LA-UR- 08-7137

Approved for public release; distribution is unlimited.

Title: Combining Multi-objective Optimization and Bayesian model Averaging to Calibrate Forecast Ensembles of Soil Hydraulic Models

Author(s): Thomas Wöhling
Jasper A. Vrugt

Submitted to: Water Resources Research

Combining Multi-Objective Optimization and Bayesian Model Averaging to Calibrate Forecast Ensembles of Soil Hydraulic Models

Thomas Wöhling*and Jasper A. Vrugt†

*Corresponding author. Lincoln Environmental Research, Lincoln Ventures Ltd., Ruakura Research Centre, Hamilton, New Zealand. Email: woehling@lvlham.lincoln.ac.nz

†Center for Nonlinear Studies (CNLS), Mail Stop B258, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Abstract

Most studies in vadose zone hydrology use a single conceptual model for predictive inference and analysis. Focusing on the outcome of a single model is prone to statistical bias and underestimation of uncertainty. In this study, we combine multi-objective optimization and Bayesian Model Averaging (BMA) to generate forecast ensembles of soil hydraulic models. To illustrate our method, we use observed tensiometric pressure head data at three different depths in a layered vadose zone of volcanic origin in New Zealand. A set of seven different soil hydraulic models is calibrated using a multi-objective formulation with three different objective functions that each measure the mismatch between observed and predicted soil water pressure head at one specific depth. The Pareto solution space corresponding to these three objectives is estimated with AMALGAM, and used to generate four different model ensembles. These ensembles are post-processed with BMA and used for predictive analysis and uncertainty estimation. Our most important conclusions for the vadose zone under consideration are: (1) the mean BMA forecast exhibits similar predictive capabilities as the best individual performing soil hydraulic model, (2) the size of the BMA uncertainty ranges increase with increasing depth and dryness in the soil profile, (3) the best performing ensemble corresponds to the compromise (or balanced) solution of the three-objective Pareto surface, and (4) the combined multi-objective optimization and BMA framework proposed in this paper is very useful to generate forecast ensembles of soil hydraulic models.

Keywords: Bayesian model averaging, vadose zone modeling, soil hydraulic models, inverse parameter estimation, multi-objective optimization

1 Introduction

Faced with the complexity, spatial and temporal variability of processes occurring in natural systems, and the difficulty of performing controlled experiments, a variety of numerical simulation models have been developed to predict the behavior of environmental systems. Even the most elaborate model, however, cannot reflect the true complexity and heterogeneity of the processes occurring in the field. To some degree it must always conceptualize and aggregate complex interactions driven by a number of spatially distributed and highly interrelated energy, mass transport, and biogeochemical processes by the use of only relatively simple mathematical equations. There is significant uncertainty associated with the correct formulation of these processes underlying the system of interest (Beven and Binley, 1992; Gupta et al., 1998; Kuczera et al., 2006; Vrugt and Robinson, 2007a). Quantification of this uncertainty is necessary to better understand what is well and what is not very well understood about the processes and systems that are being studied.

Single deterministic soil-hydraulic models are often used for studying flow and transport through the vadose zone (e.g.Mertens et al. 2005; Guber et al. 2006; Sansoulet et al. 2008, and others). In particular, the Mualem-van Genuchten (van Genuchten, 1980) (MVG) model has become the standard choice for analyzing unsaturated porous media. This model is relatively simple to use, and many contributions to the hydrologic literature have shown that it works well for a range of problems and soil types. Moreover, direct (laboratory procedures) and indirect approaches (pedotransfer functions) are widely available to obtain estimates of the MVG parameters for the specific site under consideration. Notwithstanding this progress made, the use of a single model for predictive inference and analysis is prone to statistical bias (Hoeting et al., 1999; Neuman, 2003; Raftery et al., 2003, 2005) as it implicitly rejects alternative and other plausible soil hydraulic models for the vadose zone under consideration. Arguably, there is significant advantage to using multiple different models simultaneously for predictive analysis and inference as their individual ability to fit the experimental data will infer important information about the key hydrological processes affecting flow (and transport) through the unsaturated zone of interest.

Ensemble Bayesian model averaging has recently been proposed as a methodology to explicitly

handle conceptual model uncertainty in the interpretation and analysis of environmental systems. This method combines the predictive capabilities of multiple different models and jointly assesses their uncertainty. The probability density function (pdf) of the quantity of interest predicted by Bayesian model averaging is essentially a weighted average of individual pdf's predicted by a set of different models that are centered around their forecasts (Raftery et al., 2005; Vrugt et al., 2006a). The weights assigned to each of the models reflect their contribution to the forecast skill over the training period. Typically, the ensemble mean outperforms all or most of the individual members of the ensemble (Raftery et al., 2005). Bayesian model averaging has been successfully applied to forecasting of surface temperature (Raftery et al., 2005), surface temperature and sea level pressure (Vrugt et al., 2006a), streamflow (e.g. Kuczera et al. 2006; Vrugt and Robinson 2007b; Ajami et al. 2007), and permeability structures in groundwater hydrology (Neuman, 2003; Ye et al., 2004). Guber et al. (2006) have used ensembles of pedotransfer functions to simulate water content time series. Recently, Ye et al. (2008) have provided a comprehensive test of model selection criteria in multi-model analysis.

In this study, we combine the strengths of multi-objective optimization and Bayesian model averaging (BMA) to better quantify predictive uncertainty of models of flow through unsaturated porous media. In our analysis, we consider seven different soil hydraulic models including water retention and unsaturated hydraulic conductivity function formulations based on uniform flow, hysteresis and dual-porosity. In the first step, each of these models is calibrated by posing the parameter estimation problem into a multi-objective framework. The resulting optimization problem is solved by means of the AMALGAM evolutionary search algorithm (Vrugt and Robinson, 2007a). Then, the Pareto trade-off surface of each of these models is analyzed and combined to generate different forecast ensembles. In the second step, these different ensembles are post-processed with BMA to analyze and quantify predictive uncertainty. We illustrate our approach using observations of tensiometric pressure head at three different depths in a layered vadose zone at a field site in New Zealand.

2 Materials and Methods

2.1 Bayesian Model Averaging

In a previous study (Wöhling et al., 2008) we analyzed the ability of the MVG soil hydraulic model (1980) to reproduce field data of tensiometric pressure with parameter sets estimated by multiobjective optimization. The use of a single model for predictive inference and analysis implicitly rejects other possible plausible conceptual models of the system under study, and therefore may underestimate uncertainty (Raftery et al., 2003). Bayesian model averaging (BMA, Leamer, 1978; Kass and Raftery, 1995; Hoeting et al., 1999) provides a way to combine inferences and predictions of several different conceptual models and to jointly assess their predictive uncertainty. If an ensemble of k different statistical models $M = \{M_1, M_2, ..., M_k\}$ is considered and the quantity of interest is Δ, then its posterior distribution given the observation data y is (Hoeting et al., 1999):

$$p(\Delta \mid y) = \sum_{i=1}^{k} p(\Delta \mid M_i, y)\, p(M_i \mid y) \tag{1}$$

where $p(\Delta \mid M_i, y)$ is the forecast pdf based on the model M_i alone, and $p(M_i \mid y)$ is the posterior probability of model M_i under the assumption that it is correct for the training data and reflects how well model M_i fits the data (Raftery et al., 2003). All probabilities are implicitly conditional on the set of models under consideration M. The posterior model probabilities are positive and add up to one and can thus be viewed as weights, reflecting the models relative contributions to predictive skill over the training period. Thus, Eq. (1) is a weighted average of the posterior distributions under each of the k models, weighted by their posterior model probabilities.

Raftery et al. (2005) recently extended BMA to ensembles of dynamical models and demonstrated how it can be used to postprocess forecast ensembles from dynamic weather models. To explicate the BMA method developed by Raftery et al. (2005), each ensemble member forecast f_i is associated with a conditional pdf, $g(\Delta \mid f_i)$, which can be interpreted as the pdf of Δ given

f_i. From Eq. (1), the BMA predictive model can be expressed as

$$p(\Delta \mid f_i, \ldots, f_k) = \sum_{i=1}^{k} w_i \, g_i(\Delta \mid f_i) \tag{2}$$

where w_i denotes the posterior probability of forecast i being the best one.

The original ensemble BMA method described in Raftery et al. (2005) assumes that the conditional pdf's $g_i(\Delta \mid f_i)$ of the different ensemble members can be approximated by a normal distribution centered at a linear function of the original forecast, with mean $a_i + b_i f_i$ and standard variation σ_i:

$$\Delta \mid f_i \sim N\left(a_i + b_i f_i, \sigma_i^2\right) \tag{3}$$

The values for a_i and b_i are bias-correction terms that are derived by simple linear regression of Δ on f_i for each of the individual ensemble members.

BMA Predictive Mean and Variance

The BMA predictive mean is the conditional expectation of Δ given the forecasts:

$$E(\Delta \mid f_i, \ldots, f_k) = \sum_{i=1}^{k} w_i \, (a_i + b_i f_i) \tag{4}$$

and the associated variance can be computed as (Raftery et al., 2005; Vrugt and Robinson, 2007b)

$$Var(\Delta_{st} \mid f_{i,st}, \ldots, f_{kst}) = \sum_{i=1}^{k} w_i \left[(a_i + b_i f_{ist}) - \sum_{j=1}^{k} w_j \, (a_j + b_j f_{jst}) \right]^2 + \sum_{i=1}^{k} w_i \sigma_i^2 \tag{5}$$

where f_{ist} denotes the i-th forecast in the ensemble for location s and time t. We assume a normal predictive distribution in our proposed BMA approach. Although a normal distribution seems to be inappropriate for any quantity primarily driven by precipitation, Vrugt and Robinson (2007a) showed that this assumption works well for streamflow simulation and forecasting. Using different statistical distributions to describe $g_i(\Delta \mid f_i)$ for the individual models of the ensemble resulted in very similar conclusions as those presented here for the normal conditional pdfs. We therefore do not discuss these results here.

Posterior Forecast Probability

The estimation of the posterior probability of the individual forecasts or weights, w_i, are required for the implementation of BMA. Raftery et al. (2005) estimated w_i, $i = 1, \ldots, k$; σ^2 by a maximum likelihood approach. Assuming independence of forecast errors in space and time, the log-likelihood function corresponding to the predictive model Eq. (2) can be written as

$$\ell\left(w_i, \ldots, w_k, \sigma^2\right) = \sum_{s,t} \log\left[\sum_{i=1}^{k} w_i g_i\left(\Delta_{st} \mid f_{ist}\right)\right] \tag{6}$$

where the summation is over s and t to include all observations in the training set. Eq. (6) must be maximized to obtain the BMA weights and variances. In this study we follow the approach of Vrugt and Robinson (2007a), who used the Shuffled Complex Evolution Metropolis algorithm (SCEM-UA) algorithm for the maximization of Eq. (6). The SCEM-UA algorithm is a general purpose optimization algorithm that uses adaptive Markov Chain Monte Carlo (MCMC) sampling (Vrugt et al., 2003b) to estimate the traditional best parameter combination and its underlying posterior probability density function within a single optimization run. The method uses a predefined number of different Markov Chains to independently explore the search space. These chains communicate with each other through an external population of points, which are used to continuously update the size and shape of the proposal distribution in each chain. The MCMC evolution is repeated until the R-statistic of Gelman and Rubin (1992) indicates convergence to a stationary posterior distribution. More information about the SCEM-UA algorithm can be found in Vrugt et al. (2003b) and so will not be repeated here.

2.2 Bayesian Model Averaging of Soil Hydraulic Models

Field Data

We used field data from the *Spydia* experimental site in the northern Lake Taupo catchment, New Zealand. The vadose zone materials at *Spydia* encompass a young volcanic soil (0 - 1.6 m depth), unwelded Taupo Ignimbrite (TI, 1.6 - 4.4 m), and two older buried soils (Palaeosols, 4.4 to 5.8 m depth). Tensiometric pressure head was measured in the vadose zone at 15 min intervals using Tensiometer probes (type UMS T4e, Germany, accuracy ± 0.5 kPa) installed at

five different depths (0.4, 1.0, 2.6, 4.2, and 5.1 m) and three locations per depth. The pressure head measurements at each depth were averaged before they were used in our calculations. Daily values of potential evaporation were calculated by the Penman-Monteith equation (Allen et al., 1998) using data from the nearby Waihora meteorological station (500 m distance). Precipitation was recorded on site using a 0.2 mm bucket gauge and upscaled to hourly values for use in our calculations. A detailed description of the *Spydia* experimental data can be found in Wöhling et al. (2008).

