On the value of conditioning data to reduce conceptual model uncertainty in groundwater modeling

R. Rojas,[1] L. Feyen,[2] O. Batelaan,[1,3] and A. Dassargues[1,4]

O. Batelaan, Department of Hydrology and Hydraulic Engineering, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium.

A. Dassargues, Hydrogeology and Environmental Geology, Department of Architecture, Geology, Environment, and Constructions (ArGEnCo), Université de Liège, B.52/3 Sart-Tilman, B-4000 Liège, Belgium.

L. Feyen, Land management and natural hazards unit, Institute for Environment and Sustainability (IES), Joint Research Centre (JRC), European Commission (EC), Via E. Fermi 2749, TP261, I-21027, Ispra (Va), Italy.

R. Rojas, Applied geology and mineralogy, Department of Earth and Environmental Sciences, Katholieke Universiteit Leuven, Celestijnenlaan 200E, B-3001 Heverlee, Belgium. (Rodrigo.RojasMujica@geo.kuleuven.be).

Now at: Land management and natural hazards unit, Institute for Environment and Sustainability (IES), Joint Research Centre (JRC), European Commission (EC), Via E. Fermi 2749, TP261, I-21027, Ispra (Va), Italy. (Rodrigo.Rojas@jrc.ec.europa.eu)

[1]Applied geology and mineralogy,

Department of Earth and Environmental

Abstract. Recent applications of multi-model methods have demonstrated their potential in quantifying conceptual model uncertainty in groundwater modeling applications. To date, however, little is known about the value of conditioning to constrain the ensemble of conceptualizations, to differentiate among retained alternative conceptualizations, and to reduce conceptual model uncertainty. We address these questions by conditioning multi-model simulations on measurements of hydraulic conductivity and observations of

Sciences, Katholieke Universiteit Leuven, Leuven, Belgium.

[2]Land management and natural hazards unit, Institute for Environment and Sustainability (IES), Joint Research Centre (JRC), European Commission (EC), Ispra, Italy.

[3]Department of Hydrology and Hydraulic Engineering, Vrije Universiteit Brussel, Brussels, Belgium.

[4]Hydrogeology and Environmental Geology, Department of Architecture, Geology, Environment, and Constructions (ArGEnCo), Université de Liège, Liège, Belgium.

system-state variables and evaluating the effects on (i) the posterior multi-model statistics and (ii) the contribution of conceptual model uncertainty to the predictive uncertainty. Multi-model aggregation and conditioning is performed by combining the generalized likelihood uncertainty estimation (GLUE) method and Bayesian model averaging (BMA). As an illustrative example we employ a 3-dimensional hypothetical system under steady-state conditions, for which uncertainty about the conceptualization is expressed by an ensemble ($\mathbf{M}$) of 7 models with varying complexity. Results show that conditioning on heads allowed for the exclusion of the two simplest models, but that their information content is limited to further differentiate among the retained conceptualizations. Conditioning on increasing numbers of conductivity measurements allowed for a further refinement of the ensemble $\mathbf{M}$ and resulted in an increased precision and accuracy of the multi-model predictions. For some groundwater flow components not included as conditioning data, however, the gain in accuracy and precision was partially offset by strongly deviating predictions of a single conceptualization. Identifying the conceptualization producing the most deviating predictions may guide data collection campaigns aimed at acquiring data to further eliminate such conceptualizations. Including groundwater flow and river discharge observations further allowed for a better differentiation among alternative conceptualizations and drastic reductions of the predictive variances. Results strongly advocate the use of observations less commonly available than groundwater heads to reduce conceptual model uncertainty in groundwater modeling.

1. Introduction and Scope

Groundwater modeling is a key component of sustainable groundwater management. Reliable and accurate model predictions are therefore needed to ensure an acceptable level of confidence in the model results. It has recently been suggested that predictive uncertainty in groundwater modeling is largely dominated by uncertainties arising from the definition of alternative conceptual models and that parametric uncertainty solely does not compensate for conceptual model uncertainty [*Neuman*, 2003; *Ye et al.*, 2004; *Bredehoeft*, 2005; *Højberg and Refsgaard*, 2005; *Poeter and Anderson*, 2005; *Refsgaard et al.*, 2006; *Meyer et al.*, 2007; *Rojas et al.*, 2008a; *Seifert et al.*, 2008].

The debate as to whether or not postulate simplified or complex/elaborated models to explain a groundwater system [see e.g. *Neuman and Wierenga*, 2003; *Gómez-Hernández*, 2006; *Hill*, 2006; *Hill and Tiedeman*, 2007; *Hunt et al.*, 2007; *Renard*, 2007], the advances in computational power, as well as the increasing awareness among scientists to address uncertainty in model predictions [see e.g. *Walker and Marchau*, 2003; *Refsgaard et al.*, 2005; *Van der Sluijs*, 2005; *Pappenberger and Beven*, 2006; *Refsgaard et al.*, 2007] have stimulated a growing tendency of postulating alternative conceptualizations [e.g. *Harrar et al.*, 2003; *Meyer et al.*, 2004; *Højberg and Refsgaard*, 2005; *Meyer et al.*, 2007; *Troldborg et al.*, 2007; *Rojas et al.*, 2008a; *Seifert et al.*, 2008; *Ijiri et al.*, 2009; *Rojas et al.*, 2008b, 2010]. Rather than relying on a single conceptual model, it seems more appropriate to consider a range of plausible system representations and analyze the combined multi-model output to assess the predictive modeling uncertainty. Whereas uncertainty estimations based on a single conceptualization are more likely to be biased and under-

dispersive, uncertainty estimations based on an ensemble of models are less (artificially) conservative and are more likely to capture the unknown true predicted value [*Neuman*, 2003; *Rojas et al.*, 2008a].

Several multi-model methods have recently been proposed. They seek to obtain an average prediction from a set of plausible conceptual models by linearly combining individual model predictions. The weights to aggregate multiple model outputs can be equal (model average) in the simplest case, can be determined through regression-based approaches [e.g. *Abrahart and See*, 2002; *Georgakakos et al.*, 2004], or can be linked to model performance [e.g. *Neuman*, 2003; *Poeter and Anderson*, 2005; *Refsgaard et al.*, 2006; *Ajami et al.*, 2007; *Rojas et al.*, 2008a]. Amongst multi-model methods based on model performance, the classical idea of Bayesian Model Averaging (BMA) [*Leamer*, 1978; *Box*, 1980; *Draper*, 1995; *Kass and Raftery*, 1995; *Hoeting et al.*, 1999] has recently gained popularity [e.g. *Neuman*, 2003; *Ajami et al.*, 2005; *Vrugt et al.*, 2006; *Ajami et al.*, 2007; *Duan et al.*, 2007; *Vrugt and Robinson*, 2007; *Ajami et al.*, 2008; *Tsai and Li*, 2008; *Wöhling and Vrugt*, 2008; *Hsu et al.*, 2009; *Li and Tsai*, 2009; *Singh et al.*, 2009; *Ye et al.*, 2009]. In short, BMA weights the predictions of competing models by their corresponding posterior model probability, representing each (conceptual) model's relative skill to reproduce system behavior in the observation period. 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 Zhang*, 2003; *Ye et al.*, 2004; *Ajami et al.*, 2005].

Despite the development of these multi-model techniques, performing the aggregation of multiple model predictions or assessing uncertainty arising from the definition of alter-

native conceptualizations is not common practice in groundwater flow or solute transport modeling [see e.g. *Sohn et al.*, 2000; *Harrar et al.*, 2003; *Højberg and Refsgaard*, 2005; *Troldborg et al.*, 2007; *Seifert et al.*, 2008; *Ijiri et al.*, 2009]. Rather, modelers tend to limit themselves to use the "best" conceptual model available, even though observations may be reproduced equally well by more than one conceptual model. Constraints on available resources certainly limit the use of alternative conceptual models for prediction purposes, thus, promoting using the "best" available conceptualization. However, it is amply recognized that new data may have an impact on the conceptual understanding and that the selected conceptualization should be open for improvement as a result. In addition, the (mis)perception of technical/computational limitations among practitioners as well as the existence of administrative/economical constraints hamper the implementation of uncertainty analyses [*Bredehoeft*, 2003, 2005; *Renard*, 2007]. *Pappenberger and Beven* [2006] discuss in a more general context reasons that justify avoiding uncertainty analysis in mechanistic environmental modeling.

The ultimate goal of an uncertainty analysis is to quantify the degree of confidence in the model results given the uncertainties involved in the modeling task [*Cacuci*, 2003; *Saltelli et al.*, 2008; *Hill and Tiedeman*, 2007]. Ideally, the quantification of the (total) predictive uncertainty including the contributions originating from different sources (e.g. forcing data, parameters and conceptual models) should be assessed. This information allows the analyst to focus on possible strategies to increase her/his confidence in the model results, i.e. to decrease predictive uncertainty.

One strategy to decrease the uncertainty in groundwater model predictions is to reproduce (or honor) measurements of spatially distributed key parameters. It is well-known

that the spatial distribution of hydraulic conductivity forms a large source of uncertainty in groundwater modeling [see e.g. *Freeze*, 1975; *Dagan*, 1989; *Gelhar*, 1993; *Dagan and Neuman*, 1997; *Rubin*, 2003; *Moore and Doherty*, 2005]. Therefore, methods aimed at obtaining conditional realizations of hydraulic conductivity (or transmissivity) fields to reduce model and prediction uncertainty arising from this source are abundant and well documented in the literature. Some of these approaches are gradient-based inverse techniques [e.g. *Carrera and Neuman*, 1986a, b, c; *Tiedeman et al.*, 1997, 1998, 2003, 2004; *Foglia et al.*, 2009], others use direct measurements of hydraulic conductivity or transmissivity to generate conditional realizations of the K- or T-field [e.g. *Delhomme*, 1979; *Hill et al.*, 1998; *Moore and Doherty*, 2005], some use linearized stochastic inverse solutions of the groundwater flow equation based on cokriging (e.g. of transmissivity or hydraulic conductivity and head measurements) [see e.g. *Kitanidis and Vomvoris*, 1983; *Hoeksema and Kitanidis*, 1984; *Dagan*, 1985; *Rubin and Dagan*, 1987; *Gutjahr and Wilson*, 1989; *Ezzedine and Rubin*, 1996; *Fienen et al.*, 2008, 2009], whereas others employ Monte Carlo-based inverse modeling techniques to condition on observations of system-state variables such as heads, concentrations, and/or travel time [e.g. *Sahuquillo et al.*, 1992; *Gutjahr et al.*, 1994; *LaVenue et al.*, 1995; *Poeter and McKenna*, 1995; *RamaRao et al.*, 1995; *Gómez-Hernández et al.*, 1997; *Oliver et al.*, 1997; *Hanna and Yeh*, 1998; *Hendricks Franssen et al.*, 2003; *Alcolea et al.*, 2006; *Pasquier and Marcotte*, 2006; *Capilla and Llopis-Albert*, 2009; *Llopis-Albert and Capilla*, 2009].

Another possible strategy to condition simulations is to consider the full likelihood response surface within a Bayesian framework. Rather than retaining only those simulations that "closely" reproduce the observations, e.g. by minimizing discrepancies between

observations and simulated equivalents, or rather than using the maximum likelihood estimate [e.g. *Carrera and Neuman*, 1986a, c], this strategy acknowledges some departure between observations and simulated equivalents that expresses the ability of system simulators to represent the system. This approach is closer to the philosophy underpinning the generalized likelihood uncertainty estimation (GLUE) method proposed by *Beven and Binley* [1992], in which a quantitative measure of performance (also known as likelihood measure) is used to assess the acceptability of system simulators given a set of observations. Updating of the model likelihood distributions as new calibration data become available is handled easily within a Bayesian framework. Several authors have applied this approach in groundwater modeling to obtain simulations conditioned on observations of heads [e.g. *Feyen et al.*, 2001; *Morse et al.*, 2003], river discharges [e.g. *Jensen*, 2003], concentrations [e.g. *Sohn et al.*, 2000; *Hassan et al.*, 2008], or travel times [e.g. *Feyen et al.*, 2003]. Despite the lack of true conditioning, in the sense of reproducing observations of system-state variables exactly, we believe that the Bayesian method is particularly interesting in situations where new data become available, e.g. in transient groundwater flow modeling, as it provides a formal mechanism for combining previous information with new information. Even so, it is to be expected that in such situations the conditioning techniques described above will yield different sets of conditional simulations, depending on the time series or data set used in the conditioning process. Moreover, in any real application there is the issue as to whether the simulated equivalents of system-state variables are really comparable to the observations because of the scale effect at the element scale of the model and because of measurement errors.

Notwithstanding the outstanding level of sophistication and the robust theoretical grounds of the aforementioned techniques, the effects of conditioning on parameter values and observations of system-state variables within a multi-model framework have been poorly covered to date. The conditioning techniques and study cases described above rely exclusively on a single conceptualization. To the best of our knowledge, *Sohn et al.* [2000] are the first and until present the only to report on the value of conditioning data to assess *conceptual model and parameter uncertainty* in the context of groundwater modeling. They presented a two-step Bayesian Monte Carlo (BMC) method to assess the value of conditioning to head and concentration measurements for predicting TCE concentrations. To perform the Bayesian updating, a "Bayes window" method that is analogous to GLUE was used. In the work of *Sohn et al.* [2000], however, no attempt was made to aggregate multi-model predictions, to quantify the contribution of conceptual model uncertainty to the predictive uncertainty, or to assess the value of conditioning data on multi-model predictions.

In this paper we employ the multi-model method developed by *Rojas et al.* [2008a] to assess the value of conditioning data on multi-model posterior statistics. This method aims at explicitly quantifying uncertainty in the forcing data (input), model parameters and, especially, the conceptualization of the system through the combination of GLUE and BMA. For a range of conceptual models, 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 integrated model likelihoods. The latter are used to weight the predictions of the respective conceptual models in the BMA ensemble predictions. Using a hypothetical 3-dimensional groundwater system, *Ro-*

jas et al. [2008a] illustrated the method using unconditional realizations of the hydraulic conductivity field and conditioning on 16 head observations. The use of prior information about the plausibility of alternative conceptualizations (expressed as prior model probabilities) and its effect on ensemble predictions was investigated by *Rojas et al.* [2009]. In that work it was shown that posterior GLUE-BMA statistics were very sensitive to prior model probabilities and that including proper prior knowledge about the plausibility of alternative conceptualizations considerably improved the predictive performance of the GLUE-BMA approach.

We extend upon the works of *Rojas et al.* [2008a, 2009] and present what appears to be the first comprehensive analysis of the value of conditioning data on posterior multi-model statistics and predictions, and conceptual model uncertainty estimations. Two conditioning processes are considered: (i) spatial conditioning of the hydraulic conductivity field on sets of conductivity measurements with increasing sampling density and (ii) conditioning of simulated equivalents on heads, groundwater flows, and river discharge observations. For illustrative purposes, we employ a modified version of the 3-dimensional hypothetical system described in *Rojas et al.* [2008a] under steady-state conditions as study case. Uncertainty about the representation of this hypothetical system is expressed by an ensemble $\mathbf{M}$ of 7 alternative conceptual models with varying complexity. Additionally, we improve the sampling scheme implemented in *Rojas et al.* [2008a, 2009] by replacing the uniform sampling of the input and parameter space with a Markov Chain Monte Carlo (MCMC) method, more specifically, the Metropolis-Hastings (M-H) algorithm [see e.g. *Metropolis et al.*, 1953; *Hastings*, 1970; *Chib and Greenberg*, 1995; *Gilks et al.*, 1995; *Gelman et al.*, 2004]. For a series of groundwater flow components the effect of conditioning on the

posterior model probabilities (i.e. model weights used for multi-model aggregation), the posterior GLUE-BMA statistics, and the contribution of conceptual model uncertainty to the predictive uncertainty is analyzed.

It is worth emphasizing that the way the system simulators (i.e. conceptual model + parameter set) are evaluated here against the observations differs from the way hydraulic conductivity measurements are incorporated. Rather than retaining only those simulations that closely reproduce the observations, or rather than using the maximum likelihood estimate, we take into account the whole likelihood surface (cut off at limits of tolerable error). There exist techniques, like the self-calibrating (SC) method [*Sahuquillo et al.*, 1992; *Hendricks Franssen et al.*, 2003], the pilot point (PP) method [*LaVenue et al.*, 1995; *RamaRao et al.*, 1995], the Markov Chain Monte Carlo (MCMC) method [*Oliver et al.*, 1997], the Gradual Deformation (GD) method [*Hu*, 2000], gradient-based inverse techniques [*Hill et al.*, 1998; *Moore and Doherty*, 2005], which are able to generate hydraulic conductivity fields conditional to hydraulic conductivity and system-state data such as heads and flows.

2. Methodology

To render the present paper self-contained sections 2.1, 2.2 and 2.3 briefly describe GLUE, BMA and the procedure to integrate both methodologies in the frame of the method of *Rojas et al.* [2008a]. Sampling from the conditional posterior distributions is performed using the Metropolis-Hastings (M-H) algorithm. A brief description of the M-H algorithm as applied here is presented in section 2.4. More elaborated descriptions of the M-H algorithm can be found in *Chib and Greenberg* [1995], *Gilks et al.* [1995] and *Gelman et al.* [2004].

