent optimized values of Prior Set 1 (□), Prior Set 2 (◇) and Prior Set 3 (○) described in section 4.3.

Figure 5: Contours of total variance (a-b-c) and between-model variance (d-e-f) (expressed as a percentage of total variance) for: a) recharge inflows x 10^4 [m^3 d^{-1}]^2; b) river gains x 10^4 [m^3 d^{-1}]^2, and c) EVT outflows x 10^5 [m^3 d^{-1}]^2 in the space of prior model probabilities of alternative conceptual models 1Lhtg-L3 and 3Lhtg when remaining models approach the non-informative case. Symbols represent optimized values of Prior Set 1 (□), Prior Set 2 (◇) and Prior Set 3 (○) described in section 4.3.

Figure 6: Summary statistics for the GLUE-BMA posterior predictive distributions for groundwater budget terms a) WBC inflows, b) recharge inflows, c) WBC outflows, d) river gains and e) EVT outflows for the optimized discrete sets Prior Set 1 (black), Prior Set 2 (red) and Prior Set 3 (light-grey) described in section 4.3. Open circles represent observed values obtained from the 3-dimensional hypothetical setup. Q_1 and Q_3 represent the first and third quartile, respectively. Maximum values are truncated to enhance visual comparison.

Figure 7: GLUE-BMA posterior mean (diamonds) estimated using equation 3 and the corresponding error bars expressing ± 1 standard deviation (estimated using equation 4) for the sixteen observation wells depicted in Figure 1 for the optimized discrete sets Prior Set 1 (black), Prior Set 2 (red) and Prior Set 3 (light-grey) described in section 4.3. Open circles represent observed values obtained from the 3-dimensional hypothetical setup.

Tables

Table 1: Parameters describing the hydraulic conductivity spatial correlation structure for the different layers of the 3-dimensional hypothetical setup (based on Rubin (2003), Tables 2.1 and 2.2, p34-36).

Layer	Model Parameters		
	μ_K [m d^{-1}]	$\sigma_{Ln\ K}$	$I_{Ln\ K}$
1	0.1	2.0	400
2	0.01	0.5	800
3	1	1.5	600

Table 2: Range of prior uniform distributions for unknown parameters.

Parameters			Range	
			Minimum	Maximum
Recharge rate	(RECH)	[m d^{-1}]	0	5.8e-04
Constant head west boundary condition (WBC)	(CH)	[m]	25	75
Elevation surface (EVT)	(SURF)	[m]	30	50
Extinction depth (EVT)	(EXTD)	[m]	0	25
Evapotranspiration rate (EVT)	(EVTR)	[m d^{-1}]	0	7.0e-03
River conductance	(RIVC)	[m^2 d^{-1}]	1.0e-02	1000

Table 3: Summary of integrated model likelihoods and posterior model probabilities for 7 alternative conceptual models and three optimized sets of prior model probabilities.

Prior Sets		Conceptual models							Entropy	Likelihood ratio
		1Lhtg-L1	1Lhtg-L2	1Lhtg-L3	1Lhtg-AVG	2Lhtg	2LQ3Dhtg	3Lhtg		
	$p(\mathbf{D}\|M_k)$	0	0	902.6	935.6	990.4	1046.9	1079.4		
Prior Set 1	$p(M_k)$	14.29	14.29	14.29	14.29	14.29	14.29	14.29	1.95	1.00
	$p(M_k\|\mathbf{D})$	0	0	18.22	18.88	19.99	21.13	21.78	1.61	
Prior Set 2	$p(M_k)$	4.08	4.08	4.08	8.15	12.23	31.79	35.60	1.58	1.34
	$p(M_k\|\mathbf{D})$	0	0	3.87	8.02	12.73	34.99	40.40	1.32	
Prior Set 3	$p(M_k)$	23.81	23.81	23.81	11.90	9.52	4.76	2.38	1.74	0.70
	$p(M_k\|\mathbf{D})$	0	0	43.31	22.45	19.01	10.05	5.18	1.40	

Table 4: Total variance and between-model variance for groundwater budget terms (expressed in [m^3 d^{-1}]) as function of the optimized prior probability sets described in section 4.3. Values in parentheses express percentage reduction with respect to the Prior Set 1.

