Conceptual model uncertainty in groundwater modeling: combining generalized likelihood uncertainty estimation and Bayesian model averaging

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

[a*] Corresponding author.

Applied geology and mineralogy, Department of Earth and Environmental Sciences,

Katholieke Universiteit Leuven, Celestijnenlaan 200 E, 3001 Heverlee, Belgium

Tel.: +32 016 326449; fax: +32 016 326401.

E-mail address: Rodrigo.RojasMujica@geo.kuleuven.be

[b] Land management and natural hazards unit, Institute for Environment and Sustainability,

DG Joint Research Centre, European Commission, TP261, 21020 Ispra (Va), Italy

E-mail address: luc.feyen@jrc.it

[c] Hydrogeology and environmental geology, Department of Architecture, Geology,

Environment, and Civil Engineering (ArGEnCo),

Université de Liège, B.52/3 Sart-Tilman, B-4000 Liège, Belgium

E-mail address: alain.dassargues@geo.kuleuven.be, alain.dassargues@ulg.ac.be

Abstract

Uncertainty assessments in groundwater modeling applications typically attribute all sources of uncertainty to errors in parameters and inputs, neglecting what may be the primary source of uncertainty, namely, errors in the conceptualization of the system. Confining the set of plausible system representations to a single model leads to under-dispersive and prone to bias predictions. In this work we present a general and flexible approach that combines Generalized Likelihood Uncertainty Estimation (GLUE) and Bayesian Model Averaging (BMA) to assess uncertainty in model predictions that arise from errors in model structure, inputs and parameters. In a prior analysis, a set of plausible models are selected and the joint prior input and parameter space is sampled to form potential simulators of the system. For each model the likelihood measures of acceptable simulators, assigned to them based on their ability to reproduce observed system behavior, are integrated over the joint input and parameter space to obtain the integrated model likelihood. The latter is used to weight the predictions of the respective model in the BMA ensemble predictions. For illustrative purposes we applied the methodology to a three-dimensional hypothetical setup. Results showed that predictions of groundwater budget terms varied considerably among competing models, although that a set of 16 head observations used for conditioning did not allow differentiating between the models. BMA provided consensus predictions that were more conservative. Conceptual model uncertainty contributed up to 30% of the total uncertainty. The results clearly indicate the need to consider alternative conceptualizations to account for model uncertainty.

Keywords: GLUE, BMA, Multi-model prediction, Monte Carlo methods, uncertainty assessment

1. Introduction and scope

With increasing human and climate pressures on groundwater resources, accurate and reliable predictions of groundwater flow and pollutant transport are essential for sustainable groundwater management practices. However, typically, the geological structure is only partially known and point measurements of subsurface properties or groundwater heads are sparse and prone to error. Consequently, incomplete or biased process representation, errors in the specification of initial and boundary conditions, as well as errors in the model parameters, render the predictions of groundwater dynamics and pollutant transport uncertain.

Over the last decades, considerable efforts have been put in developing methods to determine optimal groundwater parameter values and in quantifying model prediction uncertainty associated with uncertainty in these parameter estimates. This has resulted in a variety of inverse techniques for groundwater modeling applications. We do not wish to provide a complete overview of parameter estimation methods but refer the reader to Sun (1994) and Carrera et al. (2005) for excellent reviews. Despite its extensive application, the major weakness of parameter-calibration approaches is that all sources of uncertainty are attributed to parameter errors. This often results in biased parameter estimates that compensate for errors in model structure, input data and measurement errors.

Typically, these methods ignore conceptual or structural uncertainty by confining the range of plausible system representations to a single hydrological model. This often leads to overconfidence in the predictive capabilities of the model and in predictive uncertainty analyses that are under-dispersive and prone to statistical bias. In recent years, a number of authors have acknowledged that conceptual model uncertainty has received less formal attention in groundwater applications than it should (e.g., Neuman, 2003, Neuman and Wierenga, 2003; Bredehoeft, 2003, 2005; Carrera et al., 2005; Poeter and Anderson, 2005; Refsgaard et al., 2006). Bredehoeft (2005) summarizes the main issues concerning conceptual model specification as follows: (1) modelers tend to consider their conceptual models as immutable; (2) frequently, errors in model predictions turn around a poor choice of the conceptual model; (3) data will fit more than one conceptual model equally well; consequently, (4) a good calibration of a model does not ensure a correct conceptual model; and (5) parametric uncertainty does not compensate for uncertainties derived from the conceptual model specification.

These concerns have motivated researchers in the hydrological sciences to consider multi-model methods, which seek to obtain consensus predictions from a set of plausible models by linearly combining individual model predictions. The weights to aggregate multiple model outputs can be equal (model average) in the simplest case, or can be determined through regression-based approaches (e.g., Abrahart and See, 2002; Georgakakos et al., 2004). However, the weights in such combinations are not connected to model performance and can take any arbitrary value, hence lacking physical interpretation. An approach in which weights are intrinsically connected to model performance has been proposed by Poeter and Anderson (2005). This approach combines predictions of multiple competing models using Akaike's weights (Akaike, 1974; Burnham and Anderson, 1998). However, it lacks of a consistent way to incorporate previous knowledge about parameters and conceptual models in the multi-model prediction. A similar method that partially overcomes the restriction of including previous knowledge about multiple model structures has been proposed by Refsgaard et al., (2006). In this approach, a suite of conceptual models are independently calibrated and a pedigree analysis is performed to assess the overall tenability of the models. Nonetheless, the pedigree analysis does not provide an indication of the relative quality of the different model structures and, consequently, it is difficult to include it in a quantitative uncertainty analysis in terms of probabilities (Refsgaard et al., 2006).

Bayesian Model Averaging (BMA) (Draper, 1995; Hoeting et al., 1999), on the other hand, employs probabilistic techniques to derive consensus predictions from a set of alternative models. In short, BMA weights the predictions of competing models by their corresponding posterior model probability, representing each model's relative skill to reproduce system behavior 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 to assess the joint predictive distribution of several competing models. MLBMA is an approximation of BMA that relies on maximum likelihood parameter estimation and expanding around these values through Monte Carlo simulation. Subsequently, the posterior model probabilities are approximated using the Kashyap information criterion (Kashyap, 1982). MLBMA does not require exhaustive Monte Carlo simulations and obviates the need of (though it can incorporate) prior information about model parameters, which is often difficult to obtain (Ye et al., 2005). 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.

An alternative methodology that rejects the idea of a unique optimal simulator of the natural system is the Generalized Likelihood Uncertainty Estimation (GLUE) method (Beven and Binley, 1992; Beven, 1993). GLUE is based on the concept of equifinality, which acknowledges that there exist many combinations of model structures and parameter sets that provide (equally) good reproductions of the observed system response. For each possible simulator a likelihood measure is defined based on the degree of correspondence between simulated and observed records of system responses. Simulators that perform better than a subjectively chosen threshold criteria are retained and consequently used to provide an ensemble of likelihood weighted predictions of the system under future forcing conditions. The technique has found its application mainly in rainfall-runoff and flood inundation modeling (see e.g., Beven and Freer (2001) and Beven (2005b) for a complete list of references). In recent years, GLUE has also been applied in several groundwater studies (e.g., Feyen et al., 2001; Binley and Beven, 2003; Morse et al., 2003). Even though equifinality, as defined by Beven (1993; 2005a), arises because of the combined effects of errors in the forcing data, system conceptualization, measurements and parameter estimates, as yet, it has only been applied in the context of a single deterministic conceptual model (Refsgaard et al., 2006), thereby, neglecting model structural uncertainty.

In this work, we combine GLUE with BMA to explicitly account for uncertainty that originates from errors in the model conceptualization, forcing data (e.g., recharge rate, boundary conditions) and parameter values. Within the GLUE framework, we explore the global likelihood response surface of all possible combinations of plausible model structures, forcing data and parameter values in order to select those simulators that perform well. For each model structure, the posterior model probability is obtained by integrating the likelihood measures over the retained simulators for that model structure. The posterior model probabilities are subsequently used in BMA to weight the predictions of the competing models when assessing the joint predictive uncertainty.

The method presented is very flexible since (i) there is no restriction on the diversity of conceptual models or on the level of uncertainty in the forcing data or parameters that can be included; (ii) it allows for different ways of expressing the likelihood of a simulator (including a formal Bayesian one) based on the distribution of the residuals, hence allowing different types of knowledge to be incorporated (quantitative as well as qualitative); and (iii) it is Bayesian in nature, which provides a formal framework to incorporate previous knowledge about the model structures and parameters, or to update the estimates should new information become available. The main drawback of the methodology is the computational burden. Due to the presence of multiple local optima in the global likelihood response surface, good performing or behavioral simulators might be well distributed across the hyperspace dimensioned by the set of model structures, input and parameter vectors. This necessitates that the global likelihood surface is extensively sampled.

The remainder of this paper is organized as follows. In section 2, we provide a condensed overview of the GLUE and BMA methodologies, followed by a description of the procedure to integrate both methods. Section 3 details a three-dimensional hypothetical setup that is used to illustrate the integrated uncertainty assessment methodology. Implementation details are described in section 4. In this section we elaborate on the different conceptualizations as well as on input and parameter uncertainty. Results are discussed in section 5 and a summary of conclusions is presented in section 6.