A period of 546 days (April 11, 2006 to October 9, 2007) was used for all our calculations. Since the available data comprises two wet (winter) seasons and only one dry (summer) season, the model was calibrated for the first winter/spring season (April 11, 2006 to January, 18, 2007) and evaluated with a representative 96 days time period of the second wet season (July 5, 2007 to October 9, 2007).

Models in the Study

We used the HYDRUS-1D model (Šimůnek et al., 2005) to simulate water flow in the *Spydia* vadose zone. HYDRUS-1D utilizes the Galerkin finite element method based on the mass conservative iterative scheme proposed by Celia et al. (1990). The model solves the one-dimensional Richards' equation:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z} K \left(\frac{\partial h}{\partial z} + 1 \right) - S \tag{7}$$

where θ is the volumetric water content [L^3L^{-3}], t represents time [T], z is the vertical coordinate (positive upward) [L], h denotes the pressure head [L], K is the unsaturated hydraulic conductivity function [LT^{-1}], and S is a sink term representing processes such as plant water uptake [$L^3L^{-3}T^{-1}$].

Soil hydraulic functions need to be specified to solve Eq. (7). The seven soil hydraulic models employed in our BMA approach encompass not only different formulations of the same physical relationships but also different conceptual models. The first four models listed below are based on the concept of uniform flow. This concept assumes the porous medium as a system of impermeable particles separated by pores through which water flow takes place. Hysteresis of the functional relationships might be considered to account for wetting-drying cycles. In

contrast, non-equilibrium flow models assume the particles to have their own micro-porosity. The fifth model in the list below assumes that water can move into and out of the micro-pore domain whereas time-constant (immobile) water content is assumed in the micro-pore domain by the sixth model. The different model concepts are expected to have a different ability to reproduce a given set of field data. Non-equilibrium models are more flexible than uniform flow models and typically perform better when macro-pores or preferential flow paths are present in the porous media under investigation.

The individual models used in our approach are:

1. The modified Mualem-van Genuchten model (non-hysteretic) (MVG, Vogel et al. 2001):

$$S_e = \frac{\theta - \theta_r}{\theta_s - \theta_r} = \begin{cases} \theta_r + \frac{1}{[1+|\alpha h|^n]^m} & h < h_s \\ 1 & h \geq h_s \end{cases} \tag{8}$$

$$K(S_e) = K_s S_e^l \left[1 - \left(1 - S_e^{1/m}\right)^m\right]^2 \tag{9}$$

where S_e is the effective water content, θ_r and θ_s denote the residual and saturated water content, respectively [L^3L^{-3}], α [L^{-1}] and n [-] are parameters that define the shape of the water retention function, K_s represents the saturated hydraulic conductivity [LT^{-1}], l is the pore-connectivity parameter by Mualem (1976), and $h_s = -0.02$ m is the assumed air entry value. In this study, we further assume that $m = 1 - 1/n$ and $n > 1$.

2. The modified Mualem-van Genuchten model utilizing hysteresis in the water retention function (MVG-HR) (Šimůnek et al., 2005). This model uses a separate drying and wetting curve of the retention function with the mathematical formulation provided by Eq. (8) using two sets of parameter values, (θ_r^d, θ_s^d, where the subscripts d and w indicated wetting and drying, respectively. Following Šimůnek et al. (2005), we assume that $\theta_r^d = \theta_r^w$ and $n^d = n^w$. The HYDRUS-1D implementation of the hysteretic MVG model also requires an estimate of the empirical parameter θ_m (Šimůnek et al., 2005). We estimate this value following Vogel et al. (2001) as

$$\theta_m = \theta_r + \left(\theta_s^d - \theta_r\right)\left[1 + \left(\left|\alpha^d h_s\right|\right)^n\right]^m. \tag{10}$$

Thus, the hysteretic model MVG-HR model requires three more parameters (θ_s^w, α^w, K_s^w) than the original MVG formulation.

3. The Brooks and Corey (1966) model (BC):

$$S_e = \begin{cases} |\alpha h|^n & h < -1/\alpha \\ 1 & h \geq -1/\alpha \end{cases} \tag{11}$$

$$K(h) = K_s S_e^{2/n+l+2} \tag{12}$$

4. The two-parameter log normal distribution model of Kosugi (1996) (KM):

$$S_e = \frac{\theta - \theta_r}{\theta_s - \theta_r} = \begin{cases} \frac{1}{2}\, ercf\left[\frac{ln(h/\alpha)}{\sqrt{2}n}\right] & h < 0 \\ 1 & h \geq 0 \end{cases} \tag{13}$$

$$K(S_e) = \begin{cases} K_s S_e^{1/2}\left\{\frac{1}{2}\, ercf\left[\frac{ln(h/\alpha)}{\sqrt{2}n} + n\right]\right\}^2 & h < 0 \\ K_s & h \geq 0 \end{cases} \tag{14}$$

where α and n are substitutes of the original Kosugi (1996) notation (Šimůnek et al., 2005).

5. The dual-porosity model of Durner (1994) (MVG-DP) which divides the porous medium into two overlapping regions using Mualem - van Genuchten type functions for each of the two regions:

$$S_e = w_1 \left[1 + (\alpha_1 |h|)^{n_1}\right]^{-m_1} + w_2 \left[1 + (\alpha_2 |h|)^{n_2}\right]^{-m_2} \tag{15}$$

$$K(S_e) = K_s \frac{\left(w_1 S_{e,1} + w_2 S_{e,2}\right)^l \left\{w_1 \alpha_1 \left[1 - \left(1 - S_{e,1}^{1/m_1}\right)^{m_1}\right] + w_2 \alpha_2 \left[1 - \left(1 - S_{e,2}^{1/m_2}\right)^{m_2}\right]\right\}^2}{\left(w_1 \alpha_1 + w_2 \alpha_2\right)^2} \tag{16}$$

where w_i are the weighting factors for the two sub-regions of the porous medium ($w_1 + w_2 = 1$) and α_i, n_i, m_i ($= 1 - 1/n_i$), and l are empirical parameters of the corresponding functions ($i = 1, 2$).

6. The two-region, dual porosity model by Šimůnek et al. (2003) which partitions the liquid phase into a mobile (inter-aggregate) and immobile (intra-aggregate) region. Water flow

in the mobile region $\partial\theta_{mo}/\partial t$ is described by the Richards' Eq. (7) with an additional term on the right hand side of the equation, $-\Gamma_w$, representing the transfer rate of water from the inter- to the intra-aggregate pores. The moisture dynamics in the matrix is described by a simple mass balance equation (Šimůnek et al., 2005)

$$\frac{\partial\theta_{imo}}{\partial t} = S_{imo} - \Gamma_w \tag{17}$$

where S_{imo} is the sink term for the immobile region. In this study, we assumed that flow of water between the mobile and immobile region can be described with a simple linear exchange equation (Šimůnek et al., 2005), with $\Gamma_w = \omega\,[S_{e,mo} - S_{e,imo}]$, where $S_{e,mo}$ and $S_{e,imo}$ denote the effective fluid saturations of the mobile and immobile regions, respectively. The residual $\theta_{r,imo}$ and saturated $\theta_{s,imo}$ water content of the immobile region are two additional parameters in this model formulation that need to be estimated against observations.

Table 1 summarizes the calibration parameters in each of the six soil hydraulic models. The initial and boundary conditions used to solve Eq. (7) are:

$$h(z,t) = h_i(z) \qquad \text{at } t = 0, \tag{18}$$

$$h(z,t) = h_L(t) \qquad \text{at } z = L, \tag{19}$$

and

$$\begin{aligned} -K\left(\tfrac{\partial h}{\partial z} + 1\right) &= q_0(t) - \tfrac{dh}{dt} && \text{at } z = 0, \text{ for } h_A \le h \le h_s \\ h(0,t) &= h_A && \text{for } h < h_A \\ h(0,t) &= h_s && \text{for } h > h_s \end{aligned} \tag{20}$$

where $h_i(z)$ is the initial pressure head derived from linear interpolation of observed tensiometric pressure at the 0.4, 1.0, 2.6, and 4.2 m depths, $h_L(t)$ is the prescribed (observed) pressure head at the bottom boundary L = - 4.2 m (depth of the model is 4.2 m), $q_0(t)$ is the net infiltration rate (i.e. precipitation minus evaporation) and h_A and h_s are the minimum and maximum pressure head allowed at the soil surface. Eq. (20) describes the atmospheric boundary condition at the soil-air interface (Šimůnek et al., 1996) which switches between a prescribed flux condition and

a prescribed head condition, depending on the prevailing transient pressure head conditions near the surface. The plant water uptake, S in Eq. (7), is simulated by the Feddes model (1978) using HYDRUS-1D default parameters for grass and a rooting depth of 0.35 m. Because our study considers a relatively coarse textured soil with high infiltration capacity, we neglect infiltration-excess overland flow and use the limits of $h_A = -200$ m and $h_s = -0.02$ m. The initial pressure heads measured at April 11, 2006 were -0.41, -1.38, -1.18 and -0.85 m at the 0.4, 1.0, 2.6, and 4.2 m depths, respectively.

The HYDRUS-1D model was set up for three horizons. The first being the more recent materials (0 - 0.69 m depths), and the other two simulation layers being the disturbed Taupo Ignimbrite (0.69 - 1.6 m) and the in-situ Taupo Ignimbrite (1.6 - 4.2 m), respectively. Additional to the six soil hydraulic models with three layers, we also included the MVG model with four horizons (MVG-4). This was done to investigate the effect of lumping different layers of stratification into larger numerical horizons, and was accomplished by dividing the upper layer (0.69 - 1.6 m) into two individual layers: the first being the Ap and Bs horizons (0 - 0.38 m) and the second the BC and C1 horizons (0.38 - 0.69 m).

In summary, the ensemble used in our BMA approach consists of the predictions of seven different soil hydraulic models, hereafter also referred to as ensemble members. For both the 3- and 4-layer stratifications, we used a computationally efficient uniform discretization scheme with $\Delta x = 0.02$ m in the vertical domain. This results in a total of 211 nodes in the HYDRUS-1D model.

Multi-objective Calibration of Soil Hydraulic Models

The hydraulic models used in this study require the estimation of different parameters to quantify the soil water retention and unsaturated soil hydraulic conductivity functions for the various layers throughout the soil profile. For each model, these parameters are estimated using inverse modeling by minimizing the difference between observed and modeled tensiometric pressure head at three different observation depths. Similar to our previous work (Wöhling

et al., 2008), we use a multi-objective formulation with three different criteria:

$$\min F(u) = \begin{bmatrix} F_1(u) \\ F_2(u) \\ F_3(u) \end{bmatrix} \tag{21}$$

where F_1- F_3 are defined as the root-mean square error (RMSE, e.g. Hall 2001) of the fit between the simulated and observed pressure heads at the 0.4, 1.0, and 2.6 m depths in the vadose zone profile, and u is a vector of np model parameters to be optimized (Table 1).

The inverse problem expressed in Eq. (21) is solved with the AMALGAM evolutionary search algorithm (Vrugt and Robinson, 2007a). Among several different state-of-the-art multi-objective optimization algorithms, this method was shown to be the most efficient for the problem considered herein (Wöhling et al., 2008). The AMALGAM algorithm combines simultaneous multi-method search and self-adaptive offspring creation to ensure a reliable and computationally efficient solution to multiobjective optimization problems. The only algorithmic parameter to be defined by the user is the population size, s. In all the calculations reported here, we used a value of $s = 100$. To create the initial sample to be iteratively improved with AMALGAM, we used uniform sampling within the parameter bounds specified as follows: θ_s,θ_w,$\theta_{s,mo}$,$\theta_{s,imo}$ = 0.3 - 0.7 $[m^3\,m^{-3}]$; α, α_1, α_2 = 1 - 20 $[m^{-1}]$; n, n_1, n_2 = 1.1 - 9.0 [-]; K_s, K_{sw} = 10-7 - 10-3 $[m\,s^{-1}]$; l = 0.1 - 1.0; ω, w_2 = 0 - 1 [-]. To reduce the number of parameters to be optimized, θ_r and $\theta_{r,imo}$ in the MVG and DPIM model, respectively, were set to zero. This assumption is not going to influence the analysis, as there is very little sensitivity to these two parameters within the range of pressure heads spanned by the calibration data used in this study. The individual optimization runs were set up for the 282 days calibration period reported above and a 10 days initialization period was considered for the calculation of the performance measures. The runs were terminated after 50,000 HYDRUS-1D model evaluations. A detailed description of AMALGAM has been presented in Vrugt and Robinson (2007a) and is therefore not repeated here.