Groundwater budget terms	Prior Set 1		Prior Set 2		Prior Set 3	
	Total Variance	Between-model variance	Total Variance	Between-model variance	Total Variance	Between-model variance
WBC inflow	463319.2	46854.7	280622.5 (39.4)	17666.0 (62.3)	667218.6 (-44.0)	65680.3 (-40.2)
Recharge inflow	516007.8	86870.7	312683.0 (39.4)	43341.4 (50.1)	681708.6 (-32.1)	79460.8 (8.5)
WBC outflow	7624.7	342.8	3016.7 (60.4)	173.4 (49.4)	12172.6 (-59.6)	319.1 (6.9)
River gains	33893.9	8951.4	19218.0 (43.3)	3871.3 (56.8)	48020.7 (-41.7)	10321.4 (-15.3)
EVT outflow	158321.9	23414.6	82788.2 (47.7)	11495.5 (50.9)	235107.8 (-48.5)	22315.2 (4.7)

Figures

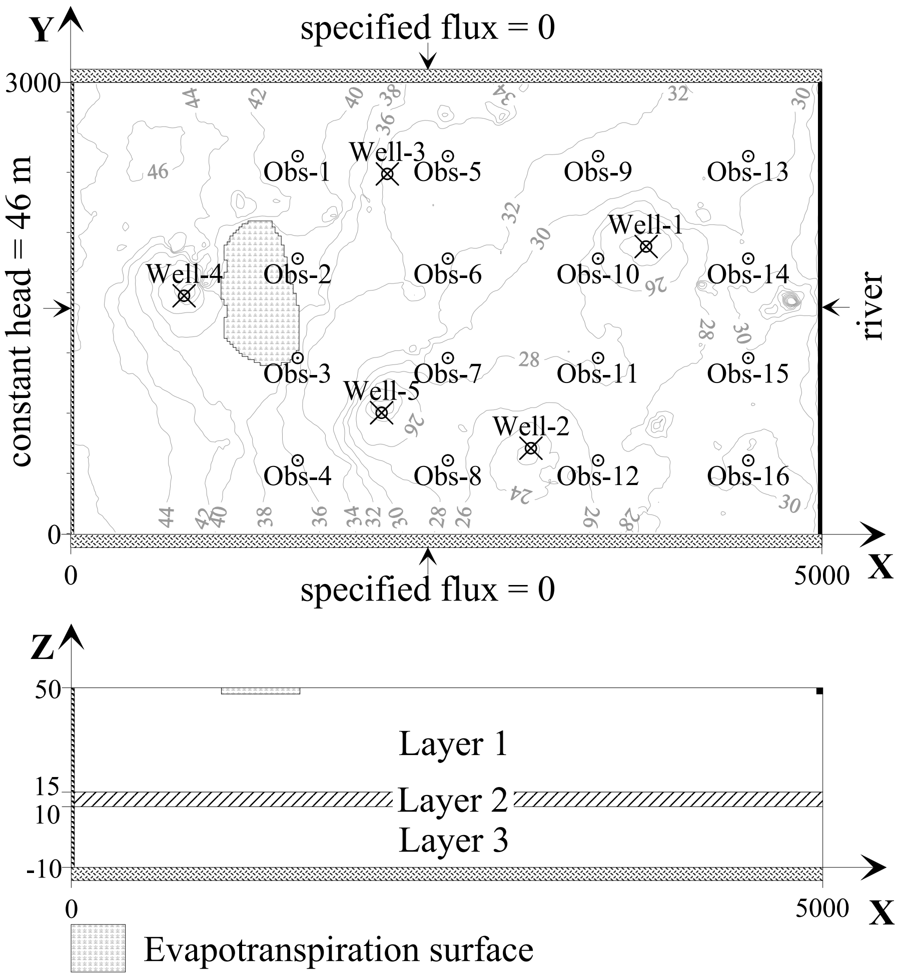

Figure 1: Three-dimensional hypothetical setup including (⊙) observation wells and (⊗) pumping wells overlain by the groundwater head distribution in the first layer.