2. Methodology for integrated uncertainty assessment

2.1. Generalized Likelihood Uncertainty Estimation (GLUE) methodology

GLUE is a Bayesian Monte Carlo simulation technique based on the concept of equifinality (Beven and Binley, 1992; Beven and Freer, 2001). It rejects the idea of a single correct representation of the system in favor of many acceptable or behavioral system representations that should be considered in the evaluation of uncertainty associated with predictions (Beven, 2005b). For each simulator sampled from a prior set of possible system representations a likelihood measure is calculated that reflects the ability of the simulator to simulate the system responses, given the available training data. Simulators that perform below a rejection criterion are discarded from the further analysis and the likelihood measures of retained simulators are rescaled so as to render the cumulative likelihood equal to one. Ensemble predictions are based on the predictions of the retained set of simulators, weighted by their respective rescaled likelihood.

The likelihood or "goodness of fit" used in GLUE must be seen in a much wider sense than the formal likelihood functions used in traditional statistical estimation theory. The likelihoods used in GLUE are a measure of the ability (performance) 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 being the correct representation of the system (Binley and Beven, 2003). However, the GLUE methodology is fully coherent with a formal Bayesian approach when the use of a classical likelihood function is justifiable based on the nature of the residuals (see e.g., Romanowicz et al., 1994).

Some critiques have recently been raised concerning the subjective nature of some decisions that have to be made in order to implement the GLUE methodology (see e.g., Mantovan and Todini (2006) and the reply of Beven et al., (2007)). These subjective decisions involve the definition of a suitable likelihood function and the definition of the rejection level in order to distinguish between "behavioral" and "non-behavioral" simulators. To evaluate the impact of these subjective decisions in the analysis, we implement three different likelihood functions in this study, namely, a formal statistical, a GLUE type, and a Fuzzy type measure.

Let us consider a set of plausible model structures $\mathbf{M}=\left\{\mathrm{M}_1,\mathrm{M}_2,...,\mathrm{M}_k,...,\mathrm{M}_K \mid K<\infty\right\}$, a set of parameter vectors $\boldsymbol{\Theta}=\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,...,\boldsymbol{\theta}_l,...,\boldsymbol{\theta}_L\right)$ and a set of input variable vectors $\boldsymbol{\Upsilon}=\left(\mathbf{Y}_1,\mathbf{Y}_2,...,\mathbf{Y}_m,...,\mathbf{Y}_M\right)$, and denote the observed and simulated system variable vectors as $\mathbf{D}=\left(\mathrm{D}_1,\mathrm{D}_2,...,\mathrm{D}_n,...,\mathrm{D}_N\right)$ and $\mathbf{D}^*=\left(\mathrm{D}_1^*,\mathrm{D}_2^*,...,\mathrm{D}_n^*,...,\mathrm{D}_N^*\right)$, respectively. Then,

$L\left(\mathrm{M}_k,\boldsymbol{\theta}_l,\mathbf{Y}_m\middle|\mathbf{D}\right)$ represents the likelihood of the *k*th model structure parameterized with parameter vector $\boldsymbol{\theta}_l$ and forced by input data vector $\mathbf{Y}_m$ to represent the true system, given the observations in $\mathbf{D}$.

As a first likelihood measure, we consider a Gaussian likelihood function (1), which is based on the assumption that the residuals follow a normal distribution centered on zero. For a given number of observations, *N*, the Gaussian likelihood is given by

$$L\left(\mathrm{M}_k,\boldsymbol{\theta}_l,\mathbf{Y}_m\middle|\mathbf{D}\right)=(2\pi)^{-N/2}\left|C_{\mathbf{D}}\right|^{-1/2}\exp\left(-\frac{1}{2}\left(\mathbf{D}-\mathbf{D}^*\right)^T C_{\mathbf{D}}^{-1}\left(\mathbf{D}-\mathbf{D}^*\right)\right) \tag{1}$$

where, $C_{\mathbf{D}}$ is the covariance matrix of the observed system variables.

The second measure implemented is the model efficiency likelihood function (2) (Freer and Beven, 1996; Feyen et al., 2001; Jensen, 2003), which is based on the Nash-Sutcliffe efficiency criterion (Nash and Sutcliffe, 1970) with shaping factor *S*, and is given by

$$L\left(\mathrm{M}_k,\boldsymbol{\theta}_l,\mathbf{Y}_m\middle|\mathbf{D}\right)=\left(1-{\sigma_\varepsilon^2}\big/{\sigma_{\mathbf{D}}^2}\right)^S \tag{2}$$

where σ_ε^2 and $\sigma_{\mathbf{D}}^2$ are the variance of the residuals and of the observations, respectively. We used a shaping factor *S* equal to one, in which case the model efficiency likelihood function is equivalent to the coefficient of determination (R^2).

As a third measure we implemented a triangular likelihood function (3), belonging to the so-called Fuzzy type measures (Jensen, 2003), given by

$$L_n\left(\mathrm{M}_k,\boldsymbol{\theta}_l,\mathbf{Y}_m\middle|\mathrm{D}_n\right)=\frac{\mathrm{D}_n^*-a}{b-a}I_{a,b}+\frac{c-\mathrm{D}_n^*}{c-b}I_{b,c} \tag{3}$$

where

$$I_{a,b}=\begin{cases}1 & \text{if } a<\mathrm{D}_n^*\le b\\ 0 & \text{otherwise}\end{cases}$$
$$I_{b,c}=\begin{cases}1 & \text{if } b<\mathrm{D}_n^*< c\\ 0 & \text{otherwise}\end{cases}$$

and the limits a and c define the tolerable error. The triangular likelihood function in (3) gives a point likelihood measure $L_n\left(\mathrm{M}_k,\boldsymbol{\theta}_l,\mathbf{Y}_m\middle|\mathrm{D}_n\right)$ for each observation data n. It was combined by a geometric mean inference function to obtain a global likelihood value $L\left(\mathrm{M}_k,\boldsymbol{\theta}_l,\mathbf{Y}_m\middle|\mathbf{D}\right)$.

2.2. Bayesian Model Averaging (BMA)

BMA provides a coherent framework for combining predictions from multiple competing conceptual models to provide a more realistic and reliable description of the total prediction uncertainty. It is a statistical procedure that infers 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.

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

$$p\left(\Delta\,|\,\mathbf{D}\right)=\sum_{k=1}^{K}p\left(\Delta\,|\,\mathbf{D},\mathrm{M}_k\right)p\left(\mathrm{M}_k\,|\,\mathbf{D}\right) \tag{4}$$

Equation (4) is an average of the predictive distributions of Δ under each model considered, $p(\Delta \mid \mathbf{D}, \mathrm{M}_k)$, weighted by its 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

$$p(\mathrm{M}_k \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \mathrm{M}_k)\, p(\mathrm{M}_k)}{\sum_{k'=1}^{K} p(\mathbf{D} \mid \mathrm{M}_{k'})\, p(\mathrm{M}_{k'})} \tag{5}$$

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

$$p(\mathbf{D} | \mathrm{M}_k) = \iint p(\mathbf{D} | \mathrm{M}_k, \boldsymbol{\theta}_l, \mathbf{Y}_m)\, p(\boldsymbol{\theta}_l, \mathbf{Y}_m | \mathrm{M}_k)\, d\boldsymbol{\theta}_l\, d\mathbf{Y}_m \tag{6}$$

where $p(\mathbf{D} | \mathrm{M}_k, \boldsymbol{\theta}_l, \mathbf{Y}_m)$ is the likelihood of model structure M_k parameterized with parameter vector $\boldsymbol{\theta}_l$ and forced by input data vector $\mathbf{Y}_m$ given the observations in $\mathbf{D}$, and $p(\boldsymbol{\theta}_l, \mathbf{Y}_m | \mathrm{M}_k)$ is the joint prior probability distribution of $(\boldsymbol{\theta}_l, \mathbf{Y}_m)$ given model M_k.

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

$$E[\Delta \mid \mathbf{D}] = \sum_{k=1}^{K} E[\Delta \mid \mathbf{D}, \mathrm{M}_k]\, p(\mathrm{M}_k \mid \mathbf{D}) \tag{7}$$

$$\begin{aligned} Var[\Delta \mid \mathbf{D}] = & \sum_{k=1}^{K} Var[\Delta \mid \mathbf{D}, \mathrm{M}_k]\, p(\mathrm{M}_k \mid \mathbf{D}) \\ & + \sum_{k=1}^{K} \left(E[\Delta \mid \mathbf{D}, \mathrm{M}_k] - E[\Delta \mid \mathbf{D}] \right)^2 p(\mathrm{M}_k \mid \mathbf{D}) \end{aligned} \tag{8}$$

In essence, the BMA prediction is the weighted average of predictions from a suite of alternative models, with the weights equal to the likelihood that a model represents the true unknown model. From equation (8) 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.

2.3. Combining GLUE and BMA

Combining the GLUE and BMA methods involves the following sequence of steps

1. On the basis of prior and expert knowledge about the site, a suite of alternative conceptualizations is proposed, following, for instance, the methodology proposed by Neuman and Wierenga (2003).
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 representations or 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 stable distributions for each of the conceptual models M_k.

