Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling

Rodrigo Rojas[*a,1], Samalie Kahunde[b], Luk Peeters[a], Okke Batelaan[a,c] , Luc Feyen[d], and Alain Dassargues[a,e]

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

[b] Interuniversity Programme in Water Resources Engineering (IUPWARE), Katholieke Universiteit Leuven and Vrije Universiteit Brussel. Pleinlaan 2, B-1050 Brussels, Belgium. Tel: + 32 2 6293039, Fax: + 32 2 6293022.
E-mail address: Samalie.Kahunde@student.kuleuven.be

[c] Department of Hydrology and Hydraulic Engineering, Vrije Universiteit Brussel. Pleinlaan 2, B-1050 Brussels, Belgium. Tel: + 32 2 6293039, Fax: + 32 2 6293022.
E-mail address: Okke.Batelaan@geo.kuleuven.be

[d] Land management and natural hazards unit, Institute for Environment and Sustainability (IES), DG- Joint Research Centre (JRC), European Commission (EC).
Via Enrico Fermi 2749, TP261, I-21027, Ispra (Va), Italy.
Tel: +39 0332789258, Fax: +39 0332 786653
E-mail address: Luc.Feyen@jrc.ec.europa.eu

[e] Hydrogeology and Environmental Geology, Department of Architecture, Geology, Environment, and Constructions (ArGEnCo), Université de Liège. B.52/3 Sart-Tilman, B-4000 Liège, Belgium. Tel.: + 32 4 3662376, Fax: + 32 4 3669520.
E-mail address: Alain.Dassargues@geo.kuleuven.be

[1] Now at: Land management and natural hazards unit, Institute for Environment and Sustainability (IES), DG- Joint Research Centre (JRC), European Commission (EC).
Via Enrico Fermi 2749, TP261, I-21027, Ispra (Va), Italy.
Tel: +39 0332 78 97 13, Fax: +39 0332 78 66 53
E-mail address: Rodrigo.Rojas@jrc.ec.europa.eu

Manuscript submitted to Journal of Hydrology on September 2nd , 2009

Abstract

Groundwater models are often used to predict the future behaviour of groundwater systems. These models may vary in complexity from simplified system conceptualizations to more intricate versions. It has been recently suggested that uncertainties in model predictions are largely dominated by uncertainties arising from the definition of alternative conceptual models. Different external factors such as climatic conditions or groundwater abstraction policies, on the other hand, may also play an important role. Rojas et al. (2008) proposed a multimodel approach to account for predictive uncertainty arising from forcing data (inputs), parameters, and alternative conceptualizations. In this work we extend upon this approach to include uncertainties arising from the definition of alternative future scenarios and we improve the methodology by including a Markov Chain Monte Carlo sampling scheme. We apply the improved methodology to a real aquifer system underlying the Walenbos Nature Reserve area in Belgium. Three alternative conceptual models comprising different levels of geological knowledge are considered. Additionally, three recharge settings (scenarios) are proposed to evaluate recharge uncertainties. A joint estimation of the predictive uncertainty including parameter, conceptual model, and scenario uncertainties is estimated for groundwater budget terms. Finally, results obtained using the improved approach are compared with the results obtained from methodologies that include a calibration step and which use a model selection criterion to discriminate between alternative conceptualizations. Results showed that conceptual model and scenario uncertainties significantly contribute to the predictive variance for some budget terms. Besides, conceptual model uncertainties played an important role even for the case when a model was preferred over the others. Predictive distributions showed to be considerably different in shape, central moment, and spread among alternative conceptualizations and scenarios analyzed. This reaffirms the idea that relying on a single conceptual model driven by a particular scenario, will likely produce bias and under-dispersive estimations of the predictive uncertainty. Multimodel methodologies based on the use of model selection criteria produced ambiguous results. In the frame of a multimodel approach, these inconsistencies are critical and can not be neglected. These results strongly advocate the idea of addressing conceptual model uncertainty in groundwater modeling practice. Additionally, considering alternative future recharge uncertainties will permit to obtain more realistic and, possibly, more reliable estimations of the predictive uncertainty.

Keywords
Groundwater flow modelling, conceptual model uncertainty, scenario uncertainty, GLUE, Bayesian Model Averaging, Markov Chain Monte Carlo.

1. Introduction and scope

Groundwater models are often used to predict the behaviour of groundwater systems under future stress conditions. These models may vary in the level of complexity from simplified groundwater system representations to more elaborated models accounting for detailed descriptions of the main processes and geological properties of the groundwater system. Whether to postulate simplified or complex/elaborated models for solving a given problem has been subject of discussion for several years (Gómez-Hernández, 2006; Hill, 2006; Neuman and Wierenga, 2003). Parsimony is the main argument for those in favour of simpler models (see e.g. Hill and Tiedeman, 2007) whereas a *more realistic representation of the unknown true system* (see e.g. Rubin, 2003; Renard, 2007) seems the main argument favouring more elaborated models. To some extent, this debate has contributed to the growing tendency among hydrologists of postulating alternative conceptual models to represent optional dynamics explaining the flow and solute transport in a given groundwater system (Harrar et al., 2003; Meyer et al., 2004; Højberg and Refsgaard, 2005; Troldborg et al., 2007; Seifert et al., 2008).

It has been recently suggested that uncertainties in groundwater model predictions are largely dominated by uncertainty arising from the definition of alternative conceptual models and that parametric uncertainty solely does not allow compensating for conceptual model uncertainty (Bredehoeft, 2003; Neuman, 2003; Neuman and Wierenga, 2003; Ye et al., 2004; Bredehoeft, 2005; Højberg and Refsgaard, 2005; Poeter and Anderson, 2005; Refsgaard et al., 2006; Meyer et al., 2007; Refsgaard et al., 2007; Seifert et al., 2008; Rojas et al., 2008). Additionally, this last situation is exacerbated for the case when predicted variables are not included in the data used for calibration (Højberg and Refsgaard, 2005; Troldborg et al., 2007). This suggests that it is more appropriate to postulate alternative conceptual models and analyze the combined multimodel predictive uncertainty than relying on a single hydrological conceptual model. Working with a single conceptualization is more likely to produce biased and under-dispersive uncertainty estimations whereas working with a multimodel approach, uncertainty estimations are less (artificially) conservative and they are more likely to capture the unknown true predicted value.

Practice suggests, however, that once a conceptual model is successfully calibrated and validated, for example, following the method described by Hassan (2004), its results are rarely questioned and the conceptual model is assumed to be correct. As a consequence, the conceptual model is only revisited when sufficient data have been collected to perform a post-audit analysis (Anderson and Woessner, 1992), which often may take several years, or when new collected data and/or scientific evidence challenge the definition of the original conceptualization (Bredehoeft, 2005). In this regard, Bredehoeft (2005) presents a series of examples where unforeseen elements or the collection of new data challenged well established conceptual models. This situation clearly states the gap between practitioners and the scientific community in addressing predictive uncertainty estimations in groundwater modelling in presence of conceptual model uncertainty.

Different external factors such as climatic conditions or groundwater abstraction policies, on the other hand, increase the uncertainty in groundwater model predictions due to unknown future conditions. This source of uncertainty has since long been recognized as an important source of predictive uncertainty, however, practical applications mainly focus on uncertainty derived from parameters and inputs (forcing data), neglecting conceptual model and scenario uncertainties (Rubin, 2003; Gaganis and Smith, 2006). Recently, Rojas and Dassargues (2007) analyzed the groundwater balance of a regional aquifer in northern Chile considering different projected groundwater abstraction policies in combination with stochastic groundwater recharge values. Meyer et al. (2007) presented a combined estimation of conceptual model and scenario uncertainties in the framework of Maximum Likelihood Bayesian Model Averaging (MLBMA) (Neuman, 2003) for a groundwater flow and transport modelling study case.

In recent years, several methodologies to account for uncertainties arising from inputs (forcing data), parameters and the definition of alternative conceptual models have been proposed in the literature (Beven and Binley, 1992; Neuman, 2003; Poeter and Anderson, 2005; Refsgaard et al., 2006; Ajami et al., 2007; Rojas et al., 2008). Two appealing methodologies in the case of groundwater modelling are the MLBMA method (Neuman, 2003) and the information-theoretic based method of Poeter and Anderson (2005). Both methodologies are based on the use of a model selection criterion, which is derived as a by-product of traditional calibration methods such as maximum likelihood or weighted least

squares. The use of a model selection criterion allows ranking alternative conceptual models, eliminating some of them, or weighing and averaging model predictions in a multimodel framework. In our case, we are interested in weighing and averaging predictions from alternative conceptual models to obtain a combined estimation of the predictive uncertainty. The most commonly used model selection criteria correspond to Akaike Information Criterion (AIC) (Akaike, 1974), modified Akaike Information Criterion (AICc) (Hurvich and Tsai, 1989), Bayesian Information Criterion (BIC) (Schwartz, 1978) and Kashyap Information Criterion (KIC) (Kashyap, 1982). Ye et al. (2008a) gives an excellent discussion on the merits and demerits of alternative model selection criteria in the context of variogram multimodel analysis. In MLBMA, KIC is the suggested criterion whereas for the information-theoretic based method of Poeter and Anderson (2005), AICc is preferred. Even though Ye et al. (2008a) appear to have settled the controversy on the use of alternative model selection criteria, the use of different model selection criteria to weigh and combine multimodel predictions in groundwater modelling may lead to controversial and misleading results.

Apart from common problems of parameter non-uniqueness (insensitivity) and 'locality behaviour' of the calibration approaches mentioned above, Refsgaard et al. (2006) pointed out an important disadvantage of including a calibration stage in a multimodel framework. In the case of multimodel approaches including a calibration step, errors in the conceptual models (which per definition can not be excluded) will be compensated by biased parameter estimates in order to optimize model fit in the calibration stage. This has been confirmed by Troldborg et al. (2007) for a real aquifer system in Denmark.

Recently, Rojas et al. (2008) proposed an alternative methodology to account for predictive uncertainty arising from inputs (forcing data), parameters and the definition of alternative conceptual models. This method combines the Generalized Likelihood Uncertainty Estimation (GLUE) method (Beven and Binley, 1992) and Bayesian Model Averaging (BMA) (Draper, 1995; Kass and Raftery, 1995; Hoeting et al., 1999). The basic idea behind this methodology is the concept of equifinality, that is, many alternative conceptual models together with many alternative parameter sets will produce equally likely good results when compared to observed data (Beven and Freer, 2001; Beven, 2006). Equifinality, as defined by Beven (1993, 2006), arises because of the combined effects of errors in the forcing data, system conceptualization, measurements and parameter estimates. In the method of Rojas et al. (2008) series of "behavioural" parameters are selected for each alternative model

producing a cumulative density function (cdf) for parameters and variables of interest. Using the performance values obtained from GLUE, weights for each conceptual model are estimated, and results obtained for each model are combined following BMA in a multimodel frame. An important aspect of the method is that *it does not rely on a unique parameter optimum or conceptual model to assess the joint predictive uncertainty, thus, avoiding compensation of conceptual model errors due to biased parameter estimates*. A complete description of the methodology and potential advantages are discussed in Rojas et al. (2008).

Rojas et al. (2008) used a traditional Latin Hypercube Sampling (LHS) scheme (McKay et al., 1979) to implement the combined GLUE-BMA methodology. This sampling scheme has been regularly used in GLUE applications. Blasone et al. (2008a, 2008b) demonstrated that the efficiency of the GLUE methodology can be boosted up by including a Markov Chain Monte Carlo (MCMC) sampling scheme. MCMC is a sampling technique that produces a Markov Chain with stationary probability distribution equal to a desired distribution through iterative Monte Carlo simulation. This technique is particularly suitable in Bayesian inference when the analytical forms of posterior distributions are not available or in cases of high dimensional posterior distributions.

In this work we extend upon the methodology of Rojas et al. (2008) to include the uncertainty in groundwater model predictions due to the definition of alternative conceptual models and alternative recharge settings. For that, we follow an approach similar to that described in Meyer et al. (2007) and patterned after Draper (1995). Additionally, we improve on the sampling scheme of the combined GLUE-BMA methodology by implementing an MCMC sampling scheme. We apply the improved methodology to a real aquifer system underlying and feeding the Walenbos Nature Reserve area in Belgium (Fig. 1). We postulate three alternative conceptual models comprising different levels of geological knowledge for the groundwater system. Average recharge conditions are used to calibrate each conceptual model under steady-state conditions. Two additional recharge settings corresponding to ± 2 standard deviations from average recharge conditions are proposed to evaluate the uncertainty in the results due to the definition of alternative recharge values. A combined estimation of the predictive uncertainty including parameter, conceptual model, and scenario uncertainties is estimated for a set of groundwater budget terms such as river gains and river losses, drain outflows, and groundwater inflows and outflows from the Walenbos area. Finally, results obtained using the combined GLUE-BMA methodology are compared with the results

obtained using multimodel methodologies that include a calibration step and a model selection criterion to discriminate between models.

The remainder of this paper is organized as follows. In section 2, we provide a condensed overview of GLUE, BMA and MCMC theory followed by a description of the procedure to integrate these methods. Section 3 details the study area where the integrated uncertainty assessment methodology is applied. Implementation details such as the different conceptualizations, recharge uncertainties and the summary of the modelling procedure are described in section 4. Results are discussed in section 5 and a summary of conclusions is presented in section 6.

2. Material and methods

Sections 2.1, 2.2 and 2.3 elaborate on the basis of GLUE, BMA, and MCMC methodologies, respectively, for more details the reader is referred to Rojas et al. (2008, 2009).

2.1. Generalized likelihood uncertainty estimation (GLUE) methodology

GLUE is a Monte Carlo simulation technique based on the concept of equifinality (Beven and Freer, 2001). It rejects the idea of a single correct representation of a system in favour of many acceptable system representations (Beven, 2006). For each potential system simulator, sampled from a prior set of possible system representations, a likelihood measure (e.g. gaussian, trapezoidal, model efficiency, inverse error variance, etc.) is calculated, which reflects its ability to simulate the system responses, given the available observed dataset **D**. 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.

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

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

2.2. Bayesian model averaging (BMA)

BMA provides a coherent framework for combining predictions from multiple 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 a model over the others, instead, observed dataset **D** give the competing models different weights (Wasserman, 2000).

Following the notation of Hoeting et al. (1999), if Δ is a quantity to be predicted, the full BMA predictive distribution of Δ for a set of alternative conceptual models $\mathbf{M}=\left(\mathrm{M}_1, \mathrm{M}_2, ..., \mathrm{M}_k, ..., \mathrm{M}_K\right)$ under different scenarios $\mathbf{S}=\left(\mathrm{S}_1, \mathrm{S}_2, ..., \mathrm{S}_i, ..., \mathrm{S}_I\right)$ is given by Draper (1995)

$$p\left(\Delta \mid \mathbf{D}\right)=\sum_{i=1}^{I}\sum_{k=1}^{K} p\left(\Delta \mid \mathbf{D}, \mathrm{M}_k, \mathrm{S}_i\right) p\left(\mathrm{M}_k \mid \mathbf{D}, \mathrm{S}_i\right) p\left(\mathrm{S}_i\right) \quad (1)$$

Equation 1 is an average of the posterior distributions of Δ under each alternative conceptual model and scenarios considered, $p\left(\Delta \mid \mathbf{D}, \mathrm{M}_k, \mathrm{S}_i\right)$, weighted by their posterior model probability, $p\left(\mathrm{M}_k \mid \mathbf{D}, \mathrm{S}_i\right)$, and by scenario probabilities, $p\left(\mathrm{S}_i\right)$. The posterior model probabilities conditional on a given scenario reflect how well model *k* fits the observed dataset **D** and can be computed using Bayes' rule

$$p(\mathrm{M}_k \mid \mathbf{D}, \mathrm{S}_i) = \frac{p(\mathbf{D} \mid \mathrm{M}_k)\, p(\mathrm{M}_k \mid \mathrm{S}_i)}{\sum_{l=1}^{K} p(\mathbf{D} \mid \mathrm{M}_l)\, p(\mathrm{M}_l \mid \mathrm{S}_i)} \tag{2}$$

where $p(\mathrm{M}_k \mid \mathrm{S}_i)$ is the prior probability of model *k* under scenario *i*, and $p(\mathbf{D} \mid \mathrm{M}_k)$ is the integrated likelihood of the model *k*. An important assumption in the estimation of posterior model probabilities (equation 2) is the fact that the dataset **D** is independent of future scenarios. That is, the probability of observing the dataset **D** is not affected by the occurrence of any future scenario S_i (Meyer et al., 2007). In a strict sense, however, model likelihoods may depend on future scenarios given the correlation of recharge and hydraulic conductivity. Accounting for this dependency would make difficult to clearly assess the intrinsic value of the conceptual models or the "extra worth" of the data itself to explain the observed system responses. This assessment is beyond the scope of this article and for the sake of clarity the assumption of independence of **D** and, as consequence, of model likelihoods and posterior model probabilities from the future scenarios will be retained.