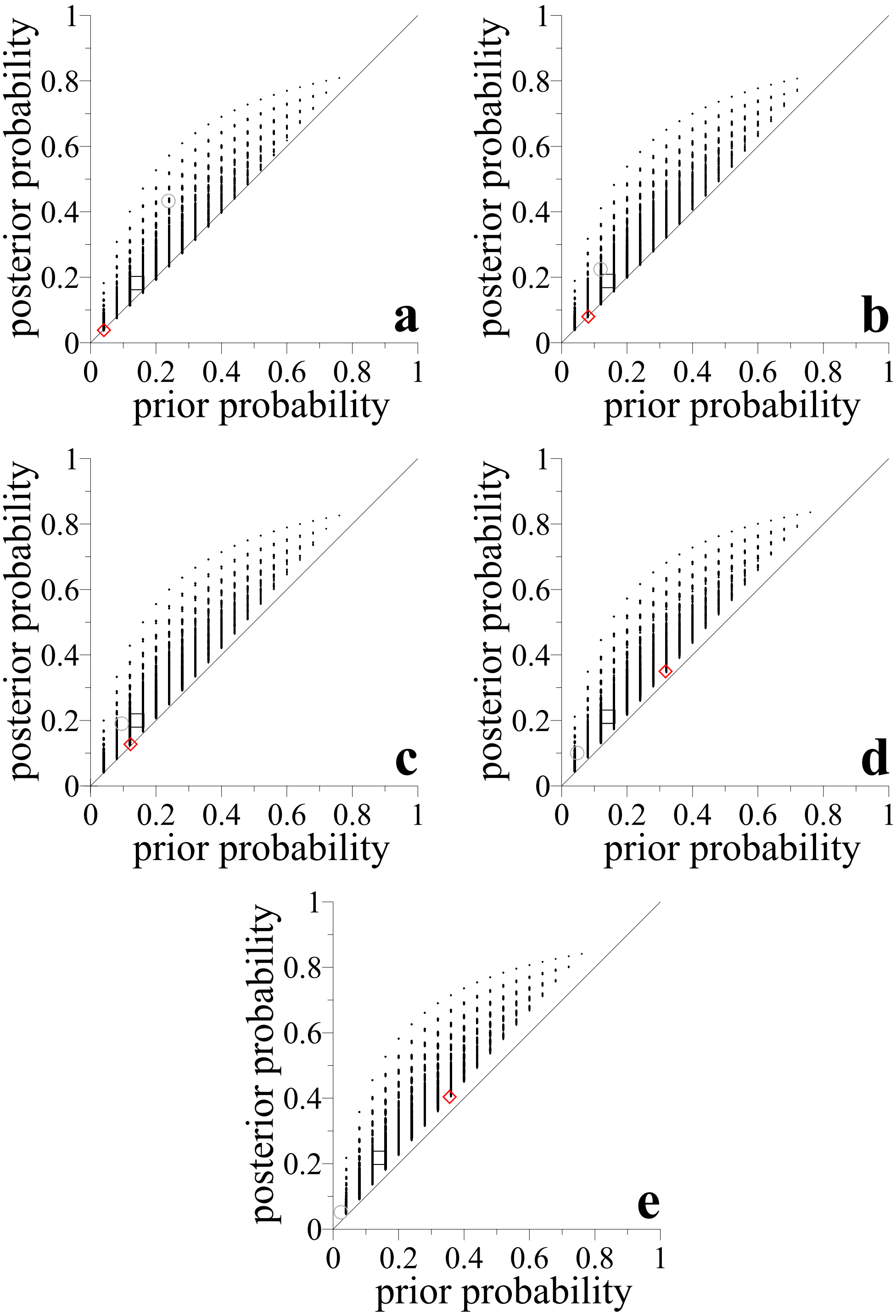

Figure 2: Posterior model probabilities for alternative conceptual models: a) 1Lhtg-L3, b) 1Lhtg-AVG, c) 2Lhtg, d) 2LQ3Dhtg and e) 3Lhtg for various sets of discrete prior model probabilities. Symbols represent optimized values of Prior Set 1 (□), Prior Set 2 (◇) and Prior Set 3 (○) described in section 4.3.