8. The integrated likelihood of each conceptual model M_k (equation 6) is approximated by summing the likelihood weights of the retained simulators in subset A_k, or

$$p\left(\mathbf{D}\middle|\mathrm{M}_k\right) \approx \sum_{l,m \in A_k} L\left(\mathrm{M}_k, \boldsymbol{\theta}_l, \mathbf{Y}_m \middle| \mathbf{D}\right) \tag{9}$$

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

$$p\left(\mathrm{M}_k\middle|\mathbf{D}\right) \approx \frac{\sum_{A_k} L\left(\mathrm{M}_k, \boldsymbol{\theta}_l, \mathbf{Y}_m \middle| \mathbf{D}\right) p\left(\mathrm{M}_k\right)}{\sum_{j=1}^{K} \sum_{l,m \in A_j} L\left(\mathrm{M}_j, \boldsymbol{\theta}_l, \mathbf{Y}_m \middle| \mathbf{D}\right) p\left(\mathrm{M}_j\right)} \tag{10}$$

10. After normalization of the likelihood weighted predictions under each individual model (such that the cumulative likelihood under each model equals one) a multi-model prediction is obtained with equation (4) using the weights obtained in (10). The leading moments of this distribution are obtained with equations (7) and (8).

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

3. Three-dimensional hypothetical case

For illustrative purposes, we employ a hypothetical setup for which the true conditions are known. The three-dimensional example system is similar to the reference case described in Poeter and Anderson (2005) and is presented in Figure 1. Lateral dimensions are 5000 m (E-W) by 3000 m (N-S) discretized 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 ln$K$ 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 ln$K$ for the different layers are presented in Table 1.

Simulation of steady-state flow employs 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 and pump a total of 2450 m^3 d^{-1} from the lowermost layer (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 altered considerable by the cones of depression around the pumping wells, local effects of spatially varying hydraulic conductivity and to a lesser extent by the evapotranspiration zone. 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

4.1. Alternative conceptual models

Theoretically, all possible models of relevance could be included in $\mathbf{M}$. However, the number of potentially feasible models may be exceedingly large, rendering their exhaustive inclusion in $\mathbf{M}$ infeasible (Hoeting et al., 1999). We adopt the idea of Ockham's Window (Madigan and Raftery, 1994) to consider a relatively small set of the most parsimonious models in view of the knowledge about the system and their ability to explain the data. As a consequence, the joint predictions do not represent all possibilities but only a limited range, conditional on the ensemble of conceptual models used to describe the groundwater system.

We consider the following seven conceptualizations with increasing complexity to describe the three-dimensional hypothetical setup presented in section 3: (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 three-dimensional 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 conceptualizations comprise a total aquifer thickness of 60 m and are forced by identical types of boundary conditions, although the magnitudes or rates of the latter are set variable (see 4.3).

4.2. Parameterization

The focus of this work is on the assessment of conceptual uncertainty. Therefore, we confine the dimensionality of the analysis 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 (see Table 2). Realizations of the hydraulic conductivity field of the different layers in the alternative conceptualizations are generated with the same mean K and spatial correlation law as the respective layers in the three-dimensional hypothetical setup (values listed in Table 1). For the 1Lhtg-AVG conceptualization the averages of these values are used. Although the nature of the underlying structure is assumed to be known, hence only the realization space is sampled, uncertainty in the mean hydraulic conductivity and spatial correlation function can be accounted for using, for example, the Bayesian methods presented in Feyen et al., (2002).

4.3. Prior distributions

We assign equal prior probabilities to the seven conceptualizations and adopt uniform prior distributions for the unknown inputs and parameters. The definition of such non-informative prior distributions is based on what is known as the principle of insufficient reason or the Bayes-Laplace postulate. According to this principle, in the absence of evidence to the contrary, all possibilities should have the same initial probability (Bernardo and Smith, 2000). Using these priors, we expect that the information in the data, expressed by the likelihood function, should dominate the form of the resulting posterior distribution. The ranges that describe the prior uniform distributions of the unknown variables are presented in Table 2.

4.4. Simulation

Parameter and input vectors, sampled from the prior distributions using a Latin Hypercube Sampling (LHS) scheme, are combined with hydraulic conductivity realizations for the respective layers and consequently evaluated under each conceptual model. On the basis of 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 the retained simulators, the performance is assessed using equations (1), (2) and (3). Point likelihood values estimated using equation (3) are combined into a global likelihood value using a geometric mean inference function (Jensen, 2003). For each conceptual model, predictive distributions of the state variables are obtained from the ensemble of likelihood weighted predictions (rescaled such that for each conceptual model the likelihoods sum up to one). Sampling from the prior input and parameter space continued until the first and second moment of these predictive distributions stabilized.

5. Results and discussion

Since it is impossible to show the complete set of results for all combinations of likelihood functions, variables, head observations, groundwater budget terms and alternative conceptualizations, in the following sections the most relevant results are summarized.

5.1. Convergence

For the conceptual models 1Lhtg-L1 and 1Lhtg-L2 none of the simulations were accepted, as all of them failed to meet the criteria of a maximum allowable departure of 5 m from the observed heads. Hence, no results are presented for these models since they are discarded from the posterior analysis.

Figure 2 shows, for the analysis based on the Gaussian likelihood function, the convergence of the mean of the predictive distribution for the following groundwater budget terms: west boundary condition (WBC) inflows (Figure 2a), recharge inflows (Figure 2b), west boundary condition (WBC) outflows (Figure 2c), river gains (Figure 2d), and evapotranspiration (EVT) outflows (Figure 2e). For all variables, convergence of the first moment was achieved in less than 10,000 retained simulations, whereby groundwater budget terms converged to different values in function of the conceptual model. Convergence of the second moment of the predictive distributions, although not shown here, was also achieved within less than 10,000 retained simulations. It is important to note that the second moments converged to smaller values when the alternative conceptual model approaches the true three-dimensional hypothetical setup. These findings support the idea that predicting state variables relying on a single conceptual model is prone to statistical bias and may produce an overconfident estimation of predictive uncertainty. Similar patterns of convergence of the first and second moments were observed for the other likelihood functions.

5.2. Likelihood response surfaces

Figure 3 shows the global likelihood response surface projected in one dimension for the six unknown variables (see Table 2). The vertical dashed lines represent the true values used in the three-dimensional hypothetical setup. It is seen from this figure that model performance is highly sensitive to variables RECH and CH, as expressed by the well defined regions of attraction centered on the true values (Figures 3a and 3b). For the other variables, well performing simulators are found across the whole prior space. However, for EVTR and SURF, zones of higher attraction are distinguished near their respective true values.

Figure 4 shows the normalized global likelihood response surface projected in two dimensions for different combinations of normalized variables for model conceptualizations 1Lhtg-AVG, 2Lhtg, and 3Lhtg. For each combination the highest five normalized likelihood values, which nearly all have a normalized likelihood larger than 0.95, are indicated by the numbered white crosses. The two-dimensional projections reveal the complex nature of the global likelihood response surface, with multiple localized zones of attraction and maximum likelihood values located in different regions of the joint input and parameter space. This reaffirms the idea of equifinality, i.e., that there exist multiple acceptable or behavioral simulators that perform equally well and that can be spread over large regions of the model, input and parameter space.

The plots in Figure 4 show that for increasing model complexity the regions of attractions become more pronounced, or less diffuse, especially for the parameters to which model performance is most sensitive. Plates a, b and c indicate that the two most sensitive parameters (RECH and CH) are inversely correlated, with a tendency to a more defined relationship with increasing model complexity. The other parameters do not show any strong correlation, mainly due to the low sensitivity of the model performance to these parameters.

Results for the other likelihood functions, although not shown here, are very similar to those presented in Figures 3 and 4. Hence, for the problem at hand, it can be concluded that the shape of the likelihood response surface does not depend on the choice of likelihood function.

Figure 5 shows a one-dimensional projection of the likelihood response surface against the model output variable river gains for three alternative conceptualizations (1Lhtg-AVG, 2Lhtg and 3Lhtg) for the three likelihood functions used. It represents the weights (y-axis) that are given to the different simulated values of river gain (x-axis) in the ensemble predictive distribution of each model. As stated before, it is clear that the choice of likelihood does not significantly impact the results. Increasing model complexity, on the other hand, results in a slight increase of the maximum likelihood values, reduces the diffusivity of the likelihood response surface and, for most groundwater budget terms, results in a more correct estimate of the true values. Hence, although simpler models may result in simulations that are nearly as good as the more complex models in terms of reproducing the set of head observations in the training period, they typically lead to more bias and a larger predictive spread.

5.3. Posterior model probabilities

The integrated likelihoods and the posterior model probabilities of each alternative conceptual model, approximated using equations (9) and (10), are presented in Table 3. The posterior model probabilities represent the ability of each of the alternative models to reproduce the observed data in the training period.

As previously stated, models 1Lhtg-L1 and 1Lhtg-L2 produced no results as none of the simulations were able to meet the acceptance criteria. Hence, their integrated model likelihood was set to zero and they were discarded in the calculation of the model ensemble predictive distribution. In Table 3 it is seen that posterior model probability of the other alternative conceptual models increases slightly from 0.18 to 0.22 with increasing level of model complexity. The small difference in posterior model probability implies that, for the given setup, the head observations do not allow to make a further distinction in performance between the five retained conceptualizations. These results confirm that in real applications, where the true hydrological concept is unknown and conditioning data are typically limited to (a sparse set of) head observations, confining the model space to a single model is often not supported by the data, hence advocate the idea of considering multiple conceptualizations. To overcome this problem, other sources of qualitative or quantitative conditioning data that allow a further differentiation between the models may be considered.