Our multi-objective formulation will result in a set of Pareto optimal solutions that represent trade-offs among the three different objectives. These solutions have the property that moving from one to another along the trade-off surface results in the improvement of one objective

while causing deterioration in at least one other objective (Gupta et al., 1998; Deb, 2001; Vrugt et al., 2003a). Consistent with our earlier approach (Wöhling et al., 2008), we isolate four different parameter combinations from the Pareto surface that we believe are most informative and useful for postprocessing with the BMA method. The first three Pareto points, $P_1 - P_3$, are the best solutions with respect to each of the individual objectives (subsequently referred to as Pareto extremes). The fourth solution, P_4, is the balanced solution, i.e. where the overall RMSE of all three objectives is at its minimum. This point is hereafter also referred to as the compromise solution.

Two additional criteria are used to measure the fit between observed and simulated tensiometric data of the selected Pareto solutions: the coefficient of determination R^2, and the coefficient of efficiency, C_e, by Nash-Sutcliffe (ASCE, 1993). C_e is a widely used fitting criterion and may assume a negative value if the mean square error of the best prediction exceeds the variance of the observations (Hall, 2001). Model predictions are considered satisfactorily if the values of R^2 and C_e are close to unity. We further used the coefficient of correlation r_c to calculate the correlation between the forecasts of the individual models.

Model Ensembles used in the BMA Approach

Pressure heads were forecasted at the three different depths during the evaluation period using each individual soil hydraulic model. To simplify the analysis, these pressure heads were interpolated to hourly values. We use four different model ensembles in our BMA approach. Each ensemble consists of seven members, namely the MVG, MVG-HR, BC, KM, MVG-DP, DPIM, and MVG-4 soil hydraulic models. The hydraulic head predictions in these four ensembles are different because different Pareto solutions were used to create the forecasts. Ensemble 1 contains the predictions of the seven models using the best solution to objective F_1 of the Pareto surface, P_1(MVG), P_1(MVG-HR), P_1(BC), P_1(KM), P_1(MVG-DP), P_1(DPIM), and P_1(MVG-4). Similarly, ensemble 2 and 3 were generated using the Pareto extremes P_2 and P_3 for each individual model, whereas ensemble 4 contains the predictions of the individual soil hydraulic models for the compromise Pareto solutions, P_4.

We follow five different steps in our analysis. A flowchart summarizing our approach and analysis is presented in Figure 1with the individual steps numbered in circles.

First (number 1 in Fig. 1), we investigate the performance of the BMA approach for the four different model ensembles using combined performance measures for all three depths, i.e. the RMSE, R^2, and C_e fitting criteria are calculated for the 0.4, 1.0, and 2.6 m depths. These criteria are hereafter referred to as RMSE_c (subscript c stands for combined), R_c^2, and $C_{e,c}$. This investigation is used to explore how the performance of the BMA model depends on the parametrization of the individual models. In the second step (2 in Fig. 1), we investigate the BMA performance at the individual depths. The best attainable pressure head forecasts at the 0.4, 1.0, and 2.6 m depths are post-processed with BMA using the parametrization of the corresponding Pareto extremes. In the remainder of this paper, these forecasts are subsequently referred to as single depth forecasts. Then we compare the results of the single depth and the combined depth forecasts. In the third step (3 in Fig. 1), we analyze the sensitivity of the BMA results to the use of a single or multiple different variances of the conditional pdf's of the individual ensemble members. In the fourth step (4 in Fig. 1), we explore the importance of the individual soil hydraulic models in the ensemble by comparing the BMA results of ensemble 1, 2 and 3 with results of a reduced ensemble size created by only selecting a sub-set of models. For instance, in ensemble 1a, 2a and 3a, we exclude the model that receives the highest BMA weight amongst all the models. Similarly, ensemble 1b, 2b and 3b are formed by excluding the two best performing soil hydraulic models. Hence, ensemble 1a, 2a and 3a contain six individual time series of model predictions, whereas ensemble 1b, 2b and 3b consist of only five members. Finally, in the last step (5 in Fig. 1) we discuss how the choice of the calibration period affects the BMA results.

3 Results and Discussion

3.1 Optimized Parameter Sets

The MVG, MVG-HR, BC, KM, MVG-DP, DPIM, and MVG-4 soil hydraulic models were calibrated with AMALGAM using the observed tensiometric pressure heads at the 0.4, 1.0, and 2.6 m depths in the *Spydia* vadose zone during the calibration period. One of the drawbacks of inverse modeling for vadose zone problems is the computational requirements. For instance, a single optimization run using 50,000 HYDRUS-1D model evaluations required several days

using the Matlab® R2007a (64 bit) - Microsoft Windows^{TM} XP Professional (64 bit) modeling environment on a Dell^{TM} Precision 390 workstation with a Quad-Core Intel® Core^{TM}2 Extreme processor QX6700 (2.67 GHz) and 2GB of RAM. Note, however that AMALGAM often converged to the Pareto solution set within less than the maximum total of 50,000 HYDRUS-1D model evaluations. Between 16,000 and 27,000 model evaluations were required for the MVG-HR, KM, BC, and MVG-DP models to approximate the three-objective Pareto solution set, whereas about 44,000 HYDRUS-1D model evaluations were needed with AMALGAM to converge for MVG-4. Use of multiple different computational nodes within a distributed computing environment would significantly reduce the time needed for calibration (Vrugt et al., 2006b).

To illustrate the outcome of a typical AMALGAM optimization run, consider Figure 2 that presents the F_1-F_2, F_1-F_3, and F_2-F_3 bi-criterion trade-off fronts of the full three-dimensional Pareto surface for the MVG-model. In each panel, the rank 1 solutions are indicated with gray circles. Note, that AMALGAM has sampled the Pareto surface quite densely and uniformly with emphasis on sampling the front of objectives F_1 and F_2. This front shows considerable trade-off demonstrating an inability of the MVG model to simultaneously provide good fits to the pressure head observations at the 0.4 and 1.0 m depths. More discussion on this can be found in Wöhling et al. (2008). In contrast, the $F_1 - F_3$ and $F_2 - F_3$ fronts exhibit a more rectangular trade-off pattern (Figures 2b, 2c), illustrating that it is possible to minimize both of these objectives simultaneously using only a single combination of values of the hydraulic parameters in the MVG model. As described above, the Pareto solutions that are separately indicated in each panel with $P1$, $P2$, $P3$ and $P4$ are used for further analysis with BMA.

The Pareto optimal parameter solutions derived with AMALGAM for the MVG-HR, BC, KM, MVG-DP, DPIM, and MVG-4 models were analyzed in the same way as described here for the MVG model. Because of space limitations, however, these results are not reported here. This information can be obtained from the corresponding author upon request.

Tables 2 and 3 summarize the performance of the seven hydraulic models with the various Pareto solutions derived with AMALGAM. Table 2 lists the RMSE, R^2, and C_e performance criteria for the simulations with the Pareto extremes ($P_1 - P_3$), i.e. with the parametrization yielding the best attainable fit to each of the three objectives $F_1 - F_3$. Note that we only report

the fit for the optimized (single) objective solutions. The RMSE of the fit between observed and simulated pressure heads at the three depth varies between 0.052 and 0.117 m. This average error can be considered quite small with magnitude similar to the average standard deviation of the pressure head measurements of the three different tensiometers at each depth. This shows the ability of AMALGAM to find good calibrated parameter values for each of the seven soil hydraulic models. However, even these relatively small RMSE values are based on simulations that include parts with significant deviations between observed and modeled tensiometric pressure heads. This will be shown and discussed later. Our results further show (not in the Table) that a good fit of a particular model at one depth is typically accompanied with a significantly worse fit at the other two measurement depths.

At the 0.4 m depth, the MVG-DP model has the best predictive capability with summary statistics of RMSE = 0.069 m, R^2 = 0.95 and C_e = 0.95, respectively. Seemingly, the tensiometers in the top part of the soil are affected by the presence of preferential flow paths that quickly move water to deeper layers in the profile. A dual-porosity model such as the MVG-DP model is expected to better account for this rapid movement of water than a single porosity (MVG) model. The best fit at the 1.0 m depth (measured with objective F_2) was obtained by the MVG-4 model with the lowest RMSE value in the analysis (RMSE = 0.052 m, cf. Table 2). This shows that a 4-layer numerical stratification better represents the actual soil profile at the *Spydia* field site. This is in agreement with soil textural information, as highlighted before. It should be noted, however, that all soil hydraulic models, but the BC model attained a very good fit to the observations at the 1.0 m depth (Table 2). The pressure head at the 2.6 m depth is responding least dynamically to rainfall at the surface. The dual porosity (MVG-DP) and the MVG-HR (which considers hysteresis) soil hydraulic models are the best predictors at this depth with associated RMSE values of 0.058 / 0.059 m and high coefficients of determination and efficiency. The 4-layer stratification used in MVG-4 performs similarly well (RMSE = 0.072m) which indicates that a better representation of the horizontal layering in HYDRUS-1D also improves the pressure head predictions at the 2.6 m depth.

Table 3 lists the RMSE$_c$, R_c^2, $C_{e,c}$, and Bias summary statistics for the fit between the observed and predicted pressure head for the combined 0.4, 1.0, and 2.6 m depths during the calibration period. At this time, please only consider the fitting criteria of the individual model forecasts in

ensemble 4. These forecasts correspond to the compromise solutions (P_4) of the Pareto surface. The $RMSE_c$ ranges from 0.101 to 0.153 m for the individual models. As expected, these values are somewhat higher than the single depth RMSE values previously reported in Table 2. Overall, MVG-4 is the best performing model with the best summary statistics. As shown previously, the 4-layer stratification used in the MVG-4 model results in the most accurate predictions of the pressure heads throughout the *Spydia* vadose zone. The overall performance of this model is even better than the dual-porosity MVG-DP model that explicitly accounts for the presence of preferential flow and therefore provides the closest fit to measured hydraulic heads at the 0.4m depth in the top part of the soil.

To illustrate the probabilistic properties of the soil hydraulic models, consider Figures 3b-d which depict the observed pressure head (bold solid lines) and the individual model predictions (thin lines) at the 0.4, 1.0, and 2.6 m depths during the calibration period. The calibrated model predictions correspond to the compromise solution (P_4) of the three objective $\{F_1, F_2, F_3\}$ Pareto surface. The response to rainfall at the surface (Figure 3a) is most dynamic at the 0.4 m depth (Figure 3b) whereas the response is more damped deeper down in the profile. (Figure 3c-d). This is due to the time required for the water to move through the soil and reach the deeper layers. It is interesting to observe that the spread of the individual model predictions is quite narrow at the 1.0 m depth and wider at the 0.4 and 2.6 m depths. The calibrated models appear to provide quite different predictions which generally bracket the observations. This is a desirable characteristic for accurate ensemble forecasting with the BMA method.

3.2 BMA

In this section, we first illustrate how BMA works by showing the prediction of tensiometric pressure head at one location and time (results presented in Table 4 and Figure 4). Then we describe the aggregated results for the entire evaluation period and the three depths.

Illustration of the BMA Approach

Consider the pressure head forecast at the $s = 0.4$ m depth on October 5, 2007, 5:00 pm ($t = 542.3$ days). Table 4 shows the individual model forecasts, the bias corrected forecasts,

the BMA weights, the BMA variances, and the verifying observation. Figure 4 depicts the BMA predictive pdf (solid line) which is the weighted sum of the seven normal pdf's of the individual ensemble members (dashed, dashed-dotted, dotted lines, and combinations thereof). The BMA pdf distribution is bimodal, indicating that there are disagreeing forecasts. In our example, the MVG-DP model predicted a somewhat larger pressure head as compared to the other models. The verifying observation is captured within the ensemble range as is the case in 82.8% of the times for the evaluation data set for the 0.4 m depth. The distinct shape of the BMA pdf in Figure 4 is defined predominantly by the conditional pdf's of the MVG-DP model (dashed-dotted line) which has received the largest BMA weight (66.7%), the KM model (dotted line) with the second largest BMA weight (17.9%), and by the MVG-4 model (large dashed line). It is interesting that the MVG-4 model has a considerable influence on the shape of the BMA pdf despite its relatively low weight of 3.8%. This is caused by its relatively small standard deviation (Table 4) causing a significant spike in the BMA predictive pdf. Note that the MVG and BC models exhibit very small BMA weights and a large variance (cf. Table 4). These models' contribution to the overall BMA predictive pdf is therefore negligible. The relatively large 95% uncertainty ranges (shaded area in Figure 4) are a result of the spread or disagreement of the individual model forecasts (Table 4).