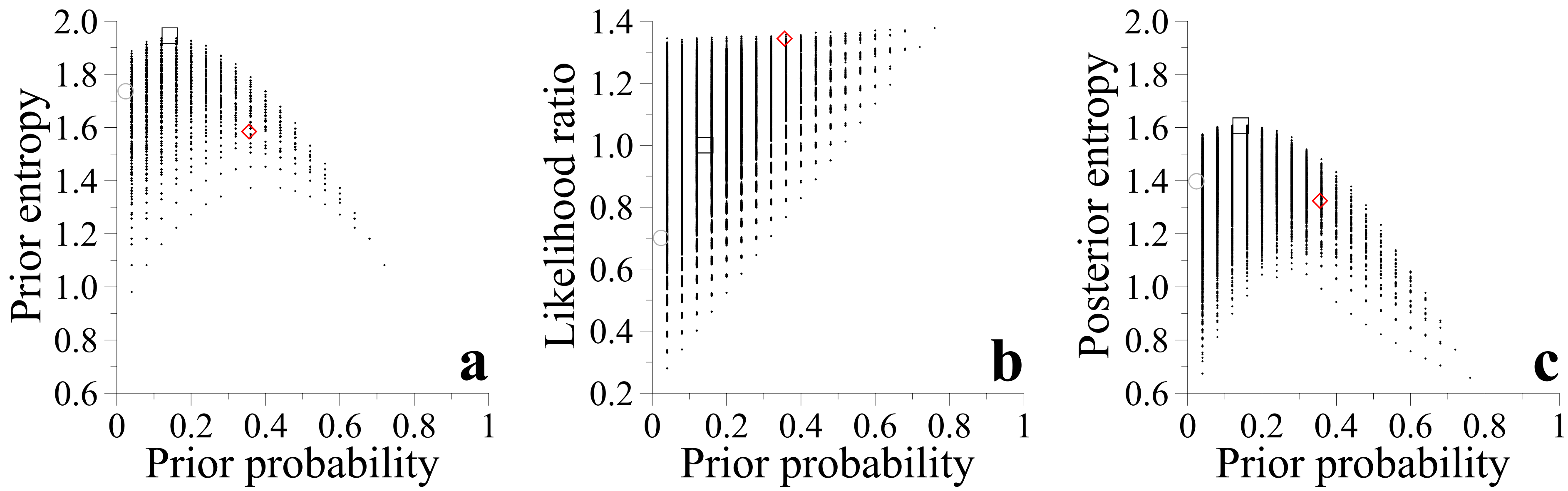

Figure 3: Sensitivity analysis in function of prior model probabilities for alternative conceptual model 3Lhtg for: a) prior entropy, b) likelihood ratio (respect to the Prior Set 1) and c) posterior entropy. Symbols represent optimized values of Prior Set 1 (□), Prior Set 2 (◇) and Prior Set 3 (○) described in section 4.3.