5.4. Predictive distributions

The posterior model probabilities are then used to combine the predictive distributions of the five retained conceptual models using equation (4). The moments of the multi-model ensemble predictive distribution are obtained through equations (7) and (8).

The cumulative predictive distributions of the groundwater budget terms for the five retained conceptual models and the Bayesian model averaging are presented in Figure 6 for the analysis based on the Gaussian likelihood function. The vertical dashed lines indicate the true values observed from the three-dimensional hypothetical setup. Summarizing statistics of the respective predictive distributions in Figure 6 are presented in Figure 7. Shown here for the groundwater budget terms are the min, max and median values, as well as the inter-quartile confidence intervals. Here the true values are represented by the horizontal lines.

Results for the individual models show that the predictive distributions of the budget terms vary substantially in shape, central moment and spread between the different conceptualizations. In general, it is seen that when the alternative conceptual model approaches the true three-dimensional hypothetical setup, confidence intervals become smaller and predictions are less biased, i.e., the median of the predictive distribution more closely reproduces the observed value of the budget terms. These results show that although the posterior model probabilities of the retained models differ only marginally their predictions can vary substantially. Whereas for some of the (mainly simpler) models the true groundwater budget values are not contained by the inter-quartile ranges of the predictions, they are always captured by the inter-quartile range of the BMA ensemble predictions. This reaffirms that relying on a single conceptual model is prone to statistical bias and may produce an overconfident estimation of predictive uncertainty. The BMA on the other hand provides consensus predictions and yields a more reliable estimation of the predictive uncertainty.

The contribution of model uncertainty to the total predictive uncertainty is estimated using equation (8) and is presented in Figure 8. Here, the total predictive variance is divided in within-model and between-model variance for the five groundwater budget terms. Both components are expressed as a percentage of the total variance. It is seen from this figure that predictive variance due to the uncertainty in the conceptual model (between-model) ranges from 5% for WBC outflows to approximately 30% for river gains, with practically no difference between the results for the different likelihood functions. Information about the susceptibility to conceptual model uncertainty of the different groundwater budget terms provides useful information for possible improvement of the model concept or to guide further data collection campaigns to optimally reduce conceptual uncertainty.

6. Conclusions

We presented a methodology to assess uncertainty in predictions of groundwater models arising from errors in the model structure, forcing data and parameter estimates. 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 integration over the input and parameter space the likelihood of the different simulators. The integrated likelihoods are consequently used in Bayesian model averaging to weight the model predictions to obtain ensemble predictions.

The adopted approach is flexible in the sense that (i) there is no limitation in the number or complexity of conceptual models that can be included, or to what degree input and parameter uncertainty can be incorporated, (ii) any quantitative or qualitative (e.g., pumping well never dries out) information about the system can be used to distinguish between different simulators, (iii) the closeness between the predictions and system observations can be defined in a variety of ways, including a formal statistical measure, and (iv) likelihoods, model probabilities and predictive distributions can be easily updated when new information becomes available. The major drawback of the approach is the computational burden inherent to any Monte Carlo method.

For illustrative purposes the methodology was applied to a three-dimensional hypothetical setup consisting of two aquifers separated by an aquitard, in which the flow field was considerably affected by pumping wells and spatially variable hydraulic conductivity. A set of 16 head observations sampled from this setup was used as conditioning data. The proximity of the simulations to these observations was evaluated using three different likelihood functions, including a formal statistical one. Seven alternative conceptualizations with increasing complexity were adopted and only uncertainty in parameters and inputs that were common to all conceptual models were considered. Two of the simpler one-layer models were discarded from the further analysis as they failed to meet a subjectively chosen criterion of closeness between the simulated and observed heads. For the other conceptualizations convergence of the first and second moment of the predicted variable distributions was achieved in less than 10,000 retained simulations.

The global likelihood response surface showed to be very complex, with multiple regions of high likelihood and local maxima in different regions of the joint model, input and parameter space. This confirms the concept of equifinality, i.e., that there exist many acceptable system representations that cannot be easily rejected and that should be considered in assessing the uncertainty associated with predictions. The likelihood response surfaces showed very little dependence on the choice of the likelihood function adopted. As such, the selection of the likelihood function did not have a significant impact on the further analysis and the general patterns observed in the results were identical for the three likelihood functions.

The integrated likelihoods of the five retained models increased slightly with increasing model complexity. The small differences in posterior model probability indicate that the set of 16 head observations did not allow a further discrimination between the five retained models. Nevertheless, predictive distributions of groundwater budget terms showed to be considerably different in shape, central moment and spread among the models. When the alternative conceptual model approached the true three-dimensional hypothetical setup, confidence intervals were in general smaller and predictions were less biased. BMA, on the other hand, provided consensus predictions yielding a more reliable estimation of the predictive uncertainty. The contribution of model uncertainty to the total predictive uncertainty varied between 5 to 30% depending on the groundwater budget term. The relative contribution of model uncertainty for the different groundwater budget terms provides useful information for updating the model concept or guiding data collection to optimally reduce conceptual uncertainty.

The results of this study strongly advocate the idea to address conceptual model uncertainty in the practice of groundwater modeling. With a hypothetical example it was shown that a set of head observations, which in reality may often be the only information available about the system dynamics, did not allow discriminating between a set of five models ranging from a simple one-layer model to a conceptualization approaching the true three-dimensional setup. Nevertheless, predictions of groundwater budget terms differed considerably among these models. The use of a single model may result in smaller uncertainty intervals, hence an increased confidence in the model simulations, but is very likely prone to statistical bias. Also, in the presence of conceptual model uncertainty, which per definition can not be excluded, this gain in accuracy in the short-term may have serious implications when using the model for long-term predictions in which the system is subject to new stresses. It is therefore advisable to explore a number of alternative conceptual models to obtain consensus predictions that are more conservative, hence that are more likely to bracket the true system responses.

It is expected that including other qualitative or quantitative sources of conditioning data, such as conductivity data, geological profiles, transient groundwater head information, or recharge estimates will allow a better differentiation between alternative models to further reduce model uncertainty. These topics will be subject of future research.

Acknowledgments

The first author wishes to thank the Katholieke Universiteit (K.U.) Leuven for providing financial support in the framework of PhD IRO-scholarships. We also wish to acknowledge the assistance provided by Jan de Laet, and Wim Obbels with running codes on the K.U. Leuven supercomputer (VIC CLUSTER) and by Jorge Gonzalez to implement the R scripts. Comments and suggestions by Marijke Huysmans and Okke Batelaan for improving the original manuscript are greatly appreciated.

References

Abrahart, R., and L. See (2002), Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments, Hydrology and Earth System Sciences, 6(4), 655-670.

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

Akaike, H. (1974), A new look at the statistical model identification, IEEE Transactions on Automatic Control AC 19, 716-723.

Bernardo, J. and A. Smith (2000), *Bayesian theory*, 608 pp., John Wiley & Sons Inc, Chichester.

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

Beven, K. (1993), Prophecy, reality and uncertainty in distributed hydrological modeling, Advances in Water Resources, 16(1), 41-51.

Beven, K. and J. Freer (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. (2005a), On the concept of model structural error, Water Science & Technology, 52(6), 167-175.

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

Beven, K., P. Smith and J. Freer (2007), Comment on "Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology" by Pietro Mantovan and Ezio Todino, Journal of Hydrology, doi: 10.1016/j.jhydrol.2007.02.023.

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

Bredehoeft, J. (2003), From models 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.

Burnham, K. and D. Anderson (1998), *Model selection and inference. A practical information-theoretic approach*, 353 pp., Springer-Verlag, New York.

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

Deutsch, C. and A. Journel (1998), *GSLIB Geostatistical software library and user's guide,* 2nd ed., Oxford University Press, New York.

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

Feyen, L., K. Beven, F. De Smedt and J. Freer (2001), Stochastic capture zone delineation within the GLUE-methodology: conditioning on head observations, Water Resources Research, 37(3), 625-638.

Feyen, L., P. Ribeiro, F. De Smedt, P. Diggle (2002), Bayesian methodology to stochastic capture zone determination: conditioning on transmissivity measurements, Water Resources Research, 38(9), doi:10.1029/2001WR000950.

Freer, J. and K. Beven (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, Water Resources Research, 32(7), 2161-2173.

Georgakakos, K., D. Seo, H. Gupta, J. Schaake and M. Butts (2004), Characterizing streamflow simulation uncertainty through multimodel ensembles, Journal of Hydrology, 298(1-4), 222-241.

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

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

Jensen, J. (2003), Parameter and uncertainty estimation in groundwater modelling, PhD thesis, 143 pp., Aalborg University, Aalborg, Dec 2002.

Kashyap, R. (1982), Optimal choice of AR and MA parts in autoregressive moving average models, IEEE Transactions Pattern Analysis and Machine Intelligence, PAMI 42(9), 99-104.

Madigan, D. and A. Raftery (1994), Model selection and accounting for model uncertainty in graphical models using Occam's window, Journal of the American Statistical Association, 89(428), 1535-1546.

Mantovan, P. and E. Todini (2006) Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, Journal of Hydrology, 330(1-2), 368-381.

Morse, G., G. Pohll, J. Huntington and R. Rodriguez (2003), Stochastic capture zone analysis of an arsenic-contaminated well using the generalized likelihood uncertainty estimator (GLUE) methodology, Water Resources Research, 39(6), 1151, doi:10.1029/2002WR001470

Nash, J. and J. Sutcliffe (1970) River flow forecasting through conceptual models. Part I – A discussion of principles, Journal of Hydrology, 10(3), 282-290.