2.1. Generalized Likelihood Uncertainty Estimation (GLUE) Methodology

GLUE is a Monte Carlo (MC) simulation-based technique that rejects the idea of a single correct representation of a system in favor of many acceptable system representations [*Beven*, 2006]. This idea is based on the concept of equifinality, i.e. many combinations of model structures, parameter sets and forcing data may provide (equally) good reproductions of the observed system response when compared to a limited dataset [*Beven and Freer*, 2001; *Beven*, 2006]. Equifinality arises because of the combined effects of errors (limitations) in the forcing (input) data, system conceptualization, measurements, and parameter estimates, thus acting as triggers for non-identifiability, non-uniqueness and stability problems. In summary, for each potential simulator of the system (i.e. conceptual model + parameter/forcing data vector), sampled from a prior set of possible system representations, a likelihood measure (e.g. Gaussian, trapezoidal, model efficiency, inverse error variance) is calculated. This likelihood measure reflects a simulator's ability to reproduce the observed set of system responses (i.e. observations). Simulators that perform below a subjectively defined rejection criterion are discarded from 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.

Following the notation of *Rojas et al.* [2008a], let us consider a set of plausible model structures $\mathbf{M} = (M_1, M_2, \ldots, M_k, \ldots, M_K)$, a set of parameter vectors $\mathbf{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_l, \ldots, \boldsymbol{\theta}_L)$ and a set of input (forcing data) variable vectors $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m, \ldots, \mathbf{Y}_M)$, and denote the observations and simulated equivalents vectors as $\mathbf{D} = (D_1, D_2, \ldots, D_n, \ldots, D_N)$ and $\mathbf{D}^* = (D_1^*, D_2^*, \ldots, D_n^*, \ldots, D_N^*)$, respectively.

Then $L(M_k, \boldsymbol{\theta}_l, \mathbf{Y}_m|\mathbf{D})$ 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 ($\mathbf{D}$). *Rojas et al.* [2008a] observed, for the particular conditions of their synthetic study case, no significant differences in the estimation of posterior model probabilities, predictive capacity, and conceptual model uncertainty when a Gaussian (equation 1), a model-efficiency-based, or a Fuzzy-type likelihood functions were used. The analysis in this work is therefore confined to a Gaussian likelihood function given by

$$L(M_k, \boldsymbol{\theta}_l, \mathbf{Y}_m|\mathbf{D}) = (2\pi)^{-N/2} \left|C_{\mathbf{D}}\right|^{-1/2} \times \exp\left(-\frac{1}{2}\left(\mathbf{D}-\mathbf{D}^*\right)^T C_{\mathbf{D}}^{-1}\left(\mathbf{D}-\mathbf{D}^*\right)\right) \quad (1)$$

where $C_{\mathbf{D}}$ is the covariance matrix of the errors of the observations.

2.2. Bayesian Model Averaging (BMA)

BMA provides a coherent framework for combining predictions from multiple competing conceptual models to attain a more realistic and reliable description of the predictive uncertainty. It is a statistical procedure that infers average predictions by weighing individual 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 others, instead, competing models are assigned different weights based on the dataset $\mathbf{D}$ [*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 [*Draper*, 1995]

$$p(\Delta|\mathbf{D}) = \sum_{k=1}^{K} p(\Delta|\mathbf{D}, M_k) p(M_k|\mathbf{D}). \quad (2)$$

Equation (2) is an average of the predictive distributions of Δ under each alternative conceptual model, $p(\Delta|\mathbf{D}, M_k)$, weighted by their posterior model probability, $p(M_k|\mathbf{D})$. This latter term reflects how well model M_k fits the data $\mathbf{D}$ and can be computed using Bayes' rule

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k)p(M_k)}{\sum_{l=1}^{K} p(\mathbf{D}|M_l)p(M_l)} \tag{3}$$

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

$$p(\mathbf{D}|M_k) = \int p(\mathbf{D}|M_k, \boldsymbol{\theta}_l)p(\boldsymbol{\theta}_l|M_k)d\boldsymbol{\theta}_l \tag{4}$$

where $p(\mathbf{D}|M_k, \boldsymbol{\theta}_l)$ is the likelihood of model structure M_k parametrized with parameter vector $\boldsymbol{\theta}_l$ given the observations in $\mathbf{D}$, and $p(\boldsymbol{\theta}_l|M_k)$ is the prior probability distribution of $(\boldsymbol{\theta}_l)$ given model M_k. For this work $p(\boldsymbol{\theta}_l|M_k) \equiv p(\boldsymbol{\theta}_l)$ since priors are equivalent for all conceptual models (see section 5).

The leading moments of the BMA prediction of Δ (equations 5 and 6), which follows from general properties of the conditional mean and variance and which are implicitly conditional on the discrete ensemble of proposed models $\mathbf{M}$, are given by [*Draper*, 1995]

$$\begin{aligned} E[\Delta|\mathbf{D}] &= E_{\mathbf{M}}[E(\Delta|\mathbf{D}, \mathbf{M})] \\ &= \sum_{k=1}^{K} E[\Delta|\mathbf{D}, M_k]p(M_k|\mathbf{D}) \end{aligned} \tag{5}$$

$$
\begin{aligned}
Var\left[\Delta|\mathbf{D}\right] &= E_{\mathbf{M}}[Var\left(\Delta|\mathbf{D},\mathbf{M}\right)] + Var_{\mathbf{M}}[E\left(\Delta|\mathbf{D},\mathbf{M}\right)] \\
&= \sum_{k=1}^{K} Var\left[\Delta|\mathbf{D},M_k\right]p(M_k|\mathbf{D}) + \sum_{k=1}^{K}(E\left[\Delta|\mathbf{D},M_k\right] - E\left[\Delta|\mathbf{D}\right])^2 p(M_k|\mathbf{D}). \quad (6)
\end{aligned}
$$

From equation (6) it is seen that the variance of the BMA prediction of Δ consists of two terms; the first representing the within-model variance, and the second representing the between-model variance.

2.3. Multi-Model Approach to Account for Conceptual Model Uncertainties

The procedure to combine both methods is summarized as follows

1. Propose, on the basis of prior and expert knowledge about the site, a suite of alternative conceptualizations defining $\mathbf{M}$. Prior model probabilities are assigned for each member of $\mathbf{M}$. The latter can be achieved following, for example, *Meyer et al.* [2007], *Ye et al.* [2005, 2008a], and *Rojas et al.* [2009]. Alternatively, these prior model probabilities could potentially be used to penalize models to fully comply with the principle of parsimony.

2. Define prior ranges (or pdf's) for the sampling of parameter vectors, which is performed for each model structure. To keep the analysis at a neutral level, multi-uniform prior distributions are assumed to proceed with the sampling. Alternatively, prior ranges or pdf's could be defined for input (forcing) data. Should sound and proper prior knowledge were available about parameters and/or forcing data, an attempt to include it as non-uniform prior distributions should be made as far as possible [see e.g. *Ghosh et al.*, 2006].

3. Define a likelihood measure (for this work equation 1) and rejection criteria to assess model performance. Rejection criteria can be based on exploratory runs [e.g. *Rojas et al.*, 2008a, b], subjectively chosen threshold limits [e.g. *Feyen et al.*, 2001] or set as a minimum level of performance [e.g. *Binley and Beven*, 2003].

4. Sample, for each member of $\mathbf{M}$, parameter values (and alternatively forcing data) using the Metropolis-Hastings (M-H) algorithm to generate simulators of the system. In this work, sampling of hydraulic conductivity realizations is handled differently to guarantee spatial representativeness of the conditioning cases and to minimize the computational demands (see section 4).

5. Calculate a value for the likelihood measure (equation 1) for each simulator. On the basis of the rejection criteria, add the corresponding simulator to the subset A_k of retained simulators for model M_k or discard it by setting its likelihood to zero.

6. Repeat steps 4-5 until the hyperspace of possible simulators is adequately sampled. That is, when for each model M_k the first two moments of the conditional distributions of predictions (equations 5 and 6) based on the retained likelihood weighted simulators converge to stable values, and the R-score of *Gelman et al.* [2004] for parameters and variables of interest converges to values close to 1. The R-score expresses the ratio of within- to between-chain variability and, thus, approximate convergence of the M-H algorithm is diagnosed when the variability between chains is not larger than that within chains [*Sorensen and Gianola*, 2002].

7. Approximate the integrated likelihood of each model M_k (equation 4 which is needed to evaluate equation 3) by summing up the GLUE-based likelihood weights of the retained

simulators in the subset A_k, that is,

$$p(\mathbf{D}|M_k) \approx \sum_{l,m \in A_k} L(M_k, \boldsymbol{\theta}_l, \mathbf{Y}_m|\mathbf{D}). \quad (7)$$

8. Calculate posterior model probabilities for each member of the ensemble $\mathbf{M}$ using equation (3).

9. Approximate $p(\Delta|\mathbf{D}, M_k)$, after normalizing the likelihood weighted predictions under each individual model (such that the cumulative likelihood under each model equals 1). A multi-model prediction probability distribution is calculated with equation (2) with leading moments of this distribution given by equations (5) and (6).

It is worth mentioning that in the present application of the GLUE-BMA method only parameter and forcing data vectors described in section 5 are updated following the M-H algorithm. For the case of (un)conditional hydraulic conductivity realizations, a different approach was implemented to render the analysis computationally tractable (see section 4).

As discussed in *Rojas et al.* [2008a], important aspects of the GLUE-BMA method are that (1) it does not rely on a unique optimum parameter set for each conceptual model to assess the joint predictive uncertainty, thus, avoiding (minimizing the risk of) compensation of conceptual model errors due to forced improvements in model fit; (2) model-weights for aggregating multi-model predictions are obtained considering the full sampled hyperspace dimensioned by the model, parameter and forcing data vectors; (3) there is no implicit assumption about the conditional pdf's obtained for each alternative conceptualization; and (4) the possibility of including different types of conditioning data and prior knowledge follows naturally from the implementation of GLUE. A potential use of this prior information is to penalize models with a higher number of parameters

compared to the number of observations. If analyst's prior knowledge about the alternative conceptualizations is sound enough, this penalizing term could be obtained by defining non-uniform prior model probabilities. How efficiently defined this non-uniform prior model probabilities to comply with the principle of parsimony, however, is beyond the scope of this article and will be the subject of future research. Some guidelines can be found in the works of *Ye et al.* [2005] and *Rojas et al.* [2009].

2.4. Markov Chain Monte Carlo (MCMC) Simulation

Following a formal Bayesian inference approach, the predictive distribution used in equation (2) is given by [e.g. *Krzysztofowicz*, 1999; *Mantovan and Todini*, 2006]

$$p(\Delta|\mathbf{D}, M_k) = \int p(\Delta|\mathbf{D}, M_k, \boldsymbol{\theta}_l) p(\boldsymbol{\theta}_l|\mathbf{D}, M_k) d\boldsymbol{\theta}. \tag{8}$$

Typically, for hydrological problems the posterior parameter distribution $p(\boldsymbol{\theta}|\mathbf{D}, M_k)$ is highly dimensional and complex, with strong non-linear parameter interdependences. Hence, it is not easily amenable to direct sampling or analytical integration and it is necessary to resort to Monte Carlo (MC) methods to approximate the distribution. Since the form of the joint posterior distribution is not known, a Markov Chain Monte Carlo (MCMC) [see e.g. *Gilks et al.*, 1995; *Sorensen and Gianola*, 2002; *Gelman et al.*, 2004] approach is adopted to infer $p(\boldsymbol{\theta}|\mathbf{D}, M_k)$. More specifically, the Metropolis-Hastings (M-H) search strategy [*Metropolis et al.*, 1953; *Hastings*, 1970; *Chib and Greenberg*, 1995] is used to generate a sequence of parameter sets $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N\}$ that adapts to the target posterior distribution.

The idea of the M-H algorithm is to stochastically generate a series of samples through iterative Monte Carlo (MC) sampling such that, asymptotically, the stationary distribu-

tion of this series is the target posterior distribution [*Sorensen and Gianola*, 2002]. The M-H algorithm uses a proposal distribution, $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{i-1})$, to generate a new proposed sample [see e.g. *Chib and Greenberg*, 1995]. The generation of the Markov Chain is, thus, achieved in a two-step process: a proposal step and an acceptance step [*Sorensen and Gianola*, 2002]. In the proposal step the next parameter vector is proposed as a candidate point from the proposal distribution. Then, in the acceptance step, the joint density of the current Markov Chain state and the proposal distribution is corrected to ensure reversibility of the chain [see e.g. *Tierney*, 1994]. As a result, there is a natural tendency to accept parameters with higher posterior probabilities than the current parameter vector [*Gallagher and Doherty*, 2007].

Implementation details of the M-H algorithm are amply discussed in the literature [see e.g. *Geyer*, 1992; *Gilks et al.*, 1995; *Cowles and Carlin*, 1996; *Brooks and Gelman*, 1998; *Makowski et al.*, 2002; *Sorensen and Gianola*, 2002; *Gelman et al.*, 2004; *Ghosh et al.*, 2006; *Robert*, 2007], and so they will not be repeated here.

3. Three-Dimensional Hypothetical Setup

A modified version of the 3-dimensional hypothetical setup described in *Rojas et al.* [2008a] is used to assess the value of conditioning data (see Figure 1). The main differences lie in the reference hydraulic conductivity fields of the different layers and the number of pumping wells. We assumed statistically homogeneous deposits with a constant mean hydraulic conductivity K. Smaller-scale variability was represented using Random Space Functions (RSF), adopting an isotropic exponential covariance function for lnK in all three layers (see Table 1). The "true" spatial distribution of the hydraulic conductivity was generated using the sequential Gaussian simulation (sgsim) algorithm [see e.g.

Goovaerts, 1997] of the Geostatistical Software Library (GSLIB) [*Deutsch and Journel*, 1998]. Vertical hydraulic conductivity was (randomly) defined for each grid cell between 5% and 15% of the horizontal conductivity values obtained from sgsim.

Simulation of steady-state flow employed MODFLOW-2000 [*Harbaugh et al.*, 2000]. A uniform recharge of $1.4 \times 10^{-4} md^{-1}$ was applied to the top layer. At the west boundary a constant head $h = 46m$ was defined. At the east boundary a $10m$-wide river was defined with a constant stage of $25m$ and constant water depth of $5m$. Underlying the river, a $5m$-thick sediments layer with a vertical hydraulic conductivity of $0.1md^{-1}$ was defined. Eight wells were distributed in the area pumping a total of $2500m^3d^{-1}$ from the uppermost and lowermost aquifers (Figure 1). An evapotranspiration (EVT) zone, delineated by the polygon in Figure 1a, was defined with a surface elevation (SURF) at $43m$, an evapotranspiration rate (EVTR) of $1.37 \times 10^{-3} md^{-1}$ and an extinction depth (EXTD) of $5m$.

This setup was run in forward mode, obtaining the "true" groundwater head distribution for layers 1 and 3 (see Figures 1a and 1c). At the 16 observation well locations heads were selected from the "true" head distribution for layers 1 and 3. Additionally, groundwater inflows ($655m^3d^{-1}$) at the west boundary condition (WBC), river gains ($192m^3d^{-1}$), and EVT outflows ($63m^3d^{-1}$) were obtained from the forward run. The set of 32 head values as well as the groundwater inflows (GWF) and river gains (RIV) were used as conditioning data to estimate the likelihood weights in the evaluation of the different simulators.

4. Conditioning Procedure

We considered two conditioning mechanisms: (i) spatial conditioning of the hydraulic conductivity realizations on hydraulic conductivity measurements, and (ii) the condition-

ing of simulated equivalents on observations of system-state variables within a GLUE framework. A summary of the conditioning cases analyzed in this paper is presented in Table 2.

To analyze the effect of spatial conditioning of the hydraulic conductivity fields, four cases were considered. The first case (Unconditional) corresponds to unconditional realizations of the hydraulic conductivity field. Cases Conditional-I, Conditional-II, and Conditional-III correspond to realizations conditioned on 10, 20, and 40 hydraulic conductivity measurements, respectively (Figure 2). Measurement points were defined such that smaller sets were a subset of the larger ones to avoid the effect of varying measurement locations among the sampling densities [*Feyen et al.*, 2002].