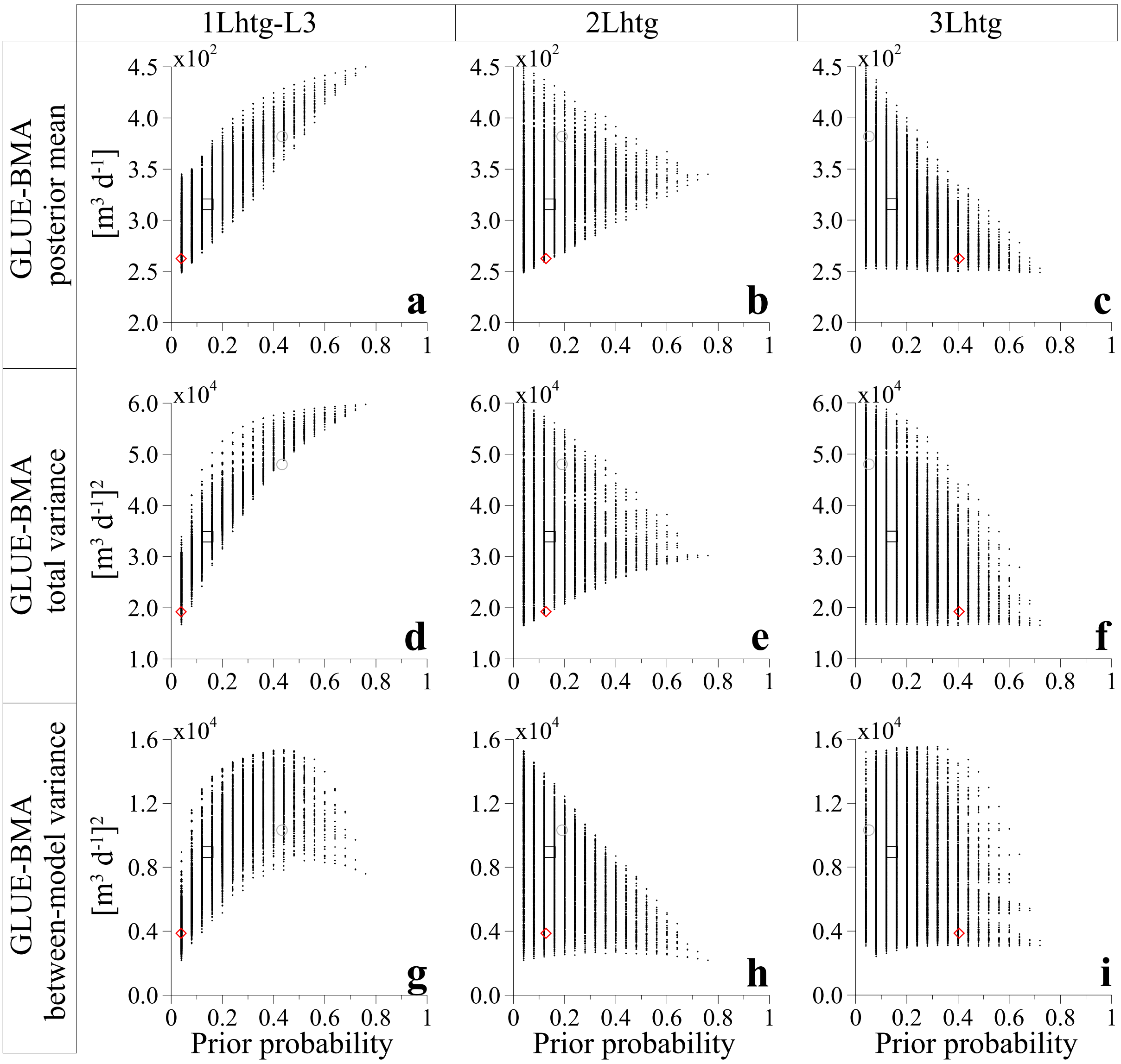

Figure 4: Leading moments for the posterior predictive distribution of river gains as function of the prior model probabilities of three alternative conceptual models 1Lhtg-L3 (a-d-g), 2Lhtg, (b-e-h) and 3Lhtg (c-f-i). Symbols represent optimized values of Prior Set 1 (□), Prior Set 2 (◇) and Prior Set 3 (○) described in section 4.3.