The ensemble members' importance over the training period is reflected by the optimized BMA weights, w_i for the individual soil hydraulic models. It seems reasonable to assume that the rank order of the weights should be similar to the reverse order of the RMSEs, based on the assumption that better models should get a higher weight in the BMA model. Indeed in our example, the MVG-DP forecast has the highest BMA weight (Table 4) and the lowest RMSE (Table 2). This is not an unexpected result since macropore flow paths are typically present close to the surface and in the active root zone. Therefore the dual-porosity flow model MVG-DP can be expected to perform better than a uniform flow model. However, the uniform flow model KM ranks second in BMA weights, but sixth in reversed RMSE order. This seems counter-intuitive, but is explained by the presence of significant correlation among the predictions of the individual model forecasts. The (linear) correlation coefficients among the models, r_c, vary between 0.95 and 0.98, demonstrating a strong similarity in tensiometric head predictions between the different soil hydraulic models. Such a strong correlation among

model predictions makes the information from some members of the ensemble fairly redundant. This results in optimized BMA weights that for at least a few soil hydraulic models might appear counter-intuitive at a first glance. This phenomena was also observed in other BMA studies (for example for streamflow forecasting). For similar cases as reported here, we would suggest that the BMA approach tends to assign a relative high weight to the best of the highly correlated models and explores additional information content of other, less correlated models. Another good example of this is the MVG-4 model which has the second best (lowest) RMSE, but ranks only fifth in its BMA weight. The additional information provided by this model is small because its forecasts are highly correlated with the MVG-DP model that has received the highest weight among the ensemble members. One should therefore be careful in drawing conclusions about the usefulness of individual models based on their optimized BMA weight. Note, that the DPIM and MVG-HR models rank third and fourth in both BMA weight and reversed RMSE order.

In the following, we provide the aggregated results for the BMA method for the three depths of pressure head observations at the *Spydia* site.

Combined depth Pressure Head Forecasts with individual Pareto Solutions

The aggregated results for the combined depth pressure head forecasts are summarized in Table 5 and Figure 5 and will be discussed here. The computational requirements of BMA are relatively small. In our analysis a typical BMA run would take not more than a couple of minutes using the computer architecture described above, which is less time than required for an average forward simulation of the HYDRUS-1D model. Table 5 presents summary statistics of the combined pressure head forecasts using the four different model ensembles for the evaluation period. The results are presented for the individual ensemble members (models) and the BMA predictive model (BMA pdf mean).

The performance of the soil hydraulic models during the evaluation period is best for the compromise solutions (ensemble 4) which is confirmed by relatively high R_c^2 and $C_{e,c}$ values up to 0.89 and 0.84, respectively (Table 5). On the other hand, the R_c^2 and $C_{e,c}$ values did not exceed 0.72 and 0.57 for ensembles 1 - 3 that contain the hydraulic head predictions of the respective single objective solutions of the Pareto surface. In addition, the $C_{e,c}$ values of

the individual soil hydraulic models in ensembles 1 to 3 often attain negative values, which is indicative of a rather poor fit to the observed pressure head data.

The best overall performance is obtained for ensemble 4, consisting of the model predictions of the compromise solutions of the Pareto surface. The performance of the individual soil hydraulic models in this ensemble, and the associated BMA model is significantly better than the performance observed for ensemble 1, 2 and 3. This is confirmed by significantly better values of the RMSE_c, Bias, R_c^2 and $C_{e,c}$ performance scores for the individual members of ensemble 4 for both the calibration and evaluation period. This is an expected result since the compromise solutions (as aggregated in ensemble 4) should provide a better fit to the combined data than the Pareto extremes (represented in ensembles 1 - 3). We observe that the BMA mean is better than any of the individual models of the ensembles for both the calibration and evaluation period (cf. Tables 3 and 5). A notable exception is the performance of the BMA mean in ensemble 4 during the evaluation period. The BMA predictive pdf does not exhibit the appropriate coverage during the evaluation period. Only 83% of the observations are captured at the 95% uncertainty interval. In addition, the RMSE_c value is significantly higher for the evaluation period as compared to the calibration period (Table 5). This indicates that the forecast spread of 0.40 (m) of the BMA model is too small during the evaluation period, and needs reconsideration. We anticipate that statistically more reliable results can be obtained when the BMA model is fitted using a calibration data that spans a larger variety of rainfall and drying events.

It is interesting to note that the optimized values of the BMA weights of the individual models differ quite substantially between the four different ensembles. A strong preference to the MVG-4 model ($w = 0.753$) is given in ensemble 1 with the remaining weight primarily allocated to the MVG-HR model (Table 5). Note that the models in ensemble 1 represent the calibration with regard to the 0.4 m depth and that the dual-porosity model MVG-DP performed best for this depth (Table 2). However, the four layer equilibrium flow model MVG-4 provides a better representation of the data in the deeper layers (as measured in the 1.0 m and 2.6 m depths) of the vadose zone profile as compared to the MVG-DP model. Similarly, the models performing best at the 1.0 and 2.6 m depths do not obtain the highest weights in ensembles 2 and 3. In ensemble 2, a high preference ($w = 0.802$) is given to the MVG-DP model, whereas the other

models receive very small weights. The MVG-HR and MVG-4 models receive the highest BMA weights in ensemble 3 with values of $w = 0.464$ and $w = 0.455$, respectively. In ensemble 4 most of the individual soil hydraulic model receive appreciable BMA weights. This shows that each member contributes to the overall BMA model, suggesting independent and additional information of each individual soil hydraulic model. This is desirable and explains why this ensemble is best suited for prediction of pressure heads throughout the soil profile.

Figures 5b-d illustrate the observed pressure head (bold solid lines) and the model forecasts of ensemble 4 (thin lines) at the 0.4, 1.0, and 2.6 m depths during the evaluation period. Also shown are the BMA mean (bold dashed lines) and the associated 95% prediction uncertainty bounds (shaded area). The spread of the pressure head predictions of the different soil hydraulic models increases with increasing depth and dryness in the soil profile. This is caused by error propagation and the functional shape of the water retention curve that predicts large variations in hydraulic head at lower water content values. The BMA prediction uncertainty intervals generally capture the observed pressure heads and are especially tight at the 0.4 and 1.0 m depths. Unfortunately, virtually all observations fall outside the 95% prediction uncertainty bounds at the 2.6 m depth during the period between days 450 - 490. During this time interval, the compromise Pareto solution is simply unable to provide pressure heads predictions that are consistent with the observations. A significant bias toward higher pressure head values is observed. A better result could have been obtained if we would have explicitly included this event in our calibration procedure with AMALGAM. This will be discussed in the final section of this paper.

Single depth Pressure Head Forecasts with individual Pareto Solutions

The aggregated results for the single depth pressure head forecasts are summarized in Table 6 and Figure 6 and will be discussed here. Table 6 presents summary statistics of the single depth pressure head forecasts of ensembles 1, 2 and 3 over the evaluation period for the individual models and the BMA mean. We like to reiterate that the single depth solutions correspond to a specific depth and that these solutions represent the best attainable fit to the tensiometric head data as determined with AMALGAM. Therefore, the RMSE values of the models in ensemble 1, 2 and 3 (Table 6) should generally be lower than their corresponding RMSE_c counterparts

derived for the combined depth forecasts (cf. Table 5). The only exception to this is the BC model in ensemble 3. This is probably caused by the fact that the BC model attains a much better fit at the 0.4 and 1.0 m depths (represented by ensembles 1 and 2) as compared to the 2.6 m depth (ensemble 3) and that the combined depth forecasts represent an average fit to all three depths.

The best fit to the observed pressure head data is found at the 1.0 m depth where RMSE values of the individual models and the BMA mean range between 0.065 and 0.115 m. While the BMA model outperformed all of the individual ensemble members during the calibration period, this is not necessarily the case during the evaluation period. This can be seen by comparison of the performance statistics of BMA against the summary statistics of the individual soil hydraulic models. As discussed for the combined depth forecasts, this might have to do with the selection of the calibration period.

The performance of the various soil hydraulic models in ensemble 1 and 2 during the evaluation period is very similar to their performance during the calibration period. This is confirmed by relatively small differences in $RMSE_c$ values ranging between -0.036 and 0.009 m (cf. Tables 2 and 6). In contrast, larger $RMSE_c$ differences of up to -0.265 m were observed for the MVG, BC, and MVG-HR models at the 2.6 m depth (ensemble 3). The importance of the individual soil hydraulic models is strongly dependent on the depth of forecasting as represented by ensemble 1, 2 and 3. The MVG-DP model attains the largest BMA weight at the 0.4 m ($w = 0.667$) and 2.6 m ($w = 0.438$) depths. The good performance of the non-equilibrium flow model close to the surface is caused by the presence of preferential flow paths as discussed previously. Interestingly, the dual-porosity model performs also best for the 2.6 m depth, where pressure head change gradients are much smaller as compared to those at the shallower depths. The hysteretic uniform flow model MVG-HR attains the largest BMA weight ($w = 0.528$) at the 1.0 m depth. This is a surprising result because one would expect the MVG-DP model to receive more weight at this depth. In fact, the performance of MVG-DP and MVG-HR is very similar with nearly identical RMSE values (Table 2). Strong correlation between the pressure head predictions of these two models again influence the selection of the BMA weights.

Note that the single depth BMA weights in ensemble 1, 2 and 3 are also significantly different than their corresponding counterparts of the combined depth solutions, which were previously

reported in Table 5.

Similar to the performance of the individual ensemble members, the single depth BMA models perform better than their corresponding combined depth models. The RMSE values are 0.102, 0.080, and 0.122 m for the 0.4, 1.0, and 2.6 m depths, respectively (Table 6). To further illustrate these results, consider Figures 6b-d which provides time series plots of the BMA mean (bold dashed lines), the observations (bold solid lines), and the ensemble 1 - 3 forecasts (thin lines) at the 0.4, 1.0, and 2.6 m depths during the evaluation period. Particularly at the 1.0 m depth, (Figure 6c), the average width of the 95% prediction uncertainty bounds (shaded area) is significantly smaller than the average widths of the combined forecasts (Figure 5c). Despite this smaller spread, the prediction uncertainty ranges generally encompass the observations. About 73% of all observations are covered by the 95% uncertainty bounds at the 0.4 m depth with the remainder of the measurements appearing only slightly outside the uncertainty bounds. At the 1.0 m depth, the coverage is relatively high at 89% (Table 6 and Figure 5c) with observations primarily falling outside the uncertainty ranges immediately following rainfall events at simulation days 467, and 476 - 477, respectively. Note that at the 1.0 m depth the upper bound of the 95% uncertainty interval derived with the BMA model is in close vicinity of the observed pressure heads for most times during the evaluation period. At the 2.6 m depth the observations fall consistently outside the uncertainty bounds between days 485 and 492 resulting in a coverage of about 74%.

The results presented here are significantly better than those obtained previously in Table 5 and Figure 5d using the BMA model for the combined depths. For instance, the uncertainty ranges have increased considerably for the first 40 days of the evaluation period (days 450 - 490), but have become more consistent with the observed tensiometric pressure head data (Figure 6d). After day 490, the spread of the BMA uncertainty bounds narrows because the tensiometric pressure head predictions of the individual soil hydraulic models become in closer agreement.

Use of a single BMA Variance for each Model in the Ensemble

So far we have used different BMA variances for the conditional pdf's of the individual soil hydraulic models. Here we illustrate the performance of the BMA model using a single variance for each of the individual ensemble members. To this end, we replace the last term $\sum_{i=1}^{k} w_i \sigma^2$

in Eq. (5) with a single variance σ^2. The results of the analysis are listed in Table 7.

In general, the results presented here are very similar to those obtained previously with the use of multiple different BMA variances for the individual models (cf. Tables 6 and 7). For ensemble 1, the MVG-DP and KM models again rank first and second in their respective BMA weights, and the RMSE, R^2, and C_e summary statistics of the BMA model are very similar to those presented previously in 6. The average width of the 95% uncertainty interval (0.22 m) and the coverage (69%) during the evaluation period were slightly smaller than their corresponding values when using multiple different BMA variances. Qualitatively similar results were obtained for ensemble 2. For ensemble 3, the MVG-DP and MVG-HR models rank first and second in importance with BMA weights of w =0.513 and w =0.327, respectively. However, the BMA weight of the MVG-HR model is significantly larger than its weight derived previously when using multiple different BMA variances (Table 6). Not surprising therefore, is that the RMSE values for both the calibration and evaluation period are slightly different than those obtained previously. The most significant difference is that the coverage of the prediction uncertainty bounds of ensemble 3 has significantly improved from about 74% in the case of multiple different BMA variances to approximately 90% for the analysis considered here.