Hydraulic conductivity fields characterized by a spatial correlation structure are not easily amenable to updating when combined with other parameters in the development of the Markov chains. Ideally, hydraulic conductivity realizations could be generated and a chain could be developed for each individual realization. However, given the high number of hydraulic conductivity realizations to properly represent spatial variability and the number of chains required to ensure convergence of posterior statistics, such approach is computationally still intractable. An alternative to overcome this limitation is working with "representative" hydraulic conductivity realizations and develop a given number of chains for them. That is, keep the spatial realizations fixed and develop a series of chains for non-spatial parameters. To obtain such representative realizations we proceeded as follows: (a) 100 hydraulic conductivity realizations for the corresponding layers of each conceptual model were generated, based on the spatial correlation structures defined in Table 1 (i.e. 100 realizations for one-layer models, 200 realizations for two-layer models,

and 300 realizations for model 3L), (b) these realizations created 100 variants of each of the 7 conceptualizations (i.e. 700 variants in total), (c) each of these 100 variants was run in forward mode driven by the "true" parameter values used in the 3-dimensional hypothetical setup (see section 3), (d) a likelihood value was obtained (equation 1) for each of these 100 variants, (e) the variant with a likelihood closest to the median likelihood of the 100-ensemble variants was considered to be representative. These steps were repeated for the 4 conditioning cases and for each conceptual model. In total, 2800 preliminary runs were done to obtain 28 (7 models $\times$ 4 conditioning cases) representative variants. For this set of preliminary runs we employed $C_{\mathbf{D}} = \mathbf{I}$ (see equation 1) as the variants were driven by the true parameter values obtained from the 3-D hypothetical setup. This was done to ensure that only spatial variability played the major role in the likelihood estimation for the definition of the conditioning case. Further development of the Markov chains was based on this (fixed) representative variants.

Conditioning on observations of system-state variables was performed within the GLUE framework. Simulations were conditioned on heads (h_{sim}), groundwater inflow at WBC (GWF_{sim}), and river gain (RIV_{sim}) observations. Model performances against the observations were assessed using 3 Gaussian likelihood functions (assuming no correlation in observation errors, i.e. off-diagonal terms of $C_{\mathbf{D}}$ equal zero): (a) one for the groundwater heads (h_{obs}) centered on each of the 32 values with a standard deviation (σ_{hobs}) equal to $2.5m$; (b) one for the groundwater flow observation (GWF_{obs}) centered on the observed value ($655m^3d^{-1}$) with a standard deviation (σ_{GWFobs}) equal to 20% of this value ($131m^3d^{-1}$), and (c) one for the river discharge observation (RIV_{obs}) centered on the observed value ($192m^3d^{-1}$) with a standard deviation (σ_{RIVobs}) equal to 10% of this value

($19m^3d^{-1}$). These standard deviations were also used to define the diagonal elements of matrix $C_{\mathbf{D}}$ used to calculate the corresponding likelihood values for each simulator. It was assumed that groundwater inflow observations are more uncertain than river discharge observations, which explains the larger spread of the likelihood function for this variable. We acknowledge that standard deviations defined for RIV and GWF observations are (to some extent) lower than the values that could be achieved at field sites. The ratio of standard deviations, however, serves the purpose of illustrating the inclusion of relatively certain and uncertain observations. Lower and upper rejection thresholds, i.e. limits of tolerable error, were set at the corresponding observations ± 3 times the standard deviations ($2.5m$, $131m^3d^{-1}$, or $19m^3d^{-1}$) defined for each Gaussian likelihood function (equation 1). Simulators whose likelihood values were not contained within the ranges defined by these tolerable errors were discarded from further analysis by assigning their likelihood to zero.

Observations used to assess model performance were not corrupted by any noise (e.g. aimed at explicitly accounting for measurement errors). In this regard, results of *Hill et al.* [1998] suggest that standard deviations such as those used in this work are a good approximation in case of synthetic problems.

In summary, we consider 4 cases of spatial conditioning to hydraulic conductivity measurements and 3 cases of conditioning on observations of system-state variables. Hence, in total, 12 cases were analyzed (see Table 2).

5. Implementation details

An ensemble **M** of 7 alternative conceptualizations was considered to represent the 3-dimensional hypothetical setup, ranging from simple one-layer models to more complex

three-layer systems approaching the true setup (see Table 3). The spatial distribution of hydraulic conductivity in the one-layer conceptualizations 1L-L1, 1L-L2 and 1L-L3 follows the spatial correlation structure of layers 1, 2 and 3, respectively. Conceptualization 1L-AVG, on the other hand, is characterized by a hydraulic conductivity distribution that follows the average spatial correlation structure of the three layers in the hypothetical setup (i.e. averaging parameters defined in Table 3). In the two-layer (2L) conceptualization the hydraulic conductivity distributions follow the spatial correlation structure of layers 1 and 3 of the hypothetical setup. The 2LQ3D conceptualization corresponds to model 2L but includes an implicit representation of the aquitard depicted in Figure 1b. This aquitard was represented as a constant value for the vertical hydraulic conductivity using the LPF package of MODFLOW-2000. The last conceptualization represents a three-layer (3L) system that accounts explicitly for the aquitard. For this setup, the hydraulic conductivity distributions of the layers follow the spatial correlation structures of the corresponding layers of the hypothetical setup. For the seven conceptualizations vertical hydraulic conductivity was (randomly) defined for each grid cell between 5% and 15% of the horizontal conductivity values, with the exception of model 2LQ3D which used a constant value of $1 \times 10^{-3} md^{-1}$. Table 3 summarizes the 7 conceptualizations defining $\mathbf{M}$.

Table 4 shows prior ranges for the 6 parameters included in the M-H algorithm for developing the chains. For pragmatic reasons, we employed the sampling ranges defined in *Rojas et al.* [2008a]. To exclusively asses the value of conditioning data we employed multi-uniform sampling distributions. *Rojas et al.* [2008a] showed that the most sensitive parameters of a similar 3-dimensional hypothetical setup were the uniformly distributed

recharge rates (RECH) and the constant head elevation (CH) at the west boundary condition (WBC). Therefore, in this work we focus on obtaining proper convergence of the M-H algorithm for these two parameters. This was done by closely monitoring the convergence of the R-score [*Gelman et al.*, 2004] to values close to 1. The evolution of the chains for the remaining 4 parameters was monitored as well.

We employed Gaussian likelihood function to assess the model performance (equation 1). As previously explained, rejection thresholds were defined on the observations as the observed value ± 3 times the standard deviations used for conditioning (see section 4).

To implement the M-H algorithm, a 6–dimensional multivariate normal distribution centered on the current state of the chain ($\boldsymbol{\theta}_{i-1}$) was selected as proposal distribution, i.e. $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{i-1}) \sim N(\boldsymbol{\theta}_{i-1}|\Sigma_{\boldsymbol{\theta}})$, where $\Sigma_{\boldsymbol{\theta}}$ is the (diagonal) variance matrix used as "jumping rule" [see e.g. *Gilks et al.*, 1995] for sampling new parameter candidates ($\boldsymbol{\theta}^*$). An acceptance rate (defined as the proportion of accepted parameter candidates in the set of the last n proposed samples) of 40% was defined. A fixed acceptance rate was chosen to minimize the potential effects of working with chains obtained with different acceptance rates. In preliminary runs, the variance terms included in $\Sigma_{\boldsymbol{\theta}}$ were adjusted by trial-and-error until the acceptance rate of 40% was achieved. Although RECH and CH were the most sensitive parameters, $\Sigma_{\boldsymbol{\theta}}$ accounted for all six parameters to allow for flexibility in the trial-and-error process. Based on preliminary runs and the results obtained by *Rojas et al.* [2008a, b], a pre-defined number of 10 chains was used for each of the 7 conceptualizations (Table 3) under each of the 12 conditioning cases (Table 2). Starting locations for the individual chains were randomly defined from the prior parameter ranges defined in Table 4. To avoid negative effects induced by autocorrelation within successive steps

of a chain, the final chains were thinned before calculating summary statistics [*Sorensen and Gianola*, 2002]. Finally, based on the results of *Rojas et al.* [2008a] the total sample of parameters should contain at least $10,000$ elements to ensure the convergence of the first two moments of the posterior distributions.

Thinned sampled obtained from the M-H algorithm were used to estimate GLUE-based model likelihoods (equation 7) and posterior model probabilities for each conceptualization (equation 3). Normalized GLUE-based likelihoods were employed to approximate $p(\Delta|\mathbf{D}, M_k)$. Finally, a multi-model prediction accounting for conceptual model uncertainty was obtained using equation (2).

6. Results

6.1. Validation of the M-H algorithm results

Several aspects of the implementation of the M-H algorithm were checked to validate the results. Based on visual inspection of the plotting series and values of the R-score for exploratory runs, the length of the *burn-in* samples was defined at a maximum of $0.2P$ (with P being the total chain length), following *Gilks et al.* [1995]. After discarding the *burn-in* samples (to minimize the influence of the starting locations of the multiple chains), this resulted in chains of $6,000$ samples. R-scores ranged between 1.001 and 1.062 for parameter RECH and between 1.001 and 1.032 for parameter CH. For the predictions of interest (groundwater inflows from WBC, river gains, recharge inflows and EVT outflows) R-scores ranged between 1.001 and 1.061. This indicates proper mixing of the chains for all parameters and variables of interest. As an example, Figure 3 shows the results for model 3L under case Conditional-III. Figure 3a shows 10 independent chains after discarding the *burn-in* samples. A good mixing (overlap) of the chains is observed indicating proper

convergence. Additionally, Figure 3b shows the effect of thinning the original sample of $60,000$ elements after every 6 iterations. The autocorrelation factor quickly vanishes and its value is evenly distributed around 0. Similar patterns were observed for parameter CH and the simulated system-state variables of concern. Therefore, the thinned parameter samples of $10,000$ elements for each model under each conditioning case were considered to be an independent sample from the target posterior distributions. These individual samples were used to assess the effect of conditioning data on posterior GLUE-BMA statistics.

6.2. Likelihood Response Surfaces and the Value of Conditioning Data

6.2.1. Hydraulic conductivity data

The effect of increasing the number of hydraulic conductivity measurements on the global likelihood response surface is shown in Figure 4. These likelihood surfaces are obtained by applying equation 1 using the $C_{\mathbf{D}}$ matrix defined in section 4. Figure 4 depicts the results for parameter RECH for cases Unconditional and Conditional-III for alternative models 1L-L3, 2L and 3L. It is seen from this figure that when alternative conceptualizations approach the hypothetical setup the global likelihood response surface becomes better defined, less disperse and less biased compared to the "true" value for parameter RECH. This implies that a more correct geological description improves the precision of the global likelihood response surface. Results for the case Unconditional, however, show that a proper description of geology, as represented in Table 1, does not guarantee unbiased estimations. Conditioning of the hydraulic conductivity field reduces the bias in the parameter estimates and decreases the range of variation (uncertainty), i.e. improves both precision and accuracy. For conceptualization 1L-L3, however, a

reduction in bias is observed while the range of variation slightly increases. The latter is due to the fact that the information content in the conditioning data is not fully coherent with this simplified one-layer setup. Moreover, absolute values of the global likelihood measure increase considerably with increasing levels of spatial conditioning, which affects the posterior model probabilities as shown in Table 5. For parameter CH and the simulated groundwater flow components, the global likelihood response surfaces showed very similar patterns as those of the RECH parameter, and are therefore not shown here.

Interestingly, Figure 4 shows a series of outliers (spurious secondary optima) in various directions. *Kavetski and Kuczera* [2007] suggest that intricate likelihood response surfaces (or objective functions in their terminology) are not always inherent features of the hydrological models, but are numerical artifacts or model nonlinearities not very important for accurate simulation of the relevant system. This could partially be explained by models showing a threshold-type behavior which induces model nonlinearities, nontrivial likelihood response surfaces, and multiple (spurious) optima [*Kavetski and Kuczera*, 2007].

6.2.2. Heads, groundwater flow and river discharge observations

The value of including head, groundwater flow (GWF) and river discharge (RIV) observations on the global likelihood response surface is illustrated in Figure 5, which presents the results for model 3L under the case Unconditional. This case was selected because the value of including observations of system-state variables might partially be masked by the spatial conditioning on hydraulic conductivity measurements. Because global likelihood values were obtained by aggregating point likelihood values using a geometric mean inference, global likelihood values among columns of Figure 5 can not be directly compared.

The spread and accuracy of the global likelihood response surfaces, however, can be fully compared.

Comparison of the global likelihood response surfaces in Figure 5 shows that including the GWF observation has a drastic impact on the spread of the groundwater inflows estimated at the west boundary condition (WBC) (Figure 5b). The GWF observation also considerably reduces the spread of the global likelihood response surface of the evapotranspiration outflows (Figure 5h), and to a lesser extent also that of the river gains (Figure 5e). Similarly, including the RIV observation has the largest impact on the spread of the river gains (Figure 5f) but also improves the estimation of the other system-state variables (Figure 5c and Figure 5i). The truncation of the global likelihood response surfaces at upper and lower bounds (see e.g. Figures 5b, 5c and 5f) when conditioning on GWF and RIV observations is explained by the definition of the limits of tolerable error (rejection criteria in GLUE) for the Gaussian likelihood functions defined in section 4. When incorporating the GWF and RIV observations the rejection criterion is no longer only based on heads. Hence, many simulators that still produced acceptable simulations based on head observations fail to meet the acceptance criteria for the GWF and RIV observations and, as a result, are rejected and assigned a likelihood value equal to zero. The fact that simultaneously including GWF and RIV observations considerable reduces (by approximately 75%) the spread for the evapotranspiration outflows (Figure 5i) compared to when including only groundwater head observations (Figure 5g), confirms that conditioning on observations of system-state variables less commonly available than groundwater heads may significantly reduce the predictive uncertainty of other simulated system-state

variables not included as conditioning data [see e.g. *Anderman et al.*, 1996; *Feyen et al.*, 2003]. A similar pattern was observed across all 12 conditioning cases analyzed.

6.3. Posterior Model Probabilities and the Value of Conditioning Data

Table 5 shows the full matrix of integrated model likelihoods and posterior model probabilities for all 7 conceptualizations and 12 conditioning cases analyzed. Although in this work we employ a slightly modified version of the hypothetical setup used in *Rojas et al.* [2008a] and we double the number of head observations for conditioning, the results shown in the first two rows of Table 5 are in full agreement with those observed in *Rojas et al.* [2008a]. Hence, considering 16 extra head observations from the lowermost aquifer (layer 3) (Figure 1c) that dominates the system dynamics, compared to the set of 16 head observations from the uppermost aquifer in *Rojas et al.* [2008a], does not allow further discrimination among conceptualizations. Rather, the same five models are selected. Moreover, assigned posterior model probabilities for the retained 5 conceptualizations remain unchanged compared to the case when using only 16 head observations [*Rojas et al.*, 2008a]. This indicates that the information content of heads is limited in its ability to differentiate among models or to refine the ensemble $\mathbf{M}$ of proposed conceptualizations. Although it can be argued that the latter might be caused by the fact that the hypothetical setup is mainly driven by Dirichlet conditions, or that head observations are located into a layer that is cut off by the aquitard, head observations are located in both aquifers to account for the lowermost aquifer and variations in head induced by the presence of the aquitard (see section 3).

These results suggest that to effectively reduce the number of conceptualizations or to further differentiate among them, other sources of information apart from head observations must be considered.

Results from Table 5 confirm that GLUE-BMA tends to produce more evenly distributed posterior model weights compared to model selection criteria-based multi-model frameworks. This avoids concentrating all posterior model weights in a rather small number of alternative conceptualizations, thus, minimizing the risk of under-estimation and biased uncertainty estimations in multi-model frameworks. These results are in full agreement with the findings of *Rojas et al.* [2008b], *Singh et al.* [2009] and *Ye et al.* [2009].

6.3.1. Hydraulic conductivity data

Comparison in Table 5 of the respective rows for the different spatial conditioning cases shows that the integrated model likelihoods increase with conditioning density. Comparing cases Unconditional and Conditional-III, this increase is more pronounced for conceptualizations closer to the hypothetical setup indicating that spatial conditioning allows a better discrimination among the alternative conceptual models by concentrating more weight to models closely describing the hypothetical setup. Comparing cases Conditional-II and Conditional-III, the additional conditioning makes the model discrimination slightly worse. The latter might potentially be explained by the particularities of the hypothetical setup and/or the spatial distribution of the sampling points of the hydraulic conductivity measurements depicted in Figure 2. For the two retained one-layer models, however, increasing the number of hydraulic conductivity data does not necessarily result in higher model likelihoods, as the information content in the conditioning data may not be fully coherent with the simplified setup. This indicates that spatial

conditioning helps in rejecting those conceptualizations contradicting the (unknown true) dynamics of the groundwater system. As an example, conceptualization 1L-AVG is no longer supported when the number of spatial conditioning data is high enough (between 11 and 20 K measurements). This is due to the fact that using an average spatial correlation structure and average conditioning values smoothes out the effects of each layer. The latter clearly emphasizes the importance of using the correct conditioning data for conditioning and deriving the spatial correlation structure to constrain potential simulators of the unknown system. In practical terms, however, (vertically) averaged values of K over alternating well-screen sections may often be the only information available about the hydraulic conductivity in an aquifer.

6.3.2. Heads, groundwater flow and river discharge observations

The effect of including GWF (Heads + GWF) and RIV (Heads + GWF + RIV) observations is also shown for each conditioning case in Table 5. Results show that conditioning to observations of system-state variables less commonly available than groundwater heads allows for a better differentiation among the alternative conceptualizations. A tendency to assign more weight to models closely reproducing the hypothetical setup is observed across all conditioning cases. On the contrary, posterior model probabilities of simplified conceptualizations tend to decrease when including observations of system-state variables. This suggests that the information contained in these data (GWF and RIV) does not support simple approximations of the hypothetical setup.