As a result, model likelihoods do not depend on the scenarios and, in contrast, prior model probabilities may be a function of future scenarios.

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

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

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

From equation 4 it is seen that the variance of the full BMA prediction consists of three terms: (I) within-models and within-scenarios variance, (II) between-models and within-scenarios variance and, (III) between-scenarios variance (Meyer et al., 2007).

2.3. Markov Chain Monte Carlo simulation

As discussed in Rojas et al (2008), due to the presence of multiple local optima in the global likelihood response surfaces, good performing simulators might be well distributed across the hyperspace dimensioned by the set of conceptual models, and forcing data (inputs) and parameter vectors. This necessitates that the global likelihood response surface is extensively sampled to ensure convergence of the posterior moments of the predictive distributions. In the context of the proposed (GLUE-BMA) methodology, we resorted to Markov Chain Monte Carlo (MCMC) to partly alleviate the computational burden of a traditional sampling scheme (e.g. Latin Hypercube Sampling).

The origins of MCMC methods can be traced back to the works of Metropolis et al. (1953) and the generalization by Hastings (1970). These works gave rise to a general MCMC method, namely, the Metropolis-Hastings (M-H) algorithm. The idea of this technique is to generate a Markov Chain for the model parameters using iterative Monte Carlo simulation that has, in an asymptotic sense, the desired posterior distribution as its stationary distribution (Sorensen and Gianola, 2002). Reviews and a more elaborate overview of alternative algorithms to implement MCMC are given in Gilks et al. (1995), Sorensen and Gianola (2002), Gelman et al. (2004) and Robert (2007).

The M-H algorithm stochastically generates a series with samples of parameters $\boldsymbol{\theta}_i, i=1,...,N$ through iterative Monte Carlo long enough such that, asymptotically, the stationary distribution of this series is the target posterior distribution, $p\left(\boldsymbol{\theta}|\mathbf{D}\right)$. This algorithm can be summarized as follows:

(1) set a starting location for the chain $\boldsymbol{\theta}_0$;

(2) set $i = 1,\ldots N$;

(3) generate a candidate parameter vector $\boldsymbol{\theta}^*$ from a proposal distribution $q\left(\boldsymbol{\theta}^*|\bullet\right)$;

(4) calculate $\alpha = \dfrac{p\left(\boldsymbol{\theta}^*|\mathbf{D}\right)q\left(\bullet|\boldsymbol{\theta}^*\right)}{p\left(\boldsymbol{\theta}_{i-1}|\mathbf{D}\right)q\left(\boldsymbol{\theta}^*|\bullet\right)}$;

(5) draw a random number $u \in [0,1]$ from a uniform probability distribution;

(6) if $\min\{1,\alpha\} > u$, then set $\boldsymbol{\theta}_i = \boldsymbol{\theta}^*$ otherwise $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i\text{-}1}$;

(7) repeat steps (3) through (6) N times.

The generation of the Markov Chain is, thus, achieved in a two-step process: a proposal step (step #3) and an acceptance step (step #6) (Sorensen and Gianola, 2002). Note that the proposal distribution $q\left(\theta^{*}\middle|\bullet\right)$ may (or may not) depend on the current position of the chain, θ_{i-1}, and may (or may not) be symmetric (Chib and Greenberg, 1995). These two properties are often modified to obtain alternative variants of the M-H algorithm (see e.g. Tierney, 1994). From the M-H algorithm, there is a natural tendency for parameters with higher posterior probabilities than the current parameter vector to be accepted, and those with lower posterior probabilities to be rejected (Gallagher and Doherty, 2007).

Several relevant aspects regarding the implementation of the M-H algorithm are worthwhile noticing. These aspects are related to (1) whether a single long-sized chain or several medium-sized parallel chains should be run, (2) the definition of the starting location for the chain $\left(\theta_{0}\right)$, (3) the nature of the proposal distribution $q\left(\theta^{*}\middle|\bullet\right)$, (4) the total number of iterations (N) to ensure a proper mixing of the chains and exploration of the support for the posterior probabilities and, (5) the number of *burn-in* initial samples (M) to reduce the influence of the starting location. Although there are no absolute rules to deal with these aspects some suggestions can be found in the literature. Brooks and Gelman (1998) and Gelman et al. (2004) suggest running several medium-sized parallel chains to ensure convergence of the posterior distribution, proper mixing of the chains in the parameter space, as well as to limit the dependence of the simulated chains on their starting locations. To determine the length N of the chains some convergence tests have been proposed (Cowles and Carlin, 1996). A formal test described in Gelman et al. (2004) consists in stopping iterations when within-chain variance is similar to between-chain variance for parameters and variables of interest. This is achieved when the R-score of Gelman et al. (2004) for multiple chains converges to values close to one. Gilks et al. (1995) suggest that the choice of the starting location is not critical as long as enough *burn-in* samples (M) are selected. To determine the *burn-in* length literature suggests values between $0.01N$ and $0.5N$ (Geyer, 1992; Gilks et al., 1995; Gelman et al., 2004). The selection of the proposal distribution remains one of the most critical aspects. Common practice is to use a multivariate normal distribution centred on the previous parameter vector, i.e. $q\left(\theta^{*}\middle|\theta_{i\text{-}1}\right)\sim \mathrm{N}\left(\theta_{i\text{-}1}\middle|\Sigma_{\theta}\right)$. The variance matrix Σ_{θ} is used as a

jumping rule to achieve acceptance rates (defined as the fraction of accepted parameter candidates in a window of the last n samples) in the range 20-70 % (Makowski et al., 2002; Robert, 2007). Another commonly used option is to use a d-dimensional uniform distribution over prior parameter ranges (Sorensen and Gianola, 2002). It is worth noticing, however, that many functional forms are available to define the proposal distribution $q\left(\boldsymbol{\theta}^{*}\middle|\bullet\right)$ and this is the main strength of the M-H algorithm.

2.4. Multimodel approach to account for conceptual model and scenario uncertainties

Combining GLUE and BMA in the frame of the method proposed by Rojas et al. (2008) to account for conceptual model and scenario uncertainties involves the following sequence of steps

1. On the basis of prior and expert knowledge about the site, a suite of alternative conceptualizations is proposed, following, for instance, the methodology proposed by Neuman and Wierenga (2003). In this step, a decision on the values of prior model probabilities should be taken (Meyer et al., 2007; Ye et al., 2005; Ye et al., 2008b). Additionally, a suite of scenarios to be evaluated and their corresponding prior probabilities should be defined at this stage.
2. Realistic prior ranges are defined for the forcing data (inputs) and parameter vectors under each plausible model structure.
3. A likelihood measure and rejection criterion to assess model performance are defined (Jensen, 2003; Rojas et al., 2008). A rejection criterion can be defined from exploratory runs of the system, based on subjectively chosen threshold limits (Feyen et al., 2001) or as an accepted minimum level of performance (Binley and Beven, 2003).
4. For the suite of alternative conceptual models, parameter values are sampled using a Markov Chain Monte Carlo (MCMC) algorithm (Gilks et al., 1995) from the prior ranges defined in (3) to generate possible representations or simulators of the system. A likelihood measure is calculated for each simulator, based on the agreement between the simulated and observed system response.
5. For each conceptual model M_k, the model likelihood is approximated using the likelihood measure. A subset A_k of simulators with likelihoods $p\left(\mathbf{D}\middle|\mathrm{M}_k,\boldsymbol{\theta}_l\right)\approx L\left(\mathrm{M}_k,\boldsymbol{\theta}_l,\mathbf{Y}_m\middle|\mathbf{D}\right)$ is retained based on the rejection criterion.
6. Steps 4-5 are repeated until the hyperspace of possible simulators is adequately sampled, i.e. when the first two moments for the conditional distributions of parameters based on

the likelihood weighted simulators converge to stable values for each of the conceptual models M_k, and when the R-score (Gelman et al., 2004) for multiple Markov Chains converges to values close to one.

7. The integrated likelihood of each conceptual model M_k (equation 2) is approximated by summing the likelihood weights of the retained simulators in the subset A_k, that is, $p\left(\mathbf{D}\middle|\mathrm{M}_k\right)\approx\sum_{i\in A_k} L_i\left(\mathrm{M}_k,\boldsymbol{\theta}_l,\mathbf{Y}_m\middle|\mathbf{D}\right)$.
8. The posterior model probabilities are then obtained by normalizing the integrated model likelihoods over the whole ensemble $\mathbf{M}$ such that they sum up to one using equation (2).
9. After normalization of the likelihood weighted predictions under each individual model for each alternative scenario (such that the cumulative likelihood under each model and scenario equals one), an approximation to $p\left(\Delta\,|\,\mathbf{D},\mathrm{M}_k,\mathrm{S}_i\right)$ is obtained, and a multimodel prediction is obtained with equation (1). The leading moments of this distribution are obtained with equations (3) and (4) considering all scenarios.

Posterior model probabilities obtained in step (8) are used in the prediction stage for the alternative conceptual models under alternative scenarios. Thus, the more demanding steps of the methodology (step 4 and step 5) are done only once to obtain the posterior model probabilities. This is based on the assumption that the observed dataset **D** is independent of future scenarios. That is, the probability of observing the dataset **D** is not affected by the occurrence of any future scenario S_i (Meyer et al., 2007).

2.5. Multimodel methods and model selection criteria

As previously stated, multimodel methodologies using model selection or information criteria have been proposed by Neuman (2003) and Poeter and Anderson (2005). These model or information criteria are obtained as by-products of the calibration of groundwater models using, e.g. maximum likelihood or weighted least squares methods. As suggested by Ye et al. (2008a), equation (2) can be approximated by

$$p\left(\mathrm{M}_k\,|\,\mathbf{D},\mathrm{S}_i\right)\approx\frac{\exp\left(-\frac{1}{2}\Delta IC_k\right)p\left(\mathrm{M}_k\,|\,\mathrm{S}_i\right)}{\sum_{l=1}^{K}\exp\left(-\frac{1}{2}\Delta IC_l\right)p\left(\mathrm{M}_l\,|\,\mathrm{S}_i\right)} \tag{5}$$

where $\Delta IC_k = IC_k - IC_{\min}$, IC_k being any of the model selection or information criteria described in section 1 for a given model *k*, and IC_{min} the minimum value obtained across models $M_k, k = \{1,...,K\}$. These posterior model probabilities are then used to estimate the leading moments of the BMA prediction (equations 3 and 4) considering alternative conceptual models and alternative scenarios.

Alternative model selection or information criteria differ in mathematical expressions, in the way they penalize the inclusion of extra model parameters, or how they value prior information about model parameters. These differences produce dissimilar results for equation (5) even for the case of a common dataset **D** to all models. This may lead to controversial and misleading results when posterior model probabilities obtained using equation (5) are used to obtain the leading moments of the BMA predictions (equations 3 and 4).

3. Study area

3.1. General description

The Walenbos Nature Reserve is located in the northern part of Belgium, 30 km North-East of Brussels, in the valley of the brook 'Brede Motte' (Fig. 1). It is a forested wetland of regional importance highly dependant on groundwater discharges, especially, in shallow depressions (De Becker and Huybrechts, 1997). Previous studies showed that groundwater discharging in the wetland infiltrated over a large area, mainly south of the wetland and it consists of groundwater of different aquifers (Batelaan et al., 1993; 1998).

The study area is bounded by two main rivers, the Demer River in the North and the Velp River in the South. Other minor rivers are observed within the study area: the Motte River, which drains the wetland towards the North, the Molenbeek River and the Wingebeek River (Fig.1). The Demer and the Velp rivers have an elevation of 10 m above sea level (asl) and 35 m asl., respectively. Between these two rivers the area consists of undulating hills and plateaus reaching a maximum elevation of 80 m asl. Within the Walenbos Nature Reserve area, the slightly raised central part divides the wetland into an Eastern and Western subbasin.

Larger and smaller rivers are administratively classified into categories for water management purposes (HAECON and Witteveen en Bos, 2004). The Demer is navigable and of category 0

while the Velp is smaller and of category 1. The Wingebeek, Motte, and Molenbeek are category 2 rivers. From these categories, initial properties (e.g. bed sediment thickness, river width, depth, etc.) for the main rivers are obtained and, consequently, used to estimate values of river conductance.

There are several observation wells within the study area from different monitoring networks of the Flemish Environment Agency (VMM) and the Research Institute for Nature and Forest (INBO). The data are made available through the Database of the Subsurface for Flanders (DOV, 2008). In this study 51 observation wells are used (Fig. 1), most of them concentrated in the Walenbos area.

3.2. Geology and hydrogeology

Fig. 2 shows the geological map of the study area. Additionally, Table 1 gives the lithostratigraphic description of the formations present in the study area. The geology of the study area consists of an alteration of sandy and more clayey formations, generally dipping to the north and ranging in age from the Early Eocene to the Miocene. The Hannut formation are clayey or sandy silts with locally a siliceous limestone. The formation only crops out south of the Velp River. The Kortrijk formation is a marine deposit consisting mainly of clayey sediments. This formation is covered by the Brussel formation, a heterogeneous alteration of coarse and fine sands, locally calcareous and/or glauconiferous. The Early Oligocene Sint Huibrechts Hern formation is a glauconiferous or micaeous, clayey fine sand, which is locally very fossiliferous. The Borgloon formation represents a transition to a more continental setting and consists of a layer of clay lenses followed by an alteration of sand and clay layers. The Bilzen formation represents a marine deposit consisting of fine sands, glauconiferous at the base. The Bilzen sands are followed by a clay layer, the Boom formation. On top of the Boom formation, the Bolderberg formation is found which consists of medium fine sands, locally clayey. The youngest deposits consist of coarse, glauconiferous sands of the Diest formation. These sands are deposited in a high energetic, shallow marine setting and have locally eroded underlying formations. In the Walenbos area, for example, the Diest formation is directly in contact with the Brussel formation. The Kortrijk, Brussel and Sint Huibrechts Hern formations are present in the entire study area, while the younger layers disappear towards the south or are eroded in the valleys. The study area is covered with Quaternary sediments, consisting of loamy eolian deposits on the interfluves and alluvial deposits in the

river valleys. The geological characteristics of the study area are described in detail in Laga et al. (2001) and Gullentops et al. (2001).

The Hydrogeological Code for Flanders (HCOV) is used to identify different hydrogeological units (Meyus et al., 2000; Cools et al., 2006). The hydrogeological conceptualization of the aquifer system surrounding and underlying the Walenbos Nature Reserve area was schematized as one-, three- and five-layers with the top of the Kortrijk formation as the bottom boundary for all conceptualizations considered (Fig. 3 and Table 2). These geological models were developed to assess the worth of extra "soft" geological knowledge about the geometry of the groundwater system underlying the Walenbos Nature Reserve. In this way, alternative layering structures for the aquifer are assessed in terms of improving the model performance.

4. Implementing the multimodel approach

Three alternative conceptual models comprising different level of geological knowledge are proposed (Fig. 3). Each model is assigned a prior model probability of 1/3. A complete analysis on the sensitivity of the multimodel methodology to these prior model probabilities is given in Rojas et al. (2009). All proposed conceptual models are bounded by the Kortrijk formation as low permeability bottom and the topographical surface for the top of the system. Model 1 (M1) corresponds to the simplest representation considering one hydrostratigraphic unit, Model 2 (M2) comprises three hydrostratigraphic units and Model 3 (M3) corresponds to the most complex system comprising five hydrostratigraphic units. Details are presented in Table 2 and Fig. 3.

Groundwater models for the three conceptualizations are constructed using MODFLOW-2005 (Harbaugh, 2005). The groundwater flow regime is assumed as steady-state conditions. The model area is ca. 11×22 km^2. Using a uniform cell size of 100 m the modelled domain is discretized into 110×220 cells. The total number of cells varies from model to model since the number of layers to account for different hydrostratigraphic units changed. At the North and South, respectively, the Demer and Velp rivers are defined as boundary conditions using the river package of MODFLOW-2005. Physical properties of both rivers (e.g. width, thickness of bed sediments and river stage) are obtained from models built within the frame of the Flemish Groundwater Model (HAECON and Witteveen en Bos, 2004). All grid cells located to the North of the Demer and to the South of the Velp, respectively, are set as

inactive (i.e. no-flow). East and west limits of the modelled domain are defined as no-flow boundary conditions. To account for possible groundwater discharge zones in the study area, the drain package is used for all active cells in the uppermost layer of each model. The elevation of the drain element for each cell is defined as the topographic elevation minus 0.5 m, in order to account for an average drainage depth of ditches and small rivulets (Batelaan and De Smedt, 2004).