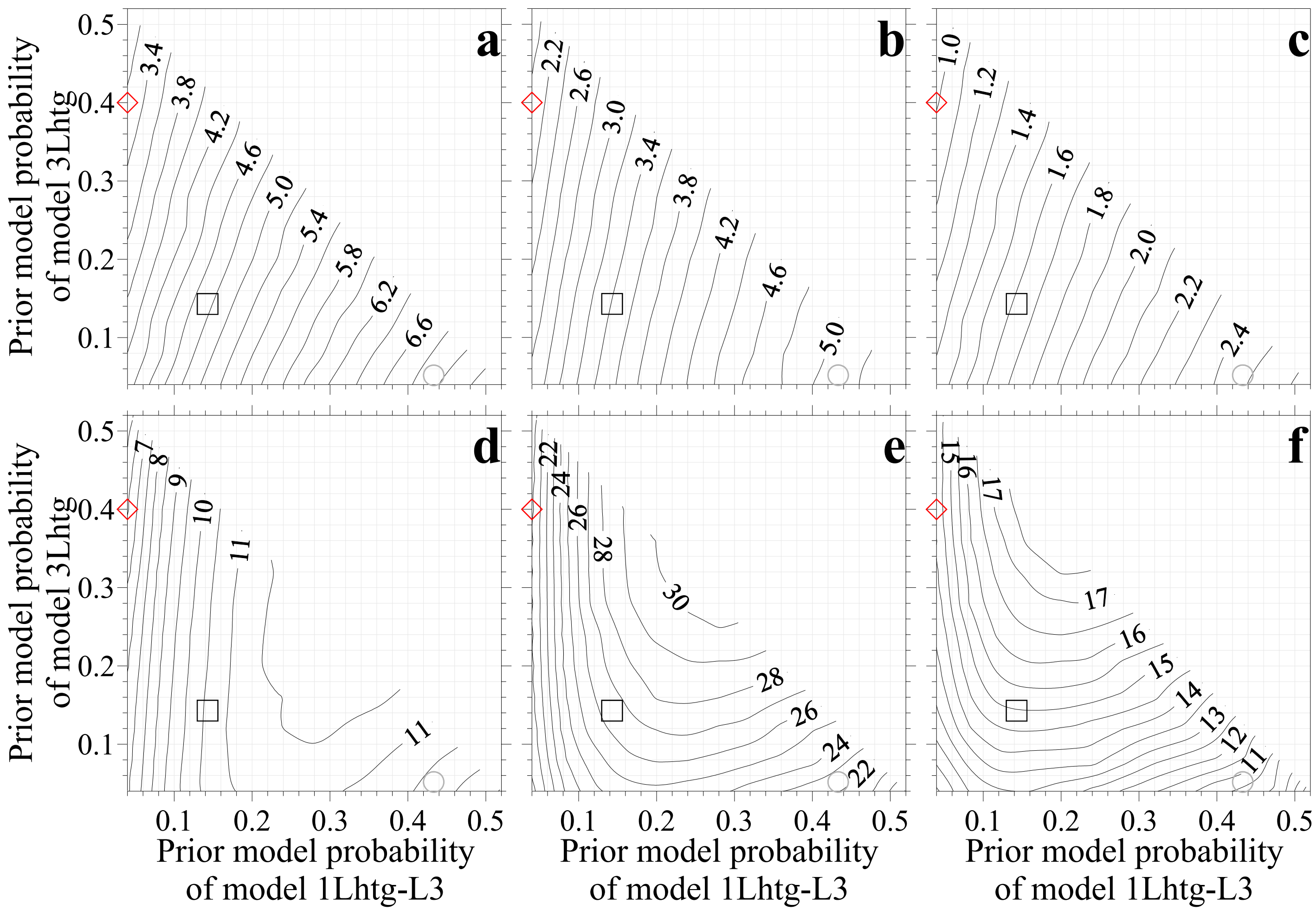

Figure 5: Contours of total variance (a-b-c) and between-model variance (d-e-f) (expressed as a percentage of total variance) for: a) WBC inflows x 10^5 $[m^3\ d^{-1}]^2$; b) river gains x 10^4 $[m^3\ d^{-1}]^2$, and c) EVT outflows x 10^5 $[m^3\ d^{-1}]^2$ in the space of prior model probabilities of alternative conceptual models 1Lhtg-L3 and 3Lhtg when remaining models approach the non-informative case. Symbols represent optimized values of Prior Set 1 (□), Prior Set 2 (◇) and Prior Set 3 (○) described in section 4.3.