These results illustrate that to further constrain the space of potential simulators in the framework of the GLUE-BMA method, the information content of heads should be complemented with observations of other system-state variables. The latter is in full agree-

ment with general recommendations suggested for the traditional calibration of groundwater models which are supported by a long tradition of modeling exercises employing nonlinear regression techniques [see e.g. *Cooley et al.*, 1986; *Cooley and Naff*, 1990; *Poeter and Hill*, 1997; *Hill et al.*, 1998; *Hill and Tiedeman*, 2007].

Note that, although it is not the aim of this work, results as those presented in Table 5 could potentially be used to design optimal sampling and/or data collection schemes to further discriminate among alternative conceptualizations [see e.g. *McLaughlin and Wood*, 1988; *Wagner*, 1995; *Tiedeman et al.*, 2003, 2004; *Tonkin et al.*, 2007].

6.4. Groundwater Model Predictions Accounting for Conceptual Model Uncertainty

6.4.1. Hydraulic conductivity data

Figure 6 shows the effect of increasing the number of hydraulic conductivity data on recharge inflows and predictions of evapotranspiration outflows. Results are shown for these system-state variables because they were not used as conditioning data, hence, allowing for a better assessment of the spatial conditioning on hydraulic conductivity. Figure 6a shows that for the Unconditional case cumulative predictive distributions for recharge inflows vary strongly in shape, spread and central moment across the alternative conceptualizations. Conditioning to a set of 40 hydraulic conductivity measurements (Figure 6c), on the other hand, produces predictive distributions that are less disperse and more similar in shape and central moments, with the exception of conceptualization 1L-L3. For the latter, the central moment remains quite dissimilar from the other distributions and fails to closely reproduce the true value. Rather, a shift from overestimation to underestimation of the true value can be observed. This shift is explained by the inverse

correlation observed between the recharge inflows and the groundwater inflows from the WBC [*Rojas et al.*, 2008a] (see also Figure 7a).

For the evapotranspiration outflows (Figure 6b and d), spatial conditioning shows a less pronounced effect on the predictive distributions, with conceptualizations 2L and 1L-L3 even showing a larger spread compared to the Unconditional case. The latter can be explained by (i) the strong influence of pumping well W4-(L3) located in the surroundings of the evapotranspiration zone (see Figure 1a) and (ii) the shift to higher evapotranspiration outflows for the case Conditional-III induced by higher groundwater heads (compared to the case Unconditional) within the polygon defining the evapotranspiration zone. This is reaffirmed in Figure 7, where a shift from lower to higher groundwater inflows from the WBC (Figure 7a) is observed, explaining the underestimation of the true recharge inflows observed in Figure 6a for conceptualization 1L-L3. In addition, Figure 7b shows higher groundwater heads downstream of well W4-(L3) compared to the Unconditional case, thus, explaining the tendency to higher evapotranspiration outflows for the case Conditional-III. The same applies for model 2L as it is driven by layer L3 as well.

These results show that, even for the case of heavily (spatially) conditioned conceptualizations, predictions can potentially be completely biased if driven by the "wrong" conceptualization. This will clearly affect the quality of multi-model predictions, and may even eliminate the gain in accuracy and precision of individual model predictions from conceptualizations more closely representing the unknown dynamics of the groundwater system. The latter is closely linked to the case when predictions from a heavily conditioned simple conceptualization are compared to a poorly conditioned but more explanatory conceptualization. A larger spread from predictive distributions derived from

a single "wrong" conceptualization produces a high value for the between-model variance term of equation (6), resulting in a BMA prediction that is more uncertain. As an example, for evapotranspiration outflows (Figure 6d), the presence of conceptual models 2L and 1L-L3 even counterbalances the gain in precision and accuracy obtained from the spatial conditioning mechanism for the other conceptual models. In this case, it would be useful to collect data that are informative on the evapotranspiration process to constrain the disagreement among predictive distributions after spatial conditioning.

The fact that multi-model predictions are largely affected by individual predictions obtained from a single (wrong) conceptual model and that this model could not be eliminated based on the available data $\mathbf{D}$ raises the paramount question on how to identify "wrong" conceptualizations and what type of data are more informative in doing so. In the context of this work, the application of the GLUE-BMA methodology allowed the refinement of the original ensemble $\mathbf{M}$ by discarding 2 out of 7 conceptualizations. In addition, hydraulic conductivity measurements showed to be highly informative in terms of eliminating an additional conceptualization that contradicted the basic dynamics of the hypothetical setup. Moreover, collecting flow-related measurements showed a clear tendency to assign less weight to "simpler" conceptualizations. We could therefore hypothesize that collecting more flow-related measurements might potentially constrain even more the ensemble $\mathbf{M}$.

It might be argued that conducting the model averaging considering solely the best three models (2L, 2LQ3D and 3L) appears as a legitimate option. This, however, contradicts the main purpose of the methodology applied which is to assess the uncertainty arising from the definition of alternative and valid conceptual models. Working with the best models

only may result in underestimation of predictive uncertainty and biased predictions, which are precisely the two problems that model averaging intends to avoid [*Ye et al.*, 2009].

6.4.2. Heads, groundwater flow and river discharge observations

The effect of including GWF and RIV observations on the predictive distributions of simulated system-state variables is shown in Figure 8. For both Unconditional and Conditional-III cases the spread of the cumulative BMA predictive distributions reduces with increased conditioning, with the minimum spread observed for the conditioning case including heads, groundwater flow and river discharge observations (Heads + GWF + RIV). GWF and RIV observations implicitly convey more valuable information on flow-related system-state variables, therefore, improving their accuracy. Flow-related measurements contain a different type of information than head measurements, and thus combining the latter two is better than adding more of one type.

From the hypothetical setup (which is assumed as the unknown true system) it is known that groundwater outflows from the system through the WBC are physically not possible given the measured heads. When conditioning on heads solely, this physical constraint is ignored and groundwater outflows from the system are observed for both Unconditional and Conditional-III cases (Figure 8b and Figure 8f). By conditioning on GWF and RIV observations a more accurate physical representation of the system is observed. The latter shows the relevance of including additional conditioning data to ensure a physically meaningful modeling of the groundwater system.

6.5. Predictive Variance

As shown earlier, conditioning on increasing numbers of hydraulic conductivity measurements resulted in the elimination of model 1L-AVG. Since we were interested in ex-

clusively assess the worth of including hydraulic conductivity measurements to reduce the predictive variance, we compared conditioning cases on the basis of the conceptualizations included in the ensemble **M**. This was done to avoid potential effects of discarding model 1L-AVG, e.g. masking the actual worth of hydraulic conductivity measurements on the estimations of the predictive variance. This means the analysis was done pair wise, i.e. cases Unconditional with Conditional-I, and cases Conditional-II with Conditional-III.

Figure 9 shows the comparison between the most conditioned cases (Conditional-II and Conditional-III) for recharge inflows and evapotranspiration outflows. Results reveal that a doubling of the number of hydraulic conductivity measurements drastically decreases the predictive variance, especially for the recharge inflows. In addition, using observations of groundwater flow and river discharge further reduces the predictive variance in both cases compared to conditioning solely on groundwater heads. Similar patterns were observed for the other predicted variables and parameters (see Tables 6).

In all cases analyzed, including more data for conditioning strongly reduces the contribution of within-model variance to the predictive variance. The effect on between-model variance is less pronounced and does not show a clear pattern. For example, for cases Conditional-II and Conditional-III including the GWF observation reduces the relative contribution of conceptual model uncertainty to the predictive variance compared to the case using only groundwater heads. Including the RIV observation, on the contrary, results in an increase in the relative contribution of conceptual model uncertainty to the predictive variance. Moreover, for evapotranspiration outflows after conditioning on extra hydraulic conductivity measurements, it is seen that between-model variance slightly increases for the conditioning based only on groundwater heads (first column Figure 5d).

As discussed earlier, this is explained by the presence of conceptualizations 2L and 1L-L3, for which predictions strongly deviate from those of the other models when conditioning on extra hydraulic conductivity measurements.

7. Discussion and Conclusions

In this work we assessed the value of conditioning to various types of data in a multi-model methodology aimed at explicitly accounting for conceptual model uncertainty in groundwater modeling. We considered conditioning on increasing sets of hydraulic conductivity and on observations of 3 system-state variables, namely, heads, groundwater flow passing through a boundary condition, and river discharge. In total 12 conditioning cases were analyzed. The analysis was applied to a 3-dimensional hypothetical groundwater system under steady-state conditions. Uncertainty about the conceptualization of this flow system was represented by an ensemble ($\mathbf{M}$) of 7 alternative conceptual models. For each conditioning case and conceptualization, integrated model likelihoods and posterior model probabilities were derived, which were then used to obtain multi-model statistics and predictions explicitly accounting for conceptual model uncertainty.

We acknowledge that we have conveniently included observations of groundwater inflows and river discharges in the conditioning mechanism. For the first, it might be argued that this type of observation is seldom available in practical applications and even if it is, the level of uncertainty in its estimation might be so high rendering its information content (or worth) relatively poor. We do believe, however, that even the inclusion of an uncertain estimation of groundwater inflows, e.g. derived by applying fundamental laws of groundwater hydraulics, will help reduce identifiability (or equifinality) problems by constraining the space of potential simulators of the system. River discharge observations

are typically less uncertain compared to groundwater flow observations, and as such, they may convey more valuable information about the dynamics of the groundwater system. In addition, as no noise was added to the observations to account for error measurements or potential correlations, the analysis and conclusions of this work are restricted to that assumption.

We recognize that the GLUE-BMA method does not penalize more complex models, i.e. models with a higher number of parameters, through the likelihood function (equation 1). It is likely that models with more parameters will have higher likelihoods as the latter are obtained from model fit solely. In the event both prior model probabilities and integrated model likelihoods are equal for alternative conceptualizations, posterior model probabilities will be identical independent of the number of parameters for each conceptualization. Penalizing more complex models to comply with the principle of parsimony, however, can be achieved by defining non-uniform prior model probabilities. These non-uniform distributions could be defined on the basis of model complexity, plausibility, or any other criteria followed by the analyst. The inclusion of this knowledge rests on the Bayesian paradigm, which is the formal approach for combining prior knowledge and the evidence provided by observed data. Expert-based prior knowledge expressed as quantitative relationships among the alternative conceptualizations optimized, for example, using a constrained maximum entropy approach could be an alternative to define sound non-uniform prior model probabilities [see e.g. *Ye et al.*, 2005; *Rojas et al.*, 2009].

An option to penalize more complex models, and alternatively assess the information content of the data, is to use model selection criteria to approximate posterior model weights in multi-model frameworks [see e.g. *Neuman*, 2003; *Poeter and Anderson*, 2005;

Meyer et al., 2007]. However, two main drawbacks of this option are worthwhile discussing. First, different model selection criteria will lead to different posterior model weights and, as a consequence, to drastic differences in uncertainty assessments [see e.g. *Rojas et al.*, 2008b; *Singh et al.*, 2009; *Ye et al.*, 2009]. This is due mainly to the differences in how alternative criteria penalize model complexity, value prior information on parameter estimates, or interpret the quality of the available data. Despite *Ye et al.* [2008b] have presented an insightful discussion about merits and demerits of alternative model selection criteria in the context of multi-model approaches, the dilemma still remains about using one criterion over the others, thus, hampering the use of multi-model frameworks relying on model selection criteria [see e.g. *Tsai and Li*, 2008; *Singh et al.*, 2009; *Ye et al.*, 2009; *Tsai and Li*, 2010; *Ye et al.*, 2010]. Second, criteria-based multi-model approaches tend to concentrate posterior model weight in a rather small number of conceptualizations, thus, promoting under-dispersion and (potentially) more drastic bias in uncertainty estimations, the two problems that model averaging precisely intends to avoid. As GLUE-BMA-based model weights are more evenly distributed across alternative conceptualizations [see e.g. *Singh et al.*, 2009; *Ye et al.*, 2009], the risk of under-dispersed and biased uncertainty estimations is lower compared to criteria-based multi-model approaches.

Working with a suite of plausible conceptualizations we explicitly attempted to disentangle the effects of conceptual model uncertainty on the predictive uncertainty. The application of the GLUE-BMA method resulted in discarding 2 out of 7 conceptualizations based on the head observations, which was subsequently updated by ruling out 1 conceptual model based on spatial conditioning. The inclusion of flow-related variables allowed for a better discrimination among the conceptualizations. However, given the

available data, some simpler models could not be eliminated based on their posterior model weight, even though they resulted in higher between-model variances for flow-related variables that were not included in the conditioning. This raises the question as to how valid are the predictions for such variables from simpler models, which cannot be resolved on the basis of model weights solely. To decide on the validity of individual model predictions, or to identify conceptualizations that may be too simplistic or erroneous representations of the true flow system, their contribution to the conceptual model uncertainty and, by association, to the predictive uncertainty must be established. The application of the GLUE-BMA method allows identifying conceptualizations producing strongly deviating predictions from the ensemble average. Hence, even though for cases when alternative conceptualizations cannot be (strongly) differentiated on the basis of available data, knowledge about the relative contribution of conceptual model uncertainty to predictive uncertainty may be useful to guide, for example, data collection campaigns or to decide on conceptualizations worth to be explored in more detail. This is a key advantage compared to approaches focusing on the estimation of the overall variance solely. Data collection should then be aimed at acquiring data that can further eliminate such conceptualizations. For example, if alternative conceptualizations differ in some (key) aspects or processes, e.g. layering or zonation, collecting data of quantities related to those aspects or processes may help towards conceptual model discrimination and refinement of $\mathbf{M}$.

In addition, considering an ensemble of conceptual models avoids problems with overfitted individual models, under-dispersive uncertainty estimations, and (potentially) biased parameter estimates in order to compensate for unknown errors in the conceptualization

of the system. A clear disadvantage of this approach is the fact that by expanding the sampling space to the conceptual model dimension, the global likelihood response surface necessitates to be extensively sampled. This, however, can partly be alleviated by including more efficient parameter sampling schemes such as the Metropolis-Hastings sampling method used herein. Despite this, recent applications have shown that the method is fully applicable to local- and regional-scale aquifer systems [*Rojas et al.*, 2008b; *Ye et al.*, 2009; *Rojas et al.*, 2010].

Although it might be argued that the validity of the results from this work is limited as they are driven by particularities of the hypothetical setup used, results about the value of spatial conditioning on global likelihood response surfaces and posterior model probabilities (model weights) are in full agreement with results obtained for a multi-model analysis in a regional aquifer system in North Chile [*Rojas et al.*, 2010]. We could therefore hypothesize that conditioning on observations of system-state variables for real applications will lead to similar results as those presented here.

The main findings of this work can be summarized as follows

1. The inclusion of 16 extra head observations as conditioning data (in comparison to the work done by *Rojas et al.* [2008a]) did not significantly alter posterior model probabilities, hence, did not allow further discrimination among conceptualizations. Although it can be argued that this might be caused by intrinsic properties of the hypothetical setup (flow system dominated by Dirichlet boundaries), in practical applications a limited set of head observations may often be the only information available about the dynamics of the groundwater system, thus, rendering multi-modeling analyses particularly challenging. It is therefore strongly advised to complement the information content of heads with

measurements of key parameters and observations of other system-state variables to further discriminate among alternative conceptual models or to constrain the ensemble **M** of proposed conceptualizations.

2. Conditioning to hydraulic conductivity measurements allowed for a better differentiation among the different conceptualizations and for the elimination of 1 conceptual model. It also resulted in a more accurate and precise global likelihood response surface and in a considerable increase in the integrated model likelihoods. These results are in full agreement with those obtained for a multi-model analysis of a real aquifer [*Rojas et al.*, 2010]. The gain in accuracy and precision by the spatial conditioning of the hydraulic conductivity fields, however, was partially counterbalanced by deviating predictions of a particular member of the ensemble (**M**) that could not be excluded given the conditioning data available. The presence of such "wrong" conceptualizations resulted in high between-model variances. As a consequence, no clear relationship between the degree of spatial conditioning and the contribution of conceptual model uncertainty to the predictive variance could be detected.

3. Including flow-related observations in the conditioning resulted in a strong reduction of the total predictive spread and more accurate predictions of head dependent variables. In particular, river discharge observations allowed for a better discrimination among the conceptualizations compared to only the groundwater flow observation. This shows that flow-related observations with a global character have a high potential to discriminate large-scale conductivity features, such as zonation or layering, and therefore seem highly valuable to reduce conceptual model uncertainty.

4. Although it was not the topic of this study, the analysis presented shows its potential to be used in the design of optimal data collection campaigns to further discriminate among alternative conceptualizations. For example, from the results of this work, an optimal dataset would require between 11 and 20 hydraulic conductivity measurements, 16 groundwater head, 1 groundwater flow and 1 river discharge observations. This scheme would produce similar model weights (used for multi-model aggregation) as a sampling scheme considering twice the number of hydraulic conductivity measurements and twice the number of groundwater head observations.