The focus of this work is on the assessment of conceptual model and recharge (scenario) uncertainties. Therefore, we confine the dimensionality of the analysis by considering uncertainty only in the conductance parameters related to the Demer and Velp rivers, conductance of drains, and hydraulic conductivities of the alternative hydrostratigraphic units (see Table 2 and Table 3). Additionally, the spatial zonation of the hydraulic conductivity field is kept constant and only the mean values for each hydrostratigraphic unit are sampled using the M-H algorithm. Parameter ranges are defined based on data from previous studies and they are presented in Table 3 (HAECON and Witteveen en Bos, 2004). It is worth noticing that in the frame of the proposed methodology, heterogeneous fields following the theory of Random Space Functions (RSF) are easily implemented (Rojas et al., 2008).

Average recharge conditions $\left(\overline{\mathrm{R}}\right)$ over a grid of 100 × 100 m accounting for average hydrological conditions is obtained from Batelaan et al. (2007). Spatially distributed recharge values are calculated with WetSpass (Batelaan and De Smedt, 2007), which is a physically based water balance model for calculating the quasi-steady-state spatially variable evapotranspiration, surface runoff, and groundwater recharge at a grid-cell basis. The average recharge condition constitutes the base situation for the estimation of the posterior model probabilities used in the multimodel approach. Additionally, to account for recharge uncertainties (scenarios), two optional recharge situations are defined based on a deviation corresponding to $\pm 2\sigma_{\overline{\mathrm{R}}}$ from the average recharge conditions $\left(\overline{\mathrm{R}}\right)$. We used $\pm 2\sigma_{\overline{\mathrm{R}}}$ to make an intuitive link with the expression of 95% confidence interval for potential recharge values. The definition of these three recharge settings is based on long-term simulations of the average hydrological conditions accounting for more than 100 years of meteorological data (see Batelaan and De Smedt, 2007). Although in a strict sense, the plausibility of these average recharge values might have been evaluated as they took place in the past similarly to the dataset **D**, this is not possible as **D** considered a limited and variable time series of head

measurements. The key assumptions for the analysis performed in this work are, first, the nature of the steady-state condition of **D**. This steady-state condition is valid for present-time situation only since the time series available with observed heads are considerably less than the series of meteorological data used to estimate average recharge conditions (S2). Second, it is the fact that there is no guarantee that similar (climate) recharge conditions will be observed for the next 100 years. The latter will have a clear influence on the definition of coherent prior probabilities for each scenario.

Based on the assumption previously discussed, recharge uncertainties are treated as scenario uncertainties in the context of the proposed GLUE-BMA method (equations 1-4). To avoid conflicting terminology, however, both terms scenario uncertainties and recharge uncertainties are used interchangeably hereafter.

Based on long-term simulations three recharge conditions (scenarios) are defined: S1 $\left(\overline{\mathrm{R}}-2\sigma_{\overline{\mathrm{R}}}\right)$, S2 $\left(\overline{\mathrm{R}}\right)$, and S3 $\left(\overline{\mathrm{R}}+2\sigma_{\overline{\mathrm{R}}}\right)$. Average values for S1, S2 and S3 are 93.1 mm yr^{-1}, 205.4 mm yr^{-1} and 319.5 mm yr^{-1}, respectively. Based on the previous assumption of future recharge conditions, each scenario is assigned a prior scenario probability of 1/3. This is based on the fact that for future recharge conditions, average or tail values are equally likely to be observed.

A Gaussian likelihood measure is implemented to assess model performance, i.e. to assess the ability of the simulator to reproduce the observed dataset **D**. Observed heads (h_{obs}) for the 51 observation wells depicted in Fig. 1 are compared to simulated heads (h_{sim}) to obtain a likelihood measure. Observed heads correspond to a representative value (average) for steady state-conditions for different time series in the period 1989-2008. Observation wells vary in depth and also the length and depth of the screening is variable. Although some local confined conditions controlled by the Boom formation are observed in the study area, the observed dataset **D** accounted for phreatic conditions solely. This might lower the information content of the dataset **D** to effectively discriminate between models. A limited set of head observations, however, may often be the only information available about the system dynamics to perform a modelling exercise and/or model discrimination. From preliminary runs a departure of ± 5 m from the observed head in each observation well is defined as rejection criterion. That is, if $h_{obs}-5m < h_{sim} < h_{obs}+5m$ a Gaussian likelihood measure is

calculated, otherwise the likelihood is zero. This rejection criterion is defined in order to achieve enough parameter samples for the exploration of the posterior probability space and to ensure convergence of the different Markov Chains used in the M-H algorithm. For details about the implementation of the rejection criterion in the frame of the proposed approach the reader is referred to Rojas et al. (2008).

Five parallel Markov Chains, starting from randomly selected points defined in the prior parameter ranges (Table 3), are implemented to proceed with the M-H algorithm for each conceptual model. Four-, six-, and eight-dimensional uniform distributions with initial prior ranges defined in Table 3 are defined as the $q\left(\boldsymbol{\theta}^{*}|\bullet\right)$ proposal distributions for M1, M2 and M3, respectively. The variance of the proposal distributions is modified by trial-and-error to achieve acceptance rates in the range 20 – 40 %. For each proposed parameter set a new gaussian likelihood value is calculated in function of the agreement between observed and simulated groundwater heads at the 51 observation wells depicted in Fig. 1. These proposed parameter sets are accepted or rejected according to step #6 of section 2.3. As previously stated, the mixing of the chains and the convergence of the posterior probability distributions is monitored using the R-score (Gelman et al., 2004). The resulting total parameter sample (after discarding the *burn-in* samples) can be considered as a sample from the posterior distribution given the observed dataset **D** for each alternative conceptual model. This simulation procedure is repeated for models M1, M2 and M3 for average recharge conditions $\left(\overline{\mathrm{R}}\right)$ to obtain the posterior model probabilities (equation 2).

Using the discrete samples from the M-H algorithm the integrated likelihood of each conceptual model, $p\left(\mathbf{D}|\mathrm{M}_{k}\right)$ in equation 2, is approximated by summing over all the retained likelihood values for M_{k}. The posterior model probabilities are then obtained by normalizing over the whole ensemble $\mathbf{M}$ under average recharge conditions.

For each series of predicted variables of interest, e.g., river losses and river gains from the Velp and Demer, drain outflows, and groundwater inflows and outflows from the Walenbos area, a cumulative predictive distribution, $p\left(\Delta\,|\,\mathbf{D},\mathrm{M}_{k},\mathrm{S}_{i}\right)$, is approximated by normalizing the retained likelihood values for each conceptual model under each scenario such that they sum up to one.

The leading moments of the full BMA predictive distribution accounting for parameter, conceptual model and scenario uncertainties are then obtained using equations (3) and (4).

5. Results and discussion

Since it is not possible to show the complete set of results for all variables, groundwater budget terms and alternative conceptualizations, in the following sections the most relevant results are summarized.

5.1. Validation of the M-H algorithm results

The proposed methodology mainly worked by sampling new parameter sets for each proposed conceptual model following an M-H algorithm with the aim of obtaining posterior parameter probability distributions. Several aspects of the implementation of the M-H algorithm such as the acceptance rate, the definition of the *burn-in* samples, the proper mixing of alternative chains and the convergence of the first two moments were checked to validate the results obtained using the improved methodology.

The average acceptance rates for the Markov Chains found for models M1, M2 and M3 (for the 20 000 parameter samples) were 25 %, 23 % and 27 %, respectively. All values lie in the ranges as suggested in literature (Makowski et al., 2002).

Fig. 4 shows, as an example, five chains for the parameters included in model M2. This figure shows the values for the proposed parameter versus the number of sampling iteration. It is seen that full mixing of the five chains is achieved for values greater than 1 000 parameter samples for all six parameters (plates a through f). As a result, the first 1 000 iterations were set as *burn-in* samples and they were discarded as they were slightly influenced by the starting values of the chains. Although not shown here, similar results were obtained for models M1 and M3, with 1 000 initial samples defined as *burn-in*.

As previously stated, the mixing of the chains and the convergence of the posterior probability distributions of parameters and variables of interest were monitored using the R-score. Gelman et al. (2004) suggest values for the R-score near 1, with values below 1.1 acceptable for different problems. The R-score was calculated for the whole series of parameters and variables of interest for the three alternative conceptual models M1, M2, and M3. The largest

R-score for 5 000 parameter samples was 1.02 indicating a good mixing of the five chains and, hence, suggesting convergence of the posterior probability distributions (e.g. see how all five chains completely overlap in Fig. 4 after a value of 1 000 for the case of M2, covering the same support for the posterior probability distributions). Subsequently, a discrete parameter sample comprising 20 000 values is obtained by combining the results of the five chains.

Although not shown here, convergence of the first two moments for the posterior distributions of parameters obtained from the total discrete parameter sample was also confirmed for the three alternative conceptual models.

Therefore, the resulting discrete samples of parameters from models M1, M2 and M3 can be considered as a sample from the target posterior distributions under the respective conceptual model.

5.2. Likelihood response surfaces

From the proposed methodology, each parameter set was linked to a likelihood value. The resulting marginal scatter plots of parameter likelihoods for models M1 and M2 are shown in Fig. 5 and Fig. 6, respectively. Also, included in these figures are the results of a weighted least squares calibration using UCODE-2005 (Poeter et al., 2005). It is worth mentioning that several calibration trials (six for Model M1, ten for model M2 and more than twenty for model M3) starting at different initial parameter values contained in the ranges defined in Table 3 were launched. For the sensitive parameters all the calibration trials converged to rather similar optimum parameter values, however, some minor differences were observed due to irregularities in the likelihood response surface. For insensitive parameters, on the other hand, different trials converged to different values. For the sake of clarity, only the final calibrated parameter set is included in the comparison with the GLUE-BMA results.

From Fig. 5 and Fig. 6 it is seen that likelihood values were rather insensitive to the conductance of drains and rivers (plates a, b c in Fig. 5 and Fig. 6). High likelihood values were observed for almost the whole prior parameter sampling range being very difficult to identify a well-defined attraction zone for these three parameters. This insensitivity was also reflected in the significant difference between the values obtained using least squares calibration and the highest likelihood points obtained in the context of the proposed method. Clearly, least squares calibration did not succeed in identifying the point and/or even the

range where the highest likelihood values for these parameters were observed. This is a well-known drawback of least squares calibration methods in the presence of highly insensitive parameters.

For parameters defining the mean hydraulic conductivity for each model layer, on the contrary, well-defined attraction zones were identified by the proposed methodology (plates d in Fig. 5 and plates d, e and f in Fig. 6). For these parameters, results obtained from least squares calibration were almost identical to the highest likelihood points identified in the frame of the proposed methodology.

Although not shown here, the same patterns were observed for model M3 for the case of the three insensitive parameters and the parameters defining the mean hydraulic conductivity for layer 1 (HK-1), layer 4 (HK-4) and layer 5 (HK-5) (Table 2). For these last three parameters, well-defined attraction zones were identified and results of least squares calibration were fairly similar to the highest likelihood points identified in the frame of the proposed methodology. However, two exceptions are worth mentioning. In Fig. 7 the marginal scatter plots of calculated likelihood for the hydraulic conductivity of layer 2 (HK-2) (plate a) and layer 3 (HK-3) (plate b) for model M3, are shown. These layers correspond to the Boom formation and Ruisbroek formation, respectively (Table 2). These marginal scatter plots show that likelihood values are fairly insensitive to these two parameters (HK-2 and HK-3) for M3. However, a clear attraction zone for values greater than 0.001 m d^{-1} is observed for both parameters. This contrasts with the results obtained using the least squares calibration method. The most severe difference is for the case of the parameter HK-2 (plate a) where the highest likelihood point identified with the proposed methodology and the result from the least squares calibration differed by more than six orders of magnitude. It is worth mentioning that convergence of UCODE-2005 was highly sensitive to the initial values of HK-2 and HK-3. After a significant number of trials, meaningful initial parameter values for HK-2 and HK-3 were set to 0.01 m d^{-1} and 4.6 m d^{-1}, respectively. These initial values allowed for convergence of UCODE-2005, however, they produce rather dissimilar calibrated values compared to the highest likelihood points obtained with the proposed methodology. On the contrary, for the case of the proposed methodology the parameters were sampled from the prior range defined in Table 3 following the acceptance/rejection rule described in step # 6 of section 2.3. Therefore, this procedure allowed identifying clear zones of attraction for these two parameters although their insensitivity remained observed.

This critical difference in both approaches (GLUE-BMA and WLS calibration) may be explained by the meaning and the type of information conveyed by the dataset **D** in this application. For pragmatic reasons, the dataset **D** did not include observation wells located in local confined aquifers distributed over the study area since the interest was on the general functioning of the aquifer system and not on local conditions. In general, these local confined aquifers are controlled by the presence of the hydrostratigraphic unit defined by the Boom Formation and, thus, by parameter HK-2. Therefore, if the dataset **D**, which corresponds to head measurements accounting for phreatic conditions solely, does not contain any relevant information on confined areas it is difficult to account for the relevance and the actual value of parameter HK-2. As a consequence, parameter HK-2 becomes redundant and the zone of attraction defined in Fig. 7a is defined for an "equivalent" parameter accounting only for a phreatic system. This situation was easily assimilated by the GLUE-BMA methodology whereas the WLS method faced convergence problems since initial values for parameter HK-2 were defined in the observed range for the hydraulic conductivity values of the Boom Formation.

Despite these differences between WLS and GLUE-BMA, both methods performed equally well in terms of model performance. As an example, the root mean squared error (RMSE) for model M1 using WLS and GLUE-BMA was 1.884 and 1.876, respectively. For model M2 both WLS and GLUE-BMA gave an RMSE of 1.890 whereas for model M3 the RMSE of WLS and GLUE-BMA was 1.761 and 1.741, respectively.

5.3. Posterior model probabilities

Table 4 presents the posterior model probabilities obtained using equation (2) for average recharge conditions as a result of the proposed methodology. It is seen from this table that the integrated likelihoods for models M1, M2 and M3 differ slightly. As a consequence, and since posterior model probabilities are proportional to the integrated likelihoods when prior model probabilities are set equal (i.e. when there is no clear preference for a given conceptual model), posterior model probabilities also differ marginally.

For this case, information provided by the observed dataset **D** (in the process of updating the prior model probabilities) is marginal and does not allow discriminating significantly between models once **D** has been observed. This suggests that, for the problem at hand and for the

level of information content of **D**, prior model probabilities will likely play a significant role in determining the posterior model probabilities. In this regard, prior model probabilities could be thought of as "prior knowledge" about the alternative conceptual models. This prior knowledge is ideally based on expert judgement, which Bredehoeft (2005) considers the basis for conceptual model development. In this way, expert "subjective" prior knowledge about optional conceptualizations in combination with the information provided by the dataset **D**, may allow some degree of discrimination between models through updated posterior model probabilities. As shown in Ye et al. (2008b), however, even for the case when an expert assigns substantially different prior model probabilities, aggregating the prior model probabilities values from several authors gives a relatively uniform prior model probability distribution. It would be interesting to investigate the joint effect of data and expert judgement on the prior model probabilities. For a complete analysis on the sensitivity of the results of the proposed methodology to different prior model probabilities, which is beyond the scope of this article, the reader is referred to Rojas et al. (2009).

Another possible strategy is to increase the information content of **D** by collecting new data that may be particularly useful in discriminating between models (e.g. river discharges, tracer travel times and observed groundwater flows). With extra data, the level of "conditioning" of the results is increased and (hopefully) the integrated model likelihoods will differ for alternative conceptual models. In practice, however, a set of observed groundwater heads may often be the only information available about the system dynamics to estimate posterior model probabilities for a set of alternative model conceptualizations. This clearly put the challenge of assigning model weights (i.e. posterior model probabilities) considering often a minimum level of information.

5.4. Groundwater model predictions accounting for conceptual model and scenario uncertainties

Using the posterior model probabilities obtained for average recharge conditions (Table 4) and the cumulative predictive distributions obtained for each model, a multimodel cumulative predictive distribution is obtained for scenarios S1, S2 and S3.