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

Neuman, S. and P. Wierenga (2003), A comprehensive strategy of hydrogeologic modelling and uncertainty analysis for nuclear facilities and sites, Report NUREG/CR-6805, US Nuclear Regulatory Commission, Washington, USA.

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

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

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

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

Rubin, Y. (2003), *Applied stochastic hydrogeology*, 391 pp., Oxford University Press, New York.

Sun, N-Z. (1994), *Inverse problems in groundwater modelling. Theory and applications of transport in porous media*, 352 pp., Kluwer Academic Publishers, Netherlands.

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

Ye, M., S. Neuman, P. Meyer and K. Pohlmann (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.

Figure 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: Convergence of the first moment of the predictive distributions of the groundwater budget terms as a function of the Number of retained Monte Carlo Simulations (NMCS) for the Gaussian likelihood function (GAUSS): (a) west boundary condition (WBC) inflows, (b) recharge inflows, (c) west boundary condition (WBC) outflows, (d) river gains and (e) evapotranspiration (EVT) outflows.

Figure 3: One-dimensional projection of the global likelihood response surface (based on the Gaussian likelihood function) for the six parameters for conceptual model 3Lhtg. Vertical dashed lines represent the parameter values used in the three-dimensional hypothetical setup.

Figure 4: Two-dimensional projection of the normalized likelihood response surface (based on the Gaussian likelihood function) for the normalized parameters RECH vs. CH, RECH vs. EVTR and, RECH vs. RIVC for the alternative conceptual models 1Lhtg-AVG, 2Lhtg and 3Lhtg. Numbered crosses represent the locations of the five highest likelihood values.

Figure 5: Results for the river gains for the alternative conceptual models 1Lhtg-AVG (a-d-g), 2Lhtg (b-e-h) and 3Lhtg (c-f-i), and the Gaussian – GAUSS (a-c), Triangular – TRIANG (d-e) and Model efficiency – MODEFF (g-i) likelihood functions. Vertical dashed-lines represent the observed values from the three-dimensional hypothetical setup.

Figure 6: Cumulative probability distributions of the groundwater budget terms for the five alternative conceptual models and the Bayesian model averaging (BMA) based on the

Gaussian likelihood function: (a) west boundary condition (WBC) inflows, (b) recharge inflows, (c) west boundary condition (WBC) outflows, (d) river gains and (e) evapotranspiration (EVT) outflows. Vertical dashed-lines represent observed values from the three-dimensional hypothetical setup.

Figure 7: Total variance estimated using equation (8) for the groundwater budget terms based on the Gaussian (GAUSS), triangular (TRIANG) and model efficiency (MODEFF) likelihood function. From left to right: west boundary condition (WBC) inflows, recharge inflows, west boundary condition (WBC) outflows, river gains and evapotranspiration (EVT) outflows.

Figure 8: Summary statistics of the predictive distributions of the alternative conceptual models and multi-model BMA prediction for the groundwater budget terms: a) west boundary condition (WBC) inflows, (b) recharge inflows, (c) west boundary condition (WBC) outflows, (d) river gains and (e) evapotranspiration (EVT) outflows. Horizontal lines represent the values obtained from the three-dimensional hypothetical setup. Q_1 and Q_3 represent the first and third quartile, respectively.

Tables

Table 1: Parameters describing the hydraulic conductivity spatial correlation structure for the different layers of the three-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

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

Table 3: Summary of the integrated likelihood and posterior model probabilities for the alternative conceptual models

	Likelihood function	Conceptual model							Total
		1Lhtg-L1	1Lhtg-L2	1Lhtg-L3	1Lhtg-AVG	2Lhtg	2LQ3Dhtg	3Lhtg	
$p(\mathbf{D}\|M_k)$	GAUSS	0	0	902.6	935.6	990.4	1046.9	1079.4	4954.9
	TRIANG	0	0	4385.5	4608.1	4997.4	5365.2	5407.3	24763.5
	MODEFF	0	0	5952.6	6191.6	6579.3	6944.5	6994.7	32662.6
$p(M_k)$		1/7	1/7	1/7	1/7	1/7	1/7	1/7	1.0
$p(M_k\|\mathbf{D})$	GAUSS	0	0	0.1822	0.1888	0.1999	0.2113	0.2178	1.0
	TRIANG	0	0	0.1771	0.1861	0.2018	0.2167	0.2184	1.0
	MODEFF	0	0	0.1822	0.1896	0.2014	0.2126	0.2141	1.0

Figures

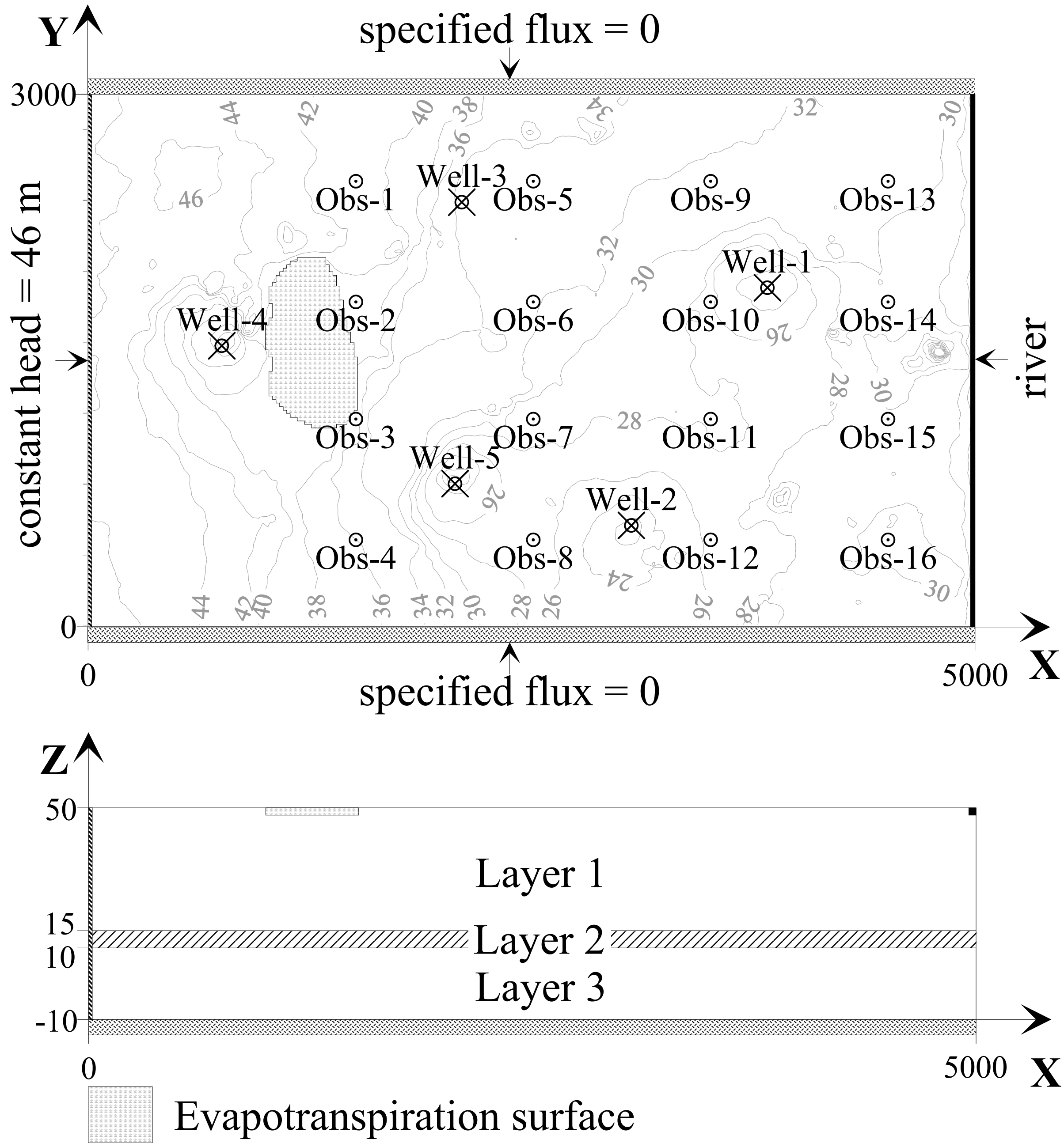

Figure 1: Three-dimensional hypothetical setup including (⊙) observation wells and (⊗) pumping wells overlain by the groundwater head distribution in the first layer.