Acknowledgments. The first author thanks the Katholieke Universiteit Leuven (K.U.Leuven), Belgium, for providing financial support in the framework of IRO-PhD scholarships. This research was conducted utilizing high performance computational resources provided by the Katholieke Universiteit Leuven through the VIC Cluster, http://ludit.kuleuven.be/hpc. This work was greatly improved by the thorough and constructive reviews provided by the Associate Editor, Shlomo Neuman, and three anonymous reviewers.

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), Multimodel combination techniques for hydrologic forecasting: Application to distributed model intercomparison project results, *Journal of Hydrometeorology*, *7*(4), 755–768, doi:10.1175/JHM519.1.

Ajami, N., Q. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resources Research*, *43*(W01403), doi: 10.1029/2005WR004745.

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

Alcolea, A., J. Carrera, and A. Medina (2006), Pilot points method incorporating prior information for solving the groundwater flow inverse problem, *Advances in Water Resources*, *29*(11), 1678–1689, doi:10.1016/j.advwatres.2005.12.009.

Anderman, E., M. Holl, and E. Poeter (1996), Two-dimensional advective transport in ground-water flow parameter estimation, *Ground Water*, *34*(6), 1001–1009, doi: 10.1111/j.1745-6584.1996.tb02165.x.

Beven, K. (2006), A manifesto for the equifinality thesis, *Journal of Hydrology*, *320*(1–2), 18–36, doi:10.1016/j.jhydrol.2005.07.007.

Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrological Processes*, *6*(5), 279–283, doi: 10.1002/hyp.3360060305.

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, doi:10.1016/S0022-1694(01)00421-8.

Binley, A., and K. Beven (2003), Vadose zone flow model uncertainty as conditioned on geophysical data, *Ground Water*, *41*(2), 119–127, doi:10.1111/j.1745-

6584.2003.tb02576.x.

Box, G. (1980), Sampling and Bayes' inference in scientific modelling and robustness, *Journal of the Royal Statistical Society, Series A*, *143*(4), 383–430.

Bredehoeft, J. (2003), From models to performance assessment: The conceptualization problem, *Ground Water*, *41*(5), 571–577, doi:10.1111/j.1745-6584.2003.tb02395.x.

Bredehoeft, J. (2005), The conceptualization model problem–surprise, *Hydrogeology Journal*, *13*(1), 37–46, doi:10.1007/s10040-004-0430-5.

Brooks, S., and A. Gelman (1998), General methods for monitoring convergence of iterative simulations, *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Cacuci, D. (2003), *Sensitivity and uncertainty analysis: Theory*, first ed., 304 pp., Chapman & Hall/CRC, Florida.

Capilla, J., and C. Llopis-Albert (2009), Gradual conditioning of non-Gaussian transmissivity fields to flow and mass transport data: 1. Theory, *Journal of Hydrology*, *371*(1–4), 66–74, doi:10.1016/j.jhydrol.2009.03.015.

Carrera, J., and S. Neuman (1986a), Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information, *Water Resources Research*, *22*(2), 199–210.

Carrera, J., and S. Neuman (1986b), Estimation of aquifer parameters under transient and steady state conditions: 2. Uniqueness, stability, and solution algorithms, *Water Resources Research*, *22*(2), 211–227.

Carrera, J., and S. Neuman (1986c), Estimation of aquifer parameters under transient and steady state conditions: 3. Application to synthetic and field data, *Water Resources Research*, *22*(2), 228–242.

Chib, S., and E. Greenberg (1995), Understanding the Metropolis–Hastings algorithm, *The American Statistician*, *49*(4), 327–335.

Cooley, R., and R. Naff (1990), Regression modeling of ground-water flow, *Techniques of Water-Resources Investigations Book 3, Chap. B4*, United States Geological Survey.

Cooley, R., L. Konikow, and R. Naff (1986), Nonlinear regression groundwater flow modeling of a deep regional aquifer system, *Water Resources Research*, *22*(13), 1759–1778.

Cowles, M., and B. Carlin (1996), Markov chain Monte Carlo convergence diagnostics: A comparative review, *Journal of the American Statistical Association*, *91*(434), 883–904.

Dagan, G. (1985), Stochastic modelling of groundwater flow by unconditional and conditional probabilities: the inverse problem, *Water Resources Research*, *21*(1), 65–72.

Dagan, G. (1989), *Flow and transport in porous formations*, first ed., 470 pp., Springer-Verlag, Berlin.

Dagan, G., and S. Neuman (1997), *Subsurface flow and transport: A stochastic approach*, first ed., 256 pp., Cambridge University Press, Cambridge.

Delhomme, J. (1979), Spatial variability and uncertainty in groundwater flow model parameters: A geostatistical approach, *Water Resources Research*, *15*(2), 269–280.

Deutsch, C., and A. Journel (1998), *GSLIB: Geostatistical Software Library and User's Guide*, second ed., 384 pp., Oxford University Press, New York.

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

Duan, Q., N. Ajami, X. Gao, and S. Sorooshian (2007), Multi–model ensemble hydrologic prediction using Bayesian model averaging, *Advances in Water Resources*, *30*(5), 1371–1386, doi:10.1016/j.advwatres.2006.11.014.

Ezzedine, S., and Y. Rubin (1996), A geostatistical approach to the conditional estimation of spatially distributed solute concentration and notes on the use of tracer data in the inverse problem, *Water Resources Research*, *32*(4), 853–861.

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, and P. Diggle (2002), Bayesian methodology to stochastic capture zone determination: Conditioning on transmissivity measurements, *Water Resources Research*, *38*(9), doi:10.1029/2001WR000950.

Feyen, L., J. Gómez-Hernández, P. Ribeiro, K. Beven, and F. De Smedt (2003), A Bayesian approach to stochastic capture zone delineation incorporating tracer arrival times, conductivity measurements, and hydraulic head observations, *Water Resources Research*, *39*(5), 1126–1138, doi:10.1029/2002WR001544.

Fienen, M., T. Clemo, and P. Kitanidis (2008), An interactive Bayesian geostatistical inverse protocol for hydraulic tomography, *Water Resources Research*, *44*(W00B01), doi:10.1029/2007WR006730.

Fienen, M., R. Hunt, D. Krabbenhoft, and T. Clemo (2009), Obtaining parsimonious hydraulic conductivity fields using head and transport observations: A Bayesian geostatistical parameter estimation approach, *Water Resources Research*, *45*(W08405), doi:10.1029/2008WR007431.

Foglia, L., M. Hill, S. Mehl, and P. Burlando (2009), Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function, *Water Resources Research*, *45*(W06427), doi:10.1029/2008WR007255.

Freeze, R. (1975), A stochastic-conceptual analysis of one-dimensional groundwater flow in non-uniform, homogeneus media, *Water Resources Research*, *11*(5), 725–741.

Gallagher, M., and J. Doherty (2007), Parameter estimation and uncertainty analysis for a watershed model, *Environmental Modelling & Software*, *22*(7), 1000–1020, doi: 10.1016/j.envsoft.2006.06.007.

Gelhar, L. (1993), *Stochastic subsurface hydrology*, first ed., 390 pp., Prentice-Hall, Inc., New Jersey.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004), *Bayesian data analysis*, second ed., 696 pp., Chapman & Hall/CRC, New York.

Georgakakos, K., D. Seo, H. Gupta, J. Schaake, and M. Butts (2004), Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *Journal of Hydrology*, *298*(1–4), 222–241, doi:10.1016/j.jhydrol.2004.03.037.

Geyer, C. (1992), Practical Markov chain Monte Carlo, *Statistical Science*, *7*(4), 473–483.

Ghosh, J., M. Delampady, and T. Samanta (2006), *An introduction to Bayesian analysis–Theory and methods*, first ed., 352 pp., Springer, New York.

Gilks, W., S. Richardson, and D. Spiegelhalter (1995), *Markov Chain Monte Carlo in practice*, first ed., 512 pp., Chapman & Hall/CRC, Boca Raton, Florida, USA.

Gómez-Hernández, J. (2006), Complexity, *Ground Water*, *44*(6), 782–785, doi: 10.1111/j.1745-6584.2006.00222.x.

Gómez-Hernández, J., A. Sahuquillo, and J. Capilla (1997), Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data: I. Theory, *Journal of Hydrology*, *203*(1–4), 162–174, doi:doi:10.1016/S0022-1694(97)00098-X.

Goovaerts, P. (1997), *Geostatistics for natural resources evaluation*, first ed., 483 pp., Oxford University Press, New York.

Gutjahr, A., and J. Wilson (1989), Co-kriging for stochastic flow models, *Transport in Porous Media*, *4*(6), 585–598.

Gutjahr, A., B. Bullard, S. Hatch, and L. Hughson (1994), Joint conditional simulations and the spectral approach for flow modeling, *Stochastic Hydrology and Hydraulics*, *8*(1), 79–108, doi:10.1007/BF01581391.

Hanna, S., and T. Yeh (1998), Estimation of co-conditional moments of transmissivity, hydraulic head and velocity fields, *Advances in Water Resources Research*, *22*(1), 87–95.

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, *Open file rep., 00-92*, United States Geological Survey, Reston, Virginia, USA.

Harrar, W., T. Sonnenberg, and H. Henriksen (2003), Capture zone, travel time, and solute transport predictions using inverse modelling and different geological models, *Hydrogeology Journal*, *11*(5), 536–548, doi:10.1007/s10040-003-0276-2.

Hassan, A., H. Bekhit, and J. Chapman (2008), Uncertainty assessment of a stochastic groundwater flow model using GLUE analysis, *Journal of Hydrology*, *362*(1–2), 89–109, doi:10.1016/j.jhydrol.2008.08.017.

Hastings, W. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, *57*(1), 97–109, doi:10.1093/biomet/57.1.97.

Hendricks Franssen, H., J. Gómez-Hernández, and A. Sahuquillo (2003), Coupled inverse modelling of groundwater flow and mass transport and the worth of concentration data,

Journal of Hydrology, *281*(4), 281–295, doi:10.1016/S0022-1694(03)00191-4.

Hill, M. (2006), The practical use of simplicity in developing ground water models, *Ground Water*, *44*(6), 775–781, doi:10.1111/j.1745-6584.2006.00227.x.

Hill, M., and C. Tiedeman (2007), *Effective groundwater model calibration: With analysis of data, sensitivities, predictions and uncertainty*, first ed., 480 pp., John Wiley & Sons, Inc., New Jersey.

Hill, M., R. Cooley, and D. Pollock (1998), A controlled experiment in ground water flow model calibration, *Ground Water*, *36*(3), 520–535, doi:10.1111/j.1745-6584.1998.tb02824.x.

Hoeksema, R., and P. Kitanidis (1984), An application of the geostatistical approach to the inverse problem in two-dimensional groundwater modeling, *Water Resources Research*, *20*(7), 1003–1020.

Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999), Bayesian model averaging: A tutorial, *Statistical Science*, *14*(4), 382–401.

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

Hsu, K., H. Moradkhani, and S. Sorooshian (2009), A sequential Bayesian approach for hydrologic model selection and prediction, *Water Resources Research*, *45*(W00B12), doi:10.1029/2008WR006824.

Hu, L. (2000), Gradual deformation and iterative calibration of Gaussian-related stochastic models, *Mathematical Geology*, *32*(1), 87–108.

Hunt, R., J. Doherty, and M. Tonkin (2007), Are models too simple? Arguments for increased parameterization, *Ground Water*, *45*(3), 254–262, doi:10.1111/j.1745-

6584.2007.00316.x.

Ijiri, Y., H. Saegusa, A. Sawada, M. Ono, K. Watanabe, K. Karasaki, C. Doughty, M. Shimo, and K. Fumimura (2009), Evaluation of uncertainties originating from the different modeling approaches applied to analyze regional groundwater flow in the Tono area of Japan, *Journal of Contaminant Hydrology*, *103*(3–4), 168–181, doi: 10.1016/j.jconhyd.2008.10.010.

Jensen, J. (2003), Parameter and uncertainty estimation in groundwater modelling, Ph.D. thesis, Department of Civil Engineering. Aalborg University, Denmark, series Paper Nr. 23.

Kass, R., and A. Raftery (1995), Bayes factors, *Journal of the American Statistical Association*, *90*(430), 773–795.

Kavetski, D., and G. Kuczera (2007), Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration, *Water Resources Research*, *43*(W03411), doi:10.1029/2006WR005195.

Kitanidis, P., and E. Vomvoris (1983), A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations, *Water Resources Research*, *19*(3), 677–690.

Krzysztofowicz, R. (1999), Bayesian theory of probabilistic forecasting via deterministic hydrologic model, *Water Resources Research*, *35*(9), 2739–2750.

LaVenue, A., B. RamaRao, G. de Marsily, and M. Marietta (1995), Pilot points methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields 2. Application, *Water Resources Research*, *31*(3), 495–516.

Leamer, E. (1978), *Specification searches: Ad hoc inference with nonexperimental data*, first ed., 384 pp., John Wiley & Sons, New York.

Li, X., and F. Tsai (2009), Bayesian model averaging for groundwater head prediction and uncertainty analysis using multimodel and multimethod, *Water Resources Ressearch*, *45*(W09403), doi:10.1029/2008WR007488.

Llopis-Albert, C., and J. Capilla (2009), Gradual conditioning of non-Gaussian transmissivity fields to flow and mass transport data: 2. Demonstration on a synthetic aquifer, *Journal of Hydrology*, *371*(1–4), 53–65, doi:10.1016/j.jhydrol.2009.03.014.

Makowski, D., D. Wallach, and M. Tremblay (2002), Using a Bayesian approach to parameter estimation; comparison of the GLUE and MCMC methods, *Agronomie*, *22*(2), 191–203, doi:10.1051/agro:2002007.

Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *Journal of Hydrology*, *330*(1–2), 368–381, doi: 10.1016/j.jhydrol.2006.04.046.

McLaughlin, D., and E. Wood (1988), A distributed parameter approach for evaluating the accuracy of groundwater model predictions: 2. Application to groundwater flow, *Water Resources Research*, *24*(7), 1048–1060.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equation of state calculations by fast computing machines, *The Journal of Chemical Physics*, *21*(6), 1087–1092, doi:10.1063/1.1699114.

Meyer, P., M. Ye, S. Neuman, and K. Cantrell (2004), Combined estimation of hydrogeologic conceptual model and parameter uncertainty, *Report nureg/cr-6843 pnnl-14534*, US Nuclear Regulatory Commission, Washington, US.

Meyer, P., M. Ye, M. Rockhold, S. Neuman, and K. Cantrell (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, US.

Moore, C., and J. Doherty (2005), Role of the calibration process in reducing model predictive error, *Water Resources Research*, *41*(W05020), doi:10.1029/2004WR003501.

Morse, B., 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), 1151–1163, doi: 10.1029/2002WR001470.

Neuman, S. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stochastic Environmental Research and Risk Assessment*, *17*(5), 291–305, doi: 10.1007/s00477-003-0151-7.

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

Oliver, D., L. Cunha, and A. Reynolds (1997), Markov Chain Monte Carlo methods for conditioning a permeability field to pressure data, *Mathematical Geology*, *29*(1), 61–91, doi:10.1007/BF02769620.

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

Pasquier, P., and D. Marcotte (2006), Steady- and transient-state inversion in hydrogeology by successive flux estimation, *Advances in Water Resources*, *29*(12), 1934–1952, doi:10.1016/j.advwatres.2006.02.001.

Poeter, E., and D. Anderson (2005), Multimodel ranking and inference in ground water modelling, *Ground Water*, *43*(4), 597–605, doi:10.1111/j.1745-6584.2005.0061.x.

Poeter, E., and M. Hill (1997), Inverse models: A necessary step in Ground–Water modeling, *Ground Water*, *35*(2), 250–260, doi:10.1111/j.1745-6584.1997.tb00082.x.

Poeter, E., and S. McKenna (1995), Reducing uncertainty associated with groundw-water flow and transport predictions, *Ground Water*, *33*(6), 899–904, doi:10.1111/j.1745-6584.1995.tb00034.x.

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

RamaRao, B., A. LaVenue, G. de Marsily, and M. Marietta (1995), Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields. 1. Theory and computational experiments, *Water Resources Research*, *31*(3), 475–493.

Refsgaard, J., J. Van der Sluijs, A. Højberg, and P. Vanrolleghem (2005), Uncertainty analysis. Harmoni-CA Guidance N0. 1, *Tech. rep.*

Refsgaard, J., J. Van der Sluijs, J. Brown, and P. Van der Keur (2006), A framework for dealing with uncertainty due to model structure error, *Advances in Water Resources*, *29*(11), 1586–1597, doi:10.1016/j.advwatres.2005.11.013.

Refsgaard, J., J. Van der Sluijs, A. Højberg, and P. Vanrolleghem (2007), Uncertainty in the environmental modelling process–A framework and guidance, *Environmental Mod-*

elling & Software, *22*(11), 1543–1556, doi:10.1016/j.envsoft.2007.02.004.

Renard, P. (2007), Stochastic hydrogeology: What professionals really need?, *Ground Water*, *45*(5), 531–541, doi:10.1111/j.1745-6584.2007.00340.x.