Fig. 8 shows the cumulative predictive distributions for a series of groundwater budget terms and the combined BMA prediction accounting only for conceptual model uncertainty for scenario S2. From this figure it is seen that, although posterior model probabilities differ

slightly (Table 4), indicating a low information content of the dataset **D**, there are significant differences in the predictions of models M1, M2 and M3. For river losses and river gains from the Demer (plates a and b) and Walenbos outflows and inflows (plates f and g), both the most likely predicted values (P_{50}) and the 95 % ($P_{2.5}$ - $P_{97.5}$) prediction intervals drastically differ between alternative conceptual models. This indicates that conceptual model uncertainty considerably dominates both the most likely predictions and the predictive uncertainty under S2. On the other hand, the most likely predicted values for river losses and river gains from the Velp (Fig. 8 plates c and d) and drain outflows (plate e) are rather similar, yet the 95 % prediction intervals span clearly different ranges. This indicates that although the most likely predicted values for models M1, M2 and M3 are quite similar, their predictive uncertainty is largely dominated by conceptual model uncertainty.

Additionally, Table 5 summarizes the most likely predicted values and the 95 % predictive intervals for models M1, M2 and M3, under scenarios S1, S2 and S3 for the same groundwater budget terms described in Fig. 8. This table shows that for scenarios S1 and S3, uncertainties due to the specification of alternative conceptual models also play an important role. Conceptual model uncertainty is more relevant (under S1) for river gains and river losses from the Demer and the Velp and, marginally, for drain outflows. This is explained by the fact that during low recharge conditions (S1) rivers contribute more water to the groundwater system due to lower simulated groundwater heads in the neighbouring areas. This lowering in heads also explains why the drain outflows are only marginally affected by the conceptual model uncertainty. For scenario S3 all predictive intervals for the groundwater budget terms are affected by the selection of an alternative conceptual model. This is expected as for high recharge conditions (S3) it is likely that all groundwater flow components will be affected by an alternative conceptualization.

A slight tendency to larger predictive intervals for M3, then M2 and, finally, M1 is observed for all recharge conditions. This is expected as an increase in model complexity, expressed as an increase in the number of model parameters, allows for more parametric uncertainty to be incorporated. This suggests that model M3 will produce the main contribution to conceptual model uncertainty due to wider predictive intervals.

If groundwater budget terms are transversely analyzed it is seen that predictive intervals for river losses are dominated by scenario S1 whereas predictive intervals for rivers gains are

dominated by scenario S3. For the drain outflows and groundwater inflows and outflows from the Walenbos area, scenario S3 shows the largest predictive intervals. These tendencies are more pronounced for model M3 compared to models M2 and M1, reaffirming the idea stated in the previous paragraph.

Each BMA cumulative distribution accounting for the alternative conceptual models is combined under each scenario (e.g. Fig. 8 shows the case for S2 only). Subsequently, each scenario prediction is combined following equation (1) to obtain a full BMA prediction accounting for conceptual model and scenario uncertainties. Fig. 9 presents the results for the full BMA prediction. From this figure it is seen that the most likely predicted values obtained with the full BMA predictive distribution are rather similar to the results obtained with scenario S2. This suggests that the main impact of including S1 and S3 is in the estimation of the predictive uncertainty rather than in the estimation of the most likely predicted value. This is evident for the case of drain outflows (plate e) where the P_{50} for the full BMA and S2 are practically identical while the predictive intervals completely span different ranges. This suggests that for the drain outflows, scenario uncertainties will represent the main contribution to the predictive variance. The most likely predicted values and the 95 % prediction intervals for the full BMA predictive distribution are summarized in Table 6.

5.5. Contribution to predictive variance

As presented in equation (4), predictive variance can be subdivided into three sources, namely, (I) within-models and within-scenarios (forcing data + parameters uncertainty), (II) between-models and within-scenarios (conceptual model uncertainty) and, (III) between-scenarios (scenario uncertainty). Fig. 10 shows the predictive variance for the groundwater budget terms described in previous paragraphs. Each source contribution is expressed as a percentage of the predictive variance. Within-models contribution is more significant for river losses from the Velp (67 %) and river gains from the Demer (66 %). The contribution attributed to between-models is more important for the groundwater outflows from Walenbos (75 %) and for river losses from the Demer (69 %). Between-scenarios contributes up to ca. 100 % the predictive variance for the drain outflows and up to 78 % for the groundwater inflows to the Walenbos area.

These results clearly show that considering fairly reasonable and observable recharge conditions have a considerable impact on the estimations of the predictive variance. However,

due to the fact that future scenarios are driven by unpredictable future conditions, it is particularly difficult to implement suitable strategies aiming to diminish their contribution to the predictive variance. On the contrary, when alternative scenarios are linked to fully or partially known future conditions, e.g., groundwater abstraction scenarios (Rojas and Dassargues, 2007), prior scenario probabilities could be defined based on expert judgement or following a similar approach to that described in Ye et al. (2008b). In the case of within- and between-models variance it is likely that new collected information/data may help in decreasing their corresponding uncertainty contributions. For the within-models variance, it would be particularly interesting to collect data on the river dynamics to aim decreasing the uncertainties in model predictions for the river gains and losses in the Demer and Velp, respectively. As for the case of between-models variance, new information/data on river dynamics together with a better understanding of the groundwater flow dynamics in the Walenbos area would be helpful in decreasing the contribution of conceptual model uncertainty to predictive variance.

5.6. Criteria-based multimodel methodologies

Alternatively, models M1, M2 and M3 were calibrated using a weighted least squares method included in UCODE-2005 (Poeter et al., 2005). Parametric uncertainty for each model was assessed using Monte Carlo simulation in a similar way to that described in Ye et al. (2006). Results of UCODE-2005 were used to approximate the posterior model probabilities using equation (5) for a series of four model selection criteria, namely, AIC, AICc, BIC and KIC (see section 2.4). These posterior model probabilities were then used to estimate the full BMA prediction (equation 1), its leading moments (equations 3 and 4), and the contribution to predictive variance in the same fashion as in the case of GLUE-BMA.

Table 7 summarizes the results of the least squares calibration using UCODE-2005. From this table it is seen that models M1, M2 and M3 are ranked differently depending on the model selection criterion used. This is in full agreement with the results obtained by Ye et al. (2008). Whereas AIC and AICc rank models identically, posterior model probabilities obtained with equation (5) are rather different for these two criteria. In the case of BIC, most of the posterior weight is assigned to model M1 (97 %), indicating that models M2 and M3 will have just marginal contributions in the estimation of the full BMA predictive distribution. Additionally, models M2 and M3 are ranked differently by BIC compared to AIC and AICc. The reason for this is the fact that BIC penalizes more drastically more complex models when the

observation sample size is larger than 9, i.e. D_i, i>9, thus, putting more importance on parsimony. For KIC a completely different ranking is obtained as a result. Using the latter, M3 is preferred over the other models accounting for a posterior weight of ca. 80 %. Remarkably, this ranking is completely opposite to the one obtained using AIC and AICc. Ye et al. (2008a) argue that the presence of the Fisher information term strongly influences the results of KIC. This allows KIC sometimes to prefer more complex models based not only on goodness of fit and number of parameters but also on the quality of the available dataset **D**. This property is not shared by AIC, AICc or BIC since the Fisher information term is not present in their definitions. Although Ye et al. (2008a) appear to have settled the controversy about the use of alternative model selection criteria in the frame of multimodel methodologies, the use of different model selection criteria will rank differently alternative conceptual models and, consequently, alternative conceptualizations will be given different posterior model probabilities using the approximation expressed in equation (5). In the framework of a multimodel approach, this is critical.

Results from Table 7 also confirm the nature of the dataset **D** used to assess model performance. As discussed earlier, **D** accounted only for phreatic conditions (head measurements of local confined areas were discarded) in order to assess the meso-scale groundwater flows to the Walenbos Nature Reserve. Table 7 shows that SWSR of model M2 is larger than that of M1, although model M2 has two more parameters. In addition, the calibrated values of HK-1 for models M1, M2, and M3 are rather similar 2.8 m d^{-1}, 2.9 m d^{-1}, and 2.6 m d^{-1}, respectively (see e.g. Fig 5. and Fig. 6). This is in agreement with the type of information conveyed by the dataset **D** (phreatic/shallow groundwater not affected by deep aquifers or local confined conditions).

In addition, significant differences from the values obtained in Table 4 are observed. These differences are explained by the estimation method of the posterior model probabilities. Values reported in Table 4 are calculated from the summation of individual likelihood values obtained from sampling the full hyperspace dimensioned by model structures, and forcing data (inputs) and parameter vectors. On the contrary, values reported in Table 7 are approximated using an exponential-type formula (equation 5). Thus, small fluctuations on the model selection criterion and, as a consequence, in the delta terms used in equation (5), will have a large influence on the resulting posterior model weights.

Fig. 11 shows the full BMA predictive distributions for groundwater budget terms obtained from criteria-based multimodel methodologies and the GLUE-BMA methodology. As expected, BMA predictive distributions obtained with alternative model selection criteria are somewhat different between them. Differences in the most likely predictive values are, in general, the largest between the values obtained using KIC and BIC. This is expected since these two criteria assigned much of the posterior weights to individual and completely opposite models; whereas BIC favours M1, KIC prefers M3 (Table 7). This reaffirms the idea that relying on a single conceptual model is likely to produce biased predictions. For the drain outflows differences between the most likely predicted values obtained from alternative multimodel methodologies are minimum.

The most significant impact of using alternative model selection criteria to approach posterior model probabilities is on the estimation of the predictive variance and the corresponding contributions from parameters, conceptual models, and scenarios. Since contributions to the predictive variance are weighted by the corresponding posterior model probabilities it is expected that the three components of the predictive variance described in section 2.2 (equation 4) will differ for results obtained using different model selection criteria.

Table 8 summarizes the predictive variances obtained using different model selection criteria. From this table it is observed that when the posterior weight of a given (and identical) conceptual model increases, which is equivalent to select a single conceptual model over the others, the values of the predictive variance decrease. This is explained by the fact that conceptual model uncertainty is neglected and, as a consequence, deviations from the average estimations as expressed by the second term (II) of equation (4) are not taken into account. For example, using AIC, AICc and BIC model M1 is assigned a posterior weight of 0.596, 0.845 and 0.972, respectively, thus, showing an increasing preference for model M1. Considering the river losses from the Velp, the predictive variances estimated using these posterior model probabilities correspond to 2.1×10^5 $(m^3\ d^{-1})^2$, 1.4×10^5 $(m^3\ d^{-1})^2$ and 9.9×10^4 $(m^3\ d^{-1})^2$, respectively. This reaffirms the idea that when a (single) conceptual model is preferred over the others, an underestimation of the predictive uncertainty is obtained. This is in full agreement with the results for a synthetic study case obtained by Rojas et al. (2009).

Additionally, Fig. 12 shows the predictive variance estimated using posterior model probabilities obtained from AIC (plate a), AICc (plate b), BIC (plate c), and KIC (plate d).

The predictive variance has been subdivided per source of uncertainty and each contribution has been expressed as a percentage of the predictive variance shown in Table 8. It is worth noting that for the case of BIC, which assigned 97 % of the posterior weight to model (M1), thus, showing a considerably preference for M1, 36 % of the predictive variance of the groundwater outflows from Walenbos comes from conceptual model uncertainty whereas for the river losses from the Demer, this contribution reaches 20%. The same two groundwater budget terms show the largest contributions of conceptual model uncertainty for AIC (plate a), AICc (plate b) and the GLUE-BMA method (Fig. 10). In the case of KIC (plate d) river gains from the Demer and groundwater outflows from Walenbos show the largest contribution of conceptual model uncertainty. Although the patterns showing the largest contributions of conceptual model uncertainty are rather similar for different model selection criteria, the values of these contributions substantially differed. For example, the contribution of conceptual model uncertainty to predictive variance for the Walenbos outflows ranged between 36% and 85% whereas for the river losses from the Demer the contribution varied between 20% and 76%. This clearly shows that using different model selection criteria may produce misleading and conflicting results.

A comparison of the capture zones obtained using the calibrated parameters from UCODE-2005 and the highest likelihood points from GLUE-BMA (Fig. 13) illustrates a relevant point. Capture zones are obtained with MODPATH (Pollock, 1994) using a forward particle tracking method from estimates of the average linear velocity using a constant effective porosity n_e of 0.1. These velocities are estimated from the simulated heads obtained with MODFLOW-2005. In general, the simulated flow fields, either obtained using calibrated parameters from UCODE-2005 or highest likelihood parameter from GLUE-BMA, are rather similar. This produces fairly similar capture zones between these approaches and between models M1, M2 and M3 despite the fact that posterior model probabilities may significantly differ between models. This is explained by the fact that the dataset **D** used to calibrate alternative conceptual models is based on the same head observations (Fig. 1), consequently, predictions of any variable closely linked to (or contained in) the data used for calibration will have a relatively low contribution of conceptual model uncertainty to predictive variance. However, as it is seen from the previous results, predicted variables not included in data set used for calibration are likely to have a significant contribution of conceptual model uncertainty. This is the case for variables like river gains and river losses from the Demer or

the Velp. These results are in full agreement with Harrar et al. (2003), Højberg and Refsgaard (2005) and Troldborg et al. (2007) whose results show that the relevance of conceptual model uncertainty increases when predicted variables are not included in the data set used for calibration.

6. Conclusions

In this work, we presented a multimodel approach to estimate the contributions to the predictive uncertainty arising from the definition of alternative conceptual models and optional recharge conditions. The proposed multimodel approach combines the GLUE and BMA methods, and it is an improved version of the approach originally developed by Rojas et al. (2008). The improvement consisted in replacing the traditional Latin Hypercube Sampling scheme of GLUE by a MCMC sampling scheme which, significantly, reduced computational times and increased the efficiency of the approach. We accounted for conceptual model and scenario (recharge) uncertainties in the modelling of several groundwater budget terms in the groundwater system of the Walenbos Nature Reserve in Belgium. For that, three conceptual models were proposed based on different levels of geological knowledge and two additional recharge settings accounting for deviations from average recharge conditions were used.

The study area is a hydrogeologically particular setup with deeply incised valleys promoting the contact between alternating aquifers and different hydrostratigraphic units. The fact that the wetness and the surface waters available at the Walenbos Nature Reserve are due solely to groundwater discharges (see e.g. Batelaan et al. 1998) is of vital importance and make the studied area an ecologically valued zone. Although we worked with relatively similar conceptual models, the predictive uncertainties in these essential groundwater flows showed to be very important for the Walenbos area. Therefore, whether the impacts of the differences between the alternative conceptual models are significant or not should be seen in the context of the present application.

The main findings of this work can be summarized as follows:

1. The adopted approach is flexible since (i) there is no limitation in the number or complexity of conceptual models that can be included, or to what degree input and parameter uncertainty can be incorporated, (ii) quantitative or qualitative information about the system can be used to distinguish between different simulators, (iii) the

closeness between the predictions and system observations can be defined in a variety of ways, and (iv) likelihoods, model probabilities and predictive distributions can be easily updated when new information becomes available. By definition, the results of the proposed methodology are conditional on the ensemble of proposed models and, therefore, the 'quality' of the uncertainty assessment is linked to the 'quality' of the sampling of conceptual models included in the ensemble (Neuman, 2003).

2. A set of 51 head observations did not allow a further discrimination between the three conceptual models proposed ending up in small differences in posterior model probabilities. This indicates that the information content of the head observations was rather low and that, for this case, the values of prior model probabilities may play an important role in the case they are not all taken equal. These prior model probabilities should be considered as the analyst's prior perception about the plausibility of the alternative conceptual models. In this context, the combination of prior expert knowledge about the conceptual models and the information given by the data will produce a better distinction between alternative conceptualizations. As shown by Rojas et al. (2009), the inclusion of proper and correct prior knowledge about the alternative conceptualizations will reduce the predictive uncertainty.

3. Despite the small differences in posterior model probabilities, predictive distributions showed to be considerably different in shape, central moment and spread among the alternative conceptualizations and scenarios analyzed. This reaffirms the idea that relying on a single conceptual model driven by a particular scenario, will likely produce biased and under-dispersive estimations of the predictive uncertainty.