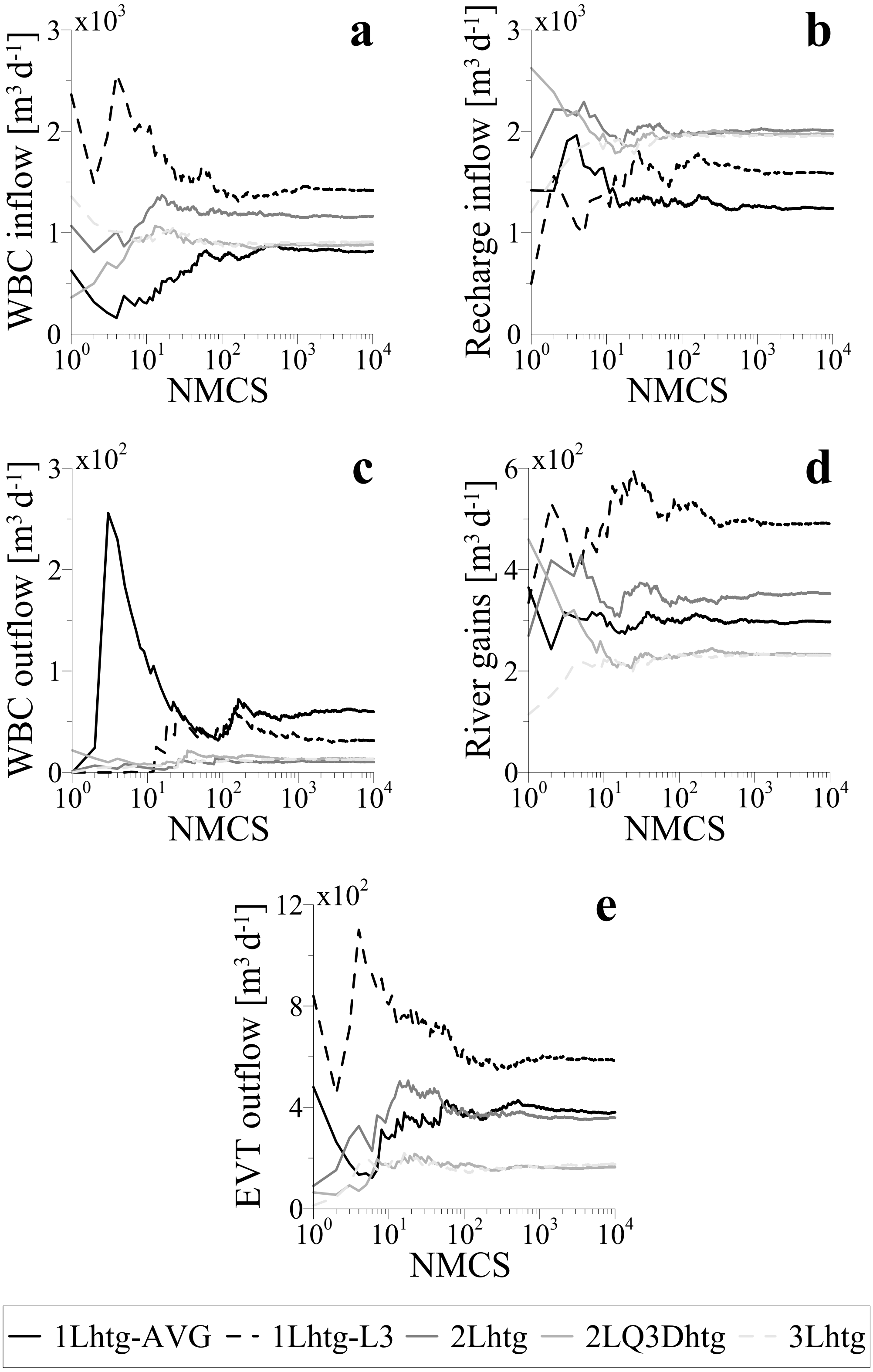

Figure 2: Convergence of the first moment of the predictive distributions of the groundwater budget terms as a function of the Number of retained Monte Carlo Simulations (NMCS) for the Gaussian likelihood function (GAUSS): (a) west boundary condition (WBC) inflows, (b) recharge inflows, (c) west boundary condition (WBC) outflows, (d) river gains and (e) evapotranspiration (EVT) outflows.

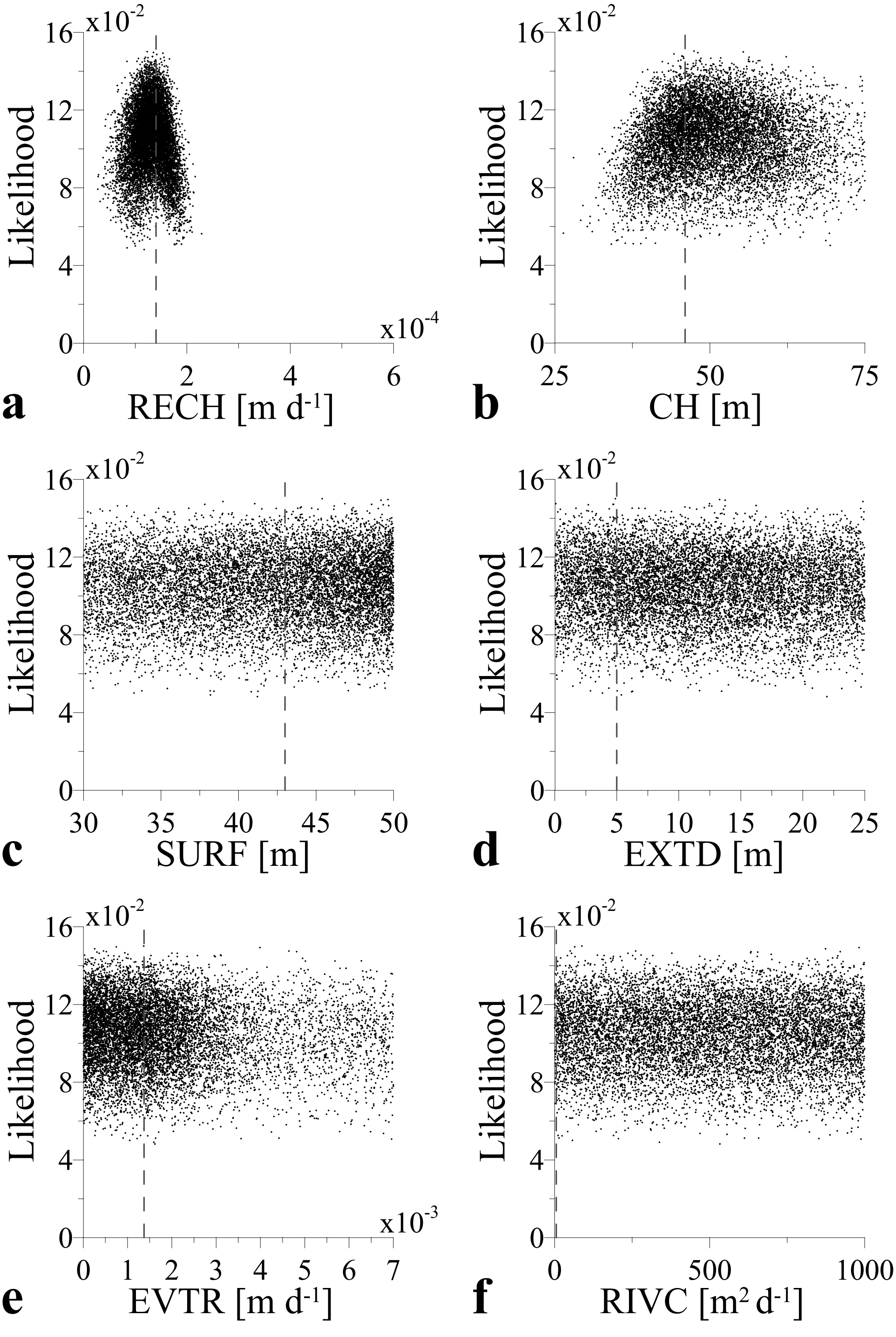

Figure 3: One-dimensional projection of the global likelihood response surface (based on the Gaussian likelihood function) for the six parameters for conceptual model 3Lhtg. Vertical dashed lines represent the parameter values used in the three-dimensional hypothetical setup.