Robert, C. (2007), *The Bayesian Choice–From decision–theoretic foundations to computational implementation*, second ed., 577 pp., Springer-Verlag, New York.

Rojas, R., L. Feyen, and A. Dassargues (2008a), Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, *Water Resources Research*, *44*(W12418), doi:10.1029/2008WR006908.

Rojas, R., S. Kahunde, L. Peeters, O. Batelaan, L. Feyen, and A. Dassargues (2008b), Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling, *Journal of Hydrology*, *Submitted, under review.*

Rojas, R., L. Feyen, and A. Dassargues (2009), Sensitivity analysis of prior model probabilities and the value of prior knowledge in the assessment of conceptual model uncertainty in groundwater modelling, *Hydrological Processes*, *23*(8), 1131–1146, doi: 10.1002/hyp.7231.

Rojas, R., O. Batelaan, L. Feyen, and A. Dassargues (2010), Assessment of conceptual model uncertainty for the regional aquifer Pampa del Tamarugal - North Chile, *Hydrology and Earth System Sciences*, *14*(2), 171–192.

Rubin, Y. (2003), *Applied stochastic hydrogeology*, first ed., 416 pp., Oxford University Press, New York.

Rubin, Y., and G. Dagan (1987), Stochastic identification of transmissivity and effective recharge in steady groundwater flow 1. Theory, *Water Resources Research*, *23*(7), 1185–1192.

Sahuquillo, A., J. Capilla, J. Gómez-Hernández, and J. Andreu (1992), Conditional simulation of transmissivity fields honouring piezometric data, in *Hydraulic Engineering Software IV. Fluid Flow Modelling*, edited by W. Blain and E. Cabrera, pp. 210–214, Kluwer Academic Publishers.

Saltelli, A., S. Tarantola, F. Campolango, and M. Ratto (2008), *Sensitivity analysis in practice: A guide to assessing scientific models*, first ed., 232 pp., John Wiley & Sons, Chichester.

Seifert, D., T. Sonnenberg, P. Scharling, and K. Hinsby (2008), Use of alternative conceptual models to assess the impact of a buried valley on groundwater vulnerability, *Hydrogeology Journal*, *16*(4), 659–674, doi:10.1007/s10040-007-0252-3.

Singh, A., S. Mishra, and G. Ruskauff (2009), Model averaging techniques for quantifying conceptual model uncertainty, *Ground Water*, *in press*, doi:10.1111/j.1745-6584.2009.00642.x.

Sohn, M., M. Small, and M. Pantazidou (2000), Reducing uncertainty in site characterization using Bayes Monte Carlo methods, *Journal of Environmental Engineering–ASCE*, *126*(10), 893–902, doi:10.1061/(ASCE)0733-9372(2000)126:10(893).

Sorensen, D., and D. Gianola (2002), *Likelihood, Bayesian, and MCMC methods in quantitative genetics*, vol. I, first ed., 740 pp., Springer-Verlag, New York.

Tiedeman, C., D. Goode, and P. Hsieh (1997), Numerical simulation of ground-water flow through glacial deposits and crystalline bedrock in the Mirror Lake area, Grafton County, New Hampshire, *Professional Paper 1572*, United States Geological Survey, U.S., Reston, Virginia.

Tiedeman, C., D. Goode, and P. Hsieh (1998), Characterizing a ground water basin in a New England mountain and valley terrain, *Ground Water*, *36*(4), 611–620, doi: 10.1111/j.1745-6584.1998.tb02835.x.

Tiedeman, C., M. Hill, F. D'Agnese, and C. Faunt (2003), Methods for using groundwater model predictions to guide hydrogeologic data collection, with application to the Death Valley regional groundwater flow system, *Water Resources Research*, *39*(1010), doi: 10.1029/2001WR001255.

Tiedeman, C., D. Ely, M. Hill, and G. O'Brien (2004), A method for evaluating the importance of system state observations to model predictions, with application to the death valley regional groundwater flow system, *Water Resources Research*, *40*(W12411), doi:10.1029/2004WR003313.

Tierney, L. (1994), Markov chains for exploring posterior distributions, *The Annals of Statistics*, *22*(4), 1701–1728.

Tonkin, M., C. Tiedeman, D. Ely, and M. Hill (2007), OPR-PPR, a computer program for assessing data importance to model predictions using linear statistics, *Techniques and Methods 6-E2*, U.S. Geological Survey.

Troldborg, L., J. Refsgaard, K. Jensen, and P. Engesgaard (2007), The importance of alternative conceptual models for simulation of concentrations in a multi–aquifer system, *Hydrogeology Journal*, *15*(5), 843–860, doi:10.1007/s10040-007-0192-y.

Tsai, F., and X. Li (2008), Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window, *Water Resources Research*, *44*(W09434), doi:10.1029/2007WR006576.

Tsai, F., and X. Li (2010), Reply to comment by Ming Ye et al. on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window", *Water Resources Research*, *46*(W02802), doi: 10.1029/2009WR008591.

Van der Sluijs, J. (2005), Uncertainty as a monster in the science–policy interface: Four coping strategies, *Water Science & Technology*, *52*(6), 87–92.

Vrugt, J., and B. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resources Research*, *43*(W01411), doi:10.1029/2005WR004838.

Vrugt, J., M. Clarck, C. Diks, Q. Duan, and B. Robinson (2006), Multi-objective calibration of forecast ensembles using Bayesian model averaging, *Geophysical Research Letters*, *33*(L19817), doi:10.1029/2006GL027126.

Wagner, B. (1995), Sampling design methods for groundwater modeling under uncertainty, *Water Resources Research*, *31*(10), 2581–2591.

Walker, W., and V. Marchau (2003), Dealing with uncertainty in policy analysis and policy making, *Integrated Assessment*, *4*(1), 1–4.

Wasserman, L. (2000), Bayesian model selection and model averaging, *Journal of Mathematical Psychology*, *44*(1), 92–107, doi:10.1006/jmps.1999.1278.

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

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.

Ye, M., K. Pohlmann, and J. Chapman (2008a), Expert elicitation of recharge model probabilities for the Death Valley regional flow system, *Journal of Hydrology*, *354*(1–4), 102–115, doi:10.1016/j.jhydrol.2008.03.001.

Ye, M., P. Meyer, and S. Neuman (2008b), On model selection criteria in multimodel analysis, *Water Resources Research*, *44*(W03428), doi:10.1029/2008WR006803.

Ye, M., K. Pohlman, J. Chapman, G. Pohll, and D. Reeves (2009), A model-averaging method for assessing groundwater conceptual model uncertainty, *Ground Water*, *in press*, doi:10.1111/j.1745-6584.2009.00633.x.

Ye, M., D. Lu, S. Neuman, and P. Meyer (2010), Comment on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window" by Frank T.-C. Tsai and Xiaobao Li, *Water Resources Research*, *46*(W02801), doi:10.1029/2009WR008501.

Table 1. Parameters describing the spatial correlation structure (isotropic exponential covariance) of hydraulic conductivity for the different layers of the 3-dimensional hypothetical setup. μ_K is the mean hydraulic conductivity for the corresponding layer, σ_{lnK} is the standard deviation of the lnK values, and I_{lnK} is the integral scale defining the correlation length (practical range) (based on Tables 2.1 and 2.2 from *Rubin* [2003]).

Layer	Model Parameters		
	μ_K $[md^{-1}]$	σ_{lnK}	I_{lnK}
1	0.1	2.0	400
2	0.01	0.5	800
3	1	1.5	600

Table 2. Summary of the 12 conditioning cases analyzed in this work.

		System-state variables conditioning		
		HEADS	GWF	RIV
Spatial conditioning	Unconditional	32 head meas.	32 head meas. + 1 groundwater flow	32 head meas. + 1 groundwater flow + 1 river discharge
	Conditional-I	32 head meas. + 10 k meas.	32 head meas. + 1 groundwater flow + 10 k meas.	32 head meas. + 1 groundwater flow + 1 river discharge + 10 k meas.
	Conditional-II	32 head meas. + 20 k meas.	32 head meas. + 1 groundwater flow + 20 k meas.	32 head meas. + 1 groundwater flow + 1 river discharge + 20 k meas.
	Conditional-III	32 head meas. + 40 k meas.	32 head meas. + 1 groundwater flow + 40 k meas.	32 head meas. + 1 groundwater flow + 1 river discharge + 40 k meas.

Table 3. Description of 7 alternative conceptualizations defining **M** used to approximate the 3-dimensional hypothetical setup.

Conceptualization	Spatial correlation structure[1]	Description
1L-L1	Layer 1	One-layer model
1L-L2	Layer 2	One-layer model
1L-L3	Layer 3	One-layer model
1L-AVG	Average of layers 1, 2 and 3	One-layer model
2L	Layer 1 and 3	Two-layer model not considering the aquitard
2LQ3D	Layer 1 and 3	Two-layer model implicitly accounting for the aquitard
3L	Layer 1, 2 and 3	Three-layer model explicitly accounting for the aquitard

[1] Parameters defining the spatial correlation structure for each layer are presented in Table 1.

Table 4. Prior range to select starting locations of multiple chains of M-H algorithm.

Parameter			Range	
		True value	Minimum	Maximum
Recharge rate $[md^{-1}]$	(RECH)	1.4×10^{-4}	0	5.8×10^{-4}
Elevation west boundary condition $[m]$	(CH)	46	25	75
Elevation surface $[m]$	(SURF)	43	30	50
Extinction depth $[m]$	(EXTD)	5	0	25
Evaporation rate $[md^{-1}]$	(EVTR)	1.37×10^{-3}	0	7.0×10^{-3}
River conductance $[m^2d^{-1}]$	(RIVC)	5	1.0×10^{-2}	1000

Table 5. Summary of the integrated model likelihoods (equation 7) and posterior model probabilities (in parentheses) for the 7 conceptualizations comprised in **M** and the 12 conditioning cases.

Conditioning			Conceptual models							
			1L-L1	1L-L2	1L-L3	1L-AVG	2L	2LQ3D	3L	Total
K	Observations	$p(M_k)$	(1/7)	(1/7)	(1/7)	(1/7)	(1/7)	(1/7)	(1/7)	1.0
0	Heads	$p(\mathbf{D}\|M_k)$	0	0	746.8	765.0	802.9	804.9	852.1	3971.3
			0	0	(0.188)	(0.193)	(0.202)	(0.203)	(0.215)	(1.0)
	Heads + GWF	$p(\mathbf{D}\|M_k)$	0	0	620.0	590.3	739.5	751.2	799.0	3500.0
			0	0	(0.177)	(0.169)	(0.211)	(0.215)	(0.228)	(1.0)
	Heads + GWF + RIV	$p(\mathbf{D}\|M_k)$	0	0	592.5	507.5	680.9	696.0	783.7	3260.5
			0	0	(0.182)	(0.156)	(0.209)	(0.213)	(0.240)	(1.0)
10	Heads	$p(\mathbf{D}\|M_k)$	0	0	785.6	722.1	851.2	923.5	926.8	4209.1
			0	0	(0.187)	(0.172)	(0.202)	(0.219)	(0.220)	(1.0)
	Heads + GWF	$p(\mathbf{D}\|M_k)$	0	0	607.3	541.3	825.0	867.5	867.9	3709.0
			0	0	(0.164)	(0.146)	(0.222)	(0.234)	(0.234)	(1.0)
	Heads + GWF + RIV	$p(\mathbf{D}\|M_k)$	0	0	547.2	486.8	712.8	854.7	860.9	3462.4
			0	0	(0.158)	(0.141)	(0.206)	(0.247)	(0.249)	(1.0)
20	Heads	$p(\mathbf{D}\|M_k)$	0	0	754.1	0	861.9	942.7	962.7	3521.3
			0	0	(0.214)	0	(0.245)	(0.268)	(0.273)	(1.0)
	Heads + GWF	$p(\mathbf{D}\|M_k)$	0	0	652.4	0	865.7	918.9	926.8	3363.9
			0	0	(0.194)	0	(0.257)	(0.273)	(0.276)	(1.0)
	Heads + GWF + RIV	$p(\mathbf{D}\|M_k)$	0	0	539.4	0	798.6	911.7	925.0	3174.6
			0	0	(0.170)	0	(0.252)	(0.287)	(0.291)	(1.0)
40	Heads	$p(\mathbf{D}\|M_k)$	0	0	981.2	0	1148.8	1206.2	1208.8	4545.0
			0	0	(0.216)	0	(0.253)	(0.265)	(0.266)	(1.0)
	Heads + GWF	$p(\mathbf{D}\|M_k)$	0	0	781.5	0	1028.9	1063.7	1065.2	3939.2
			0	0	(0.198)	0	(0.261)	(0.270)	(0.270)	(1.0)
	Heads + GWF + RIV	$p(\mathbf{D}\|M_k)$	0	0	595.4	0	823.8	959.2	979.7	3358.2
			0	0	(0.177)	0	(0.245)	(0.286)	(0.292)	(1.0)

Table 6. Predictive variance for parameters and simulated system-state variables of interest. Values expressed in $[md^{-1}]^2$ for RECH, $[m]^2$ for CH, and $[m^3d^{-1}]^2$ for flow components. Shaded cells indicate a contribution of between-model variance larger than 33% of the predictive variance (i.e. within- + between-model variance).

		RECH		CH		GW inflows		Recharge inflows		River gains		EVT outflows	
K	Observations	within-model	between-model	within-model	between-model	within-model	between-model	within-model	between-model	within-model	between-model	within-model	between-model
20	Heads	2.18×10^{-9}	9.18×10^{-10}	5.30×10^{1}	1.00×10^{0}	3.69×10^{5}	3.09×10^{4}	4.85×10^{5}	2.05×10^{5}	3.34×10^{3}	2.04×10^{5}	3.51×10^{3}	
	Heads + GWF	5.60×10^{-10}	3.03×10^{-10}	1.70×10^{1}	3.84×10^{0}	3.93×10^{4}	1.24×10^{2}	1.25×10^{5}	6.75×10^{4}	2.04×10^{4}	1.10×10^{4}	9.21×10^{3}	7.49×10^{0}
	Heads + GWF + RIV	1.29×10^{-10}	6.23×10^{-10}	1.30×10^{1}	5.50×10^{0}	2.10×10^{4}	2.98×10^{3}	2.87×10^{4}	1.39×10^{5}	9.96×10^{2}	4.37×10^{0}	3.08×10^{3}	3.04×10^{1}
40	Heads	7.65×10^{-10}	2.43×10^{-10}	1.78×10^{1}	2.24×10^{-1}	2.32×10^{5}	4.02×10^{4}	1.70×10^{5}	5.41×10^{4}	1.95×10^{4}	1.40×10^{3}	1.46×10^{5}	7.66×10^{3}
	Heads + GWF	3.56×10^{-10}	7.45×10^{-11}	9.44×10^{0}	2.10×10^{0}	3.51×10^{4}	1.56×10^{3}	7.93×10^{4}	1.66×10^{4}	1.13×10^{4}	3.71×10^{3}	4.52×10^{3}	3.74×10^{1}
	Heads + GWF + RIV	7.79×10^{-11}	1.82×10^{-10}	7.08×10^{0}	3.07×10^{0}	1.60×10^{4}	7.93×10^{3}	1.74×10^{4}	4.06×10^{4}	9.82×10^{2}	1.01×10^{0}	9.60×10^{2}	5.14×10^{0}

Figure 1. Three-dimensional hypothetical setup including 16 ($\odot$) observation wells and 8 ($\otimes$) pumping wells overlain by the "true" groundwater head distribution and hydraulic conductivity realizations used in the forward run (grid $25m \times 25m$) for (a) layer 1 and (c) layer 3.

Figure 2. Sampling locations for the 3 conditioning cases: Conditional-I (10 K measurements), Conditional-II (20 K measurements) and Conditional-III (40 K measurements).

Figure 3. Results from the M-H algorithm for parameter RECH for model 3L under case Conditional-III: (a) 10 independent chains after discarding the *burn-in* samples and (b) autocorrelogram of original ($60,000$ elements) and thinned samples ($10,000$ elements).

Figure 4. Scatter plots of global likelihood values for parameter RECH for alternative conceptual models 1L-L3 (a,d), 2L (b,e) and 3L (c,f) for cases Unconditional (a-c) and Conditional-III (d-f). Vertical dashed lines represent the true value used in the 3-dimensional hypothetical setup.