4. The contribution of conceptual model uncertainty varied between 1 % and 75 % of the predictive uncertainty depending on the groundwater budget term. Additionally, the contribution of scenario uncertainty varied between 5 % and ca. 100 % of the predictive uncertainty depending on the budget term. The relative contribution of conceptual model uncertainty for the different groundwater budget terms provides useful information for updating the model concept or guiding data collection to optimally reduce conceptual uncertainty. If there had been better data available (e.g. dynamic heads, discharge values, travel time, hydraulic conductivity measurements, etc.) parametric uncertainty would have been reduced and possibly conceptual model uncertainty would have been a relatively

larger fraction of the predictive uncertainty. In addition, a better dataset **D** would likely allow a better discrimination between alternative conceptual models.
For scenario uncertainty contributions, on the other hand, useful information to reduce its contribution may be difficult to collect due to unknown and unpredictable future conditions. However, if future scenarios are linked to potential groundwater abstraction policies (Rojas and Dassargues, 2007), expert knowledge about the scenarios, in the form of prior scenario probabilities, could be included to optimally reduce the contribution of scenario uncertainty to predictive uncertainty.

5. Critical differences between the proposed approach and a traditional least squares calibration method were observed. The proposed approach successfully identified attraction zones (and the highest likelihood points) for all parameters which were contained within feasible and meaningful ranges. On the contrary, for relatively insensitive parameters across the three alternative conceptual models, the least squares method did not succeed in locating the highest likelihood point and, in the most critical case, the calibrated value was found outside the attraction zone defined by the proposed approach. This is due to equifinality and the fact that the dataset **D** did not contain enough information to identify unique parameter values.

6. The use of different model selection criteria to approximate posterior model probabilities in the frame of a multimodel methodology resulted in alternative conceptual models being ranked differently, in the calculation of dissimilar posterior model probabilities, in different estimations of the predictive uncertainty and in different estimations for the corresponding contributions to the predictive uncertainty from conceptual models and optional scenarios. In the frame of a multimodel approach, these issues are critical and can not be neglected.

7. Interestingly, for the extreme case when a single model was preferred over the others, a rather significant contribution of conceptual model uncertainty (36 %) to the predictive uncertainty was observed for the groundwater outflows from the Walenbos area. This clearly states that even for slight contributions from alternative models to the posterior weights, in this case 3 % from models M2 and M3, conceptual model uncertainty may play an important role and can not be neglected.

8. Results obtained from criteria-based multimodel methodologies reaffirms the idea that relying on predictions obtained using a single conceptual model is likely to produce biased estimations of the predictive uncertainty. Additionally, results obtained from alternative model selection criteria may be ambiguous in indicating the contributions of conceptual model and scenario uncertainties producing serious implications in planning future data collection campaigns.

9. Results from the proposed methodology as wells as results from traditional parameter calibration show that the relevance of conceptual model uncertainty increases when predicted variables are not included in the data used for calibration. This is in full agreement with the results of Harrar et al. (2003), Højberg and Refsgaard (2005) and Troldborg et al. (2007).

10. The results of this study strongly advocate the idea to address conceptual model uncertainty in the practice of groundwater modeling. Additionally, to account for unforeseen future circumstances, including scenario uncertainty permits to obtain more realistic, and possibly, more reliable estimations of the predictive uncertainty. The use of a single model may result in smaller uncertainty intervals, hence an increased confidence in the model simulations, but is very likely prone to statistical bias. Also, in the presence of conceptual model uncertainty, which per definition can not be excluded, this gain in accuracy in the short-term may have serious implications when the model is used for long-term predictions in which the system is subject to new stresses. It is therefore advisable to explore a number of alternative conceptual models and scenarios to obtain predictions that are more realistic, hence, that are more likely to include the unknown true system responses.

Acknowledgements

The first author thanks the Katholieke Universiteit Leuven (K.U.Leuven) for providing financial support in the framework of PhD IRO-scholarships. We also wish to thank Roberta-Serena Blasone for helpful comments on implementing MCMC in the frame of the GLUE methodology.

References

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

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716-723.

Anderson, M., Woessner, W., 1992. Applied groundwater modelling - Simulation of flow and advective transport. Academic Press, San Diego. 381 pp.

Batelaan, O., De Smedt, F., 2004. SEEPAGE, a new MODFLOW DRAIN Package. Ground Water 42 (4), 576-588. doi:10.1111/j.1745-6584.2004.tb02626.x.

Batelaan, O., De Smedt, F., 2007. GIS-based recharge estimation by coupling surface-subsurface water balances. Journal of Hydrology 337 (3-4), 337-355. doi:10.1016/j.hydrol.2007.02.001.

Batelaan, O., De Smedt, F., De Becker, P., Huybrechts, W., 1998. Characterization of a regional groundwater discharge area by combined analysis of hydrochemistry, remote sensing and groundwater modelling. In: Shallow Groundwater Systems. Dillon, P., Simmers, I. (Editors), International contributions to hydrogeology 18. A.A. Balkema, Rotterdam. 75-86 pp.

Batelaan, O., De Smedt, F., Otero Valle, M., Huybrechts, W., 1993. Development and application of a groundwater model integrated in the GIS GRASS. In: Kovar, K., Nachtebel, H. (Editors), Application of geographic information systems in hydrology and water resources management. IAHS Publ. No 211. 581-590 pp.

Batelaan, O., Meyus, Y., De Smedt, F., 2007. De grondwatervoeding van Vlaanderen. Water 28, 64-71. Available on: http://www.tijdschriftwater.be/water28-15HI.pdf.

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

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

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology 249 (1-4), 11-29. doi:10.1016/S0022-1694(01)00421-8.

Binley, A., Beven, K., 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.

1 Blasone, R-S., Madsen, H., Rosbjerg, D., 2008a. Uncertainty assessment of integrated
2 distributed hydrological models using GLUE with Markov chain Monte Carlo sampling.
3 Journal of Hydrology 353 (1-2), 18-32. doi:10.1016/j.jhydrol.2007.12.026.

4 Blasone, R-S., Vrugt, J., Madsen, H., Rosbjerg, D., Robinson, B., Zyvoloski, G., 2008b.
5 Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain
6 Monte Carlo sampling. Advances in Water Resources 31 (4), 630-648.
7 doi:10.1016/j.advwatres.2007.12.003.

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

10 Bredehoeft, J., 2005. The conceptualization model problem - surprise. Hydrogeology Journal
11 13 (1), 37-46. doi:10.1007/s10040-004-0430-5.

12 Brooks, S., Gelman, A., 1998. General methods for monitoring convergence of iterative
13 simulations. Journal of Computational and Graphics Statistics 7 (4), 434-455.

14 Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. The
15 American Statistician 49 (4), 327-335.

16 Cools, J., Meyus, Y., Woldeamlak, S., Batelaan, O., De Smedt, F., 2006. Large-scale GIS-
17 based hydrogeological modeling of Flanders: a tool for groundwater management.
18 Environmental Geology 50 (8), 1201-1209. doi:10.1007/s00254-006-0292-3.

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

21 De Becker, P., Huybrechts, W., 1997. Het Walenbos - Ecohydrologische Atlas. Rep. Nr. IN
22 97/03. 76 pp.

23 DOV, (2008). Databank Ondergrond Vlaanderen.
24 http://dov.vlaanderen.be/dovweb/html/engels.html.

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

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

30 Gaganis, P., Smith, L., 2006. Evaluation of the uncertainty of groundwater model predictions
31 associated with conceptual errors: A per-datum approach to model calibration. Advances in
32 Water Resources 29 (4), 503-514. doi:10.1016/j.advwatres.2005.06.006.

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

36 Gelman, A., Carlin, J., Stern, H., Rubin, D., 2004. Bayesian data analysis. Chapman &
37 Hall/CRC, Boca Raton. 668 pp.

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

2 Gilks, W., Richardson, S., Spiegelhalter, D., 1995. Markov Chain Monte Carlo in Practice.
3 Chapman & Hall/CRC, Boca Raton. 486 pp.

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

6 Gullentops, F., Bogemans, F., De Moor, G., Palissen, E., Pissart, A., 2001. Quaternary
7 lithostratigraphic units (Belgium). Geologica Belgica 4 (1-2), 153-164.

8 HAECON, Witteveen en Bos, 2004. Ontwikkelen van regionale modellen ten behoeve van het
9 vlaams grondwater model (VGM) in GMS/MODFLOW: Perceel Nr. 3 Brulandkrijtmodel
10 (Development of regional models for the Flemish Groundwater Model (VGM) in
11 GMS/MODFLOW. AMINAL, afdeling WATER. 159 pp.

12 Harbaugh, A., 2005. MODFLOW-2005, the U.S. geological Survey modular ground-water
13 model-the Ground-Water Flow Process. Techniques and Methods 6-A16. U.S. Geological
14 Survey. 253 pp.

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

18 Hassan, A, 2004. A methodology for validating numerical ground water models. Ground
19 Water 42 (3), 347-362. doi:10.1111/j.1745-6584.2004.tb02683.x.

20 Hastings, W., 1970. Monte Carlo sampling methods using Markov chains and their
21 applications. Biometrika 57 (1), 97-109.

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

24 Hill, M., Tiedeman, C., 2007. Effective groundwater model calibration: With analysis of data,
25 sensitivities, predictions, and uncertainty. First ed., 480 pp. John Wiley & Sons Inc. New
26 Jersey.

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

29 Højberg, A., Refsgaard, J. C., 2005. Model uncertainty -- parameter uncertainty versus
30 conceptual models. Water Science & Technology 52 (6), 177-186.

31 Hurvich, C., Tsai, C., 1989. Regression and time series model selection in small sample.
32 Biometrika 76 (2), 99-104.

33 Jensen, J., 2003. Parameter and uncertainty estimation in groundwater modelling. PhD thesis.
34 Department of Civil Engineering. Aalborg University. Denmark. Series Paper Nr. 23. 139
35 pp.

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

1 Kass, R., Raftery, A., 1995. Bayes factors. Journal of the American Statistical Association 90
2 (430), 773-795.

3 Laga, P., Louwye, S., Geets, S., 2001. Paloegene and Neogene lithostratigraphie units
4 (Belgium). Geologica Belgica 4 (1-2), 135-152.

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

8 McKay, D., Beckman, R., Conover, W., 1979. A comparison of three methods for selecting
9 values of input variables in the analysis of output from a computer code. Technometrics 21
10 (2), 239-245.

11 Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953. Equation of state
12 calculations by fast computing machines. The Journal of Chemical Physics 21 (6), 1087-
13 1092.

14 Meyer, P., Ye, M., Neuman, S., Cantrell, K., 2004. Combined estimation of hydrogeologic
15 conceptual model and parameter uncertainty. Report NUREG/CR-6843 PNNL-14534. US
16 Nuclear Regulatory Commission. 42 pp.

17 Meyer, P., Ye, M., Rockhold, M., Neuman, S., Cantrell, K., 2007. Combined estimation of
18 hydrogeologic conceptual model parameter and scenario uncertainty with application to
19 uranium transport at the Hanford Site 300 area. Report NUREG/CR-6940 PNNL-16396.
20 U.S. Nuclear Regulatory Commission. 124 pp.

21 Meyus, Y., Batelaan, O., De Smedt, F., 2000. Concept Vlaams Grondwater Model (VGM):
22 Technicsh concept van het VGM; deelraport 1: Hydrogeologische codering van de
23 ondergrond van Vlaanderen (HCOV) (Technical concept of the Flemish groundwtaer
24 model: Report 1: Hydrogeological coding of the subsoil of Flanders) [In Dutch].
25 AMINAL, afdeling Water.

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

29 Neuman, S., Wierenga, P., 2003. A comprehensive strategy of hydrogeologic modelling and
30 uncertainty analysis for nuclear facilities and sites. Report NUREG/CR-6805. US Nuclear
31 Regulatory Commission. 225 pp.

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

34 Poeter, E., Hill, M., Banta, E., Mehl, S., Christensen, S., 2005. UCODE_2005 and six other
35 computer codes of runiversal sensitivity analysis, calibration, and uncertainty evaluation.
36 Technical Methods 6-A11. U.S. Geological Survey. 283 pp.

37 Pollock, D., 1994. User's guide for MODPATH/MODPATH-PLOT, Version 3: A particle
38 tracking post-processing package for MODFLOW, the U.S. Geological Survey finite-
39 difference ground-water flow model. Open-File Report 94-464. U.S. Geological Survey.
40 249 pp.

Refsgaard, J. C., van der Sluijs, J., Brown, J., van der Keur, P., 2006. A framework for dealing with uncertainty due to model structure error. Advances in Water Resources 29 (11), 1586-1897. doi:10.1016/j.advwatres.2005.11.013.

Refsgaard, J. C., van der Sluijs, J., Højberg, A., Vanrolleghem, P., 2007. Uncertainty in the environmental modelling process - A framework and guidance. Environmental Modelling & 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. Springer Texts in Statistics, Springer-Verlag, New York. 602 pp.

Rojas, R., Dassargues, A., 2007. Groundwater flow modelling of the regional aquifer of the Pampa del Tamarugal, northern Chile. Hydrogeology Journal 15 (3), 537-551. doi:10.1007/s10040-006-0084-6.

Rojas, R., Feyen, L., Dassargues, A., 2008. Conceptual model uncertainty in groundwater modelling: combining generalized likelihood uncertainty estimation and Bayesian model averaging. Water Resources Research, 44, W12418, doi:10.1029/2008WR006908.

Rojas, R., Feyen, L., Dassargues, A., 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.

Romanowicz, R., Beven, K., Tawn, J. 1994. Evaluation of prediction uncertainty in non-linear hydrological models using a Bayesian approach. In: Statistics for the environment 2 - Water related issues, Barnett, V. and Turkman, F. (Editors). John Wiley & Sons Inc., Chichester. 297-317 pp.

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

Schwartz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6 (2), 461-464.

Seifert, D., Sonnenberg, T., Scharling, P., Hinsby, K., 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.

Sorensen, D., Gianola, D., 2002. Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer-Verlag, New York. 740 pp.

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

Troldborg, L., Refsgaard, J, Jensen, K, Engesgaard, P., 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.

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

3 Ye, M., Meyer, P., Neuman, S., 2008a. On model selection criteria in multimodel analysis.
4 Water Resources Research 44. doi:10.1029/2008WR006803.

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

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

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

14 Ye, M., Pohlmann, K., Chapman, J., Shafer, D., 2006. On evaluation of recharge model
15 uncertainty: a priori and a posteriori. In: International High-Level Radioactive Waste
16 Management Conference. Las Vegas, Nevada U.S. 12 pp.

Figure captions

Figure 1: Location of the study area, river network, and location of 51 observation wells used as dataset **D** for the application of the multimodel methodology.

Figure 2: Geological map of the study area.

Figure 3: Layer setup for three alternative conceptual models M1, M2 and M3.

Figure 4: Series (chains) obtained from the M-H algorithm for the six parameters: (a) drain conductance, (b) conductance of the Velp River bed, (c) conductance of the Demer River bed, (d) hydraulic conductivity layer 1 (HK-1), (e) hydraulic conductivity layer 2 (HK-2), and (f) hydraulic conductivity layer 3 (HK-3) included in the conceptual model M2.

Figure 5: Marginal scatter plots of calculated likelihood using the M-H algorithm for parameters: (a) drain conductance, (b) conductance of the Velp River bed, (c) conductance of the Demer River bed, (d) hydraulic conductivity layer 1 (HK-1) for model M1. Vertical black line represents solution obtained from calibration using least squares (UCODE-2005). Red diamond represents point of highest likelihood in the context of the GLUE methodology.

Figure 6: Marginal scatter plots of calculated likelihood using the M-H algorithm for parameters: (a) drain conductance, (b) conductance of the Velp River bed, (c) conductance of the Demer River bed, (d) hydraulic conductivity layer 1 (HK-1), (e) hydraulic conductivity layer 2 (HK-2), and (f) hydraulic conductivity layer 3 (HK-3) for model M2. Vertical black line represents solution obtained from calibration using least squares (UCODE-2005). Red diamond represents point of highest likelihood in the context of the GLUE methodology.

Figure 7: Marginal scatter plots of calculated likelihood using the M-H algorithm for parameters: (a) hydraulic conductivity layer 2 (HK-2), and (b) hydraulic conductivity layer 3 (HK-3) for model M3. Vertical black line represents solution obtained from calibration using least squares (UCODE-2005). Red diamond represents point of highest likelihood in the context of the GLUE methodology.

Figure 8: Cumulative predictive distributions for groundwater budget terms for alternative conceptual models M1, M2, M3 and the BMA cumulative prediction accounting exclusively for conceptual model uncertainty under scenario S2.