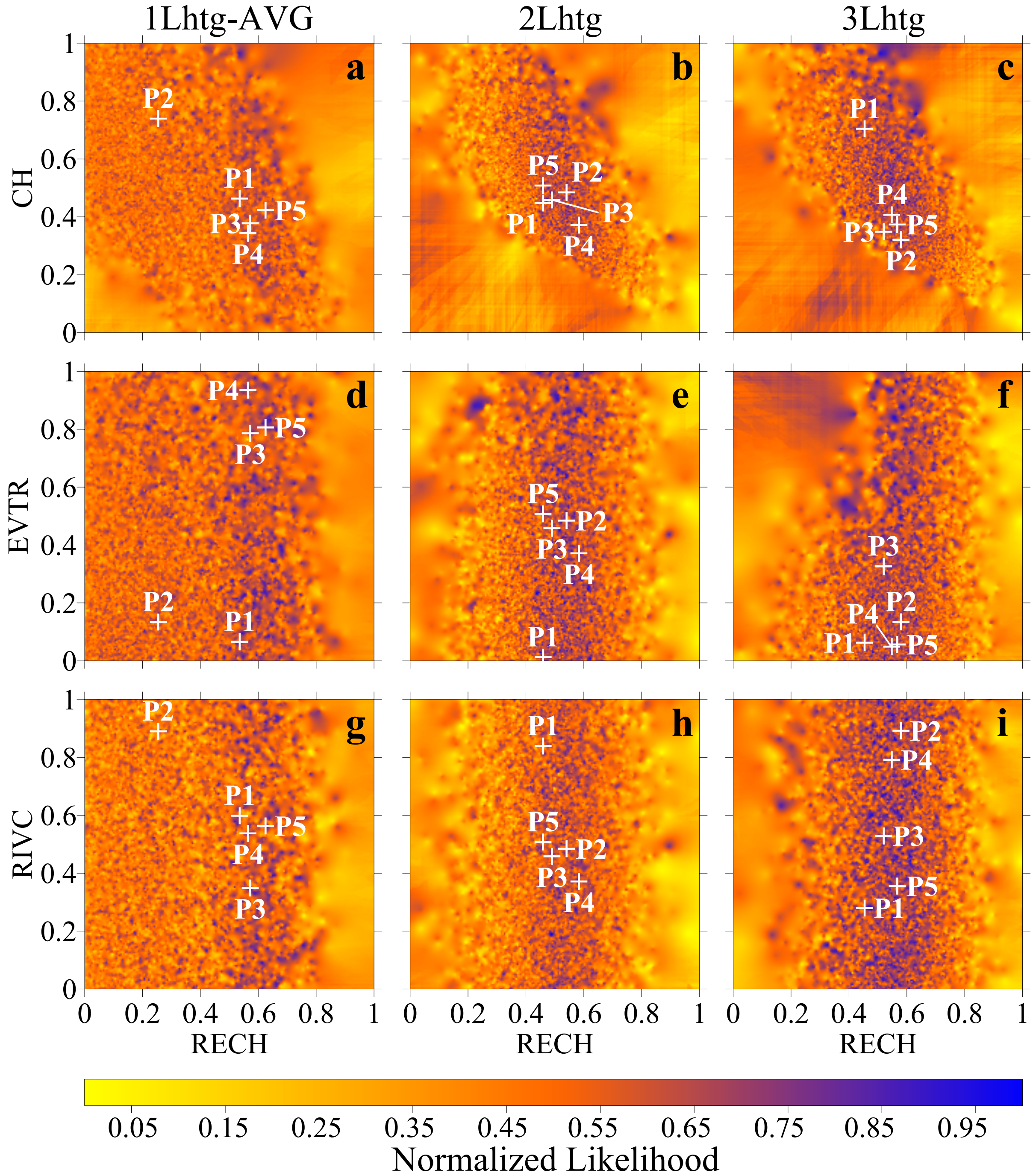

Figure 4: Two-dimensional projection of the normalized likelihood response surface (based on the Gaussian likelihood function) for the normalized parameters RECH vs. CH, RECH vs. EVTR and, RECH vs. RIVC for the alternative conceptual models 1Lhtg-AVG, 2Lhtg and 3Lhtg. Numbered crosses represent the locations of the five highest likelihood values.

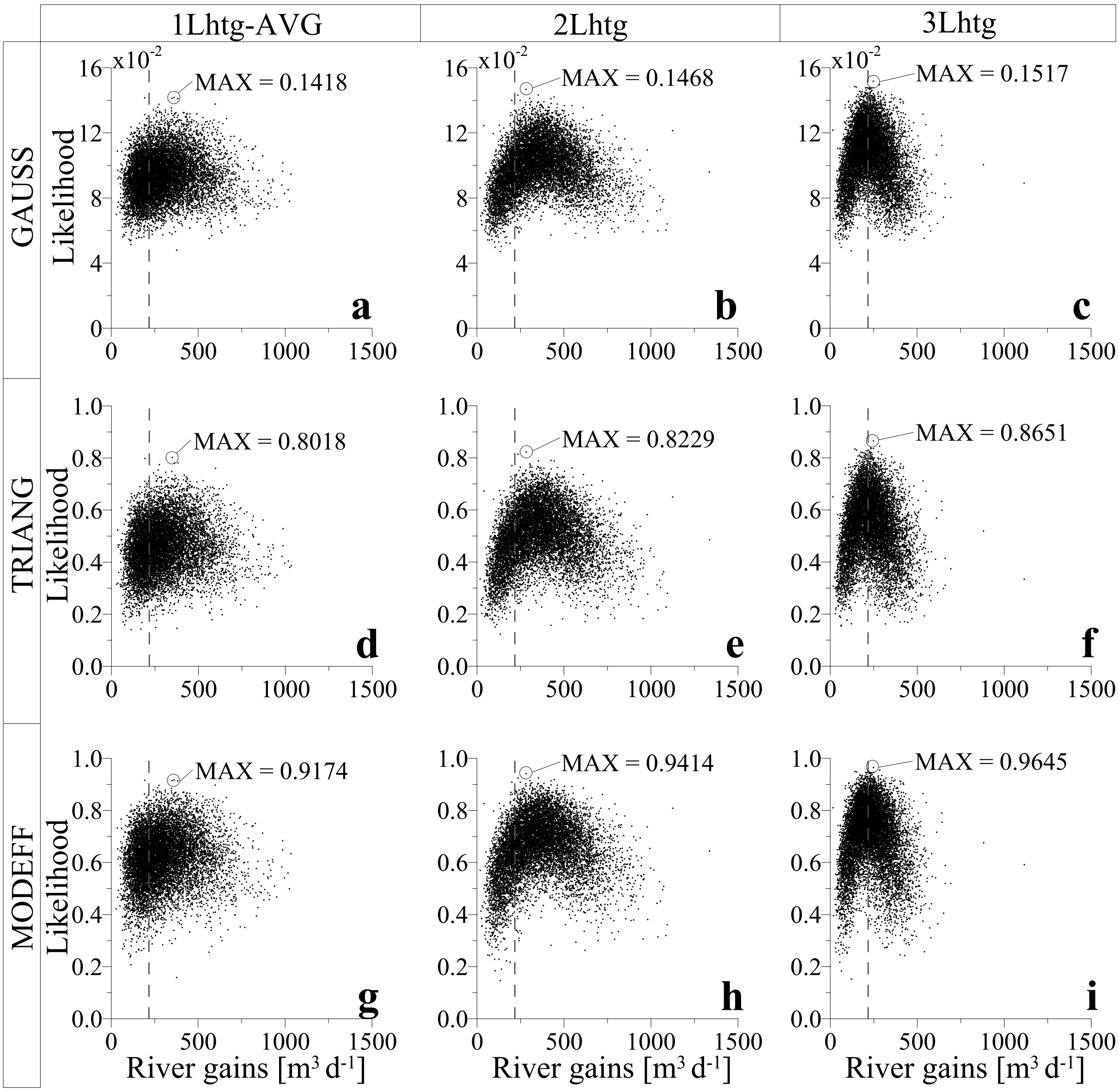

Figure 5: Results for the river gains for the alternative conceptual models 1Lhtg-AVG (a-d-g), 2Lhtg (b-e-h) and 3Lhtg (c-f-i), and the Gaussian – GAUSS (a-c), Triangular – TRIANG (d-e) and Model efficiency – MODEFF (g-i) likelihood functions. Vertical dashed-lines represent the observed values from the three-dimensional hypothetical setup.

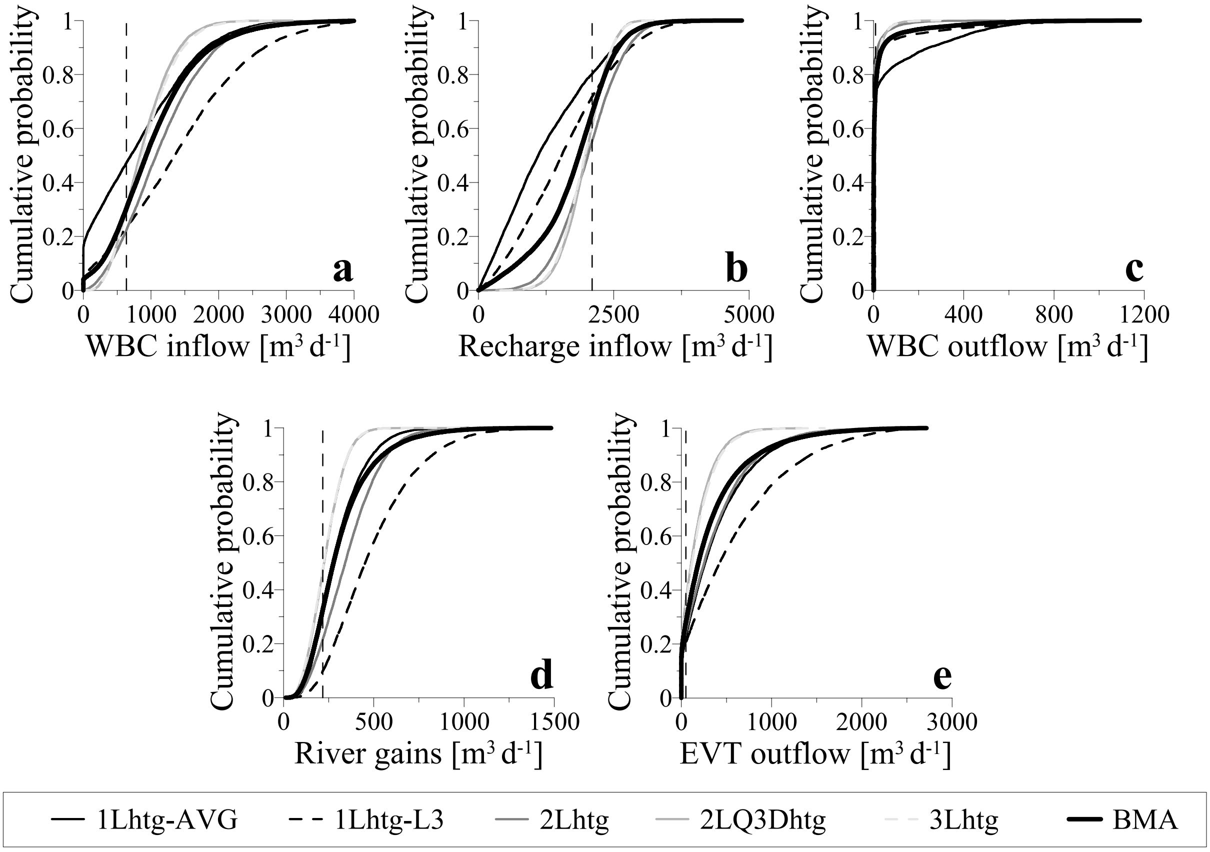

Figure 6: Cumulative probability distributions of the groundwater budget terms for the five alternative conceptual models and the Bayesian model averaging (BMA) based on the Gaussian likelihood function: (a) west boundary condition (WBC) inflows, (b) recharge inflows, (c) west boundary condition (WBC) outflows, (d) river gains and (e) evapotranspiration (EVT) outflows. Vertical dashed-lines represent observed values from the three-dimensional hypothetical setup.