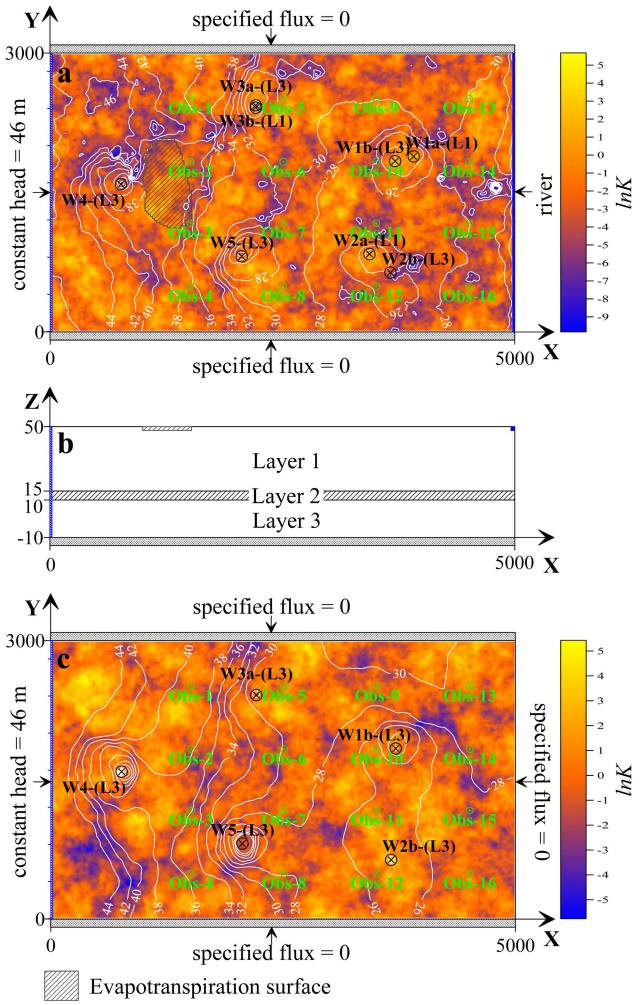
Figure 5. Scatter plots of global likelihood values for groundwater inflows from the WBC (a-c), river gains (d-f) and evapotranspiration (EVT) outflows (g-i) for conceptualization 3L under case Unconditional. Vertical dashed lines represent the reference values obtained from the 3-dimensional hypothetical setup ($655m^3d^{-1}, 192m^3d^{-1}, and 63m^3d^{-1}$). For EVT outflows (g-i) most likelihood values are located on the y-axis or around zero.

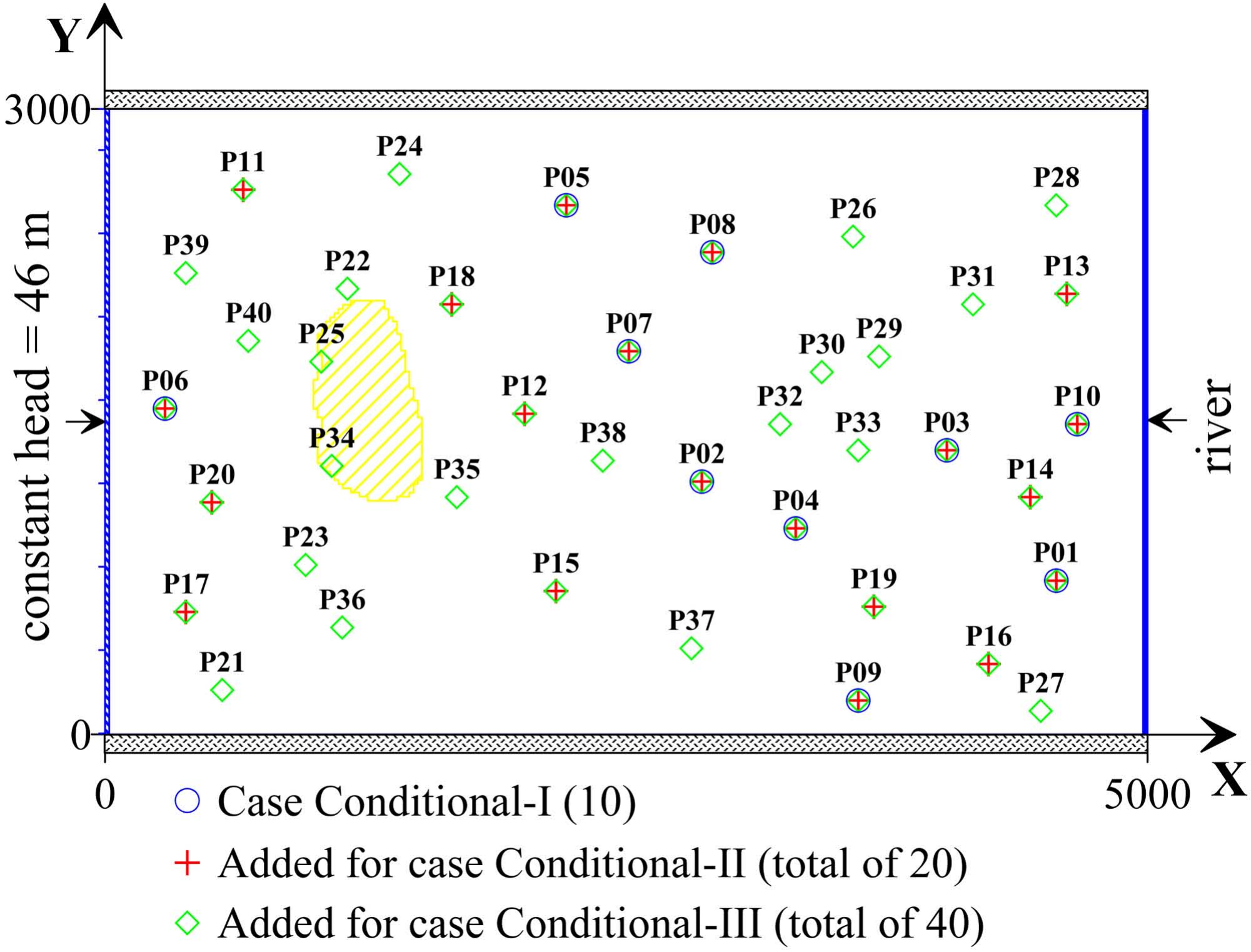
Figure 6. Cumulative probability distributions of recharge inflows (a,c) and evapotranspiration (EVT) outflows (b,d) for alternative conceptual models and Bayesian Model Averaging (BMA) for case Unconditional (a-b) and Conditional-III (c-d) based on groundwater heads conditioning only. Vertical dashed lines represent the reference values obtained from the 3-dimensional hypothetical setup ($2100m^3d^{-1} and 63m^3d^{-1}$).

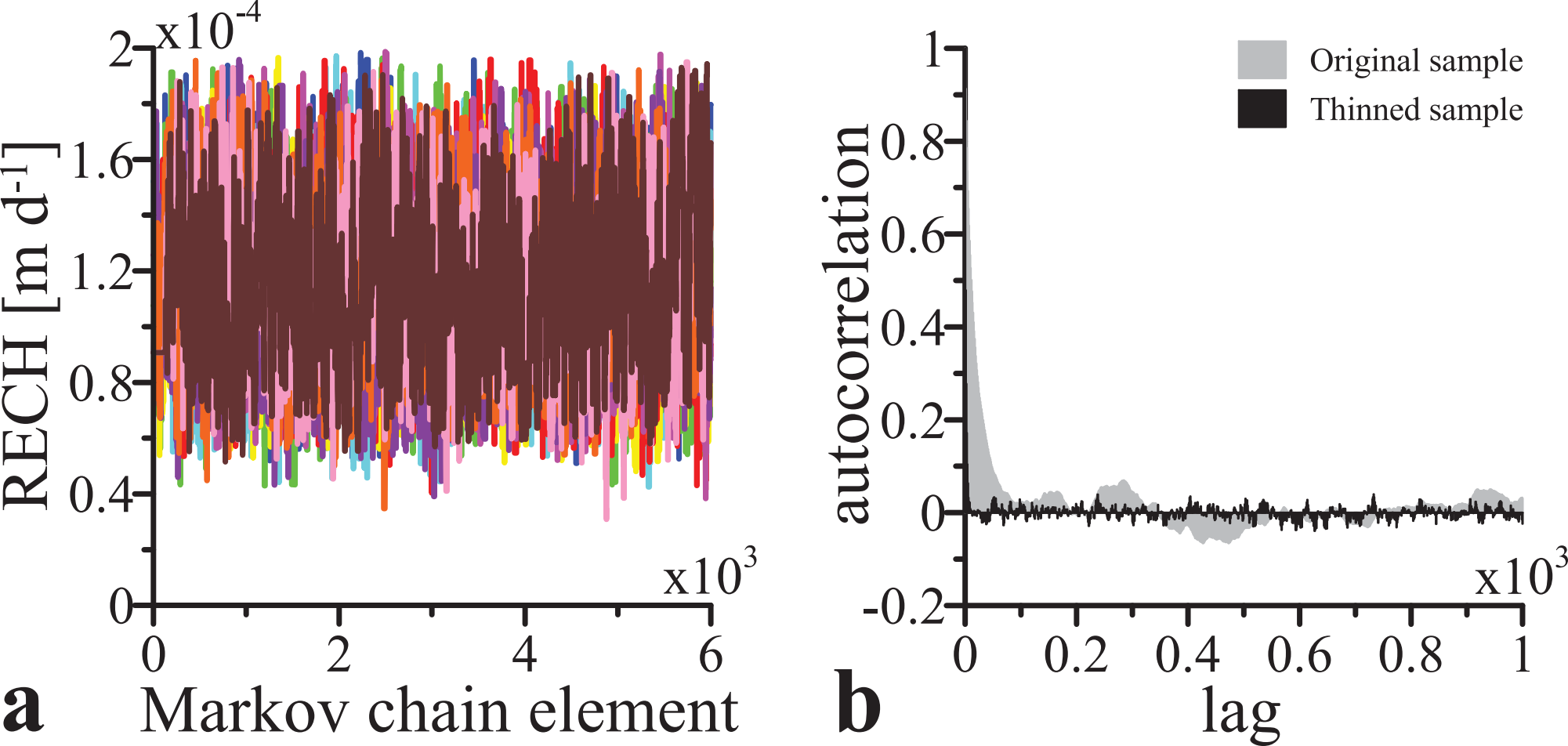
Figure 7. (a) Cumulative probability distributions of groundwater inflows from the west boundary condition (WBC) for the cases Unconditional and Conditional-III for conceptual model 1L-L3; (b) longitudinal groundwater head profiles crossing the middle point of the evapotranspiration zone depicted in Figure 1 for cases Unconditional and Conditional-III. Vertical dashed line represents the reference value obtained from the 3-dimensional hypothetical setup ($655 m^3 d^{-1}$).

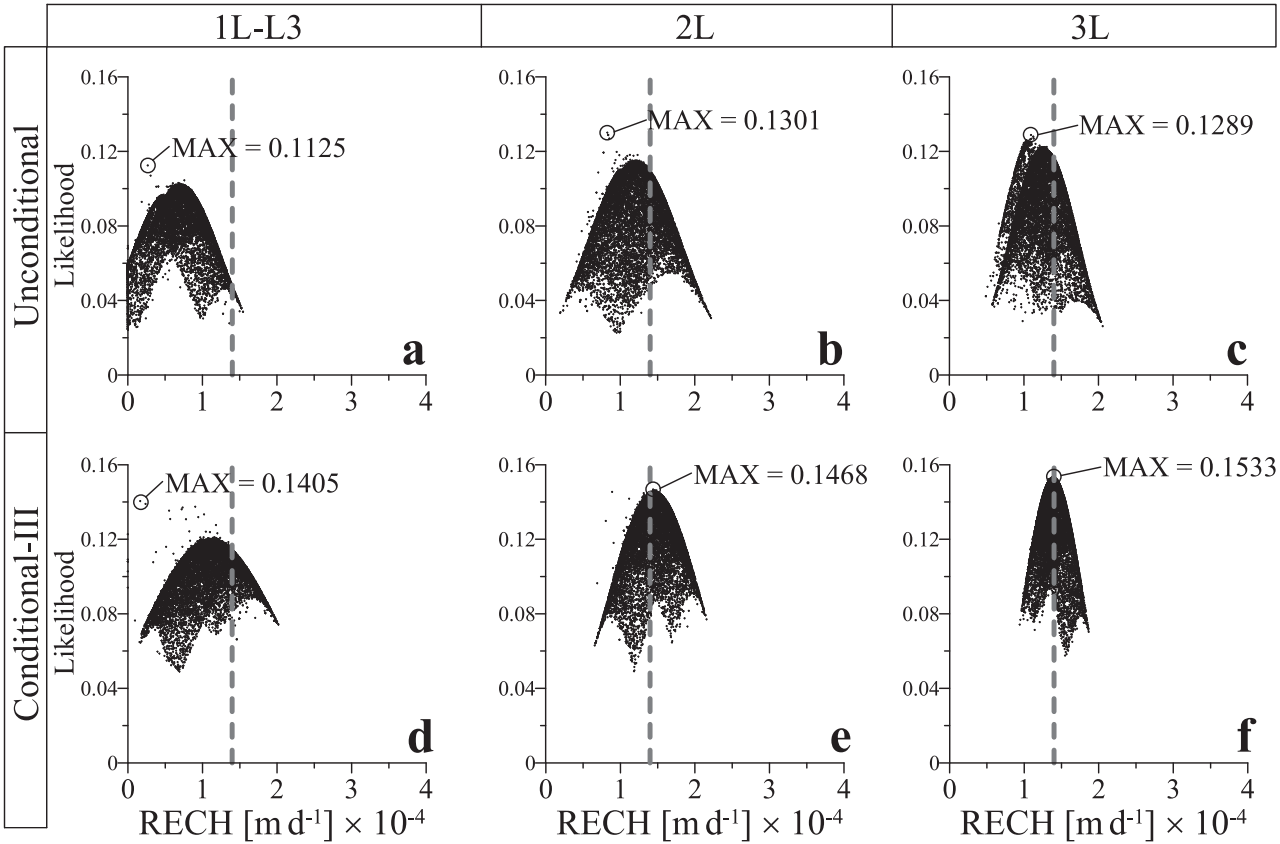
Figure 8. Cumulative BMA probability distributions of groundwater inflows from the WBC (a,e), groundwater outflows from the west boundary condition (WBC) (b,f), river gains (c,g), and evapotranspiration (EVT) outflows (d,h) for cases Unconditional (a-d) and Conditional-III (e-h) based on groundwater heads, one GWF observation and one RIV observation. Vertical dashed lines represent the reference values obtained from the 3-dimensional hypothetical setup ($655 m^3 d^{-1}, 0 m^3 d^{-1}, 192 m^3 d^{-1}, and 63 m^3 d^{-1}$).

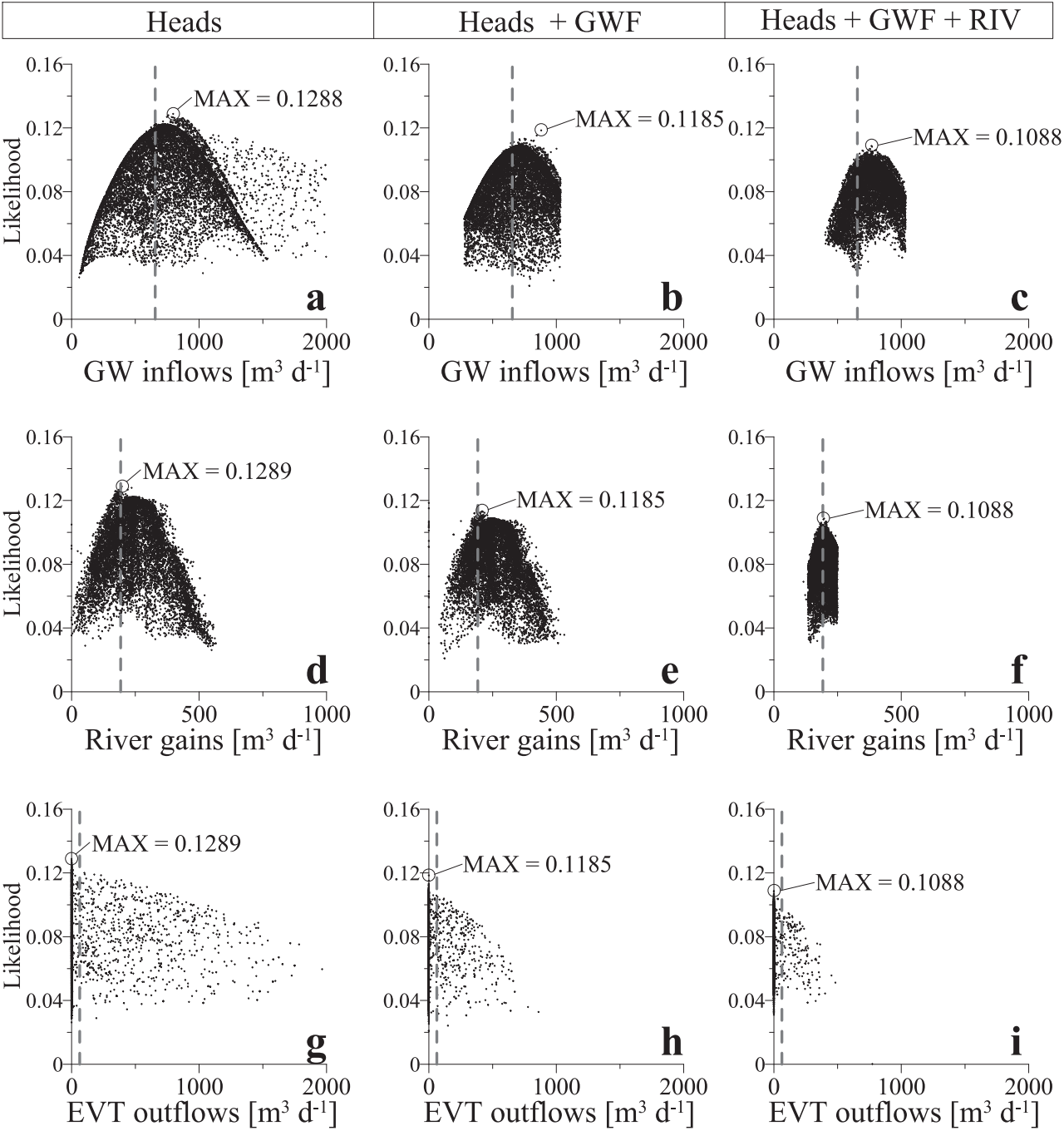
Figure 9. Predictive variance estimated using equation (6) for recharge inflows (a-b) and evapotranspiration (EVT) outflows (c-d) for conditioning cases Conditional-II (a, c) and Conditional-III (b, d). Height of columns shown in Table 6.