Figure 9: BMA cumulative predictive distributions for groundwater budget terms for alternative scenarios S1, S2 and S3 and the Full BMA cumulative prediction accounting for conceptual model and scenario uncertainties.

Figure 10: Sources of variance expressed as a percentage of the predictive variance calculated using equation (4) for groundwater flow components. (L stands for losses, G stands for gains, I stands for inflows and O stands for outflows).

Figure 11: Comparison of full BMA cumulative predictive distributions for groundwater budget terms between criteria-based multimodel methodologies and GLUE-BMA.

Figure 12: Sources of variance expressed as a percentage of the predictive variance calculated using equation (4) for groundwater flow components for criteria-based multimodel methodologies: (a) AIC-based, (b) AICc-based, (c) BIC-based, and (d) KIC-based. (L stands for losses, G stands for gains, I stands for inflows and O stands for outflows).

Figure 13: Forward particle tracking defining the capture zone for steady-state (calibrated) results obtained from UCODE-2005 (first row) and highest likelihood point in GLUE-BMA (second row) for models M1 (a and d), M2 (b and e) and M3 (c and f).

1 **Tables**

2 Table 1: Lithostratigraphic description of formations present in the study area

Time	Lithostratigraphy		Lithology
	Group	Formation	
Quaternary	Eolian deposits		Loam and Sandy loam
	Alluvial deposits		Sand, Silt, Clay, possible Gravel to base
Miocene	Diest		Coarse sand with glauconite and iron sand toe banks
	Bolderberg		Fine sand with mica
Oligocene	Rupel	Boom	Clay with septarien
		Bilzen	Fine sand with shell rests
	Tongeren	Borgloon	Clay and coarse sand
		Sint Huibrechts Hern	Fine sand with glauconite and mica
Eocene	Zenne	Brussels	Fine sand
	Ieper	Kortrijk	Clay & traces of fine sand
Paleoceen	Landen	Hannut	Fine to silty sand

1 Table 2: Hydrostratigaphic unit setup for conceptual models M1, M2 and M3

Formation	Hydraulic conductivity parameter		
	Model M1	Model M2	Model M3
Eolian and alluvial deposits	HK-1	HK-1	HK-1
Diest			
Bolderberg			
Boom		HK-2	HK-2
Bilzen			HK-3
Borgloon			
Sint Huibrechts Hern			HK-4
Brussels		HK-3	HK-5

Table 3: Range of prior uniform distributions for unknown parameters common to the three conceptual models M1, M2 and M3

Parameter		Range	
		Minimum	Maximum
River Demer conductance	($m^2 d^{-1}$)	0	1.0e04
River Velp conductance	($m^2 d^{-1}$)	0	1.0e04
Drain conductance	($m^2 d^{-1}$)	0	1.0e04
Hydraulic conductivities Layer 1 to Layer 5	($m d^{-1}$)	0	50

Table 4: Integrated model likelihoods, prior model probabilities, and posterior model probabilities obtained for average recharge conditions (scenario S2) for alternative conceptual models

	Conceptual models		
	M1	M2	M3
Integrated model likelihood $p(\mathbf{D}\|M_k)$	2210.5	1966.5	2058.1
Prior model probability $p(M_k)$	0.33	0.33	0.33
Posterior model probabilities $p(M_k\|\ \mathbf{D})$	**0.355**	**0.315**	**0.330**

Table 5: Prediction intervals (95 %) and most likely predicted value based on the cumulative predictive distributions obtained from the GLUE-BMA methodology for different groundwater budget terms for scenarios S1, S2 and S3 for conceptual models M1, M2 and M3. All values expressed in $m^3 d^{-1}$

		Conceptual models								
		M1			M2			M3		
		$P_{2.5}$	P_{50}	$P_{97.5}$	$P_{2.5}$	P_{50}	$P_{97.5}$	$P_{2.5}$	P_{50}	$P_{97.5}$
S1	Demer losses	1499	4083	4488	910	2337	3756	118	1129	1327
	Demer gains	3339	7488	8102	2801	5993	8167	609	4955	7302
	Velp losses	1025	1321	1392	419	1270	2891	125	1030	2499
	Velp gains	1662	1951	2002	1142	1989	3258	749	1943	3222
	Drain outflows	46557	46886	48440	45277	46610	48234	44445	46303	49540
	Walenbos outflows	985	1027	1066	297	559	756	271	593	825
	Walenbos inflows	3589	3658	3709	2241	2808	3255	2951	3346	3749
S2	Demer losses	1018	3636	4033	543	1946	3373	42	881	1608
	Demer gains	3194	8021	8668	2760	6534	8837	622	5567	8231
	Velp losses	634	876	930	193	760	1880	63	563	1464
	Velp gains	2914	3433	3525	2134	3600	5706	1188	3556	5642
	Drain outflows	97504	97947	100127	95854	97392	99762	94613	97128	101384
	Walenbos outflows	1055	1090	1124	303	629	848	287	624	875
	Walenbos inflows	4923	5050	5158	3399	4051	4648	4348	4846	5377
S3	Demer losses	946	3125	3499	416	1534	2796	11	632	1275
	Demer gains	4283	8880	9590	3755	7473	9811	873	6623	9678
	Velp losses	537	757	805	170	635	1637	54	450	1196
	Velp gains	3457	4111	4214	2663	4260	6316	1399	4217	6540
	Drain outflows	153395	153902	156704	151157	153342	156248	150140	153214	158841
	Walenbos outflows	1130	1165	1203	353	678	910	328	664	930
	Walenbos inflows	5844	5995	6127	4188	4853	5579	5246	5832	6414

Table 6: Prediction intervals (95 %) and most likely predicted value based on the full BMA cumulative predictive distribution obtained from the GLUE-BMA methodology for different groundwater budget terms. All values expressed in $m^3\ d^{-1}$

	$P_{2.5}$	P_{50}	$P_{97.5}$
Demer losses	143	1865	4309
Demer gains	1690	6957	9420
Velp losses	129	843	2280
Velp gains	1180	3377	6059
Drain outflows	45334	97775	156344
Walenbos outflows	319	739	1186
Walenbos inflows	2517	4675	6161

Table 7: Summary of posterior model probabilities for alternative model selection criteria and the proposed methodology for models M1, M2 and M3

	Conceptual model		
	M1	M2	M3
Nr. Observations	51	51	51
SWSR*	180.95	182.18	158.18
MLOFO**	64.59	64.93	57.73
Ln |F|***	-122.75	-117.88	-102.18
$p(M_k)$	1/3	1/3	1/3
AIC	74.59	78.93	75.73
RANK (AIC)	1	3	2
$p(M_k\|\mathbf{D})$ (AIC)	**0.596**	**0.068**	**0.337**
AICc	75.92	81.54	80.12
RANK (AICc)	1	3	2
$p(M_k\|\mathbf{D})$ (AICc)	**0.845**	**0.051**	**0.104**
BIC	84.25	92.46	93.11
RANK (BIC)	1	2	3
$p(M_k\|\mathbf{D})$ (BIC)	**0.972**	**0.016**	**0.012**
KIC	-5.99	-6.68	-10.48
RANK (KIC)	3	2	1
$p(M_k\|\mathbf{D})$ (KIC)	**0.085**	**0.119**	**0.796**

* SWSR: Sum of weighted squared residuals.

** MLOFO: Maximum likelihood objective function observations obtained from UCODE-2005 (Poeter et al. 2005).

*** Ln |F|: Natural log of the determinant of the Fisher Matrix.

Table 8: Predictive variance estimated using posterior model probabilities based on alternative model selection criteria (AIC, AICc, BIC, KIC) and the GLUE-BMA proposed methodology. All values in $(m^3\ d^{-1})^2$

	AIC	AICc	BIC	KIC	GLUE-BMA
Demer losses	2.606×10^6	1.485×10^6	6.822×10^5	1.632×10^6	1.588×10^6
Demer gains	5.636×10^6	3.242×10^6	1.596×10^6	5.743×10^6	3.751×10^6
Velp losses	2.056×10^5	1.408×10^5	9.892×10^4	2.556×10^5	2.610×10^5
Velp gains	1.556×10^6	1.191×10^6	9.761×10^5	2.257×10^6	1.574×10^6
Drain outflows	1.924×10^9	1.905×10^9	1.898×10^9	1.961×10^9	1.912×10^9
Walenbos outflows	1.264×10^5	6.679×10^4	2.261×10^4	6.868×10^4	7.300×10^4
Walenbos inflows	1.393×10^6	1.285×10^6	1.157×10^6	9.235×10^5	1.151×10^5

Figures

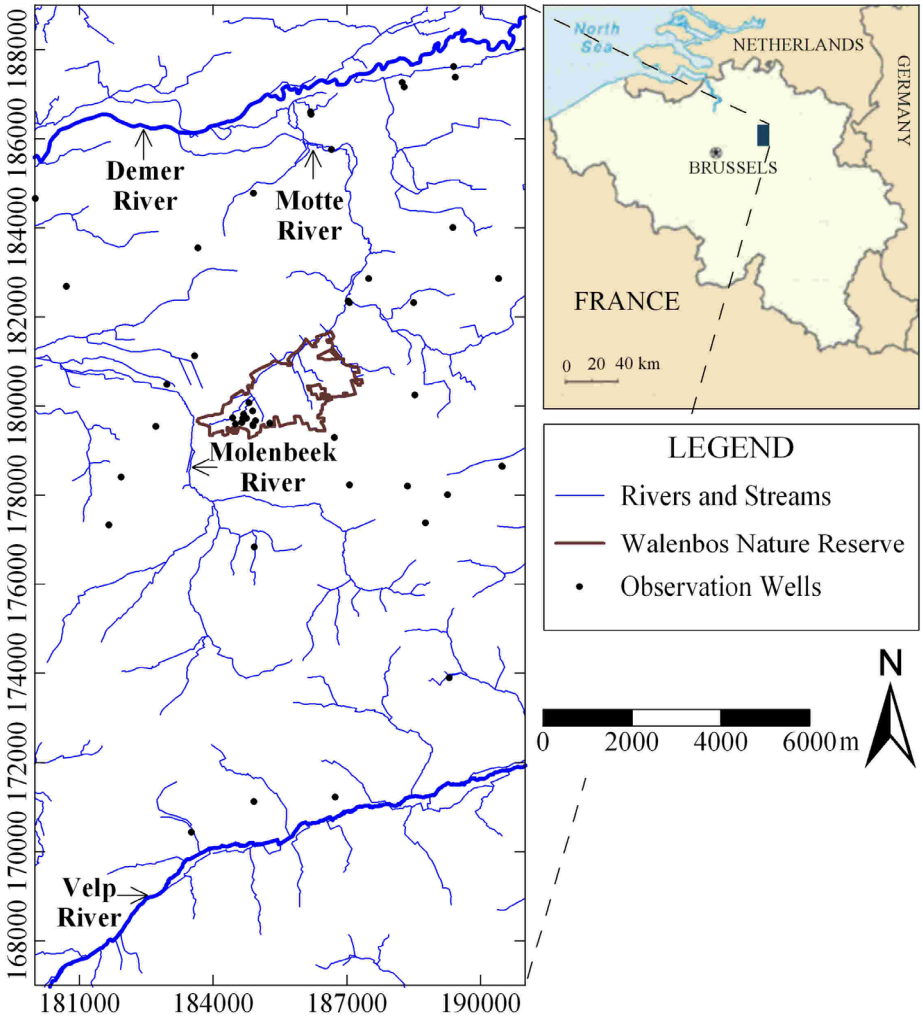


Figure 1: Location of the study area, river network, and location of 51 observation wells used as dataset **D** for the application of the multimodel methodology

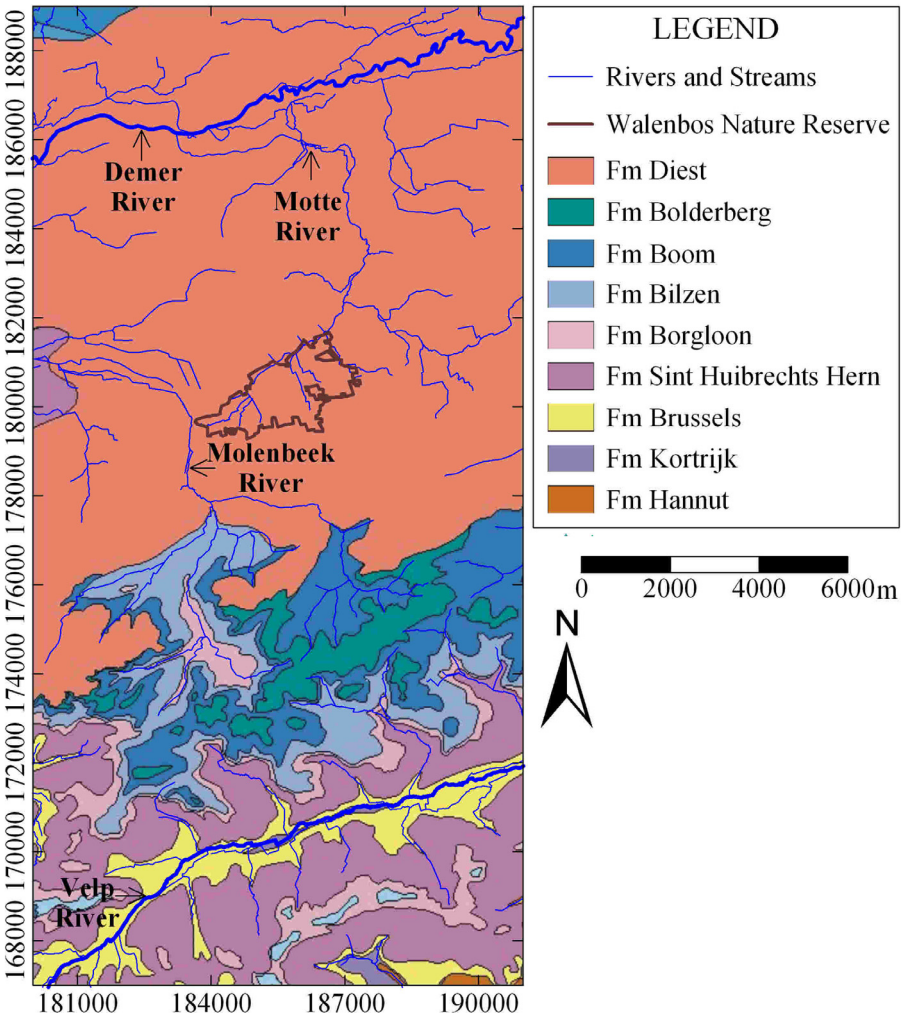
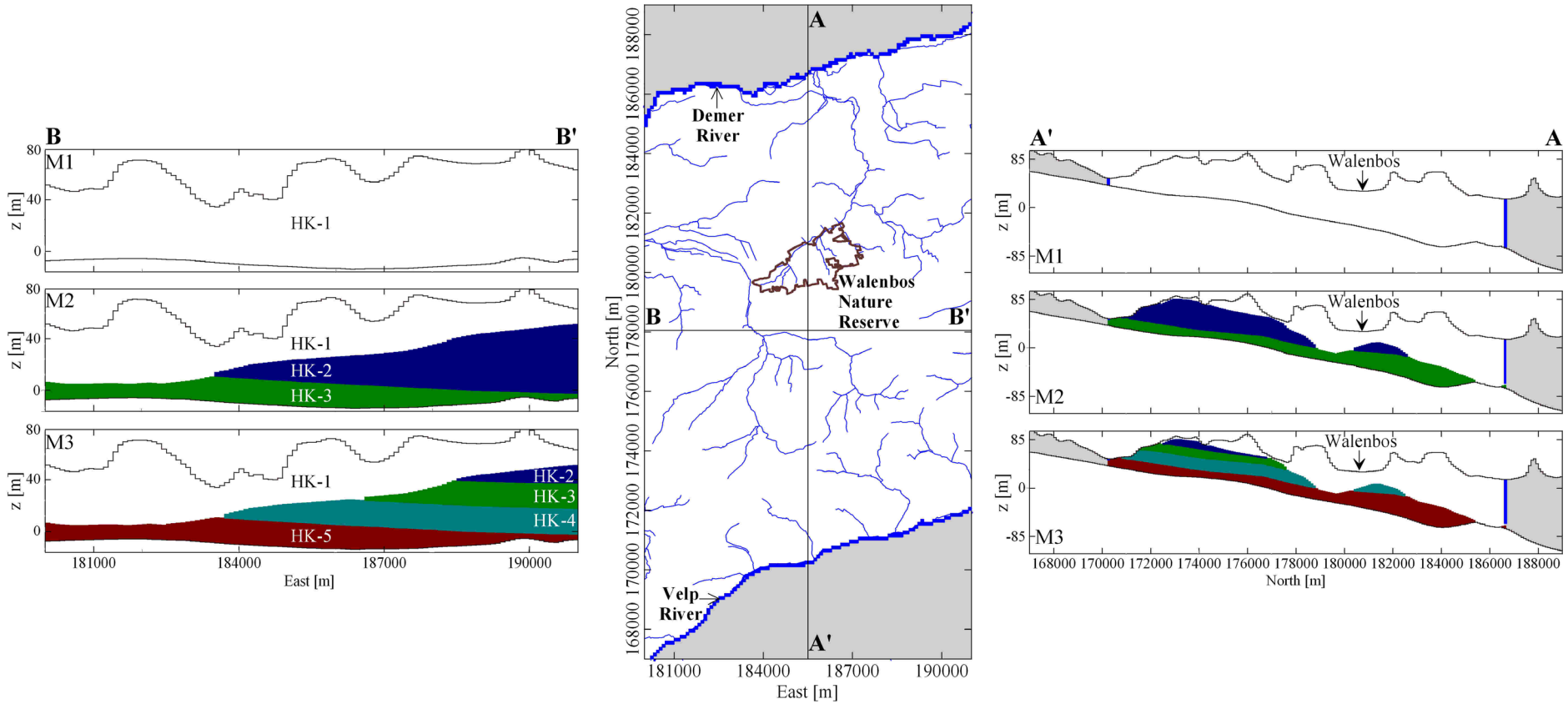


Figure 2: Geological map of the study area.