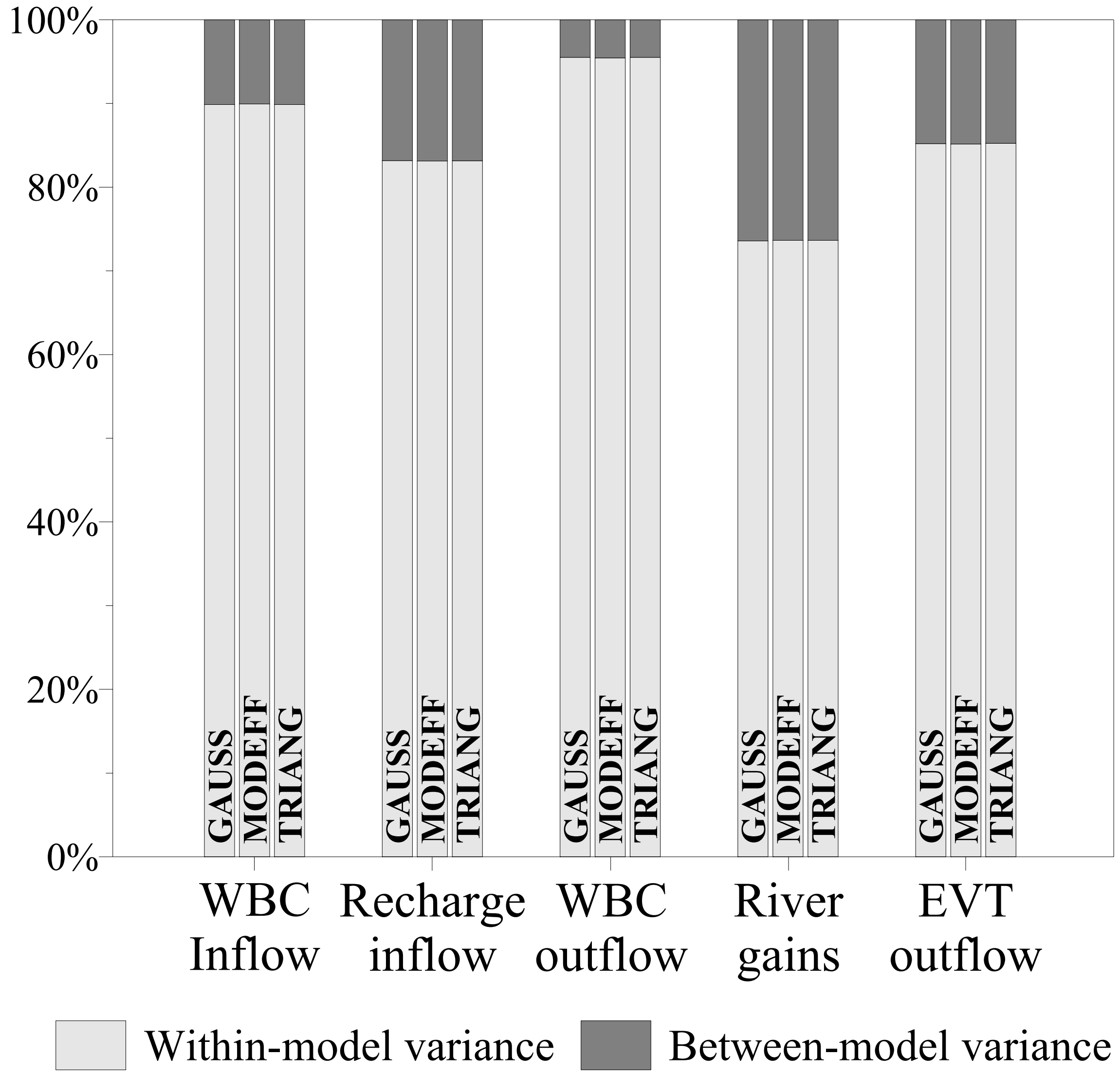

Figure 7: Total variance estimated using equation (8) for the groundwater budget terms based on the Gaussian (GAUSS), triangular (TRIANG) and model efficiency (MODEFF) likelihood function. From left to right: west boundary condition (WBC) inflows, recharge inflows, west boundary condition (WBC) outflows, river gains and evapotranspiration (EVT) outflows.

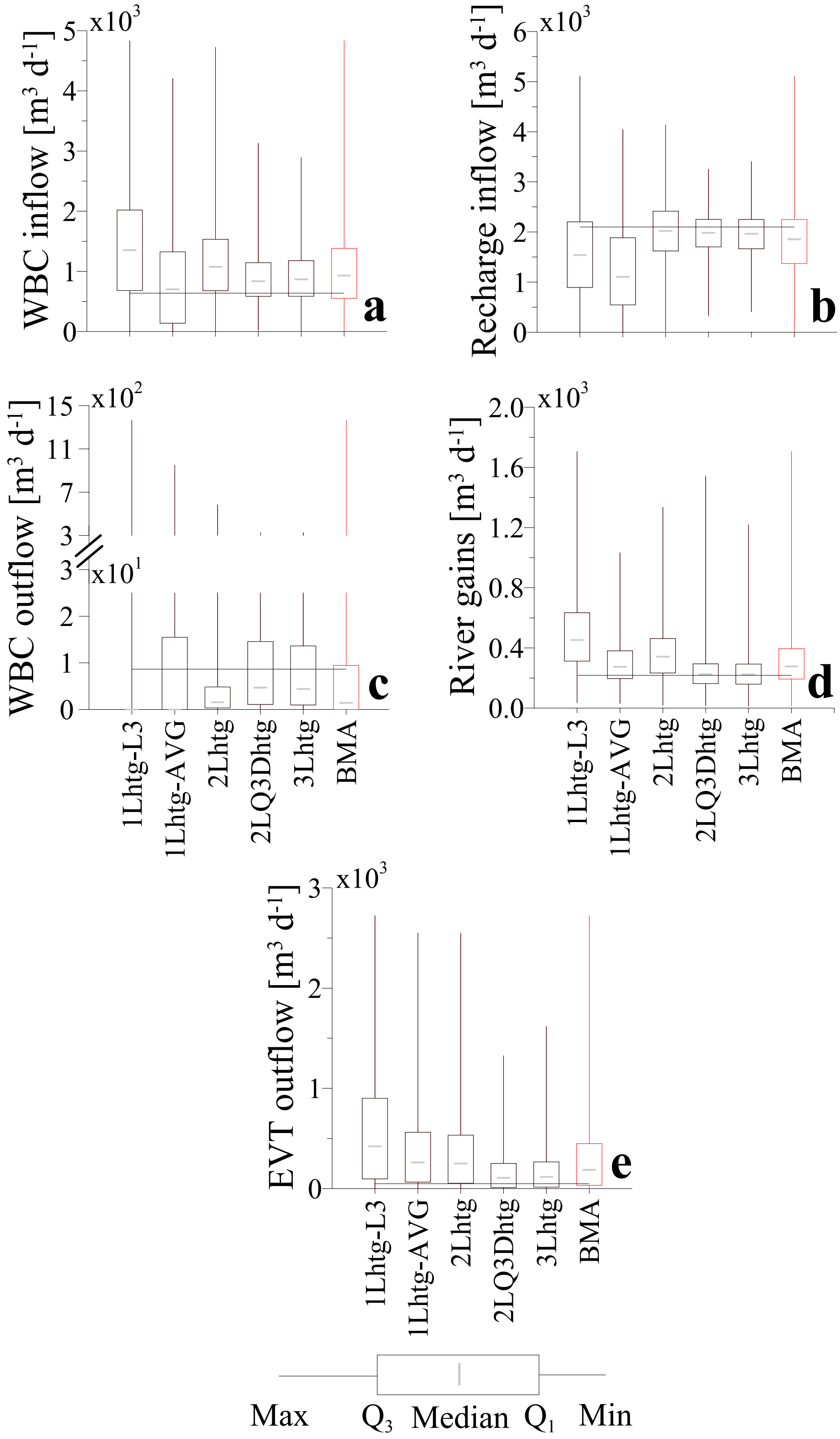

Figure 8: Summary statistics of the predictive distributions of the alternative conceptual models and multi-model BMA prediction for the groundwater budget terms: a) west boundary condition (WBC) inflows, (b) recharge inflows, (c) west boundary condition (WBC) outflows, (d) river gains and (e) evapotranspiration (EVT) outflows. Horizontal lines represent the values obtained from the three-dimensional hypothetical setup. Q_1 and Q_3 represent the first and third quartile, respectively.