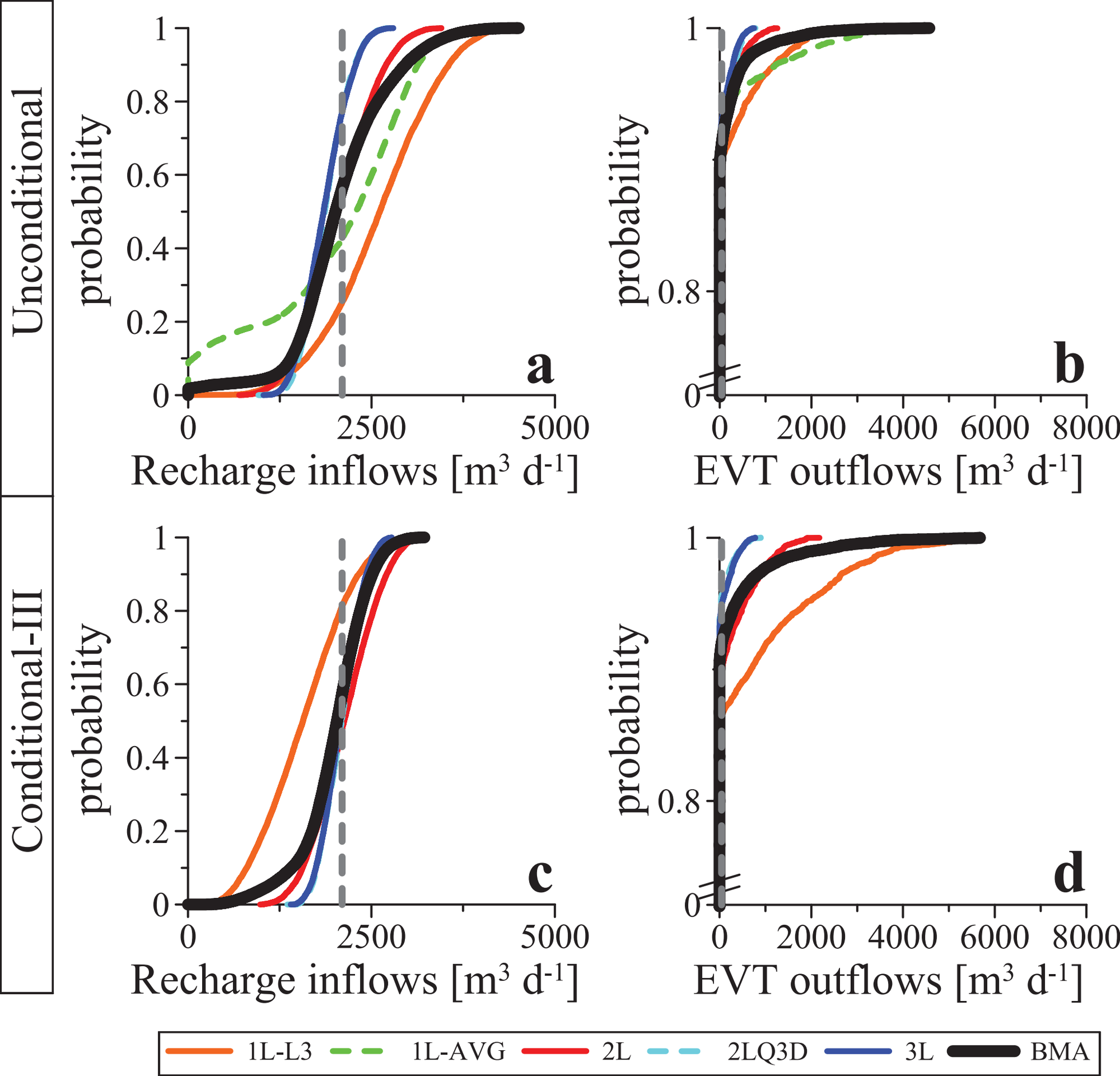

Y
specified flux = 0
3000
a
W3a-(L3)
W3b-(L1)
Obs-1
Obs-5
Obs-9
Obs-13
W1b-(L3)
W1a-(L1)
Obs-2
Obs-6
Obs-10
Obs-14
constant head = 46 m
W4-(L3)
river
W5-(L3)
W2a-(L1)
W2b-(L3)
Obs-3
Obs-7
Obs-11
Obs-15
Obs-4
Obs-8
Obs-12
Obs-16
0
5000
X
specified flux = 0
lnK
Z
50
b
Layer 1
15
Layer 2
10
Layer 3
-10
c
specified flux = 0
Evapotranspiration surface

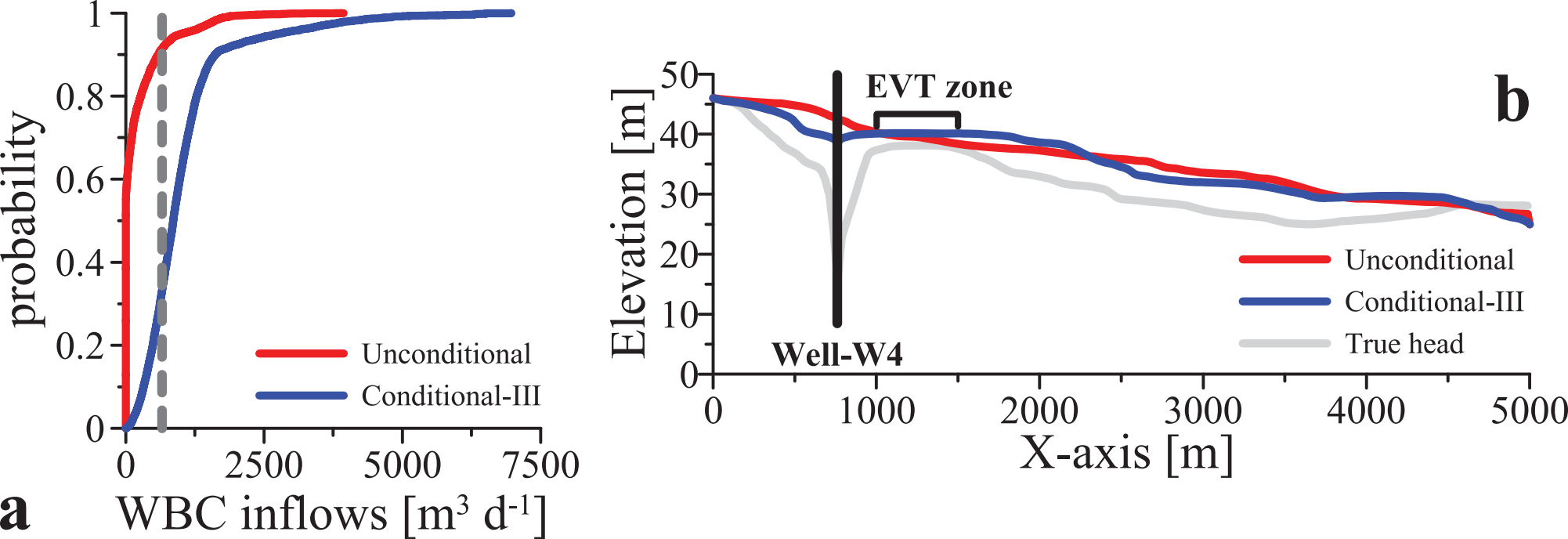

Y
3000
0
constant head = 46 m
river
0
5000
X
P11
P24
P05
P28
P26
P08
P39
P22
P18
P31
P13
P40
P25
P07
P30
P29
P06
P12
P32
P10
P34
P38
P33
P03
P02
P35
P20
P14
P04
P23
P01
P15
P19
P17
P36
P37
P16
P21
P09
P27
Case Conditional-I (10)
Added for case Conditional-II (total of 20)
Added for case Conditional-III (total of 40)

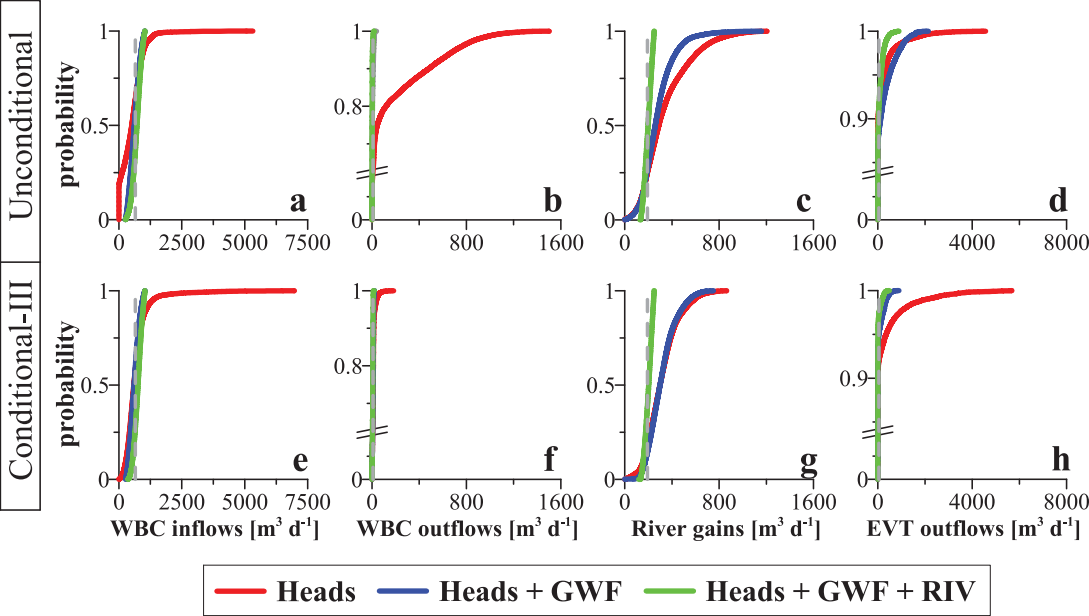

x10-4
2
1.6
1.2
0.8
0.4
0
RECH [m d-1]
x10^3
0
2
4
6
a
Markov chain element
1
0.8
0.6
0.4
0.2
0
-0.2
autocorrelation
Original sample
Thinned sample
x10^3
0
0.2
0.4
0.6
0.8
1
b
lag

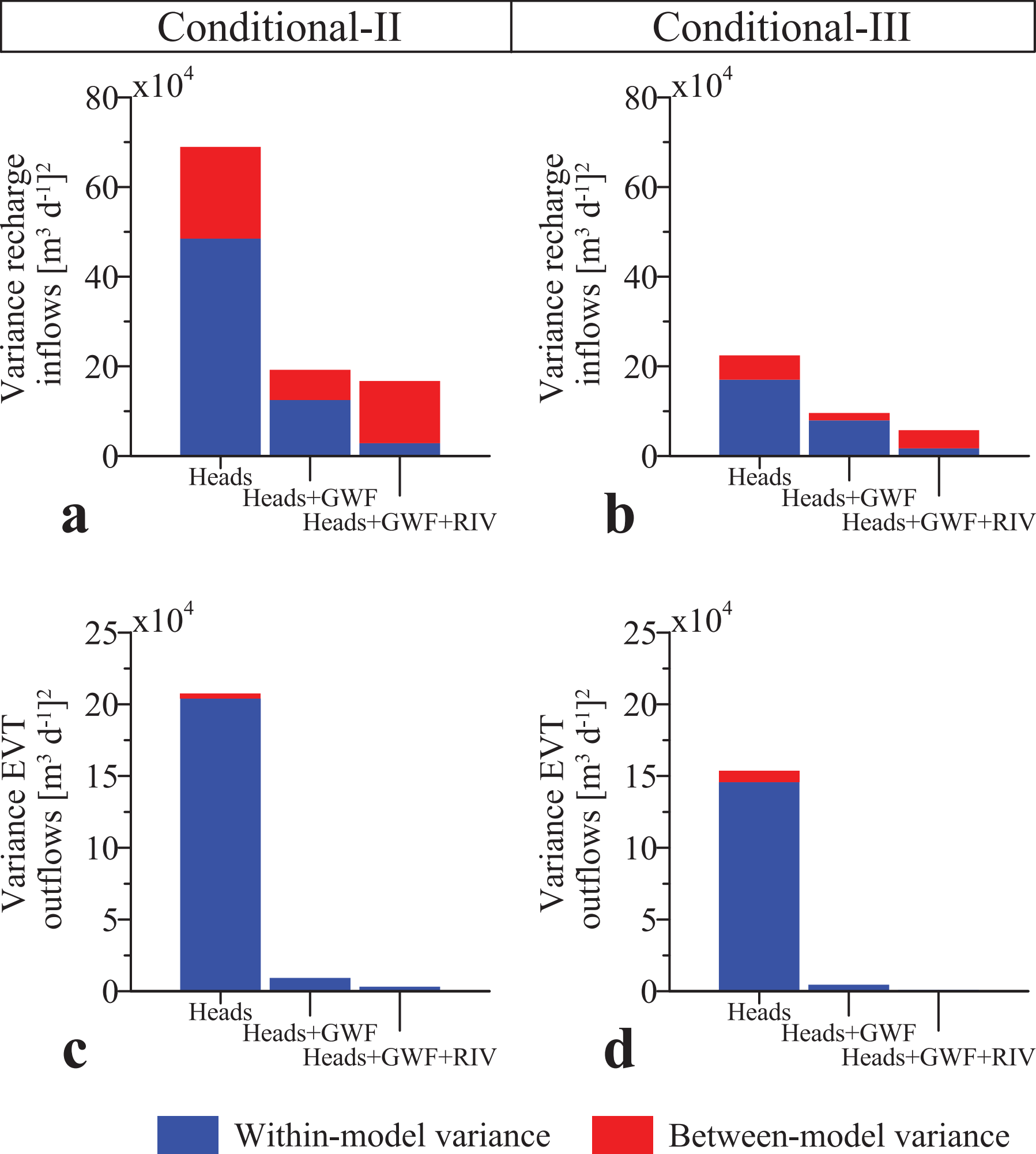

1L-L3
2L
3L
Unconditional
Conditional-III
Likelihood
MAX = 0.1125
MAX = 0.1301
MAX = 0.1289
MAX = 0.1405
MAX = 0.1468
MAX = 0.1533
a
b
c
d
e
f
0.16
0.12
0.08
0.04
0
1
2
3
4
RECH [m d-1] × 10-4


Heads
Heads + GWF
Heads + GWF + RIV
Likelihood
MAX = 0.1288
MAX = 0.1185
MAX = 0.1088
GW inflows [m³ d⁻¹]
a
b
c
MAX = 0.1289
MAX = 0.1185
MAX = 0.1088
River gains [m³ d⁻¹]
d
e
f
MAX = 0.1289
MAX = 0.1185
MAX = 0.1088
EVT outflows [m³ d⁻¹]
g
h
i


Unconditional
Conditional-III
probability
Recharge inflows [m³ d⁻¹]
EVT outflows [m³ d⁻¹]
a
b
c
d
0
0.2
0.4
0.6
0.8
1
2500
5000
2000
4000
6000
8000
1L-L3
1L-AVG
2L
2LQ3D
3L
BMA


probability
1
0.8
0.6
0.4
0.2
0
Unconditional
Conditional-III
0
2500
5000
7500
a
WBC inflows [m³ d⁻¹]
Elevation [m]
50
40
30
20
10
0
EVT zone
Well-W4
b
Unconditional
Conditional-III
True head
0
1000
2000
3000
4000
5000
X-axis [m]


Unconditional
Conditional-III
probability
a
b
c
d
e
f
g
h
WBC inflows [m³ d⁻¹]
WBC outflows [m³ d⁻¹]
River gains [m³ d⁻¹]
EVT outflows [m³ d⁻¹]
Heads
Heads + GWF
Heads + GWF + RIV


Conditional-II
Conditional-III
x10^4
Variance recharge inflows [m^3 d^-1]^2
80
60
40
20
0
Heads
Heads+GWF
Heads+GWF+RIV
a
b
x10^4
Variance EVT outflows [m^3 d^-1]^2
25
20
15
10
5
0
Heads
Heads+GWF
Heads+GWF+RIV
c
d
Within-model variance
Between-model variance