The results presented in Table 7 are very similar to those presented in Table 6, suggesting that the optimized BMA results are fairly insensitive to the choice of a single or multiple different BMA variances of the individual soil hydraulic models in the ensemble. Similar results have been found in other BMA modeling studies.

Impact of Ensemble size on Performance of BMA Model

In this section, we discuss the single depth BMA model forecasts for the ensembles 1a - 3a and 1b - 3b. These ensembles represent sub-sets of the original ensembles 1-3 where the models with the best predictive characteristics (i.e. the models which received the largest BMA weights) have sequentially been removed. For example, the MVG-DP model received the largest BMA weight of the seven models in ensemble 1 (Table 6) and was therefore excluded in ensemble 1a. This ensemble therefore contains the predictions of the remaining 6 soil hydraulic models. Further, the KM model was the second best performing model in ensemble 1 (and hence the best model of ensemble 1a), and was therefore removed from ensemble 1a to create ensemble

1b that now consists of 5 different soil hydraulic models.

The information content of an ensemble should generally deteriorate with decreasing size of the ensemble, k. Hence, it seems logical to assume that ensembles 1 - 3 should contain more information than ensembles 1a - 3a, which in turn should be more informative than ensembles 1b - 3b. The smallest possible ensemble that still contains all the necessary information to make good predictions and reliable estimates of uncertainty is warranted. A small ensemble has important computational advantages, since it requires calibrating and running the smallest possible number of soil hydraulic models. Here we investigate the influence of ensemble size on the performance of the BMA model.

Table 8 lists summary statistics of the performance of the BMA model generated using the information contained in ensembles 1a - 3a, and 1b - 3b. The quality of the fit between observed and BMA predicted pressure heads generally decreases with decreasing size of the ensemble. The RMSE values increase from 0.067 to 0.075 and 0.086 m, when moving from ensemble 1 to 1a and 1b, respectively. This is what is to be expected. The largest increase in RMSE of about 0.025 m was observed at the 2.6 m depth (ensemble 3b). This deterioration in performance is still relatively small, considering that we have sequentially removed the best two ensemble members. The information content of the full ensemble is only slightly better than the information content of the reduced ensemble because of highly correlated predictions of the individual soil hydraulic models.

The average width of the 95% uncertainty intervals and associated coverage of the observations increase with decreasing ensemble size during the evaluation period. The largest increase in width was about 0.09 m. The increase in coverage compared to the original seven member ensemble is the largest at the 2.6 m depth (93% coverage for ensemble 3b vs. 74% for ensemble 3). This is accompanied by a relatively large average width of 0.46 m (ensemble 3b, Table 8). The largest coverage of the uncertainty bounds (95%) was observed for the BMA forecasts of ensemble 2b (1.0 m depth). Interestingly, this is associated with a relatively small width of the uncertainty ranges of approximately 0.21 m. These results show that the model forecasts are most closely centered around the observations for the 1.0 m depth. Larger prediction uncertainty ranges are observed at the soil surface and 2.6 m depths.

Importance of the Choice of the Calibration Period

In the analysis presented above we decided to calibrate the individual soil hydraulic models and the BMA model for the first wet season and to evaluate their performance during the second wet season. This decision was made because water flow in the highly porous volcanic vadose zone at the *Spydia* field site occurs primarily under wet conditions, and the soil hydraulic conductivity declines rapidly with decreasing soil water pressure heads (Wöhling et al. 2008). To investigate the effect of the choice of calibration period on the final accuracy and reliability of the soil hydraulic and BMA model forecasts, we conducted a second analysis using another calibration and evaluation period. We now included dry and wet conditions in the calibration data set using a 296 days period between December 17, 2006 and October 9, 2007 (296 days). The evaluation period was selected to span the period between May 1, 2006 and December 16, 2006 (230 days), using a 20 day spin-up period for state-value initialization. We re-calibrated the seven soil-hydraulic models and the BMA model using this new calibration period. Hereafter, we refer to this second calibration data set as Approach B and to our initial results as Approach A. The most important results of this analysis are summarized here.

The optimized parameter sets for Approach B resulted in a substantially lower quality of fit to the calibration data at the 0.4, 1.0 and 2.6 m depths as compared to Approach A. This poorer fit was caused by an inability of the individual models to accurately reproduce the tensiometric head data during the dry period - which was excluded in Approach A. Our data interpretation and analysis suggests that this rather large misfit is caused by water repellency, which is not included in any of the seven soil hydraulic models.

The combined pressure head forecasts using the four different model ensembles for the evaluation period also showed a poorer fit to the observed tensiometer data with RMSE values ranging between 0.190 and 0.417 m and R^2 values between 0.13 and 0.63. Moreover, the individual soil hydraulic models in ensemble 4 (consisting of the forecasts of the compromise solutions), did not perform better than the respective models in the ensembles 1, 2 and 3 as was the case in Approach A. In addition, the average width of the 95% uncertainty bounds was noticeably larger than obtained previously in Approach A (Table 5). Similar results were found for the single depth pressure head forecasts for the evaluation period.

We conclude that the choice of the calibration period has a strong influence on the results of the analysis. It determines the accuracy of the individual model forecasts on one hand, and the accuracy and uncertainty estimates of the BMA model on the other hand. For accurate forecasting, it is desirable to use a calibration period of the soil hydraulic and BMA model that spans the largest possible range of drying and wetting events. However, water repellency is not included in any of our model formulations, and so it is better not to include rainfall events during prolonged dry periods in the calibration of the individual soil hydraulic models. This avoids the parameters in these models to take unrealistic optimized values so as to compensate for this missing physical process.

4 Summary and Conclusions

Uncertainty estimation is currently receiving a surge in attention because researchers are trying to better understand what is well and what is not very well understood about the environmental systems that are being studied and as decision makers push to better quantify accuracy and precision of model predictions. In this paper, we have presented a combined multi-objective optimization and Bayesian Model Averaging (BMA) framework to calibrate forecast ensembles of soil hydraulic models. To illustrate our methodology, we used pressure head data from three different depths in a layered vadose zone of volcanic origin in New Zealand. A multi-objective formulation was used to calibrate the individual soil hydraulic models. The resulting Pareto solution space was estimated with the AMALGAM multi-method global optimization algorithm and used to generate different model ensembles. The most important conclusions of our study are:

1. The 4-layer uniform flow model MVG-4 provides the most accurate predictions of the combined pressure heads throughout the *Spydia* vadose zone. Its performance is superior to the dual-porosity MVG-DP model that explicitly accounts for the presence of preferential flow and provides the best fit to measured pressure heads at the 0.4m depth.

2. The mean pressure head forecast of the BMA model has similar predictive capabilities as the best performing soil hydraulic model in the ensemble. This is because the various hydraulic models have been calibrated well against the observed pressure head data.

3. The optimized values of the BMA weights do not necessarily follow the reverse RMSE order of the individual models. This is because of cross-correlations between predictions of the individual models in the ensemble. One should therefore be particularly careful in drawing conclusions about the usefulness of individual ensemble members based on their optimized BMA weight.

4. The best BMA model at each particular depth is made up of the ensemble of forecasts corresponding to the respective Pareto extremes. The best BMA model at one depth however, receives relatively poor performance in predicting tensiometric pressure heads at the other two depths.

5. The overall best ensemble and BMA model is obtained when selecting the compromise solution of the Pareto trade-off surface. This is a balanced solution that minimizes the overall RMSE of observed and simulated pressure heads at the three different measurement depths.

6. Removing the best two soil hydraulic models of the ensemble only slightly deteriorated the performance of the BMA model with a small increase in the spread of the 95% prediction uncertainty bounds. Significant correlation between the predictions of the individual soil hydraulic models in the ensemble causes a large amount of redundancy in information.

7. The selection of the calibration period greatly affects the final optimized BMA weights and variances. The results of the BMA model are fairly insensitive to the choice of a single or multiple different values for the variances of the conditional pdf's of the individual ensemble members.

8. The prediction uncertainty bounds of the BMA model generally increase with increasing depth and dryness in the soil profile.

9. The combined multi-objective optimization and BMA framework proposed in this paper is very useful to generate forecast ensembles of soil hydraulic models and appropriately quantify predictive uncertainty of flow through unsaturated porous media. Accurate uncertainty quantification is important for decision makers and end-users.

Acknowledgments

Thomas Wöhling likes to thank the New Zealand Foundation for Research. Science and Technology (FRST) for funding this work as part of LVL's Groundwater Quality Protection Programme (8137-ASXS-LVL). Jasper A. Vrugt is supported by a J. Robert Oppenheimer Fellowship from the LANL postdoctoral program. We would like to thank the constructive reviews by the Associate Editor, Jirka Simunek, and three reviewers including Ty Ferré and Ming Ye that have improved the current version of this paper.

References

Ajami, N. K., Duan, Q., and Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resour. Res.*, 43(1):W01403.

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). Crop evapotranspiration. FAO Irrigation and Drainage Paper No. 56, Rome (Italy).

ASCE (1993). Task committee on definition of watershed models of the watershed management committee, Irrigation and drainage division.: Criteria for evaluation of watershed models. *J. Irrig. Drain. Div.*, 119(3):429–442.

Beven, K. and Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, 6:279–298.

Brooks, H. and Corey, A. (1966). Properties of porous media affecting fluid flow. *Journal of the Irrigation and Drainage Division - Proceedings of the American Society of Civil Engineers (IR2)*, 92:61–88.

Celia, M. A., Bouloutas, E. T., and Zarba, R. L. (1990). A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resour. Res.*, 26(7):1483–1496.

Deb, K. (2001). Multi-objective optimization using evolutionary algorithms. John Wiley and Sons, Chicester, UK.

Durner, W. (1994). Hydraulic conductivity estimation for soils with heterogeneous pore structure. *Water Resour. Res.*, 30:211–223.

Feddes, R. A., Kowalik, P., and Zaradny, H. (1978). Simulation of field water use and crop yield. PUDOC, Wageningen, Netherlands. ISBN 90-220-0676-X.

Guber, A., Pachepski, Y., van Genuchten, M., Rawls, W., Šimůnek, J., Jacques, D., Nicholson, T. J., and Cady, R. A. (2006). Field-scale water flow simulations using ensembles of Pedotransfer functions for soil water retention. *Vadose Zone Journal*, 5:234–247.

Gupta, H. V., Sorooshian, S., and Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, 34(4):751–764. DOI:10.1029/97WR03495.

Hall, J. M. (2001). How well does your model fit the data? *Journal of Hydroinformatics*, 3(1):49–55.

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

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

Kosugi, K. (1996). Lognormal distribution model for unsaturated soil hydraulic properties. *Water Resour. Res.*, 32:2697–2703.

Kuczera, G., Kavetski, D., Franks, S., and Thyer, M. (2006). Towards a bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, 331(1-2):161–177.

Leamer, E. E. (1978). Specification searches: Ad hoc inference with non-experimental data. New York: Wiley.

Mertens, J., Madsen, H., Kristensen, M., Jaques, D., and Feyen, J. (2005). Sensitivity of soil parameters in unsaturated zone modelling, and the relation between effective, laboratory, and in-situ estimates. *Hydrological Processes*, 19(8):1611–1633.

Mualem, Y. (1976). A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, 12(3):513–522.

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

Raftery, A. E., Balabdaoui, F., Gneiting, T., and Polakowsk, M. (2003). Using bayesian model averaging to calibrate forecast ensembles. Technical Report 440, Department of Statistics, University of Washington, Seattle, Washington.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using bayesian model averaging to calibrate forecast ensembles. *American Meteorological Society*, 133:1155–1174.

Sansoulet, J., Cabidoche, Y.-M., Cattan, P., Ruy, S., and Simunek, J. (2008). Spatially distributed water fluxes in an andisol under banana plants: Experiments and three-dimensional modeling. *Vadose Zone J*, 7(2):819–829.

van Genuchten, M. Th. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, 44(5):892–898.

Vogel, T., van Genuchten, M. Th., and Cislerova, M. (2001). Effect of the shape of the soil hydraulic functions near saturation on variably-saturated flow predictions. *Advances in Water Resources*, 24(2):133–144.