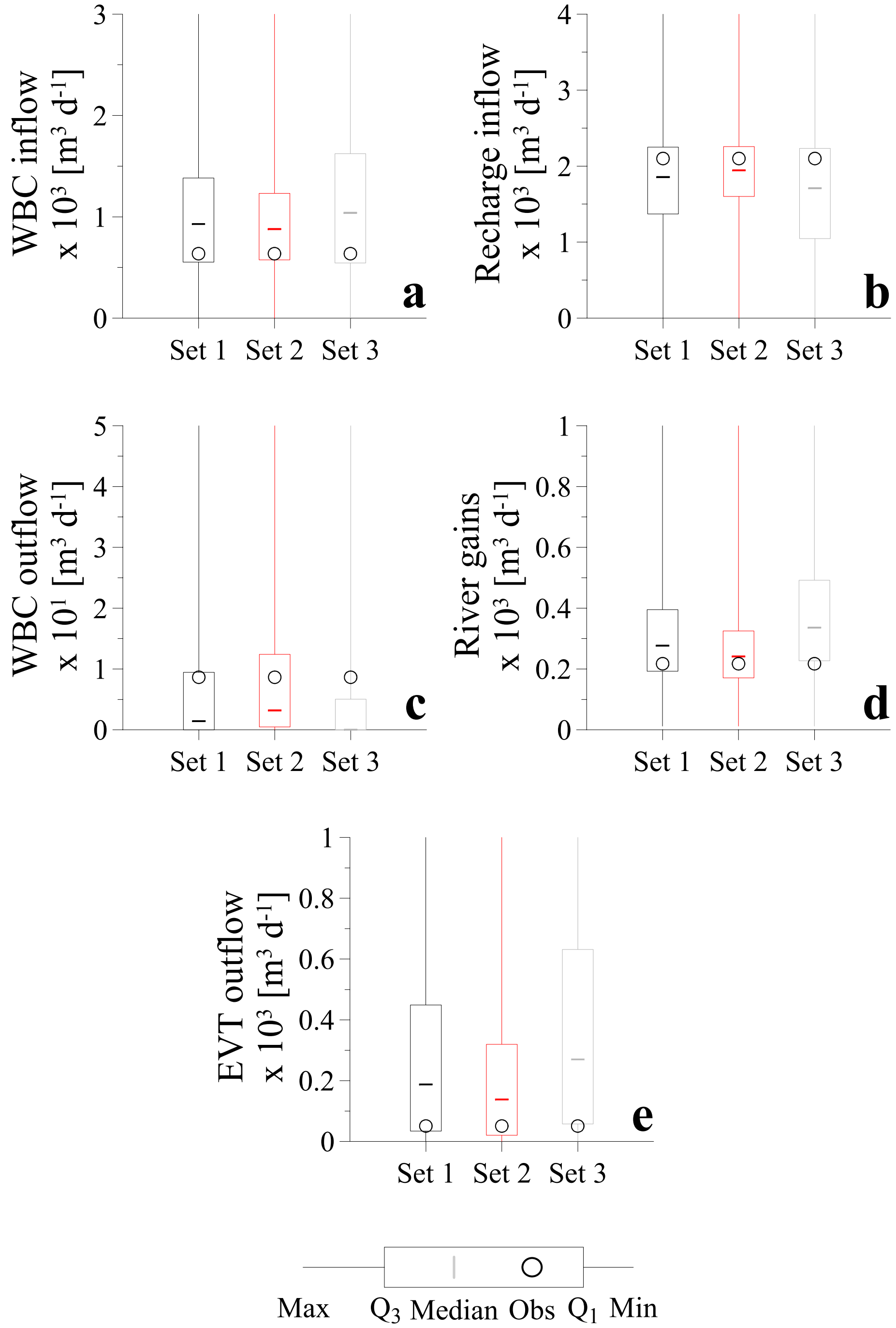

Figure 6: Summary statistics for the GLUE-BMA posterior predictive distributions for groundwater budget terms a) WBC inflows, b) recharge inflows, c) WBC outflows, d) river gains and e) EVT outflows for the optimized discrete sets Prior Set 1 (black), Prior Set 2 (red) and Prior Set 3 (light-grey) described in section 4.3. Open circles represent observed values obtained from the 3-dimensional hypothetical setup. Q_1 and Q_3 represent the first and third quartile, respectively. Maximum values are truncated to enhance visual comparison.