1

2 Figure 3: Model setup for three alternative conceptual models M1 (upper row), M2 (middle row) and M3 (lower row). Details for each

3 hydrostratigraphic unit are described in Table 2

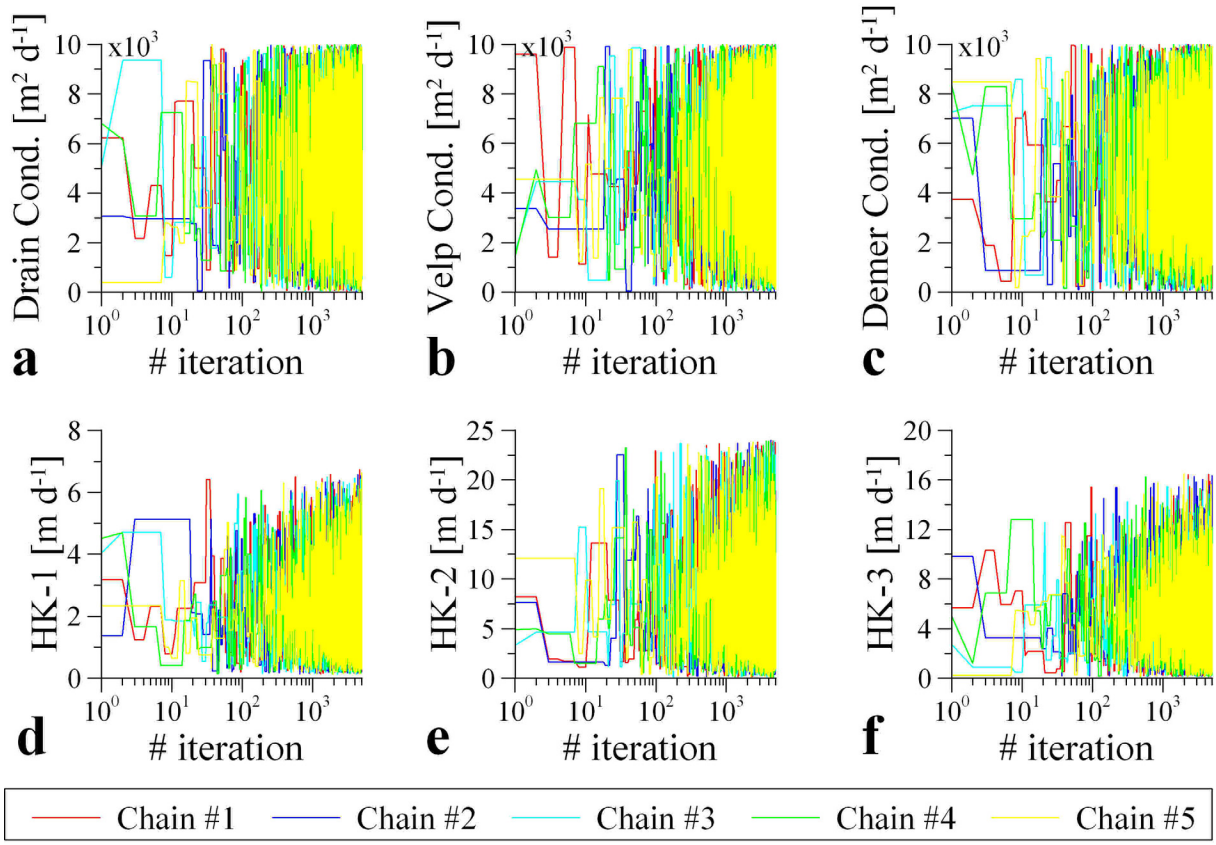


Figure 4: Series (chains) obtained from the M-H algorithm for the six parameters: (a) drain conductance, (b) conductance of the Velp River bed, (c) conductance of the Demer River bed, (d) hydraulic conductivity layer 1 (HK-1), (e) hydraulic conductivity layer 2 (HK-2), and (f) hydraulic conductivity layer 3 (HK-3) included in the conceptual model M2

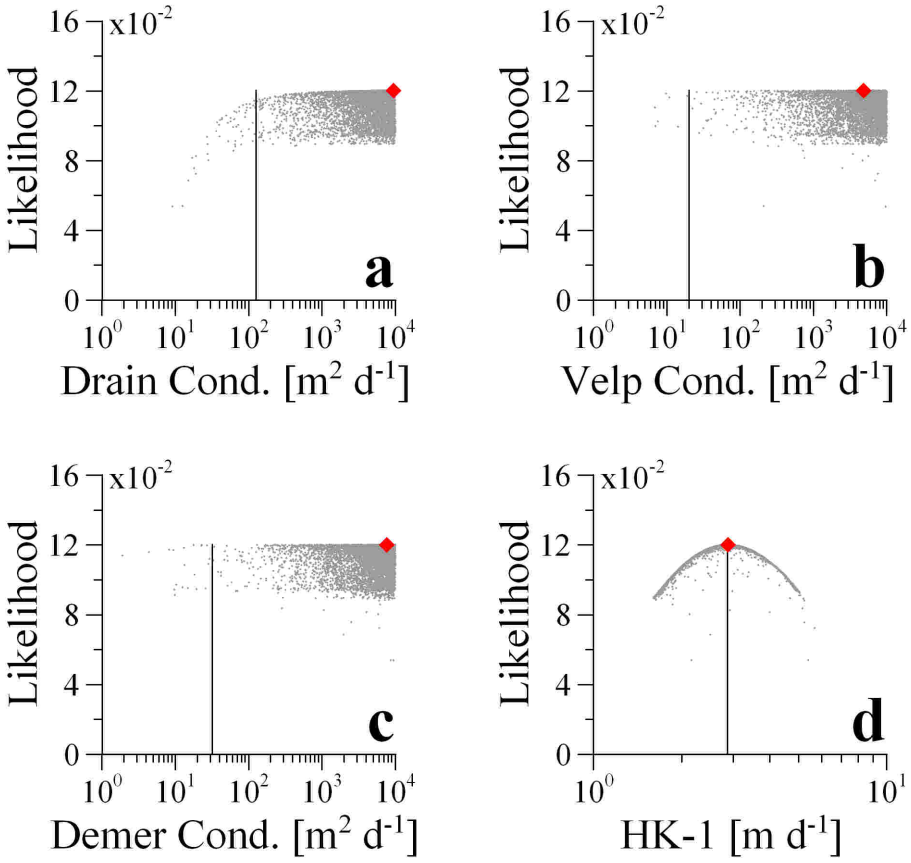


Figure 5: Marginal scatter plots of calculated likelihood using the M-H algorithm for parameters: (a) drain conductance, (b) conductance of the Velp River bed, (c) conductance of the Demer River bed, (d) hydraulic conductivity layer 1 (HK-1) for model M1. Vertical black line represents solution obtained from calibration using least squares (UCODE-2005). Red diamond represents point of highest likelihood in the context of the GLUE methodology

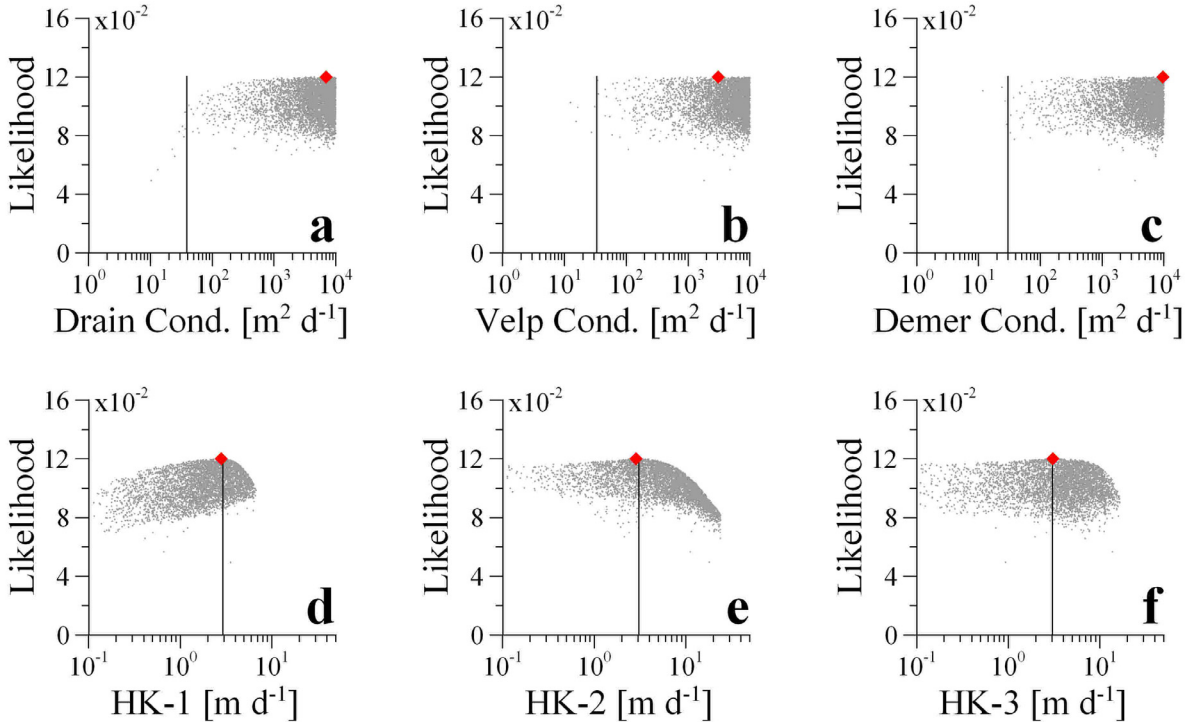


Figure 6: Marginal scatter plots of calculated likelihood using the M-H algorithm for parameters: (a) drain conductance, (b) conductance of the Velp River bed, (c) conductance of the Demer River bed, (d) hydraulic conductivity layer 1 (HK-1), (e) hydraulic conductivity layer 2 (HK-2), and (f) hydraulic conductivity layer 3 (HK-3) for model M2. Vertical black line represents solution obtained from calibration using least squares (UCODE-2005). Red diamond represents point of highest likelihood in the context of the GLUE methodology

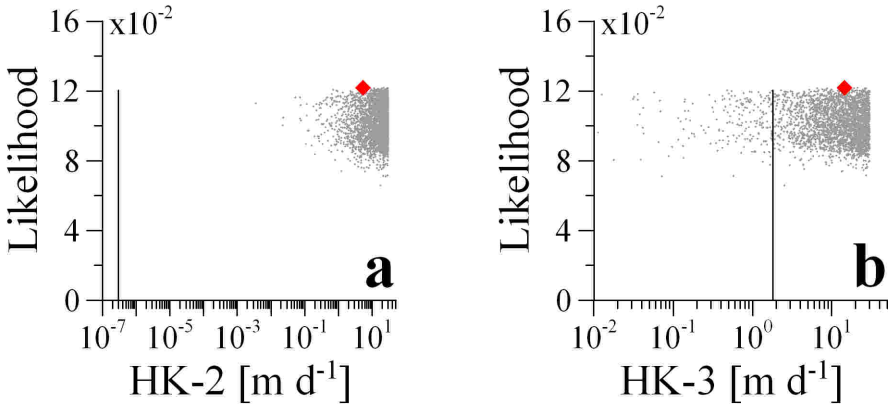


Figure 7: Marginal scatter plots of calculated likelihood using the M-H algorithm for parameters: (a) hydraulic conductivity layer 2 (HK-2), and (b) hydraulic conductivity layer 3 (HK-3) for model M3. Vertical black line represents solution obtained from calibration using least squares (UCODE-2005). Red diamond represents point of highest likelihood in the context of the GLUE methodology

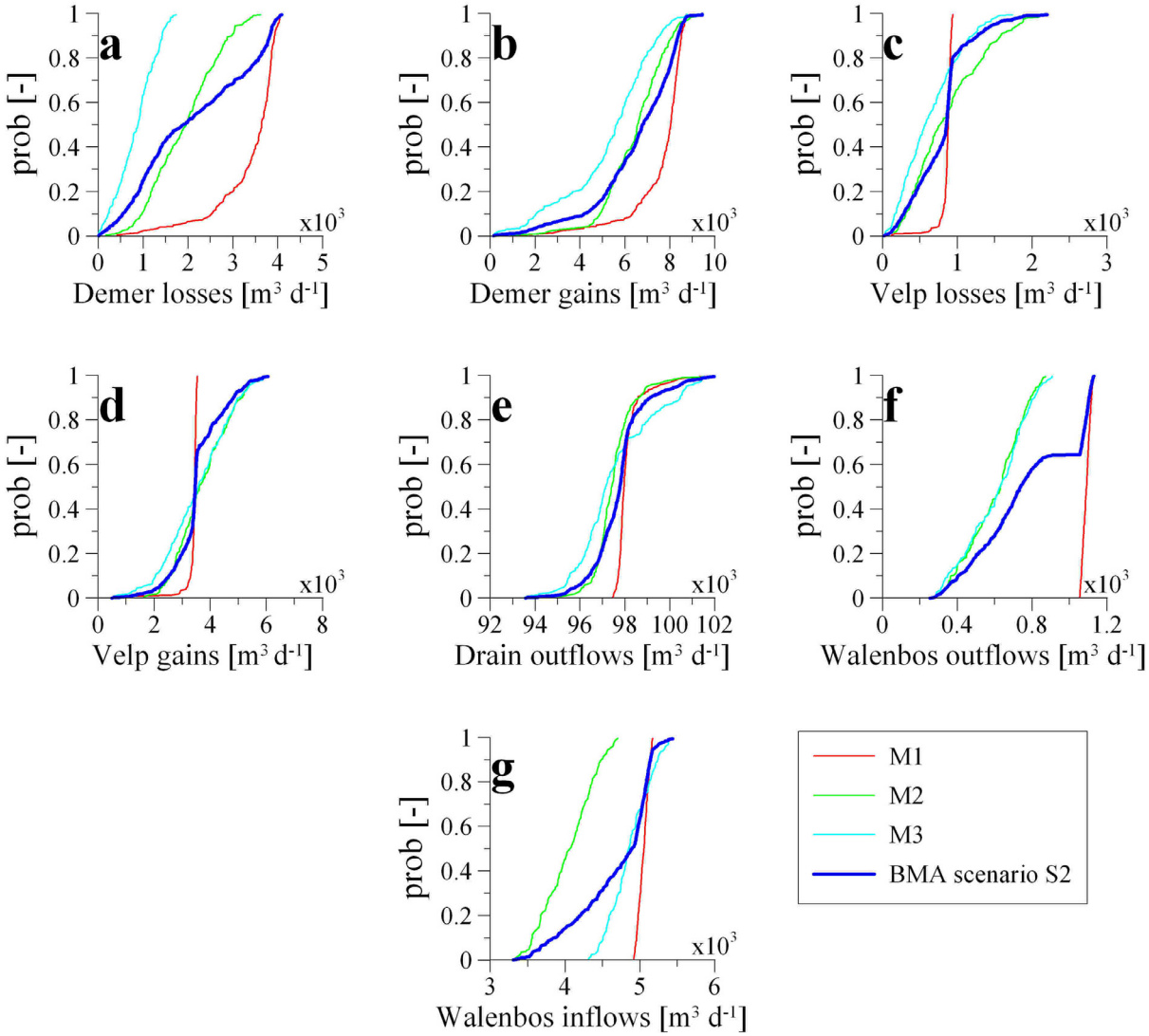


Figure 8: Cumulative predictive distributions for groundwater budget terms for alternative conceptual models M1, M2, M3 and the BMA cumulative prediction accounting exclusively for conceptual model uncertainty for scenario S2