Vrugt, J. A., Clark, M. P., Diks, C. G. H., Duan, Q., and Robinson, B. A. (2006a). Multi-objective calibration of forecast ensembles using bayesian model averaging. *Geophysical research letters*, 33:L19817.

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., and Sorooshian, S. (2003a). Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research*, 39(5):1–19. DOI: 10.1029/2002WR001746.

Vrugt, J. A., Gupta, H. V., Bouten, W., and Sorooshian, S. (2003b). A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, 39(8).

Vrugt, J. A., Ó Nualláin, B., Robinson, B. A., Bouten, W., Dekker, S. C., and Sloot, P. M. A. (2006b). Application of parallel computing to stochastic parameter estimation in environmental models. *Computers & Geosciences*, 32(8):1139 - 1155.

Vrugt, J. A. and Robinson, B. A. (2007a). Improved evolutionary optimization from genetically adaptive multimethod search. In *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, volume 104, pages 708–711.

Vrugt, J. A. and Robinson, B. A. (2007b). Treatment of uncertainty using ensemble methods:

Comparison of sequential data assimilation and bayesian model averaging. *Water Resources Research*, 43:W01411.

Šimůnek, J., Jarvisb, N. J., van Genuchten, M. Th., and Gaŕdenaš, A. (2003). Review and comparison of models for describing non-equilibrium and preferential flow and transport in the vadose zone. *Journal of Hydrology*, 272:14–35.

Šimůnek, J., van Genuchten, M. Th., and Šejna, M. (2005). *The HYDRUS-1D Software Package for Simulating the One-Dimensional Movement of Water, Heat, and Multiple Solutes in Variably-Saturated Media. Version 3.0.* Department of Environmental Sciences, University of California Riverside, Riverside, CA, 92521, USA.

Šimůnek, J., Šejna, M., and van Genuchten, M. Th. (1996). Hydrus-2d. simulating water flow and solute transport in two-dimensional variably saturated media. version 1.0. User manual, IGWMC-TPS 53 International Ground Water Modeling Center, Colorado School of Mines.

Wöhling, Th., Vrugt, J. A., and Barkle, G. F. (2008). Comparison of three multiobjective algorithms for inverse modeling of vadose zone hydraulic properties. *Soil Science Society of America Journal*, 72(2):305–319.

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

Ye, M., Neuman, S., and Meyer, P. (2004). Maximum likelihood Bayesian abveraging of spatial variablility models in unsaturated fractured tuff. *Water Resources Research*, 40:W05113.

Figure captions

Figure 1: Flowchart of the combined multi-objective optimization and Bayesian modeling averaging approach used in our study. A detailed explanation of the various boxes, and numbers appears in the text.

Figure 2: Pareto optimal solutions (solid circles) of the three-dimensional Pareto trade-off space for the MVG model; (a) the $F_1 - F_2$plane, (b) the $F_1 - F_3$ plane, (c) the $F_2 - F_3$ plane of the objective space. The single objective solutions ($\circ$ symbol, $P_1 - P_3$) and the compromise solution (+ symbol, P_4) are also indicated in each panel.

Figure 3: Pressure head predictions of the individual models for the calibration period using the compromise solution - parameter sets: a) daily rainfall, b) - d) the pressure head forecasts at the 0.4, 1.0, and 2.6 m depths, respectively.

Figure 4: BMA predictive probability density function (solid line) and the conditional pdf's (dashed, dashed-dotted, and dotted lines; the abbreviations of the individual models are given in the text) for the pressure head forecast at the 0.4 m depth on October 5, 2007, 5:00 pm. The 95% uncertainty bounds (shaded area), the individual model forecasts (dots) and the verifying observation (x) are also indicated.

Figure 5: Pressure head forecasts of the individual models of the BMA ensemble 4 for the evaluation period: a) daily rainfall and b)-d) the pressure head forecasts at the 0.4, 1.0, and 2.6 m depths, respectively. The observations (thick solid line), the BMA mean (thick dashed line), and the 95% prediction uncertainty bounds (shaded area) are also shown.

Figure 6: Pressure head forecasts of the BMA ensemble members for the evaluation period: a) daily rainfall and the pressure head forecasts b) at the 0.4 m depth using ensemble 1, c) at the 1.0 m depth using ensemble 2, and d) at the 2.6 m depth using ensemble 3. Also shown are the observations (thick solid line), the BMA mean (thick dashed line), and the 95% prediction uncertainty bounds (shaded area).

Table 1: Number of parameters, np, to be estimated for the seven soil hydraulic models used in this study

Model	Parameters	np
MVG, BC, KM	θ_s, n, α, l, K_s	15
MVG-HR	θ_s^d, n^d, α^d, l, K_s^d,θ_s^w, α^w, K_s^w	24
MVG-DP	θ_s,n_1, n_2, α_1, α_2, l, K_s, w_2	24
DPIM	$\theta_{s,mo}$, n, α, l, K_s, $\theta_{s,imo}$, ω	21
MVG-4	θ_s, n, α, l, K_s	20

Table 2: Measures of fit between the observed and simulated pressure head at the 0.4, 1.0, and 2.6 m depths (represented by the objectives $F_1 - F_3$), using single objective Pareto efficient parameter sets ($P_1 - P_3$) in the simulations with the various models in the study. Performance criteria are shown for the optimized objective and the values are calculated for the calibration period. The best RMSE values in the ensembles are indicated in bold fonts.

	RMSE [m]			R^2			C_e		
Model	F_1/P_1	F_2/P_2	F_3/P_3	F_1/P_1	F_2/P_2	F_3/P_3	F_1/P_1	F_2/P_2	F_3/P_3
MVG	0.095	0.079	0.106	0.91	0.93	0.71	0.90	0.93	0.70
BC	0.117	0.096	0.108	0.86	0.91	0.68	0.86	0.89	0.68
KM	0.096	0.065	0.099	0.92	0.95	0.75	0.92	0.95	0.73
MVG-HR	0.090	0.058	0.059	0.91	0.96	0.91	0.91	0.96	0.91
MVG-DP	**0.069**	0.059	**0.058**	0.95	0.96	0.91	0.95	0.96	0.91
DPIM	0.089	0.074	0.088	0.92	0.94	0.80	0.92	0.94	0.79
MVG-4	0.087	**0.052**	0.072	0.92	0.97	0.86	0.92	0.97	0.86

Table 3: Measures of fit between the observed and simulated pressure head for the 0.4, 1.0, and 2.6 m depths combined using the Pareto extremes ($P_1 - P_3$) and the compromise solution parameter sets (P_4) in the simulations with the individual models. The criteria are calculated for the calibration period. The best $RMSE_c$ values in the ensembles are indicated in bold fonts.

Pareto point	Model	$RMSE_c$ [m]	R_c^2	$C_{e,c}$	Bias [%]
P_1	MVG	0.222	0.54	0.33	-15.5
	BC	0.261	0.33	0.07	-2.1
	KM	0.307	0.39	-0.28	-1.3
	MVG-HR	**0.213**	0.53	0.38	-12.8
	MVG-DP	0.335	0.32	-0.52	4.1
	DPIM	0.409	0.25	-1.27	-36.7
	MVG-4	0.216	0.71	0.36	-20.0
P_2	MVG	0.235	0.68	0.25	15.6
	BC	0.835	0.16	-8.48	63.6
	KM	0.349	0.51	-0.66	-11.9
	MVG-HR	0.183	0.73	0.54	9.7
	MVG-DP	**0.135**	0.80	0.75	4.0
	DPIM	0.274	0.56	-0.02	-8.4
	MVG-4	0.405	0.26	-1.24	17.8
P_3	MVG	0.428	0.07	-1.50	-35.4
	BC	0.338	0.23	-0.56	-27.5
	KM	0.383	0.27	-1.00	-32.0
	MVG-HR	**0.306**	0.38	-0.27	-21.9
	MVG-DP	0.362	0.15	-0.79	-27.9
	DPIM	0.365	0.12	-0.82	-27.7
	MVG-4	0.352	0.33	-0.69	-29.5
P_4	MVG	0.147	0.74	0.71	-3.0
	BC	0.149	0.70	0.70	0.6
	KM	0.153	0.74	0.68	-3.4
	MVG-HR	0.112	0.84	0.83	-0.4
	MVG-DP	0.120	0.81	0.80	-0.5
	DPIM	0.144	0.73	0.72	-3.1
	MVG-4	**0.101**	0.86	0.86	1.0

Table 4: Ensemble forecasts of pressure head, bias corrected forecasts, BMA weights and variances, and verifying observation for the 0.4 m depth of the *Spydia* vadose zone at time $t =$ 542.3 days (October 5, 2007, 5:00 pm).

	MVG	BC	KM	MVG-HR	MVG-DP	DPIM	MVG-4
Forecast [m]	-0.440	-0.480	-0.360	-0.440	-0.300	-0.440	-0.430
Bias corrected forecast [m]	-0.425	-0.441	-0.390	-0.435	-0.299	-0.439	-0.433
BMA weight	0.001	0.004	0.179	0.055	0.667	0.056	0.038
BMA variance	0.377	0.007	0.062	0.072	0.036	0.132	0.010
Observation [m]				-0.342			

Table 5: Summary statistics of the combined depth pressure head forecasts using the individual models and the BMA predictive model in ensembles 1-4 for the evaluation period. The statistics are also listed for the BMA model during the calibration period, BMA (cali). The best $RMSE_c$ values in the ensembles are indicated in bold fonts.

	Model	Evaluation period				BMA		95% interval:	
Ensemble		$RMSE_c$ [m]	R_c^2	$C_{e,c}$	Bias [%]	w_i	σ_i	Coverage [%]	Width [m]
1	MVG	0.269	0.39	0.28	-14.1	0.000	0.31		
	BC	0.286	0.25	0.18	-0.8	0.000	0.28		
	KM	0.317	0.70	-0.01	10.3	0.000	0.21		
	MVG-HR	0.297	0.24	0.11	-15.3	0.246	0.07		
	MVG-DP	0.349	0.19	-0.22	1.0	0.000	0.29		
	DPIM	0.497	0.01	-1.47	-39.2	0.000	0.23		
	MVG-4	0.232	0.66	0.46	-17.1	0.753	0.15		
	BMA	**0.216**	0.60	0.53	3.8	-	-	86.0	0.55
	BMA (cali)	0.143	0.73	0.72	-	-	-	95.1	0.55
2	MVG	0.237	0.55	0.44	14.0	0.000	0.06		
	BC	0.861	0.38	-6.42	77.0	0.000	0.23		
	KM	0.332	0.30	-0.10	-14.9	0.000	0.43		
	MVG-HR	0.239	0.46	0.43	7.5	0.170	0.17		
	MVG-DP	0.222	0.51	0.51	2.1	0.802	0.07		
	DPIM	0.252	0.48	0.37	-7.2	0.000	0.43		
	MVG-4	0.382	0.55	-0.46	26.0	0.028	0.18		
	BMA	**0.219**	0.54	0.52	0.04	-	-	84.1	0.47
	BMA (cali)	0.118	0.81	0.81	-	-	-	96.2	0.47
3	MVG	0.356	0.33	-0.27	-32.5	0.000	0.43		
	BC	0.319	0.41	-0.02	-28.4	0.067	0.35		
	KM	0.314	0.59	0.01	-23.1	0.000	0.33		
	MVG-HR	0.266	0.56	0.29	-18.2	0.464	0.12		
	MVG-DP	0.275	0.51	0.24	-20.1	0.014	0.01		
	DPIM	0.271	0.57	0.27	-14.5	0.000	0.07		
	MVG-4	0.291	0.58	0.15	-23.9	0.455	0.18		
	BMA	**0.207**	0.72	0.57	8.07	-	-	93.8	0.76
	BMA (cali)	0.201	0.49	0.45	-	-	-	95.1	0.75
4	MVG	0.125	0.88	0.84	1.3	0.010	0.01		
	BC	0.276	0.24	0.24	0.8	0.006	0.20		
	KM	0.143	0.89	0.79	7.1	0.000	0.25		
	MVG-HR	0.182	0.74	0.67	0.4	0.237	0.14		
	MVG-DP	0.228	0.48	0.48	-1.0	0.097	0.02		
	DPIM	0.193	0.64	0.63	-0.4	0.138	0.03		
	MVG-4	0.137	0.87	0.81	5.0	0.512	0.05		
	BMA	**0.154**	0.83	0.76	2.5	-	-	82.9	0.40
	BMA (cali)	0.099	0.87	0.87	-	-	-	95.2	0.39

Table 6: Summary statistics of the single depth pressure head forecasts using the individual models and the BMA predictive model in ensembles 1-3 for the evaluation period. The statistics are also listed for the BMA model during the calibration period, BMA (cali). The best RMSE values in the ensembles are indicated in bold fonts.