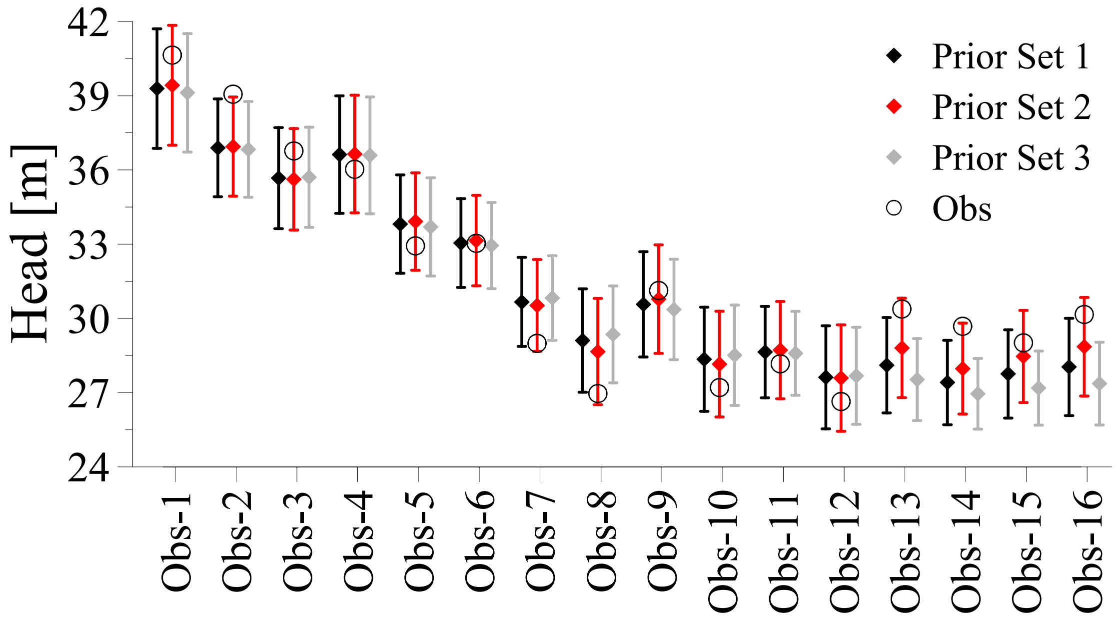

Figure 7: GLUE-BMA posterior mean (diamonds) estimated using equation 3 and the corresponding error bars expressing ± 1 standard deviation (estimated using equation 4) for the sixteen observation wells depicted in Figure 1 for the optimized discrete sets Prior Set 1 (black), Prior Set 2 (red) and Prior Set 3 (light-grey) described in section 4.3. Open circles represent observed values obtained from the 3-dimensional hypothetical setup.