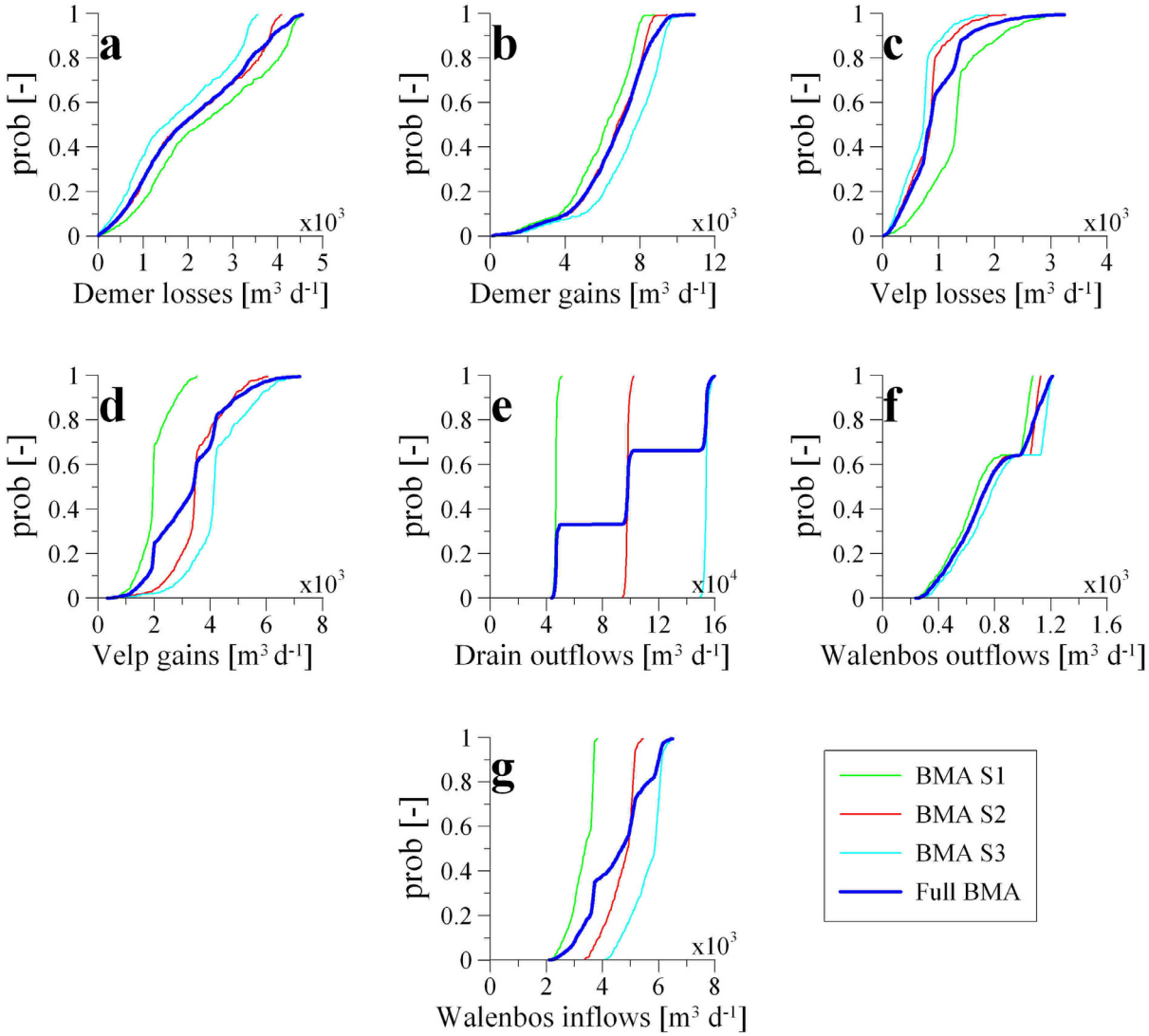


Figure 9: BMA cumulative predictive distributions for groundwater budget terms for alternative scenarios S1, S2 and S3 and the Full BMA cumulative prediction accounting for conceptual model and scenario uncertainties

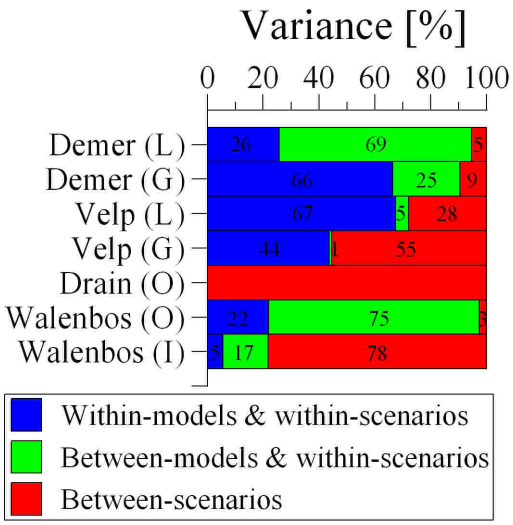


Figure 10: Sources of variance expressed as a percentage of the predictive variance calculated using equation (4) for groundwater flow components. (L stands for losses, G stands for gains, I stands for inflows and O stands for outflows)

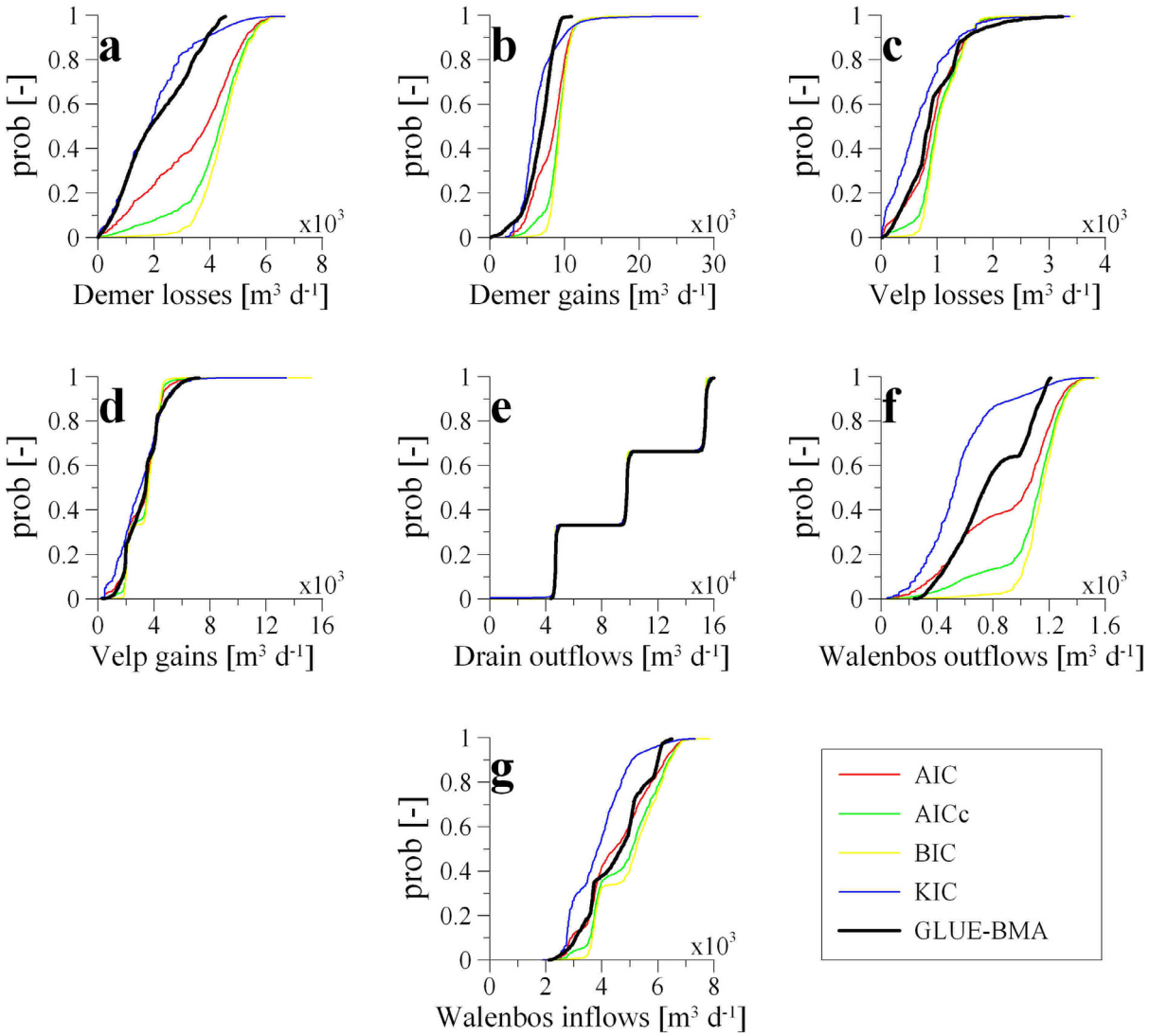


Figure 11: Comparison of full BMA cumulative predictive distributions for groundwater budget terms between criteria-based multimodel methodologies and GLUE-BMA

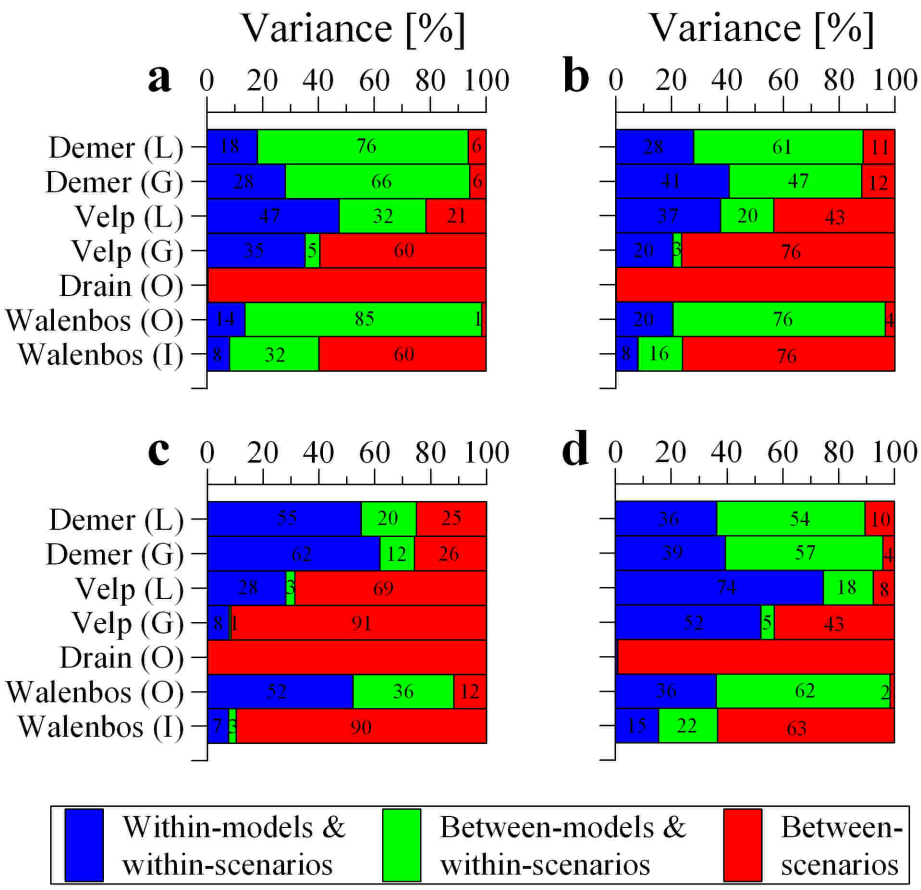


Figure 12: Sources of variance expressed as a percentage of the predictive variance calculated using equation (4) for groundwater flow components for criteria-based multimodel methodologies: (a) AIC-based, (b) AICc-based, (c) BIC-based, and (d) KIC-based. (L stands for losses, G stands for gains, I stands for inflows and O stands for outflows)

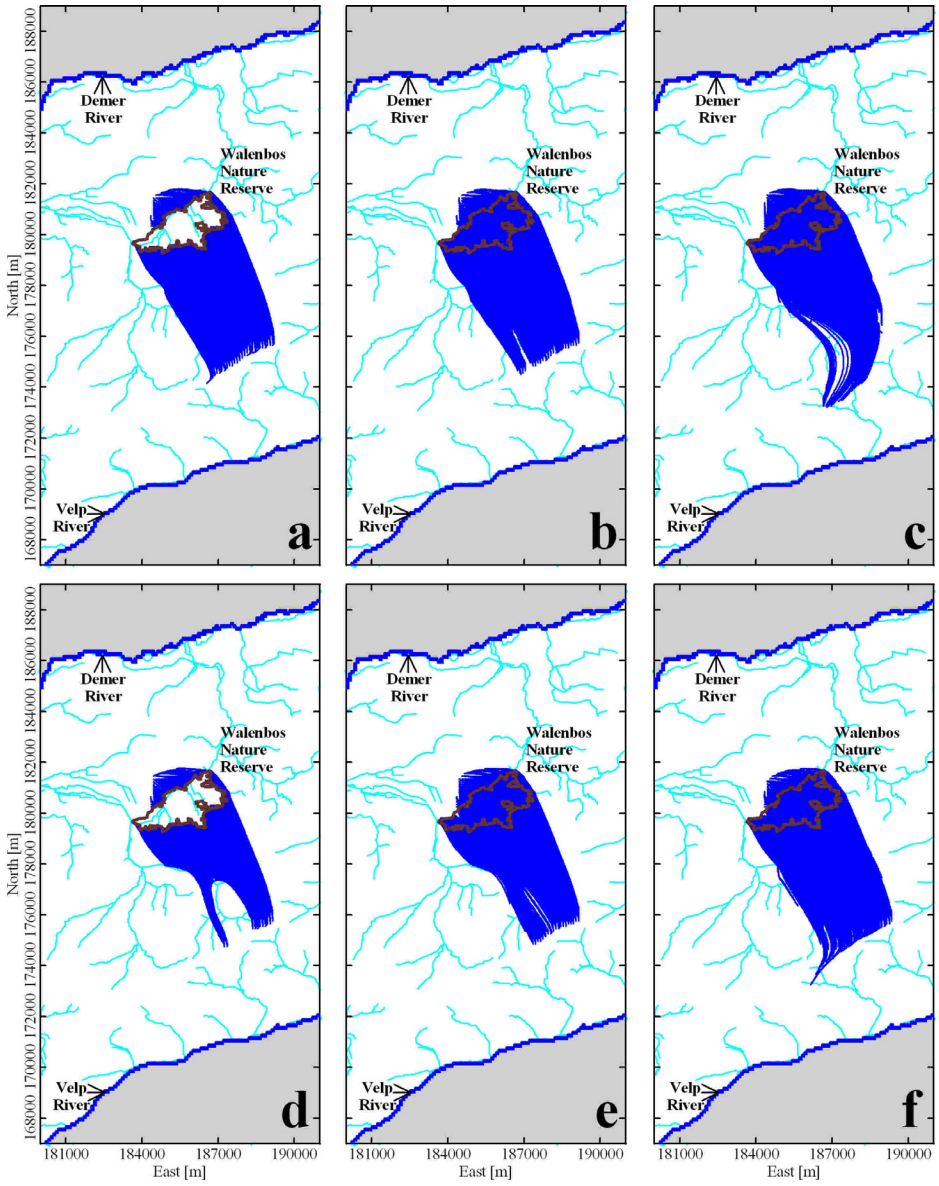


Figure 13: Forward particle tracking defining the capture zone for steady-state (calibrated) results obtained from UCODE-2005 (first row) and highest likelihood point in GLUE-BMA (second row) for models M1 (a and d), M2 (b and e) and M3 (c and f)