	Model	Evaluation period				BMA		95% interval:	
Ense-mble		RMSE [m]	R^2	C_e	Bias [%]	w_i	σ_i	Coverage [%]	Width [m]
1	MVG	0.108	0.95	0.78	13.9	0.000	0.377		
	BC	0.133	0.94	0.67	17.8	0.004	0.007		
	KM	0.112	0.84	0.77	9.9	0.179	0.062		
	MVG-HR	0.111	0.93	0.77	13.9	0.055	0.072		
	MVG-DP	**0.097**	0.91	0.82	11.5	0.667	0.036		
	DPIM	0.098	0.94	0.82	11.7	0.056	0.132		
	MVG-4	0.098	0.91	0.82	10.9	0.038	0.010		
	BMA	0.102	0.91	0.80	12.6	-	-	73.2	0.24
	BMA (cali)	0.067	0.95	0.95	-	-	-	94.2	0.23
2	MVG	0.083	0.93	0.88	8.1	0.000	0.107		
	BC	0.115	0.92	0.76	13.2	0.000	0.688		
	KM	0.070	0.95	0.91	7.0	0.000	0.113		
	MVG-HR	0.090	0.91	0.86	8.3	0.528	0.027		
	MVG-DP	0.067	0.97	0.92	8.3	0.025	0.131		
	DPIM	**0.065**	0.95	0.93	5.5	0.056	0.005		
	MVG-4	0.080	0.97	0.89	10.3	0.391	0.048		
	BMA	0.080	0.95	0.89	8.9	-	-	88.6	0.14
	BMA (cali)	0.049	0.97	0.97	-	-	-	96.3	0.18
3	MVG	0.313	0.54	0.19	-19.1	0.102	0.010		
	BC	0.373	0.14	-0.15	-19.4	0.000	0.057		
	KM	0.136	0.90	0.85	5.3	0.000	0.142		
	MVG-HR	0.217	0.67	0.61	-6.2	0.190	0.092		
	MVG-DP	0.121	0.91	0.88	-5.1	0.438	0.017		
	DPIM	**0.118**	0.97	0.88	5.0	0.190	0.020		
	MVG-4	0.141	0.87	0.84	-2.4	0.080	0.023		
	BMA	0.122	0.89	0.88	-3.9	-	-	74.5	0.40
	BMA (cali)	0.055	0.92	0.92	-	-	-	96.1	0.21

Table 7: Summary statistics for the single depth BMA predictive models using a single BMA variance for the hydraulic models in ensemble 1, 2 and 3. The words 'cali.' and 'eval.' are abbreviations for the calibration and evaluation period. Coverage and average width are given for the 95% uncertainty bounds.

BMA model		Ensemble Number		
		1	2	3
weight, w	MVG	0.000	0.000	0.001
	BC	0.002	0.000	0.000
	KM	0.130	0.001	0.000
	MVG-HR	0.039	0.581	0.327
	MVG-DP	0.786	0.081	0.513
	DPIM	0.036	0.000	0.156
	MVG-4	0.007	0.337	0.003
Variance, σ^2		0.051	0.041	0.046
RMSE, cali. [m]		0.067	0.049	0.055
RMSE, eval. [m]		0.102	0.080	0.132
Coverage, cali. [%]		94.0	95.6	95.8
Coverage, eval. [%]		69.1	85.7	90.2
Average width, cali. [m]		0.23	0.18	0.21
Average width, eval. [m]		0.22	0.18	0.32
Bias, eval. [%]		12.5	8.9	12.6

Table 8: Summary statistics for the single depth BMA predictive models excluding best-fitting model forecast in the ensembles 1a - 3a and 1b - 3b. The abbreviation for the excluded models are: b =KM, d =MVG-HR, e =MVG-DP, f =DPIM, and g =MVG-4. Further, 'cali.' and 'eval.' are abbreviations for the calibration and evaluation period, respectively.

BMA model	Ensemble Number					
	1a	1b	2a	2b	3a	3b
Excluded model	e	b, e	d	d, g	e	d, e
Model number, k	6	5	6	5	6	5
RMSE, cali. [m]	0.075	0.086	0.053	0.058	0.062	0.084
RMSE, eval. [m]	0.109	0.101	0.074	0.068	0.123	0.099
Coverage, cali. [%]	0.94	0.94	0.95	0.94	0.95	0.95
Coverage, eval. [%]	0.80	0.87	0.92	0.95	0.92	0.93
Average width, cali. [m]	0.30	0.34	0.21	0.22	0.27	0.32
Average width, eval. [m]	0.27	0.33	0.20	0.21	0.43	0.46
Bias, eval. [%]	13.7	12.6	9.3	8.0	-1.5	-2.9

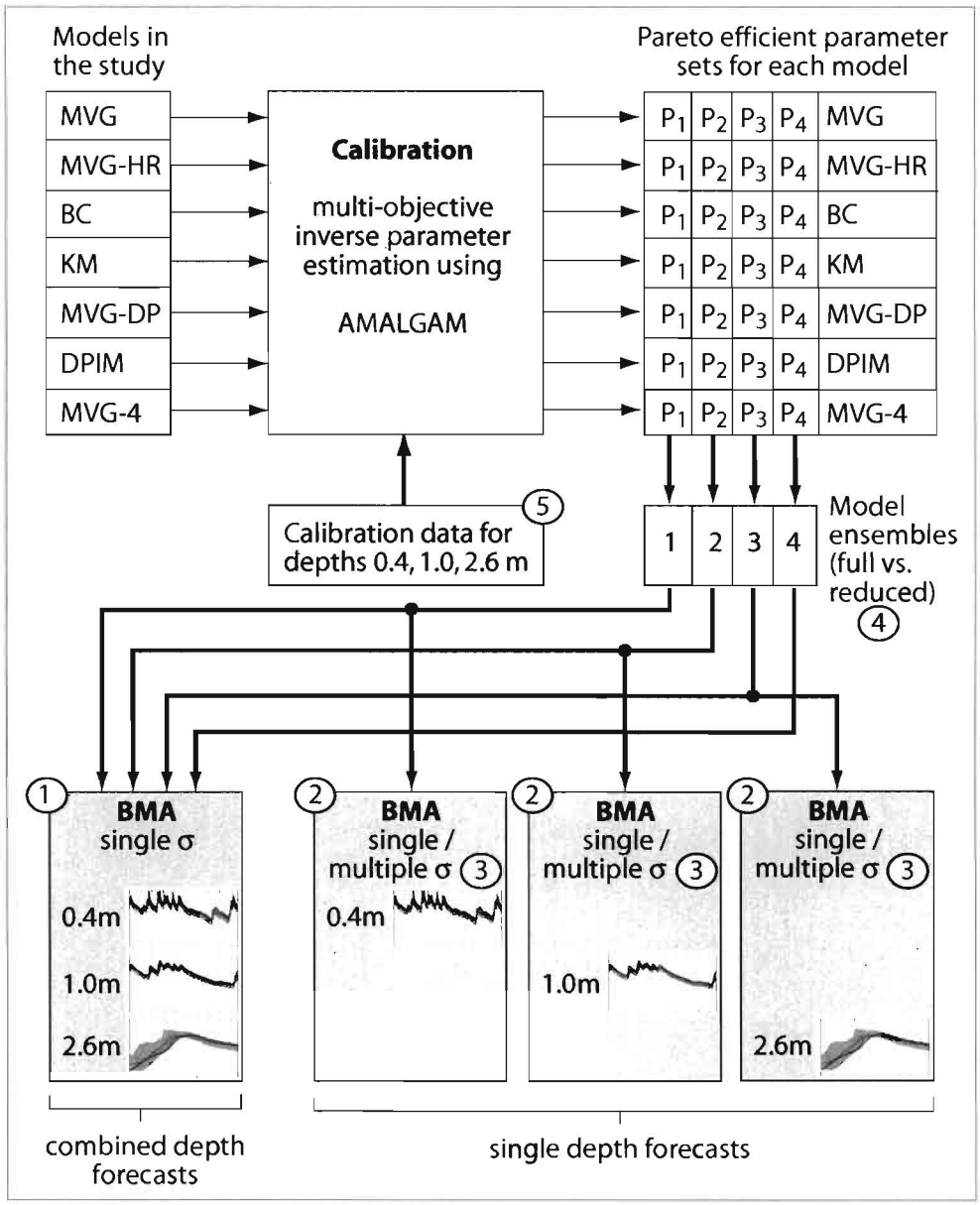

Figure 1: Flowchart of the combined multi-objective optimization and Bayesian modeling averaging approach used in our study. A detailed explanation of the various boxes, and numbers appears in the text.

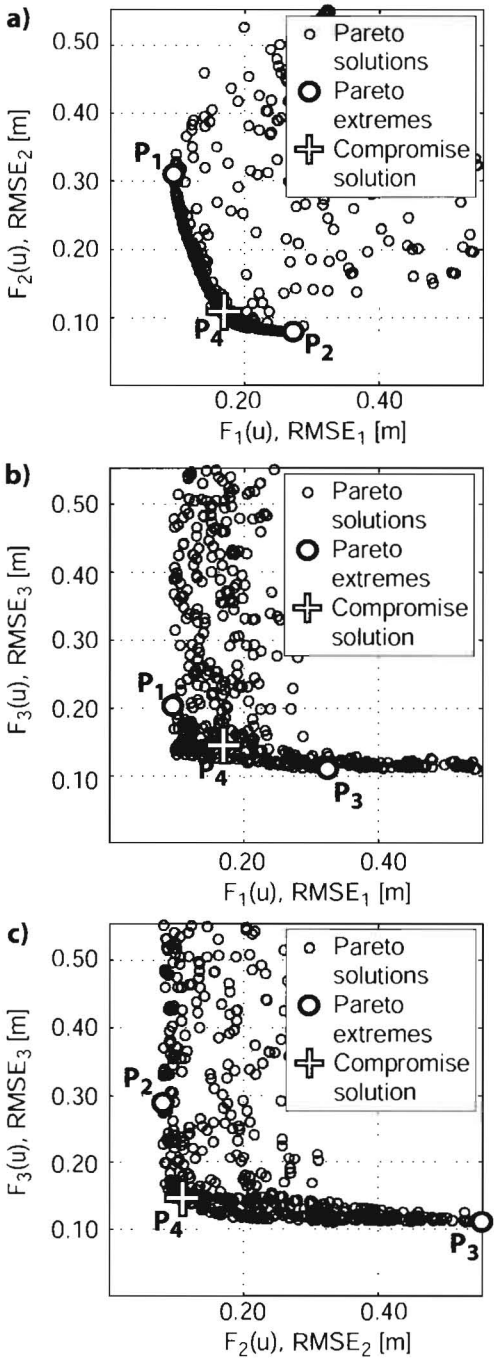

Figure 2: Pareto optimal solutions (solid circles) of the three-dimensional Pareto trade-off space for the MVG model; (a) the $F_1 - F_2$plane, (b) the $F_1 - F_3$ plane, (c) the $F_2 - F_3$ plane of the objective space. The single objective solutions ($\circ$ symbol, $P_1 - P_3$) and the compromise solution ($+$ symbol, P_4) are also indicated in each panel.

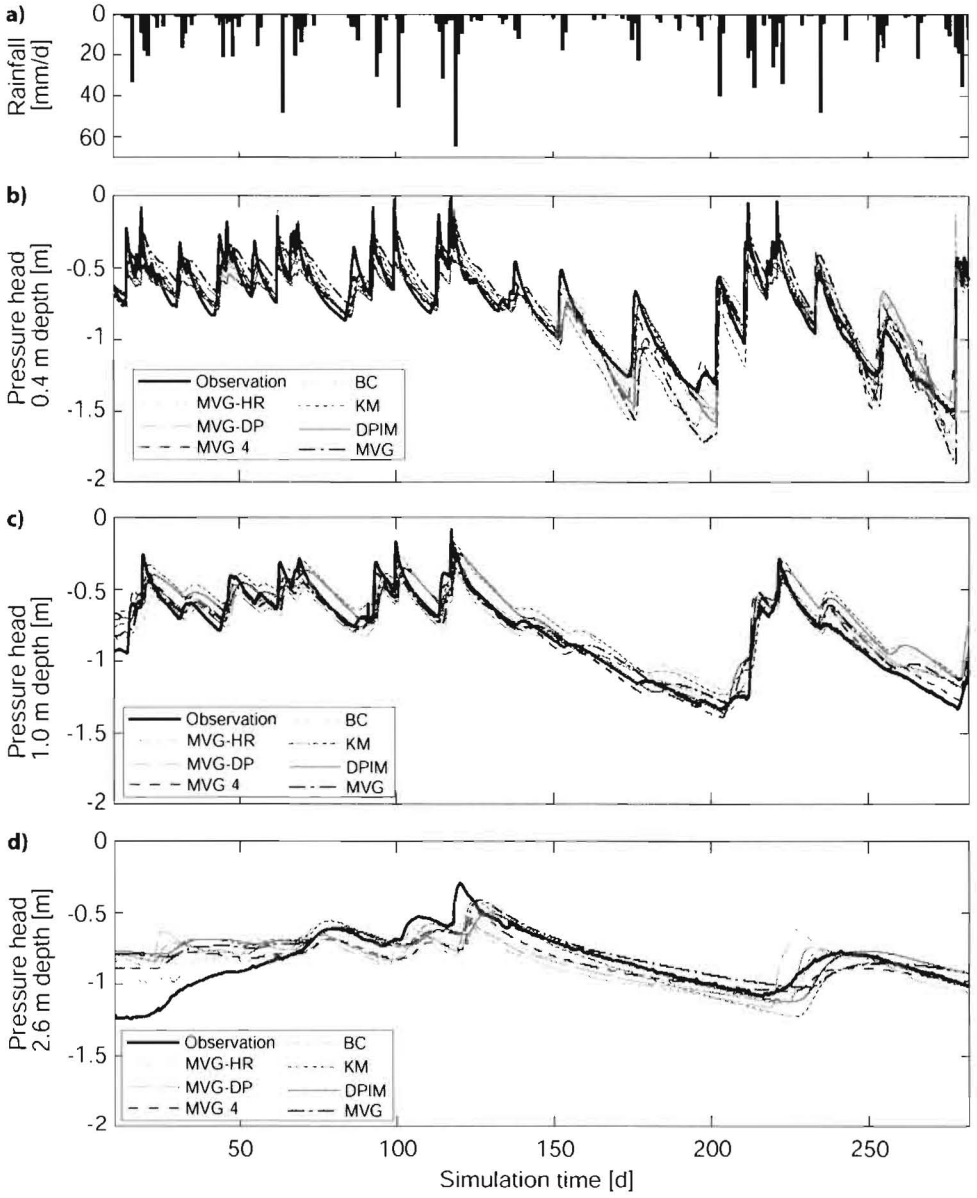

Figure 3: Pressure head predictions of the individual models for the calibration period using the compromise solution - parameter sets: a) daily rainfall, b) - d) the pressure head forecasts at the 0.4, 1.0, and 2.6 m depths, respectively.

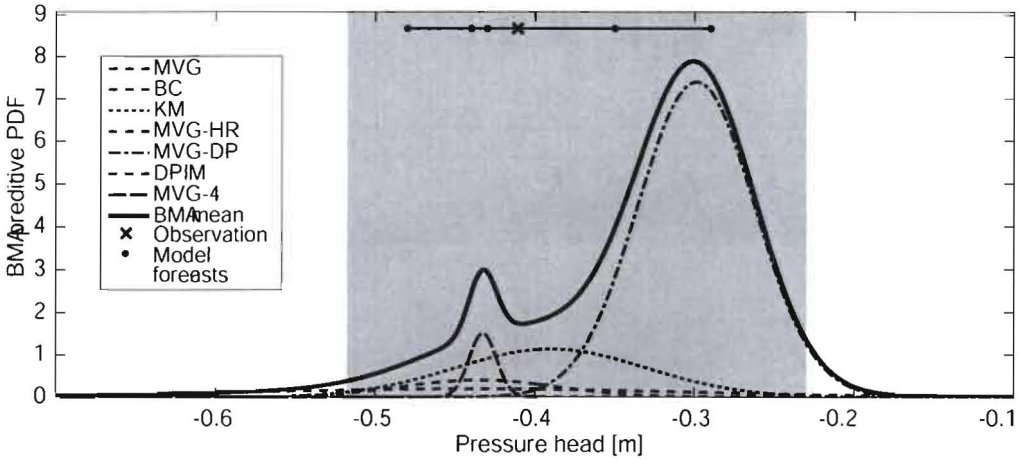

Figure 4: BMA predictive probability density function (solid line) and the conditional pdf's (dashed, dashed-dotted, and dotted lines; the abbreviations of the individual models are given in the text) for the pressure head forecast at the 0.4 m depth on October 5, 2007, 5:00 pm. The 95% uncertainty bounds (shaded area), the individual model forecasts (dots) and the verifying observation (x) are also indicated.

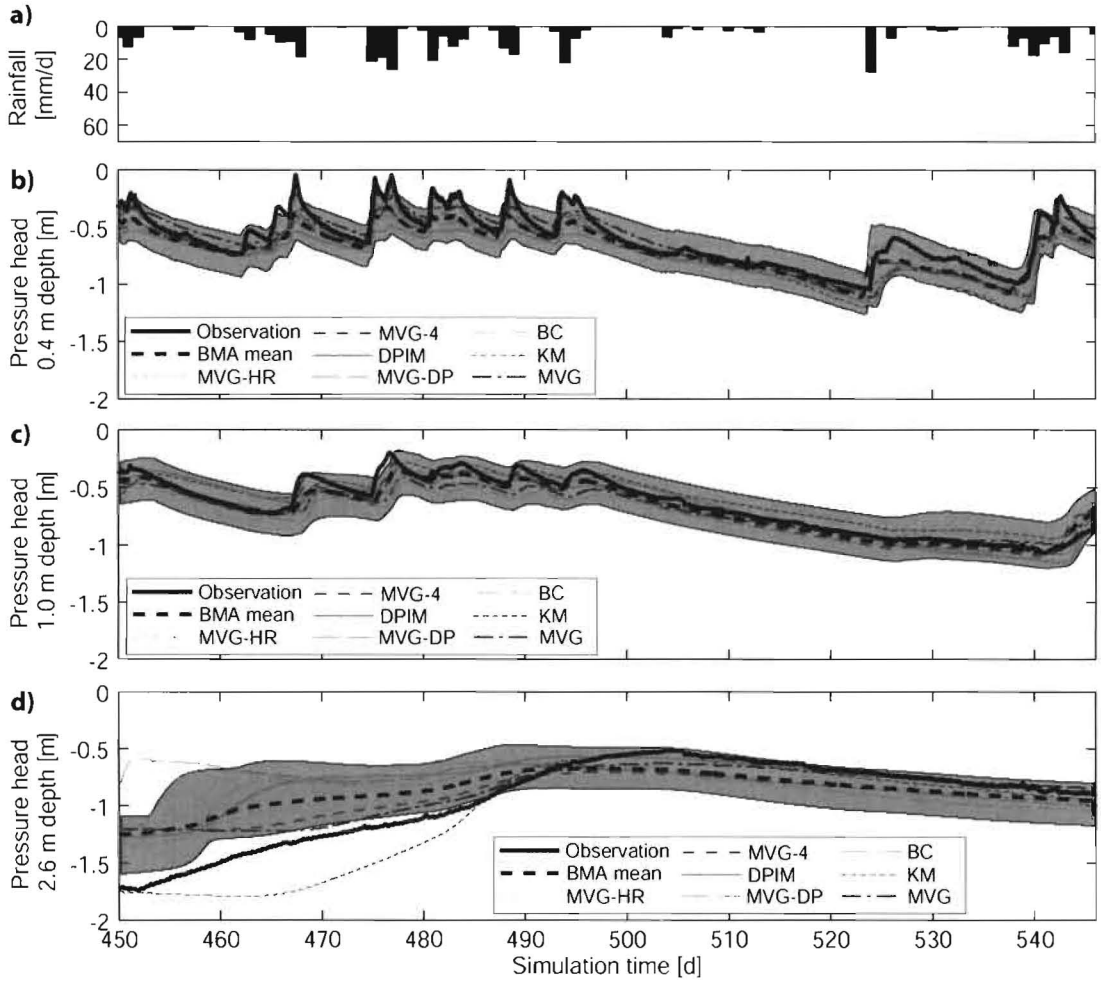

Figure 5: Pressure head forecasts of the individual models of the BMA ensemble 4 for the evaluation period: a) daily rainfall and b)-d) the pressure head forecasts at the 0.4, 1.0, and 2.6 m depths, respectively. The observations (thick solid line), the BMA mean (thick dashed line), and the 95% prediction uncertainty bounds (shaded area) are also shown.

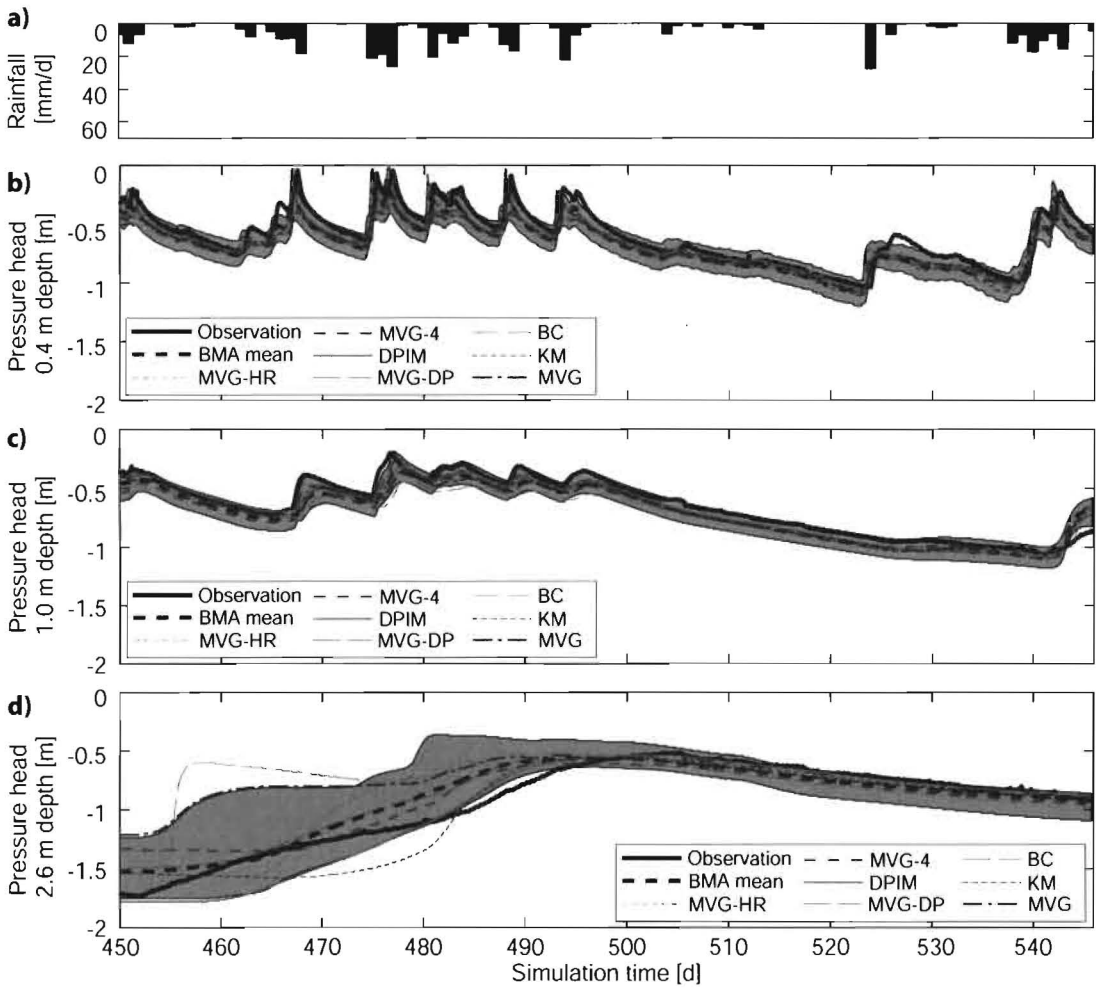

Figure 6: Pressure head forecasts of the BMA ensemble members for the evaluation period: a) daily rainfall and the pressure head forecasts b) at the 0.4 m depth using ensemble 1, c) at the 1.0 m depth using ensemble 2, and d) at the 2.6 m depth using ensemble 3. Also shown are the observations (thick solid line), the BMA mean (thick dashed line), and the 95% prediction uncertainty bounds (shaded area).